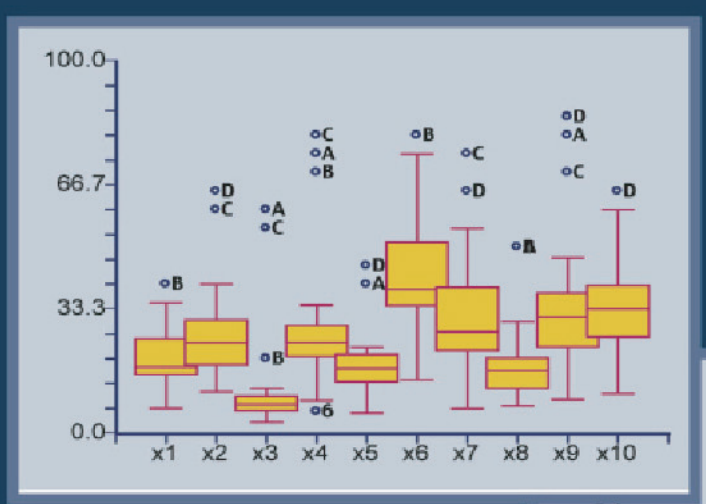
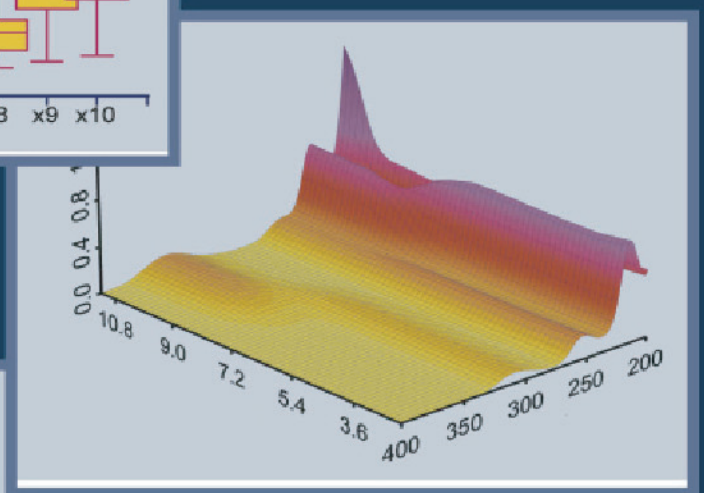
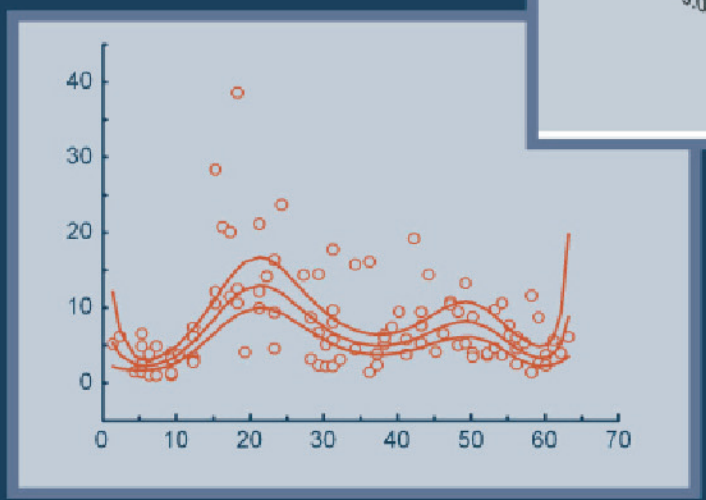
MILAN MELOUN • JIŘÍ MILITKÝ • ACADEMIA

KOMPENDIUM STATISTICKÉHO ZPRACOVÁNÍ DAT

METODY A ŘEŠENÉ ÚLOHY VČETNĚ CD







Kompendium statistického zpracování dat

Metody a řešené úlohy včetně CD

MILAN MELOUN, JIŘÍ MILITKÝ

Kompendium statistického zpracování dat

Metody a řešené úlohy včetně CD

Academia 2002

Lektorovali:

Ing. Jan Balcárek, Ph.D.
Ing. Adolf Goebel, Ph.D.
Prof. Ing. Oldřich Pytela, DrSc.

Obsah

Předmluva autorů

Aplikace statistických metod, využívajících počítače v klasických i nových přírodovědeckých, lékařských, technických a gnozeologických oborech patří mezi perspektivní směr na pomezí vědních disciplín, matematické statistiky a informatiky. Vedl ke vzniku nových oborů, jako jsou chemometrie, biometrie, psychometrie, ekonometrie, technometrie a další. Umožňuje nejen extrakci informací z experimentálních dat ale také tvorbu modelů a zobecnění získané kombinací výsledků z různých zdrojů.

Zaměření knihy

Statistická analýza dat nabývá na stále větším významu a stává se jedním ze základních přístupů v řadě přírodovědných, lékařských, technických a sociálních věd. Osvojení si metod statistické analýzy na příkladech samostatného diagnostikování a odkrývání tajů dat z praxe se jeví jako účinný způsob studia. Jedině řešením nových a nevšedních praktických úloh se můžeme dopracovat mistrovství v tomto oboru. Praktické úlohy analýzy dat obsahují totiž řadu nečekaných překážek:

~ rozsahy zpracovávaných dat nejsou obyčejně velké,
~ v datech existují výrazné nelinearity, multikolinearita, neaditivity a skryté vazby, anomálie a vlivné body, které komplikují analýzu dat,
~ rozdělení experimentálních dat jen zřídka odpovídá Gaussovu,
~ v datech se vyskytují vybočující měření a heterogenita,
~ parametry hledaných modelů mají definovaný fyzikální význam a jejich odhady musí proto vyhovovat svou velikostí a znaménkem.

Je žádoucí zkoumat předem statistické zvláštnosti dat průzkumovou čili exploratorní analýzou a ověřovat základní předpoklady o datech, hodnotit kvalitu výsledků s ohledem na základní schéma "data-model-statistická metoda".

Okruh čtenářů

Předložené kompendium představuje sbírku metod s úlohami pro počítačem podporované zpracování experimentálních dat. Kniha je určena především studentům vysokých škol technického, přírodovědného, ale i humanitního a ekonomického směru. Poslouží i k samostudiu. Představuje podrobnou "kuchařku", dle které si každý pohodlně vyhodnotí svá data. V knize nejsou definice, věty a důkazy matematického pojetí procvičované látky. Kniha byla sepsána tak, aby byla v praxi srozumitelná i středoškolsky vzdělaným pracovníkům. Cílem je dát návod k úspěšnému a statisticky rozumnému vyhodnocení dat

v laboratorní praxi. Spolu se statistickým softwarem (ADSTAT, QC-Expert, NCSS2000, STATGRAPHICS, STATISTICA, MINITAB, SCAN, S-Plus atd.) tvoří účinnou pomůcku analýzy experimentálních dat v laboratořích výzkumných ústavů, podnikových a státních kontrolních laboratořích a zkušebnách ke kontrole kvality. Kompendium je vhodnou pomůckou pro manažery zdravotnických, veterinárních a vodohospodářských laboratoří, potravinářské a zemědělské inspekce, chemické a farmaceutické výroby. Uplatnění najde také u pracovníků kontroly životního prostředí všech odvětví průmyslu, energetiky a zemědělství, u technologů, pracovníků řízení jakosti, a především u vedoucích pracovníků. S databází na přiloženém CD je kompendium bezprostředně využitelné v chemické laboratoři a chemometrii, v ekonomice a ekonometrii, sociologii, medicíně, biologii a při sledování kvality životního prostředí. Je určeno lidem, které láká počítačem podporované odkrývání informací, uložených v experimentálních datech.

Struktura kapitol

Úlohy jsou částnečně převzaty z literatury ale především z laboratorní a průmyslové praxe účastníků postgraduálního studia chemometrie z posledních deseti let. Úlohy se vesměs týkají netradičního pojetí statistické analýzy dat - *interaktivní analýzy*.

Kniha je organizována podle jednotného schématu: v první části každé kapitoly je stručně pojednáno o základech a technikách analýzy dat. Tam, kde jde o klasické statistické postupy, resp. známé metody, jde pouze o stručný úvod do řešené problematiky. U méně známých postupů a metod je výklad poněkud podrobnější. Vždy však postačuje k pochopení výstupů z příkladů. Následuje řešení vzorových úloh včetně postupu a numerických výsledků a jejich interpretace. Poslední částí každé kapitoly jsou praktické příklady z různých oborů s uvedením vybraných výsledků, indikujících správnost z hlediska především numerického. Jednotlivé kapitoly lze studovat samostatně, i když členění kapitol je zvoleno tak, aby se čtenář postupně seznamoval s navazujícími metodami a technikami.

Je třeba zdůraznit, že kompendium obsahuje především kolekci úloh a problémů každodenní laboratorní praxe, odlišných od školních modelových úloh, s cílem naučit čtenáře vyšetřovat data i v případech, kdy výpočty nevedou k jednoznačným nebo snadným závěrům. V každé kapitole jsou úlohy rozděleny dle svého obsahového zaměření do pěti skupin: úlohy s kódem **B** obsahují *biochemická a farmakologická data,* **C** pak *chemická a fyzikální data,* **E** *environmentální, potravinářská a zemědělská data,* **H** *hutní, metalurgická a mineralogická data*, **S** *sociologická, ekonomická a ostatní data.* Pouze v 6. kapitole (lineární regresní modely) je třídění úloh jiné: úlohy s kódem **J** obsahují *data jednoduchých lineárních regresních modelů*, **V** značí *data k validaci nové analytické metody,* **L** *data k polynomům*, **K** *data pro kalibraci,* **M** *data pro vícerozměrné lineární regresní modely*.

Postup analýzy dat

Jádrem výkladu každé metody je rozbor problému na charakteristické vzorové úloze. Po stručném popisu metody včetně potřebných vzorců následuje vždy vzorová úloha, která obvykle obsahuje: 1. *Zadání:* formulaci vlastního problému s jasným úkolem úlohy. 2. *Data:* zadání vstupních dat. 3. *Program:* volbu vhodného počítačového programu. 4. *Řešení:* detailní výklad postupu metody a vysvětlení řešení. 5. *Závěr:* formulaci dosaženého závěru úlohy. Soubory potřebných dat jsou ke všem úlohám na doprovodném

kompaktním disku, zatímco v textu knihy jsou z úsporných důvodů data pouze ve zkrácené podobě. Pokud by čtenáři nestačil stručný popis užité metody, najde detailní výklad matematického pozadí všech metod v učebnici M. Meloun, J. Militký: STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT, Plus, Praha 1994, a nebo v dalším vydání East Publishing, Praha 1998. Především náročné regresní úlohy (6. a 8. kapitola) byly v tomto kompendiu pro svou rozsáhlost uvedeny velmi stručně.

Pro řešení příkladů lze použít řady různých programových balíků i programů. Řada z nich však nebude obsahovat ani všechny probírané metody a dále ani testování základního tripletu "data - model - statistická metoda". Námi zaváděný postup komplexnější statistické analýzy dat je standardně použit například ve statistickém balíku ADSTAT. Ve skutečnosti je statistická analýza dat iterativní proces, kdy výsledkem bývá často návrh dalšího zkoumání a případně i návrh nových experimentů. Tento proces však nebylo možné v knize v plném rozsahu ukázat, takže snahou autorů bylo uvést alespoň výsledky jednoho kroku statistické analýzy, které jsou typické nebo nějak zajímavé a nevyžadují ani externí informace, ani detailní znalost oboru. V nejednoznačných případech bylo snahou použít jednoduché a názorné řešení plynoucí z daných dat. Tento dnes zaváděný *datově orientovaný postup* má pochopitelně nevýhodu v tom, že informace v datech mohou být nejen neúplné, ale také nepostačující pro statistické závěry s rozumnou mírou nejistoty. U řady převzatých příkladů nebylo také možné nijak ověřit kvalitu měřicího procesu ani metrologickou správnost měření. Některé závěry proto neplatí obecně a souvisejí pouze s danými daty. Na druhé straně jsou však takové příklady dobře použitelné z hlediska metodiky zpracování dat.

Zadávání dat

V laboratoři získáváme data přímo z experimentu ve formě databáze, spreadsheetu nebo ASCII souboru. Z dat se pak snažíme získat maximum užitečné informace a odhalit všechna ukrytá tajemství. Na doprovodném kompaktním disku jsou data k jednotlivým kapitolám jednak v ASCII souborech a jednak v speadsheetu pro program MS Excel. Většina statistického software je schopna načíst data přímo či importovat data v excelovské podobě nebo data "přeneseme" z MS Excelu do libovolného softwaru přes schránku (clipboard). Úlohy jsou v knize číslovány: např. ***C201a*** znamená první úlohu druhé kapitoly, oddíl **C** *chemická a fyzikální data*, část dat *a*. Takto deklarovaný výběr dat představuje vektor čili sloupec v tabulce MS Excelu, uvedený záhlavím ***C201a***, a prvky tohoto vektoru jsou jednotlivé řádky.

(1) Analýza jednorozměrného výběru značí analýzu jednoho vektoru: zadáme sloupce čísel ***C201a***.

(2) Analýza matice dat ve vícerozměrné analýze dat nebo v regresi: obdélníková matice obsahuje stejně dlouhé sloupce prvků, představujících *proměnné,* zatímco řádky v této matici představují *objekty*. Data zadáváme po sloupcích, např. sloupec ***B402i*** obsahuje název objektu, ***B402x1*** první proměnnou, ***B402x2*** druhou proměnou, ***B402x3*** třetí proměnnou atd.

(3) Matice u vícefaktorové analýzy rozptylu, ANOVA: prvky matice u dvoufaktorové ANOVA (či 3D-krabice u trojfaktorové) je třeba popsat jejich faktory. Sloupce ***B526f1, B526f2, B526y*** představují faktorové souřadnice prvků *y*.

Diagnóza analýzou dat

Diagnóza interaktivní analýzou dat znamená nastavení zrcadla tajům, ukrytým v datech, která zůstávala v minulosti standardními přístupy bez počítače nedostupná, a proto vědomě zanedbávána. Jsou jimi především

(1) Skryté problémové hodnoty, jako např. hrubé chyby, systematické chyby, odlehlé hodnoty nebo extrémy, kdy musíme rozhodnout, zda problémové hodnoty z analýzy odstraníme, opravíme, či je ponecháme beze změny a zvolíme k další statistické analýze metody méně citlivé, robustní.
(2) Nezávislost dat znamená, že prvky analyzovaného výběru nejsou spojeny žádným skrytým vztahem a byly získány nezávisle na sobě. To pochopitelně připouští pouze náhodné ovlivnění člověkem, přístrojem nebo postupem odběru dat.
(3) Výběr obsahuje chybějící data, tabulka má "díry". Data je pak třeba upravit tak, aby přesto poskytla spolehlivé výsledky.
(4) Průzkum v datech provádíme grafickými diagnostikami s cílem odhalit v nich symetrii rozdělení, typ rozdělení, lokální koncentraci dat, homogenitu dat, anomálie a velikost šumu.
(5) Výběr obsahuje málo dat. Když v laboratoři získáme málo dat, je třeba v první řadě doporučit další měření. Existují také metody (např. Hornův postup), které za cenu ztráty informací poskytnou řešení i pro tyto případy.
(6) Efektivní analýza dat předpokládá i zvláštní hodnoty, které jsou v datech velmi vlivné. Vlivné body totiž významně ovlivňují hledané parametry. Vlivným bodům je třeba věnovat zvláštní péči již při sběru dat.
(7) Oddělení vlivu jednotlivých proměnných je základním problémem regrese a vícerozměrných metod.

Kontrolní hodnoty u úloh

Kontrolní hodnoty řešení úloh jsou z pedagogických důvodů uvedeny u většiny úloh kromě těch, kde je řešení buď vícestupňové, nebo kde se na základě získaných informací provádí výběr metody či následuje úprava dat nebo modelu. Kontrolní hodnoty nejsou uváděny tam, kde jsou výstupy značně obsáhlé nebo grafické (např. vícerozměrná data, interpolace), resp. závislé na softwaru (např. nelineární regresi), nebo kde jsou výsledky tvořivého charakteru (např. metody vícerozměrné analýzy dat, interpolace a aproximace). Uživatel by je měl chápat jako kontrolní mezivýsledky a měl by se ve svém řešení dopracovat těchto, resp. podobných odhadů.

Doprovodný kompaktní disk (CD)

K usnadnění procvičování úloh je ke sbírce přiložen kompaktní disk, který obsahuje data ke všem úlohám. Vstupní data jsou jednak v ASCII kódu a jednak v matici pro MS Excel. Z něho se pak dají importovat data do většiny programů. Data jsou dostupná i na Internetu na osobní stránce autorů

//meloun.upce.cz/chemometrics/sbírka.

Poděkování

Není možné poděkovat všem spolupracovníkům, studentům licenčního studia chemometrie a doktorandům, kteří nám pomáhali či přispěli praktickými úlohami, radami či konstruktivní kritikou. Zvláštní dík patří lektorům za celkové přehlédnutí rukopisu a řadu podnětných praktických připomínek a výpočtů. Výroba této monografie byla umožněna finační podporou vědeckého záměru Ministerstva školství mládeže a tělovýchovy č. MSM253100002, za což autoři vyslovují svůj dík.

Milan Meloun
Jiří Militký

O autorech

MILAN MELOUN (*1943), prof. RNDr. DrSc., vystudoval přírodovědeckou fakultu Univerzity J. E. Purkyně (dnešní Masarykova) v Brně 1965. Je profesorem analytické chemie a chemometrie na katedře analytické chemie Univerzity Pardubice. Vyučoval analytickou chemii a statistiku dva roky na Bagdádské univerzitě v Iráku a tři semestry přednášel chemometrii na Královské technice ve Stockholmu. Je autorem a spoluautorem 70 originálních sdělení, 16 monografií a 10 vysokoškolských učebnic, řady patentů a zlepšovacích návrhů a na konferencích přednesl více než 160 přednášek. Je členem redakčních rad zahraničních odborných časopisů Talanta a Analytica Chimica Acta, tajemníkem sekce Chemometrie při České společnosti chemické.

Jeho práce jsou zaměřeny na počítačovou analýzu instrumentálních dat. Knižně se uvedl spolu s Josefem Havlem dvoudílnou monografií *Computation of Solution Equilibria*, která po doplnění o extrakční rovnováhy vyšla v roce 1988 v anglickém nakladatelství Ellis Horwood, Chichester. V jeho pracích představuje počítač spojovací článek mezi statistikou a analytickou chemií. Výsledkem je analytická chemometrie, předmět, který přednáší od roku 1978 podle svých učebnic. V Bagdádu napsal učebnici *Data Analysis by Statistical and Computing Technique*, University Baghdad Press, 1980, a na Královské technice ve Stockholmu pak sbírku příkladů *Introduction to Chemometrics*, která obsahuje analýzu dat programovým systémem STATGRAPHICS. Vlastní pojetí analýzy experimentálních dat se promítá i do kapi-toly *Chemometrics in the Instrumental Laboratory* uvedené v monografii, editované Jaroslavem Churáčkem *Advanced Instrumental Methods of Chemical Analysis*, Academia, Praha 1993, nebo v kapitole *Hodnocení analytických výsledků* ve Vláčilově sbírce *Příklady z chemické a instrumentální analýzy*, SNTL, Praha 1983.

Zkušenosti spoluautora Jiřího Militkého přinesly řadu novějších postupů ze statistické analýzy dat, průzkumové analýzy a především interaktivní přístup k analýze dat na osobním počítači. Společně tak vzniklo první vydání učebnice *Chemometrie - Zpracování experimentálních dat na IBM PC.* Text byl přeložen do angličtiny a po doplnění o kapitoly vícerozměrné statistiky Michelem Forinou vyšel postupně jako dvojdílná učebnice u nakladatele Ellis Horwood, Chichester 1991, pod titulem Milan Meloun, Jiří Militký a Michele Forina: *Chemometrics for Analytical Chemistry - Volume I. PC-Aided Statistical Data Analysis,* a *Volume II. PC-Aided Regression and Related Methods.*

Na Univerzitě v Pardubicích přednáší Milan Meloun chemometrii, základy počítačové typografie pomocí textových editorů rodiny WordPerfect, práci s daty v tabulkových procesorech, organizuje licenční studium a krátkodobé intenzivní kurzy

chemometrie pro aplikaci v průmyslu. V těchto formách studia jsou užívány především jeho vlastní učebnice, z nichž na předním místě třeba jmenovat knihu Milan Meloun a Jiří Militký: *Statistické zpracování experimentálních dat,* PLUS, Praha 1994, a další vydání v East Publishing, Praha 1998.

Do kolekce počítačových znalostí každého studenta přírodních nebo technických věd patří vedle zpracování experimentálních dat i zvládnutí solidního textového editoru a tabulkového procesoru. Mezeru v literatuře ve své době vyplnily dvě učebnice, spíše sbírky řešených úloh, ve světě rozšířeného textového editoru WordPerfect (*Učíme se WordPerfect 5.1, ELVIRA, Praha 1993,* a druhá *Učíme se WordPerfect 6.0*, *ELVIRA, Praha 1994*) a tabulkového procesoru QUATTRO (*Učíme se QUATTRO PRO 3.0, VŠCHT, Pardubice 1992* a *Učíme se QUATTRO PRO 4.0 CZ, Multisys, Pardubice 1993*). Oblibu těchto učebnic přinesl jejich styl a účinnost, vyučují totiž software formou řešených příkladů. Po nastudování desítek, stovek povinných příkladů a vyřešení velkého počtu úloh zvládne student ve velmi krátké době aktivně potřebnou látku a je dokonale připraven k užívání softwaru.

JIŘÍ MILITKÝ (*1949), prof. Ing. CSc., ukončil fakultu textilní, specializace textilní chemie na VŠST v Liberci roku 1973. V letech 1974 až 1976 pracoval ve Státním výzkumném ústavu textilním Liberec v oddělení matematického modelování textilních struktur. V letech 1976 až 1989 pracoval ve Výzkumném ústavu zušlechtovacím Dvůr Králové n. L., kde se věnoval převážně zpracování experimentálních dat s využitím výpočetní techniky. Od roku 1990 je vedoucím katedry textilních materiálů na Technické univerzitě Liberec. V roce 1989 byl jmenován docentem, v 1993 byl jmenován řádným profesorem. Je členem několika vědeckých a odborných společností, The Textile Institute, JČMF, FEANI. Pracuje ve výboru sekce Chemometrie při České společnosti chemické a ve výboru České Statistické Společnosti. Je akademikem ukrajinské akademie inženýrských věd.

Jeho publikační činnost zahrnuje oblasti textilního inženýrství, modelování kinetických procesů v pevné fázi a zpracování experimentálních dat. Je autorem nebo spoluautorem 606 vědeckých příspěvků (publikací, monografií, referátů a posterů). Jeho první kniha *Modifikovaná PES vlákna* (spoluatoři Jiří Kryštůfek, Jiří Vaníček a Oldřich Hartych) vyšla v SNTL v roce 1984. Zcela přepracované a rozšířené vydání bylo publikováno nakladatelstvím Elsevier v roce 1991. S Jiřím Kryštůfkem zpracoval knihu *Barvení akrylových vláken a směsí*, která vyšla v SNTL Praha v roce 1987. Ve spolupráci s Milanem Melounem publikoval učebnice a monografie z oblasti využití interaktivních statistických metod v chemometrii. Jiří Militký publikoval celkem 8 knih, z nichž tři jsou zaměřeny do oblasti zpracování experimentálních dat s využitím výpočetní techniky. Moderní metody interaktivní statistické analýzy dat zpracoval do rozsáhlého seriálu příruček *Statistické metody v textilní praxi I - IV*, vydaného v letech 1982 až 1985 v Domě techniky Pardubice. Přehled metod regrese a matematického modelování publikoval v seriálu skript *Tvorba matematických modelů I - VI*, vydaných v letech 1983 až 1989 v Domě techniky Ostrava. Vytvořil systém programů pro zpracování experimentálních dat v jazyce HPL. Tyto programy jsou charakteristické tím, že kromě stránky *statistické* vycházející vždy nejdříve z ověřování předpokladů o modelech, datech a použité metodě, využívají také progresivních *numerických* postupů (zejména v oblasti lineární a nelineární

regrese). Tyto algoritmy se později staly jádrem originálního programového systému ADSTAT.

Prezentoval příspěvky na řadě konferencí o *počítačové statistice* (Edinburg, Řím, Kodaň, Dubrovník, Vídeň, Neuchatel, Tampere atd.), *chemometrii* (Montreal, Boloňa, Taormina atd.) a *souvisejících vědních disciplínách* (Nice, Perugia, Ithaca, Honolulu, Kyoto, Mt Fuji, Interlaken, Bukurešt, Lodž, Budapešt, Stockholm, Norimberk, Hakone, Bolton, Espoo, Melbourne, Hong Kong, Shanghai, Hannoi, Teneriffe, Madeira atd.).

KAREL KUPKA (*1962), Ing., vystudoval specializaci analytické a fyzikální chemie na VŠCHT Pardubice. V roce 1990 byl spoluzakladatelem společnosti TriloByte, pro statistickou analýzu a software, kde pracuje dodnes. Je spoluautorem statistických systémů pro ADSTAT a autorem statistického systému QC-Expert (pro MS Windows) zaměřenými na analýzu dat a pokročilé metody řízení jakosti v technologii a výzkumu. Vedle řady odborných článků a knihy *Statistické řízení jakosti* je autorem příspěvků na řadě konferencí v ČR, SR, Evropě, USA, Japonsku.

VÍTĚZSLAV CENTNER (*1968), Ing. Ph.D., vystudoval specializaci analytické chemie na VŠCHT Pardubice u prof. Melouna a doktorát ukončil na Farmaceutické fakultě Vrije Universiteit Brusel u Prof. Massarta. Doktorskou práci na téma "Methods and Diagnostics in Multivariate Calibration" obhájil v roce 1998. Od té doby se intenzivně zabývá vícerozměrnou analýzou dat, vícerozměrnou kalibrací a blízkou infračervenou spektroskopií (NIR). Pracoval na několika projektech EU pro farmaceutické a chemické firmy. V současnosti pracuje jako nezávislý konzultant pro oblast zajištění jakosti v akreditovaných analytických laboratořích, validace, optimalizace procesů a NIR spektroskopii. Je autorem nebo spoluautorem 10 publikací a 32 vědeckých referátů a posterů (Tervuren, Amsterdam, Budapešť, Leuven, Taragona, Brusel, Wursburg).

1

CHYBY, VARIABILITA A NEJISTOTY INSTRUMENTÁLNÍCH MĚŘENÍ

Účelem měření je stanovit velikost *měřené veličiny*, charakterizující určitou specifickou vlastnost. Specifikace měřené veličiny může vyžadovat i údaje o dalších veličinách, jako jsou čas, teplota a tlak. Jednotlivá měření jsou obvykle zatížena celou řadou různých šumů, označovaných jako *chyby*. Výsledky měření jsou vyjádřeny pomocí vhodného odhadu střední hodnoty μ a odpovídající *nejistoty*, související se šumem, nebo-li modelem chyb[1].

Klasická statistika, vycházející z definice pravděpodobnosti jako limity relativní četnosti, poskytuje celý aparát pro vyjádření nejistoty jako intervalu spolehlivosti parametru μ. Vyjádření nejistot je filozoficky blíže subjektivní definici pravděpodob-nosti jako stupně důvěry či víry. Tato pravděpodobnost však souvisí spíše s nedostatkem znalostí než s výsledkem opakovaného experimentu.

1.1 Chyby měřicích přístrojů

Kvalita měřicích přístrojů a výsledků měření se standardně vyjadřuje pomocí odpovídajících nepřesností, označovaných *chyby*. Chyby měření mohou být způsobeny řadou faktorů.

Dělení chyb. Obecně lze chyby rozdělit podle *místa vzniku* v měřicím řetězci do čtyř základních skupin[2]:

1. *Instrumentální chyby* jsou způsobeny konstrukcí měřicího přístroje a souvisí s jeho přesností. U řady přístrojů jsou identifikovány, ale také garantovány výrobcem.
2. *Metodické chyby* souvisejí s použitou metodikou stanovení výsledků měření, jako je odečítání dat, organizace měření, eliminace vnějších vlivů atd.
3. *Teoretické chyby* souvisejí s použitým postupem měření. Jde zejména o principy měření, fyzikální modely měření, použité parametry, fyzikální konstanty atd.
4. *Chyby zpracování dat* jsou numerické chyby metody a chyby způsobené užitím nevhodného statistického vyhodnocení.

Podle *příčin vzniku* lze chyby rozdělit do tří skupin:

1. *Náhodné chyby*, které kolísají při opakování měření náhodně co do velikosti i znaménka, působí nepředvídatelně a jsou popsány určitým pravděpodobnostním rozdělením. Jsou výsledkem vlivu celé řady příčin, které lze jen obtížně odstranit, popř. alespoň omezit.

2. *Systematické chyby* působí na výsledek měření předvídatelným způsobem. Bývají funkcí času nebo parametrů měřicího procesu. Mívají stejná znaménka. Konstantní systematické chyby snižují nebo zvyšují numerický výsledek všech měření o stejnou velikost. Často se navenek neprojevují a lze je odhalit až při porovnání s výsledky z jiného přístroje. Existují i systematické chyby s časovým trendem, způsobené stárnutím nebo opotřebováním měřicího přístroje. Systematické chyby měřicího přístroje se dělí na *aditivní* (chyba nastavení nulové hodnoty) a *multiplikativní* (chyba citlivosti). Typ a velikost chyby přístroje bývá garantována výrobcem.

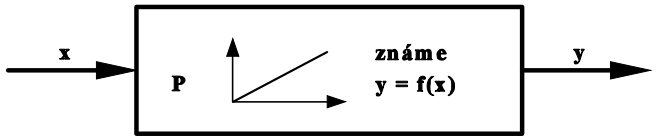
3. *Hrubé chyby*, označované jako vybočující, resp. odlehlé hodnoty, jsou způsobeny výjimečnou příčinou, náhlým selháním měřicí aparatury, nesprávným záznamem výsledku. Způsobují, že se dané měření výrazně liší od ostatních.

Chyby se dále dělí na

(a) *chyby výsledků měření* jako nejistoty hodnot výsledků měření, charakterizované například intervalem spolehlivosti, a

(b) *chyby měřicího přístroje,* resp. *procesu* měření, jako jedna z charakteristik kvality měření, udávající obyčejně přípustnou odchylku od skutečné hodnoty.

Chyba měřicího přístroje je pouze jednou součástí ovlivňující chybu výsledků měření a vhodnou volbou přesnosti měřicího přístroje se dá její vliv silně omezit. Vstupem měřicího přístroje je *měřená veličina x* a výstupem je *výsledek měření y*. Způsob transformace $y = f(x)$ je znám, obr. 1.



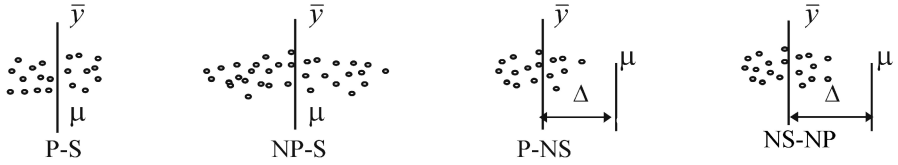
Obr. 1 Schéma měřicího přístroje.

Pro stanovení chyb přístroje je nutné znát skutečnou hodnotu měřené veličiny μ, tzv. *etalon*, nebo mít k dispozici další, *velmi přesný přístroj* a získat odhad μ velmi dokonalým měřením. V obou případech je k dispozici hodnota μ nebo její odhad $\hat{\mu}$. Výsledky opakovaných měření pak umožňují určit *míry přesnosti a správnosti* měření.

Model měření lze vyjádřit ve tvaru $y_i = g(\varepsilon_i, \mu)$, kde ε_i jsou šumové složky (externí zdroje nejistot). Funkce $g(\varepsilon_i, \mu)$ souvisí s modelem působení šumových složek, který může být aditivní, multiplikativní nebo kombinovaný. Pro aditivní model měření $y_i = \mu + \varepsilon_i$, $\{y_i, i = 1, ..., n\} \to \bar{y}, s^2$ je *rozptyl,* tj. měřítko přesnosti měření, a *průměrná odchylka* $\bar{\Delta} = \bar{y} - \mu$ je pak měřítkem správnosti, obr. 2. Využitím hodnoty μ je možno definovat různé typy odchylek od správné hodnoty μ, které vyjadřují chyby měřicího přístroje. Rozlišujeme *absolutní odchylku* zvanou také absolutní chyba $\Delta_i = x_i - \mu$ a dále *relativní odchylku* zvanou relativní chyba $\delta_i = 100\ \Delta/x$, [%]. *Celková odchylka* čili celková chyba Δ_i je složena ze dvou složek, a to ze *systematické chyby* Δ_s a *náhodné chyby* $\Delta_{N,i}$ dle vztahu

$$\Delta_i = \Delta_s + \Delta_{N,i} = \bar{y} - \mu + y_i - \bar{y}.$$

U přístrojů se obyčejně garantují různé druhy mezních chyb[3].

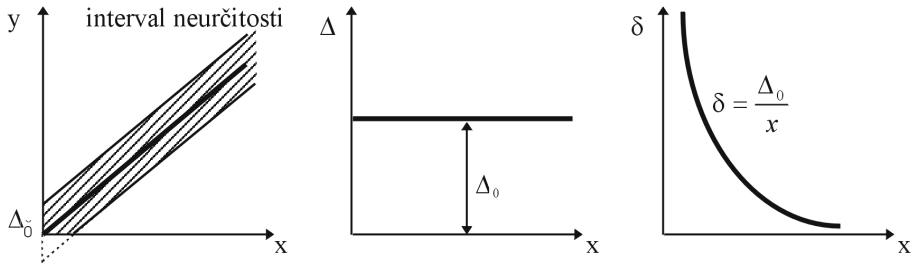


Obr. 2 Správnost a přesnost měření: P přesné, S správné, NP nepřesné, NS nesprávné.

Mezní chyba Δ_0 měřicího přístroje je jeho nejvyšší přípustná chyba, kterou ostatní odchylky měřicího přístoje za daných podmínek prakticky nikdy nepřekročí.

Redukovaná mezní chyba $\delta_{0,R}$ měřicího přístroje pro určitou hodnotu měřené veličiny x_i a stanovené podmínky je dána poměrem mezní chyby Δ_0 a měřicího rozsahu R, dle vzorce δ_0

R, $\delta_{0,R} = 100\Delta_0/R$ [%]. Měřicí rozsah R je algebraický rozdíl krajních hodnot stupnice, $R = x_{max} - x_{min}$.



Obr. 3 Konstantní absolutní chyba měření Δ_0 a relativní chyba měření δ v závislosti na měřené hodnotě veličiny x.

Třída přesnosti měřicího přístroje. Je klasifikačním znakem přesnosti v celém měřicím rozsahu přístroje. Vyjadřuje se číslem, které je vždy větší, nebo nanejvýš stejné, jako největší absolutní hodnota z redukovaných mezních chyb, zjištěných za daných podmínek v celém měřicím rozsahu přístroje. Určení třídy přesnosti záleží na typu chyby, kterou přístroj vykazuje. Dle druhu přítomné chyby pak rozlišujeme tři skupiny přístrojů:

(a) **Přístroje s konstantní absolutní chybou**, které vykazují tzv. *aditivní chybu,* tj. chybu nulové hodnoty, obr. 3. V případě čistě aditivních chyb měření se užívá *redukovaná relativní odchylka* (zde přímo rovna třídě přesnosti přístroje)

$$\delta_0 = 100\,\frac{\Delta_0}{x_{max} - x_{min}} = 100\,\frac{\Delta_0}{R},$$

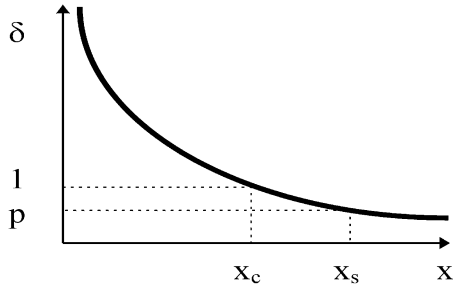
kde R je rozmezí stupnice. U přístrojů, kde působí chyby měření aditivně, klesá relativní odchylka δ hyperbolicky s hodnotou x.

Metrologické vlastnosti měřicích přístrojů charakterizují také další veličiny: *prahem*

citlivosti x_c se označuje vstupní hodnota, pro kterou je absolutní chyba rovna x_c, čili $\Delta_0 = x_c$, tj. relativní chyba $\delta(x_c) = 100$ %. Při znalosti třídy přesnosti δ_0 a rozmezí R se práh citlivosti vyčíslí podle vztahu $x_c = \delta_0 R/100$. Pro zajištění dostatečně malé hodnoty relativní chyby měřicího přístroje se definuje *spodní mez pracovního intervalu* x_s tak, aby relativní chyba $\delta(x_s)$ byla právě p %, obyčejně 4 nebo 10 %. Platí, že

$$x_s = 100 \frac{\Delta_0}{p} = 100 \frac{x_c}{p} .$$

Aditivní chyby měřicího přístroje omezují rozsah použití přístroje v oblasti malých hodnot vstupní veličiny x. Vztah mezi prahem citlivosti x_c a spodní mezí pracovního intervalu x_s vyjadřuje obr. 4.

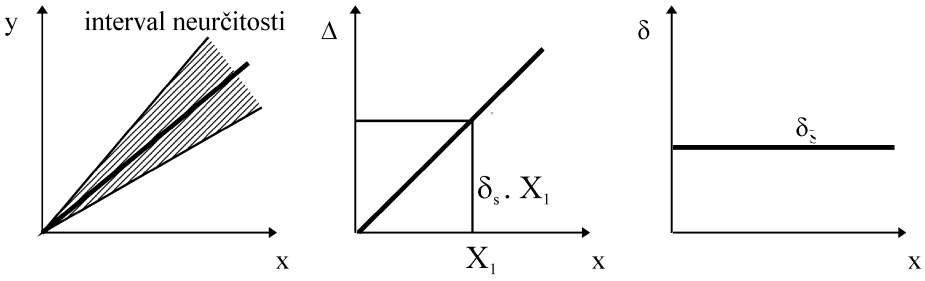


Obr. 4 Vztah mezi prahem citlivosti x_c a spodní mezí pracovního intervalu x_s.

(b) **Přístroje s konstantní relativní chybou,** kdy v případě čistě *multiplikativních chyb* měření je *relativní chyba citlivosti* (zde rovna přímo třídě přesnosti přístroje)

$$\delta_s = 100 \Delta_0/x$$

konstantní. V tomto případě je absolutní odchylka lineárně rostoucí funkcí vzhledem k veličině x, obr. 5.

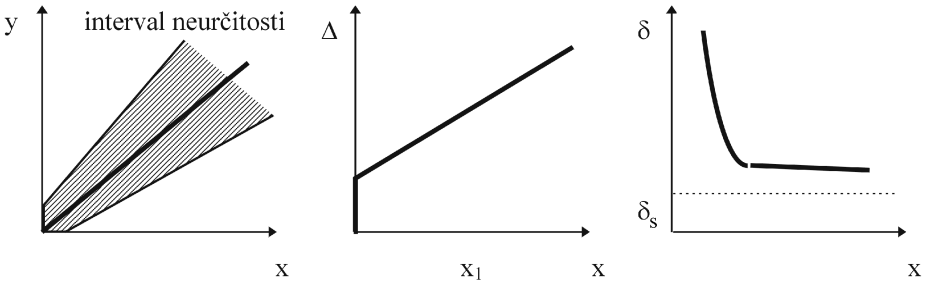


Obr. 5 Konstantní relativní chyba měření δ_s.

(c) **Přístroje s kombinovanými chybami**. U kombinovaných chyb měření lze *celkovou chybu* rozepsat na součet aditivní složky Δ_0 a multiplikativní složky $\delta_s x$ podle

rovnice ' Δ_0 % δ_s . *Interval neurčitosti,* zvaný také *pás neurčitosti,* je potom tvořen součtem příspěvků aditivního a multiplikativního pásu neurčitosti. *Celková redukovaná relativní chyba* δ_R ' δ_0 % ($\delta_s x/R$), zde monotónně roste s růstem x. Na rozdíl od případů čistě aditivní chyby tady růst δ_R začíná tím později, čím je poměr δ_s/δ_0 větší. K vyjádření třídy přesnosti δ_k se v těchto případech užívají dva údaje: *redukovaná relativní chyba* δ_0 a *chyba vzniklá na horní hranici měřicího rozsahu* δ_s, dle vzorce δ_k ' δ_0 % δ_s .

Jde o případ kombinovaného působení obou chyb, aditivní a multiplikativní, obr. 6.



Obr. 6 Kombinované působení aditivní a multiplikativní chyby měření $\delta_0 + \delta_s$.

Vzorová úloha 1.1 *Absolutní a relativní chyba pH-metru*

Skleněná elektroda k měření pH má odpor $R_1 = 5 \times 10^8$ ohmů při 25 EC a vstupní impedance milivoltmetru je $R_2 = 2 \times 10^{11}$ ohmů. Jaká je absolutní chyba Δ a relativní δ chyba měření napětí, když bylo změřeno napětí $U = 0.624$ V? Pro napětí na skleněné elektrodě platí přitom vzorec $U_x = U\,(R_1 + R_2)/R_2$.

Řešení: $U_x = 0.624\ (5 \cdot 10^8 + 2 \cdot 10^{11})/(2 \cdot 10^{11}) = 0.6256$ V,

$\Delta = 0.6256 - 0.624 = 0.0016$ V,

$\delta = 0.0016 \times 100 / 0.6256 = 0.26$ %.

Závěr: Napětí U_x je 0.6256 V, absolutní chyba Δ je 1.6 mV a relativní δ je 0.26 %.

Vzorová úloha 1.2 *Třída přesnosti a práh citlivosti ampérmetru*

Do jaké třídy přesnosti patří a s jakým prahem citlivosti pracuje miliampérmetr rozsahu $R = 60$ mA, jestliže pro skutečnou hodnotu proudu $\mu = 50$ mA byla naměřena střední hodnota $\bar{x} = 49.6$ mA?

Řešení: $\Delta_0 = 50.0 - 49.6 = 0.4$ mA,

$\delta_0 = 0.4 \times 100 / 60 = 0.67$ % a po zaokrouhlení 1 %,

$x_c = 0.67 \times 60 / 100 = 0.402$ mA a po zaokrouhlení 0.4 mA.

Závěr: Třída přesnosti je 1 % a práh citlivosti 0.4 mA.

Vzorová úloha 1.3 *Mezní absolutní a relativní chyba ampérmetru*

Na ampérmetru je uveden údaj hodnot δ_k/δ_0, numericky 1.5/0.5, a maximální rozsah $R = 50$ mA. Určete mezní absolutní chybu Δ_0 a relativní chybu δ_0 měření pro hodnoty okolo $x = 10$ mA.

Řešení: Ampérmetr vykazuje kombinovanou chybu. Celková relativní chyba je

$$\delta_0 = 1.5\ \% + 0.5\ (\frac{50}{10} - 1) = 3.5\ .\ 3\ \%\ \text{(po zaokrouhlení)}$$

a mezní absolutní chyba je

$$\Delta_0 = \frac{1.5 \cdot 10\ \% + 0.5\ (50 - 10)}{100} = 0.35\ .\ 0.3\,\text{mA}\ \text{(po zaokrouhlení).}$$

Závěr: Výsledek měření se proto zapíše ve tvaru 10 ± 0.3 mA.

1.2 Způsoby vyjádření odhadů chyb měření

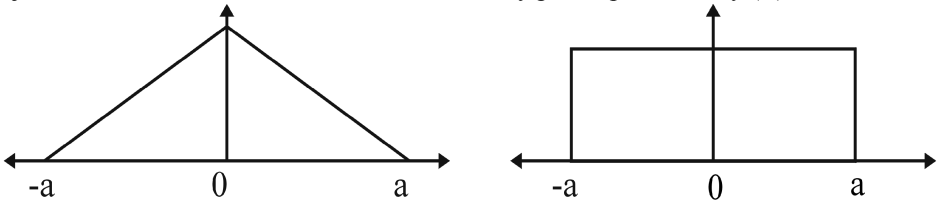
K vyjádření absolutních chyb měření Δ se nejčastěji využívá pravděpodobnostní přístup, vycházející ze znalosti pravděpodobnostního zákona rozdělení chyb, vyjádřeného hustotou pravděpodobnosti $f(\Delta)$. To umožňuje zahrnutí systematické složky chyb jako *střední hodnoty chyb* a náhodné složky chyb jako relativní šířky rozdělení, vyjádřené *rozptylem* či ostatními mírami rozptýlení. Lze použít také nepravděpodobnostní přístup, založený na *intervalové analýze.*

Pro praktický odhad chyb měření Δ, a to i přístrojových chyb, lze užít celou řadu charakteristik, počítaných z výběrových hodnot chyb $\Delta_i = x_i - \mu_s$, kde μ_s je buď hodnota standardu, nebo odhad skutečné hodnoty μ.

Základní charakteristiky polohy a rozptýlení chyb vycházejí ze znalosti hustoty pravděpodobnosti chyb $f(\Delta)$:

1. *Momenty* jsou jednak *obecné*, typu $M_K(x) = \int x^K f(x)\,\mathrm{d}x$, jednak *centrální* typu $c_K(x) = \int [x - M_1(x)]^K f(x)\,\mathrm{d}x$.

2. *Speciální míry rozptýlení* se určují z vhodného aproximujícího rozdělení. Pokud je znám pouze interval chyb $-a \le \Delta \le a$, resp. pouze *mezní odchylka a*, volí se buď trojúhelníkové, nebo rovnoměrné rozdělení hustoty pravděpodobnosti $f(\Delta)$, obr. 7.



Obr. 7 Hustota pravděpodobnosti pro trojúhelníkové a rovnoměrné rozdělení.

V obou případech je *střední hodnota chyb* $E(\Delta) = 0$ a *směrodatná odchylka* pro

(a) pro rovnoměrné rozdělení rovna $\sigma_R = a/\sqrt{3} = 0.5774\,a$, a pro

(b) trojúhelníkové rozdělení $\sigma_T = a/2\sqrt{6} = 0.2041\,a$.

Jako míra rozptýlení se pak bere buď σ_R, nebo σ_T podle toho, které z těchto rozdělení lépe vystihuje daný problém (jde vlastně o výpočet *nejistoty typu B*).

3. *Pravděpodobnost* $P(a \# \Delta \# b)$, s jakou chyby leží ve zvoleném intervalu [*a, b*].

4. *Kvantily*, tj. hodnoty chyb $\tilde{\Delta}_\alpha$, pro které platí, že $P(\Delta \# \tilde{\Delta}_\alpha) = \alpha$. To znamená, že α % všech chyb leží pod hodnotou $\tilde{\Delta}_\alpha$.

Je zřejmé, že pravděpodobnosti a kvantily spolu vzájemně souvisí. Například předpoklad symetrického rozdělení umožňuje stanovení mezí [*a, b*] jako speciálních kvantilů, pro které je pravděpodobnost $P(\Delta \# a) = \alpha/2$ a pravděpodobnost $P(\Delta \# b) = 1 - \alpha/2$.

Obecně lze pro tyto charakteristiky definovat jistý interval $[a^-, b^+]$ jejich možných hodnot. Pro případ, že je známa hustota pravděpodobnosti *f(Δ)*, lze tuto úlohu řešit pomocí standardních statistických metod (tj. *intervalů spolehlivosti*). Další možností je použití metodiky stanovení neurčitosti výsledků měření. Při nepravděpodobnostním přístupu lze využít také intervalové analýzy.

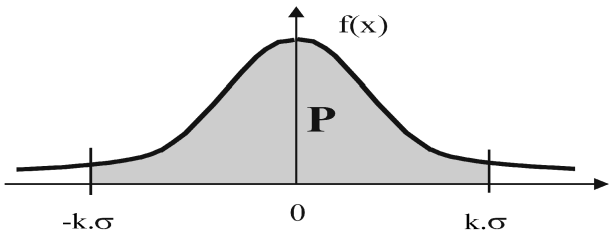
1.2.1 Momentové odhady chyb

Obecně lze při znalosti chyb Δ_i, $i = 1, ..., n$, určit odpovídající rozptyl σ^2_Δ. Na základě představy, že měření je vyjádřeno jednoduchým aditivním modelem $x_i = \mu + \Delta_i$, lze snadno určit rozptyl měřené veličiny *x* jako σ^2. Pokud platí, že $\bar{\Delta} = 0$ (střední hodnota chyb je rovna nule, $E(\Delta) = 0$, tzn. že systematická složka chyby je nulová), je $\sigma_\Delta = \sigma$, kde

$$\sigma = \sqrt{\frac{1}{n - 1} (x_i - \bar{x})^2}.$$

Pravděpodobnostní interval chyb. Předpokládejme, že chyby mají symetrickou hustotu pravděpodobnosti *f(Δ)* se střední hodnotou $E(\Delta) = 0$. Nechť je známa také distribuční funkce $F(\Delta)$. Pro pravděpodobnostní interval, v nemž leží 100(1 - α)% všech chyb, platí

$$P = (-k\sigma \# \Delta \# k\sigma) = F(k\sigma) - F(-k\sigma) = 1 - 2F(-k\sigma) = 1 - \alpha.$$



Obr 8. Pravděpodobnostní interval chyb.

Zde *-k* představuje α kvantil, *k* je 1 - α kvantil standardizovaného rozdělení chyb a *σ* je směrodatná odchylka. Pro řadu rozdělení platí, že pro $P = 0.9$ je $*k* = 1.64$, takže *pravděpodobnostní interval náhodné chyby* Δ se vyjádří nerovností

$$-1.64\,\sigma \# \Delta \# 1.64\,\sigma .$$

Toleranční interval chyb. Je-li znám pouze odhad směrodatné odchylky *s* a je-li střední hodnota chyb opět nulová $E(\Delta) = 0$, vyjádří se *toleranční interval náhodné chyby* Δ nerovností

$$-k_T s \le \Delta \le k_T s ,$$

kde za předpokladu normálního rozdělení chyb bude

$$k_T = u_{(1+P)/2} \sqrt{\frac{n-1}{\chi^2_\alpha (n-1)}}$$

a $u_{(1+P)/2}$ je kvantil normovaného normálního rozdělení, χ^2_α je α-kvantil χ^2 rozdělení. Platí pravidlo, že *toleranční intervaly jsou vždy širší než intervaly pravděpodobnostní.*

Kombinace rozptylů. Výsledný rozptyl σ^2_V se vyčíslí na základě propagace rozptylů z *m* zdrojů

$$\sigma^2_V = \sum_{i=1}^{m} \sigma_i^2 + 2 \sum_{i=1}^{m} \sum_{j=i+1}^{m} \operatorname{cov}(x_i, x_j),$$

kde σ_i^2 je rozptyl, způsobený *i*-tým zdrojem, $\operatorname{cov}(x_i, x_j)$ je kovariance mezi *i*-tým a *j*-tým zdrojem. Potom platí, že pro

a) vzájemně *nezávislé rozptyly*, kdy $\operatorname{cov}(x_i, x_j) = 0$, vyjde až na konstantu *kvadratický průměr rozptylů* dle vzorce $\sigma_{VN} = \sqrt{\sum_{i=1}^{m} \sigma_i^2}$,

b) *lineárně závislé rozptyly,* kdy $\operatorname{cov}(x_i, x_j) = \sigma_i \sigma_j$ bude až na konstantu rovna *aritmetickému průměru rozptylů* dle vzorce $\sigma_{VL} = \sum_{i=1}^{m} \sigma_i$. Ze známé trojúhelníkové nerovnosti plyne, že $\sqrt{\sum \sigma^2} < \sum \sigma_i$.

V souladu s tím, že se připouští vždy horší varianta, je vhodné volit v případech, kdy nejsou o korelacích mezi zdroji chyb žádné informace, jako celkovou směrodatnou odchylku σ_{VL}.

Volba měřicího přístroje vzhledem k chybám měření. Celková chyba měření σ_V pro případ, že *variabilita měřeného materiálu* vyjádřená *rozptylem* σ^2 a *rozptyl měřicího přístroje* τ^2 pocházejí z nezávislých zdrojů, se vyčíslí výrazem

$$\sigma_V^2 = \sigma^2 + \tau^2 .$$

Celková chyba měření bude přitom záviset na volbě přístroje:

1. Pro *velmi přesný přístroj* je složka chyby τ zanedbatelně malá a proto platí $\sigma_V = \sigma$. Opakováním lze zlepšit přesnost měření.

2. Pro *vhodně zvolený přístroj* platí $\tau \doteq \sigma/3$ a pak bude celková chyba měření rovna

$$\sigma_V = \sqrt{\sigma^2 + \sigma^2/3} = \sigma \sqrt{10/9} \doteq \sigma .$$

3. Pro *srovnatelné chyby* platí $\tau \doteq \sigma$. Pak bude celková chyba měření rovna $\sigma_V = \sqrt{2}\,\sigma = 1.4142\,\sigma$.

4. Pro *nepřesný přístroj* platí $\sigma_V \gg \tau$, opakováním měření tedy nelze přesnost zlepšit.

Pravidla o chybách měření. Při měření je vhodné respektovat pravidla o chybách:

(a) Při měření veličin x_1 a x_2, které se pro získání výsledku sčítají či odčítají $y = x_1 \pm x_2$ dbáme, aby oba sčítance x_1 a x_2 byly měřeny se stejnou absolutní přesností. Je-li chyba jedné z nich mnohem větší, rozhoduje pak sama o chybě výsledku y.

(b) Je-li výsledkem sčítání či odčítání malá hodnota $y = x_1 \pm x_2$, je výsledek zatížen velkou relativní chybou. Tomu se vyhýbáme a malé hodnoty se snažíme měřit přímo.

(c) Při měření veličin x_1 a x_2, které pro získání výsledku násobíme $y = x_1 x_2$ nebo dělíme $y = x_1 / x_2$, by měly mít obě veličiny x_1 a x_2 stejnou relativní přesnost. V případě součinu mocnin s různými exponenty je výhodnější, jsou-li relativní chyby nepřímo úměrné příslušným exponentům, aby součin exponentu a relativní chyby byl přibližně konstantní.

Vzorová úloha 1.4 *Relativní a absolutní systematická chyba pipety*

Pipeta o objemu 5 ml byla kontrolována vážením a po přepočtu byly získány hodnoty objemu v ml: 4.969, 4.945, 5.058, 5.021, 4.945, 5.006, 4.972, 5.022, 5.013 a 4.986. Určete relativní a absolutní systematickou chybu pipety a proveďte analýzu dat.

Řešení: Objem pipety $\bar{x}$ je 4.9937 ml s rozptylem $s^2(x) = 0.0013$. *Odhad absolutní systematické chyby pipety* ($\hat{a} = \bar{x} - \mu$) je -0.0063 ml. *Odhad relativní systematické chyby pipety* ($\delta = 100\ (\hat{a}/\bar{x})$) je -0.13 %. Jelikož $\mu = 5.000$ je pevná hodnota, bude *rozptyl* $s^2(a) = s^2(\bar{x}) = s^2(x)/n$ roven hodnotě 0.000134. Za předpokladu normálního rozložení chyb bude

a) *95% interval spolehlivosti systematické chyby*

$$\hat{a} - t_{0.95}(10 - 1) \times s(a) \le a \le \hat{a} + t_{0.95}(10 - 1) \times s(a)$$

kde kvantil Studentova rozdělení $t_{0.95}(9) = 2.263$ a dosazením do nerovnosti bude

$$-0.0325 \le a \le 0.0199 \ .$$

b) *95% toleranční interval systematické chyby* se spolehlivostí $(1 - \alpha) = 0.99$ je roven

$$\hat{a} - k_T \times s(a) \le a \le \hat{a} + k_T \times s(a) \ ,$$

kde pro k_T platí vztah $k_T = 1.96 \sqrt{\frac{9}{2.088}} = 4.069$ a po dosazení do nerovnosti bude

$$-0.0534 \le a \le 0.0408 \ .$$

c) Je-li rozptyl náhodných chyb vážení objemu vody roven $s^2(x)$, bude *95 % toleranční interval* se spolehlivostí 0.99

$$-0.1489 \le \Delta \le 0.1489$$

a *mezní kvantilová chyba* pipety

$$\Delta_{0.9} = 1.65\ s(x) = 1.65 \times 0.0366 = 0.0604 \ .$$

Závěr: Protože 95% interval spolehlivosti systematické chyby i toleranční interval systematické chyby pokrývají hodnotu nula, lze považovat *systematickou chybu pipety*

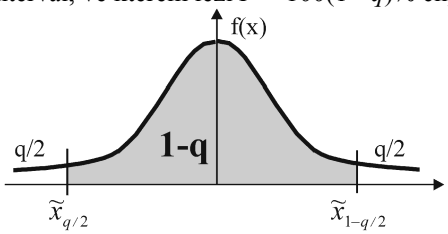
$\hat{a}$ = 0.0063 ml za statisticky nevýznamnou a objem pipety se vyjádří jako 4.994 ± 0.060 ml.

1.2.2 Kvantilové odhady chyb

Uvažujme stejnou situaci jako u momentových odhadů, tj. *směrodatná odchylka měření* σ je přímo rovna *střední kvadratické chybě* σ_Δ, čili $\sigma = \sigma_\Delta$. Jednou z jednoduchých charakteristik rozptýlení je tzv. *interkvantilová odchylka*

$$K_{1\&q} = (\tilde{x}_{1\&q/2} - \tilde{x}_{q/2}).$$

Zde $\tilde{x}_\alpha$ představuje obecně kvantil rozdělení chyb, ve kterém leží 100α% všech chyb. Hodnota K_{1-q} definuje interval, ve kterém leží $P = 100(1 - q)$% chyb.



Obr 9. Kvantilový interval chyb.

Pro zvolenou pravděpodobnost, čili statistickou jistotu $P = 1 - q$, je pak *mezní chyba měření* rovna $\Delta_{\Delta P} = K_{1\&q}$ a odpovídá intervalu obsahujícímu 100(1 - *q*)% všech chyb. Její velikost obecně závisí na hodnotě *q* a na konkrétním zákonu rozdělení chyb. Pro vybrané hodnoty pravděpodobnosti *P* je v praxi zavedeno specifické označení dotyčné chyby:

(a) *Střední chyba* ($P = 0.5$): $\sigma_{\Delta\ 0.5} = (\tilde{x}_{0.75} - \tilde{x}_{0.25})/2$, užívá se pro normální rozdělení platí $\sigma_{\Delta\ 0.5} = 0.68\ \sigma$.

(b) *Pravděpodobná chyba* ($P = 0.683$): $\sigma_{\Delta\ 0.683} = (\tilde{x}_{0.8415} - \tilde{x}_{0.1585})/2$, užívá se pro normální rozdělení platí $\sigma_{\Delta\ 0.683} = \sigma$.

(c) *Chyba pro libovolné neznámé rozdělení* ($P = 0.9$): $\sigma_{\Delta\ 0.9} = (\tilde{x}_{0.95} - \tilde{x}_{0.05})/2$.

Pro řadu rozdělení platí, že $\sigma_{\Delta\ 0.9} = 1.64\ \sigma$, a proto je pro sčítání dílčích kvantilových chyb vhodné využít vztahu $\sigma_{\Delta\ 0.9} = \sqrt{\sum \sigma^2_{\Delta\ 0.9,i}}$. V případě, kdy chyby měření mají normální rozdělení, lze psát $\sigma_{\Delta P} = u_{(1\%P)/2}\ \sigma$, kde $u_{(1+P)/2}$ je 100(1+*P*/2)% kvantil normovaného normálního rozdělení. Pro ostatní rozdělení platí, že mezní kvantilovou chybu $\sigma_{\Delta P}$ lze vyjádřit vztahem $\sigma_{\Delta P} = h\ \sigma$. Velikost *h* souvisí se špičatostí g_2 daného rozdělení chyb podle vztahu

$$h \doteq 1.62[3.8(g_2 - 1.6)^{2/3}]^Z, \quad \text{kde} \quad Z = \log[\log(1/(1 - P))].$$

Je třeba ale zdůraznit, že kvantilové chyby $\sigma_{\Delta P}$ nelze obecně sčítat.

Vzorová úloha 1.5 *Kvantilové odhady chyb přístroje*

Na základě předběžných experimentů byl zjištěn rozptyl měřicího přístroje $\sigma^2 = 0.5$. Stanovte 95% mezní kvantilovou chybu $\sigma_{\Delta 0.95}$ pro případ, kdy je známo, že měřicí přístroj

má (a) normálně rozdělené chyby, (b) rovnoměrně rozdělené chyby.
Řešení: K výpočtu mezních chyb se vypočte $Z = 1.14287$ pro $P = 0.95$:

a) Pro normální rozdělení je $g_2 = 3$. Dosazením bude odhad $h = 1.936$. Vyčíslením je pak kvantilový odhad chyby $\sigma_{\Delta 0.95} = 0.968$. Uveďme, že skutečná hodnota h je pro tento případ 1.96 (odečteno ze statistických tabulek).

b) Pro rovnoměrné rozdělení je $g_2 = 1.8$ a dosazením vyjde odhad $h = 1.669$. Vyčíslením bude pak kvantilový odhad chyby $\sigma_{\Delta 0.95} = 0.835$.
Závěr: Typ rozdělení chyb výrazně ovlivňuje kvantilový odhad chyby. Pro normální rozdělení je $\sigma_{\Delta 0.95} = 0.968$ a pro rovnoměrné rozdělení je tento odhad menší $\sigma_{\Delta 0.95} = 0.835$.

1.2.3 Nepravděpodobnostní intervalové odhady chyb

V některých případech chybí pravděpodobnostní informace o příčinách nejistoty proměnných. Pro každou proměnnou x_j se dá pouze vyjádřit ohraničení ve tvaru $[a_j^D \# x_j \# a_j^H]$, kde a_j^D je dolní mezní hodnota a a_j^H je horní mezní hodnota proměnné x_j. Tímto intervalem lze definovat tzv. *intervalové proměnné.* Pro intervalové proměnné platí obecně $A \subset [a_1, a_2]$ pro $s \in \{a_1 \# a_2\}$. Pro kombinaci intervalových proměnných A a B lze použít intervalové aritmetiky. Platí, že

$$A \cup B \subset [a_1, a_2] \cup [b_1, b_2] \subset [a_1 \cup b_1, a_2 \cup b_2],$$
$$A \cap B \subset [a_1, a_2] \cap [b_1, b_2] \subset [a_1 \cap b_2, a_2 \cup b_1],$$
$$A - B \subset [a_1, a_2] - [b_1, b_2] \subset$$
$$[\min(a_1 - b_1, a_2 - b_2, a_2 - b_1, a_2 - b_2),$$
$$\max(a_1 - b_1, a_1 - b_2, a_2 - b_1, a_2 - b_1, a_2 - b_2)],$$
$$A / B \subset [a_1, a_2] / [b_1, b_2] \subset [a_1, a_2] \cdot [1/\cap b_2, 1/b_1].$$

Tato aritmetika se dá použít v jednodušších situacích pro stanovení intervalů chyb měření.

1.3 Šíření chyb a nejistot

1.3.1 Metoda Taylorova rozvoje

Případ jedné, přímo měřené veličiny *x*. Uvažujme nejdříve případ jedné přímo měřené veličiny x, kdy výsledky měření jsou $\{x_i\}$, i = 1, . . . n, a z nich se určují odhady $\bar{x}$, s_x^2. Výsledek nepřímých měření je vyjádřen známou funkcí $y = f(x)$. Obecně zde značí "f(.)" nelineární funkci, a proto platí, že $\bar{y} \neq f(\bar{x})$. Odhad $\bar{y}$, s_y^2 se provádí s využitím Taylorova rozvoje $f(x)$ v okolí $\bar{x}$

$$f(x) = f(\bar{x}) + \frac{1}{1!}\frac{\mathrm{d}f(x)}{\mathrm{d}x}(x - \bar{x}) + \frac{1}{2!}\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}(x - \bar{x})^2 + \ldots ,$$

$$E(f(x)) \;/\; \bar{y} \;.\; f(\bar{x}) \;\%\; \frac{\mathrm{d}f(x)}{\mathrm{d}x} E(x \;\&\; \bar{x}) \;\%\; \frac{1}{2}\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} E(x \;\&\; \bar{x})^2,$$

$$\bar{y} \;.\; f(\bar{x}) \;\%\; \frac{1}{2}\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} s_x^2,$$

$$D(f(x) \;\&\; f(\bar{x})) \;/\; s_y^2 \;.\; D\left[\frac{\mathrm{d}f(x)}{\mathrm{d}x}(x \;\&\; \bar{x})\right],$$

$$s_y^2 \;.\; \left\{\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right\}^2 s_x^2.$$

Případ více měřených proměnných. Výsledkem více nepřímých měření je pak funkční vztah $f(x_1,...,x_m)$. Ze známých výsledků více přímých měření se určí

$$\bar{x}_1,\; s_{x_1}^2,\; ...,\; \bar{x}_m,\; s_{x_m}^2 \quad .$$

Označme zde vektor průměrů symbolem $\bar{\boldsymbol{x}}$ ' $(\bar{x}_1,\; ...,\; \bar{x}_m)$. Pro Taylorův rozvoj pak platí

$$f(\boldsymbol{x}) \quad . \quad f(\bar{\boldsymbol{x}}) \;\%\; \sum_{i}^{m} \frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}x_i} (x_i \;\&\; \bar{x}_i) \;\%$$

$$\%\; \frac{1}{2} \sum_{i}^{m} \frac{\mathrm{d}^2 f(\boldsymbol{x})}{\mathrm{d}x_i^2} (x_i \;\&\; \bar{x}_i)^2 \%\; \sum_{i}^{m\&1} \sum_{j>i}^{m} \frac{\mathrm{d}^2 f(\boldsymbol{x})}{\mathrm{d}x_i\, dx_j} (x_i \;\&\; \bar{x}_i)\,(x_j \;\&\; \bar{x}_j) \;\%\; ...,$$

$$\bar{y} \quad . \quad f(\bar{\boldsymbol{x}}) \;\%\; \frac{1}{2} \sum_{i}^{m} \frac{\mathrm{d}^2 f(\boldsymbol{x})}{\mathrm{d}x_i^2} s_{x_i}^2 \;\%\; \sum_{i}^{m\&1} \sum_{j'i\%1}^{m} \frac{\mathrm{d}^2 f(\boldsymbol{x})}{\mathrm{d}x_i\, dx_j} \mathrm{cov}(x_i,\, x_j) \quad ,$$

$$s_y^2 \quad ' \quad \sum_{i}^{m} \left[\frac{\mathrm{d}f(x)}{\mathrm{d}x_i}\right]^2 s_{x_i}^2 \;\%\; 2 \sum_{i}^{m\&1} \sum_{j>i}^{m} \left[\frac{\mathrm{d}f(x)}{\mathrm{d}x_i}\frac{\mathrm{d}f(x)}{\mathrm{d}x_j}\right] \mathrm{cov}\;(x_i,\, x_j) \quad .$$

kde $\mathrm{cov}(x_i, x_j)$ je kovariance mezi veličinami x_i a x_j. Existují pak krajní případy pro s^2_y.

Vzorová úloha 1.6 *Šíření chyb v metodě izotopového zřeďování*

Arsen ve vzorku byl stanoven metodou izotopového zřeďování. Byla změřena měrná aktivita $a_2 = 37000\ \mathrm{s}^{-1}$ a po standardním přídavku As o hmotnosti $m = 5 \cdot 10^{-7}$ g byla měrná aktivita $a_1 = 5300000\ \mathrm{s}^{-1}$. Stanovte relativní chybu obsahu arsenu ve vzorku, pokud je relativní chyba vážení $\delta(m) = 0.03$ % a relativní chyba stanovení aktivity $\delta(a_1) = \delta(a_2) = 1$ %.

Řešení: Pro množství arsenu ve vzorku platí

$$m_x = m \frac{a_1 - a_2}{a_2} .$$

Předpokládejme, že m, a_1, a_2 jsou vzájemně nekorelované, takže dosazením dostaneme

$$\bar{m}_x = m \frac{a_1 - a_2}{a_2} + m\, a_1 \frac{s^2(a_2)}{a_2^3}$$

$$= 7.112 \cdot 10^{-5} + 7.162 \cdot 10^{-9} = 7.112 \cdot 10^{-5}\,\mathrm{g}$$

Pro rozptyl lze psát

$$s^2(m_x) = \left(\frac{a_1}{a_2} - 1\right)^2 s^2(m) + \left(\frac{m}{a_2}\right)^2 s^2(a_1) + \left(\frac{m\, a_1}{a_2^2}\right)^2 s^2(a_2)$$

$$= \left(\frac{a_1}{a_2} - 1\right)^2 m^2\, \delta^2(m) + \left(\frac{m\, a_1}{a_2}\right)^2 [\delta^2(a_1) + \delta^2(a_2)]$$

$$= 3.2 \cdot 10^{-18} + 1.0259 \cdot 10^{-12} = 1.0259 \cdot 10^{-12}.$$

Závěr: Relativní chyba je $\delta(m_x) = 100\, s(m_x)/m_x = 1.424$ %.

Vzorová úloha 1.7 *Korelace chyb objemů v laboratorních operacích*

Množství $m = 0.1$ g Zn bylo rozpuštěno v HCl a převedeno do objemu $V = 1000$ ml. Objem tohoto roztoku $V_1 = 100$ ml byl dále zředěn doplněním v odměrce $V_2 = 1000$ ml. Pro instrumentální analýzu bylo odpipetováno $V_3 = 5$ ml a dále naředěno do objemu $V_4 = 25$ ml. Určete koncentraci roztoku a její relativní chybu, je-li směrodatná odchylka vážení $s(m) = 0.3$ mg, odměrného nádobí $s(V) = s(V_2) = 0.2$ ml, $s(V_1) = 0.05$ ml, $s(V_3) = 0.005$ ml a $s(V_4) = 0.025$ ml.

Řešení: Koncentrace c se vyčíslí podle vztahu $c = m V_1 V_3/(V V_2 V_4)$. Chyby objemů V_2 a V_4 budou silně korelované s chybami objemů V_1 a V_3. Uvažujme nejprve ideální případ, kdy jsou korelační koeficienty $r_{V_1 V_2} = r_{V_3 V_4} = 1$, zatímco ostatní veličiny jsou nekorelované. Pak vyjde

$$\delta^2(c) = \left(\frac{s(m)}{m}\right)^2 + \left(\frac{s(V)}{V}\right)^2 + \left(\frac{s(V_1)}{V_1}\right)^2 + \left(\frac{s(V_2)}{V_2}\right)^2 +$$

$$+ \left(\frac{s(V_3)}{V_3}\right)^2 + \left(\frac{s(V_4)}{V_4}\right)^2 - 2\, \frac{s(V_1)}{V_1}\, \frac{s(V_2)}{V_2} - 2\, \frac{s(V_3)}{V_3}\, \frac{s(V_4)}{V_4}.$$

Po dosazení získáme $\delta(c) = 0.302$ %. V případě, že bude zanedbána jak korelace mezi V_1 a V_2, tak i mezi V_3 a V_4, čili korelační koeficient $r_{V_1 V_2} = r_{V_3 V_4} = 0$, bude $\delta(c) = 0.336$ %. Dosazením příslušných derivací se vyčíslí střední hodnota koncentrace $\bar{c}$,

podle rovnice

$$\bar{c} \doteq \frac{m}{V}\frac{V_1}{V_2}\frac{V_3}{V_4} + m\,V_1\,V_3\left[\frac{s^2(V)}{V_3\,V_2\,V_4} + \frac{s^2(V_2)}{V_2\,V\,V_4} + \frac{s^2(V_4)}{V_4\,V\,V_2}\right] -$$

$$- \frac{m\,V_3}{V\,V_2^2\,V_4}\,s(V_1)\,s(V_2) - \frac{m\,V_1}{V\,V_2\,V_4^2}\,s(V_3)\,s(V_4),$$

ve které první člen je roven $2\cdot10^{-6}$, druhý $2.16\cdot10^{-12}$ a třetí $2.2\cdot10^{-12}$. Při zanedbání dvou nejmenších členů bude průměrná koncentrace $\bar{c} = 2\cdot10^{-3}$ g l^{-1}, $s(\bar{c}) = 6.73\cdot10^{-6}$ g. l^{-1}.
Závěr: Korelace mezi odebíranými objemy V_1 a V_3 a doplňovanými objemy V_2 , V_4 snižuje celkovou relativní chybu koncentrace, způsobenou navažováním a zřeďováním roztoků.

Vzorová úloha 1.8 *Výpočet jemnosti vlákna z hmotností a délek vláken*

Cílem je výpočet jemnosti $\bar{T} = g/L$ při znalosti střední hodnoty hmotnosti $\bar{g}$, jejího rozptylu s_g^2 a dále střední hodnoty délky vlákna $\bar{L}$ a jejího rozptylu s_L^2 za předpokladu, že měření jsou nekorelovaná, cov(*g, L*) = 0. Výpočet se provede dle vztahu

$$\bar{T} \doteq \frac{\bar{g}}{\bar{L}} + \frac{\bar{g}}{\bar{L}^3}s_L^2 = \frac{\bar{g}}{\bar{L}}\left(1 + \frac{s_L^2}{\bar{L}^2}\right).$$

Závěr: Střední hodnota jemnosti vlákna $\bar{T}$ závisí pouze na přesnosti měření délky, tj. rozptylu délky vlákna s_L^2.

Případ mocninné transformace $\bar{y} = f(x) = x^P$. Pokud *x* mělo symetrické rozdělení s konstantním rozptylem σ_x^2, bude rozdělení nesymetrické s nekonstantním rozptylem

$$\sigma_y^2 \doteq \left(\frac{\delta x^P}{\delta x}\right)^2 \sigma_x^2 = P^2\,x^{2(P-1)}\,\sigma_x^2 \doteq .$$

Z Taylorova rozvoje pak vyjde

$$\bar{y} \doteq \bar{x}^P + \frac{P(P-1)}{2}\,\bar{x}^{(P-2)}\,s_x^2 = \bar{x}^P\left[1 + \frac{P(P-1)}{2}s^2\right] .$$

Použijeme symetrizační transformaci

$$Z = f(x)^{1/P} = y^{1/P} ,$$

kde *Z* již má symetrické rozdělení a tedy přibližně platí, že $\bar{Z} = \frac{1}{n}\sum Z_i$. Z toho pak plyne přibližný výraz

$$\bar{y} \;=\; \left[\frac{1}{n} \;\sum\; y_i^{1/P}\right]^P \quad .$$

Platí, že pro $P = 1$ jde o *aritmetický průměr*, pro $P = -1$ o *harmonický průměr*, pro $P = 2$ o *kvadratický průměr*. Dle typu mocniny lze zvolit odpovídající průměr.

Vzorová úloha 1.9 *Určení střední hodnoty jemnosti vláken*
Vychází se z *n*-tice úseků příze délky *L* o hmotnostech g_i. Úsek *nL* má hmotnost $g = \sum_{i=1}^{n} g_i$ a existuje metrické číslo $C = nL/g$. Pro *i*-tý úsek pak platí, že jeho metrické číslo bude $C_i = L/g_i$. Cílem je určit z dílčích jemností C_i střední hodnotu jemnosti vláken pomocí "průměrného" metrického čísla $\bar{C}$.
Řešení:
(a) *Běžný (nesprávný) postup* je takový, že použijeme aritmetický průměr metrického čísla $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$. Po dosazení vyjde, že

$$\bar{C} = \frac{L}{n}\sum_{i=1}^{n}\frac{1}{g_i} = \frac{L}{n\,g_i}\sum_{i=1}^{n}\frac{1}{1+(g_i-\bar{g})/\bar{g}} = \bar{C}[1+v_g^2].$$

(b) *Symetrizační transformace* $C_i \to \frac{1}{g_i}$ $\forall$ $P = -1$. Volí se harmonický průměr

$$\bar{C}_H = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{C_i}\right]^{-1} \qquad \bar{C}_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{g_i}{L}} = \frac{Ln}{\sum_{i=1}^{n} g_i} = \bar{C}\,.$$

(c) *Logickou úvahou* $C_i = \frac{L}{g_i}$ máme $g_i = \frac{L}{C_i}$. Protože je $\bar{C} = \frac{nL}{\sum_{i=1}^{n} g_i}$ vyjde

$$\bar{C} = \frac{nL}{\sum_{i=1}^{n} g_i} = \frac{nL}{L\sum_{i=1}^{n}\frac{1}{C_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{C_i}} = \bar{C}_H\,.$$

Závěr: Pro případ, kdy výsledek měření je úměrný reciproké hodnotě měřené veličiny je třeba použít harmonický průměr.

1.3.2 Metoda dvoubodové aproximace

Postup je založen na náhradě rozdělení pravděpodobnosti funkce *f(x)* dvoubodovým rozdělením se stejnou střední hodnotou a rozptylem[7]. Pro *odhad střední hodnoty* $\bar{y}$ platí

$$\bar{y} \doteq \frac{f(\bar{x}+s(x)) + f(\bar{x}-s(x))}{2}$$

a pro *odhad rozptylu*

$$s^2(y) \doteq \frac{[f(\bar{x}+s(x)) - f(\bar{x}-s(x))]^2}{4}\,.$$

Je-li *f(**x**)* funkcí *m nezávislých*, náhodných a vzájemně nekorelovaných veličin x_i, $i = 1,..., m$, je možné užít vztahů

pro *odhad střední hodnoty* $\bar{y}$.
$$\sum_{i}^{m} \frac{f(\bar{x}_i + s(x_i)) + f(\bar{x}_i - s(x_i))}{2\ m}$$

a pro *odhad rozptylu* $s^2(y)$.
$$\sum_{i}^{m} \frac{[f(\bar{x}_i + s(x_i)) - f(\bar{x}_i - s(x_i))]^2}{4\ m} \quad .$$

Vzorová úloha 1.10 *Určení chyby viskozity dvoubodovou aproximací*

Vypočtěte chybu viskozity glycerolu Stokesovou metodou pro experimentální data: poloměr kuličky $r = (0.0112 \pm 0.0001)$ m, hustota kuličky $\rho_0 = 1.335 \cdot 10^3$ kg m^{-3}, hustota glycerolu $\rho = 1.28 \cdot 10^3$ kg m^{-3}, dráha kuličky $l = (31.23 \pm 0.05)$ cm, kterou kulička vykoná za dobu $t = (62.1 \pm 0.2)$ s, a tíhové zrychlení $g = 9.801$ m s^{-1}.

Řešení: Viskozita η, určovaná Stokesovou metodou, se vyčíslí podle vztahu

$$\eta = \frac{2\ g\ r^2\ (\rho_0 - \rho)\ t}{9\ l} \quad .$$

Protože nejde o součtový nebo součinový výraz, nelze jednoduše určit relativní chybu. Metodou dvoubodové aproximace se vyčíslí hodnoty: $\bar{\eta} = 0.0299$ Pa. s, $s(\eta) = 5.422 \cdot 10^{-4}$ Pa. s a relativní chyba $\delta(\eta) = 1.82$ %.

Závěr: Rozdělení viskozity η je přibližně symetrické.

1.3.3 Metoda simulací Monte Carlo

Při určování středních hodnot $\bar{y}$ a rozptylů $\sigma^2(y)$ jako funkcí náhodných veličin počítačem je výhodné použít techniky simulačních experimentů metodou Monte Carlo. Postup lze shrnout do pěti kroků[8]:

1. *Zadání funkce f(x)*. V úlohách přírodních věd bývá funkce *f(x)* většinou známa. Výhodou simulace je, že funkce *f(x)* nemusí být vyjádřena v explicitním tvaru.

2. *Rozdělení měřených veličin*. Předpokládá se, že měřené veličiny jsou nezávislé a mají normální rozdělení pravděpodobnosti. Stačí zadání veličin že $\bar{x}_i$, $s(x_i)$, $i = 1, ..., m$. Pokud nejsou tyto veličiny k dispozici, postačují dvě krajní hodnoty intervalu $[A_i, B_i]$, ve kterém lze očekávat výskyt měřené veličiny x_i. Aproximativní hustotu pravděpodobnosti $f(x_i)$ lze vyjádřit pomocí parabolického rozdělení

$$f(x_i) = \frac{6}{(B - A)^2}\ (x_i - A)\ (B - x_i) \quad \text{pro } A \# x_i \# B.$$

Složitější bude situace, kdy se mezi vstupními měřenými veličinami x_i vyskytnou korelace. Pak je třeba specifikovat simultánní rozdělení všech hodnot x_i, $i = 1, ..., m$, což bude jednoduché pouze pro případ normálního rozdělení.

3. *Generace náhodných čísel*. Na počítačích existují kvalitní generátory pseudonáhodných čísel s rovnoměrným rozdělením $R[0, 1]$. Při znalosti dvou nezávislých náhodných čísel R_j , R_{j+1} s rovnoměrným rozdělením lze pomocí Boxovy-Müllerovy transformace určit dvě nezávislá náhodná čísla N_j , N_{j+1} s normovaným normálním rozdělením podle

$$N_j = \sqrt{-2 \ln R_j}\, \sin(2\pi R_{j+1}) \quad \text{a} \quad N_{j+1} = \sqrt{-2 \ln R_j}\, \cos(2\pi R_{j+1}) \ .$$

Pak j-tá simulovaná hodnota i-té veličiny x_i bude vyjádřena jako $x_{i,j}^{(} = N_j\, s(x_i) + \bar{x}_i$.

Pro parabolické rozdělení bude simulovaná hodnota $x_{i,j}$* řešením kubické rovnice

$$0.5\, x_{i,j}^{(2} - x_{i,j}^{(} - \frac{x_{i,j}^{(3}}{3} + \alpha = R_i\, \beta, \quad \text{kde} \quad \alpha = A^3 - A^2 + A \quad \beta = \frac{6}{(B - A)^2}.$$

4. *Volby počtu simulací*. Pravidla pro určení nezbytného počtu simulací jsou stejná jako pravidla pro určování velikosti výběru. Jde-li o odhad střední hodnoty a pokud lze definovat požadovanou šířku $100(1 - \alpha)\%$ intervalu spolehlivosti D, platí pro minimální počet simulací vztah $n_{\min} = \left[\frac{4\, u_{1-\alpha/2}^2}{D^2}\, s^2(y)\right] + 1$, kde $u_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení a $s^2(y)$ je odhad rozptylu určený například z prvních 50 simulací.

5. *Sumarizace výsledků*. S ohledem na maximální univerzálnost je výhodné nalézt empirickou hustotu pravděpodobnosti rozdělení souboru simulovaných hodnot y_j*, $j = 1, ..., n_{\min}$, a pak vyčíslit odhady parametrů polohy a rozptýlení.

Vzorová úloha 1.11 *Hromadění chyb při určení rozpustnosti stříbrné soli*

Součin rozpustnosti stříbrné soli AgX má hodnotu $K_S = (4.0 \pm 0.4) \cdot 10^{-8}$. Jaká je chyba vypočtené rovnovážné koncentrace stříbrných iontů $[Ag^+]$ ve vodě?

Řešení: Rozpustnost $[Ag^+]$ se vypočte podle vztahu $[Ag^+] = \sqrt{K_S}$.

1. *Metoda Taylorova rozvoje*. Přímým dosazením, kdy se vyčíslí hodnota rozpustnosti

$$[Ag^+] = \sqrt{K_S} - 0.125\, K_S^{-3/2}\, s^2(K_S) = 2 \cdot 10^{-4} - 2.5 \cdot 10^{-7} = 1.9975 \cdot 10^{-4}\ \text{mol.l}^{-1}$$

a rozptyl rozpustnosti $s^2([Ag^+]) = 0.25\, K_S^{-1}\, s^2(K_S) = 10^{-10}$ a relativní chyba rozpustnosti $\delta([Ag^+]) = 5\ \%$.

2. *Metoda dvoubodové aproximace* vede k hodnotám $[Ag^+] = 1.997 \cdot 10^{-4}$ mol . l^{-1}, $s^2([Ag^+]) = 1.003 \cdot 10^{-10}$, $s([Ag^+]) = 1.001 \cdot 10^{-5}$ mol . l^{-1}, $\delta([Ag^+]) = 5\ \%$.

3. *Metoda simulací Monte Carlo* vede k hodnotám $[Ag^+] = 1.997 \cdot 10^{-4}$ mol . l^{-1}, $s^2([Ag^+]) = 1.038 \cdot 10^{-10}$, $s([Ag^+]) = 1.019 \cdot 10^{-5}$ mol . l^{-1}, $\delta([Ag^+]) = 5\ \%$.

Závěr: Všechny tři metody poskytují shodné výsledky.

Vzorová úloha 1.12 *Korelace v hromadění chyb*

Gravimetrické stanovení obsahu oxidu železitého v železné rudě obsahující asi 50% Fe_2O_3 se provede na analytických váhách s chybou vážení $s(m)$ = 0.3 mg a navážkou vzorku m = 0.105 g. Určete chybu gravimetrického stanovení, pokud navážka vzorku m a vyvážka popela m_0 jsou v relaci, a to m_0 . 0.5 m.

Řešení: Pro hmotnostní zlomek w stanovovaného Fe_2O_3 v rudě v procentech platí $w = 100\ m_0 / m$. Jelikož jsou navážka a vyvážka silně korelovány, r_{m_0m} … 0, dosazením získáme vztah

$$\delta(w) \ ' \ \sqrt{\delta^2(m_0) \ \% \ \delta^2(m) \ \& \ 2 \ \delta(m_0) \ \delta(m) \ r_{m_0m}} \ .$$

V případě úplné lineární závislosti navážky a vyvážky, bude $r_{m_0m} = 1$ a

$$\delta(w) \ ' \ 100 \ \sqrt{\left(\frac{0.3}{52.5} \right)^2 \ \% \ \left(\frac{0.3}{105} \right)^2 \ \& \ 2 \ \frac{0.3}{52.5} \ \frac{0.3}{105}} \ ' \ 0.286 \ \%.$$

Naopak, pokud by vyvážka nezávisela na navážce, tj. $r_{m_0m} = 0$, vyšlo by $\delta(w)$ = 0.639 %. V případě částečné korelace $r_{m0m} = 0.5$ vyjde $\delta(w)$ = 0.49 %.

Střední hodnota $\bar{w}$ a její rozptyl $s^2(w)$ budou rovněž ovlivněny korelací mezi m_0 a m. Bude-li $s(m_0)$. $s(m)$. 0.3 a měření byla n-krát opakována, pak dostaneme

$$\bar{w} \ . \ 100 \frac{m_0}{m} \ \% \ \frac{s^2(m)}{m^3} \ \& \ \frac{r_{m_0m} \ s(m_0) \ s(m)}{m^2} \ .$$

Je-li $0 < r_{m_0m} < 1$, bude příspěvek třetího členu vždy zanedbatelný a $\bar{w}$. 50 %. Dosazením bude

$$s^2(w) \ . \ 10^4 [\ \frac{s^2(m_0)}{m^2} \ \% \ \frac{m_0^2 \ s^2(m)}{m^4} \ \& \ 2 \ \frac{m_0 \ r_{m_0m}}{m^3} \ s(m_0) \ s(m) \ \% \ \frac{s^2(m_0) \ s^2(m)}{m^4}]$$

a při volbě $r_{m_0m} = 1$ bude $s^2(w)$. 0.103, zatímco pro $r_{m_0m} = 0$ bude $s^2(w)$. 0.102. Pro případ $r_{m_0m} = 0$ je relativní chyba $\delta(w)$ = 0.64 % a tatáž je i pro $r_{m_0m} = 1$.

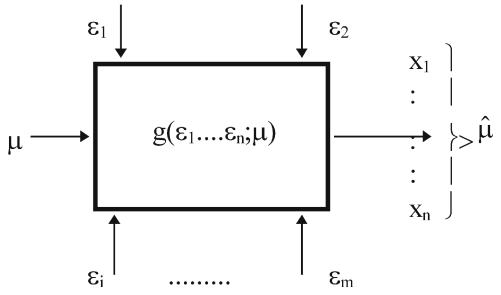
Závěr: Kladná korelace mezi navážkou a vyvážkou snižuje relativní chybu metody. Pro dostatečně veliké navážky vzhledem k chybě vážení se na odhadech střední hodnoty $\bar{w}$ a rozptylu $s^2(w)$ projeví stupeň korelace jen nevýrazně. To je způsobeno vedle vysoké relativní přesnosti měření také přibližností obou užitých vztahů.

1.4 Nejistoty výsledků měření

V této části jsou porovnány postupy k určování nejistot a s tím spojená v literatuře nově zaváděná terminologie s terminologií statistické analýzy. Zvolený přístup výpočtu nejistot je silně závislý na předpokladech o vzniku a vlastnostech jednotlivých nejistot. *Nejistoty* v nově zaváděné terminologii představují vlastně *intervaly spolehlivosti* ve statistické analýze. Základní rozdíl je pak v užívání neexperimentálních informací o zdrojích variability.

1.4.1 Porovnání přístupů k výpočtu nejistot

A. Přímá měření: pro případ řady zdrojů chyb měřicího systému pro přímá měření je situace znázorněna na obr. 10.



Obr. 10 Blokové schéma měřicího systému pro přímá měření.

Zde μ je měřená veličina, g_1, g_2, ... , g_m jsou šumové složky, nazývané externí zdroje nejistot a funkce $g(g_1, g_2, ..., g_m, \mu)$ souvisí s modelem působení šumových složek, který může být aditivní, multiplikativní a kombinovaný.

1. Analýza nejistot podle standardní statistické analýzy, cit. [15, 16].

a) *Odhad veličiny μ* (bodový), viz kap. 3.

b) *Odhad rozptylu* σ^2 (bodový), viz kap. 3.

c) *Odhad intervalu spolehlivosti (IS)* pro *μ,* viz kap. 3.

d) *Odhad vychýlení b = E(μ - μ̂),* viz kap. 3.

Obecně je třeba připustit, že odhad $\hat{\mu}$ je vychýlený, tj. střední hodnota $E(\hat{\mu}) \dots \mu$. Pak je mírou celkové variability *střední kvadratická chyba MSE*, pro kterou platí

$$MSE = E(\mu - \hat{\mu})^2 = E[\mu - E(\hat{\mu})]^2 + E[\hat{\mu} - E(\hat{\mu})]^2 = D(\hat{\mu}) + b^2$$

a která je rovna součtu rozptylu $D(\hat{\mu})$ a čtverce vychýlení b^2.

2. Analýza nejistot podle nové terminologie, cit. [11, 12].

a) *Odhad veličiny* μ (bodový), i když v nové literatuře není obvykle model speciálně uveden, předpokládá se aditivní model $x_i = \mu + \sum_{i=1}^{m} \varepsilon_i$, kde šumy mají nulové střední hodnoty $E(\varepsilon_i) = 0$ a konstantní rozptyly $D(\varepsilon_i) = \sigma_i^2$. Pak rezultuje známý odhad střední hodnoty ve formě aritmetického průměru, $\hat{\mu} = \bar{x}$.

b) *Odhad rozptylu* $D(\hat{\mu})$ předpokládá nezávislost (případně pouze lineární závislost všech složek šumu) g_i . Jsou užívány následující míry rozptýlení:

Standardní nejistota typu A, u_{Ai} , tj. směrodatná odchylka *měřené* šumové složky, se počítá jako odmocnina z výběrového rozptylu.

Standardní nejistota typu B, u_{Bi} , tj. směrodatná odchylka *neměřené* a v experimentu nesledované šumové složky, která se odhaduje jako směrodatná odchylka, odpovídající jejímu apriorně vybranému rozdělení. Řada odhadů pro apriorní rozdělení rovnoměrné, trojúhelníkové, lichoběžníkové a normální je uvedena v literatuře[11, 12].

Místo odhadu $D(\hat{\mu})$ se používá *kombinovaná standardní nejistota* u_c vycházející z platnosti výše uvedeného aditivního modelu

$$u_c^2 = \Sigma\ u_{Ai}^2 + \Sigma\ u_{Bi}^2 .$$

Pro závislé zdroje nejistot se přičítají ještě kovariance.

c) *Odhad intervalu spolehlivosti (IS)* pro μ: předpokládá se přibližná normalita, zřejmě plynoucí z centrální limitní věty. *Rozšířená nejistota*, formálně totiž poloviční šířka intervalu spolehlivosti pro μ, je pak $U = 2\,u_c$. Problémem však je, že v řadě případů některé standardní nejistoty dominují a pak již představu normality z centrální limitní věty nelze aplikovat.

d) *Odhad vychýlení* $b = E(\mu - \hat{\mu})$: vůbec se neuvažuje. Předpokládá se, že vychýlení U je odstraněno v rámci metody měření. V práci Phillipse a ost.[14] je navržen postup výpočtu nejistot pro případy, kdy vychýlení b eliminováno není. Jednoduše se vytvoří dvě rozšířené nejistoty

$$U_+ = U - b \text{ pro } U - b > 0, \text{ resp. } U_+ = 0,$$
$$U_- = U + b \text{ pro } U + b > 0, \text{ resp } U_- = 0.$$

To pak pochopitelně vede k nesymetrickému intervalu spolehlivosti.

B. Nepřímá měření: Výsledek analýzy $f(\mu_1, ..., \mu_m)$ je vytvořen známou funkcí skutečných výsledků přímých měření $\mu_1, ..., \mu_m$, (např. měříme poloměr a chceme znát plochu příčného řezu kruhových vláken). K dispozici jsou odhady parametrů $(\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_m)$ a příslušné odhady rozptylů, resp. čtverců nejistot,

$$D(\hat{\mu}_1),\ D(\hat{\mu}_2),\ ...,\ D(\hat{\mu}_m).$$

1. Analýza nejistot podle standardní statistické analýzy, cit. [15, 16].

a) *Odhad y z odhadů* $\hat{\mu}_i$, $i = 1, ..., m$, viz kap. 1.3.

b) *Odhad rozptylu* $D(\hat{y})$, viz kap. 1.3.

c) *Odhad intervalu spolehlivosti pro y,* viz kap. 1.3.

2. Analýza nejistot podle nové terminologie, cit. [11, 12].

a) *Odhad y z odhadů* $\hat{\mu}_i$, $i = 1, ..., m$. Není řešen přímo, ale velmi aproximativně se předpokládá $\hat{y} \doteq f(\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_m)$.

b) *Odhad rozptylu* $D(\hat{y})$. Je to vlastně *rozšířená nejistota u(y)*. Vychází se z předpokladu, že $f(\boldsymbol{x})$ lze nahradit linearizací Taylorovým rozvojem v okolí $\boldsymbol{\mu}$

$$y \doteq f(\boldsymbol{x}) \,.\, f(\boldsymbol{\mu}) + \sum_{i=1}^{m} \left(\frac{\mathrm{d}f(.)}{\mathrm{d}x_i} \right) (x_i - \mu_i) \quad ,$$

$$D(y) \,.\, \sum_{i=1}^{m} \left(\frac{\mathrm{d}f(.)}{\mathrm{d}x_i} \right)^2 D(x_i) + \mathrm{cov}(...),$$

$$u^2(y) \,.\, \sum_{i=1}^{m} \left(\frac{\mathrm{d}f(.)}{\mathrm{d}x_i} \right)^2 u^2(x_i) + \mathrm{cov}(...).$$

$D(y)$ se obvykle nesprávně označuje jako *zákon šíření nejistot*. V případě, že zdroje nejistot jsou lineárně závislé, provádí se korekce s využitím kovariancí cov(...). Linearizace však může být v řadě případů velmi nepřesná.

c) *Odhad intervalu spolehlivosti* pro *y*. Předpokládá se téměř vždy, ale nekorektně, přibližná normalita. Nelineární funkce normálně rozdělených náhodných veličin totiž normální rozdělení nemá. Polovina 95% intervalu spolehlivosti, čili *rozšířená nejistota*, je potom $U = 2\,u(y)$. Zde 2, či přesněji 1.96, představuje *kvantil normovaného normálního rozdělení.* Pro nelineární transformaci však rezultují nesymetrická rozdělení, což vede k nesymetrickému intervalu spolehlivosti. Ve speciálních případech (např. stopová analýza v analytické chemii) to může výrazně ovlivnit závěry, protože pro pozitivně zešikmená rozdělení vyjde totiž ve směru k nižším hodnotám korektnější interval užší a ve směru k vyšším hodnotám širší.

1.4.2 Kritické poznámky k výpočtu nejistot

a) Poznámky terminologické. Následující převodní tabulka ukazuje, že termíny doporučované v některé literatuře odpovídají vlastně běžným pojmům, užívaným ve statistice:

Nová terminologie:	**Statistika:**
Standardní nejistota A	*směrodatná odchylka měřené šumové složky s*
Standardní nejistota B	*směrodatná odchylka (odhadnutá) šumové složky s*
Kombinovaná nejistota	*směrodatná odchylka funkce y, s(y)*
Rozšířená nejistota	*polovina intervalu spolehlivosti IS*
Faktor pokrytí	*kvantil normovaného normálního rozdělení α*

b) Poznámky statistické. Výpočty v některých odkazech literatury vycházejí z předpokladů, které však nejsou v procesu navrhovaného výpočtu nikterak ověřovány:

a) aditivní model měření, resp. působení šumových složek (zdrojů nejistot),

b) konstantní rozptyl měření (resp. zdrojů nejistot),

c) normalita nelineární funkce normálně rozdělených proměnných (pro určení rozšířené nejistoty, resp. intervalu spolehlivosti IS),

d) nekorelovanost měření,

e) malá nelinearita funkce *f(x)*, umožňující použití její linearizace.

Problémem je nekorektnost při konstrukci a interpretaci *rozšířené nejistoty U*, (resp. intervalu spolehlivosti IS). Klasická statistika vede totiž k tomu, že pro $n \geq 4$ je $100(1 - \alpha)\%$ interval spolehlivosti parametru μ roven výrazu

$$\hat{\mu} \pm u_{1-\alpha/2} \sqrt{D(\hat{\mu})}.$$

Při výpočtu nejistot není *kombinovaná nejistota* u_c^2 pouze odhadem rozptylu $D(\hat{\mu})$, ale obsahuje ještě další složky. Pak vyjde *rozšířená nejistota U systematicky vyšší* než polovina intervalu spolehlivosti, hodnota 2 zde nezajišťuje přibližně 95% pokrytí a interpretace takového intervalu je nesnadná.

c) Poznámky výpočetní. Místo náhrady derivací diferencemi, jak se často doporučuje v různých příručkách, by bylo podstatně jednodušší užít *simulace* nebo tzv. *Bootstrap odhadů* zejména pak, užívá-li se počítač.

Vzorová úloha 1.13 *Nejistota aritmetických operací přibližných čísel*

Vypočtěte nejistotu výsledku *y* po provedení řady operací s přibližnými čísly.

Řešení:

$$y = \frac{4.10(\pm 0.02) \times 0.0050(\pm 0.0001)}{1.97(0.04)} = 0.0104 \pm 0.0003,$$

$$y = \frac{(14.3(\pm 0.2) - 11.6(\pm 0.2)) \times 50.0(0.1))}{42.3(0.4)} = 3.2 \pm 0.3.$$

Závěr: K výpočtu nejistot bylo užito vzorců metody propagace nejistot.

1.4.3 Přístup intervalové analýzy k nejistotám

V praxi obvykle není známo rozdělení měřených veličin x, resp. chyb měření Δ, takže analýza nejistot založená na pravděpodobnostních předpokladech je silně omezená. Při intervalové analýze se k "průměrnému" výsledku měření x musí definovat *mezní odchylka* (*chyba*) d a výsledek vyjádřit jako interval. Zde X označuje tzv. *intervalovou proměnnou* $X = [\bar{x} - d, \bar{x} + d]$.

V případě, že se vyjadřuje interval neurčitosti absolutní chyby, je $\bar{x} = 0$ a platí, že X = [-d, +d]. V obecnějším případě může být hraniční chyba funkcí úrovně měřené veličiny $d = d(\bar{x})$. Pak dostáváme intervalovou proměnnou jako funkci

$$X(\bar{x}) = [\bar{x} - d(\bar{x}), \bar{x} + d(\bar{x})].$$

Účelem je pro *výsledek nepřímých měření* $y = f(x_1, ..., x_n)$ stanovit odpovídající interval neurčitosti (mezní chybu), cit. [15]:

$$Y = [y^-, y^+] = f(X_1, ..., X_n) = \{f(x_1, ..., x_n) \text{ pro } x_1 \in X_1, x_2 \in X_2, ... \} .$$

Hodnoty y^- , y^+ jsou maximem a minimem výrazu

$$f(X_1, ..., X_n) = f(\bar{x}_1 + \Delta x_1, ..., \bar{x}_n + \Delta x_n) \text{ , kde } \Delta x_i = \mu_i - d_i \text{ ,}$$

kde μ_i je skutečná hodnota veličiny x. Na základě linearizace pomocí Taylorova rozvoje lze dospět ke vztahu

$$f(X_1, ..., X_n) = \bar{y} + \sum_{i=1}^{n} \frac{\mathrm{d}f}{\mathrm{d}x_i}(\bar{x}_1, ..., \bar{x}_n)\, Dx_i \text{ ,}$$

a tedy $$[y^-, y^+] = \bar{y} \pm d,$$

kde $\bar{y} = f(\bar{x}_1, ..., \bar{x}_n)$ a $d = \sum_{i=1}^{n} \left| \frac{\mathrm{d}f}{\mathrm{d}x_i}(\bar{x}_1, ..., \bar{x}_n) \right| d_i$.

Tento algoritmus je sice zjednodušený, ale umožňuje práci s hraničními chybami d, které nemají pravděpodobnostní charakter. Zajímavé je, že pro případ lineární funkce $f(x)$, kdy jsou derivace $\frac{df}{\mathrm{d}x_i} = 1$, je celková odchylka d součtem dílčích odchylek, což v pravděpodobnostní interpretaci znamená variantu lineárně korelovaných šumových složek.

Vzorová úloha 1.14 *Výpočet nejistoty teploty měřené rtuťovým teploměrem*

Cílem je stanovit nejistotu měření teploty rtuťovým teploměrem dle specifikace *nejistot typu B*. Příklad ilustruje jednak různé možnosti výpočtu nejistot, jednak i zásadní fakt, že lze dokonce stanovit nejistotu bez znalosti konkrétního měření.

Data: zdroje nejistot typu *B* jsou x_1 chyba teploměru dle údajů výrobce [± 0.1 EC], x_2 nejistota kalibrace dle údajů výrobce [± 1 EC], x_3 nejistota odečtu teploty, odhad [± 0.25 EC].

Řešení:

a) Za předpokladu ***rovnoměrného rozdělení*** nejistot v daném intervalu:

nejistota pro zdroj x_1 je $\sigma_{x1} = 0.5774 \cdot 0.1 = 0.05774$,

nejistota pro zdroj x_2 je $\sigma_{x2} = 0.5774 \cdot 1 = 0.5744$,

nejistota pro zdroj x_3 je $\sigma_{x3} = 0.14435$, a bude potom

kombinovaná nejistota (čili celková chyba) *pro nekorelované zdroje nejistot*

$$\sigma_c = \sqrt{\sigma_{x1}^2 + \sigma_{x2}^2 + \sigma_{x3}^2} = 0.59796 \ ,$$

rozšířená nejistota $U = 2\,\sigma_c = 1.1958$ a po zaokrouhlení 1.2.

Kombinovaná nejistota (celková chyba) *pro korelované zdroje nejistot* bude

$$\sigma_c = \sigma_{x1} + \sigma_{x2} + \sigma_{x3} = 0.77949 \quad \text{a}$$

rozšířená nejistota $U = 2\,\sigma_c = 1.5588$ a po zaokrouhlení 1.6.

b) Za předpokladu ***trojúhelníkového rozdělení*** nejistot v daném intervalu:

nejistota pro zdroj x_1 je rovna $\sigma_{x1} = 0.2041 * 0.1 = 0.02041$,

nejistota pro zdroj x_2 je rovna $\sigma_{x2} = 0.20410 * 1 = 0.20410$,

nejistota pro zdroj x_3 je rovna $\sigma_{x3} = 0.05102$ a bude potom

kombinovaná nejistota (celková chyba) *pro nekorelované zdroje nejistot* je rovna $\sigma_c = 0.21136$,

rozšířená nejistota je $U = 2\,\sigma_c = 0.42227$ a po zaokrouhlení bude 0.4.

Kombinovaná nejistota (celková chyba) *pro korelované zdroje* je $\sigma_c = 0.2755$ a

rozšířená nejistota je $U = 2\,\sigma_c = 0.5510$ a po zaokrouhlení bude 0.6.

c) Nepravděpodobnostní odhad nejistot (*intervalové proměnné*)

Celková odchylka $d = 0.1 + 1.0 + 0.25 = 1.35$ a interval neurčitosti je roven

$$\left[\underline{y}, \overline{y}\right] = \bar{y} \pm d = \pm 1.35$$

Závěr: Volba rozdělení nejistot hraje ve výpočtu nejistot rozhodující roli. Navíc je velmi pravděpodobné, že zdroje nejistot x_1 a x_2 budou zde mít spíše systematický než náhodný charakter.

1.4.4 Zaokrouhlování čísel

Zaokrouhlováním se nedopouštíme větší chyby než poloviny jednotky odpovídající řádu poslední ponechané číslice. *Relativní chyba zaokrouhleného čísla* je menší nebo rovna

$$s_{\mathrm{rel}} \# \frac{1}{2\,A\;10^{n\,\&\,1}} \ ,$$

kde A je první platná číslice a n je počet platných číslic v zaokrouhleném čísle.

Vzorová úloha 1.15 *Zaokrouhlování čísel na 2, 3 a 4 platná místa*
Jaké relativní chyby se dopustíme, když číslo 10500 zaokrouhlíme na 2, 3 a 4 platná místa? Jaká bude chyba, když první platná číslice bude 9?
Řešení: U čísla $10500 = 1.05 \times 10^4$ je $A = 1.05$ a při zaokrouhlení na *dvě platná místa* je $n = 2$ a hodnota 10500 má pak relativní chybu

$$s_{\mathrm{rel}} \# \frac{1}{2\ .\ 1\ .\ 10^{2\,\&\,1}}\ (\times\ 100\%)\ '\ \frac{1}{20}\ (\times\ 100\%)\ '\ 5\% \ ,$$

zatímco při zaokrouhlení na *tři platná místa* je $n = 3$ a hodnota 10500 má relativní chybu

$$s_{\mathrm{rel}} \# \frac{1}{2\ .\ 1\ .\ 10^{3\,\&\,1}}\ (\times\ 100\%)\ '\ \frac{1}{200}\ (\times\ 100\%)\ '\ 0.5\% \ .$$

Závěr: Platí pravidlo, že u čísel, jejichž první platná číslice je 9, jsou relativní chyby při zaokrouhlení na *dvě platná místa* menší než $s_{\mathrm{rel}} = 100\ \%/(2\ .\ 9\ .\ 10^{2\&1}) = 0.56\ \%$, dále pak při zaokrouhlení na *tři platná místa* menší než $s_{\mathrm{rel}} = 100\ \%/(2\ .\ 9\ .\ 10^{3\&1}) = 0.056\ \%$ a konečně při zaokrouhlení na *čtyři platná místa* menší než $s_{\mathrm{rel}} = 100\%/(2\ .\ 9\ .\ 10^{4\&1}) = 0.0056\ \%$.

1.5 Úlohy

Úlohy jsou rozděleny do čtyř kapitol: B1 (farmakologická a biochemická data), C1 (chemická a fyzikální data), E1 (environmentální, potravinářská a zemědělská data), H1 (hutní a mineralogická data), S1 (ekonomická, sociologická a ostatní data). Variability chemických veličin a analytických operací jsou v zadáních úloh vyjádřeny dvojím způsobem: (a) svou směrodatnou odchylkou v závorce (s), např. zápis "naváženo 16.0000 (0.0003) g NaCl" značí směrodatnou odchylku analytických vah s = 0.0003 g, nebo (b) intervalovým odhadem ± 2s, např. zápis "naváženo 16.0000 ± 0.0006 g NaCl" značí zde $2s = 0.0006$ g a $s = 0.0003$ g.

1.5.1 Analýza farmakologických a biochemických dat

Úloha B1.01 *Vliv rozptýlení kalibračního faktoru na nejistotu koncentrace*
Konec titračního stanovení byl vyhodnocen spektrofotometricky jako hodnota absorbance při 540 nm. Bylo provedeno 8 paralelních stanovení glukózy v krvi. Data vykazují normální rozdělení s průměrem 0.2346 a směrodatnou odchylkou 0.01456. Koncentraci glukózy je vyčíslena vztahem *koncentrace = faktor × absorbance*. Faktor byl určen dříve a jeho střední hodnota je 23.5 $mmol.l^{-1}$. Ukažte, jaký vliv na variabilitu koncentrace glukózy má variabilita faktoru za předpokladu, že výsledky stanovení faktoru mají normální rozdělení. Výpočet proveďte pro tyto případy: (a) $\bar{x}$ = 23.5, s = 0.0235, (b) $\bar{x}$ = 23.5, s = 0.235.

Úloha B1.02 *Nejistota stanoveného obsahu merkuriamidochloridu*
Stanovení obsahu merkuriamidochloridu v "medicinálním precipitátu" je založeno na reakci

$$HgNH_2Cl + 4\ KI + H_2O \rightarrow K_2(HgI_4) + KCl + KOH + NH_3$$

a uvolněný KOH a NH_3 se titrují odměrným roztokem kyseliny chlorovodíkové. Vypočtěte procentuální obsah $HgNH_2Cl$ ve vzorku navážky x_4 = 0.2085 (0.0003) g, jestliže bylo spotřebováno x_1 = 16.20 (0.03) ml x_2 = 0.1 (0.001) M HCl a molekulová hmotnost $HgNH_2Cl$ je x_3 = 252.06 (0.00001). K výpočtu se doporučuje použít vztahu $c[\%] = x_1 x_2 x_3/(20x_4)$.

1.5.2 Analýza chemických a fyzikálních dat

Úloha C1.01 *Výpočet nejistoty koncentrace roztoku chloridu sodného*
V 500 (0.12) cm^3 roztoku je rozpuštěno 16.0000 (0.0003) g NaCl o molekulové hmotnosti 58.44 (0.00001). Jaká je rozšířená nejistota koncentrace (mol. dm^{-3}) roztoku?

Úloha C1.02 *Výpočet nejistoty koncentrace roztoku modré skalice*
V 1000 (0.2) ml roztoku je rozpuštěno 12.500 (0.0003) g $CuSO_4 . 5H_2O$ o molekulové hmotnosti 249.686 (0.01). Jaká je rozšířená nejistota koncentrace (mol. dm^{-3}) roztoku?

Úloha C1.03 *Výpočet nejistoty koncentrace roztoku chloridu barnatého*
V 800 (0.15) ml roztoku je rozpuštěno 39.08 (0.0003) g $BaCl_2 . 2H_2O$ o molekulové hmotnosti 244.28 (0.01). Jaká je rozšířená nejistota koncentrace (mol. dm^{-3}) roztoku?

Úloha C1.04 *Výpočet nejistoty koncentrace nasyceného roztoku síranu stříbrného*
Vypočtěte rozšířenou nejistotu koncentrace nasyceného roztoku síranu stříbrného při 18 EC, jestliže roztok obsahuje 0.58 (0.005) % Ag o atomové hmotnosti 107.868 (0.01).

Úloha C1.05 *Výpočet nejistoty koncentrace roztoku uhličitanu sodného*
Určete rozšířenou nejistotu koncentrace (mol. dm^{-3}) roztoku uhličitanu sodného, je-li v 500 ml (0.12) rozpuštěno 8.0 (0.0003) g Na_2CO_3 o molekulové hmotnosti 105.99 (0.01).

Úloha C1.06 *Výpočet nejistoty koncentrace hydroxidu draselného stanovené acido-bazickou titrací*

Určete rozšířenou nejistotu koncentrace (mol. dm^{-3}) hydroxidu draselného KOH užitého při zpětné titraci, jestliže se navážka 0.2580 (0.0003) g $CaCO_3$ rozpustila v 50 (0.02) ml roztoku 0.2046 (0.0001) M HCl a přebytečná kyselina byla neutralizována 20.0 (0.03) ml KOH.

Úloha C1.07 *Nejistota obsahu fosforečnanu sodného určeného acidobazickou titrací*
Určete průměrnou hodnotu, rozptyl a rozšířenou nejistotu procentického obsahu terciárního fosforečnanu sodného, jestliže se na navážku 0.5629 (0.0003) g Na_3PO_4 spotřebovalo acidobazickou titrací na methyloranž 26.93 (0.03) ml roztoku 0.0977 (0.001) M HCl. Molekulová hmotnost Na_3PO_4 je 163.9408 (0.0001).

Úloha C1.08 *Výpočet nejistoty koncentrace fluoridu sodného v kuchyňské soli*
Kuchyňská sůl, NaCl, je obohacována přídavkem fluoridu sodného NaF. Jeho obsah je laboratorně kontrolován. V připraveném roztoku soli se proměří fluoridovou iontově selektivní elektrodou ISE koncentrace fluoridů a přepočte se vzhledem k navážce, z níž byl roztok připraven, dle vzorce c_F = [F] 1000/m_s. Byl užit vzorek navážky m_s = 5.019 (0.0003) g. Pomocí ISE elektrody byla stanovena koncentrace fluoridových iontů [F] = 1.3 (0.1) mg. l^{-1}. Jaká je koncentrace fluoridů a její rozšířená nejistota v původním vzorku?

Úloha C1.09 *Výpočet nejistoty koncentrace oxidu fosforečného*
Při přípravě standardního roztoku, obsahujícího v 1 ml 0.01 mg oxidu fosforečného P_2O_5 se postupuje následujícím postupem: po vysušení při 105 EC navážíme 1.9175 (0.0003) g KH_2PO_4 p. a., který rozpustíme v odměrné baňce 1000 (0.4) ml destilovanou vodou a doplníme po rysku. Z odměrné baňky odpipetujeme 10 (0.02) ml a naředíme na 1000 (0.4) ml. Určete průměrnou koncentraci standardního roztoku P_2O_5 a její nejistotu.

Úloha C1.10 *Výpočet nejistoty výsledku stanovení obsahu nerozpuštěných látek*
Ke gravimetrickému stanovení obsahu nerozpuštěných látek bylo použito V_0 = 100 (0.5) ml vody. Hmotnost filtru s nerozpuštěnými látkami byla m_2 = 27.1230 (0.0003) mg a hmotnost samotného filtru m_1 = 27.1214 (0.0003) mg. Obsah nerozpuštěných látek se vypočte dle vzorce $c = 10^6 (m_2 - m_1)/V_0$. Odhadněte rozšířenou nejistotu obsahu nerozpuště-ných látek.

Úloha C1.11 *Výpočet nejistoty obsahu volného kyanidu v kyanidové lázni*
Při titraci 50 (0.02) ml stříbřící kyanidové lázně se na volný kyanid spotřebovalo 31.40 (0.03) ml 0.0875 (0.0005) M $AgNO_3$. Vypočtěte nejistotu obsahu volného kyanidu draselného v gramech na 1000 (0.2) ml kyanidové lázně, když molekulová hmotnost KCN je 65.120 (0.001).

Úloha C1.12 *Výpočet nejistoty obsahu kyanidu draselného při argentometrické titraci*
Určete nejistotu procentního obsahu KCN v neznámém vzorku, když se na navážku 0.3826 (0.0003) g vzorku spotřebovalo při argentometrické titraci 27.18 (0.03) ml roztoku 0.09633 (0.001) M $AgNO_3$. Molekulová hmotnost KCN je 65.12 (0.00001).

Úloha C1.13 *Výpočet nejistoty rozpustnosti olovnatých iontů*
Součin rozpustnosti fosforečnanu olovnatého je $K_S = 8.0 \cdot 10^{-43}$ ($0.8 \cdot 10^{-43}$). Jaká je nejistota

vypočtené rovnovážné koncentrace olovnatých iontů [Pb^{2+}] v nasyceném roztoku $Pb_3(PO_4)_2$ ve vodě?

Úloha C1.14 *Stanovení nejistoty obsahu sulfidu barnatého jodometricky*
Při jodometrickém stanovení obsahu BaS v uhličitanu barnatém se vypočte procento BaS v neznámém vzorku dle vzorce % BaS = 0.84702 $(V_1\, c_1 - V_2\, c_2)/m$, kde na vzorek uhličitanu navážky m = 1.9958 (0.0003) g byl objem V_1 = 10.00 (0.04) ml roztoku jodu o koncentraci c_1 = 1.1460 (0.0008) mol. dm^{-3} a spotřeby V_2 = 9.75 (0.05) ml roztoku sirnatanu o koncentraci c_2 = 1.0312 (0.0002) mol. dm^{-3}. Jaká je nejistota obsahu sulfidu barnatého?

Úloha C1.15 *Výpočet nejistoty obsahu měďnatých iontů chelatometrickou titrací*
Na 50 (0.02) ml vzorku síranu měďnatého se spotřebovalo 28.6 (0.01) ml 0.05002 (0.00001) M chelatonu 3. Vypočtěte, jaká je rozšířená nejistota počtu gramů mědi, obsažených v 1000 (0.02) ml roztoku síranu měďnatého, když atomová hmotnost mědi je 63.546 (0.0001).

Úloha C1.16 *Výpočet nejistoty koncentrace chelatonu 3 titrací zinečnatou solí*
Navážka 5.00 (0.0003) g čistého zinku byla rozpuštěna v HCl a zředěna 500 (0.12) ml. Na titraci 25.0 (0.03) ml tohoto roztoku bylo spotřebováno 58.0 (0.05) ml chelatonu 3. Jaká je nejistota koncentrace chelatonu 3 (mol. dm^{-3}), když atomová hmotnost zinku je 65.39 (0.05)?

Úloha C1.17 *Výpočet nejistoty koncentrace dusitanů v 70% kyselině sírové*
Stanovení dusitanů v 70% kyselině sírové se provádí permanganometricky na kyselinu šťavelovou. Koncentrace dusitanů se vypočte dle vztahu c = 1.172 $(V_1\, c_1 - V_2\, c_2)$, kde V_1 = 7.0 (0.150) ml je objem přebytku $KMnO_4$ o koncentraci c_1 = 0.9973 (0.0008) mol. dm^{-3}, V_2 = 1.3 (0.03) ml je objem spotřebované kyseliny šťavelové o koncentraci c_2 = 1.0072 (0.0008) mol. dm^{-3}. Vypočtěte nejistotu koncentrace dusitanů.

Úloha C1.18 *Výpočet nejistoty obsahu hořčíku metodou AAS*
Určete nejistotu obsahu hořčíku ve vzorku tvárné slitiny metodou AAS dle vztahu

$$\%\ \mathrm{Mg} = 100\ \%\ (m\, V_1\, A)/(V_2\, V_3),$$

kde m = 1.000 (0.0003) g je hmotnost navážky vzorku, A = 0.475 (0.0043) je naměřená absorbance, V_1 = 100.0 (0.085) ml je objem odměrky při rozpouštění vzorku, V_2 = 10.0 (0.014) ml je pipetovaný objem, V_3 = 100.0 (s = 0.085) ml je objem odměrky při dalším ředění.

Úloha C1.19 *Nejistota titračního stanovení chloridů ve vodě*
Vypočtěte nejistotu výsledku titračního stanovení chloridů ve vodě, když hmotnostní koncentrace chloridů ve vzorku se vypočte podle vzorce

$$c_m[\mathrm{Cl}^-] = \frac{V_e\, f_t\, c[\mathrm{Hg(NO_3)_2}]\ 1000}{V_0},$$

kde hmotnostní koncentrace chloridů ve vzorku $c_m[Hg(NO_3)_2]$ se vypočte podle vzorce

$$c_{\mathrm{m}}[\mathrm{Hg(NO_3)_2}] \ ' \ \frac{m[\mathrm{Hg(NO_3)_2 \ . \ 1/2\,H_2O}]}{V \ . \ M[\mathrm{Hg(NO_3)_2 \ . \ 1/2\,H_2O}]} \ ,$$

kde hmotnost navažované látky $m[\mathrm{Hg(NO_3)_2 \ . \ ½\ H_2O}]$ = 8.5 g (0.001), objem výsledného roztoku V = 1 litr (0.0002), molekulová hmotnost $M[\mathrm{Hg(NO_3)_2 \ . \ ½\ H_2O}]$ = 333.61 g/mol (0.01), spotřeba odměrného roztoku dusičnanu rtuťnatého při titraci vzorku V_e = 5.00 ml (0.05), titrační přepočítávací faktor pro merkurimetrické stanovení chloridů f_t = 2 (0.000001), původní objem vzorku při titraci V_0 = 100.0 ml (0.5), atomová hmotnost chloru A_{Cl} = 35.453 g/mol.

Úloha C1.20 *Porovnání propagované nejistoty a směrodatné odchylky*
Porovnejte nejistotu výsledku spektrofotometrického stanovení amonných iontů ve vodě, určenou technikou propagace chyb se směrodatnou odchylkou, vyčíslenou opakovaným stanovením vzorku. Koncentrace amonných iontů se vypočte pomocí vzorce $c_m\ (\mathrm{NH_4^{\%}}) \ ' \ k\ A$, kde hmotnostní koncentrace amonných iontů ve vzorku se značí $c_m(\mathrm{NH_4^+})$ [mg. $\mathrm{l^{-1}}$], konstanta kalibrační křivky k = 0.73 mg. $\mathrm{l^{-1}}$ (0.01) a absorbance A = 0.849 (0.005). Opakovaným stanovením koncentrace $\mathrm{NH_4^+}$ ve vzorku odpadních vod byla zjištěna následující data: 0.63, 0.62, 0.61, 0.62, 0.62, 0.63, 0.61, 0.62, 0.63 mg. $\mathrm{l^{-1}}$.

Úloha C1.21 *Výpočet nejistoty koncentrace kyseliny chlorovodíkové acidobazickou titrací*
Vypočtěte nejistotu stanovené koncentrace kyseliny chlorovodíkové HCl, která je titrována čerstvě připraveným roztokem hydroxidu sodného NaOH, standardizovaného na hydrogenftalan draselný KHP. Předpokládá se úroveň koncentrace HCl okolo 0.1 mol. $\mathrm{l^{-}}$. Koncentrace HCl v mol. $\mathrm{l^{-1}}$ se vyčíslí vztahem

$$c_{\mathrm{HCl}} \ ' \ \frac{1000 \ m_{\mathrm{KHP}} \ P_{\mathrm{KHP}} \ V_{\mathrm{KHP}} \ V_{\mathrm{N2}}}{V_{\mathrm{f}} \ F_{\mathrm{KHP}} \ V_{\mathrm{N1}} \ V_{\mathrm{HCl}}} \ ,$$

kde navážka KHP m_{KHP} = 5.1050 g (variační koeficient $\delta = 1.7 \times 10^{-5}$), čistota KHP je P_{KHP} = 0.999 ($\delta = 5.8 \times 10^{-4}$), spotřeba KHP je V_{KHP}= 24.85 ml ($\delta = 1.3 \times 10^{-3}$), objem NaOH pro titraci KHP je V_{N2} = 25.0 ($\delta = 8.4 \times 10^{-4}$), objem zásobního roztoku KHP je V_f = 250.0 ($\delta = 4.8 \times 10^{-4}$), molekulová hmotnost KHP je F_{KHP} = 204.2236 ($\delta = 2.3 \times 10^{-5}$), objem NaOH pro titraci HCl je V_{N1} = 25.0 ($\delta = 8.4 \times$ 10^{-4}), spotřeba HCl je V_{HCl} = 25.45 ml ($\delta = 1.3 \times 10^{-3}$). Z hodnot relativních směrodatných odchylek čili variačních koeficientů je zřejmé, že oba objemy V_{KHP} a V_{HCL} přispívají kcelkové nejistotě každý 1.3×10^{-3}. Jsou-li tyto dvě hodnoty kombinovány, činí příspěvek 1.8×10^{-3}, což odpovídá 80% odhadované nejistoty.

1.5.3 Analýza environmentálních, potravinářských a zemědělských dat

Úloha E1.01 *Výpočet nejistoty obsahu vody v etylacetátu Fischerovou metodou*
Určete velikost nejistoty při stanovení obsahu vody ve vzorku etylacetátu Fischerovou metodou v hmotnostních procentech z následujících dat: navážka etylacetátu m = 1.4021 (0.0003) g, spotřeba Fischerova činidla V = 0.108 (0.001) ml, titr t = 4.5797 (0.001) mg/ml, faktor f = 0.1.

Úloha E1.02 *Nejistota chromatografického stanovení*
Vypočtěte nejistotu chromatografického GC/MS stanovení obsahu referenčního kongeneru polychlorovaných bifenylů PCB-28 (klasifikace dle Ballschmidtera) ve vodním vzorku v oblasti meze stanovitelnosti pro následující experimentální data: objem vzorku vody pro extrakci PCB V = 1000 ± 0.2 ml, zakoncentrování vzorku na v = 100 ± 2 µl po jeho extrakci a čištění, minimální stanovitelná plocha GC/MS píku P = 255000 ± 45000, plocha chromatografického píku odpovídající 2 ng/µl referenčního standardu PCB-28 činí S = 3750000 ± 220000.

Úloha E1.03 *Výpočet nejistoty obsahu kyseliny octové v konzumním octu*
Při titraci x_1 = 0.7886 (0.0003) g hydrogenftalanu draselného $KHC_8H_4O_4$ se spotřebovalo x_4 = 27.12 (0.03) ml NaOH. Molekulová hmotnost hydrogenftalanu draselného je x_5 = 204.23 (0.001) a kyseliny octové CH_3COOH x_2 = 60.053 (0.0001). Vypočtěte nejistotu objemového procenta kyseliny octové v konzumním octu, když se na x_6 = 10.0 (0.005) ml konzumního octa spotřebovalo x_3 = 49.35 (0.03) ml odměrného roztoku NaOH. Hustota x_8 = 100% kyseliny octové je x_7 = 1.0498 g. cm^{-3}. Vztah je $Z = (x_1\ x_2\ x_3\ x_8)/(x_4\ x_5\ x_6\ x_7)$.

1.5.4 Analýza hutnických a mineralogických dat

Úloha H1.01 *Stanovení nejistoty množství kadmia v hmotě keramického nádobí*
U zkoušky ke stanovení množství olova a kadmia, tj. množství vyluhovaného z povrchu keramického nádobí 4% vodným roztokem kyseliny octové se užívá metoda atomové absorpční spektrometrie. U nádob, které je možné zcela naplnit loužícím roztokem, je množství vylouženého kovu vyjádřeno jako koncentrace c_0 [mg. l^{-1}] loužícího roztoku. U nádob, které není možno zcela zaplnit, norma vyžaduje, aby výsledek byl vyjádřen jako množství kadmia či olova r, vyloužené z jednotky povrchu dle vztahu $r = c_0\ V_L/a_V$, kde c_0 je vypočtená koncentrace ve výluhu, V_L je objem loužícího roztoku a a_V je povrch nádoby. Koncentrace je stanovena za užití dvou standardních roztoků. První roztok má koncentraci kovu nižší než je očekávaná měřená koncentrace a druhý vyšší. Výraz pro výpočet koncentrace c_0 je

$$c_0 = \left\{\frac{A_0 - A_1}{A_2 - A_1}\left[c_2 - \frac{c_2}{f_5}\right] + \frac{c_2}{f_5}\right\} d\, f_{\text{kys}}\, f_{\text{cas}}\, f_{\text{tepl}} \quad ,$$

kde A_0 je optická hustota kovu ve výluhu vzorku 53.0 (0.62), kde A_1 je absorbance kovu ve

standardu s nižší koncentrací 21.8 (0.39), kde A_2 je absorbance kovu ve standardu s vyšší koncentrací 101.4 (0.22). Absorbance byly měřeny opakovaně 10krát a v datech jsou uvedeny aritmetické průměry a směrodatné odchylky průměrné hodnoty. Roztoky standardů byly připraveny postupným ředěním zásobního roztoku: c_0 je obsah kovu ve výluhu vzorku, c_1 je obsah kovu ve standardu s nižší koncentrací c $_2$5 = 0.1 (0.0017) mg. l^{-1}, c_2 je obsah kovu ve standardu s vyšší koncentrací 0.5 (0.0017) mg. l^{-1}, d zřeďovací koeficient 1.000 (0.000), f_{kys} je koeficient přípravy 4% kyseliny octové 1.0 (0.0064) %, f_{cas} koeficientu času potřebného na vyloužení kovu z keramiky, značí průměrnou změnu koncentrace asi 0.3 %/h což činí korekci na c_0 o hodnotu 1 ± (c_2 × 0.003) = 1 ± 0.0015. Vyluhovaní kovu z keramiky je ovlivněno teplotou. Pro normou dovolený rozsah 2 EC byl získán koeficient f_{tepl} = 1 ± 0.1, po převedení na směrodatnou odchylku s = 0.1/%3 = 0.06. Dále množství kovu r [mg/dm^2], vyluhovaného jednotkou povrchu a_V = 2.37 (0.069) dm^2 je objemem kyseliny octové V_L = 332 (0.007) ml.

1.5.5 Analýza fyzikálních dat

Úloha S1.01 *Výpočet nejistoty výsledku u operací s přibližnými čísly*
Určete rozšířenou nejistotu U výsledku a zaokrouhlete výsledek na správný počet platných desetinných míst u následujících výrazů:
(a) $y = 6.75(\pm 0.03) - 0.843(\pm 0.001) - 7.021(\pm 0.001)$,
(b) $y = 67.1(\pm 0.3) \times 1.03(\pm 0.02) \times 10^{-17}$,
(c) $y = (143(\pm 6) - 64(\pm 3))/(1249(\pm 1) - 77(\pm 8))$,
(d) $y = \log(6.02(\pm 0.02) \times 10^{23})$,
(e) $y = \text{antilog}(0.99(\pm 0.05))$.

Úloha S1.02 *Výpočet nejistoty výsledného odporu v sériovém zapojení*
Vypočtěte celkový odpor R spolu s jeho nejistotou při sériovém zapojení čtyř odporů, u kterých známe jejich relativní chybu (v závorce v %): R_1 = 100.12 (0.01 %) ohmů, R_2 = 249.61 (0.008 %) ohmů, R_3 = 1001.2 (0.01 %) ohmů, R_4 = 10003.0 (0.01 %) ohmů.

Úloha S1.03 *Nejistota kalibrace teploměru*
Při kalibraci teploměru byl posuzován vliv ocelové jímky tak, že během zahřívání kalibrační lázně byla sledována dynamická nejistota správně měřicího teploměru. V jistém okamžiku bylo zjištěno, že během 20 s se zvýšil údaj teploměru o 0.4 EC, teplota lázně je 67.0 EC a údaj teploměru je 65.6 EC. Určete časovou konstantu teploměru v jímce, víte-li, že teplota byla měřena s nejistotou 0.1 EC; nejistotu měření času můžeme vzhledem k automati-zovanému sběru dat zanedbat a pro změnu údaje teploměru platí vztah

$$\frac{\mathrm{d}\zeta}{\mathrm{d}t} = \frac{1}{\tau}(\zeta_l - \zeta_t) \quad ,$$

kde ζ_t vyjadřuje údaj teploměru, ζ_l teplotu lázně, t čas a τ časovou konstantu.

Úloha S1.04 *Výpočet nejistoty meze skluzu*
Fyzikální parametr mez skluzu se stanovuje při tahové zkoušce tělíska, zhotoveného podle

ISO normy: mikrometrem se změří tlouštka d_1 = 1.870 (0.001) a šířka d_2 = 6.130 (0.001) pracovní části zkušebního tělíska v 10 bodech a průměrná hodnota se užije pro výpočet průřezu tělíska. Pak se tělísko upne do čelistí tahového stroje a po nastavení předepsané rychlosti vzdalování pohyblivé části (50 mm/min) a příslušné síly (1000 nebo 2000 N) se provede zkouška. Ze záznamu křivky zatížení - protažení - se odečetla mez skluzu Z = 278 (0.005) N v bodě maxima. Vedle meze skluzu vypočtěte i její nejistotu. Mez skluzu se přepočte dle vztahu $MK = Z/(d_1 \,.\, d_2)$.

1.6 Kontrolní hodnoty (ADSTAT, NCSS2000)

1.6.1 Analýza farmakologických a biochemických dat

B1.01 $\bar{x}$ = *XXXXX* .

B1.02 $\bar{x}$ = *97.90 [%], s = 1.01 [%]* .

1.6.2 Analýza chemických a fyzikálních dat

C1.01 $\bar{x}$ = *0.5476, s = 1.32E-4.*

C1.02 $\bar{x}$ = *0.05008, s = 1.01E-5.*

C1.03 $\bar{x}$ = *0.200, s = 3.8E-5.*

C1.04 $\bar{x}$ = *2.688E-5, s = 2.32E-7.*

C1.05 $\bar{x}$ = *0.1510, s = 3.67E-5.*

C1.06 $\bar{x}$ = *0.2537, s = 1.30E-2.*

C1.07 $\bar{x}$ = *25.543, s = 0.263.*

C1.08 $\bar{x}$ = *260 mg/kg, s = 20.0 mg/kg.*

C1.09 $\bar{x}$ = *0.010 g/dm^3 P_2O_5, s = 2.08E-5.*

C1.10 $\bar{x}$ = *16.0 mg, s = 4.2.*

C1.11 $\bar{x}$ = *3.578, s = 2.079E-2.*

C1.12 $\bar{x}$ = *44.59, s = 0.47.*

C1.13 $\bar{x}$ = *4.474E-9, s = 0.0897E-9.*

C1.14 $\bar{x}$ = *5.97E-4, s = 2.9E-5.*

C1.15 $\bar{x}$ = *1.818, s = 1.094E-3.*

C1.16 $\bar{x}$ = *6.594E-2, s = 9.88E-5.*

C1.17 $\bar{x}$ = *6.647E-3, s = 1.79E-4.*

C1.18 $\bar{x}$ = *4.75E-2, s = 4.38E-6.*

C1.19 $\bar{x}$ = *90.33, s = 1.01, δ = 1.12 %.*

C1.20 $\bar{x}$ = *0.62 mg. l^{-1}, s = 0.01 mg. l^{-1} (z opakovaných hodnot),* $\bar{x}$ = *0.62 mg. l^{-1}, s = 0.01 mg. l^{-1}.*

C1.21 $\bar{x}$ = *0.0975 mol. l^{-1}, s = 0.00022 mol. l^{-1} (= u(c_{HCl})), δ = 0.23 %, U(c_{HCl}) = ± 0.00044 mol. l^{-1}.*

1.6.3 Analýza environmentálních, potravinářských a zemědělských dat

E1.01 $\bar{x}$ *= 3.53E-2, s = 3.28E-4.*

E1.02 $\bar{x}$ *= 0.0136 ppb, s = 2.54E-3 ppb.*

E1.03 $\bar{x}$ *= 4.0213 obj. %, s = 0.00567.*

1.6.4 Analýza hutnických a mineralogických dat

H1.01 $c_0 = 0.26$ mg. l^{-1}, $s = 0.0035$ mg. l^{-1}, $r = 0.036$ mg/dm^2, $s = 0.0024$ mg/dm^2,
$\boldsymbol{U}(r) = 2 \times 0.0024 = \pm 0.0048$ mg/dm^2.

1.6.5 Analýza fyzikálních dat

S1.01 *(a)* 0.57 ± 0.03, *(b)* $(6.9 \pm 0.1) \times 10^{-16}$, *(c)* $(6.0 \pm 0.5) \times 10^{-2}$, *(d)* 23.800 ± 0.001, *(e)* 10 ± 1.

S1.02 $R = 11354 \pm 1$ ohm.

S1.03 $\bar{x}$ *= 74.38, s = 18.88.*

S1.04 $\bar{x}$ *= 24.253, s = 0.0136.*

1.7 Doporučená literatura

[1] Taylor J. R.: *An Introduction to Error Analysis.* University Science Books, Mill Valey, California 1982.

[2] Lyon A. J.: *Dealing with Data*, Pergamon Press, London 1970.

[3] Zelený F.: *Základní vlastnosti měřicích přístrojů*, SNTL Praha 1976.

[4] Novickij P. V., Zograf I. A.: *Oceňka pogrešnostej rezultatov izmerenij*. Atomizdat, Moskva 1985.

[5] Hahn G. J., Nelson W.: *Technometrics* **12**, 95 (1970).

[6] Mandel J.: *The Statistical Analysis of Experimental Data*, Interscience, New York 1964.

[7] Manly B. F. J.: *Biom. J.* **28**, 949 (1986).

[8] Müller J. W.: *Nucl. Instr. Meth.* **163**, 241 (1979).

[9] Schwartz L. M.: *Anal. Chem.* **47**, 963 (1975).

[10] Shapiro S. S., Gross A. J.: *Statistical Modelling Techniques*. Marcel Dekker Inc., New York 1981.

[11] *Quantifying Uncertainty in Analytical Measurement,* EURACHEM 1995.

[12] Taylor B., Kuyatt CH. E. : *Guidelines for Evaluation and Expressing the Uncertainty of NIST Measurement Results,* NIST Tech. Note 1297, 1994.

[13] Agostini D. G.: *Probability and Measurement Uncertainty in Physic,* Rept. DESY 95-242, Roma December 1995.

[14] Phillips S. D., Eberhart K. R., Parry B.: *Guidelines for Expressing the Uncertainty of Measurement Results Containing Uncorrected Bias,* J. Res. Natl. Inst. of Standards **102**, 577 (1997).

[15] Meloun M., Militký J., Forina M.: *Chemometrics for Analytical Chemistry, Volume 1,* Ellis Horwood, Chichester, 1992.

[16] Meloun M., Militký J.: *Statistické zpracování experimentálních dat,* Plus Praha 1994 (1. vydání), EAST PUBLISHING, Praha 1998 (2. vydání).

[17] Elishakoff I.: *Convex Modeling - a Generalization of Interval Analysis for Non-*

probabilistic Treatment of Uncertainty, Proc. Int. Conf. APIC 95, El Paso, 1995 (a supplement to the international Journal of Reliable Computing).
[18] Ratschek, H. : *SIAM J. Numer. Anal.* **17,** 656 (1980).

2
PRŮZKUMOVÁ ANALÝZA JEDNOROZMĚRNÝCH DAT

Experimentální data se v analytické laboratoři často vyznačují nekonstantním rozptylem, malým počtem, asymetrickým rozdělením a porušením základních předpokladů, kladených na výběr. Uveďme nejprve 3 etapy obecné osnovy analýzy výběru dat.

A. V průzkumové analýze dat se vyšetřují *statistické zvláštnosti dat,* jako je lokální koncentrace dat, tvarové zvláštnosti rozdělení dat a přítomnost podezřelých hodnot. Odhalí se také anomálie a odchylky rozdělení výběru od typického rozdělení, obyčejně normálního (Gaussova). Interaktivní statistická analýza na počítači tento postup ulehčuje, většina statistického softwaru nabízí řadu diagnostických grafů a diagramů. Pokud je rozdělení dat nevhodné pro standardní statistickou analýzu (tj. většinou je asymetrické), provádí se nejprve vhodná úprava dat. Pokud bylo indikováno zešikmené rozdělení nebo rozdělení s dlouhými konci, pomocníkem je mocninná a Boxova-Coxova transformace. Transformace je vhodná především při asymetrii rozdělení původních dat, ale také při nekonstantnosti rozptylu.

B. Pro případ rutinních měření se ověří *základní předpoklady,* kladené na výběr, jako jsou nezávislost prvků, homogenita výběru, dostatečný rozsah výběru a rozdělení výběru. Jsou-li závěry tohoto kroku optimistické, následuje vyčíslení klasických odhadů polohy a rozptýlení, tj. obyčejně aritmetického průměru a rozptylu. Dále se vyčíslí intervaly spolehlivosti, následované testováním statistických hypotéz. V pesimistickém případě následuje další pokus o úpravu dat.

C. V konfirmatorní analýze je nabízena paleta rozličných odhadů polohy, rozptýlení a tvaru, jež lze rozdělit do dvou skupin: na *klasické odhady* a na *robustní odhady* (necitlivé na odlehlé prvky výběru, resp. další předpoklady o datech). Z nabídky odhadů parametrů vybírá uživatel uvážlivě ty, jež mají statistický smysl a odpovídají závěrům průzkumové analýzy dat a ověření předpokladů o výběru.

Postup statistické analýzy jednorozměrných dat[1, 3, 5], prováděné v interaktivním režimu na počítači, lze shrnout do bloků, i když lze jednotlivé bloky provádět samostatně: blok A, blok B, blok A+B, blok B+C a konečně všechny bloky A+B+C.

Přehled operací analýzy jednorozměrných dat

A. Průzkumová (exploratorní) analýza dat (EDA):
- Odhalení stupně symetrie a špičatosti výběrového rozdělení;
- Indikace lokální koncentrace výběru dat;
- Nalezení vybočujících a podezřelých prvků ve výběru;
- Porovnání výběrového rozdělení dat s typickými rozděleními;
- Mocninná transformace výběru;
- Boxova-Coxova transformace výběru.

B. Ověření předpokladů o datech:
- Ověření nezávislosti prvků výběru;
- Ověření homogenity rozdělení výběru;
- Určení minimálního rozsahu výběru;
- Ověření normality rozdělení výběru.

C. Konfirmatorní analýza dat (CDA) - odhady parametrů (polohy, rozptýlení a tvaru)
1. Klasické odhady (bodové a intervalové) z výběru;
2. Robustní odhady (bodové a intervalové) z výběru.

Vzorová úloha 2.1 *Analýza dat normálního a logaritmicko-normálního rozdělení.*

Analýza simulovaných dat výběru, pocházejícího z (a) rozdělení normálního ***norm*** *N(10, 0.1)* a z (b) rozdělení logaritmicko-normálního ***log***. *l(5, 2)*.

Data:

(a) Výběr ***norm****:*

10.0010	9.9290	10.0370	9.9490	10.1850	9.9590	10.0630	9.8790	10.0500	9.8460
...	...	...	...	...	...	...	...	...	...
10.0370	10.0110	9.9310	9.9870	9.9550	10.0130	10.0020	10.1150	10.0250	

(b) Výběr ***log*****:**

2.408	**5.389**	**2.259**	**2.439**	**2.173**	**1.157**	**0.892**	**0.498**	**0.351**	**1.229**
...	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**
2.816	**0.666**	**4.972**	**0.451**	**1.316**	**3.241**	**0.316**	**2.200**	**8.291**	**0.815**

Řešení: u jednotlivých diagnostických diagramů a grafů budou uvedeny vždy dvě ukázky, jednak pro výběr ze symetrického normálního rozdělení *norm N*(10, 0.1) a jednak pro výběr z asymetrického logaritmicko-normálního rozdělení *log LN*(5, 2). Čtenář může porovnat, jak jednotlivé diagnostické pomůcky monitorují symetrické a silně asymetrické rozdělení.

2.1 Průzkumová (exploratorní) analýza dat EDA

Prvním krokem v analýze jednorozměrných dat je průzkumová čili exploratorní analýza. Vychází se z *pořádkových statistik* výběru, tj. z prvků výběru uspořádaných vzestupně $x_{(1)} \# x_{(2)} \# \ldots \# x_{(n)}$. Platí, že střední hodnota *i*-té pořádkové statistiky $E(x_{(i)})$ je rovna 100 P_i procentnímu kvantilu výběrového rozdělení $Q(P_i)$ a symbol $P_i \ . \ i/(n + 1)$ označuje *pořadovou pravděpodobnost*. Připomeňme, že $100P_i$ *procentní výběrový kvantil* je hodnota, pod kterou leží 100 P_i procent prvků výběru. Vynesením hodnot $x_{(i)}$ proti P_i, $i = 1, \ldots, n$, se získá hrubý odhad *kvantilové funkce* $Q(P_i)$. Ta je inverzní k *funkci distribuční* $F(x_i)$ a charakterizuje jednoznačně rozdělení výběru. Pro libovolnou hodnotu α z intervalu [0, 1] lze vyčíslit 100α% kvantil $\tilde{x}_\alpha$ pomocí lineární interpolace

$$x_\alpha \ ' \ (n \ \% \ 1) \left(\alpha \ \& \ \frac{i}{n \ \% \ 1} \right) (x_{(i\%1)} \ \& \ x_{(i)}) \ \% \ x_{(i)} \ ,$$

kde pro index *i* musí být splněna nerovnost $\frac{i}{n \ \% \ 1} \ \# \ \alpha \ \# \ \frac{i \ \% \ 1}{n \ \% \ 1}$. Pro rozptyl kvantilu $\tilde{x}_\alpha$, určeného z výběru velikosti *n*, platí vztah $D(\tilde{x}_\alpha) \ ' \ \frac{\alpha \ (1 \ \& \ \alpha)}{n \ [f(\tilde{x}_\alpha)]^2}$, kde $f(x_\alpha)$ je výběrová hodnota hustoty pravděpodobnosti v bodě x_α.

V průzkumové analýze se často používá speciálních *kvantilů L* pro pořadové pravděpodobnosti $P_i = 2^{-i}$, $i = 1, 2, \ldots$, které se také nazývají *písmenové hodnoty*.

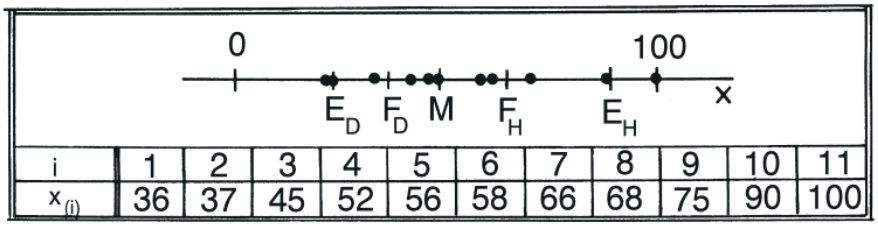
i	*i*-tý kvantil	Pořadová pravděpodobnost P_i	Písmenová hodnota *L*
1	Medián	$2^{-1} = 1 / 2$	*M*
2	Kvartily	$2^{-2} = 1 / 4$	*F*
3	Oktily	$2^{-3} = 1 / 8$	*E*
4	Sedecily	$2^{-4} = 1 / 16$	*D*

Symbol u_{P_i} označuje kvantil normovaného normálního rozdělení $N(0, 1)$. Kromě *mediánu* ($i = 1$) existují pro každé $i > 1$ dvojice kvantilů, a to *dolní* a *horní písmenová hodnota* L_D a L_H. Dolní písmenová hodnota je pro pořadovou pravděpodobnost $P_i = 2^{-i}$, zatímco horní je pro $P_i = 1 - 2^{-i}$.

Pro odhad písmenových hodnot lze použít jednoduché techniky *pořadí* a *hloubek*. Pořádková statistika $x_{(i)}$ má *rostoucí pořadí* $R_{P_i} = i$ a *klesající pořadí* $K_{P_i} = n + 1 - i$. *Hloubka* H_i je pak menší číslo z obou pořadí $H_i \ ' \ \min(R_{P_i}, K_{P_i})$. Pro hloubku mediánu platí $H_M \ ' \ (n \ \% \ 1)/2$. Pokud je tato hloubka celé číslo, je medián $\tilde{x}_{0.5} \ ' \ M \ ' \ x_{(H_M)}$. V opačném případě se provádí lineární interpolace mezi $x_{(n/2)}$ a $x_{(n/2+1)}$. Hloubky dolních písmenných hodnot jsou $H_L \ ' \ [1 \ \% \ \mathrm{int}(H_{L\&1})]/2$, kde *L* jsou indexy *F, E, D* a int(*x*) značí celočíselnou část čísla *x*. Pokud je $L = F$, bere se $L - 1 = M$. Jestliže je H_L celé číslo, bude dolní kvantil $L_D = x_{(H_L)}$ a horní kvantil $L_H = x_{(n+1-H_L)}$. Je-li H_L číslo necelé, provádí se lineární interpolace podle vztahů

$$L_D = \frac{x_{\mathrm{int}(H_L)} + x_{\mathrm{int}(H_L)+1}}{2} \quad \text{a} \quad L_H = \frac{x_{n+1-\mathrm{int}(H_L)} + x_{n+2-\mathrm{int}(H_L)}}{2}.$$

Tento postup se pro menší hodnoty H_L, kdy jsou kvantily blízko hodnot $x_{(1)}$ a $x_{(n)}$, považuje za robustnější. Počet písmenových hodnot závisí na rozsahu výběru. Pro velikost výběru n ze určit n_L písmenových hodnot včetně mediánu. Platí, že $n_L = 1.44 \ln(n+1)$.



i	1	2	3	4	5	6	7	8	9	10	11
$x_{(i)}$	36	37	45	52	56	58	66	68	75	90	100

(a)

L	H_L	L_D (n=...)	$(P)_L$	L_H	R_L
M	H_M		M		
F	H_F	F_D	(P_F)	F_H	R_F
E	H_E	E_D	(P_E)	E_H	R_E
D	H_D	D_D	(P_D)	D_H	R_D
...	...	...	...	...	...
...	1	$x_{(1)}$	(P_c)	$x_{(n)}$	R_c

(b) n=11

M		58			
F	3.5	48.5	(60)	71.5	23
E	2	37	(63.5)	90	53
	1	36	(68)	100	64

Obr. 2.1 Graficko-tabelární schéma sumarizace dat: (a) obecné schéma písmenově-číslicového zápisu výběru, (b) sedmipísmenový zápis výběru {36, 37, 45, 52, 56, 58, 66, 68, 75, 90, 100}.

Mezi základní statistické zvláštnosti rozdělení dat patří symetrie výběrového rozdělení a jeho relativní délky konců ve srovnání s normálním rozdělením. K vyjádření symetrie a špičatosti v různých vzdálenostech od mediánu se užívají jednoduché funkční charakteristiky založené na písmenových hodnotách. Ze vztahu pro délku konců T_L v následující tabulce lze snadno určit jejich teoretické velikosti pro vybraná symetrická rozdělení, a to hodnoty (T_E, T_D): *normální rozdělení* (0.534; 0.822), *rovnoměrné rozdělení* (0.405; 0.559) a *Laplaceovo rozdělení* (0.693; 1.098).

Pro rozdělení zešikmená k vyšším hodnotám jsou hodnoty šikmosti S_L záporné, při zešikmení k nižším hodnotám jsou kladné (viz tabulka 2.1). Pro rozdělení s delšími konci, než má normální rozdělení, rostou hodnoty pseudosigmy G_L se vzdáleností od mediánu. Když hodnoty G_L klesají s rostoucí vzdáleností od mediánu, má výběrové rozdělení kratší konce než normální. K posouzení statistických zvláštností dat se používá různých grafů, využívajících charakteristik z tabulky. Pro větší výběry se v grafech znázorňují pouze funkce písmenových hodnot, zatímco pro menší výběry se využívá všech kvantilů $\tilde{x}_{P_i} = x_{(i)}$, obyčejně při volbě

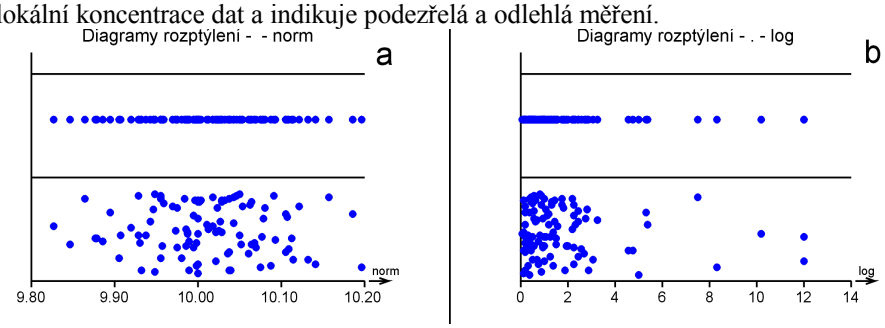
$$P_i = \frac{i - 0.333}{n + 0.333}.$$

Tabulka 2.1 Charakteristiky šikmosti a špičatosti výběrového rozdělení

Název	Definice	Charakterizuje	Platí pro L
Polosuma Z_L	$0.5\ (L_D + L_H)$	symetrii při $Z_L = 0$	*F, E, D, ...*
Rozpětí R_L	$(L_H - L_D)$	rozptýlení	*F, E, D, ...*
Šikmost S_L	$(M - Z_L) / R_L$	symetrii při $S_L = 0$	*F, E, D, ...*
Pseudosigma*) G_L	$R_L / (-2\ u_{P_i})$	špičatost (Gaussovo G_L = konst.)	*F, E, D, ...*
Délky konců T_L	$\ln\ (R_L / R_F)$	špičatost	*E, D*

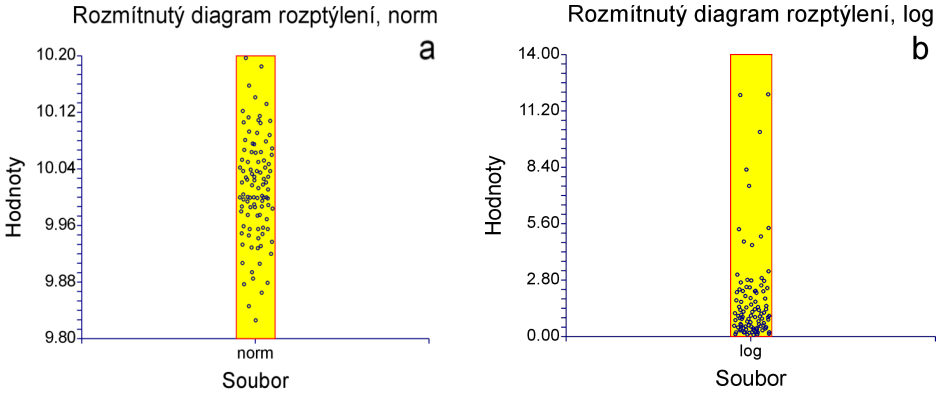
*) u_{P_i} je kvantil standardizovaného normálního rozdělení pro $P_i = 2^{-i}$.

Diagram rozptýlení (osa *x*: hodnoty x_i, osa *y*: libovolná úroveň, např. $y = 0$). Představuje jednorozměrnou projekci kvantilového grafu do osy *x*. I při své jednoduchosti ukazuje na lokální koncentrace dat a indikuje podezřelá a odlehlá měření.



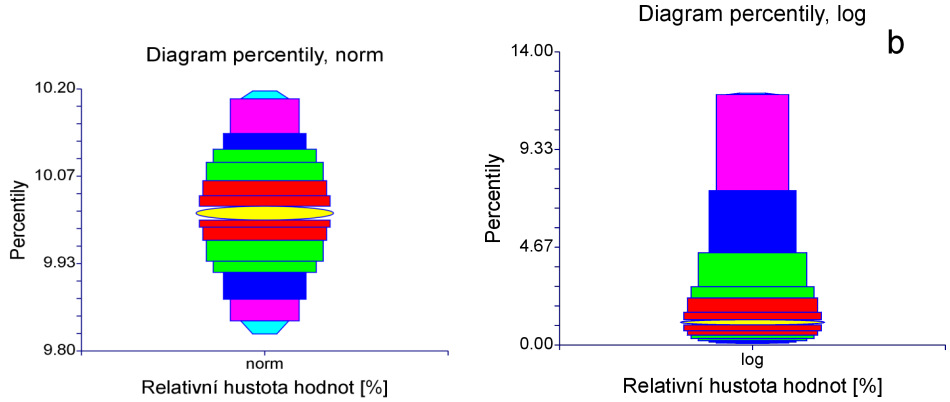
Obr. 2.2 Diagram rozptýlení a rozmítnutý diagram rozptýlení pro výběry (shora dolu): (a) *norm,* symetrického (normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Rozmítnutý diagram rozptýlení (osa *x*: hodnoty *x*, osa *y*: interval náhodných čísel). Diagram představuje rovněž projekci kvantilového grafu, body jsou však pro lepší přehlednost vhodně rozmítnuté.



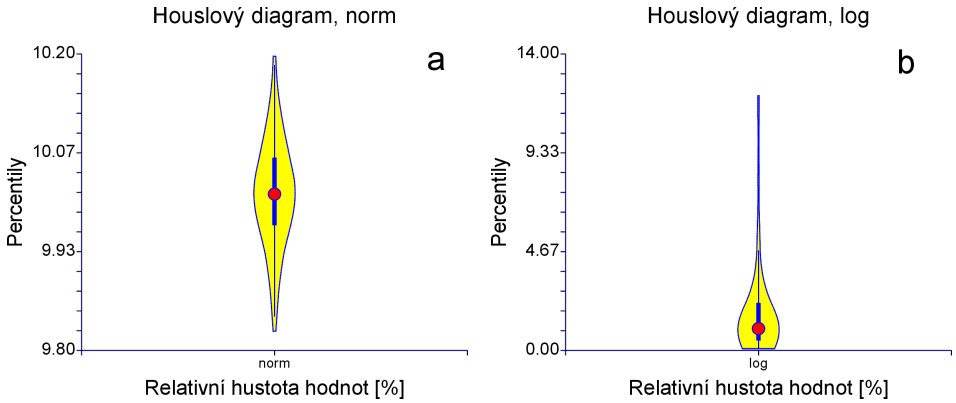
Obr. 2.3 Rozmítnutý diagram rozptýlení pro výběry: (a) *norm,* symetrického (normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000.***

Diagram percentilů (osa *x:* proměnná, osa *y:* percentily).
Diagram zobrazuje vybrané percentily. Jsou to obvykle intervaly 0-2, 2-5, 5-10, 10-15, 15-25, 25-35, 35-45, 45-55, 55-65, 65-75, 75-85, 85-90, 90-95, 95-99, 99-100. Z výsledného obrazce lze usoudit na symetrii rozdělení nebo na jeho tvar.



Obr. 2.4 Diagram některých percentilů pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000.***

Houslový diagram (osa *x*: název výběru proměnné, osa *y*: percentily, hodnoty proměnné). Diagram je kombinací krabicového grafu a dvou vertikálních, zrcadlově k sobě zobrazených grafů hustoty. Jeden graf hustoty roste směrem doprava a druhý doleva. Diagram zobrazuje píky a údolí stejně jako graf hustoty pravděpodobnosti. Medián je zobrazen černým kolečkem a začátek a konec úsečky zobrazuje dolní a horní kvantil. Houslový diagram se jmenuje dle tvaru připomínajícího housle. Normální rozdělení se projeví v symetrickém tvaru houslí, zatímco logaritmicko-normální v silně asymetrickém tvaru.

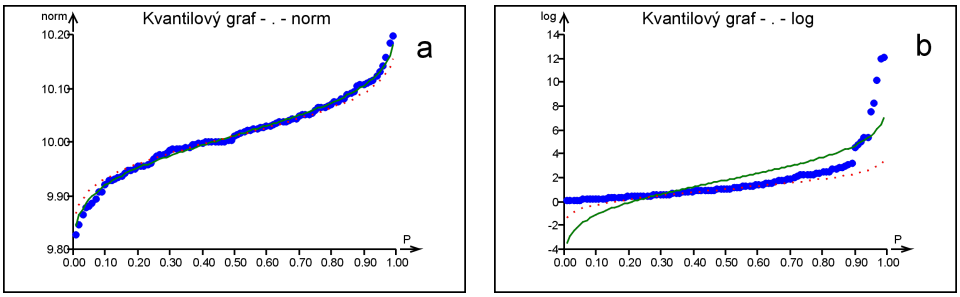


Obr. 2.5 Houslový diagram pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***NCSS2000.***

Kvantilový graf (osa *x*: pořadová pravděpodobnost P_i, osa *y*: pořádková statistika $x_{(i)}$). Umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, které může být

symetrické, zešikmené k vyšším nebo nižším hodnotám. Ke snadnějšímu porovnání s normálním rozdělením se do tohoto grafu zakreslují i kvantilové funkce normálního rozdělení, $N_{P_i} = \hat{\mu} + \hat{\sigma}\ u_{P_i}$, pro $0 \# P_i \# 1$:

(1) *klasických odhadů* parametrů polohy a rozptýlení, tj. aritmetického průměru a směrodatné odchylky $\hat{\mu} = x$ a $\hat{\sigma} = s$, a dále
(2) *robustních odhadů*, tj. mediánu M, $\hat{\mu} = M$ a $\hat{F} = R_F/1.349$, kde $R_F = F_H - F_D$ je interkvartilové rozpětí.

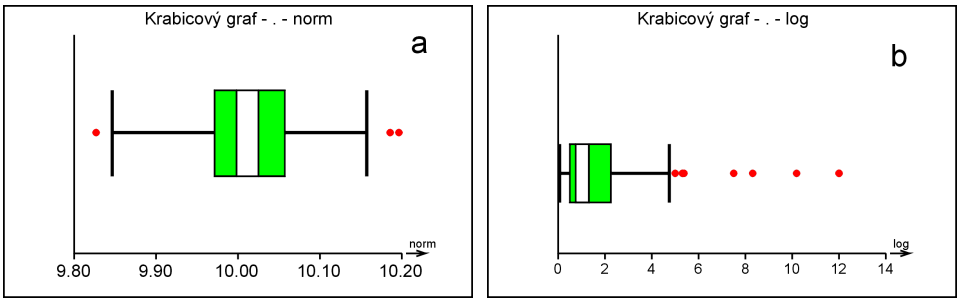


Obr. 2.6 Kvantilový graf (robustní --- a klasický ...) pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Krabicový graf (osa x: úměrná hodnotám x, osa y: interval úměrný hodnotě $\sqrt{n}$). Pro částečnou sumarizaci dat lze využít krabicového grafu, který umožňuje znázornění robustního odhadu polohy, mediánu M, dále posouzení symetrie v okolí kvartilů, posouzení symetrie u konců rozdělení a konečně identifikaci odlehlých dat. Krabicový graf je obdélník o délce $R_F = F_H - F_D$ s vhodně zvolenou šířkou, která je úměrná hodnotě $\sqrt{n}$. V místě mediánu M je vertikální čára. Od obou protilehlých stran tohoto obdélníku pokračují úsečky. Ty jsou ukončeny *vnitřními hradbami* B_H, B_D, pro které platí

$$B_H \ ' \ F_H \ \% \ 1.5\ R_F \ , \quad B_D \ ' \ F_D \ \& \ 1.5\ R_F \ .$$

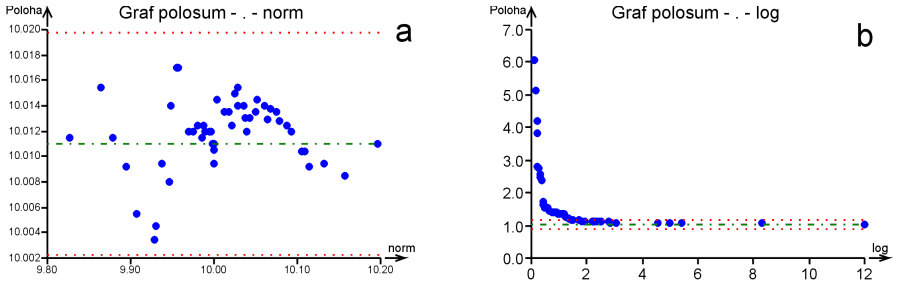
Prvky výběru, ležící mimo interval vnitřních hradeb $[B_H, B_D]$ jsou považovány za podezřelé, obvykle vybočující body; v grafu jsou znázorněny kroužky.



Obr. 2.7 Vrubový krabicový graf pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

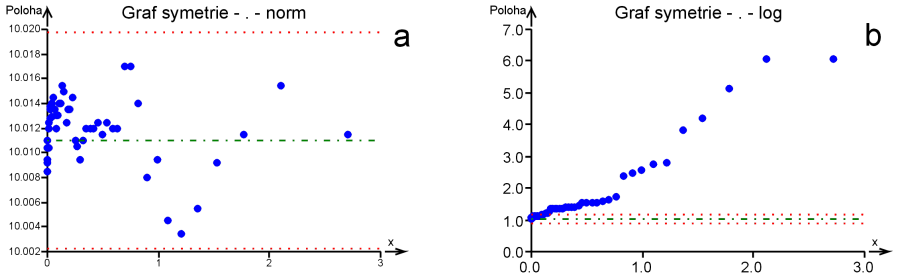
Vrubový krabicový graf (osa x: úměrná hodnotám x_i, osa y: interval úměrný hodnotě $\sqrt{n}$). Obdobou krabicového grafu je vrubový krabicový graf, který umožňuje také posouzení variability mediánu. Ta je totiž vyjádřena dolní a horní mezí intervalu spolehlivosti IS mediánu, $I_D \# M \# I_H$. Interval spolehlivosti IS bývá znázorněn v okolí mediánu bílým proužkem.

Graf polosum (osa x: pořádkové statistiky $x_{(i)}$, osa y: $Z_i = 0.5\ (x_{(n+1-i)} + x_{(i)})$). Pro symetrické rozdělení je grafem polosum horizontální přímka, určená rovnicí $y = M$. U tohoto grafu je důležité, že zde body oscilují okolo horizontální přímky a vykazují tak náhodný shluk (mrak) a měřítko osy y je silně detailní. Naopak, asymetrické rozdělení vykazuje nenáhodný trend a body pak neoscilují okolo horizontální přímky a měřítko osy y není detailní.

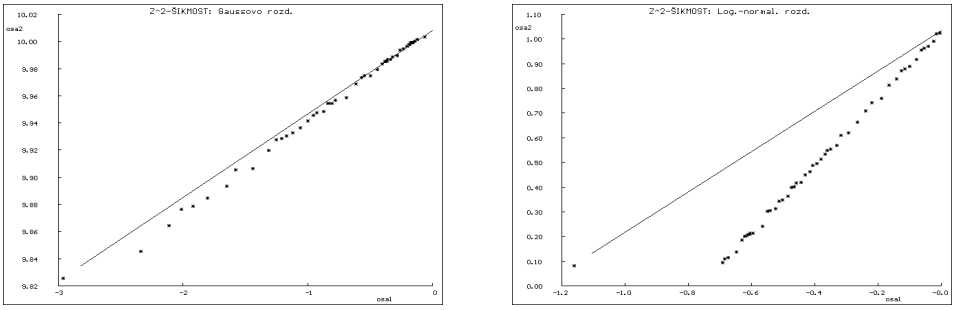


Obr. 2.8 Graf polosum pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Graf symetrie (osa x: $M - x_{(i)}$, osa y: $x_{(n+1-i)} - M$). Symetrická rozdělení jsou charakterizována přímkou $y = M$. Pro asymetrické rozdělení tato přímka nemá nulovou směrnici a v tomto grafu je směrnice odhadem parametru šikmosti. Asymetrické rozdělení vykazuje body uspořádané v trendu nějaké křivky.



Obr. 2.9 Graf symetrie pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***



Obr. 2.10 Graf šikmosti pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Graf šikmosti

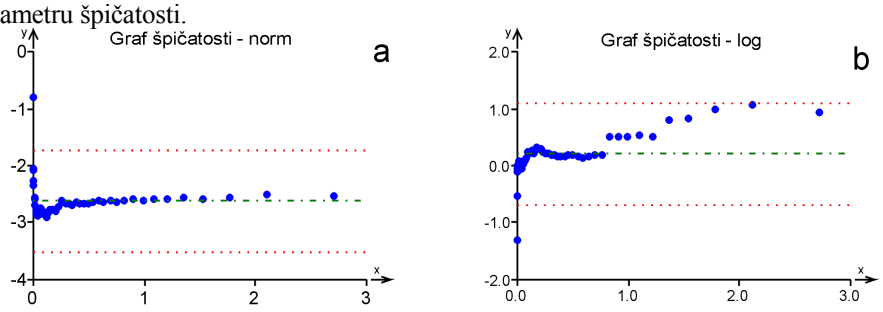
$$(\text{osa x: } u^2_{P_i}/2 \text{ pro } P_i \text{ ' } i/(n \text{ \% } 1),\ \text{osa } y\text{: } Z_i \text{ ' } 0.5\ (x_{(n\%1\&i)}\%x_{(i)})).$$

Pro případ symetrického rozdělení rezultuje u grafu šikmosti přímková závislost s nulovým úsekem a jednotkovou směrnicí. Body leží těsně na této přímce. U asymetrického rozdělení body neleží na této přímce a vykazují jinou směrnici.

Graf špičatosti

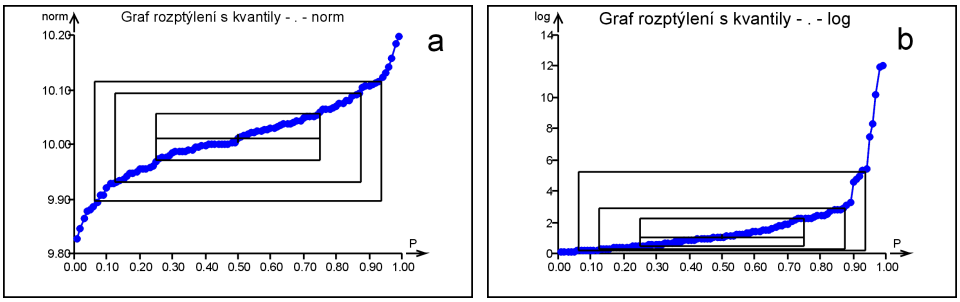
$$(\text{osa } x\text{: } u^2_{P_i}/2 \text{ pro } P_i \text{ ' } i/(n \text{ \% } 1),\ \text{osa } y\text{: } \ln(x_{(n\%1\&i)}/\&2u_{P_i}) \ .$$

Pro normální rozdělení je grafem horizontální přímka a body leží převážně na této přímce. Pokud body tvoří nenáhodný trend, odpovídá hodnota směrnice této aktuální přímky parametru špičatosti.



Obr. 2.11 Graf špičatosti pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Graf rozptýlení s kvantily (osa x: P_i, osa y: $x_{(i)}$). Základem je odhad kvantilové funkce výběru, který se získá spojením bodů $\{x_{(i)}, P_i\}$ lineárními úseky. Pro symetrická rozdělení má kvantilová funkce sigmoidální tvar. Pro rozdělení zešikmená k vyšším hodnotám je konvexně rostoucí a pro rozdělení zešikmená k nižším hodnotám konkávně rostoucí.



Obr. 2.12 Graf rozptýlení s kvantily pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Do grafu se zakreslují tři pomocné kvantilové obdélníky:

(a) *Kvartilový obdélník F:* na ose *y* kvartily F_D a F_H a na ose *x* pořadové pravděpodobnosti $P_2 = 2^{-2} = 0.25$ a $1 - 2^{-2} = 0.75$.

(b) *Oktilový obdélník E*: na ose *y* oktily E_D a E_H a na ose *x* pořadové pravděpodobnosti $P_3 = 2^{-3} = 0.125$ a $1 - 2^{-3} = 0.875$.

(c) *Sedecilový obdélník D*: na ose *y* sedecily D_D, D_H a na *x* ose pořadové pravděpodobnosti $P_4 = 2^{-4} = 0.0625$ a $1 - 2^{-4} = 0.9375$.

Graf rozptýlení s kvantily poskytuje následující závěry vyšetření dat:

1. *Symetrické unimodální rozdělení výběru* obsahuje obdélníky symetricky uvnitř sebe.

2. *Nesymetrická rozdělení* mají vzdálenosti mezi dolními hranami obdélníků *F, E* a *D* pro rozdělení zešikmené k vyšším hodnotám výrazně kratší než mezi horními.

3. *Odlehlá pozorování* jsou indikována tím, že na kvantilové funkci mimo obdélník *D* se objeví náhlý vzrůst, kdy hodnota směrnice roste téměř nade všechny meze.

4. *Vícemodální rozdělení* jsou indikována tím, že na kvantilové funkci uvnitř obdélníku *F* je několik úseků s téměř nulovými směrnicemi.

Histogram (osa *x*: proměnná *x*, osa *y*: úměrná hustotě pravděpodobnosti). Jde o obrys sloupcového grafu, kde jsou na ose *x* jednotlivé třídy, definující šířky sloupců. Výšky sloupců odpovídají empirickým hustotám pravděpodobnosti. Kvalitu histogramu ovlivňuje ve značné míře volba počtu tříd *L*. Pro přibližně symetrická rozdělení výběru lze vyčíslit počet tříd *L* podle vztahu $L = \mathrm{int}(2\sqrt{n})$, kde funkce int(*x*) označuje celočíselnou část čísla *x*. V širokém rozmezí velikostí výběrů *n* užijeme výraz $L = \mathrm{int}(2.46\,(n-1)^{0.4})$. Pro malé a střední výběry se konstruují jádrové odhady hustoty podle vztahu

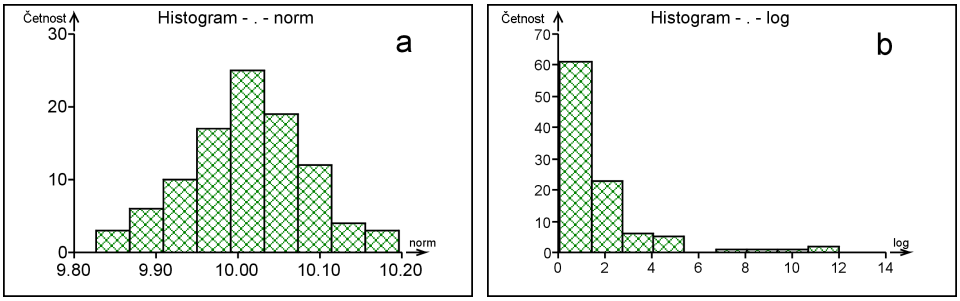
$$\hat{f}(x) = \frac{1}{n\,h} \sum_{i=1}^{n} K\left[\frac{(x - x_i)}{h}\right],$$

kde šířka pásu *h* určuje stupeň vyhlazení. Jádrová funkce *K(x)* je symetrická kolem nuly a má vlastnosti hustoty pravděpodobnosti. Vhodná je tzv. *bikvadratická funkce*

$$K(x) = 0.9375\,(1 - x^2)^2 \quad \text{pro} \quad -1 \le x \le 1,$$

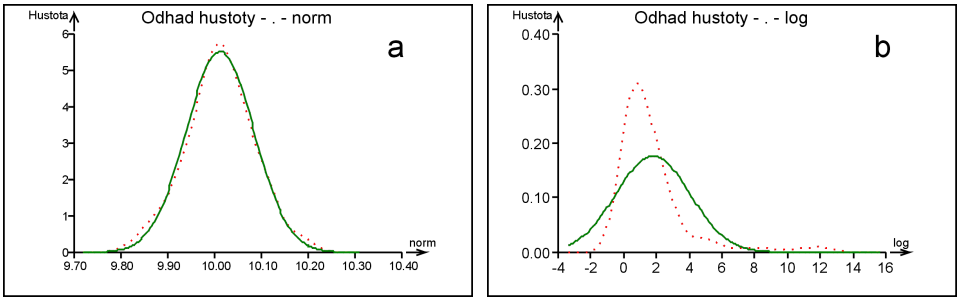
$$K(x) = 0 \quad \text{pro } x \text{ mimo interval} \quad -1 \le x \le 1.$$

O kvalitě odhadu hustoty pravděpodobnosti rozhoduje volba parametru h. Pro výběry velikosti n z přibližně normálního rozdělení se známým rozptylem σ^2 je optimální šířka pásu h_{opt} ' 2.34 σ $n^{\&0.2}$.



Obr. 2.13 Histogram pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

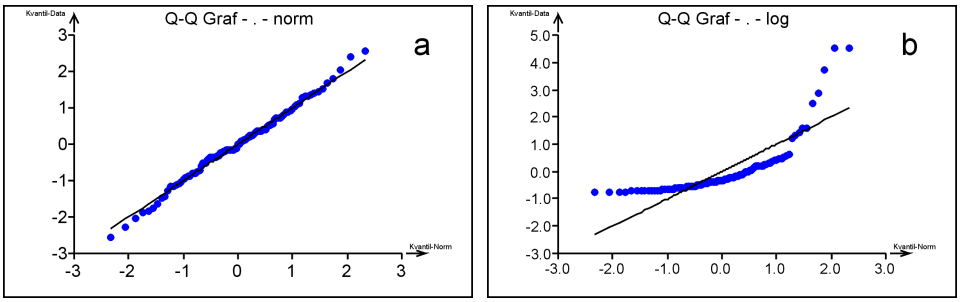
Jádrový odhad hustoty pravděpodobnosti (osa x: proměnná x, osa y: hustota pravděpodobnosti $\hat{f}(x)$).



Obr. 2.14 Jádrový odhad hustoty pravděpodobnosti pro výběry. Empirická křivka rozdělení (čárkovaně) a aproximační křivka Gaussova rozdělení (plná čára): (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Kvantilově-kvantilový graf (graf *Q-Q*) (osa x: $Q_T(P_i)$, osa y: $x_{(i)}$). Umožňuje posoudit shodu výběrového rozdělení, jež je charakterizováno kvantilovou funkcí $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$.

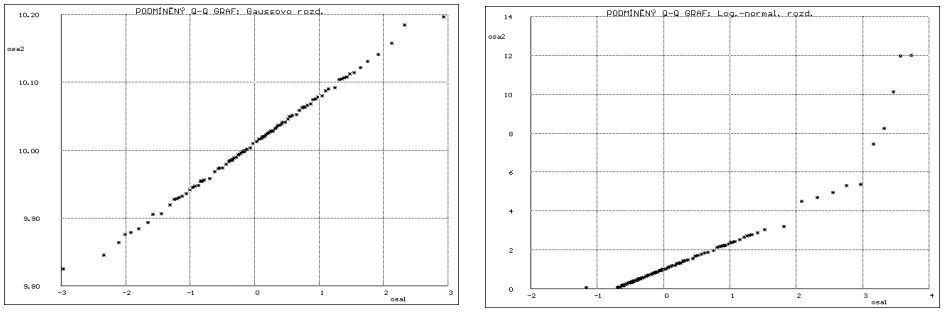
Jako odhad kvantilové funkce výběru se využívají pořádkové statistiky $x_{(i)}$. Při shodě výběrového rozdělení se zvoleným teoretickým rozdělením platí přibližná rovnost kvantilů $x_{(i)}$. $Q_T(P_i)$, kde P_i je pořadová pravděpodobnost, a závislost $x_{(i)}$ na $Q_T(P_i)$ je přibližně přímka. Korelační koeficient r_{xy} je pak kritériem těsnosti proložení této přímky při hledání typu neznámého rozdělení.



Obr. 2.15 Rankitový čili kvantil-kvantilový graf (*Q-Q* graf) pro ověření shody s teoretickým normálním rozdělením: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Rankitový graf (osa x: kvantil normovaného Gaussova rozdělení u_{P_i}, osa y: $x_{(i)}$). Pro porovnání rozdělení výběru s rozdělením normálním se *Q-Q* graf nazývá *grafem rankitovým*. Umožňuje také orientační zařazení výběrového rozdělení do skupin podle šikmosti, špičatosti a délky konců.

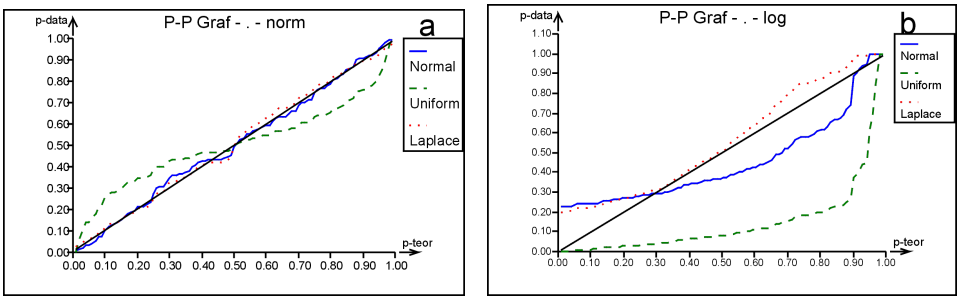
Podmíněný rankitový graf (osa x: Φ^{-1} $[0.5\ (U_{(i-1)} + U_{(i+1)})]$, osa y: $x_{(i)}$). K ověření normality výběrového rozdělení se užívá podmíněný rankitový graf. Symbol $\Phi^{-1}(U)$ značí standardizovanou kvantilovou funkci normálního rozdělení. Hodnoty $U = P_i$ jsou přímo kvantily u_{P_i}. Pořádkové statistiky $U_{(i)}$ jsou uspořádané náhodné proměnné U_i definované vztahem $U_i \sim \Phi\left[x_i \,\&\, \hat{\mu}_R / \hat{\sigma}_R\right]$, kde symbol $\Phi(x)$ značí distribuční funkci standardizovaného normálního rozdělení. Robustní odhad polohy je roven mediánu $\hat{\mu}_R = \tilde{x}_{0.5}$ a robustní směrodatná odchylka se vyčíslí vztahem $\hat{\sigma}_R = 0.75(\tilde{x}_{0.75} - \tilde{x}_{0.25})$. Pro úplnou definici se volí $U_{(0)} = 0$ a $U_{(n+1)} = 1$. Přibližná lineární závislost je v podmíněném rankitovém grafu důkazem normality testovaného rozdělení výběru. Z grafu normálního rozdělení je patrná výrazně menší lokální variabilita ve srovnání s rankitovými grafy.



Obr. 2.16 Podmíněný rankitový graf pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Pravděpodobnostní graf (***P-P* graf**), (osa x: P_i, osa y: $F_T(S_{(i)})$. Pravděpodobnostní grafy jsou alternativou ke *Q-Q* grafům. Slouží k porovnání distribuční funkce výběru, vyjádřené

přes pořadovou pravděpodobnost, se standardizovanou distribuční funkcí zvoleného teoretického rozdělení. Standardizovaná proměnná je zde definována vztahem $S_{(i)}$ ' $(x_{(i)}$ & $Q)/R$, kde Q je *parametr polohy* a R je *parametr rozptýlení.* V případě shody výběrového rozdělení se zvoleným teoretickým rozdělením vyjde *P-P* graf lineární s jednotkovou směrnicí a nulovým úsekem.

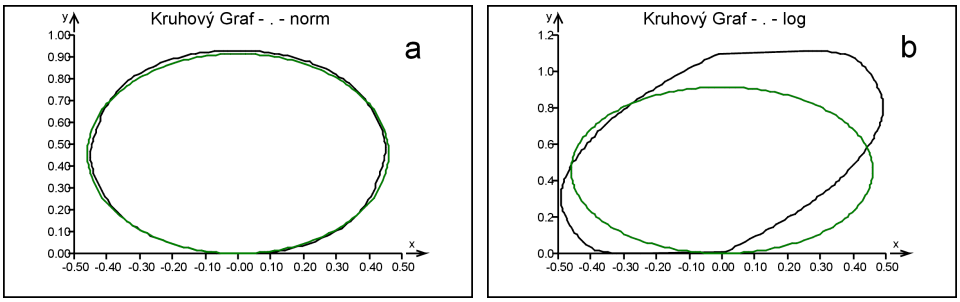


Obr. 2.17 Pravděpodobnostní graf (*P-P* graf) pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Na rozdíl od *Q-Q* grafů je při konstrukci *P-P* grafů nezbytné znát teoretické rozdělení až do hodnot všech parametrů. Obyčejně se určují odhady parametrů Q a R a navíc i dalších parametrů rozdělení s využitím momentové metody, resp. metody maximální věrohodnosti. Například pro normální rozdělení je $\hat{R} = s,\ \hat{Q} = \bar{x}$. Při porovnání *Q-Q* a *P-P* grafů platí, že

a) *P-P grafy* jsou citlivé na odchylky od teoretického rozdělení ve střední části,
b) *Q-Q grafy* jsou citlivé na odchylky od teoretického rozdělení v oblasti konců.

Kruhový graf slouží k vizuálnímu ověření hypotézy, že výběr pochází ze symetrického (nejčastěji Gaussova) rozdělení. V takovém případě je grafem regulární, konvexní polygon, blízký kružnici. Odchylky od kružnice ukazují na jiné než symetrické rozdělení výběru: (a) protáhlý elipsovitý tvar, s hlavní osou umístěnou úhlopříčně, ukazuje na asymetrické rozdělení, (b) elipsovitý tvar podél osy x ukazuje na rovnoměrné rozdělení.



Obr. 2.18 Kruhový graf pro výběry: (a) *norm,* symetrického (Gaussova, normálního), a (b) *log,* asymetrického (logaritmicko-normálního) rozdělení, ***ADSTAT.***

Vychází se z faktu, že transformací $Z_{(i)} \leftarrow F_e(x_{(i)})$ vyjdou náhodné veličiny $Z_{(i)}$ rozdělené přibližně rovnoměrně na intervalu [0, 1]. Při konstrukci kruhového grafu se definuje soustava vektorů $\bar{V}_i$ o stejné délce $l_0 \leftarrow 1/\sqrt{N\,(N-1)/2}$ a směru $\pi\, Z_{(i)}$. Pro složky x a y vektoru $\bar{V}_i$ platí $V_{x_i} \leftarrow l_0 \cos(\pi\, Z_{(i)})$, $V_{y_i} \leftarrow l_0 \sin(\pi\, Z_{(i)})$. Úhly se uvažují v radiánech. Vlastní kruhový graf pak vznikne, když se postupně (od počátku) spojují vektory $\bar{V}_1, \bar{V}_2, \ldots \bar{V}_N, -\bar{V}_1, -\bar{V}_2, \ldots, -\bar{V}_N$. Výsledný obrazec je $2N$ vrcholový konvexní polygon. Odchylky od ideálního tvaru ukazují na nevhodnost specifikace F_e, resp. jeho parametrů. Obyčejně se jako F_e volí distribuční funkce normálního rozdělení $\Phi(x_{(i)})$ a kruhový graf pak slouží pro ověřování normality.

Exploratorní analýza dat u výběru *norm* (ADSTAT)

(1) Linearita kvantilově-kvantilového (*Q*-*Q*) grafu $y = \beta_0 + \beta_1 x$:

Rozdělení	**Směrnice β_1**	**Úsek β_0**	**Korelační koeficient r_{xy}**
Laplaceovo	0.05246	10.012	0.99039
Normální	0.07289	10.012	0.99707
Exponenciální	0.06796	9.944	0.90972
Rovnoměrné	0.24233	9.891	0.97036
Lognormální	0.03253	9.960	0.82901
Gumbelovo	0.05623	10.044	0.97527

(2) Kvantilové míry polohy a rozptýlení:

Kvantil *L*	***P***	**Spodní L_D**	**Horní L_H**	**Rozpětí R_L**
Medián *M*	0.5	10.011	-	-
Kvartil *F*	0.25	9.9690	10.060	0.09100
Oktil *E*	0.125	9.9300	10.105	0.17450
Sedecil *D*	0.0625	9.8872	10.120	0.23225

(3) Vybrané míry polohy, rozptýlení a tvaru rozdělení, vyčíslené z kvantilů:

Kvantil *L*	***P***	**Polosuma Z_L**	**Šikmost S_L**	**Délka konců T_L**	**Pseudosigma G_L**
Kvartil *F*	0.25	10.015	-0.038462	0.000000.000000	
Oktil *E*	0.125	10.017	-0.035817	-19.1851	0.51708
Sedecil *D*	0.0625	10.003	0.032831	-16.0430	0.80089

Korelační koeficient r_{xy} dosahuje nejvyšší hodnoty pro normální rozdělení, a tím dokazuje, že výběr *norm* pochází ze souboru s normálním rozdělením.

2.2 Ověření předpokladů o datech

V praxi se nejčastěji předpokládá, že data tvoří *náhodný výběr* $\{x_i\}$, $i = 1, \ldots, n$, velikosti n. *Reprezentativní náhodný výběr* je charakterizován třemi důležitými předpoklady, které je třeba před vlastní analýzou vždy ověřit. Jsou to nezávislost jednotlivých prvků, homogenita a případná normalita rozdělení prvků výběru.

1. předpoklad: **Prvky výběru x_i jsou vzájemně nezávislé**
Pokud se podmínky pro měření dat mění s časem, projeví se tyto vznikem trendu mezi prvky výběru, uspořádanými v časovém sledu. K identifikaci časové závislosti prvků výběru nebo závislosti související s pořadím jednotlivých měření, testuje se významnost autokorelačního koeficientu prvního řádu ϱ_a podle testovacího von Neumannova kritéria

$$t_n = \frac{T_1\sqrt{n+1}}{\sqrt{1-T_1}}, \text{ kde } T_1 = (1-\frac{T}{2})\sqrt{\frac{n^2-1}{n^2-4}},$$

a T je von Neumannův poměr $T = \frac{\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$. Pokud jsou prvky výběru vzájemně nezávislé a platí nulová hypotéza H_0: $\varrho_a = 0$, má veličina t_n Studentovo rozdělení s $(n+1)$ stupni volnosti. Alternativní hypotézou je H_A: $\varrho_a \neq 0$. Platí, že pro případ $|t_n| > t_{1-\alpha/2}(n+1)$ je nutno nulovou hypotézu H_0 o nezávislosti prvků výběru na hladině významnosti α zamítnout.

2. předpoklad: **Výběr je homogenní**

Homogenní výběr znamená, že všechny jeho prvky x_i, $i = 1, ..., n,$ pocházejí ze stejného rozdělení s konstantním rozptylem σ^2. K nehomogenitě naměřených dat dochází všude tam, kde se vyskytuje výrazná nestejnoměrnost měřených vlastností vzorků nebo se náhle mění podmínky experimentů. Speciálním případem jsou odlehlá měření. Nehomogenita může být způsobena také nevhodnou specifikací souboru. Pokud lze daný výběr rozdělit podle nějakých logických kritérií do několika podskupin, je možno zpracovat statisticky každou podskupinu zvlášť a pak na základě testů shody středních hodnot v podskupinách rozhodnout, zda je toto dělení významné. Omezíme se na případ, kdy se v datech vyskytují vybočující hodnoty. Tyto hodnoty se co do velikosti výrazně liší od ostatních a lze je běžně identifikovat v grafech průzkumové analýzy. Odlehlá měření silně zkreslují odhady polohy a zejména rozptylu s^2, takže zcela znehodnocují další statistickou analýzu.

Problém vybočujících měření je velmi komplikovaný. Při jejich ověřování se používá řada idealizovaných předpokladů. Je nutné znát jejich předpokládaný počet, jejich rozdělení a rozdělení zbývajících prvků výběru. Navíc je třeba sestrojit model, podle kterého se odlehlá měření chovají. Testování vybočujících měření bez doplňkových informací je proto málo spolehlivé.

Jednoduchou technikou, kdy se pouze předpokládá, že "správná" data mají normální rozdělení, je *modifikace dolní vnitřní hradby* $B_D^{(}$ a *horní vnitřní hradby* $B_H^{(}$,

$$B_D^{(} = \tilde{x}_{0.25} - K(\tilde{x}_{0.75} - \tilde{x}_{0.25}), \quad B_H^{(} = \tilde{x}_{0.75} + K(\tilde{x}_{0.75} - \tilde{x}_{0.25}).$$

Parametr K se volí tak, aby pravděpodobnost $P(n, K)$, že z výběru velikosti n a pocházejícího z normálního rozdělení nebude žádný prvek mimo vnitřní hradby $[B_D^{(}, B_H^{(}]$, byla dostatečně vysoká, např. 0.95.

Při volbě $P(n, K) = 0.95$ lze v rozmezí $8 \leq n \leq 100$ použít aproximace $K \doteq 2.25 - 3.6/n$. Pro takto určený parametr K se všechny prvky výběru, ležící mimo hradby $[B_D^{(}, B_H^{(}]$, považují za vybočující. Výhodou je robustnost postupu. Není třeba znát počet vybočujících bodů ani jejich rozdělení a neprojevují se ani různé efekty "maskování".

3. předpoklad: **Rozdělení výběru je normální**

Na předpokladu normality je založena celá standardní statistická analýza dat. V *testu*

kombinace výběrové šikmosti a špičatosti dle Jarque-Berra se užívá testovací kritérium

$$\chi^2_{\exp} \;=\; \frac{\hat{g}_1^2}{D(\hat{g}_1)} + \frac{[\hat{g}_2 - E(\hat{g}_2)]^2}{D(\hat{g}_2)} \;,$$

kde jsou výběrová šikmost $\hat{g}_1$ a její rozptyl $D(\hat{g}_1)$, resp. výběrová špičatost $\hat{g}_2$, její střední hodnota $E(\hat{g}_2)$ a rozptyl $D(\hat{g}_2)$. Za předpokladu normality má veličina $\chi^2_{\exp}$ asymptoticky $\chi^2(2)$-rozdělení. Prokáže-li se proto, že $\chi^2_{\exp} > \chi^2_{1-\alpha}(2)$, je nutno hypotézu o normalitě rozdělení výběru zamítnout.

Střední hodnota výběru, pocházejícího z normálního rozdělení je $E(\hat{g}_1) = 0$. Pro asymptotický rozptyl tohoto odhadu platí

$$D(\hat{g}_1) \;\doteq\; \frac{(n-2)}{(n+1)(n+3)} \;.$$

Momentový odhad *špičatosti* g_2 je

$$\hat{g}_2 \;=\; \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} \;.$$

Střední hodnota tohoto odhadu pro výběry pocházející z normálního rozdělení je

$$E(\hat{g}_2) \;=\; 3 - \frac{6}{n+1}$$

a pro asymptotický rozptyl tohoto odhadu platí

$$D(\hat{g}_2) \;\doteq\; \frac{24\, n\, (n-2)\, (n-3)}{(n+1)^2\, (n+3)\, (n+5)} \;.$$

Při stanovení libovolného bodového odhadu parametru je třeba určit vždy i jeho rozptyl. K docílení stejné "přesnosti" odhadů je třeba při užití méně efektivního odhadu provést větší počet měření n. Například u dat pocházejících z normálního rozdělení se musí při použití mediánu $\tilde{x}_{0.5}$ provést 1.6krát více měření než při použití aritmetického průměru $\bar{x}$, aby se docílilo stejné přesnosti odhadu.

Základní předpoklady o prvcích výběru *norm* (ADSTAT)

(a) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$:	10.012
Odhad rozptylu s^2:	5.223E-03
Odhad směrodatné odchylky s:	0.0723
Odhad šikmosti $\hat{g}_1$:	-0.04
Odhad špičatosti g_2:	3.08
(b) Test normality: tabulkový kvantil $\chi^2_{1-\alpha}(2)$:	5.992
Odhad $\chi^2_{\exp}$ statistiky:	0.112
Závěr: Předpoklad normality přijat. Spočtená hladina významnosti $\alpha = 0.9456$.	

(c) Test nezávislosti: tabulkový kvantil $t_{1-\alpha/2}(n+1)$:	1.984
Odhad von Neumannovy statistiky t_n:	1.218
Závěr: Předpoklad nezávislosti přijat. Spočtená hladina významnosti $\alpha = 0.113$.	
(d) Detekce odlehlých bodů: metodou modifikované vnitřní hradby	
Dolní vnitřní hradba B_D:	9.783
Horní vnitřní hradba B_H:	10.245
Závěr: Ve výběru nejsou odlehlé body.	

2.3 Transformace dat

Pokud se na základě analýzy dat zjistí, že rozdělení výběru dat se systematicky odlišuje od rozdělení normálního, vzniká problém, jak data vůbec vyhodnotit. Často je pak vhodná transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě rozdělení.

1. *Stabilizace rozptylu* vyžaduje nalezení transformace $y = g(x)$, ve které je již rozptyl $\sigma^2(y)$ konstantní. Pokud je rozptyl původní proměnné x funkcí typu $\sigma^2(x) = f_1(x)$, lze rozptyl $\sigma^2(y)$ určit

$$\sigma^2(y) \doteq \left[\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right]^2 f_1(x) = C ,$$

kde C je konstanta. Hledaná transformace $g(x)$ je pak řešením diferenciální rovnice

$$g(x) \doteq C \int \frac{\mathrm{d}x}{\sqrt{f_1(x)}} .$$

2. *Zesymetričtění rozdělení* výběru je možné provést užitím prosté *mocninné transformace*

$$y = g(x) = \begin{cases} x^{\lambda} & \lambda > 0 \\ \ln x & \text{pro } \lambda = 0 \\ -x^{-\lambda} & \lambda < 0 \end{cases} .$$

Mocninná transformace však nezachovává měřítko, není vzhledem k hodnotě λ všude spojitá a hodí se pouze pro kladná data. Optimální odhad exponentu λ se hledá s ohledem na optimalizaci charakteristik asymetrie (šikmosti) a špičatosti. K určení optimálního λ lze užít i *rankitového grafu*, kdy pro optimální exponent λ budou kvantily $y_{(i)}$ ležet na přímce, nebo *selekčního grafu dle Hinese a Hinesové*.

Selekční graf dle Hinese a Hinesové (osa x: $\tilde{x}_{0.5}/\tilde{x}_{1-P_i}$, osa y: $\tilde{x}_{Pi}/\tilde{x}_{0.5}$). Diagnostickou pomůckou pro odhad optimálního exponentu λ je selekční graf dle Hinese a Hinesové. Vychází z požadavků symetrie jednotlivých kvantilů kolem mediánu

$$\left(\frac{\tilde{x}_{P_i}}{\tilde{x}_{0.5}}\right)^{\lambda} \cdot \left(\frac{\tilde{x}_{0.5}}{\tilde{x}_{1-P_i}}\right)^{-\lambda} = 2 ,$$

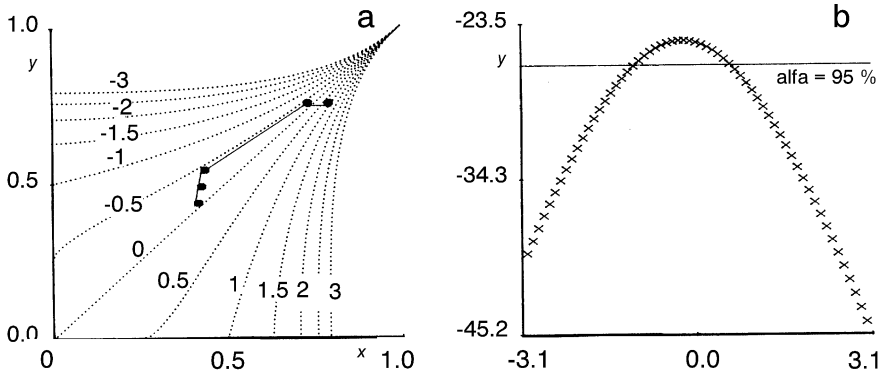
kde jako kvantily jsou obvykle voleny písmenové hodnoty. K porovnání průběhu jednotlivých bodů s ideálním pro zvolené λ se do grafu zakreslují řešení rovnice $y^{\lambda} + x^{-\lambda} = 2$ pro $0 \# x \# 1$ a $0 \# y \# 1$:

a) pro $\lambda = 0$ je řešením přímka $y = x$,

b) pro $\lambda < 0$ je řešením vztah $y = (2 - x^{-\lambda})^{1/\lambda}$,

c) pro $\lambda > 0$ je řešením vztah $x = (2 - y^{\lambda})^{-1/\lambda}$.

Podle umístění experimentálních bodů v okolí nomogramu teoretických křivek selekčního grafu lze vizuálně odhadovat velikost λ a posuzovat tak kvalitu transformace v různých vzdálenostech od mediánu.



Obr. 2.19 (a) Selekční graf dle Hinese a Hinesové, a (b) graf logaritmu věrohodnostní funkce na λ pro výběr z lognormálního rozdělení, ***ADSTAT.***

Pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti je vhodná *Boxova-Coxova transformace*

$$y = g(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \ln x & (\lambda = 0) \end{cases},$$

Boxova-Coxova transformace je použitelná pouze pro kladná data. Rozšíření této transformace na oblast, kdy rozdělení dat začíná od prahové hodnoty x_0, spočívá v náhradě x rozdílem $(x - x_0)$, který je vždy kladný.

Graf logaritmu věrohodnostní funkce (osa x: λ, osa y: ln L). Pro odhad parametru λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti s tím, že pro $\lambda = \hat{\lambda}$ je rozdělení transformované veličiny y normální, $N(\mu_y, \sigma^2(y))$. Po úpravách bude logaritmus věrohodnostní funkce ve tvaru

$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i ,$$

kde $s^2(y)$ je výběrový rozptyl transformovaných dat y. Průběh věrohodnostní funkce ln $L(\lambda)$ lze znázornit ve zvoleném intervalu, např. $-3 \# \lambda \# 3$, a identifikovat maximum křivky, jejíž souřadnice x indikuje odhad $\hat{\lambda}$.

Dva průsečíky křivky ln $L(\lambda)$ s rovnoběžkou s osou x indikují $100(1-\alpha)\%$ interval

spolehlivosti parametru λ. Čím bude tento interval spolehlivosti $+\lambda_D$, λ_H, širší, tím je mocninná Boxova-Coxova transformace méně výhodná. Pokud obsahuje interval $+\lambda_D$, λ_H, i hodnotu λ = 1, není transformace ze statistického hlediska přínosem.

Zpětná transformace: po vhodné transformaci vyčíslíme $\bar{y}$, $s^2(y)$ a potom pomocí zpětné transformace využitím Taylorova rozvoje v okolí $\bar{y}$ odhadneme retransformované parametry $\bar{x}_R$ a $s^2(\bar{x}_R)$ původních dat. Uvedený postup vesměs vede k lepším odhadům polohy $\bar{x}_R$ a rozptylu $s^2(\bar{x}_R)$ a je vhodný zvláště v případech asymetrického rozdělení výběru.

Mocninná a Boxova-Coxova transformace u výběru *norm* a **úlohy B2.04** (ADSTAT)

(1) Odhady klasických parametrů:	*norm*	**B2.04**
Odhad aritmetického průměru	10.012	0.177
Odhad směrodatné odchylky	0.072271	0.159
Odhad šikmosti	-0.037	1.54
Odhad špičatosti	3.08	5.36
(2) Prostá mocninná transformace:		
Odhad optimálního exponentu	2.67	0.53
Odhad průměru transformovaných dat	465.69	0.355
Odhad směrodatné odchylky transformovaných dat	8.96	0.191
Odhad šikmosti transformovaných dat	0.0006	0.28
Odhad špičatosti transformovaných dat	3.08	3.13
Opravený odhad průměru původních dat	10.012	0.143
Opravený odhad směrodatné odchylky původních dat	0.07226	0.145
Spodní mez intervalu spolehlivosti původních dat	9.998	0.102
Horní mez intervalu spolehlivosti původních dat	10.027	0.190
(3) Boxova-Coxova transformace:		
Odhad optimálního exponentu	2.67	0.53
Odhad průměru transformovaných dat	174.26	-1.210
Odhad směrodatné odchylky transformovaných dat	3.361	0.358
Odhad šikmosti transformovaných dat	0.0006	0.28
Odhad špičatosti transformovaných dat	3.08	3.13
Opravený odhad průměru původních dat	10.012	0.143
Opravený odhad směrodatné odchylky původních dat	0.07226	0.145
Spodní mez intervalu spolehlivosti původních dat	9.998	0.102
Horní mez intervalu spolehlivosti původních dat	10.027	0.190

2.4 Průběh průzkumové analýzy dat

Průběh vlastní průzkumové, exploratorní analýzy dat (EDA) je možné libovolně kombinovat dle dosavadních informací o vyšetřovaných datech. Omezíme se na zpracování dvojího druhu dat, jednak *rutinních dat*, o kterých jsou známy vlastnosti, jako je např. rozdělení, a jednak *neznámých dat*, o kterých nejsou známy dosud žádné předběžné informace a hrozí nebezpečí nesplnění předpokladů o datech.

A. Postup analýzy rutinních dat

Při zpracování rutinních výsledků měření předpokládáme, že známe rozdělení dat. Předpokládá se, že rozdělení dat je normální a data asi splňují předpoklady nezávislosti a homogenity. Účelem je

a) testování nezávislosti prvků výběru - autokorelace,
b) testování homogenity výběru,
c) testování normality rozdělení výběru.

Z grafických metod se k předběžné analýze rutinních dat nejčastěji užívá *rankitového grafu* a *grafu rozptýlení s kvantily*. Nejsou-li však o rozdělení dat dostupné žádné informace, nebo očekává-li se výrazně nenormální rozdělení, je vhodné provést

a) průzkumovou analýzu dat využitím řady grafických diagnostik,
b) určení výběrového rozdělení a jeho konstrukce.

Pokud nebylo nalezeno vhodné aproximující rozdělení, provádí se *mocninná transformace*, která by měla zlepšit rozdělení dat. Kombinace metod závisí na konkrétních datech a konkrétních požadavcích analýzy.

B. Postup při nesplnění předpokladů o datech

1. Nesplnění předpokladu nezávislosti prvků. Pokud prvky měření nejsou nezávislé, vzrůstá nebezpečí, že odhady budou systematicky vychýleny a nadhodnoceny pro pozitivní hodnotu autokorelačního koeficientu ρ_a. Nezbývá, než hlouběji analyzovat logické příčiny a snažit se o jejich odstranění, zkontrolovat celý měřicí řetězec a provést nová měření.

2. Nesplnění předpokladu normality výběru. Rozdělení dat je buď jiné než normální, nebo jsou v datech odlehlá měření. V případě nenormálního rozdělení dat může jít o odchylky pouze v délce konců, nebo se jedná *o zešikmená rozdělení.* V případě symetrických rozdělení, lišících se od normálního délkou konců, lze použít pro odhad parametrů polohy a rozptýlení jednoduché robustní techniky. U zešikmených rozdělení je vždy výhodné začít hledáním mocninné transformace. Pokud byla mocninná transformace úspěšná a byla nalezena optimální mocnina λ, provádí se další analýza v této transformaci a nakonec se vyčíslí zpětná transformace do původních proměnných.

Pro *zešikmená rozdělení*, charakterizovaná třetím centrálním momentem m_3, lze definovat modifikovanou náhodnou veličinu

$$T_C = \left[(\bar{x} - \mu) + \frac{m_3}{6\,\sigma^2\,n} + \frac{m_3}{3\,\sigma^4}\,(\bar{x} - \mu)^2\right]\frac{\sqrt{n}}{s},$$

která má Studentovo rozdělení s $(n - 1)$ stupni volnosti. Při praktických výpočtech se rozptyl σ^2 nahrazuje nevychýleným odhadem s^2 a třetí centrální moment m_3 jeho nevychýleným odhadem

$$\hat{m}_3 = \frac{n}{(n - 1)\,(n - 2)} \sum_{i=1}^{n} (x_i - \bar{x})^3 .$$

Při konstrukci *konfidenčního intervalu střední hodnoty* $L_D \le \mu \le L_H$ se užívá vztahů pro dolní a horní meze

$$L_D = \bar{x} + \frac{1}{2\,C_2} - \frac{\sqrt{d_1}}{2\,C_2}, \quad L_H = \bar{x} + \frac{1}{2\,C_2} - \frac{\sqrt{d_2}}{2\,C_2},$$

kde $$C_1 = \frac{\hat{m}_3}{6\, s^2\, n}, \qquad C_2 = \frac{\hat{m}_3}{3\, s^4},$$

$$C = t_{1-\alpha/2}(n-1)\, \frac{s}{\sqrt{n}},$$

$$d_1 = 1 - 4\, C_2\, (C_1 - C), \qquad d_2 = 1 - 4\, C_2\, (C_1 + C).$$

Využitím tohoto konfidenčního intervalu pro střední hodnotu zešikmených rozdělení lze také provádět testy významnosti parametru polohy.

3. Přítomnost vybočujících hodnot: Na základě logické analýzy je třeba nejdříve zvážit, zda nejde o zešikmené rozdělení. Body, které se jeví vybočující pro symetrické (speciálně normální) rozdělení, mohou být pro zešikmená rozdělení naopak přijatelné. Pokud jde o vybočující pozorování, lze použít dvou alternativ.

První alternativa spočívá v jejich vyloučení z další analýzy, což však není vždy zcela nejvhodnější. Pokud jsou odlehlá měření výsledkem řídce se vyskytujících jevů, může tím totiž dojít ke ztrátě informace úplně. Proto lze tyto hodnoty vyloučit jedině při doplnění o nová experimentální data.

Druhá alternativa spočívá v použití robustních metod. Tento postup však nemusí být vždy korektní. Robustnost spočívá v přiblížení se k přijatému modelu měření bez ohledu na jeho platnost. Pokud se analýzy vybočujících měření účastní experimentátor, měl by rozhodnout, která měření jsou evidentní hrubé chyby (jako je selhání přístroje, špatný zápis dat) a která jsou jen podezřelá. Evidentní hrubé chyby je vhodné z další analýzy vyloučit, ale podezřelá měření je lépe ponechat. Robustními metodami se jejich vliv na odhady parametrů výrazně oslabí.

4. Nedostatečný rozsah výběru: Nejjednodušší je v tomto případě provést dodatečná měření. Platí, že čím jsou data méně rozptýlená, tím menší počet jich stačí k zajištění dostatečné přesnosti odhadu. Pokud nelze provést dodatečné experimenty, je možné použít techniky vhodné pro malé výběry (viz Hornův postup v 3. kap.).

Tento postup je vhodný zejména pro analýzu rutinních měření, kde jsou o chování dat předběžné informace. Když se analyzují výsledky nových měření nebo neznámé výběry, je vždy třeba začít průzkumovou analýzou dat a stanovit statistické zvláštnosti výběru.

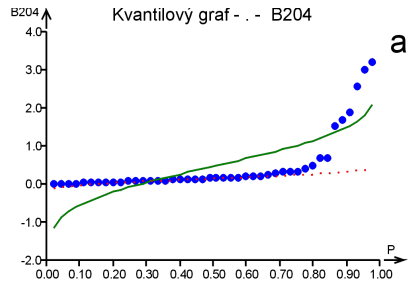
Vzorová úloha 2.2: *Průzkumová analýza velkého výběru*

Na úloze **B2.04** *Kontrola obsahu ergosterinu v kalciferolu* ukážeme postup průzkumové analýzy dat. Při výrobě calciferolu se provádí kontrola meziproduktu 3,5 DNB esteru calciferolu metodou HPLC. Sleduje se také obsah přítomného ergosterinu jako nečistoty, jejíž střední hodnota by neměla přesáhnout 0.4 %. Metodou průzkumové analýzy dat vyšetřete, zda jsou splněny požadavky, kladené na náhodný výběr a zda je splněn i požadavek čistoty calciferolu. Určete typ rozdělení. Které diagnostiky shodně indikují vybočující hodnoty? Jak velké procento hodnot dosahuje obsahu 0.4 %? Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru.

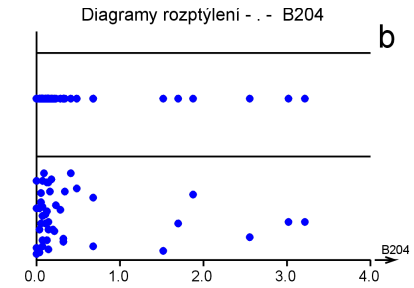
Řešení:

1. Zkoumání zvláštností dat: grafické diagnostiky indikují vedle stupně symetrie a špičatosti rozdělení také odlehlé body.

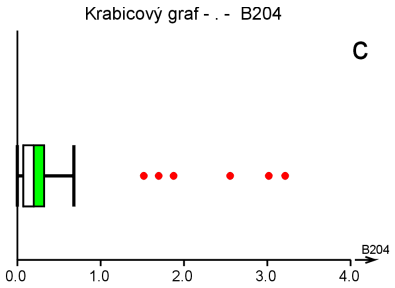
(a) *Odhalení stupně symetrie a špičatosti rozdělení:* celkem 12 grafických diagnostik indikuje symetrii a špičatost rozdělení s těmito závěry:



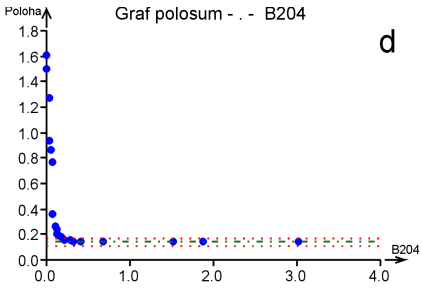
Obr. 2.20a Kvantilový graf.



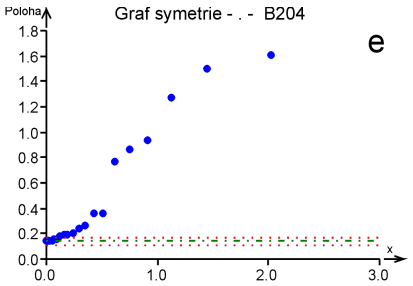
Obr. 2.20b Diagram rozptýlení a rozmítnutý diagram rozptýlení.



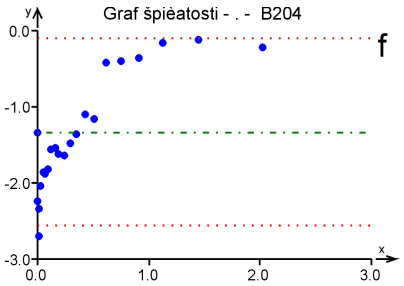
Obr. 2.20c Vrubový krabicový graf.



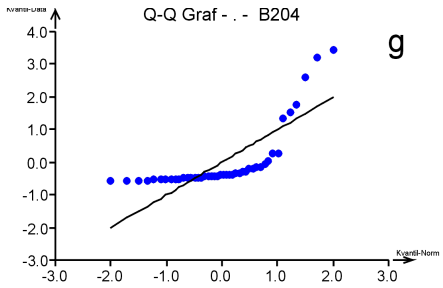
Obr. 2.20d Graf polosum.



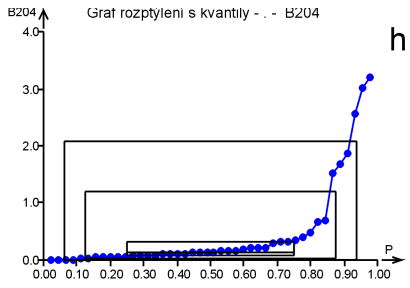
Obr. 2.20e Graf symetrie.



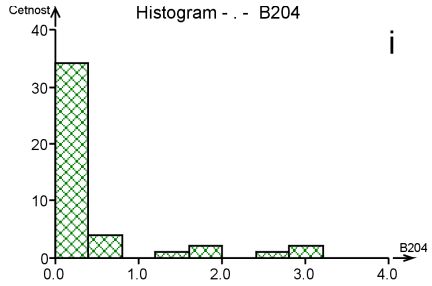
Obr. 2.20f Graf špičatosti.



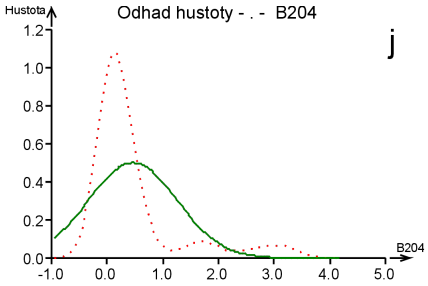
Obr. 2.20g Kvantil-kvantilový (*Q-Q*) graf.



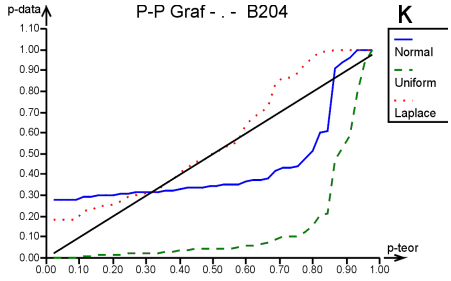
Obr. 2.20h Graf rozptýlení s kvantily.



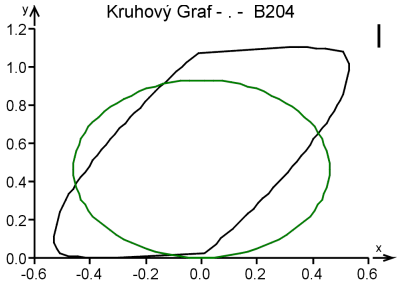
Obr. 2.20i Histogram.



Obr. 2.20j Graf hustoty pravděpodobnosti.



Obr. 2.20k *P-P* graf.



Obr. 2.20*l* Kruhový graf.

1. *Kvantilový graf* (obr. 2.20a): je patrný velký rozdíl mezi symetrickým Gaussovým a empirickým rozdělením. Tvar křivky je charakteristický pro asymetrické rozdělení, silně zešikmené k vyšším hodnotám.
2. *Diagram rozptýlení a rozmítnutý diagram rozptýlení* (obr. 2.20b): ukazuje na 4 odlehlé body v horní části diagramu a 1 odlehlý bod v dolní části diagramu.
3. *Vrubový krabicový graf* (obr. 2.20c): v horní části je detekováno 6 odlehlých bodů. Krabice je rozdělena na dvě části mediánem.
4. *Graf polosum* (obr. 2.20d): indikuje velkou část bodů jako vybočujících ze symetrického rozdělení. Body, ležící na mediánové rovnoběžce s osou *x* jsou ze symetrického rozdělení, ostatní nikoliv.
5. *Graf symetrie* (obr. 2.20e): indikuje valnou část bodů jako vybočujících ze symetrického rozdělení nebo patřících do asymetrického rozdělení.

6. Graf špičatosti (obr. 2.20f): většina bodů neleží na rovnoběžce s osou *x* pro symetrické rozdělení, a proto jde o rozdělení asymetrické.

7. Kvantilově-kvantilový (Q-Q) graf (obr. 2.20g): jelikož většina bodů neleží na přímce jde o asymetrické rozdělení.

8. Graf rozptýlení s kvantily (obr. 2.20h): asymetrie kvantilových obdélníků obdélníků dokazuje silně asymetrické rozdělení. Body ležící vně sedecilového obdélníku indikuje tato pomůcka jako odlehlé.

9. Histogram (obr. 2.20i): zřetelně ukazuje na asymetrické rozdělení zešikmené k vyšším hodnotám.

10. Jádrový odhad hustoty pravděpodobnosti (obr. 2.20j): ve srovnání s Gaussovým rozdělením je patrné silné zešikmení k vyšším hodnotám. Empirickou křivku nelze aproximovat symetrickým Gaussovým rozdělením.

11. Pravděpodobnostní P-P graf (obr. 2.20k): empirická křivka nesouhlasí s žádnou křivkou symetrického rozdělení (normálního, rovnoměrného a Laplaceova). Rozdělení je asymetrické.

12. Kruhový graf (obr. 2.20*l*): tvar elipsy dokazuje silně asymetrické rozdělení zešikmené k vyšším hodnotám.

Kvantily a písmenové hodnoty **Úlohy B2.04** (ADSTAT)

Kvantily a písmenové hodnoty:

Procento	Kvantil	Procento	Kvantil
5	1.0000E-04	10	3.2300E-02
15	4.3900E-02	20	5.4800E-02
25	6.8250E-02	30	7.6700E-02
35	9.4300E-02	40	1.1520E-01
45	1.3000E-01	50	1.4000E-01
55	1.5000E-01	60	1.5820E-01
65	1.8230E-01	70	2.0100E-01
75	2.3175E-01	80	2.9440E-01
85	3.1910E-01	90	3.3430E-01
95	4.7800E-01		

Písmenové hodnoty:

Kvantil	Písmeno	Pravděpodobnost	Spodní mez L_D	Horní mez L_H
Sedecil	*D*	0.0625	1.0000E-04	4.3500E-01
Oktil	*E*	0.1250	3.6750E-02	3.2062E-01
Kvartil	*F*	0.2500	6.8250E-02	2.3175E-01
Medián	*M*	0.5000	1.4000E-01	1.4000E-01

Kvantilové míry:

Kvantil	*F (0.25)*	*E (0.125)*	*D(0.0625)*
Rozpětí R_L	1.6350E-01	2.8387E-01	4.3490E-01
Polosuma Z_L	1.5000E-01	1.7869E-01	2.1755E-01
Délka konců T_L	0.0000E+00	5.5172E-01	9.7830E-01
Šikmost S_L	1.1443E-01	5.8392E-02	-4.8843E-03
PseudoSigma G_L	1.2129E-01	1.2342E-01	1.4212E-01

Kvantily a písmenové hodnoty umožňují posoudit jednak symetrii výběrového rozdělení a jednak procento prvků ve výběru. Pro *procento* 45 je *kvantil* 0.1300, což znamená, že pod hodnotou 0.1300 leží 45 % a nad ní 55 % prvků výběru. *Písmenové hodnoty* a *kvantilové míry* umožňují sestrojení *graficko-tabelárního schématu písmenově-číslicového zápisu*

výběru či sumarizace dat. Lišící se hodnoty *kvantilových polosum* (kvartilové, oktilové, sedecilové) indikují asymetrické rozdělení, v případě symetrického rozdělení by totiž všechny polosumy dosahovaly stejné hodnoty.

Sedmipísmenový zápis výběru:

	Dolní kvantil L_D	(Polosuma)	Horní kvantil L_H	
Median M		0.1400		Rozpětí R_L
Kvartil F	0.06825	(0.1500)	0.23175	0.16350
Oktil E	0.03675	(0.17869)	0.32062	0.28387
Sedecil D	0.0001	(0.21755)	0.43500	0.43490

(b) *Indikace lokální koncentrace dat a rozdělení výběru:* rozdělení je asymetrické, s dlouhým horním koncem s větší koncentrací bodů ve spodní části hodnot. Z analýzy kvantilově-kvantilového *Q-Q* grafu vyplývá, že nejvyšší hodnoty korelačního koeficientu je dosaženo především pro exponenciální rozdělení. Výběrové rozdělení je zde aproximováno exponenciálním.

Linearita v kvantilově-kvantilovém grafu **Úlohy B2.04** (ADSTAT)

Linearita kvantil-kvantilovém (Q-Q) grafu $y = \beta_0 + \beta_1 x$:			
Rozdělení	**Směrnice β_0**	**Úsek β_1**	**Korelační koeficient r_{xy}**
Laplaceovo	0.11077	0.17671	0.92263
Normální	0.14976	0.17671	0.92054
Exponenciální	0.16708	0.01259	0.98963
Rovnoměrné	0.48940	-0.06799	0.89212
Log.-normální	0.09055	0.03413	0.96956

(c) *Nalezení vybočujících prvků ve výběru:* z grafických diagnostik průzkumové (exploratorní) analýzy výběru bylo nalezeno 3 až 6 podezřelých bodů, které by mohly být chápany v symetrickém rozdělení jako odlehlé. Jelikož však jde o asymetrické rozdělení exponenciální, nemá smyslu zde indikovat odlehlé body.

2. Ověření předpokladů o datech: *Reprezentativní náhodný výběr* je charakterizován třemi důležitými předpoklady, které je třeba před vlastní analýzou ověřit. Jsou to nezávislost, homogenita a případná normalita výběru.

(a) *Ověření normality rozdělení:* Na předpokladu normality je založena celá statistická analýza dat. Test kombinace výběrové šikmosti a špičatosti ukázal, že normalita výběrového rozdělení je zamítnuta.

(b) *Ověření nezávislosti prvků výběru:* K identifikaci časové závislosti prvků výběru nebo závislosti související s pořadím jednotlivých měření se testuje významnost autokorelačního koeficientu prvního řádu podle von Neumannova testovacího kritéria. U analyzovaného výběru byla nezávislost prvků ve výběru prokázána.

(c) *Ověření homogenity rozdělení výběru:* Homogenní výběr znamená, že všechny jeho prvky pocházejí ze stejného rozdělení s konstantním rozptylem. Odlehlá měření silně zkreslují odhady polohy a zejména rozptylu s^2, takže zcela znehodnocují další statistickou analýzu. Testování vybočujících měření bez doplňkových informací průzkumové analýzy

dat je málo spolehlivé. Kritérium vnitřních mezí určilo 2 odlehlé body, a to body č. 14 a 23. Doplněním informací z průzkumové analýzy lze identifikovat exponenciální rozdělení výběru, které již odlehlé body mít nebude, takže toto vyloučení dvou bodů nemá statistický smysl.

(d) *Určení minimálního rozsahu výběru:* Uvažujeme-li 25% relativní chybu směrodatné odchylky, bude minimální rozsah výběru n = 18, pro 10% relativní chybu směrodatné odchylky pak n = 110 a pro 5% relativní chybu směrodatné odchylky bude n = 437.

Základní předpoklady o výběru **Úlohy B2.04** (ADSTAT)

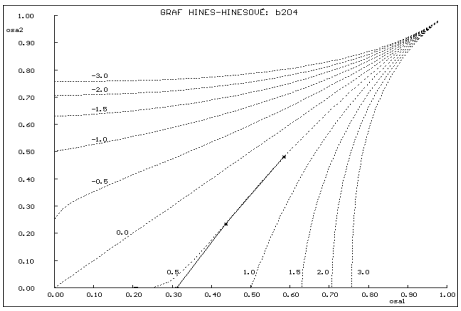
(a) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$:	0.177
Odhad rozptylu s^2:	2.536E-02
Odhad směrodatné odchylky s:	0.159
Odhad šikmosti $\hat{g}_1$:	1.54
Odhad špičatosti g_2:	5.36
(b) Test normality: tabulkový kvantil $\chi^2_{1-\alpha}(2)$:	5.992
Odhad χ^2_{exp} statistiky:	35.87
Závěr: Předpoklad normality zamítnut na spočtené hladině významnosti α = 1.6254E-08	
(c) Test nezávislosti: tabulkový kvantil $t_{1-\alpha/2}(n+1)$:	2.0141
Odhad von Neumannovy statistiky t_n:	0.4049
Závěr: Předpoklad nezávislosti přijat na spočtené hladině významnosti α = 0.3437	
(d) Detekce odlehlých hodnot: metodou modifikované vnitřní hradby	
Dolní vnitřní hradba B_D:	-0.3054
Horní vnitřní hradba B_H:	0.6800
Závěr: Ve výběru jsou 2 odlehlé hodnoty.	
Bod číslo 14 (horní):	0.6700
Bod číslo 23 (horní):	0.6800
Odhady parametrů s vynechanými odlehlými hodnotami:	
Odhad aritmetického průměru $\bar{x}$:	0.153
Odhad rozptylu s^2:	1.391E-02
Odhad směrodatné odchylky s:	0.118
Odhad šikmosti $\hat{g}_1$:	0.89
Odhad špičatosti g_2:	3.40

3. Transformace dat: Jelikož se na základě průzkumové analýzy dat zjistilo, že rozdělení výběru dat se systematicky odlišuje od rozdělení normálního, vyvstává zde problém, jak data vůbec vyhodnotit. V takovém případě je vhodná transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a v případě Boxovy-Coxovy transformace i k normalitě.

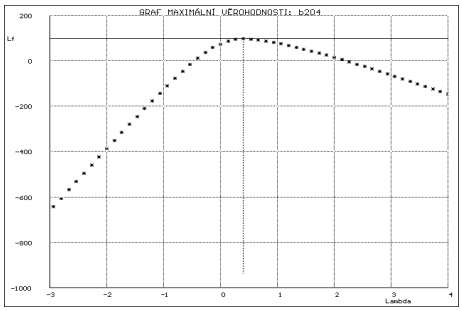
(a) *Prostá mocninná transformace:* Zesymetričtění rozdělení výběru je možné provést užitím prosté mocninné transformace, která sice nezachovává měřítko, není vzhledem k hodnotě λ všude spojitá a hodí se pouze pro kladná data.

Pro odhad exponentu λ se hledají optimální hodnoty charakteristik asymetrie (šikmosti) a špičatosti. K určení optimálního λ lze ale také užít orientační grafické metody, selekčního grafu dle Hinese a Hinesové, obr. 2.21a. Podle umístění experimentálních bodů v okolí

teoretických křivek selekčního grafu byla odhadnuta $\hat{\lambda}$. 0.5.

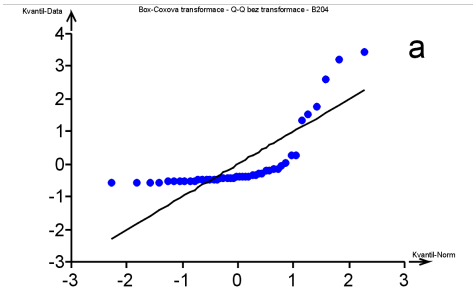


Obr. 2.21a Hinesův-Hinesové selekční graf pro výběr z exponenciálního rozdělení, ***ADSTAT.***

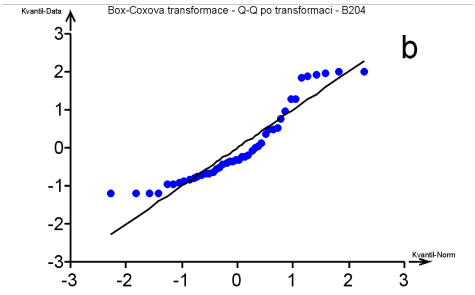


Obr. 2.21b Graf logaritmu věrohodnostní funkce na λ pro výběr z exponenciálního rozdělení, ***ADSTAT.***

(b) *Boxova-Coxova transformace:* Pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti se užívá Boxovy-Coxovy transformace. Pro odhad parametru λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti (obr. 2.21b) s tím, že pro $\lambda = \hat{\lambda}$ je rozdělení transformované veličiny y nejblíže normálnímu, $N(\mu_y, \sigma^2(y))$.



Obr. 2.22a *Q-Q* graf původních dat, ***ADSTAT***.



Obr. 2.22b *Q-Q* graf dat po Boxově-Coxově transformaci, ***ADSTAT.***

Průběh věrohodnostní funkce ln $L = f(\lambda)$ lze znázornit ve zvoleném intervalu, např. $-3 \# \lambda \# 3$, a identifikovat maximum křivky v grafu věrohodnostní funkce tak, že souřadnice x indikuje odhad $\hat{\lambda}$. Dva průsečíky křivky ln $L(\lambda)$ s rovnoběžkou s osou x indikují 100(1-α)% interval spolehlivosti parametru λ, tj. $+\lambda_D, \lambda_H$, . Jelikož tento interval spolehlivosti neobsahuje číslo +1, jsou mocninná a Boxova-Coxova transformace ze statistického hlediska výhodné a má smyl je užívat.

Zpětná transformace: Po vhodné transformaci určíme $\bar{y}$, $s^2(y)$ a potom pomocí zpětné transformace využitím Taylorova rozvoje v okolí $\bar{y}$ = 0.35465 odhadneme retransformované parametry původních dat $\bar{x}_R = 0.14318$ a $s^2_R = 0.020931$. Uvedený postup

vede vesměs k lepším odhadům polohy a rozptylu a je vhodný zvláště v případech asyme-

trického (exponenciálního) rozdělení výběru.
Závěr: Výběr pochází z exponenciálního rozdělení, a proto nejlepší odhad střední hodnoty získáme transformací dat, a to $\bar{x}_R = 0.143$ a směrodatnou odchylku $s = 0.145$. Konečně 95%ní interval spolehlivosti bude $L_D = 0.102$ a $L_H = 0.190$.

2.5 Úlohy

Úlohy jsou uspořádány do pěti kapitol: B2 (farmakologická a biochemická data), C2 (chemická a fyzikální data), E2 (environmentální, potravinářská a zemědělská data), H2 (hutní a mineralogická data) a S2 (ekonomická a sociologická data). Data jsou uváděna ve zkrácené podobě, celé datové soubory jsou dostupné na přiloženém kompaktním disku.

2.5.1 Analýza farmakologických a biochemických dat

Úloha B2.01 *Typ rozdělení obsahu léčiva v krvi u náhodně vybraných pacientů*
Byl sledován obsah léčiva v krvi u náhodně vybraných pacientů, (str. 143 v cit.[13]). Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru a rozhodněte o typu rozdělení. Vyšetřete předpoklady o náhodnosti a normalitě výběru a sestrojte histogram. Obsahují data za předpokladu, že pocházejí z normálního rozdělení, nějaké odlehlé body? Jaký maximální obsah léčiva v krvi má 75% pacientů? Odhadněte, jaká je hloubka prvku 4.00?
Data: Obsah léčiva v krvi [mg. l^{-1}]:

3.86	4.06	3.67	3.97	3.76	3.61	3.76	4.26	3.52	3.96
..	..	..	..	..	..	..	..	..	..
4.08	4.04	3.78	3.98	3.81	3.92	3.73	4.16	4.18	3.57

Úloha B2.02 *Symetrie rozdělení obsahu účinné látky v tabletě*
V jistém přípravku byl stanoven obsah účinné látky A v mg na tabletu. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru a rozhodněte o typu rozdělení. Určete kvantilové parametry polohy a rozptýlení a prověřte předpoklady o výběru, kladené na reprezentativní náhodný výběr. Je rozdělení výběru symetrické? Obsahuje výběr nějaké odlehlé body? Určete, jaká je hloubka dolního a horního kvartilu?
Data: Obsah účinné látky A [mg/tbl]:

0.6544	0.6121	0.6438	0.6510	0.6592	0.6525	0.6515	0.6545	0.6519	0.6504
0.6416	0.6656	0.6626	0.6546	0.6342	0.6413	0.6588	0.6531	0.6461	0.6196

Úloha B2.03 *Ověření symetrie rozdělení obsahu fenitrothionu v METATION E50*
Stanovení obsahu fenitrothionu v přípravku METATION E50 bylo provedeno metodou plynové chromatografie. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmost S_L, pseudosigmu G_L a délky konců T_L pro kvartily a oktily a ukažte jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti, (symboly viz cit.[19]). Vyšetřete předpoklad symetrie a určete parametry polohy a rozptýlení. Které diagnostiky ukazují, že ve výběru jsou odlehlá měření? Jaké rozdělení prokazuje kruhový

graf? Na jaký tvar rozdělení ukazuje koeficient šikmosti a špičatosti? Jaký maximální obsah vykazuje 90% přípravků METATION E50?

Data: Obsah fenitrothionu [mg. l^{-1}]:

49.82	49.91	50.00	50.10	50.13	50.19	50.22	50.23	50.25	50.29
..	..	..	..	..	..	..	..	..	..
50.57	50.72	50.92	50.40						

Úloha B2.04 *Typ rozdělení a odlehlé body obsahu ergosterinu v kalciferolu*

Při výrobě kalciferolu se provádí kontrola meziproduktu 3,5 DNB esteru kalciferolu metodou HPLC. Sleduje se také obsah přítomného ergosterinu jako nečistoty, jehož střední hodnota by neměla přesáhnout 0.4 %. Metodou průzkumové analýzy dat vyšetřete, zda jsou splněny požadavky, kladené na náhodný výběr a zda je splněn i požadavek čistoty kalciferolu. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmost S_L, pseudosigmu G_L a délky konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti, (symboly viz cit.[19]). Které diagnostiky shodně indikují vybočující hodnoty? Jak velké procento hodnot dosahuje obsahu 0.4 % a nižšího?

Data: Obsah ergosterinu [%]:

0.1200	0.2240	0.0660	0.0770	0.3400	0.1690	0.0430	0.0450	0.0001	0.0330
..	..	..	..	..	..	..	..	..	..
0.1300	0.0580	0.0001	0.0001						

Úloha B2.05 *Určení procenta aberantních buněk při cytogenetické analýze krve*

Součástí biologického monitoringu v rámci akce Monitoring zdraví a životního prostředí je i cytogenetická analýza krve. Předložená data jsou počty aberantních buněk (% AB. B.) ve vzorcích krve, odebraných v jednom okrese. Aberantní buňka představuje buňku, v jejíž genetické informaci došlo ke změně, např. ke zlomu chromozonu. Bylo vždy měřeno 100 buněk a počty aberantních buněk jsou celá čísla. Určete typ diskrétního rozdělení a odhad střední hodnoty.

Data: Procento aberantních buněk [% AB. B]:

2	2	3	4	2	2	2	3	0	3	2	4	2	1	4	3	3	0	1	2
1	6	0	4	2	1	1	4	0	1	3	1	3	4	1	1	0	5	4	3
1	2	3	1	2	4	3	2	4											

Úloha B2.06 *Maximální obsah zinku v krvi u 75 % vybraných dětí*

Součástí biologického monitoringu je i sledování koncentrace esenciálních prvků v lidském organismu. Data představují obsah zinku v mg. l^{-1} krve dětí z jednoho okresu za období jednoho roku. Jsou splněny předpoklady o náhodnosti a normalitě výběru? K analýze použijte písmenové hodnoty M, F_D, F_H, E_D, E_H, D_D a D_H a zkonstruujte pak i bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Určete míry polohy rozptýlení a tvaru. Kolik lze ve výběru indikovat vyšších odlehlých hodnot? Jaký maximální obsah zinku v krvi dosahuje 75 % dětí?

Data: Koncentrace zinku v krvi dětí [mg. l^{-1} Zn] krve:

3.44	5.13	4.33	6.18	5.11	4.88	5.23	4.97	4.38	4.92
..	..	..	..	..	..	..	..	..	..

5.87

Úloha B2.07 *Předpoklady o výběru cyclosporinu u dvou rozdílných standardů*
Během studie biologické dostupnosti cyclosporinu A byla denně sledována stabilita podmínek analýzou dvou standardních roztoků o známé koncentraci cyclosporinu A, a to 20 ng/ml a 80 ng/ml. Průzkumovou analýzou dat vyšetřete oba výběry, jejich rozdělení, předpoklady o výběru a odlehlé body. K analýze použijte písmenové hodnoty *M,* F_D , F_H, E_D , E_H, D_D a D_H a zkonstruujte pak i bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Jaké rozdělení indikuje *Q-Q* graf a kruhový graf? Je v tomto případě nutná transformace dat? Jsou rozdělení obou výběrů stejného charakteru?
Data:
B207a: Výběr o 20 ng/ml cyclosporinu A:

19.99	19.84	19.92	19.89	19.98	20.13	19.90	20.01	19.98	20.06
..	..	..	..	..	..	..	..	..	..
19.68	20.04	19.88	20.09	19.81	19.83	19.75	19.95	19.90	20.10

B207b: Výběr o 80 ng/ml cyclosporinu A:

80.18	80.07	79.56	80.28	80.02	79.85	80.15	79.76	79.96	80.03
..	..	..	..	..	..	..	..	..	..
80.24	80.18	79.95	80.03	79.71	80.12	80.02	80.12	79.86	80.00

Úloha B2.08 *Vyšetření hladiny penicilinu v séru pacientů po 50 minutách od aplikace*
Při studii biologické dostupnosti léků byla stanovena hladina koncentrace penicilinu v séru zdravých dobrovolníků vysokotlakou kapalinovou chromatografií. Proveďte statistické vyšetření výběru dat hladiny penicilinu v séru u skupiny zdravých dobrovolníků 50 minut po podání. Jsou splněny předpoklady reprezentativního náhodného výběru? Vyčíslete písmenové hodnoty *M,* F_D , F_H , E_D , E_H , D_D a D_H a pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. O jaké rozdělení se jedná?
Data: Hladina penicilinu [mg. l^{-1}]:

2.100	0.900	1.980	1.890	1.114	1.200	1.345	1.560	1.584	1.567
..	..	..	..	..	..	..	..	..	..
2.150	1.256	1.980	1.650	1.750	2.030	1.660	1.230	1.563	1.720

Úloha B2.09 *Vyšetření hladiny penicilinu v séru pacientů po 90 minutách od aplikace*
Při studii biologické dostupnosti léků byla stanovena hladina penicilinu v séru zdravých dobrovolníků vysokotlakou kapalinovou chromatografií. Proveďte statistické vyšetření výběru dat hladiny penicilinu v séru u skupiny zdravých dobrovolníků 90 minut po podání. Zkonstruujte bariérově-číslicové schéma formou pětipísmenového zápisu výběru. Jsou ve výběru nějaké odlehlé hodnoty?
Data: Hladina penicilinu [mg. l^{-1}]:

0.732	0.732	0.712	0.753	0.654	0.720	0.70l	0.762	0.770	0.704
..	..	..	..	..	..	..	..	..	..
0.725	0.756	0.740	0.722	0.745	0.778	0.721	0.762	0.752	0.735

Úloha B2.10 *Vyloučení odlehlých hodnot obsahu chenodeoxykalciferolu ve vitaminu D*
V kontrolní laboratoři je průběžně sledován obsah chenodeoxykalciferolu ve vitaminu D,

kalciferolu. Byla sledována produkce za 1 měsíc. Vyčíslete písmenové hodnoty *M,* F_D , F_H, E_D , E_H , D_D a D_H metodou pořadí a hloubek. Podle závěrů průzkumové analýzy EDA odhadněte kvantilové odhady parametrů polohy a rozptýlení. Existují v tomto výběru odlehlé hodnoty, které je možno vyloučit? Jde o symetrické rozdělení?

Data: Obsah chemodeoxykalciferolu ve vitaminu D [%]:

0.6	1.0	0.9	0.9	0.6	0.9	0.6	0.3	0.7	0.5	0.9	0.7	0.8	0.8	0.5	0.6	0.6	0.9	1.5	0.7
0.7	1.0	0.8	0.9	0.8	0.7	0.7	0.7	0.7	0.8	1.0	0.9	0.8	0.8	0.8	1.0	1.2	0.7	0.7	

Úloha B2.11 *Porovnání sedimentace krve ve skleněných a v plastikových trubičkách*
U 302 vzorků nesrážlivé krve z laboratorního provozu byla stanovena sedimentace ve skleněných a plastikových trubičkách za 1 a za 2 hodiny. Měřením byly získány čtyři skupiny dat, odpovídající rychlostem sedimentace v mm za 1 a 2 hodiny ve skle (*B211a* a *B211b*) a v plastické hmotě (*B211c* a *B211d*). Porovnejte rozdělení čtyř výběrů. Jaký kvantilový odhad polohy zvolíte, když ve výběru naleznete odlehlé hodnoty? Dáte přednost robustním odhadům nebo mocninné transformaci? Určete kvantilové míry polohy a rozptýlení.

Data: Hodnoty sedimentace [mm] jsou v pořadí *B211a, B211b, B211c, B211d*:

34, 66, 33, 61;	7, 18, 7, 20;	22, 49, 19, 49;	9, 24, 5, 20;	2, 5, 3, 6;
.....	.....	.....	.....	.....
11, 28, 11, 32;	5, 14, 6, 16;	2, 5, 4, 10;	4, 10, 4, 12;	5, 14, 6, 15;
9, 22, 10, 27;	2, 5, 4, 12;			

Úloha B2.12 *Vyšetření rozdělení přírůstku hmotnosti chovných jalovic*
Při vyhodnocování výsledků nové metody chovu jalovic byly zjištěny přírůstky hmotnosti v kg za určité období u 100 jalovic. Komentujte rozdělení výběru a rozhodněte, který kvantilový odhad výběrového parametru polohy a rozptýlení zde bude nejlepší. Vyšetřete statistické zvláštnosti (stupeň symetrie, špičatosti, lokální koncentrace dat, vybočující hodnoty)? Vyšetřete tvar a symetrii rozdělení na základě grafu polosum, symetrie, špičatosti a diferenčního kvantilového grafu. Kolik procent jalovic dosáhlo maximálního přírůstku 80 kg?

Data: Přírůstky hmotnosti jalovic [kg]:

59.8	61.6	62.1	62.3	62.8	65.2	65.4	65.7	66.7	67.4
..	..	..	..	..	..	..	..	..	..
92.8	93.7	94.2	94.7	94.9	96.1	96.9	98.1	99.1	100.4

Úloha B2.13 *Vyšetření rozdělení porodní délky živě narozených hochů*
Vyšetřete rozdělení výběru 150 živě narozených hochů a navrhněte vhodný odhad polohy. Jaké procento hochů dosáhlo maximální délky 50 cm? Vyčíslete písmenové hodnoty *M,* F_D, F_H , E_D , E_H , D_D a D_H. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Odhadněte tvar rozdělení na základě písmenových hodnot v předešlém schématu. O jaké jde rozdělení?

Data: Data jsou rozdělena dle tříd {délka novorozence [cm]; počet novorozenců této délky [četnost]}:

35; 1	37; 1	40; 1	43; 1	44; 1	45; 2	46; 3	47; 5	48; 12
49; 20	50; 32	51; 29	52; 20	53; 11	54; 7	55; 2	56; 1	57; 1

Úloha B2.14 *Vyšetření rozdělení porodní délky živě narozených děvčat*

Vyšetřete rozdělení výběru 33 živě narozených děvčat a navrhněte vhodný odhad polohy. Jaké procento děvčat dosáhlo maximální délky 50 cm? Vyčíslete písmenové hodnoty M, F_D, F_H, E_D, E_H, D_D a D_H. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. O jaké jde rozdělení?

Data: Data jsou rozdělena dle tříd {délka novorozence [cm]; počet novorozenců této délky [četnost]}:

42; 1	44; 1	46; 1	48; 3	49; 4	50; 8	51; 6	52; 4	53; 3	54; 1	56; 1

Úloha B2.15 *Počet dnů pracovní neschopnosti u onemocnění rýmou*

Uvažujme diskrétní kvantitativní znak, a to trvání pracovní neschopnosti ve dnech u onemocnění horních cest dýchacích, tj. rýmou. Bylo zaznamenáno 51 případů pracovní neschopnosti. Určete vhodný typ diskrétního rozdělení a odhadněte jeho parametry.

Data: Počet případů pracovní neschopnosti [dny]:

20	9	13	11	7	13	8	12	13	10	9	8	9	12	11	13	9	8	10	11
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
13	12	10	8	7	8	12	10	11	9	11									

Úloha B2.16 *Rozdělení výběru sedimentace červených krvinek*

Z naměřených hodnot sedimentace červených krvinek u 40 lidí určete typ diskrétního rozdělení výběru a vhodný odhad střední hodnoty a míry variability. Jsou v datech vyšší odlehlé hodnoty, indikující nemocného pacienta? O jaké rozdělení se jedná?

Data: Hodnoty sedimentace krve [mm/h]:

4	5	10	6	6	7	5	28	5	7	11	9	6	5	9	9	5	7	8	5
6	8	6	7	12	9	6	27	8	7	10	9	6	5	9	9	7	7	8	9

Úloha B2.17 *Typ rozdělení výběru koncentrace fibrinogenu v krevní plazmě*

Nalezněte typ rozdělení koncentrace fibrinogenu v krevní plazmě u 50 pacientů a určete vyšší odlehlé hodnoty nemocných pacientů, a to především těch, jež přesahují normu 3 až 5 g. l^{-1}. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují vybočující hodnoty?

Data: Úroveň fibrinogenu v krevní plazmě[g. l^{-1}]:

6.68	3.24	4.16	3.21	3.64	3.61	2.22	3.88	4.08	3.07
..	..	..	..	..	..	..	..	..	..
4.99	3.76	4.05	4.15	3.28	3.76	4.11	3.72	3.57	3.43

Úloha B2.18 *Typ rozdělení u tří různých typů globulinu v séru*

Zdravý pacient vykazuje tyto typické hodnoty bílkovin séra: 4.6 g. l^{-1} α-globulinů, 3.4 g. l^{-1} β-globulinů a 11.7 g. l^{-1} γ-globulinů. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Určete rozdělení a vhodné kvantilové odhady střední hodnoty a variability pro výběr 40 pacientů. Obsahují výběry nějaké odlehlé hodnoty? Jaké procento pacientů dosáhlo hodnot typického zdravého pacienta? Je nutná mocninná transformace dat?

Data:

B218a: α_1-globuliny v séru [g. l^{-1}]:

3.51	3.73	3.59	3.50	3.37	4.08	3.37	3.86	3.90	3.59
..	..	..	..	..	..	..	..	..	..
3.56	3.80	3.81	3.57	3.35	3.51	3.99	3.72	3.60	3.41
B218b: β-globuliny v séru [g. l^{-1}]:									
7.22	6.58	6.91	7.22	7.32	7.49	7.35	7.55	6.82	6.84
..	..	..	..	..	..	..	..	..	..
6.97	6.87	7.55	7.02	7.41	6.39	6.72	7.32	6.89	7.50
B218c: γ-globuliny v séru [g. l^{-1}]:									
12.7	11.5	12.5	11.7	12.5	12.4	12.6	11.9	14.0	12.5
..	..	..	..	..	..	..	..	..	..
11.7	12.2	12.1	10.7	13.2	11.8	11.2	11.5	11.6	12.4

Úloha B2.19 *Odlehlé hodnoty obsahu celkového HDL cholesterolu u zdravých pacientů*
Určete rozdělení výběru 40 pacientů a nalezněte odlehlé, vyšší hodnoty nemocných pacientů, když lékařská norma pro zdravé pacienty uvádí hodnoty celkového cholesterolu HDL v rozmezí 1.6 - 2.7 g. l^{-1}. Jaká je střední hodnota uvedeného výběru a míra rozptýlení? Je rozdělení symetrické? Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Dáte přednost robustnímu odhadu střední hodnoty nebo mocninné transformaci?

Data: HDL cholesterol celkový [g. l^{-1}]:

1.61	1.80	2.62	2.55	3.02	2.15	2.33	3.82	1.47	1.41
..	..	..	..	..	..	..	..	..	..
1.97	1.34	2.00	2.59	1.92	1.71	2.36	2.47	7.52	2.11

Úloha B2.20 *Symetrie rozdělení obsahu bílkoviny v moči*
Jsou rozdělení výběru symetrická a dále střední hodnota obsahu bílkoviny v moči u výběru zdravých a výběru nemocných pacientů stejná? Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti u každého výběru (symboly viz cit.[19]). Které diagnostiky shodně indikují vybočující hodnoty? Jaké procento pacientů dosáhlo maximální hodnoty u výběru uzdravených a u výběru nemocných pacientů?

Data: Obsah bílkoviny v moči [mg/24 h]: *B220a* zdraví lidé, *B220b* nemocní lidé:

B220a:	44.8	44.4	40.0	39.5	56.9	47.4	41.8	50.5
..	..	..	..	..	..	..	..	..
43.3	53.3	35.4	38.1	58.3				
B220b:	72.74	32.42	55.07	67.10	56.19	38.01	66.45	82.68
..	..	..	..	..	..	..	..	..
48.99	49.34	74.13	87.90	65.59				

Úloha B2.21 *Odlehlé hodnoty obsahu účinné látky tablet Blocalcin 60 mg*
V laboratoři kontroly léčiv je stanovován obsah účinné látky kardiovaskulárních tablet Blocalcin 60 mg. Vyhodnoťte obsah účinné látky, stanovený u 32 náhodně odebraných tablet. Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti, (symboly

viz cit.[19]). Jde o Gaussovo rozdělení? Jsou ve výběru odlehlé hodnoty? Diskutujte, zda k odhadu střední hodnoty využijete robustního odhadu polohy nebo mocninné transformace.

Data: Obsah tablety [mg]:

60.44	59.60	58.40	59.68	59.22	59.73	59.97	58.91	58.28
..	..	..	..	..	..	..	..	..
58.59	59.77	59.08	59.10	59.08				

Úloha B2.22 *Typ rozdělení obsahu kadmia ve vlasech*

U 98 náhodně vybraných lidí byla provedena analýza vlasů pro stanovení obsahu kadmia. Vlasy byly po promytí a odmaštění mineralizovány v přístroji Apion směsí kyslíku a amoniaku. Obsah kadmia byl v mineralizátu stanoven metodou AAS a přepočten na původní navážku vzorku vlasů. Nalezněte typ rozdělení, odlehlé body a odhad střední hodnoty. Využijte také schéma sumarizace dat. Zdůvodněte, jakému odhadu dáte přednost.

Data: Obsah kadmia ve vlasech [mg/kg]:

0.558	0.063	0.442	0.049	0.041	0.044	0.380	0.630	0.179	0.179
..	..	..	..	..	..	..	..	..	..
0.164	0.064	0.229	0.065	0.157	0.060	0.951	0.374.		

Úloha B2.23 *Sedmipísmenový zápis výběru koncentrace albuminu u dárců krve*

Nalezněte typ rozdělení koncentrace albuminu u 50 dárců krve. Vyčíslete písmenové hodnoty M, F_D, F_H, E_D, E_H, D_D *a* D_H a pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L , rozpětí R_L , šikmosti S_L , pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Splňuje výběr předpoklady o náhodnosti výběru? O jaké rozdělení se jedná? Určete střední hodnotu a míru rozptýlení výběru.

Data: Koncentrace albuminu v krvi [g. l^{-1}]:

41,9	42.5	42.3	44.9	47.1	41.6	42.0	46.6	41.4	41.7
..	..	..	..	..	..	..	..	..	..
44.9	46.3	46.6	47.1	41.8	41.5	42	42.4	42.8	42.7

Úloha B2.24 *Typ rozdělení koncentrace HDL-cholesterolu v krvi pacientů*

Nalezněte typ rozdělení koncentrace HDL cholesterolu v krvi, stanoveného u reprezentativního náhodného výběru 50 pacientů. Určete odlehlé hodnoty nemocných pacientů, když lékařská norma zdravých jedinců je 0.90 až 2.00 mmol. l^{-1}. Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Dáte raději přednost robustnímu odhadu střední hodnoty nebo mocninné transformaci? Zdůvodněte, jaký zvolíte odhad míry polohy výběru a míry rozptýlení ?

Data: Koncentrace HDL cholesterolu [mmol. l^{-1}]:

1.96	1.11	1.51	1.71	1.14	3.30	1.23	1.05	2.01	1.25
..	..	..	..	..	..	..	..	..	..
1.37	0.92	1.39	1.64	0.98	1.82	0.87	2.19	1.46	1.82

Úloha B2.25 *Tvar rozdělení koncentrace kyseliny močové v krvi dárců krve*

Určete typ rozdělení stanovení kyseliny močové u 50 dárců krve. Vyšetřete tvar rozdělení

na základě grafu polosum, symetrie a špičatosti, (symboly viz cit.[19]). Jsou ve výběru odlehlé hodnoty? Jaká je střední hodnota a míra rozptýlení uvedeného výběru? Zdůvodněte svoji volbu.

Data: Koncentrace kyseliny močové v krvi [μmol. l^{-1}]:

295	325	320	316	314	310	306	304	298	297
..	..	..	..	..	..	..	..	..	..
295	354	353	349	348	340	339	329	326	290

Úloha B2.26 *Exploratorní analýza koncentrace triacylglycerolů v krevním séru*

Nalezněte typ rozdělení koncentrace triacylglycerolů v krevním séru u náhodného výběru 50 pacientů. Testujte, zda data splňují základní požadavky o reprezentativním náhodném výběru a pokuste se o zesymetričtění rozdělení využitím mocninné transformace. Rozhodněte, zda míry polohy a rozptýlení bude výhodnější stanovit momentovými nebo kvantilovými odhady. Zjistěte nejlepší odhad střední hodnoty a míry variability nalezeného rozdělení. Určete procento nemocných pacientů, je-li lékařská norma 0.20 - 2.25 mmol. l^{-1}.

Data: Koncentrace triacylglycerolů v krvi [mmol. l^{-1}]:

3.27	1.23	1.23	0.48	0.71	1.81	1.31	1.95	2.27	2.62
..	..	..	..	..	..	..	..	..	..
1.33	0.76	1.01	1.38	1.71	2.41	0.54	1.20	1.82	1.74

Úloha B2.27 *Exploratorní analýza hmotnosti prázdných tub u tubovací linky*

Při validaci provozu tubovací linky byla měřena hmotnost prázdných tub, určených pro balení sterilního masťového přípravku. Analýzou hmotnosti prázdných tub posuďte kvalitu tubovací linky. Vyčíslete písmenové hodnoty M, F_D , F_H , E_D, E_H, D_D a D_H, pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L , šikmosti S_L , pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Splňuje výběr předpoklady o náhodnosti výběru? O jaké rozdělení se jedná? Určete střední hodnotu a míru rozptýlení výběru. Dosahuje míra rozptýlení nepřijatelně vysoké hodnoty?

Data: Hmotnost prázdných tub [g]:

3.49	3.18	3.08	3.13	3.74	3.18	3.08	3.15	3.19	3.17	3.08	3.15
..	..	..	..	..	..	..	..	..	..	..	..
3.18	3.06	3.11	3.08	3.17	3.08	3.14					

Úloha B2.28 *Homogenita hmotnosti výběru náhodně vybraných potkanů*

Ke sledování nově zaváděného farmaka eroprilu bylo třeba vybrat reprezentativní náhodný výběr 40 potkanů přibližně stejné hmotnosti. Vyšetřete předpoklady o výběru a sledujte především homogenitu výběru a symetrii výběrového rozdělení. Pomocí rankitového grafu rozhodněte, zda jde o Gaussovo rozdělení? Určete rovněž kvantilovou míru polohy a rozptýlení.

Data: Hmotnost 40 potkanů [g]:

243	246	258	236	254	237	248	246	270	253	264	238
..	..	..	..	..	..	..	..	..	..	..	..
249	256	250	246								

Úloha B2.29 *Typ rozdělení a homogenita hladin prolaktinu u výběru žen*
U náhodného výběru 74 pacientek byla metodou ELISA změřena (a) koncentrace hormonu prolaktin [µg. l^{-1}] a (b) koncentrace růstového hormonu [µg. l^{-1}]. Exploratorní analýzou dat vyšetřete, zda jde o nezávislá a homogenní data s normálním rozdělením. Určete kvantilové odhady polohy a rozptýlení a rozhodněte, zda je vhodné užít transformaci dat.
Data:
(a) *B229a*: koncentrace prolaktinu [µg. l^{-1}],

1.63	4.00	4.83	4.91	4.84	5.56	4.89	2.11	2.21	5.33
..	..	..	..	..	..	..	..	..	..
9.70	24.71	24.39	7.58						

(b) *B229b:* koncentrace růstového hormonu [µg. l^{-1}],

22.20	0.47	0.66	0.44	0.46	0.47	0.46	4.90	22.80	0.47	0.49	9.90	0.47
..	..	..	..	..	..	..	..	..	..	..	..	..
2.10	14.60	1.80	0.64	16.80	1.30	10.50	21.50	0.48				

2.5.2 Analýza chemických a fyzikálních dat

Úloha C2.01 *Statistické zvláštnosti výběru obsah isopropylaminu v surovém produktu*
Testujte, zda data obsahu isopropylaminu v surovém produktu, shromážděná za týden v chemickém závodě, splňují základní požadavky na výběr. Z písmenových hodnot usuzujte na symetrii výběrového rozdělení. Jaká je hloubka horního a dolního kvartilu? Co zvolíte za vhodný odhad střední hodnoty obsahu isopropylaminu, momentový odhad nebo kvantilový? Které diagnostiky shodně indikují odlehlé body? Kolik procent hodnot obsahu isopropylaminu leží pod hodnotou 4.9?
Data: Obsah isopropylaminu v surovém produktu [%]:

2.05	5.07	5.43	5.02	5.24	4.92	5.50	7.56	4.61	5.14
..	..	..	..	..	..	..	..	..	..
4.56	4.47	4.30	9.63	4.07	4.46	4.70	4.42	4.80	4.82

Úloha C2.02 *Rozdělení obsahu acetonu v surovém produktu isopropylaminu*
Proveďte rozbor dat obsahu acetonu v surovém isopropylaminu. Testujte, zda uvedený výběr, shromážděný za týden v chemickém závodě, splňuje předpoklady o výběru. Je konstantní délka intervalu histogramu optimální i pro zešikmená rozdělení? Porovnejte histogram s jádrovým odhadem hustoty pravděpodobnosti a určete typ rozdělení. Na základě korelačního koeficientu shody výběrového *Q-Q* grafu s teoretickým rozhodněte, z jakého rozdělení výběr pochází. Jsou v datech vybočující hodnoty?
Data: Obsah acetonu v surovém isopropylaminu [%]:

1.61	0.36	0.02	0.03	0.03	0.03	0.12	0.07	0.36	0.11
..	..	..	..	..	..	..	..	..	..
0.15	0.12	0.20	0.16	0.90	0.22	0.14	0.22	0.26	0.15

Úloha C2.03 *Základní předpoklady o výběru obsahu anilinu v surovém cyklohexylaminu*
Cyklohexylamin je vyráběn hydrogenací anilinu. Po dobu jednoho měsíce byl sledován obsah anilinu v procentech v surovém produktu cyklohexylaminu po proběhlé hydrogenaci.

Vyčíslete písmenové hodnoty *M*, F_D , F_H , E_D , E_H , D_D a D_H a pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L , pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Splňuje výběr předpoklady o náhodnosti výběru? O jaké rozdělení se jedná? Rozhodněte, zda míry polohy a rozptýlení bude výhodnější stanovit momentovými nebo kvantilovými odhady.

Data: Obsah anilinu v surovém isopropylaminu [%]:

0.10	0.58	0.21	0.68	0.13	0.46	0.22	0.85	0.04	0.37
..	..	..	..	..	..	..	..	..	..
0.73	0.09	0.73	0.86	1.26					

Úloha C2.04 *Symetrie rozdělení obsahu kyseliny chromotropové v organické látce*

V laboratoři chemického závodu je průběžně sledován obsah kyseliny chromotropové ve vyráběné organické látce. Vyčíslete písmenové hodnoty *M*, F_D , F_H , E_D , E_H , D_D a D_H a pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L , pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Lze posoudit symetrii rozdělení na základě dolních a horních kvartilů, oktilů a sedecilů? Splňuje výběr předpoklady o náhodnosti výběru? O jaké rozdělení se jedná? Určete střední hodnotu a míru rozptýlení výběru.

Data: Obsah kyseliny chromotropové v organické látce [%]:

46	39	65	40	48	38	59	49	48	37	50	40	50	45	37
41	48	50	49	43	47	40	53	45	60	58	41	54	45	58

Úloha C2.05 *Statistické zvláštnosti výběru obsahu zbylého anilinu v dicyklohexylaminu*

Dicyklohexylamin je vyráběn hydrogenací anilinu. Proveďte rozbor zbylého obsahu anilinu v surovém dicyklohexylaminu. Testujte, zda uvedená data, shromážděná za týden v chemickém závodě, splňují základní požadavky na náhodný výběr. Posuďte symetrii rozdělení především dle grafu polosum a grafu symetrie. Kolik procent hodnot obsahu anilinu leží nad hodnotou 0.20? Kolik procent hodnot výběru leží pod dolním oktilem E_D a nad horním oktilem E_H? Jsou tato procenta shodná?

Data: Obsah anilinu v dicyklohexylaminu [%]:

0.03	0.10	0.15	0.14	0.06	0.08	0.39	0.14	0.19	0.08
..	..	..	..	..	..	..	..	..	..
0.09	0.36	0.19	0.09	0.25	0.14				

Úloha C2.06 *Procento velikosti granule umělého hnojiva mimo požadovaný interval*

V laboratoři technické kontroly chemického závodu se sleduje, zda velikost granulí kombinovaného hnojiva je menší než 1 mm. Testujte, zda výběr dat splňuje základní předpoklady, zejména předpoklad nezávislosti a normality. Vysvětlete tvar rozdělení dle kvantilových odhadů šikmosti a špičatosti. Jde o symetrické rozdělení? Kolik procent hodnot výběru velikosti granule leží pod hodnotou 0.3 a kolik nad 0.5?

Data: Velikost granule [mm]:

0.4	0.4	0.4	0.7	0.3	0.4	0.4	0.4	0.2	0.2	0.4	0.3	0.2	0.3	0.3	0.2	0.4	0.2	0.4
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
1.1	0.4	0.4	0.7	2.5	0.9	0.2	0.2	0.7	0.8									

Úloha C2.07 *Typ rozdělení velikosti výtěžku kontinuálního reaktoru*
Byl sledován procentuální výtěžek kontinuálního reaktoru v průběhu 25 dnů. Typ výběrového rozdělení hustoty pravděpodobnosti určete pomocí *Q-Q* grafu tak, že shodu s teoretickým rovnoměrným, normálním, exponenciálním, Laplaceovým a logaritmicko-normálním rozdělením ověříte korelačním koeficientem. Jaký odhad navrhujete pro míru polohy a rozptýlení, momentový nebo kvantilový? Komentujte požadavky, kladené na reprezentativní náhodný výběr a prověřte, zda-li jsou všechny splněny.
Data: Výtěžek reaktoru [%]:

64.97	64.60	62.37	64.12	61.97	63.22	61.60	62.85	61.12	64.60
..	..	..	..	..	..	..	..	..	..
61.97	67.87	63.22	64.97	62.85					

Úloha C2.08 *Homogenita vedlejšího produktu při výrobě čpavku*
Při výrobě čpavku vzniká jako vedlejší nežádoucí produkt plynný dusík, oddělovaný v čisticí jednotce. V průběhu směny bylo nasbíráno 60 měření obsahu dusíku v procentech. Vyšetřete předpoklady o výběru a ověřte, zda data pocházejí z Laplaceova nebo z normálního rozdělení. Kolik procent hodnot obsahu dusíku leží nad hodnotou 27.0? Kolik procent hodnot výběru leží pod dolním kvartilem a kolik nad horním kvartilem? Jsou tyto počty shodné, svědčící o symetrii rozdělení?
Data: Obsah dusíku [%]:

24.5	24.2	28.3	29.8	26.4	29.0	27.0	27.0	22.4	25.3
..	..	..	..	..	..	..	..	..	..
20.6	20.0	21.2	21.4	29.6	29.4	29.0	29.0	28.5	28.7

Úloha C2.09 *Ověření nezávislosti výběru analytického signálu*
Bylo sledováno kolísání signálu u měření na infračerveném spektrofotometru UR-10 (Zeiss, Jena). Ověřte, zda lze považovat jednotlivé údaje signálu za nezávislé a zda pocházejí z Gaussova rozdělení. Indikujte odlehlé hodnoty porovnáním více diagnostik. Která diagnostika nejlépe indikuje symetrii výběrového rozdělení? Existuje shoda alespoň tří diagnostik průzkumové analýzy dat o symetrii výběrového rozdělení?
Data: Infračervený signál:

2144	2234	2244	2196	2285	2255	2166	2068	2183	2166
..	..	..	..	..	..	..	..	..	..
2226	2198	2196	2179	2179	2286	2225	2216	2175	

Úloha C2.10 *Symetrie výběrového rozdělení bodu tání včelího vosku*
White, Reithof a Kushnir studovali vlastnosti včelího vosku za účelem detekce umělých, synteticky připravených vosků, které včelaři do pravého včelího vosku často míchají (str. 313 v cit.[17]). Přítomnost mikrokrystalů umělých vosků zvyšuje totiž bod tání včelího vosku. Jelikož se bod tání vosku mění od úlu k úlu, shromáždili autoři body tání pravého včelího vosku z 59 úlů za účelem výpočtu střední hodnoty. Zkonstruujte jádrový odhad hustoty pravděpodobnosti. Porovnejte závěry o tvaru rozdělení a především o jeho symetrii

analýzou *Q-Q* grafu a kruhového grafu. Vedou všechny diagnostiky ke stejným závěrům? Jakou hodnotu bodu tání vykazuje dolní a horní kvartil? Z kvantilového grafu odečtěte kolik procent hodnot výběru leží pod bodem tání 63.0 EC a kolik nad 64.0 EC?

Data: Bod tání pravého včelího vosku [°C]:

64.42	63.27	63.10	63.34	63.50	63.83	63.63	63.27	63.30	63.83
..	..	..	..	..	..	..	..	..	..
63.41	63.60	63.13	63.69	63.05	62.85	63.31	63.66	63.60	

Úloha C2.11 *Tvar rozdělení výběru obsahu uhlovodíků v pravém včelím vosku*

White, Riethof a Kushnir stanovili obsah uhlovodíků v pravém včelím vosku (str. 313 v cit.[17]). Uměle vytvořené syntetické vosky obsahují uhlovodíků podstatně více, někdy až 85 %. Zvýšený obsah uhlovodíků ve včelím vosku dokazuje, že včelař přidal do včelího vosk umělý. Testujte základní předpoklady o náhodném výběru a vyšetřete přítomnost především vyšších vybočujících hodnot. Pokuste se o zesymetričtění rozdělení výběru využitím mocninné transformace. Z kvantilového grafu odečtěte kolik procent hodnot výběru leží nad hodnotou obsahu uhlovodíků 15.0 %?

Data: Obsah uhlovodíků v pravém včelím vosku [%]:

14.27	14.80	12.28	17.09	15.10	12.92	13.66	14.68	14.32	14.01
..	..	..	..	..	..	..	..	..	..
15.03	13.68	13.65	14.57	13.83	15.40	15.31	15.02	15.49	

Úloha C2.12 *Podezřelá a odlehlá měření sublimačního tepla rhodia a iridia*

Hampson a Walker stanovili sublimační teplo rhodia a iridia (str. 341 v cit.[17]). Vyšetřete výběr jejich experimentálních hodnot, testujte nezávislost, homogenitu a normalitu rozdělení. Jde o symetrické rozdělení? Která diagnostika nejlépe indikuje asymetrii tohoto rozdělení? Určete počet vybočujících hodnot a hodnot pouze podezřelých. Diskutujte, který kvantilový odhad polohy bude pro tento výběr nejlepší.

Data: Sublimační teplo [kcal/mol], *C212a* rhodium, *C212b* iridium:

C212a:	126.4	135.7	132.9	131.5	131.1	131.1	131.9	132.7	
..	..	..	..	..	..	..	..	..	
133.5	132.3	132.7	132.9	134.1					
C212b:	136.6	145.2	151.5	162.7	159.1	159.8	160.8	173.9	160.1
..	..	..	..	..	..	..	..	..	..
160.2	160.1	160.0	159.7	159.5	159.5	159.6	159.5		

Úloha C2.13 *Odlehlé hodnoty u stanovení chloridů ve vodě*

V hydrochemických laboratořích firmy se zavedla alternativní metoda stanovení aniontů ve vodách pomocí iontové párové chromatografie. Z 30 opakovaných stanovení chloridů ve vzorku mineralizované vody určete parametry polohy a rozptýlení a prověřte předpoklady o náhodném výběru. Které diagnostiky indikace odlehlých bodů dospěly ke stejnému závěru? O jaký typ rozdělení se jedná? Dle typu rozdělení rozhodněte, zda odlehlé hodnoty ve výběru ponecháte. Je vhodnější využít transformaci dat? Kolik procent hodnot výběru leží nad 10.0 mg. l^{-1}? Jaká je hloubka tohoto prvku ve výběru?

Data: Koncentrace chloridů v mineralizované vodě Cl^- [mg. l^{-1}]:

10.0	10.0	10.1	10.3	10.0	9.4	10.0	10.3	10.3	10.1
..	..	..	..	..	..	..	..	..	..
10.1	9.9	10.2	10.6	10.0	10.7	10.7	10.2	10.1	10.1

Úloha C2.14 *Procento obsahu alifatických uhlovodíků v toluenu nad požadovanou mezí*
V toluenu se běžně stanovuje obsah alifatických uhlovodíků v procentech metodou plynové chromatografie. Určete parametry polohy a rozptýlení a prověřte předpoklady o výběru. Je vhodné užít transformaci dat? Je rozdělení výběru symetrické? Kolik je ve výběru odlehlých bodů, indikovaných shodně více diagnostikami? Z kvantilového grafu odečtěte kolik procent hodnot výběru leží nad požadovanou mezí obsahu 0.060 %? Jaká je hloubka tohoto prvku ve výběru?

Data: Obsah alifatických uhlovodíků v toluenu [%]:

0.05	0.03	0.01	0.04	0.04	0.04	0.04	0.05	0.02	0.03
..	..	..	..	..	..	..	..	..	..
0.07	0.05	0.06	0.08	0.09	0.19	0.07	0.08	0.07	0.06

Úloha C2.15 *Procento hodnot obsahu fenolů pod požadovanou mezí*
Obsah fenolů, těkajících s vodní parou v kontrolním roztoku, byl stanoven sérií opakovaných analýz. Určete o jaký typ rozdělení jde? Určete parametry polohy a rozptýlení a prověřte předpoklady o výběru, především nezávislost prvků, ve výběru na hladině významnosti $\alpha = 0.05$. Určete, zda naměřená koncentrace je shodná s očekávanou hodnotou 0.45 mg. l^{-1}. Kolik odlehlých bodů lze ve výběru bezpečně indikovat? Z kvantilového grafu odečtěte kolik procent hodnot výběru leží pod požadovanou mezí 0.45 mg. l^{-1}?

Data: Obsah fenolů v kontrolním roztoku [mg. l^{-1}]:

0.458	0.302	0.386	0.424	0.501	0.355	0.420	0.385	0.424	0.423
..	..	..	..	..	..	..	..	..	..
0.406	0.464	0.407	0.465	0.408	0.466	0.409	0.467	0.405	0.468

Úloha C2.16 *Odlehlé hodnoty a střední doba želatinace u pryskyřice UMAFORM B-118*
U fenolformaldehydové pryskyřice UMAFORM B-118 byly pro 60 šarží stanoveny doby želatinace za experimentálně stejných podmínek. Ke stanovení byl užit ocelový blok, vyhřívaný na 150 °C a jedna kapka pryskyřice. Proveďte rozbor hodnot naměřeného času pro jednotlivé šarže. Testujte, zda uvedený výběr, shromážděný za půl roku v chemickém závodě, splňuje základní předpoklady, kladené na reprezentativní náhodný výběr.

Data: Doba želatinace [s]:

42	40	36	37	38	43	43	30	38	32	44	39	45	44	43	39	45	39	34	32
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
38	35	44	34	40	37	36	39	45	40	41	40	38	40	36	33	35	32	43	40

Úloha C2.17 *Podezřelé a odlehlé hodnoty efektivního průměru pigmentů*
Analýzou vzorku Colanylschwarz PR 100 na přístroji Particle Sizer BI-90 byly stanoveny hodnoty efektivního průměru dispergovaných částic pigmentu ve vodném prostředí. Úkolem je testovat normalitu, nezávislost, homogenitu rozdělení výběru a vyšetřit, zda jsou ve výběru přítomny odlehlé hodnoty. Je vhodné odlehlé hodnoty vypustit? Je rozdělení symetrické? Kolik procent hodnot výběru leží nad 190 μm a kolik pod touto hodnotou?

Data: Efektivní průměr dispergované částice [μm]:

195	181	187	196	175	186	189	176	182	177	207	189	184	196	188	181	188	190	184	180
183	188	207	179	171	192	170	178	194	158	169	182	199	177	186	183	162	161	167	137

Úloha C2.18 *Odlehlé hodnoty výběru obsahu I-kyseliny*
V laboratoři oddělení kontroly a řízení jakosti se sleduje kvalita vyráběné I-kyseliny. Jedním z jakostních ukazatelů je obsah I-kyseliny v sušeném produktu. Data zahrnují dvouměsíční produkci. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L , pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? Z kvantilového grafu odečtěte, pod jakou hodnotou obsahu I-kyseliny leží všech 30 % hodnot výběru a pod jakou 10 %?
Data: Obsah I-kyseliny v sušeném produktu [%]:

92.80	92.30	93.50	91.80	91.50	91.60	90.80	91.20	97.90	90.50
..	..	..	..	..	..	..	..	..	..
93.20	93.00	92.70	91.30	93.70	93.50	92.80	92.60	91.90	93.30
93.30									

Úloha C2.19 *Odlehlé hodnoty stanovovaného obsah chlorfenpropmetylu v pesticidu*
Zkontrolujte obsah účinné látky chlorfenpropmetylu [%] v různých šaržích přípravku Fatex EK 80 metodou plynové chromatografie. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L , pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_l) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? Které diagnostiky poskytnou spolehlivý závěr o rozdělení tohoto výběru? Ověřte všechny požadavky, kladené na reprezentativní náhodný výběr.
Data: Obsah chlorfenpropmetylu v pesticidním přípravku [%]:

80.7	80.0	80.7	82.1	80.5	82.0	82.0	81.7	88.9	81.0
..	..	..	..	..	..	..	..	..	..
80.9	81.3	89.4	80.3						

Úloha C2.20 *Indikace odlehlých hodnot obsahu chlorfenpropmetylu*
Zkontrolujte, zda se vyskytují odlehlé hodnoty ve stanoveném obsahu účinné látky chlorfenpropmetylu [hm %] v různých šaržích přípravku Fatex EK 80 metodou plynové chromatografie. Z grafických diagnostik EDA určete také rozdělení tohoto výběru. Co lze vyčíst z bariérově-číslicového schéma, zkonstruovaného formou sedmipísmenového zápisu dat? Jaká je hloubka prvku 80.0 % ve výběru?
Data: Obsah chlorfenpropmetylu v pesticidním přípravku [%]:

60.7	80.0	80.7	82.1	80.5	82.0	82.0	81.7	88.9	81.0
..	..	..	..	..	..	..	..	..	..
80.9	81.3	89.4	111.3						

Úloha C2.21 *Symetrie rozdělení výběru obsahu elementární síry ve fenantrenu*
Na přístroji pro elementární analýzu byla provedena měření obsahu znečišťující síry ve vzorku fenantrenu. Prozkoumejte, zda je třeba odstranit odlehlé hodnoty? Vyšetřete tvar

rozdělení především na základě grafu polosum, symetrie, špičatosti a diferenčního kvantilového grafu. Pomocí *Q-Q* grafu a korelačního koeficientu těsnosti proložené přímky určete pak typ výběrového rozdělení. Je nalezené rozdělení symetrické? Kolik procent hodnot leží pod hodnotou obsahu síry 0.7 %? Co říká hloubka tohoto prvku ve výběru?

Data: Obsah síry S [%] ve fenantrenu:

0.618	0.726	0.885	0.688	0.741	0.746	0.676	0.721	0.702	0.735
0.611	0.784	0.706	0.817	0.781	0.757	0.763	0.756	0.792	

Úloha C2.22 *Variabilita hodnot molárního absorpčního koeficientu Saturnové černi OB*
Spektrofotometrickou analýzou typové substance Saturnové černi OB byly získány hodnoty molárních absorpčních koeficientů v průběhu 3 měsíců. Vyšetřete základní předpoklady o reprezentativním náhodném výběru. Které diagnostiky ukázaly, že je třeba odstranit odlehlé hodnoty? Který odhad užijete pro míru rozptýlení? Dle míry rozptýlení zaokrouhlete střední hodnotu molárního absorpčního koeficientu na patřičný počet platných cifer.

Data: Molární absorptivita [mol^{-1}. cm^{-1}. dm]:

6731.77	6790.63	6765.66	6740.65	6791.47	6731.65	6748.11	6749.40	6681.19	6665.60
..	..	..	..	..	..	..	..	..	..
6703.24	6716.05	6689.66	6668.77	6710.00	6724.28				

Úloha C2.23 *Statistické zvláštnosti výběru obsahu standardu rtuti na přístroji AMA254*
Je třeba provést vyhodnocení kontrolního stanovení standardu rtuti (0.1 mg. l^{-1}) za účelem kalibrace přistroje AMA254 a určit rozdělení tohoto výběru dat. Jde o Laplaceovo či normální rozdělení? Vyhodnoťte kvantilové parametry polohy a rozptýlení a ověřte předpoklady, kladené na reprezentativní náhodný výběr. Které kvantilové diagnostiky odhalily spolehlivě odlehlé hodnoty? Ukazuje graf maximální věrohodnosti nutnost aplikace mocninné transformace?

Data: Koncentrace standardu rtuti [mg. l^{-1}]:

0.1070	0.0940	0.0719	0.0912	0.1000	0.1020	0.1020	0.1030	0.0956	0.0928
..	..	..	..	..	..	..	..	..	..
0.0928	0.0920	0.1030	0.0954	0.0968	0.1020	0.106	0.0975	0.1030	0.1070

Úloha C2.24 *Variabilita viskozity nitrocelulozy v průběhu 2 měsíců*
Jednou složkou při výrobě střelného prachu je nitrocelulóza. Při posuzování kvality je sledovaným znakem viskozita, měřená v MPa.s. Při zpracování nitrocelulozy se během 2 měsíců odebralo 83 vzorků a změřila jejich viskozita. Proveďte průzkumovou analýzu, ověření předpokladů o náhodném výběru. Jaká je míra variability hodnoty viskozity? Jaká je hloubka prvku 5.10 MPa.s ve výběru? Je třeba aplikovat transformaci dat?

Data: Viskozita nitrocelulozy [MPa. s]:

5.43	5.39	6.33	5.92	5.43	5.83	6.37	5.61	6.05	6.07
..	..	..	..	..	..	..	..	..	..
5.33	6.00	5.15							

Úloha C2.25 *Typ rozdělení procentuálního obsahu vody ve výrobku*
Stanovte střední hodnotu procentuálního obsahu vody ve výrobku LAV 27 % N za první čtvrtletí roku. Pomocí *Q-Q* grafu vyšetřete typ rozdělení, jeho symetrii a odhalte odlehlé hodnoty. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu

výběru. Které tři grafické diagnostiky nejlépe prokázaly symetrii rozdělení? Jaká je hloubka prvku 0.10 % ve výběru?

Data: Obsah vody ve výrobku [%]:

0.12	0.13	0.14	0.13	0.11	0.11	0.11	0.09	0.10	0.14
..	..	..	..	..	..	..	..	..	..
0.11	0.12	0.12	0.12	0.13	0.11	0.13	0.13	0.14	0.12

Úloha C2.26 *Symetrie rozdělení obsahu dusíku ve výrobku LAV 27 % N*

Stanovte procentuální obsah dusíku ve výrobku LAV 27 % N za první čtvrtletí roku na základě nalezeného typu rozdělení. Jde o Gaussovo rozdělení? Jsou ve výběru odlehlé hodnoty? Jak spolehlivě dokážete určit symetrii rozdělení? Jaká je hloubka prvku 27.0 % ve výběru a z kvantilového grafu odečtěte kolik procent hodnot leží pod tímto prvkem?

Data: Obsah dusíku ve výrobku LAV 27 %N [%]:

27.0	27.1	27.0	26.9	26.9	27.1	27.0	26.8	26.8	26.8
..	..	..	..	..	..	..	..	..	..
26.8	26.7	26.8	26.9	27.0	26.7	26.9	26.6	26.6	26.9

Úloha C2.27 *Délky konců rozdělení výběru nezreagované suroviny v reakční směsi*

V posledním kroku třístupňové syntézy se stanovuje množství nezreagovaného 4-methylkatecholdimethylacetátu v reakční směsi. Aplikujte postup průzkumové analýzy dat, rozeberte grafické diagnostiky a učiňte své závěry. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? Vysvětlete, k čemu vám poslouží znalost délky konců rozdělení? Jakou vlastnost výběru můžeme indikovat z písmenových hodnot?

Data: Obsah nezreagovaného 4-methylkatecholdimethylacetátu [%]:

2.4	3.2	3.8	3.1	3.1	2.9	4.0	2.9	2.6	3.3	3.4	3.9
3.3	2.9	2.7	3.5	2.8	3.3	3.0	2.9	3.2	2.8	3.2	3.6

Úloha C2.28 *Typ rozdělení obsahu kyseliny sírové ve 4-methylkatecholdimethylacetátu*

V dílčím kroku výroby 4-methylkatecholdimethylacetátu se pro okyselení reakční směsi na požadované pH 1 - 2 používá koncentrovaná kyselina sírová, teoretickém množství 56.3 litrů. Na získaná data aplikujte postup průzkumové analýzy dat, rozeberte kvantilové diagnostiky a učiňte své závěry o chování dat. Pomocí *Q-Q* grafu a korelačního koeficientu rozhodněte, zda jde o normální nebo rovnoměrné rozdělení?

Data: Obsah kyseliny sírové ve 4-methylkatecholdimethylacetátu [l]:

57.2	55.7	56.2	56.8	55.8	57.1	55.9	58.0	56.0
..	..	..	..	..	..	..	..	..
57.3	56.1	56.5	57.0	57.8	56.5	56.3	56.0	57.2

Úloha C2.29 *Odlehlé body a symetrie rozdělení obsahu CNFHES v sušině*

Laboratorně byla zkoumána technologie výroby barviva modř. Vzorky izolovaných produktů nitrace z pokusů označené 4-chlor-3-nitrofenyl-(2-hydroxyethyl)sulfon, zkráceně

CNFHES, byly hydrolyzovány varem s NaOH a uvolněné chloridy titrovány. Z analytických stanovení byla tak získána série obsahů CNFHES v sušině. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? Porovnejte histogram s jádrovým odhadem hustoty pravděpodobnosti. Rozhodněte, zda je konstantní délka intervalu histogramu optimální i pro zešikmená rozdělení?

Data: Obsah CNFHES v sušině [%]:

88.75	83.10	97.60	90.25	89.00	90.10	94.40	89.70	93.30	96.90
93.00	94.40	86.60	92.80	88.75	93.90	94.70	94.70	90.00	94.90
92.00									

Úloha C2.30 *Analýza symetrie a typu rozdělení výtěžku produktů nitrace*

Z analyticky zjištěných obsahů 4-chlor-3-nitrofenyl-(2-hydroxyethyl)sulfonu, zkráceně CNFHES v produktech nitrace z jednotlivých pokusů a z hmotnosti sušin produktů byly vypočteny výtěžky. Získaný soubor dat byl podroben statistické analýze a výsledky použity k ekonomickému hodnocení přípravy. Porovnejte závěry o tvaru rozdělení a jeho symetrii z *Q-Q* grafu a krabicového grafu. Vedou oba grafy ke stejným závěrům?

Data: Obsah CNFHES v produktech nitrace [%]:

85.57	92.77	81.12	92.12	88.49	93.22	84.39	83.34	81.50	91.06
86.40	90.14	81.23	87.32	84.74	90.86	79.52	89.75	80.18	72.87
90.20									

Úloha C2.31 *Reprezentativní náhodný výběr obsahu modrých látek*

Při pokusné výrobě Modré báze H-3R byl titrací dusitanem zjišťován obsah modrých látek ve filtrátech po vykyselení. Titrační analýza filtrátů z každé provozní operace byla prováděna v provozní laboratoři a pro kontrolu současně i v centrální laboratoři. Oba výběry byly testovány zda mají shodné rozdělení. Vyšetřete nejprve, zda oba výběry splňují požadavky, kladené na reprezentativní náhodný výběr. Který kvantilový odhad polohy bude nejvhodnější?

Data: Obsah modrých látek [%], *C231a* v provozní laboratoři, *C231b* v centrální laboratoři:

3.470	2.980	2.880	2.230	3.467	2.220	2.880	3.380	2.430	2.630
..	..	..	..	..	..	..	..	..	..
2.300	2.670	3.160	2.800	2.909	2.330	3.040	1.180		

Úloha C2.32 *Velikost rozptýlení hodnot obsahu vlhkosti v granulích polymeru*

Důležitým zpracovatelským parametrem akrylonitrilbutadienstyrenových polymerů je obsah vlhkosti (požadavek 0.270 ± 0.006 %), která může být příčinou povrchových vad folií z ABS plastů. Z 31 výrobních šarží FORSANU 573 byl odebrán vzorek granulí, u kterého byla Fischerovou metodou stanovena voda. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty?

Zjistěte, zda daný výběr splňuje předepsané požadavky zákazníka. Jaké je rozptýlení hodnot obsahu vlhkosti v granulích polymeru a který kvantilový odhad pro ně užijete?

Data: Obsah vody v granulích polymeru[%]:

0.272	0.271	0.270	0.269	0.268	0.275	0.273	0.270	0.271	0.271	0.269	0.271	0.270
..	..	..	..	..	..	..	..	..	..	..	..	..
0.268	0.276	0.270	0.274	0.271								

Úloha C2.33 *Regulační diagram obsahu kadmia v referenčním materiálu*

Regulační diagramy jsou vizuálním nástrojem pro kontrolu výrobního procesu. K sestrojení centrální linie, reprezentované střední hodnotou μ, horního (+2s) a dolního (-2s) varovného limitu a horního (+3s) a dolního (-3s) regulačního limitu byl analyzovaný vzorek certifikovaného referenčního materiálu CRM, ve kterém byl certifikovaný obsah kadmia $\mu_0 = 1.19$ μg. l^{-1}. Analyzujte soubor hodnot za účelem získaní kvantilového odhadu střední hodnoty. Jde o symetrické rozdělení? Jsou ve výběru nějaké odlehlé hodnoty, jež leží mimo regulační meze diagramu?

Data: Obsah kadmia μ_0 [μg. l^{-1}]:

1.24	1.37	1.13	1.25	1.16	1.13	1.17	1.40	1.15	1.00	1.22	1.19	1.12	1.09
..	..	..	..	..	..	..	..	..	..	..	..	..	..
1.21	1.14	1.02	1.25	1.22	1.15	1.22	1.15	1.17	1.19	1.25	1.27		

Úloha C2.34 *Vyšetření výběru migračních časů u kapilární zónové elektroforézy*

Při měření viskozity kapilární zónovou elektroforézou byl do kapiláry dávkován jako mobilní standard roztok semihydrátu vinanu draselného o hmotnostní koncentraci 16.35 g. l^{-1}. Byly zaznamenány migrační časy. Vyšetřete symetrii rozdělení výběru migračních časů a ověřte předpoklady, kladené na výběr. Je vhodné vyloučit z výběru odlehlé hodnoty? Vyčíslete kvantilovou míru polohy a rozptýlení.

Data: Migrační čas [s]:

32.40	32.80	32.34	32.80	32.92	32.22	32.54	32.16	32.26	32.68	32.62
..	..	..	..	..	..	..	..	..	..	..
32.26	32.38	32.34	32.74	32.38	31.78	32.36				

Úloha C2.35 *Homogenita pevnosti polyesterových vláken*

V laboratoři technické kontroly je měřena pevnost polyesterových vláken náhodného výběru. Vyšetřete předpoklady, kladené na výběr a soustřeďte se především na test homogenity a odlehlé hodnoty. Vyčíslete kvantilovou míru polohy a rozptýlení. Vyplývá z exploratorní analýzy, že je třeba použít transformaci dat?

Data: Pevnost polyesterových vláken [N]:

13.16	12.71	13.14	12.93	12.77	12.83	13.06
..	..	..	..	..	..	..
12.94	12.86	12.78	12.84	12.81	12.94	13.22

Úloha C2.36 *Statistické zvláštnosti rozdělení výběru tažnosti šicí nitě*

Při hodnocení mechanických vlastností šicí nitě byla měřena především její tažnost. Exploratorní analýzou dat vyšetřete statistické zvláštnosti náhodného výběru, ověřte výběrové předpoklady a rozhodněte o volbě nejlepšího odhadu polohy.

Data: Tažnost šicí nitě [%], $n = 50$:

30.44	27.62	29.52	29.04	27.86	29.52	29.76	29.54	28.82	30.16
..	..	..	..	..	..	..	..	..	..
29.72	28.68	28.78	28.90	28.04	28.56	30.34	30.20	29.12	30.38

2.5.3 Analýza environmentálních, potravinářských a zemědělských dat

Úloha E2.01 *Kvantily k vyšetření obsahu oxidů dusíku v ovzduší*
V průběhu 24 hodin byla sledována koncentrace oxidů dusíku v ovzduší v prostorách před slévárnou strojírenského závodu. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? O jaký typ rozdělení se jedná? Testujte, zda uvedená data splňují požadavky na výběr. Dle závěrů průzkumové analýzy zvolte vhodné rozdělení, aproximující rozdělení výběru.
Data: Koncentrace oxidů dusíku v ovzduší [μg. m^{-3}]:

4.61	3.80	10.53	34.40	31.48	109.58	16.34	0.50	13.07	18.31	22.39
23.36	27.90	8.10	9.29	33.78	11.90	44.84	24.49	81.30	38.46	22.42

Úloha E2.02 *Symetrie rozdělení obsahu dusičnanů v pitné vodě*
Vyšetřete předpoklady o výběru u 50 stanovených obsahů dusičnanů v pitné vodě (str. 35 v cit.[15]) a testujte, zda je splněn předpokládaný obsah 0.50 μg. ml^{-1}. Užijte písmenově-číslicové schéma sumarizace dat k vyšetření statistických zvláštností výběru. Odhadněte vhodný typ rozdělení dat. Jsou v datech odlehlé hodnoty? Je rozdělení symetrické? Jakou hloubku mají prvky 0.46 a 0.52?
Data: Obsah dusičnanů ve vodě [μg. ml^{-1}]:

0.51	0.51	0.51	0.50	0.51	0.49	0.52	0.53	0.50	0.47
..	..	..	..	..	..	..	..	..	..
0.51	0.50	0.50	0.53	0.52	0.52	0.50	0.50	0.51	0.51

Úloha E2.03 *Velikost rozptýlení hodnot obsahu fosforu v odpadních vodách*
Analýzou odpadní vody, odebírané na odtoku, byla získána data o obsahu celkového fosforu v mg. l^{-1}. Je třeba odstranit odlehlé hodnoty? Prozkoumejte tvar rozdělení především na základě grafu polosum, symetrie, špičatosti a diferenčního kvantilového grafu. Pomocí *Q-Q* grafu a korelačního koeficientu těsnosti proložené přímky určete typ výběrového rozdělení. Je nalezené rozdělení symetrické? Liší se robustní míra interkvartilové rozpětí od nerobustní směrodatné odchylky? Kterou zde využijete?
Data: Obsah fosforu v odpadní vodě [mg. l^{-1}]:

1.25	1.27	1.27	1.24	1.25	1.23	1.26	1.28	1.27	1.27
..	..	..	..	..	..	..	..	..	..
1.25	1.25	1.25	1.24	1.27	1.25	1.24	1.24	1.25	1.28

Úloha E2.04 *Symetrie rozdělení výběru koncentrace amoniaku v upravené vodě*
Během roku bylo v upravené vodě z úpravny vod Saky provedeno 44 stanovení koncentrace

amonných iontů v mg. l^{-1}. Průzkumovou analýzou dat určete zvláštnosti rozdělení výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? O jaký typ rozdělení se jedná? Jaký odhad zvolíte za nejlepší pro míru polohy?

Data: Koncentrace amonných iontů v upravené vodě [mg. l^{-1}]:

0.06	0.12	0.40	0.19	0.15	0.14	0.31	0.11	0.41	0.02
..	..	..	..	..	..	..	..	..	..
0.28	0.69	0.33	0.32						

Úloha E2.05 *Podezřelé a odlehlé hodnoty obsahu polychlorovaných bifenylů PCB v oleji*
V referenčním vzorku oleje byl stanoven obsah polychlorovaných bifenylů PCB [mg/kg] (Delor 103, Delor 106). Vyšetřete statistické vlastnosti výběru dat a ověřte předpoklady, kladené na reprezentativní náhodný výběr. Jsou v datech podezřelé a odlehlé hodnoty? Je zde vhodnější užít robustní odhady střední hodnoty nebo mocninnou transformaci?

Data: Obsah PCB v oleji [mg. kg^{-1}]:

60.9	59.6	59.0	60.0	60.1	62.1	53.2	63.3	57.8	58.4
..	..	..	..	..	..	..	..	..	..
57.9	61.1	60.8	58.3	67.2	59.2	61.8	63.0	62.6	70.0

Úloha E2.06 *Ověření předpokladů o výběru obsahu železa v pitné vodě*
Ve vzorku pitné vody byl metodou AAS analyzován obsah železitých iontů. Norma připouští maximální obsah 0.3 mg. l^{-1}. Aplikujte postup analýzy jednorozměrných dat v pořadí: 1. Průzkumová analýza EDA, 2. Ověření předpokladů o výběru, 3. Transformace dat. Rozeberte a vysvětlete jednotlivé kvantilové diagnostiky.

Data: Obsah železitých iontů v pitné vodě [mg. l^{-1}]:

0.1910	0.1640	0.1580	0.2040	0.2090	0.2070	0.1890	0.1570	0.1820	0.2070
..	..	..	..	..	..	..	..	..	..
0.2472	0.2088	0.2050	0.2010	0.2066	0.2000	0.2030	0.2463		

Úloha E2.07 *Symetrie a typ rozdělení hodnot výběru chloridů v přírodních vodách*
Osvědčeným způsobem zajištění kvality chemických analýz vod je okružní porovnávání výsledků mezi laboratořemi. Přírodní vody představují zdroje pitné vody v různých lokalitách, ve kterých se sleduje řada složek, např. chloridy s maximálním obsahem 250 mg. l^{-1}. Byla analyzována data z okružního testu 33 laboratoří. Učiňte průzkum tvaru rozdělení především na základě grafu polosum, symetrie, špičatosti a diferenčního kvantilového grafu a určete typ tohoto rozdělení. Co říká hloubka tohoto prvku ve výběru?

Data: Obsah chloridů v přírodních vodách [mg. l^{-1}] v okružním testu:

184.00	183.00	221.60	182.92	180.00	189.50	181.00	174.60	243.00
..	..	..	..	..	..	..	..	..
186.20	187.88	186.00	182.00	210.99	177.25			

Úloha E2.08 *Ověření předpokladů náhodného výběru hodnot CHSK ve vodách*
Ve vzorcích vod byla zjišťována chemická spotřeba kyslíku. Na hladině významnosti $\alpha = 0.05$ testujte, zda je kolísání kolem střední hodnoty náhodného charakteru. Vyčíslete

kvantilové charakteristiky šikmosti a špičatosti a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? O jaký typ rozdělení se jedná? Jaký odhad zvolíte za nejlepší pro míru polohy? Indikuje graf logaritmu věrohodnostní funkce potřebu aplikace mocninné transformace?

Data: Hodnoty CHSK ve vodách [mg. l^{-1}]:

26.5	26.6	27.8	25.8	29.6	25.9	27.8	27.5	28.6
26.9	27.5	26.6	26.5	27.2	26.5	24.5	26.6	25.8

Úloha E2.09 *Odlehlé hodnoty výběru u obsahu nepolárních uhlovodíků v kalu*

Vzorek kalu byl předsušen, rozemlet na částice menší než 0.1 mm a homogenizován po dobu 24 hodin v rotačním zařízení. Vzorek byl potom rozeslán do 55 laboratoří v okružním testu, kde byl analyzován mimo jiné obsah nepolárních uhlovodíků. Určete symetrii a typ rozdělení a dále vyšetřete předpoklady o reprezentativním náhodném výběru. Jsou ve výběru odlehlé hodnoty? Dokažte, zda je k vyčíslení polohy a rozptýlení lepší užít robustních odhadů nebo dát přednost transformaci dat?

Data: Obsah nepolárních uhlovodíků v kalu [mg/kg] v okružním testu:

8100	12409	14199	15270	12701	12550	11768	10341	6422	21500
..	..	..	..	..	..	..	..	..	..
11125	15707	8309	24396	15730					

Úloha E2.10 *Vyšetření symetrie výběrového rozdělení obsahu dusičnanů v pitné vodě*

Fotometrickou metodou se salicylanem sodným byl stanoven obsah dusičnanů [mg. l^{-1}] v reálném vzorku pitné vody. Aplikujte postup exploratorní analýzy, rozeberte diagnostiky a učiňte své závěry o typu rozdělení. Jsou ve výběru odlehlé hodnoty? Prokažte, zda je v tomto případě výhodnější mocninná transformace nebo užití robustních odhadů?

Data: Obsah dusičnanů v pitné vodě [mg. l^{-1}]:

57.56	57.80	58.59	56.72	59.33	58.27	56.65	57.03	56.58	55.71
58.00	57.08	58.41	53.64	57.13	58.04	58.45	57.92	56.21	

Úloha E2.11 *Kvantily a tvar výběrového rozdělení obsahu fosforu v odpadní vodě*

Analýzou odpadní vody, odebírané na odtoku z čistírny odpadních vod, byl získán obsah celkového fosforu v mg. l^{-1}. Odhadněte typ rozdělení a ověřte předpoklady, kladené na výběr. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Je vhodné vyloučit odlehlé hodnoty nebo užít robustní odhady?

Data: Obsah fosforu v odpadní vodě [mg. l^{-1}]:

1.25	1.27	1.27	1.24	1.25	1.23	1.26	1.28	1.27	1.27
..	..	..	..	..	..	..	..	..	..
1.26	1.25	1.26	1.24	1.29	1.25	1.22	1.24	1.25	1.28

Úloha E2.12 *Odlehlé hodnoty v obsahu mědi a zinku v odpadních vodách*

V odpadních vodách se mimo jiné sleduje i obsah mědi a zinku. Vlastnímu biologickému čištění předchází proces neutralizace odpadních vod a obsah kovů je sledován i před touto

neutralizací. Vyšetřete statistické zvláštnosti výběru dat pomocí průzkumové analýzy. O jaký typ rozdělení se jedná? Jsou ve výběru nějaké odlehlé nebo podezřelé hodnoty?

Data: *E212a*, obsah Cu [mg. l^{-1}]:

1.098	1.876	2.149	0.904	1.128	0.468	0.428	0.175	0.050	0.074
..	..	..	..	..	..	..	..	..	..
1.055	2.058	0.627	1.005						

E212b, obsah Zn [mg. l^{-1}]:

1.231	0.654	5.800	0.963	1.795	3.808	0.982	0.191	0.161	0.560
..	..	..	..	..	..	..	..	..	..
0.239	0.149	0.874	0.349						

Úloha E2.13 *Určení pH zeminy pro rekultivační účely*

Pro rekultivační účely je požadována zemina, která má hodnotu pH asi 7. Z výsypky určené pro zemědělské účely byl proveden odběr vzorků za účelem stanovení hodnoty pH ve vodním výluhu. Celkem bylo analyzováno 30 vzorků. Ověřte, zda lze přijmout předpoklad normality výběru. Uvědomte si však, že pH = -log $[H^+]$ a normální rozdělení bude mít koncentrace vodíkových iontů $[H^+] = 10^{-pH}$.

Data: Hodnoty [pH] zeminy pro rekultivační účely:

6.2	6.5	6.8	7.0	7.2	7.3	7.4	7.6	6.3	6.5	6.8	6.9	7.1	7.2	7.0
6.4	7.6	7.2	7.2	6.7	6.8	7.1	7.0	6.0	6.9	5.9	5.8	6.2	7.1	7.7

Úloha E2.14 *Typ rozdělení obsahu dusičnanů ve studniční vodě*

Analýzou studničních vod v obci byly stanoveny obsahy dusičnanů. Vyšetřete tvar rozdělení na základě grafu polosum, symetrie, špičatosti a diferenčního kvantilového grafu. Pomocí *Q-Q* grafu a korelačního koeficientu těsnosti proložené přímky pak určete pak typ výběrového rozdělení. Je nalezené rozdělení symetrické? Proveďte konstrukci rozdělení výběru a zvolte vhodný kvantilový odhad střední hodnoty. Existují v datech odlehlé hodnoty nadměrného obsahu dusičnanů, které mohou signalizovat přítomnost silného lokálního zdroje znečištění?

Data: Obsah dusičnanů [mg. l^{-1}]:

32	75	16	99	80	78	28	29	170	86	80	81	26	170	81	190	235	30	19	69
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
180	78	130	23	77	32	145	130	80	81	10	130	80	81	500	85	75	25	32	86

Úloha E2.15 *Indikace odlehlých hodnot koncentrace mědi v řece Odře*

Ve vodohospodářské laboratoři se v průběhu celého roku provádí rozbor vody řeky Odry. Vyčíslete písmenové hodnoty M, F_D, F_H, E_D, E_H, D_D a D_H a pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Splňuje výběr předpoklady o náhodnosti výběru? Pokuste se určit typ rozdělení a zvolit vhodný kvantilový odhad střední hodnoty a rozptýlení koncentrace měďnatých iontů v průběhu 1 měsíce. Na základě vnitřních a vnějších hradeb rozhodněte, zda jsou v datech nějaké odlehlé a podezřelé hodnoty? Kolik

procent hodnot leží nad hodnotou 0.1 mg. l^{-1}?

Data: Koncentrace měďnatých iontů v říční vodě [mg. l^{-1}]:

0.0140	0.0130	0.0200	0.0160	0.0170	0.0170	0.0120	0.0140	0.0016	0.0160
..	..	..	..	..	..	..	..	..	..
0.0180									

Úloha E2.16 *Indikace odlehlých hodnot koncentrace amonného iontu v okružním testu*
Výsledky kruhového rozboru OR-CH-1/92 pro stanovení koncentrace amonného iontu v distribuovaném standardním roztoku, naředěném destilovanou vodou příslušné laboratoře, byly vyhodnoceny podle normy ČSN 01 0251 (ekvivalent ISO 5725). Vyčíslete písmenové hodnoty *M,* F_D , F_H , E_D , E_H , D_D a D_H a pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Splňuje výběr předpoklady o náhodnosti výběru? O jaké rozdělení se jedná? Určete střední hodnotu a míru rozptýlení výběru. Kolik procent laboratoří stanovilo vyšší koncentraci než 0.5 mg. l^{-1}?

Data: Obsah amonného iontu v okružním testu [mg. l^{-1}]:

0.510	0.455	0.485	0.520	0.200	0.455	0.485	0.520	0.340	0.455
..	..	..	..	..	..	..	..	..	..
0.515	0.995	0.450	0.480	0.515	0.455	0.480	0.517		

Úloha E2.17 *Statistické zvláštnosti výběru koncentrace hlinitých iontů v pitné vodě*
Ve vzorcích pitné vody byla měřena koncentrace hlinitých iontů. Norma připouští maximální obsah 0.3 mg. l^{-1} Al^{3+}. Vyšetřete statistické zvláštnosti rozdělení výběru a ověřte základní předpoklady, aby bylo možné zvolit vhodný odhad polohy a rozptýlení. Z kvantilového grafu odečtěte, kolik procent hodnot leží nad hodnotou 0.15 mg. l^{-1}?

Data: Koncentrace hlinitých iontů v pitné vodě [mg. l^{-1}]:

0.109	0.153	0.122	0.238	0.236	0.158	0.218	0.148	0.144	0.163
..	..	..	..	..	..	..	..	..	..
0.142	0.075	0.083	0.197	0.081	0.098	0.166	0.135	0.114	0.187

Úloha E2.18 *Nezávislost a normalita výběru obsahu MCPA v herbicidním přípravku*
U herbicidního přípravku Aminex Pur byl sledován obsah MCPA metodou kapalinové chromatografie. Učiňte průzkum tvaru rozdělení především na základě grafu polosum, symetrie, špičatosti a diferenčního kvantilového grafu. Pomocí *Q-Q* grafu a korelačního koeficientu těsnosti proložené přímky pak určete typ výběrového rozdělení. Co nám říká hloubka tohoto prvku ve výběru? Testujte, zda uvedený výběr dat splňuje základní požadavky nezávislosti a normality a z kvantilového grafu odečtěte, zda parametr polohy leží v intervalu od 25 do 28 %. Z kvantilového grafu odečtěte také kolik procent hodnot leží pod obsahem MCPA 25 % a kolik nad 28 %?

Data: Obsah MCPA v herbicidu [%]:

27.3	27.1	27.4	25.9	25.6	26.5	26.8	26.1	25.0	25.7
..	..	..	..	..	..	..	..	..	..
26.8	27.2	27.0	25.8	26.8	25.4	27.6	26.1	27.7	27.6

Úloha E2.19 *Statistické zvláštnosti výběru hodnot koncentrace kadmia v bramborách*
Na data o koncentraci kadmia v bramborách v oblastech jižní Moravy a Čech aplikujte postup vyšetření v pořadí: průzkumová analýza spojitých dat EDA, ověření předpokladů o výběru, transformace dat. Rozeberte a vysvětlete diagnostiky. Je třeba užít mocninnou transformaci? Je rozdělení výběru symetrické? Kolik procent hodnot leží nad koncentrací kadmia 0.030?

Data: Koncentrace kadmia v bramborách [mg. l^{-1}]:

0.080	0.018	0.015	0.052	0.080	0.020	0.036	0.013	0.032	0.060
..	..	..	..	..	..	..	..	..	..
0.115	0.033	0.060	0.020	0.009	0.005	0.027	0.055		

Úloha E2.20 *Střední hodnota barvy náhodného výběru 12 % světlých českých piv*
Na reprezantativním náhodném výběru 70 hodnot naměřené barvy 12 % světlých českých piv ověřte normalitu rozdělení, odhadněte střední hodnotu a stanovte, v jakém intervalu se nachází 95 % (popř. 99 %) naměřených hodnot výběru. Tento interval porovnejte s normou, která stanoví, že barva piva se má pohybovat v rozmezí 6 - 16 jednotek EBC. Barva piva b se stanoví spektrofotometricky dle vzorce $b = 25 \,.\, A_{430}$ (jednotek EBC), kde A_{430} je absorbance piva v 1 cm kyvety při 430 nm.

Data: Hodnota naměřené barvy piva b [jednotek EBC]:

8.800	9.700	8.100	10.100	8.300	10.500	10.500	8.700	11.000	11.700
..	..	..	..	..	..	..	..	..	..
13.300	10.500	10.400	13.500	11.800	12.200	11.700	14.000	9.800	8.600

Úloha E2.21 *Typ rozdělení výběru chemické spotřeby kyslíku v profilu řeky Úhlavy*
Během 7 měsíců byl každý pracovní den měřen parametr $CHSK_{Mn}$ v profilu řeky Úhlavy. Učiňte průzkum tvaru rozdělení především na základě grafu polosum, symetrie, špičatosti a diferenčního kvantilového grafu. Pomocí Q-Q grafu a korelačního koeficientu těsnosti proložené přímky určete typ výběrového rozdělení. Je nalezené rozdělení symetrické? Co nám říká hloubka tohoto prvku ve výběru? Jaká je střední hodnota tohoto výběru? Kolik procent překročilo hodnotu $CHSK_{Mn} = 8.0$?

Data: Hodnoty $CHSK_{Mn}$ v profilu řeky Úhlavy (Doudlevce):

5.3	5.4	3.4	4.0	6.4	4.4	6.8	6.3	6.2	10.0	4.9	4.2
..	..	..	..	..	..	..	..	..	..	..	..
5.1	4.6	4.3	5.9	3.8	5.3	7.8	5.8	5.7	5.1	3.7	3.3

Úloha E2.22 *Odlehlé hodnoty obsahu $KNK_{4.5}$ při analýze povrchové vody*
V rámci kontrolních rozborů byl analyzován vzorek povrchové vody. Jedním ze sledovaných parametrů byla hodnota parametru $KNK_{4.5}$. Kontroly kruhovým testem se zůčastnilo 88 laboratoří. Vyčíslete písmenové hodnoty M, F_D , F_H , E_D , E_H , D_D a D_H a pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Splňuje výběr předpoklady o náhodnosti výběru? O jaké rozdělení se jedná? Určete střední hodnotu a míru

rozptýlení výběru. Určete nejlepší odhad střední hodnoty a indikujte odlehlé hodnoty.
Data: Obsah $KNK_{4.5}$ při analýze povrchové vody:

3.27	3.34	3.21	3.28	3.60	3.29	3.13	3.21	3.28	3.32	3.07	3.29
..	..	..	..	..	..	..	..	..	..	..	..
3.27	3.27	3.34	3.30								

Úloha E2.23 *Symetrie rozdělení výběru obsahu mědi v odpadové vodě*
Metodou atomové absorpční spektrometrie byl analyzovaný vzorek odpadní vody. Je třeba vyšetřit naměřená data metodou exploratorní analýzy jednorozměrných dat a rozhodnout, zda jde o symetrické rozdělení. Kolik je ve výběru odlehlých hodnot? Poskytuje mocninná transformace stejný odhad střední hodnoty jako Boxova-Coxova? Vysvětlete, proč tomu tak je.
Dáta: Obsah mědi v odpadové vodě [mg. l^{-1}]:

0.0274	00319	0.0300	0.0318	0.0281	0.0321	0.0318	0.0286	0.0319	0.0322
..	..	..	..	..	..	..	..	..	..
0.0301	0.0337	0.0336	0.0306	0.0310	0.0326	0.0294	0.0336		

Úloha E2.24 *Kruhový test laboratoří při stanovení kadmia v grahamových rohlících*
V mezilaboratorním testu stanovení kadmia v grahamových rohlících byly stanoveny koncentrace kadmia Cd. Analyzujte data s cílem získání nejlepšího odhadu střední hodnoty. Při použití normy ISO, která využívá vylučování odlehlých výsledků klasickými postupy, byla získaná hodnota 0.01880 mg/kg Cd. Je rozdělení symetrické? Obsahuje výběr nějaké odlehlé hodnoty? Je správné vyloučit rohlík z analyzovaného náhodného výběru, obsahuje-li nadměrně odlehlou (vyšší) koncentraci kadmia? Dáte přednost robustním odhadům polohy nebo mocninné transformaci?
Data: Kód laboratoře, naměřená koncentrace kadmia v rohlících [mg/kg]:

101	0.0237,	103	0.0177,	104	0.0120,	105	0.0200,	106	0.0100,
..	..	..	..	..	..	..	..	..	..
130a	0.0297	130b	0.0197,	131	0.0280,	132	0.0117,	133	0.0177,
134	0.0270,								

Úloha E2.25 *Ověření předpokladů o výběru odezvy detektorů MS/MS a ECD*
Při testování nového přístroje byly naměřeny a porovnány odezvy detektorů MS/MS a ECD pro stanovení jednotlivých kongenerů v Deloru 103. Průzkumovou analýzou dat určete zvláštnosti rozdělení výběru, ověřte základní předpoklady o datech, proveďte transformaci dat a určete parametry rozptýlení a polohy. Kolik procent hodnot leží nad poměrem 1.0?
Data: Poměr odezvy detektorů MS/MS a ECD:

1.0408	1.1082	0.9393	1.0244	1.1522	0.8232	0.9653	1.0933	0.9130	1.0058
..	..	..	..	..	..	..	..	..	..
0.9738	0.9251								

Úloha E2.26 *Odlehlé hodnoty při kontrole spolehlivosti dávkování u dávkovacího stroje*
V průběhu jedné směny bylo sledováno dávkování práškového kolagenového nápoje plnicím strojem do dóz v nastaveném množství 250 g. Testujte, zda náhodný výběr dóz splňuje základní předpoklady o reprezentativním náhodném výběru, určete typ rozdělení a vyhodnoťte tak spolehlivost (tj. správnost a přesnost) činnosti dávkovacího stroje.

Z kvantilového grafu odečtěte, kolik procent dóz mělo obsah hmotnosti menší než 250 g? Lze hovořit o správném dávkování?

Data: Hmotnost dávkování u dávkovacího stroje *m* [g]:

250.2	250.4	250.2	251.2	250.6	250.5	250.6	251.0
..	..	..	..	..	..	..	..
249.8	250.0	250.0	249.4	249.8	249.5		

2.5.4 Analýza hutnických a mineralogických dat

Úloha H2.01 *Odlehlé hodnoty obsahu manganu ve výběru oceli*

Byl sledován procentuální obsah manganu v oceli pro 33 taveb odlité oceli C64.Vyčíslete písmenové hodnoty M, F_D , F_H , E_D , E_H , D_D a D_H a pak zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Splňuje výběr předpoklady o náhodnosti výběru? O jaké rozdělení se jedná? Určete střední hodnotu a míru rozptýlení výběru. Jaká je hloubka dolního a horního kvartilu?

Data: Obsah manganu ve výběru oceli [%]:

0.54	0.51	0.53	0.54	0.53	0.56	0.58	0.52	0.58	0.55
0.56	0.53	0.60	0.57	0.64	0.60	0.57	0.50	0.52	0.52
0.51									

Úloha H2.02 *Statistické zvláštnosti výběru obsahu křemíku v oceli*

Byl sledován procentuální obsah křemíku v oceli pro 33 taveb odlité oceli C78. Vyšetřete statistické zvláštnosti výběru tak, že zkonstruujete bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Kterými diagnostikami vyšetříte, zda jde o symetrické rozdělení? Sestrojte graf hustoty pravděpodobnosti rozdělení a odhadněte střední hodnotu. Z kvantilového grafu odečtěte, kolik procent prvků obsahu křemíku je pod hodnotou 0.15 %?

Data: Obsah křemíku ve výběru oceli [%]:

0.19	0.18	0.17	0.14	0.17	0.16	0.16	0.18	0.19	0.16
..	..	..	..	..	..	..	..	..	..
0.18	0.19	0.16							

Úloha H2.03 *Ověření typu rozdělení obsahu niklu ve výrobním procesu*

V hodinových intervalech byly ve výrobním procesu stanovovány hodnoty obsahu niklu (str. 25 v cit.[16]). Z kruhového grafu a z *Q-Q* grafu posuďte, zda lze považovat výběr za normálně nebo rovnoměrně rozdělený a bez odlehlých hodnot? Porovnejte výsledky s dalšími kvantilovými diagnostikami. Jakou hloubku má prvek 21.7 ppm ve výběru? Co to pro uživatele znamená?

Data: Obsah niklu ve výrobním procesu [ppm]:

21.2	21.8	21.4	21.0	21.3	21.6	21.1	21.7
22.2	21.9	22.3	22.1	22.4	22.2	22.1	

Úloha H2.04 *Podezřelá a odlehlé měření sublimačního tepla platiny*
Hampson a Walker stanovili sublimační teplo platiny (str. 328 v cit.[17]). Napište sedmipísmenový zápis výběru a vyšetřete statistické vlastnosti výběru. Testujte nezávislost, homogenitu a normalitu rozdělení a podezřelé odlehlé hodnoty. Jde o normální či logaritmicko-normální rozdělení? Užijete raději robustní odhady střední hodnoty nebo transformaci dat? Jak vyjádříte míru rozptýlení sublimačního tepla?
Data: Sublimační teplo platiny [kcal. mol^{-1}]:

136.2	136.6	135.8	135.4	134.7	135.0	134.1	143.3	147.8	148.8
..	..	..	..	..	..	..	..	..	..
134.2	134.9	134.8	134.5	134.3	135.2				

Úloha H2.05 *Klasifikace kvality kaolinu dle obsahu oxidu hlinitého*
Pro keramickou výrobu je požadován kaolin o obsahu alespoň 37 % oxidu hlinitého (str. 247 v cit.[18]). Rozhodněte, zda výběr splňuje základní požadavky, především normální rozdělení. Na jaký tvar rozdělení ukazují koeficienty šikmosti a špičatosti? Je nutné užít mocninnou transformaci? Jde o symetrické rozdělení? Splňuje tento výběr obsahu oxidu hlinitého Al_2O_3 v kaolinu požadavek normy 37 %? Kolik procent hodnot leží pod hodnotou 37 % Al_2O_3 a kolik nad ní?
Data: Obsah oxidu hlinitého v kaolinu [%]:

37.76	38.15	37.55	37.95	38.32	37.59	37.81	37.84	37.45	37.81
..	..	..	..	..	..	..	..	..	..
37.36	37.28								

Úloha H2.06 *Odlehlé hodnoty obsahu oxidu vápenatého CaO v pórovitém jílu*
V keramickém průmyslu je používán pórovitý jíl, u kterého je požadován obsah CaO v rozmezí 0.3 - 1.3 %. Pro sledování obsahu CaO bylo z průzkumných vrtů odebráno celkem 31 vzorků a provedeno stanovení obsahu CaO. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Jaké je rozdělení výběru? Určete, zda keramická surovina splňuje požadovaný obsah CaO. Z kvantilového grafu odečtěte, kolik procent prvků výběru leží pod hodnotou 0.3 % a kolik nad 1.3 % CaO?
Data: Obsah CaO v pórovitém jílu [%]:

1.34	1.12	1.13	1.38	0.95	1.23	0.25	1.25	1.68	1.40
..	..	..	..	..	..	..	..	..	..
1.18	1.43	1.23	2.02	1.18	1.18	1.46	1.96	1.96	1.74

Úloha H2.07 *Vyšetření rozdělení výběru přírodního pozadí ekvivalentu záření gama*
Měření hodnoty přírodního pozadí fotonového dávkového ekvivalentu záření gama bylo provedeno na referenční ploše přístrojem D300 s nastavením časové konstanty 10 s. Vyčíslete písmenové hodnoty *M*, F_D , F_H , E_D , E_H , D_D a D_H a zkonstruujte i bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Proveďte průzkumovou analýzu spojitých dat, ověření předpokladů a transformaci dat. Je rozdělení symetrické? Jsou ve výběru odlehlé body?
Data: Hodnoty záření gama [$\mu Gy.h^{-1}$]:

0.124	0.115	0.121	0.122	0.123	0.101	0.101	0.118	0.122	0.114
..	..	..	..	..	..	..	..	..	..

0.124	0.117	0.123	0.121	0.125	0.121	0.118	0.122	0.125	0.125

Úloha H2.08 *Nejlepší odhad obsahu vody, popela síry v uhlí a jeho spalného tepla a*
U 31 vzorků uhlí byly stanoveny (a) obsah vody v procentech, (b) popel v procentech, (c) spalné teplo v kcal/kg a (d) obsah síry v procentech. Vyčíslete písmenové hodnoty *M,* F_D, F_H , E_D , E_H , D_D a D_H a zkonstruujte i bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Nalezněte typ rozdělení pro každý ze čtyř výběrů. U každého výběru vyšetřete jeho symetrii a počet odlehlých hodnot. Je výhodnější použít robustních odhadů nebo mocninné transformace dat? Dle jakého kritéria rozhodnete, zda je nutné využít transformace dat? Kdy je možné využít za nejlepší odhad míry polohy hodnoty aritmetického průměru?
Data: *H208a* obsah vody [% . 100], *H208b* obsah popela [% . 100], *H208c* spalné teplo [kcal/kg] a *H208d* obsah síry [% . 100] v uhlí:

993	2928	4144	22,	1249	1446	4992	47,	1350	1206	5045	43,
..	..	..	..	..	..	..	..	..	..	..	..
665	6183	1628	17,	888	5038	2374	48,				

Úloha H2.09 *Vyšetření symetrie rozdělení dvou výběrů obsahu uhlíku a síry v oceli*
Uhlík a síra v oceli se stanovuje kvantometricky nebo spalováním s infračervenou detekcí. Porovnejte rozdělení výběru obsahu uhlíku i síry, stanoveném oběma metodami. Nalezněte vhodné kvantilové odhady střední polohy a rozptýlení pro všechny čtyři výběry. Jsou rozdělení symetrická? Doporučujete užít transformaci dat?
Data: Kvantometricky [%] *H209a* obsah uhlíku a *H209b* síry, infračervenou detekcí [%] *H209c* obsah uhlíku a *H209d* síry v oceli:

3.25	0.032	3.28	0.028,	3.28	0.038	3.34	0.040,	3.55	0.043	3.55	0.042,
..	..	..	..	..	..	..	..	..	..	..	..
3.23	0.045	3.26	0.051,	3.76	0.047	3.61	0.046,	3.28	0.066	3.17	0.060,

Úloha H2.10 *Vyšetření symetrie rozdělení výběru obsahu manganu v oceli*
Bylo odlito 21 taveb oceli značky C64 a stanoven obsah manganu. Exploratorní analýzou dat určete typ rozdělení a nalezněte nejlepší kvantilový odhad střední hodnoty a rozptýlení. Jsou všechny tavby rovnocenné, homogenní a prvky ve výběru nezávislé? Jsou ve výběru odlehlé hodnoty? Odhadněte, kolik procent hodnot leží nad obsahem manganu 0.58 %?
Data: Obsah manganu v oceli [%]:

0.54	0.51	0.53	0.54	0.53	0.56	0.58	0.52	0.58	0.55
0.56	0.53	0.60	0.57	0.64	0.60	0.57	0.50	0.52	0.52
0.51									

Úloha H2.11 *Vyšetření statistických zvláštností výběru obsahu síry v oceli*
Bylo odlito 33 taveb oceli značky C78 a stanoven obsah síry. Vyšetřete statistické zvláštnosti exploratorní analýzou dat a vyčíslete písmenové hodnoty *M,* F_D , F_H , E_D , E_H, D_D a D_H metodou pořadí a hloubek (dle vzoru Příklad 2.1, cit.[19]). Podle závěrů průzkumové analýzy odhadněte kvantilové odhady parametrů polohy a rozptýlení. Existují v tomto výběru odlehlé hodnoty? Jde o symetrické rozdělení? Jsou všechny tavby rovnocenné, homogenní a prvky ve výběru nezávislé?
Data: Obsah síry v oceli [%]:

0.006	0.004	0.006	0.012	0.014	0.012	0.013	0.007	0.006	0.010
..	..	..	..	..	..	..	..	..	..
0.005	0.004	0.004							

Úloha H2.12 *Odlehlé hodnoty obsahu olova ve finálním umělém hnojivu*
Při kontrole jakosti jednoho druhu tabletovaného hnojiva bylo v průběhu roční výroby získáno 99 hodnot obsahu olova v hnojivu. Určete, jak velký počet šarží nevyhovoval maximálnímu přípustnému množství olova, stanovenému podnikovou normou. Vyšetřete typ rozdělení a nalezněte, zda nejlepší kvantilový odhad střední hodnoty obsahu olova v roční produkci jako celku vyhovuje hygienickému limitu 30 mg/kg Pb. Jsou vzorky s extrémně vysokým obsahem olova součástí souboru s normálním rozdělením?

Data: Obsah olova v hnojivu během roční produkce [mg/kg]:

5.47	5.17	9.40	2.58	5.61	371.2	16.1	4.80	8.65	23.6	15.5
..	..	..	..	..	..	..	..	..	..	..
12.2	10.1	12.0	4.92	8.60	7.10	3.79	2.27	7.30	13.0	37.0

Úloha H2.13 *Určení typu rozdělení dvou výběrů obsahu uhlíku v oceli*
Určete typ rozdělení dvou výběrů obsahu uhlíku v oceli, určeného (a) kvantometricky a (b) spalováním s infračervenou detekcí. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? Dáte přednost robustním odhadům polohy nebo budete raději data transformovat?

Data: Dvojice obsahu uhlíku [%], stanoveného *H213a* kvantometricky a *H213b* spalováním s IČ detekcí:

3.85	3.82,	3.77	3.81,	3.71	3.80,	3.66	3.64,	3.73	3.64,	3.73	3.66,	3.72	3.67,
..	..	..	..	..	..	..	..	..	..	..	..	..	..
3.65	3.69,	3.61	3.70,	3.60	3.71,	3.64	3.73,	3.65	3.73,	3.66	3.65,	3.64	3.65,
3.59	3.64,												

Úloha H2.14 *Určení typu rozdělení dvou výběrů obsahu síry v oceli*
Určete typ rozdělení dvou výběrů obsahu síry v oceli o rozsahu n = 365, určené jednak (a) kvantometricky a jednak (b) spalováním s infračervenou detekcí. Na základě kvantilových měr rozhodněte, zda jsou obě rozdělení symetrická, homogenní a shodná? Nalezněte, které kvantilové odhady použijete za nejlepší k vyčíslení střední hodnoty a míry rozptýlení? Jsou ve výběru odlehlé hodnoty? Je nutné užít transformaci dat nebo robustní odhady?

Data: Dvojice obsahu síry [%], stanoveného *H214a* kvantometricky a *H214b* spalováním s infračervenou detekcí:

0.013	0.014,	0.014	0.016,	0.014	0.015,	0.019	0.015,	0.028	0.023,	0.024	0.019,	0.035	0.023,
..	..	..	..	..	..	..	..	..	..	..	..	..	..
0.014	0.011,	0.013	0.012,	0.011	0.010,	0.014	0.013,	0.015	0.011,	0.011	0.011,	0.009	0.011,
0.007	0.007,												

Úloha H2.15 *Určení typu rozdělení výběru oxidu vápenatého a oxidu hořečnatého ve skle*
Z 30 rozličných míst tabulového skla byly odebrány vzorky a analyzovány atomovou spektroskopií na obsah oxidu vápenatého CaO a oxidu hořečnatého MgO. Vyšetřete

statistické zvláštnosti, jako je stupeň symetrie, špičatosti, lokální koncentrace dat, odlehlé hodnoty? Vyšetřete tvar rozdělení na základě grafu polosum, symetrie, špičatosti a diferenčního kvantilového grafu. Jsou oba výběry homogenní a bez odlehlých hodnot? Komentujte rozdělení četnosti a rozhodněte, který kvantilový odhad parametru polohy a rozptýlení zde bude nejlepší.

Data: *H215a* obsah CaO, *H215b* obsah MgO [%] ve skle:

8.574 4.084,	8.529 4.043,	8.593 4.004,	8.676 4.048,	8.717 4.013,	8.680 3.993,
..	..	..	..	..	..
8.484 4.001,	8.518 4.004,	8.507 4.056,	8.563 4.048,	8.473 4.030,	8.546 4.015,

Úloha H2.16 *Symetrie rozdělení výběrů obsahu oxidu železitého a oxidu hlinitého ve skle*
Z 30 rozličných míst tabulového skla byly odebrány vzorky a analyzovány atomovou spektroskopií na obsah oxidu železitého Fe_2O_3 a oxidu hlinitého Al_2O_3. Sestrojte graf hustoty pravděpodobnosti a histogramy obou výběrů a vyšetřete symetrii rozdělení. Jsou oba výběry homogenní a bez odlehlých hodnot? Je třeba užít transformaci dat nebo robustních odhadů?

Data: *H216a* obsah Fe_2O_3 [%], *H216b* obsah Al_2O_3 [%] ve skle:

0.0413 0.620,	0.0423 0.625,	0.0413 0.620,	0.0406 0.620,	0.0409 0.635,
..	..	..	..	..
0.0393 0.645,	0.0384 0.649,	0.0392 0.659,	0.0398 0.654,	0.0393 0.649,

Úloha H2.17 *Určení typu rozdělení výběrů obsahu tří prvků v půdě metodou AAS*
Z dobře zhomogenizovaného půdního vzorku o celkové hmotnosti asi 1.5 kg, užívaného v laboratoři jako kontrolní vzorek, bylo odebráno 32 navážek a stanoveny obsahy tří prvků metodou AAS, a to zinku, manganu a mědi. Každá hodnota v datech představuje průměr tří až pěti paralelních měření. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar každého rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Splňují data požadavky na nezávislost? Je třeba užít transformaci dat?

Data: *H217a* obsah zinku, *H217b* obsah manganu, *H217c* obsah mědi [$mg.kg^{-1}$] v půdě:

3.97 74.4 3.01,	3.77 75.8 2.90,	4.10 73.3 2.80,	3.71 71.3 2.61,	3.77 74.1 2.90,
..	..	..	..	..
4.05 78.7 2.87,	4.11 79.8 3.05,			

Úloha H2.18 *Určení typu rozdělení výběrů při kontrole homogenity referenčního skla*
Pro kontrolu homogenity referenčního skla bylo z tohoto skla připraveno 30 vzorků, ve kterých byl opakovaně 4× stanoven obsah oxidu hlinitého Al_2O_3. Nalezněte, zda naměřené hodnoty představují úrovně signálu, které odpovídají obsahu Al_2O_3 při spektrálním stanovení. Sestrojte také rozdělení všech čtyř výběrů (sloupců) a odhadněte kvantilový parametr polohy. Je třeba k získání nejlepšího odhadu střední hodnoty užít transformaci dat?

Data: Signál odpovídající obsahu Al_2O_3 v referenčním skle (řádky jsou místa odběru skla a sloupce představují opakovaná měření):

565 569 565 565,	663 708 681 686,	690 695 699 645,	735 739 726 735,
....	....	....	....
627 620 623 616,	613 613 606 606,		

Úloha H2.19 *Kvantily a symetrie rozdělení obsahu oxidu křemičitého v cinvalditu*
Byly stanoveny hodnoty obsahu oxidu křemičitého v cinvalditu. Průzkumovou analýzou dat určete statistické zvláštnosti rozdělení výběru, ověřte také základní předpoklady o datech. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Určete kvantilové odhady parametrů rozptýlení a polohy. Je rozdělení výběru symetrické?
Data: Obsah oxidu křemičitého v cinvalditu[%]:

50.85	51.37	51.95	52.06	52.17	52.26	52.44	52.46	52.58	52.90
..	..	..	..	..	..	..	..	..	..
54.85	55.00	55.02	55.20	55.33	55.46	55.59	55.56	55.70	55.90
56.04									

Úloha H2.20 *Vyšetření symetrie rozdělení obsahu skandia v cinvalditu*
Pro obsah skandia v cinvalditu byly nalezeny hodnoty, uvedené v datech. Vyšetřete, zda je rozdělení výběru symetrické? Jsou ve výběru odlehlé hodnoty? Vede graf maximální věrohodnosti k důkazu nutnosti použití transformace dat? Nalezněte kvantilové míry polohy, rozptýlení a tvaru. Jaká je hloubka prvku 44.0 ve výběru?
Data: Obsah skandia v cinvalditu [%]:

11.0	21.0	27.0	32.0	33.50	34.0	35.0	35.45	37.0	37.0
..	..	..	..	..	..	..	..	..	..
49.50	50.0	51.0	52.0	52.03	54.0				

Úloha H2.21 *Sedmipísmenový zápis výběru obsahu oxidu lithného v cinvalditu*
Pro obsah oxidu lithného v cinvalditu byly nalezeny hodnoty výběru. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru a diskutujte symetrii rozdělení. Jsou ve výběru odlehlé hodnoty? Nalezněte kvantilové míry polohy, rozptýlení a tvaru. Posuďte, zda je třeba užít mocninnou transformaci k nalezení nejlepšího odhadu střední hodnoty?
Data: Obsah oxidu lithného v cinvalditu [%]:

1.27	1.83	1.85	2.15	2.17	2.22	2.23	2.26	2.26	2.28
..	..	..	..	..	..	..	..	..	..
2.82	3.22	3.65	3.73						

Úloha H2.22 *Vyšetření symetrie a určení typu rozdělení obsahu SiO_2 v kaolinu*
V laboratoři se kyždý týden sleduje proměřením vnitřního standardu kaolinu reprodukovatelnost chemické analýzy kaolinu rentgenově-fluorescenčním spektrometrem. Vyčíslete kvantilové charakteristiky šikmosti S_L a délek konců T_L v oblasti oktilů E a sedecilů D (symboly viz cit.[19]). Na jejich základě komentujte tvar rozdělení hustoty pravděpodobnosti. Jde o symetrické rozdělení? Určete průměrný obsah SiO_2 u vnitřního standardu kaolinu. Který kvantilový odhad budete považovat za nejlepší ?

Data: Obsah SiO_2 v kaolinu [%]:

49.34	49.47	49.60	49.60	49.60	49.50	49.41	49.33	49.47
..	..	..	..	..	..	..	..	..
49.66	49.56	49.61	49.55	49.46	49.47	49.57	49.48	49.37

Úloha H2.23 *Kvantily a typ rozdělení obsahu železa ve sklářském písku*
Sklárna uzavřela smlouvu na dodávky sklářského písku s těžební společností - dodavatelem. Předmětem smlouvy byl také požadavek na kvalitu sklářského písku z hlediska obsahu železa, specifikovaný normou: Jakost TS 15/08, Fe_2O_3 (max.) 0.015 hm.%. Laboratoř odběratele při náhodných kontrolách obsahu železa v dodávkách sklářské suroviny zjistila obsahy rovněž v hm.%. Ověřte u obou výběrů požadavky, kladené na reprezentativní náhodný výběr. Jsou obě výběrová rozdělení symetrická? Porovnejte rozdělení obou výběrů, do jaké míry jsou shodná a zda vykazují shodné míry polohy a rozptýlení.

Data: *H223a* obsahy železa (Fe_2O_3 v hm.%) ve sklářském písku, zjištěné laboratoří *odběratele:*

0.0100	0.0140	0.0160	0.0150	0.0160	0.0150	0.0160	0.0150	0.0150	0.0140
0.0150	0.0120	0.0170	0.0150	0.0170	0.0160	0.0110	0.0130		

H223b obsahy železa (Fe_2O_3 v hm.%) ve sklářském písku, zjištěné laboratoří *dodavatele:*

0.0140	0.0160	0.0140	0.0120	0.0160	0.0170	0.0110	0.0140	0.0110	0.0170
..	..	..	..	..	..	..	..	..	..
0.0140	0.0170	0.0160	0.0160	0.0150	0.0130	0.0170			

Úloha H2.24 *Tvar rozdělení a odlehlé hodnoty obsahu kobaltu ve sklářském písku*
V rámci přípravy referenčního materiálu na bázi laboratorně utaveného skla se stopovými obsahy vybraných prvků byl organizován kruhový test s cílem zjistit, do jaké míry obsahy těchto složek odpovídají předpokladům. Pro stanovení kobaltu, kterého se zúčastnilo 17 laboratoří, používajících různé metody, bylo dosaženo výsledků v tabulce. Na základě grafu polosum, symetrie a špičatosti vyšetřete tvar rozdělení (symboly viz cit.[19]). Dáte přednost robustnímu odhadu střední hodnoty nebo mocninné transformaci? Dosáhla některá laboratoř vysloveně odlehlých výsledků? Které diagnostiky shodně potvrdily odlehlé hodnoty?

Data: Výsledky mezilaboratorního testu referenčního skla pro obsah kobaltu [mg/kg]:

3.720	1.518	3.400	8.400	3.650	1.447	4.300	8.000	3.680	2.700	4.120	8.700
..	..	..	..	..	..	..	..	..	..	..	..
1.557	3.200	2.880	3.970								

Úloha H2.25 *Homogenita a typ rozdělení bělosti šarže kaolinu*
Pro zákazníka byla připravena expediční šarže kaolinu a z této byl odebrán náhodný výběr o 60 prvcích a u všech na přístroji ELREPHO 2000 změřena bělost [%]. Vyšetřete, zda je náhodný výběr homogenní a symetrického rozdělení a zda jsou splněny předpoklady o náhodném výběru. Vyčíslete kvantilovou míru polohy a rozptýlení.

Data: Hodnoty bělosti šarže kaolinu [%]:

79.0	79.0	79.2	79.4	79.1	79.2	79.0	79.0	79.0	79.0	79.2	79.2
..	..	..	..	..	..	..	..	..	..	..	..
79.3	79.2	79.3	79.1	79.2	79.0	79.4	79.5	79.3	79.1	79.3	79.8

Úloha H2.26 *Typ rozdělení a odlehlé hodnoty zrnitosti výběru kaolinu*
Při výrobě plavených kaolinů prochází surovina třídicími odstředivkami, které materiál třídí na jemnou a hrubou frakci. Jedním z mnoha sledovaných parametrů kvality vytřídění je zrnitost. U měřené vytříděné frakce se požaduje přítomnost zrn menších než 2 μm. Určete homogenitu a míru polohy u velikosti frakce vytříděného kaolinu v procentech.
Data: Velikost frakce kaolinu [%] o zrnu menším než 2 μm:

5.9	5.7	6.0	4.3	5.8	5.9	6.1	5.7	5.5	6.0	6.0	6.4
6.4	5.9	6.4	6.2	6.4	6.3	6.5	6.1				

Úloha H2.27 *Homogenita výběru obsahu dusíku v kontrolním vzorku oceli*
Kontinuálním měřením obsahu dusíku v kontrolním vzorku oceli metodou OES na automatickém analyzátoru s jiskrovým buzením je sledován proces stanovení dusíku za podmínek reprodukovatelnosti v počtu 5 analýz denně. Vyšetřete statistické zvláštnosti výběru, především homogenitu a vyčíslete odhad střední hodnoty.
Data: Obsah dusíku *N*.10000 [%] v oceli:

119	115	111	122	112	111	117	119	115	110	111	112
..	..	..	..	..	..	..	..	..	..	..	..
119	105	108	113								

Úloha H2.28 *Statistické zvláštnosti výběru obsahu manganu v 100 tavbách oceli*
Mangan působí v oceli jako desoxidační a desulfurační komponenta, která navíc zlepšuje pevnost oceli v tahu, tvrdost, vrubovou houževnatost a kalitelnost, a do určitého obsahu též kujnost a svařitelnost. Zhoršuje však slévatelnost, magnetické vlastnosti a snižuje tavicí teplotu a tepelnou vodivost. Obsah manganu je třeba při výrobě oceli neustále sledovat. Bylo odlito 100 taveb oceli ST52-3. Vyšetřete statistické zvláštnosti výběru obsahu manganu v těchto tavbách, určete typ rozdělení a odhadněte střední hodnotu.
Data: Obsah manganu v oceli [%]:

1.17	1.22	1.26	1.10	1.15	1.21	1.21	1.12	1.26	1.23	1.26	1.18
..	..	..	..	..	..	..	..	..	..	..	..
0.96	1.21	1.01	1.00								

2.5.5 Analýza ekonomických a sociologických dat

Úloha S2.01 *Kvantilový graf k výstupní kontrole hmotnosti balíčků rýže*
Kontrolní vážení obsahu 20 "jednokilogramových" balíčků rýže mělo prověřit poctivost pracovníků balíren (str. 21 v cit.[15]). Určete zvláštnosti a typ rozdělení dat a ověřte, zda data pochází z Gaussova rozdělení se střední hodnotou 1000. Je třeba užít transformaci dat? Z kvantilového grafu odečtěte, kolik procent balíčků rýže leží nad hodnotou 1000 g a kolik pod ní?
Data: Hmotnost jednokilogramového balíčku rýže [g]:

1004	1005	1002	1011	998	1001	997	993	1000	1002
1003	995	1004	999	1010	1006	1002	994	1000	999

Úloha S2.02 *Kvantilová míra variability věku pojištěnců a stáří automobilu*
Vedení pojišťovny požádalo oddělení marketingového výzkumu o průzkum názorů zákazníků na uvažovaný systém pojištění aut. Náhodně bylo vybráno 100 současných pojištěnců a u nich byl zjišťován kromě názoru na nový systém pojištění také věk pojištěnců a stáří auta. Proveďte průzkumovou analýzu zjištěných dat. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. U kolika diagnostik je shoda v indikaci odlehlých bodů? Vede graf maximální věrohodnosti k nutnosti použití transformace dat?

Data: *S202a*: Věk pojištěnců [roky]:

45	34	58	46	47	35	34	32	48	49	57	24	35	34	39	48	28	34	46	59
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
32	38	37	45	46	47	49	54	37	58	37	36	32	46	49	48	47	47	37	29

S202b: Stáří aut [roky]:

4	3	8	7	9	7	10	8	7	4	5	5	6	4	3	2	5	8	7	9
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
3	1	1	2	2	2	3	3	4	3	5	2	2	2	3	5	6	6	6	4

Úloha S2.03 *Kontrola odlehlých hodnot u automatem dávkované hmotnosti pytlů*
Ve dvou cementárnách byla provedena kontrola dávkovačů: (a) Zvážením 13 náhodně vybraných pytlů cementu byly získány hodnoty výběru. (b) Obdobně zvážením 9 náhodně vybraných pytlů z druhé cementárny byly získány hodnoty výběru. Proveďte průzkumovou analýzu těchto výběrů a nalezněte kvantilové míry polohy, rozptýlení a tvaru. Jsou obě rozdělení symetrická a obsahují odlehlé hodnoty?

Data: *S203a*: Hmotnost pytle cementu [kg]:

51.5	47.0	48.5	53.0	47.3	48.1	48.8
49.2	52.3	47.1	49.5	46.3	50.1	

S203b: Hmotnost pytle cementu [kg]:

50.3	50.7	49.2	50.1	49.9	51.1	49.8	48.9	50.3

Úloha S2.04 *Vyšetření rozdělení výsledků Amthaureova testu u 98 studentů*
V rámci přijímacího řízení absolvují uchazeči o studium na vysoké škole Amthauerův test struktury inteligence. Výsledky tohoto testu se vyjadřují prostřednictvím tzv. celkového hrubého skóre. Ze studentů přijatých ke studiu během 4 let byl proveden náhodný výběr 98 studentů. Proveďte průzkumovou analýzu a ověření předpokladů pro tento výběr. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Je rozdělení symetrické? U kolika diagnostik je shoda v indikaci odlehlých bodů?

Data: Výsledek Amthaureova testu u 98 studentů [skóre]:

77	105	110	88	128	104	94	104	129	96	82	120	102	80	103	101	147	112	120	104
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
99	146	99	104	109	116	124	132	125	109	134	113	118	122	127	131	110	117		

Úloha S2.05 *Kvantilová míra polohy a tabulka rozdělení četnosti u kontroly jízdenek*
Při 20 kontrolách jízdenek v pražském metru byl zaznamenán počet cestujících pokutovaných za jízdu načerno. Sestavte tabulku rozdělení četností a rozhodněte jaký typ diskrétního rozdělení vyhovuje datům. Vyčíslete také kvantilovou míru polohy.

Data: Počet cestujících s pokutou na 1 kontrolu [počet]:

1	3	2	4	2	5	3	3	4	0	3	2	7	1	4	5	2	6	3	4

Úloha S2.06 *Odlehlé hodnoty u měsíční spotřeby elektrické energie v domácnosti*
Z údajů o měsíční spotřebě elektrické energie (kWh) v 25 bytech určete zvláštnosti rozdělení dat. Vyčíslete kvantilové charakteristiky šikmosti a špičatosti: polosumu Z_L, rozpětí R_L, šikmosti S_L, pseudosigmu G_L a délek konců T_L pro kvartily a oktily a ukažte, jak charakterizují symetrii (tj. Z_L a S_L), rozptýlení (tj. R_L) a špičatost (tj. G_L a T_L). Vyšetřete tvar rozdělení na základě grafu polosum, symetrie a špičatosti (symboly viz cit.[19]). Které diagnostiky shodně indikují odlehlé hodnoty? Dosahuje míra rozptýlení nepřijatelně vysoké hodnoty? Jaká je hloubka prvku 120 ve výběru?

Data: Měsíční spotřeba elektrické energie v domácnosti [kWh]:

169	108	26	43	114	68	35	183	103	266	74	205	62
230	85	487	120	148	91	18	58	96	295	137	42	

Úloha S2.07 *Rozdělení doby čekání zákazníka*
U 50 zákazníků byla měřena doba čekání v minutách na obsluhu v prodejně. Z naměřených hodnot je třeba určit typ diskrétního rozdělení.

Data: Doba čekání zákazníka [min]:

3	10	2	2	16	7	7	14	7	16	5	1	2	13	8	5	13
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
0	5	4	1	7	8	4	9	5	11	8	10	6	1	6	7	

Úloha S2.08 *Vyšetření typu rozdělení ročního průměrného počtu zaměstnanců podniku*
Určete typ rozdělení výběru průměrného počtu zaměstnanců podniku, jsou-li dána následující personální data o průměrném počtu ke třem dnům v měsíci. Na základě vnitřních a vnějších hradeb v bariérově-číslicovém zápise rozhodněte, zda je rozdělení symetrické a zda obsahuje výběr odlehlé hodnoty? Jaký tvar má rozdělení výběru a jakého je typu?

Data: Průměrný počet zaměstnanců v podniku k datu [počet]:

5. 1.	15. 1.	25. 1.	5. 2.	15. 2.	25. 2.	5. 3.	15. 3.	25. 3.	5. 4.
211	211	210	209	210	212	213	214	215	215
..	..	..	..	..	..	..	..	..	..
5.11.	15.11.	25.11.	5.12.	15.12.	25.12.				
260	260	261	261	261	262				

Úloha S2.09 *Kvantily a typ rozdělení příjmů a výdajů domácnosti*
30 vybraných pražských domácností poskytlo data za čtvrletí roku, a to k datu 1. 1. 1961: (a) příjem, (b) výdaje za potraviny a (c) nájem v Kčs. Sestavte rozdělení četností o čtvrtletním příjmu, vydání za potraviny a konečně i za nájem a znázorněte je histogramem. Určete též relativní četnosti a kumulativní četnosti. Z písmenových hodnot a kvantilových měr usuzujte na symetrii výběrového rozdělení.

Data: *S209a* příjem, *S209b* výdaje za potraviny a *S209c* nájem u vybraných domácností [Kčs]:

9966	4158	484,	5105	2925	114,	13911	6336	343,
..	..	..	..	..	..	..	..	..
14165	5801	405,	5165	2043	105,	7260	2455	298,

Úloha S2.10 *Exploratorní analýza při zobrazení údajů o průtoku řek*
Proveďte přehledné zobrazení údajů o průtoku řek naší a Slovenské republiky v měsících leden až prosinec. Je rozdělení jednotlivých výběrů symetrické? Užijte písmenových hodnot a kvantilové míry polohy.

Data: Průtoky vybraných řek [$m^3 . s^{-1}$]:

S210a Labe:	154	138	167	125	103	63	52
	40	51	50	60	144		
...	...	...	...	...	...	...	...
S210g Bodrog:	48	94	290	159	384	129	58
	54	56	42	43	93		

Úloha S2.11 *Vyšetření typu rozdělení údajů o IQ žáků*
Data obsahují inteligenční kvocient IQ dvou skupin 866 žáků 8. tříd, jednak skupiny žáků, kteří byli přijati do gymnázií, a jednak žáků, kteří se nedostali na žádnou střední školu. Jakému rozdělení odpovídají data? Vykazují prvky výběru dat homogenitu a nezávislost? Jde o unimodální rozdělení? Lze posoudit symetrii rozdělení na základě dolních a horních kvartilů, oktilů a sedecilů? Vyčíslete kvantilové charakteristiky šikmosti S_L a délky konců T_L v oblasti oktilů E a sedecilů D. Usuzujte na typ rozdělení dle délky konců T_L.
Data: Hladina inteligenčního kvocientu IQ dvou skupin 866 žáků 8. tříd [skóre]:

94	101	109	116	128	100	75	123	82	123	94	92	113	115	103	80	92
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
123	130	106	88	120	119	105	82	116	103	73	90	125	66	124	111	

Úloha S2.12 *Určení typu a symetrie rozdělení počtu nezaměstnaných*
Sestrojte rozdělení výběru věku 500 nezaměstnaných mužů (USA 1989, data jsou získána z CPS). Aplikujte Sturgesovo pravidlo pro optimální počet tříd v histogramu $k \doteq 1 + 3.3 \log n$, kde n je velikost výběru. Vyčíslete také kvantilové míry polohy, rozptýlení a tvaru. Jaké jsou míry tvaru? Je rozdělení symetrické? Jaká je hloubka prvku 50 ve výběru? Kolik procent mužů ve výběru je mladších 50 let?
Data: Věk nezaměstnaného [roky]:

60	61	56	55	54	63	64	56	63	57	68	67	61	64	66	63	65	61
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
64	67	66	67	53	62	66	53	64	51	63	62	60	66				

Úloha S2.13 *Posouzení tvaru rozdělení a symetrie výběru délky skoků dvou sportovců*
Dva lehcí atleti-skokani, Jan Skokan (první hodnota ve dvojici dat) a John Jumper, jsou porovnáváni co do vyrovnanosti svých skoků do dálky. Na základě kvantilové míry polohy rozhodněte, zda jsou obě výběrová rozdělení symetrická a bez odlehlých hodnot. Nalezněte míry polohy, rozptýlení a tvaru a odpovězte na otázku, který ze sportovců se jeví lepší? Je výhodnější vyrovnanost skoků nebo ojedinělý rekordní skok tzn. horní odlehlá hodnota?
Data: Délky skoků [cm] u *S213a* Jana Skokana a u *S213b* Johna Jumpera ve dvojicích:

806	808,	802	805,	821	788,	795	806,	783	792,
..	..	..	..	..	..	..	..	..	..
799	799,	793	798,	790	800,	786	803,	809	791,

Úloha S2.14 *Určení typu rozdělení hodnot hustoty Země dle Cavendishe (1798)*
Měření hustoty Země H. Cavendishem v roce 1798 je na svou dobu pozoruhodné zvláště, když si uvědomíme, že dnešní měření přináší hodnotu blízkou, a to okolo 5.517. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Určete typ rozdělení výběru a především vyšetřete, zda jde o symetrické rozdělení? U kolika diagnostik je shoda v indikaci odlehlých bodů? Vede graf maximální věrohodnosti k nutnosti použití transformace dat?
Data: Hustota země [kg/dm], naměřená Cavendishem v roce 1798:

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
..	..	..	..	..	..	..	..	..	..
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Úloha S2.15 *Rozdělení výběru cen piva v 45 pražských hospodách v listopadu 1991*
Posuďte rozdělení výběru cen 10 E piva ve 45 náhodně vybraných restauracích nižší cenové skupiny v Praze v listopadu 1991. Jsou v datech i odlehlé hodnoty? U kolika diagnostik je shoda v indikaci odlehlých bodů? Vede graf maximální věrohodnosti k potvrzení nutnosti užití transformace dat?
Data: Cena piva [Kč] v 45 pražských hospodách v listopadu 1991:

4.50	4.40	4.30	4.20	4.30	4.30	4.20	4.20	4.30	4.00	4.70	4.20
..	..	..	..	..	..	..	..	..	..	..	..
4.00	4.40	4.20	4.00	4.20	4.20	4.40	4.30	4.90			

Úloha S2.16 *Určení typu rozdělení výšky výběru středoškolských studentů*
Určete střední hodnotu výšky středoškolského studenta v cm analýzou hodnot výšky 48 studentů z gymnázií. Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru a rozhodněte, zda je ve výběru odlehlá hodnota, tj. abnormálně vysoký či malý student? U kolika diagnostik dochází ke shodě v indikaci odlehlých bodů? Vede graf maximální věrohodnosti k potvrzení nutnosti použití transformace dat?
Data: Výška studenta z českých gymnázií [cm]:

165	170	170	179	170	168	174	162	167	165	170	173	183	176	165	168	171
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
162	166	170	168	155	162	169	166	160	169	165	163	168	163			

Úloha S2.17 *Určení typu rozdělení hmotnosti výběru středoškolských studentů*
Určete typ rozdělení hmotnosti středoškolského studenta v kg, analýzou hmotnosti 48 studentů z českých gymnázií. Na základě kvantilových měr polohy rozhodněte, zda existuje ve výběru odlehlá hodnota, tj. abnormálně těžký či lehký student? U kolika diagnostik se nachází shoda v indikaci odlehlých bodů? Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vede graf maximální věrohodnosti k nutnosti použití transformace dat? Kolik procent studentů váží méně než 55 kg?
Data: Hmotnost studenta z českých gymnázií [kg]:

70	68	56	62	70	58	60	47	51	63	62	57	68	82	50	57	52	58	64	55	56	52	59	62
53	65	66	58	58	56	57	53	62	54	52	54	58	64	58	52	56	50	64	55	58	51	61	60

Úloha S2.18 *Určení typu rozdělení rozměrů okvětních lístků u 150 kosatců (Fisherova úloha)*
Sestrojte jádrový odhad hustoty pravděpodobnosti rozdělení čtyř výběrů, obsahujících čtvero popisných rozměrů okvětních lístků 150 druhů květů kosatců, pocházejících ze tří základních druhů: Iris setosa, Iris versicolor, Iris virginica. Z botaniky je známo, že druh Iris versicolor je hybridem zbývajících dvou druhů. Iris setosa je diploidní květ s 38 chromozomy, Iris virginica je tetraploidní a Iris versicolor je hexaploidní s 108 chromozomy. U kolika diagnostik je shoda v indikaci odlehlých bodů? Nalezněte vhodný kvantilový odhad parametru polohy a variability ale také míry tvaru rozdělení u výběrů: (a) délky kališních lístků v cm, anglicky *lsepal,* (b) jejich šířky, *wsepal,* (c) délky korunních plátků v cm, *lpetal*, a (d) jejich šířky, *wpetal*.
Data: Rozměry okvětních lístků [cm] u 150 kosatců: *S218a* Lsepal, *S218b* Wsepal, *S218c* Lpetal, *S218d* wpetal

ve čtveřicích:

5.1	3.5	1.4	0.2,	4.9	3.0	1.4	0.2,	4.7	3.2	1.3	0.2,	4.6	3.1	1.5	0.2,
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
5.4	3.4	1.7	0.2,	5.1	3.7	1.5	0.4,								

Úloha S2.19 *Určení typu a symetrie rozdělení výnosu žita a brambor u zemědělských farem*
Porovnejte vyrovnanost 20 zemědělských farem z hlediska hektarových výnosů žita a brambor. Určete typ rozdělení obou výběrů. Která plodina má větší absolutní variabilitu? Kterou relativní míru variability využijete k požadovanému srovnání? U kolika diagnostik je shoda v indikaci odlehlých bodů? Zkonstruujte bariérově-číslicové schéma formou sedmipísmenového zápisu výběru. Vede graf maximální věrohodnosti k nutnosti použití transformace dat?
Data:
S219a Výnos žita [q/ha] u zemědělských farem:

18.0	18.2	18.5	18.6	19.0	19.1	19.3	19.4	19.5	19.7
19.8	20.0	20.2	20.4	20.5	20.8	21.0	21.1	21.2	21.5

S219b Výnos brambor [q/ha] u zemědělských farem:

92	93	94	96	98	98	99	100	100	100
100	100	101	101	102	102	103	104	106	108

Úloha S2.20 *Vyšetření symetrie rozdělení u výběru příjmů a vydání v rodině*
Vyšetřete rozdělení u výběrů příjmů a vydání v rodině na potraviny v Praze v roce 1967. Na základě písmenových hodnot, polosumy, rozpětí, pseudosigmy a relativní délky konců rozhodněte, zda je rozdělení výběru symetrické a blízké Gaussovu normálnímu rozdělení? Jsou ve výběrech odlehlé hodnoty?
Data: *S220a* Příjem [Kčs], *S220b* vydání [Kčs] na potraviny v roce 1967:

14928	6512,	13213	6925,	13809	5817,	16983	7729,	15963	2934,	20016	6076,
..	..	..	..	..	..	..	..	..	..	..	..
17875	6003,	11429	3749,	13804	4718,	15153	5226,	17962	7415,		

Úloha S2.21 *Kvantilové vyšetření symetrie rozdělení u výběrů ukazatelů kvality aut*
Americká databáze aut obsahuje řadu ukazatelů kvality auta: (a) spotřeba je uvedena v mílích na galon, (b) výkon v HP, (c) zrychlení v yardech/s a (d) hmotnost v librách. Určete typ rozdělení výběru spotřeby, výkonu, zrychlení a hmotnosti a posuďte, zda je symetrické a blízké Gaussovu? Zdůvodněte, zda je nutné užít transformaci dat nebo dáte přednost robustním odhadům střední hodnoty?
Data: Ukazatele kvality aut: *S221a* spotřeba [míle/gallon], *S221b* výkon [HP], *S221c* zrychlení [yard/s] a *S221d* hmotnost [libry]:

S221a	*S221b*	*S221c*	*S221d*	*S221e*	
Spotřeba	Výkon	Zrychlení	Hmotnost	Typ a značka auta	
43.1	48	21.5	1985	Volkswagen	
...	...	...	...	...	...
31.0	82	19.4	2720	Chevrolet	

2.6 Kontrolní hodnoty (ADSTAT, NCSS2000)

2.6.1 Analýza farmakologických a biochemických dat

B2.01 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 3.89, $\tilde{x}_{0.5}$ = 3.92, $\bar{x}_R$ = 3.90, s = 0.20, $\hat{g}_1$ = -0.12, $\hat{g}_2$ = 2.40, 3.84 < μ < 3.96 (Box).

B2.02 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.6479, $\tilde{x}_{0.5}$ = 0.6517, $\bar{x}_R$ = 0.6504, s = 0.0133, $\hat{g}_1$ = -1.36, $\hat{g}_2$ = 4.42, 0.6453 < μ < 0.6552 (Box).

B2.03 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 50.45, $\tilde{x}_{0.5}$ = 50.33, $\bar{x}_R$ = 50.33, s = 1.17, $\hat{g}_1$ = 2.73, $\hat{g}_2$ = 14.54, 49.94 < μ < 50.77 (Box).

B2.04 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 0.4599, $\tilde{x}_{0.5}$ = 0.1400, $\bar{x}_R$ = 0.2014, s = 0.7970, $\hat{g}_1$ = 2.36, $\hat{g}_2$ = 7.46, 0.1100 < μ < 0.1950 (Box).

B2.05 Poissonovo rozdělení,

B2.06 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 6, $\bar{x}$ = 4.83, $\tilde{x}_{0.5}$ = 4.85, $\bar{x}_R$ = 4.81, s = 1.10, $\hat{g}_1$ = 0.24, $\hat{g}_2$ = 4.88, 4.54 < μ < 5.07 (Box).

B2.07A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 19.98, $\tilde{x}_{0.5}$ = 19.99, $\bar{x}_R$ = 19.86, s = 0.12, $\hat{g}_1$ = -0.39, $\hat{g}_2$ = 2.27, 19.92 < μ < 20.02 (Box).

B2.07B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 80.00, $\tilde{x}_{0.5}$ = 80.02, $\bar{x}_R$ = 80.01, s = 0.19, $\hat{g}_1$ = -0.29, $\hat{g}_2$ = 2.25, 79.96 < μ < 80.07 (Box).

B2.08 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 1.56, $\tilde{x}_{0.5}$ = 1.58, $\bar{x}_R$ = 1.56, s = 0.33, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 2.15, 1.44 < μ < 1.68 (Box).

B2.09 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: **1,** $\bar{x}$ = 0.737, $\tilde{x}_{0.5}$ = 0.733, $\bar{x}_R$ = 0.739, s = 0.027, $\hat{g}_1$ = -0.68, $\hat{g}_2$ = 3.93, 0.730 < μ < 0.749 (Box).

B2.10 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 0.79, $\tilde{x}_{0.5}$ = 0.80, $\bar{x}_R$ = 0.77, s = 0.21, $\hat{g}_1$ = 0.86, $\hat{g}_2$ = 5.66, 0.71 < μ < 0.84 (Box).

B2.11A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 12.1, $\tilde{x}_{0.5}$ = 8.0, $\bar{x}_R$ = 8.8, s = 12.4, $\hat{g}_1$ = 2.80, $\hat{g}_2$ = 14.21, 7.4 < μ < 8.7 (Box).

B2.11B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 26.4, $\tilde{x}_{0.5}$ = 20.0, $\bar{x}_R$ = 21.01 s = 20.9, $\hat{g}_1$ = 1.64, $\hat{g}_2$ = 6.51, 18.6 < μ < 21.8 (Box).

B2.11C Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 14.1, $\tilde{x}_{0.5}$ = 10.0, $\bar{x}_R$ = 10.3, s = 13.9, $\hat{g}_1$ = 2.63, $\hat{g}_2$ = 12.88, 8.7 < μ < 10.2 (Box).

B2.11D Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 31.7, $\tilde{x}_{0.5}$ = 26.0, $\bar{x}_R$ = 26.4, s = 22.7, $\hat{g}_1$ = 1.41, $\hat{g}_2$ = 5.57, 23.7 < μ < 27.6 (Box).

B2.12 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 80.0, $\tilde{x}_{0.5}$ = 79.8, $\bar{x}_R$ = 79.9, s = 9.1, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 2.58, 78.2 < μ < 81.8 (Box).

B2.15 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 10.0, $\tilde{x}_{0.5}$ = 10.0, $\bar{x}_R$ = 10.0, s = 1.9, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 1.91, 9.5 < μ < 10.5 (Box).

B2.16 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 8.3, $\tilde{x}_{0.5}$ = 7.0, $\bar{x}_R$ = 7.1, s = 4.8, $\hat{g}_1$ = 3.18, $\hat{g}_2$ = 13.19, 6.4 < μ < 7.8 (Box).

B2.17 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 4.15,

$\tilde{x}_{0.5}$ = 3.95, $\overline{x}_R$ = 4.00, s = 1.02, $\hat{g}_1$ = 1.47, $\hat{g}_2$ = 6.82, 3.76 < μ < 4.26 (Box).

B2.18A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 3.59, $\tilde{x}_{0.5}$ = 3.55, $\overline{x}_R$ = 3.57, s = 0.20, $\hat{g}_1$ = 0.41, $\hat{g}_2$ = 2.53, 3.51 < μ < 3.63 (Box).

B2.18B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 7.02, $\tilde{x}_{0.5}$ = 6.95, $\overline{x}_R$ = 7.00, s = 0.48, $\hat{g}_1$ = 0.10, $\hat{g}_2$ = 2.60, 6.86 < μ < 7.16 (Box).

B2.18C Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 11.87, $\tilde{x}_{0.5}$ = 11.85, $\overline{x}_R$ = 11.86, s = 0.88, $\hat{g}_1$ = 0.09, $\hat{g}_2$ = 3.22, 11.58 < μ < 12.14 (Box).

B2.19 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\overline{x}$ = 2.34, $\tilde{x}_{0.5}$ = 2.13, $\overline{x}_R$ = 2.08, s = 1.09, $\hat{g}_1$ = 3.21, $\hat{g}_2$ = 14.96, 1.90 < μ < 2.25 (Box).

B2.20A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\overline{x}$ = 41.7, $\tilde{x}_{0.5}$ = 41.3, $\overline{x}_R$ = 42.6, s = 10.3, $\hat{g}_1$ = -0.58, $\hat{g}_2$ = 3.53, 39.4 < μ < 45.7 (Box).

B2.20B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 60.44, $\tilde{x}_{0.5}$ = 62.57, $\overline{x}_R$ = 61.81, s = 19.44, $\hat{g}_1$ = -0.46, $\hat{g}_2$ = 3.13, 55.91 < μ < 68.16 (Box).

B2.21 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 59.80, $\tilde{x}_{0.5}$ = 59.53, $\overline{x}_R$ = 59.44, s = 1.75, $\hat{g}_1$ = 4.06, $\hat{g}_2$ = 20.98, 59.15 < μ < 59.75 (Box).

B2.22 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\overline{x}$ = 0.3067, $\tilde{x}_{0.5}$ = 0.1790, $\overline{x}_R$ = 0.1890, s = 0.4270, $\hat{g}_1$ = 3.17, $\hat{g}_2$ = 13.79, 0.1410 < μ < 0.1950 (Box).

B2.23 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\overline{x}$ = 44.0, $\tilde{x}_{0.5}$ = 42.8, $\overline{x}_R$ = 43.5, s = 2.2, $\hat{g}_1$ = 0.46, $\hat{g}_2$ = 1.69, 42.7 < μ < 43.7 (Box).

B2.24 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 1.41, $\tilde{x}_{0.5}$ = 1.40, $\overline{x}_R$ = 1.34, s = 0.46, $\hat{g}_1$ = 1.29, $\hat{g}_2$ = 6.71, 1.22 < μ < 1.46 (Box).

B2.25 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\overline{x}$ = 368.5, $\tilde{x}_{0.5}$ = 356.0, $\overline{x}_R$ = 352.1, s = 73.0, $\hat{g}_1$ = 1.33, $\hat{g}_2$ = 5.00, 334.9 < μ < 367.0 (Box).

B2.26 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\overline{x}$ = 1.73, $\tilde{x}_{0.5}$ = 1.45, $\overline{x}_R$ = 1.54, s = 0.88, $\hat{g}_1$ = 1.00, $\hat{g}_2$ = 3.35, 1.33 < μ < 1.75 (Box).

B2.27 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 3.15, $\tilde{x}_{0.5}$ = 3.14, $\overline{x}_R$ = 3.13, s = 0.09, $\hat{g}_1$ = 4.45, $\hat{g}_2$ = 29.81, 3.12 < μ < 3.14 (Box).

2.6.2 Analýza chemických a fyzikálních dat

C2.01 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 8, $\overline{x}$ = 5.09, $\tilde{x}_{0.5}$ = 5.08, $\overline{x}_R$ = 5.06, s = 1.38, $\hat{g}_1$ = 0.40, $\hat{g}_2$ = 5.57, 4.76 < μ < 5.33 (Box).

C2.02 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 12, $\overline{x}$ = 0.36, $\tilde{x}_{0.5}$ = 0.24, $\overline{x}_R$ = 0.26, s = 0.38, $\hat{g}_1$ = 2.28, $\hat{g}_2$ = 9.14, 0.20 < μ < 0.29 (Box).

C2.03 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\overline{x}$ = 0.44, $\tilde{x}_{0.5}$ = 0.33, $\overline{x}_R$ = 0.34, s = 0.37, $\hat{g}_1$ = 1.60, $\hat{g}_2$ = 5.42, 0.28 < μ < 0.38 (Box).

C2.04 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\overline{x}$ = 47.43, $\tilde{x}_{0.5}$ = 47.50, $\overline{x}_R$ = 46.42, s = 7.40, $\hat{g}_1$ = 0.54, $\hat{g}_2$ = 2.56, 43.91 < μ < 49.11 (Box).

C2.05 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 6, $\overline{x}$ = 0.17,

$\tilde{x}_{0.5}$ = 0.14, $\overline{x}_R$ = 0.14, s = 0.13, $\hat{g}_1$ = 2.43, $\hat{g}_2$ = 10.45, 0.11 < μ < 0.16 (Box).

C2.06 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\overline{x}$ = 0.67, $\tilde{x}_{0.5}$ = 0.40, $\overline{x}_R$ = 0.48, s = 0.61, $\hat{g}_1$ = 2.04, $\hat{g}_2$ = 6.41, 0.39 < μ < 0.510 (Box).

C2.07 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 64.450, $\tilde{x}_{0.5}$ = 64.600, $\overline{x}_R$ = 64.230, s = 2.130, $\hat{g}_1$ = 0.35, $\hat{g}_2$ = 2.18, 63.4100 < μ < 65.1100 (Box).

C2.08 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 25.155, $\tilde{x}_{0.5}$ = 25.150, $\overline{x}_R$ = 25.110, s = 2.820, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 2.09, 24.430 < μ < 25.890 (Box).

C2.09 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 2191.9, $\tilde{x}_{0.5}$ = 2196.0, $\overline{x}_R$ = 2194.1, s = 51.4, $\hat{g}_1$ = -0.28, $\hat{g}_2$ = 2.89, 2174.8 < μ < 2213.5 (Box).

C2.10 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 63.59, $\tilde{x}_{0.5}$ = 63.53, $\overline{x}_R$ = 63.56, s = 0.35, $\hat{g}_1$ = 0.41, $\hat{g}_2$ = 2.87, 63.48 < μ < 63.65 (Box).

C2.11 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\overline{x}$ = 14.5800, $\tilde{x}_{0.5}$ = 14.5700, $\overline{x}_R$ = 14.5850, s = 0.7800, $\hat{g}_1$ = 0.03, $\hat{g}_2$ = 4.44, 14.3800 < μ < 14.7800 (Box).

C2.12A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 132.4200, $\tilde{x}_{0.5}$ = 132.6500, $\overline{x}_R$ = 132.5600, s = 1.4400, $\hat{g}_1$ = -1.57, $\hat{g}_2$ = 9.13, 132.1300 < μ < 132.9700 (Box).

C2.12B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\overline{x}$ = 158.80, $\tilde{x}_{0.5}$ = 159.80, $\overline{x}_R$ = 159.47, s = 6.22, $\hat{g}_1$ = -1.68, $\hat{g}_2$ = 8.86, 156.99 < μ < 161.51 (Box).

C2.13 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 10.1, $\tilde{x}_{0.5}$ = 10.1, $\overline{x}_R$ = 10.1, s = 0.4, $\hat{g}_1$ = -1.02, $\hat{g}_2$ = 6.52, 10.0 < μ < 10.3 (Box).

C2.14 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 0.06, $\tilde{x}_{0.5}$ = 0.05, $\overline{x}_R$ = 0.05, s = 0.04, $\hat{g}_1$ = 2.70, $\hat{g}_2$ = 11.61, 0.04 < μ < 0.06 (Box).

C2.15 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 0.423, $\tilde{x}_{0.5}$ = 0.420, $\overline{x}_R$ = 0.424, s = 0.038, $\hat{g}_1$ = -0.37, $\hat{g}_2$ = 4.21, 0.412 < μ < 0.436 (Box).

C2.16 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\overline{x}$ = 38.5, $\tilde{x}_{0.5}$ = 39.0, $\overline{x}_R$ = 38.9, s = 4.7, $\hat{g}_1$ = -0.39, $\hat{g}_2$ = 2.10, 37.9 < μ < 40.3 (Box).

C2.17 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 181.9, $\tilde{x}_{0.5}$ = 183.0, $\overline{x}_R$ = 183.1, s = 13.4, $\hat{g}_1$ = -0.87, $\hat{g}_2$ = 4.87, 179.0 < μ < 187.1 (Box).

C2.18 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 92.65, $\tilde{x}_{0.5}$ = 92.60, $\overline{x}_R$ = 92.44, s = 1.34, $\hat{g}_1$ = 1.80, $\hat{g}_2$ = 8.80, 92.05 < μ < 92.89 (Box).

C2.19 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 80.78, $\tilde{x}_{0.5}$ = 80.70, $\overline{x}_R$ = 80.46, s = 2.41, $\hat{g}_1$ = 1.58, $\hat{g}_2$ = 7.64, 79.85 < μ < 81.14 (Box).

C2.20 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\overline{x}$ = 81.47, $\tilde{x}_{0.5}$ = 80.80, $\overline{x}_R$ = 81.13, s = 6.71, $\hat{g}_1$ = 1.90, $\hat{g}_2$ = 13.00, 79.20 < μ < 83.00 (Box).

C2.21 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 0.737, $\tilde{x}_{0.5}$ = 0.741, $\overline{x}_R$ = 0.737, s = 0.065, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 3.36, 0.706 < μ < 0.768 (Box).

C2.22 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\overline{x}$ = 6735.80, $\tilde{x}_{0.5}$ = 6738.70, $\overline{x}_R$ = 6740.50, s = 46.30, $\hat{g}_1$ = -0.80, $\hat{g}_2$ = 4.17, 6729.50 < μ < 6752.70 (Box).

C2.23 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 0.0993, $\tilde{x}_{0.5}$ = 0.0995, $\overline{x}_R$ = 0.0991, s = 0.0085, $\hat{g}_1$ = 0.73, $\hat{g}_2$ = 9.40, 0.0963 < μ < 0.1017 (Box).

C2.24 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 5.38, $\tilde{x}_{0.5}$ = 5.33, $\bar{x}_R$ = 5.41, s = 0.57, $\hat{g}_1$ = -0.38, $\hat{g}_2$ = 3.91, 5.28 < μ < 5.52 (Box).

C2.25 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 0.12, $\tilde{x}_{0.5}$ = 0.12, $\bar{x}_R$ = 0.12, s = 0.02, $\hat{g}_1$ = 0.44, $\hat{g}_2$ = 3.27, 0.120 < μ < 0.126 (Box).

C2.26 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 26.81, $\tilde{x}_{0.5}$ = 26.80, $\bar{x}_R$ = 26.80, s = 0.15, $\hat{g}_1$ = 0.49, $\hat{g}_2$ = 2.65, 26.77 < μ < 26.82 (Box).

C2.27 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 3.2, $\tilde{x}_{0.5}$ = 3.2, $\bar{x}_R$ = 3.1, s = 0.4, $\hat{g}_1$ = 0.37, $\hat{g}_2$ = 2.64, 3.0 < μ < 3.3 (Box).

C2.28 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 56.6 $\tilde{x}_{0.5}$ = 56.5, $\bar{x}_R$ = 56.6, s = 0.7, $\hat{g}_1$ = 0.54, $\hat{g}_2$ = 2.41, 56.3 < μ < 56.8 (Box).

C2.29 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 91.85, $\tilde{x}_{0.5}$ = 92.80, $\bar{x}_R$ = 92.17, s = 3.54, $\hat{g}_1$ = -0.56, $\hat{g}_2$ = 3.00, 90.63 < μ < 93.71 (Box).

C2.30 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 86.04, $\tilde{x}_{0.5}$ = 86.40, $\bar{x}_R$ = 86.60, s = 5.33, $\hat{g}_1$ = -0.60, $\hat{g}_2$ = 2.78, 84.49 < μ < 88.99 (Box).

C2.31 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 2.759, $\tilde{x}_{0.5}$ = 2.790, $\bar{x}_R$ = 2.814, s = 0.486, $\hat{g}_1$ = -1.06, $\hat{g}_2$ = 5.12, 2.650 < μ < 2.990 (Box).

C2.31A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 2.964, $\tilde{x}_{0.5}$ = 2.975, $\bar{x}_R$ = 2.985, s = 0.367, $\hat{g}_1$ = -0.30, $\hat{g}_2$ = 2.08, 2.780 < μ < 3.190 (Box).

C2.31B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 2.555, $\tilde{x}_{0.5}$ = 2.663, $\bar{x}_R$ = 2.621, s = 0.514, $\hat{g}_1$ = -1.16, $\hat{g}_2$ = 4.89, 2.330 < μ < 2.870 (Box).

C2.32 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 0.271, $\tilde{x}_{0.5}$ = 0.271, $\bar{x}_R$ = 0.271, s = 0.002, $\hat{g}_1$ = 0.73, $\hat{g}_2$ = 2.90, 0.270 < μ < 0.272 (Box).

C2.33 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 1.19, $\tilde{x}_{0.5}$ = 1.19, $\bar{x}_R$ = 1.19, s = 0.09, $\hat{g}_1$ = 0.09, $\hat{g}_2$ = 3.22, 1.16 < μ < 1.21 (Box).

2.6.3 Analýza environmentálních, potravinářských a zemědělských dat

E2.01 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 26.86, $\tilde{x}_{0.5}$ = 22.41, $\bar{x}_R$ = 20.41, s = 25.54, $\hat{g}_1$ = 1.96, $\hat{g}_2$ = 6.70, 13.66 < μ < 27.94 (Box).

E2.02 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.50, $\tilde{x}_{0.5}$ = 0.50, $\bar{x}_R$ = 0.50, s = 0.02, $\hat{g}_1$ = -0.26, $\hat{g}_2$ = 2.69, 0.50 < μ < 0.51 (Box).

E2.03 Symetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 1.26, $\tilde{x}_{0.5}$ = 1.25, $\bar{x}_R$ = 1.26, s = 0.02, $\hat{g}_1$ = 7.40, $\hat{g}_2$ = 1.76, 1.25 < μ < 1.26 (Box).

E2.04 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 0.23, $\tilde{x}_{0.5}$ = 0.19, $\bar{x}_R$ = 0.21, s = 0.12, $\hat{g}_1$ = 1.40, $\hat{g}_2$ = 6.39, 0.18 < μ < 0.25 (Box).

E2.05 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 60.6, $\tilde{x}_{0.5}$ = 60.1, $\bar{x}_R$ = 60.3, s = 3.0, $\hat{g}_1$ = 1.00, $\hat{g}_2$ = 5.33, 59.4 < μ < 61.3 (Box).

E2.06 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 0.1947, $\tilde{x}_{0.5}$ = 0.1995, $\bar{x}_R$ = 0.1940, s = 0.0232, $\hat{g}_1$ = 0.18, $\hat{g}_2$ = 3.03, 0.1870 < μ < 0.2000 (Box).

E2.07 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 186.16, $\tilde{x}_{0.5}$ = 183.00, $\bar{x}_R$ = 183.64, s = 15.00, $\hat{g}_1$ = 2.04, $\hat{g}_2$ = 8.24, 179.80 < μ < 188.24 (Box).

E2.08 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 26.9, $\tilde{x}_{0.5}$ = 26.6, $\bar{x}_R$ = 26.9, s = 1.2, $\hat{g}_1$ = 0.35, $\hat{g}_2$ = 3.55, 26.3 < μ < 27.4 (Box).

E2.09 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 12163.9, $\tilde{x}_{0.5}$ = 11125.0, $\bar{x}_R$ = 11322.9, s = 5768.0, $\hat{g}_1$ = 1.64, $\hat{g}_2$ = 7.28, 10049.0 < μ < 12736.5 (Box).

E2.10 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 57.32, $\tilde{x}_{0.5}$ = 57.56, $\bar{x}_R$ = 57.49, s = 1.28, $\hat{g}_1$ = -1.13, $\hat{g}_2$ = 4.70, 56.93 < μ < 58.03 (Box).

E2.11 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 1.26, $\tilde{x}_{0.5}$ = 1.25, $\bar{x}_R$ = 1.25, s = 0.02, $\hat{g}_1$ = 1.58, $\hat{g}_2$ = 7.08, 1.25 < μ < 1.26 (Box).

E2.12A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 0.997, $\tilde{x}_{0.5}$ = 0.599, $\bar{x}_R$ = 0.654, s = 1.410, $\hat{g}_1$ = 3.98, $\hat{g}_2$ = 21.40, 0.494 < μ < 0.720 (Box).

E2.12B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.923, $\tilde{x}_{0.5}$ = 0.612, $\bar{x}_R$ = 0.655, s = 1.080, $\hat{g}_1$ = 3.32, $\hat{g}_2$ = 14.42, 0.490 < μ < 0.780 (Box).

E2.13 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 6.85, $\tilde{x}_{0.5}$ = 6.95, $\bar{x}_R$ = 6.89, s = 0.51, $\hat{g}_1$ = -0.41, $\hat{g}_2$ = 2.38, 6.71 < μ < 7.07 (Box).

E2.14 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 93.1, $\tilde{x}_{0.5}$ = 80.0, $\bar{x}_R$ = 76.4, s = 81.5, $\hat{g}_1$ = 3.81, $\hat{g}_2$ = 22.22, 66.3 < μ < 85.2 (Box).

E2.15 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 2.700, $\tilde{x}_{0.5}$ = 0.014, $\bar{x}_R$ = 0.070, s = 15.200, $\hat{g}_1$ = 5.39, $\hat{g}_2$ = 30.03, < μ < (Box).

E2.16 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.481, $\tilde{x}_{0.5}$ = 0.485, $\bar{x}_R$ = 0.483, s = 0.070, $\hat{g}_1$ = -0.31, $\hat{g}_2$ = 5.79, 0.469 < μ < 0.496 (Box).

E2.17 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}_R$ = 0.1917, $\hat{g}_2$ = , 0.1570 < μ < 0.1990 (Box).

E2.18 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 26.5, $\tilde{x}_{0.5}$ = 26.6, s = 0.8, $\hat{g}_1$ = -0.20, $\hat{g}_2$ = 1.95.

E2.19 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.036, $\tilde{x}_{0.5}$ = 0.029, $\bar{x}_R$ = 0.029, s = 0.027, $\hat{g}_1$ = 1.18, $\hat{g}_2$ = 3.67, 0.020 < μ < 0.035 (Box).

E2.20 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 10.84, $\tilde{x}_{0.5}$ = 10.70, $\bar{x}_R$ = 10.76, s = 1.80, $\hat{g}_1$ = 0.25, $\hat{g}_2$ = 2.96, 10.34 < μ < 11.19 (Box).

E2.21 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 5.4, $\tilde{x}_{0.5}$ = 4.8, $\bar{x}_R$ = 4.9, s = 1.9, $\hat{g}_1$ = 1.66, $\hat{g}_2$ = 6.76, 4.7 < μ < 5.1 (Box).

E2.22 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 3.27, $\tilde{x}_{0.5}$ = 3.28, $\bar{x}_R$ = 3.28, s = 0.12, $\hat{g}_1$ = -0.18, $\hat{g}_2$ = 4.57, 3.25 < μ < 3.30 (Box).

E2.23 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.0311, $\tilde{x}_{0.5}$ = 0.0313, $\bar{x}_R$ = 0.0310, s = 0.0020, $\hat{g}_1$ = 0.24, $\hat{g}_2$ = 2.26, 0.0304 < μ < 0.0317 (Box).

E2.24 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.0227, $\tilde{x}_{0.5}$ = 0.0190, $\bar{x}_R$ = 0.0188, s = 0.0169, $\hat{g}_1$ = 3.55, $\hat{g}_2$ = 17.05, 0.0157 < μ < 0.0210 (Box).

E2.25 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 0.9720, $\tilde{x}_{0.5}$ = 0.9743, $\bar{x}_R$ = 0.9752, s = 0.0995, $\hat{g}_1$ = -0.21, $\hat{g}_2$ = 3.14, 0.9440 < μ < 1.0060 (Box).

E2.26 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 250.0, $\tilde{x}_{0.5}$ = 250.1, $\bar{x}_R$ = 250.1, s = 0.517, $\hat{g}_1$ = 0.05, $\hat{g}_2$ = 2.72, 249.9 < μ < 250.3 (Box).

2.6.4 Analýza hutnických a mineralogických dat

H2.01 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 0.55, $\tilde{x}_{0.5}$ = 0.54, $\bar{x}_R$ = 0.54, s = 0.04, $\hat{g}_1$ = 0.73, $\hat{g}_2$ = 2.96, 0.53 < μ < 0.56 (Box).

H2.02 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 0.18, $\tilde{x}_{0.5}$ = 0.18, $\bar{x}_R$ = 0.18, s = 0.02, $\hat{g}_1$ = 0.53, $\hat{g}_2$ = 2.68, 0.17 < μ < 0.19 (Box).

H2.03 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 21.8, $\tilde{x}_{0.5}$ = 21.8, $\bar{x}_R$ = 21.8, s = 0.5, $\hat{g}_1$ = -0.23, $\hat{g}_2$ = 1.67, 21.5 < μ < 22.0 (Box).

H2.04 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 137.0, $\tilde{x}_{0.5}$ = 135.1, $\bar{x}_R$ = 135.5, s = 4.5, $\hat{g}_1$ = 1.76, $\hat{g}_2$ = 4.55, 134.8 < μ < 135.9 (Box).

H2.05 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 37.9, $\tilde{x}_{0.5}$ = 37.8, $\bar{x}_R$ = 37.9, s = 0.6, $\hat{g}_1$ = -0.57, $\hat{g}_2$ = 3.72, 33.7 < μ < 38.1 (Box).

H2.06 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 1.35, $\tilde{x}_{0.5}$ = 1.32, $\bar{x}_R$ = 1.37, s = 0.35, $\hat{g}_1$ = -0.46, $\hat{g}_2$ = 4.94, 1.24 < μ < 1.49 (Box).

H2.07 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 0.119, $\tilde{x}_{0.5}$ = 0.121, $\bar{x}_R$ = 0.120, s = 0.005, $\hat{g}_1$ = -1.71, $\hat{g}_2$ = 6.43, 0.119 < μ < 0.121 (Box).

H2.08A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 1161.9, $\tilde{x}_{0.5}$ = 1202.0, $\bar{x}_R$ = 1174.3, s = 271.9, $\hat{g}_1$ = -0.26, $\hat{g}_2$ = 2.28, 1081.1 < μ < 1274.6 (Box).

H2.08B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 6, $\bar{x}$ = 2976.2, $\tilde{x}_{0.5}$ = 2924.5, $\bar{x}_R$ = 2663.3, s = 1694.0, $\hat{g}_1$ = 0.53, $\hat{g}_2$ = 2.07, 2102.1 < μ < 3188.8 (Box).

H2.08C Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 7, $\bar{x}$ = 3861.0, $\tilde{x}_{0.5}$ = 3941.0, $\bar{x}_R$ = 4020.3, s = 1181.9, $\hat{g}_1$ = -0.56, $\hat{g}_2$ = 2.06, 3739.6 < μ < 4481.9 (Box).

H2.08D Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 54.3, $\tilde{x}_{0.5}$ = 46.5, $\bar{x}_R$ = 49.6, s = 26.9, $\hat{g}_1$ = 0.89, $\hat{g}_2$ = 3.35, 41.4 < μ < 58.5 (Box).

H2.09A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 3.36, $\tilde{x}_{0.5}$ = 3.36, $\bar{x}_R$ = 3.37, s = 0.23, $\hat{g}_1$ = -0.48, $\hat{g}_2$ = 4.31, 3.31 < μ < 3.42 (Box).

H2.09B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 0.047, $\tilde{x}_{0.5}$ = 0.047, $\bar{x}_R$ = 0.046, s = 0.013, $\hat{g}_1$ = 0.42, $\hat{g}_2$ = 2.43, 0.043 < μ < 0.049 (Box).

H2.09C Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 3.38, $\tilde{x}_{0.5}$ = 3.38, $\bar{x}_R$ = 3.39, s = 0.22, $\hat{g}_1$ = -0.10, $\hat{g}_2$ = 4.76, 3.33 < μ < 3.44 (Box).

H2.09D Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 0.045, $\tilde{x}_{0.5}$ = 0.048, $\bar{x}_R$ = 0.046, s = 0.014, $\hat{g}_1$ = -0.17, $\hat{g}_2$ = 2.69, 0.043 < μ < 0.049 (Box).

H2.10 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.55, $\tilde{x}_{0.5}$ = 0.54,

$\bar{x}_R$ = 0.54, s = 0.04, $\hat{g}_1$ = 0.72, $\hat{g}_2$ = 2.96, 0.53 < μ < 0.56 (Box).

H2.11 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 0.009, $\tilde{x}_{0.5}$ = 0.008, $\bar{x}_R$ = 0.008, s = 0.003, $\hat{g}_1$ = 0.39, $\hat{g}_2$ = 2.30, 0.007 < μ < 0.009 (Box).

H2.12 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 3715.0, $\tilde{x}_{0.5}$ = 13.1, $\bar{x}_R$ = 13.1, s = 68.3, $\hat{g}_1$ = 3.68, $\hat{g}_2$ = 18.18, 11.0 < μ < 15.6 (Box).

H2.13A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 3.79, $\tilde{x}_{0.5}$ = 3.78, $\bar{x}_R$ = 3.79, s = 0.09, $\hat{g}_1$ = -0.17, $\hat{g}_2$ = 3.21, 3.78 < μ < 3.80 (Box).

H2.13B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 3.78, $\tilde{x}_{0.5}$ = 3.77, s = 0.077, $\hat{g}_1$ = -0.28, $\hat{g}_2$ = 6.50.

H2.14A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 0.015, $\tilde{x}_{0.5}$ = 0.015, $\bar{x}_R$ = 0.015, s = 0.0033, $\hat{g}_1$ = 2.93, $\hat{g}_2$ = 14.91, 0.0145 < μ < 0.0150 (Box).

H2.14B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\bar{x}$ = 0.013, $\tilde{x}_{0.5}$ = 0.013, $\bar{x}_R$ = 0.013, s = 0.004, $\hat{g}_1$ = -0.07, $\hat{g}_2$ = 8.59, 0.013 < μ < 0.014 (Box).

H2.15A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 8.581, $\tilde{x}_{0.5}$ = 8.566, $\bar{x}_R$ = 8.562, s = 0.091, $\hat{g}_1$ = 1.56, $\hat{g}_2$ = 6.34, 8.536 < μ < 8.589 (Box).

H2.15B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 4.031, $\tilde{x}_{0.5}$ = 4.034, $\bar{x}_R$ = 4.029, s = 0.028, $\hat{g}_1$ = 0.17, $\hat{g}_2$ = 2.04, 4.019 < μ < 4.040 (Box).

H2.16A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 0.0403, $\tilde{x}_{0.5}$ = 0.0401, $\bar{x}_R$ = 0.0401, s = 0.0011, $\hat{g}_1$ = 0.82, $\hat{g}_2$ = 3.63, 0.0398 < μ < 0.0405 (Box).

H2.16B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.6450, $\tilde{x}_{0.5}$ = 0.6460, $\bar{x}_R$ = 0.6450, s = 0.0143, $\hat{g}_1$ = 0.01, $\hat{g}_2$ = 2.70, 0.0640 < μ < 0.6500 (Box).

H2.17A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 3.90, $\tilde{x}_{0.5}$ = 3.88, $\bar{x}_R$ = 3.91, s = 0.21, $\hat{g}_1$ = -0.30, $\hat{g}_2$ = 3.62, 3.83 < μ < 3.98 (Box).

H2.17B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 73.76, $\tilde{x}_{0.5}$ = 73.50, $\bar{x}_R$ = 73.59, s = 4.41, $\hat{g}_1$ = 0.42, $\hat{g}_2$ = 4.42, 72.03 < μ < 75.17 (Box).

H2.17C Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 2.92, $\tilde{x}_{0.5}$ = 2.92, $\bar{x}_R$ = 2.91, s = 0.18, $\hat{g}_1$ = 0.41, $\hat{g}_2$ = 4.32, 2.85 < μ < 2.97 (Box).

H2.18A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 7, $\bar{x}$ = 623.4, $\tilde{x}_{0.5}$ = 610.5, $\bar{x}_R$ = 615.7, s = 37.7, $\hat{g}_1$ = 1.32, $\hat{g}_2$ = 4.30, 605.4 < μ < 628.0 (Box).

H2.18B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 6, $\bar{x}$ = 628.0, $\tilde{x}_{0.5}$ = 615.0, $\bar{x}_R$ = 619.8, s = 39.2, $\hat{g}_1$ = 1.23, $\hat{g}_2$ = 3.90, 608.9 < μ < 632.4 (Box).

H2.18C Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 6, $\bar{x}$ = 623.7, $\tilde{x}_{0.5}$ = 612.5, $\bar{x}_R$ = 616.2, s = 38.6, $\hat{g}_1$ = 1.00, $\hat{g}_2$ = 3.35, 604.7 < μ < 629.1 (Box).

H2.18D Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 626.0, $\tilde{x}_{0.5}$ = 617.0, $\bar{x}_R$ = 619.6, s = 37.7, $\hat{g}_1$ = 1.06, $\hat{g}_2$ = 3.94, 608.1 < μ < 632.6 (Box).

H2.19 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 53.82, $\tilde{x}_{0.5}$ = 53.78, $\bar{x}_R$ = 53.86, s = 0.98, $\hat{g}_1$ = -0.22, $\hat{g}_2$ = 3.60, 53.64 < μ < 54.07 (Box).

H2.20 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 41.1, $\tilde{x}_{0.5}$ = 42.2, $\bar{x}_R$ = 42.1, s = 7.9, $\hat{g}_1$ = -1.43, $\hat{g}_2$ = 6.46, 40.2 < μ < 44.3 (Box).

H2.21 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 6, $\bar{x}$ = 2.47, $\tilde{x}_{0.5}$ = 2.44, $\bar{x}_R$ = 2.46, s = 0.40, $\hat{g}_1$ = 0.70, $\hat{g}_2$ = 6.99, 2.34 < μ < 2.57 (Box).

H2.22 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 49.50, $\tilde{x}_{0.5}$ = 49.49, $\bar{x}_R$ = 49.50, s = 0.08, $\hat{g}_1$ = 0.49, $\hat{g}_2$ = 3.25, 49.48 < μ < 49.51 (Box).

H2.23A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 0.0146, $\tilde{x}_{0.5}$ = 0.0150, $\bar{x}_R$ = 0.0149, s = 0.0020, $\hat{g}_1$ = -0.97, $\hat{g}_2$ = 3.16, 0.0141 < μ < 0.0157 (Box).

H2.23B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 0.0144, $\tilde{x}_{0.5}$ = 0.0140, $\bar{x}_R$ = 0.0143, s = 0.0024, $\hat{g}_1$ = 0.11, $\hat{g}_2$ = 1.86, 0.0138 < μ < 0.0149 (Box).

H2.24 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 8, $\bar{x}$ = 3.989, $\tilde{x}_{0.5}$ = 3.710, $\bar{x}_R$ = 3.543, s = 2.060, $\hat{g}_1$ = 1.77, $\hat{g}_2$ = 5.70, 3.160 < μ < 4.000 (Box).

2.6.5 Analýza ekonomických a sociologických dat

S2.01 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 1001.2, $\tilde{x}_{0.5}$ = 1001.5, $\bar{x}_R$ = 1001.0, s = 4.8, $\hat{g}_1$ = 0.24, $\hat{g}_2$ = 2.72, 998.9 < μ < 1003.3 (Box).

S2.02 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 40.1, $\tilde{x}_{0.5}$ = 39.0, $\bar{x}_R$ = 39.8, s = 8.7, $\hat{g}_1$ = 0.12, $\hat{g}_2$ = 2.41, 38.2 < μ < 41.6 (Box).

S2.03, Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 4.0, $\tilde{x}_{0.5}$ = 3.0, $\bar{x}_R$ = 3.5, s = 2.3, $\hat{g}_1$ = 0.80, $\hat{g}_2$ = 2.75, 3.1 < μ < 3.9 (Box).

S2.03A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 49.1, $\tilde{x}_{0.5}$ = 48.8, $\bar{x}_R$ = 48.8, s = 2.1, $\hat{g}_1$ = 0.52, $\hat{g}_2$ = 2.16, 47.7 < μ < 50.1 (Box).

S2.03B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 50.0, $\tilde{x}_{0.5}$ = 50.1, $\bar{x}_R$ = 50.1, s = 0.7, $\hat{g}_1$ = -0.20, $\hat{g}_2$ = 2.29, 49.5 < μ < 50.6 (Box).

S2.04 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 109.9, $\tilde{x}_{0.5}$ = 109.0, $\bar{x}_R$ = 109.9, s = 16.6, $\hat{g}_1$ = -0.02, $\hat{g}_2$ = 2.93, 106.6 < μ < 113.3 (Box).

S2.05 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 3.2, $\tilde{x}_{0.5}$ = 3.0, $\bar{x}_R$ = 3.1, s = 1.7, $\hat{g}_1$ = 0.31, $\hat{g}_2$ = 2.77, 2.3 < μ < 3.9 (Box).

S2.06 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 130.4, $\tilde{x}_{0.5}$ = 103.0, $\bar{x}_R$ = 104.5, s = 104.9, $\hat{g}_1$ = 1.8, $\hat{g}_2$ = 6.43, 75.8 < μ < 134.2 (Box).

S2.07 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 7.5, $\tilde{x}_{0.5}$ = 7.0, $\bar{x}_R$ = 7.1, s = 4.3, $\hat{g}_1$ = 0.41, $\hat{g}_2$ = 2.63, 5.9 < μ < 8.3 (Box).

S2.08 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 232.7, $\tilde{x}_{0.5}$ = 229.0, $\bar{x}_R$ = 230.2, s = 18.3, $\hat{g}_1$ = 0.29, $\hat{g}_2$ = 1.72, 224.3 < μ < 236.1 (Box).

S2.09A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 8681.1, $\tilde{x}_{0.5}$ = 8555.0, $\bar{x}_R$ = 8655.6, s = 2980.2, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 2.37, 7570.9 < μ < 9796.3 (Box).

S2.09B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 3219.7, $\tilde{x}_{0.5}$ = 2926.5, $\bar{x}_R$ = 3072.6, s = 1216.3, $\hat{g}_1$ = 0.72, $\hat{g}_2$ = 3.30, 2664.3 < μ < 3516.8 (Box).

S2.09C Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 288.6, $\tilde{x}_{0.5}$ = 250.5, $\bar{x}_R$ = 258.0, s = 156.4, $\hat{g}_1$ = 0.88, $\hat{g}_2$ = 3.24, 209.0 < μ < 309.1 (Box).

S2.10A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 95.6, $\tilde{x}_{0.5}$ = 83.0, $\overline{x}_R$ = 88.2, s = 47.6, $\hat{g}_1$ = 0.23, $\hat{g}_2$ = 1.37, 60.7 < μ < 117.0 (Box).

S2.10B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\overline{x}$ = 95.0, $\tilde{x}_{0.5}$ = 93.5, s = 34.6, $\hat{g}_1$ = 0.97, $\hat{g}_2$ = 3.76.

S2.10C Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\overline{x}$ = 30.4, $\tilde{x}_{0.5}$ = 24.0, $\overline{x}_R$ = 26.0, s = 15.5, $\hat{g}_1$ = 1.27, $\hat{g}_2$ = 3.83, 19.9 < μ < 30.5 (Box).

S2.10D Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 30.8, $\tilde{x}_{0.5}$ = 29.5, $\overline{x}_R$ = 28.2, s = 14.6, $\hat{g}_1$ = 1.11, $\hat{g}_2$ = 4.18, 21.1 < μ < 37.4 (Box).

S2.10E Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 1956.0, $\tilde{x}_{0.5}$ = 1994.0, $\overline{x}_R$ = 1951.0, s = 328.0, $\hat{g}_1$ = 0.04, $\hat{g}_2$ = 2.59, 1746.2 < μ < 2162.9 (Box).

S2.10F Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 138.2, $\tilde{x}_{0.5}$ = 130.5, $\overline{x}_R$ = 138.3, s = 51.45, $\hat{g}_1$ = -0.09, $\hat{g}_2$ = 1.71, 106.8 < μ < 172.0 (Box).

S2.10G Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\overline{x}$ = 120.8, $\tilde{x}_{0.5}$ = 75.5, $\overline{x}_R$ = 85.2, s = 109.2, $\hat{g}_1$ = 1.53, $\hat{g}_2$ = 4.03, 54.8 < μ < 105.3 (Box).

S2.11 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 106.4, $\tilde{x}_{0.5}$ = 106.0, $\overline{x}_R$ = 106.7, s = 15.4, $\hat{g}_1$ = -0.11, $\hat{g}_2$ = 2.78, 105.8 < μ < 107.8 (Box).

S2.12 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: mnoho, $\overline{x}$ = 62.3, $\tilde{x}_{0.5}$ = 63.0, $\overline{x}_R$ = 63.1, s = 4.8, $\hat{g}_1$ = -0.90, $\hat{g}_2$ = 3.16, 63.1 < μ < 63.8 (Box).

S2.13A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 795.3, $\tilde{x}_{0.5}$ = 797.0, $\overline{x}_R$ = 795.2, s = 11.8, $\hat{g}_1$ = 0.04, $\hat{g}_2$ = 2.66, 789.8 < μ < 800.8 (Box).

S2.13B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 799.8, $\tilde{x}_{0.5}$ = 799.0, $\overline{x}_R$ = 799.8, s = 6.3, $\hat{g}_1$ = -0.02, $\hat{g}_2$ = 2.15, 796.9 < μ < 802.8 (Box).

S2.14 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 5.45, $\tilde{x}_{0.5}$ = 5.46, $\overline{x}_R$ = 5.46, s = 0.22, $\hat{g}_1$ = -0.44, $\hat{g}_2$ = 3.10, 5.38 < μ < 5.54 (Box).

S2.15 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\overline{x}$ = 4.31, $\tilde{x}_{0.5}$ = 4.30, $\overline{x}_R$ = 4.27, s = 0.22, $\hat{g}_1$ = 1.03, $\hat{g}_2$ = 4.07, 4.22 < μ < 4.33 (Box).

S2.16 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\overline{x}$ = 168.6, $\tilde{x}_{0.5}$ = 168.0, $\overline{x}_R$ = 168.2, s = 5.8, $\hat{g}_1$ = 0.50, $\hat{g}_2$ = 3.49, 166.6 < μ < 169.9 (Box).

S2.17 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 58.6, $\tilde{x}_{0.5}$ = 58.0, $\overline{x}_R$ = 57.7, s = 6.5, $\hat{g}_1$ = 1.03, $\hat{g}_2$ = 4.92, 56.0 < μ < 59.47 (Box).

S2.18A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 5.84, $\tilde{x}_{0.5}$ = 5.80, $\overline{x}_R$ = 5.78, s = 0.83, $\hat{g}_1$ = 0.31, $\hat{g}_2$ = 2.43, 5.65 < μ < 5.91 (Box).

S2.18B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 3.06, $\tilde{x}_{0.5}$ = 3.00, $\overline{x}_R$ = 3.04, s = 0.44, $\hat{g}_1$ = 0.32, $\hat{g}_2$ = 3.18, 2.97 < μ < 3.11 (Box).

S2.18C Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 3.76, $\tilde{x}_{0.5}$ = 4.35, $\overline{x}_R$ = 3.87, s = 1.77, $\hat{g}_1$ = -0.27, $\hat{g}_2$ = 1.60, 3.72 < μ < 4.27 (Box).

S2.18D Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 1.200, $\tilde{x}_{0.5}$ = 1.300, $\overline{x}_R$ = 1.201, s = 0.760, $\hat{g}_1$ = -0.10, $\hat{g}_2$ = 1.66, 1.114 < μ < 1.359 (Box).

S2.19A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 19.79,

$\tilde{x}_{0.5}$ = 19.75, $\overline{x}_R$ = 19.79, s = 1.04, $\hat{g}_1$ = -0.05, $\hat{g}_2$ = 1.94, 19.32 < μ < 20.30 (Box).

S2.19B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 99.85, $\tilde{x}_{0.5}$ = 100.00, $\overline{x}_R$ = 99.94, s = 4.02, $\hat{g}_1$ = -0.15, $\hat{g}_2$ = 2.88, 98.05 < μ < 101.81 (Box).

S2.20A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 14757.0, $\tilde{x}_{0.5}$ = 14928.0, $\overline{x}_R$ = 14737.5, s = 2787.2, $\hat{g}_1$ = -0.02, $\hat{g}_2$ = 2.20, 13716.6 < μ < 15836.6 (Box).

S2.20B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 5341.2, $\tilde{x}_{0.5}$ = 5355.0, $\overline{x}_R$ = 5322.8, s = 1201.8, $\hat{g}_1$ = 0.02, $\hat{g}_2$ = 2.35, 4878.3 < μ < 5792.5 (Box).

S2.21A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 28.6, $\tilde{x}_{0.5}$ = 28.8, $\overline{x}_R$ = 28.7, s = 7.6, $\hat{g}_1$ = -0.06, $\hat{g}_2$ = 2.61, 27.5 < μ < 29.9 (Box).

S2.21B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 86.5, $\tilde{x}_{0.5}$ = 85.0, $\overline{x}_R$ = 85.9, s = 27.2, $\hat{g}_1$ = 0.25, $\hat{g}_2$ = 3.68, 81.4 < μ < 90.0 (Box).

S2.21C Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 18.35, $\tilde{x}_{0.5}$ = 16.00, $\overline{x}_R$ = 18.15, s = 11.71, $\hat{g}_1$ = 4.92, $\hat{g}_2$ = 26.50, 16.43 < μ < 19.67 (Box).

S2.21D Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 2589.9, $\tilde{x}_{0.5}$ = 2595.0, $\overline{x}_R$ = 2648.9, s = 761.2, $\hat{g}_1$ = -0.80, $\hat{g}_2$ = 5.48, 2533.2 < μ < 2766.1 (Box).

2.7 Doporučená literatura

[1] Tukey J. W.: *Exploratory Data Analysis*. Addison Wesley, Reading 1977.
[2] Chambers J., Cleveland W., Kleiner W., Tukey P.: *Graphical Methods for Data Analysis*, Duxbury Press, Boston 1983.
[3] Hoaglin D. C., Mosteler F., Tukey J. W.: *Exploring Data Tables, Trends and Shapes*. J. Wiley, New York 1985.
[4] Scott D. W., Sheater S. J.: *Commun. Statist.* **14**, 1353 1985.
[5] Lejenne M., Dodge Y., Koelin E.: *Proc. Conf. COMSTAT 82*, Toulouse, (Vol. III). Str. 173.
[6] Hoaglin D. C., Mosteler F., Tukey J. W., *Edits: Understanding Robust and Exploratory Data Analysis*. J. Wiley, New York 1983.
[7] Kafander K., Spiegelman C. H.: *Comput. Stat. and Data Anal.* **4**, 167, 1986.
[8] Hines W. G. S., Hines R. J. H.: *Amer. Statist.* **41**, 21, 1987.
[9] Hoaglin D. C. a kol.: *J. Amer. Statist. Assoc.* **81**, 991, 1986.
[10] Royston J. P.: *Appl. Statist.* **31**, 115 (1982)
[11] Gan F. F., Koehler K. J., Thompson J. C.: *Amer. Statist.* **45**, No 1, 14 (1991)
[12] D'Agostino R. B., Belanger A., D'Agostino R. B. J.: *Amer. Statist.* **44**, No 4, 316 (1990)
[13] Potocký R., Kalas J., Komorník J. a Lamoš F.: *Zbierka úloh z pravdepodobnosti a matematickej štatistiky.* ALFA-SNTL, Bratislava 1986.
[14] Cyhelský L., Hustopecký J. a Závodský P.: *Příklady k základům statistiky.* SNTL, Praha 1988.
[15] Miller J. C., Miller J. N.: *Statistics for Analytical Chemistry.* Ellis Horwood, Chichester 1984.
[16] Anderson R. L.: *Practical Statistics for Analytical Chemists.* van Nostrand Reinhold Comp., New York 1987.
[17] Rice J. A.: *Mathematical Statistics and Data Analysis.* Wadsworth & Brooks, California, 1988.
[18] Dempír J.: *Geolog. průzkum* **26**, 247 (1981).
[19] Meloun M., Militký J.: *Statistické zpracování experimentálních dat.* PLUS, Praha 1994, EAST PUBLISHING, Praha 1998.

3

STATISTICKÁ ANALÝZA JEDNOROZMĚRNÝCH DAT

Účelem statistické analýzy jednorozměrných dat je charakterizace výběrového rozdělení, odhad jeho parametrů a v širším kontextu tvorba pravděpodobnostích modelů. Použití jednotlivých výběrových charakteristik je závislé na rozdělení základního souboru, ze kterého výběr pochází. Pro charakterizaci výběru se vyčíslují základní *momentové míry (charakteristiky).* Jsou-li v datech předpokládány i odlehlé hodnoty, užívají se *robustní odhady*, zejména *kvantilové odhady*. Pokud bylo nalezeno jiné než předpokládané normální rozdělení, používají se *maximálně věrohodné odhady* jeho parametrů.

3.1 Bodový odhad parametrů polohy, rozptýlení a tvaru

A. Momentové míry polohy a rozptýlení

Míry polohy. Momentové míry polohy zahrnují různé druhy průměrů, pomocí kterých můžeme charakterizovat centrální tendenci dat. Momentové míry polohy jsou jednoduché číselné charakteristiky, které se vyčíslují ze všech prvků výběru.

Základní momentovou charakteristikou polohy je *aritmetický průměr* $\bar{x}$, který je zároveň maximálně věrohodným odhadem střední hodnoty $E(x)$ normálního rozdělení. Představuje první obecný statistický moment m_1. Z n prvků výběru $x_1, x_2, ..., x_n$, se vypočte aritmetický průměr $\bar{x}$ dle vzorce

$$\bar{x} = \frac{x_1 + x_2 + \dots x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} .$$

Tento odhad má rozptyl $D(\bar{x}) = \sigma^2/n$, kde σ^2 je rozptyl souboru, ze kterého výběr pochází. Má-li každý prvek x_i normální rozdělení s rozptylem $\sigma^2(x_i)$, lze pro odhad střední hodnoty odvodit vztah

$$\hat{x}_w = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma^2(x_i)}}{\sum_{i=1}^{n} \frac{1}{\sigma^2(x_i)}},$$

který se nazývá *vážený aritmetický průměr* s vahami $1/\sigma^2(x_i)$. Rozptyl tohoto odhadu má tvar $D(\bar{x}_w) = 1/\sum_{i=1}^{n} 1/\sigma^2(x_i)$. Obě rovnice jsou použitelné při znalosti rozptylů $\sigma^2(x_i)$ nebo jim odpovídajících "vah" pro jednotlivé prvky výběru.

Míry rozptýlení (variability). Momentové charakteristiky rozptýlení slouží jako odhad variability základního souboru. Míry rozptýlení, které charakterizují proměnlivost výběru v absolutní velikosti, tj. ve stejných jednotkách jako sledovaný prvek, nazýváme *mírami absolutního rozptýlení (variability)*. Když však srovnáváme rozptýlení výběrů lišících se svojí úrovní, užíváme *míry relativního rozptýlení (variability)*. Jsou to buď bezrozměrná čísla, nebo čísla vyjádřená v procentech.

Důležité jsou takové míry rozptýlení, jejichž velikost je závislá na velikostech všech prvků výběru. Míra rozptýlení, která měří současně rozptýlení všech prvků kolem střední hodnoty, se nazývá *rozptyl* σ^2. Je definován jako druhý centrální statistický moment m_2. Pro odhad rozptylu platí vztah

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 .$$

Rozptyl tohoto odhadu je $D(\hat{\sigma}^2) = 2\sigma^4/n$. V praktických situacích není parametr střední hodnoty μ znám a nahrazuje se aritmetickým průměrem $\mu = \bar{x}$. Takto definovaný rozptyl $\hat{\sigma}^2$ však představuje vychýlený odhad, protože $E(\hat{\sigma}^2) = K\ \sigma^2$, kde $K = (n - 1)/n$. Jako nevychýlený odhad se užívá *výběrový rozptyl*

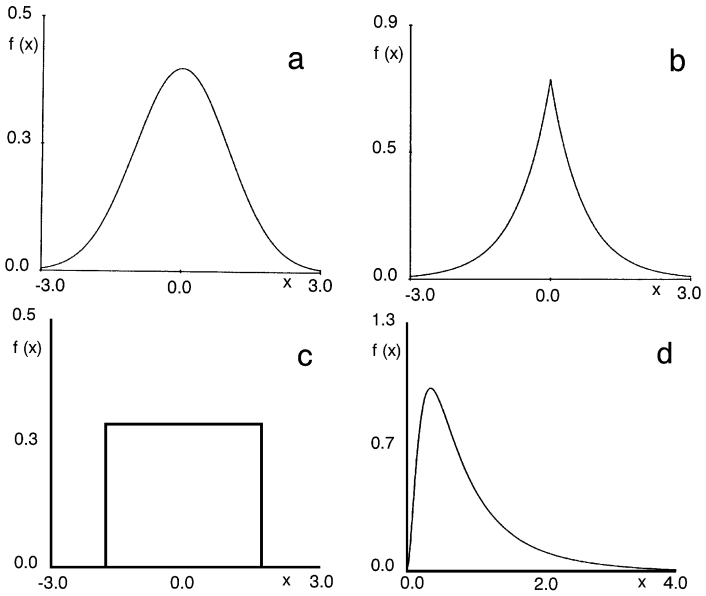
$$s^2 = \frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Z praktického hlediska je určitou nevýhodou, že výběrový rozptyl je vyjádřen ve čtvercích užité jednotky. Proto se za míru rozptýlení volí obvykle druhá odmocnina z rozptylu, označovaná jako *směrodatná odchylka* $s = \sqrt{s^2}$. Její výhodou je to, že je uvedena ve stejných jednotkách jako zkoumaný výběr.

Pro charakterizaci relativního rozptýlení dat se užívá míry relativního rozptýlení, nazvané *variační koeficient* $\delta = \sigma/\mu$ nebo-li *relativní směrodatná odchylka* (často vyjádřená v procentech) se svým rozptylem

$$D(\delta) \doteq \delta^2 (n + \delta^2 (2n + 1))/(2n(n-1)).$$

Jeho odhad $\hat{\delta}$ je roven $\hat{\delta} = s/\bar{x}$.



Obr. 3.1 Znázornění vybraných rozdělení hustot pravděpodobnosti: (a) normované normální N(0, 1), (b) standardizované Laplaceovo L(0, 1), (c) standardizované rovnoměrné R(0, 1), (d) logaritmicko-normální LN(1, 1).

Míry tvaru. Momentové charakteristiky tvaru poskytují informace o tvaru rozdělení. Užívá se *šikmost (asymetrie)* g_1, čili třetí normovaný centrální moment, a *špičatost (excess)* g_2 čili čtvrtý normovaný centrální moment. Pro normální rozdělení platí hodnoty $g_1 = 0$ a $g_2 = 3$, pro rovnoměrné $g_1 = 0$ a $g_2 = 1.8$, pro Laplaceovo $g_1 = 0$ a $g_2 = 6$ a pro exponenciální $g_1 = 2$ a $g_2 = 9$.

Momentový odhad *šikmosti* g_1 je vyjádřen vztahem

$$\hat{g}_1 = \frac{\sqrt{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{3/2}} .$$

Střední hodnota pro výběry normálního rozdělení je $E(\hat{g}_1) = 0$. Pro asymptotický rozptyl tohoto odhadu platí $D(\hat{g}_1) = (n - 2)/[(n + 1)(n + 3)]$.

Momentový odhad *špičatosti* g_2 je vyjádřen vztahem

$$\hat{g}_2 = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} .$$

Střední hodnota tohoto odhadu pro výběry z normálního rozdělení je

$$E(\hat{g}_2) \;'\; 3 \;\&\; 6/(n \;\%\; 1).$$

Pro asymptotický rozptyl tohoto odhadu platí vztah

$$D(\hat{g}_2) \quad . \quad 24\; n\; (n \;\&\; 2)\; (n \;\&\; 3)/[(n \;\%\; 1)^2\; (n \;\%\; 3)\; (n \;\%\; 5)].$$

Při stanovení libovolného bodového odhadu parametru je třeba určit vždy i jeho rozptyl. K docílení stejné "přesnosti" odhadů, vyjádřené jeho rozptylem, je třeba při užití méně efektivního odhadu provést větší počet měření n. Například u dat pocházejících z normálního rozdělení se musí při použití mediánu provést 1.6krát více měření než při použití aritmetického průměru, aby se docílilo stejné přesnosti odhadu. Naopak, u dat pocházejících z Laplaceova rozdělení, se k odhadu parametru polohy pomocí aritmetického průměru $\bar{x}$ musí použít k dosažení stejné přesnosti jako u mediánu dvojnásobný počet měření.

B. Kvantilové a robustní míry polohy a rozptýlení

Kvantilové a robustní charakteristiky jsou méně citlivé na odlehlé hodnoty než momentové. Patří sem:

Modus, $\hat{x}_M$, který je definován jako lokální maximum na hustotě pravděpodobnosti. V praxi se vyskytují většinou rozdělení unimodální, jejichž hustota pravděpodobnosti má pouze jedno maximum. Modus je vždy robustní, není citlivý na odlehlá měření.

Kvantily (kvartily, decily, percentily). *Výběrový α-kvantil* je hodnota, která rozděluje výběr prvků na dvě části, jedna obsahuje α % prvků, které jsou menší (nebo stejné) než tento kvantil, druhá část (1-α) % prvků, které jsou větší (nebo stejné) než kvantil. V případě kvartilů jde o kvantily, které dělí uspořádané prvky ve výběru na čtyři části, přičemž každá část obsahuje 25 % prvků. Kvartily jsou celkem tři: *dolní kvartil* $\tilde{x}_{0.25}$ odděluje čtvrtinu nejmenších prvků. Prostřední kvartil se jmenuje *medián* $\tilde{x}_{0.5}$ a rozděluje výběr prvků na dvě stejné části, z nichž každá obsahuje 50 % prvků. Jsou-li prvky výběru setříděny podle velikosti vzestupně $x_{(1)} \# x_{(2)} \# \ldots \# x_{(n)}$ (pořádkové statistiky), je medián pro n liché roven $\tilde{x}_{0.5} = x_{(k)}$, kde $k = (n + 1)/2$ a pro n sudé roven $\tilde{x}_{0.5} = [x_{(k)} + x_{(k+1)}]/2$, kde $k = n/2$. Medián patří mezi kvantilové odhady, které jsou robustní, tj. necitlivé na odlehlé hodnoty. Medián je maximálně věrohodným odhadem polohy u Laplaceova (oboustranného exponenciálního) rozdělení a má pro toto rozdělení minimální rozptyl $D_L(\tilde{x}_{0.5}) \;'\; \sigma^2/2\; n$. Pro normální rozdělení však již medián nemá nejmenší rozptyl. Konečně třetím kvartilem je *horní kvartil* $\tilde{x}_{0.75}$, který odděluje 75 % menších prvků od zbývajících 25 % největších.

Obdobně jsou definovány *decily* $\tilde{x}_{10}$, $\tilde{x}_{20}$, ..., $\tilde{x}_{90}$, které dělí výběr na 10 stejně obsazených částí, ve kterých je stejná relativní četnost, a *centily* $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_{99}$, které dělí výběr na 100 stejně obsazených částí.

Z kvantilových odhadů rozptýlení se používá *interkvartilové rozpětí*

$$R \;'\; (\tilde{x}_{0.75} \;\&\; \tilde{x}_{0.25}),$$

kde $\tilde{x}_{0.75}$ je odhad horního kvartilu a $\tilde{x}_{0.25}$ odhad dolního kvartilu. Využitím R lze odhadnout směrodatnou odchylku σ_R podle vztahu $s_R = 0.7413\; R$.

Jedním z nejefektivnějších robustních, a přitom jednoduchých odhadů parametru polohy, je *uřezaný průměr* $\bar{x}(\mathrm{h})$, využívající lineární kombinace pořádkových statistik

$$\bar{x}(\alpha) \;=\; \frac{1}{n \,-\, 2M} \sum_{i=M+1}^{n-M} x_{(i)} \;,$$

kde M = int (α n / 100). Parametr α určuje procento "uřezaných" pořádkových statistik na každém konci, nejvyšších a nejnižších. Optimální hodnota bývá 10 %. Vzniká tak 10 % uřezaný průměr $\bar{x}(10)$. V případě očekávaného většího počtu odlehlých měření se uřezává až na hodnotu α = 25 %. Uřezaný průměr se užívá s odhadem směrodatné odchylky, určené z winsorizovaného součtu čtverců odchylek

$$S_w(\alpha) \;=\; \sum_{i=M+2}^{n-M-1} (x_{(i)} - \bar{x}_w(\alpha))^2 + (M+1)\,[(x_{(M+1)} - \bar{x}_w(\alpha))^2 + (x_{(n-M)} - \bar{x}_w(\alpha))^2] \;,$$

kde $\bar{x}_w(\alpha)$ je *winsorizovaný průměr,* pro který platí definiční vztah

$$\bar{x}_w(\alpha) \;=\; \frac{1}{n}\left[(M+1)\,(x_{(M+1)} + x_{(n-M)}) + \sum_{i=M+2}^{n-M-1} x_{(i)}\right] \;.$$

Pro nesymetrická, značně zešikmená rozdělení je doporučen *nesymetrický uřezaný průměr* $\bar{x}(\alpha_1, \alpha_2)$, pro který platí

$$\bar{x}(\alpha_1, \alpha_2) \;=\; \frac{\sum_{i=n_1}^{n_2} x_i}{n_2 - n_1 + 1} \;,$$

kde n_1 = int(α_1 n / 100) a $n_2 = n$ - int(α_2 n / 100). Pokud jsou hodnoty α_1 a α_2 zvoleny tak, že rozdělení uřezaného výběru je již symetrické, lze určit *rozptyl nesymetricky uřezaného* průměru $\bar{x}(\alpha_1, \alpha_2)$ vztahem

$$s_n^2 \;=\; \frac{1}{h\,(h+1)}\,[n_1\,(x_{(n_1)} - \bar{x}(\alpha_1,\alpha_2))^2 + \sum_{i=n_1+1}^{n_2-1} (x_{(i)} - \bar{x}(\alpha_1,\alpha_2))^2 +$$

$$+ (n - n_2 + 1)\,(x_{(n_2)} - \bar{x}(\alpha_1,\alpha_2))^2 + ((n_1 - 1)\,(x_{(n_1)} -$$

$$- \bar{x}(\alpha_1,\alpha_2))^2 + (n - n_2)\,(x_{(n_2)} - \bar{x}(\alpha_1,\alpha_2))^2],$$

kde $h = n_2 - n_1 + 1$.

M-odhady jsou maximálně věrohodné odhady parametrů pro speciálně vybraná rozdělení. Maximalizace věrohodnostní funkce podle parametru μ_M zde vede k minimalizaci funkce $\sum_{i=1}^{n} \kappa\left(\frac{x_i - \mu_M}{\sigma}\right) \to \min$. Tvar funkce $\kappa(u)$ určuje vlastnost odhadu. Derivací tohoto vztahu a úpravou vyjde výraz pro *M-odhad střední hodnoty*

$$\hat{\mu}_M = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

kde *statistická váha* je

$$w_i = W\left(\frac{x_i - \mu_M}{\sigma}\right) \quad \text{a} \quad W(u) = \frac{d\rho(u)}{du}.$$

Mezi *M*-odhady patří však i medián a aritmetický průměr. Pro *robustní M*-odhady platí, že *váhová funkce W(u)* musí být ohraničená. Protože statistická váha w_i je funkcí μ_M, musí se výpočet provést iterativně a počáteční hodnotou může být aritmetický průměr. Mezi doporučované váhové funkce *W(u)* patří bikvadratická funkce typu

$$W(u) = \begin{cases} (1 - \left(\frac{u}{4.69}\right)^2)^2 & \text{pro } |u| < 4.69 \\ 0 & \text{pro } |u| \geq 4.69 \end{cases},$$

kde konstanta 4.69 zajišťuje, že pro normálně rozdělená data bude asymptotická efektivnost odhadu μ_M rovna 0.95. Doporučuje se použít i robustní *M-odhad směrodatné odchylky* dle výrazu

$$s_M = \sqrt{\left(\sum_{i=1}^{n} V_i [x_i - \hat{\mu}_M]^2\right) / \sum_{i=1}^{n} V_i},$$

kde

$$V_i = W\left(\sqrt{\Delta\left(\frac{x_i - \hat{\mu}_M}{s_M}\right)}\right),$$

v němž $\Delta(u)$ je *odchylková funkce*, pro kterou platí

$$\Delta(u) = \begin{cases} u^2 - \ln(u^2) - 1 & \text{pro } u \neq 0 \\ 4 & \text{pro } u = 0 \end{cases}.$$

Protože robustní *M*-odhad $\hat{\mu}_M$ představuje vlastně vážený aritmetický průměr, je jeho rozptyl vyjádřen vztahem $D(\hat{\mu}_M) = s_M^2 / \sum_{i=1}^{n} w_i$.

Výstup analýzy jednorozměrného výběru *norm*(10, 0.005) u programu ADSTAT ukazuje přehled obvykle vyčíslovaných odhadů parametrů polohy, rozptýlení a tvaru.

Analýza jednorozměrného výběru *norm* (ADSTAT)

(1) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$	10.012
Odhad rozptylu s^2	5.2232E-03
Odhad směrodatné odchylky s	0.0723
Odhad šikmosti $\hat{g}_1$	-0.04
Odhad špičatosti g_2	3.08
Dolní mez 95.0% intervalu spolehlivosti L_D	9.998
Horní mez 95.0% intervalu spolehlivosti L_H	10.026
(2) Odhady ostatních parametrů:	
Odhad modu x_M	10.000
Odhad polosumy x_P	10.011
(3) Robustní odhady parametrů:	
Medián $x_{0.5}$	10.011
Odhad směrodatné odchylky mediánu $s(\tilde{x}_{0.5})$	0.0780
Dolní mez 95.0% intervalu spolehlivosti L_D	9.991
Horní mez 95.0% intervalu spolehlivosti L_H	10.031
Odhad 10% uřezaného průměru $\bar{x}$(10%)	10.013
Odhad směrodatné odchylky s(10 %)	0.0719
Odhad winsorizovaného průměru $\bar{x}_w$(10%)	10.013
Odhad směrodatné odchylky s_w(10%)	0.0648
Dolní mez 95.0% intervalu spolehlivosti L_D	9.998
Horní mez 95.0% intervalu spolehlivosti L_H	10.027
Odhad 40% uřezaného průměru $\bar{x}$(40%)	10.011
Odhad směrodatné odchylky s(40%)	0.0726
Odhad winsorizovaného průměru $\bar{x}_w$(40%)	10.013
Odhad směrodatné odchylky s_w(40%)	0.0316
Dolní mez 95.0% intervalu spolehlivosti L_D	9.996
Horní mez 95.0% intervalu spolehlivosti L_H	10.026
Odhad *M*-odhadu střední hodnoty $\hat{\mu}_M$	10.012
Odhad směrodatné odchylky s_M	0.0719
Dolní mez 95.0% intervalu spolehlivosti L_D	9.997
Horní mez 95.0% intervalu spolehlivosti L_H	10.027
(4) Hoggovy adaptivní odhady parametrů:	
Hoggův průměr μ_M	10.012
Odhad směrodatné odchylky s_M	0.0716
Dolní mez 95.0% intervalu spolehlivosti L_D	9.998
Horní mez 95.0% intervalu spolehlivosti L_H	10.027

C. Odhady parametrů polohy a rozptýlení pro důležitá rozdělení

Odhady parametrů polohy a rozptýlení pro často se vyskytující rozdělení dat v laboratoři se vyčíslují podle následujících vztahů:

a) **Laplaceovo rozdělení.** Laplaceovo (oboustranné exponenciální) rozdělení se vyskytuje v případech, kdy jsou náhodné veličiny měřeny za podmínek kolísání rozptylu kolem určité střední hodnoty. Hustota pravděpodobnosti spojité náhodné veličiny x ležící v intervalu (- 4, 4) s Laplaceovým rozdělením má tvar

$$f(x) \quad ' \quad 0.5\,\Phi^{\&1} \exp\left(\& \frac{*x \ \& \ \Theta*}{\Phi} \right).$$

Střední hodnota Laplaceova rozdělení je $E(x) = \Theta$, *rozptyl* $D(x) = 2\,\Phi^2$, *šikmost* $g_1 = 0$ a *špičatost* $g_2 = 6$. Ve srovnání s normálním rozdělením je Laplaceovo rozdělení špičatější a má delší konce. Tak 1% kvantil Laplaceova rozdělení je roven $E(x)$ - 2.72 $\sqrt{D(x)}$, zatímco odpovídající kvantil normálního rozdělení je $E(x)$ - 2.33 $\sqrt{D(x)}$. Laplaceovo rozdělení připouští výskyt výrazněji odchýlených hodnot a využívá se jako "robustní" alternativa normálního rozdělení. Po dosazení a zlogaritmování vyjde logaritmus věrohodnostní funkce ve tvaru

$$\ln L \quad ' \quad \& \ n \ \ln(2\,\Phi) \ \& \ \Phi^{\&1} \ \mathop{\mathrm{j}}_{i}^{n}{}_{1} \ *x_i \ \& \ \Theta*.$$

Při známé hodnotě parametru Φ lze maximálně věrohodný odhad $\hat{\Theta}$ parametru Θ získat minimalizací výrazu $\mathop{\mathrm{j}}_{i}^{n}{}_{1} \ * \mathrm{x_i} - \Theta*$.

Maximálně věrohodný odhad θ je výběrový medián $\tilde{\theta} \ ' \ \tilde{x}_{0.5}$. Odhad parametru Φ se počítá dle rovnice

$$\hat{\Phi} \quad ' \quad \frac{1}{n} \ \mathop{\mathrm{j}}_{i}^{n}{}_{1} \ *x_i \ \& \ \Theta* \quad .$$

Rozptyl se vyjádří vztahem $D(\hat{\Phi}) \ ' \ \frac{\Phi^2}{n}$. *Interval spolehlivosti parametru* Φ lze konstruovat tak, že je-li známa střední hodnota Θ, lze určit 100(1 - α)% interval spolehlivosti pro Φ podle vztahu

$$\frac{2\,n\,\hat{\Phi}}{\chi^2_{1\&\alpha/2}(2\,n)} \quad \# \quad \Phi \quad \# \quad \frac{2\,n\,\hat{\Phi}}{\chi^2_{\alpha/2}(2\,n)}.$$

b) **Rovnoměrné rozdělení.** Rovnoměrné (rektangulární) rozdělení je nejjednodušším typem rozdělení pro oboustranně omezenou náhodnou veličinu, která musí ležet v zadaném intervalu $a - h \# x \# a + h$. Týká se náhodných veličin, které se v daném intervalu vyskytují se stejnou pravděpodobností. Pokud je $a = 0$ a $h = 0.5 \cdot 10^{-k}$, popisuje rovnoměrné rozdělení chyby vzniklé zaokrouhlením na k desetinných míst. Hustota pravděpodobnosti

rovnoměrného rozdělení má tvar

$$f(x) = \frac{1}{2h}, \quad a - h \le x \le a + h .$$

Střední hodnota rovnoměrného rozdělení je $E(x) = a$, *rozptyl* $D(x) = h^2/3$, *šikmost* $g_1 = 0$ a *špičatost* $g_2 = 1.8$. Logaritmus věrohodnostní funkce má tvar

$$\ln L = -n \ln(2h)$$

pro $a - h \le \min(x_1, ..., x_n) \le \max(x_1, ..., x_n) \le a + h$. Tento vztah nabývá maxima při minimální velikosti h. Je zřejmé, že $\min(x_1, ..., x_n) = x_{(1)}$ a $\max(x_1, ..., x_n) = x_{(n)}$. Maximálně věrohodný *odhad parametru rozptýlení* $\hat{h}$ je roven

$$\hat{h} = 0.5 (x_{(n)} - x_{(1)})$$

a maximálně věrohodný *odhad parametru polohy a* je roven

$$\hat{a} = 0.5 (x_{(n)} + x_{(1)}) .$$

Odhad $\hat{a}$ je totožný s polosumou $\tilde{x}_P$. Odhad $\hat{h}$ je vychýlený. *Nevychýlený odhad* $\hat{h}_0$ se získá násobením odhadu $\hat{h}$ faktorem $(n + 1)/(n - 1)$. Pro rozptyly těchto odhadů platí vztahy

$$D(\hat{h}) = \frac{2h^2}{(n-1)(n+2)}$$

a

$$D(\hat{a}) = \frac{2h^2}{(n+1)(n+2)} .$$

Pro 100(1 - α)% interval spolehlivosti libovolného parametru Θ lze užít asymptotický vztah

$$\hat{\Theta} - u_{1-\alpha/2} \sqrt{D(\hat{\Theta})} \le \Theta \le \hat{\Theta} + u_{1-\alpha/2} \sqrt{D(\hat{\Theta})} .$$

c) **Jednoparametrové exponenciální rozdělení.** Exponenciální rozdělení je jednostranně ohraničené zdola. Využívá se ho k popisu řady reálných dějů. Exponenciální rozdělení má uplynulý čas, resp. obsazený prostor před tím, než nastal náhodný jev. Je typické pro životnost součástí strojů, vzdálenost, kterou urazí molekuly plynu při nízkém tlaku až do vzájemné srážky, doby mezi dopadem částic do čítače a doby bezporuchové činnosti. Exponenciální rozdělení bývá spjato s Poissonovým rozdělením náhodných jevů. Popisuje statistické chování kladné náhodné veličiny pro $x \ge 0$. Jeho hustota pravděpodobnosti je definována vztahem

$$f(x) = \Theta^{-1} \exp\left(-\frac{x}{\Theta}\right) .$$

Střední hodnota jednoparametrového exponenciálního rozdělení je $E(x) = \Theta$, rozptyl $D(x) = \Theta^2$, šikmost $g_1 = 2$ a *špičatost* $g_2 = 9$. *Medián* je roven $\tilde{x}_{0.5} = \Theta \ln 2$. Logaritmus

věrohodnostní funkce má tvar

$$\ln L = -n \ln \Theta - \sum_{i=1}^{n} \frac{x_i}{\Theta} .$$

Po dosazení se určí maximálně věrohodný odhad $\Theta = \frac{\sum_{i=1}^{n} x_i}{n}$ a vyčíslí se odpovídající rozptyl $D(\hat{\Theta}) = \frac{\Theta^2}{n}$. Při konstrukci intervalů spolehlivosti se využívá skutečnosti, že náhodná veličina $2\hat{\Theta}n/\Theta$ má rozdělení $\chi^2(2n)$. Pro 100 *(*1 - *α)%ní interval spolehlivosti* pak platí

$$\frac{2n\hat{\Theta}}{\chi^2_{1-\alpha/2}(2n)} \le \Theta \le \frac{2n\hat{\Theta}}{\chi^2_{\alpha/2}(2n)}.$$

d) **Exponenciální rozdělení dvouparametrové.** Popisuje statistické chování náhodné veličiny, která může nabývat hodnot $x \ge \mu$, tj. je zdola ohraničená. Hustota pravděpodobnosti má tvar

$$f(x) = \Theta^{-1} \exp[(-x + \mu)/\Theta].$$

Střední hodnota dvouparametrového exponenciálního rozdělení je $E(x) = \mu + \Theta$. Vztahy pro rozptyl, šikmost a špičatost jsou stejné jako u jednoparametrového exponenciálního rozdělení. *Odhad* $\hat{\mu}$ je

$$\hat{\mu} = x_{(1)} = \min(x_1, ..., x_n).$$

Pro maximálně věrohodný odhad $\hat{\Theta}$ parametru Θ lze napsat

$$\hat{\Theta} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}) = \bar{x} - x_{(1)} .$$

Odhad $\hat{\mu}$ má střední hodnotu $E(\hat{\mu}) = \mu + \frac{\Theta}{n}$ a *rozptyl* $D(\hat{\mu}) = \Theta^2/n^2$.

Odhad $\hat{\Theta}$ má střední hodnotu $E(\hat{\Theta}) = \Theta\left(1 - \frac{1}{n}\right)$ a *rozptyl*

$$D(\hat{\Theta}) = \Theta^2 \left(\frac{1}{n} + \frac{1}{n^2} - \frac{2}{n^3}\right) .$$

Maximálně věrohodné odhady $\hat{\Theta}$ a $\hat{\mu}$ jsou vychýlené. Pro *nevychýlené odhady* $\hat{\Theta}_0$ a $\hat{\mu}_0$ lze odvodit vztahy

$$\hat{\mu}_0 = \frac{n x_{(1)} - \bar{x}}{n - 1} , \quad D(\hat{\mu}_0) = \frac{\Theta^2}{n(n - 1)} ,$$

$$\hat{\Theta}_0 \;=\; \frac{n\,(\bar{x} - x_{(1)})}{n - 1} \;, \quad D(\hat{\Theta}_0) \;=\; \frac{\Theta^2}{n - 1} \;.$$

Odhady $\hat{\mu}_0$, Θ_0 jsou však korelované s korelačním koeficientem rovným $(-1/\sqrt{n})$.

Pro 100 *(1 - α)% oboustranný interval spolehlivosti parametru* Θ platí

$$\frac{2\,(n-1)\,\hat{\Theta}_0}{\chi^2_{1-\alpha/2}(2n-2)} \;\le\; \Theta \;\le\; \frac{2\,(n-1)\,\hat{\Theta}_0}{\chi^2_{\alpha/2}(2n-2)} \;.$$

Protože má podíl $n\,(x_{(1)} - \mu) \,/\, \hat{\Theta}_0$ rozdělení F se 2 a $(2n - 2)$ stupni volnosti, je spodní mez μ_1 pro $100(1 - \alpha)\%$ interval spolehlivosti parametru μ vyjádřitelná vztahem

$$\mu_1 \;=\; x_{(1)} - \frac{\hat{\Theta}_0 \; F_{1-\alpha}(2,\; 2n - 2)}{n} \;.$$

Horní mezí je s pravděpodobností blízkou jedné nejmenší prvek výběru $x_{(1)}$. Pro určení kvantilů rozdělení $F(2, 2n - 2)$ stačí dosadit do vztahu

$$F_P(2,\; 2n - 2) \;=\; (n - 1)\,\left[(1 - P)^{-\frac{1}{n-1}} - 1\right] \;.$$

e) **Logaritmicko-normální rozdělení dvouparametrové**. Logaritmicko-normální rozdělení je nejrozšířenější alternativou rozdělení normálního pro jednostranně ohraničená data. Fyzikální veličiny (teplota, tlak, objem, hmotnost, koncentrace atd.) jsou buď kladné, nebo mají přirozeně definovaný počátek (např. absolutní nula u teploty). Pokud jsou však naměřené hodnoty v blízkosti počátku, je lépe použít např. logaritmicko-normální rozdělení. Toto rozdělení se používá všude tam, kde se měří nízké koncentrace, malé hmotnosti, malé délky atd. Typickým příkladem je v analytické chemii stopová analýza. Rovněž souhrnná chyba, která je součinem dílčích malých chyb, má logaritmicko-normální rozdělení. Náhodná veličina x s dvouparametrovým logaritmicko-normálním rozdělením souvisí s náhodnou veličinou u s normovaným normálním rozdělením vztahem

$$u \;=\; \frac{\ln(x - \mu)}{\sigma} \;,$$

kde μ, σ jsou parametry. Logaritmicko-normální rozdělení má náhodná veličina, která může nabývat pouze kladných hodnot, tj. leží v intervalu $0 < x < 4$. Využitím vztahů pro hustotu pravděpodobnosti transformované náhodné veličiny lze odvodit hustotu pravděpodobnosti logaritmicko-normálního rozdělení ve tvaru

$$f(x) \;=\; \frac{1}{(x - \Theta)\,\sigma\sqrt{2\pi}} \; \exp\left[-\frac{(\ln(x - \Theta) - \mu)^2}{2\,\sigma^2}\right] .$$

Náhodná veličina x má dvouparametrové logaritmicko-normální rozdělení, pokud má náhodná veličina ln x normální rozdělení $N(\mu, \sigma^2)$. *Střední hodnota* a *rozptyl* náhodné veličiny x se vyčíslí podle vztahů

$$E(x) = \exp(\mu + 0.5\sigma^2) \quad a \quad D(x) = \exp(2\mu)\,\omega\,(\omega - 1) \ ,$$

kde $\omega = \exp(\sigma^2)$. *Šikmost* g_1 a *špičatost* g_2 tohoto rozdělení závisí pouze na veličině ω podle rovnic

$$g_1 = \sqrt{\omega - 1}\,(\omega + 2) \text{ a } g_2 = \omega^4 + 2\omega^3 + 3\omega^2 - 3 .$$

Také *variační koeficient* δ je pro logaritmicko-normální rozdělení funkcí pouze parametru ω a platí $\delta = \sqrt{\omega - 1}$. *Modus* $\hat{x}_M$ a *medián med* lze vyjádřit vztahy $\hat{x}_M = \exp(\mu - \sigma^2)$ a $med = \exp(\mu)$. Maximálně věrohodný *odhad parametru polohy* $\boldsymbol{\mu}$ se určí vztahem

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i$$

a maximálně věrohodný *odhad parametru rozptýlení* σ^2 se vypočte vztahem

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\ln x_i - \hat{\mu})^2 \ ,$$

který je však vychýlený. *Nevychýlený odhad* $\hat{\sigma}_0^2$ se vyčíslí analogicky jako u normálního rozdělení $\hat{\sigma}_0^2 = (n/(n-1))\,\hat{\sigma}^2$.

V řadě případů je analýza v logaritmické transformaci nevyhovující. Je třeba stanovit odhady parametrů polohy a rozptýlení spolu s jejich intervaly spolehlivosti pro původní data. Jednoduše lze konstruovat intervaly spolehlivosti pro medián, který je exponenciální transformací parametru μ. Pro $100(1-\alpha)\%$ *oboustranný interval spolehlivosti mediánu* platí

$$\exp\left[\hat{\mu} - t_{1-\alpha/2}(n-1)\,\frac{\hat{\sigma}}{\sqrt{n}}\right] \le med \le \exp\left[\hat{\mu} + t_{1-\alpha/2}(n-1)\,\frac{\hat{\sigma}}{\sqrt{n}}\right] .$$

Podobně lze sestrojit i intervaly spolehlivosti pro variační koeficient, šikmost a špičatost, které jsou funkcí pouze parametru σ^2. Pro $100(1-\alpha)\%$ *oboustranný interval spolehlivosti variačního koeficientu* δ platí

$$\sqrt{\exp\frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}(n-1)} - 1} \le \delta \le \sqrt{\exp\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-1)} - 1} \ .$$

Pokud je třeba odhadnout střední hodnotu původních dat $M = E(x)$ a odpovídající rozptyl $V = D(x)$, užije se vztahů

$$\hat{M} = \exp(\hat{\mu})\; g(0.5\sigma^2) \ ,$$

$$\hat{V} = \exp(2\hat{\mu}) \left[g(2\sigma^2) - g\left(\frac{(n-2)\hat{\sigma}^2}{n-1}\right)\right] .$$

V obou vztazích je funkce $g(t)$ vyjádřena nekonečnou řadou

$$g(t) = 1 + \frac{n-1}{n}\, t + \sum_{j=2}^{\infty} \frac{(n-1)^{2j-1}}{n^{j}\,(n+1)\,(n+3)\,\ldots\,(n+2j-1)}\,\frac{t^{j}}{j!}\,.$$

f) **Logaritmicko-normální rozdělení tříparametrové**. Toto rozdělení má náhodná veličina, která může nabývat hodnot vyšších než spodní mez Θ, tzn. leží v intervalu $\Theta < x < \infty$. Náhodná veličina x má tříparametrové logaritmicko-normální rozdělení, pokud má náhodná veličina $\ln(x - \Theta)$ normální rozdělení $N(\mu, \sigma^2)$. Pro parametry polohy tříparametrového logaritmicko-normálního rozdělení platí, že jsou o Θ vyšší než odpovídající parametry dvouparametrového logaritmicko-normálního rozdělení. Parametry rozptýlení, šikmost a špičatost jsou shodné.

3.2 Intervalový odhad parametrů polohy a rozptýlení

Jelikož bodový odhad parametrů polohy a rozptýlení neříká nic o vzdálenosti od skutečné hodnoty Θ, kterou odhadujeme pomocí bodového odhadu, je třeba konstruovat *intervalový odhad parametru*. Intervalový odhad představuje interval, ve kterém se bude se zadanou pravděpodobností či statistickou jistotou $(1 - \alpha)$ nacházet skutečná hodnota daného parametru Θ. Interval parametru Θ odhadujeme dvěma číselnými hodnotami L_D a L_H, které tvoří *meze intervalu spolehlivosti*. Platí-li, že $P(L_D < \Theta < L_H) = 1 - \alpha$, interval spolehlivosti pokryje parametr Θ s předem zvolenou pravděpodobností či statistickou jistotou $P = (1 - \alpha)$ nazvanou také *koeficient spolehlivosti*. Jeho hodnota je obyčejně rovna 0.95 nebo 0.99. Parametr α se zde nazývá *hladina významnosti*.

Studentův *t*-test správnosti analytického výsledku je možné obejít využitím intervalu spolehlivosti. Nachází-li se totiž hodnota μ (tj. správná hodnota, norma, standard) v intervalu spolehlivosti $[L_D; L_H]$, je stanovení správné. Pro intervalový odhad platí:

1. Čím je rozsah výběru n větší, tím je interval spolehlivosti užší.
2. Čím je odhad přesnější (tj. čím má menší rozptyl), tím je interval spolehlivosti užší.
3. Čím je vyšší statistická jistota $(1 - \alpha)$, tím je interval spolehlivosti širší.

Míra polohy. Postup konstrukce intervalu spolehlivosti střední hodnoty μ pro výběry, pocházející z normálního rozdělení $N(\mu, \sigma^2)$, se pak rozlišuje dle velikosti výběru:

1. Velký výběr, $n \geq 30$: bodovým odhadem střední hodnoty μ je výběrový průměr $\bar{x}$ s rozdělením $N(\mu, \sigma^2/n)$. V intervalu $\bar{x} \pm 1.96\sigma/\sqrt{n}$ leží 95 % hodnot náhodných veličin výběru o rozsahu n. Potom 95% *oboustranný interval spolehlivosti střední hodnoty* je vyjádřen nerovností

$$\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\,,$$

kde hodnota 1.96 je 100(1 - 0.05/2) = 97.5% kvantil normovaného Gaussova normálního rozdělení $u_{0.975}$.

2. Střední výběr, $n \leq 30$: v praxi neznáme směrodatnou odchylku σ. Jelikož má $\sqrt{n}\,(\bar{x} - \mu)/s$ Studentovo *t*-rozdělení, bude 100(1 - α)% *oboustranný interval spolehlivosti střední hodnoty* vyjádřen nerovností

$$\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}.$$

Zde symbol $t_{1-\alpha/2}(\nu)$ označuje 100 (1 - α/2)% kvantil Studentova rozdělení s n - 1 stupni volnosti. Meze intervalu spolehlivosti závisí vedle směrodatné odchylky s i na rozsahu výběru n.

Interval spolehlivosti mediánu se vyčíslí podle přibližného vztahu

$$\tilde{x}_{0.5} - u_{1-\alpha/2}(n-1)\frac{0.707\,s}{\sqrt{n}} \le med \le \tilde{x}_{0.5} + u_{1-\alpha/2}(n-1)\frac{0.707\,s}{\sqrt{n}}.$$

Míra rozptýlení. 100 (1 - α)% *oboustranný interval spolehlivosti rozptylu* σ^2 se vypočte dle

$$\frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)\,s^2}{\chi^2_{\alpha/2}(n-1)},$$

kde $\chi^2_{1-\alpha/2}(n-1)$ je horní a $\chi^2_{\alpha/2}(n-1)$ dolní kvantil rozdělení χ^2.

3.3 Analýza malých výběrů

U malých výběrů jsou závěry vždy zatíženy značnou mírou nejistoty. Malých výběrů užijeme jen tam, kde obvykle z ekonomických či časových důvodů není možné získat větší počet dat.

(a) Zvláště malé výběry,

***n* = 2**: 100(1 - α)% *interval spolehlivosti střední hodnoty*

$$\frac{x_1 + x_2}{2} - T_\alpha\frac{|x_1 - x_2|}{2} \le \mu \le \frac{x_1 + x_2}{2} + T_\alpha\frac{|x_1 - x_2|}{2}.$$

Pro normální rozdělení $T_\alpha = \mathrm{cotg}(\alpha\,\pi/2)$, $T_{0.05} = 12.71$ a pro rovnoměrné rozdělení $T_\alpha = 1/\alpha - 1$, tj. $T'_{0.05} = 19$.

***n* = 3:** 100 (1 - α)% *interval spolehlivosti střední hodnoty*

$$\bar{x} - T^{\mathbb{N}}_\alpha\frac{s}{\sqrt{3}} \le \mu \le \bar{x} + T^{\mathbb{N}}_\alpha\frac{s}{\sqrt{3}}.$$

Pro normální rozdělení je $T^{\mathrm{t}}_\alpha \doteq 1/\sqrt{\alpha} - 3\sqrt{\alpha}/4 \ldots, T^{\mathrm{t}}_\alpha = 4.30$ a pro rovnoměrné rozdělení je $T_{0.05} = 5.74$.

(b) Malé výběry,

4 ≤ *n* ≤ 20, Hornův postup:

1. Postup je založený na pořádkových statistikách, $x_{(i)}$.

2. Hloubka pivotu je $H = (\mathrm{int}((n+1)/2))/2$ nebo $H = (\mathrm{int}((n+1)/2) + 1)/2$ podle toho, které číslo vyjde celé a dolní pivot je potom $x_D = x_{(H)}$ a horní pivot $x_H = x_{(n+1-H)}$.

3. Odhadem parametru polohy je *pivotová polosuma* $P_L = (x_D + x_H)/2$ a odhadem parametru rozptýlení je *pivotové rozpětí* $R_L = x_H - x_D$.

4. Náhodná veličina, použitelná k testování, $T_L = \dfrac{P_L}{R_L} = \dfrac{x_D + x_H}{2\,(x_H - x_D)}$ má přibližně symetrické rozdělení, jehož vybrané kvantily $t_{L,0.975}(n)$ jsou uvedeny v následující tabulce[1].

5. 95% *interval spolehlivosti střední hodnoty* se vypočte dle vztahu

$$P_L \& R_L t_{L,0.975}(n) \# \mu \# P_L \% R_L t_{L,0.975}(n).$$

Hornovy kvantily $t_{L,1-\alpha}(n)$ rozdělení T_L

n	1 - α =	0.9	0.95	0.975	0.99	0.995
4		0.477	0.555	0.738	1.040	1.331
5		0.869	1.370	2.094	3.715	5.805
6		0.531	0.759	1.035	1.505	1.968
7		0.451	0.550	0.720	0.978	1.211
8		0.393	0.469	0.564	0.741	0.890
9		0.484	0.688	0.915	1.265	1.575
10		0.400	0.523	0.668	0.878	1.051
11		0.363	0.452	0.545	0.714	0.859
12		0.344	0.423	0.483	0.593	0.697
13		0.389	0.497	0.608	0.792	0.945
14		0.348	0.437	0.525	0.661	0.776
15		0.318	0.399	0.466	0.586	0.685
16		0.299	0.374	0.435	0.507	0.591
17		0.331	0.421	0.502	0.637	0.774
18		0.300	0.380	0.451	0.555	0.650
19		0.288	0.361	0.423	0.502	0.575
20		0.266	0.337	0.397	0.464	0.519

3.4 Statistické testování

Uvedeme obecný postup statistického testování.

A. Postup testování statistické hypotézy

1. Formulace nulové hypotézy H_0 a alternativní hypotézy H_A.
2. Volba hladiny významnosti α.
3. Volba testační statistiky, např. *t*.
4. Určení kritického oboru testové charakteristiky.
5. Vyčíslení testační statistiky a jejích kvantilů.
6. Rozhodnutí, zda
   a) zamítnout hypotézu H_0 a přijmout H_A, jestliže testační statistika padne do kritického oboru,
   b) nezamítnout hypotézu H_0, jestliže testační statistika nepadne do kritického oboru.

Výsledek testování:

a) Zamítnutí hypotézy H_0 neznamená, že testovaná nulová hypotéza *neplatí*, ale znamená, že její platnosti nevěříme, protože výsledek testu poskytl objektivní důvod nevěřit. V dalším postupu pak budeme uvažovat, že H_0 neplatí a H_A platí.

b) Nezamítneme-li hypotézu H_0, neznamená to její přijetí. Výsledek testu neukázal tak velkou neshodu mezi zjištěnou skutečností a testovanou hypotézou, která by dala dostatečný

důvod k zamítnutí hypotézy.
Dva případy chybného rozhodnutí při testování:

a) Testační statistika padne mimo obor přijetí nulové hypotézy oboru O_p, tj. mimo interval $u_{\alpha/2} \# u_s \# u_{1\&\alpha/2}$,

a hypotéza H_0 přitom platí. Platí-li H_0, je pravděpodobnost padnutí u_s mimo obor O_p rovna právě hladině významnosti α. Velikost α určuje velikost *chyby I. druhu*, tj. nesprávného zamítnutí správné nulové hypotézy H_0.

b) Testační statistika padne do oboru O_P, tj. mimo interval

$$u_s < u_{\alpha/2}, \quad \text{resp.} \quad u_s > u_{1\&\alpha/2}.$$

a přitom platí alternativní hypotéza H_A. Pravděpodobnost, že u_s padne do oboru přijetí O_p, i když H_0 neplatí, představuje velikost *chyby II. druhu*, ß.

B. Testy střední hodnoty ("testy správnosti")

a) **100 (1 - α)% interval spolehlivosti.** Vypočteme intervalový odhad parametru μ (tj. polohy či rozptýlení). Padne-li zadaná hodnota μ_0 parametru μ do tohoto intervalu, nezamítá se hypotéza H_0: $\mu = \mu_0$. Padne-li μ_0 mimo tento interval, zamítá se H_0.

b) **Studentův *t*-test.** Ze základního souboru s rozdělením $N(\mu, \sigma^2)$ provedeme náhodný výběr rozsahu *n* a vypočteme výběrový průměr $\bar{x}$ a směrodatnou odchylku *s*. Jako testovou statistiku zvolíme náhodnou veličinu

$$t \quad ' \quad \frac{\bar{x} \& \mu_0}{s} \sqrt{n} \quad .$$

Kritické obory testů polohy hypotézy H_0: $\mu = \mu_0$ proti různým alternativám H_A pro hladinu významnosti α jsou uvedeny v následující tabulce. Hraniční body kritického oboru představují 100α% kvantily známých rozdělení. Místo formálního testování, zda jsou tyto kvantily větší než testové statistiky, je možné přímo vyčíslit velikost pravděpodobnosti (1 - α) (u oboustranného testu (1 - α/2)).

Nulová hypotéza H_0	**Alternativní hypotéza H_A**	**Testační charakteristika**	**Kritický obor**
	$\mu > \mu_0$		$t \$ t_{(1-\alpha)}(n\text{-}1)$
$\mu = \mu_0$	$\mu < \mu_0$	$t = (x - \mu_0)\%n /s$	$t < t_{\alpha}(n\text{-}1)$
	$\mu \ldots \mu_0$		$*t* \$ t_{(1-\alpha/2)}(n\text{-}1)$

C. Testy shody středních hodnot ("testy shodnosti")

Porovnání dvou výběrů $\{x_i\}$, i = 1, ..., n_1, a $\{y_j\}$, $j = 1, ..., n_2$, patří k častým úlohám v přírodních i technických vědách, a to při

(a) porovnání výsledků z různých instrumentálních metod nebo laboratoří,
(b) ověřování nutnosti dělení heterogenních výběrů do homogenních podskupin,
(c) hodnocení rozdílu mezi rozličnými materiály a přístroji.

Někdy lze tuto úlohu převést na testování jednoho výběru. To je totiž případ, kdy mezi prvky obou výběrů existuje jistá logická vazba. Představují-li prvky x_i vlastnosti před úpravou materiálu a prvky y_i stejné vlastnosti po úpravě materiálu ***těchže*** vzorků ($n_1 = n_2$), lze utvořit jednorozměrný výběr, $D_i = x_i - y_i$, pro který lze užít klasickou statistickou analýzu. Pokud se střední hodnota μ_D významně neliší od nuly, znamená to, že $\mu_x = \mu_y$ a efekt zpracování materiálu není pro sledovanou vlastnost statisticky významný (tzv. *párový test)*. V obecnějším případě dvou výběrů lze zjistit, zda pocházejí ze stejného rozdělení pravděpodobnosti a zda se neliší v parametrech polohy a rozptýlení.

Postup testu shodnosti středních hodnot dvou souborů:

1. Ověření normálního rozdělení obou souborů: testy a statistické diagnostiky k ověření předpokladů o výběru.

2. Test shody rozptylů:

a) Klasický Fisherův-Snedecorovův *F*-test,
b) Modifikovaný Fisherův-Snedecorův *F*-test,
c) Robustní Jackknife test F_J.

3. Test shody středních hodnot dvou souborů:

a) Klasický Studentův *t*-test T_1 pro homoskedasticitu,
b) Klasický Studentův *t*-test T_2 pro heteroskedasticitu,
c) Modifikovaný Studentův *t*-test T_3 pro výběry, odchýlené od normálního rozdělení.
d) Robustní Jackknife test polohy T_4 pro homoskedasticitu,
e) Robustní Jackknife test polohy T_5 pro heteroskedasticitu.

1. krok: **Ověření normálního rozdělení obou výběrů**

Klasické testy vycházejí z předpokladů:

a) výběry $\{x_i\}$, $i = 1, ..., n_1$, a $\{y_j\}$, $j = 1, ..., n_2$ jsou vzájemně nezávislé;
b) rozdělení obou výběrů je normální, $x_i \sim N(\mu_x, \sigma_x^2)$ a $y_j \sim N(\mu_y, \sigma_y^2)$.

Existuje řada různých metod, které jsou použitelné i v případech, kdy jsou tyto dva předpoklady narušeny. Před vlastní statistickou analýzou je výhodné vyšetřit nejprve metodami průzkumové analýzy chování obou výběrů.

2. krok: **Testy shody rozptylů**

(a) Klasický *F*-test. Umožňuje ověření nulové hypotézy H_0: $\sigma_x^2 = \sigma_y^2$ proti alternativní H_A: $\sigma_x^2 \neq \sigma_y^2$. Vychází se z předpokladu, že oba výběry jsou nezávislé a pocházejí z normálního rozdělení. Testovací kritérium má tvar

$$F \;=\; \max\left(\frac{s_x^2}{s_y^2}, \frac{s_y^2}{s_x^2}\right) \quad .$$

Platí-li hypotéza H_0 a $s_x^2 > s_y^2$, má F kritérium F-rozdělení s $\nu_1 = n_1 - 1$ a $\nu_2 = n_2 - 1$ stupni volnosti. V opačném případě se pořadí stupňů volnosti zamění. Je-li $F > F_{1\&\alpha}(\nu_1 , \nu_2)$, je nulová hypotéza H_0 o shodnosti rozptylů zamítnuta.

(b) Modifikovaný *F*-test. Předchozí klasický F-test je značně citlivý na předpoklad normality. Mají-li obě výběrová rozdělení jinou špičatost než odpovídá normálnímu, je třeba užít kvantil $F_{1-\alpha}(\nu_1, \nu_2)$ se stupni volnosti ν_1 a ν_2, vyčíslenými podle vztahů

$$\nu_1 \;=\; \frac{n_1 - 1}{1 + \frac{\hat{g}_{2c}}{2}} \quad , \quad \nu_2 \;=\; \frac{n_2 - 1}{1 + \frac{\hat{g}_{2c}}{2}} \quad ,$$

$$\text{kde} \quad \hat{g}_{2c} \;=\; \frac{2\,(n_1 + n_2)\left[\sum_{i=1}^{n_1} (x_i - \bar{x})^4 + \sum_{i=1}^{n_2} (y_i - \bar{y})^4\right]}{\left[\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2\right]^2} - 3 \quad .$$

(c) Robustní Jackknife test. Jsou-li v datech navíc odlehlé hodnoty, jeví se užitečný robustní Jackknife test. Testovací kritérium má tvar

$$F_J \;=\; \frac{n_1 (\bar{z}_1 - \bar{z})^2 + n_2 (\bar{z}_2 - \bar{z})^2}{\dfrac{\sum_{i=1}^{n_1} (z_{1i} - \bar{z}_1)^2 + \sum_{i=1}^{n_2} (z_{2i} - \bar{z}_2)^2}{n_1 + n_2 - 2}} \quad ,$$

$$\text{kde} \qquad \bar{z} \;=\; \frac{n_1 \bar{z}_1 + n_2 \bar{z}_2}{n_1 + n_2} \quad , \quad \bar{z}_j \;=\; \frac{\sum_{i=1}^{n_j} z_{ji}}{n_j} \quad , \qquad j = 1, 2 \quad .$$

Veličina z_{1i} se počítá podle vztahu $\quad z_{1i} \;=\; n_1 \ln s_x^2 - (n_1 - 1) \ln s_{1(i)}^2 \quad ,$

$$\text{kde} \qquad s_{1(i)}^2 \;=\; \frac{1}{n_1 - 2} \sum_{j \neq i}^{n_1} (x_j - \bar{x}_{(i)})^2 \quad .$$

Ve vztahu se vyskytuje průměr s vynechanou i-tou hodnotou, pro který platí

$$\bar{x}_{(i)} = \frac{1}{n_1 - 1} \sum_{j \neq i}^{n_1} x_j .$$

Při výpočtu z_{2i} se do výše uvedených vztahů dosazují hodnoty $\{y_j\}, j = 1, ..., n_2$, rozptyl s_y^2 a rozsah výběru n_2.

Platí-li nulová hypotéza H_0, má testovací kritérium F_J přibližně F-rozdělení s $\nu_1 = 2$, $\nu_2 = n_1 + n_2 - 2$ stupni volnosti. Vyjde-li, že $F_J > F_{1-\alpha}(\nu_1, \nu_2)$, je nutné zamítnout hypotézu H_0 o shodnosti obou výběrových rozptylů na hladině významnosti α.

3. krok: **Testy shody středních hodnot dvou souborů**

Studentův *t*-test umožňuje testování hypotézy H_0: $\mu_x = \mu_y$ proti alternativní H_A: $\mu_x \neq \mu_y$ i při splnění obou uvedených předpokladů o výběrech:

(a) Klasický Studentův *t*-test T_1 pro shodné rozptyly. Pro $\sigma_x^2 = \sigma_y^2$ a vykazují-li obě rozdělení Gaussovo rozdělení, má testovací kritérium tvar

$$T_1 = \frac{|\bar{x} - \bar{y}|}{\sqrt{(n_1 - 1) s_x^2 + (n_2 - 1) s_y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} .$$

Platí-li, že $T_1 > t_{1-\alpha/2}(n_1 + n_2 - 2)$, je hypotéza H_0 o shodě středních hodnot na hladině významnosti α zamítnuta.

(b) Klasický Studentův *t*-test T_2 pro různé rozptyly. Pro $\sigma_x^2 \neq \sigma_y^2$ a vykazují-li obě rozdělení Gaussovo rozdělení, má testovací kritérium tvar

$$T_2 = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}} .$$

Platí-li hypotéza H_0, má tato testová statistika Studentovo rozdělení s *"ekvivalentními"* stupni volnosti ν

$$\nu = \frac{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}{\frac{s_x^4}{n_1^2 (n_1 - 1)} + \frac{s_y^4}{n_2^2 (n_2 - 1)}} .$$

Platí-li, že $T_2 > t_{1-\alpha/2}(\nu)$, je hypotéza H_0 o shodě středních hodnot na hladině významnosti α zamítnuta.

Testovací kritérium T_1 není robustní vůči heteroskedasticitě, tj. případu, kdy data jsou ve výběrech měřena s různou přesností. V této situaci je správnější užít testovacího kritéria T_2, které je vůči heteroskedasticitě robustnější. Na druhé straně však ekvivalentní stupně volnosti ν vycházejí menší než $n_1 + n_2 - 2$, takže síla testu T_2 je nižší než síla T_1.

(c) Modifikovaný Studentův *t*-test T_3 pro výběry, odchýlené od normálního rozdělení. Jestliže jedno z rozdělení se odchyluje od normality nebo se významně liší v šikmosti od druhého, je vhodné použít modifikované testovací kritérium T_3

$$T_3 = \frac{|\bar{x} - \bar{y}| + C + D(\bar{x} - \bar{y})^2}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}},$$

kde $C = \frac{1}{6}\,\frac{\frac{\hat{g}_{1x}}{n_1^2}\frac{s_x^3}{\sqrt{n_1}} - \frac{\hat{g}_{1y}}{n_2^2}\frac{s_y^3}{\sqrt{n_2}}}{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}$ a $D = \frac{1}{3}\,\frac{\frac{\hat{g}_{1x}}{n_1^2}\frac{s_x^3}{\sqrt{n_1}} - \frac{\hat{g}_{1y}}{n_2^2}\frac{s_y^3}{\sqrt{n_2}}}{\left[\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right]^2}$.

V těchto vztazích jsou $\hat{g}_{1x}$ a $\hat{g}_{1y}$ výběrové šikmosti. Aby bylo možné užít kvantilů Studentova rozdělení pro předepsanou hladinu významnosti α, je třeba přeformulovat testovací kritérium T_3 do tvaru

$$T_3 = T_2 + B_x - B_y,$$

kde

$$B_x = \frac{\frac{\hat{g}_{1x}\, s_x^3}{6\, n_1^2 \sqrt{n_1}\left[\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right]} + \frac{\hat{g}_{1x}\, s_x^2 (\bar{x} - \bar{y})^2}{3\, n_2^2 \sqrt{n_2}\left[\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right]}}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}}$$

a B_y se vyčíslí analogicky, pouze šikmost $\hat{g}_{1x}$ se nahradí hodnotou $\hat{g}_{1y}$, rozptyl σ_x^2 hodnotou σ_y^2 a rozsah n_1 hodnotou n_2. Za předpokladu platnosti hypotézy H_0 má testovací kritérium T_3 Studentovo rozdělení s počtem stupňů volnosti ν. Test založený na kritériu T_3 je robustní vůči zešikmení výběrových rozdělení i vůči heteroskedasticitě a není u něho požadována ani shoda rozptylů, $\sigma_x^2 \ldots \sigma_y^2$. Vůči odchylkám rozdělení od normality ve špičatosti jsou uvedené *t*-testy T_1, T_2 a T_3 dostatečně robustní. Je možné použít i korekci na špičatost, což však nepřináší výrazné zlepšení.

(d) Robustní Jackknife test polohy T_4 pro homoskedasticitu. Jsou-li ve výběrech přítomna odlehlá měření, lze pro test hypotézy H_0: $\mu_1 = \mu_2$, a $\sigma_1^2 = \sigma_2^2$ upravit testovací kritérium založené na uřezaném průměru na tvar

$$T_4 = \frac{(\bar{x}(h) - \bar{y}(h))}{\sqrt{S_{w,x}(h) + S_{w,y}(h)}},$$

kde $S_{w,x}(h)$ a $S_{w,y}(h)$ se vyčíslí pro výběry $\{x_i\}$, $i = 1, \ldots, n_1$, a $\{y_j\}$, $j = 1, \ldots, n_2$. Je-li

$n_1 = n_2$, má náhodná veličina T_4 přibližně Studentovo rozdělení s 2 (k - 1) stupni volnosti. Test T_4 lze použít jen pro rozsah $n \geq 7$.

(e) Robustní Jackknife test polohy T_5 pro heteroskedasticitu. Pro případ nestejných rozptylů $\sigma_1^2 \dots \sigma_2^2$ a nestejných rozsahů $n_1 \dots n_2$ a s využitím kritéria T_2 lze formulovat robustní kritérium T_5 pro test hypotézy H_0: $\mu_x = \mu_y$

$$T_5 = \frac{\bar{x}(h) - \bar{y}(h)}{\sqrt{\frac{s_{w,x}^2}{h_1} + \frac{s_{w,y}^2}{h_2}}}, \text{ kde } s_{w,x}^2 = \frac{S_{w,x}(h)}{h_1 - 1}, \quad s_{w,y}^2 = \frac{S_{w,y}(h)}{h_2 - 1}$$

$$h_i = n_i - 2\,\mathrm{int}\left(\frac{h\,n_i}{100}\right), \quad \text{pro } i = 1, 2.$$

Testovací kritérium T_5 má přibližně Studentovo rozdělení s ν stupni volnosti, pro které platí

$$\frac{1}{\nu} = \frac{z^2}{h_1 - 1} + \frac{(1 - z)^2}{h_2 - 1}, \text{ kde } z = \frac{\frac{s_{w,x}^2}{h_1}}{\frac{s_{w,x}^2}{h_1} + \frac{s_{w,y}^2}{h_2}}.$$

Robustní testy T_4 a T_5 jsou výhodné také pro rozdělení s dlouhými konci, když je špičatost větší než 3. V případě normálního rozdělení však mají menší sílu než testy T_1 a T_2.

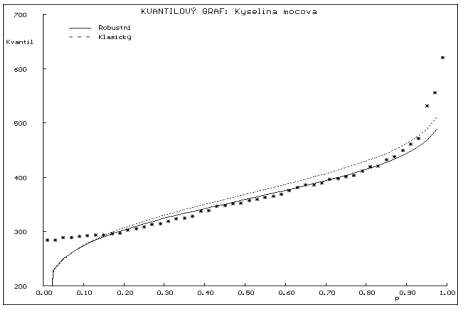
Vzorová úloha 3.1 *Analýza velkého výběru*

Na úloze **B2.25** *Koncentrace kyseliny močové v krvi dárců* ukážeme postup analýzy velkého výběru s odlehlými prvky pro určení typu rozdělení koncentrace kyseliny močové u 50 dárců krve. Jaká je míra polohy a rozptýlení uvedeného výběru?

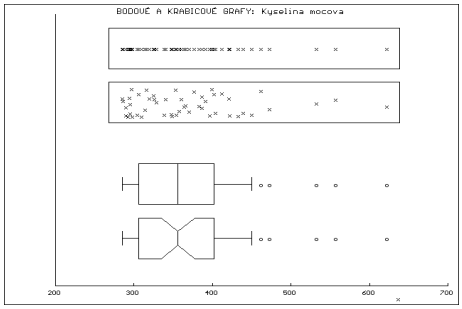
Řešení:

1. Průzkumová (exploratorní) analýza dat

Z grafických diagnostik průzkumové analýzy dat jsou uvedeny pouze čtyři nejdůležitější.

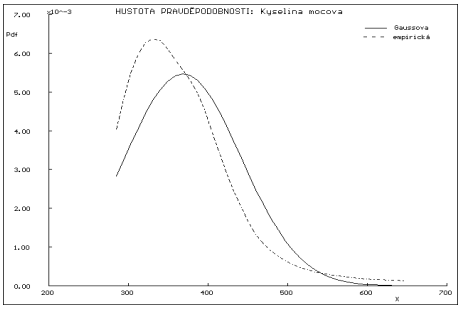


Obr. 3.2 Kvantilový graf.

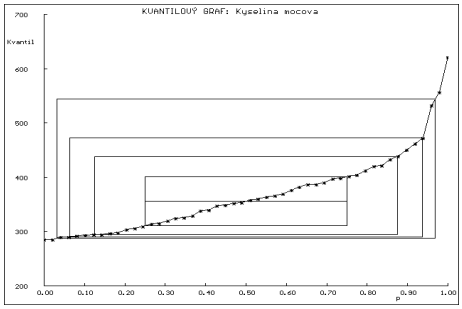


Obr. 3.3 Diagramy rozptýlení a krabicové grafy.

Grafy ukazují, že výběrové rozdělení je asymetrické a silně zešikmené. V horní části pořádkových statistik lze indikovat 3 až 5 podezřelých bodů, z nich se 3 jeví jako vysloveně odlehlé. Nelze proto použít klasických odhadů polohy a rozptýlení.



Obr. 3.4 Graf hustoty pravděpodobnosti.



Obr. 3.5 Graf rozptýlení s kvantily.

2. Ověření základních předpokladů o výběru

Předpoklad o normalitě rozdělení je zamítnut, protože hodnota testovacího kritéria χ^2_{exp} je vyšší než tabulkový kvantil $\chi^2_{1-\alpha}(2)$. Předpoklad nezávislosti je přijat, protože hodnota testovacího kritéria je nižší než tabulkový kvantil. Data nejsou homogenní, mimo modifikované vnitřní hradby $B_D^{(}$ = 109.62 μmol. l^{-1}, $B_H^{(}$ = 602.38 μmol. l^{-1} leží hodnota č. 17, a to 622.0 μmol. l^{-1}. Po odstranění této odlehlé hodnoty by byl aritmetický průměr $\overline{x}$ = 363.29 μmol. l^{-1} a směrodatná odchylka s = 63.875 μmol. l^{-1}. Protože se však jedná o biochemická data, nelze odlehlou hodnotu z dat vyloučit. Znamenalo by to zde totiž ztrátu důležité biochemické informace.

Základní předpoklady výběru **úlohy B2.25** (ADSTAT)

(a) Test normality: tabulkový kvantil $\chi^2_{1-\alpha}(2)$:	5.992
Odhad χ^2_{exp} statistiky:	29.199
Závěr: Předpoklad normality zamítnut na spočtené hladině významnosti α = 4.566E-07.	
(b) Test nezávislosti: tabulkový kvantil $t_{1-\alpha/2}(n+1)$:	2.008
Odhad von Neumannovy statistiky t_n:	1.0266
Závěr: Předpoklad nezávislosti přijat na spočtené hladině významnosti α = 0.155.	
(c) Detekce odlehlých bodů: metodou modifikované vnitřní hradby	

Dolní vnitřní hradba B^*_D:	109.62
Horní vnitřní hradba B^*_H:	602.38
Závěr: Ve výběru je 1 odlehlý bod, a to bod č. 17 (horní) o hodnotě 622.0.	
(d) Opravené parametry výběru s vynechanými odlehlými hodnotami:	
Odhad aritmetického průměru x:	363.29
Odhad směrodatné odchylky s:	63.88
Odhad šikmosti $\hat{g}_1$:	0.97
Odhad špičatosti $\hat{g}_2$:	3.81

3. Mocninná transformace

Při pokusu o transformaci dat poskytuje prostá mocninná transformace hodnotu opraveného průměru $\bar{x}_R$ = 350.91 μmol. l^{-1}, zatímco Boxova - Coxova transformace $\bar{x}_R$ = 362.17 μmol. l^{-1}, s odhadem šikmosti g_1 = 0.81 a špičatosti g_2 = 3.38, což je bližší parametrům normálního rozdělení. U mocninné transformace výpočet šikmosti a špičatosti selhává. Věrohodnější se zde proto jeví odhad, získaný metodou Boxovy-Coxovy transformace.

Boxova-Coxova transformace u dat výběru úlohy B2.25 (ADSTAT)

Boxova-Coxova transformace:	
Odhad optimálního exponentu λ	-2.13
Opravený odhad průměru původních dat $\bar{x}_R$	350.91

4. Odhady polohy, rozptylu a tvaru rozdělení

Rozdělení souboru vykazuje mírné zešikmení. Soubor obsahuje jeden výrazně odlehlý bod. Mocninná transformace selhává, Boxova-Coxova transformace přináší zlepšení parametrů šikmosti a špičatosti souboru a je robustní vůči odlehlé hodnotě. Dobrým odhadem střední hodnoty se jeví také uřezané aritmetické průměry.

Analýza jednorozměrného výběru dat **úlohy B2.25** (ADSTAT)

(1) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$	368.46
Odhad směrodatné odchylky s	73.04
Odhad šikmosti $\hat{g}_1$	1.33
Odhad špičatosti $\hat{g}_2$	5.00
Dolní mez 95.0% intervalu spolehlivosti L_D	347.70
Horní mez 95.0% intervalu spolehlivosti L_H	389.22
(2) Odhady ostatních parametrů:	
Odhad modu $\hat{x}_M$	293.00
Odhad polosumy $\hat{x}_P$	453.50
(3) Robustní odhady parametrů:	
Medián $\tilde{x}_{0.5}$	356.00
Odhad směrodatné odchylky mediánu $s(\tilde{x}_{0.5})$	85.48
Dolní mez 95.0% intervalu spolehlivosti L_D	331.52
Horní mez 95.0% intervalu spolehlivosti L_H	380.48

Odhad 40% uřezaného průměru $\bar{x}$(40%)	356.20
Odhad směrodatné odchylky s(40%)	73.69
Odhad winsorizovaného průměru $\bar{x}_w$(40%)	355.24
Odhad směrodatné odchylky s_w(40%)	31.26
Dolní mez 95.0% intervalu spolehlivosti L_D	332.98
Horní mez 95.0% intervalu spolehlivosti L_H	379.42
Odhad M-odhadu střední hodnoty $\hat{\mu}_M$	358.17
Odhad směrodatné odchylky s_M	63.09
Dolní mez 95.0% intervalu spolehlivosti L_D	339.29
Horní mez 95.0% intervalu spolehlivosti L_H	377.05
(4) Hoggovy adaptivní odhady parametrů:	
Hoggův průměr $\hat{\mu}_M$	356.46
Odhad směrodatné odchylky s_M	72.72
Dolní mez 95.0% intervalu spolehlivosti L_D	335.79
Horní mez 95.0% intervalu spolehlivosti L_H	377.13

Vzorová úloha 3.2 *Analýza malého výběru*

Na úloze **B3.01** *Střední hodnota haptoglobinu v lidském krevním séru* ukážeme Hornův postup analýzy malých výběrů.

Řešení: (a) Užijeme Hornův postup pivotů pro malé výběry ($4 < n < 20$):

1. Pořádkové statistiky: seřadíme prvky od nejmenší do největší hodnoty

i	1	2	3	4	5	6	7	8
$x_{(i)}$	0.15	0.49	1.07	1.27	1.82	1.98	3.32	3.79

2. Hloubka pivotu: vyčíslíme hloubku pivotu

$n = \mathbf{8}$, sudé

$$H = \text{int}\,\frac{\frac{n+1}{2}+1}{2} \qquad \text{int}(2.75) = 2$$

3. Pivoty: Dolní pivot $x_D = x_{(H)}$ $\quad x_{(2)} = 0.49$

Horní pivot $x_H = x_{(n+1-H)}$ $\quad x_{(7)} = 3.32$

4. Pivotová polosuma $P_L = \dfrac{x_D + x_H}{2} = 1.905$

5. Pivotové rozpětí $R_L = x_H - x_D$ $\quad 3.32 - 0.49 = 2.83$

6. 95% interval spolehlivosti střední hodnoty μ: $\quad t_{L,1-\alpha/2} = 0.564$

$$P_L - R_L\, t_{L,1-\alpha/2}(n) \le \mu \le P_L + R_L\, t_{L,1-\alpha/2}(n)$$

$$1.905 - 2.83 \times 0.564 \le \mu \le 1.905 + 2.83 \times 0.564$$

$$0.31 \le \mu \le 3.50.$$

7. Závěr: Bodový odhad míry polohy je 1.91, míry rozptýlení 2.83 a intervalový odhad míry polohy je $0.31 \le \mu \le 3.50$.

(b) Užijeme také počítačové analýzy jednorozměrných dat: z průzkumové analýzy dat a ověření předpokladů o výběru plyne závěr, že rozdělení výběru pochází z Gaussova rozdělení, prvky výběru jsou nezávislé a ve výběru nejsou odlehlé body.

Analýza jednorozměrného výběru **úlohy B3.01** (ADSTAT)

(1) Test normality: tabulkový kvantil $\chi^2_{1-\alpha}(2)$:	5.992
Odhad χ^2_{exp} statistiky:	0.809
Závěr: Předpoklad normality přijat na spočtené hladině významnosti $\alpha = 0.6674$.	
(2) Test nezávislosti: tabulkový kvantil $t_{1-\alpha/2}(n+1)$:	2.262
Odhad von Neumannovy statistiky t_n:	1.077
Závěr: Předpoklad nezávislosti přijat na spočtené hladině významnosti $\alpha = 0.155$.	
(3) Detekce odlehlých bodů: metodou modifikované vnitřní hradby	
Závěr: Ve výběru nejsou odlehlé body.	
(4) Prostá mocninná transformace:	
Odhad optimálního exponentu λ	0.53
Odhad průměru transformovaných dat $\bar{y}$	1.246
Opravený odhad průměru původních dat $\bar{x}_R$	1.510
(5) Boxova-Coxova transformace:	
Odhad optimálního exponentu λ	0.53
Odhad průměru transformovaných dat $\bar{y}$	0.461
Opravený odhad průměru původních dat $\bar{x}_R$	1.510
(6) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$	1.736
Odhad směrodatné odchylky s	1.283
Odhad šikmosti $\hat{g}_1$	0.46
Odhad špičatosti g_2	1.99
Dolní mez 95.0% intervalu spolehlivosti L_D	0.664
Horní mez 95.0% intervalu spolehlivosti L_H	2.809
(7) Robustní odhady parametrů:	
Medián $\tilde{x}_{0.5}$	1.545
Odhad směrodatné odchylky mediánu $s(\tilde{x}_{0.5})$	1.347
Odhad 40% uřezaného průměru $\bar{x}(40\%)$	1.545
Odhad směrodatné odchylky $s(40\%)$	1.347
Dolní mez 95.0% intervalu spolehlivosti L_D	-0.821
Horní mez 95.0% intervalu spolehlivosti L_H	3.912
Odhad M-odhadu střední hodnoty $\hat{\mu}_M$	1.679
Odhad směrodatné odchylky s_M	1.234
Dolní mez 95.0% intervalu spolehlivosti L_D	0.602
Horní mez 95.0% intervalu spolehlivosti L_H	2.756
(8) Hoggovy adaptivní odhady parametrů:	
Hoggův průměr $\hat{\mu}_M$	1.736
Odhad směrodatné odchylky s_M	1.283
Dolní mez 95.0% intervalu spolehlivosti L_D	0.664
Horní mez 95.0% intervalu spolehlivosti L_H	2.809

Vzorová úloha 3.3 *Test střední hodnoty (test správnosti)*

Na úloze *Test správnosti koncentrace vápníku* ukážeme užití testu správnosti. Výrobce kontrolního komerčního materiálu uvádí koncentraci vápníku 2.20 mmol. l^{-1}. Jsou naměřené výsledky správné?

Data: Koncentrace vápníku [mmol. l^{-1}] v komerčním materiálu:
2.262.162.182.152.232.252.192.182.162.20
2.192.222.192.212.252.292.262.152.18

Řešení: Z exploratorní analýzy dat byla zjištěna mírná asymetrie, posun k nižším hodnotám, v horní části řady pořádkových statistik 3 podezřelé body. Ze základních předpokladů vyplývá, že data jsou homogenní, soubor neobsahuje odlehlé hodnoty. Data mají normální rozložení a jsou nezávislá.

Analýza jednorozměrných dat koncentrace vápníku (ADSTAT)

(1) Odhady klasických parametrů:	
Odhad aritmetického průměru $\bar{x}$	2.205
Odhad směrodatné odchylky s	0.041
Odhad šikmosti g_1	0.44
Odhad špičatosti g_2	2.12
Dolní mez 95.0% intervalu spolehlivosti L_D	2.185
Horní mez 95.0% intervalu spolehlivosti L_H	2.225
(2) Robustní odhady parametrů:	
Medián $x_{0.5}$	2.190
Odhad směrodatné odchylky mediánu $s(\tilde{x}_{0.5})$	0.056
Dolní mez 95.0% intervalu spolehlivosti L_D	2.161
Horní mez 95.0% intervalu spolehlivosti L_H	2.219
Odhad 40% uřezaného průměru $\bar{x}(40\%)$	2.194
Odhad směrodatné odchylky $s(40\%)$	0.040
Dolní mez 95.0% intervalu spolehlivosti L_D	2.171
Horní mez 95.0% intervalu spolehlivosti L_H	2.219
Odhad *M*-odhadu střední hodnoty $\hat{\mu}_M$	2.203
Odhad směrodatné odchylky s_M	0.041
Dolní mez 95.0% intervalu spolehlivosti L_D	2.183
Horní mez 95.0% intervalu spolehlivosti L_H	2.224
(3) Hoggovy adaptivní odhady parametrů:	
Hoggův průměr μ_M	2.205
Odhad směrodatné odchylky s_M	0.041
Dolní mez 95.0% intervalu spolehlivosti L_D	2.185
Horní mez 95.0% intervalu spolehlivosti L_H	2.225
(4) Prostá mocninná transformace:	
Odhad optimálního exponentu λ	-4.00
Opravený odhad průměru původních dat $\bar{x}_R$	2.204

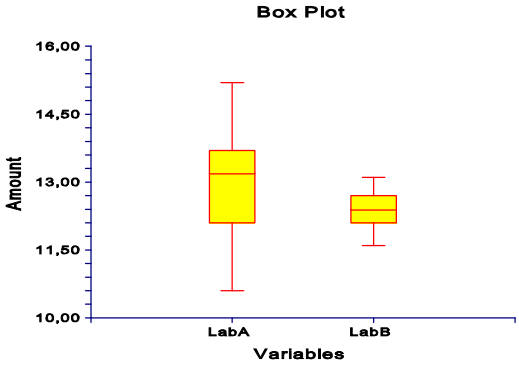
Závěr: Pro 95% statistickou jistotu byly nalezeny následující intervalové odhady: pro aritmetický průměr x je interval $2.19 < \mu < 2.23$, pro medián $\tilde{x}_{0.5}$ pak $2.16 < \mu < 2.22$. Z uvedených intervalových odhadů vyplývá, že obsah vápníku 2.20 mmol. l^{-1} leží v rozmezí zadané normy a naměřené výsledky jsou správné.

Vzorová úloha 3.4 *Test shodnosti středních hodnot*

Na vzorové úloze *Porovnání práce dvou laborantek* ukážeme aplikaci testu shodnosti. Dvě laborantky prováděly analýzu koncentrace mukoproteinů v kontrolním vzorku. Určete, zda obě dospěly ke stejným výsledkům. Test shodnosti proveďte na hladině významnosti $\alpha = 0.05$.

Data: Stanovení koncentrace mukoproteinů v g. l^{-1} laborantkou A a laborantkou B:
A: 11.6 12.1 13.2 12.1 10.6 13.3 13.7 14.4 15.2 13.6 13.7 12.4 12.5
B: 12.4 12.8 12.3 12.7 12.4 12.5 11.9 13.1 12.7 12.5 11.8 11.6 12.3

Řešení: Z ověření základních předpokladů pro jednotlivé výběry vyplývá, že data v obou výběrech jsou nezávislá, homogenní bez odlehlých bodů, test normality u obou výběrů prokázal Gaussovo rozdělení.



Obr. 3.6 Krabicový graf porovnání dvou výběrů, ***NCSS2000.***

Porovnání dvou výběrů (ADSTAT)

(1) Klasické odhady parametrů:

Parametr	**Výběr 1**	**Výběr 2**	**Celkově**
Četnost	13	13	26
Průměr	12.954	12.385	12.669
Rozptyl	1.5160	0.1764	0.81243
Šikmost	-0.07	-0.34	-0.10
Špičatost	2.58	2.50	4.21

(2) Test homogenity rozptylu H_0: s_1^2 ' s_2^2. Fisherův-Snedecorův *F*-test:

Počet stupňů volnosti n_1 - 1	12
Počet stupňů volnosti n_2 - 1	12
Tabulkový kvantil $F_{1-\alpha}(n_1 - 1, n_2 - 1)$	3.277
Experimentální F_{exp} statistika	8.594

Závěr: Rozptyly se považují za rozdílné, H_0 zamítnuta

(3) Test shody průměrů H_0: $\mu_1 = \mu_2$, Studentův *t*-test (pro různé rozptyly)

Tabulkový kvantil $t_{1-\alpha/2}(v)$	2.120
T_2 statistika	1.578

Závěr: Průměry se považují za shodné, H_0 přijata

Závěr: Na hladině významnosti $\alpha = 0.05$ potvrzuje oboustranný klasický test shodu středních hodnot obou výběrů i při významné odlišnosti obou rozptylů. Obě laborantky dosáhly stejných výsledků, i když každá s jinou variabilitou.

Vzorová úloha 3.5 *Párový test*

Na úloze **H3.22** *Zkoušení obsahu niklu v drátu a svárovém kovu* u párových dat ukážeme párový test, $n = 45$.
Řešení: Párový test **úlohy H3.22** (ADSTAT): Průměrný rozdíl $\bar{D}$ je 0.04356.

Rozptyl s_D^2	0.03004
Počet stupňů volnosti n-1	44
Tabulkový kvantil $t_{1-\alpha/2}(n\text{-}1)$	2.015
t_{exp} statistika	9.728
Závěr: Průměry se považují za rozdílné, H_0 zamítnuta na hladině významnosti 0.000.	

Závěr: Párový test zamítl hypotézu o shodě obsahu niklu v drátu a svárovém kovu.

3.5 Úlohy

Řešte úlohy jednorozměrných dat s využitím interaktivní analýzy tak, že na každý soubor dat budete aplikovat jednotlivé metody dle následujícího postupu:

Postup analýzy jednorozměrných dat:

1. Exploratorní (průzkumová) analýza dat EDA: postup je uveden v 2. kapitole.

2. Konfirmatorní analýza dat CDA: na základě exploratorní analýzy dat nalezněte

odhad míry polohy (aritmetický průměr $\bar{x}$, medián $\tilde{x}_{0.5}$, modus x_M, polosumu $\bar{x}_P$, pivotová polosumu P_L, retransformovaný průměr $\bar{x}_R$, uřezaný průměr $\bar{x}(\theta\%)$ a další),

odhad míry rozptýlení (odhad směrodatné odchylky s, interkvartilové rozpětí R, pivotové rozpětí R_L a další) a

odhad míry tvaru (odhad koeficientu šikmosti $\hat{g}_1$, odhad koeficientu špičatosti g_2),

intervalové odhady vyčíslete na hladině významnosti $\alpha = 0.05$,

pro malé výběry s četností menší než 7 aplikujte pouze Hornův postup malých výběrů. U malých výběrů $4 \# n \# 20$ poskytuje totiž správné odhady střední hodnoty jedině Hornův postup pivotů. Pivotová polosuma a pivotové rozpětí umožnují vyčíslit i intervalový odhad střední hodnoty a navíc jsou oba odhady dostatečně robustní vůči asymetrii rozdělení malého výběru a i vůči odlehlým hodnotám.

3. Statistické testování: testy správnosti a shodnosti. Intervalový odhad míry polohy a Studentův t-test správnosti jsou vhodné k posouzení správnosti analytického výsledku. Nachází-li se totiž hodnota μ_0 (tj. "pravda", správná hodnota, norma, standard) v intervalu spolehlivosti $[L_D; L_H]$, je stanovení správné. Exploratorní analýza předurčí volbu, zda k testu správnosti využijeme intervalový odhad aritmetického průměru, uřezaného průměru, retransformovaného průměru, mediánu nebo pivotové polosumy. Interaktivní statistická analýza s vhodným software umožnuje snadno a jednoznačně vyšetřit správnost analytického výsledku.

Úlohy jsou rozděleny do pěti kapitol: B3 (farmakologická a biochemická data), C3 (chemická a fyzikální data), E3 (environmentální, potravinářská a zemědělská data), H3 (hutní a mineralogická data) a S3 (ekonomická a sociologická data).

3.5.1 Analýza farmakologických a biochemických dat

Úloha B3.01 *Střední hodnota haptoglobinu v lidském krevním séru (Horn)*
Bylo provedeno měření koncentrace haptoglobinu v lidském krevním séru osmi dospělých jedinců. Vypočtěte střední hodnotu, parametr rozptýlení a 95% interval spolehlivosti střední hodnoty. Vyšetřete, zda tento výběr nepochází z logaritmicko-normálního rozdělení. Aplikujte také Hornův postup (str. 51 v cit[14]).
Data: Koncentrace haptoglobinu [g. l^{-1}] v lidském krevním séru : 1.82, 3.32, 1.07, 1.27, 0.49, 3.79, 0.15, 1.98.

Úloha B3.02 *Posouzení obsahu vápníku v krevním séru nemocných lidí (Horn)*
Obsah vápníku v krevním séru nemocných pacientů je nižší než u zdravých lidí. Dokažte, že střední hodnota výběru nemocných je významně nižší než střední hodnota výběru zdravých jedinců. Aplikujte i Hornův postup. Postupy porovnejte na hladině významnosti $\alpha = 0.05$ (str. 485 v cit.[18]).
Data: Koncentrace vápníku v krevním séru [mmol. l^{-1}], *B302a* nemocní, *B302b* zdraví:

B302a Nemocní:	2.09	1.80	1.97	2.35	2.08	1.90	2.06	2.30	2.35
B302b Zdraví:	2.15	2.13	2.27	2.52	2.11	2.24	2.26	2.34	2.68

Úloha B3.03 *Vliv glucagonu na koncentraci krevního cukru (Horn)*
Je třeba vyšetřit vliv farmaka glucagonu na snížení hladiny krevního cukru. Po 15 minutách od dávkování glucagonem byla u 8 pokusných krys hladina 300 mg/100 ml významně snížena. Byly naměřeny následující hodnoty krevního cukru v mg/100 ml krve. Jde o symetrické rozdělení? Určete střední hodnotu krevního cukru. Aplikujte i Hornův postup, (str.43 v cit.[19]). Obsahuje intervalový odhad hodnotu 300 mg/100 ml?
Data: Hladina krevního cukru [mg/100 ml]: 270, 275, 265, 250, 280, 245, 265, 260.

Úloha B3.04 *Stanovení hmotnosti účinné látky v tabletě (Horn)*
Analýzou byla stanovena hmotnost účinné látky v tabletě. Vyšetřete předpoklady o výběru. Jsou prvky výběru vybrány náhodně, není mezi nimi skrytá závislost? Pocházejí data ze souboru, který vykazuje Gaussovo rozdělení? U kolika diagnostik je shoda v indikaci odlehlých bodů? Určete parametr polohy a parametr rozptýlení Hornovým postupem a porovnejte s klasickými, či robustními odhady. Zdůvodněte nejlepší odhad střední hodnoty.
Data: Hmotnost účinné látky v tabletě [g]: 19.27, 19.25, 19.49, 19.71, 19.98, 19.26, 19.79, 19.29, 19.78, 19.36, 19.66.

Úloha B3.05 *Stanovení čistoty kalciferolu (Horn)*
V kontrolní laboratoři se stanovuje čistota kalciferolu metodou vnějšího standardu na kapalinovém chromatografu. Vyšetřete požadavky kladené na výběr a určete parametry polohy a rozptýlení. Které diagnostiky ukazují, že ve výběru jsou odlehlá měření? Je rozdělení symetrické? Na jaký tvar rozdělení ukazuje koeficient šikmosti a špičatosti? Stanovte střední hodnotu a její 95% interval spolehlivosti. Aplikujte také Hornův postup. U kterého odhadu polohy vychází interval spolehlivosti širší?
Data: Procento čistoty kalciferolu [%]: 99.0, 97.6, 99.5, 98.8, 101.0, 99.2, 98.7.

Úloha B3.06 *Test shodnosti obsahu paracetamolu dvěma analytickými metodami*

V podnikové normě analgetika Ataralgin je stanovení paracetamolu metodou HPLC doplněno o alternativní stanovení metodou multikomponentní UV-VIS spektroskopie. Pro porovnání výsledků byl paracetamol stanoven u jedné výrobní šarže oběma povolenými metodami. Vyšetřete předpoklady o výběru a sestrojte histogram. Jsou prvky každého výběru vybrány náhodně, není mezi prvky skrytá závislost? Pocházejí data obou výběrů ze souboru, který vykazuje Gaussovo rozdělení? Obsah paracetamolu v tabletě je v normě dán rozmezím 0.309 - 0.341 g/tbl. Posuďte jednak správnost obsahu paracetamolu u vyrobené tablety a jednak shodnost střední hodnoty obou výběrů.

Data: Obsah paracetamolu v Ataralginu [g/tbl]: *B306a* 1. metoda, *B306b* 2. metoda.

B306a	0.332	0.325	0.313	0.340	0.334	0.322	0.329	0.325
	0.319	0.327	0.331	0.338	0.314	0.317	0.330	
B306b	0.329	0.326	0.325	0.326	0.334	0.328	0.330	0.321
	0.325	0.330	0.338	0.331	0.313	0.323	0.331	

Úloha B3.07 *Porovnání čistoty ergosterolu z produkce dvou měsíců*

V kontrolní laboratoři se průběžně sleduje čistota ergosterolu, meziproduktu výroby kalciferolu. Testem shodnosti srovnejte výsledky čistoty (a) za měsíc říjen a (b) za listopad. Je splněna homogenita a Gaussovo rozdělení obou výběrů?

Data: Obsah ergosterolu [%]: *B307a* za říjen, *B307b* za listopad

(a) *B307a*:

92.2	93.3	91.3	90.4	92.3	93.4	93.6	94.3	93.6
95.9	94.1	93.7	94.9	93.6	94.9	95.7	93.8	96.3
96.2	96.5	97.0	97.0	86.9	94.0	93.0	93.0	

(b) *B307b*:

93.0	94.1	92.1	92.8	90.8	90.3	92.5	90.2	94.0
93.3	92.5	92.8	92.8	93.2	93.8	93.4	93.8	95.0
96.3	97.1	95.5	97.1	97.5	97.0	96.7	96.4	

Úloha B3.08 *Kontrola nástřiku autosampleru u dvou různých koncentrací (Horn)*

Při studii biologické dostupnosti cyclosporinu A byla provedena kontrola správnosti nástřiku autosampleru. Bylo provedeno deset nástřiků dvou různých koncentrací. Vyšetřete rozdělení a ověřte předpoklady o výběru. Určete střední hodnotu nástřiku a míru rozptýlení. Je míra rozptýlení pro obě koncentrace přibližně stejná? Užijte také Hornův postup.

Data: Nástřik autosampleru [mg. l^{-1}]: *B308a* 1. koncentrace, *B308b* 2. koncentrace,

(a) *B308a*

3.14	3.15	3.17	3.19	3.19	3.20	3.20	3.21	3.23	3.25

(b) *B308b*

1.55	1.56	1.58	1.59	1.60	1.60	1.60	1.62	1.61	1.65

Úloha B3.09 *Test správnosti koncentrace cyclosporinu metodou HPLC (Horn)*

Pro studii biologické dostupnosti cyclosporinu A byl zakoupen roztok této látky v metanolu. Deklarovaná koncentrace cyclosporinu A byla 20 ng/ml. Při HPLC analýzách byly naměřeny následující koncentrace. Test správnosti je třeba provést na hladině významnosti $\alpha = 0.05$. Obsahuje intervalový odhad střední hodnoty číslo 20 ng/ml?

Data: Koncentrace cyclosporinu A [ng/ml]: 19.96, 20.05, 20.00, 19.99, 20.01, 19.98, 20.00, 20.02, 20.01, 19.93.

Úloha B3.10 *Shodnost stanovených obsahů cyclosporinu dvěma laborantkami*
Na studii biologických dostupností pracují vždy dvě laborantky a je proto nezbytné ověřit, zda extrakce jimi prováděné poskytují stejné výsledky. Byly připraveny dva vzorky o rozdílné koncentraci a každá z laborantek provedla deset kompletních analýz každé koncentrace. První koncentrace byla provedena laborantkou (a) a (b), druhá koncentrace laborantkou (c) a (d). Shodnost výsledků je třeba otestovat testem shodnosti střední hodnoty na hladině významnosti $\alpha = 0.05$.

Data: Obsah cyclosporinu[mg. l^{-1}] ve vzorku:

1. koncentrace:										
a) *B310a*	1.5200	1.4900	1.5100	1.5250	1.4930	1.5090	1.5000	1.5050	1.4800	1.4820
b) *B310b*	1.5000	1.4900	1.5070	1.4930	1.4930	1.5090	1.5000	1.5050	1.4990	1.4970
2. koncentrace:										
c) *B310c*	9.8200	9.7800	9.7500	9.8000	9.8300	9.7700	9.7600	9.8000	9.7900	9.8100
d) *B310d*	9.8000	9.8100	9.7200	9.7700	9.8000	9.7400	9.7900	9.7700	9.7600	9.7800

Úloha B3.11 *Párový test stanovení obsahu cyclosporinu dvěma laborantkami*
Vzhledem k tomu, že na studii biologických dostupností pracují vždy dvě laborantky, je nezbytné ověřovat, zda extrakce jimi prováděné poskytují stejné výsledky. Bylo připraveno šest vzorků v celém rozsahu očekávaných koncentrací a každá z laborantek provedla kompletní analýzy. Výsledky je třeba vyšetřit párovým testem na hladině významnosti $\alpha = 0.05$.

Data: Obsah cyclosporinu [mg. l^{-1}] ve vzorku: *B311a* 1. laborantka, *B311b* 2. laborantka.

(a) *B311a*	2.538	1.390	0.612	0.306	0.161	0.080
(b) *B311b*	2.583	1.398	0.608	0.272	0.152	0.079

Úloha B3.12 *Správnost obsahu penicilinu v krvi metodou kapalinové chromatografie (Horn).* Vysokotlakou kapalinovou chromatografií byl v 7 měřeních stanoven obsah penicilinu v krvi. Vypočtete bodové odhady polohy a intervalové odhady Hornovou metodou pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení. Pracujte na hladině významnosti $\alpha = 0.05$. Obsahuje intervalový odhad střední hodnoty obsahu penicilinu číslo 2.20 mg. l^{-1}?

Data: Obsah penicilinu v krvi [mg. l^{-1}]: 2.20, 2.30, 2.50, 2.10, 2.30, 2.40, 2.50.

Úloha B3.13 *Správnost obsahu penicilinu v krvi u malého výběru metodou HPLC (Horn)*
Byla provedena analýza krve a byl stanoven obsah penicilinu. Vypočtěte bodové odhady polohy a intervalové odhady Hornovou metodou pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení. Vyšetření proveďte na hladině významnosti $\alpha = 0.05$ a ověřte předpokládaný obsah 0.50 mg. l^{-1}.

Data: Obsah penicilinu v krvi [mg. l^{-1}]: 0.50, 0.51, 0.48, 0.45, 0.48.

Úloha B3.14 *Test správnosti nalezeného obsahu penicilinu v krvi vůči deklarovanému (Horn).* Vzorek s deklarovaným obsahem penicilinu 2.4 mg. l^{-1} byl proměřen HPLC analýzou. Zjistěte, zda stanovená koncentrace odpovídá požadavku. Užijte intervalový odhad na hladině významnosti $\alpha = 0.05$.

Data: Obsah penicilinu v krvi [mg. l^{-1}]: 2.2, 2.3, 2.5, 2.1, 2.3, 2.4, 2.5

Úloha B3.15 *Shodnost obsahu penicilinu v krvi u dvou pacientů*
Porovnejte, zda výsledky analýz obsahu penicilinu v krvi u dvou pacientů A a B jsou shodné. Vyšetření proveďte na hladině významnosti $\alpha = 0.05$.
Data: Obsah penicilinu v krvi [mg. l^{-1}]: *B315a* pacient A, *B315b* pacient B.
B315a: 1.25, 1.27, 1.26, 1.28, 1.31, 1.30, 1.27, 1.29.
B315b: 1.24, 1.30, 1.26, 1.28, 1.25, 1.27, 1.26, 1.28.

Úloha B3.16 *Shodnost obsahu penicilinu v krvi dvojí detekcí*
Metodou HPLC byl stanoven obsah penicilinu v krvi. V prvním případě byla použita metoda (a) s detekcí při 254 nm a ve druhém případě (b) detekce při 200 nm. Testujte shodnost výsledků obou metod detekce na hladině významnosti $\alpha = 0.05$.
Data: Obsah penicilinu v krvi [mg. l^{-1}]: *B316a* detekce při 254 nm, *B316b* detekce při 200 nm.
B316a: 0.35, 0.37, 0.35, 0.35, 0.33, 0.34, 0.34, 0.36.
B316b: 0.36, 0.34, 0.34, 0.33, 0.35, 0.35, 0.36, 0.36.

Úloha B3.17 *Párový test obsahu penicilinu v krvi, určeného dvěma metodami*
Stanovení penicilinu v krvi bylo provedeno dvěma metodami, HPLC a spektrofotometricky. Zjistěte, zda oba postupy dávají stejné výsledky na hladině významnosti $\alpha = 0.05$.
Data: Obsah penicilinu v krvi [mg. l^{-1}]: *B317a* HPLC, *B317b* spektrofotometricky.
B317a: 6.58, 6.52, 6.57, 6.56, 6.56, 6.51, 6.61, 6.55.
B317b: 6.56, 6.50, 6.55, 6.54, 6.53, 6.49, 6.60, 6.57.

Úloha B3.18 *Porovnání biologického a polarografického stanovení inzulinu*
Účinnost 32 šarží inzulinu, antidiabetického proteohormonu z pankreatu, byla stanovena jednak biologicky a jednak polarograficky. Na hladině významnosti $\alpha = 0.05$ porovnejte výsledky biologické aktivity 1 mg vzorku a posuďte, vedou-li obě instrumentální metody ke stejným závěrům.
Data: Aktivita inzulinu [akt/mg] ve vzorku, *B318a*: biologicky, *B318b*: polarograficky.
B318a:

25.2	24.1	23.7	23.7	23.7	23.3	24.5	23.5	23.7
..	..	..	..	..	..	..	..	..
23.3	24.0	24.5	24.3	23.9				

B318b:

25.7	24.8	22.0	24.0	23.9	24.5	21.9	22.3	21.7
..	..	..	..	..	..	..	..	..
23.8	24.6	24.5	23.0	24.0				

Úloha B3.19 *Binomické rozdělení 200 vzorků pětic ryb na obsah rtuti*
200 vzorků, každý s obsahem 5 ryb z jednoho jezera, bylo testováno, zda obsah rtuti nepřekračuje povolenou hranici. Data obsahují četnost vzorků *f*, obsahujících ve vyšetřované pětici *X* kontaminovaných ryb (*f* může nabývat hodnot 1, 2, ..., 5). Sestrojte binomické rozdělení a určete jeho parametry.
Data: Číslo pětice ryb *X*, četnost vzorků *f*:

0	4,	1	43,	2	61,	3	56,	4	30,	5	6

Úloha B3.20 *Test správnosti a shodnosti obsahu těhotenského hormonu HCG v krvi*
Při analýze krve na obsah těhotenského hormonu choriogonadotropinu HCG metodou

enzymatické imunoanalýzy ELISA se denně provádí analýza kontrolního séra s certifikovanou hodnotou 13.60 ng/ml. Po dvou měsících práce, během nichž bylo toto stanovení provedeno 44krát, vystřídal laboranta B laborant A, který doposud provedl 24 stanovení. Vyšetřete, zda je mezi prací obou laborantů statisticky významný rozdíl. Vyšetření proveďte na hladině významnosti $\alpha = 0.05$.

Data: Obsah těhotěnského hormonu [ng/ml], *B320a*: laborant A, *B320b*: laborant B.

B320a:

11.37	14.58	8.60	16.54	13.22	14.78	13.60	13.42	15.98
..	..	..	..	..	..	..	..	..
18.45	9.540	11.16	14.40	15.84	10.24			

B320b:

14.07	17.19	13.84	16.30	13.84	18.13	14.96	10.40	12.91
..	..	..	..	..	..	..	..	..
8.620	14.97	15.82	13.41	17.12	13.01	14.14	15.33	

Úloha B3.21 *Střední hodnota poměru absorpčního maxima a minima spektra (Horn)*
U protinádorových injekcí Platidiam byla provedena zkouška čistoty pomocí spektrofotometrie a vypočten poměr absorbčního maxima a minima. K výpočtu střední hodnoty aplikujte Hornovu metodu pivotů a výsledky porovnejte se statistikami polohy.
Data: Poměr absorbčního maxima a minima spektra Platidiamu: 5.03, 5.04, 5.05, 5.10, 5.29, 5.30, 5.33.

Úloha B3.22 *Porovnání dvou metod stanovení obsahu účinné látky v Platidiamu*
U výrobku Platidiam (injekce 25 mg) byl stanoven obsah účinné látky (a) metodou kapalinové chromatografie a (b) spektrofotometricky. Určete, zda obě metody dávají shodné výsledky.
Data: Obsah účinné látky [mg] v Platidiamu, *B322a* metodou kapalinové chromatografie, *B322b* spektrofotometricky:
B322a: 25.79, 24.89, 25.38, 25.00, 25.65, 24.92, 24.95, 24.04, 24.75, 24.01, 24.39, 24.63, 24.86, 25.09, 25.27, 24.99.
B322b: 25.42, 25.00, 25.41, 25.35, 25.70, 24.65, 24.12, 24.24, 24.00, 24.15, 24.96, 24.85, 24.00, 26.07, 25.35, 25.1.

Úloha B3.23 *Test správnosti obsahu účinné látky v Platidiamu*
Spektrofotometricky byl stanoven obsah účinné látky u protinádorových injekcí Platidiam. Byl připraven roztok, který obsahoval 10 mg účinné látky a byl měřen dvanáckrát. Určete, zda byl obsah účinné látky stanoven správně.
Data: Obsah účinné látky [mg] v Platidiamu : 10.02, 9.98, 10.08, 9.89, 10.12, 10.08, 9.99, 9.99, 10.11, 10.02, 10.12, 10.06.

Úloha B3.24 *Shodnost obsahu chloridu sodného ve dvou šaržích Platidiamu*
U dvou šarží výrobku Platidiam 10 byl stanoven obsah chloridu sodného z deseti injekcí. Ověřte, zda jsou jejich průměry a rozptyly shodné na hladině významnosti $\alpha = 0.05$.
Data: Obsah chloridu sodného v šaržích injekcí Platidiamu [mg]: *B324a* je 1. šarže, *B324b* je 2. šarže.
B324a: 4.86, 4.98, 4.96, 4.90, 4.97, 4.80, 4.89, 4.81, 4.87, 4.90.
B324b: 4.87, 4.88, 4.86, 5.00, 5.08, 4.99, 5.03, 4.90, 5.11, 4.98.

Úloha B3.25 *Test správnosti a shodnosti stanovení léčiva v plazmě na dvou HPLC*

soustavách. V analytické laboratoři je obsah sledovaného léčiva v plazmě stanovován na dvou soustavách kapalinových chromatografů HPLC. Vyšetřete, zda obě soustavy poskytují správné a navzájem shodné výsledky, když bylo změřeno 20 vzorků krevní plazmy s obsahem léčiva 45 mg.

Data: Stanovené obsahy léčiva [mg]: *B325a* metodou HPLC1, *B325b* metodou HPLC2.
B325a: 45.36, 45.18, 44.96, 45.32, 44.89, 44.88, 45.36, 45.14, 44.90, 45.18, 45.16, 44.82, 45.23, 44.96, 45.28, 44.97, 45.14, 44.93, 45.29, 45.13.
B325b: 44.98, 45.20, 45.12, 45.02, 45.15, 44.98, 44.98, 45.13, 44.90, 45.09, 44.94, 45.17, 45.21, 45.24, 44.97, 44.96, 45.08, 45.07, 45.12, 45.07.

Úloha B3.26 *Střední hodnota koncentrace aminokyselin v séru (Horn)*
Metodou plynové chromatografie byly ve vzorku séra stanoveny koncentrace aminokyselin alaninu a glycinu. Porovnejte robustní odhady polohy a rozptýlení s pivotovými odhady Hornova postupu.

Data: *B326a:* koncentrace alaninu [mmol. l^{-1}], *B326b:* koncentrace glycinu [mmol. l^{-1}].
B326a: 415.7, 422.2, 415.8, 419.4, 470.8, 422.3, 421.8.
B326b: 329.8, 336.5, 327.9, 337.5, 395.0, 342.4, 335.1.

Úloha B3.27 *Test shodnosti dvou metod GC stanovení aminokyselin*
Vyšetřete shodnost výsledků koncentrace leucinu [mmol. l^{-1}] a glycinu [mmol. l^{-1}] v séru, které byly stanoveny dvěma rozličnými postupy plynové chromatografie.

Data: (a) koncentrace leucinu [mmol. l^{-1}]: metodou GC1 (*B327a*) a metodou GC2 (*B327b*). Data jsou ve dvojicích: 108.1 106.9, 107.5 109.8, 106.2 108.6, 106.5 111.2, 108.3 108.8, 107.3 109.9, 106.7 110.8, 108.3 107.2, 106.9 110.3, 107.8 107.8.
(b) koncentrace glycinu [mmol. l^{-1}]: metodou GC1 (*B327c*) a metodou GC2 (*B327d*). Data jsou ve dvojicích: 323.0 315.2, 316.8 335.8, 324.9 321.8, 314.2 328.6, 319.1 312.8, 323.8 316.8, 318.5 322.7, 327.0 318.9, 321.8 327.8, 317.6 332.8.

Úloha B3.28 *Párový test threoninu v séru metodami GC a IC*
Metodou plynové chromatografie GC a iontově výměnné chromatografie IC byla v sadě vzorků séra stanovena koncentrace aminokyseliny threoninu [mmol. l^{-1}]. Párovým testem vyšetřete, zda obě metody poskytují shodné výsledky.

Data: Koncentrace threoninu [mmol. l^{-1}] v séru, *B328a* metodou GC, *B328b* metodou IC. Data jsou ve dvojicích: 209 191, 179 185, 182 172, 168 157, 159 169, 198 169, 182 167, 191 168, 170 184, 179 158.

Úloha B3.29 *Test správnosti obsahu trimethoprimu v biseptolu*
Ke stanovení obsahu pomocné látky trimethoprimu v léčivu Biseptol byla užita metoda diferenční pulzní polarografie. K ověření byl analyzován standard se známou koncentrací trimethoprimu 57.66 μg. l^{-1}. Proveďte test správnosti 9 opakovaných měření.

Data: Koncentrace trimethoprimu [μg. l^{-1}] v biseptolu: 57.64, 56.11, 56.60, 56.23, 57.63, 56.31, 57.64, 54.51, 55.00.

3.5.2 Analýza chemických a fyzikálních dat

Úloha C3.01 *Průměrná hodnota pH (Horn)*
Hodnota pH byla poněkud kolísavá a byla proto změřena opakovaně. Z odečtených hodnot se má určit střední hodnota a 95% a 90% interval spolehlivosti. Aplikujte i Hornův postup.

Uvědomte si, že platí pH = - log $[H^+]$ a normální rozdělení bude mít koncentrace vodíkových iontů $[H^+] = 10^{-pH}$.
Data: Hodnoty pH: 5.12, 5.20, 5.15, 5.17, 5.16, 5.19, 5.15.

Úloha C3.02 *Střední hodnota poměru ploch píků na chromatogramu (Horn)*
Z poměru ploch reprodukovaných píků dvou chromatogramů v kapalinové chromatografii HPLC je třeba vyhodnotit střední hodnotu, parametr rozptýlení a 95% a 99% interval spolehlivosti střední hodnoty. Aplikujte i Hornův postup.
Data: Poměr ploch píků na chromatogramu : 0.2911, 0.2898, 0.2923, 0.3019, 0.2997, 0.2961, 0.2947, 0.2986, 0.2902, 0.2882.

Úloha C3.03 *Střední hodnota bodu ekvivalence v acidobazické titraci (Horn)*
Při opakované acidobazické titraci 10 ml 0.1M NaOH 0.1M roztokem silné kyseliny na barevný indikátor bylo zaznamenáno pět hodnot nalezeného bodu ekvivalence. Vypočtěte bodový a intervalový odhad střední hodnoty bodu ekvivalence na hladině významnosti $\alpha = 0.05$ a rozhodněte, zda lze indikovat systematickou chybu (str. 51 v cit.[14]). Aplikujte i Hornův postup.
Data: Bod ekvivalence [ml] v acidobazické *titraci* : 9.88, 10.18, 10.23, 10.39, 10.25.

Úloha C3.04 *Střední hodnota obsahu prvku v malém výběru (Horn)*
Opakovanou analýzou materiálu byl stanoven procentuální obsah prvku v malém výběru, (str. 36 v cit.[16]). Vyšetřete požadavky kladené na výběr a určete parametry polohy a rozptýlení. Které diagnostiky ukazují, že ve výběru jsou odlehlá měření? Je rozdělení symetrické? Jaké rozdělení prokazuje rankitový a kruhový graf? Na jaký tvar rozdělení ukazuje koeficient šikmosti a špičatosti? Aplikujte Hornův postup.
Data: Obsah prvku ve vzorku [%]: 0.0013, 0.0011, 0.0011, 0.0009, 0.0010, 0.0012, 0.0010, 0.0011, 0.0008, 0.0012.

Úloha C3.05 *Stanovení obsahu hydroxylu v organických sloučeninách (Horn)*
Byla vyvinuta nová metoda stanovení obsahu hydroxylu v organických sloučeninách. Hydroxyly byly stanoveny v ethylenglykolu a nonylfenolu. Určete parametry polohy a rozptýlení obsahu hydroxylu na hladině významnosti $\alpha = 0.05$ a 0.10. Aplikujte Hornův postup.
Data: Obsah hydroxylu v: *C305a* v ethylenglykolu, *C305b* v nonylfenolu.
C305a: 1767, 1767.9, 1798.0, 1818.1, 1783.0, 1716.1, 1782.0, 1782.7, 1805.4, 1776.2.
C305b: 248.8, 243.8, 261.8, 250.1, 248.0, 245.0, 246.7, 249.3, 246.9, 244.3.

Úloha C3.06 *Test správnosti nalezené atomové hmotnosti kadmia (Horn)*
Byla získána data pro výpočet atomové hmotnosti kadmia. Vyšetřete předpoklady o výběru dat. Pocházejí data ze souboru, který vykazuje Gaussovo rozdělení? U kolika grafických diagnostik je shoda v indikaci odlehlých bodů? Určete odhad střední hodnoty atomové hmotnosti a parametry rozptýlení. Intervalové odhady vypočtěte na hladině významnosti $\alpha = 0.05$. Aplikujte Hornův postup. Souhlasí nalezená střední hodnota s hodnotou 112.41? (str. 250 v cit.[17])
Data: Atomová hmotnost kadmia [g]: 112.23, 112.34, 112.30, 112.22, 112.32, 112.34.

Úloha C3.07 *Test správnosti nalezeného obsahu bizmutu fotometrickou mikrotitrací (Horn)*

Fotometrickou, chelatometrickou mikrotitrací bizmutitých iontů kyselinou ethylen-diamintetraoctovou EDTA bylo v kyselém prostředí pH = 1 získáno 14 hodnot obsahu bizmutu v mg. Teoretický obsah je 1.67 mg. Aplikujte Hornův postup. Ovlivňují odlehlé hodnoty významně parametry polohy a rozptýlení? Je titrační stanovení zatíženo soustavnou chybou?

Data: Obsah bizmutu [mg] fotometrickou mikrotitrací : 1.65, 1.65, 1.67, 1.64, 1.67, 1.70, 1.69, 1.67, 1.62, 1.65, 1.70, 1.63, 1.63, 1.66.

Úloha C3.08 *Test správnosti kalibrované mikrobyrety (Horn)*

Určete parametry polohy a rozptýlení a vyjádřete přesnost mikrobyrety, jestliže procedura kalibrace mikrobyrety o obsahu 1250.0 µl byla opakována 10krát. Aplikujte Hornův postup. Test správnosti proveďte na hladině významnosti $\alpha = 0.05$.

Data: Objem vytlačené kapaliny [µl]: 1249.8, 1249.9, 1250.0, 1250.1, 1250.1, 1250.0, 1250.0, 1250.1, 1249.9, 1250.0.

Úloha C3.09 *Stanovení obsahu nitroglycerinu v nitroglycerinové laktóze (Horn)*

Z jedné výrobní šarže nitroglycerinové laktózy bylo odebráno šest vzorků, ve kterých byl metodou kapalinové chromatografie stanoven obsah nitroglycerinu. Aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení. Je rozdělení výběru symetrické? Intervalový odhad vypočtěte na hladině významnosti $\alpha = 0.05$.

Data: Obsah nitroglycerinu [%] v nitroglycerinové laktóze : 10.05, 10.03, 10.07, 10.02, 10.05, 10.25.

Úloha C3.10 *Určení střední hodnoty efektivního průměru dispergovaných částic (Horn)*

Analýzou vzorku vodné pigmentové disperze byl na přístroji Particle Sizer BI-90 určen efektivní průměr dispergovaných částic pigmentu. Úkolem je určit střední hodnotu efektivního průměru částice a parametr rozptýlení. Aplikujte Hornův postup. Jsou v datech odhlehlé hodnoty? Je třeba užít mocninnou transformaci nebo užijete robustní odhady?

Data: Průměr částice [nm]: 219, 213, 203, 237, 230, 214, 228, 232, 241, 214.

Úloha C3.11 *Test správnosti stanovené koncentrace tenzidů (Horn)*

Standardní vzorek obsahuje 2.5 mg. l^{-1} anionaktivních tenzidů. Testujte, zda kontrolní hodnoty koncentrace standardu jsou správné. Jde o symetrické rozdělení? Vyšetřete požadavky kladené na výběr a určete parametry polohy a rozptýlení. Které diagnostiky ukazují, že ve výběru jsou odlehlá měření? Jaké rozdělení prokazuje kruhový graf? Na jaký tvar rozdělení ukazuje koeficient šikmosti a špičatosti? Aplikujte i Hornův postup.

Data: Koncentrace anionaktivních tenzidů [mg. l^{-1}]: 2.36, 2.40, 2.48, 2.50, 2.57, 2.62, 2.68.

Úloha C3.12 *Střední hodnota obsahu Cleve 1,7-kyseliny ve vzorku Cyanolu (Horn)*

Ve vzorku Cyanolu (1-amino-7-naftol) byl stanoven obsah Cleve 1,7-kyseliny, která je surovinou pro jeho výrobu. Metodou vysokoúčinné kapalinové chromatografie bylo získáno 10 hodnot. Úkolem je zjistit bodové a intervalové odhady obsahu dané nečistoty ve vzorku Hornovou metodou pivotů a porovnat je s klasickými a robustními statistikami. Jsou ve výběru odlehlé hodnoty? Dáte přednost robustním odhadům nebo mocninné transformaci? Vyšetření proveďte na hladině významnosti $\alpha = 0.05$.

Data: Obsah Cleve 1,7 kyseliny [%]: 1.060, 1.098, 0.9930, 1.085, 1.050, 1.049, 1.117, 1.010, 0.981, 1.067.

Úloha C3.13 *Určení střední hodnoty pH (Horn)*
Průměrná hodnota pH byla kolísavá a byla proto změřena opakovaně. Na získaných datech aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení. Intervalový odhad vyčíslete na hladině významnosti $\alpha = 0.05$. Uvědomte si, že platí $pH = - \log [H^+]$ a normální rozdělení bude mít koncentrace vodíkových iontů $[H^+] = 10^{-pH}$.
Data: pH: 8.71, 8.72, 8.75, 8.88, 8.89, 8.92, 8.92, 8.94, 8.94, 8.95, 8.97.

Úloha C3.14 *Zjištění homogenity látky CURSATE (Horn)*
Pro zjištění homogenity látky zvané CURSATE se odebírá z obalu deset vzorků. Tyto se pak analyzují pomocí infračervené spektrofotometrie. Ověřte základní předpoklady kladené na výběr. Aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení. Je rozdělení výběru symetrické?
Data: Obsah látky CURZATE [%]: 3.998, 4.012, 4.026, 4.026, 4.062, 4.078, 4.083, 4.094, 4.122, 4.151.

Úloha C3.15 *Střední hodnota pevnosti lepeného spoje (Horn)*
Měření pevnosti lepeného spoje se provádí na 9 tělíscích zhotovených z jedné překližované desky. Na výsledky měření tlaku v MPa aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení. Intervalový odhad vyčíslete na hladině významnosti $\alpha = 0.05$.
Data: Pevnost lepeného spoje [MPa]: 2.56, 2.48, 2.40, 1.76, 2.12, 2.56, 2.48, 2.36, 2.20.

Úloha C3.16 *Určení obsahu aminoantrachinonsulfonové kyseliny (Horn)*
Byl stanoven procentuální obsah aminoantrachinonsulfonové kyseliny v kyselině bromaminové. Jsou prvky výběru vybrány náhodně, není mezi nimi skrytá závislost? Pocházejí data ze souboru, který vykazuje Gaussovo rozdělení? U kolika diagnostik je shoda v indikaci odlehlých bodů? Určete střední hodnotu a její 95% interval spolehlivosti. Aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení. Jakým odhadům dáte přednost?
Data: Obsah aminoantrachinonsulfokyseliny *x* [%] v kyselině bromaminové: 0.35, 0.43, 0.36, 0.33, 0.38, 0.30, 0.48, 0.31, 0.35.

Úloha C3.17 *Test správnosti stanovovaného obsahu g-butyrobetainu (Horn)*
Obsah g-butyrobetainu se stanovuje metodou HPLC za použití iontové kolony a konduktometrického detektoru. Pro vyhodnocení integrace se používá metoda externího standardu o obsahu 50.0 g. Určete parametry polohy a rozptýlení využitím Hornova postupu malých výběrů a porovnejte je s klasickými a robustními odhady. Je stanovení zatíženo systematickou chybou?
Data: Obsah g-butyrobetainu [%]: 47.66, 48.73, 49.00, 49.97, 49.99, 51.94.

Úloha C3.18 *Test shodnosti dvou analytických metod - validace nové metody*
Ve vzorcích byl stanovován obsah etylesteru kyseliny octové dvěma metodami, a to jednak zmýdelněním a následnou titrací, a jednak novou metodou plynové chromatografie. Na hladině významnosti $\alpha = 0.05$ rozhodněte, zda obě metody poskytují srovnatelné výsledky.
Data: Obsah ethylesteru kyseliny octové [%]: *C318a* stanovený titračně a *C318b* plynovou chromatografií. Data

jsou ve dvojicích: 99.73 99.89, 99.65 99.92, 99.56 99.94, 99.71 99.97, 99.60 99.96, 99.75 99.97, 99.66 99.97, 99.53 99.96, 99.80 99.96, 99.77 99.93, 99.72 99.94, 99.82 99.98.

Úloha C3.19 *Test správnosti obsahu nitroglycerinu vůči normě (Horn)*
Obsah nitroglycerinu ve 2% nitroglycerinovém roztoku v lihu pro farmaceutické účely má být podle normy 1.8 - 2.2 %. Vyšetřete požadavky kladené na výběr a určete parametry polohy a rozptýlení. Které diagnostiky ukazují, že ve výběru jsou odlehlá měření? Je rozdělení symetrické? Na jaký tvar rozdělení ukazuje koeficient šikmosti a špičatosti? Vyšetřete, zda-li naměřené obsahy odpovídají normě. Využijte mocninnou transformaci. Užijte také Hornův postup.
Data: Obsah nitroglycerinu [%] v lihu pro farmaceutické účely : 2.01, 2.04, 2.03, 2.05, 2.06, 2.04, 2.04, 2.04, 2.05, 2.07.

Úloha C3.20 *Test správnosti obsahu naftolu vůči normě*
Obsah naftolu AS byl v dodávce stanoven spektrofotometricky. Podle normy je vyhovující vzorek takový, který obsahuje minimálně 94% stanovované látky. Vyšetřete předpoklady náhodného výběru. Ověřte, zda tato dodávka vyhovuje normě. Ověření proveďte na hladině významnosti $\alpha = 0.05$.
Data: Obsah naftolu AS [%]: 94.1, 93.4, 94.1, 94.9, 92.6, 94.3, 93.9, 93.3, 94.0, 94.4, 93.6, 93.3, 94.6, 95.0, 94.7, 93.5.

Úloha C3.21 *Porovnání dvou analytických metod stanovení obsahu THI*
Obsah látky THI v přípravku byl stanoven dvěma různými metodami. Ověřte, zda oba výběry pocházejí z téhož rozdělení, které je normálního charakteru.
Data: Obsah látky THI [g]: *C321a* 1. výběr, *C322b* 2. výběr:
C321a: 3.27 3.16 2.92 3.53 2.92 3.08 3.20 3.32 2.95 3.28 3.40 3.34 3.08 3.26 3.51 3.16 3.12 3.07.
C321b: 3.27 3.14 3.01 3.76 3.78 3.05 3.33 3.38 3.02 3.21 3.27 3.27 3.15 2.98 3.01 3.02 3.28 3.62.

Úloha C3.22 *Test správnosti a shodnosti obsahu síranů určených turbidimetricky a izotachoforeticky*
Standardní roztok síranů o obsahu 40.0 mg. l^{-1} SO_4^{2-} byl v delším časovém období analyzován dvěma metodami, turbidimetricky a izotachoforeticky. Pomocí intervalového odhadu a testů shody středních hodnot rozhodněte, zda obě metody poskytují stejné a správné výsledky. Vyšetření proveďte na hladině významnosti $\alpha = 0.05$.
Data: Obsah síranů [mg. l^{-1}], *C322a* turbidimetrie, *C322b* izotachoforéza:
C322a:, 47.0, 46.0, 47.0, 46.8, 51.8, 50.7, 51.4.
C322b:, 39.0, 41.0, 48.0, 49.0, 49.0, 52.0, 46.0, 46.0, 46.0, 47.0.

Úloha C3.23 *Test shodnosti hodnot efektivního průměru vzorku se standardem*
Na přístroji Particle Sizer BI-90 byly stanoveny hodnoty efektivních průměrů dispergovaných částic pigmentu typového vzorku Versanylu červeného 2241 (standard) a analyzovaného vzorku (vzorek). Menší hodnota efektivního průměru značí dokonalejší dispergaci pigmentu a tím i lepší koloristickou vydatnost barviva. Úkolem je zjistit, zda hodnota efektivního průměru vzorku je shodná nebo menší než u standardu. Test shodnosti proveďte na hladině významnosti $\alpha = 0.05$.
Data: Efektivní průměr částic pigmentů [nm]: *C323a* standard, *C323b* vzorek. Data jsou ve dvojicích: 233 232, 230 225, 227 221, 222 228, 218 223, 232 222, 227 219, 228 230, 228 231, 236 222, 229 228, 230 239, 231 239, 244 232, 240 230, 225 235, 237 222, 236 229, 230 220, 240 222, 237 225, 240 227.

Úloha C3.24 *Test shodnosti obsahu ergosterolu v kvasnicích metodami HPLC*
Pro provedení analýzy v kapalinové chromatografii lze použít metodu s interním a externím standardem. Obě metody jsou porovnány při stanovení obsahu ergosterolu ve vzorku kvasnic. Vedou ke stejným výsledkům? Test shodnosti proveďte na hladině významnosti $\alpha = 0.05$.

Data: Obsah ergosterolu v kvasnicích [mg. l^{-1}]: *C324a* interním a *C324b* externím standardem. Data jsou ve dvojicích:

2.502	2.500,	1.520	1.756,	1.420	1.258,	1.295	1.292,
..	..	..	..	..	..	..	..
0.837	0.875,	1.090	0.873,				

Úloha C3.25 *Porovnání obsahu zinku v pesticidním přípravku Neroxonu 50 s normou*
Stanovte jodometricky obsah zinku ve vzorku pesticidního přípravku Neroxon 50 [hm.%] ze zkušebny po rozkladu vzorku kyselinou chlorovodíkovou na sirovodík a jeho zachycení v metanolickém hydroxidu draselném. Jsou v datech odlehlé hodnoty? Užijete robustní odhady nebo mocninnou transformaci? Ověřte, zda střední hodnota leží v intervalu normy ±23.5 [hm. %], 26.5 [hm. %],. Vyšetření proveďte na hladině významnosti $\alpha = 0.05$.

Data: Obsah zinku [hm. %] v pesticidním přípravku Neroxonu 50 :

24.69	25.26	25.90	24.70	24.64	24.44	26.50	25.42	25.27	24.68
..	..	..	..	..	..	..	..	..	..
24.40	23.60	23.50	23.50	23.70	25.50	23.50	26.50	26.50	

Úloha C3.26 *Test shodnosti obsahu desmediphamu určeného metodami HPTLC a HPLC*
Stanovte obsah desmediphamu [hm.%] v přípravku Synbetan D metodami HPTLC a HPLC a ověřte, zda obě metody dávají stejné výsledky. Jsou oba výběry symetrického rozdělení? Mají oba výběry stejný rozptyl?

Data: Obsah desmediphanu [hm.%]: *C326a* metodou HPTLC, *C326b* metodou HPLC.
C326a: 14.81, 14.54, 15.15, 14.59, 15.39, 15.24, 15.24, 14.74, 13.27, 15.00.
C326b: 13.20, 13.61, 15.11, 15.10, 15.01, 13.73, 14.85, 15.30, 15.10, 14.93.

Úloha C3.27 *Test správnosti obsahu chlordiazonu v přípravku* BUREX 80
Stanovení obsahu chlordiazonu v přípravku BUREX 80 bylo provedeno plynovou chromatografií. Správnost výsledku porovnáme s normou státní zkušebny pro 1. jakostní kategorii, $\mu = 80$ mg na hladině významnosti $\alpha = 0.05$. Je stanovení správné? Je rozdělení výběru symetrické a bez odlehlých hodnot?

Data: Obsah chlordiazonu [mg] v přípravku BUREX 80:

79.70	80.80	80.30	80.50	81.00	79.90	80.50	80.90
..	..	..	..	..	..	..	..
80.30	80.10	80.20	81.40	80.70	81.20		

Úloha C3.28 *Shodnost obsahu fenitrithionu v METATIONu dvěma metodami*
Stanovení obsahu fenitrithionu v METATIONu se provádí plynovou chromatografií a spektrofotometricky. Obě stanovení byla provedena u téhož vzorku s obsahem $\mu = 50$ mg. K posouzení správnosti použijte intervalového odhadu na hladině významnosti $\alpha = 0.05$. Jsou obě stanovení shodná a správná? Jsou rozdělení obou výběrů symetrická a bez odlehlých hodnot?

Data: Obsah fenitrithionu [mg] v přípravku METATION:
1. výběr:

49.82	50.22	50.32	50.47	49.91	50.23	50.36	50.50	50.00
50.25	50.39	50.53	50.10	50.29	50.40	50.57	50.13	50.30
50.40	50.72	50.19	50.30	50.43	50.92			

2. výběr:

50.24	50.73	50.83	50.92	50.37	50.73	50.85	50.98	50.47
50.76	50.87	51.13	50.67	50.77	50.88	51.23	50.68	50.80
50.90	51.41	50.70	50.80	50.91	51.81			

Úloha C3.29 *Test shodnosti argentometrie a merkurimetrie při stanovení chloridů*
Pro stanovení chloridů ve vodě lze použít dvou titračních metod, argentometrické a merkurimetrické. Aplikujte test shodnosti dvou výsledků na hladině významnosti $\alpha = 0.05$. Jsou rozptyly obou rozdělení stejné a pocházejí z Gaussova rozdělení?
Data: Obsah chloridů [mg]: *C329a* argentometrie, *C329b* merkurimetrie.
C329a: 75.9, 75.9, 75.6, 75.6, 75.7, 75.9, 75.9, 76.1, 76.0, 75.8, 75.8.
C329b: 76.0, 76.0, 75.9, 75.7, 75.9, 76.1, 76.3, 75.8, 76.0, 75.9, 76.1.

Úloha C3.30 *Test správnosti a shodnosti obsahu artrazinu metodou GC a argentometricky*
Bylo provedeno stanovení účinné látky artrazinu v přípravku Zeazin S 40 o obsahu 40 mg, a to plynovou chromatografií GC a argentometrickou titrací Ag. Porovnejte výsledky na hladině významnosti $\alpha = 0.05$. Mají obě rozdělení stejné rozptyly? Jsou obě stanovení správná a shodná?
Data: Obsah artrazinu [mg] v přípravku Zeazin S 40 : *C330a* GC, *C330b* metodou argentometricky.
C330a: 39.72, 39.79, 39.86, 39.91, 39.94, 39.97, 40.02, 40.04, 40.05, 40.06, 40.10, 40.12, 40.10, 40.15, 40.21, 40.24, 40.18, 40.18, 40.26, 40.29, 40.32, 40.43, 40.60, 40.10, 40.18.
C330b: 40.19, 40.56, 40.64, 40.73, 41.45, 40.30, 40.58, 41.00, 40.66, 40.74, 40.38, 40.56, 40.68, 40.78, 40.42, 40.60, 40.70, 40.90, 40.48, 40.62, 40.70, 41.00, 40.54, 40.64, 40.72.

Úloha C3.31 *Test shodnosti výhřevnosti odpadu ze dvou typů zástavby města*
V rámci úvah o způsobu zpracování tuhého komunálního odpadu byl prozkoumán obsah 24 náhodně vybraných kontejnerů z centrální zástavby s převážně ústředním topením a 28 náhodně vybraných kontejnerů ze smíšené zástavby s lokálním i ústředním topením. Předmětem průzkumu je posoudit rozdíl ve výhřevnosti odpadu získaného v obou typech zástavby. Vyšetření proveďte na hladině významnosti $\alpha = 0.05$.
Data: Výhřevnost odpadu [kJ/kg]: *C331a* v centrální zástavbě, *C331b* ve smíšená zástavbě.
C331a:, 906, 1208, 906, 844, 1233, 1316, 919, 951, 992, 1143, 1135, 942, 671, 892, 844, 703, 660, 673, 927, 786, 1296, 917, 1004, 556.
C331b:, 966, 1071, 683, 987, 1002, 709, 877, 588, 957, 795, 643, 702, 518, 765, 631, 719, 698, 770, 632, 802, 857, 865, 612, 670, 758, 696, 711, 637.

Úloha C3.32 *Test shodnosti výsledků dvou instrumentálních metod*
Při stanovení obsahu organických zásad ve frakcích z destilace kyseliny karbolové bylo užíváno pracné chromatografické metody, kdy obsah jednotlivých zásad byl postupně sčítán. Rychlejší je stanovení organických zásad potenciometrickou titrací v nevodném prostředí. Celkem 10 vzorků frakcí z destilace bylo změřeno oběma metodami na hladině významnosti $\alpha = 0.05$. Určete, zda obě metody vedou ke shodným výsledkům a zda je možné nahradit chromatografickou metodu titračním stanovením.

Data: Obsah organických zásad ve frakcích z destilace kyseliny karbolové: *C332a*, chromatografie, *C332b*, potenciometrická titrace.
C332a: 0.40, 1.49, 0.25, 2.60, 0.45, 3.50, 0.52, 0.17, 0.39, 1.92.
C332b: 0.35, 1.92, 0.28, 2.25, 0.39, 3.37, 0.45, 0.17, 3.37, 1.78.

Úloha C3.33 *Test správnosti obsahu dusíku v p-nitroanilinu*
Standard p-nitroanilin obsahující 20.29 % dusíku byl podroben elementární analýze. Bylo získáno 15 výsledků měření. Proveďte test správnosti výsledku vůči obsahu deklarovanému ve standardu využitím intervalového odhadu. Je rozdělení výběru symetrické?
Data: Obsah dusíku v p-nitroanilinu [%]: 20.315, 20.106, 20.268, 20.256, 20.224, 20.217, 20.276, 20.223, 20.122, 20.288, 20.368, 20.330, 20.296, 20.400, 20.261.

Úloha C3.34 *Test správnosti obsahu amonných iontů vůči normě* (Horn)
Byl stanoven obsah amonných iontů ve standardním vzorku (mg. l^{-1}). Vyšetřete předpoklady o výběru. Jsou prvky výběru vybrány náhodně, není mezi nimi skrytá závislost? Pochází data ze souboru, který vykazuje Gaussovo rozdělení? U kolika diagnostik je shoda v indikaci odlehlých bodů? Zkonstruujte bariérově-číslicové schéma formou sedmipís-menového zápisu výběru. Na data aplikujte test správnosti výsledku vůči normě čili test shody s normou $\mu = 0.500$. Aplikujte i Hornův postup pro malé výběry.
Data: Obsah amonných iontů [mg. l^{-1}]: 0.505, 0.503, 0.502, 0.500, 0.500, 0.499, 0.502, 0.504, 0.502, 0.501.

Úloha C3.35 *Stanovení smykového napětí pevnosti lepených spojů* (Horn)
U určitého typu lepidla na dřevo je zkoumána pevnost lepených spojů ve smyku. Zkouška spočívá ve zvyšování smykového napětí tlakem na slepené dřevěné destičky. Měření bylo provedeno na 8 vzorcích. Proveďte odhady parametrů polohy a rozptýlení Hornovou metodou a výsledky porovnejte s klasickými a robustními odhady parametrů.
Data: Tlak smykového napětí [Mpa]: 3.2, 5.5, 6.2, 7.2, 7.8, 8.6, 8.7, 9.2.

Úloha C3.36 *Stanovení střední hodnoty obsahu vody ve výrobku LAV 27 % N (Horn)*
Stanovte procentuální obsah vody ve výrobku LAV 27 % N z týdenního sběru náhodných vzorků. Nalezněte odhad střední hodnoty. Je nutné užít transformaci dat? Je rozdělení symetrické a bez odlehlých hodnot?
Data: Obsah vody [%] ve výrobku LAV 27 % N: 0.12, 0.13, 0.14, 0.13, 0.11, 0.11, 0.11, 0.09, 0.10, 0.14, 0.12, 0.17, 0.12, 0.11, 0.11.

Úloha C3.37 *Stanovení střední hodnoty obsahu dusíku ve výrobku LAV 27 % N (Horn)*
Stanovte procentuální obsah dusíku ve výrobku LAV 27 % N z týdenního sběru náhodných vzorků. Jsou v datech odlehlé hodnoty? Nalezněte odhad střední hodnoty. Dáte přednost klasickým nebo robustním odhadům? Je nutné užít transformaci dat? Vyšetření proveďte na hladině významnosti $\alpha = 0.05$.
Data: Obsah dusíku [%] ve výrobku LAV 27 % N: 27.0, 27.1, 27.0, 26.9, 26.9, 27.1, 27.0, 26.8, 26.8, 26.8, 26.7, 26.7, 26.8, 26.8, 26.8.

Úloha C3.38 *Test správnosti a kalibrace pipety (Horn)*
Pipeta o objemu 10 ml byla kalibrována metodou vážení odpipetované vody. Bylo získáno 10 hodnot. Vyšetřete požadavky kladené na výběr a určete parametry polohy a rozptýlení. Které diagnostiky ukazují, že ve výběru jsou odlehlá měření? Je rozdělení symetrické? Na jaký tvar rozdělení ukazuje koeficient šikmosti a špičatosti? Určete intervalový odhad

skutečného objemu pipety a testujte správnost na hladině významnosti $\alpha = 0.05$. Užijte také Hornův postup.

Data: Objem pipety [ml]: 9.9993, 9.9954, 10.0044, 9.9924, 10.0044, 10.0017, 10.0022, 10.0036, 10.0037, 10.0004.

Úloha C3.39 *Test shodnosti obsahu vody v německé a ruské draselné soli*

Porovnejte procentuální obsah vody v německé a ruské draselné soli. Jsou obě rozdělení Gaussovská a se stejným rozptylem? Je nutné užít transformaci dat k odhadu střední hodnoty? Test shodnosti proveďte na hladině významnosti $\alpha = 0.05$.

Data: Obsah vody [%]: *C339a*: německá sůl, *C339b*: ruská sůl,
C339a: 0.20, 0.18, 0.20, 0.20, 0.20, 0.23, 0.26, 0.29.
C339b: 0.31, 0.39, 0.25, 0.21, 0.24, 0.56.

Úloha C3.40 *Test shodnosti obsahu K_2O v německé a ruské draselné soli*

Porovnejte procentuální obsah K_2O v německé a ruské draselné soli. Jsou rozdělení Gaussovská s konstantním rozptylem? Je nutné užít transformaci dat k odhadu střední hodnoty? Test shodnosti proveďte na hladině významnosti $\alpha = 0.05$.

Data: Obsah K_2O [%]: *C340a*: německá sůl, *C340b*: ruská sůl.
C340a: 60.48, 60.39, 60.31, 60.84, 60.67, 60.33, 60.51, 60.51.
C340b: 60.61, 60.24, 60.57, 60.82, 60.17, 60.57.

Úloha C3.41 *Test shodnosti titračního a fotometrického stanovení dusíku*

Srovnejte titrační a fotometrickou metodu stanovení obsahu dusíku ve výrobku a určete, zda výsledky získané jednotlivými metodami jsou ekvivalentní. Odpovídá vzorek, změřený metodou titrační, dotyčnému vzorku fotometrickou metodou? Mají oba výběry shodné rozptyly? Test shodnosti proveďte na hladině významnosti $\alpha = 0.05$.

Data: Obsah dusíku N [%]: *C341a*: titračně, *C341b*: fotometricky.
C341a: 29.77, 30.34, 30.59, 30.47, 30.89, 30.31, 30.27, 29.62, 30.42, 30.51, 29.72, 30.17, 30.48, 30.74, 29.79, 30.73, 30.63, 30.57, 30.24, 30.92.
C341b: 29.33, 30.12, 30.65, 30.52, 31.10, 30.15, 30.13, 29.37, 30.02, 30.10, 29.80, 30.74, 30.83, 31.17, 29.67, 30.53, 30.18, 30.65, 29.90, 30.48.

Úloha C3.42 *Test shodnosti obsahu kovu ve dvou vzorcích organického barviva*

V organických barvivech se sleduje obsah kovů. Ve dvou vzorcích téhož organického barviva bylo provedeno (a, b) deset stanovení železa a (c, d) deset stanovení manganu. Na získaná data byl aplikován test shodnosti výsledků. Na hladině významnosti $\alpha = 0.05$ zjistěte, zda jsou obsahy železa a manganu v obou vzorcích shodné. Mají oba výběry stejný rozptyl a vyhovují Gaussovu rozdělení?

Data: Obsah Fe [ppm] v organických barvivech: *C342a*: 1. vzorek, *C342b*: 2. vzorek,
C342a: 37.389, 38.245, 36.533, 37.941, 36.724, 37.552, 36.947, 38.101, 36.661, 37.812.
C342b: 38.105, 37.165, 38.002, 37.594, 37.944, 37.215, 37.717, 37.229, 37.639, 38.093.
Obsah Mn [ppm] v organických barvivech: *C342c*: 1. vzorek, *C342d*: 2. vzorek,
C342c: 0.408, 0.308, 0.358, 0.400, 0.360, 0.312, 0.399, 0.343, 0.359, 0.370.
C342d: 0.384, 0.363, 0.342, 0.372, 0.355, 0.369, 0.350, 0.377, 0.349, 0.361.

Úloha C3.43 *Párový test shody bodu ekvivalence normální a automatickou byretou*

K porovnání bodů ekvivalence titrací normální byretou a automatickou digitální byretou byly použity modelové vzorky různých koncentrací HCl. Aplikujte párový test na dvojice

bodů ekvivalence obou metod na hladině významnosti $\alpha = 0.05$.

Data: Bod ekvivalence, *C343a* normální byretou, *C343b* automatickou byretou [ml]. Data jsou ve dvojicích: 6.70 6.62, 8.20 8.28, 10.3 10.41, 12.2 12.22, 15.6 15.61, 15.7 15.62 18.1 18.11, 20.2 20.14, 20.3 20.21.

Úloha C3.44 *Určení obsahu tetrahydrofuranu v destilátu s toluenem (Horn)*
Při výrobě Watermelonketonu se zjišťuje poměrné zastoupení tetrahydrofuranu v destilátu s toluenem. Pro další zpracování je potřebné množství tetrahydrofuranu v intervalu 20 - 30 %. Aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení.

Data: Obsah tetrahydrofuranu [%] v destilátu s toluenem : 29, 27, 26, 28, 28, 28, 25, 26, 27, 27.

Úloha C3.45 *Test správnosti obsahu 4-methylkatecholu metodou GC s předpisem*
Jako výchozí surovina při výrobě Watermelon-ketonu je používán 4-methylkatechol. Dodavatel uvádí v předpisu jeho obsah 94 - 97 %. Měření obsahu byla provedena u několika šarží metodou GC. Na data aplikujte test správnosti výsledků pomocí intervalu spolehlivosti míry polohy na hladině významnosti $\alpha = 0.05$.

Data: Obsah 4-methylkatecholu [%] ve výchozí surovině: 95.2, 94.1, 94.0, 95.1, 94.1, 94.7, 94.2, 93.9, 94.8, 95.1, 96.1, 95.3, 95.1, 96.0, 94.6.

Úloha C3.46 *Určení obsahu kyseliny 4-methylkatecholdioctové*
Jako výchozí surovina při výrobě Watermelon-ketonu se používá kyselina 4-methyl-katecholdioctová s obsahem, který se má pohybovat v intervalu 93 - 100 %. Aplikujte Hornů postup k vyčíslení parametrů polohy na hladině významnosti $\alpha = 0.05$.

Data: Obsah 4-methylkatecholdioctové kyseliny [%]: 94.2 96.1 93.2 95.0 93.2 93.2 95.2 94.5 94.0 95.0.

Úloha C3.47 *Test shodnosti obsahu Watermelon-ketonu metodou HPLC a GC*
K zjištění obsahu Watermelonketonu ve vzorcích byla použita kapalinová chromatografie (HPLC) i plynová chromatografie (GC). Na hladině významnosti $\alpha = 0.05$ aplikujte test shodnosti výsledků. Je rozptyl obou rozdělení shodný?

Data: Obsah Watermelonketonu [%] ve vzorku, *C347a*: metoda HPLC, *C347b*: metoda GC.
Data jsou ve dvojicích: 99.12 99.44, 99.07 99.20, 99.73 99.09, 99.79 98.89, 99.14 98.96, 99.01 98.51, 99.02 99.70, 99.00 99.54, 98.91 99.20, 98.86 99.05.

Úloha C3.48 *Stanovení pěnivosti piva na přístroji NIBEM (Horn)*
Na přístroji NIBEM firmy Haffmans byla stanovena pěnivost vzorku piva a bylo získáno šest hodnot. Stanovte 95% interval spolehlivosti pro střední hodnotu pomocí Hornova postupu a porovnejte ji s výsledkem klasického postupu analýzy jednorozměrných dat.

Data: Pěnivost piva x [s/30 mm]: 215, 227, 321, 241, 238, 235.

Úloha C3.49 *Test shodnosti obsahu extraktu mladiny a alkoholu v pivu klasickou destilační metodou a analyzátorem SCABA*. Ve čtyřiceti různých pivech a ležácích byl stanoven obsah extraktu původní mladiny časově náročnou klasickou destilační metodou a rychlou metodou na automatickém analyzátoru SCABA. Porovnejte výsledky obou metod. Dále vyšetřete náhradu klasické metody rychlometodou pro případ stanovení alkoholu v nealkoholickém pivu, když bylo provedeno deset paralelních stanovení na jednom vzorku piva. Posuďte, zda je možno nahradit klasickou destilační metodu automatickým analyzátorem SCABA.

Data: Obsah mladiny v pivu: *C349a*: výběr A

9.58	11.630	12.060	11.670	9.97	11.990	8.090	10.040	9.840	11.870
11.790	9.180	11.870	11.750	11.65	9.000	12.000	11.530	11.820	10.1
12.030	11.920	11.770	9.840	11.930	11.990	11.710	13.010	11.920	10.070
11.86	11.090	11.870	11.580	11.880	9.250	11.850	12.020	11.810	12.150

Obsah mladiny v pivu: *C349b*: výběr B

9.62	11.650	12.110	11.730	9.970	11.900	8.040	10.040	9.880	11.83
11.680	9.090	11.770	11.750	11.640	8.940	12.130	11.560	11.660	10.24
11.910	11.830	11.800	9.920	11.960	12.090	11.810	13.130	11.940	10.070
11.81	11.100	11.920	11.540	11.900	9.270	11.940	11.970	11.810	12.1

Obsah alkoholu v pivu: *C349c*: výběr A

3.91	3.920	3.910	3.87	3.92	3.910	3.930	3.920	3.950	3.910

Obsah alkoholu v pivu: *C349d*: výběr B

3.91	3.860	3.850	3.830	3.83	3.840	3.850	3.840	3.820	3.850

Úloha C3.50 *Test správnosti obsahu olova v sedimentu (Horn)*
V rámci projektu PHARE byl analyzován vzorek certifikovaného materiálu sedimentu a stanovován obsah olova. Vzorek byl mineralizován směsí kyselin HNO_3 a $HClO_4$ v mikrovlnné pícce. Deklarovaný obsah olova je 146±3 μg/g. Celkem bylo provedeno 8 rozkladů a stanovení olova. Určete střední hodnotu Hornovou metodou pivotů a porovnejte s klasickým odhadem.
Data: Obsah olova [μg/g] v sedimentu : 146.8, 142.3, 158.0, 145.8, 145.3, 126.6, 123.0, 151.4.

Úloha C3.51 *Test shodnosti dvou metod stanovení síranů ve vodách*
Porovnání dvou metod stanovení síranů ve vodách na deseti vzorcích. Srovnávala se starší metoda vážková a novější titrační. Na soubor dat aplikujte test shodnosti.
Data: Obsah síranů [mg. l^{-1}] ve vodách: *C351a* vážkově, *C351b* titračně.
C351a: 365.3, 72.0, 118.9, 132.3, 98.3, 74.2, 130.1, 125.4, 178.8, 162.0.
C351b: 348.5, 76.0, 114.1, 129.5, 101.0, 75.5, 129.0, 123.1, 175.0, 159.5.

Úloha C3.52 *Párový test stanovení dimethylaminu v butadienu*
Určete rozdíl mezi spektrofotometrickým a chromatografickým stanovením dimethylaminu v butadienu. Ze získaných 10 výsledků rozhodněte na hladině významnosti $\alpha = 0.05$, zda je rozdíl mezi oběma metodami významný.
Data: Obsah DMA [ppm] v butadienu: *C352a* chromatografie, *C352b* spektrofotometrie.
C352a: 10, 10.10, 10.2, 9.8, 9.9, 10, 9.9, 10.1, 9.8, 10.1.
C352b: 9.7, 9.8, 10.3, 9.5, 9.4, 9.6, 9.7, 9.3, 9.5, 9.2.

Úloha C3.53 *Test shodnosti stanovení OH skupin v kapalném kaučuku dvěma metodami*
Obsah OH skupin u 20 vzorků kapalných kaučuků byl stanoven acetylační a isokyanátovou metodou. Rozhodněte, zda obě metody poskytují na hladině významnosti $\alpha = 0.05$ stejné výsledky.
Data: Obsah OH [mmol OH/g] v kapalném kaučuku : *C353a* metodou isokyanátovou a *C353b* acetylační.
C353a: 0.735, 0.740, 0.734, 0.736, 0.738, 0.736, 0.738, 0.737, 0.739, 0.740, 0.740, 0.734, 0.735, 0.737, 0.735, 0.736, 0.730, 0.741, 0.742, 0.737.
C353b: 0.726, 0.730, 0.741, 0.742, 0.729, 0.734, 0.738, 0.735, 0.736, 0.732, 0.731, 0.737, 0.740, 0.738, 0.730, 0.740, 0.735, 0.741, 0.732, 0.737.

Úloha C3.54 *Střední hodnota obsahu nadouvadla v polystyrenu (Horn)*
Obsah nadouvadla (pentanu) byl stanoven ve vzorcích pěnového polystyrenu plynovou chromatografií. Aplikujete Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními odhady.
Data: Obsah pentanu [%] v polystyrenu : 4.93, 5.29, 5.41, 5.67, 5.72.

Úloha C3.55 *Test správnosti obsahu vázaného styrenu v kopolymeru (Horn)*
Stanovení obsahu vázaného styrenu v kopolymeru bylo testováno pomocí refraktometru. Modelově připravený vzorek obsahoval 30 % styrenu. Refraktometrické stanovení bylo provedeno dvanáctkrát. Určete, zda byl obsah styrenu na hladině významnosti $\alpha = 0.05$ stanoven správně.
Data: Obsah styrenu [%] v kopolymeru: 29.7, 30.4, 30.0, 29.8, 30.1, 29.9, 29.6, 30.1, 30.2, 30.4, 31.0, 29.0.

Úloha C3.56 *Test správnosti nalezeného obsahu amonných iontů*
Při mezilaboratorním testu bylo prováděno fotometrické stanovení amonných iontů. Bylo provedeno 14 stanovení a z nich byl určen odhad skutečné hodnoty neznámého vzorku. Skutečná hodnota standardu byla 3.60 mg amonného iontu v 1 litru. Posuďte, zda hodnota zjištěná laboratoří je správná.
Data: Obsah amoných iontů [mg. l^{-1}]: 3.61, 3.68, 3.62, 3.57, 3.66, 3.59, 3.55, 3.61, 3.57, 3.65, 3.58, 3.54, 3.66, 3.63.

Úloha C3.57 *Párový test obsahu fosforu metodami ICP-OES a AAS*
V 11 vzorcích oleje byl stanoven obsah fosforu metodami ICP-OES a AAS. Vyšetřete, zda párová data poskytují stejné výsledky.
Data: Obsah fosforu v oleji [mg. l^{-1}], *C357a* metodou ICP-OES, *C357b* metodou AAS. Data jsou ve dvojicích: 2.20 2.40, 2.50 2.50, 2.20 2.30, 2.80 2.40, 2.40 2.50, 2.80 3.30, 2.90 2.80, 2.90 2.90, 2.80 2.60, 2.70 2.50, 2.70 2.60, 2.50 2.30, 2.70 2.40, 2.90 2.80, 2.50 2.50, 2.80 3.00, 2.80 3.30, 3.00 3.10.

Úloha C3.58 *Střední hodnota plošné hmotnosti zdravotnické textilie (Horn)*
U plošné zdravotnické textilie, chirurgické síťky, byla měřena plošná hmotnost. U malého výběru dat užijeme Hornův postup pivotů a porovnáme výsledky s robustními odhady polohy a rozptýlení.
Data: Plošná hmotnost chirurgické síťky [g/m^2]: 111.9, 105.7, 108.1, 113.9, 112.3, 110.0, 110.0.

Úloha C3.59 *Test správnosti hodnoty jemnosti polyesterového hedvábí*
U nové dodávky polyesterového hedvábí byla měřena jemnost. Vyšetřete, zda naměřená jemnost odpovídá hodnotě 95.3 dtex±2 %, udávané výrobcem.
Data: Jemnost polyesterového hedvábí [dtex]: 86.1, 85.9, 84.1, 84.1, 86.8, 87.3, 85.8, 86.2, 85.9, 86.1, 85.5, 85.2, 85.5, 85.5, 84.2, 84.2, 87.1, 87.2, 85.4, 85.6.

Úloha C3.60 *Test shodnosti pevnosti šicích nití před extrakcí a po ní*
Byla měřena pevnost šicích nití v režném stavu před extrakci chloroformem a po ní. Testujte, zda má extrakce statisticky významný vliv na pevnost nitě.

Data: Pevnost šicích nití [N], *C360a* před extrakcí, *C360b* po extrakci. Data jsou ve dvojicích: 13.16 12.71, 13.14 12.93, 12.77 12.83, 13.01 13.06, 13.06 12.94, 12.78 12.99, 12.9 13.12, 12.77 12.79, 12.94 12.86, 12.78 12.84, 12.81 12.94, 13.19 13.22.

Úloha C3.61 *Párový test měření tloušťky plošných textilií dvěma způsoby*
Tloušťka plošných textilií byla měřena jednak digitálním tloušťkoměrem a jednak mikrometrem. Aplikujte test shodnosti a zjistěte, zda oba způsoby měření poskytují srovnatelné výsledky.
Data: Tloušťka plošných textilií [mm]: *C361a* digitálním tloušťkoměrem, *C361b* mikrometrem.
C361a: 0.35, 0.28, 0.52, 0.29, 0.56, 0.27, 0.33, 0.51, 0.25, 1,68.
C361b: 0.295, 0.232, 0.285, 0.228, 0.380, 0.220, 0.288, 0.415, 0.185, 1.37.

3.5.3 Analýza environmentálních, potravinářských a zemědělských dat

Úloha E3.01 *Čistota vody v řece testem BSK_5 (Horn)*
Čistota vody v řece byla denně sledována v průběhu 10 dní dle biologické spotřeby kyslíku BSK_5. Jsou v uvedených datech odlehlé hodnoty svědčící o dni, ve kterém blízký chemický zavod vypouštel do řeky nečistoty? Jaká je střední hodnota a parametr rozptýlení BSK_5 v průběhu 10 dnů a jaká je po vyloučení odlehlých hodnot? Aplikujte i Hornův postup. Je výhodné užít robustní odhady polohy a rozptýlení?
Data: Hodnoty BSK_5 vody v řece: 6.50, 5.80, 16.7, 6.40, 7.00, 6.30, 7.00, 9.20, 6.70, 6.70.

Úloha E3.02 *Detekce odlehlých hodnot při stanovení dusíku v půdě (Horn)*
Obsah dusíku v půdě, pocházející ze čpavku, dusičnanů a dusitanů, lze stanovit jako celkový plynný dusík v molech na litr. Obsah byl sledován 1krát týdně v průběhu osmi týdnů. Nachází se v datech jedna či více odlehlých hodnot? Určete střední hodnotu a rozptýlení koncentrace dusíku v půdě (str. 99 v cit.[15]). Aplikujte i Hornův postup.
Data: Obsah dusíku [$mol.l^{-1}$] v půdě: 0.0127, 0.1630, 0.0159, 0.0243, 0.0176, 0.0170, 0.0147, 0.0168.

Úloha E3.03 *Detekce odlehlých hodnot obsahu chemikálie ve vodě (Horn)*
V opakovaných výsledcích chemické analýzy vody se asi vyskytuje odlehlá hodnota. Na hladině významnosti $\alpha = 0.05$ vyšetřete, je-li nutné tuto podezřelou hodnotu z výběru vyloučit a vyšetřete rovněž do jaké míry dojde k ovlivnění střední hodnoty výběru a parametru rozptýlení. Užijte také Hornův postup. Intervalový odhad střední hodnoty vyčíslete na hladině významnosti $\alpha = 0.05$ (str. 44 v cit.[16]).
Data: Obsah chemikálie ve vodě [%]: 0.26, 0.21, 0.21, 0.20, 0.21, 0.19, 0.18, 0.17, 0.18, 0.19.
Úloha E3.04 *Střední hodnota obsahu fluoranthenu ve vodě (Horn)*
Fluoranthen je typický zástupce polyaromatických uhlovodíků ve vodě. Skupina těchto látek se stanovuje pomocí vysokoúčinné kapalinové chromatografie s fluorimetrickou detekcí. Na datech stanovení fluoranthenu je třeba určit parametry polohy a rozptýlení Hornovy metody pivotů a výsledky porovnat s klasickými a robustními statistikami polohy a rozptýlení.
Data: Koncentrace fluoranthenu [$ng.l^{-1}$] ve vodě: 24.1, 24.1, 24.3, 24.7, 25.1, 25.1, 25.1, 25.1, 25.4, 25.4, 25.6, 25.7, 26.2, 26.3.

Úloha E3.05 *Test správnosti obsah amoniakálního dusíku v odpadní vodě (Horn)*

Byl stanoven obsah amoniakálního dusíku v odpadní vodě v mg. l^{-1}. Hornovou metodou je třeba z naměřených hodnot stanovit střední hodnotu a 95% interval spolehlivosti. Dosahuje střední hodnota hodnoty 2.45 mg. l^{-1}? Ověřte především pomocí interva-lového odhadu.

Data: Obsah amoniakálního dusíku [mg. l^{-1}] v odpadní vodě: 2.45, 2.46, 2.46, 2.47, 2.48, 2.48, 2.49.

Úloha E3.06 *Střední hodnota obsahu p,p'-DDE v kontaminované zemině metodou GC (Horn)*

Stanovte reziduální obsah p,p'-DDE (v ppb) v kontaminované zemině metodou plynové chromatografie. K odhadu střední hodnoty obsahu p,p'-DDE užijte Hornův postup.

Data: Reziduální obsah p,p'-DDE [ppb] v kontaminované zemině: 9, 12, 18, 15, 18, 17.

Úloha E3.07 *Test správnosti obsahu síranu při analýze vody*

Pro kontrolu 35 vodohospodářských laboratoří v analýze vody byl připraven umělý vzorek s přesným obsahem síranů 94.8 mg. l^{-1}. Testujte, zda intervalový odhad střední hodnoty obsahuje tuto hodnotu. Jsou ve výběru nějaké odlehlé hodnoty?

Data: Obsah síranu SO_4^{2-} [mg. l^{-1}] ve vodě: 96.9, 110.00, 86.00, 101.70, 86.00, 85.00, 85.80, 86.50, 87.30, 103.9, 93.40, 86.90, 99.47, 115.90, 95.00, 92.80, 98.20, 98.00, 88, 97.10, 96.06, 80.20, 105.60, 122.50, 90.00, 68.00, 116.00, 89.4, 89.00, 92.50, 102.00, 83.12, 117.60, 91.35, 100.00.

Úloha E3.08 *Kontrola obsahu benzenu v ovzduší*

Při kontrole obsahu benzenu v ovzduší se přes skleněnou sorbční trubičku, plněnou Carbopackem B, prosává vzduch po dobu 40 minut. Bylo získáno 6 vzorků a obsah stanoven chromatograficky. Testujte, zda intervalový odhad střední hodnoty obsahuje hodnotu 5 mg/m^3. Aplikujte Hornův postup.

Data: Obsah benzenu mg/m^3: 3.34 4.76 4.87 5.11 3.89 4.83 5.02.

Úloha E3.09 *Střední hodnota obsahu manganu v rostlinném materiálu (Horn)*

Analýzou rostlinného materiálu byl stanoven obsah manganu v mg/kg. Výsledky analýzy vyhodnoťte Hornovým postupem a pak porovnejte s klasickými a robusními statistikami polohy a rozptýlení. Je rozdělení symetrické a obsahuje odlehlé hodnoty?

Data: Obsah manganu [mg/kg] v rostlinném materiálu: 16.4, 13.2, 11.2, 10.8, 16.2, 16.5, 11.1, 12.7, 16.2, 11.1.

Úloha E3.10 *Test správnosti obsahu síranu v pitné vodě (Horn)*
Ve vzorku povrchové vody, který byl 7krát vedle sebe naředěn v poměru 1:10, byly analyzovány sírany metodou RQ-flex (Merck). Prove(te analýzu využitím Hornova postupu. Deklarovaný obsah má být 50 mg. l^{-1}.
Data: Obsah síranů v povrchové vodě [mg. l^{-1}]: 50.0, 50.4, 55.0, 55.0, 55.5, 56.0, 56.5.

Úloha E3.11 *Střední hodnota obsahu olova v pitné vodě (Horn)*
V pitné vodě byl stanoven obsah olova v mg. l^{-1} voltametrickou metodou. Aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení. Jde o symetrické rozdělení? Ovlivňují odlehlé hodnoty výrazně výsledky ?
Data: Obsah olova v pitné vodě [mg. l^{-1}]: 48.3, 50.4, 48.3, 51.6, 49.2, 52.1, 50.0, 52.5.

Úloha E3.12 *Test správnosti obsahu amoniakálního dusíku v odpadní vodě (Horn)*
V odpadní vodě byl stanoven amoniakální dusík v mg. l^{-1}. Z naměřených hodnot je třeba stanovit střední hodnotu a 95% interval spolehlivosti analýzou malého výběru. Prove(te také analýzu využitím Hornova postupu. Testujte na hladině významnosti $\alpha = 0.05$, zda je splněn předpokládaný obsah 2.45 mg. l^{-1}.
Data: Obsah amoniakálního dusíku v odpadní vodě [mg. l^{-1}]: 2.45, 2.46, 2.46, 2.47, 2.48, 2.48, 2.49.

Úloha E3.13 *Stanovení střední hodnoty obsahu benzenu ve vodě (Horn)*
Využitím Hornova postupu analýzy malých výběrů vypočtěte střední hodnotu obsahu benzenu ve vodě, naměřeného metodou GC-FID (HP 5890). Vyšetření proveďte na hladině významnosti $\alpha = 0.05$.
Data: Obsah benzenu ve vodě [μg. l^{-1}]: 4.56, 3.82, 5.21, 4.92, 4.35.

Úloha E3.14 *Test shodnosti potenciometrického a jodometrického stanovení obsahu kyslíku*
Pro stanovení obsahu rozpuštěného kyslíku ve vodě lze použít dvou metod, a to stanovení potenciometricky pomocí Clarkovy kyslíkové elektrody (oximetrie) nebo jodometrickou titraci, tj. modifikovanou Winklerovou metodou. Na dvou výběrech dat stanovení kyslíku ve vodě aplikujeme test shodnosti na hladině významnosti $\alpha = 0.05$. Jsou rozptyly obou výběrů stejné? Jsou obě rozdělení gaussovská?
Data: Obsah kyslíku ve vodě [mg. l^{-1}]: *E314a,* potenciometricky, *E314b* jodometricky,
E314a: 7, 7.0, 6.9, 6.9, 6.8, 6.9, 6.9, 7.1, 7.0, 7.0, 6.4, 6.6, 6.7, 6.8, 6.9, 7.1, 7.0, 6.9, 6.8, 7.1, 6.9, 6.6, 6.6, 6.6, 6.6, 6.8, 6.8, 7.3, 6.9, 6.7.
E314b: 6.9, 6.9, 6.6, 6.7, 6.5, 6.7, 6.6, 6.6, 6.6 , 6.6, 6.6, 6.6, 6.6, 6.8, 6.9, 7.1, 7.0, 6.7, 7.4, 7.0, 7.1, 7.0, 6.7, 7.1, 6.9, 6.9, 7.1, 7.3, 7.1, 7.2.

Úloha E3.15 *Porovnání dvojího stanovení CHSK v odpadních vodách*
Stanovení *CHSK*(Cr) v odpadních vodách bylo prováděno destilační metodou a současně tepelným rozkladem v termoreaktoru. Určete, zda obě metody vedou ke stejným výsledkům. Testujte shodnost rozptylů obou výběrů na hladině významnosti $\alpha = 0.05$.
Data: Hodnota *CHSK (Cr)* [mg. l^{-1}] v odpadních vodách: *E315a* titračně, *E315b* tepelně.
E315a: 270, 274, 271, 274, 270, 273, 274, 273, 273, 273, 272, 273, 274, 272, 274.
E315b: 267, 270, 268, 268, 270, 271, 267, 268, 269, 270, 268, 270, 271, 268, 271.

Úloha E3.16 *Test správnosti obsahu chloru v upravené vodě*
V průběhu jednoho dne byl na úpravně vody sledován po hodinách obsah chlóru v upravené vodě. Jsou ve výběru odlehlé hodnoty? Dáte přednost robustním odhadům nebo mocninné transformaci? Určete parametry polohy a rozptýlení výběru a odpovídající 95%ní intervaly spolehlivosti. Dle normy je doporučená hodnota chloru 0.3 mg. l^{-1}. K testování využijte intervalového odhadu na hladině významnosti $\alpha = 0.05$.
Data: Obsah chlóru [mg. l^{-1}] v upravené vodě: 0.1, 0.15, 0.25, 0.15, 0.30, 0.25, 0.25, 0.30, 0.35, 0.55, 0.70, 0.70, 0.80, 0.65, 0.55, 0.50, 0.30, 0.35, 0.30, 0.25, 0.25, 0.20, 0.15.

Úloha E3.17 *Test shodnosti obsahu chloroformu ve vodě dvojím zařízením*
Byly použity dva výběry po sedmi hodnotách opakovaných stanovení obsahu chloroformu ve vodě po zpracování na dvou různých stripovacích zařízeních označených jako A a B, (stanovení provedena GC-FID). Hodnoty jsou uváděny v µg. l^{-1}. Testujte homogenitu rozptylů a pak shodu středních hodnot obou výběrů na hladině významnosti $\alpha = 0.05$. Jsou oba výběry gaussovské?
Data: Obsah chloroformu [µg. l^{-1}] ve vodě, *E317a*: zařízení A, *E317b*: zařízení B.
E317a: 28.80, 20.20, 23.11, 19.84, 22.11, 19.95, 25.30.
E317b: 22.78, 23.46, 22.00, 19.83, 22.19, 22.11, 22.48.

Úloha E3.18 *Test shodnosti obsahu selenu v odpadní vodě metodami ICP a AAS*
V modelových vzorcích odpadních vod byl jednorázově měřen obsah selenu paralelně na ICP spektrometru a atomovém absorpčním spektrometru po generaci hydridu. Na naměřená data aplikujte párový test na hladině významnosti $\alpha = 0.05$.
Data: Obsah Se [mg. l^{-1}] v odpadní vodě: *E318a* metoda ICP, *E318b* metoda AAS.
E318a: 47.40, 47.40, 47.48, 47.66, 47.92, 48.18, 49.99, 51.68, 51.68, 52.33.
E318b: 48.71, 48.76, 48.92, 48.98, 49.00, 49.08, 49.73, 49.73, 49.95, 51.51.

Úloha E3.19 *Test správnosti dodržení přípustné koncentrace fluoridů v pitné vodě*
Předpis normy ČSN 757111-Pitná voda připouští v pitné vodě maximální koncentraci fluoridů 1.5 mg. l^{-1}. Na výběru 33 dat a hladině významnosti $\alpha = 0.05$ je třeba rozhodnout, byla-li dodržena maximálně přípustná střední hodnota koncentrace fluoridů ve vodě nebo zda je nižší. Jaký nejlepší odhad střední hodnoty užijete? Je rozdělení symetrické a bez odlehlých hodnot?
Data: Koncentrace fluoridů [mg. l^{-1}] v pitné vodě: 1.5, 1.54, 1.43, 1.44, 1.46, 1.44, 1.48, 1.46, 1.51, 1.42, 1.91, 1.55, 1.49, 1.55, 1.45, 1.53, 1.49, 1.50, 1.52, 1.54, 1.55, 1.46, 1.47, 1.44, 1.53, 1.52, 1.52, 1.48, 1.46, 1.58, 1.51, 1.46, 1.47.

Úloha E3.20 *Test shodnosti výsledků stanovení rtuti na dvou přístrojích*
Rtuť byla stanovována ve vzorku odpadního kalu na dvou přístrojích TMA 254, A a B. Na každém přístroji bylo provedeno 10 analýz tohoto materiálu. Statistickým testováním je třeba posoudit, zda výsledky obsahu rtuti jsou shodné. Pro testování byla zvolena hladina významnosti $\alpha = 0.05$.
Data: Obsah rtuti v kalu [ppm]: *E320a* přístroj A, *E320b* přístroj B.
E320a: 1.039, 1.071, 1.032, 1.129, 1.046, 1.113, 1.069, 1.022, 1.096, 1.044.
E320b: 1.131, 1.226, 1.233, 1.043, 0.970, 1.183, 1.178, 1.014, 1.201, 1.227.

Úloha E3.21 *Test shodnosti dvou analytických metod stanovení dusičnanů*
K určení obsahu dusičnanů v 10 vzorcích povrchové vody byla použita metoda kapalinové chromatografie (A) a RQ - flex (Merck) metoda (B). Ze získaných 10 hodnot určete, zda obě metody vedou ke stejným výsledkům. Pro testování byla zvolena hladina významnosti $\alpha = 0.05$.
Data: Obsah dusičnanů [mg. l^{-1}] v povrchové vodě, *E321a:* metoda A, *E321b:* metoda B.
E321a: 47.0, 47.5, 48.0, 46.5, 46.0, 47.0, 47.5, 47.0, 46.0, 48.0.
E321b: 53.0, 52.0, 53.0, 54.0, 53.0, 54.0, 54.0, 53.0, 52.0, 53.0.

Úloha E3.22 *Test shodnosti stanoveného obsahu amonných iontů na dvou přístrojích*
Byl stanoven obsah amonných iontů ve standardním vzorku v mg. l^{-1} na dvou různých přístrojích, spektrofotometrech A a B. Na data aplikujte test shodnosti obou obsahů. Jsou rozptyly obou výběrů stejné? Pro testování byla zvolena hladina významnosti $\alpha = 0.05$.
Data: Obsah amonných iontů [mg. l^{-1}] ve standardním vzorku: *E322a* přístroj A, *E322b* přístroj B.
E322a: 0.500, 0.500, 0.501, 0.502, 0.498, 0.499, 0.501, 0.498, 0.499, 0.497.
E322b: 0.505, 0.503, 0.502, 0.500, 0.500, 0.499, 0.502, 0.504, 0.502, 0.501.

Úloha E3.23 *Stanovení střední hodnoty obsahu dusíku v půdě (Horn)*
Obsah dusíku v půdě, pocházející ze čpavku, dusičnanů a dusitanů, lze stanovit jako celkový plynný dusík v molech na litr. V rámci určitého období bylo provedeno 8 měření. Odhadněte parametry Hornovou metodou a výsledky porovnejte s klasickými a robustními odhady parametrů.
Data: Obsah dusíku v půdě [mol. l^{-1}]: 0.0127, 0.0147, 0.0159, 0.0168, 0.0170, 0.0176, 0.0243, 0.0163.

Úloha E3.24 *Stanovení střední hodnoty obsahu olova a chromu v odpadních vodách (Horn)*. V období září - prosinec 1995 byla v okolí jistého závodu sledována koncentrace olova Pb a chromu Cr v odpadních vodách. Na tato data aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy a rozptýlení.
Data: Stanovení koncentrace olova a chromu v odpadních vodách [mg. l^{-1}], *E324a* Pb, *E324b* Cr.
E324a: 0.271, 0.278, 0.326, 0.328, 0.338, 0.392, 0.444, 0.460, 0.528, 0.555, 0.575, 0.575, 0.925.
E324b: 0.057, 0.060, 0.085, 0.085, 0.097, 0.100, 0.101, 0.126, 0.158, 0.228, 0.230, 0.246, 0.321.

Úloha E3.25 *Test správnosti stanovení antimonu metodou AAS (Horn)*
Antimon lze stanovit atomovou absorpční spektrometrií po generaci jeho hydridu. Ověřte, zda lze touto metodou stanovit koncentraci 1000 ng Sb. Aplikujte test správnosti výsledků metodou intervalu spolehlivosti míry polohy na hladině významnosti $\alpha = 0.05$. Uměle připravený vzorek obsahoval 1000 ng/100 ml Sb a byl měřen 12krát.
Data: Koncentrace antimonu [ng/100ml Sb] ve vzorku: 983, 993, 995, 996, 997, 999, 1000, 1003, 1004, 1008, 1009, 1012.

Úloha E3.26 *Střední hodnota u obsahu BSK za devítiměsíční období (Horn)*
V měsících leden až září 1997 byla v říčním profilu stanovena 1krát měsíčně hodnota BSK. Soubor je třeba statisticky vyhodnotit a určit míry polohy a rozptýlení.
Data: Hodnota BSK_5 v říčním profilu:

Měsíc	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.
BSK_5 [mg. l^{-1}]	2.5	2.7	1.4	1.3	2.9	3.2	2.6	2.2	1.6

Úloha E3.27 *Test shodnosti počtu koliformích bakterií v toku řeky dvěmi laboratořemi*

V měsících duben až červenec 1997 byl v říčním profilu Úhlava (Doudlevce) v pracovních dnech stanovován počet koliformních bakterií v toku. Stanovení prováděly paralelně vedle sebe 2 laboratoře. Zhodnoťte statisticky oba soubory a porovnejte jejich shodnost.

Data: Počet koliformních bakterií v toku řeky, *E327a*: A, *E327b*: B. Data jsou ve dvojicích:

49	48	28	57	10	20	42	46	97	80
..	..	..	..	..	..	..	..	..	..
89	52	25	24	34	74	74	62	160	160

Úloha E3.28 *Stanovení střední hodnoty obsahu železa v pitné vodě (Horn)*

Obsah železa ve vodě byl zjišťován fotometrickou metodou s kyselinou sulfosalicylovou. Ze získaných devíti hodnot vyhodnoďte klasické a robustní odhady polohy a rozptýlení. Použijte Hornovou metodou pivotů. Získané výsledky porovnejte a zhodnoťte.

Data: Obsah železa Fe [mg. l^{-1}]: 0.169, 0.207, 0.163, 0.178, 0.226, 0.159, 0.180, 0.221, 0.169.

Úloha E3.29 *Stanovení střední hodnoty dusičnanů v pitné vodě fotometricky*

Po dobu dvou měsíců byl sledován obsah dusičnanů v novém zdroji pitné vody. Určete statistické vlastnosti výběru dat a případné odlehlé body. Jaké jsou nejlepší odhady polohy a rozptýlení ?

Data: Obsah dusičnanů [mg. l^{-1}] v pitné vodě fotometricky: 22.00, 24.70, 23.40, 21.10, 22.80, 22.85, 23.80, 24.10, 22.20, 25.60, 24.80, 23.00, 23.50, 24.15, 24.40, 22.70, 25.40, 25.00, 23.2, 23.60, 23.00, 25.20, 24.80, 22.00, 21.30, 22.50, 24.60, 24.5, 23.95, 23.00, 22.80, 23.20, 23.60, 24.20, 24.20, 23.70, 24, 24.00, 23.55.

Úloha E3.30 *Párový test obsahu železa ve vodě, určeného ve dvou laboratořích*

Koncentrace železa v pitné vodě je sledována ve dvou laboratořích A a B. Vzorkování proběhlo v jeden den, voda byla nabírána z různých vodovodů. Určete, zda je možné považovat stanovení v obou laboratořích za shodná.

Data: Obsah železa [mg. l^{-1}] ve vodě dvěma laboratořemi *E330a* A a *E330b* B.
E330a: 0.05, 0.06, 0.06, 0.16, 0.13, 0.04, 0.24, 0.06, 0.95, 1.20, 0.11, 0.18, 0.32.
E330b: 0.06, 0.08, 0.07, 0.16, 0.15, 0.06, 0.26, 0.07, 0.79, 1.15, 0.26, 0.07, 0.30.

Úloha E3.31 *Test shodnosti obsahu dusičnanů ve vodě dvěma laboratořemi*

Koncentrace dusičnanů v pitné vodě je sledována dvěma laboratořemi A a B. Bylo provedeno 10 stanovení obsahu dusičnanů ve vodě, pocházející z jednoho zdroje. Odběr vzorků byl proveden jeden den a vychází se z předpokladu, že obsah dusičnanů se v rozvodu nemění. Určete, zda lze považovat stanovení obou laboratoří za shodná.

Data: Obsah dusičnanů [mg NO_3^- l^{-1}] ve vodě dvěma laboratořemi *E331a* A a *E331b* B.
E331a: 42.0, 44.0, 46.0, 46.0, 54.0, 52.0, 56.0, 46.0, 40.0, 46.0.
E331b: 49.6, 58.7, 51.1, 58.7, 56.8, 53.9, 55.1, 56.6, 54.5, 47.7.

Úloha E3.32 *Střední hodnota hustoty dřeva smrku (Horn)*

Vypočtěte střední hodnotu hustoty dřeva smrku [kg/m^3] využitím Hornova postupu a výsledky porovnejte s klasickými a robustními odhady.

Data: Hustota dřeva smrku [kg/m^3]: 481, 475, 495, 470, 504, 498, 488.

Úloha E3.33 *Test správnosti měření výšky stromu novým výškoměrem*
Testujte správnost měření výšky stromu novým výškoměrem, když kontrolní výška 20 m byla změřena 30krát. Posuďte, zda výškoměr měří správně. Ověřte předpoklady, kladené na výběr a symetrii rozdělení a rozhodněte, který odhad užijete.
Data: Naměřené kontrolní výšky 20 m stromu novým výškoměrem: 20.3, 20.0, 21.6, 19.7, 20.9, 20.2, 20.2, 19.5, 20.0, 20.0, 20.4, 19.7, 20.2, 19.9, 19.9, 20.1, 20.5, 21.3, 20.9, 21.2, 19.0, 19.5, 20.2, 19.5, 20.3, 19.5, 20.5, 22.1, 19.5, 20.0.

Úloha E3.34 *Test shodnosti výčetní tloušťky dvou porostů*
Byly měřeny výčetní tloušťky (tj. tloušťka kmene stromu ve výšce 1.3 m nad patou kmene) dvou porostů. Testujte, zda jsou oba porosty stejně tloušťkově vyspělé.
Data: Výčetní tloušťky [cm] dvou porostů: *E334a* porost I, *E334b* porost II.
E334a:, 33.5, 28.6, 36.2, 41.4, 41.0, 43.7, 24.1, 33.8, 40.5, 29.6, 31.5, 26.5, 25.8, 30.1, 31.1, 24.4, 32.2, 33.0, 35.7, 33.2, 33.4, 33.1, 41.7, 34.6, 34.1.
E334b:, 29.5, 30.4, 21.6, 24.4, 28.5, 26.8, 25.5, 22.6, 26.1, 19.2, 24.6, 28.7, 20.2, 21.0, 27.5, 25.8, 23.3, 26.5, 31.2, 28.3.

Úloha E3.35 *Párový test měřených výšek dvěma výškoměry*
Při zkoušce spolehlivosti výškoměru bylo provedeno 15 kontrolních měření při porovnání s druhým standardním výškoměrem. Testujte shodnost naměřených výšek oběma výškoměry párovým testem.
Data: Naměřená výška [m] standardu: *E335a* výškoměr I, *E335b* výškoměr II.
E335a: 26.8, 22.1, 19.5, 18.5, 29.4, 33.2, 16.3, 18.5, 35.4, 29.5, 19.2, 26.2, 18.4, 28.6, 17.5.
E335b: 26.2, 22.4, 19.5, 19.0, 29.3, 33.5, 16.2, 18.9, 36.0, 29.8, 18.9, 26.0, 18.5, 29.0, 18.0.

Úloha E3.36 *Test shodnosti hodnot BSK_5 dvěma metodami*
Koncentrace BSK_5 v odpadní vodě je stanovována dvěma metodami, standardní zřeďovací metodou a pomocí biosenzoru. Oběma metodami byla provedeno 14 stanovení. Proveďte test shodnosti výsledků obou metod.
Data: Koncentrace kyslíku [mg. l^{-1}] v odpadní vodě: *E336a* standardní zřeďovací metodou, *E336b* biosenzorem,
E336a: 17.0, 14.0, 17.0, 17.0, 17.0, 14.0, 14.0, 14.0, 17.0, 17.0, 17.0, 14.0, 14.0, 17.0.
E336b: 14.6, 15.3, 17.8, 14.5, 13.8, 16.2, 17.6, 14.1, 14.7, 15.2, 17.5, 14.3, 13.9, 17.1.

Úloha E3.37 *Test shodnosti obsahu vápníku ve dvou vzorcích*
Ve jednom odběru odpadní vody byl ve dvou vzorcích stanovován obsah vápníku, jako jednoho z hlavních ukazatelů znečištění odpadních vod. Vyšetřete, zda jsou obsahy vápníku v obou vzorcích shodné.
Data: Obsah vápníku v odpadní vodě [mg. l^{-1}]: *C337a* v 1. vzorku, *C337b* ve 2. vzorku.
C337a: 134.4, 134.7, 138.1, 139.1, 132.1, 131.3, 134.3, 134.3, 138.5, 137.8, 133.6, 133.1, 138.5, 136.8.
C337b: 137.3, 137.2, 138.3, 136.6, 135.9, 134.7, 134.3, 134.9, 137.3, 138.7, 134.9, 133.8, 132.5, 133.1.

3.5.4 Analýza hutnických a mineralogických dat

Úloha H3.01 *Test správnosti stanoveného obsahu chromu v kontrolním roztoku (Horn)*
Bylo provedeno sedm měření obsahu chromu v kontrolním roztoku. Na základě aplikace Hornovy metody malých výběrů rozhodněte na hladině významnosti $\alpha = 0.05$, zda interval spolehlivosti obsahuje správnou hodnotu 0.20 mg. l^{-1}. Výsledky porovnejte s klasickými i robustními odhady polohy a rozptylu. Jsou přítomny odlehlé hodnoty?
Data: Obsah chromu [mg. l^{-1}] v kontrolním roztoku : 0.18, 0.14, 0.19, 0.19, 0.18, 0.20, 0.19.

Úloha H3.02 *Stanovení střední hodnoty rázové houževnatosti (Horn)*
Stanovte střední hodnotu rázové houževnatosti z naměřených hodnot malého výběru u UMALURU U 95/42 Hornovým postupem. Měření bylo prováděno na digitálním kladivu označení ZWICK. Je rozdělení symetrické a bez odlehlých hodnot?
Data: Hodnota rázové houževnatosti [kJ. m^{-2}]: 8.808, 5.549, 7.384, 7.200, 7.172, 5.755, 6.786.

Úloha H3.03 *Stanovení střední hodnoty obsahu paladia v Pd-Pt sítích (Horn)*
Byl proveden rozbor pěti platinových sít (Pd-Pt) na obsah paladia. Určete odhad polohy a rozptýlení malého výběru dat Hornovým postupem.
Data: Obsah paladia [%] v Pd-Pt sítích: 65.1, 69.7, 77.8, 80.2, 80.2, 81.4.

Úloha H3.04 *Stanovení střední hodnoty obsahu mědi v pesticidu Cursate K (Horn)*
Stanovte obsah mědi ve vzorku Státní zkušebny po mineralizaci směsí kyselin dusičné a sírové a nasledné jodometrické titraci. Užijte Hornův postup analýzy malého výběru. Je rozdělení symetrické a homogenní? Lze indikovat nějaké odlehlé hodnoty?
Data: Obsah mědi [hm.%] v pesticidu Cursate K : 44.79, 43.79, 44.04, 44.38, 44.11, 44.65, 44.54, 45.70, 45.79, 45.59.

Úloha H3.05 *Test shodnosti obsahu zinku stanoveného dvěma metodami*
Při častém stanovování obsahu zinku v roztoku byly zkoušeny dvě metody kontrolního stanovení bodu ekvivalence chelatometricky a odměrným roztokem hexakyano-železnatanu draselného. Aplikujte test shodnosti obsahu zinku u obou metod na hladině významnosti $\alpha = 0.05$.
Data: Obsah zinku dle bodu ekvivalence [ml]:
H305a: 21.2, 22.0, 23.2, 22.1, 22.4, 23.0, 23.0, 21.8, 21.9, 22.2, 22.3, 21.8.
H305b: 21.0, 22.2, 23.3, 22.1, 22.2, 22.8, 22.9, 21.9, 21.9, 22.3, 22.4, 21.9.

Úloha H3.06 *Stanovení střední hodnoty odporu v penetraci zeminy (Horn)*
Ve vzorku zeminy bylo provedeno stanovení odporu v penetraci. Vyšetřete předpoklady o výběru. Jsou prvky výběru vybrány náhodně, není mezi nimi skrytá závislost? Pocházejí data ze souboru, který vykazuje Gaussovo rozdělení? U kolika diagnostik průzkumové analýzy je shoda v indikaci odlehlých bodů? Aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy rozptýlení.
Data: Odpor v penetraci zeminy [N. cm^{-1}]: 44.9, 45.9, 47.6, 47.6, 48.5.

Úloha H3.07 *Stanovení pevnosti v ohybu po vysušení cihlářské suroviny (Horn)*
U vzorku cihlářské suroviny byla provedena zkouška pevnosti v ohybu po vysušení. Aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními statistikami polohy rozptýlení. Jsou v datech odlehlé hodnoty?
Data: Pevnost v ohybu po vysušení cihly [N. cm^{-1}]: 7.93, 8.30, 8.47, 8.88, 8.95.

Úloha H3.08 *Test správnosti obsahu křemene v minerálu*
Na difraktometru bylo zkoušeno kvantitativní stanovení minerálu křemene. Uměle připravený vzorek minerálu obsahoval 10 % křemene a byl proměřen dvanáctkrát. Na hladině významnosti $\alpha = 0.05$ určete, zda byl obsah křemene stanoven správně, tj. zda hodnota 10.0 % leží v intervalu spolehlivosti.
Data: Obsah křemene [%]: 8.7, 10.2, 10.07, 9.75, 9.65, 10.37, 10.14, 10.50, 9.48, 11.22, 9.49, 9.86.

Úloha H3.09 *Obsah oxidu hlinitého z průzkumného vrtu*
Za účelem využití nadložních sedimentů severočeské pánve v cihlářské výrobě byl proveden odběr vzorků z průzkumného vrtu. Ve vzorcích byl sledován obsah Al_2O_3. Vyčíslete míry polohy obsahu.
Data: Obsah Al_2O_3 [%] ve vzorcích zeminy:

H309	24.48	21.84	22.70	22.30	22.70	22.88	21.74	21.52	23.88
21.74	21.44	23.08	23.39	22.16	22.73	22.66	22.50	22.05	20.40
23.31	21.20	20.90	25.60	22.28	21.74	22.30	22.42	23.52	

Úloha H3.10 *Test shodnosti obsahu oxidu křemičitého ze dvou průzkumných vrtů*
Za účelem využití nadložních sedimentů severočeské pánve v cihlářské výrobě byl proveden odběr vzorků ze dvou průzkumných vrtů. Ve vzorcích A a B byl sledován obsah SiO_2. Ověřte, zda lze považovat obsahy za shodné?
Data: Obsah SiO_2 [%] ve vzorcích zeminy: *H310a* vzorek A, *H310b* vzorek B.

H310a:	58.54	58.61	60.95	57.53	59.92	59.08	60.80	56.83	59.50
59.92	57.49	59.98	58.71	59.09	59.69	60.91	59.74	59.66	59.58
57.90	59.51	58.99	58.41	59.86	61.92	59.61	61.15	58.18	
H310b: B	60.39	58.54	60.14	59.82	61.97	59.18	60.74	59.34	59.37
60.80	58.89	57.70	58.86	60.69	62.01	60.80	61.10	60.10	62.26
60.65	61.99	60.24	60.35	60.79	62.45	62.22	61.57	60.99	

Úloha H3.11 *Párový test obsahu CaO stanoveného dvěma analytickými metodami*
Ve dvaceti vzorcích zeminy bylo provedeno stanovení obsahu CaO dvěma analytickými metodami, titrační a vážkovou. Na hladině významnosti $\alpha = 0.05$ určete, zda je možné pracnou metodu vážkovou nahradit rychlejší metodou titrační. K posouzení užijte párový test.
Data: Obsah CaO [mg. l^{-1}] ve vzorcích zeminy: *H311a:* titračně, *H311b:* vážkově.
H311a: 0.2388, 0.2200, 0.2413, 0.2014, 0.2370, 0.1835, 0.2228, 0.1975, 0.2455, 0.2319, 0.1944, 0.2438, 0.1944, 0.2291, 0.2388, 0.2358, 0.2399, 0.2015, 0.1893, 0.1944.
H311b: 0.2317, 0.2428, 0.2318, 0.2428, 0.2373, 0.2456, 0.2442, 0.2249, 0.2538, 0.2442, 0.2249, 0.2400, 0.2402, 0.2538, 0.2290, 0.2359, 0.2395, 0.2456, 0.1914, 0.2428.

Úloha H3.12 *Test shodnosti hmotnosti pytlů cementu ze dvou dávkovačů*
Ve dvou cementárnách byla provedena kontrola dávkovačů. Zvážením 13 náhodně vybraných pytlů cementu byly získány hodnoty výběru A. Obdobně zvážením 9 náhodně vybraných pytlů z druhé cementárny byly získány hodnoty výběru B. Na hladině významnosti $\alpha = 0.05$ odhadněte střední hodnotu hmotnosti pytle u každého výběru a proveďte test shodnosti.
Data: Hmotnost pytlů [kg] cementu: *H312a*: ze 13 pytlů, *H312b*: z 9 pytlů.
H312a: 51.5, 47, 48.5, 53.0, 47.3, 48.1, 48.8, 49.2, 52.3, 47.1, 49.5, 46.3, 50.1.
H312b: 50.3, 50.7, 49.2, 50.1, 49.9, 51.1, 49.8, 48.9, 50.3.

Úloha H3.13 *Stanovení střední hodnoty obsahu síry v cinvalditu (Horn)*
Určete míry polohy, rozptýlení výběru obsahu síry v cinvalditu v jednotkách ppm. Jsou v datech odlehlé hodnoty? Je rozdělení symetrické? Je třeba užít mocninné transformace?
Data: Obsah síry v cinvalditu [ppm]: 290, 190, 200, 245, 300, 311, 320, 376, 400, 400, 400.

Úloha H3.14 *Stanovení obsahu oxidu uhličitého v cinvalditu (Horn)*
Určete míry polohy, rozptýlení a tvaru výběru obsahu oxidu uhličitého v cinvalditu v procentech. Na základě ověření předpokladů o výběru rozhodněte, zda je třeba užít robustních odhadů nebo mocninné transformace?
Data: Obsah oxidu uhličitého CO_2 v cinvalditu [%]: 0.19, 0.20, 0.21, 0.26, 0.33, 0.34, 0.40, 0.58.

Úloha H3.15 *Test shodnosti dvou metod stanovení obsahu fluoru v cinvalditu*
Určete míry polohy, rozptýlení a tvaru rozdělení výběru obsahu fluoru v cinvalditu v procentech dvěma metodami, iontově-selektivní elektrodou ISE a rentgenovou fluoroscenční spektrometrií FX, a výsledky porovnejte. Na hladině významnosti $\alpha = 0.05$ aplikujte test shodnosti a vyšetřete, zda bylo dosaženo shodných výsledků?
Data: Obsah fluoru v cinvalditu [%], *H315a:* ISE, *H315b:* FX.
H315a: 4.53, 4.60, 4.95, 5.05, 5.24, 5.30, 5.31, 5.33, 5.37, 5.43, 5.45, 5.47, 5.60, 5.65, 5.68, 5.70, 5.70, 5.80, 6.10, 6.54.
H315b: 4.05, 4.60, 4.83, 4.84, 5.45, 5.66, 5.89, 5.91, 6.00, 6.58.

Úloha H3.16 *Test správnosti obsahu oxidu vápenatého ve vápenci (Horn)*
Obsah oxidu vápenatého CaO ve vápenci činí 56.03 %. Určete míry polohy, rozptýlení a tvaru výběru obsahu oxidu vápenatého ve vápenci, když analytické výsledky obsahu CaO byly získány chelatometrickou titrací. Na hladině významnosti $\alpha = 0.05$ aplikujte test správnosti a vyšetřete, zda chelatometrická titrace je zatížena systematickou chybou.
Data: Obsah CaO v uhličitanu vápenatém [%]: 55.58, 55.72, 55.85, 55.9, 56.15.

Úloha H3.17 *Přesnost sériového stanovení zlata (Horn)*
Pro směrodatnou odchylku stanovení stopového zlata v mineralogickém materiálu byl odvozen vztah $\sigma_{max} \doteq 0.22\, c^{\frac{5.75}{\log c \,\%\, 8.42}}$, kde c je obsah zlata vyjádřený v ppm. Odpovídá numerická hodnota rozptylu hodnot předloženého výběru výrazu σ^2_{max}?
Data: Obsah zlata stopové analýzy [ppm] v mineralogickém materiálu: 2.7, 2.8, 3.0, 3.0, 3.3.

Úloha H3.18 *Test shodnosti obsahu oxidu hlinitého u dvou metod úpravy vzorku*
Srovnejte úpravu vzorku kaolinu tavením a lisováním pro stanovení obsahu oxidu hlinitého v procentech a určete, zda výsledky jednotlivými metodami jsou shodné. Mají oba výběry shodné rozptyly? Test shodnosti proveďte na hladině významnosti $\alpha = 0.05$.
Data: Obsah oxidu hlinitého [%] v kaolinu: *H318a* tavením, *H318b* lisováním.
H318a: 35.02, 35.26, 35.26, 34.92, 35.27, 35.37, 35.01, 35.14, 35.22, 34.85, 35.33, 35.20, 35.29.
H318b: 34.73, 34.68, 34.69, 34.34, 34.75, 35.10, 35.10, 35.04, 35.22, 34.74, 35.17, 34.87, 34.92.

Úloha H3.19 *Párový test obsahu oxidu hlinitého ve dvou laboratořích*
K porovnání stejné metody chemické analýzy kaolinu roentgenově-fluorescenční analýzou na připraveném vzorku kaolinu tavením byly použity výsledky stanoveného obsahu oxidu hlinitého z průměrných měsíčních vzorků, naměřené v laboratoři závodu a v laboratoři odběratele. Aplikujte párový test na hladině významnosti $\alpha = 0.05$.
Data: Obsah oxidu hlinitého [%] v kaolinu: *H319a* závod, *H319b* odběratel.
H319a: 35.02, 35.26, 35.26, 34.92, 35.27, 35.37, 35.01, 35.14, 35.22, 34.85, 35.33, 35.20, 35.29.
H319b: 35.24, 34.88, 35.23, 34.95, 35.14, 35.48, 35.55, 35.15, 35.46, 35.09, 35.29, 35.33, 35.39.

Úloha H3.20 *Test správnosti stanovení obsahu SiO_2*
V laboratoři se analyzuje každý měsíc standardní vzorek kaolinu KKA na rentgenově-fluorescenčním spektrometru. Na obsahu oxidu křemičitého ve standardním vzorku KKA [%] aplikujte test správnosti výsledku vůči známé hodnotě dané normou 50.87 %.
Data: Obsah SiO_2 [%] v kaolinu: 50.67, 50.79, 50.66, 50.75, 50.81, 50.54, 51.05, 51.13, 51.11, 50.37, 51.16, 50.70, 50.99.

Úloha H3.21 *Střední hodnota pevnosti v ohybu keramického válečku (Horn)*
Zkoušení pevnosti keramického válečku v ohybu se provádí na 5 tělíscích, zhotovených z jednoho keramického válečku. Na výsledky měření pevnosti v ohybu v MPa aplikujte Hornovu metodu pivotů a výsledky porovnejte s klasickými a robustními odhady.
Data: Pevnost keramického válečku v ohybu [MPa]: 50.64, 49.28, 54.54, 50.18.

Úloha H3.22 *Párový test obsahu niklu v drátu a svárovém kovu*
Posuzovány jsou párové hodnoty obsahu niklu Ni [%] v drátu a ve svarovém kovu, přičemž svařování probíhalo v inertní atmosféře argonu. Párovým testem určete, zda lze nahradit zkoušení svarového kovu výsledky z původního drátu. Analyzujte rovněž hodnotu rozdílu párových hodnot a aplikujte test správnosti na nulovou hodnotu tohoto rozdílu.
Data: Párová data obsahu niklu [%], *H322a* v drátu, *H322b* ve svaru, *H322c* rozdíl páru.

25.49	25.55	-0.06,	25.79	25.23	0.56,	25.32	25.56	-0.24,
..	..	..	..	..	..	..	..	..
13.52	13.39	0.13,	13.67	13.45	0.22,	13.27	13.17	0.10,

Úloha H3.23 *Párový test stanovení koncentrace železa dvěma metodami*
Koncentrace železa v hliníkové slitině byla stanovena dvěma metodami, pomocí jiskrového výboje a pomocí plazmy. Oběma metodami bylo paralelně proměřeno 55 vzorků. Určete, zda obě metody vedou ke stejným výsledkům.
Data: Koncentrace železa [%] v hliníkové slitině: *H323a* metoda jiskrového výboje, *H323b* plazma.

H323a:	1.29	1.39	1.37	1.36	1.31	1.30	1.44	1.31	1.45	1.28	1.33
..	..	..	..	..	..	..	..	..	..	..	..

1.32	1.32	1.34	1.41	1.37	1.27	1.31	1.32				
H323b:	1.32	1.37	1.36	1.34	1.30	1.34	1.45	1.35	1.45	1.37	1.33
..	..	..	..	..	..	..	..	..	..	..	..
1.35	1.31	1.34	1.45	1.36	1.31	1.30	1.29				

Úloha H3.24 *Střední hodnota nasákovosti dlaždic*
Na výstupní kontrole keramického závodu se sleduje nasákovost dlažby. Jaká je střední hodnota nasákovosti dlažby 200 × 200 mm, vyjádřená v %, když byla změřena u náhodného výběru 99 dlaždic. Vyšetřete předpoklady, kladené na výběr, dále symetrii rozdělení a určete počet odlehlých hodnot. Který odhad užijete pro střední hodnotu? Je třeba užít transformaci dat?
Data: Nasákovost 99 dlaždic [%]:

2.6	2.0	2.2	3.0	3.2	3.0	3.2	3.4	3.2	2.6	2.3	2.5
...	...	...	...	...	...	...	...	...	...	...	...
1.1	1.1	2.0	1.5	1.8	1.8	1.8	1.7	1.8	2.2	2.1	1.8

Úloha H3.25 *Střední hodnota obsahu SiO_2 v jílu (Horn)*
Sledování obsahu SiO_2 v jílu se při výrobě dlaždic provádí jedenkrát měsíčně a výběry jsou proto malé. Určete střední hodnotu obsahu SiO_2 využitím Hornova postupu a výsledky porovnejte s klasickými a robustními odhady polohy.
Data: Obsah SiO_2 [%] v jílu: 66.55, 67.68, 71.46, 66.40, 68.47.

Úloha H3.26 *Test správnosti a shodnosti obsahu mědi v rudě dvěma metodami*
Měď v rudě byla stanovena dvěma metodami, chelatometricky a jodometricky. Vyšetřete správnost stanoveného obsahu oběma metodami, když standard obsahoval 40.85 % mědi. Na hladině významnosti $\alpha = 0.05$ proveďte také test shodnosti obsahů mědi oběma metodami.
Data: Obsah mědi v rudě [%]: *H326a* chelatometricky, *H326b* jodometricky. Data jsou ve dvojicích:
40.2 42.01, 40.8 41.55, 41.2 40.98, 40.4 42.08, 41.0 41.85, 40.1 41.00.

Úloha H3.27 *Párový test u ztráty žíháním vápence, sledované dvěma metodami*
Párová data obsahovala výsledky stanovení ztráty žíháním vápence jednak gravimetricky a jednak komerčním přístrojem CWA5003. Dosahují obě analytické metody shodných výsledků?
Data: Ztráta žíháním vápence [%]: *H327a* gravimetricky, *H327b* přístrojem CWA5003. Data jsou ve dvojicích:
43.55 43.75, 43.43 43.85, 43.52 43.73, 43.44 43.72, 43.60 43.55, 43.41 43.71, 43.46 43.68, 43.46 43.76, 43.45 43.64, 42.91 43.69, 43.43 43.78, 43.47 43.77, 43.45 43.76, 43.34 43.90, 43.45 43.95, 43.48 43.88.

Úloha H3.28 *Ověření správnosti hodnot pevnosti v tahu u kolejnic dané jakosti*
Byly ověřovány hodnoty pevnosti v tahu u kolejnic dané jakosti a rozměru za období 1 měsíce. Každý vzorek kolejnice o průměru 10 mm se postupně zatěžoval silou F až k přetržení. Byla zaznamenána závislost síly F na poměrném prodloužení tyče a maximální hodnota zatěžovací křivky pak odpovídala pevnosti v tahu R_m [MPa]. Vyšetřete, zda hodnota pevnosti vyšetřované oceli leží v intervalu 920 až 980 MPa, předepsaném normou.
Data: Hodnota pevnosti [MPa] v tahu u kolejnic dané jakosti: 975, 983, 968, 938, 954, 957, 924, 942, 947, 933, 945, 944, 941, 960, 951, 955, 956, 922, 954, 939, 944, 953, 940, 949, 952, 926, 931, 937, 945, 942, 947, 956, 968,

946, 935, 942, 948, 949, 947, 956, 961, 943, 939, 941, 947, 952, 957, 951, 965, 938, 954, 945, 934, 967, 946, 942, 923, 955.

Úloha H3.29 *Shodnost obsahu dusíku v oceli, stanoveného dvěma metodami*
Vzorky oceli oválného tvaru a výšky 1.5 mm, odebrané ze stejného odběrového místa na kyslíkově-konvertorové ocelárně, byly analyzovány termoevolučními (LECO) a optickoemisními (OES) metodami. U termoevoluční metody se třísky oceli spálí v keramickém kelímku a plynné oxidy se vedou inertním plynem argonem ke zdroji infračerveného záření, které oxidy uhlíku absorbují. Plynná směs se vede do vodivostní cely, kde se na základě změny tepelné vodivosti vyhodnotí kvantitativní hodnota. Vyšetřete, zda obě instrumentální metody vedou ke stejným výsledkům.
Data: Obsah dusíku *N*.10000[%] v oceli: *H329a* metodou LECO, *H329b* metodou OES.
H329a: 49 48 55 53 67 46 48 57 56 44 55 59 65 40 41 45 49 49 54 55 61 60 49 52 53 55 54.
H329b: 60 54 61 62 80 50 59 70 69 49 64 69 75 48 48 52 58 59 66 69 68 73 51 58 53 58 65.

Úloha H3.30 *Párový test obsahu niobu v oceli dvěma analytickými metodami*
Niob v oceli je stanovován jednak klasickou metodou emisní spektrometrie s individuálně vázanou plazmou a dále také automatickým emisním spektrometrem. Vyšetřete párová data a rozhodněte, zda obě analytické metody vedou ke stejným závěrům.
Data: Obsah niobu *Nb*.1000 v oceli [%]: *H330a* metodou ICP, *H330b* metodou OES.
H330a: 76 88 87 92 93 91 93 88 93 72 87 70 85 91 90.
H330b: 76 86 86 88 89 86 88 86 91 74 87 75 90 92 89.

3.5.5 Analýza ekonomických a sociologických dat

Úloha S3.01 *Test shodnosti délky anglického a českého slova*
V datech je uvedena délka anglických slov vybraných z knihy (a) Amise (1980) a (b) Škvoreckého (1990). Určete typ diskrétního rozdělení, parametry polohy a rozptýlení.
Data: Délka slova [počet písmen]:
S301a: Amis, Jake's Thing, Penguin Books 1980: 3, 2, 4, 6, 2, 6, 2, 3, 9, 6, 4, 2, 8, 5, 3, 3, 4, 6.
S301b: Škvorecký, Dvě legendy, Primus 1990: 5, 8, 3, 3, 7, 7, 6, 2, 5, 5, 9, 8, 6, 8, 2, 7, 6, 7, 5, 12.

Úloha S3.02 *Určení systematické chyby přístroje testem správnosti*
Skupina studentů měřila vzdálenost dvou orientačních bodů, které byly vybrány tak, aby byly přesně ve vzdálenosti 100.00 m. Studenti naměřili 26 následujících údajů. Na hladině významnosti $\alpha = 0.05$ metodou intervalu spolehlivosti míry polohy otestujte zda přístroj, kterým bylo měření provedeno, nemá systematickou chybu. Jsou ve výběrech horní odlehlé hodnoty, to znamená dosažené rekordy?
Data: Měřená 100 m vzdálenost: 100.02, 100.01, 99.98, 100.09, 100.01, 100.01, 100.01, 99.99, 100.05, 100.03, 100.01, 99.96, 100.02, 100.04, 100.00, 99.98, 99.99, 100.03, 100.07, 100.01, 100.00 , 99.96, 100.01, 100.02, 100.02, 100.00.

Úloha S3.03 *Test shodnosti výdělků taxikářů ve dvou městech*
O hodinových výdělcích taxikářů ze dvou měst máme k dispozici údaje dvou náhodných výběrů. Ověřte na 5% hladině významnosti prohlášení magistrátu, že výše výdělků taxikářů je v obou městech stejná. Jsou ve výběrech odlehlé hodnoty a jsou obě rozdělení symetrická?

Data: Hodinový výdělek [Kč]: *S303a*: 1. město, *S303b*: 2. město

S303a:

168	133	144	106	154	175	141	148	75	125	133	85	168	99	134	183
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
143	88	151	111	145	165	123	155	131	98	148	44				

S303b:

148	127	174	132	125	139	158	140	108	146	125	154	167	132	128	127
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
104	141	171	148	145	151	140	113	147	146	105	141	128	167	152	131

Úloha S3.04 *Párový test kontroly účtované ceny a správné ceny zboží*

Při kontrole na trhu ovoce a zeleniny bylo přesně zváženo 30 nákupů u náhodně vybraných zákazníků. Zjištěné rozdíly v účtované a správné ceně jsou vidět z následujících dat. Lze tvrdit s 5% rizikem omylu, že ceny jsou zkresleny v neprospěch zákazníka? Užijte párový test.

Data: Cena zboží v nákupu [Kč], *S304a* účtovaná cena, *S304b* správná cena. Data jsou ve dvojicích.

7.20	7.20,	3.00	42.90,	12.60	12.40,	7.70	7.80,	42.00	41.40,
..	..	..	..	..	..	..	..	..	..
7.20	7.00,	17.50	17.20,	12.50	12.40,	75.00	74.90,	28.70	28.70

Úloha S3.05 *Test správnosti naměřené rychlosti střely*

Z výrobní série střeliva 1 typu odebereme náhodně 20 ks nábojů, u nichž změříme rychlost střely. Je požadováno, aby rychlost střely byla 300 ± 5 m/s. Rychlost střely se měří pomocí času potřebného k tomu, aby střela protnula 2 světelné paprsky vzdálené od sebe 50 m. Rychlost se potom vypočítá ze vztahu $v = s/t$. Úkolem je rozhodnout, zda série střeliva vyhovuje stanovené normě. K testu správnosti užijte intervalový odhad.

Data: Rychlost střely [m . s^{-1}]: 291, 295, 295, 296, 296, 297, 298, 299, 300, 300, 300, 301, 302, 302, 303, 304, 305, 309, 312, 315.

Úloha S3.06 *Test správnosti pevnosti cihel za sucha a po nasáknutí*

Při hodnocení cihlářských výrobků byly získány údaje týkající se dutinových příčkovek o rozměru 290 × 140 × 65 mm. Jedním z důležitých ukazatelů kvality cihel je stálost pevnosti za sucha i po nasáknutí. Měření bylo provedeno na náhodných výběrech 70 cihel. Úkolem je zjistit, zda existuje rozdíl mezi pevností za sucha a po nasáknutí. Test shodnosti provedete na hladině významnosti $\alpha = 0.05$.

Data: Pevnost cihly [Mpa/10] za sucha a po nasáknutí: *S306a*, za sucha, *S306b*, po nasáknutí.

S306a: 40, 61, 43, 69, 96, 79, 69, 50, 62, 77, 80, 65, 81, 61, 87, 50, 54, 76, 74, 86, 59, 57, 70, 73, 77, 51, 75, 61, 91, 69, 52, 54, 78, 83, 62, 100, 86, 80, 85, 87, 64, 56, 60, 56, 42, 65, 54, 42, 49, 51, 76, 50, 53, 44, 5, 2, 81, 80, 90, 64, 68, 134, 71, 88, 36, 58, 47, 54, 78, 49, 73.

S306b: 33, 49, 44, 96, 41, 92, 107, 48, 46, 75, 70, 49, 70, 59, 48, 60, 52, 70, 75, 66, 53, 84, 71, 59, 67, 49, 60, 88, 52, 82, 56, 106, 76, 52, 89, 73, 49, 80, 65, 80, 58, 53, 44, 48, 51, 51, 55, 48, 49, 46, 66, 40, 45, 84, 52, 55, 67, 97, 72, 51, 62, 61, 80, 96, 53, 77, 77, 92, 76, 64.

Úloha S3.07 *Test shodnosti cen barelu ropy pomocí variačního rozpětí*
Pomocí variačního rozpětí srovnejte variabilitu cen ropy v US$, dovážené do Evropy jednak ze zemí kartelu OPEC, jednak od ostatních vývozců, když jsou dány údaje o cenách dodávek v určitém týdnu. Test shodnosti proveďte na hladině významnosti $\alpha = 0.05$.

Data: Cena barelu ropy [$]: *S307a*: OPEC, *S307b*: ostatní
S307a: OPEC 18.50, 20.00, 19.75, 19.00, 19.50, 20.50, 19.00, 19.25, 18.75, 18.00.
S307b: 19.50, 19.00, 17.50, 19.00, 18.75, 20.00, 20.50, 18.00.

Úloha S3.08 *Zjišťování sklizňových ztrát*
Zjišťování sklizňových ztrát u obilí se provádí následujícím způsobem: prvky výběru jsou plošky o výměře 1 m^2, na nichž se sbírají nesklizené klasy a jednotlivá zrna a zjišťuje se jejich přesná hmotnost. Z údajů v gramech odhadněte střední hodnotu základního souboru, který je představován pozemkem. Určete mez, o které můžeme s 95% jistotou prohlásit, že ji průměrné ztráty na 1 m^2 v základním souboru nepřekročí. Jsou ve výběru odlehlé hodnoty?

Data: Sklizňová ztráta obilí [g/m^2]: 3.2, 6.1, 8.0, 6.5, 5.5, 5.5, 3.2, 6.2, 8.7, 5.6, 7.8, 5.6, 9.5, 6.2, 8.5, 6.3, 6.2, 6.8, 6.6, 6.0.

Úloha S3.09 *Test shodnosti nákladů na údržbu vozidel*
Porovnejte náklady v Kč na údržbu vozidel s benzinovým pohonem a naftovým pohonem v jednom zemědělském podniku, sledované v daném časovém úseku. Získané výsledky považujeme za náhodný výběr z normálního rozdělení. Sestrojte 95% interval spolehlivosti pro rozdíl mezi průměrnými náklady u vozidel obou typů. Nezapomeňte však nejprve vyšetřit homogenitu shodu rozptylů.

Data: Náklady na údržbu vozidel [Kč]: *S309a:* benzinový pohon, *S309b:* naftový pohon
S309a: 1243, 578, 1098, 761, 507, 259, 773, 824, 516, 863, 1122.
S309b: 1226, 1472, 889, 506, 1325, 1290, 1765, 1218.

Úloha S3.10 *Test shodnosti vlivu dvou krmných směsí na přírůstek zvířat*
Při krmení náhodně vybraných zvířat bylo dosaženo rozdílných přírůstků s krmnou směsí A a B za určité časové období. Formulujte závěr o výhodnosti směsí pomocí intervalového odhadu rozdílu ve středních hodnotách přírůstků. Nejprve je třeba prověřit i normalitu obou rozdělení a heteroskedasticitu.

Data: Přírůstek hmotnosti dobytka [kg/ks]: *S310a* krmná směs A, *S310b* krmná směs B.
S310a: A, 26, 29, 24, 29, 20, 22, 16, 23, 20, 21.
S310b: B, 17, 16, 20, 18, 19, 24, 22, 22, 17, 25.

Úloha S3.11 *Test shodnosti kvality lněného vlákna ze dvou pozemků*
Jednou z rozhodujících vlastností ke kontrole jakosti lněného vlákna je hmotnost stonku. Abychom prokázali účinnost jisté metody ošetření lnu, bylo provedeno 20 měření na ošetřeném pozemku a 20 měření na neošetřeném. Proveďte vhodný test na hladině významnosti $\alpha = 0.05$ za předpokladu normality a homoskedasticity.

Data: Hmotnost stonku lnu [g] ze dvou pozemků: *S311a* ošetřený pozemek, *S311b* neošetřený pozemek.
S311a: 47.5, 57.7, 47.1, 38.8, 45.2, 49.8, 43.4, 50.8, 41.5, 38.8, 47.1, 51.2, 52.6, 46.1, 36.0, 44.8, 47.1, 68.3, 62.8, 45.7.

S311b: 49.2, 44.1, 44.1, 38.1, 40.9, 32.1, 36.8, 39.5, 67.2, 41.8, 46.2, 42.3, 50.6, 52.5, 50.6, 49.7, 39.5, 47.4, 57.1, 49.7.

Úloha S3.12 *Test shodnosti počtu rozsvícení zářivek dvou dodavatelů*
Odběratel dostává zářivky od dvou dodavatelů. Při hodnocení kvality zářivek se u malých náhodných výběrů sleduje mimo jiné i počet rozsvícení, který snese zářivka bez poškození. Bývá to okolo 2000 rozsvícení. Od dodavatele A bylo testováno 15 zářivek a od B pak 9 zářivek. Na hladině významnosti $\alpha = 0.05$ ověřte hypotézu, že životnost zářivek v naznačeném smyslu je u obou dodavatelů shodná. Testujte i normalitu a homogenitu.
Data: Počet rozsvícení zářivky od dvou dodavatelů: *S312a* dodavatel A, *S312b* dodavatel B.
S312a: 2139, 2041, 1968, 1903, 1952, 1980, 2089, 1915, 2389, 2163, 2072, 1712, 2018, 1792, 1849.
S312b: 1947, 1602, 1906, 2031, 2072, 1812, 1942, 2074, 2132.

Úloha S3.13 *Test správnosti a shodnosti hmotnosti porcí z balicího automatu před seřízením a po něm*. Seřízením balicího automatu se mělo dosáhnout snížení kolísavosti hmotnosti balených porcí hodnoty 250 g. Na hladině významnosti $\alpha = 0.05$ proveďte potřebný test za předpokladu normality rozdělení a homoskedasticity. Výsledky kontrolních měření před seřízením a po seřízení jsou v gramech. Pracuje zařízení správně a shodně před seřízením i po něm? Mají oba výběry shodný rozptyl a Gaussovo rozdělení?
Data: Hmotnost balených porcí [g] z balicího automatu : *S313a* před seřízením, *S313b* po seřízení.
S313a: 243.2, 244.8, 253.1, 247.5, 251.0, 251.7, 254.0, 252.5, 252.8, 250.1, 247.3, 250.9, 253.2, 252.7.
S313b: 250.4, 250.2, 251.1, 249.9, 250.2, 251.1, 250.8, 249.3, 250.2, 250.3, 250.1, 250.2, 250.0, 249.9, 249.7, 250.1.

Úloha S3.14 *Porovnání intenzity osvětlení naměřené dvěma pracovníky*
Opakovaným měřením intenzity osvětlení školní učebny byla dvěma pracovníky z rozličných kontrolních laboratoří získána data. Vedou obojí měření ke stejným středním hodnotám - shodným výsledkům intenzit osvětlení školní učebny?
Data: Intenzita osvětlení *E* [lx]: *S314a* 1. pracovník, *S314b* 2. pracovník.

50.00	35.0,	209.0	222.0,	230.0	248.0,	246.0	255.0,	248.0	263.0,
..	..	..	..	..	..	..	..	..	..
267.0	0.0,	271.0	0.0,	272.0	0.0,	289.0	0.0,	304.0	0.0,
287.0	0.0,								

Úloha S3.15 *Párový test hodnot intenzity osvětlení dvěma přístroji*
Měření intenzity osvětlení školní učebny bylo provedeno dvěma přístroji luxmetr Krochmann a Mavolux. Párovým testem vyšetřete, zda oba luxmetry naměřily stejné hodnoty.
Data: Intenzita osvětlení *E* [lx] dvěma přístroji: *S315a* Krochmann, *S315b* Mavolux.

211.0	208.0,	247.0	249.0,	255.0	251.0,	264.0	257.0,	273.0	279.0,	255.0	261.0,
..	..	..	..	..	..	..	..	..	..	..	..
269.0	278.0,	279.0	294.0,	310.0	312.0,	317.0	321.0,	266.0	269.0 ,		

3.6 Kontrolní hodnoty (ADSTAT, NCSS2000)

3.6.1 Analýza farmakologických a biochemických dat

B3.01 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 1.736, $\tilde{x}_{0.5}$ = 1.545, Box-Cox: $\bar{x}$ = 1.551, Hornův P_L = 1.905, Hornův R_L = 2.83, s = 1.28, $\hat{g}_1$ = 0.46, $\hat{g}_2$ = 1.99, Závěr: 0.31 < μ < 3.50 (Horn).

B3.02 Homoskedasticita, průměry jsou rozdílné: **B3.02A** Nemocní: Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 2, $\bar{x}$ = 2.10, $\tilde{x}_{0.5}$ = 2.08, Box-Cox: $\bar{x}$ = 2.09, Hornův P_L = 2.14, Hornův R_L = 0.33, s = 0.20, $\hat{g}_1$ = 0.03, $\hat{g}_2$ = 1.77, Závěr: 1.83 < μ < 2.44 (Horn), **B3.02B** Zdraví: Symetrické rozdělení, zamítnuto Gaussovo rozdělení, počet odlehlých hodnot: 2, $\bar{x}$ = 2.30, $\tilde{x}_{0.5}$ = 2.26, Box-Cox: $\bar{x}$ = 2.25, Hornův P_L = 2.24, Hornův R_L = 0.19, s = 0.19, $\hat{g}_1$ = 0.97, $\hat{g}_2$ = 2.81, Závěr: 2.07 < μ < 2.42 (Horn).

B3.03 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 263.8, $\tilde{x}_{0.5}$ = 265.0, Box-Cox: $\bar{x}$ = 264.4, Hornův P_L = 262.5, Hornův R_L = 25.0, s = 11.9, $\hat{g}_1$ = -0.30, $\hat{g}_2$ = 2.02, Závěr: 248.4 < μ < 276.6 (Horn).

B3.04 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 2, $\bar{x}$ = 19.53, $\tilde{x}_{0.5}$ = 19.49, Box-Cox: $\bar{x}$ = 19.49, Hornův P_L = 19.53, Hornův R_L = 0.51, s = 0.26, $\hat{g}_1$ = 0.30, $\hat{g}_2$ = 1.64, Závěr: 19.25 < μ < 19.80 (Horn).

B3.05 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 2, $\bar{x}$ = 99.11, $\tilde{x}_{0.5}$ = 99.00, Box-Cox: $\bar{x}$ = 99.03, Hornův P_L = 99.10, Hornův R_L = 0.80, s = 1.02, $\hat{g}_1$ = 0.53, $\hat{g}_2$ = 3.18, Závěr: 98.52 < μ < 99.68 (Horn).

B3.06 Homogenita, průměry jsou shodné: **B3.06A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 1, $\bar{x}$ = 0.3264, $\tilde{x}_{0.5}$ = 0.3270, Box-Cox: $\bar{x}$ = 0.3265, Hornův P_L = 0.3255, Hornův R_L = 0.0130, s = 0.0082, $\hat{g}_1$ = -0.10, $\hat{g}_2$ = 2.07, Závěr: 0.319 < μ < 0.332 (Horn), **B3.06B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 2, $\bar{x}$ = 0.3273, $\tilde{x}_{0.5}$ = 0.3280, Box-Cox: $\bar{x}$ = 0.3278, Hornův P_L = 0.3280, Hornův R_L = 0.0060, s = 0.0059, $\hat{g}_1$ = -0.61, $\hat{g}_2$ = 3.86, Závěr: 0.325 < μ < 0.331 (Horn).

B3.07 Homogenita, párový test - střední hodnoty jsou shodné: **B3.07A** Asymetrické rozdělení, zamítnuto Gaussovo rozdělení, počet odlehlých hodnot: 2, $\bar{x}$ = 93.88, $\tilde{x}_{0.5}$ = 93.75, Box-Cox: $\bar{x}$ = 94.14, s = 2.21, $\hat{g}_1$ = -1.08, $\hat{g}_2$ = 5.04, Závěr: 93.34 < μ < 94.96 (Box-Cox), **B3.07B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 5, $\bar{x}$ = 94.00, $\tilde{x}_{0.5}$ = 93.60, Box-Cox: $\bar{x}$ = 93.93, s = 2.15, $\hat{g}_1$ = 0.08, $\hat{g}_2$ = 2.08, Závěr: 93.09 < μ < 94.82 (Box-Cox).

B3.08A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 3.19, $\tilde{x}_{0.5}$ = 3.19, Box-Cox: $\bar{x}$ = 3.19, Hornův P_L = 3.19, Hornův R_L = 0.04, s = 0.03, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 2.31, Závěr: 3.16 < μ < 3.22 (Horn).

B3.08B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 1, $\bar{x}$ = 1.57, $\tilde{x}_{0.5}$ = 1.60, Box-Cox: $\bar{x}$ = 1.60, Hornův P_L = 1.60, Hornův R_L = 0.03, s = 0.03, $\hat{g}_1$ = 0.13, $\hat{g}_2$ = 2.73, Závěr: 1.58 < μ < 1.62 (Horn).

B3.09 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 2, $\bar{x}$ = 19.995, $\tilde{x}_{0.5}$ = 20.000, Box-Cox: $\bar{x}$ = 19.997, Hornův P_L = 19.995, Hornův R_L = 0.030, s = 0.033, $\hat{g}_1$ = -0.43,

$\hat{g}_2$ = 2.99, Závěr: 19.975 < μ < 20.015 (Horn).

B3.10AB 1. laborantka: nehomogenita, průměry jsou shodné, **B3.10A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 1.501, $\tilde{x}_{0.5}$ = 1.503, Box-Cox: $\bar{x}$ = 1.501, Hornův P_L = 1.500, Hornův R_L = 0.020, s = 0.015, $\hat{g}_1$ = 0.05, $\hat{g}_2$ = 1.87, Závěr: 1.490 < μ < 1.513 (Horn), **B3.10B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 2, $\bar{x}$ = 1.499, $\tilde{x}_{0.5}$ = 1.500, Box-Cox: $\bar{x}$ = 1.499, Hornův P_L = 1.499, Hornův R_L = 0.012, s = 0.006, $\hat{g}_1$ = 0.11, $\hat{g}_2$ = 1.88, Závěr: 1.491 < μ < 1.507 (Horn).

B3.10CD 2. laborantka: homogenita, průměry jsou shodné: **B3.10C** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 9.791, $\tilde{x}_{0.5}$ = 9.795, Box-Cox: $\bar{x}$ = 9.791,

Hornův P_L = 9.790, Hornův R_L = 0.040, s = 0.026, $\hat{g}_1$ = -0.11, $\hat{g}_2$ = 1.92,

Závěr: 9.763 < μ < 9.817 (Horn), **B3.10D** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 9.774, $\tilde{x}_{0.5}$ = 9.775, Box-Cox: $\bar{x}$ = 9.777, Hornův P_L = 9.780,

Hornův R_L = 0.040, s = 0.028, $\hat{g}_1$ = -0.58, $\hat{g}_2$ = 2.39, Závěr: 9.753 < μ < 9.807 (Horn).

B3.11 Homogenita, průměry jsou shodné: **B3.11A** Asymetrické rozdělení, zamítnuto Gaussovo rozdělení, počet odlehlých hodnot: asi 2, $\bar{x}$ = 0.848, $\tilde{x}_{0.5}$ = 0.459, Box-Cox: $\bar{x}$ = 0.563, Hornův P_L = 0.776, Hornův R_L = 1.229, s = 0.955, $\hat{g}_1$ = 1.06, $\hat{g}_2$ = 2.57, Závěr: -0.497 < μ < 2.048 (Horn), **B3.11B,** počet odlehlých hodnot: , $\bar{x}$ = , $\tilde{x}_{0.5}$ = , Box-Cox: $\bar{x}$ = 0.553, Hornův P_L = 0.775, Hornův R_L = 1.246, -0.515 < μ < 2.065.

B3.12 Asymetrické rozdělení, zamítnuto Gaussovo rozdělení, počet odlehlých hodnot: asi 2, $\bar{x}$ = 2.33, $\tilde{x}_{0.5}$ = 2.30, Box-Cox: $\bar{x}$ = 2.33, Hornův P_L = 2.35, Hornův R_L = 0.30, s = 0.15, $\hat{g}_1$ = -0.20, $\hat{g}_2$ = 1.85, Závěr: 2.134 < μ < 2.566 (Horn).

B3.13 Asymetrické rozdělení, zamítnuto Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 0.484, $\tilde{x}_{0.5}$ = 0.480, Box-Cox: $\bar{x}$ = 0.490, Hornův P_L = 0.490, Hornův R_L = 0.020, s = 0.023, $\hat{g}_1$ = -0.41, $\hat{g}_2$ = 2.07, Závěr: 0.448 < μ < 0.532 (Horn).

B3.14 Asymetrické rozdělení, zamítnuto Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 2.33, $\tilde{x}_{0.5}$ = 2.30, Box-Cox: $\bar{x}$ = 2.33, Hornův P_L = 2.35, Hornův R_L = 0.30, s = 0.15, $\hat{g}_1$ = -0.20, $\hat{g}_2$ = 1.85, Závěr: 2.134 < μ < 2.566 (Horn).

B3.15 Homogenita, průměry jsou shodné: **B3.15A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 1.279, $\tilde{x}_{0.5}$ = 1.275, Box-Cox: $\bar{x}$ = 1.277, Hornův P_L = 1.280, Hornův R_L = 0.040, s = 0.020, $\hat{g}_1$ = 0.18, $\hat{g}_2$ = 1.91, Závěr: 1.257 < μ < 1.303 (Horn), **B3.15B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 1.268, $\tilde{x}_{0.5}$ = 1.265, Box-Cox: $\bar{x}$ = 1.266, Hornův P_L = 1.265, Hornův R_L = 0.030, s = 0.019, $\hat{g}_1$ = 0.25, $\hat{g}_2$ = 2.26,

Závěr: 1.248 < μ < 1.282 (Horn).

B3.16A Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\bar{x}$ = 0.349, $\tilde{x}_{0.5}$ = 0.350, Box-Cox: $\bar{x}$ = 0.348, Hornův P_L = 0.350, Hornův R_L = 0.020, s = 0.012, $\hat{g}_1$ = 0.24, $\hat{g}_2$ = 2.40, Závěr: 0.339 < μ < 0.361 (Horn).

B3.16B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 1, $\bar{x}$ = 0.349, $\tilde{x}_{0.5}$ = 0.350, Box-Cox: $\bar{x}$ = 0.350, Hornův P_L = 0.350, Hornův R_L = 0.020, s = 0.011, $\hat{g}_1$ = -0.39, $\hat{g}_2$ = 1.86, Závěr: 0.339 < μ < 0.361 (Horn).

B3.17 Homogenita, průměry jsou shodné: **B3.17A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\overline{x}$ = 6.558, $\tilde{x}_{0.5}$ = 6.560, Box-Cox: $\overline{x}$ = 6.557, Hornův P_L = 6.550, Hornův R_L = 0.060, s = 0.032, $\hat{g}_1$ = -0.01, $\hat{g}_2$ = 2.34, Závěr: 6.516 < μ < 6.584 (Horn), **B3.17B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\overline{x}$ = 6.543, $\tilde{x}_{0.5}$ = 6.545, Hornův P_L = 6.535, Hornův R_L = 0.070, s = 0.036, $\hat{g}_1$ = -0.02, $\hat{g}_2$ = 2.15, Závěr: 6.496 < μ < 6.574 (Horn).

B3.18 Heterogenita, průměry jsou shodné: **B3.18A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\overline{x}$ = 23.53, $\tilde{x}_{0.5}$ = 23.60, Box-Cox: $\overline{x}$ = 23.51, s = 0.74, $\hat{g}_1$ = 0.08, $\hat{g}_2$ = 2.64, Závěr: 23.25 < μ < 23.78 (Box-Cox), **B3.18B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: asi 2, $\overline{x}$ = 23.44, $\tilde{x}_{0.5}$ = 23.50, Box-Cox: $\overline{x}$ = 23.42, s = 1.07, $\hat{g}_1$ = 0.04, $\hat{g}_2$ = 2.22, Závěr: 23.05 < μ < 23.82 (Box-Cox).

B3.19A Poissonovo,

B3.19B Poissonovo,

B3.20 Homogenita, průměry jsou shodné: **B3.20A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: 0, $\overline{x}$ = 12.840, $\tilde{x}_{0.5}$ = 13.600, Box-Cox: $\overline{x}$ = 13.010, s = 2.890, $\hat{g}_1$ = -0.37, $\hat{g}_2$ = 2.70, Závěr: 11.840 < μ < 14.190 (Box-Cox), **B3.20B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: asi 4, $\overline{x}$ = 15.100, $\tilde{x}_{0.5}$ = 14.520, Box-Cox: $\overline{x}$ = 15.000, s = 2.260, $\hat{g}_1$ = 0.34, $\hat{g}_2$ = 3.31, Závěr: 14.230 < μ < 15.770 (Box-Cox).

B3.21 Asymetrické rozdělení, zamítnuto Gaussovo rozdělení, počet odlehlých hodnot: asi 2, $\overline{x}$ = 5.163, $\tilde{x}_{0.5}$ = 5.100, Box-Cox: $\overline{x}$ = 5.137, Hornův P_L = 5.170, Hornův R_L = 0.260, s = 0.137, $\hat{g}_1$ = 0.24, $\hat{g}_2$ = 1.17, Závěr: 4.983 < μ < 5.357 (Horn).

B3.22 Homogenita, průměry jsou shodné**: 322A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: asi 4, $\overline{x}$ = 24.913, $\tilde{x}_{0.5}$ = 24.935, Box-Cox: $\overline{x}$ = 24.928, Hornův P_L = 24.950, Hornův R_L = 0.640, s = 0.494, $\hat{g}_1$ = -0.20, $\hat{g}_2$ = 2.71, Závěr: 24.672 < μ < 25.228 (Horn), **B3.22B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: asi 3, $\overline{x}$ = 24.898, $\tilde{x}_{0.5}$ = 24.980, Box-Cox: $\overline{x}$ = 24.885, Hornův P_L = 24.780, Hornův R_L = 1.260, s = 0.647, $\hat{g}_1$ = -0.01, $\hat{g}_2$ = 1.90, Závěr: 24.232 < μ < 25.328 (Horn).

B3.23 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: asi 2, $\overline{x}$ = 10.038, $\tilde{x}_{0.5}$ = 10.040, Box-Cox: $\overline{x}$ = 10.046, Hornův P_L = 10.050, Hornův R_L = 0.120, s = 0.070, $\hat{g}_1$ = -0.59, $\hat{g}_2$ = 2.67, Závěr: 9.992 < μ < 10.108 (Horn).

B3.24 Homogenita, průměry jsou rozdílné: **B3.24A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: asi 1, $\overline{x}$ = 4.894, $\tilde{x}_{0.5}$ = 4.895, Box-Cox: $\overline{x}$ = 4.894, Hornův P_L = 4.910, Hornův R_L = 0.100, s = 0.063, $\hat{g}_1$ = -0.05, $\hat{g}_2$ = 1.90, Závěr: 4.843 < μ < 4.977 (Horn), **B3.24B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých hodnot: asi 2, $\overline{x}$ = 4.970, $\tilde{x}_{0.5}$ = 4.985, Box-Cox: $\overline{x}$ = 4.963, Hornův P_L = 4.955, Hornův R_L = 0.150, s = 0.089, $\hat{g}_1$ = 0.15, $\hat{g}_2$ = 1.71, Závěr: 4.855 < μ < 5.055 (Horn).

3.6.2 Analýza chemických a fyzikálních dat

C3.01 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 5.160, $\tilde{x}_{0.5}$ = 5.160, $\overline{x}_R$ = 5.160, Box-Cox transf. není nutná, Horn P_L = 5.170, Horn R_L = 0.040, s = 0.030, $\hat{g}_1$ = -0.10, $\hat{g}_2$ = 2.18, Závěr: 5.141 < μ < 5.199.

C3.02 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.294, $\tilde{x}_{0.5}$ = 0.294, $\bar{x}_R$ = 0.294, Box-Cox transf. není nutná, Horn P_L = 0.294, Horn R_L = 0.008, s = 0.005, $\hat{g}_1$ = 0.32, $\hat{g}_2$ = 1.74, Závěr: 0.289 < μ < 0.300.

C3.03 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 10.190, $\tilde{x}_{0.5}$ = 10.230, $\bar{x}_R$ = 10.210, Box-Cox transf. není nutná, Horn P_L = 10.220, Horn R_L = 0.070, s = 0.070, $\hat{g}_1$ = -0.83, $\hat{g}_2$ = 2.63, Závěr: 10.0684 < μ < 10.3616.

C3.04 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.0011, $\tilde{x}_{0.5}$ = 0.0011, $\bar{x}_R$ = , Box-Cox transf. není nutná, Horn P_L = 0.0011, Horn R_L = 0.0002, s = , $\hat{g}_1$ = -0.30, $\hat{g}_2$ = 2.37, Závěr: 0.0010 < μ < 0.0012.

C3.05A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 1779.6, $\tilde{x}_{0.5}$ = 1782.3, $\bar{x}_R$ = 1783.2, Box-Cox transf. není nutná, Horn P_L = 1783.0, Horn R_L = 30.1, $\hat{g}_1$ = -1.00, $\hat{g}_2$ = 4.06, Závěr: 1762.8 < μ < 1803.1, **C3.05B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 248.5, $\tilde{x}_{0.5}$ = 247.5, $\bar{x}_R$ = 247.2, Box-Cox transf. je nutná, P_L = 247.2, Horn R_L = 4.3, $\hat{g}_1$ = 1.85, $\hat{g}_2$ = 5.73, Závěr: 244.3 < μ < 250.0.

C3.06 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 112.29, $\tilde{x}_{0.5}$ = 112.31, $\bar{x}_R$ = 112.29, Box-Cox transf. není nutná, Horn P_L = 112.29, Horn R_L = 0.11, $\hat{g}_1$ = -0.50, $\hat{g}_2$ = 1.50, Závěr: 112.17 < μ < 112.40.

C3.07 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 1.660, $\tilde{x}_{0.5}$ = 1.660, $\bar{x}_R$ = 1.660, Box-Cox transf. není nutná, Horn P_L = 1.655, Horn R_L = 0.030, $\hat{g}_1$ = 0.23, $\hat{g}_2$ = 2.03, Závěr: 1.639 < μ < 1.671.

C3.08 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 1250.0, $\tilde{x}_{0.5}$ = 1250.0, $\bar{x}_R$ = 1250.0, Box-Cox transf. není nutná, Horn P_L = 1250.0, Horn R_L = 0.2, $\hat{g}_1$ = -0.51, $\hat{g}_2$ = 2.36, Závěr: 1249.9 < μ < 1250.1.

C3.09 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 10.08, $\tilde{x}_{0.5}$ = 10.05, $\bar{x}_R$ = 10.05, Box-Cox transf. je nutná, P_L = 10.05, Horn R_L = 0.04, $\hat{g}_1$ = 1.63, $\hat{g}_2$ = 3.91, Závěr: 10.01 < μ < 10.09.

C3.10 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 223.1, $\tilde{x}_{0.5}$ = 223.5, $\bar{x}_R$ = 223.1, Box-Cox transf. není nutná, Horn P_L = 223.0, Horn R_L = 18.0, s = 12.3, $\hat{g}_1$ = -0.07, $\hat{g}_2$ = 1.84, Závěr: 211.0 < μ < 235.0.

C3.11 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 2.516, $\tilde{x}_{0.5}$ = 2.500, $\bar{x}_R$ = 2.512, Box-Cox transf. není nutná, Horn P_L = 2.510, Horn R_L = 0.220, s = 0.1150, $\hat{g}_1$ = 0.04, $\hat{g}_2$ = 1.78, Závěr: 2.352 < μ < 2.668.

C3.12 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 1.0510, $\tilde{x}_{0.5}$ = 1.0550, $\bar{x}_R$ = 1.0530, Box-Cox transf. není nutná, Horn P_L = 1.0480, Horn R_L = 0.0750, s = 0.0450, $\hat{g}_1$ = -0.22, $\hat{g}_2$ = 1.96, Závěr: 0.9974 < μ < 1.0976.

C3.13 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 8.872, $\tilde{x}_{0.5}$ = 8.920, $\bar{x}_R$ = 8.894, Box-Cox transf. není nutná, Horn P_L = 8.845,

Horn R_L = 0.190, s = 0.097, $\hat{g}_1$ = -0.83, $\hat{g}_2$ = 2.01, Závěr: 8.742 < μ < 8.949.

C3.14 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 4.065, $\tilde{x}_{0.5}$ = 4.070, $\overline{x}_R$ = 4.061, Box-Cox transf. není nutná, Horn P_L = 4.060, Horn R_L = 0.068, s = 0.050, $\hat{g}_1$ = 0.26, $\hat{g}_2$ = 1.98, Závěr: 4.015 < μ < 4.105.

C3.15 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 2.32, $\tilde{x}_{0.5}$ = 2.40, $\overline{x}_R$ = 2.41, Box-Cox transf. není nutná, Horn P_L = 2.34, Horn R_L = 0.28, s = 0.26, $\hat{g}_1$ = -1.17, $\hat{g}_2$ = 3.39, Závěr: 2.08 < μ < 2.60.

C3.16 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 0.366, $\tilde{x}_{0.5}$ = 0.350, $\overline{x}_R$ = 0.352, Box-Cox transf. není nutná, Horn P_L = 0.355, Horn R_L = 0.050, s = 0.058, $\hat{g}_1$ = 0.87, $\hat{g}_2$ = 2.78, Závěr: 0.309 < μ < 0.400.

C3.17 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 49.55, $\tilde{x}_{0.5}$ = 49.49, $\overline{x}_R$ = 49.40, Box-Cox transf. není nutná, Horn P_L = 49.36, Horn R_L = 1.26, s = 1.46, $\hat{g}_1$ = 0.45, $\hat{g}_2$ = 2.45, Závěr: 48.06 < μ < 50.66.

C3.18 Heteroskedasticita, průměry jsou významně rozdílné: **C3.18A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 0.9969, $\tilde{x}_{0.5}$ = 0.9972, $\overline{x}_R$ = 0.9970, Box-Cox transf. není nutná, Horn P_L = 0.9968, Horn R_L = 0.0017, s = 0.0009, $\hat{g}_1$ = -0.38, $\hat{g}_2$ = 2.02, Závěr: 0.9960 < μ < 0.9977, **C3.18B** počet odlehlých bodů: 0, $\overline{x}$ = 0.9995, $\tilde{x}_{0.5}$ = 0.9996, $\overline{x}_R$ = 0.9995, Box-Cox transf. není nutná, Horn P_L = 0.9995, Horn R_L = 0.0004, s = 0.0002, $\hat{g}_1$ = -0.96, $\hat{g}_2$ = 3.13, Závěr: 0.9993 < μ < 0.9997.

C3.19 Prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 2.043, $\tilde{x}_{0.5}$ = 2.040, $\overline{x}_R$ = 2.044, Box-Cox transf. není nutná, Horn P_L = 2.045, Horn R_L = 0.010, s = 0.016, $\hat{g}_1$ = -0.35, $\hat{g}_2$ = 3.16, Závěr: 2.038 < μ < 2.052.

C3.20 Prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 93.98, $\tilde{x}_{0.5}$ = 94.05, $\overline{x}_R$ = 94.01, Box-Cox transf. není nutná, Horn P_L = 94.00, Horn R_L = 1.20, s = 0.66, $\hat{g}_1$ = -0.25, $\hat{g}_2$ = 2.36, Závěr: 93.48 < μ < 94.52.

C3.21 Homogenita, průměry jsou shodné: **C3.21A** Prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 3.198, $\tilde{x}_{0.5}$ = 3.180, $\overline{x}_R$ = 3.189, Box-Cox transf. není nutná, Horn P_L = 3.200, Horn R_L = 0.240, s = 0.182, $\hat{g}_1$ = 0.16, $\hat{g}_2$ = 2.25, Závěr: 3.092 < μ < 3.308, **C3.21B,** počet odlehlých bodů: 0, $\overline{x}$ = 3.253, $\tilde{x}_{0.5}$ = 3.240, $\overline{x}_R$ = 3.183, Box-Cox transf. je nutná, P_L = 3.175, Horn R_L = 0.310, s = 0.250, $\hat{g}_1$ = 0.95, $\hat{g}_2$ = 2.92, Závěr: 3.035 < μ < 3.315.

C3.22 Homogenita, průměry jsou shodné: **C3.22A,** počet odlehlých bodů: 0, $\overline{x}$ = 48.67, $\tilde{x}_{0.5}$ = 47.00, $\overline{x}_R$ = 47.83, Box-Cox transf. není nutná, Horn P_L = 49.10, Horn R_L = 4.60, s = 2.50, $\hat{g}_1$ = 0.28, $\hat{g}_2$ = 1.22, Závěr: 45.79 < μ < 52.41, **C3.22B,** počet odlehlých bodů: 0, $\overline{x}$ = 46.30, $\tilde{x}_{0.5}$ = 46.50, $\overline{x}_R$ = 46.72, Box-Cox transf. není nutná, Horn P_L = 47.50, Horn R_L = 3.00, s = 3.83, $\hat{g}_1$ = -0.64, $\hat{g}_2$ = 2.75, Závěr: 45.50 < μ < 49.50.

C3.23 Homogenita, průměry jsou rozdílné: **C3.23A,** počet odlehlých bodů: 0, $\overline{x}$ = 231.8, $\tilde{x}_{0.5}$ = 230.5, $\overline{x}_R$ = nelze, Box-Cox transf. není nutná, s = 6.4, $\hat{g}_1$ = -0.07, $\hat{g}_2$ = 2.54,

Závěr: 228.9 < μ < 234.7, **C3.23B,** počet odlehlých bodů: 0, $\bar{x}$ = 227.3, $\tilde{x}_{0.5}$ = 227.5000, $\bar{x}_R$ = 226.5, Box-Cox transf. není nutná, s = 5.8, $\hat{g}_1$ = 0.50, $\hat{g}_2$ = 2.37, Závěr: 224.7 < μ < 229.9.

C3.24 Homogenita, průměry jsou významně rozdílné: **C3.24A,** počet odlehlých bodů: 0, $\bar{x}$ = 1.018, $\tilde{x}_{0.5}$ = 1.090, $\bar{x}_R$ = 1.021, Box-Cox transf. není nutná, s = 0.420, $\hat{g}_1$ = -0.11, $\hat{g}_2$ = 1.79, Závěr: 0.843 < μ < 1.193, **C3.24B,** počet odlehlých bodů: 0, $\bar{x}$ = 0.919, $\tilde{x}_{0.5}$ = 0.873, $\bar{x}_R$ = 0.860, Box-Cox transf. není nutná, s = 0.500, $\hat{g}_1$ = 0.31, $\hat{g}_2$ = 1.90, Závěr: 0.714 < μ < 1.124.

C3.25 Počet odlehlých bodů: 0, $\bar{x}$ = 25.03, $\tilde{x}_{0.5}$ = 24.74, $\bar{x}_R$ = 24.97, Box-Cox transf. není nutná, s = 0.93, $\hat{g}_1$ = 0.35, $\hat{g}_2$ = 2.96, Závěr: 24.76 < μ < 25.30.

C3.26 Homogenita, průměry jsou shodné: **C3.26A,** počet odlehlých bodů: 0, $\bar{x}$ = 14.797, $\tilde{x}_{0.5}$ = 14.905 $\bar{x}_R$ = 14.921, Box-Cox transf. je nutná, Horn P_L = 14.915, Horn R_L = 0.650, s = 0.610, $\hat{g}_1$ = -1.62, $\hat{g}_2$ = 5.03, Závěr: 14.481 < μ < 15.349, **C3.26B,** počet odlehlých bodů: 0, $\bar{x}$ = 14.594, $\tilde{x}_{0.5}$ = 14.970, $\bar{x}_R$ = 14.793, Box-Cox transf. není nutná, Horn P_L = 14.415, Horn R_L = 1.370, s = 0.766, $\hat{g}_1$ = -0.89, $\hat{g}_2$ = 2.04, Závěr: 13.500 < μ < 15.330.

C3.27 Počet odlehlých bodů: 0, $\bar{x}$ = 80.51, $\tilde{x}_{0.5}$ = 80.50, $\bar{x}_R$ = 80.50, Box-Cox transf. není nutná, s = 0.41, $\hat{g}_1$ = 0.18, $\hat{g}_2$ = 2.92, Závěr: 80.33 < μ < 80.69.

C3.28 Homogenita, průměry jsou rozdílné: **C3.28A,** počet odlehlých bodů: 0, $\bar{x}$ = 50.32, $\tilde{x}_{0.5}$ = 50.31, $\bar{x}_R$ = 50.32, Box-Cox transf. není nutná, s = 0.24, $\hat{g}_1$ = 0.17, $\hat{g}_2$ = 3.47, Závěr: 50.22 < μ < 50.43, **C3.28B,** počet odlehlých bodů: 0, $\bar{x}$ = 50.85, $\tilde{x}_{0.5}$ = 50.82, $\bar{x}_R$ = 50.82, Box-Cox transf. není nutná, s = 0.3200, $\hat{g}_1$ = 0.96, $\hat{g}_2$ = 5.08, Závěr: 50.70 < μ < 50.95.

C3.29 Heterogenita, průměry jsou shodné: **C3.29A,** počet odlehlých bodů: 0, $\bar{x}$ = 75.84, $\tilde{x}_{0.5}$ = 75.90, $\bar{x}_R$ = 75.84, Box-Cox transf. není nutná, Horn P_L = 75.80, Horn R_L = 0.20, s = 0.16, $\hat{g}_1$ = -0.14, $\hat{g}_2$ = 2.23, Závěr: 75.69 < μ < 75.91, **C3.29B,** počet odlehlých bodů: 0, $\bar{x}$ = 75.93, $\tilde{x}_{0.5}$ = 76.00, $\bar{x}_R$ = 75.96, Box-Cox transf. není nutná, Horn P_L = 76.00, Horn R_L = 0.20, s = 0.16, $\hat{g}_1$ = 0.31, $\hat{g}_2$ = 2.97, Závěr: 75.89 < μ < 76.11.

C3.30 Homogenita, průměry jsou rozdílné: **C3.30A,** počet odlehlých bodů: 0, $\bar{x}$ = 40.11, $\tilde{x}_{0.5}$ = 40.10, $\bar{x}_R$ = 40.1100Box-Cox transf. není nutná, s = 0.19, $\hat{g}_1$ = 0.23, $\hat{g}_2$ = 3.39, Závěr: 40.03 < μ < 40.19, **C3.30B,** počet odlehlých bodů: 0, $\bar{x}$ = 40.66, $\tilde{x}_{0.5}$ = 40.64, $\bar{x}_R$ = 40.64, Box-Cox transf. není nutná, s = 0.25, $\hat{g}_1$ = 1.02, $\hat{g}_2$ = 5.39, Závěr: 40.54 < μ < 40.74.

C3.31 Homogenita, průměry jsou rozdílné: **C3.31A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 934.3, $\tilde{x}_{0.5}$ = 918.0, $\bar{x}_R$ = 923.5, Box-Cox transf. není nutná, s = 206.9, $\hat{g}_1$ = 0.20, $\hat{g}_2$ = 2.34, Závěr: 841.1 < μ < 1011.3, **C3.31B,** počet odlehlých bodů: 0, $\bar{x}$ = 761.5, $\tilde{x}_{0.5}$ = 715.0, $\bar{x}_R$ = 742.7, Box-Cox transf. není nutná, s = 140.0, $\hat{g}_1$ = 0.57, $\hat{g}_2$ = 2.48, Závěr: 693.8 < μ < 795.6.

C3.32 Homogenita, průměry jsou shodné: **C3.32A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 1.169, $\tilde{x}_{0.5}$ = 0.485, $\bar{x}_R$ = 0.801, Box-Cox transf. je nutná, Horn P_L = 1.155, Horn R_L = 1.530, s = 1.161, $\hat{g}_1$ = 0.95, $\hat{g}_2$ = 2.51, Závěr: 0.133 < μ < 2.177, **C3.32B,** počet odlehlých bodů: 0,

$\bar{x}$ = 1.433, $\tilde{x}_{0.5}$ = 1.115, $\bar{x}_R$ = 1.150, Box-Cox transf. není nutná, Horn P_L = 1.300, Horn R_L = 1.900, s = 1.277, $\hat{g}_1$ = 0.47, $\hat{g}_2$ = 1.69, Závěr: 0.031 < μ < 2.569.

C3.33 $\bar{x}$ = 20.26, $\tilde{x}_{0.5}$ = 20.27, $\bar{x}_R$ = 20.27, Box-Cox transf. není nutná, Horn P_L = 20.27, Horn R_L = 0.09, s = 0.08, $\hat{g}_1$ = -0.40, $\hat{g}_2$ = 2.89, Závěr: 20.23 < μ < 20.31.

C3.34 $\bar{x}$ = 0.5018, $\tilde{x}_{0.5}$ = 0.5020, $\bar{x}_R$ = 0.5017, Box-Cox transf. není nutná, Horn P_L = 0.5015, Horn R_L = 0.0030, s = 0.0019, $\hat{g}_1$ = 0.20, $\hat{g}_2$ = 2.14, Závěr: 0.5000 < μ < 0.5035.

C3.35 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 7.05, $\tilde{x}_{0.5}$ = 7.50, $\bar{x}_R$ = 7.37, Box-Cox transf. není nutná, Horn P_L = 7.10, Horn R_L = 3.20, s = 2.00, $\hat{g}_1$ = -0.80, $\hat{g}_2$ = 2.61, Závěr: 5.30 < μ < 8.90.

C3.36 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.121, $\tilde{x}_{0.5}$ = 0.120, $\bar{x}_R$ = 0.118, Box-Cox transf. není nutná, Horn P_L = 0.120, Horn R_L = 0.020, s = 0.020, $\hat{g}_1$ = 0.93, $\hat{g}_2$ = 3.93, Závěr: 0.110 < μ < 0.129.

C3.37 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 26.88, $\tilde{x}_{0.5}$ = 26.80, $\bar{x}_R$ = 26.86, Box-Cox transf. není nutná, Horn P_L = 26.90, Horn R_L = 0.20, s = 0.13, $\hat{g}_1$ = 0.38, $\hat{g}_2$ = 1.93, Závěr: 26.81 < μ < 26.99.

C3.38 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 10.0008, $\tilde{x}_{0.5}$ = 10.0020, $\bar{x}_R$ = 10.0016, Box-Cox transf. není nutná, Horn P_L = 10.0015, Horn R_L = 0.0044, s = 0.0040, $\hat{g}_1$ = -1.04, $\hat{g}_2$ = 2.86, Závěr: 9.9986 < μ < 10.0044.

C3.39 Homogenita, průměry jsou shodné: **C3.39A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.22, $\tilde{x}_{0.5}$ = 0.20, $\bar{x}_R$ = 0.21, Box-Cox transf. není nutná, Horn P_L = 0.23, Horn R_L = 0.06, s = 0.04, $\hat{g}_1$ = 0.90, $\hat{g}_2$ = 2.48, Závěr: 0.20 < μ < 0.26. **C3.39B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.33, $\tilde{x}_{0.5}$ = 0.28, $\bar{x}_R$ = 0.29, Box-Cox transf. není nutná, Horn P_L = 0.32, Horn R_L = 0.15, s = 0.13, $\hat{g}_1$ = 1.00, $\hat{g}_2$ = 2.66, Závěr: 0.16 < μ < 0.47.

C3.40 Homogenita, průměry jsou shodné: **C3.40A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 60.51, $\tilde{x}_{0.5}$ = 60.50, $\bar{x}_R$ = 60.47, Box-Cox transf. není nutná, Horn P_L = 60.50, Horn R_L = 0.34, s = 0.18, $\hat{g}_1$ = 0.75, $\hat{g}_2$ = 2.58, Závěr: 60.31 < μ < 60.69, **C3.40B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 60.50, $\tilde{x}_{0.5}$ = 60.57, $\bar{x}_R$ = 60.51, Box-Cox transf. není nutná, Horn P_L = 60.43, Horn R_L = 0.37, s = 0.25, $\hat{g}_1$ = -0.23, $\hat{g}_2$ = 1.78, Závěr: 60.04 < μ < 60.81.

C3.41 Homogenita, průměry jsou shodné: **C3.41A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 30.359, $\tilde{x}_{0.5}$ = 30.445, $\bar{x}_R$ = 30.402, Box-Cox transf. není nutná, Horn P_L = 30.400, Horn R_L = 0.460, s = 0.383, $\hat{g}_1$ = -0.56, $\hat{g}_2$ = 2.37, Závěr: 30.217 < μ < 30.583, **C3.41B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: , $\bar{x}$ = 30.272, $\tilde{x}_{0.5}$ = 30.165, $\bar{x}_R$ = 30.278, Box-Cox transf. není nutná, Horn P_L = 30.275, Horn R_L = 0.750, s = 0.516, $\hat{g}_1$ = -0.10, $\hat{g}_2$ = 2.32, Závěr: 29.977 < μ < 30.573.

C3.42AB Fe: homogenita, průměry jsou shodné: **C3.42A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 37.39, $\tilde{x}_{0.5}$ = 37.47, $\bar{x}_R$ = nelze, ,P_L = 37.33, Horn R_L = 1.22, s = 0.64,

$\hat{g}_1$ = -0.08, $\hat{g}_2$ = 1.48, Závěr: 36.52 < μ < 38.15, **C3.42B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 37.67, $\tilde{x}_{0.5}$ = 37.68, $\bar{x}_R$ = 37.69, Box-Cox transf. není nutná, Horn P_L = 37.62, Horn R_L = 0.77, s = 0.37, $\hat{g}_1$ = -0.22, $\hat{g}_2$ = 1.56, Závěr: 37.10 < μ < 38.13.

C3.42CD Mn: nehomogenita, průměry jsou shodné: **C3.42C** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.362, $\tilde{x}_{0.5}$ = 0.360, $\bar{x}_R$ = 0.363, Box-Cox transf. není nutná, Horn P_L = 0.371, Horn R_L = 0.056, s = 0.035, $\hat{g}_1$ = -0.22, $\hat{g}_2$ = 1.99,

Závěr: 0.334 < μ < 0.408, **C3.42D** Symetrické rozdělení, prokázáno Gaussovo rozdělení,
počet odlehlých bodů: 0, $\bar{x}$ = 0.362, $\tilde{x}_{0.5}$ = 0.362, $\bar{x}_R$ = 0.361, Box-Cox transf. není nutná, Horn P_L = 0.361, Horn R_L = 0.022, s = 0.013, $\hat{g}_1$ = 0.10, $\hat{g}_2$ = 1.93, Závěr: 0.346 < μ < 0.376.

C3.43 Homogenita, průměry jsou shodné: **C3.43A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 14.14, $\tilde{x}_{0.5}$ = 15.60, $\bar{x}_R$ = 14.27, Box-Cox transf. není nutná, Horn P_L = 14.20, Horn R_L = 7.80, s = 5.05, $\hat{g}_1$ = -0.16, $\hat{g}_2$ = 1.64, Závěr: 7.06 < μ < 21.34,

C3.43B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 14.14,
$\tilde{x}_{0.5}$ = 15.61, $\bar{x}_R$ = 14.29, Box-Cox transf. není nutná, Horn P_L = 14.26, Horn R_L = 7.70, s = 5.02,
$\hat{g}_1$ = -0.18, $\hat{g}_2$ = 1.66, Závěr: 7.21 < μ < 21.31.

C3.44 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: , $\bar{x}$ = 27.1, $\tilde{x}_{0.5}$ = 27.0,
$\bar{x}_R$ = 27.1, Box-Cox transf. není nutná, Horn P_L = 27.0, Horn R_L = 2.0, s = 1.2,
$\hat{g}_1$ = -0.20, $\hat{g}_2$ = 2.24, Závěr: 25.7 < μ < 28.3.

C3.45 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 94.82,
$\tilde{x}_{0.5}$ = 94.80, $\bar{x}_R$ = 94.74, Box-Cox transf. není nutná, Horn P_L = 94.65,
Horn R_L = 1.10, s = 0.69, $\hat{g}_1$ = 0.37, $\hat{g}_2$ = 2.23, Box-Cox, Závěr: 94.14 < μ < 95.16.

C3.46 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 94.36,
$\tilde{x}_{0.5}$ = 94.35, $\bar{x}_R$ = 94.28, Box-Cox transf. není nutná, Horn P_L = 64.10,
Horn R_L = 1.80, s = 0.99, $\hat{g}_1$ = 0.20, $\hat{g}_2$ = 2.01, Závěr: 92.90 < μ < 95.30.

C3.47 Homogenita, průměry jsou shodné: **C3.47A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 99.165, $\tilde{x}_{0.5}$ = 99.045, $\bar{x}_R$ = 99.069, Box-Cox transf. je nutná, P_L = 99.070, Horn R_L = 0.140, s = 0.330, $\hat{g}_1$ = 1.25, $\hat{g}_2$ = 2.97, Závěr: 98.977 < μ < 99.164,

C3.47B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 99.158,
$\tilde{x}_{0.5}$ = 99.145, $\bar{x}_R$ = 99.167, Box-Cox transf. není nutná, Horn P_L = 99.200, Horn R_L = 0.480, s = 0.345,
$\hat{g}_1$ = -0.18, $\hat{g}_2$ = 2.60, Závěr: 98.88 < μ < 99.52.

C3.48 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 246.2,
$\tilde{x}_{0.5}$ = 236.5, $\bar{x}_R$ = 235.6, Box-Cox transf. je nutná, P_L = 234.0, Horn R_L = 14.0,
s = 37.8, $\hat{g}_1$ = 1.54, $\hat{g}_2$ = 3.80, Závěr: 219.5 < μ < 248.5.

C3.49AB Mladina: Homogenita, průměry jsou shodné: **C3.49A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 11.22, $\tilde{x}_{0.5}$ = 11.78, $\bar{x}_R$ = 11.47, Box-Cox transf. je nutná, s = 1.12, $\hat{g}_1$ = -1.11, $\hat{g}_2$ = 3.16, Závěr: 11.21 < μ < 11.75, **C3.49B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0 , $\bar{x}_R$ = 11.45, Box-Cox transf. je nutná, Závěr:
11.17 < μ < 11.74,

C3.49CD Alkohol: Homogenita, průměry jsou rozdílné: **C3.49C** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 3.915, $\tilde{x}_{0.5}$ = 3.915, $\bar{x}_R$ = 3.917, Box-Cox transf. není nutná, s = 0.020, $\hat{g}_1$ = -0.64, $\hat{g}_2$ = 4.25, Box-Cox, Závěr: 3.908 < μ < 3.922, **C3.49D** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 3.848, $\tilde{x}_{0.5}$ = 3.845, $\bar{x}_R$ = 3.843, Box-Cox transf. je nutná, s = 0.025, $\hat{g}_1$ = 1.56, $\hat{g}_2$ = 5.06, Box-Cox, Závěr: 3.827 < μ < 3.853.

C3.50 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 142.4, $\tilde{x}_{0.5}$ = 145.6, $\bar{x}_R$ = 143.9, Box-Cox transf. není nutná, Horn P_L = 139.0, Horn R_L = 24.8, s = 11.9, $\hat{g}_1$ = -0.60, $\hat{g}_2$ = 2.21, Závěr: 125.0 < μ < 153.0.

C3.51 Homogenita, průměry jsou shodné: **C3.51A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 145.7, $\tilde{x}_{0.5}$ = 127.8, $\bar{x}_R$ = 124.4, Box-Cox transf. je nutná, P_L = 130.2, Horn R_L = 63.7, s = 84.3, $\hat{g}_1$ = 1.90, $\hat{g}_2$ = 5.84, Závěr: 87.6 < μ < 172.7,

C3.51B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 143.1, $\tilde{x}_{0.5}$ = 126.1, $\bar{x}_R$ = 120.9, Box-Cox transf. je nutná, P_L = 130.3, Horn R_L = 58.7, s = 78.8, $\hat{g}_1$ = 1.91, $\hat{g}_2$ = 5.83, Závěr: 91.2 < μ < 169.3.

C3.52 Homogenita, průměry jsou rozdílné: **C3.52A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 9.990, $\tilde{x}_{0.5}$ = 10.000, $\bar{x}_R$ = 9.995, Box-Cox transf. není nutná, Horn P_L = 10.000, Horn R_L = 0.200, s = 0.137, $\hat{g}_1$ = -0.09, $\hat{g}_2$ = 1.79, Závěr: 9.866 < μ < 10.134, **C3.52B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 9.600, $\tilde{x}_{0.5}$ = 9.550, $\bar{x}_R$ = 9.549, Box-Cox transf. není nutná, Horn P_L = 9.550, Horn R_L = 0.300, s = 0.309, $\hat{g}_1$ = 1.00, $\hat{g}_2$ = 3.75, Závěr: 9.350 < μ < 9.750.

C3.53 Heterogenita, průměry jsou shodné: **C3.53A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.737, $\tilde{x}_{0.5}$ = 0.737, $\bar{x}_R$ = 0.737, Box-Cox transf. není nutná, Horn P_L = 0.738, Horn R_L = 0.005, s = 0.003, $\hat{g}_1$ = -0.33, $\hat{g}_2$ = 3.09, Box-Cox, Závěr: 0.736 < μ < 0.740, **C3.53B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 0.735, $\tilde{x}_{0.5}$ = 0.736, $\bar{x}_R$ = 0.735, Box-Cox transf. není nutná, Horn P_L = 0.736, Horn R_L = 0.009, s = 0.005, $\hat{g}_1$ = -0.26, $\hat{g}_2$ = 2.04, Box-Cox, Závěr: 0.732 < μ < 0.739.

C3.54 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 5.40, $\tilde{x}_{0.5}$ = 5.41, $\bar{x}_R$ = 5.44, Box-Cox transf. není nutná, Horn P_L = 5.48, Horn R_L = 0.38, s = 0.32, $\hat{g}_1$ = -0.49, $\hat{g}_2$ = 1.97, Závěr: 4.68 < μ < 6.28.

C3.55 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 30.0, $\tilde{x}_{0.5}$ = 30.0, $\bar{x}_R$ = 30.2, Box-Cox transf. není nutná, Horn P_L = 30.0, Horn R_L = 0.7, s = 0.5, $\hat{g}_1$ = -0.08, $\hat{g}_2$ = 3.50, Závěr: 29.7 < μ < 30.4.

C3.56 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 3.609, $\tilde{x}_{0.5}$ = 3.610, $\bar{x}_R$ = 3.607, Box-Cox transf. není nutná, Horn P_L = 3.610, Horn R_L = 0.080, s = 0.044, $\hat{g}_1$ = -0.05, $\hat{g}_2$ = 1.85, Závěr: 3.568 < μ < 3.652.

3.6.3 Analýza environmentálních, potravinářských a zemědělských dat

E3.01 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 7.83, $\tilde{x}_{0.5}$ = 6.70, $\bar{x}_R$ = 6.86, Box-Cox transf. je nutná, P_L = 6.70, Horn R_L = 0.60, s = 3.24, $\hat{g}_1$ = 2.32, $\hat{g}_2$ = 6.94, Závěr: 6.30 < μ < 7.10 (Horn).

E3.02 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\tilde{x}_{0.5}$ = 0.0169, $\bar{x}_R$ = 0.0191, Box-Cox transf. je nutná, P_L = 0.0195, Horn R_L = 0.0096, s = 0.0517, $\hat{g}_1$ = 2.25, $\hat{g}_2$ = 6.09, Závěr: 0.0141 < μ < 0.0249 (Horn).

E3.03 Asymetrické rozdělení, $\bar{x}_R$ = 0.20, $\tilde{x}_{0.5}$ = 0.20, Horn P_L = 0.195, Horn R_L = 0.03, s = 0.03, $\hat{g}_1$ = 1.25,$\hat{g}_2$ = 4.21, Závěr: 0.175 < μ < 0.215 (Horn).

E3.04 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 25.16, $\tilde{x}_{0.5}$ = 25.10, $\bar{x}_R$ = 25.16, Box-Cox transf. není nutná, P_L = 25.15, Horn R_L = 0.90, s = 0.69, $\hat{g}_1$ = -0.05, $\hat{g}_2$ = 2.27, Závěr: 24.68 < μ < 25.62 (Horn).

E3.05 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 2.47, $\tilde{x}_{0.5}$ = 2.47, $\bar{x}_R$ = 2.47, Box-Cox transf. není nutná, P_L = 2.47, Horn R_L = 0.02, s = 0.01, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 1.75, Závěr: 2.46 < μ < 2.48 (Horn).

E3.06 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 14.8, $\tilde{x}_{0.5}$ = 16.0, $\bar{x}_R$ = 15.5, Box-Cox transf. není nutná, P_L = 15.0, Horn R_L = 6.0, s = 3.7, $\hat{g}_1$ = -0.66, $\hat{g}_2$ = 1.94, Závěr: 8.8 < μ < 21.2 (Horn).

E3.07 Gaussovo rozdělení, $\bar{x}$ = 95.35, $\bar{x}_R$ = 92.83, $\tilde{x}_{0.5}$ = 93.40, s = 11.54, $\hat{g}_1$ = 0.42, $\hat{g}_2$ = 3.26, Závěr: 91.39 < μ < 99.31, intervalový odhad obsahuje požadovanou hodnotu 94.8 mg. l^{-1}.

E3.08 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 4.55, $\tilde{x}_{0.5}$ = 4.83, $\bar{x}_R$ = 4.71, Box-Cox transf. je nutná, P_L = 4.46, Horn R_L = 1.13, s = 0.67, $\hat{g}_1$ = -1.03, $\hat{g}_2$ = 2.44, Závěr: 3.64 < μ < 5.27 (Horn).

E3.09 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 13.54, $\tilde{x}_{0.5}$ = 12.95, $\bar{x}_R$ = 13.18, Box-Cox transf. není nutná, P_L = 13.65, Horn R_L = 5.10, s = 2.51, $\hat{g}_1$ = 0.17, $\hat{g}_2$ = 1.24, Závěr: 10.24 < μ < 17.06 (Horn).

E3.10 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 54.06, $\tilde{x}_{0.5}$ = 55.00, $\bar{x}_R$ = 54.71, Box-Cox transf. není nutná, P_L = 53.20, Horn R_L = 5.60, s = 2.69, $\hat{g}_1$ = -0.82, $\hat{g}_2$ = 1.88, Závěr: 49.17 < μ < 57.23 (Horn).

E3.11 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 50.3, $\tilde{x}_{0.5}$ = 50.2, $\bar{x}_R$ = 50.2, Box-Cox transf. není nutná, P_L = 50.2, Horn R_L = 3.8, s = 1.7, $\hat{g}_1$ = 0.04, $\hat{g}_2$ = 1.54, Závěr: 48.1 < μ < 52.3 (Horn).

E3.12 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 2.47, $\tilde{x}_{0.5}$ = 2.47, $\bar{x}_R$ = 2.47, Box-Cox transf. není nutná, P_L = 2.47, Horn R_L = 0.02, s = 0.01, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 1.75, Závěr: 2.46 < μ < 2.48 (Horn).

E3.13 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 4.57, $\tilde{x}_{0.5}$ = 4.56, $\bar{x}_R$ = 4.60, Box-Cox transf. není nutná, P_L = 4.64, Horn R_L = 0.57, s = 0.53, $\hat{g}_1$ = -0.24, $\hat{g}_2$ = 1.92, Závěr: 3.44 < μ < 5.82 (Horn).

E3.14 Homogenita, průměry jsou stejné: **E3.14A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 6.85, $\tilde{x}_{0.5}$ = 6.90, $\overline{x}_R$ = 6.86, Box-Cox transf. není nutná, s = 0.19, $\hat{g}_1$ = -0.15, $\hat{g}_2$ = 2.94, Závěr: 6.79 < μ < 6.93 (Box), **E3.14B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\overline{x}$ = 6.86, $\tilde{x}_{0.5}$ = 6.90, $\overline{x}_R$ = 6.83, Box-Cox transf. není nutná, s = 0.24, $\hat{g}_1$ = 0.37, $\hat{g}_2$ = 2.10, Závěr: 6.74 < μ < 6.92 (Box).

E3.15 Homogenita, průměry jsou rozdílné: **E3.15A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\overline{x}$ = 272.7, $\tilde{x}_{0.5}$ = 273.0, $\overline{x}_R$ = 272.9, Box-Cox transf. je nutná, P_L = 273.0, Horn R_L = 3.0, s = 1.4, $\hat{g}_1$ = -0.84, $\hat{g}_2$ = 2.51, Závěr: 272.1 < μ < 273.9 (Horn), **E3.15B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 269.1, $\tilde{x}_{0.5}$ = 269.0, $\overline{x}_R$ = 269.0, Box-Cox transf. není nutná, P_L = 269.0, Horn R_L = 3.0, s = 1.4, $\hat{g}_1$ = 0.03, $\hat{g}_2$ = 1.58, Závěr: 268.1 < μ < 269.9 (Horn).

E3.16 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 0.363, $\tilde{x}_{0.5}$ = 0.300, $\overline{x}_R$ = 0.318, Box-Cox transf. je nutná, s = 0.200, $\hat{g}_1$ = 0.78, $\hat{g}_2$ = 2.39, Závěr: 0.248 < μ < 0.393 (Box).

E3.17 Homogenita, průměry jsou shodné: **E3.17A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 22.76, $\tilde{x}_{0.5}$ = 22.11, $\overline{x}_R$ = 21.92, Box-Cox transf. není nutná, P_L = 22.63, Horn R_L = 5.35, s = 3.33, $\hat{g}_1$ = 0.84, $\hat{g}_2$ = 2.46, Závěr: 18.77 < μ < 26.48 (Horn), **E3.17B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\tilde{x}_{0.5}$ = 22.19, $\overline{x}_R$ = 22.31, Box-Cox transf. je nutná, s = 1.1300, $\hat{g}_1$ = -1.17, $\hat{g}_2$ = 3.73, Závěr: 21.83 < μ < 22.95.

E3.18 Heteroskedasticita, průměry jsou shodné: **E3.18A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\overline{x}$ = 49.47, $\tilde{x}_{0.5}$ = 48.05, $\overline{x}_R$ = 48.57, Box-Cox transf. není nutná, P_L = 22.39, Horn R_L = 0.78, s = 2.03, $\hat{g}_1$ = 0.57, $\hat{g}_2$ = 1.56, Závěr: 46.77 < μ < 52.39 (Horn), **E3.18B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 49.44, $\tilde{x}_{0.5}$ = 49.04, $\overline{x}_R$ = Box-Cox transf. není nutná, P_L = 49.33, Horn R_L = 0.81, s = 0.85, $\hat{g}_1$ = 1.53, $\hat{g}_2$ = 4.53, Závěr: 48.78 < μ < 49.87 (Horn).

E3.19 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 1.505, $\tilde{x}_{0.5}$ = 1.490, $\overline{x}_R$ = 1.488, Box-Cox transf. je nutná, s = 0.080, $\hat{g}_1$ = 3.53, $\hat{g}_2$ = 18.02, Závěr: 1.471 < μ < 1.505 (Box).

E3.20 Nehomogenita, průměry jsou rozdílné: **E3.20A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 1.066, $\tilde{x}_{0.5}$ = 1.058, $\overline{x}_R$ = 1.060, Box-Cox transf. není nutná, P_L = 1.068, Horn R_L = 0.057, s = 0.036, $\hat{g}_1$ = 0.53, $\hat{g}_2$ = 1.96, Závěr: 1.029 < μ < 1.106 (Horn), **E3.20B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 1.141, $\tilde{x}_{0.5}$ = 1.181, $\overline{x}_R$ = 1.160, Box-Cox transf. není nutná, P_L = 1.135, Horn R_L = 0.183, s = 0.097, $\hat{g}_1$ = -0.71, $\hat{g}_2$ = 1.94, Závěr: 1.012 < μ < 1.257 (Horn).

E3.21 Homogenita, průměry jsou rozdílné: **E3.21A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 47.05, $\tilde{x}_{0.5}$ = 47.00, $\overline{x}_R$ = 47.07, Box-Cox transf. není nutná, P_L = 47.00, Horn R_L = 1.00, s = 0.72, $\hat{g}_1$ = -0.18, $\hat{g}_2$ = 1.90, Závěr: 46.33 < μ < 47.67 (Horn), **E3.21B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 53.10, $\tilde{x}_{0.5}$ = 53.00, $\overline{x}_R$ = 53.11, Box-Cox transf. není nutná, P_L = 53.50, Horn R_L = 1.00, s = 0.74, $\hat{g}_1$ = -0.14, $\hat{g}_2$ = 2.04, Závěr: 52.83 < μ < 54.17 (Horn).

E3.22 Homogenita, průměry jsou rozdílné: **E3.22A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 0.4995, $\tilde{x}_{0.5}$ = 0.4995, $\overline{x}_R$ = 0.4995, Box-Cox transf. není nutná, P_L = 0.4995, Horn R_L = 0.0300, s = 0.0016, $\hat{g}_1$ = 0.00, $\hat{g}_2$ = 1.95, Závěr: 0.4975 < μ < 0.5015 (Horn), **E3.22B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 0.5018, $\tilde{x}_{0.5}$ = 0.5020, $\overline{x}_R$ = 0.5017, Box-Cox transf. není nutná, P_L = 0.5015, Horn R_L = 0.0300, s = 0.0019, $\hat{g}_1$ = 0.20, $\hat{g}_2$ = 2.14, Závěr: 0.4995 < μ < 0.5035 (Horn).

E3.23 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 0.0169, $\tilde{x}_{0.5}$ = 0.0166, $\overline{x}_R$ = 0.0163, Box-Cox transf. je nutná, P_L = 0.0161, Horn R_L = 0.0290, s = 0.0034, $\hat{g}_1$ = 1.27, $\hat{g}_2$ = 4.24, Závěr: 0.0145 < μ < 0.0178 (Horn).

E3.24A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 0.461, $\tilde{x}_{0.5}$ = 0.444, $\overline{x}_R$ = 0.422, Box-Cox transf. je nutná, P_L = 0.442, Horn R_L = 0.227, s = 0.178, $\hat{g}_1$ = 1.29, $\hat{g}_2$ = 4.50, Závěr: 0.304 < μ < 0.580 (Horn), **E3.24B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\overline{x}$ = 0.146, $\tilde{x}_{0.5}$ = 0.101, $\overline{x}_R$ = 0.124, Box-Cox transf. je nutná, P_L = 0.156, Horn R_L = 0.143, s = 0.084, $\hat{g}_1$ = 0.81, $\hat{g}_2$ = 2.38, Závěr: 0.070 < μ < 0.243 (Horn).

E3.25 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 999.92, $\tilde{x}_{0.5}$ = 999.50, $\overline{x}_R$ = 1000.40, Box-Cox transf. není nutná, P_L = 1001.50, Horn R_L = 13.00, s = 8.00, $\hat{g}_1$ = -0.42, $\hat{g}_2$ = 2.86, Závěr: 995.2 < μ < 1007.8 (Horn).

E3.26A Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\overline{x}$ = 49.17, $\tilde{x}_{0.5}$ = 48.05, $\overline{x}_R$ = 48.57, Box-Cox transf. není nutná, P_L = 49.58, Horn R_L = 4.20, s = 2.00, $\hat{g}_1$ = 0.57, $\hat{g}_2$ = 1.56, Závěr: 46.77 < μ < 52.39 (Horn), **E3.26B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 49.44, $\tilde{x}_{0.5}$ = 49.05, $\overline{x}_R$ = 49.19, Box-Cox transf. je nutná, P_L = 49.33, Horn R_L = 0.81, s = 0.85, $\hat{g}_1$ = 1.53, $\hat{g}_2$ = 4.53, Závěr: 48.78 < μ < 49.87 (Horn).

E3.27 Homogenita, průměry jsou shodné: **E3.27A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 10, $\overline{x}$ = 97.68, $\tilde{x}_{0.5}$ = 79.00, $\overline{x}_R$ = 76.24, Box-Cox transf. je nutná, s = 84.00, $\hat{g}_1$ = 2.12, $\hat{g}_2$ = 8.01, Závěr: 63.50 < μ < 87.05 (Box), **E3.27B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 8, $\overline{x}$ = 80.48, $\tilde{x}_{0.5}$ = 68.00, $\overline{x}_R$ = 66.18, Box-Cox transf. je nutná, s = 60.74, $\hat{g}_1$ = 1.87, $\hat{g}_2$ = 7.41, Závěr: 55.76 < μ < 75.24 (Box).

E3.28 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 0.19, $\tilde{x}_{0.5}$ = 0.18, $\overline{x}_R$ = 0.18, Box-Cox transf. není nutná, P_L = 0.19, Horn R_L = 0.04, s = 0.03, $\hat{g}_1$ = 0.60, $\hat{g}_2$ = 1.79, Závěr: 0.15 < μ < 0.22 (Horn).

E3.29 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 23.60, $\tilde{x}_{0.5}$ = 23.60, $\overline{x}_R$ = 23.65, Box-Cox transf. není nutná, s = 1.08, $\hat{g}_1$ = -0.31, $\hat{g}_2$ = 2.67, Závěr: 23.32 < μ < 24.01 (Box).

E3.30 Párový test: stanovení jsou shodná: **E3.30A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 0.27, $\tilde{x}_{0.5}$ = 0.13, $\overline{x}_R$ = 0.15, Box-Cox transf. je nutná, P_L = 0.15, Horn R_L = 0.18, s = 0.37, $\hat{g}_1$ = 1.80, $\hat{g}_2$ = 4.69, Závěr: 0.04 < μ < 0.26 (Horn),

E3.30B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.27, $\tilde{x}_{0.5}$ = 0.15, $\bar{x}_R$ = 0.16, Box-Cox transf. je nutná, P_L = 0.17, Horn R_L = 0.19, s = 0.33, $\hat{g}_1$ = 1.86, $\hat{g}_2$ = 5.17, Závěr: 0.05 < μ < 0.28 (Horn).

E3.31 Homogenita, průměry jsou rozdílné: **E3.31A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 47.2, $\tilde{x}_{0.5}$ = 46.0, $\bar{x}_R$ = 46.5, Box-Cox transf. není nutná, P_L = 48.0, Horn R_L = 8.0, s = 5.2, $\hat{g}_1$ = 0.47, $\hat{g}_2$ = 2.09, Závěr: 42.7 < μ < 53.3 (Horn),

E3.31B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 54.3, $\tilde{x}_{0.5}$ = 54.8, $\bar{x}_R$ = 54.7, Box-Cox transf. není nutná, P_L = 54.0, Horn R_L = 5.7, = 3.8, $\hat{g}_1$ = -0.47, $\hat{g}_2$ = 2.03, Závěr: 50.1 < μ < 57.8 (Horn).

3.6.4 Analýza hutnických a mineralogických dat

H3.01 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 0.181, $\tilde{x}_{0.5}$ = 0.190, $\bar{x}_R$ = 0.186, Horn P_L = 0.185, Horn R_L = 0.010, s = 0.020, $\hat{g}_1$ = -1.52, $\hat{g}_2$ = 4.13, Závěr: 0.178 < μ < 0.192 (Horn).

H3.02 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1). $\bar{x}$ = 6.951, $\tilde{x}_{0.5}$ = 7.172, $\bar{x}_R$ = 6.873, Horn P_L = 6.570, Horn R_L = 1.630, s = 1.090, $\hat{g}_1$ = 0.28, $\hat{g}_2$ = 2.43, Závěr: 5.397 < μ < 7.742 (Horn).

H3.03 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 75.7, $\tilde{x}_{0.5}$ = 79.0, $\bar{x}_R$ = 77.2, Horn P_L = 75.0, Horn R_L = 10.5, s = 6.7, $\hat{g}_1$ = -0.76, $\hat{g}_2$ = 1.88, Závěr: 64.1 < μ < 85.8 (Horn).

H3.04 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 44.74, $\tilde{x}_{0.5}$ = 44.60, $\bar{x}_R$ = 44.64, Horn P_L = 44.85, Horn R_L = 1.50, s = 0.72, $\hat{g}_1$ = 0.36, $\hat{g}_2$ = 1.72, Závěr: 43.86 < μ < 45.84 (Horn).

H3.05 Homogenita, průměry jsou shodné: **H3.05A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 22.29, $\tilde{x}_{0.5}$ = 22.20, $\bar{x}_R$ = 22.33, Horn P_L = 22.50, Horn R_L = 1.00, s = 0.72, $\hat{g}_1$ = -0.32, $\hat{g}_2$ = 2.35, Závěr: 22.02 < μ < 22.98 (Horn), **H3.05B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 22.19, $\tilde{x}_{0.5}$ = 22.05, $\bar{x}_R$ = 22.08, Horn P_L = 22.15, Horn R_L = 0.50, s = 0.41, $\hat{g}_1$ = 0.92, $\hat{g}_2$ = 2.60, Závěr: 21.91 < μ < 22.39 (Horn).

H3.06 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 46.9, $\tilde{x}_{0.5}$ = 47.6, $\bar{x}_R$ = 47.0, Horn P_L = 46.8, Horn R_L = 1.7, s = 1.5, $\hat{g}_1$ = -0.38, $\hat{g}_2$ = 1.65, Závěr: 43.2 < μ < 50.3 (Horn).

H3.07 Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 8.51, $\tilde{x}_{0.5}$ = 8.47, $\bar{x}_R$ = 8.52, Horn P_L = 8.59, Horn R_L = 0.58, s = 0.42, $\hat{g}_1$ = -0.22, $\hat{g}_2$ = 1.68, Závěr: 7.38 < μ < 9.80 (Horn).

H3.08 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 9.95, $\tilde{x}_{0.5}$ = 9.97, $\bar{x}_R$ = 9.95, Horn P_L = 9.93, Horn R_L = 0.88, s = 0.63, $\hat{g}_1$ = 0.03, $\hat{g}_2$ = 3.36, Závěr: 9.51 < μ < 10.36 (Horn).

H3.09 Homogenita, průměry jsou shodné: **H3.09A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 22.58, $\tilde{x}_{0.5}$ = 22.36, $\bar{x}_R$ = 22.38, s = 1.0900, $\hat{g}_1$ = 0.73, $\hat{g}_2$ = 4.06, Závěr: 21.99 < μ < 22.79 (Box-Cox), **H3.09B** Symetrické rozdělení,

prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 22.48, $\tilde{x}_{0.5}$ = 22.36, $\bar{x}_R$ = 22.38, s = 1.0900, $\hat{g}_1$ = 0.74, $\hat{g}_2$ = 4.06, Závěr: 22.06 < μ < 22.90 (Průměr).

H3.10 Homogenita, průměry jsou rozdílné: **H3.10A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 59.36, $\tilde{x}_{0.5}$ = 59.55, $\bar{x}_R$ = 59.36, s = 1.19, $\hat{g}_1$ = -0.04, $\hat{g}_2$ = 2.75, Závěr: 58.91 < μ < 59.83 (Box-Cox), **H3.10B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 60.50, $\tilde{x}_{0.5}$ = 60.67, $\bar{x}_R$ = 60.56, s = 1.2300, $\hat{g}_1$ = -0.29, $\hat{g}_2$ = 2.45, Závěr: 60.10 < μ < 61.04 (Box-Cox).

H3.11 Nehomogenita, párovým testem jsou průměry rozdílné: **H3.11A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 0.2191, $\tilde{x}_{0.5}$ = 0.2260, $\bar{x}_R$ = 0.2207, Horn P_L = 0.2166, Horn R_L = 0.0444, s = 0.0217, $\hat{g}_1$ = -0.29, $\hat{g}_2$ = 1.42, Závěr: 0.1990 < μ < 0.2342 (Horn), **H3.11B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 0.2371, $\tilde{x}_{0.5}$ = 0.2401, $\bar{x}_R$ = 0.2393, Horn P_L = 0.2380, Horn R_L = 0.0125, s = 0.0134, $\hat{g}_1$ = -1.96, $\hat{g}_2$ = 7.78, Závěr: 0.233 < μ < 0.243 (Horn).

H3.12 Nehomogenita, průměry jsou shodné: **H3.12A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 49.13, $\tilde{x}_{0.5}$ = 48.80, $\bar{x}_R$ = 48.80, Horn P_L = 48.70, Horn R_L = 2.80, s = 2.11, $\hat{g}_1$ = 0.52, $\hat{g}_2$ = 2.16, Závěr: 47.00 < μ < 50.40 (Horn), **H3.12B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 50.03, $\tilde{x}_{0.5}$ = 50.10, $\bar{x}_R$ = 50.06, Horn P_L = 50.05, Horn R_L = 0.50, s = 0.69, $\hat{g}_1$ = -0.20, $\hat{g}_2$ = 2.29, Závěr: 49.59 < μ < 50.51 (Horn).

H3.13 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 288.3, $\tilde{x}_{0.5}$ = 311.0, $\bar{x}_R$ = 307.7, Horn P_L = 300.0, Horn R_L = 200.0, s = 115.5, $\hat{g}_1$ = -0.97, $\hat{g}_2$ = 3.22, Závěr: 191.0 < μ < 409.0 (Horn).

H3.14 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 0.314, $\tilde{x}_{0.5}$ = 0.295, $\bar{x}_R$ = 0.282, Horn P_L = 0.300, Horn R_L = 0.200, s = 0.132, $\hat{g}_1$ = 1.00, $\hat{g}_2$ = 3.05, Závěr: 0.187 < μ < 0.413 (Horn).

H3.15 Homogenita, průměry jsou shodné: **H3.15A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 6.1, $\tilde{x}_{0.5}$ = 5.5, $\bar{x}_R$ = 5.5, s = 3.2, $\hat{g}_1$ = 4.11, $\hat{g}_2$ = 18.31, **H3.15B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 3.039, $\tilde{x}_{0.5}$ = 4.050, $\bar{x}_R$ = 2.502, Horn P_L = 5.415, Horn R_L = 1.170, s = 3.170, $\hat{g}_1$ = 0.36, $\hat{g}_2$ = 1.91, Závěr: 4.886 < μ < 5.944.

H3.16 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 55.84, $\tilde{x}_{0.5}$ = 55.85, $\bar{x}_R$ = 55.82, Horn P_L = 55.81, Horn R_L = 0.18, s = 0.21, $\hat{g}_1$ = 0.31, $\hat{g}_2$ = 2.12, Závěr: 55.43 < μ < 56.19 (Horn).

H3.17 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 2.96, $\tilde{x}_{0.5}$ = 3.00, $\bar{x}_R$ = 2.93, Horn P_L = 2.90, Horn R_L = 0.20, s = 0.23, $\hat{g}_1$ = 0.41, $\hat{g}_2$ = 2.07, Závěr: 2.48 < μ < 3.32 (Horn).

H3.18 Homogenita, průměry jsou rozdílné: **H3.18A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 35.16, $\tilde{x}_{0.5}$ = 35.22, $\bar{x}_R$ = 35.19, Horn P_L = 35.15, Horn R_L = 0.25, s = 0.16, $\hat{g}_1$ = -0.67, $\hat{g}_2$ = 2.17, Závěr: 34.99 < μ < 35.30 (Horn),

H3.18B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 34.87, $\tilde{x}_{0.5}$ = 34.87, $\bar{x}_R$ = 34.89, Horn P_L = 34.92, Horn R_L = 0.37, s = 0.25, $\hat{g}_1$ = -0.42, $\hat{g}_2$ = 2.60, Závěr: 34.69 < μ < 35.14 (Horn).

H3.19 Párový test: hodnoty jsou rozdílné: **H3.19A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 35.164, $\tilde{x}_{0.5}$ = 35.220, $\bar{x}_R$ = 35.190, Horn P_L = 35.145, Horn R_L = 0.250, s = 0.164, $\hat{g}_1$ = -0.67, $\hat{g}_2$ = 2.17, Závěr: 34.993 < μ < 35.297 (Horn),

H3.19B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 35.240, $\tilde{x}_{0.5}$ = 35.240, $\bar{x}_R$ = 35.254, Horn P_L = 35.265, Horn R_L = 0.250, s = 0.202, $\hat{g}_1$ = -0.26, $\hat{g}_2$ = 2.19, Závěr: 35.113 < μ < 35.417 (Horn).

H3.20 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 50.83, $\tilde{x}_{0.5}$ = 50.79, $\bar{x}_R$ = 50.83, Horn P_L = 50.86, Horn R_L = 0.38, s = 0.25, $\hat{g}_1$ = -0.14, $\hat{g}_2$ = 2.03, Závěr: 50.63 < μ < 51.09 (Horn).

H3.21 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 51.16, $\tilde{x}_{0.5}$ = 50.41, $\bar{x}_R$ = 50.53, Horn P_L = 51.91, Horn R_L = 5.26, s = 2.32, $\hat{g}_1$ = 0.95, $\hat{g}_2$ = 2.20, Závěr: 48.03 < μ < 55.79 (Horn).

H3.22 Párový test: hodnoty jsou rozdílné: **H3.22A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 7, $\bar{x}$ = 13.95, $\tilde{x}_{0.5}$ = 12.83, $\bar{x}_R$ = 12.30, s = 5.21, $\hat{g}_1$ = 1.63, $\hat{g}_2$ = 4.12, Závěr: 11.45 < μ < 12.85 (Box-Cox), **H3.22B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 7, $\bar{x}$ = 13.91, $\tilde{x}_{0.5}$ = 13.16, $\bar{x}_R$ = 12.28, s = 5.11, $\hat{g}_1$ = 1.60, $\hat{g}_2$ = 4.05, Závěr: 11.37 < μ < 12.71 (Box-Cox), **H3.22C** Diference paru: Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 0.0436, $\tilde{x}_{0.5}$ = -0.0500, $\bar{x}_R$ = -0.0282, Horn P_L = Horn R_L = s = 0.3690, $\hat{g}_1$ = 1.12, $\hat{g}_2$ = 3.98, Závěr: -0.1190 < μ < 0.0670 (Box-Cox).

3.6.5 Analýza ekonomických a sociologických dat

S3.01 Homogenita, střední hodnoty jsou rozdílné: **S3.01A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 4.3, $\tilde{x}_{0.5}$ = 4.0, $\bar{x}_R$ = 3.9, Horn P_L = 4.5, Horn R_L = 3.0, s = 2.1, $\hat{g}_1$ = 0.70, $\hat{g}_2$ = 2.55, Závěr: 3.1 < μ < 5.9 (Horn), **S3.01B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 6.1, $\tilde{x}_{0.5}$ = 6.0, $\bar{x}_R$ = 5.9, Horn P_L = 6.5, Horn R_L = 3.0, s = 2.5, $\hat{g}_1$ = 0.25, $\hat{g}_2$ = 3.17, Závěr: 5.3 < μ < 7.7 (Horn).

S3.02 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 100.01, $\tilde{x}_{0.5}$ = 100.01, $\bar{x}_R$ = 100.01, s = 0.03, $\hat{g}_1$ = 0.60, $\hat{g}_2$ = 3.84, Závěr: 99.99 < μ < 100.02 (Box-Cox).

S3.03 Heterogenita, průměry jsou rozdílné: **S3.03A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 123.90, $\tilde{x}_{0.5}$ = 124.50, $\bar{x}_R$ = 123.70, s = 33.40, $\hat{g}_1$ = 0.02, $\hat{g}_2$ = 2.82, Závěr: 115.10 < μ < 132.40 (Box-Cox), **S3.03B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 135.2, $\tilde{x}_{0.5}$ = 135.0, $\bar{x}_R$ = 135.6, s = 22.6, $\hat{g}_1$ = -0.12, $\hat{g}_2$ = 3.41, Závěr: 129.9 < μ < 141.2 (Box-Cox).

S3.04 Homogenita, párovým testem průměry jsou rozdílné: **S3.04A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 26.80, $\tilde{x}_{0.5}$ = 14.75, $\bar{x}_R$ = 20.29, s = 23.23, $\hat{g}_1$ = 0.84, $\hat{g}_2$ = 2.37, Závěr: 13.18 < μ < 24.80 (Box-Cox),

S3.04B Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 26.11, $\tilde{x}_{0.5}$ = 14.90, $\bar{x}_R$ = 19.88, s = 22.27, $\hat{g}_1$ = 0.92, $\hat{g}_2$ = 2.64,
Závěr: 13.18 < μ < 24.24 (Box-Cox), **S3.04C** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 7, $\bar{x}$ = 23.96, $\tilde{x}_{0.5}$ = 12.50, $\bar{x}_R$ = 17.38, Horn P_L = 21.35, Horn R_L = 27.30,
s = 21.34, $\hat{g}_1$ = 1.00, $\hat{g}_2$ = 2.47, Závěr: 8.62 < μ < 34.07 (Horn), **S3.04D** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 5, $\bar{x}$ = 26.09, $\tilde{x}_{0.5}$ = 16.00, $\bar{x}_R$ = 20.41,
Horn P_L = 26.90, Horn R_L = 32.00, s = 20.19, $\hat{g}_1$ = 0.75, $\hat{g}_2$ = 2.06, Závěr: 11.99 < μ < 41.81 (Horn).

S3.05 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\bar{x}$ = 301.0, $\tilde{x}_{0.5}$ = 300.0, $\bar{x}_R$ = 300.2, Horn P_L = 300.0, Horn R_L = 1.0, s = 5.9, $\hat{g}_1$ = 0.75, $\hat{g}_2$ = 3.23,
Závěr: 296.82 < μ < 303.18 (Horn).

S3.06 Homogenita, průměry jsou shodné: **S3.06A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 67.0, $\tilde{x}_{0.5}$ = 65.0, $\bar{x}_R$ = 65.0, s = 17.2, $\hat{g}_1$ = 0.80, $\hat{g}_2$ = 4.63,
Závěr: 61.2 < μ < 69.0 (Box-Cox)., **S3.06B** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 12, $\bar{x}$ = 64.4, $\tilde{x}_{0.5}$ = 60.5, $\bar{x}_R$ = 61.9, s = 17.2000, $\hat{g}_1$ = 0.59, $\hat{g}_2$ = 2.53, Závěr: 58.0 < μ < 65.6 (Box-Cox).

S3.07 Nehomogenita, průměry jsou shodné: **S3.07A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 19.23, $\tilde{x}_{0.5}$ = 19.12, $\bar{x}_R$ = 19.20, Horn P_L = 19.25, Horn R_L = 1.00,
s = 0.74, $\hat{g}_1$ = 0.10, $\hat{g}_2$ = 2.33, Závěr: 18.58 < μ < 19.92 (Horn),
S3.07B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 19.03, $\tilde{x}_{0.5}$ = 19.00, $\bar{x}_R$ = 19.04, Horn P_L = 19.00, Horn R_L = 2.00, s = 0.99, $\hat{g}_1$ = -0.08, $\hat{g}_2$ = 2.11, Závěr: 17.87 < μ < 20.13 (Horn).

S3.08 Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 6.4, $\tilde{x}_{0.5}$ = 6.2, $\bar{x}_R$ = 6.4, Horn P_L = 6.7, Horn R_L = 2.2, s = 1.6, $\hat{g}_1$ = -0.14, $\hat{g}_2$ = 3.17,
Závěr: 5.83 < μ < 7.57 (Horn).

S3.09 Homogenita, průměry jsou rozdílné: **S3.09A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 776.7, $\tilde{x}_{0.5}$ = 773.0, $\bar{x}_R$ = 774.8, Horn P_L = 807.0, Horn R_L = 582.0, s = 299.7, $\hat{g}_1$ = -0.03, $\hat{g}_2$ = 2.12, Závěr: 489.8 < μ < 1124.2 (Horn), **S3.09B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 1211.4, $\tilde{x}_{0.5}$ = 1258.0, $\bar{x}_R$ = 1244.8, Horn P_L = 1180.5, Horn R_L = 583.0, s = 377.1, $\hat{g}_1$ = -0.55, $\hat{g}_2$ = 2.88, Závěr: 851.7 < μ < 1509.3 (Horn).

S3.10 Homogenita, průměry jsou shodné: **S3.10A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\bar{x}$ = 23.0, $\tilde{x}_{0.5}$ = 22.5, $\bar{x}_R$ = 22.9, Horn P_L = 23.0, Horn R_L = 6.0, s = 4.1, $\hat{g}_1$ = 0.09, $\hat{g}_2$ = 2.22, Závěr: 18.99 < μ < 27.00 (Horn), **S3.10B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\bar{x}$ = 20.0, $\tilde{x}_{0.5}$ = 19.5, $\bar{x}_R$ = 19.6, Horn P_L = 19.5, Horn R_L = 5.0, s = 3.1, $\hat{g}_1$ = 0.29, $\hat{g}_2$ = 1.74, Závěr: 16.2 < μ < 22.8 (Horn).

S3.11 Homogenita, průměry jsou shodné: **S3.11A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 3, $\bar{x}$ = 48.12, $\tilde{x}_{0.5}$ = 47.10, $\bar{x}_R$ = 46.94, Horn P_L = 47.30, Horn R_L = 7.80, s = 7.86, $\hat{g}_1$ = 0.95, $\hat{g}_2$ = 3.78, Závěr: 44.20 < μ < 50.40 (Horn), **S3.11B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\bar{x}$ = 45.97, $\tilde{x}_{0.5}$ = 45.15, $\bar{x}_R$ = 45.11, Horn P_L = 45.05, Horn R_L = 11.10, s = 7.90, $\hat{g}_1$ = 0.75, $\hat{g}_2$ = 3.90, Závěr: 40.64 < μ < 49.46 (Horn).

S3.12 Homogenita, průměry jsou shodné: **S3.12A** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 1, $\overline{x}$ = 1998.8, $\tilde{x}_{0.5}$ = 1980.0, $\overline{x}_R$ = 1986.5, Horn P_L = 1996.0, Horn R_L = 186.0, s = 165.1, $\hat{g}_1$ = 0.50, $\hat{g}_2$ = 3.46, Závěr: 1909.3 < μ < 2082.7 (Horn),

S3.12B Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 0, $\overline{x}$ = 1946.4, $\tilde{x}_{0.5}$ = 1947.0, $\overline{x}_R$ = 1973.5, Horn P_L = 1989.0, Horn R_L = 166.0, s = 162.8, $\hat{g}_1$ = -0.99, $\hat{g}_2$ = 3.24, Závěr: 1837.1 < μ < 2140.9 (Horn).

S3.13 Nehomogenita, průměry jsou shodné: **S3.13A** Asymetrické rozdělení, neprokázáno Gaussovo rozdělení, počet odlehlých bodů: 2, $\overline{x}$ = 250.34, $\tilde{x}_{0.5}$ = 251.35, $\overline{x}_R$ = 251.00, Horn P_L = 250.15, Horn R_L = 0.50, s = 3.36, $\hat{g}_1$ = -0.95, $\hat{g}_2$ = 2.65, Závěr: 247.40 < μ < 252.93 (Horn), **S3.13B** Symetrické rozdělení, prokázáno Gaussovo rozdělení, počet odlehlých bodů: 4, $\overline{x}$ = 250.22, $\tilde{x}_{0.5}$ = 250.20, $\overline{x}_R$ = 250.20, Horn P_L = 250.15, Horn R_L = 0.50, s = 0.47, $\hat{g}_1$ = 0.38, $\hat{g}_2$ = 3.08, Závěr: 249.93 < μ < 250.37 (Horn).

3.7 Doporučená literatura

[1] Hensgaard. D.: *Commun. Statist.* **B8**, 359 (1979).
[2] Tukey J. W., McLaughlin: *Sankya* **125**, 331 (1963).
[3] Johnson N. L., Kotz S.: *Continuous Univariate Distributions*. Mifflin 1970.
[4] Hogg R. V.: *J. Amer. Statist. Assoc.* **69**, 909 (1964).
[5] Du Mond Ch., Lenth R. V.: *Technometrics* **29**, 211 (1987).
[6] Blackman N. M., Machol R. E.: *IEEE Trans. on Inform. Theory* **IT-33**, 373 (1987).
[7] Horn J.: *J. Amer. Statist. Assoc.* **78**, 930 (1983).
[8] Efron B.: *Canad. J. Statist.* **9**, 139 (1981).
[9] Posten H. O., Yeh H. C., Owen D. B.: *Commun. Statist.* **A11**, 109 (1982).
[10] Cressie N. A. C., Whitford H. J.: *Biom. J.* **28**, 131 (1986).
[11] Yuen K., Dixon W. J.: *Biometrika* **60**, 369 (1973).
[12] Owen D. B.: *Handbook of Statistical Tables*. Addison Wesley Publ., Reading 1963.
[13] Green J. R., Margerison D.: *Statistical Treatment of Experimental Data*, Elsevier, Amsterdam 1978.
[14] Miller J. C. a Miller J. N.: *Statistics for Analytical Chemistry*. Ellis Horwood, Chichester, 1984.
[15] Himmelblau D. M.: *Process Analysis by Statistical Methods*. Wiley, New York 1969.
[16] Liteanu C., Rica I.: *Statistical Theory and Methodology of Trace Analysis.* Ellis Horwood, Chichester 1980.
[17] Anderson R. L.: *Practical Statistics for Analytical Chemists*. van Nostrand Reinhold Comp., New York 1987.
[18] Eason G. a kol.: *Mathematics and Statistics for the Bio-Sciences*. Ellis Horwood, Chichester 1980.
[19] Stoodly K.: *Applied and Computation Statistics*. Ellis Horwood, Chichester 1984.
[20] Meloun M., Militký J.: *Statistické zpracování experimentálních dat.* East Publishing, Praha 1998.

4

STATISTICKÁ ANALÝZA VÍCEROZMĚRNÝCH DAT

V technické, biologické ale také lékařské praxi se často vedle informací, obsažených v náhodném skaláru ξ, vyskytují i informace obsažené v náhodném vektoru $\boldsymbol{\xi}$ s m složkami $\xi_1, ..., \xi_m$. Příklady vícerozměrných dat jsou

a) vyjádření vlastností produktů, jako jsou potraviny, oleje, slitiny atd., pomocí řady různých analytických metod,

b) hodnocení spekter pomocí poloh a ploch absorpčních pásů, sloužící k charakterizaci a identifikaci chemických sloučenin,

c) sledování složení surovin, produktů, odpadů, v závislosti na čase nebo na místě výskytu,

d) regulace jakosti na základě různých procesních proměnných,

e) stanovení charakteristiky produktu na základě měření souvisejících proměnných, např. spekter (vícerozměrná kalibrace).

Často je účelem vícerozměrné analýzy zkoumání vztahů mezi složkami náhodného vektoru. Pro tento účel se volí koncepce latentních proměnných (hlavních komponent, faktorů, kanonických proměnných) y, které jsou lineární kombinací původních proměnných x s vhodně volenými vazbami. Latentní proměnná y je kombinací m-tice sledovaných (měřených či jinak získaných) proměnných $x_1, x_2, ..., x_m$ ve tvaru

$$y = w_1 x_1 + w_2 x_2 + ... + w_m x_m \quad .$$

Jednotlivé vícerozměrné metody využívají různých způsobů stanovení vahových koeficientů $w_1, w_2, ..., w_m$.

V této kapitole je věnována pozornost především postupům, patřícím do oblasti průzkumové (exploratorní) analýzy a charakterizace vícerozměrných dat. Detailní popis najde čtenář v doporučované učebnici[22] k této sbírce.

4.1 Popis vícerozměrných dat

Zdrojová matice, tj. matice výchozích dat (popisující např. řadu aut různých značek), obsahuje **proměnné** v m sloupcích (např. obsah motoru, výkon, spotřebu paliva, hmotnost vozu, zrychlení, výšku, šířku, délku atd.) a **objekty** v n řádcích (např. auta různých výrobců

a značek), na nichž jsou tyto proměnné (vlastnosti) měřeny. Protože měřené proměnné mají různé jednotky, a často se řádově liší, bývá zdrojová matice před zpracováním ještě upravována, *škálována*, a to buď (a) *centrováním*, kdy se od prvků sloupce odečte jejich sloupcový aritmetický průměr, nebo (b) *standardizací* čili *normováním*, kdy se prvky centrovaných sloupců ještě vydělí svou sloupcovou směrodatnou odchylkou.

Standardní statistická analýza je založena na předpokladu, že hodnoty x_{ij} tvoří *náhodný výběr.* Tento výběr je tvořen *n*-ticí řádkových vektorů $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,m})$, které lze chápat jako řádky zdrojové matice anebo souřadnice *n* bodů v *m*-rozměr-ném prostoru původních experimentálních proměnných. Výběr lze vyjádřit maticí rozměru ($n \times m$)

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_i^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,m} \\ \vdots & & \vdots & & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,m} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,j} & \cdots & x_{n,m} \end{bmatrix} .$$

Řádek zdrojové matice čili *i*-tý vektor $\boldsymbol{x}_i^{\mathrm{T}} = (x_{i,1}, ..., x_{i,m})$ nazýváme *objektem* (např. auto určitého typu) a můžeme ho chápat jako jeden bod v *m*-rozměrném prostoru. Tento objekt je charakterizován svými *proměnnými*, a to buď *kvantitativními,* metrickými, tj. číselnými hodnotami, nebo proměnnými *kvalitativními*, nemetrickými.

Metrické proměnné se vyskytují ve čtyřech škálách:

(a) *Proměnné v absolutní škále* mají přirozený počátek a jeden parametr měřítka, např. obsah uhlíku v %, rychlostní konstanta.

(b) *Proměnné v poměrové škále* mají zachován podíl hodnot charakteristik $c = x_2/x_1$, např. vztah vůči standardní sloučenině, vztah vůči jevu s definovaným nulovým počátkem, parametr σ v Hammettově rovnici.

(c) *Proměnné v intervalové škále* mají zachován podíl rozdílů $c = x_2 - x_1$. Jedná se o poměrovou škálu s přirozeným počátkem pro obě srovnávané hodnoty, např. poměr absorbancí indikátoru, vztažený na absorbanci nulové linie.

(d) *Proměnné v rozdílové škále* jsou vztahovány k různému počátku, např. hodnoty časových škál, stáří atd.

Nemetrické proměnné se vyskytují ve škálách:

(a) *Proměnné v nominální škále* jsou nejméně informativní. Je kvantifikována pouze rovnost nebo různost tříd. Obsahují kód, např. barvu výrobku, vyjádřenou kódem 1 až 16, rodinný stav (svobodný 1, ženatý 2, rozvedený 3, vdovec 4).

(b) *Proměnné v ordinální škále* jsou seřazené do tříd. Je definována relace větší anebo menší mezi třídami, a také kvantifikován rozdíl, např. žebříček umístění, pořadové číslo.

(c) *Proměnné v alternativní (binární) škále* vyjadřují rovnost či nerovnost vůči nějakému kritériu. Mají binární charakter, který můžeme popsat dvojicí 1 (ano), 0 (ne).

Třídu nebo **shluk objektů** chápeme jako množinu objektů se společnými nebo alespoň podobnými proměnnými, znaky (např. auta typu BMW). Blízkost či podobnost objektů posuzujeme na základě *míry vzdálenosti objektů* v *m*-rozměrném prostoru proměnných.

Mírou vzdálenosti objektů pro *kvantitativní* proměnné jsou běžně základní metriky:

Eukleidova metrika, čili *geometrická vzdálenost*, je standardním typem vzdálenosti, který je definován vztahem

$$d_{\mathrm{E}}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{\sum_{j=1}^{m} (x_{kj} - x_{lj})^2},$$

Hammingova metrika, čili *Manhattanská vzdálenost*, je definována vztahem

$$d_{\mathrm{H}}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sum_{j=1}^{m} |x_{kj} - x_{lj}|,$$

zobecněná Minkowskiho metrika vztahem

$$d_{\mathrm{M}}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt[n]{\sum_{j=1}^{m} |x_{kj} - x_{lj}|^n},$$

kde pro $n = 1$ jde o Hammingovu metriku a pro $n = 2$ o Eukleidovu. Čím je n větší, tím více je zdůrazňován rozdíl mezi vzdálenými objekty. Všechny tyto metriky neuvažují závislost mezi proměnnými. Zahrneme-li do vztahu pro vzdálenost i vazby mezi proměnnými, vyjádřené kovarianční maticí $\boldsymbol{C}$, dostaneme statistickou míru, zvanou *Mahalanobisova metrika*

$$d_{\mathrm{MA}}(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sqrt{(\boldsymbol{x}_k - \boldsymbol{x}_l)^{\mathrm{T}}\, \boldsymbol{C}^{-1}\, (\boldsymbol{x}_k - \boldsymbol{x}_l)}.$$

Ta se společně s Eukleidovou metrikou nejvíce používá v praxi. Ve všech uvedených případech jsou si dva objekty tím bližší, čím je jejich vzdálenost menší.

Mírou podobnosti dvou objektů či proměnných x_i a x_j může být *Pearsonův párový korelační koeficient r*. Objekty jsou si tím podobnější, čím je párový korelační koeficient větší a bližší jedničce. V případě ordinální škály je analogickou mírou podobností *Spearmanův korelační koeficient.* Podobnost binárních nebo nominálních proměnných vyjadřují různé koeficienty asociace. Označíme-li počet případů negativní shody typu 0-0 písmenem *a*, počet případů s neshodou typu 1-0 písmenem *b*, počet případů s neshodou typu 0-1 písmenem *c* a počet případů s pozitivní shodou typu 1-1 písmenem *d*, můžeme definovat tyto koeficienty podobnosti:

(a) *Sokalův-Michenerův koeficient asociace*

$$S_{SM} = \frac{a + d}{a + b + c + d},$$

(b) *Russelův-Raoův koeficient asociace*

$$S_{RR} = \frac{d}{a + b + c + d},$$

(c) *Hamannův koeficient asociace*

$$S_H = \frac{a + d - b - c}{a + b + c + d} ,$$

a také lze konstruovat *obdobu korelačního koeficientu*

$$r_B = \frac{a d - b c}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} .$$

Míra podobnosti mezi objekty, charakterizovanými různými typy proměnných, se vypočte jako vážený průměr jednotlivých měr podobnosti. Na základě měr podobnosti objektů se konstruují míry podobnosti mezi objekty a třídami a míry podobnosti mezi třídami. Jako nejčastější míra podobnosti se používá vzdálenost tříd $d(x_k, x_l)$. Analogicky zde užijeme způsobů vyjádření vzdálenosti objektů, protože objekt můžeme chápat jako třídu o jednom objektu. Čím větší je vzdálenost, tím menší je podobnost:

(a) *Vzdálenost nejbližšího souseda:* nejbližší jsou ty třídy či shluky, které mají nejmenší vzdálenost mezi dvěma nejbližšími objekty dvou pozorovaných tříd.

(b) *Vzdálenost nejvzdálenějšího souseda:* nejbližší jsou ty třídy či shluky, které mají nejmenší vzdálenost mezi dvěma nejvzdálenějšími objekty.

(c) *Vzdálenost mezi těžišti tříd:* nejbližší jsou ty třídy či shluky, které mají nejmenší vzdálenost mezi svými těžišti.

(d) *Vzdálenost průměrné vazby:* nejbližší jsou ty třídy či shluky, které mají nejmenší průměrnou vzdálenost mezi všemi objekty jedné a všemi objekty druhé třídy.

4.2 Obecný postup analýzy vícerozměrných dat

Postup analýzy vícerozměrných dat záleží na typu dat a na druhu požadované informace, jež se z dat má získat.

Typ dat:

Otázky: Před vlastní analýzou je třeba zodpovědět tři základní otázky:

(1) Je možné rozdělit vyšetřované proměnné na *závislé* a *nezávislé*?

(2) Kolik proměnných se uvažuje jako *závisle proměnných*?

(3) V jaké škále jsou jednotlivé proměnné měřeny, tj. *kardinální* čili číselné, *ordinální* čili pořadové nebo *nominální* čili znakové. Kardinální škála se označuje jako *metrická* a ostatní dvě, ordinální a nominální, jako škály *nemetrické*.

Odpovědi:

(1) Pokud je odpověď na první otázku kladná, volí se techniky pro stanovení *vztahu* mezi závisle proměnnými a vhodnou kombinací nezávisle proměnných.

(2) Pokud je odpověď na první otázku záporná, volí se techniky pro stanovení *vzájemných vazeb*, tj. provádí se simultánní analýza všech proměnných.

Typ informace:

Jednotlivé techniky pro *stanovení závislosti* se dále dělí podle počtu závisle proměnných a podle škály měření. Schematicky lze vztahy mezi jednotlivými technikami analýzy vícerozměrné závislosti zapsat ve formě těchto přiřazení:

(a) **Kanonická korelace (CC)**:

$$y_1 \,\%\, y_2 \,\%\, \ldots \,\%\, y_m \quad \mathrm{Z} \quad x_1 \,\%\, x_2 \,\%\, \ldots \,\%\, x_m$$
(metrická, nemetrická) (metrická, nemetrická) ,

(b) **Vícerozměrná analýza rozptylu (MANOVA)**:

$$y_1 \,\%\, y_2 \,\%\, \ldots \,\%\, y_m \quad \mathrm{Z} \quad x_1 \,\%\, x_2 \,\%\, \ldots \,\%\, x_m$$
(metrická) (nemetrická) ,

(c) **Analýza rozptylu (ANOVA)**:

$$y_1 \quad \mathrm{Z} \quad x_1 \,\%\, x_2 \,\%\, \ldots \,\%\, x_n$$
(metrická) (nemetrická) ,

(d) **Diskriminační analýza (DA)**:

$$y_1 \quad \mathrm{Z} \quad x_1 \,\%\, x_2 \,\%\, \ldots \,\%\, x_n$$
(nemetrická) (metrická) ,

(e) **Vícerozměrná regrese a kalibrace**:

$$y_1 \quad \mathrm{Z} \quad x_1 \,\%\, x_2 \,\%\, \ldots \,\%\, x_m$$
(metrická) (metrická, nemetrická) ,

(f) **Analýza "conjoint"**:

$$y_1 \quad \mathrm{Z} \quad x_1 \,\%\, x_2 \,\%\, \ldots \,\%\, x_m$$
(metrická, nemetrická) (nemetrická) ,

(g) **Strukturní rovnice**:

$$\begin{array}{lll} y_1 & \mathrm{Z} & x_{11} \,\%\, x_{12} \,\%\, \ldots \,\%\, x_{1m} \\ y_2 & \mathrm{Z} & x_{21} \,\%\, x_{22} \,\%\, \ldots \,\%\, x_{2m} \\ & & \ldots\ldots \\ y_n & \mathrm{Z} & x_{n1} \,\%\, x_{n2} \,\%\, \ldots \,\%\, x_{nm} \end{array}$$
(metrická) (metrická, nemetrická)

Dělení od zcela obecné techniky (kanonická korelace) až k velmi speciálnímu případu (strukturní rovnice) umožňuje výběr konkrétní analýzy dat s ohledem na cíl analýzy a počet a typ závisle, resp. nezávisle proměnných.

4.3 Charakteristiky vícerozměrných náhodných veličin

Intenzita vztahu mezi proměnnými. K charakterizaci polohy j-té proměnné ξ_j, tj. j-tého sloupce zdrojové matice $\boldsymbol{X}$ se používá **střední hodnota** $E(\xi_j) = \mu_j$ a pro charakterizaci rozptýlení **rozptyl** $D(\xi_j) = \sigma_j^2$. Dále je třeba definovat *míru intenzity* vztahu mezi proměnnými ξ_i a ξ_j, $j = i$. Vhodnou charakteristikou je *druhý smíšený centrální moment*, nazývaný **kovariance** $\mathrm{cov}(\xi_i, \xi_j)$, definovaný vztahem

$$cov(\xi_i, \xi_j) \quad ' \quad E(\xi_i\, \xi_j) \;\&\; E(\xi_i)\, E(\xi_j) \quad .$$

Kovariance má vlastnosti:

a) Její znaménko ukazuje na trend stochastické vazby mezi *j*-tým a *i*-tým sloupcem matice.

b) Je v absolutní hodnotě shora ohraničená součinem $\sigma_i \sigma_j$, tj.

$$* \mathrm{cov}(\xi_i, \xi_j)* \# \sigma_i \sigma_j .$$

c) Je symetrickou funkcí svých argumentů.

d) Nemění se posunem počátku, ale změna měřítka se projeví úměrně jeho velikosti. Pro čísla a_1, a_2, b_1, b_2 pak platí, že

$$\mathrm{cov}(a_1 \xi_i \% b_1,\ a_2 \xi_j \% b_2) \quad ' \quad a_1 a_2 \mathrm{cov}(\xi_i, \xi_j) \quad .$$

e) Pro nekorelované náhodné veličiny je $\mathrm{cov}(\xi_i, \xi_j) = 0$ a mohou nastat dva případy:

1. $E(\xi_i \xi_j) = 0$ a zároveň $E(\xi_i) = E(\xi_j) = 0$, což je případ *centrovaných ortogonálních* náhodných veličin, ne nutně nezávislých.
2. $E(\xi_i \xi_j) = E(\xi_i) E(\xi_j)$, což je případ *nezávislých* náhodných veličin.

f) Je *mírou intenzity lineární závislosti*.

Nevýhodou kovariance je fakt, že její hodnoty závisí na měřítku, ve kterém jsou vyjádřeny proměnné ξ_i a ξ_j. Její velikost je omezena součinem $\sigma_i \sigma_j$. Je proto přirozené provést standardizaci dělením tímto součinem. Vzniklá veličina $\rho_{ij} = \rho(\xi_i, \xi_j)$ se nazývá *Pearsonův párový korelační koeficient*

$$\rho(\xi_i, \xi_j) \quad ' \quad \rho_{ij} \quad ' \quad \frac{\mathrm{cov}(\xi_i, \xi_j)}{\sigma_i \sigma_j} \quad .$$

Je zřejmé, že párový korelační koeficient leží v rozmezí $-1 \# \rho_{ij} \# 1$. Pokud je $\rho_{ij} > 0$, jde o *pozitivně korelované* náhodné veličiny, a pokud je $\mathsf{k}_{ij} < 0$, jde o *negativně korelované* náhodné veličiny.

Pearsonův párový korelační koeficient má vlastnosti:

a) Rovnost $*\rho_{ij}* = 1$ ukazuje, že mezi ξ_i a ξ_j existuje přesně lineární vztah.

b) Pokud jsou náhodné veličiny ξ_i a ξ_j vzájemně nekorelované, je $\rho_{ij} = 0$.

c) V případě, že ξ_i a ξ_j pocházejí z vícerozměrného normálního rozdělení a $\rho_{ij} = 0$, znamená to, že jsou *vzájemně nezávislé*.

d) Platí, že i pro nelineárně závislé náhodné veličiny může být $\rho_{ij} = 0$.

e) Korelační koeficient ρ_{ii} náhodné veličiny ξ_i samotné se sebou je roven jedné.

f) Korelační koeficient je invariantní vůči lineární transformaci náhodných proměnných ξ_i , ξ_j. Pro čísla a_1 , a_2 , b_1 , b_2 platí vztah

$$\rho(a_1 \xi_i \% b_1,\ a_2 \xi_j \% b_2) \quad ' \quad \mathrm{sign}(a_1 a_2)\, \rho(\xi_i, \xi_j),$$

kde sign(x) je znaménková funkce, pro kterou platí

$$\mathrm{sign}(x) \quad ' \quad \begin{cases} \&1 & \text{pro } x < 0 \\ 0 & \text{pro } x \ ' \ 0 \\ 1 & \text{pro } x > 0 \end{cases} .$$

Ověření normality. Jako při analýze jednorozměrných dat, hraje také u vícerozměrných výběrů hlavní roli předpoklad, zda data pocházejí z *normálního rozdělení*. Tento předpoklad usnadňuje zejména statistickou analýzu vektoru středních hodnot nebo kovarianční matice.

Podobně jako v jednorozměrném případě, existuje i zde řada testů, které jsou více či méně citlivé vůči různým typům narušení normality. Nenormalita může být například způsobena vybočujícími objekty $\boldsymbol{x}_i$ či pouze některými vybočujícími hodnotami x_{ij} Mezi nejjednodušší metody ověřování normality patří testy založené na vícerozměrné šikmosti $g_{1,m}$ a vícerozměrné špičatosti $g_{2,m}$ V tomto případě se testuje simultánní platnost nulových hypotéz H_{01}: $g_{1,m} = 0$ a H_{02}: $g_{2,m} = m(m + 2)$.

Odhady parametrů polohy a rozptýlení. Z vícerozměrného výběru objektů o velikosti n, definovaného n-ticí m-rozměrných objektů $\boldsymbol{x}^{\mathrm{T}}{}_i = (x_{i,1}, ..., x_{i,m})^{\mathrm{T}}$, $i = 1, ..., n$, je možno stanovit *výběrový vektor středních hodnot* $\hat{\boldsymbol{\mu}}$ určený vztahem

$$\hat{\boldsymbol{\mu}} \quad ' \quad \frac{1}{n} \sum_{i}^{n} \boldsymbol{x}_i^{\mathrm{T}} \quad .$$

Podobně pro *odhad kovarianční matice* $\boldsymbol{S}^{\boldsymbol{0}}$ platí rovnice

$$\boldsymbol{S}^{\mathbf{0}} \quad ' \quad \frac{1}{n} \sum_{i}^{n} (\boldsymbol{x}_i \,\&\, \hat{\boldsymbol{\mu}})\,(\boldsymbol{x}_i \,\&\, \hat{\boldsymbol{\mu}})^{\mathrm{T}} \quad .$$

Míra polohy náhodného vektoru se charakterizuje pomocí *vektoru středních hodnot* $\boldsymbol{\mu}^{\mathbf{T}} \quad ' \quad [E(\xi_1), \ldots, E(\xi_m)]$, a míra rozptýlení pomocí *kovarianční matice* řádu $m \times m$

$$\boldsymbol{C} \quad ' \quad \begin{bmatrix} D(\xi_1) & \mathrm{cov}(\xi_2, \xi_1) & \ldots & \mathrm{cov}(\xi_i, \xi_1) & \ldots & \mathrm{cov}(\xi_m, \xi_1) \\ \mathrm{cov}(\xi_1, \xi_2) & D(\xi_2) & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(\xi_1, \xi_m) & \mathrm{cov}(\xi_2, \xi_m) & \ldots & \mathrm{cov}(\xi_i, \xi_m) & \ldots & D(\xi_m) \end{bmatrix} .$$

Místo kovarianční matice se používá také její normovaná verze, tj. *korelační matice*

$$\boldsymbol{R} = \begin{bmatrix} 1 & k_{21} & \cdots & k_{i1} & \cdots & k_{m1} \\ k_{12} & 1 & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ k_{1m} & k_{2m} & \cdots & k_{im} & \cdots & 1 \end{bmatrix} .$$

Korelační matice má na diagonále samé jedničky a mimodiagonální prvky jsou jednotlivé *Pearsonovy párové korelační koeficienty*. Kovarianční matice $\boldsymbol{C}$ i korelační matice $\boldsymbol{R}$ jsou symetrické.

Pro vektor výběrových středních hodnot platí

$$E(\hat{\boldsymbol{\mu}}) = \boldsymbol{\mu} \quad \text{a} \quad D(\hat{\boldsymbol{\mu}}) = \frac{1}{n}\boldsymbol{C} .$$

Odhad $\hat{\boldsymbol{\mu}}$ je tedy nevychýlený. Pro odhad kovarianční matice platí, že

$$E(\boldsymbol{S}^{0}) = \frac{n-1}{n}\boldsymbol{C}$$

a jde o vychýlený odhad. Proto se používá *výběrová korigovaná kovarianční matice*

$$\boldsymbol{S} = \frac{n}{n-1}\boldsymbol{S}^{0} ,$$

která je již nevychýleným odhadem kovarianční matice $\boldsymbol{C}$. Matice $\boldsymbol{S}^{0}$ je *výběrová kovarianční matice*. Odhady $\hat{\boldsymbol{\mu}}$ a $\boldsymbol{S}^{0}$ jsou maximálně věrohodné, tzn. že náhodný výběr, charakterizovaný maticí $\boldsymbol{X}$ pochází z normálního rozdělení $N(\boldsymbol{\mu}, \boldsymbol{C})$. Za stejných podmínek má $\hat{\boldsymbol{\mu}}$ rozdělení $N(\boldsymbol{\mu}, \boldsymbol{C}/n)$.

Pokud máme dva vektory, $\boldsymbol{\xi}_1$ a $\boldsymbol{\xi}_2$, které jsou nezávislé a stejně rozdělené se střední hodnotou $\boldsymbol{\mu}$ a kovarianční maticí $\boldsymbol{C}$, je *vícerozměrná šikmost* dána vztahem

$$g_{1,m} = E[(\boldsymbol{\xi}_1 - \boldsymbol{\mu})^{\mathrm{T}}\, \boldsymbol{C}^{-1}\, (\boldsymbol{\xi}_2 - \boldsymbol{\mu})]^3$$

a pro *vícerozměrnou špičatost* platí

$$g_{2,m} = E[(\boldsymbol{\xi}_1 - \boldsymbol{\mu})^{\mathrm{T}}\, \boldsymbol{C}^{-1}\, (\boldsymbol{\xi}_1 - \boldsymbol{\mu})]^2 .$$

K vyjádření funkcí $g_{1,m}$ a $g_{2,m}$ lze využít i vícerozměrných centrálních momentů. Speciálně pro případ vícerozměrného normálního rozdělení pak platí, že

$$g_{1,m} = 0 \quad \text{a} \quad g_{2,m} = m(m+2) .$$

Vzorová úloha 4.1 *Popisné charakteristiky vícerozměrných náhodných veličin*

Na úloze **B4.02** *Účinky neuroleptik při tlumení rozličných psychóz* si ukážeme pomůcky vícerozměrné analýzy dat. K analýze užijeme také škálovaná data.
Řešení: byl použit program NCSS2000.

1. Popisné statistiky: mezi popisné statistiky patří především míry polohy a rozptýlení.

Proměnná	***n***	$\bar{x}$	***s***
B402X1	20	20.10	33.90
B402X2	20	18.68	33.84
B402X3	20	3.01	5.21
B402X4	20	10.44	36.65

Tabulka umožňuje porovnat klasické odhady měr polohy a rozptýlení čtyř sledovaných proměnných, tj. za předpokladu normality. Proměnné *B402X1* a *B402X2* mají podobné hodnoty aritmetického průměru a velmi blízké míry rozptýlení, směrodatné odchylky. Zbývající dvě proměnné se pak výrazně liší od dvou předešlých.

2. Kovarianční matice *C*:

Proměnná	*B402X1*	*B402X2*	*B402X3*	*B402X4*
B402X1	1140.90	1127.90	148.50	1044.50
B402X2	1127.90	1136.40	140.15	1051.20
B402X3	148.50	140.15	27.32	159.96
B402X4	1044.50	1051.20	159.96	1340.40
Vícerozměrná šikmost g_1: 32.14, vícerozměrná špičatost g_2: 46.708				

Kovariance mezi dvěma proměnnými. Na diagonále jsou rozptyly dotyčné proměnné. Pod tabulkou jsou vícerozměrná šikmost a vícerozměrná špičatost.

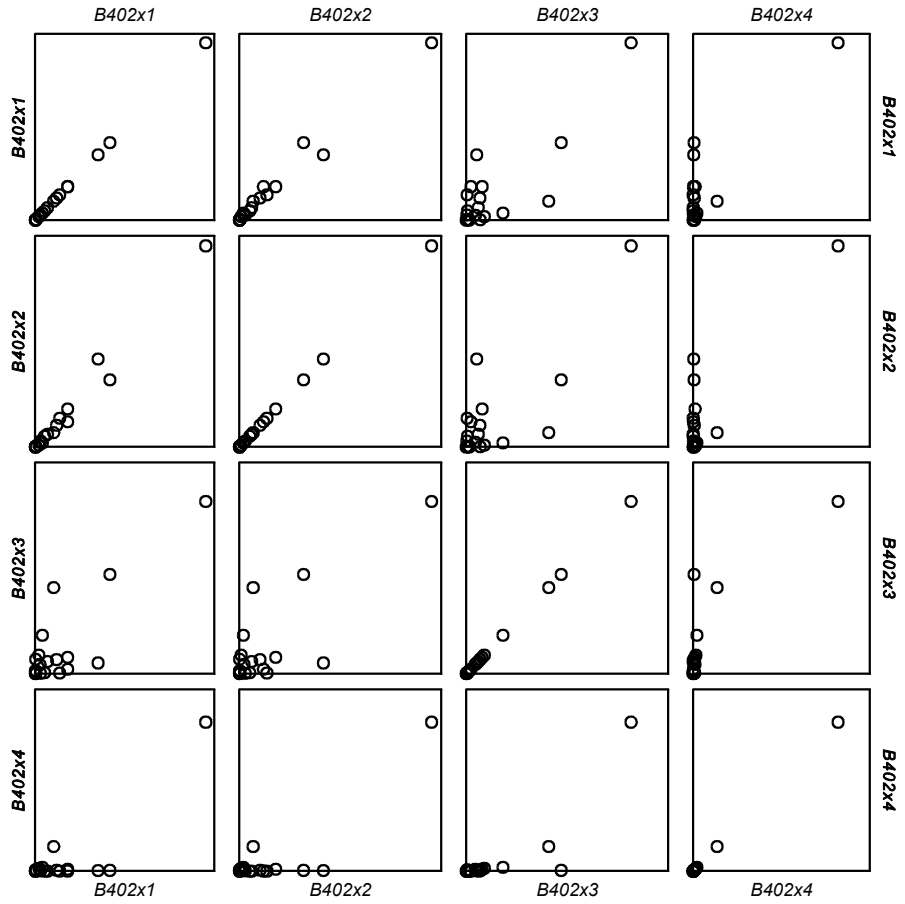
3. Korelační matice *R*:

Proměnná	*B402X1*	*B402X2*	*B402X3*	*B402X4*
B402X1	1.0000	0.9905	0.8359	0.8445
B402X2	0.9905	1.0000	0.7864	0.8518
B402X3	0.8359	0.7864	1.0000	0.8238
B402X4	0.8445	0.8518	0.8238	1.0000

Korelace mezi dvěma proměnnými vystihuje míru lineární závislosti mezi proměnnými. Platí zásada: je-li korelace mezi proměnnými malá, není vůbec třeba užít metod hlavních komponent PCA nebo faktorové analýzy FA.

4.4 Exploratorní analýza struktury objektů (EDA)

Průzkumová analýza vícerozměrných dat je stejně jako u jednorozměrných dat založena na vyšetření grafických diagnostik. K tomuto účelu se využívá různých technik zobrazování vícerozměrných dat. Pro případ, kdy jsou jednotlivé sloupce matice $\boldsymbol{X}$ málo korelované postačují *rozptylové diagramy* pro jednotlivé kombinace složek vektoru $\boldsymbol{x}$ a pro nekorelované pak sloupce matice $\boldsymbol{X}$.



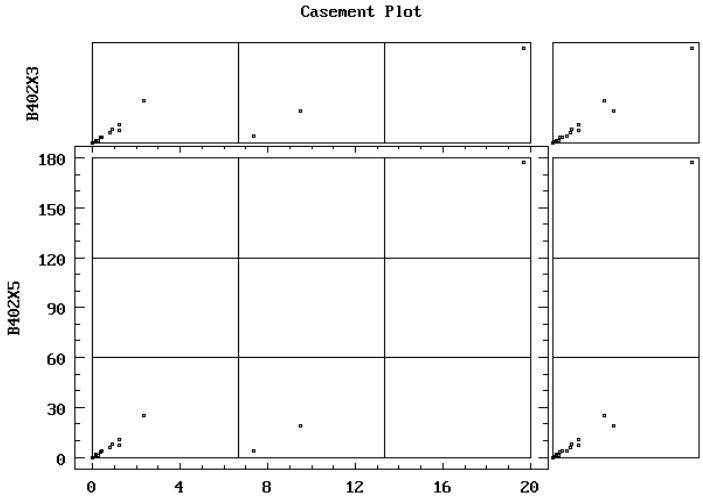
Obr. 4.1 Rozptylový diagram pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* nestandardizovaných dat B402, ***SCAN.*** *Je patrná podobnost objektů a vysoká korelovanost zejména prvních dvou proměnných. Jsou patrné i odlehlé objekty, představované body vzdálenými od ostatních*.

Rychlé posouzení podobnosti mezi jednotlivými objekty čili řádky datové matice usnadňují především *symbolové grafy*. Jednotlivé proměnné jsou v nich "kódovány" s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů, *symbolů*. Každému objektu $\boldsymbol{x}_i$ (např. autu) tak odpovídá jistý obrazec zvaný *symbol*. Vlastnosti dat se posuzují s ohledem na vizuální rozdíly mezi symboly. Tím lze v jednom grafu rozlišit *více proměnných* x_j , $j = 1, ..., m$. Prvním krokem před vlastním zobrazením do symbolů je obvykle standardizace. Mezi základní typy zobrazovaných symbolů patří *profily*, *polygony*, *tváře*, *křivky* a *stromy*.

Profily představují dvourozměrné zobrazení m-rozměrných objektů. Každý objekt $\boldsymbol{x}_i$ je charakterizován m proměnnými, zobrazenými zde vertikálními úsečkami. Jejich velikost je úměrná hodnotě odpovídající proměnné x_{ij} , $j = 1, ..., m$. Profil pak vzniká spojením koncových bodů těchto úseček. Je vhodné použít standardizované proměnné dle vzorce

$$x_{ij}^{(} \quad ' \quad \frac{x_{ij}}{(\max_{i} *x_{ij}*)} \quad ,$$

kde max$*x_{ij}*$ je maximální hodnota absolutní velikosti proměnné $\boldsymbol{x}_j$ vektoru $\boldsymbol{x}_i^T$ přes všechny body, $i = 1, ..., n$. Profily jsou jednoduché a umožňují snadné určení rozdílů mezi jednotlivými objekty $\boldsymbol{x}_i$ a $\boldsymbol{x}_k$. Snadno lze takto identifikovat vybočující objekt.



Obr. 4.2 Korelační diagram (Casement Plot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* nestandardizovaných dat B402, ***STATGRAPHICS.*** *Je patrná vysoká korelovanost čtyř sledovaných proměnných. V pravém horním rohu jsou patrné odlehlé objekty.*

Polygony jsou vlastně profily v polárních souřadnicích, kdy každá proměnná objektu $\boldsymbol{x}_i^{\mathbf{T}}$, $i = 1, ..., n,$ odpovídá délce paprsku vycházejícího ze společného středu. Paprsky dělí kružnici ekvidistantně, proměnné jsou standardizovány do intervalu [0, 1]. Mezi polygony patří *graf slunečních paprsků* a *hvězdicový graf.*

(a) **Graf slunečních paprsků** má tvar "sluníčka", které se skládá z paprsků, začínajících ve společném bodě, a úseček spojujících paprsky, které tak tvoří polygon. Zde každá proměnná x_{ij} objektu $\boldsymbol{x}_i^{\mathbf{T}}$ odpovídá délce paprsku vycházejícího ze středu sluníčka. Paprsky jsou rozmístěny ekvidistantně, ve stejných vzdálenostech na kružnici, a proto se provádí lineární transformace do intervalu $[a, 1]$, kde a je zvolená spodní mez, obyčejně $a = 0$. Pro tuto transformaci platí, že

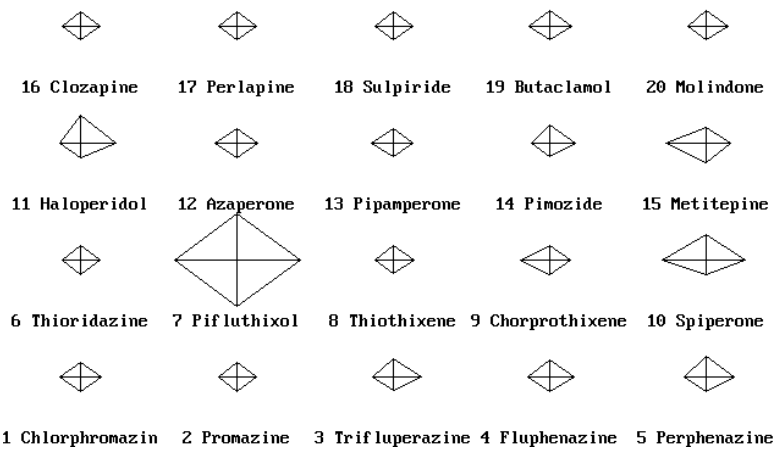
$$x_{ij}^{(} \quad ' \quad \frac{(1 \;\&\; a)\,(x_{ij} \;\&\; \min_{i} x_{ij})}{\max_{i} x_{ij} \;\&\; \min_{i} x_{ij}} \;\%\; a \quad ,$$

kde min x_{ij} je minimální a max x_{ij} maximální hodnota j-té proměnné objektu $\boldsymbol{x}_i^{\mathbf{T}}$ přes všechny objekty $\boldsymbol{x}_i^{\mathbf{T}}$, $i = 1, ..., n$. K určení směrů jednotlivých paprsků se definuje jejich úhel α_j, pro který platí

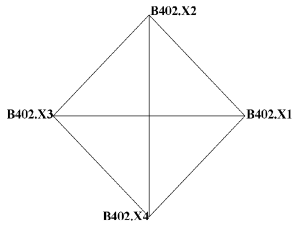
$$\alpha_j = \frac{2\pi(j-1)}{m}, \qquad j = 1, \ldots, m .$$

Za společný střed paprsků se obyčejně volí počátek souřadnic. Pokud má být maximální délka paprsků rovna R, je polygon pro objekt $\boldsymbol{x}_i^{\mathrm{T}}$ spojnicí m bodů p_{ij} o souřadnicích $p_{ij} = (x_{ij} R\cos\alpha_j, x_{ij} R\sin\alpha_j)$. Aby vznikl uzavřený obrazec, spojují se ještě první a poslední bod p_{i1} a p_{im}. Vzájemné porovnání polygonů slouží k vizuálnímu posouzení podobnosti objektů. V případě velkého počtu proměnných, např. $m > 6$, bývá však výsledný obrázek polygonů nepřehledný.

(b) **Hvězdicový graf** vypadá na první pohled jako předchozí graf sluníčka. Sestává z paprsků, reprezentujících relativní hodnoty proměnných u jednotlivých objektů, které se pro každý objekt spojují v jednom centrálním bodě. Stejně směřující paprsky u různých objektů se liší svojí délkou. *Nejkratší paprsek* indikuje, že u objektu nabývá příslušná proměnná nejmenší hodnoty z celého výběru. Podobně *nejdelší paprsek* informuje o nejvyšší hodnotě příslušné proměnné. Délky ostatních paprsků se pohybují podle relativní velikosti hodnot proměnné u příslušného objektu mezi těmito dvěma krajními mezemi.

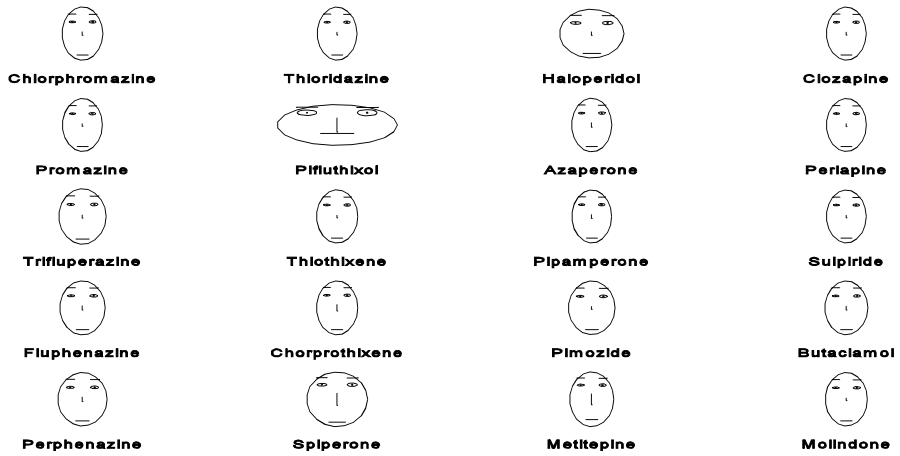


Obr. 4.3a Hvězdičkový graf (Stars Plot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* standardizovaných dat B402, ***STATGRAPHICS.***

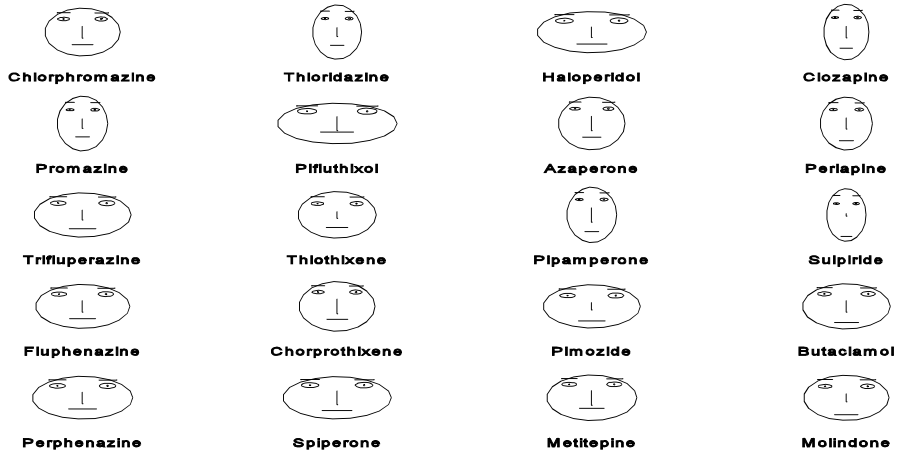


Obr. 4.3b Klíč ke hvězdičkovému grafu pro 4 proměnné *B402X1, B402X2, B402X3, B402X4* standardizovaných dat B402, ***STATGRAPHICS.***

Tváře charakterizují každou proměnnou x_{ij} objektu $\boldsymbol{x}_i^{\mathrm{T}}$ nějakým znakem. Mezi znaky patří tvar tváře, délka nosu, velikost očí, tvar úst atp. Tvar tváře závisí na použitém pořadí proměnných, které ovlivňuje snadnost interpretace dat.



Obr. 4.4 Tváře nestandardizovaných dat B402 pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4,* ***S-Plus.*** Lze nalézt řadu vzájemně podobných tváří, ukazujících na podobnost objektů. Tvář Pifluthixolu se jeví silně odlišná od ostatních.



Obr. 4.5 Tváře zlogaritmovaných dat B402 pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4*, ***S-Plus***. *Logaritmováním dat se rozdíly mezi objekty poněkud setřou a odlehlé objekty nejsou při porovnání podobnosti tak výrazné odlišné. Tvář Pifluthixolu se však stále jeví odlišná od ostatních.*

Křivky využívají transformace každého objektu $\boldsymbol{x}_i^{\mathrm{T}}$ do spojité křivky, která je lineární kombinací všech jeho proměnných. Andrews[12] volí pro vyjádření křivky f_i odpovídajícího objektu $\boldsymbol{x}_i^{\mathrm{T}}$ konečnou Fourierovu řadu

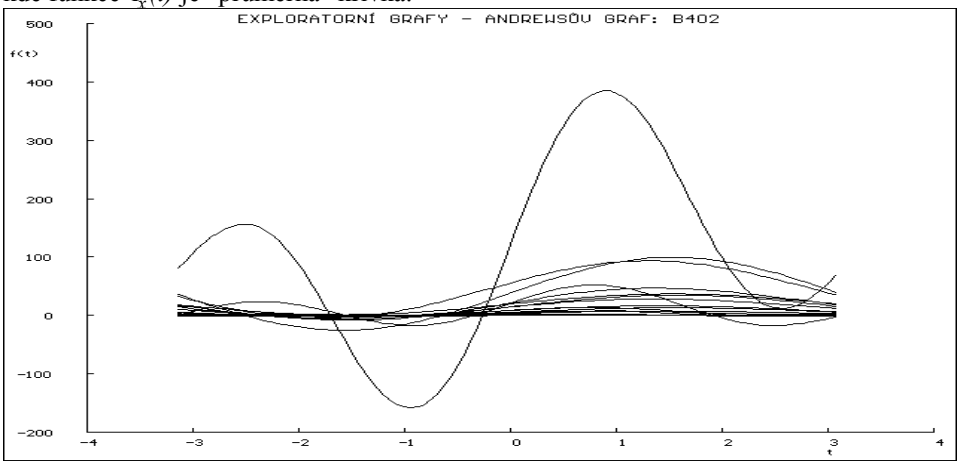
$$f_{\boldsymbol{x}_i}(t) \equiv f_i \equiv \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(t) + x_{i3}\cos(t) + x_{i4}\sin(2t) + x_{i5}\cos(2t) + \ldots$$

Křivky f_i, $i = 1, \ldots, n$, se vynášejí jako funkce proměnné t v intervalu $-\pi \le t \le \pi$. Funkce f_i mají řadu výhodných vlastností:

a) Funkce f_i zachovávají *průměr*. To znamená, že pokud je $\bar{\boldsymbol{x}}$ průměrem z celkového počtu n vícerozměrných dat $\boldsymbol{x}_i$, je funkce rovna

$$f_{\bar{x}}(t) \quad ' \quad \frac{1}{n} \sum_{i}^{n} f_{x_i}(t),$$

kde funkce $f_{\bar{x}}(t)$ je "průměrná" křivka.



Obr. 4.6 Andrewsův graf křivek dat pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* nestandardizovaných dat B402, ***S-Plus***, *Graf ukazuje na značnou podobnost celé řady objektů. Jeden objekt je však výrazně odlišný od ostatních, jde o odlehlý objekt.*

b) Funkce f_i zachovávají *vzdálenosti*. To znamená, že celková vzdálenost mezi křivkami f_i a f_j, definovaná jako integrální kvadratická odchylka, odpovídá vzdálenosti mezi objekty $\boldsymbol{x}_i^{\mathbf{T}}$ a $\boldsymbol{x}_j^{\mathbf{T}}$. Blízké křivky ukazují na nepříliš vzdálené objekty.

c) Pro zvolenou hodnotu t_0 je funkce $f_{x_i}(t_0)$ projekcí objektu $\boldsymbol{x}_i$ na vektor $\boldsymbol{p}_0$ o složkách

$$\boldsymbol{p}_0 \quad ' \quad (\frac{1}{\sqrt{2}}, \sin(t_0), \cos(t_0), \sin(2t_0), \cos(2t_0), ...).$$

Tato projekce do jednoho bodu umožňuje odhalení vybočujících objektů či skupin objektů, které mohou být ve více dimenzích špatně identifikovatelné. Křivka $f_{x_i}(t)$ je složena ze všech projekcí na daném intervalu hodnot t.

d) Funkce f_i zachovávají *rozptyl*. To znamená, že pokud jsou proměnné x_j objektu $\boldsymbol{x}_i^{\mathbf{T}}$ nekorelované náhodné veličiny se stejným rozptylem σ^2, je

$$D(f_i) \quad ' \quad \sigma^2 (0.5 \; \% \; \sin^2(t) \; \% \; \cos^2(t) \; \% \; \sin^2(2t) \; \% \; \cos^2(2t) \; \% \; ...).$$

Pro liché m je $D(f_i) = 0.5\ \sigma^2$ m a pro sudé m je $0.5\ \sigma^2\ (m - 1) < D(f_i) < 0.5\ \sigma^2\ (m + 1)$. Rozptyl funkce f_i je téměř konstantní v celém rozmezí veličiny t.

V praktických úlohách je běžné, že jednotlivé proměnné jsou silně korelované a mají nestejné rozptyly. Pak je výhodné převést objekty původních dat $\boldsymbol{x}_i$ na objekty $\boldsymbol{y}_i$, kde y_{ij} odpovídá transformaci do j-té hlavní komponenty. Veličiny y_{ij} jsou již nekorelované. Snadno lze provést i jejich standardizaci tak, aby měly konstantní rozptyly. Nevýhodou křivek je to, že jejich tvar závisí na pořadí složek. Na druhé straně lze pomocí křivek snadno indikovat vybočující objekty nebo skupiny objektů a konstruovat i konfidenční křivky. Pro větší počty objektů ($n > 10$) dochází ke splývání křivek, což ztěžuje jejich interpretaci. Pak je možné vynášet pouze zvolené podskupiny objektů.

Stromy čili **dendrogramy** jsou vhodné pro případy, kdy je počet proměnných m objektu $\boldsymbol{x}_i^{\mathbf{T}}$ veliký. Jednotlivé složky $\boldsymbol{x}_j$ představují délku větví schematického stromu. Jeho struktura větví vzniká na základě předběžného hierarchického shlukování proměnných (viz *shluková analýza*). Předběžná shluková analýza se dá použít také při výběru pořadí složek objektu $\boldsymbol{x}$ při konstrukci ostatních symbolových grafů.

4.5 Určení struktury a vazeb v proměnných a objektech

Zdrojová matice má rozměr $n \times m$. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje

(a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafů,
(b) nalézt *vybočující objekty*, resp. jejich proměnné,
(c) stanovit, zda lze použít předpoklad *lineárních vazeb,*
(d) ověřit *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda se hledají
(a) *struktura a vazby v proměnných* nebo
(b) *struktura a vazby v objektech*:

(1) Hledání struktury *v proměnných* v metrické škále: *faktorová analýza FA*, *analýza hlavních komponent PCA* a *shluková analýza*.
(2) Hledání struktury *v objektech* v metrické škále: *shluková analýza*.
(3) Hledání struktury *v objektech* v metrické i v nemetrické škále: *vícerozměrné škálování.*
(4) Hledání struktury *v objektech* v nemetrické škále: *korespondenční analýza.*
(5) Většina metod vícerozměrné statistické analýzy umožňuje *zpracování lineárních vícerozměrných modelů*, kde závisle proměnné se uvažují jako lineární kombinace nezávisle proměnných, resp. vazby mezi proměnnými jsou lineární. V řadě případů se také uvažuje normalita metrických proměnných.

Určením struktury a vzájemných vazeb mezi proměnnými ale i mezi objekty se zabývají techniky redukce proměnných na latentní proměnné, metoda *analýzy hlavních komponent (PCA)* a *metoda faktorové analýzy (FA).* Důležitou metodou určení vzájemných vazeb mezi proměnnými je i *kanonická korelační analýza CA*, která se používá ke zkoumání závislosti mezi dvěma skupinami proměnných, přičemž jedna ze skupin se považuje za proměnné nezávislé a druhá za skupinu proměnných závislých.

4.5.1 Analýza hlavních komponent (PCA)

Cílem metody je transformace dat z původních proměnných x_i, j=1, ..., m, do menšího počtu latentních proměnných y_j. Tyto proměnné mají vhodnější vlastnosti, je jich výrazně méně, vystihují téměř celou *proměnlivost* původních proměnných a jsou vzájemně nekorelované (korelační koeficient mezi latentními proměnnými y_1 , ..., y_m je 0). Latentní proměnné jsou u této metody nazvány *hlavními komponentami* a jsou to lineární kombinace původních proměnných: *první hlavní komponenta* y_1 popisuje největší část proměnlivosti čili rozptylu původních dat, *druhá hlavní komponenta* y_2 zase největší část rozptylu neobsaženého

v y_1 atd. Matematicky řečeno, *první hlavní komponenta* je takovou lineární kombinací vstupních proměnných, která zahrnuje největší proměnlivost mezi všemi lineárními kombinacemi. Má tvar

$$y_1 = \sum_{j=1}^{m} v_{1j} x_j = \boldsymbol{v_1}^{\mathrm{T}} \boldsymbol{x},$$

kde objekt $\boldsymbol{x}$ obsahuje proměnné $x_1, ..., x_m$. Pro vektor koeficientů $\boldsymbol{v_1}^{\mathrm{T}} = (v_{11}, ..., v_{1m})^{\mathrm{T}}$ platí, že proměnlivost vyjádřená rozptylem $D(y_1) = \boldsymbol{v_1}^{\mathrm{T}} \boldsymbol{S} \boldsymbol{v_1}$ je maximální, přičemž $\boldsymbol{S}$ značí kovarianční matici původních dat. Zcela analogicky jsou konstruovány další hlavní komponenty, jejichž celkový počet je roven menšímu ze dvou čísel, a to *n* (počet objektů) nebo *m* (počet proměnných). Protože platí, že součet rozptylů všech hlavních komponent je roven součtu rozptylů vstupujících původních proměnných, můžeme z podílu rozptylů jednotlivých hlavních komponent usuzovat na část proměnlivosti vysvětlenou dotyčnou hlavní komponentou. Jestliže součet prvních (nejvyšších) *A* podílů proměnlivosti je dostatečně blízký jedné, resp. 100 % (obvykle však stačí 80 % - 90 %), postačí brát v úvahu právě těchto prvních *A* hlavních komponent pro "dostatečné" vysvětlení variability původních proměnných. Rozdíl mezi souřadnicemi objektů v původních proměnných a v hlavních komponentách čili ztráta informace projekcí do menšího počtu rozměrů se nazývá *špatnou mírou těsností proložení* modelu PCA nebo také *chybou modelu PCA*. I při velkém počtu původních proměnných (*m*) může být *A* velmi malé, často 2 až 5. Volba počtu užitých komponent *A* představuje vlastní *model hlavních komponent PCA*. Vysvětlení užitých hlavních komponent, jejich pojmenování a vysvětlení vztahu původních proměnných x_j, $j = 1, ..., m$, k hlavním komponentám y_k, $k = 1, ..., A$, tvoří dominantní součásti zvoleného modelu hlavních komponent PCA.

Vlastní *matematický postup PCA* je následující: maximalizací při zavedení normalizační podmínky $\boldsymbol{v_1}^{\mathrm{T}} \boldsymbol{v} = 1$ vyjde, že

$$(\boldsymbol{S} - \lambda_1 \boldsymbol{I}) \boldsymbol{v_1} = \boldsymbol{0},$$

kde $\boldsymbol{0}$ označuje nulový vektor, λ_1 je největší *vlastní číslo* a $\boldsymbol{v}_1$ je odpovídající *vlastní vektor* kovarianční matice $\boldsymbol{S}$ *a* $\boldsymbol{I}$ je jednotková matice. Po dosazení vyjde

$$D(y_1) = \boldsymbol{v_1}^{\mathrm{T}} \boldsymbol{S} \boldsymbol{v_1} \lambda_1.$$

Analogicky lze odvodit, že vektor koeficientů $\boldsymbol{v}_2$ ve vztahu $y_2 = \sum_{j=1}^{m} v_{2j} x_j$, maximalizující $D(y_2)$ za podmínky, že $\mathrm{cov}(y_1, y_2) = 0$, odpovídá vlastnímu vektoru, příslušejícímu druhému největšímu vlastnímu číslu λ_2.

Provedeme-li rozklad kovarianční matice $\boldsymbol{S}$ na vlastní čísla $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$, jsou odpovídající vlastní vektory $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_m$ přímo koeficienty hlavních komponent $y_1, ..., y_m$. Hlavní komponenty mají řadu zajímavých vlastností. Lze je interpretovat jako hlavní osy *m*-rozměrného elipsoidu $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{S}^{-1} \boldsymbol{x} = \mathrm{konst}$.

K odstranění závislosti na jednotkách původních proměnných se lépe užívá standardizovaných proměnných $\boldsymbol{x}^*$ s prvky $x_j^* = (x_j - \bar{x}_j)/\sigma_j$. Pro *j*-tou hlavní komponentu pak platí

$$y_j^* = \sum_{k=1}^{m} v_{jk}^* x_k^*,$$

kde $\boldsymbol{v}_j^{(}$ je vlastní vektor *korelační* matice $\boldsymbol{R}$ odpovídající *j*-tému největšímu vlastnímu číslu $\lambda_j^{(}$. Hlavní komponenty $y_j^{(}$, určené z korelační matice $\boldsymbol{R}$, jsou však hůře interpretovatelné. Platí, že $\boldsymbol{v}_j^{(\mathrm{T}}\,\boldsymbol{v}_j^{(} = \lambda_j$, nikoliv rovno jedné. Pro účely zobrazení vícerozměrných dat různého měřítka jsou však vhodnější standardizované $y_j^{(}$ než původní y_j.

PCA umožňuje rozklad matice dat $\boldsymbol{X}$ na *strukturní část* $\boldsymbol{TP}^{\mathrm{T}}$ a *šumovou část* $\boldsymbol{E}$ dle vztahu $\boldsymbol{X} = \boldsymbol{TP}^{\mathrm{T}} + \boldsymbol{E}$, kde $\boldsymbol{T}$ je *matice komponentního skóre* a $\boldsymbol{P}$ je *matice komponentních vah,* $\boldsymbol{E}$ je *matice reziduí*. Modelem hlavních komponent PCA nazýváme součin $\boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}}$. Matice reziduí $\boldsymbol{E}$ není součástí modelu, týká se modelem PCA nevysvětlované části dat $\boldsymbol{X}$. Je to část dat, která není zahrnuta v modelu $\boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}}$ a představuje *míru netěsnosti proložení* původních reálných dat modelem PCA. Model hlavních komponent PCA

$$\boldsymbol{X} = t_1\,p_1^{\mathrm{T}} + t_2\,p_2^{\mathrm{T}} + \ldots + t_A\,p_A^{\mathrm{T}} + \boldsymbol{E}$$

se vypočte postupem:

1. Vypočte se t_1 a p_1 z $\boldsymbol{X}$ a vyčíslí se $\boldsymbol{E_1} = \boldsymbol{X} - t_1 p_1^{\mathrm{T}}$.
2. Vypočte se t_2 a p_2 z $\boldsymbol{E_1}$ a vyčíslí se $\boldsymbol{E_2} = \boldsymbol{E_1} - t_2 p_2^{\mathrm{T}}$.
3. Vypočte se t_3 a p_3 z $\boldsymbol{E_2}$ a vyčíslí se $\boldsymbol{E_3} = \boldsymbol{E_2} - t_3 p_3^{\mathrm{T}}$ a pokračuje se tak dlouho, až se vyčíslí všech A komponent, $A = \min(n, m)$. Výhodou postupu je snížení počtu proměnných, snížení dimenzionality a dále rozdělení původní matice dat $\boldsymbol{X}$ na část strukturní $\boldsymbol{TP}^{\mathrm{T}}$ a část šumovou $\boldsymbol{E}$. Když bychom použili celý PCA model, pak je $\boldsymbol{E}$ rovno nule. Musíme ale hledat optimální počet využitelných hlavních komponent A tak, abychom dosáhli nejlepšího proložení a aby matice $\boldsymbol{E}$ byla téměř nulová a její absolutní velikost byla srovnatelná s experimentální chybou dat. To je ústřední myšlenka vícerozměrné analýzy dat: uživatel musí navrhnout počet hlavních komponent A tak, aby byla matice reziduí co nejmenší. Velká hodnota $\boldsymbol{E}$ znamená špatný model, malá hodnota $\boldsymbol{E}$ dobrý model. Termíny malý, velký, dobrý, špatný jsou však pouze kvalitativní. Vyhodnocení $\boldsymbol{E}$ je relativní k centrovanému počátku, tj. střednímu objektu, který má souřadnice v aritmetickém průměru každé proměnné a představuje počátek (0, 0, ..., 0). Je vhodné říkat, že jde o nulu hlavních komponent. První operace je totožná s centrováním proměnných původní matice dat $\boldsymbol{X}$. To znamená, že pro $A = 0$ bude reziduálová matice $\boldsymbol{E_0}$ totožná s centrovanou maticí $\boldsymbol{X}$. Zde $\boldsymbol{E_0}$ hraje důležitou roli jako referenční stav, ke kterému budeme přirovnávat velikost klesající hodnoty $\boldsymbol{E}$, takže pro $A = 0$ bude $\boldsymbol{E_0}$ = 100 % a člen $\boldsymbol{TP}^{\boldsymbol{T}} = 0$ %. Velikost $\boldsymbol{E}$ představuje chybový výraz, který vyčíslíme běžným statistickým způsobem. Buď vyjádříme rezidua jako rezidua objektů (v řádcích) nebo rezidua proměnných (ve sloupcích).

Rezidua objektů (v řádcích): rozptyl reziduí *i*-tého objektu se týká průměru středově centrovaných dat a je dán vztahem

$$e_i = \sqrt{\sum_{k=1}^{m} e_{ik}^2}$$

a odpovídá *vzdálenosti* mezi *i*-tým objektem a hyperplochou A hlavních komponent. Je to vzdálenost, která byla minimalizována nejmenšími čtverci, když se určovaly hlavní komponenty. Proto je rozptyl reziduí objektu mírou vzdálenosti mezi prostorem proměnných objektu a reprezentací objektu v prostoru hlavních kompo-nent. Platí přitom pravidlo: čím menší bude tato vzdálenost, tím těsnější bude reprezentace objektu v prostoru

hlavních komponent (čili PCA model) vůči původnímu objektu. Řádky v matici ***E*** jsou nepřímo úměrné těsnosti proložení původních dat PCA modelem.

Celkový rozptyl reziduí objektu je suma rozptylu reziduí všech objektů e_{tot} dle vztahu

$$e_{tot} = \sqrt{\sum_{i=1}^{n} e_i^2} \quad .$$

Graf reziduí jednotlivých objektů. Rozptyl reziduí jednoho *i*-tého objektů představuje vzdálenost mezi objektem a modelem. Je vhodné zobrazovat tuto veličinu pro všechny objekty a odhalit tak odlehlé objekty či jiné anomálie.

Graf celkového rozptylu reziduí. Metodou hlavních komponent PCA se vyčíslí $\boldsymbol{E}_0$, $\boldsymbol{E}_1$, $\boldsymbol{E}_2$, ... a z nich $e_{tot,1}$, $e_{tot,2}$, ... a ty se vynesou do indexového grafu úpatí proti indexu A. Protože je snaha o pojmenování a vysvětlení hlavních komponent, je správný počet hlavních komponent A velmi důležitý: je-li počet A příliš malý, "podceněný model" způsobí povrchní popis datové struktury. Je-li počet A příliš velký, "přeceněný model" je ještě horší, protože zahrnuje do své struktury také část šumu. Existuje pravidlo, že velká proměnlivost v datech odpovídá hlavnímu jevu, například proměňované koncentraci, zatímco malá proměnlivost odpovídá spíše šumu. V analýze dat se snažíme odhalit stěžejní jev, o kterém předpokládáme, že dominuje v datech. V PCA platí pravidlo, že "velké" hlavní komponenty odpovídají nejdůležitější informaci, kterou vlastně hledáme, zatímco "malé" hlavní komponenty odpovídají spíše šumu a jsou pro strukturu dat obvykle nepodstatné. Malé hlavní komponenty mohou být proto zahrnuty do matice ***E***. Cílem PCA je odfiltrování šumu z dat a soustředění se na strukturní, bezšumovou část dat.

Graficky lze výsledek analýzy hlavních komponent zobrazit v několika grafech hlavních komponent následujícím způsobem:

(a) **Indexový graf úpatí vlastních čísel** (Scree Plot) je vlastně sloupcový diagram vlastních čísel nebo reziduálního rozptylu proti stoupající hodnotě indexu, pořadového čísla A (obr. 4.7). Zobrazuje relativní velikost jednotlivých vlastních čísel. Řada autorů ho s oblibou využívá k určení počtu A "užitečných" hlavních komponent. Cattel vysvětluje scree jako zlomové místo mezi kolmou stěnou a vodorovným dnem. Vybrané "užitečné" hlavní komponenty (nebo také faktory) pak tvoří kolmou stěnu a "neužitečné" hlavní komponenty (nebo faktory) představují vodorovné dno. Užitečné komponenty jsou tak odděleny zřetelným zlomovým místem a souřadnice x tohoto zlomu je hledaná hodnota indexu. Jiným, hrubším kritériem je pravidlo, podle kterého využíváme ty hlavní komponenty, jejichž vlastní číslo je větší než jedna. Graf úpatí se však jeví objektivnějším.

(b) **Graf komponentních vah, zátěží** (Plot Components Weights) zobrazí komponentní váhy pro první dvě hlavní komponenty (obr. 4.8). V tomto grafu se porovnávají vzdálenosti mezi proměnnými. Krátká vzdálenost mezi dvěma proměnnými znamená silnou korelaci. Lze nalézt i shluk podobných proměnných, jež spolu korelují. Tento graf můžeme považovat za most mezi původními proměnnými a hlavními komponentami, protože ukazuje, jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent. Někdy se podaří hlavní komponenty y_1, y_2, ... pojmenovat, vysvětlit a přidělit jim fyzikální, chemický nebo biologický význam. Pak lze názorně vysvětlit, jak

jednotlivé původní proměnné x_j, j = 1,..., m, přispívají do první hlavní komponenty y_1 nebo do druhé hlavní komponenty y_2. Některé původní proměnné x_j přispívají kladnou vahou, některé zápornou. Bývá zajímavé sledovat kovarianci původních proměnných x_j v prostorovém 3D grafu komponentních vah y_1, y_2 a y_3. Jsou-li proměnné x_j, j = 1,..., m, blízko sebe v prostorovém shluku, jde o silnou pozitivní kovarianci. Kovariance však nemusí ještě nutně znamenat korelaci. Výklad grafu komponentních vah lze obecně shrnout do následujících bodů:

1. *Důležitost původních proměnných* x_j, j = 1,..., m: proměnné x_j s vysokou mírou proměnlivosti v datech objektů mají vysoké hodnoty komponentní váhy. Ve 2Ddiagramu prvních dvou hlavních komponent pak leží hodně daleko od počátku. Proměnné s malou důležitostí leží blízko počátku. Když určíme *důležitost proměnných*, určíme tím také proměnlivost proměnných: jestliže například y_1 objasňuje 70 % proměnlivosti a y_2 jenom 5 % (přečteno z indexového grafu úpatí vlastních čísel), jsou původní proměnné x_j, j= 1,..., m, s vysokou vahou v y_1 tím pádem mnohem důležitější než proměnné x_j s vysokou vahou v y_2. Proměnné s úhlem 0E mezi průvodiči jsou zcela pozitivně korelované, proměnné s úhlem 90E jsou zcela nekorelované zatímco proměnné s úhlem 180E jsou sice nekorelované, říkáme však lépe že jsou negativně korelované.
2. *Korelace a kovariance:* původní proměnné x_j, j= 1,..., m, jsou blízko sebe, anebo proměnné x_j s malým úhlem mezi svými průvodiči proměnných a na stejné straně vůči počátku mají vysokou kladnou kovarianci a vysokou kladnou korelaci. Naopak, původní proměnné x_j daleko od sebe, anebo s velikým úhlem mezi průvodiči proměnných, jsou negativně korelovány.
3. *Spektroskopická data:* ve spektroskopických datech je 1-rozměrný graf komponentních vah často nejvhodnější. I zde platí pravidlo, že vysoké komponentní váhy představují vysokou důležitost proměnných x_j (vlnových délek).

(c) **Rozptylový diagram komponentního skóre** (Scatterplot) zobrazuje *komponentní skóre* čili hodnoty obyčejně prvních dvou hlavních komponent u všech objektů (obr. 4.9). Dokonalé rozptýlení objektů v rovině obou hlavních komponent vede k rozlišení objektů při jejich popisu pomocí y_1 a y_2. V rovině lze snadno nalézt shluk vzájemně podobných objektů a dále objekty odlehlé a silně odlišné od ostatních. Diagram komponentního skóre však může být i ve 3 či více hlavních komponentách a v rovin-ném grafu se pak sleduje pouze jeho průmět do roviny. Tento diagram se užívá k identifikaci odlehlých objektů, identifikaci trendů, tříd, shluků objektů, k objasnění podobnosti objektů atd. Je nemožné analyzovat všechny diagramy, protože jich je velmi mnoho: uvažujme například $m < n$ a pro m = 10 proměnných existuje
$m(m\text{-}1)/2$ = 45 diagramů, pro m = 11 pak 55 diagramů, pro m = 12 pak 66 diagramů, atd. Obvykle vybíráme diagramy y_1 vs. y_2, y_1 vs. y_3, y_1 vs. y_4 atd. Držíme se první hlavní komponenty y_1, protože v ní bývá největší míra proměnlivosti v datech. Interpretace rozptylového diagramu komponentního skore lze shrnout do těchto bodů:

1. *Umístění objektů*. Objekty daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.
2. *Podobnost objektů*. Objekty blízko sebe si jsou podobné, objekty daleko od sebe jsou si nepodobné.

3. *Objekty v shluku*. Objekty umístěné zřetelně v jednom shluku jsou si podobné a přitom nepodobné objektům v ostatních shlucích. Dobře oddělené shluky prozrazují, že lze nalézt vlastní model pro samotný shluk. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.
4. *Osamělé objekty*. Izolované objekty mohou být odlehlé objekty, které jsou silně nepodobné ostatním objektům. Pravidlo platí, pokud se nejedná o zdánlivou nehomogenitu danou sešikmením dat a odstranitelnou transformací proměnných.
5. *Odlehlé objekty.* V ideálním případě bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je něco špatného v modelu, obyčejně je přítomen silně odlehlý objekt. Odlehlé objekty jsou totiž schopny zbortit celý diagram, ve srovnání se silně vybočujícím objektem jsou ostatní objekty nakumulovány do jediného úzkého shluku. Po odstranění vybočujícího objektu se ostatní objekty roztřídí po celé ploše diagramu a teprve vypovídají o existujících shlucích.
6. *Pojmenování objektů.* Výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a mezi pojmenovanými hlavními komponentami. Snadno obkroužíme shluky podobných objektů nebo nakreslením spojky mezi objekty vystihneme tak jejich fyzikální či biologický vztah.
7. *Vysvětlení místa objektu.* Umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami původních proměnných ve dvojném grafu a pomocí původních proměnných pak i vysvětleno.

(d) **Dvojný graf** (Biplot) kombinuje předchozí dva grafy (obr. 4.10). Úhel mezi průvodiči dvou proměnných x_j a x_k je nepřímo úměrný velikosti korelace mezi těmito dvěma proměnnými. Čím je menší úhel, tím je větší korelace. Každý průvodič má své souřadnice na první a na druhé hlavní komponentě. Délka této souřadnice je úměrná příspěvku původní proměnné x_j do hlavní komponety, čili je úměrná komponentní váze. Kombinace obou grafů v jediném přináší cenné srovnání, jeden graf působí zde doplňkově vůči druhému. Když se ve dvojném grafu nachází objekt v blízkosti určité proměnné x_j, znamená to, že tento objekt "obsahuje" hodně právě této proměnné a je s ní v interakci. Interakce proměnných a objektů umožňuje také vysvětlit umístění objektů vpravo od nuly na ose y_1 (či vlevo od nuly) pomocí pozice proměnných v tomto grafu, resp. umístění nahoře od nuly (či dole od nuly) na ose y_2.

(e) **Indexový graf úpatí rozptylu reziduí** (Residual Scree Plot). V PCA se vyčíslí $\boldsymbol{E_0}$, $\boldsymbol{E_1}$, $\boldsymbol{E_2}$, ..., z nich $e_{\text{tot},1}$, $e_{\text{tot},2}$, ... a ty se vynesou do indexového grafu úpatí proti indexu A. Poslední bod bude pro menší ze dvou čísel n nebo m, $A = \min(n, m)$. Zlom na křivce úpatí $e_{\text{tot},i} = f(i)$ ukazuje na optimální počet hlavních komponent A, na nejlepší dimenzionalitu. V tomto bodě končí struktura a začíná šum. Graf je co do použití naprostou obdobou indexového grafu úpatí vlastních čísel.

(f) **Graf reziduí jednotlivých objektů.** Rozptyl reziduí jednoho i-tého objektu představuje vzdálenost mezi objektem a modelem. Je vhodné zobrazovat tuto veličinu pro všechny objekty a odhalit odlehlé objekty či jiné anomálie.

Diagnostikování častých problémů v PCA: v analýze metodou hlavních komponent se často setkáváme s následujícími problémy:

1. *Data neobsahují předpokládanou informaci.* Vysvětlení grafů a diagramů metody PCA nemá smysl, protože data neobsahují informaci, popisující studovaný problém.
2. *Užito příliš málo hlavních komponent.* V modelu PCA bylo použito příliš málo latentních proměnných. Nedostatečné vysvětlení dat vede ke ztrátě informace. Problém se může vyřešit opětovným rozborem grafu úpatí vlastních čísel.
3. *Užito příliš mnoho hlavních komponent* V modelu PCA bylo zahrnuto příliš mnoho latentních proměnných, což může vyvolat vážnou chybu, protože šum je zahrnut do modelu.
4. *Neodstranění odlehlých objektů.* Odlehlé objekty mohou být důvodem hrubých chyb v datech. Do modelu jsou vtahovány spíše hrubé chyby než zajímavé proměnlivosti v datech objektů.
5. *Odstraněné odlehlé objekty obsahovaly důležitou informaci.* Ztrátou určitých objektů se vytratila důležitá informace z dat a nalezený model je proto zkreslený.
6. *Komponentní skóre je nedostatečně analyzováno.* Nedostatečným rozborem důležitého rozptylového diagramu byly zanedbány důležité rysy v datech.
7. *Vysvětlení komponentních vah se špatným počtem hlavních komponent.* Může vést k vážnému zkreslení výkladu. Může totiž dojít k vyjmutí důležitých proměnných, protože se zdají býti odlehlými. Tento graf je mostem mezi prostorem původních proměnných a prostorem hlavních komponent PC. Když zvolíme špatný prostor PC, tento most už nám mnoho nepomůže.
8. *Přecenění standardních diagnostik v software.* Je třeba hodně rozvažovat a přemýšlet o úloze samé a specifickém problému řešeném před pohodlným přebíráním počítačových výsledků.
9. *Užití špatného předzpracování dat.* Chybná předúprava dat (ve škálování užitého centrování nebo standardizace, transformace logaritmická, mocninná, Boxova-Coxova atd.) může vést ke zkresleným závěrům a neporozumění úloze. Způsob předúpravy dat je obecně dán typem úlohy a druhem instrumen-tálních dat a může vést ke ztrátě informace.

Postup metody hlavních komponent (PCA)

Problém musí být správně a přesně definován. Je odpovědností řešitele, aby data obsahovala dostatek relevantní informace k řešení problému. Ani nejlepší počítačová metoda nemůže kompenzovat nedostatek informace v datech. Maticový graf korelace proměnných slouží k získání počáteční informace o celém datovém souboru. Odhalí, zda data potřebují škálování. Při prvním seznámení s daty se v rámci exploratorní analýzy, kam také patří metoda hlavních komponent PCA, aplikuje první výpočet touto metodou. Data je obvykle potřeba škálovat nebo alespoň centrovat. Lze vyzkoušet i ostatní formy předúpravy dat. V tomto stadiu se vždy vyčíslují všechny hlavní komponenty. První diagramy komponentního skóre slouží k odhalení odlehlých hodnot, tříd, shluků a trendů. Jsou-li objekty roztříděny do dobře oddělených shluků, je třeba určit způsob, jak je z dat oddělit a shluky pak analyzovat odděleně. Nikdy se nepokoušíme odhalovat a odstraňovat odlehlé proměnné, mohlo by pak dojít k odstranění cenné informace. Po redukci datového souboru na několik podsouborů, kdy shluky jsou modelovány odděleně, se znovu aplikuje metoda hlavních komponent PCA na jednotlivé podsoubory.

1. *Vyšetření indexového grafu úpatí vlastních čísel:* z hrany v tomto diagramu se určí nejlepší počet hlavních komponent.

2. *Výpočet vlastních vektorů pro hlavní komponenty:* vedle číselných hodnot se užívá i názorný čárový diagram hodnot vlastních čísel vektorů, který přehledně informuje o zastoupení původních proměnných x_j , j = 1, ..., m, v hlavních kompo-nentách.

3. *Výpočet komponentních vah:* matice párových korelačních koeficientů ukazuje na korelaci původních proměnných s hlavními komponentami. Čárový diagram názorně vysvětluje korelační strukturu mezi oběma druhy proměnných. Uživatel nyní vybere pouze prvních A hlavních komponent a vytvoří tak *PCA model.*

4. *Vyšetření grafu komponentních vah.*

5. *Vyšetření rozptylového diagramu komponentního skóre.*

6. *Vyšetření dvojného grafu.*

7. *Vyšetření grafu úpatí reziduí objektů:* rezidua objektů a rezidua proměnných by měla prokazovat dostatečnou těsnost proložení. Není-li tomu tak, je třeba se navrátit k předúpravě dat a celý výpočet PCA opakovat.

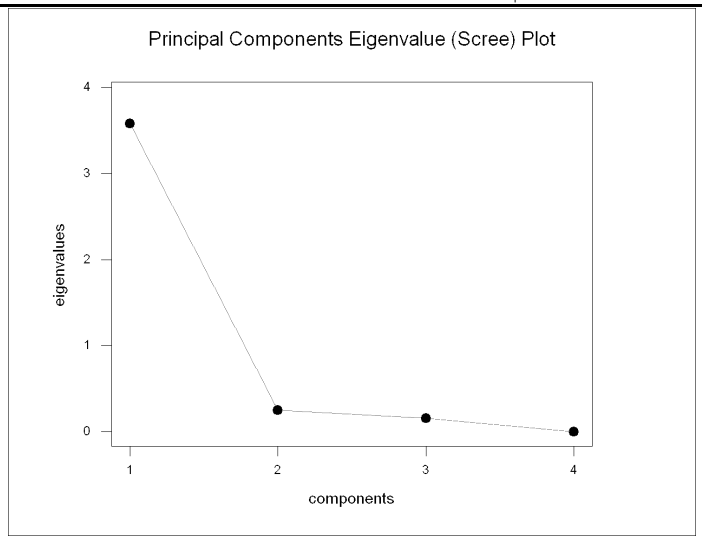
Vzorová úloha 4.2 *Postup metody hlavních komponent*

Na úloze **B4.02** *Účinky neuroleptik při tlumení rozličných psychóz* si ukážeme pomůcky vícerozměrné analýzy dat. K analýze užijeme také škálovaná data.

Řešení: k analýze byl použit program NCSS2000. Výstup metody PCA programu NCSS2000 pro nestandardizovaná data úlohy B402 obsahuje:

1. Vyšetření indexového grafu úpatí vlastních čísel:

Index	Vlastní číslo λ_i	Individualní procento	Kumulativní procento	Kumulativní čárový graf úpatí
1	3394.339	92.62	92.62	|||||||||||||||
2	252.286	6.88	99.50	||
3	15.8825	0.43	99.94		
4	2.295	0.06	100.00		



Obr. 4.7 Indexový graf úpatí vlastních čísel (Scree Plot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* ze standardizovaných dat B402, *SCAN.*

Vlastní čísla slouží k určení počtu A "využitelných" hlavních komponent, jež si zvolíme v analýze k dalšímu užívání. Procento a kumulativní procento popisuje proměnlivost v původních proměnných, popsanou dotyčnou hlavní komponentou. K dalšímu popisu proměnlivosti bereme obvykle tolik hlavních komponent, aby jimi bylo popsáno 90 až 99 % celkové proměnlivosti. V tomto případě stačí užít první dvě. **Indexový graf úpatí vlastních čísel** je vlastně sloupcový diagram velikosti vlastních čísel proti stoupající hodnotě indexu, pořadového čísla. Zobrazuje relativní velikost jednotlivých vlastních čísel. Užitečné komponenty jsou tak odděleny zřetelným zlomovým místem a souřadnice *x* tohoto zlomu je hledanou hodnotou indexu.

2. (a) Vlastní vektory pro hlavní komponenty:

Proměnná	y_1	y_2	y_3	y_4
B402X1	-0.5684	0.4339	0.532	-0.4534
B402X2	-0.5688	0.4042	-0.584	0.4148
B402X3	-0.0758	-0.0286	0.6125	0.7863
B402X4	-0.5896	-0.8047	-0.0282	-0.0642

Vlastní vektory jsou váhy, jež umožňují kombinovat komponentní proměnné, které byly předem normovány vzorcem $(x_i - \bar{x})/\sigma_i$. Např. první hlavní komponenta y_1 je *vážený průměr normovaných proměnných,* kdy váha každé proměnné je dána odpovídajícím prvkem prvního vlastního vektoru

$$y_1 = v_{11}x_1 + v_{12}x_2 + \ldots + v_{1m}x_m.$$

Koeficienty v této rovnici vystihují relativní důležitost každé proměnné při tvorbě hlavní komponenty. Vlastní vektory bývají často normovány, takže rozptyl komponentního skóre je roven jedné.

(b) Čárový diagram absolutních hodnot vlastních vektorů:

Proměnná	y_1	y_2	y_3	y_4
B402X1	|||||||||||	||||||||	|||||||||	||||||||
B402X2	|||||||||||	||||||||	|||||||||	||||||||
B402X3	||	|	||||||||||	||||||||||||||
B402X4	|||||||||||	||||||||||||||	|	||

Diagram absolutních hodnot vlastních vektorů vysvětluje strukturu a ukazuje, která proměnná silně koreluje shlavní komponentou.

3. (a) Komponentní váhy:

Proměnná	y_1	y_2	y_3	y_4
B402X1	-0.9769	0.2033	0.0625	-0.0203
B402X2	-0.9793	0.1897	-0.0688	0.01857
B402X3	-0.8486	-0.0874	0.4688	0.2288
B402X4	-0.9372	-0.3487	-0.0031	-0.0026

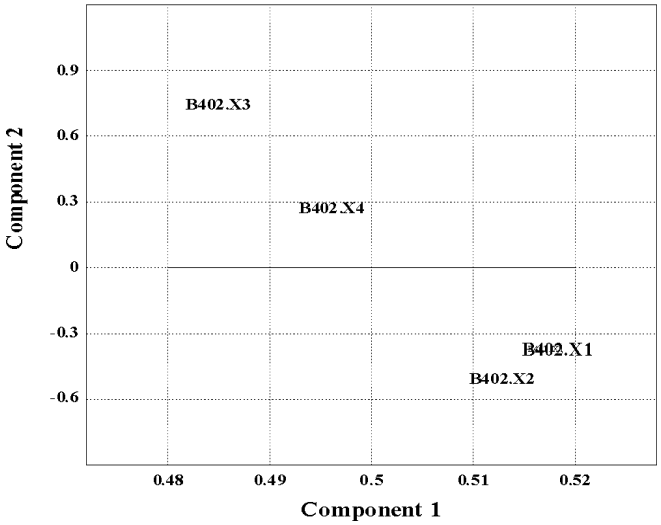
Ukazuje korelace mezi původními proměnnými a hlavními komponentami.

(b) Čárový diagram absolutních hodnot komponentních vah:

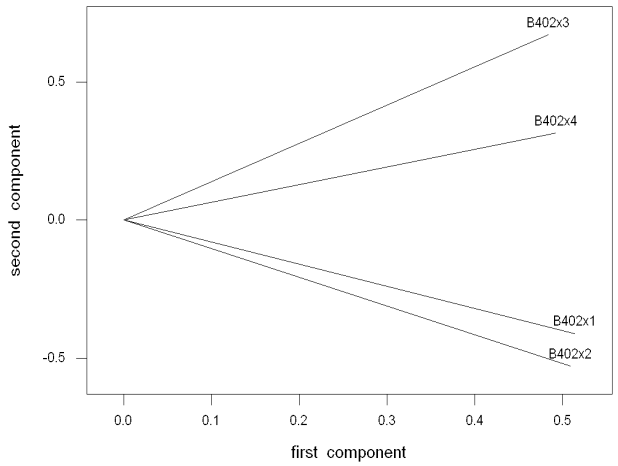
Proměnná	y_1	y_2	y_3	y_4
B402X1	||||||||||||||||||	|||||	||	|
B402X2	||||||||||||||||||	||||	||	|
B402X3	|||||||||||||||	||	||||||||	|||||
B402X4	||||||||||||||||||	|||||||	|	|

Diagram zobrazuje absolutní hodnoty komponentních vah a vysvětluje korelační strukturu, tj. která původní proměnná silně koreluje s hlavní komponentou. Je zřejmé, že 1. hlavní komponenta silně koreluje se všemi proměnnými, zatímco 3. nebo 4. hlavní komponenta koreluje pouze s *B402X3.*

4. Vyšetření grafu komponentních vah: grafická analýza předešlého kroku.

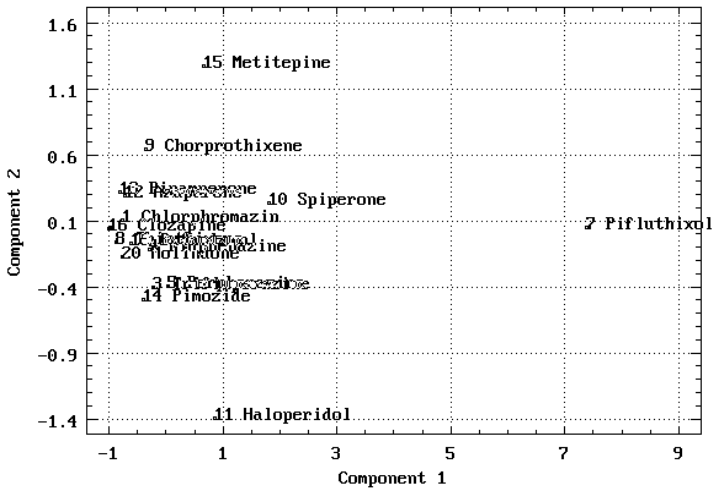


Obr. 4.8a Graf komponentních vah (Components Weights Plot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* ze standardizovaných dat B402, *STATGRAPHICS. Proměnné B402X1 a B402X2 leží v diagramu blízko sebe a proto spolu korelují. Proměnné B402X3 a B402X4 jsou dál od sebe, proto méně korelují. Méně korelují rovněž se dvěma proměnnými B402X1 a B402X2, nacházejí se daleko od nich.*

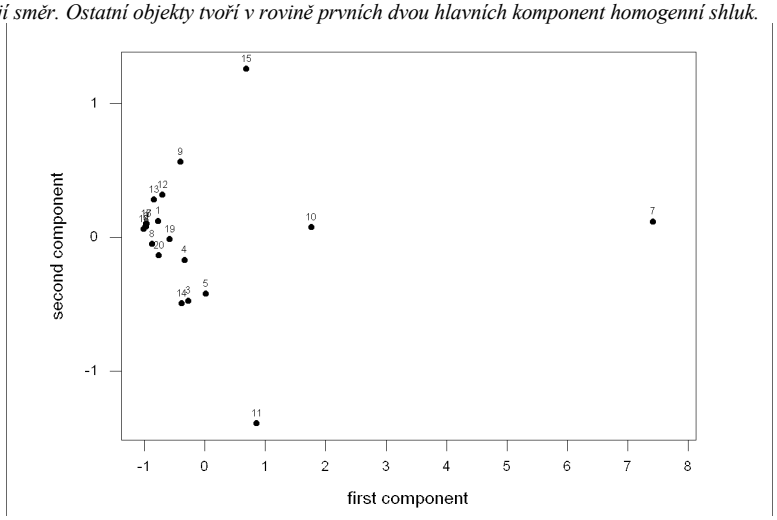


Obr. 4.8b Graf komponentních vah (Components Weights Plot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* ze standardizovaných dat B402. *Výklad je stejný jako u obr. 4.8a SCAN.*

5. Vyšetření rozptylového diagramu komponentního skóre: nejdůležitější diagram metody hlavních komponent ukazuje celou vyšetřovanou strukturu objektů, tzn. shluky objektů, izolované objekty, odlehlé objekty, anomálie atd. Objekty mohou být označeny textovým popisem nebo číselně, indexem.

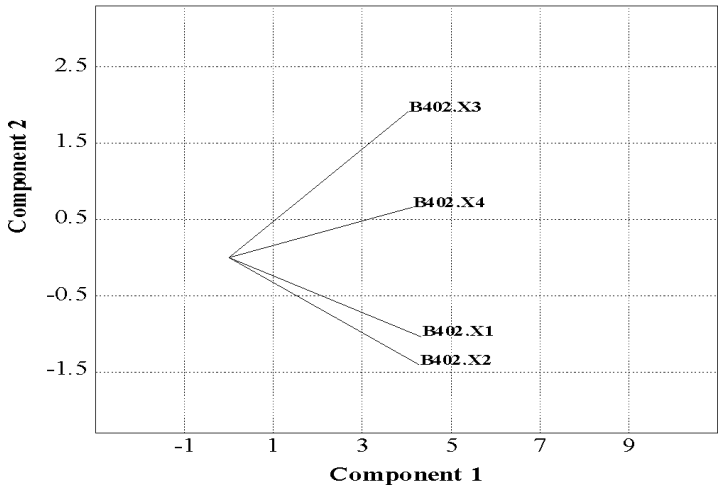


Obr. 4.9a Rozptylový diagram komponentního skóre (Scatterplot) pro 20 objektů a 4 proměnné B402X1, B402X2, B402X3, B402X4 standardizovaných dat B402, ***STATGRAPHICS***. *Kromě tří objektů: 7, 11 a 15 leží zbývajících 17 objektů v jediném shluku. Objekty 7, 11 a 15 jsou odlehlé body. Nejvíce odlišný objekt od ostatních je 7, protože ten je odlehlý na hlavní komponentě 1 popisující většinu rozptylu. První hlavní komponenta 1 popisuje rozdíl mezi Pifluthixolem a ostatními objekty. Na druhé straně objekty 11 a 15 jsou extrémy na druhé hlavní komponentě a udávají její směr. Ostatní objekty tvoří v rovině prvních dvou hlavních komponent homogenní shluk.*

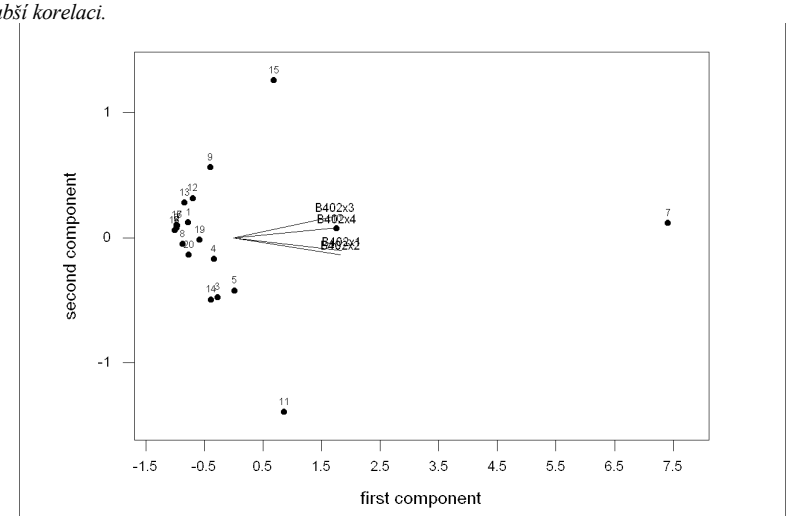


Obr. 4.9b Rozptylový diagram komponentního skóre (Scatterplot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* standardizovaných dat B402. *Výklad je stejný jako u obr. 4.9a*, ***SCAN***.

6. Vyšetření dvojného grafu: je důležité sledovat interakci objektů a proměnných. Je - li některý objekt umístěn ve dvojném grafu na stejném místě nebo alespoň poblíž místa proměnné, jsou spolu v interakci. Interakce poslouží interpretaci objektů.



Obr. 4.10a Dvojný graf (Biplot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* standardizovaných dat úlohy B402, ***STATGRAPHICS***. *Úhel mezi dvěma průvodiči dvou proměnných je nepřímo úměrný velikosti korelace mezi těmito proměnnými. Mezi průvodiči B402X1 a B402X2 je malý úhel, což svědčí o silné korelaci. Úhly mezi těmito dvěma průvodiči a průvodiči B402X3 a B402X4 jsou pak větší, což ukazuje na jejich slabší korelaci.*



Obr. 4.10b Dvojný graf (Biplot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* standardizovaných dat úlohy B402. *Výklad je stejný jako u obr. 4.10a,* ***SCAN***.

7. Vyšetření grafu úpatí reziduí objektů: těsnost proložení PCA modelu danými daty se posuzuje statistickou analýzou reziduí. Rezidua by měla vykazovat dostatečnou těsnost.

4.5.2 Faktorová analýza (FA)

Podobně jako metoda hlavních komponent patří také faktorová analýza mezi metody redukce počtu původních proměnných. Ve faktorové analýze předpokládáme, že každou vstupující proměnnou můžeme vyjádřit jako lineární kombinaci nevelkého počtu společných skrytých faktorů a jediného chybového faktoru. Na rozdíl od komponentní analýzy se při faktorové analýze snažíme vysvětlit závislost proměnných. K nevýhodám metody patří zejména nutnost zadat *počet společných faktorů* ještě před prováděním vlastní analýzy.

Předpokládáme-li, že $\boldsymbol{x}^{\mathbf{T}} = (x_1, x_2, ..., x_m)^{\mathrm{T}}$ je jeden objekt pozorovaných proměnných s korelační $\boldsymbol{R}$ nebo kovarianční maticí $\boldsymbol{C}$, potom můžeme všechny objekty $\boldsymbol{X}$, rozměru $n \times m$ zapsat jako *model faktorové analýzy (FA)* ve tvaru

$$x_1 = l_{11}f_1 + l_{12}f_2 + \ldots + l_{1m}f_m + \gamma_1 ,$$

$$x_2 = l_{21}f_1 + l_{22}f_2 + \ldots + l_{2m}f_m + \gamma_2 ,$$

$$\ldots\ldots\ldots$$

$$x_n = l_{n1}f_1 + l_{n2}f_2 + \ldots + l_{nm}f_m + \gamma_n ,$$

kde $f_1, f_2, ..., f_m$ jsou *faktory*, které vyvolávají korelace mezi proměnnými a $\gamma_1, \gamma_2, ..., \gamma_n$ jsou *chybové faktory*, které přispívají k rozptylu jednotlivých proměnných. Koeficienty l_{ik} nazýváme *faktorové zátěže i*-tého objektu u *k*-tého společného faktoru f_k a představují prvky matice faktorových zátěží. Model můžeme přepsat v maticové podobě jako

$$\boldsymbol{x} = \boldsymbol{L}\,\boldsymbol{f} + \boldsymbol{\varepsilon}.$$

Pro ortogonální faktorový model lze kovarianční matici vektoru vstupujících proměnných čili sloupců zdrojové matice napsat ve formě tzv. *základní faktorové věty* ve tvaru

$$\boldsymbol{C} = \boldsymbol{L}\,\boldsymbol{L}^{\mathbf{T}} + \boldsymbol{\Gamma}^2,$$

kde $\boldsymbol{L}$ je *matice faktorových zátěží*, dále $\boldsymbol{L}\,\boldsymbol{L}^{\mathrm{T}}$ představuje vlastně kovarianční matici vektoru $\boldsymbol{L}\,\boldsymbol{f}$ a konečně $\boldsymbol{\Gamma}^2$ je *matice jedinečností*.

Matice jedinečností $\boldsymbol{\Gamma}^2$ čili *kovarianční matice chybových faktorů* je maticí diagonální, protože předpokládáme nekorelované chyby. Uvědomíme-li si dále, že diagonální prvky matice $\boldsymbol{\Gamma}^2$ představují rozptyly jednotlivých sloupců zdrojové matice, lze psát

$$\boldsymbol{S}^2 = \boldsymbol{H}^2 + \boldsymbol{\Gamma}^2 ,$$

kde $\boldsymbol{S}^2$ je diagonální matice rozptylů faktorů. Proměnlivost každého faktoru, vyjádřenou sloupcem zdrojové matice, můžeme rozdělit na součet dvou složek: *komunalitu* $\boldsymbol{H}^2$, která představuje proměnlivost společnou všem faktorům a *jedinečnost* $\boldsymbol{\Gamma}^2$, která představuje část proměnlivosti čili rozptylu, nevysvětlenou faktory.

Komunalita $\boldsymbol{H}$ vyjadřuje míru proměnlivosti a je vahou, s jakou jednotlivé faktory přispívají do rozptylu odpovídající proměnné. Čtverec komunality je suma faktorových zátěží faktorů.

Jedinečnost $\boldsymbol{\Gamma}^2$ bývá dále rozdělována na část *specifity* $\boldsymbol{\Gamma}^2_s$ a část *nespolehlivosti* $\boldsymbol{\Gamma}^2_n$.

Specifita představuje tu část proměnlivosti, kterou nelze vysvětlit ani chybou experimentu, ani společnými faktory, zatímco *nespolehlivost* představuje experimentální chybu při měření faktorů.

Uvedený způsob rozkladu proměnlivosti představuje základní hledisko pro klasifikaci metod faktorové analýzy. Metoda hlavních komponent je zvláštním případem faktorové analýzy, kdy je model definován tvarem $\boldsymbol{S}^2 = \boldsymbol{H}^2$ a předpokládá se, že prostřednictvím hlavních komponent lze proměnlivost zdrojové matice beze zbytku reprodukovat. Jde tedy o vhodnou ortogonální transformaci, která zachovává všechnu původní proměnlivost, a to beze zbytku. Hovoříme pak z hlediska faktorové analýzy o *úplné komponentní analýze*. Jestliže při reprodukci pomocí hlavních komponent reprodukujeme pouze podstatnou část proměnlivosti (ale ne všechnu), jedná se ve faktorové analýze o *neúplnou komponentní analýzu*.

Pro odhad parametrů faktorového modelu se často užívá metody hlavních komponent, která je aplikována na redukovanou kovarianční matici. Předpokládá se, že jsou známy nějaké počáteční odhady chybových rozptylů, které jsou odečítány od diagonálních prvků výběrové korelační matice $\boldsymbol{R}$. Takto upravenou kovariační matici rozkládáme na součin matic $\boldsymbol{L}_1\ \boldsymbol{L}_1^{\mathrm{T}}$, kde $\boldsymbol{L}_1$ představuje výchozí matici odhadů faktorových zátěží. Postup pokračuje iterativně a po několika krocích konverguje ke konečné matici odhadů faktorových zátěží. Pokud neznáme výchozí odhady chybových rozptylů (resp. komunalit), je možné je určit speciálním postupem.

Porovnání FA a PCA:

1. Obě metody nemá cenu použít, když jsou původní proměnné x_j^{T}, $j = 1, \ldots, m$, nekorelované. FA pak nemá co objasnit a PCA povede k hlavním komponentám totožným s původními proměnnými.
2. FA postuluje model pro data, PCA nikoliv.
3. FA se pokouší objasnit kovariance a korelace původních proměnných pomocí několika málo společných faktorů. PCA objasňuje pouze rozptyl původních proměn-ných.
4. PCA: když zvýšíme počet použitých proměnných A o 1 na A+1, původní komponenty se nezmění. FA: když přidáme další faktor, ostatní faktory se podstatně změní.
5. PCA: výpočet je přímočarý, jednoduchý. FA: výpočet faktorového skóre je daleko komplexnější a byla pro něj navržena řada postupů.
6. Obyčejně není žádný vztah mezi hlavními komponentami PC a korelační maticí $\boldsymbol{R}$ anebo kovarianční maticí $\boldsymbol{C}$.

Vzorová úloha 4.3 *Vyčíslení faktorů z korelační matice*

Spearman (1904) analyzoval známky 200 žáků ze tří předmětů. Po vyčíslení korelační matice $\boldsymbol{R}$ uvažoval jeden faktor a následující faktorový model:

$$x_1 = \underline{\lambda}_1 f + u_1,$$
$$x_2 = \underline{\lambda}_2 f + u_2,$$
$$x_3 = \underline{\lambda}_3 f + u_3.$$

V tomto případě můžeme pojmenovat faktor f jako všeobecnou inteligenci žáka a specifické proměnné u_1, u_2, u_3 mají malé rozptyly, když jsou dotyčné proměnné x_i těsně spjaty s faktorem f. Z korelační matice plyne, že

$\lambda_1\lambda_2 = 0.83$, $\lambda_1\lambda_3 = 0.78$, $\lambda_2\lambda_3 = 0.67$,
$u_1 = 1 - \lambda_1^2$, $u_2 = 1 - \lambda_2^2$, $u_3 = 1 - \lambda_3^2$
a řešením vyjde $\lambda_1 = 0.99$, $\lambda_2 = 0.84$, $\lambda_3 = 0.79$,
$u_1 = 0.83$, $u_2 = 0.78$, $u_3 = 0.67$.

Vzorová úloha 4.4 *Ukázka pojmů a podstaty faktorové analýzy*

Na úloze **B4.02** *Účinky neuroleptik při tlumení rozličných psychóz* si ukážeme pomůcky vícerozměrné analýzy dat. K analýze užijeme také škálovaná data.

Řešení: byl použit program NCSS2000. Výstup metody Factor Analysis programu NCSS2000 pro nestandardizovaná data úlohy B402 obsahuje:

1. Popisné statistiky měr polohy a rozptýlení:

Proměnná	***n***	$\bar{x}$	***s***	*Komunalita H*
B402X1	20	20.05	33.89997	1.004443
B402X2	20	18.6	33.84236	1.005217
B402X3	20	2.95	5.206019	0.883469
B402X4	20	10.35	36.64951	0.846859

Klasické odhady parametrů polohy a rozptýlení pro jednotlivé proměnné informují o faktu, že proměnné byly správně vybrány. Komunalita ukazuje, jak dobře je tato proměnná predikována vybranými faktory.

2. (a) Korelační matice:

Proměnná	*B402X1*	*B402X2*	*B402X3*	*B402X4*
B402X1	1.000000	0.990529	0.835934	0.844519
B402X2	0.990529	1.000000	0.786439	0.851776
B402X3	0.835934	0.786439	1.000000	0.823784
B402X4	0.844519	0.851776	0.823784	1.000000

φ = 0.857883, Ln(Det|*R*|) = -7.336319, Bartlettův test = 123.49, SV = 6, Spočtená hladina významnosti α = 0.000000

Tabulka přináší korelace k posouzení celkové korelační struktury dat. Je zde několik případů vysokého korelačního koeficientu. Jsou-li všechny korelace nízké, menší než 0.3, není žádný důvod k užití faktorové analýzy. **Gleasonova-Staelinova míra redundance φ** = 0.8579 je veliká. Měří sílu vztahu mezi proměnnými. Nulová hodnota φ značí nulovou korelaci mezi proměnnými, zatímco hodnoty blízké jedné indikují silnou korelaci. I když je φ < 0.5, stále ještě *může být* nějaká struktura v datech. K vyčíslení Gleasonovy-Staelinovy míry φ se užívá vzorec

$$\varphi \;=\; \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} r_{ij}^2 \;-\; m}{m(m-1)}}\;.$$

Ln(Det*R*) značí přirozený logaritmus determinantu korelační matice. Při použití kovariance půjde o přirozený logaritmus determinantu kovariační matice. **Bartlettův test, SV, Spočtená hladina významnosti α**: jde o Bartlettův test sféricity k testování nulové hypotézy, že korelační matice je jednotková matice, všechny mimodiagonální prvky jsou nuly. Je-li velikost spočtené hladiny významnosti α větší než zadaná hodnota 0.05, neměli bychom aplikovat na tato data faktorovou analýzu ani metodu hlavních komponent. Test platí pro velké výběry (n>150) a užívá χ^2 rozdělení s $m(m-1)/2$ stupni volnosti: test lze užít pouze pro korelační, nikoliv však kovarianční matici. Testovací kritérium je vyčísleno vztahem

$$\chi^2 \cdot \frac{(11 \% 2m \&6n)}{6} \ln *\boldsymbol{R}* \cdot$$

(b) Čárový diagram absolutních hodnot korelační matice:

Proměnná	*B402X1*	*B402X2*	*B402X3*	*B402X4*
B402X1		|||||||||||||||	|||||||||||||	|||||||||||||
B402X2	|||||||||||||||		|||||||||||||	|||||||||||||
B402X3	|||||||||||||	|||||||||||||		|||||||||||||
B402X4	|||||||||||||	|||||||||||||	|||||||||||||	

Diagram zobrazuje absolutní hodnoty korelací a ukazuje největší a nejmenší korelaci proměnných.

3. Vyšetření indexového grafu úpatí vlastních čísel (Scree Plot):

Index	**Vlastní číslo λ_i**	**Individualní procento**	**Kumulativní procento**	**Kumulativní čárový graf úpatí**
1	3.507191	93.92	93.92	||||||||||||||||
2	0.187628	5.02	98.94	||
3	0.045168	1.21	100.15		
4	-0.005689	-0.15	100.00		

Jde o vlastní čísla matice $\boldsymbol{LL^T}$**.** Často se užívají jako rozlišovací kritérium při výběru počtu faktorů. Užívá se těch faktorů, jejichž vlastní čísla jsou větší než 1. Suma vlastních čísel je rovna počtu proměnných. Odtud platí, že první faktor obsahuje informaci obsaženou v 3.507191 původních proměnných. Zatímco všechna vlastní čísla jsou v PCA kladná, vlastní čísla ve FA mohou být i záporná. Obyčejně se tyto faktory vypouští a analýza se potom opakuje. **Individuální procento:** první sloupec přináší procento celkové proměnlivosti v proměnných, vystižené tímto faktorem a druhý sloupec pak **Kumulativní procento**. **Kumulativní čárový graf úpatí** určí hranu, index, rovnající se počtu užitých faktorů.

4. (a) Vlastní vektory pro jednotlivé faktory:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	-0.525227	-0.361706	0.523702
B402X2	-0.519584	-0.545676	-0.236211
B402X3	-0.473506	0.692071	0.400954
B402X4	-0.479542	0.304046	-0.713566

Jde o vlastní vektory matice $\boldsymbol{L\,L^T}$.

(b) Čárový diagram absolutních hodnot vlastních vektorů pro jednotlivé faktory:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	||||||||||||	||||||||	||||||||||||
B402X2	||||||||||||	||||||||||	|||||
B402X3	|||||||||	||||||||||||	||||||||
B402X4	|||||||||	|||||||	|||||||||||||

Diagram absolutních hodnot vlastních vektorů umožňuje rychle posoudit velikost vlastních vektorů, totiž, která původní proměnná x_j silně koreluje s dotyčným faktorem. Tak se znázorní struktura obou faktorů.

5. (a) Faktorové váhy pro jednotlivé faktory:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	-0.983619	-0.156677	0.111301

B402X2	-0.973051	-0.236365	-0.050201
B402X3	-0.886759	0.299778	0.085213
B402X4	-0.898062	0.131701	-0.151652

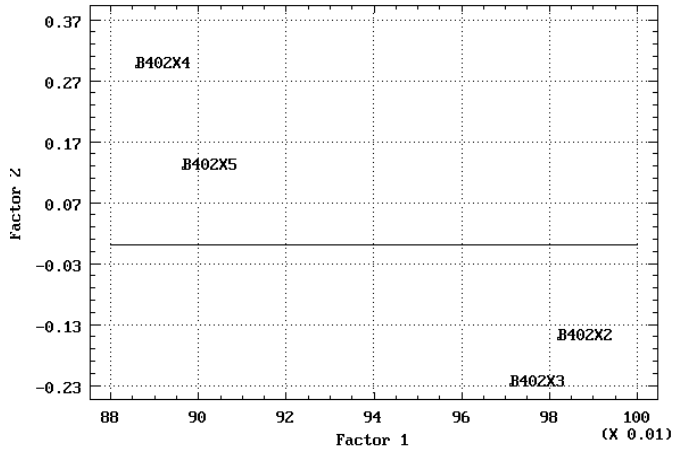
Tabulka numericky znázorňuje korelace mezi proměnnými a faktory.

(b) Čárový diagram absolutních hodnot faktorových vah pro jednotlivé faktory:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	||||||||||||||||	||||	|||
B402X2	||||||||||||||||	|||||	||
B402X3	|||||||||||||||	||||||	||
B402X4	|||||||||||||||	|||	||||

Diagram znázorňuje absolutní hodnotu faktorových zátěží a vyjadřuje korelační strukturu jednotlivých původních proměnných s dotyčnými faktory. Faktor je obyčejně ovlivněn všemi původními proměnnými. Faktor1 je nejvíce ovlivněn *B402X1* a *B402X2*. Faktor2 pak nejvíce *B402X3* a také *B402X2* a nejméně proměnnými *B402X1* a *B402X4* .

6. Graf faktorových vah:



Obr. 4.12 Graf faktorových vah pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* pro data úlohy B402, ***STATGRAPHICS.*** *Proměnné B402X1 a B402X2 leží v diagramu blízko sebe, a proto silně korelují. Proměnné B402X3 a B402X4 jsou poněkud dál od sebe, proto méně korelují. Méně korelují se zbývajícími dvěma proměnnými B402X1 a B402X2, jsou totiž umístěny daleko od nich.*

7. Příspěvky daného faktoru do komunality:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*	*Kommunalita*
B402X1	0.967507	0.024548	0.012388	1.00
B402X2	0.946828	0.055869	0.002520	1.00
B402X3	0.786341	0.089867	0.007261	0.88
B402X4	0.806515	0.017345	0.022998	0.85

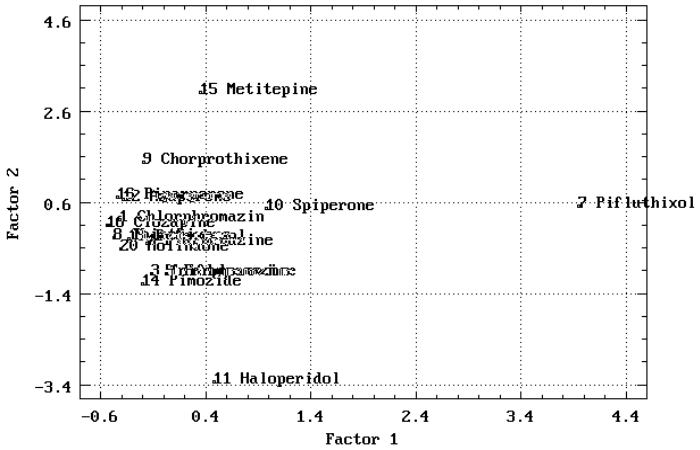
Komunalita představuje podíl proměnlivosti proměnné, vyjádřené dotyčným faktorem. Je podobná hodnotě R^2, kterou dostaneme, když budeme původní proměnné regresovat vybranými faktory. Tabulka obsahuje příspěvek daného faktoru do komunality. Diagram přináší příspěvky vybraných faktorů do komunality.

8. Faktorová skóre jednotlivých faktorů:

Proměnná	*Faktor1*	*Faktor2*	*Faktor3*
B402X1	-0.2804579	-0.8350396	2.464167
B402X2	-0.2774445	-1.259754	-1.111442
B402X3	-0.2528401	1.597723	1.886601
B402X4	-0.256063	0.7019241	-3.357533

V tabulce jsou koeficienty, které jsou užity k vytvoření faktorového skóre. Faktorová skóre jsou hodnoty faktorů pro jednotlivé řádky dat. Tyto koeficienty skóre jsou podobné vlastním vektorům. Protože byly předem normovány, přináší skóre jednotkový rozptyl a nikoliv rovný vlastním číslům. To způsobuje, že každý z faktorů má stejný rozptyl. Uživatel může použít tato skóre, jestliže chce vypočítat faktorové skóre pro nové řádky, jež nebyly zatím zařazeny do analýzy.

9. Rozptylový diagram faktorového skóre: diagram ukazuje na závislost faktoru proti faktoru. Prvních *k* faktorů (kde *k* je počet největších vlastních čísel) ukazuje na hlavní strukturu, která byla nalezena v datech. Zbytek faktorů ukazuje odlehlé hodnoty a lineární závislosti.



Obr. 4.13 Rozptylový diagram faktorových skore pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* ze standardizovaných dat úlohy B402, **STATGRAPHICS.** *Kromě tří objektů 7, 11 a 15 leží zbývajících 17 objektů v jediném shluku. Objekty 7, 11 a 15 tvoří každý samostatný shluk. Co do podobnosti ve čtyřech vlastnostech, vystižených dvěma hlavními komponentami v rovině, lze hovořit o 4 shlucích: první 15 Metitepine, druhý 7 Pifluthixol, třetí 11 Haloperidol, čtvrtý zbytek.*

4.5.3 Kanonická korelační analýza

Kanonická korelační analýza je vícerozměrná metoda, která se používá ke zkoumání závislosti mezi dvěma skupinami proměnných. První ze dvou skupin se považuje za *soubor závisle proměnných* $\boldsymbol{y}$ a druhá za *soubor nezávisle proměnných* $\boldsymbol{x}$. Toto rozdělení je ale čistě účelové z důvodu výkladu a nemá žádný vliv na řešení problému. Jde v podstatě o rozšíření metody vícenásobné lineární regrese a korelační analýzy. Zatímco ve vícenásobné lineární regresi hledáme nejlepší kombinaci *m* nezávisle proměnných $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_m}$ k výpočtu *jediné* závisle proměnné *y* hledáme v kanonické korelační analýze lineární vztah $\boldsymbol{U_1} = a_1 \boldsymbol{y_1} + a_2 \boldsymbol{y_2} + ... + a_p \boldsymbol{y_p}$ mezi skupinou *p,* čili *více než jediné*, závisle proměnných $\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_p}$, a dále linearní vztah $\boldsymbol{V_1} = b_1 \boldsymbol{x_1} + b_2 \boldsymbol{x_2} + ... + b_p \boldsymbol{x_p}$ mezi skupinou *m* nezávisle proměnných $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_p}$. Podstata metody spočívá v tom, že se v každé skupině proměnných vyhledávají

koeficienty a a b tak, aby pro všech n objektů vyčíslené kanonické proměnné U_{1i} a V_{1i}, $i = 1, ..., n$, vykazovaly maximální párový korelační koeficient. Po jejich nalezení se pak hledají další lineární kombinace čili kanonické proměnné $\boldsymbol{U_2}$ a $\boldsymbol{V_2}$, které mají druhý největší korelační koeficient za podmínky, že $\boldsymbol{U_2}$ a $\boldsymbol{V_2}$ jsou nekorelované s prvními kanonickými proměnnými $\boldsymbol{U_1}$ a $\boldsymbol{V_1}$.

V kanonické korelační analýze jsou koeficienty a a b hledány tak, aby maximalizovaly korelaci mezi proměnnými U_1 a V_1. Po nalezení nejlepších odhadů a a b se U_1 nazývá *první kanonická proměnná* závisle proměnných $\boldsymbol{y}$ a V_1 *první kanonická proměnná* nezávisle proměnných $\boldsymbol{x}$. Obě kanonické proměnné mají průměr roven nule. Korelace mezi U_1 a V_1 se nazývá *první kanonická korelace* a čtverec této korelace je nazýván *vlastní číslo.*

První kanonická korelace je tudíž největší možná korelace mezi lineárními kombinacemi závisle proměnných $\boldsymbol{y}$ a lineárními kombinacemi nezávisle proměnných $\boldsymbol{x}$. První kanonická korelace představuje analogii vícenásobnému korelačnímu koeficientu ve vícenásobné lineární regresi mezi *jedinou* závisle proměnnou $\boldsymbol{y}$ a souborem nezávisle proměnných $\boldsymbol{x}$. Rozdíl proti vícenásobné lineární regresi je pouze v tom, že u kanonické korelace je *několik* závisle proměnných $\boldsymbol{y}$ a dále je nutno navíc hledat lineární kombinaci mezi nimi.

Vzorová úloha 4.5 *Ukázka pojmů a podstaty kanonické korelační analýzy*

Kanonická korelační analýza se často užívá v psychologii a pedagogice, např. k validování testu inteligence. Postup je pak takový, že dva testy, standardní test a nový test jsou aplikovány na jedny a tytéž osoby. Na dva testy, každý o 10 bodovaných otázkách (0 až 100 bodů), odpovědělo 15 studentů a byly tak získány dvě matice dat *TEST1* (rozměru 15 ×10) a *TEST2* (rozměru 15 ×10). Kanonická korelace nalezne pro 15 studentů hodnoty váženého průměru z 10 bodovaných odpovědí standardního testu $U_{1,i}$, $i = 1, ..., 15$. Tyto pak koreluje s 15 hodnotami váženého průměru 10 bodovaných odpovědí nového a validovaného testu $V_{1,i}$, $i = 1, ..., 15$. Váhy jsou konstruovány tak, že maximalizují korelaci mezi těmito dvěma průměry. Jde o korelaci mezi těmito dvěma testy, když máme k dispozici 15 dvojic průměrů $\{U, V\}$. Vyčíslená korelace se nazývá *první kanonický korelační koeficient.*

Můžeme sestrojit i jiný soubor vážených průměrů (a to třeba jen pro vybrané otázky), nesouvisející s prvním souborem, a vypočítat jejich korelaci. Proces se opakuje tolikrát, až se počet kanonických korelací rovná počtu proměnných v menší ze dvou skupin.

Budeme nadále rozlišovat *původní proměnné x, y* a *kanonické proměnné V, U.* Kanonické proměnné jsou proměnné, které byly sestaveny z vážených průměrů původních proměnných, např. z odpovědí na 10 otázek testu (původní proměnné) se vytvoří kanonická proměnná, která představuje pro každého studenta jediné číslo jako výsledek dotyčného testu. Soubor kanonických proměnných $\boldsymbol{U}$ vznikl z původních proměnných $\boldsymbol{y}$**.** Soubor kanonických proměnných $\boldsymbol{V}$ vznikl z původních proměnných $\boldsymbol{x}$. V průběhu kanonické korelace by mělo být vzato v úvahu následujících několik bodů:

1. *Určení počtu párů kanonických proměnných:* počet možných párů je roven menšímu číslu z počtu proměnných v každém souboru.

2. *Kanonické proměnné je nutno také interpretovat:* stejně jako ve faktorové analýze pracujeme i zde s matematicky umělými proměnnými, které je často obtížné fyzikálně vysvětlit.

3. *Důležitost každé proměnné musí být vyhodnocena ze dvou hledisek:* musíme určit intenzitu vztahu mezi kanonickou proměnnou U a původní proměnnou y nebo proměnnými V a x, ze které byla kanonická proměnná vytvořena. Musíme rovněž vyjádřit intenzitu vztahu mezi oběma kanonickými proměnnými V a U.

4. *Pozornost je třeba věnovat velikosti výběru*: v sociálních vědách potřebujeme obvykle 10 experimentálních hodnot na jeden neznámý parametr, v přírodních vědách trochu méně.

Normalita a odlehlé body. Kanonická korelace nemá silné požadavky na normalitu. Odlehlé hodnoty však mohou zničit průběh výpočtu či přinést velké komplikace fyzikálně, biologicky či jinak.

Linearita. Kanonická korelační analýza předpokládá pouze lineární závislost mezi proměnnými. Pečlivě je třeba vyšetřit grafy každého páru proměnných a prověřit linearitu a odlehlé body. Kanonická korelace je založena na korelaci mezi dvěma soubory proměnných. Korelační matice všech proměnných lze pak rozdělit na čtyři části:

1. $\boldsymbol{R}_{xx}$. Jde o korelaci mezi proměnnými $\boldsymbol{x}$.
2. $\boldsymbol{R}_{yy}$. Jde o korelaci mezi proměnnými $\boldsymbol{y}$.
3. $\boldsymbol{R}_{xy}$. Jde o korelaci mezi proměnnými $\boldsymbol{x}$ a $\boldsymbol{y}$.
4. $\boldsymbol{R}_{yx}$. Jde o korelaci mezi proměnnými $\boldsymbol{y}$ a $\boldsymbol{x}$.

Kanonická korelace může být vyjádřena s využitím metody SVD (Singlular Value Decomposition) matice $\boldsymbol{C}$, kde $\boldsymbol{C} = \boldsymbol{R}_{yy}^{-1}\,\boldsymbol{R}_{yx}\,\boldsymbol{R}_{xx}^{-1}\,\boldsymbol{R}_{xy}$. V SVD rozkladu matice $\boldsymbol{C}$ vztahem $\boldsymbol{C} = \hat{a}_y^{T}\,\boldsymbol{\lambda}\,\hat{a}_y$ je diagonální matice $\boldsymbol{\lambda}$ vlastních čísel vytvořena z vlastních čísel matice $\boldsymbol{C}$. Pak j-té vlastní číslo λ_j matice $\boldsymbol{C}$ je rovno čtverci j-té kanonické korelace, která se nazývá r^2_j. Odtud j-tá kanonická korelace je druhou odmocninou z j-tého vlastního čísla matice $\boldsymbol{C}$.

Dva soubory kanonických koeficientů (podobně jako regresních koeficientů) se užívají pro každou kanonickou korelaci: jeden pro proměnné $\boldsymbol{x}$ a druhý pro proměnné $\boldsymbol{y}$. Tyto kanonické koeficienty jsou definovány

$$\boldsymbol{a} = (\boldsymbol{R}_{yy}^{-1/2})^{T}\,\hat{\boldsymbol{a}}_y \quad , \qquad \boldsymbol{b} = \boldsymbol{R}_{xx}^{-1}\,\boldsymbol{R}_{xy}\,\boldsymbol{a}\,\lambda_j^{-1/2},$$

kde $\hat{\boldsymbol{a}}_y$ je normovaná matice vlastních vektorů pro y. Kanonické skóre pro $\boldsymbol{V}$ a $\boldsymbol{U}$ vzniklo vynásobením standardizovaných dat (od prvků se odečte průměr a výsledek se vydělí směrodatnou odchylkou) maticí kanonických koeficientů $V = \boldsymbol{Z}_x\,\boldsymbol{b}$ a $U = \boldsymbol{Z}_y\,\boldsymbol{a}$, kde $\boldsymbol{Z}_x$ a $\boldsymbol{Z}_y$ představují standardizovaná data $\boldsymbol{X}$ a $\boldsymbol{Y}$.

Abychom pomohli interpretaci kanonických proměnných, vyčíslíme také *matice zátěží* dle vztahů:

$$\boldsymbol{L}_x = \boldsymbol{R}_{xx}\,\boldsymbol{b} \quad \text{a} \quad \boldsymbol{L}_y = \boldsymbol{R}_{yy}\,\boldsymbol{a} \quad .$$

Jsou to vlastně korelace mezi původními proměnnými a kanonickými proměnnými.

Postup kanonické korelační analýzy

1. *Bodové odhady parametrů polohy a rozptýlení všech proměnných:* vyčíslí se aritmetický průměr a směrodatná odchylka pro všechny proměnné.

2. *Korelační koeficienty všech původních proměnných:* vyčíslí se párové korelační koeficienty mezi všemi proměnnými.
3. *Kanonické korelace:* vedle kanonických korelačních koeficientů obsahuje řadu pomocných statistik k interpretaci kanonické korelace.
4. *Objasněná proměnlivost v datech:* obsahuje procento proměnlivosti v každém souboru proměnných, vysvětlovaných jiným souborem proměnných.
5. *Standardizované kanonické parametry pro kanonické proměnné Y a X:* koeficienty slouží k interpretaci proměnných v hodnotě váhy u každé proměnné.
6. *Korelace párů původní proměnné vs. kanonická proměnná:* napomůže snadnější interpretaci kanonických proměnných. Je-li kanonická proměnná silně korelovaná s původní proměnnou, má pak i stejnou či podobnou interpretaci.
7. *Tabulka kanonického skóre pro všechny objekty:* obsahuje kanonické skóre každého souboru proměnných pro každý řádek úplných dat. Hodnoty lze také vynést do grafu.
8. *Grafy kanonického skóre pro všechny objekty:* grafy ukazují na vztah mezi každým párem kanonických proměnných. Korelační koeficient v prvním grafu je *první kanonický korelační koeficient.*

Vzorová úloha 4.6: *Postup kanonické korelační analýzy*

V úloze **S4.18** *Testy IQ* bylo vyšetřeno 15 respondentů (čili 15 objektů) pěti rozličnými testy a vyčíslena hodnota *IQ* (čili dohromady šesti původními proměnnými) za účelem zjištění objektivní hodnoty výsledného inteligenčního kvocientu. Každý z testů obsahoval 10 bodovaných otázek (0 až 100 bodů), na které odpovědělo 15 studentů, matice *TEST1* až *TEST5* a *IQ* byly velikosti (15 ×10). Kanonická korelace nalezne 15 hodnot váženého průměru z 10 bodovaných odpovědí každého testu a koreluje je s 15 hodnotami váženého průměru 10 bodovaných odpovědí jiného testu. Jde o korelaci vždy mezi dvojicí testů, když je k dispozici 15 dvojic vážených průměrů $\{X, Y\}$. Pokuste se vyšetřit tři vybrané testy v závislosti na prvních třech testech čili pokuste se popsat závislostí (TEST4, TEST5, *IQ*) = *f(TEST1, TEST2, TEST3)*.

Řešení: výstup Canonical correlation (NCSS2000) pro nestandardizovaná data

1. Popisné statistiky polohy a rozptýlení:

Typ	Proměnná	Průměr	Směrodatná odchylka	Úplné řádky bez chybějících hodnot
U	*Test4*	65.53333	13.95332	15
U	*Test5*	69.93333	16.15314	15
U	*IQ*	104.3333	11.0173	15
V	*Test1*	67.93333	17.39239	15
V	*Test2*	61.4	19.39735	15
V	*Test3*	72.33334	14.73415	15

Obsahuje popisné statistiky pro všechny proměnné. Kontroluje, zda průměry dosahují "přijatelných" hodnot a zda počet úplných “neděravých” řádků je správný.

2. Korelační koeficienty párů všech původních proměnných:

	Test4	*Test5*	*IQ*	*Test1*	*Test2*	*Test3*
Test4	1.000000	-0.172864	0.371404	0.753937	0.719623	-0.140941

Test5	-0.172864	1.000000	-0.058064	0.013967	-0.281449	0.347335
IQ	0.371404	-0.058064	1.000000	0.225648	0.240651	0.074070
Test1	0.753937	0.013967	0.225648	1.000000	0.100018	-0.260801
Test2	0.719623	-0.281449	0.240651	0.100018	1.000000	0.057232
Test3	-0.140941	0.347335	0.074070	-0.260801	0.057232	1.000000

Obsahuje jednoduché korelace čili Pearsonovy korelační koeficienty mezi všemi proměnnými.

3. Kanonické korelace:

Index proměnné	**Kanonická korelace**	***D***	**F-test**	**Čitatel SV**	**Jmen. SV**	**Spočtená hlad. významnosti**	**Wilkovo Lambda**
1	0.995600	0.991219	16.58	9	22	0.000000	0.006819
2	0.467461	0.218519	0.67	4	20	0.617695	0.776503
3	0.079810	0.006370	0.07	1	11	0.795498	0.993630

F-test testuje, zda tato kanonická korelace a všechny následné jsou nulové.

Obsahuje kanonické korelace a veškeré podpůrné informace, potřebné k interpretaci. **Index proměnné** je pořadové číslo kanonické korelace. Je třeba si uvědomit, že první korelace bude vždy největší. **Kanonická korelace** je hodnota kanonického korelačního koeficientu. Koeficient má stejné vlastnosti jako jiné korelace. Rozsah je od -1 do +1, přičemž 0 značí nízkou korelaci a absolutní hodnota blízká jedné pak perfektní korelaci. ***D*** značí čtverec kanonického korelačního koeficientu (čili koeficient determinace) a udává hodnotu těsnosti proložení lineárního modelu kanonické proměnné *Y* na odpovídající *X* kanonické proměnné. ***F*-test:** hodnota *F*-testu při testování statistické významnosti Wilkova lambda, odpovídajícího řádku a všech hodnot pod tímto řádkem. V tomto případě první *F*-hodnota testuje významnost první, druhé a třetí kanonické korelace, zatímco druhá *F*-hodnota testuje významnost pouze druhé a třetí. **Čitatel SV:** počet stupňů volnosti v čitateli. **Jmenovatel SV:** počet stupňů volnosti ve jmenovateli. **Spočtená hladina významnosti:** hodnota spočtené hladiny významnosti čili pravděpodobnosti pro výše vyčíslené *F*-testační kritérium. Hodnota blízko nule ukazuje na významnou kanonickou korelaci. Hranice $\alpha = 0.05$ bývá často užívána k určení statistické významnosti, tj. hodnoty pravděpodobnosti větší než 0.05 ukazující na statistickou nevýznamnost. **Wilkovo lambda:** hodnota Wilkova lambda pro kanonickou korelaci tohoto řádku představuje vlastně vícerozměrné zobecnění *D*. Wilkovo lambda je interpretováno opačně než *D*, tedy hodnota blízká nule ukazuje na vysokou korelaci a hodnota blízká 1 na nízkou korelaci.

4. Objasněná proměnlivost v datech:

Index kanonické proměnné	**Proměnlivost v těchto proměnných**	**Objasněno těmito proměnnými**	**Procento objasnění jednotlivě**	**Procento objasnění kumulativně**	**Kanonický koeficient determinace**
1	*U*	*U*	37.6	37.6	0.9912
2	*U*	*U*	32.1	69.7	0.2185
3	*U*	*U*	30.3	100.0	0.0064
1	*U*	*V*	37.2	37.2	0.9912
2	*U*	*V*	7.0	44.3	0.2185
3	*U*	*V*	0.2	44.5	0.0064
1	*V*	*U*	37.1	37.1	0.9912
2	*V*	*U*	5.4	42.5	0.2185
3	*V*	*U*	0.2	42.8	0.0064
1	*V*	*V*	37.4	37.4	0.9912
2	*V*	*V*	24.8	62.2	0.2185
3	*V*	*V*	37.8	100.0	0.0064

Obsahuje procento proměnlivosti v každém souboru proměnných, vysvětlovaných jiným souborem proměnných. **Index kanonické proměnné:** pořadové číslo (index) kanonické proměnné. Nesmíme zapomenout, že maximální počet proměnných se rovná minimálnímu počtu proměnných v každém souboru. **Proměnlivost v těchto proměnných:** je stejné jako následující. **Objasněno těmito proměnnými:** každý řádek tabulky obsahuje výsledek,

jak dokonale je soubor proměnných vysvětlen dotyčnou kanonickou proměnnou. Tento sloupec označuje, který soubor proměnných je právě komentován. **Procento objasnění jednotlivě:** tento sloupec ukazuje procento změny v označeném souboru proměnných, které je vysvětleno touto kanonickou proměnnou. **Procento objasnění kumulativně:** tento sloupec ukazuje kumulativní procento změny v označeném souboru proměnných, které je vysvětleno touto kanonickou proměnnou a ostatními výše. **Kanonický koeficient determinace:** čtverec kanonického korelačního koeficientu.

5. Standardizované kanonické parametry pro kanonické proměnné *U*:

	U_1	U_2	U_3
Test4	1.021375	0.104989	0.370860
Test5	-0.005995	0.990267	0.224017
IQ	-0.065358	0.229775	-1.050237

6. Standardizované kanonické parametry pro kanonické proměnné *V*:

	V_1	V_2	V_3
Test1	0.690657	0.592485	0.510311
Test2	0.655584	-0.428196	-0.636097
Test3	-0.008941	0.919574	-0.485199

Koeficienty jsou užity k určení standardních skóre pro kanonické proměnné V a U . Slouží k interpretaci proměnných v hodnotě váhy, dané u každé proměnné při konstrukci kanonické proměnné. Jsou analogické standardizovaným parametrům β ve vícenásobné lineární regresi.

7. Korelace párů původní proměnné vs. kanonická proměnná:

	U_1	U_2	U_3	V_1	V_2	V_3
Test4	0.998137	0.019146	-0.057927	0.993745	0.008950	-0.004623
Test5	-0.178759	0.958777	0.220890	-0.177972	0.448190	0.017629
IQ	0.314333	0.211270	-0.925505	0.312950	0.098760	-0.073865
Test1	0.755221	0.144834	0.045750	0.758559	0.309832	0.573230
Test2	0.720964	-0.147861	-0.048910	0.724151	-0.316308	-0.612826
Test3	-0.150877	0.346177	-0.052251	-0.151544	0.740547	-0.654694

Ukazuje korelace párů mezi původní proměnnou a kanonickou proměnnou. Určení, které proměnné jsou vysoce korelované s odpovídající kanonickou proměnnou, napomůže snadnější interpretaci kanonických proměnných. Např. U_1 je vysoce korelovaná s *TEST4*. Proto předpokládáme, že U_1 má stejnou interpretaci jako *TEST4*.

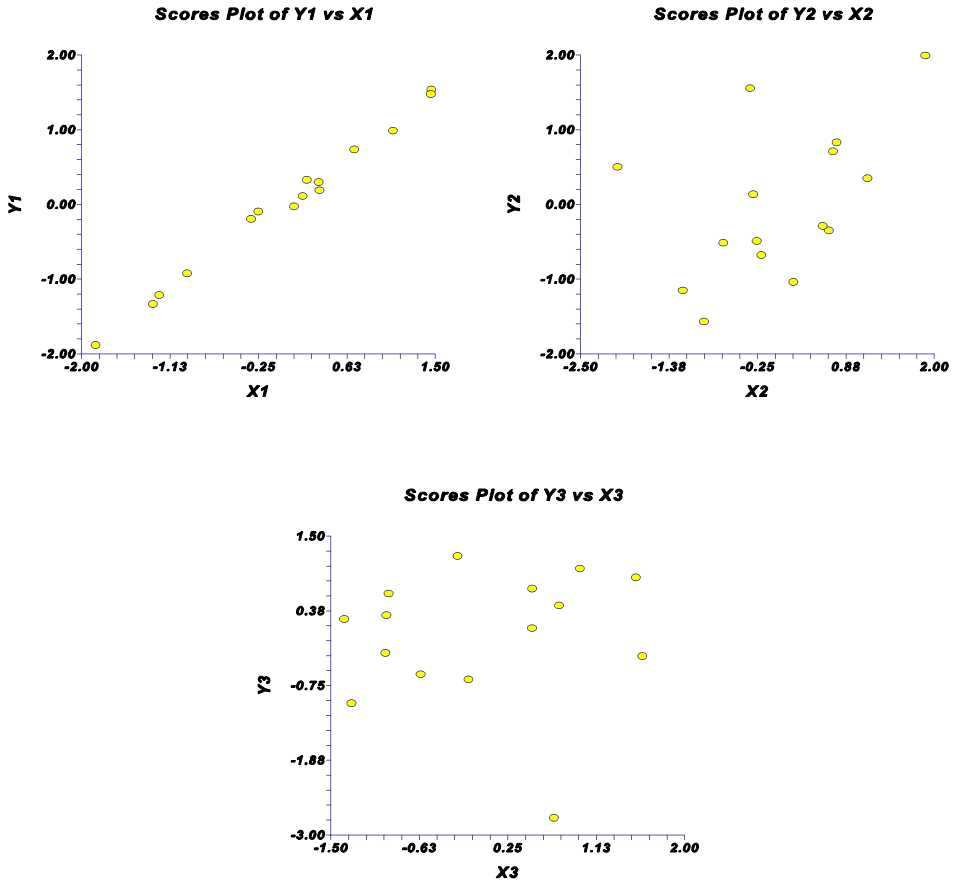
6. Tabulka kanonického skóre pro všechny objekty:

Řádek	U_1	U_2	U_3	V_1	V_2	V_3
1	-0.193124	-0.348044	-0.308495	-0.323303	0.660431	1.582089
2	-1.214743	0.350598	0.877022	-1.232224	1.150186	1.517131
3	-0.026336	0.135325	0.250782	0.103271	-0.304012	-1.369888
4	1.536744	1.992049	-0.657871	1.461462	1.887123	-0.138798
5	0.189923	0.709643	0.455333	0.354314	0.711949	0.757851
6	0.986597	-0.677646	0.115011	1.081350	-0.201044	0.489839
7	0.299464	-0.490602	0.708912	0.345665	-0.258540	0.491428
8	-0.922687	0.503305	1.011073	-0.954587	-2.031644	0.963769
9	-1.881691	-0.288458	0.308479	-1.862181	0.579830	-0.951854
10	-1.333760	0.829021	-1.015632	-1.294283	0.756978	-1.297593
11	0.111861	-1.151067	-2.741954	0.188193	-1.199877	0.707092
12	0.329061	1.555086	-0.579356	0.228934	-0.342184	-0.612825
13	0.736439	-1.037650	0.634374	0.698925	0.206974	-0.929772
14	1.477329	-0.513679	1.201759	1.456751	-0.684236	-0.247278

15	-0.095076	-1.567882	-0.259437	-0.252288	-0.931936	-0.961191

Obsahuje kanonické skóre každého souboru proměnných pro každý řádek úplných dat. Jde o hodnoty, které lze rovněž vynést do grafu.

7. Grafy kanonického skóre pro všechny objekty: grafy ukazují na vztah mezi každým párem kanonických proměnných. Korelační koeficient r_1 dat v prvním grafu (U_1 versus V_1 v grafech označený jako *Y1* versus *X1*) je první kanonický korelační koeficient.



Obr. 4.14 Grafy tří párů kanonických skóre pro všechny objekty.

4.6 Klasifikace objektů

Hledáním struktury a vzájemných vazeb v objektech se zabývají klasifikační metody vícerozměrné statistické analýzy. *Klasifikační metody* jsou postupy, pomocí kterých se jeden objekt zařadí do jedné existující třídy (*diskriminační analýza DA*), nebo pomocí nichž

lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd či shluků (*analýza shluků CLU*). Postup klasifikace je založen na určitých předpokladech o vlastnostech klasifikovaných objektů, např. normální rozdělení náhodného vektoru, charakterizujícího objekty; pak hovoříme o *parametrických klasifikačních metodách.* Není-li klasifikace založena na znalostech rozdělení náhodného vektoru, mluvíme o *neparametrických klasifikačních metodách.* Významnou roli při hledání struktury a vazeb mezi objekty na základě jejich podobnosti tvoří také *vícerozměrné škálování MDS.*

4.6.1 Diskriminační analýza DA

Diskriminační analýza patří mezi metody zkoumání závislosti mezi skupinou *p* nezávisle proměnných, nazvaných *diskriminátory,* tj. sloupců zdrojové matice na jedné straně a jednou kvalitativní závisle proměnnou na druhé straně. Umožňuje zařazení objektu do jedné z již existujících tříd. Ve vstupních datech jsou svými hodnotami diskriminátorů u všech objektů dány *zařazené objekty do primárních tříd*. Dále jsou dány *nezařazené objekty*, pro které budeme hledat zařazení do třídy. Objekt zařadíme do třídy na základě jeho největší míry podobnosti, např. nejmenší Mahalanobisovy vzdálenosti.

Diskriminační (zařazovací) pravidla: při diskriminační analýze se snažíme vyčíslit hodnotu *diskriminační funkce*, která nám usnadní zařazení do primární třídy. Takto vyčíslené hodnoty funkce používáme také ke třídění *nezařazených objektů* do předem známých primárních tříd, a to na základě *p* diskriminátorů $x_1 , x_2 , ..., x_p$. Každá primární třída je charakterizována svou funkcí hustoty pravděpodobnosti $f_j(\boldsymbol{x})$, kde $\boldsymbol{x}^{\mathrm{T}} = [x_1 , x_2 , ..., x_p]$. Existuje citlivé pravidlo pro zařazení, diskriminaci objektu vektoru $\boldsymbol{x}$, do třídy G_j

$$f_j(\boldsymbol{x}) \quad \max_{i \; 0 \; 1,...,g} f_i(\boldsymbol{x}) \quad .$$

Uveďme příklady diskriminace:

1. Existuje jednoduchá binární proměnná x a dvě třídy G_1 a G_2. Nejprve předpokládejme, že pravděpodobnost $P(x = 0) = P(x = 1) = 1/2$ a dále pravděpo-dobnost $P(x = 0) = 1/4$ a pravděpodobnost $P(x = 1) = 3/4$. Pravidlo zařadí objekt $x = 0$ do G_1 a objekt $x = 1$ do G_2.

2. Předpokládejme spojitou jednoduchou proměnnou x a opět dvě třídy G_1 a G_2. Ve třídě G_1 má proměnná normální rozdělení se střední hodnotou μ_1 a rozptylem σ^2_1 a ve třídě G_2 má proměnná rovněž normální rozdělení se střední hodnotou μ_2 a rozptylem σ^2_2, přičemž budeme předpokládat $\mu_1 < \mu_2$ a $\sigma^2_1 > \sigma^2_2$. Pomocí diskriminačního pravidla $f_j(\boldsymbol{x})$ bude objekt o skóre $\boldsymbol{x}$ zařazen do třídy G_1, když bude platit $f_1(\boldsymbol{x}) > f_2(\boldsymbol{x})$. Nahrazením skutečnou hustotou pravděpodobnosti normálního rozdělení dostaneme pravidlo k zařazení objektu x do třídy G_1:

$$\frac{\sigma_1}{\sigma_2} \exp\left\{ \& \frac{1}{2} \left[\frac{(x \; \& \; \mu_1)^2}{\sigma_1^2} \; \& \; \frac{(x \; \& \; \mu_2)^2}{\sigma_2^2} \right] \right\} > 1$$

a po zlogaritmování a úpravě bude toto pravidlo ve tvaru

$$x^2 \left[\frac{1}{\sigma_1^2} - \frac{1}{\sigma_1^2}\right] - 2x \left[\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right] + \left[\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right] < 2 \ln \frac{\sigma_1}{\sigma_2} \quad .$$

Dle tohoto pravidla dojde k rozdělení hodnot x do dvou tříd: první třída G_1 bude obsahovat malé hodnoty x a druhá třída G_2 velké hodnoty x. Ve zvláštním případě $\sigma_1 = \sigma_2$ dostaneme pravidlo pro zařazení do třídy G_1 ve znění $|x - \mu_2| > |x - \mu_1|$. Bude-li navíc $\mu_1 < \mu_2$, objekt se skóre x padne do třídy G_1, když bude platit, že $x < (\mu_1 + \mu_2)/2$.

Zobecnění diskriminačního pravidla: G_1 je třída objektů s vícerozměrným normálním rozdělením a střední hodnotou μ_1 a G_2 obdobně třída objektů se střední hodnotou μ_2. Předpokládejme, že kovarianční matice obou tříd jsou stejné a užijeme proto pro ně společné označení $\boldsymbol{S}$. Obecné pravidlo zařazení objektu o vektoru $\boldsymbol{x}$ do třídy G_1 bude

$$(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\, \boldsymbol{S}^{-1} \left(\boldsymbol{x} - \frac{\boldsymbol{\mu_1} + \boldsymbol{\mu_2}}{2}\right) > 0 \quad .$$

Když třídy mají známé hustoty pravděpodobnosti rozličných rozdělení $\pi_1, \pi_2, ..., \pi_p$, bude pravidlo o zařazení do třídy upraveno takto: jde-li o 2 třídy, bude pravidlo ve tvaru

$$(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\, \boldsymbol{S}^{-1} \left(\boldsymbol{x} - \frac{\boldsymbol{\mu_1} + \boldsymbol{\mu_2}}{2} \right) > \ln \frac{\pi_1}{\pi_2} \quad .$$

Lineární diskriminační funkce (LDA): z diskriminačních funkcí je neznámější *Fisherova lineární diskriminační funkce* tvaru

$$z_i = a_{i1}\, x_1 + a_{i2}\, x_2 + a_{i3}\, x_3 + ... + a_{ip}\, x_p \ ,$$

kde p je počet proměnných primárních tříd čili počet diskriminátorů a $x_1, x_2, ..., x_p$ jsou standardizované hodnoty těchto proměnných. Parametry z_i nazýváme *standardi-zované klasifikační koeficienty* Fisherovy diskriminační funkce $\boldsymbol{a}^{\mathrm{T}} = [a_1, a_2, ..., a_p]$, které byly nalezeny tak, že poměr rozptylu mezi třídami $\boldsymbol{B}$ a rozptylu uvnitř tříd $\boldsymbol{S}$

$$V = \boldsymbol{a}^{\mathrm{T}}\, \boldsymbol{B}\, \boldsymbol{a} \,/ (\boldsymbol{a}^{\mathrm{T}}\, \boldsymbol{S}\, \boldsymbol{a})$$

je maximální. Zde $\boldsymbol{B}$ je kovarianční matice třídních průměrů a $\boldsymbol{S}$ je celková kovarianční matice uvnitř tříd. Vektor $\boldsymbol{a}$, který maximalizuje poměr V, se vypočte ze vztahu

$$(\boldsymbol{B} - \lambda\, \boldsymbol{S})\, \boldsymbol{a} = 0 .$$

V případě pouze dvou tříd budou klasifikační koeficienty diskriminační funkce $\boldsymbol{a}^{\mathrm{T}} = [a_1, a_2, ..., a_p]$ vypočteny jednoduchým vztahem $\boldsymbol{a} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$.

Vzorová úloha 4.7 *Užití lineární diskriminační funkce*

Předpokládejme, že máme data o 2 třídách objektů tibetských lebek v úloze **B4.14** *Aglomerativní hierarchické shlukování při analýze lebek Tibeťanů*: prvních 13 bylo nalezeno v hrobech v Sikkimu a okolí, zatímco druhých 15 lebek na bojištích okolo Lhasy. První třída vede ke středním hodnotám $\vec{\boldsymbol{x}}_1^{\mathrm{T}}$ = [174.82, 139.35, 132.00, 69.82, 130.35] a kovarianční matici

$$S_1 = \begin{vmatrix} 45.53 & & & & \\ 25.22 & 57.81 & & & \\ 12.39 & 11.88 & 36.09 & & \\ 22.15 & 7.52 & -0.31 & 20.94 & \\ 27.97 & 48.06 & 1.41 & 16.77 & 66.21 \end{vmatrix} .$$

Druhá třída vede ke středním hodnotám $\vec{\boldsymbol{x}}_1^{\mathrm{T}}$ = [185.73, 138.73, 134.77, 76.47, 137.50] a kovarianční matici

$$S_2 = \begin{vmatrix} 74.42 & & & & \\ -9.52 & 37.35 & & & \\ 22.74 & -11.26 & 36.32 & & \\ 17.79 & 0.70 & -10.72 & 15.30 & \\ 11.13 & 9.46 & 7.20 & 8.66 & 17.96 \end{vmatrix} .$$

Koeficienty diskriminační funkce jsou vyčísleny vztahem

$$\boldsymbol{a} \;'\; \boldsymbol{S}^{\&1}(\bar{\boldsymbol{x}}_1 \;\&\; \bar{\boldsymbol{x}}_2) = [-0.09,\ 0.16,\ 0.01,\ -0.18,\ -0.18]$$

a vedou k průměrům u obou tříd: $\bar{z}_1 = -28.71$ a $\bar{z}_2 = -32.21$. Hraniční bod, dle kterého se budou nezařazené objekty třídit do první nebo druhé třídy se vyčíslí jako polosuma obou průměrů $(\bar{z}_1 + \bar{z}_2)/2 = (-28.71 + (-32.21))/2 = -30.46$.

Diskriminace: vezmeme data pro lebku prvního Tibetana z dat všech lebek a pokusíme se ji diskriminovat čili zařadit do 1. nebo 2. třídy. Vyčísleme pro ni hodnotu lineární diskriminační funkce

$$z_1 = -0.09 \times 190.5 + 0.16 \times 152.5 + 0.01 \times 145.0 - 0.18 \times 73.5 - 0.18 \times 136.5 = -29.74,$$

a protože -29.74 je menší než hraniční bod -30.46, patří lebka prvního Tibeťana do první třídy.

Kvadratická diskriminační funkce (QDA). Jsou-li střední hodnoty dvou souborů μ_1 a μ_2 shodné, ale soubory se liší v kovariančních maticích $\boldsymbol{S}_1$ a $\boldsymbol{S}_2$, nelze použít lineární diskriminační funkci, což dokumentuje příklad

$$\text{Soubor } G_1\text{:}\quad \boldsymbol{\mu}_1^{\mathrm{T}} = [0,\ 0], \qquad \boldsymbol{S}_1 \;'\; \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix},$$

$$\text{Soubor } G_2\text{:}\quad \boldsymbol{\mu}_2^{\mathrm{T}} = [0,\ 0], \qquad \boldsymbol{S}_2 \;'\; \begin{pmatrix} 4.0 & 0.0 \\ 0.0 & 4.0 \end{pmatrix}.$$

Užije se kvadratická diskriminační funkce. Objekt o vektoru $\boldsymbol{x}$ bude patřit do třídy G_1, když bude splněna nerovnost

$$\boldsymbol{\mu}_1^{T}(\boldsymbol{S}_2^{\&1} \;\&\; \boldsymbol{S}_1^{\&1})\boldsymbol{x} \;\&\; 2\boldsymbol{x}^{T}(\boldsymbol{S}_2^{\&1}\,\boldsymbol{\mu}_2 \;\&\; \boldsymbol{S}_1^{\&1}\,\boldsymbol{\mu}_1) \;\%$$
$$\%\ (\boldsymbol{\mu}_2^{T}\,\boldsymbol{S}_2^{\&1}\,\boldsymbol{\mu}_2 \;\&\; \boldsymbol{\mu}_1^{T}\,\boldsymbol{S}_1^{\&1}\,\boldsymbol{\mu}_1) \;\$\; \ln\frac{{}^{*}\boldsymbol{S}_1{}^{*}}{{}^{*}\boldsymbol{S}_2{}^{*}} \;\%\; 2\,\ln\frac{\pi_1}{\pi_2},$$

kde $\boldsymbol{S}_1$ a $\boldsymbol{S}_2$ jsou kovarianční matice pro 1. a 2. třídu, G_1 a G_2.

Diskriminace mezi více než dvěma třídami. Pro tři třídy budou tři lineární diskriminační funkce nabývat následujících tvarů:

$$h_{12} \;'\; (\bar{\boldsymbol{x}}_1 \;\&\; \bar{\boldsymbol{x}}_2)^{T}\,\boldsymbol{S}^{\&1}\left[\boldsymbol{x} \;\&\; \frac{\bar{\boldsymbol{x}}_1 \;\%\; \bar{\boldsymbol{x}}_2}{2}\right],$$

$$h_{13} \;'\; (\bar{\boldsymbol{x}}_1 \;\&\; \bar{\boldsymbol{x}}_3)^{T}\,\boldsymbol{S}^{\&1}\left[\boldsymbol{x} \;\&\; \frac{\bar{\boldsymbol{x}}_1 \;\%\; \bar{\boldsymbol{x}}_3}{2}\right],$$

$$h_{23} \;'\; (\bar{\boldsymbol{x}}_2 \;\&\; \bar{\boldsymbol{x}}_3)^{T}\,\boldsymbol{S}^{\&1}\left[\boldsymbol{x} \;\&\; \frac{\bar{\boldsymbol{x}}_2 \;\%\; \bar{\boldsymbol{x}}_3}{2}\right].$$

kde $\boldsymbol{S}$ je vážená kovarianční matice všech tříd. Klasifikační pravidla zařazení objektu do dotyčné třídy jsou

umístění objektu do první třídy G_1 nastane, když $h_{12}(\boldsymbol{x}) > 0$ a $h_{13}(\boldsymbol{x}) > 0$,
umístění objektu do druhé třídy G_2 nastane, když $h_{12}(\boldsymbol{x}) < 0$ a $h_{23}(\boldsymbol{x}) > 0$,
umístění objektu do třetí třídy G_3 nastane, když $h_{13}(\boldsymbol{x}) > 0$ a $h_{23}(\boldsymbol{x}) < 0$.

Kvalita zařazení objektů do tříd (diskriminace). Předpokládejme, že máme data o K třídách s N_k , $k = 1, ..., K,$ objekty v každé třídě, N představuje celkový počet objektů (např. $N = N_1 + N_2 + N_3 = 150$). Každý objekt je popsán p diskriminátory. Každý i-tý objekt je prezentován prvkem x_{ki}. Nechť $\bar{\boldsymbol{x}}$ představuje vektor průměrů těchto diskriminátorů ve všech třídách a $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$ pak vektor průměrů objektů v k-té třídě. Definujme sumy čtverců $\boldsymbol{S_T}$, $\boldsymbol{S_W}$, $\boldsymbol{S_B}$ odchylek od středních hodnot vztahy

$$\boldsymbol{S_T} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} (x_{ki} - \bar{x})(x_{ki} - \bar{x})^T ,$$

$$\boldsymbol{S_W} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} (x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^T ,$$

$$\boldsymbol{S_B} = \boldsymbol{S_T} - \boldsymbol{S_W}$$

a definujme stupně volnosti, *df1* a *df2,* vztahy $df1 = K - 1$ a $df2 = N - K$. Diskriminační funkce je váženým průměrem hodnot nezávisle proměnných. Váhy jsou přitom voleny tak, že výsledný vážený průměr rozděluje objekty do tříd. Vysoké hodnoty průměru pocházejí z jedné třídy, nízké hodnoty průměru pocházejí z jiné třídy. Problém spočívá v nalezení vah tak, aby dobře diskriminovaly objekty do tříd. Řešení spočívá v nalezení vlastních vektorů $\boldsymbol{V}$ matice $\boldsymbol{S}_W^{-1}\,\boldsymbol{S}_B$. Kanonické koeficienty jsou totiž prvky těchto vlastních vektorů. *Mírou těsnosti proložení* je potom Wilkovo kritérium λ, definované vztahem

$$\lambda = \frac{|\boldsymbol{S_W}|}{|\boldsymbol{S_T}|} = \prod_{j=1}^{m} \frac{1}{1+\lambda_j} ,$$

kde λ_j je j-té vlastní číslo, odpovídající vlastnímu vektoru, a m je minimum ze dvou čísel, K-*1* a p.

Kanonická korelace mezi j-tou diskriminační funkcí a nezávisle proměnnými čili diskriminátory je vztažena k těmto vlastním číslům

$$r_{cj} = \sqrt{\frac{\lambda_j}{1 + \lambda_j}} .$$

Řada rozličných matic potřebných v diskriminační analýze je definována vztahy:

celková kovarianční matice $\boldsymbol{T} = \frac{1}{N-1}\,\boldsymbol{S_T}$,

kovarianční matice uvnitř tříd $\boldsymbol{W} = \frac{1}{N-K}\,\boldsymbol{S_W}$,

kovarianční matice mezi třídami $\boldsymbol{B} = \frac{1}{K-1}\,\boldsymbol{S_B}$,

lineární diskriminační funkce $z_k = \boldsymbol{W}^{-1}\,\bar{\boldsymbol{x}}_{\boldsymbol{k}}$,

standardizované kanonické koeficienty $v_{ij}\,\sqrt{w_{ij}}$,

kde v_{ji} jsou prvky $\boldsymbol{V}$ a w_{ij} prvky matice $\boldsymbol{W}$. Korelace mezi nezávisle proměnnými a kanonickými proměnnými jsou dány vztahem

$$Cor_{jk} = \frac{1}{\sqrt{w_{jj}}} \sum_{i=1}^{p} v_{ik}\, w_{ji} \quad .$$

Logistická diskriminace. Fisherova lineární diskriminace je optimální, když dva soubory mají vícerozměrné normální rozdělení se stejnými kovariančními maticemi. Tato diskriminační funkce se jeví také dostatečně robustní na odchylky od normality. Existuje však řada případů silné nenormality, např. přítomnost binárních proměnných. Pak je možné užít logistický model k výpočtu pravděpodobnosti, že objekt je členem dotyčné třídy:

$$P(G_1^*\boldsymbol{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} \quad ,$$

$$P(G_2^*\boldsymbol{x}) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} \quad .$$

Neznámé parametry $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ jsou odhadovány na základě maximální věrohodnosti. Důležité je, že odhad je zcela nezávislý na funkci hustoty třídní pravděpodobnosti. Po vyčíslení odhadů b_0 , b_1 , b_2 , ..., b_p neznámých parametrů $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ se uplatní klasifikační pravidlo zařazení objektu do třídy G_1, platí-li

$$b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p > 0,$$

což odpovídá pravděpodobnosti $P(G_1^*\boldsymbol{x}) > P(G_2^*\boldsymbol{x})$.

Vzorová úloha 4.8 *Užití logistické diskriminace*

Logistickou diskriminaci budeme demonstrovat na **Úloze B4.12** *Aplikace logistické diskriminační analýzy u rakoviny prostaty*. Režim léčení je závislý na rozšíření rakoviny na lymfatické uzliny. Rozhodující metodou vyšetření je laparotomie, vyjádřená proměnnou *B412x6*: je-li výsledek laparotomického vyšetření 0 negativní výsledek a je-li roven 1 pozitivní výsledek nodálního rozšíření rakoviny. Brownův postup následujícího vyšetření pěti diskriminantů u 53 pacientů by měl do jisté míry nahradit právě toto obtížnější laparotomické vyšetření. Brown ve své studii použil databázi: *i* je index pacienta, *B412x1* věk pacienta, *B412x2* hladina sérové kyselé fosfatázy v Kingových-Armstrongových jednotkách, *B412x3* výsledek roentgenového vyšetření (0 = negativní, 1 = pozitivní), *B412x4* velikost tumoru rektálním vyšetřením (0 = malý, 1 = velký), *B412x5* závěr pathologického bodování z biopsie (0 = méně vážný, 1 = velmi vážný).

Diskriminace: odhady parametrů (včetně svých směrodatných odchylek v závorce) k vyčíslení logistické diskriminační funkce jsou b_0 1.52 (3.56), b_1 0.10 (0.06), b_2 2.64 (1.33), b_3 1.68 (0.80), b_4 2.04 (0.83), b_5 0.35 (0.80). Tyto odhady vedou k formulaci

klasifikačního pravidla, zda má pacient rakovinu lymfatických uzlin či ne. Pacient rakovinu lymfatických uzlin nemá a je diskriminován do 1. třídy, je-li splněna nerovnost

$$1.52 - 0.10\,x_1 + 2.64\,x_2 + 1.68\,x_3 + 2.04\,x_4 + 0.35\,x_5 > 0.$$

Není-li splněna tato nerovnost, je pacient diskriminován do 2. třídy s rakovinou lymfatických uzlin. Dosadíme-li do této nerovnosti hodnoty prvního pacienta z databáze, dostaneme

$$1.52 - 0.10 \times 66 + 2.64 \times 0.48 + 1.68 \times 0 + 2.04 \times 0 + 0.35 \times 0 = -3.81.$$

Protože výsledek -3.81 není větší než nula, je pacient diskriminován do 1. třídy bez rakoviny lymfatických uzlin, což potvrdilo konečně i laparotomické vyšetření.

Posouzení správnosti diskriminace: po aplikaci diskriminační funkce k zařazení objektů do tříd je třeba posoudit správnost diskriminace. Aplikaci diskriminace na data objektů vyhodnotíme jejich chybné zařazení do tříd:

(a) Křížová tabulka diskriminace. Ukážeme křížovou tabulku zařazených objektů na konkrétním příkladu, například databáze lebek Tibeťanů. Sestavíme křížovou tabulku původního (správného) umístění objektů (lebek) do tříd a nalezeného zařazení do tříd diskriminací. Výsledkem bude *tabulka správnosti klasifikace* diskriminační analýzou, kde nesprávné zařazení je zvýrazněno tučným písmem:

		Známo (správné třídy)	
		1	2
Nalezeno diskriminací	1	14	**3**
	2	**3**	12

Nesprávného umístění je 100 % . 6/32 = 19 %. Výhodou této techniky je právě její jednoduchost, nevýhodou příliš optimistické závěry, ke kterým většinou metoda dospěje.

(b) Postupné vypouštění "vždy jednoho objektu". Spolehlivější výsledky přináší modifikace předešlého způsobu. Vytvoříme primární třídy pro n - 1 objektů a vyšetřujeme zařazení jediného dosud nezařazeného objektu. Postup n-krát opakujeme tak, že postupně vyšetřujeme zařazení všech objektů testovaného souboru. Užijeme-li i zde databáze lebek Tibeťanů, obdržíme tabulku správnosti klasifikace diskriminační analýzou, kde nesprávné zařazení je zvýrazněno tučným písmem:

		Známo (správné třídy)	
		1	2
Nalezeno diskriminací	1	12	**5**
	2	**6**	9

Nesprávného umístění je 100 % . 11/32 = 34 %, což je téměř dvojnásobek než u předešlé příliš optimistické metody.

Volba proměnných: otázkou v diskriminační analýze je, zda volba proměnných je schopna provést zařazení objektů do tříd čili diskriminaci. Byla navržena řada postupů jak provést volbu těch nejúčinnějších proměnných. Principem většiny metod je zajištění dostatečné separability tříd a volba takových proměnných, které vedou k maximalizaci

nějaké míry. Jindy se volí postup, který začne se všemi původními proměnnými a postupně se vypouštějí takové, které vedou k nedostatečné redukci separace.

K ilustraci užijeme databáze lebek Tibeťanů z **úlohy B4.14** *Aglomerativní hierarchické shlukování při analýze lebek Tibeťanů*. Užijeme pouze jednu proměnnou, *B414x4* výšku horní části obličeje [mm]. Dostaneme velmi jednoduché klasifikační pravidlo: lebka bude zařazena do 1. třídy tehdy, když výška horní části obličeje bude menší než 73.14 mm. Optimistický odhad chybné klasifikace je 25%.

Krokový postup u logistické diskriminace **úlohy B4.12** *Aplikace logistické diskriminační analýzy u rakoviny prostaty* vede k volbě tří nejúčinnějších proměnných: *B412x2* hladina sérové kyselé fosfatázy v Kingových-Armstrongových jednotkách, *B412x3* výsledek roentgenového vyšetření (0 = negativní, 1 = pozitivní), *B412x4* velikost tumoru rektálním vyšetřením (0 = malý, 1 = velký).

Postup klasifikace diskriminační analýzou

1. *Bodové odhady parametrů polohy a rozptýlení všech diskriminátorů:* vyčíslí se (a) aritmetické průměry ve třídách, (b) směrodatné odchylky ve třídách, (c) celková korelační a kovarianční matice všech diskriminátorů, (d) mezitřídní korelace a kovariance za použití průměrů místo hodnot objektů, (e) vnitrotřídní korelace a kovariance za použití dat, ve kterých byly odečteny průměry tříd a provede se zhodnocení dosažených výsledků.
2. *Vyšetření vlivu jednotlivých diskriminátorů:* vliv jednotlivých diskriminátorů na výsledky diskriminační analýzy se sleduje pomocí testačních statistik při odstranění odpovídajícího diskriminátoru.
3. *Odhady neznámých parametrů b_0 , b_1 , ..., b_p lineární diskriminační funkce pro každou třídu:* odhady neznámých parametrů b_0 , b_1 , ..., b_p jsou mezivýpočtem k vyčíslení diskriminačního skóre.
4. *Odhady regresních parametrů b_0 , b_1 , ... , b_p lineárního regresní modelu pro každou třídu*: těmito regresními parametry predikované hodnoty budou ležet mezi nulou a jedničkou. Zařazení se provede na základě třídy s nejvyšším skóre, blízkým jedničce.
5. *Klasifikace objektů diskriminační funkcí (diskriminace do tříd):* provede se (a) vyčíslení klasifikačních počtů objektů v jednotlivých třídách po diskriminaci do tříd, (b) přehled chybně klasifikovaných objektů tak, že vedle skutečné třídy je predikovaná třída a procento pravděpodobnosti výskytu objektu v predikované třídě, (c) přehled klasifikovaných objektů - skutečná (primární) třída, predikovaná třída všech objektů a procento pravděpodobnosti výskytu objektu v predikované třídě.
6. *Kanonická korelační analýza:* (a) analýza kanonických proměnných: první soubor obsahuje diskriminátory a druhý soubor třídní proměnné, (b) odhady parametrů u kanonických proměnných, (c) kanonické proměnné u třídních průměrů, (d) standardizované kanonické koeficienty slouží k výpočtu kanonického skóre, což jsou vážené průměry objektů, (e) korelace původních a kanonických proměnných představuje zátěže (korelace) původních proměnných na kanonické proměnné. Tím se usnadní vysvětlení dotyčné kanonické proměnné.
7. *Lineární diskriminační skóre všech objektů:* jsou vyčísleny hodnoty predikovaných skóre lineárních diskriminačních proměnných pro všechny objekty.

8. *Regresní skóre všech objektů:* hodnoty predikovaných skóre regresních proměnných pro všechny objekty jsou založeny na regresních koeficientech.
9. *Kanonické skóre:* hodnoty predikovaných skóre kanonických proměnných pro všechny objekty jsou založeny na kanonických koeficientech.
10. *Volba proměnných:* z velké palety diskriminátorů se vybírají pouze ty, které jsou dostatečně účinné, maximálně 8 proměnných. Výběr se provádí krokově: k nejlepšímu diskriminátoru se nalezne druhý nejlepší tak, že se prověří zda diskriminace bude tak dokonalá, jako když byl jeden diskriminátor odebrán. U nové proměnné se ověřuje, zda její F má hodnotu pravděpodobnosti menší než $\alpha = 0.05$.
11. *Výklad grafů:* výsledkem diskriminační analýzy je grafické zařazení do tříd. Zobrazení se provede na třech grafech: (a) zobrazení lineárních diskriminačních skóre, (b) zobrazení regresního skóre, a (c) zobrazení kanonického skóre.

Vzorová úloha 4.9 *Užití postupu diskriminační analýzy*

V úloze **S2.18** *Fisherova úloha rozměrů okvětních lístků u 150 kosatců* analyzujte předložený výběr kosatců, obsahujících čtvero popisných rozměrů okvětních lístků (čili diskriminátorů) u 150 květů kosatců (čili objektů), pocházejících ze tří základních tříd: (1) Iris setosa, (2) Iris versicolor, (3) Iris virginica. Z botaniky je známo, že druh Iris versicolor je hybridem zbývajících dvou druhů. Iris setosa je diploidní květ s 38 chromozomy, Iris virginica je tetraploidní a Iris versicolor je hexaploidní s 108 chromozomy. Květy kosatců jsou popsány čtyřmi diskriminátory: délkou kališních lístků v mm anglicky *lsepal,* šířkou *wsepal,* dále délkou korunních plátků v mm *lpetal* a šířkou *wpetal*. Budeme proto formulovat úlohu: jsou dána data o K třídách, $K = 3$, tři druhy čili třídy kosatců: *Setosa, Versicolor* a *Virginica* s N_k, $k = 1, ..., K,$ objekty v každé třídě, pro *Setosu* $k = 1$ *je* $N_1 = 50$, pro *Versicolor* $k = 2$ *je* $N_2 = 50$ a pro *Virginica* $k = 3$ *je* $N_3 = 50$, N představuje celkový počet objektů, $N = N_1 + N_2 + N_3 = 150$. Každý objekt je popsán p diskriminátory, $p = 4$, a to *Sepal Length, Sepal Width, Petal Length, Petal Width*. Každý i-tý objekt je prezentován prvkem x_{ki}. Nechť $\bar{\boldsymbol{x}}$ představuje vektor průměrů diskriminátorů ve všech třídách dohromady a $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$ je vektor průměrů objektů v k-té třídě. Cílem diskriminační analýzy je vyšetřit a ověřit botanické třídění a odpovědět na otázku, zda botanické třídění kosatců Iris do tří tříd je správné. Nelze zařadit 150 kosatců do jiného počtu tříd?

Řešení: Výstup z bloku Discriminant Analysis (NCSS2000) pro Fisherovu úlohu:

1. **Výpočet bodových odhadů parametrů polohy a rozptýlení všech diskriminátorů:**
   (a) **Aritmetický průměr** [mm] u tříd G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica) a celkově:

	G_1	G_2	G_3	
Proměnná	**Setosa**	**Versicolor**	**Virginica**	**Celkově**
SepalLength	50.06	59.36	65.88	58.43333
SepalWidth	34.28	27.7	29.74	30.57333
PetalLength	14.62	42.6	55.52	37.58
PetalWidth	2.46	13.26	20.26	11.99333
Počet	50	50	50	150

Tabulka obsahuje průměry každého diskriminátoru, a to v každé třídě kosatců. Poslední řádek obsahuje počet objektů ve třídě. Nadpisy sloupců jsou názvy dotyčné třídy kosatců. **Celkově** znamená všechny třídy dohromady.

(b) **Směrodatné odchylky** [mm] u tříd G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica) a celkově:

	G_1	G_2	G_3	
Proměnná	**Setosa**	**Versicolor**	**Virginica**	**Celkově**
SepalLength	3.524897	5.161712	6.358796	8.280662
SepalWidth	3.790644	3.137983	3.224966	4.358663
PetalLength	1.73664	4.69911	5.518947	17.65298
PetalWidth	1.053856	1.977527	2.7465	7.622377
Počet	50	50	50	150

Tabulka obsahuje směrodatné odchylky každého diskriminátoru, a to v každé třídě kosatců. Poslední řádek obsahuje počet objektů ve třídě. Nadpisy sloupců jsou názvy dotyčné třídy kosatců. **Celkově** znamená všechny třídy dohromady. Diskriminační analýza je postavena na předpokladu, že kovarianční matice jsou stejné pro každou třídu. Tato tabulka umožňuje posoudit předpoklad, zda totiž jsou směrodatné odchylky ve třídách zhruba stejné.

(c) **Celkové korelace a kovariance:**

	Proměnná			
Proměnná	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength	68.56935	-4.243401	127.4315	51.62707
SepalWidth	-0.117570	18.99794	-32.96564	-12.16394
PetalLength	0.871754	-0.428440	311.6278	129.5609
PetalWidth	0.817941	-0.366126	0.962865	58.10063

Tabulka obsahuje korelace a kovariance, vytvořené když jsou ignorovány smíšené proměnné *diskriminátorů.* Korelace jsou v dolní levé části, kovariance jsou v pravé horní části matice. Rozptyly jsou na diagonále matice.

(d) **Mezitřídní korelace a kovariance:**

	Proměnná			
Proměnná	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength	3160.607	-997.6334	8262.42	3563.967
SepalWidth	-0.745075	567.2466	-2861.98	-1146.633
PetalLength	0.994135	-0.812838	21855.14	9338.7
PetalWidth	0.999768	-0.759258	0.996232	4020.667

Tabulka obsahuje korelace a kovariance, vytvořené za použití průměrů místo jednotlivých objektů. Korelace jsou v dolní levé části, mezitřídní kovariance jsou na diagonále matice a v horní pravé části matice. Všimněte si, že když by byly jenom dvě třídy kosatců, všechny korelace by byly rovny jedné, protože byly vytvořeny pouze ze dvou řádků, totiž ze dvou třídních průměrů.

(e) **Vnitrotřídní korelace a kovariance:**

	Proměnná			
Proměnná	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
SepalLength	26.50082	9.272109	16.75143	3.840136
SepalWidth	0.530236	11.53878	5.524354	3.27102
PetalLength	0.756164	0.377916	18.51878	4.266531
PetalWidth	0.364506	0.470535	0.484459	4.188163

Tabulka obsahuje korelace a kovariance, vytvořené z dat, ve kterých byly třídní průměry odečteny. Korelace jsou v dolní levé části, vnitrotřídní kovariance jsou na diagonále a v pravé horní části matice.

2. Vyšetření vlivu jednotlivých diskriminátorů:

	Při odstranění této proměnné			**Pro tuto samotnou proměnnou**			R^2
Proměnná	**Lambda**	***F*-test**	**Spočtená α**	**Lambda**	***F*-test**	**Spočtená α**	**ostatní *X***
SepalLength	0.938463	4.72	0.010329	0.381294	119.26	0.000000	0.858612

SepalWidth	0.766480	21.94	0.000000	0.599217	49.16	0.000000	0.524007
PetalLength	0.669206	35.59	0.000000	0.058628	1180.2	0.000000	0.968012
PetalWidth	0.743001	24.90	0.000000	0.071117	960.01	0.000000	0.937850

Tabulka ukazuje na vliv jednotlivých diskriminátorů proměnných na výsledky diskriminační analýzy. **Proměnná:** jméno diskriminátoru. **Lambda při odstranění této proměnné:** hodnota Wilkova lambda, vypočtená k testování důsledku odstranění této diskriminační proměnné. ***F*-test při odstranění této proměnné:** hodnota *F*-kritéria, vyčísleného k testování statistické významnosti Wilkova lambda. **Spočtená hladina významnosti při odstranění této proměnné:** vypočtená hladina významnosti výše uvedeného *F*-testu při odstranění této diskriminační proměnné. Test je totiž statisticky významný a diskriminátor je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$. **Lambda pro tuto samotnou proměnnou:** jde o hodnotu Wilkova lambda, kterou dostaneme za použití této jediné nezávisle proměnné. ***F*-test pro tuto samotnou proměnnou:** jde o testační kritérium, vyčíslené k testování statistické významnosti Wilkova lambda. **Spočtená hladina významnosti pro tuto samotnou proměnnou:** uvedený *F*-test je statisticky významný a diskriminátor je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$.

3. **Odhady neznámých parametrů b_0 , b_1 , ..., b_p lineární diskriminační funkce pro každou třídu** G_1 *(Setosa),* G_2 *(Versicolor),* G_3 *(Virginica)*:

	G_1	G_2	G_3
Proměnná	**Setosa**	**Versicolor**	**Virginica**
Absolutní člen	-85.20985	-71.754	-103.2697
SepalLength	2.354417	1.569821	1.244585
SepalWidth	2.358787	0.707251	0.3685279
PetalLength	-1.643064	0.5211451	1.276654
PetalWidth	-1.739841	0.6434229	2.107911

Tabulka obsahuje odhady neznámých parametrů b_0 , b_1 , ..., b_p lineární diskriminační funkce. Tyto parametry jsou také nazývány diskriminačními koeficienty. Technika předpokládá, že diskriminátory v každé třídě kosatců vykazují vícerozměrné normální rozdělení se shodnými variančně-kovariančními maticemi ve třídách. Technika je dostatečně robustní i při nesplnění těchto předpokladů. Tabulka obsahuje celkem tři klasifikační funkce, jednu pro každou třídu. Každá funkce je prezentována vertikálně hodnotami ve sloupci. Když vytvoříme vážený průměr diskriminátorů užitím těchto koeficientů jako vah (a přidáním konstanty jako absolutního členu), dostaneme diskriminační skóre.

4. **Odhady regresních parametrů b_0 , b_1 , ..., b_p lineárního regresní modelu pro každou třídu** G_1 *(Setosa),* G_2 *(Versicolor),* G_3 *(Virginica)*:

	G_1	G_2	G_3
Proměnná	**Setosa**	**Versicolor**	**Virginica**
Absolutní člen	0.1182229	1.577059	-0.6952819
SepalLength	6.602977E-03	-2.015369E-03	-4.587608E-03
SepalWidth	2.428479E-02	-4.456162E-02	2.027684E-02
PetalLength	-2.246571E-02	2.206692E-02	3.987911E-04
PetalWidth	-5.747273E-03	-4.943066E-02	5.517793E-02

Tabulka obsahuje regresní parametry b_0 , b_1 , ..., b_p lineárního regresního modelu pro každou třídu G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica), které byly vyčísleny následujícím postupem: (1) Vytvoříme tři indikátorové proměnné, jedna je pro každou ze tří druhů kosatců (Setosa, Versicolor a Virdinica). Každá indikátorová proměnná je položena rovna jedné. (2) Proložíme vícenásobnou regresí nezávisle proměnných každý ze tří kosatců. (3) Obdržíme odhady regresních parametrů, uvedené v tabulce. Těmito regresními parametry pak predikované hodnoty budou ležet mezi nulou a jedničkou. Určení, ke které třídě jedinec patří se provede tak, že se vybere třída s nejvyšším skóre.

5. **Klasifikace objektů diskriminačního funkcí** (diskriminace objektů do tříd):
   (a) **Tabulka klasifikačních počtů pro kosatce u diskriminace do tříd** G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica) **a celkově:**

Predikovaná	G_1	G_2	G_3	
Známá	**Setosa**	**Versicolor**	**Virginica**	**Total**
Setosa	50	0	0	50
Versicolor	0	34	16	50
Virginica	0	7	43	50
Celkově	50	41	59	150
Redukce v klasifikační správnosti v důsledku proměnných $\boldsymbol{X}$ = 77.0 %.				

Tabulka ukazuje, jak navržené diskriminační funkce klasifikují objekty v datech. Bylo-li dosaženo perfektní klasifikace, obdržíme v matici mimo diagonálu nuly. Řádky tabulky představují aktuální třídy kosatců, zatímco sloupce představují predikované třídy kosatců. **Redukce v klasifikační správnosti:** obsahuje procento redukce v klasifikační správnosti, dosažené diskriminačními funkcemi vůči očekávané hodnotě, když byly objekty klasifikovány náhodně.

(b) **Přehled chybně klasifikovaných objektů v řádcích u diskriminace do tříd** G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica):

			Procento zařazení do jednotlivé třídy		
Řádek	**Známá**	**Predikovaná**	**Třída 1**	**Třída 2**	**Třída 3**
5	Virginica	Versicolo	-1.8	58.6	43.1
9	Versicolo	Virginica	10.3	20.2	69.5
22	Versicolo	Virginica	18.8	22.6	58.6
28	Versicolo	Virginica	22.1	35.5	42.4
29	Versicolo	Virginica	22.1	27.4	50.6
38	Versicolo	Virginica	10.6	38.3	51.1
45	Virginica	Versicolo	-31.4	66.4	65.0
57	Virginica	Versicolo	-18.6	83.9	34.7
62	Versicolo	Virginica	24.4	34.0	41.6
66	Versicolo	Virginica	11.9	37.9	50.2
70	Versicolo	Virginica	12.1	41.5	46.3
78	Virginica	Versicolo	-7.3	58.4	48.9
91	Virginica	Versicolo	-16.1	83.8	32.3
95	Versicolo	Virginica	23.7	14.3	62.0
106	Versicolo	Virginica	20.7	30.7	48.7
111	Virginica	Versicolo	-21.4	63.8	57.6
112	Virginica	Versicolo	-23.9	71.8	52.1
114	Versicolo	Virginica	17.1	35.6	47.2
117	Versicolo	Virginica	22.1	38.9	39.0
130	Versicolo	Virginica	30.9	32.4	36.8
131	Versicolo	Virginica	14.0	39.6	46.4
142	Versicolo	Virginica	21.4	38.6	40.0
148	Versicolo	Virginica	6.8	36.8	56.4

V řádku se u každého chybně klasifikovaného objektu nachází vždy název známé třídy kosatců a predikované třídy kosatců. Následuje 100× zvětšená hodnota pravděpodobnosti (v procentech), že objekt se nachází v dané třídě kosatců. Procento pravděpodobnosti se jeví totiž názornější než normovaný odhad v rozmězí 0 a 1. Hodnota blízko 100 % ukazuje, že objekt patří do dotyčné třídy. ***P(i)*:** při užití lineární diskriminační techniky se vyčíslí pravděpodobnosti, že tento řádek patří do i-té třídy: nechť f_i, $i = 1, ..., K$, je hodnota lineární diskriminační funkce a $\max(f_k)$ je maximální skóre ze všech tříd. Označme $P(G_i)$ celkovou pravděpodobnost, klasifikující jednotlivce do třídy i. Hodnota $P(i)$ se vypočte dle vztahu

$$P(i) \text{ ' } \frac{\exp[f_i \text{ \& } \max(f_k)]\ P(G_i)}{\sum_{j=1}^{K} \exp[f_i \text{ \& } \max(f_k)]\ P(G_i)} \quad .$$

Když užijeme regresní klasifikační techniku, bude představovat predikovanou hodnotu regresní rovnice. Implicitně je *Y* v regresní rovnici rovno 1 nebo 0 v závislosti, zda objekt do *i*-té třídy kosatců patří či ne. Proto predikovaná hodnota blízko nuly ukazuje, že objekt nepatří do *i*-té třídy, zatímco blízko 1 ukazuje na silný důkaz, že objekt patří do *i*-té třídy. V žádném případě nemůže vyčíslena hodnota být větší než 1 a menší než 0.

(c) **Zařazení objektů predikovanou klasifikací pomocí diskriminační funkce do tříd** G_1 (Setosa), G_2 (Versicolor), G_3 (Virginica):

			Procento zařazení do jednotlivé třídy		
Řádek	**Známá**	**Predikovaná**	**Třída 1**	**Třída 2**	**Třída 3**
1	Setosa	Setosa	92.4	21.6	-14.0
2	Virginica	Virginica	-16.4	34.9	81.5
3	Versicolo	Versicolo	10.8	47.2	42.0
..	.....	.....	.....	.....	.....
..	.....	.....	.....	.....	.....
150	Setosa	Setosa	101.8	5.4	-7.2

Tabulka obsahuje pro každý objekt kosatců vždy skutečnou, čili známou třídu kosatců, predikovanou třídu kosatců a procento pravděpodobnosti zařazení do dotyčné třídy kosatců.

6. Kanonická korelační analýza:

(a) **Analýza kanonických proměnných:**

	Inv(W)B	**Ind.**	**Total**	**Kanon.**	**Kanon.**		**Čitatel**	**Jmenov.**	**Spočtená**	**Wilkovo**
Fn	**vlast.číslo**	**Pcnt**	**Pcnt**	**korel.**	**korel2**		***F*-test**	**SV**	**SV**	**α** **Lambda**
1	32.191929	99.1	99.1	0.9848	0.9699	199.1	8.0	288.0	0.0000	0.023439
2	0.285391	0.9	100.0	0.4712	0.2220	13.8	3.0	145.0	0.0000	0.777973

F-test testuje, zda tato funkce a další jsou statisticky významné.

Tabulka obsahuje výsledky kanonické korelační analýzy diskriminačního problému. U kanonické korelační analýzy jsou dva soubory proměnných, které jsou zde definovány následovně: první soubor obsahuje diskriminátory. Třídní proměnná definuje druhý, jiný soubor, který je generován vytvořením indikátorové proměnné pro každou třídu, kromě poslední. **Inv(W)B vlastn. číslo:** vlastní čísla matice $\mathbf{W^{-1}B}$ ukazují, jak mnoho je celková proměnlivost vysvětlena různými diskriminačními funkcemi. První diskriminační funkce totiž odpovídá prvnímu vlastnímu číslu, atd. Počet vlastních čísel je roven minimu počtu diskriminátorů a K-1, kde K je počet tříd kosatců. **Ind. Pcnt:** procento, jež toto vlastní číslo představuje z celku vlastních čísel. **Total Pcnt:** kumulativní procento tohoto a všech předešlých vlastních čísel. **Kanon korel.:** kanonický korelační koeficient. **Kanon korel2:** čtverec kanonického korelačního koeficientu je podobný R^2 ve vícenásobné regresi. ***F*-test:** hodnota *F*-kritéria, testujícího Wilkovo lambda, které odpovídá tomuto řádku a řádkům níže. V tomto případě testuje *F*-kritérium statistickou významnost obou, první a druhé, kanonické korelace, zatímco druhá *F*-hodnota testuje významnost pouze druhé korelace. **Čitatel SV:** počet stupňů volnosti pro čitatele v tomto *F*-testu. **Jmenov. SV:** počet stupňů volnosti pro jmenovatele v tomto *F*-testu. **Spočtená α:** spočtená hladina významnosti pro *F*-test. Je-li tato hodnota α menší než uživatelem zadané 0.05, je test statisticky významný. **Wilkovo lambda:** hodnota Wilkova lambda pro tento řádek se užívá k testování statistické významnosti diskriminační funkce, odpovídající tomuto řádku a řádkům níže. Wilkovo lambda je vícerozměrným zobecněním R^2. Uvedený *F*-test je aproximativním testem Wilkova lambda.

(b) **Odhady parametrů u kanonických proměnných:**

	Kanonická proměnná	
Proměnná	**Proměnná1**	**Proměnná2**
Absolutní člen	-2.105106	6.661473

SepalLength	-0.082938		-0.002410
SepalWidth	-0.153447	-0.216452	
PetalLength	0.220121		0.093192
PetalWidth	0.281046		-0.283919

Obsahuje koeficienty k výpočtu kanonického skóre. Kanonická skóre jsou vážené průměry objektů a tyto koefienty jsou pak váhy s přidaným absolutním členem.

(c) **Kanonické proměnné u třídních průměrů:**

	Kanonická funkce	
Iris	**Funkce 1**	**Funkce 2**
Setosa	-7.6076	-0.215133
Versicolor	1.82505	0.7278996
Virginica	5.78255	-0.5127666

Tabulka obsahuje výsledky kanonických koeficientů pro průměry u každé třídy.

(d) **Standardizované kanonické koeficienty:**

	Kanonická proměnná	
Proměnná	**Proměnná 1**	**Proměnná 2**
SepalLength	-0.426955	-0.012408
SepalWidth	-0.521242	-0.735261
PetalLength	0.947257	0.401038
PetalWidth	0.575161	-0.581040

Tabulka obsahuje standardizované kanonické koeficienty.

(e) **Korelace původních a kanonických proměnných:**

	Kanonická proměnná	
Proměnná	**Proměnná 1**	**Proměnná 2**
SepalLength	0.222596	-0.310812
SepalWidth	-0.119012	-0.863681
PetalLength	0.706065	-0.167701
PetalWidth	0.633178	-0.737242

Tabulka obsahuje zátěže (korelace) původních proměnných na kanonické proměnné. Každý výstup je korelací mezi kanonickou proměnnou a diskriminátorem. Tato tabulka usnadní interpretovat dotyčné kanonické proměnné.

7. Lineární diskriminační skóre všech objektů :

Řádek	**Iris**	**Skóre1**	**Skóre2**	**Skóre3**
1	Setosa	83.86837	38.65921	-6.790054
2	Virginica	1.230765	91.857	104.5692
..	.....	.....	.....	.....
150	Setosa	98.72371	46.71882	-0.3055334

Tabulka obsahuje jednotlivé hodnoty lineárních diskriminačních skóre pro všechny objekty, tj. pro všech 150 kosatců.

8. Regresní skóre všech objektů:

Řádek	**Iris**	**Skóre1**	**Skóre2**	**Skóre3**
1	Setosa	0.923755	0.215832	-0.139588
2	Virginica	-0.163732	0.348623	0.815109
3	Versicolo	0.107759	0.471953	0.420288

..	.....	.....	.....	.....
..	.....	.....	.....	.....
150	Setosa	1.018238	0.053607	-0.071844

Tabulka obsahuje jednotlivé hodnoty predikovaných skóre, založené na regresních koeficientech. I když tyto hodnoty jsou predikované indikátorové proměnné, může nastat případ, že hodnota bude menší než nula a větší než 1.

9. Kanonická skóre všech objektů:

Řádek	Iris	Skóre1	Skóre2
1	Setosa	-7.671967	0.134894
2	Virginica	6.800150	-0.580895
3	Versicolo	2.548678	0.472205
..	.....	.....	.....
..	.....	.....	.....
150	Setosa	-8.314449	-0.644953

Tabulka obsahuje skóre kanonických proměnných pro každý řádek u všech objektů, tj. 150 kosatců.

10. Automatická volba účinných diskriminátorů:

Dosavadní tabulky jsou postaveny na čtyřech diskriminátorech: Petal Length, Petal Width, Sepal Length a Sepal Width. Stěžejním úkolem v diskriminační analýze je však výběr diskriminátorů. Často máme velikou paletu možných diskriminátorů, ze kterých potřebujeme vybrat menší výběr, asi tak maximálně 8 účinných proměnných, který se bude chovat jako původní velký soubor.

Iterace	Činnost v kroku	Nezávisle proměnná	% změny v lambda	*F*-test	Spočtená hladina α	Wilkovo lambda
0	None					1.000000
1	Entered	PetalLength	94.14	1180.16	0.000000	0.058628
2	Entered	SepalWidth	37.09	43.04	0.000000	0.036884
3	Entered	PetalWidth	32.29	34.57	0.000000	0.024976
4	Entered	SepalLength	6.15	4.72	0.010329	0.023439
..	...	.....	...	...	.....	.....

Detail ve 4. kroku automatického výběru proměnné:

Status	Nezávisle proměnná	% změny v lambda	*F*-test	Spočtená hladina α	R^2 ostatních *X*
In	SepalLength	6.15	4.72	0.010329	0.858612
In	SepalWidth	23.35	21.94	0.000000	0.524007
In	PetalLength	33.08	35.59	0.000000	0.968012
In	PetalWidth	25.70	24.90	0.000000	0.937850

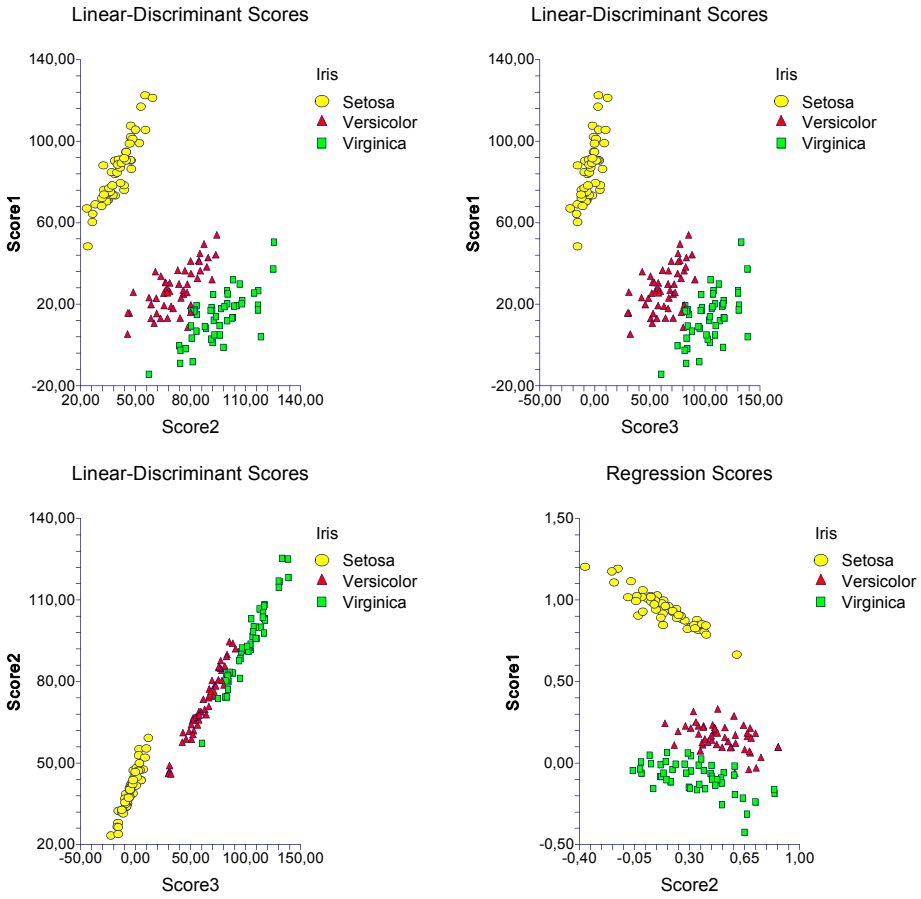
Celkové Wilkovo lambda = 0.023439

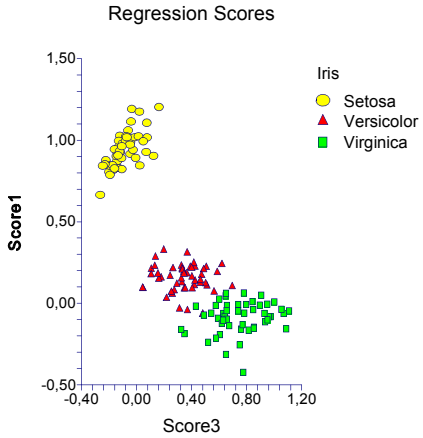
Tabulka *Automatický výběr diskriminátorů* se provádí krokově: nejprve se nalezne nejlepší diskriminátor a potom druhý nejlepší. Když byly nalezeny první dva, prověří se, zda diskriminace bude tak dokonalá, jako když byl jeden diskriminátor odebrán. Postupný (či krokový) proces přidávání nejlepšího zbývajícího diskriminátoru s následným ověřením, zda by jeden aktivní diskriminátor mohl být odebrán, pokračuje, dokud není žádný nový diskriminátor k dispozici. U nového diskriminátoru se ověřuje, zda jeho *F*-hodnota má pravděpodobnost menší než uživatelem zadaná vstupní hodnota hladiny významnosti α = 0.05. **Přehled výběru proměnných:** obsahuje protokol o činnosti v každém kroku. **Iterace:** uvádí pořadové číslo (index) kroku. **Činnost v tomto kroku:** uvádí zda diskriminátor byl zaveden do souboru aktivních diskriminátorů nebo odstraněn z tohoto souboru. **% změny v lambda:** procento snížení v hodnotě lambda, jež je výsledkem tohoto kroku. Všimněte si, že Wilkovo lambda je analogické ($1-R^2$) ve vícenásobné regresi. Abychom zlepšili model, budeme žádat snížit Wilkovo lambda. Např. od iterace 2 k iteraci 3 se lambda sníží z hodnoty 0.036884 na 0.024976. To je 32.29% snížení hodnoty lambda. ***F*-test:** jde o *F*-

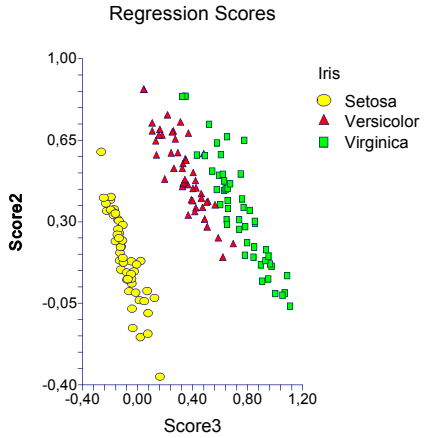
kritérium k testování statistické významnosti tohoto diskriminátoru. Je-li diskriminátor zaveden, testuje se hypotéza, že diskriminátor je třeba přidat. Je-li diskriminátor odstraněn, testuje se hypotéza, že diskriminátor je třeba odstranit. **Spočtená hladina významnosti *α*:** od uvedeného *F*-testu. **Wilkovo lambda:** víceparametrické rozšíření R^2 redukuje (1-R^2) ve dvojtřídě. Může být vysvětleno právě opačně než R^2. Mění se v intervalu od 1 do 0. Hodnoty blízko 1 vedou k nízké prediktibilitě, zatímco hodnoty blízko 0 k vysoké. Wilkovo lambda odpovídá právě aktivním diskriminátorům.

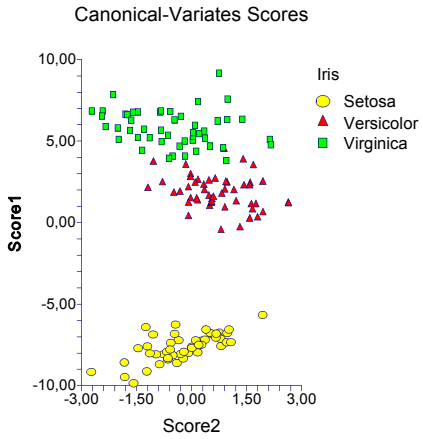
11. Výklad grafů diskriminace všech objektů do tříd:

Nabízí se několik zobrazení (a) lineárních diskriminačních skóre, (b) regresních skóre nebo (c) kanonických skóre: Na základě diagramů těchto tří druhů skóre pak snáze vytvoří svou interpretaci. Diagramy totiž poskytnou vizuální vysvětlení, jak diskriminační funkce klasifikují objekty v datech. Předložený diagram ukazuje hodnoty prvního a druhého kanonického skóre. Z grafu je patrné klasifikační pravidlo: první kanonická funkce postačuje k diskriminování mezi kosatci, protože třídy kosatců mohou být snadno odděleny vertikální osou. Existuje software (S-Plus), který umožňuje 3D zobrazení s rotací podél os v prostoru. Potom by bylo vytvoření a rozlišení tříd kosatců ještě názornější.









Obr. 4.15 Graf lineárního diskriminačního skóre (Linear Discriminant Scores - 1 vs. 2, 1 vs. 3, 2 vs. 3).

Obr. 4.16 Graf regresního skóre (Regression Scores - 1 vs. 2, 1 vs. 3, 2 vs. 3)

Obr. 4.17 Graf kanonických proměnných (Canonical Scores - 1 vs. 2).

4.6.2 Analýza shluků CLU

Analýza shluků (**Clu**ster analysis, CLU) patří mezi metody, které se zabývají vyšetřováním podobnosti *vícerozměrných objektů* (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich klasifikací do tříd čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Podle způsobu shlukování se postupy dělí na *hierarchické* a *nehierarchické*. Hierarchické se dělí dále na *aglomerativní* a *divizní.*

Hierarchické postupy jsou založeny na postupném spojování objektů a jejich shluků do dalších, větších shluků. Nejprve se vypočte základní matice vzdáleností mezi objekty. U *aglomerativního shlukování* se dva objekty, jejichž vzdálenost je nejmenší, spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku a naopak tento shluk je zařazen jako celek. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. *Divizní postup* je obrácený. Vychází se z množiny všech

objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stadiu jednotlivých objektů. Výhodou hierarchických metod je nepotřebnost informace o optimálním počtu shluků v procesu shlukování; tento počet se určuje až dodatečně. Při shlukování vznikají dva základní problémy:

(a) *způsob měření vzdáleností mezi objekty*. I když existuje celá řada měr vzdáleností (vícerozměrných metrik), nejčastěji se užívá *euklidovská metrika*, která je přirozeným zobecnením běžného pojmu vzdálenosti;

(b) *volba vhodné shlukovací procedury* dle zvoleného způsobu metriky.

Metody metriky shlukování jsou

Metoda průměrová (Average): vzdálenost dvou shluků se počítá jako průměr z možných mezishlukových vzdáleností dvou objektů, kdy se mezishlukovou vzdáleností objektů rozumí vzdálenost dvou objektů, z nichž každý patří do jiného shluku. Nejbližší jsou shluky, které mají nejmenší průměrnou vzdálenost mezi všemi objekty jednoho a všemi objekty druhého shluku. Dendrogramy mají strukturu podobnou dendrogramům metody nejvzdálenějšího souseda, pouze spojení je provedeno při obvykle vyšších vzdálenostech.

Metoda centroidní (Centroid): vzdálenost shluků se počítá jako euklidovská vzdálenost jejich těžišť. Nejbližší jsou ty shluky, které mají nejmenší vzdálenost mezi těžišti.

Metoda nejbližšího souseda (Single, Nearest): kritériem pro vytváření shluků je minimum z možných mezishlukových vzdáleností objektů. Metoda tvoří nový shluk na základě nejkratší vzdálenosti mezi shluky (či objekty) a neumí proto rozlišit špatně separované shluky. Je zde silná tendence ke tvorbě řetězců. Jsou-li objekty na opačních koncích řetězce zcela nepodobné, řetězování může vést až ke zcela mylným závěrům. Na druhé straně je to jedna z mála metod, která umí roztřídit a rozlišit i neeliptické shluky.

Metoda nejvzdálenějšího souseda (Complete, Furthest): počítá vzdálenost dvou shluků jako maximum z možných mezishlukových vzdáleností objektů. Probíhá podobně jako metoda Single s jednou důležitou výjimkou: vzdálenost (či nepodob-nost) mezi shluky je určována vzdáleností (či nepodobností) mezi dvěma nejvzdá-lenějšími objekty, každý přitom je z jiného shluku. Proto všechny objekty ve shluku jsou klasifikovány na základě maximální vzdálenosti či minimální podobnosti vůči objektům ve druhém shluku.

Metoda mediánová (Median): jde o jisté vylepšení centroidní metody, neboť se snaží odstranit rozdílné "váhy", které centroidní metoda dává různě velkým shlukům.

Wardova metoda je založena na minimalizaci ztráty informace při spojení dvou tříd. V každém kroku je uvažován takový možný pár objektů (či shluků), aby suma čtverců odchylek od střední hodnoty $ESS = \sum_{i=1}^{n} (x_i - \bar{x})^2$ dosáhla při vzniku shluku svého minima.

Nehierarchické shlukovací metody: u *metody typických bodů* (Seeded) uživatel na základě svých věcných znalostí určí, které objekty mají být "typickými" představiteli nově vytvořených shluků a systém rozdělí objekty do shluků podle jejich euklidovské vzdálenosti od těchto typických objektů. V nehierarchických shlukovacích metodách je počet shluků obvykle předem dán, i když se v průběhu výpočtu může změnit. Zůstává-li počet shluků zachován, hovoříme o nehierarchických metodách s *konstantním počtem shluků*, v opačném případě o nehierarchických metodách s *optimalizovaným počtem shluků*. Nehierarchické metody zahrnují dvě základní varianty - optimalizační metody a analýzu módů, medoidů. *Optimalizační nehierarchické metody* hledají optimální rozklad přeřazováním objektů ze shluku do shluku s cílem minimalizovat nebo maximalizovat nějakou charakteristiku rozkladu. Metody, označované jako *analýza módů, medoidů*,

představují hledání rozkladu do shluků, kde shluky jsou chápány jako místa se zvýšenou koncentrací objektů v *m*-rozměrném prostoru proměnných.

Místo výchozí matice vzdáleností může být v některých případech ke shlukování použita i *korelační matice.*

(a) Hierarchické shlukování

Analýza shluků patří mezi metody, zabývající se vyšetřováním podobnosti vícerozměrných objektů (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich roztříděním do tříd čili *shluků.* Hodí se zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Analýzou shluků budeme sledovat a vyšetřovat jednak podobnost objektů, analyzovanou pomocí *dendrogramu objektů*, a jednak podobnost proměnných analyzovanou pomocí *dendrogramu proměnných.*

Dendrogram, diagram shluků nebo vývojový strom se objeví pouze v případě zadání hodnot původních proměnných a nikoli při zadání maticí vzdáleností. Výsledkem je zobrazení hodnot ve dvojrozměrném prostoru, kde osy tvoří zadané proměnné. Objeví se také "obkroužení" objektů v jednotlivých shlucích.

Dendrogram podobnosti objektů je standardní výstup hierarchických shlukovacích metod, ze kterého je patrná struktura objektů ve shlucích.

Dendrogram podobnosti proměnných odhaluje nejčastěji dvojice či trojice (obecně *m*-tice) proměnných, které jsou si velmi podobné a silně spolu korelují. Odhaluje proměnné, které jsou ve společném shluku, které jsou si tím pádem značně podobné a které jsou také vzájemně nahraditelné. To má značný význam při plánování experimentu a respektování úsporných ekonomických kritérií. Některé vlastnosti (či proměnné) není třeba vůbec měřit, protože jsou snadno nahraditelné jinými a nepřispívají do celku velkou vypovídací schopností.

Míra věrohodnosti: dendrogram lze sestrojit celou řadou technik. Prvním kritériem věrohodnosti čili těsnosti proložení při volbě "nejlepšího dendrogramu", jež nejlépe odpovídá struktuře objektů a proměnných mezi objekty, je *kofenetický korelační koeficient CC*. Je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu. Je-li tato hodnota větší než 0.75, je obvykle nulová hypotéza o dané struktuře zamítnuta. Hodnota 0.9 svědčí, že dendrogram vůbec neodpovídá skutečné struktuře dat.

Druhým kritériem těsnosti proložení je *kritérium delta* Δ, které měří stupeň přetvoření, distorze spíše než stupeň podobnosti. Kritérium delta je definováno vztahem

$$\Delta_A \;'\; \left[\frac{\sum_{j<k}^{N} *d_{jk} \;\&\; d_{jk}^{(*1/A}}{\sum_{j<k}^{N} (d_{jk}^{(})^{1/A}} \right]^A ,$$

kde $A = 0.5$ nebo 1 a d^*_{ij} je vzdálenost získaná z dendrogramu. Jsou žádoucí hodnoty *delta* blízké nule. Řada autorů ukázala, že metoda průměrová vede obvykle k nejlepšímu dendrogramu.

Postup shlukové analýzy

1. *Volba vstupní databáze*: zadává se typ dat (a) proměnných (sloupců) analyzovaných objektů (řádků), (b) sloupců matice vzdáleností, (c) sloupců korelační matice.
2. *Volba druhu veličin:* zadává se typ užitých veličin v datech, která mohou být (a) intervalová, (b) ordinální, (c) nominální, (d) symetrická binární, (e) asymetrická binární, (f) poměrová.
3. *Název objektů:* zadání pojmenování či jmen jednotlivých objektů, umístěných v řádcích, které se mohou objevit v dendrogramu místo indexů (pořadových čísel) objektů.
4. *Typ shlukovací techniky:* volba metody z možností: jednoduchá průměrová (Average), skupinového průměru, centroidní (Centroid), nejbližšího souseda (Single, Nearest), nejvzdálenějšího souseda (Complete, Furthest), mediánová (Median), Wardova, a flexibilní.
5. *Volí se druh užité vzdálenosti:* vzdálenosti mohou být Eukleidova metrika čili geometrická vzdálenost, Hammingova metrika čili Manhattanská vzdálenost, zobecněná Minkowskiho metrika a Mahalanobisova metrika.
6. *Postup linkování a zařazení do shluků:* tabelární výpočet vzdáleností (nebo podobností) mezi objekty a shluky a postupné vytváření dendrogramu. Postupy jsou (1) metodou hierarchického shlukování, (2) shlukování metodou nejbližších středů, (3) shlukování metodou středů-medoidů, a (4) metodou fuzzy shlukování.
7. *Výpočet skutečných a predikovaných vzdáleností v dendrogramu:* jsou porovnány skutečné vzdálenosti mezi objekty a vypočtené vzdálenosti (predikované) v dendrogramu, jejich rozdíl a konečně i procentuální vyjádření tohoto rozdílu.
8. *Hledání nejlepší techniky tvorby dendrogramu:* dle bodu 4. a 5. lze k sestrojení optimálního dendrogramu kombinovat řadu technik. Rozhodčím kritériem věrohodnosti jsou především kofenetický korelační koeficient *CC*, obě míry těsnosti proložení *delta*, ale také další kritéria: mezishluková suma čtverců WSS_K, procento variace PV_K, silueta *s*, průměrná silueta *SC*, Wilkova statistika λ, rozdělovací koeficienty Dunnův *F(U)* a Kaufmanův *D(U)*.
9. *Vysvětlení nejlepšího dendrogramu podobností objektů:* interpretace optimálního dendrogramu podobnosti jednotlivých objektů je prvním a nejdůležitějším cílem shlukové analýzy.
10. *Vysvětlení nejlepšího dendrogramu podobností proměnných:* interpretace optimál-ního dendrogramu podobnosti jednotlivých proměnných odhalí souvislosti ve struktuře objektů analyzované databáze a je druhým důležitým cílem shlukové analýzy.

Vzorová úloha 4.10 *Nalezení shluků hráčů podobných vlastností*

Použijeme dat úlohy **S4.21** *Shluky 12 superhvězd košíkové*. Následující tabulka dat obsahuje informace o osmi hráčských vlastnostech a aktivitách 12 superhvězd košíkové v sezóně 1989. Cílem je najít shluky hráčů podobných vlastností, a naopak, odhalit jejich aktivity a vlastnost, ve které se hráč neshoduje s ostatními hráči.

Řešení: Výklad výstupu programu NCSS2000:

1. Titulní stránka:

Shlukovací metoda:	Průměrová
Typ vzdálenosti:	Eukleidovská
Typ škály:	Směrodatná odchylka

Titulní stránka popisuje techniku shlukování a typ vzdálenosti.

2. Tvorba shluků a zařazení objektů do shluků:

Shluk řádek	Shluk	Oddíl Hráč
2	1	Barry R.
3	1	Baylor E.
4	1	Bird L.
7	1	Erving J.
9	1	Jordan M.
10	1	Robertson O.
12	1	West J.
1		Jabbar K. A.
5		Chamberlain W.
6		Cousy B.
8		Johnson M.
11		Russell B.

Každý řádek je spojen s číslem shluku. Tabulka je uspořádána dle čísel shluků. Čísla shluků, která nemohou být klasifikována, jsou ponechána prázdná. Sestava shluku je závislá na výseči shluků, která byla zadána v datech.

3. Tabulka linkování shluků v dendrogramu:

Spoj	Počet shluků	Velikost vzdálenosti	Diagram vzdálenosti	Propojené řádky
11	1	2.054454	IIIIIIIIIIIIIIIIIIIIIIIIIIIII	1,2,4,7,3,9,10,12,8,6,11,5
10	2	1.609330	IIIIIIIIIIIIIIIIIIIIII	1,2,4,7,3,9,10,12,8,6,11
9	3	1.458223	IIIIIIIIIIIIIIIIII	1,2,4,7,3,9,10,12,8,6
8	4	1.282432	IIIIIIIIIIIIIIII	1,2,4,7,3,9,10,12,8
7	5	1.210352	IIIIIIIIIIIIIII	1,2,4,7,3,9,10,12
6	6	0.863494	IIIIIIIIIII	2,4,7,3,9,10,12
5	7	0.842134	IIIIIIIIII	9,10,12
4	8	0.795145	IIIIIIIIII	2,4,7,3
3	9	0.684606	IIIIIIII	10,12
2	10	0.614431	IIIIIII	2,4,7
1	11	0.565813	IIIIIII	4,7
Kofenetická korelace:			0.891119	
Delta(0.5):			0.122037	
Delta(1.0):			0.160527	

Průběh spojování do shluků je zobrazen v opačném pořadí. Uvádí se také hladina vzdálenosti, na které dochází ke shlukování. To umožňuje vybrat si vhodnou hladinu, na které provedeme výseč shluků. Když vezmeme hladinu vzdálenosti rovnu 1, bude výseč u spojení 7 a 8 a povede to na 5 shluků. Podíváme-li se podrobněji, je zřejmé, že jde o 3 shluky a 2 odlehlé objekty. Oba odlehlé objekty jsou přesto nazvány shluky, i když každý shluk obsahuje pouze 1 objekt. **Spoj:** pořadové číslo shluku. Počet shluků však obsahuje také odlehlé hodnoty. **Velikost vzdálenosti:** vzdálenost mezi dvěma spojovanými shluky. Za normálních okolností se tato hodnota monotóně zvyšuje. Slouží k určování vhodného počtu shluků. **Diagram vzdálenosti:** čárový diagram hodnot vzdálenosti. Vybereme takový počet shluků, který vykazuje náhlý skok v diagramu. **Propojené řádky:** jde o čísla objektů čili řádků, propojených v tento aktuální shluk. Všimněte si, že propojení jsou zobrazena v opačném pořadí: v našem příkladě se objekt 7 a 10 spojují jako první, k nim se pak přidává objekt 4 atd. **Kofenetická korelace:** kofenetický korelační koeficient *CC* mezi skutečnou vzdáleností a dendrogramem predikovanou vzdáleností je založen na dotyčné hierarchické konfiguraci. Hodnoty *CC* = 0.75 a vyšší znamenají, že shlukování je možné považovat za užitečné. **Delta(0.5, 1):** obě kritéria jsou mírami těsnosti proložení. Užijeme je jako rozhodčí kritéria k hledání a rozlišení hledání a mezi technikami shlukování. Čím více je delta bližší nule, tím je shlukovací technika lepší a

věrohodnější.

4. Tabulka vzdáleností skutečných a predikovaných v dendrogramu:

První řádek	Druhý řádek	Skutečná vzdálenost	Vzdálenost ve shluku	Rozdíl vzdálenosti	Procento rozdílu
1	2	1.202793	1.210352	-0.007559	-0.63
1	3	1.227878	1.210352	0.017526	1.43
1	4	0.910046	1.210352	-0.300306	-33.00
1	5	1.596586	2.054454	-0.457869	-28.68
..	..	.....	.....	.....	.....
11	12	1.761248	1.609330	0.151919	8.63

Ukazuje skutečné a v dendrogramu predikované vzdálenosti pro každý pár objektů. Obsahuje také jejich rozdíl a rozdíl v procentech. Obvykle se tento oddíl zkracuje, či vůbec vynechává, protože i pro malý počet objektů jeho délka velice narůstá.

5. Hledání nejlepší metody shlukování:

V tabulce je uvedeno postupně 7 metod shlukování a za rozhodčí kritéria jsou použita kofenetický korelační koeficient a obě kritéria delta. Nejlepší technika tvorby dendrogramu vykazuje nejvyšší hodnotu kofenetického korelačního koeficientu a nejnižší hodnotu blízkou nule u kritérií delta. V dané úloze se ukázala jako nejlepší technika Průměrové metody (Group Average, Unweighted Pair-Group).

1. Metoda shlukování: **Skupinový průměr**, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.891119, *Delta(0.5):* 0.122037, *Delta(1.0):* 0.160527,

2. Metoda shlukování: Jednoduchý průměr, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.883564, *Delta(0.5):* 0.133286: *Delta(1.0),* : 0.173097,

3. *Metoda shlukování:* Těžiště, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.858393, *Delta(0.5):* 0.369357, *Delta(1.0):* 0.454509,

4. Metoda shlukování: Nejbližšího souseda, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.848369, *Delta(0.5):* 0.294508, *Delta(1.0):* 0.379311,

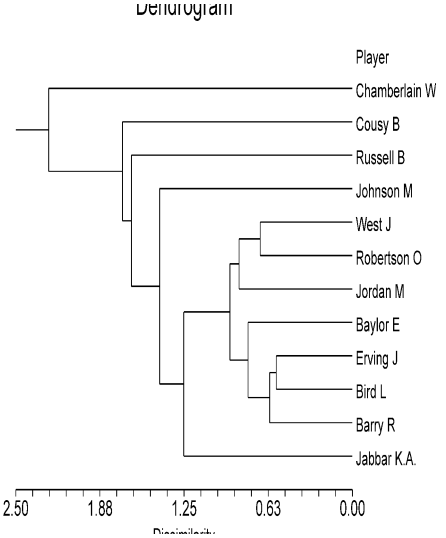
5. *Metoda shlukování:* Nejvzdálenějšího souseda*, Typ vzdálenosti:* Eucleid., směrodatná odchylka *Kofenetická korelace:* 0.762902, *Delta(0.5):* 0.283987, *Delta(1.0):* 0.361304,

6. Metoda shlukování: Median, *Typ vzdálenosti:*Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.847619, *Delta(0.5):* 0.297623, *Delta(1.0):* 0.354229,

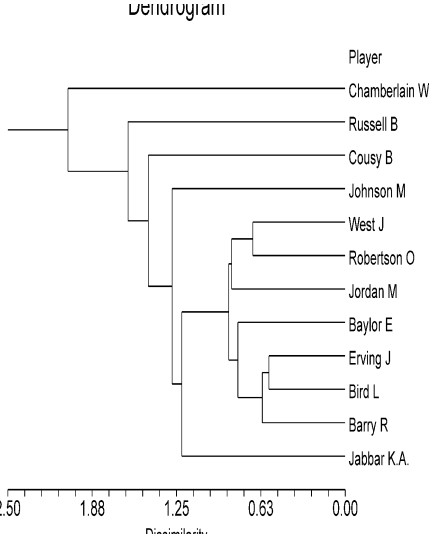
7. *Metoda shlukování:* Wardova metoda, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.687581, *Delta(0.5):* 0.536833, *Delta(1.0):* 0.622271,

6. Grafické zobrazení dendrogramu podobnosti objektů:

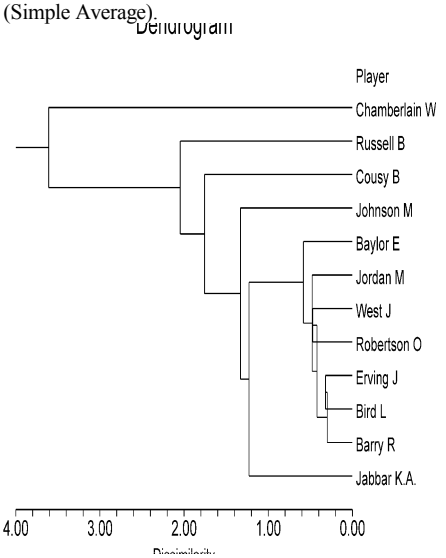
Dendrogram podobnosti objektů názorně ukazuje tvorbu a rozlišení shluků a představuje vlastně rozhodující výsledek shlukové analýzy vícerozměrných dat.



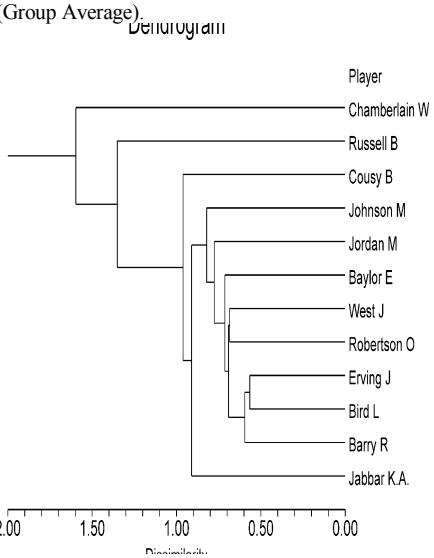
Obr. 4.18a Metoda váženého průměru (Simple Average).



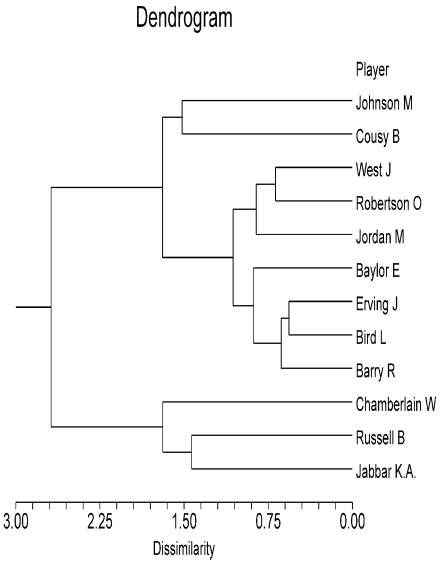
Obr. 4.18b Metoda nevážených průměrů (Group Average).



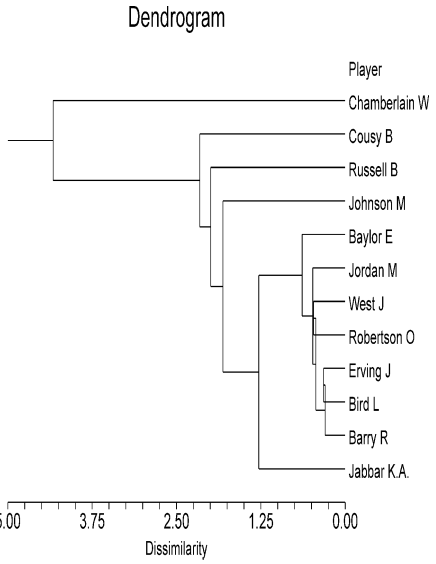
Obr. 4.18c Metoda těžiště (Centroid).



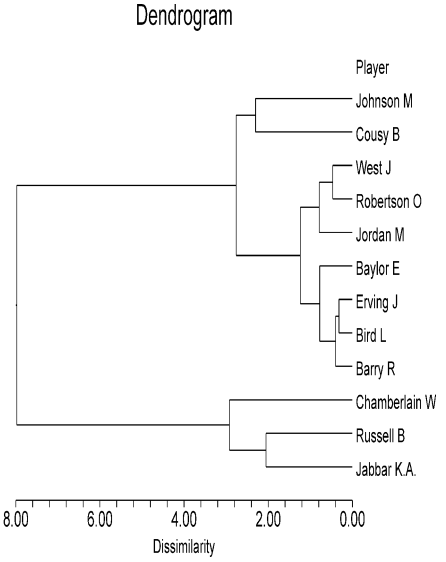
Obr. 4.18d Metoda nejbližšího souseda (Single Linkage).



Obr. 4.18e Metoda nejvzdálenějšího souseda (Complete Link.).



Obr. 4.18f Metoda mediánová (Median).



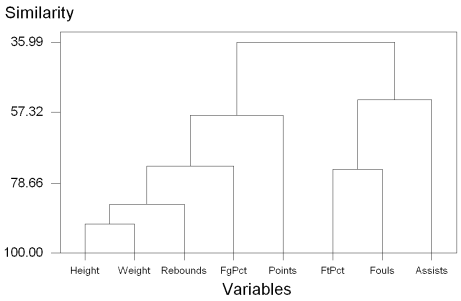
Obr. 4.18g Metoda Wardova (Wards' Minimum Variance).

Interpretace dendrogramu je snadná: objekty blízko sebe jsou propojeny spojovací úsečkou hodně vpravo, mají malou vzdálenost, čili značnou podobnost. Objekty propojené hodně vlevo mají malou podobnost a mezi sebou vykazují velkou vzdálenost, např. hráč Bob Cousy se velice liší od všech ostatních hráčů. Míra podobnosti nebo naopak míra vzdálenosti dvou objektů se může přečíst přímo na ose. Počet vhodných shluků může být snadno určen zakreslením svislice do diagramu. Vztyčíme-li, například, kolmici v bodě 1.0 na ose vzdáleností *x*, dostaneme 5 shluků. Jeden shluk obsahuje 2 objekty, jeden shluk obsahuje 7 objektů a tři shluky obsahují každý pouze po 1 objektu. Nejvhodnější techniku shlukování vybereme na základě dvou rozhodčích kritérií, kofenetické korelace

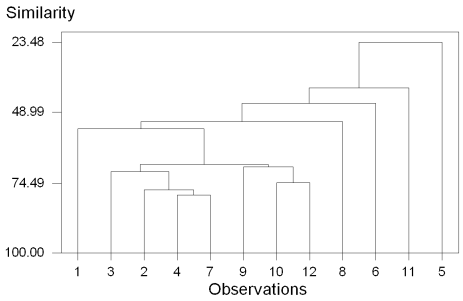
CC a kritéria delta. Na sedmi dendrogramech jsou uvedeny shluky, vytvořené sedmi rozličnými technikami. Protože rozhodčí kritéria nalezla jako nejlepší techniku průměrovou, budeme tento dendrogram považovat za optimální a v naší analýze za výsledný. Uživatel může porovnat jemné rozdíly mezi dendrogramy a posoudit jejich věrohodnost dle velikosti kofenetické korelace a kritérií delta.

7. Grafické zobrazení dendrogramu podobnosti proměnných:

Dendrogram podobnosti proměnných názorně ukazuje rozlišení proměnných ve shlucích. Jeho interpretace je snadná: proměnné blízko sebe jsou propojeny spojovací úsečkou hodně nízko, mají malou vzdálenost čili značnou vzájemnou podobnost. Proměnné propojené hodně vysoko mají malou podobnost a mezi sebou vykazují velkou vzdálenost. Stejná interpretace podobnosti je i u dendrogramu objektů, kde jsou tentokrát objekty označeny nikoliv svými jmény, ale indexy.



Obr. 4.19a Dendrogram proměnných metodou průměrovou, **MINITAB.**



Obr. 4.19b Dedrogram objektů (hráčů košíkové) metodou průměrovou, **MINITAB.**

Vzorová úloha 4.11 *Vytvoření dendrogramu objektů neuroleptika*

Vytvoření dendrogramu neuroleptik u **Úlohy B4.02** *Účinky neuroleptik při tlumení rozličných psychóz.* Neuroleptika redukují nežádoucí účinky přebytečného dopaminu. Liší se však ve svých účincích: potlačují nervozitu, záchvaty, třes, ospalost, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění atd. Účelem je provést klasifikaci neuroleptik do shluků podobných účinků s ohledem na čtyři proměnné.

Řešení:

U úlohy uvedeme pro stručnost pouze zkrácený výstup programu NCSS2000. Po výběru optimální metody tvorby dendrogramu uvedeme dendrogram podobnosti proměnných a dendrogram podobnosti objektů. Nejvyšší hodnota kofenetického korelačního koeficientu a nejnižší hodnota kritéria *delta* svědčí o optimální shlukovací metodě tvorby dendrogramu metodou průměrovou

1. Metoda shlukování: **Skupinový průměr,** *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.987356, *Delta(0.5):* 0.137455, *Delta(1.0):* 0.125290;

2. Metoda shlukování: Jednoduchý průměr, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.988876, *Delta(0.5):* 0.177810, *Delta(1.0):* 0.188781;

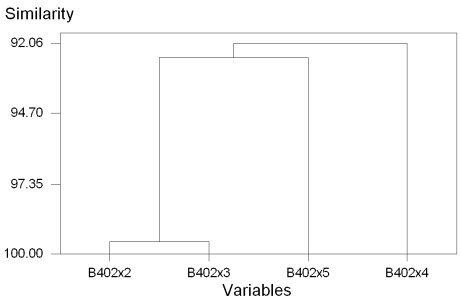
3. Metoda shlukování: Těžiště, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.984750, *Delta(0.5):* 0.175238, *Delta(1.0):* 0.166599;

4. Metoda shlukování: Nejbližšího souseda, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.988598, *Delta(0.5):* 0.474238, *Delta(1.0):* 0.391993;

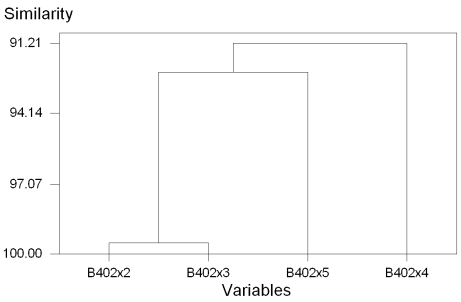
5. Metoda shlukování: Median, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:*

0.984215, *Delta(0.5):* 0.452308, *Delta(1.0):* 0.428346;

6. Metoda shlukování: Wardova metoda, *Typ vzdálenosti:* Eucleid., směrodatná odchylka, *Kofenetická korelace:* 0.979285, *Delta(0.5):* 0.549394, *Delta(1.0):* 0.492716.

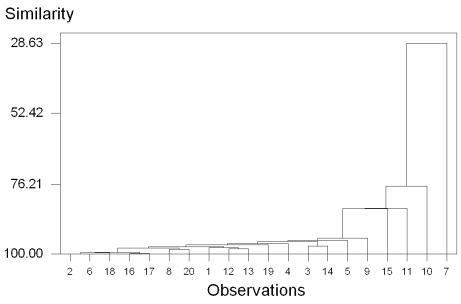


Obr. 4.20a Metoda nejbližšího souseda v dendrogramu proměnných, ***MINITAB.***

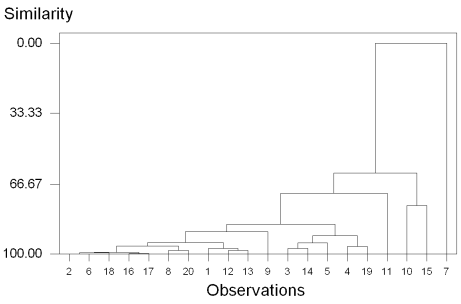


Obr. 4.20b Metoda průměru v dendrogramu proměnných, **MINITAB.**

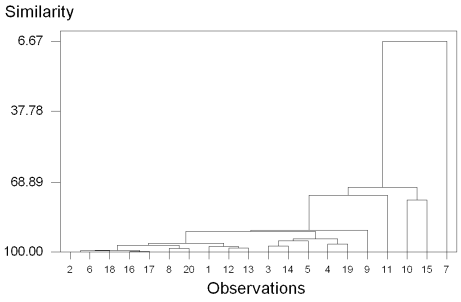
Metoda průměru v dendrogramu podobnosti objektů ukazuje na 6 shluků: první shluk obsahuje 10 objektů 18, 16, 17, 8, 20, 1, 12, 13, druhý shluk 5 objektů 3, 14, 5, 4, 19, třetí shluk 2 objekty 10 a 15 a zbývající tři shluky obsahují vždy po jednom objektu, a to 9, 11 a 7. Jako silně vybočující objekt se jeví objekt 7. Dendrogramy získané ostatními technikami dospěly převážně ke stejným shlukům a podobné struktuře.



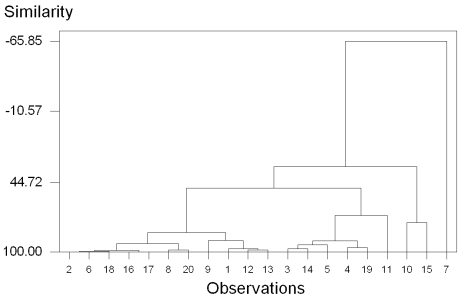
Obr. 4.21a Dendrogram objektů (neuroleptik) metodou nejbližšího souseda, *CC* = 0.9886, Δ(0.5) = 0.474, Δ(1.0) = 0.392.



Obr. 4.21b Dendrogram objektů metodou nejvzdálenějšího souseda, *CC* = 0.9828, Δ(0.5) = 0.179, Δ(1.0) = 0.183.



Obr. 4.21c Dendrogram objektů (neuroleptik) metodou průměru, $CC = 0.9889$, $\Delta(0.5) = 0.178$, $\Delta(1.0) = 0.189$.



Obr. 4.21d Dendrogram objektů (neuroleptik) Wardovou metodou, $CC = 0.979$, $\Delta(0.5) = 0.549$, $\Delta(1.0) = 0.493$.

(b) Shlukování metodou nejbližších středů (K-Means)

Při vytváření malého počtu shluků z velkého počtu objektů se jeví nejužitečnější shlukovací metodou. Vyžaduje spojité proměnné a především bez odlehlých hodnot. Diskrétní data mohou být rovněž analyzována, ale mohou způsobit problémy.

Princip metody spočívá v rozdělení n objektů o m proměnných do k shluků tak, že mezishluková suma čtverců je přitom minimalizována. Jelikož počet možných uspořádání je enormě veliký, není praktické očekávat vždy nejlepší řešení. Algoritmus nalezne spíše optimum lokální než globální. Je to takové uspořádání shluků, kdy již přemístění objektu z jednoho shluku do druhého nezpůsobí snížení sumy čtverců. Algoritmus pracuje opakovaně, startuje vždy z jiného počátečního uspořádání. Nakonec vybere optimální řešení ze všech možných dosažených uspořádání shluků.

Uživatel zadává počet shluků, jež mají být nalezeny. Pak jsou vytvořeny prostorové shluky nalezením souboru středů shluků tak, že každý objekt je přiřazen do jednoho shluku, načež jsou určeny nové shluky a celý proces se opakuje.

Předpokládejme n objektů rozdělených do k shluků. Pak k-tý shluk obsahuje n_K objektů. Každý objekt je v jednom řádku popsán m proměnnými. Chybějící hodnota i-té proměnné v j-tém řádku u k-tého shluku je označena δ_{ijk}. Data x_{ij} jsou předem standardizována a označena z_{ij}.

Počáteční přiblížení ovlivňuje konečné uspořádání shluků. Proto algoritmus pro každý pokus zcela náhodně přiřazuje každý objekt jednomu shluku. Toto uspořádání je pak optimalizováno. Pokus nastartovat proces z rozličných náhodných uspořádání vysoko zvýší pravděpodobnost nalezení globálního optima počtu shluků.

Kritérium věrohodnosti: jde o kritérium těsnosti proložení, které je založeno na srovnání rozličných konfigurací shluků a vychází z *mezishlukové sumy čtverců* WSS_K definované vztahem

$$WSS_K = \frac{nm}{nm - m} \sum_{k=1}^{k} \sum_{i=1}^{m} \sum_{j=1}^{n_k} (1 - \delta_{ijk})(y_{ij} - c_{ik})^2 \ ,$$

kde c_{ik} je střední hodnota (průměr) *i*-té proměnné v *k*-tém shluku. *Procento variace* (proměnlivosti) je definováno vztahem:

$$PV_K \text{ ' } 100\% \frac{WSS_K}{WSS_1} \quad .$$

Chybějící data: lze řídit vypouštění objektů s chybějícími hodnotami proměnných pomocí procenta chybějících hodnot u proměnných. Objekty, které mají více chybějících proměnných než dovolené procento, jsou z další analýzy vypuštěny.

Vzorová úloha 4.12 *Klasifikace objektů do shluků*

Vstupní data jsou z úlohy **S4.16** *Shluky 12 superhvězd košíkové*. Tabulka dat obsahuje informace o osmi hráčských vlastnostech a aktivitách 12 superhvězd košíkové v sezóně 1989. Cílem je najít shluky hráčů podobných vlastností a naopak odhalit jejich aktivity a vlastnost, ve které se hráč neshoduje s ostatními hráči.

Řešení: výklad výstupu programu NCSS2000.

1. **Vyčíslení minimálního počtu iterací:**

Číslo iterace	**Počet shluků**	**Procento proměnlivosti**	**Čárový diagram procenta**
2	2	65.54	|||||||||||||||||
4	3	46.48	||||||||||||
8	4	29.17	||||||||

Oddíl pomáhá určit optimální počet shluků. **Počet vytvořených shluků**. **Procento proměnlivosti:** dává sumu čtverců pro daný počet shluků v tomto řádku ve formě procenta, kdyby vůbec nedošlo ke shlukování. **Čárový diagram procenta proměnlivosti:** grafické zobrazení procenta variace čárovým diagramem.

2. **Iterování:**

Číslo iterace	**Počet shluků**	**Procento proměnlivosti**	**Čárový diagram procenta**
1	2	72.03	|||||||||||||||||||
2	2	65.54	|||||||||||||||||
3	2	66.09	|||||||||||||||||
4	3	46.48	||||||||||||
5	3	49.04	||||||||||||
6	3	46.48	||||||||||||
7	4	31.81	|||||||||
8	4	29.17	||||||||
9	4	29.17	||||||||

Oddíl je zvláště užitečný k určení dostatečného počtu náhodných startovacích uspořádání. Když bylo specifikováno dost startovacích uspořádání, obvykle dvě či tři představují optimum pro každý počet shluků. To se pak posuzuje dle dosaženého minima procenta variace. Když se vůbec neobjeví, je třeba zvýšit počet startovacích uspořádání a výpočet opakovat. **Počet shluků:** v tomto uspořádání. **Procento proměnlivosti:** dává sumu čtverců pro počet shluků tohoto řádku vyjádřený jako procento sumy čtverců, když ke shlukování nedojde. Tak, jak se přidává více a více shluků, tato hodnota může poklesnout. Po volbě optimálního počtu shluků, bod selhání tohoto procenta dramaticky poklesne. **Čárový diagram procenta:** dává čárový diagram znázornění procenta variace.

3. Středy shluků:

Proměnná	**Shluk 1**	**Shluk 2**	**Shluk 3**
Height	78.25	85.5	77
FgPct	48.6375	54.95	40.75

Points	25.575	27.35	16.75
Rebounds	8.225	17.05	13.9
Počet	8	2	2

Tabulka ukazuje střední hodnoty proměnných pro jednotlivé shluky. Poslední řádek přináší počet objektů v dotyčném shluku.

4. Směrodatné odchylky shluků:

Proměnná	**Shluk 1**	**Shluk 2**	**Shluk 3**
Height	2.171241	0.7071068	6.363961
FgPct	3.357694	1.343503	4.596194
Points	3.770089	3.889087	2.333452
Rebounds	2.544321	8.273149	12.30366
Počet	8	2	2

Tabulka ukazuje směrodatné odchylky proměnných pro jednotlivé shluky. Poslední řádek přináší počet objektů v dotyčném shluku.

5. Testování *F*-poměru:

Proměnné	**SV1**	**SV2**	**Mezi průměrnými čtverci**	**Uvnitř průměrných čtverců**	***F*-test**	**Spočtená hladina významnosti α**
Height	2	9	48.125	8.222222	5.85	0.023532
FgPct	2	9	101.6469	11.31653	8.98	0.007170
Points	2	9	72.7475	13.34056	5.45	0.028096
Rebounds	2	9	75.04459	29.46	2.55	0.132844

Oddíl přináší výsledky jednofaktorové analýzy rozptylu pro každou proměnnou, a to při použití běžně definovaných shluků jako faktoru. Tabulka pomáhá vyšetřit důležitost každé proměnné ve shlukovacím procesu.

6. Vzdálenosti:

Řádek	**Shluk**	**Vzdálenost 1**	**Vzdálenost 2**	**Vzdálenost 3**
1 Jabbar K.A.	2	2.4609	1.1263	4.0315
2 Barry R.	1	0.9139	3.1499	1.9940
3 Baylor E.	1	1.4427	3.1724	2.2139
4 Bird L.	1	0.8398	1.8867	2.7392
5 Chamberlain W.	2	3.2456	1.1263	4.4712
6 Cousy B.	3	2.9971	5.3790	1.9512
7 Erving J.	1	0.4724	2.4891	2.5912
8 Johnson M.	1	1.6497	2.5426	2.8064
9 Jordan M.	1	1.5532	2.8939	4.0067
10 Robertson O.	1	0.3409	2.9490	2.5629
11 Russell B.	3	3.3878	3.5197	1.9512
12 West J.	1	1.0971	3.6374	2.8439

Tabulka přináší relativní vzdálenosti každého objektu ke středu shluku. Tím se prokáže, jak ostře byly shluky vytvořeny: jestliže vzdálenost každého objektu ke středu shluku je mnohem menší než vzdálenost z objektu ke středům ostatních shluků, je uspořádání do shluků výtečné a účinné. Jestliže však nejmenší vzdálenost je téměř stejná jako vzdálenosti ke středům ostatních shluků, je pak otázkou, ke ktererému shluku daný objekt vlastně patří. Takové uspořádání pak není příliš žádoucí.

7. Oddíl jednotlivých vzdáleností:

Vzdálenosti pro shluk 1:

Řádek	**Shluk**	**Vzdálenost 1**	**Vzdálenost 2**	**Vzdálenost 3**

2 Barry R.	1	0.9139	3.1499	1.9940
3 Baylor E.	1	1.4427	3.1724	2.2139
4 Bird L.	1	0.8398	1.8867	2.7392
7 Erving J.	1	0.4724	2.4891	2.5912
8 Johnson M.	1	1.6497	2.5426	2.8064
9 Jordan M.	1	1.5532	2.8939	4.0067
10 Robertson O.	1	0.3409	2.9490	2.5629
12 West J.	1	1.0971	3.6374	2.8439
Počet = 8				

Vzdálenosti pro shluk 2:

Řádek	Shluk	Vzdálenost 1	Vzdálenost 2	Vzdálenost 3
1 Jabbar K. A.	2	2.4609	1.1263	4.0315
5 Chamberlain W.	2	3.2456	1.1263	4.4712
Počet = 2				

Vzdálenosti pro shluk 3:

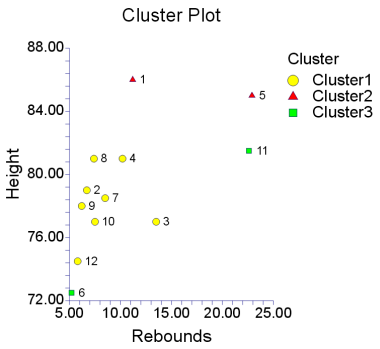
Řádek	Shluk	Vzdálenost 1	Vzdálenost 2	Vzdálenost 3
6 Cousy B.	3	2.9971	5.3790	1.9512
11 Russell B.	3	3.3878	3.5197	1.9512
Počet = 2				

Vysvětlení tohoto oddílu je stejné jako předešlého, kromě řádku, ve kterém byl zobrazen jediný shluk. Je pak snadnější poznat, který objekt připadá do kterého shluku.

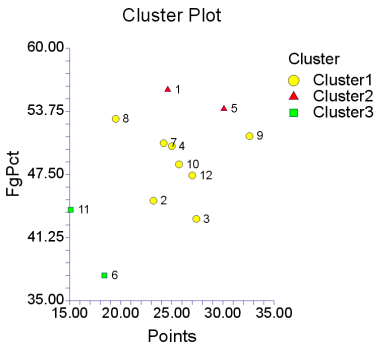
8. Grafy:

Obr. 4.22a Rozptylový diagram proměnných *Height-FgPct.*

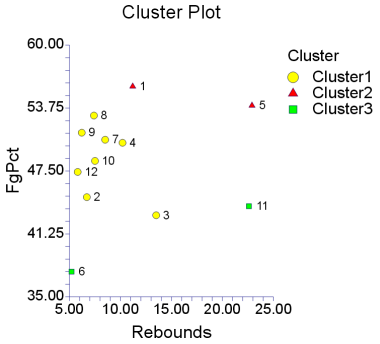
Obr. 4.22b Rozptylový diagram proměnných *Height-Points.*



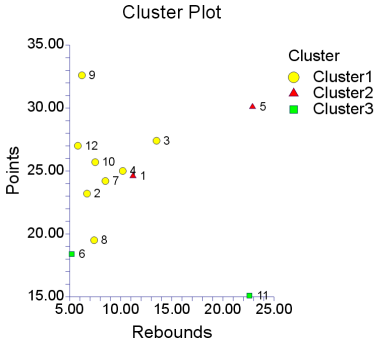
Obr. 4.22c Rozptylový diagram proměnných *Height-Rebounds*.



Obr. 4.22d Rozptylový diagram proměnných *FgPct-Points*.



Obr. 4.22e Rozptylový diagram proměnných *FgPct-Rebounds*.



Obr. 4.22f Rozptylový diagram proměnných *Points-Rebounds*.

Série indexových rozptylových diagramů, vždy pro dvojici proměnných, ukazuje rozdělení objektů do shluků. Diagramy jsou vlastně cílem shlukové analýzy, protože pomohou odhalit odlehlé objekty, anomálie a řadu dalších strukturních problémů.

(c) Shlukování metodou středů-medoidů

Medoid, čili střed shluku, je střední objekt, pro který platí, že průměrná vzdálenost k ostatním objektům v tomto shluku je minimální. Je-li požadováno *k* shluků, bude existovat také *k* medoidů. Po nalezení medoidů jsou data klasifikována do shluků vždy okolo nejbližšího medoidu. Medoidy a shluky se vytvářejí na základě *vzdáleností* čili *nepodobností* (dissimilarities).

Proměnné shluku. Druhů proměnných je celá řada: *Intervalové* jsou spojité kladné či záporné, v lineární škále, např. výška, hmotnost, cena, teplota, čas atd. *Ordinální* jsou pořadová čísla stupnice, hodnotící nějakou vlastnost, např. silný nesouhlas (5), nesouhlas (4), neutrální (3), souhlas (2) a silný souhlas (1). *Poměrové* jsou kladné hodnoty, kdy vzdálenost mezi dvěma čísly je stejná, když i jejich poměr je stejný, např. mezi 3 a 30 je stejná jako mezi 30 a 300, chemická koncentrace, intenzita záření, absorbance atd. *Nominální* jsou proměnné, které vyjadřují pouze kvalitu a nikoliv kvantitu, např. PSČ, rasa, barva, název města atd. *Symetrické binární:* mají dvě možnosti, obvykle ano (1), ne (0).

Asymetrické binární: přítomnost či nepřítomnost zřídka se vyskytující události, kdy nepřítomnost není tak důležitá, např. osoba má jizvu na tváři, a tím je lépe identifikovatelná.

Späthova metoda. Metoda minimalizuje účelovou funkci přemísťováním objektů z jednoho shluku do druhého. Začíná u počátečního uspořádání shluků, algoritmus pak najde lokální minimum inteligentním přesouváním objektů ze shluku do shluku. Jakmile se nepřemístí už žádný objekt, metoda terminuje. Lokální minimum však nemusí být globálním. Aby program překonal toto omezení, zopakuje se několikrát hledání vždy z jiného startovacího uspořádání a nejlepší uspořádání shluků je nakonec bráno za výsledné. Za účelovou funkci se bere celková vzdálenost mezi všemi objekty ve shluku podle vzorce $D = \sum_{k=1}^{K} \sum_{i \in c_K} \sum_{j \in c_K} d_{ij}$, kde K je celkový počet shluků, d_{ij} je vzdálenost mezi i-tým a j-tým objektem a c_K je soubor všech objektů ve shluku k.

Metoda PAM (Partition Around Medoids). Algoritmus se opět pokouší minimalizovat celkovou vzdálenost D ve dvou krocích:

1. Nalezne se reprezentativní soubor k objektů. První objekt má nejkratší vzdálenost ke všem ostatním objektům, čili představuje *střed*. Pak se *k-1* objektů nachází tak, že hodnota D je co možná nejmenší.
2. Možné alternativy polohy k objektů jsou vybírány iteračním způsobem. Algoritmus vyhledává dosud nezařazené objekty a přemísťuje je tak, aby se hodnota D snižovala. Iterace skončí, jakmile změny nezpůsobí další snížení hodnoty D.

Silueta: poskytuje klíčovou informaci o dobrém a špatném shluku. Hodnota siluety s se vypočte postupem:

1. Objekt i je ve shluku A a má průměrnou vzdálenost a ke všem objektům ve svém shluku. Je-li ve shluku A jediný objekt, je $a = 0$.
2. Sousední shluk B obsahuje objekty, které jsou nejbližší k objektu i ve shluku A a b je průměrná vzdálenost mezi objektem i a všemi objekty ve shluku B.
3. *Silueta s* objektu i se vyčíslí následovně: když shluk A obsahuje pouze jeden objekt, je $s = 0$. Když $a < b$, *je* $s = 1 - a/b$. Když $a > b$, je $s = b/a - 1$. Když $a = b$, je $s = 0$.

Silueta se vyčíslí pro každý objekt. Hodnota siluety se mění od -1 do +1 a je mírou úspěšné klasifikace do shluků při porovnání vzdáleností uvnitř shluku A se všemi vzdálenostmi objektů nejbližšího souseda B dle pravidla:

1. Je-li s blízko +1, objekt i je dobře klasifikován do shluku A, protože jeho vzdálenosti k ostatním objektům v tomto shluku jsou podstatně kratší než vzdálenosti k objektům nejbližšího souseda B.

2. Je-li s blízko nule, objekt i se nachází kdesi uprostřed mezi shluky A a B, a čistě náhodou byl přiřazen do shluku A.

3. Je-li s blízko -1, objekt i je špatně klasifikován. Vzdálenosti k ostatním objektů ve svém shluku jsou mnohem větší než vzdálenosti k objektům nejbližšího souseda B. Otázkou pak je, proč byl vlastně zařazen do shluku A.

Určení počtu shluků. Přehlednou statistikou je průměrná silueta s, počítaná přes všechny objekty. Tato hodnota sumarizuje jak těsně shlukové uspořádání tato prokládá

analyzovaná data. Snadný způsob nalezení správného počtu shluků spočívá v nalezení takového počtu, který maximalizuje průměrnou siluetu. Označme *maximální hodnotu průměrné siluety* všech shluků *k* symbolem *SC* a pak budeme rozlišovat následující typy shlukových uspořádání:

SC	Vysvětlení uspořádání do shluků
od 0.71 do 1.00	Silná a dobrá struktura.
od 0.51 do 0.70	Ještě přijatelná struktura.
od 0.26 do 0.50	Slabá struktura, asi umělá. Je třeba najít novou, lepší.
od -1.00 do 0.25	Naprosto nevhodná struktura.

Diagnostika dobrého shlukování. Čárový diagram siluet, uspořádány podle stoupajícího počtu shluků a hodnoty siluety, dobře prokazuje nejlepší uspořádání shluků. Důležitým kritériem je kladná hodnota siluety *s*, která by měla být také větší než 0.50. Je- li silueta pro některé shluky menší než 0.50 nebo dokonce záporná, jsou takové shluky nepravděpodobné a měli bychom hledat jiné.

Další analýza struktury objektů. Po úspěšném nalezení počtu shluků a nejlepšího shlukového uspořádání by měla následovat diskriminační analýza, která statisticky testuje, jak dobře byly objekty (řádky) roztříděny do shluků. Testování se provádí pomocí Wilkovy statistiky λ. Vedle diskriminační analýzy jsou sestrojeny různé rozptylové diagramy, ve kterých je počet shluků použit jako důležitá proměnná. Teprve diagramy odhalí a vysvětlí pravý smysl klasifikační analýzy do shluků.

Vzorová úloha 4.13 *Odhalení struktury objektů rozličnými metodami shlukování*

Vstupní data jsou z úlohy **S4.16** *Shluky 12 superhvězd košíkové*. Tabulka dat obsahuje informace o osmi hráčských vlastnostech a aktivitách 12 superhvězd košíkové v sezóně 1989. Cílem je najít shluky hráčů podobných vlastností, a naopak odhalit jejich aktivity a vlastnost, ve které se hráč neshoduje s ostatními hráči.

Řešení: Výklad výstupu programu NCSS2000. Při shlukovací analýze těchto dat bude použito dvou rozličných metod medoidního shlukování programu NCSS2000:

(i) Výstup algoritmu medoidního shlukování za použití Späthovy metody:

1. Titulní stránka: *Metoda Späthova*

Proměnné:	*Height* až do *Rebounds*
Metoda:	Späthova,
Objektivní funkce:	Silueta,
Typ vzdálenosti:	Eukleidovský,
Typ škály:	Směrodatná odchylka,

2. Iterační průběh hledání optimálního počtu shluků:

Počet shluků	Minimalizovat tuto průměrnou vzdálenost	Nastavená průměrná vzdálenost	Maximalizovat tuto průměrnou siluetu
2	35.977405	5.996234	0.135735
2	34.352873	5.725479	0.185579
2	34.862052	5.810342	0.170356
2	36.031237	6.005206	0.101405
3	19.525066	4.881267	0.094407
3	21.106317	5.276579	0.033435
3	19.005957	4.751489	0.045621
3	22.202362	5.550590	-0.026350
4	12.547872	4.182624	-0.013869
4	12.318440	4.106147	0.044989
4	12.210147	4.070049	0.018876
4	14.209356	4.736452	-0.097672
5	9.344940	3.893725	-0.099737
5	10.556815	4.398673	-0.189487
5	8.274123	3.447551	-0.045335
5	8.049819	3.354091	-0.004580

Průběh ukazuje hodnoty účelových funkcí *D, D_{adjust}* a *s* pro každou iteraci a proměnný počet shluků metodou Späthovou. Tabulka je zvláště užitečná při zadávání správného počtu opakování: jsou-li totiž dvě či tři shluková uspořádání při stejném požadovaném počtu shluků totožná, potom byl zvolen dostatečně velký zadaný počet opakování. Jinak bylo třeba zvýšit tuto hodnotu a analýzu opakovat. V tomto příkladu dojdeme k závěru, že optimální *k* je rovno 2. Protože jsme hodnotu siluety *s* = 0.185579 obdrželi pouze jednou, je třeba změnit počet opakování na 10 a výpočet opakovat. **Průměrná vzdálenost:** tato hodnota může být přepočtena tak, že uvede procento vzdálenosti, vztažené vůči maximální vzdálenosti v matici vzdáleností. Tím se zlepší interpretace této veličiny. **Nastavená průměrná vzdálenost:** jde o hodnotu nastavené průměrné vzdálenosti dle vzorce $D = \frac{K}{N} \sum_{k=1}^{K} \sum_{i \in c_K} \sum_{j \in c_K} d_{ij}$. Tato hodnota může být stejně jako předešlá vyjádřena v procentech maximální vzdálenosti. **Průměrná hodnota siluety všech objektů:** čili všech řádků, je důležitým kritériem úspěšnosti shlukování a především určení optimálního počtu shluků. Maximální hodnota siluety znamená nejlepší vhodný počet shluků.

3. Přehled iterací:

Počet shluků	Minimalizovat tuto průměrnou vzdálenost	Nastavená průměrná vzdálenost	Maximalizovat tuto průměrnou siluetu
2	34.352873	5.725479	0.185579
3	19.525066	4.881267	0.094407
4	12.318440	4.106147	0.044989
5	8.049819	3.354091	-0.004580

Přehled uvádí závěrečné výsledky účelových funkcí pro každý počet shluků. Uvádí závěry hledání vhodného počtu shluků. Tento počet odpovídá maximální hodnotě siluety v posledním sloupci. Takový řádek poskytne také důležitou hodnotu nalezeného průměru všech vzdáleností. Toto číslo by mělo dosahovat svého minima.

4. Souřadnice medoidů shluků:

Proměnná	Shluk 1	Shluk 2
Height	86	77

Weight	230	210
FgPct	55.9	48.5
FtPct	72.1	83.8
Points	24.6	25.7
Rebounds	11.2	7.5
Řádek	1 Jabbar K.A	10 Robertson

Tabulka uvádí souřadnice čili proměnné medoidů všech shluků, což pomůže vysvětlit a kvalitativně rozlišit jednotlivé shluky. Poslední řádek přináší číslo nebo jméno objektu, který je označen jako medoid. Všimněme si, že hráči ve shluku č. 1 jsou typicky o 9 palců vyšší než ve shluku č. 2 a mají o 4 doskoky pod košem více právě kvůli své výhodné výšce. Shluk č. 1 proto obsahuje především vysoké podkošové hráče, zatímco shluk č. 2 pak ostatní hráče.

5. Objektové (řádkové) podrobnosti:

Řádek	Nejbližší shluk	Vzdál. soused	Průměrná vzdálenost k němu	Průměr siluety souseda	Silueta	Čárový diagram
5 Chamberlai W.	1	2	58.85	74.48	0.2098	|IIIIIIIIII
11 Russell B.	1	2	58.40	57.83	-0.0098	|
1 Jabbar K. A.	1	2	47.59	44.28	-0.0696	|
3 Baylor E.	1	2	47.74	31.69	-0.3363	|
9 Jordan M.	1	2	56.34	32.62	-0.4210	|
Cluster Average	1	(5)	53.79	48.18	-0.1254	
2 Barry R.	2	1	24.20	47.65	0.4921	|IIIIIIIIIIIIIIIIIIIIIII
10 Robertson O.	2	1	21.94	42.43	0.4830	|IIIIIIIIIIIIIIIIIIIII
12 West J.	2	1	29.13	51.74	0.4370	|IIIIIIIIIIIIIIIIIII
7 Erving J.	2	1	23.03	40.90	0.4369	|IIIIIIIIIIIIIIIIIII
6 Cousy B.	2	1	45.02	68.58	0.3435	|IIIIIIIIIIIIII
8 Johnson M.	2	1	30.31	45.50	0.3339	|IIIIIIIIIIIIII
4 Bird L.	2	1	27.19	40.44	0.3275	|IIIIIIIIIIIIII
Shlukový průměr	2	(7)	28.69	48.18	0.4077	
Celkový průměr		(12)	39.15	48.18	0.1856	
Maximální vzdálenost:	3.012578					

Tabulka přináší cenné informace o každém objektu v řádku, který byl zařazen do shluků. Tabulka je uspořádána podle klesajících hodnot siluety v každém shluku. **Řádek:** číslo nebo jméno objektu v řádku. V tomto výstupu je vysvětlen každý řádek databáze. **Číslo shluku:** číslo, do kterého jsou objekty roztříděny. **Nejbližší soused:** identifikační číslo nejbližšího shluku vůči tomuto objektu v řádku. Tato informace je užita při výpočtu hodnoty siluety. **Průměrná vzdálenost k němu:** průměrná vzdálenost mezi tímto objektem a ostatními objekty v uvažovaném shluku. Jde o číslo **a** ve výpočtu siluety. **Průměrná vzdálenost souseda:** průměrná vzdálenost mezi tímto objektem a objekty v nejbližším sousedním shluku. Jde o číslo ***b*** ve výpočtu siluety. **Silueta:** hodnota ***s*** musí být kladná a pokud možno vyšší než 0.5, jinak je shluk nevhodný, špatný.

(ii) Medoidního shlukování metodou PAM:

1. Titulní stránka: *Metoda PAM (Partition Around Medoids):*

Proměnné:	*Height* až do *Rebounds*
Metoda:	Kaufmanova-Rousseeuwova,
Objektivní funkce:	Silueta,
Typ vzdálenosti:	Eukleidovská,
Typ škály:	Směrodatná odchylka,

2. Iterační průběh hledání optimálního počtu shluků:

Počet shluků	Minimalizovat tuto průměrnou vzdálenost	Nastavená průměrná vzdálenost	Maximalizovat tuto průměrnou siluetu
2	37.478948	6.246491	0.382164
3	44.870219	11.217555	0.340904
4	31.340599	10.446866	0.270905
5	21.523501	8.968126	0.198372

3. Souřadnice medoidů shluků:

Proměnná	Shluk 1	Shluk 2
Height	86	77
Weight	230	210
FgPct	55.9	48.5
FtPct	72.1	83.8
Points	24.6	25.7
Rebounds	11.2	7.5
Řádek	1 Jabbar K.A	10 Robertson O.

4. Objektové (řádkové) podrobnosti:

Řádek	Nejbližší shluk	Vzdál. soused	Průměrná vzdálenost k němu	Průměr siluety souseda	Silueta	Čárový diagram
5 Chamberlain W.	1	2	51.58	72.62	0.2898	|IIIIIIIIIIII
11 Russell B.	1	2	55.70	58.55	0.0488	|II
1 Jabbar K. A.	1	2	49.24	44.65	-0.0931	|
Cluster Average	1	(3)	52.17	58.61	0.0818	
10 Robertson O.	2	1	22.01	55.89	0.6061	|IIIIIIIIIIIIIIIIIIIIIIIIIIII
2 Barry R.	2	1	25.49	59.86	0.5742	|IIIIIIIIIIIIIIIIIIIIIIII
12 West J.	2	1	28.86	67.56	0.5728	|IIIIIIIIIIIIIIIIIIIIIIII
7 Erving J.	2	1	23.89	50.51	0.5270	|IIIIIIIIIIIIIIIIIIIII
9 Jordan M.	2	1	32.91	63.46	0.4814	|IIIIIIIIIIIIIIIIIII
4 Bird L.	2	1	27.41	48.69	0.4370	|IIIIIIIIIIIIIIIIII
6 Cousy B.	2	1	46.64	79.96	0.4167	|IIIIIIIIIIIIIIIII
3 Baylor E.	2	1	32.10	51.99	0.3827	|IIIIIIIIIIIIIII
8 Johnson M.	2	1	32.59	49.56	0.3425	|IIIIIIIIIIIII
Shlukový průměr	2	(9)	30.21	58.61	0.4823	
Celkový průměr		(12)	35.70	58.61	0.3822	
Maximální vzdálenost:		3.012578				

(d) Fuzzy shlukování

Fuzzy shlukování zobecňuje všechny shlukovací metody tím, že umožňuje shlukování jednoho objektu do více než jednoho shluku, zatímco v běžném shlukování je každý objekt členem pouze jednoho shluku. Předpokládejme, že máme K shluků a budeme definovat soubor proměnných m_{i1} , m_{i2} , ..., m_{iK} , které představují pravděpodobnost, že objekt i je klasifikován do k-tého shluku. V běžném shlukovacím algoritmu je jedna z těchto proměnných rovna jedné a zbytek roven nule. To představuje skutečnost, že takový algoritmus klasifikuje každý objekt do jednoho a právě jednoho shluku.

Ve fuzzy shlukování je "účast objektu", čili přítomnost objektu, rozdělena do všech shluků. Proměnná m_{ik} může zde být rovna 1 nebo 0 a suma těchto hodnot musí být rovna 1. Nazveme tento process *fuzzifikací shlukové konfigurace.* Proces má výhodu, že nenutí objekt aby byl zařazen pouze do jediného specifického shluku. Nevýhodou však je, že se zde objevuje mnohem více informací, které musí být vysvětleny. Fuzzy algoritmus minimalizuje účelovou funkci *C*, která je funkcí neznámých účastí ve shluku a dále funkcí i vzdáleností dle vztahu

$$C = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} m_{ik}^2 \, m_{jk}^2 \, d_{ij}}{2 \sum_{j=1}^{N} m_{jk}^2} ,$$

kde m_{ik} představuje neznámou účast objektu *i* v *k*-tém shluku *k* a d_{ij} je vzdálenost mezi objekty *i* a *j*. Účasti ve shluku jsou předmětem omezení a musí být nezápornými čísly a dále účasti pro jeden objekt musí být v sumě rovny 1. To znamená, že účasti mají stejná omezení, jako by to byly pravděpodobnosti, že individuum patří do jisté skupiny.

Míra věrohodnosti: jedním z nejobtížnějších úkolů ve shlukové analýze je nalezení vhodného *počtu* shluků. Velikost "fuzzifikace" v řešení se dá změřit *Dunnovým rozdělovacím koeficientem*, který představuje míru, jak těsně padne fuzzy řešení na odpovídající *pevné shluky*. Za pevné shluky budeme považovat klasifikaci každého objektu do shluku, který má největší účast. Dunnův rozdělovací koeficient se vyjádří vzorcem

$$F(U) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} m_{ik}^2 .$$

Koeficient leží v intervalu od *1/K* do 1. Hodnoty *F(U) = 1/K* se dosáhne, když všechny účasti jsou rovny *1/K*. Hodnota *F(U)* = 1 platí, když pro každý objekt je účast jednotková a zbytek je roven nule. Dunnův rozdělovací koeficient může být také normován tak, že jeho hodnota se mění od 0 (úplně fuzzy) do 1 (pevný shluk). Normovaná verze má tvar:

$$Fc(U) = \frac{F(U) - (1/K)}{1 - (1/K)} .$$

Další koeficient představuje *Kaufmanův rozdělovací koeficient*

$$D(U) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} (h_{ik} - m_{ik})^2 .$$

Koeficient má hodnotu danou intervalem od *D(U)* = 0 (pevné shluky) do *D(U)* = 1-(1/*K)* (úplně fuzzy). Normovaná verze tohoto koeficientu má tvar

$$Dc(U) = \frac{D(U)}{1 - (1/K)} .$$

Oba normované koeficienty *Fc(U)* a *Dc(U)* poskytují dohromady dobrou indikaci optimálního počtu shluků. celočíselná hodnota *K* by měla být volena tak, že *Fc(U)* bude nabývat malé a *Dc(U)* velké hodnoty.

Vzorová úloha 4.14 *Klasifikace objektů barev fuzzy shlukováním*

V úloze S4.17 je dáno 22 objektů barev, které vznikly různým podílem červené a modré barvy. Je třeba provést klasifikaci objektů barev do shluků.
Rešení: Výstup programu NCSS2000:

1. Titulní stránka: *Fuzzy metoda*

Proměnné:	Red, Blue
Typ vzdálenosti:	Eukleidovská,
Typ škály:	Směrodatná odchylka

2. Přehled závěrečných výsledků určení počtu shluků:

Počet shluků	**Průměrná vzdálenost**	**Průměrná silueta**	***F(U)***	***Fc(U)***	***D(U)***	***Dc(U)***
2	5.928694	0.418844	0.6775	0.3551	0.1656	0.3312
3	2.788827	0.704435	0.7098	0.5647	0.0877	0.1315
4	2.112581	0.492543	0.5481	0.3975	0.2030	0.2707
5	1.655932	0.351359	0.4871	0.3589	0.2682	0.3352

I když se tato tabulka objevuje až na samém konci výstupu, začneme s jejím výkladem hned na začátku. Ukazuje totiž na hledaný počet shluků, zde pro 3 shluky dosahuje průměrná silueta své maximální hodnoty, dále *Fc(U)* rovněž své maximální hodnoty a *Dc(U)* své minimální hodnoty. **Průměrná vzdálenost:** jde o hodnotu průměrné vzdálenosti. Tato hodnota může být přepočítána jako relativní procento vůči maximální vzdálenosti v matici vzdáleností. **Průměrná silueta:** jde o průměrnou siluetu ze všech objektů v řádcích. Optimální počet shluků je ten, pro který tato veličina nabývá své maximální hodnoty. ***F(U), Fc(U), D(U), Dc(U)*** jsou rozdělovací koeficienty. Optimálního počtu shluků je dosaženo pro maximální *F(U)* a minimální *D(U)*. Po určení optimálního počtu shluků bude řešení studováno detailně. Jelikož zde vyšlo, že optimální jsou 3 shluky, budeme si nadále všímat jenom výsledků pro 3 shluky a ostatní části výstupu pro počty shluků 2, 4, a 5 vynecháme.

3. Středy (medoidy) shluků:

Proměnná	**Shluk 1**	**Shluk 2**	**Shluk 3**
Červená (Red)	2	14	7
Modrá (Blue)	9	10	2
Řádek	3	10	18

Tato tabulka uvádí středy, čili medoidy, nejbližšího pevného shlukování. Tabulka pomůže vysvětlit podstatu shluků v této úloze. Poslední řádek udává číslo řádku (a někdy také jméno) objektu každého shlukového středu - medoidu.

4. Přehled účastí ve shlucích:

Přehled účasti ve 3 shlucích:

Row	**Shluk**	**Shluková účast**	**Suma čtverců účasti**	**Čárový diagram čtverců účasti**	**Silueta**	**Diagram siluety**
3	1	0.9354	0.8770	|IIIIIIIIIIIIIIIIIIIIIII	0.6907	|IIIIIIIIIIIIIIIII
5	1	0.8779	0.7782	|IIIIIIIIIIIIIIIIIII	0.6681	|IIIIIIIIIIIIIIII
2	1	0.8728	0.7698	|IIIIIIIIIIIIIIIIIII	0.6892	|IIIIIIIIIIIIIIIII

1	1	0.8705	0.7662	|IIIIIIIIIIIIIIIIIIII	0.6509	|IIIIIIIIIIIIIIIII
4	1	0.8535	0.7395	|IIIIIIIIIIIIIIIIIII	0.6060	|IIIIIIIIIIIIIII
6	1	0.4227	0.3561	|IIIIIIIII	0.1246	|IIII
13	1	0.3670	0.3374	|IIIIIIII	-0.0544	|
10	2	0.8705	0.7662	|IIIIIIIIIIIIIIIIIIII	0.8282	|IIIIIIIIIIIIIIIIIIII
8	2	0.8680	0.7622	|IIIIIIIIIIIIIIIIIII	0.8117	|IIIIIIIIIIIIIIIIII
11	2	0.8556	0.7425	|IIIIIIIIIIIIIIIII	0.8057	|IIIIIIIIIIIIIIIIII
9	2	0.8508	0.7352	|IIIIIIIIIIIIIIIII	0.7905	|IIIIIIIIIIIIIIIIII
12	2	0.8402	0.7188	|IIIIIIIIIIIIIIIII	0.8106	|IIIIIIIIIIIIIIIIII
7	2	0.8231	0.6931	|IIIIIIIIIIIIIIII	0.7660	|IIIIIIIIIIIIIIIII
18	3	0.9209	0.8511	|IIIIIIIIIIIIIIIIIIIII	0.8627	|IIIIIIIIIIIIIIIIIIII
21	3	0.8663	0.7595	|IIIIIIIIIIIIIIIII	0.8398	|IIIIIIIIIIIIIIIIII
19	3	0.8652	0.7577	|IIIIIIIIIIIIIIIII	0.8324	|IIIIIIIIIIIIIIIIII
17	3	0.8647	0.7570	|IIIIIIIIIIIIIIIII	0.8210	|IIIIIIIIIIIIIIIIII
15	3	0.8483	0.7312	|IIIIIIIIIIIIIIII	0.8023	|IIIIIIIIIIIIIIIII
20	3	0.8258	0.6972	|IIIIIIIIIIIIIIII	0.8050	|IIIIIIIIIIIIIIIII
22	3	0.8258	0.6971	|IIIIIIIIIIIIIIII	0.8145	|IIIIIIIIIIIIIIIII
16	3	0.8012	0.6617	|IIIIIIIIIIIIIII	0.7764	|IIIIIIIIIIIIIIII
14	3	0.8000	0.6603	|IIIIIIIIIIIIIII	0.7556	|IIIIIIIIIIIIIIII

Tabulka přináší informaci o každém objektu (každém řádku). Je roztříděna dle hodnoty siluety v každém shluku. Všimněme si, jak dobře jsou identifikovány dva odlehlé objekty, č. 6 a č. 13. **Řádek:** číslo nebo i jméno objektu, řádku. **Shluk:** číslo shluku, do kterého je objekt klasifikován. **Shluková účast:** jde o maximum shlukové účasti. Je to účast pro shluky, do kterých byl tento řádek (objekt) zařazen při pevném shlukování. **Suma čtverce účasti:** všechny účasti pro daný objekt (řádek) jsou umocněny na druhou a sečteny. Když je objekt úplně přiřazen do jediného shluku, je tato hodnota rovna 1. Když je řádek klasifikován stejnou měrou do každého shluku, bude hodnota 1/*K*. Řádky s vysokými hodnotami jsou proto ve shluku blízko středu. Řádky s nízkými hodnotami jsou odlehlé hodnoty. **Čárový diagram čtverce účasti**: čárový diagram sumy čtverců hodnot účasti pomůže odhalit řádky, které nejsou dobře zařazeny do shluků. **Silueta:** hodnota siluety by měla být kladná a větší než 0.5. **Čárový diagram hodnot siluety:** pomůže rozlišit řádky, objekty, které nejsou dobře zařazeny do shluků.

5. Účast objektů ve shlucích: ukazuje na pravděpodobnost účasti každého řádku v každém shluku.

Řádek	Shluk	Pravděpod. v 1	Pravděpod. v 2	Pravděpod. v 3
1	1	0.8705	0.0572	0.0723
2	1	0.8728	0.0601	0.0671
3	1	0.9354	0.0289	0.0357
4	1	0.8535	0.0613	0.0852
5	1	0.8779	0.0558	0.0663
6	1	0.4227	0.3620	0.2154
7	2	0.0860	0.8231	0.0909
8	2	0.0659	0.8680	0.0661
9	2	0.0673	0.8508	0.0819
10	2	0.0635	0.8705	0.0660
11	2	0.0648	0.8556	0.0797
12	2	0.0745	0.8402	0.0853
13	1	0.3670	0.2821	0.3509
14	3	0.1135	0.0865	0.8000
15	3	0.0803	0.0714	0.8483

16	3	0.0973	0.1015	0.8012
17	3	0.0751	0.0602	0.8647
18	3	0.0415	0.0377	0.9209
19	3	0.0662	0.0686	0.8652
20	3	0.0950	0.0792	0.8258
21	3	0.0694	0.0642	0.8663
22	3	0.0858	0.0884	0.8258

4.7 Vícerozměrné škálování MDS

Vícerozměrné škálování (**M**ulti**D**imensional **S**caling, MDS) je technika vytvoření diagramu relativního umístění objektů v rovině dvojrozměrného grafu na základě dat vzdáleností mezi objekty, tzv. *matice proximity* (blízkosti). Diagram může obsahovat jeden, dva, tři a zřídka i více rozměrů, dimenzí. Technika vyčíslí metrické klasické (CMDS) nebo nemetrické (NNMDS) řešení a vychází buď přímo z experimentálních hodnot ***X***, z korelační matice ***R*** nebo z matice podobností ***S*** či vzdáleností ***D***. Vzdálenost mezi oběma objekty je Eukleidovská, počítaná na základě Pythagorovy věty, $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$, kde m je počet proměnných a x_{ik} jsou data i-tého řádku a k-tého sloupce. I když vynášíme vzdálenosti do dvojrozměrného grafu, může být d_{ij} vyčísleno na základě většího počtu proměnných $m \geq 2$. Matice vzdáleností je potom trojúhelníková a zajímá nás jenom její horní část. S růstem objektů však roste i počet dimenzí, takže pro *tři* objekty je to *dvoj*-rozměrná rovina, pro *čtyři* objekty pak *troj*-rozměrný prostor atd.

Kritérium maximální věrohodnosti. Jak těsně prokládá model vzdáleností daná experimentální data se hodnotí *testem těsnosti proložení* s využitím statistického kritéria *stress*, založeného na rozdílu mezi skutečnou vzdáleností d_{ij} a modelem predikovanou hodnotou $\hat{d}_{ij}$,

$$stress = \sqrt{\frac{\sum_{k=1}^{m} (d_{ij} - \hat{d}_{ij})^2}{\sum_{k=1}^{m} d_{ij}^2}},$$

kde $\hat{d}_{ij}$ je predikovaná vydálenost, založená na MDS modelu. Predikovaná hodnota závisí především na počtu užitých dimenzí a algoritmu, a to metrickém či nemetrickém. Je-li *stress* číslo nízké, blízké nule, jeví se MDS proložení jako nejlepší.

Počet dimenzí. Důležitým úkolem v MDS je určení počtu dimenzí v MDS modelu. Každá dimenze zde představuje latentní proměnnou. Cílem MDS je udržet počet dimenzí na co možná nejmenší hodnotě. Obvykle volí uživatel dvoj- maximálně trojrozměrný prostor. Vychází-li vyšší počet dimenzí, není MDS technika k analýze dotyčných dat vhodná. Počet dimenzí se volí na základě co nejmenší hodnoty kritéria *stress*. Někteří autoři

si pomáhají indexovým grafem relativní velikosti vlastních čísel, která jsou vyčíslována pro rostoucí počet dimenzí, tzv. *grafem úpatí*. Postup a inter-pretace jsou pak stejné jako u metod PCA nebo FA.

Vstupní data. Data mohou být trojího typu, mohou obsahovat (1) vzdálenosti mezi objekty ***D***, (2) podobnost mezi objekty ***S*** nebo (3) hodnoty proměnných (sloupce) pro jednotlivé objekty (řádky) ***X***.

Vzdálenost (disimilarita) d_{ij} představující vzdálenost mezi objekty, může být měřena přímo, jako např. vzdálenost dvou měst. MDS užívá vzdálenost v datech přímo a matice vzdáleností ***D*** je symetrická.

Podobnost (similarita) s_{ij} vyjadřuje, jak blízko se nacházejí dva objekty. MDS umožňuje načíst míry podobnosti pro každý pár objektů. Matice podobností ***S*** je opět symetrická. Podobnost lze konvertovat do veličiny vzdálenosti vzorcem

$$d_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}},$$

kde d_{ij} představuje vzdálenost a s_{ij} podobnost.

Hodnoty x_{ij} proměnných pro jednotlivé objekty představují spíše standardní míry. Z nich se vypočte nejprve korelační matice ***R*** a potom matice Eukleidovských či Mahalanobisových vzdáleností ***D***.

Klasická metrická metoda MDS. Je dána matice vzdáleností ***D***, která vystihuje meziobjektové vzdálenosti objektů ***X*** v prostoru spíše nižšího rozměru dle vzorce

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}.$$

Jednotlivé kroky klasické MDS jsou následující:

1. Z ***D*** se vypočte $\boldsymbol{A} = \{-0.5\ d_{ij}^2\}$.
2. Z ***A*** se vypočte $\boldsymbol{B} = \{a_{ij} - a_{i.} - a_{.j} + a_{..}\}$, kde $a_{i.}$ je průměr všech a_{ij} přes j.
3. Nalezne se m největších vlastních čísel $\lambda_1 > \lambda_2 > ... > \lambda_m$ matice ***B*** a odpovídající vlastní vektory $\boldsymbol{L} = \boldsymbol{L}_{(1)}, \boldsymbol{L}_{(2)}, ..., \boldsymbol{L}_{(m)}$, které jsou normovány, takže $\boldsymbol{L}^{\mathrm{T}}_{(i)}\, \boldsymbol{L}_{(i)} = \lambda_i$. Předpokládáme, že m je voleno tak, že vlastní hodnoty jsou relativně velké a kladné.
4. Souřadnicemi objektů jsou řádky matice ***L***.

Klasické řešení je optimalizováno metodou nejmenších čtverců: přímé řešení ***L*** minimalizuje sumu čtverců vzdáleností mezi skutečnými prvky matice ***D***, tj. d_{ij} a predikcemi $\hat{d}_{ij}$, založenými na ***L***. Předpokládejme, že experimentální hodnoty vzdálenosti d_{ij} jsou zatíženy náhodnou chybou ε_{ij} dle vzorce $d_{ij} = \delta_{ij} + \varepsilon_{ij}$, kde ε_{ij} představuje kombinaci náhodných chyb z měření, distorze vzdáleností, když MDS model zcela neodpovídá konfiguraci navržených m vzdáleností. Navrhněme model závislosti mezi vzdáleností dvou objektů vztahem $d_{ij} = \beta_0 + \beta_1\, \delta_{ij} + \varepsilon_{ij}$ a potom nalezením nejlepších odhadů b_0 pro β_0 a b_1 pro β_1 obdržíme odhad vypočtené vzdálenosti $d_{ij} = b_0 + b_1\, \delta_{ij}$. Optimalizační procedura vychází z účelové funkce

$$U = \sum_{i<j}^{n} (d_{ij} - \hat{d}_{ij})^2 \rightarrow \min.$$

Aby byla zajištěna úplná invariantnost vůči transformaci proměnných, užívá se modifikovaná účelová funkce U_{mod} dle vztahu

$$U_{\mathrm{mod}} = \frac{\sum_{i<j}^{n} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j}^{n} d_{ij}^2}$$

a především její druhá odmocnina, zvaná *stress* $= \sqrt{U_{\mathrm{mod}}}$. Je proto výhodné hledat optimální počet dimenzí, která se vezmou k vyčíslení predikce MDS vzdálenosti $\hat{d}_{ij}$ pomocí minimální hodnoty veličiny *stress*. Pro *stress* < 0.05 je těsnost proložení ještě přijatelná a pro *stress* < 0.01 je těsnost proložení výtečná.

Nemetrická MDS. V dosavadnín postupu se předpokládalo, že vzdálenosti jsou vyčísleny metricky. Jsou však situace, kdy jedna hodnota nevystihuje dostatečně skutečnost: např. při porovnávání barev na stupnici může být jedna barva zářivější než druhá, a tento fakt však nikterak neovlivní polohu barvy na stupnici. Predikované vzdálenosti d_{ij} jsou vyčíslovány *monotónní regresí:* experimentální vzdálenosti jsou uspořádány vzestupně do řady

$$d_{i1,j1} \leq d_{i2,j2} \leq \ldots \leq d_{iN,jN}, \text{ kde } N = n(n-1)/2$$

a d_{ij} jsou odhadovány tak, aby splnily podmínku *slabé monotonicity (WM)*

$$d_{i1,j1} \leq \hat{d}_{i2,j2} \leq \ldots \leq \hat{d}_{iN,jN}, \text{ nebo}$$

nebo podmínku *silné monotonicity (SM)*

$$d_{i1,j1} < \hat{d}_{i2,j2} < \ldots < \hat{d}_{iN,jN}.$$

Prvním krokem k získání počátečních odhadů predikovaných vzdáleností d_{ij} bývá vždy metrické vyčíslení. Pak následuje nemetrický přístup monotónní regrese. Indexový graf úpatí veličiny *stress* je užitečnou pomůckou i u nemetrické metody. Hledá se jednak zlom na tomto grafu a jednak se vyšetřuje, kdy veličina *stress* nabyde hodnot menších než 0.05, resp. 0.01. Takový index, čili počet dimenzí, se pak jeví jako optimální. Obdobně, jako metrická metoda CMDS, ústí i nemetrická NNMDS ve vícerozměrnou škálovací mapu, na které se sleduje roztřídění vyšetřovaných objektů.

Vzorová úloha 4.15 *Vícerozměrné škálování u analýzy podobnosti*

Vícerozměrné škálování ukážeme na datech **Úlohy E4.17** *Vícerozměrná škálování u analýzy podobnosti 10 výrobků Coly.* Často je třeba vyhodnotit úroveň veřejného mínění, shodu předmětu zájmu s jiným předmětem. Vícerozměrným škálováním posoudíme podobnost 10 výrobků Coly, a to na základě výsledků ankety: 50 respondentů hodnotilo

a vzájemně porovnalo 10 výrobků Coly (objekty) způsobem "každý s každým" a při dokonalé podobnosti byla přidělena nulová vzdálenost mezi dvěma objekty, zatímco při naprosté nepodobnosti vzdálenost 100. Z hodnot párových vzdáleností od 50 respondentů byla vždy vypočtena střední hodnota a zapsána do buňky vytvořené symetrické čtvercové matice. Z této matice se ve vstupních datech užije pouze trojúhelníková část, tj. prvky nad (nebo pod) diagonálou nul. Je třeba provést *dvojrozměrné škálování* a z výsledného grafu usoudit na podobné a nepodobné výrobky Coly. Aplikujte metodu klasického metrického škálování CMDS a porovnejte se závěry nemetrického škálování NNMDS, str. 95, ref.[30].

Řešení:

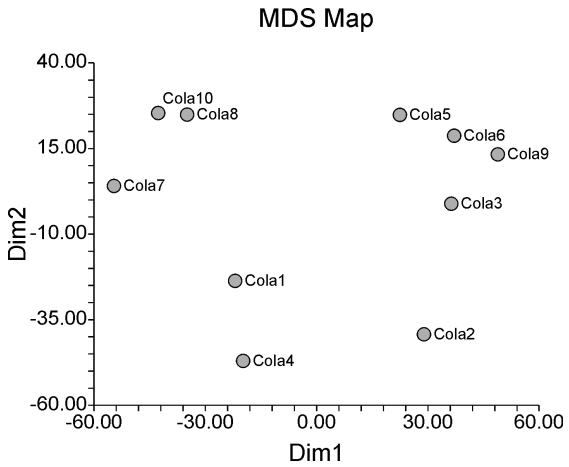
Veřejnost by v původních datech měla odhadnout míru shodnosti či podobnosti vždy mezi dvěma druhy Coly tak, že odhadne vzdálenost (čili nepodobnost) mezi nimi: naprosto stejné nápoje budou mít mezi sebou vzdálenost nula a naprosto odlišné vzdálenost 100. Těchto 45 hodnot trojúhelníkové matice vede k diagramu na obr. 4.18.

Oddíl vlastních čísel (NCSS2000)

Číslo	**Vlastní číslo**	**Individual procento**	**Kumulativní procento**	**Čárový diagram**
1	13328.04	34.21	34.21	|IIIIIIIIIIIIIII
2 (Used)	6737.70	17.29	51.50	|IIIIIIII
3	5307.69	13.62	65.13	|IIIIII
4	3387.91	8.70	73.82	|IIII
5	1670.85	4.29	78.11	|II
6	14.90	0.04	78.15	|
7	0.00	0.00	78.15	|
8	-2135.51	5.48	83.63	|III
9	-2894.18	7.43	91.06	|IIII
10	-3483.58	8.94	100.00	|IIII
Celkově	38960.35			
Stress	0.071646			
Pseudo R^2	57.006058			

Dvojrozměrný diagram rozložení 10 druhů nápoje Cola představuje hlavní výsledek vícerozměrného škálování. Často je nazýván *vícerozměrnou škálovací mapou* a umožňuje interpretovat matici vzdáleností mezi objekty ve dvojrozměrném diagramu. Neexistuje reálná orientace tohoto diagramu, diagramem je totiž možné libovolně otáčet okolo počátku. Důležité jsou relativní polohy objektů vůči sobě a pak hlavně poloha shluků objektů.

Obr. 4.23 ukazuje, že jednotlivé druhy Coly jsou zřetelně roztříděny v rovině. Cola 3, 5, 6 a 9 tvoří jeden shluk druhů podobných vlastností, dále Cola 8 a 10 spolu s Colou 7 pak druhý shluk. Odlišné jsou Cola 1, 2, a 4, které se od předešlých dvou velmi odlišují, navíc Cola 1 se značně liší od Coly 2 a Cola 2 se značně liší od Coly 4. První shluk má dominantu Colu 9 a druhý shluk pak Colu 10. Okolo těchto dominant jsou soustředěny ostatní.



Obr. 4.23 MDS dvojrozměrný diagram rozložení 10 druhů nápoje cola (dvojrozměrná škálovací mapa).

Závěr: Cílem MDS bývá také pojmenování obou os vícerozměrné škálovací mapy, a pokud je to možné, i následující výklad poloh objektů na mapě.

4.8 Vícerozměrná kalibrace

Podobně, jako jednorozměrná kalibrace, i vícerozměrná kalibrace se používá především v analytické chemii. Bude vysvětlena na příkladu spektroskopie: cílem je popis závislosti mezi blokem naměřených dat, např. spektrem $\boldsymbol{X}$, a veličinami kalibračního modelu, např. koncentracemi $\boldsymbol{C}$ látek ve směsi. Aplikace totiž umožňuje stanovit hodnoty koncentrací jednotlivých složek v neznámém vzorku ($\boldsymbol{c}^*$) z jeho naměřeného spektra ($\boldsymbol{x}^*$). Existují dvě alternativy vícerozměrné kalibrace stejné jako u jednorozměrné, a to tzv. *klasická kalibrace* a *kalibrace inverzní.*

4.8.1 Klasická vícerozměrná kalibrace

Uplatňuje se v případě, kdy všechny sledované látky směsi jsou známy stejně jako látky, které s nimi mohou interferovat. Navíc jsou látky i interferenty fyzicky k dispozici. Pak jsou připraveny umělé standardy a proměřeny při uvažovaných m vlnových délkách. Matice absorbancí $\boldsymbol{X}$ je výsledkem přítomnosti p látek o koncentracích $\boldsymbol{C}$ s absorpčními koeficienty $\boldsymbol{\beta}$ dle vztahu

$$\boldsymbol{X} = \boldsymbol{C}\,\boldsymbol{\beta} + \boldsymbol{E}_x$$

$$(n,\ m)\quad (n,\ p)\quad (p,\ m)\quad (n,\ m)$$

$$\begin{vmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{vmatrix} = \begin{vmatrix} c_{11} & \dots & c_{1p} \\ c_{21} & \dots & c_{2p} \\ \dots & \dots & \dots \\ c_{n1} & \dots & c_{np} \end{vmatrix} \times \begin{vmatrix} \beta_{11} & \dots & \beta_{1m} \\ \beta_{21} & \dots & \beta_{2m} \\ \dots & \dots & \dots \\ \beta_{p1} & \dots & \beta_{pm} \end{vmatrix} + \begin{vmatrix} e_{11} & \dots & e_{1m} \\ e_{21} & \dots & e_{2m} \\ \dots & \dots & \dots \\ e_{n1} & \dots & e_{nm} \end{vmatrix}$$

kde x_{ij} je absorbance vzorku i při vlnové délce j, c_{ir} je koncentrace látky r ve vzorku i, β_{ij} je molární absorpční koeficient látky r při vlnové délce j a e_{ij} je chyba absorbance pro vzorek i při vlnové délce j.

Rovnice je řešitelná metodou nejmenších čtverců MNČ, přičemž se minimalizuje součet čtverců reziduí v matici $\boldsymbol{E}_x$. Výsledkem je odhad neznámých parametrů β klasického kalibračního modelu

$$\boldsymbol{\beta} = (\boldsymbol{C}^T\,\boldsymbol{C})^{-1}\,\boldsymbol{C}^T\,\boldsymbol{X}$$

$$(p,\ m)\quad (p,\ n)\ (n,\ p)\ (p,\ n)\ (n,\ m)$$

Každý r-tý řádek matice $\boldsymbol{\beta}$ reprezentuje absorpční spektrum látky r při jednotkové koncentraci a jednotkové délce kyvety čili molární absorpční koeficient.

Znalost $\boldsymbol{\beta}$ dovoluje stanovit koncentraci všech p látek přítomných v neznámém vzorku, tj. $\boldsymbol{c}^*$, na základě jeho naměřeného spektra $\boldsymbol{x}^*$ dle vztahu

$$\boldsymbol{c}^{(T} = (\boldsymbol{\beta}\,\boldsymbol{\beta}^T)^{-1}\,\boldsymbol{\beta}\,\boldsymbol{x}^{(T}$$

$$(p,\ 1)\quad (p,\ m)\ (m,\ p)\ (p,\ m)\ (m,\ 1)$$

Ve srovnání s jednorozměrnou kalibrací má klasická vícerozměrná kalibrace následující výhody: (1) absorbanční měření $\boldsymbol{X}$, použitá ke kalibraci, nemusí být selektivní vzhledem k $\boldsymbol{C}$, čistá spektra stanovovaných látek se mohou totiž libovolně překrývat, (2) přesnost stanovení koncentrací neznámých vzorků je lepší, (3) dovoluje automatickou korekci na změny průběhu spektrální základní linie a (4) dovoluje odhadnout *čistá* a *reziduální spektra* stanovovaných látek. Znalost obou těchto spekter může vést k lepšímu pochopení studovaného kalibrovaného systému.

Hlavní nevýhodou klasické vícerozměrné kalibrace je omezení na případy, kdy všechny látky směsi musí být známy včetně interferentů a musí být fyzicky k dispozici. Vzhledem

ke vzácnému splnění těchto předpokladů se klasická vícerozměrná kalibrace v praxi příliš nepoužívá. Naopak, značně se v praxi rozšířila tzv. *inverzní vícerozměrná kalibrace.*

4.8.2 Inverzní vícerozměrná kalibrace

Na rozdíl od klasické vícerozměrné kalibrace jsou požadavky pro aplikaci inverzní vícerozměrné kalibrace minimální. Vyžaduje znalost koncentrace sledované látky v kalibračních vzorcích. Kalibrační vzorky mohou být komplexní směsi se složitou matricí a neznámými interferenty, např. přírodní, potravinářské nebo petrochemické materiály. Pro vzorky jsou zaznamenána spektra absorbance $\boldsymbol{X}$ a sestrojen kalibrační model, popisující vztah mezi koncentrací sledované látky $\boldsymbol{c}$ a absorbancí $\boldsymbol{X}$ rovnicí

$$\boldsymbol{c} \;=\; \boldsymbol{X}\;\boldsymbol{\beta} \;+\; \boldsymbol{e}_c$$

$$(n,\,1)\quad (n,\,m)\quad (m,\,1)\quad (n,\,1)$$

$$\begin{vmatrix} c_1 \\ c_2 \\ \dots \\ c_n \end{vmatrix} = \begin{vmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{vmatrix} \times \begin{vmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_m \end{vmatrix} + \begin{vmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{vmatrix}$$

kde c_i je koncentrace sledované látky v kalibračním vzorku i, x_{ij} je absorbance tohoto vzorku při j-té vlnové délce, β_j je regresní koeficient čili molární absorpční koeficient pro vlnovou délku j a e_i je koncentrační reziduum vzorku i, čili rozdíl mezi koncentrací stanovenou referenční metodou a koncentrací vypočtenou proložením kalibračních dat modelem.

Místo maticového zápisu lze kalibrační model prezentovat i vektorově

$$\boldsymbol{c} \;=\; \beta_1\,\boldsymbol{x_1} \;+\; \beta_2\,\boldsymbol{x_2} \;+\; \dots \;+\beta_m\,\boldsymbol{x_m} \;+\; \boldsymbol{e}_c \quad .$$

Odhad parametrů modelu metodou nejmenších čtverců MNČ je dán vztahem

$$\boldsymbol{\beta} \;=\; (\boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{c}$$

$$(m,\,1)\quad (m,\,n)\;(n,\,m)\;(m,\,n)\;(n,\,1)$$

Řešení existuje pouze v případě, když počet kalibračních vzorků n je větší než počet vlnových délek m spektra, $m>n$. Jinak nelze invertovat součin $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ a tím ani odhadnout parametry $\boldsymbol{\beta}$. Moderní analytické přístroje umožňují naměřit spektrum při stovkách vlnových délek a zmíněný předpoklad je pak zásadním omezením použitelnosti metody. Byly proto navrženy dvě strategie jak postupovat, pokud uvedená podmínka není splněna:

1) Z původního spektra vhodnou metodou vybrat omezený počet vlnových délek tak, aby podmínka byla splněna.

2) Zkomprimovat původní spektra získaná při stovkách vlnových délek do formy latentních proměnných, a ty pak použít ke kalibraci namísto původních absorbančních proměnných.

1. Selekce originálních proměnných

Klasickou metodou k selekci proměnných z původního výběru všech nezávislých proměnných spektra je postupná *kroková vícenásobná lineární regrese*. Alternativních metod je několik, přičemž nejznámější metodou je *genetický algoritmus*.

Postupná kroková vícenásobná lineární regrese má následující postup:

(i) V prvním kroku se využitím korelačního koeficientu v absorbanční matici $\boldsymbol{X}$ nalezne proměnná $\boldsymbol{x}_j$, která je nejvíce korelována s kalibrovanou koncentrací $\boldsymbol{c}$. Pak se sestrojí jednorozměrný kalibrační model

$$\boldsymbol{c} = \beta_0 + \beta_j \boldsymbol{x}_j + \boldsymbol{e}_c \quad .$$

Testuje se statistická významnost odhadnuté směrnice tohoto modelu β_j. Pokud je parametr β_j shledán statisticky významným, selekce proměnných v modelu pokračuje.

(ii) Ve druhém kroku je vybrána ta proměnná $\boldsymbol{x}_l$, která má nejvyšší parciální korelační koeficient s koncentrací $\boldsymbol{c}$, a to je ta, která má nejvyšší korelační koeficient s vektorem reziduí $\boldsymbol{e}_c$, získaným z první regrese. Postup se nazývá *dopředná selekce parametrů lineárního regresního modelu*. Pak se sestrojí lineární regresní model

$$\boldsymbol{c} = \beta_0 + \beta_j \boldsymbol{x}_j + \beta_l \boldsymbol{x}_l + \boldsymbol{e}_c$$

a testuje se statistická významnost obou odhadnutých regresních parametrů β_j a β_l. Nevýznamné členy jsou z rovnice eliminovány. Postup se pak nazývá *zpětná eliminace parametrů v lineárním regresním modelu*.

(iii) Dopředná selekce a zpětná eliminace parametrů regresního modelu jsou opakovány tak dlouho, až model již nelze zlepšit přidáním nějaké proměnné a současně všechny proměnné zahrnuté v modelu jsou statisticky významné. Takové řešení je pak považováno za optimální.

Genetický algoritmus je obecný optimalizační nástroj, simulující proces přírodní evoluce, kde přežívá ten jedinec, který se nejlépe hodí do okolního prostředí. Tento jedinec se pak dále rozvíjí a adaptuje. Analogicky ve vícerozměrné kalibraci hledají genetické algoritmy optimální kombinaci původních proměnných. Začíná se od náhodně zvolených kombinací těchto proměnných. U počátečních kombinací se testuje schopnost popsat variabilitu závisle proměnné koncentrace **c**. S pravděpodobností odvozenou od této schopnosti jsou pak nejlepší řešení podrobena dvěma operacím: míšení a mutaci. *Míšení* je řízené promíchání nejlepších řešení z posledního cyklu. *Mutace* simuluje proces náhodné změny, která se v přírodě vyskytuje s nižší frekvencí. Celá procedura má iterativní charakter a je založena na srovnání vícenásobných lineárních modelů pro všechny testované kombinace.

2. Redukce originálních proměnných metodami s latentními proměnnými

Latentní proměnné jsou lineárními kombinacemi všech původních proměnných (viz kap. 4.5.1). Sloupcový vektor *i*-té latentní proměnné y_i je matematicky definován jako

$$y_i = w_{i1} \boldsymbol{x}_1 + w_{i2} \boldsymbol{x}_2 + \ldots + w_{i1} \boldsymbol{x}_i + \ldots + w_{im} \boldsymbol{x}_m + \ldots \quad ,$$

kde i nabývá hodnot od 1 do min(n, m). Koeficienty w_{ij} odpovídají zátěži j-té původní proměnné do i-té latentní proměnné a $\boldsymbol{x}_j$ je sloupcový vektor absorbancí, měřených při vlnové délce j. Soubor latentních proměnných dává matici latentních proměnných $\boldsymbol{Y}$.

Smyslem konstrukce latentních proměnných je zkomprimovat informaci obsaženou ve velkém počtu původních proměnných v absorbanční matici do několika málo nových proměnných. Důležitou vlastností latentních proměnných je jejich vzájemná ortogonalita čili kolmost. Tato vlastnost zaručuje, že metoda nejmenších čtverců MNČ, aplikovaná na soubor latentních proměnných, poskytne matematicky stabilní řešení a přítomnost či absence určité latentní proměnné neovlivní hodnoty regresních parametrů ostatních členů. Nejpoužívanější metodou konstrukce latentních proměnných je *metoda hlavních komponent* PCA a *metoda částečných nejmenších čtverců* PLS (anglicky Partial Least Squares):

A. Využití metody hlavních komponent. Nejběžnější metodou, umožňující konverzi původních proměnných na latentní, je *metoda hlavních komponent* PCA a vzniklé proměnné se nazývají *hlavní komponenty.* Kalibrace závisle proměnné $\boldsymbol{c}$ pomocí metody nejmenších čtverců MNČ, aplikovaná na malý, správně zvolený podsoubor hlavních komponent, se nazývá *regrese na hlavních komponentách* (Principal Component Regression PCR). Metodou hlavních komponent lze zdrojovou absorbanční matici měření $\boldsymbol{X}$ rozložit dle vztahu

$$\underset{(n,\ m)}{\boldsymbol{X}} \ ' \ \underset{(n,\ n)}{\boldsymbol{V}} \ \underset{(n,\ n)}{\boldsymbol{\Gamma}} \ \underset{(n,\ m)}{\boldsymbol{W}^{\mathrm{T}}} \quad ,$$

kde $\boldsymbol{V}$ je matice *normovaných skóre,* diagonální matici vlastních čísel $\boldsymbol{\Gamma}$ nazýváme *maticí zátěží,* matice $\boldsymbol{W}$ pak zajišťuje transformaci původních proměnných na latentní. Vlastní čísla obsažená v $\boldsymbol{\Gamma}$ udávají důležitost jednotlivých vektorů matice $\boldsymbol{V}$ vzhledem ke zdrojové matici $\boldsymbol{X}$, tj. příspěvek jednotlivých latentních proměnných $\boldsymbol{y}_i$ do celkové variability $\boldsymbol{X}$: $\boldsymbol{y}_1$ popisuje nejvíce variability, $\boldsymbol{y}_2$ méně atd. Součin $\boldsymbol{V}$ a $\boldsymbol{\Gamma}$ dává matici nenormovaných skóre $\boldsymbol{Y}$

$$\mathrm{Y} \ ' \ \boldsymbol{V} \ \boldsymbol{\Gamma}$$

a i-tý sloupec matice $\boldsymbol{Y}$ obsahuje hodnoty skóre pro všechny kalibrační vzorky na i-té latentní proměnné. Jelikož matice $\boldsymbol{Y}$ má rozměr (n, n), lze z ní nebo z podsouboru k sloupcových vektorů aplikací metody nejmenších čtverců sestrojit model pro koncentraci $\boldsymbol{c}$ dle vztahu:

$$\underset{(n,\ 1)}{\boldsymbol{c}} \ ' \ \underset{(n,\ k)}{\boldsymbol{Y}} \ \underset{(k,\ 1)}{\boldsymbol{q}} \ \% \ \underset{(n,\ 1)}{\boldsymbol{e}_c} \quad .$$

Tato rovnice má řešení

$$\underset{(k,\ 1)}{\boldsymbol{q}} \ ' \ \underset{(k,\ n)}{(\underset{(n,\ k)}{\boldsymbol{Y}^{\mathrm{T}}} \underset{(k,\ n)}{\boldsymbol{Y}})^{\&1}} \ \underset{(n,\ 1)}{\boldsymbol{Y}^{\mathrm{T}}} \ \underset{(n,\ 1)}{\boldsymbol{c}} \quad .$$

Vedle metody hlavních komponent PCA je zde regrese metodou nejmenších čtverců MNČ druhým základním krokem v metodě regrese na hlavních komponentách PCR.

Protože několik málo k prvních hlavních komponent vysvětluje většinu variability v matici $\boldsymbol{X}$, vyšší komponenty k+1 atd. není třeba při konstrukci modelu vůbec uvažovat. Ty totiž popisují *šum v datech*.

Hledání a optimalizace počtu k latentních proměnných je jedním z nejdůležitějších úkolů ve vícerozměrné inverzní kalibraci. Jelikož model regresí na hlavních komponentách PCR není sestrojen na základě původních, ale latentních proměnných, je před stanovením koncentrace neznámého vzorku $\boldsymbol{c}^*$ z jeho spektra třeba vypočítat skóre vzorku na jednotlivých latentních proměnných. Výpočet se provede pomocí zátěží $\boldsymbol{W}$

$$\underset{(1,\,k)}{\boldsymbol{y}^{\mathrm{T}}} = \underset{(1,\,m)}{\boldsymbol{x}^{\mathrm{T}}}\,\underset{(m,\,k)}{\boldsymbol{W}} \,.$$

Získaná skóre se nakonec použijí k odhadu koncentrace stanovovaného vzorku

$$\underset{(1,\,1)}{c^{\mathrm{T}}} = \underset{(1,\,k)}{\boldsymbol{y}^{\mathrm{T}}}\,\underset{(k,\,1)}{\boldsymbol{q}} \,.$$

B. Částečné nejmenší čtverce (PLS). Hlavní nevýhodou metody regrese na hlavních komponentách PCR je fakt, že používá ke konstrukci modelu všech k prvních hlavních komponent. Jelikož některé z nich mohou popisovat zdroj variability v datech, který však není svázán s koncentrací $\boldsymbol{c}$, může při použití takového modelu dojít k nesprávné predikci. Tento problém lze potlačit selekcí latentních proměnných na základě korelace s koncentrací $\boldsymbol{c}$ nebo aplikací *metody částečných nejmenších čtverců* PLS. Tato metoda se při konstrukci latentních proměnných zaměřuje na tu část variability ve spektrech $\boldsymbol{X}$, která je korelována s koncentrací $\boldsymbol{c}$. Popis činnosti vhodného iterativního algoritmu částečných nejmenších čtverců NIPALS obsahuje následující kroky:

(1) Centrování zdrojové absorbanční matice $\boldsymbol{X}$ a kalibrované proměnné $\boldsymbol{c}$.

(2) Rozklad absorbanční matice $\boldsymbol{X}$ za použití vah $\boldsymbol{w}$ maximalizujících kovarianci $\boldsymbol{X}^{\mathrm{T}}\,\boldsymbol{c}$

$$\underset{(n,\,m)}{\boldsymbol{X}} = \underset{(n,\,1)}{\boldsymbol{c}}\,\underset{(1,\,m)}{\boldsymbol{w}_{\mathrm{iter}}^{\mathrm{T}}} + \underset{(n,\,m)}{\boldsymbol{E}_X} \,,$$

kde *iter* značí číslo iterace. Rovnici lze řešit metodou nejmenších čtverců s výsledkem

$$\underset{(m,\,1)}{\boldsymbol{w}_{\mathrm{iter}}} = \underset{(1,\,n)}{(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{c})}\,\underset{(n,\,1)}{\boldsymbol{X}^{\mathrm{T}}\,\boldsymbol{c}} \,.$$

Vektor $\boldsymbol{w}_{\mathrm{iter}}$ se skládá z vah pro každou vlnovou délku j. V první iteraci je tento krok identický s klasickou vícerozměrnou kalibrací za předpokladu, že je známá koncentrace pouze jedné látky. Veličinu $\boldsymbol{w}_{\mathrm{iter}}$ lze interpretovat jako spektrum dotyčné látky.

(3) Normalizace vah $\boldsymbol{w}_{\mathrm{iter}}$: první normalizovaný vektor vah $\boldsymbol{w}_1$ je úměrný váženému průměru centrovaných spekter $\boldsymbol{X}$.

(4) Rozklad absorbanční matice $\boldsymbol{X}$ s cílem získat $\boldsymbol{X}$ skóre, označená $\boldsymbol{Y}$. Jde o krok ekvivalentní prvnímu kroku v metodě regrese na hlavních komponentách

$$\underset{(n,\,m)}{\boldsymbol{X}} = \underset{(n,\,1)}{\boldsymbol{y}_m}\,\underset{(1,\,m)}{\boldsymbol{w}_{\mathrm{iter}}^{\mathrm{T}}} + \underset{(n,\,m)}{\boldsymbol{E}_X} \,,$$

s řešením metodou nejmenších čtverců MNČ

$$\underset{(n,\ 1)}{\boldsymbol{y}_{\mathbf{iter}}} \leftarrow \underset{(1,\ p)}{(\boldsymbol{w}_{\mathbf{iter}}^{\mathbf{T}}\ \boldsymbol{w}_{\mathbf{iter}})^{-1}}\ \underset{(n,\ p)}{\boldsymbol{X}}\ \underset{(p,\ 1)}{\boldsymbol{w}_{\mathbf{iter}}} .$$

Tento krok je srovnatelný s predikcí v klasické kalibraci. Protože vektor vah $\boldsymbol{w}_{\text{iter}}$ je funkcí $\boldsymbol{c}$, $\boldsymbol{Y}$ v metodě částečných nejmenších čtverců PLS, reprezentuje společnou část $\boldsymbol{X}$ a $\boldsymbol{c}$.

(5) Stanovení *vnitřního vztahu metody částečných nejmenších čtverců* mezi $\boldsymbol{y}_{\mathbf{iter}}$ a $\boldsymbol{c}$

$$\underset{(1,\ n)}{\boldsymbol{c}^{\mathbf{T}}} \leftarrow \underset{(1,\ 1)}{q_{\text{iter}}}\ \underset{(1,\ n)}{\boldsymbol{y}_{\mathbf{iter}}^{\mathbf{T}}} + \underset{(1,\ n)}{\boldsymbol{e}_c} .$$

Rovnice je řešitelná metodou nejmenších čtverců MNČ a je ekvivalentní regresnímu kroku v regresi na hlavních komponentách PCR s výsledkem

$$\underset{(1,\ 1)}{q_{\text{iter}}} \leftarrow \underset{(1,\ n)}{(\boldsymbol{y}_{\mathbf{iter}}^{\mathbf{T}}\ \boldsymbol{y}_{\mathbf{iter}})^{-1}}\ \underset{(n,\ 1)}{\boldsymbol{c}^{\mathbf{T}}}\ \underset{(n,\ 1)}{\boldsymbol{y}_{\mathbf{iter}}} .$$

Na rozdíl od metody regrese na hlavních komponentách PCR, q_{iter} se určí iterativní metodou.

(6) Stanovení $\boldsymbol{X}$ zátěží $\mathbf{p}_{\mathbf{iter}}$:

$$\underset{(n,\ m)}{\boldsymbol{X}} \leftarrow \underset{(n,\ 1)}{\boldsymbol{y}_{\mathbf{iter}}}\ \underset{(1,\ m)}{\boldsymbol{p}_{\mathbf{iter}}^{\mathbf{T}}} + \underset{(n,\ m)}{\boldsymbol{E}_X}$$

a rovnice je řešitelná metodou nejmenších čtverců MNČ

$$\underset{(m,\ 1)}{\boldsymbol{p}_{\mathbf{iter}}} \leftarrow \underset{(1,\ n)}{(\boldsymbol{y}_{\mathbf{iter}}^{\mathbf{T}}\ \boldsymbol{y}_{\mathbf{iter}})^{-1}}\ \underset{(n,\ 1)}{\boldsymbol{X}^{\mathbf{T}}}\ \underset{(n,\ 1)}{\boldsymbol{y}_{\mathbf{iter}}} ,$$

zátěže $\boldsymbol{p}_{\mathbf{iter}}$ mají stejnou funkci jako váhy $\boldsymbol{w}_{\mathbf{iter}}$ a jsou zapotřebí pouze ke korekci faktu, že $\boldsymbol{y}_{\mathbf{iter}}$ nejsou známy a musely být nahrazeny odhadem.

(7) iter-tá aproximace spekter častečnými nejmenšími čtverci PLS a koncentrace je dána vztahy

$$\underset{(n,\ m)}{\boldsymbol{E}_x} \leftarrow \underset{(n,\ m)}{\boldsymbol{X}} - \underset{(n,\ 1)}{\boldsymbol{y}_{\mathbf{iter}}}\ \underset{(1,\ m)}{\boldsymbol{p}_{\mathbf{iter}}^{\mathbf{T}}} ,$$

$$\underset{(n,\ 1)}{\boldsymbol{e}_c} \leftarrow \underset{(n,\ 1)}{\boldsymbol{c}} - \underset{(1,\ 1)}{q_{iter}}\ \underset{(n,\ 1)}{\boldsymbol{y}_{iter}} ,$$

rezidua $\boldsymbol{E}_x$ a $\boldsymbol{e}_c$ přebírají roli původního $\boldsymbol{X}$ a $\boldsymbol{c}$ a celá procedura se iterativně opakuje od kroku (ii). To se provádí k-krát, resp. tak dlouho, dokud nedojde ke konvergenci. Konvergence nastává, když veškerá kovariance mezi $\boldsymbol{E}_x$ a $\boldsymbol{e}_c$ je popsána modelem.

Postup výpočtu koncentrací neznámých vzorků se provede následovně:

(a) Zkonstruuje se výsledný model částečnými nejmenšími čtverci PLS, tj. odhadnou se parametry β, popisující přímo vztah mezi $\boldsymbol{c}$ a $\boldsymbol{X}$

$$\underset{(m,\ 1)}{\hat{\boldsymbol{\beta}}} = \underset{(m,\ k)}{\hat{\boldsymbol{W}}^{\mathrm{T}}}\ \underset{(k,\ m)}{(\hat{\boldsymbol{P}}\ \hat{\boldsymbol{W}}^{\mathrm{T}})^{-1}}\ \underset{(m,\ k)}{\hat{\boldsymbol{q}}} \qquad (k,\ 1)\ ,$$

kde $\hat{\boldsymbol{W}}$ obsahuje k-tý řádek vah $\boldsymbol{w}_{\mathrm{iter}}$ a $\boldsymbol{P}$ obsahuje k řádků zátěží $\boldsymbol{p}_{\mathrm{iter}}$.

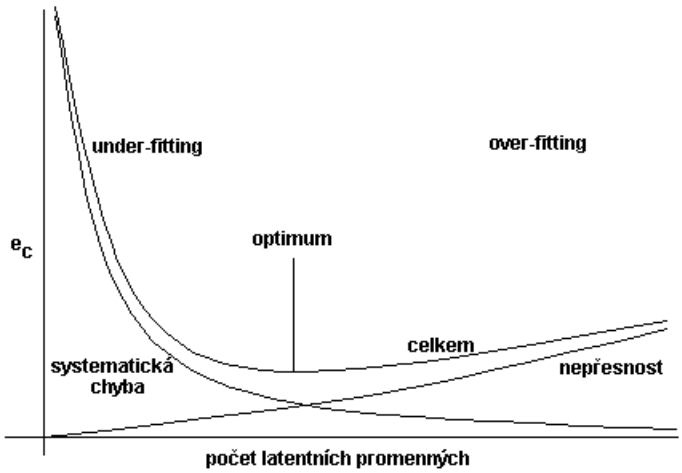
(b) Bude se centrovat spektrum neznámého vzorku $\boldsymbol{x}^*$ použitím sloupcového průměru kalibračních dat.

(c) Stanoví se koncentrace c^* tohoto vzorku dle vzorce

$$\underset{(1,\ 1)}{c^*} = \underset{(1,\ m)}{\boldsymbol{x}^{*\mathrm{T}}}\ \underset{(m,\ 1)}{\boldsymbol{b}} + \underset{(1,\ 1)}{\bar{c}}\ ,$$

kde $\bar{c}$ je průměrná koncentrace kalibračních vzorků.

C. Validace optimálního modelu a určení jeho predikční schopnosti. Ve vícerozměrné inverzní kalibraci se pod pojmem validace chápe jednak stanovení optimálního počtu proměnných k, jež mají být zahrnuty v modelu, a jednak odhad predikční schopnosti modelu. Cílem optimalizace počtu proměnných je nalézt vhodný kompromis mezi příliš těsným (over-fitting) a nedostatečným (under-fitting) proložením kalibračních dat. Nedostatečným proložením je míněna situace, kdy významná část koncentrace $\boldsymbol{c}$ není popsána modelem, protože počet latentních proměnných uvažovaný v modelu je příliš malý. Na druhé straně, příliš těsné proložení znamená, že je modelován i nesystematický rozptyl v datech (tj. šum), a to z důvodu zahrnutí příliš mnoha komponent do modelu. Následkem toho je zhoršená věrohodnost predikované koncentrace. Obě uvedené situace i hledaný kompromis, označený *optimum*, jsou na obr. 4.24.



Obr. 4.24 Predikční chyba *versus* počet latentních proměnných v modelu.

V praxi lze použít dva postupy stanovení optimálního počtu latentních proměnných:

(i) *metodu příčné validace* (cross-validation, CV), nebo

(ii) *metodu predikčního testování externí validaci,* EV. Externí validace je postavena na použití tzv. tréninkového a monitorovacího souboru, které vznikají rozdělením původních kalibračních dat.

Nevýhodou druhého postupu je právě zmíněné dělení kalibračního souboru. To musí být provedeno tak, aby vznikly dva reprezentativní podsoubory. Dělení navíc vyžaduje, aby celkový počet kalibračních vzorků byl velký.

Cílem stanovení predikční schopnosti modelu je kvantifikace chyby, kterou je odhad koncentrace průměrně zatížen. Ke stanovení predikční schopnosti lze použít jak metodu příčné validace, tak predikčního testování externí validací. Jelikož náklady spjaté s měřením monitorovacího souboru pro predikční testování bývají vysoké, častěji se užívá metoda příčné validace. Typickou charakteristikou sloužící ke kvantifikaci chyby predikce je *střední kvadratická chyba predikce*, kterou lze vyčíslit dvojí technikou:

(a) Příčnou validací (CV). Metoda je založena na simulaci predikční fáze při skutečné kalibraci. Kalibrační data jsou nejprve rozdělena na určitý počet skupin (*gr*). V extrémním případě je *gr* = *n* a metoda se nazývá *"odlož-jeden-mimo" příčná validace* (leave-one-out validation). V prvním kroku je sestrojen kalibrační model na základě všech dat kromě první skupiny. Tento model je použit na stanovení koncentrace vzorků z první skupiny. Konstrukce modelu a predikce se poté opakuje *gr*-krát, takže koncentrace každého vzorku je stanovena právě jednou.

Srovnání stanovených (c*) a referenčních (c_{ref}*) koncentrací kalibračních vzorků poskytuje odhad očekávané chyby koncentrace pro neznámé vzorky v budoucnu, střední kvadratickou chybu predikce RMSECV:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n} (c_i - c_{i,ref})^2}{n}} ,$$

kde i je index kalibračního vzorku od 1 do n nebo do gr.

(b) Predikčním testováním externí validací (EV). Predikční testování lze použít jak k optimalizaci počtu proměnných v modelu, tak k odhadu průměrné chyby predikce. Kalibrační soubor je rozdělen do dvou skupin, na tzv. tréninkový a monitorovací podsoubor. Kalibrační model je sestrojen na základě tréninkových dat. Je stanovena koncentrace pro vzorky v monitorovacím souboru a vypočtena střední kvadratická chyba predikce RMSEP:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (c_i - c_{i,ref})^2}{n_t}} ,$$

kde i je index monitorovacího vzorku od 1 do n_t.

D. Aplikace vícerozměrné kalibrace v analytické chemii. Klasická vícerozměrná kalibrace se používá především v UV/VIS spektroskopii, a to v případech, kdy je cílem kvantifikovat množství látek ve směsi, obsahující známé složky. Na druhé straně, inverzní vícerozměrná kalibrace se využívána hlavně v oblasti NIR, UV/VIS, ICP a Ramanovy spektroskopie, a to při analýzách směsí obsahujících neznámé látky v neznámém množství.

Technika, která je na možnostech vícerozměrné kalibrace vysloveně závislá, je blízká infračervená spektroskopie NIR, pokrývající spektrální oblast 700 - 2500 nm. Na rozdíl od IR jsou totiž NIR spektra naprosto neselektivní a neumožňují aplikaci jednorozměrné kalibrace.

S rozvojem počítačů a vícerozměrné kalibrace vystoupila metoda NIR do popředí zájmu analytických chemiků a dnes se už běžně používá k analýzám v petrochemii, potravinářství i farmacii a chemii. V petrochemii se konkrétně využívá jako náhrada kalibrovaného motoru při stanovení oktanových čísel benzinů, dále při stanovení složení benzinů, destilačních charakteristik, hydroxylového čísla polyeter-polyolů apod. V potravinářství se aplikuje zejména při stanovení vody, proteinů a uhlohydrátů v potravinách a nápojích. Existuje však množství dalších specifických aplikací. Ve farmacii a chemii lze NIR využít k identifikaci surovin, produktů a obalových materiálů, dále ke stanovení obsahu hlavní látky, vody a polymorfismu, ale i ke stanovení množství mikroorganismů.

Vzorová úloha 4.16 *Postup vícerozměrné kalibrace*

Postup vícerozměrné kalibrace ukážeme na úloze **C4.10** *Vícerozměrný kalibrační model kvality bezolovnatého benzinu*. Dle následujících kroků na základě naměřených NIR spekter sestrojte vícerozměrný kalibrační model pro jednu z charakteristik kvality, tj. koncentraci jedné ze složek bezolovnatého benzinu. Model pak užijte ke kontrole kvality benzinů z kontinuální produkce:

(a) Sestrojte nejlepší jednorozměrný kalibrační model a použijte ho ke srovnání s výsledky vícerozměrné kalibrace.

(b) Prozkoumejte vícerozměrná spektrálních data metodou hlavních komponent (PCA). Interpretujte rozptylové grafy komponentního skóre, grafy zátěží a matici vlastních čísel. Nalezněte i odlehlá spektra.

(c) Sestrojte vícerozměrný kalibrační model metodou PCR a PLS. Interpretujte graf chyby predikce *versus* počet latentních proměnných. Zvolte optimální počet proměnných v modelu. Interpretujte souvislost mezi latentními proměnnými a odlehlými body.

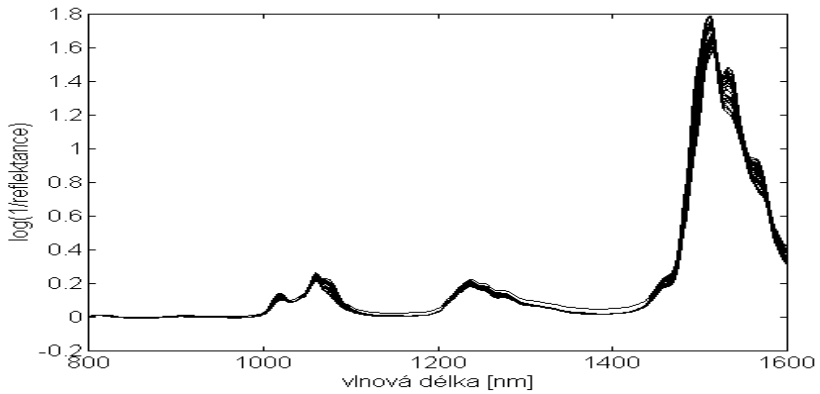
(d) Z absorbanční matice $\boldsymbol{X}$ eliminujte odlehlá spektra, jež prokazatelně zhoršují kvalitu kalibračního modelu. Sestrojte finální PCR a PLS model.

(e) Ke kalibraci aplikujte krokovou vícerozměrnou lineární regresi, sloužící zde jako jako alternativa k metodám s latentními proměnnými. Diskutujte statistickou významnost jednotlivých proměnných v modelu s ohledem na výsledky t-testu. Metodou příčné validace nalezněte na závěr optimální model. Porovnejte takto dosažené výsledky s výsledky z metody regrese na hlavních komponentách PCR a častečnými nejmenšími čtverci PLS.

(f) Zhodnoťte všechny výsledky a doporučte metodiku, která bude aplikována v praxi.

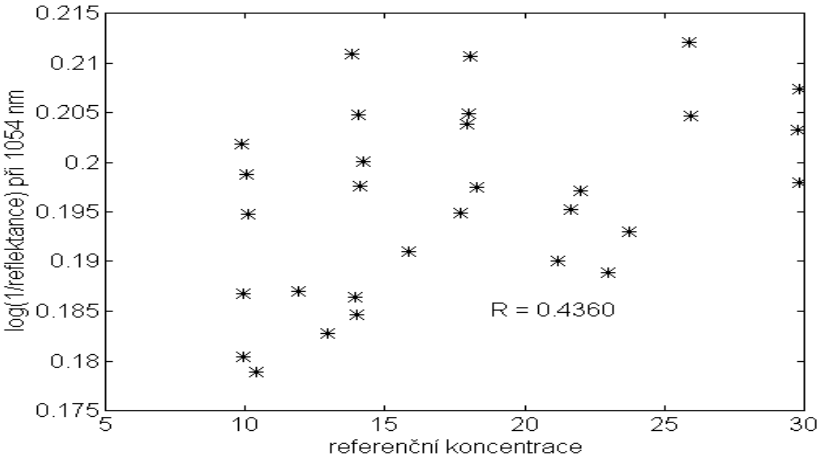
Data: Naměřená NIR spektra analyzovaných bezolovnatých benzinů úlohy C4.10 jsou na obr. 4.25.

	l_1	l_2	l_3	l_4	l_5
vzorek1	0.0033390	0.0047277	0.0062653	0.0077811	0.0090141
...	...	...	...	...	...
...	...	...	...	...	...
vzorek 30	0.0056775	0.0058778	0.0066916	0.0076192	0.0087291



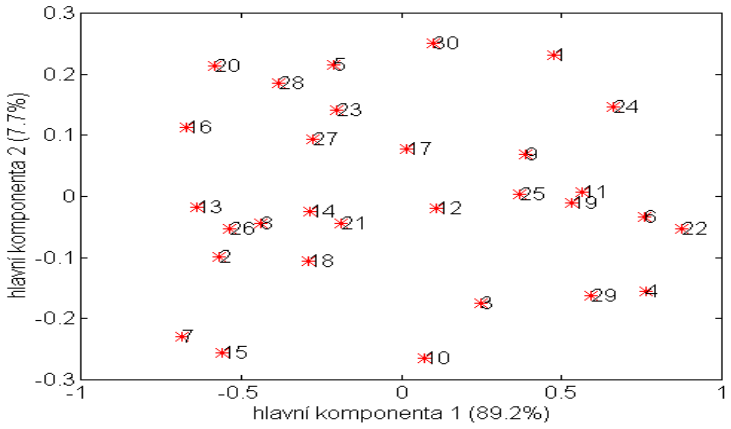
Obr. 4.25 NIR spektra vzorků benzinů naměřená v oblasti 800-1600 nm.

Řešení: Vzhledem k velkému rozsahu dat, a to 700 spekter při 30 vlnových délkách, je v následující tabulce uvedeno pouze prvních 5 hodnot signálu pro první a poslední spektrum datového souboru C4.10.

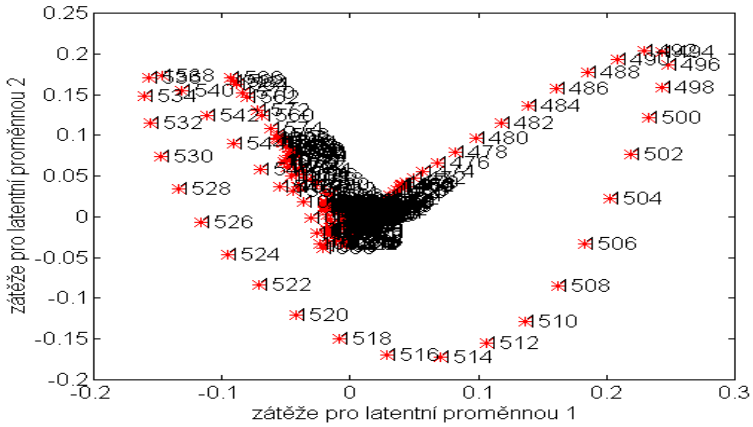


Obr. 4.26 Odezva vzorku při vlnové délce 1054 nm *versus* jeho referenční koncentrace.

Z obrázku obr. 4.25 je patrné, že nejméně jedno spektrum je odlišné od ostatních, a to zejména ve spektrální oblasti 1100-1220 nm a 1250-1470 nm. Toto spektrum by mělo být identifikovatelné v grafu hlavních komponent.



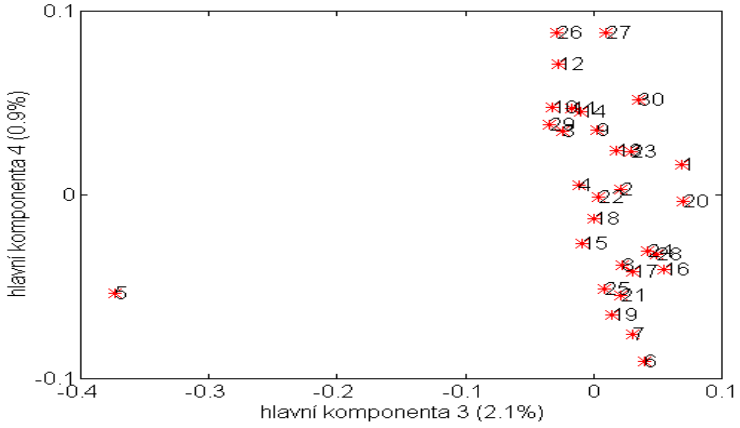
Obr. 4.27 Graf hlavních komponent 1 a 2. Čísla objektů v grafu odpovídají číslům vzorků.



Obr. 4.28 Graf zátěží. Čísla objektů v grafu odpovídají číslům vzorků.

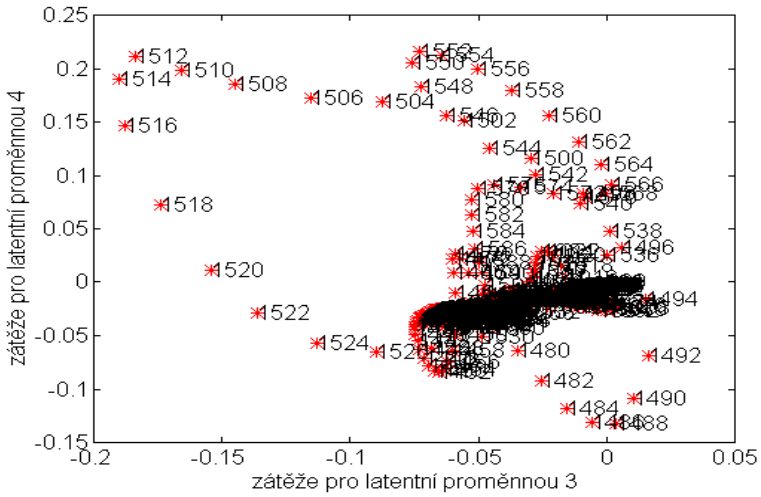
(a) *Jednorozměrná kalibrace:* použitím korelačního koeficientu byla vybrána vlnová délka 1054 nm, která poskytuje nejlepší jednorozměrný model. Jak je patrné z obr. 4.26 tento model není však v laboratoři použitelný. Korelační koeficient mezi měřením a kalibrovanou koncentrací je 0.4360 a vypočtená střední kvadratická chyba predikce RMSEP je 6.00. Pokud by neexistovala možnost použít vícerozměrnou kalibraci, analytický problém by nebyl řešitelný.

(b) *Analýza hlavních komponent (PCA)*: z matice vlastních čísel plyne, že první hlavní komponenta vysvětluje 89.2% celkového rozptylu zdrojové spektrální matice ***X***, druhá hlavní komponenta 7.7 %, třetí hlavní komponenta 2.1% a čtvrtá hlavní komponenta 0.9%. Graf komponentního skóre na první hlavní komponentě *versus* komponentního skóre na druhé hlavní komponentě je prezentován na obr. 4.27. Graf odpovídajících zátěží je na obr. 4.28. Hodnoty v grafu odpovídají vlnovým délkám.



Obr. 4.29 Graf hlavních komponent 3 a 4. Bod č. 5 je odlehlý na komponentě 3.

Největší váhu v při konstrukci první hlavní komponenty mají originální proměnné kolem vlnové délky 1496 resp. 1536 nm a při konstrukci druhé latentní proměnné originální proměnné okolo vlnových délek 1492 a 1514 nm, obr. 4.28.

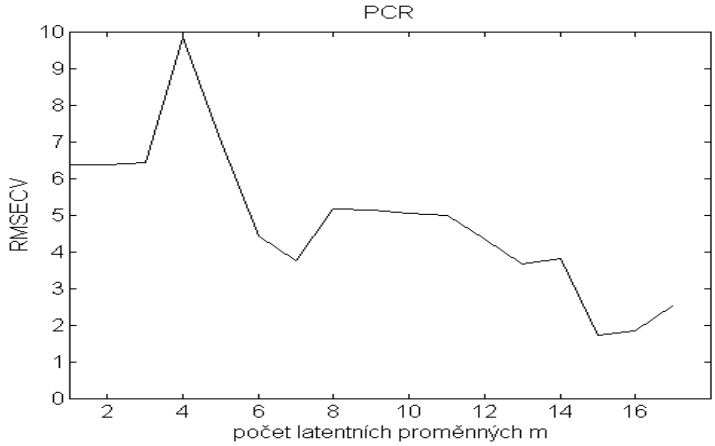


Obr. 4.30 Zátěže pro hlavní komponentu 3 a 4.

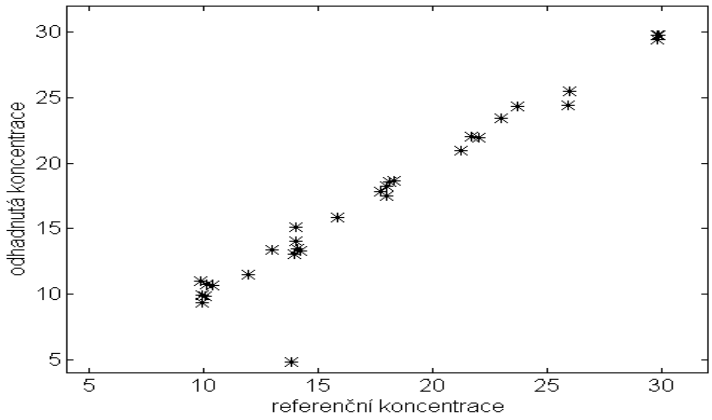
Toto zjištění je v souladu s originálními spektry z obr. 4.25, protože uvedené vlnové délky odpovídají maximům absorpčních pásů a skutečně reprezentují maximum rozptylu v datech. Graf prvních dvou hlavních komponent kromě toho ukazuje rovnoměrné rozložení bodů v prostoru. Žádná abnormalita v grafu nebyla zjištěna. Naproti tomu, rozptylový graf komponentních skóre na třetí hlavní komponentě *versus* čtvrtá hlavní komponenta, obr. 4.29, ukazuje jednoznačnou odlehlost bodu 5 na třetí latentní proměnné.

Odlehlost spektra č. 5 je způsobena hlavně rozdílem v odezvách kolem vlnové délky 1514 nm, obr. 4.25.

(c) *PCR a PLS kalibrace*: obr. 4.31 ukazuje závislost velikosti střední kvadratické chyby predikce na počtu latentních proměnných v modelu. Na křivce jsou patrná dvě minima. První, při 7 latentních proměnných, odpovídá chybě predikce RMSCV > 3.5, což je ještě velká chyba. Druhé, při 15 latentních proměnných, reprezentuje nerobustní řešení, protože tolik proměnných v modelu znamená modelování velké části šumu. Odlehlý bod č. 5 ovlivňuje predikci tak zásadním a negativním způsobem, že musí být vyloučen a model sestrojen znovu.

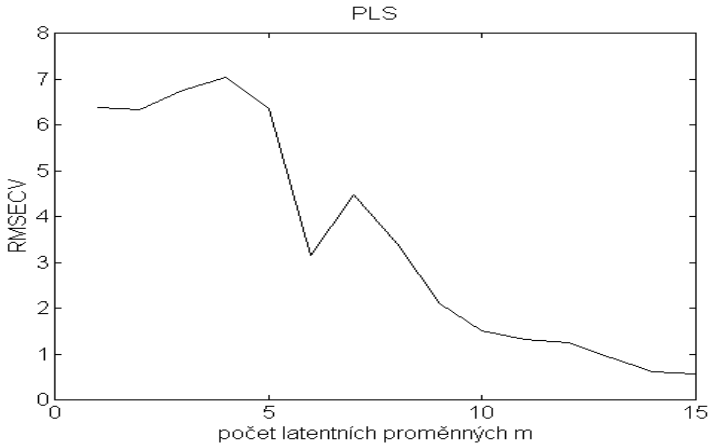


Obr. 4.31 Střední kvadratická chyba predikce RMSECV *versus* počet latentních proměnných.

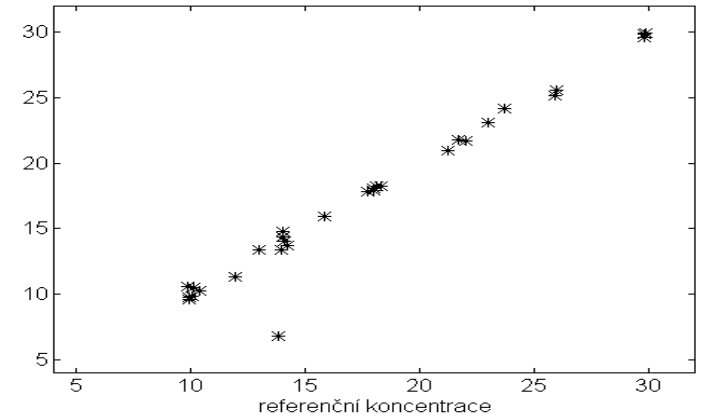


Obr. 4.32 Odhadnutá *versus* referenční koncentrace vzorků. Predikované hodnoty byly stanoveny metodou příčné validace.

Obrázek odhadnutých *versus* referenčních koncentrací, obr. 4.32, potvrzuje uvedený fakt: odhadnutá koncentrace pro bod 5 je výrazně nižší, než je její odpovídající referenční hodnota. Obr. 4.33 ukazuje, že predikce dosažená metodou částečných nejmenších čtverců PLS na datech zahrnujících bod 5 je podobná regresi hlavních komponent PCR.



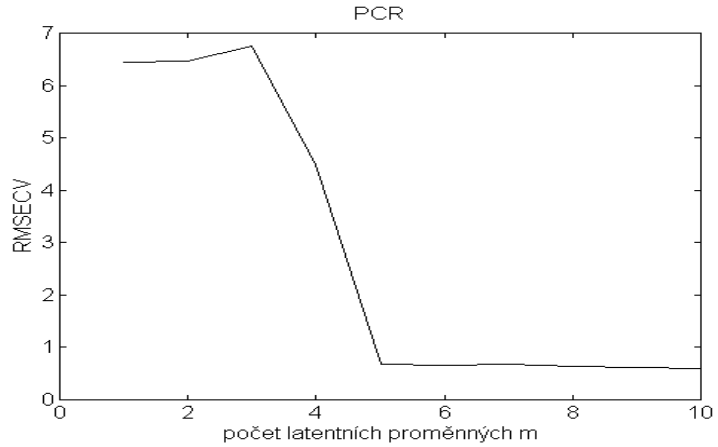
Obr. 4.33 Střední kvadratická chyba predikce RMSECV *versus* počet latentních proměnných.



Obr. 4.34 Graf predikovaných versus referenčních koncentrací kalibračních vzorků. Predikované hodnoty byly stanoveny metodou příčné validace.

Obr. 4.34 potvrzuje, že bod č. 5 by měl být z kalibračních dat eliminován.

(d) Jelikož absorbanční spektrum č. 5 prokazatelně zhoršuje kvalitu kalibračního modelu, bylo odstraněno z matice ***X***. Obr. 4.35 ukazuje rozptylový graf komponentních skóre na třetí hlavní komponentě *versus* skóre na čtvrté hlavní komponentě po této eliminaci. Výsledkem je téměř rovnoměrné rozložení bodů v prostoru. Jelikož odlehlý bod představoval poměrný silný zdroj rozptylu, po jeho odstranění relativní významnost třetí hlavní komponenty poklesla z 2.1 na 1.0%, viz obr. 4.27 a obr. 4.29.

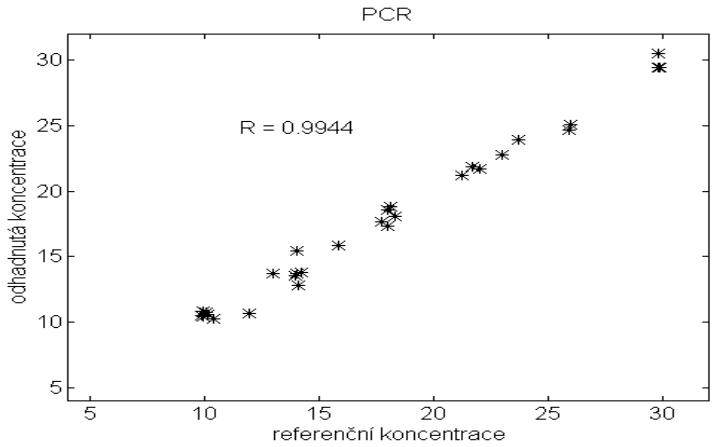


Obr. 4.35 Optimalizace počtu proměnných ve finálním modelu. RMSECV *versus* počet latentních proměnných použitý ke konstrukci modelu.

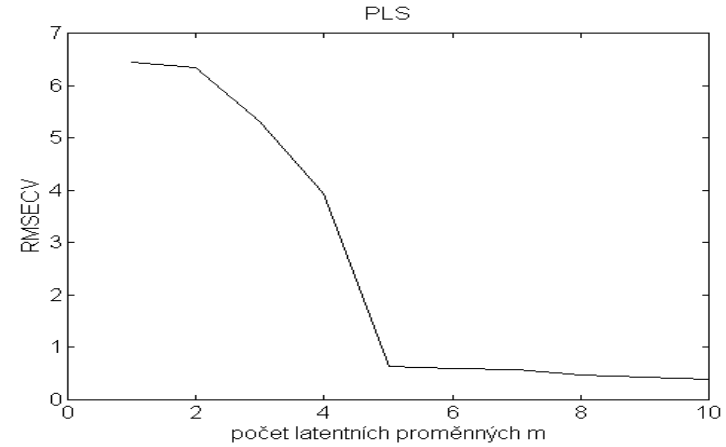
Po eliminaci odlehlého bodu je třeba zopakovat postup optimalizace počtu latentních proměnných v modelu metodou příčné validace. Graf chyby predikce *versus* počet hlavních komponent v PCR modelu je zobrazen v obr. 4.35. Graf ukazuje přijatelnější průběh než v případě dat s odlehlým měřením, obr. 4.31. Po dosažení počátečního maxima střední kvadratická chyba predikce prudce klesá. Minimum je dosaženo při 5 hlavních komponentách, což je mnohem nižší počet než 15 na obr. 4.31.

Graf predikovaných *versus* referenčních koncentrací dosažených s optimálním PCR modelem, obsahujícím 5 latentních proměnných je prezentován v obr. 4.36. Zmizel odlehlý bod č. 5 a střední kvadratická chyba predikce

poklesla na hodnotu 0.67. Korelační koeficient mezi predikovanými a referenčními koncentracemi dosáhl hodnoty 0.9944.



Obr. 4.36 Koncentrace odhadnuté pomocí PCR modelu s 5 hlavními komponentami.



Obr. 4.37 Optimalizace počtu latentních proměnných v PLS modelu.

PLS kalibrace poskytuje podobné výsledky jako PCR, obr. 4.37. Optimální model obsahuje 5 latentních proměnných. Odpovídající chyba predikce je 0.64. Graf částečných nejmenších čtverců odhadnutých PLS *versus* referenčních koncentrací je podobný grafu na obr. 4.36.

Finální modely, určené regresí na hlavních komponentách PCR, resp. metodou částečných nejmenších čtverců PLS jsou sestrojeny použitím 5 latentních proměnných. Oba modely poskytují odhad koncentrace neznámých vzorků, zatížený absolutní chybou menší než 0.7.

(e) *Kalibrace krokovou vícenásobnou lineární regresí:* postupná selekce proměnných byla provedena dle následujících kroků:

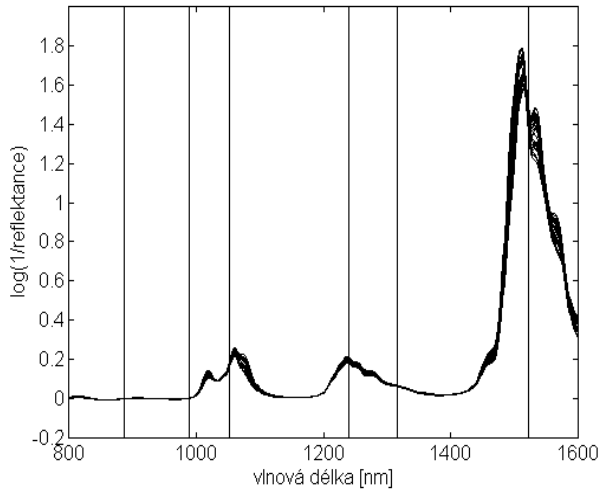
1. krok: byla vybrána proměnná č. 127, odpovídající vlnové délce 1052 nm. S použitím Studentova t-testu bylo potvrzeno, že tato proměnná vysvětluje významnou část rozptylu kalibrované koncentrace ($t_{exp} = 2.98$ je větší než kritický kvantil $t_{crit} = 2.05$). Dopředná selekce proměnných v modelu proto pokračuje.

2. krok: nejvyšší korelační koeficient s vektorem koncentračních reziduí poskytuje měření č. 259, tj. vlnová délka 1316 nm. Studentův t-test potvrdil, že obě proměnné (č. 127 a 259) jsou významné, protože pro obě je experimentální hodnota $t_{exp} > 15$. Dopředná selekce proměnných v modelu proto pokračuje.

3. krok: další kandidátskou proměnnou je měření 96, tj. vlnová délka 990 nm. I zde byla při aplikaci t-testu prokázána významnost všech 3 proměnných v modelu.
4. krok: byla vybrána proměnná č. 221, odpovídající vlnové délce 1240 nm. Potvrzena významnost všech členů v modelu. Selekce pokračuje.
5. krok: byla vybrána proměnná č. 45, tj. vlnová délka 888 nm. Potvrzena významnost všech proměnných v modelu.
6. krok: byla nalezena proměnná č. 362, tj. 1522 nm. Potvrzena významnost členů. Dopředná selekce pokračuje.
7 krok: byla vybrána proměnná č. 8 při 814 nm. Významnost této proměnné nebyla potvrzena. Ukončení algoritmu.

Finální model krokovou vícenásobnou lineární regresí obsahuje proměnné č. 45, 96, 127, 221, 259 a 362, což odpovídá vlnovým délkám 888, 990, 1052, 1240, 1316 a 1522. Poloha proměnných v NIR spektru je zobrazena v obr. 4.38.

Z obr. 4.38 je patrné, že selekce proměnných s pomocí matematických metod je nezastupitelná. Prosté vizuální porovnání spekter analytikovi nedovoluje vybrat kombinaci proměnných, která by vedla ke smysluplným kalibračním výsledkům.



Obr. 4.38 NIR spektra benzinů použitá ke kalibraci s vyznačením 6 vlnových délek vybraných Stepwisovou metodou vícenásobné lineární regrese.

Příčná validace potvrdila, že všech 6 proměnných, vybraných krokovou metodou lineární regrese je významných. Minimální střední kvadratická chyba predikce je RMSECV = 0.446. Kdyby v modelu nebyla obsažena poslední vybraná proměnná, tj. č. 362, střední kvadratická chyba predikce by dosáhla hodnoty 0.67.

(f) *Závěr*: Použitím metody částečných nejmenších čtverců PCA bylo zjištěno, že naměřená blízká infračervená spektra NIR obsahují jedno odlehlé měření. Kalibrační výsledky toto zjištění potvrdily. PCR i PLS model se po odstranění odlehlého spektra výrazně zlepšil a zjednodušil. Počet hlavních komponent klesl z 15 na 5 komponent.

Hlavní komponenty použité ke kalibraci byly analyzovány. Grafy zátěží pro sledované komponenty vykazují maxima při vlnových délkách odpovídajících hlavním absorpčním pásům v původním spektru. To je očekávané zjištění a dokladuje logickou konstrukci hlavních komponent.

K optimalizaci počtu proměnných v modelu metodou regrese hlavních komponent PCR a částečných nejmenších čtverců PLS byla použita metoda příčné validace. Absolutní chyba predikce, obdržená s modelem obsahujícím 5 latentních proměnných, je menší než 0.7. Korelace predikovaných koncentrací s referenčními je velmi dobrá. Korelační koeficient je vyšší než 0.994.

Aplikace krokové metody vícenásobné lineární regrese prokázala, že i kalibrační model založený na podsouboru původních proměnných může dávat dobré výsledky. Ve studovaném případě dokonce lepší než PCR nebo PLS. Z důvodu menší robustnosti takového modelu k posunu vlnových délek, ke změnám v citlivosti přístroje a k linearitě odezvy by se v praxi tento model uplatnil pouze v krátkém časovém horizontu. Pokud by cílem kalibrace bylo používat model v průběhu např. 1 roku, pak by byl upřednostněn robustnější model PCR nebo PLS.

4.9 Úlohy

Při vyšetřování jednotlivých úloh je třeba postupovat dle následujícícho postupu:

Postup analýzy vícerozměrných dat

1. *Standardizace:* vícerozměrné analýze obvykle předchází standardizace čili škálování proměnných.

2. *Odhady parametrů polohy, rozptýlení, tvaru a intenzita vztahu mezi proměnnými:* vyčíslení výběrové střední hodnoty každé proměnné, odhad kovarianční matice $\boldsymbol{S}$ a její normované podoby - korelační matice $\boldsymbol{R}$, odhadu vícerozměrné šikmosti $g_{1,m}$ a vícerozměrné špičatosti $g_{2,m}$. Matice $\boldsymbol{R}$ obsahuje Pearsonovy párové korelační koeficienty ρ_{ij}, které se diskutují. Užitečný je především diagram korelační matice.

3. *Exploratorní analýza dat EDA:* (a) posoudí podobnost objektů pomocí vizuálních rozptylových diagramů typu casement plot, draftsman plot, dále symbolových a profilových grafů (hvězdičky, sluníčka, obličeje, křivky, stromy), (b) nalezne vybočující objekty nebo vybočující proměnné, mnohdy k nevhodné analýze, (c) stanoví, zda platí předpoklad lineárních vazeb, (d) testuje všechny předpoklady o datech (normalitu, nekorelovanost, homogenitu). Ověřování normality je založeno na vícerozměrné šikmosti $g_{1,m}$ a vícerozměrné špičatosti $g_{2,m}$, kdy se testuje simultánní platnost nulových hypotéz H_{01}: $g_{1,m} = 0$ a H_{02}: $g_{2,m} = m(m + 2)$.

4. *Určení vhodného počtu latentních proměnných:* matice $\boldsymbol{S}$ nebo $\boldsymbol{R}$ se rozloží na vlastní čísla λ_i a vlastní vektory $\boldsymbol{v}_i$. Z indexového grafu úpatí vlastních čísel (Scree plot) se určí vhodný počet latentních proměnných (pro zobrazení v rovině se obvykle dává přednost prvním dvěma latentním proměnným), které ještě dostatečně popisují proměnlivost v

datech. Když se latentní proměnné podaří pojmenovat a dát jim i fyzikální, biologický či jiný věcný význam, jedná se o faktory. V opačném případě jde o hlavní komponenty.

5. *Určení struktury v proměnných (PCA a FA):* hledání struktury a vzájemných vazeb (korelace) proměnných se provede v grafu komponentních vah (Plot of components weights, loadings). Hledání struktury v objektech a třídění objektů do shluků se provede v rozptylovém diagramu komponentního skóre (Plot of principal components). Dvojný graf (Biplot) je přehledným spojením obou předešlých grafů a ukáže interakci objektů a proměnných.

6. *Určení struktury a vzájemných vazeb v objektech:* klasifikační postupy zařadí v diskriminační analýze analyzovaný objekt do jednoho již existujícího a předem zadaného shluku. Neutříděnou skupinu objektů lze uspořádat do shluků a výsledek třídění zobrazit dendrogramem v analýze shluků. V hierarchickém postupu je třeba k vytvoření shluků vybrat vzdálenost mezi objekty (Eukleidovskou, Manhattanovskou, Mahalanobisovu) a jednu z nabídnutých metod: průměrovou, centroidní, nejbližšího souseda, nejvzdálenějšího souseda, mediánovou, Wardovu. Nehierarchické postupy rozdělí objekty do shluků, v nichž jsou předem umístěni typičtí reprezentanti.

7. *Soulad nalezené struktury objektů a vzájemných vazeb v dendrogramu a PCA (či FA) grafech:* je třeba vyšetřit a komentovat nalezenou strukturu a vazby *jednotlivých proměnných,* nalezenou jednak v PCA (či FA) a jednak v dendrogramu podobnosti proměnných analýzou vzniklých shluků. Dále je třeba komentovat také strukturu a vazby *klasifikovaných objektů,* nalezenou v PCA a v dendrogramu podobnosti objektů.

Využitím modulu Vícerozměrná data programového systému ADSTAT, resp. programu STATGRAPHICS, SCAN, MINITAB, STATISTICA, S-Plus atd. je třeba analyzovat dále uvedené úlohy. Úlohy jsou rozděleny do pěti kapitol: B4 (farmakologická a biochemická data), C4 (chemická a fyzikální data), E4 (environ-mentální, potravinářská a zemědělská data), H4 (hutní a mineralogická data) a S4 (ekonomická a sociologická data).

4.9.1 Analýza farmakologických a biochemických dat

Úloha B4.01 *Chromatografické chování farmakologických sloučenin* (EDA, PCA, CLU)
Byly měřeny hodnoty R_F pro 20 sloučenin s 18 eluenty[22]. Žádné eluční činidlo však neprovedlo úplné rozdělení. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalita, nekorelovanost, homogenita). Vyšetřete, zda lze nalézt minimální výběr elučních činidel, které by daly dostatek informace pro kvalitativní analýzu. Postupujte dle kroků *Postupu analýzy vícerozměrných dat*.

Data: Datový soubor GIUSEPPE obsahuje 100 × R_F pro 20 sloučenin (řádky, jména byla zkrácena na max. 8 písmen) a 18 elučních činidel (sloupce), představujících zde proměnné: *B401i* název vzorku, *B401x1* Toluen : aceton : ethanol : 30 % amoniak = 45 : 45 : 7 : 3, *B401x2* Ethylacetát : benzen : methanol : 30 % amoniak = 60 : 35 : 6.5 : 2.5, *B401x3* Benzen : dioxan : ethanol : 30 % amoniak = 50 : 40 : 7.5 : 2.5, *B401x4* Methanol : 30 % amoniak = 100 : 1.5, *B401x5* Benzen : 2-propanol : methanol : 30 % amoniak = 70 : 30 : 20 : 5, *B401x6* Ethylacetát : methanol : 30 % amoniak = 85 : 10 : 5, *B401x7* Cyklohexan : toluen : diethylamin = 65 : 25 : 10, *B401x8* Cyklohexan : toluen : diethylamin = 75 : 15 : 10, *B401x9* Cyklohexan : benzen : metanol : diethylamin = 70 : 20 : 10 : 5, *B401x10* Chloroform : aceton : diethylamin = 50 : 40 : 10, *B401x11* Cyklohexan : chloroform : diethylamin = 50 : 40 : 10, *B401x12* Benzen : ethylacetát : diethylamin = 50 : 40 : 10, *B401x13* Xylen : methylethylketon : methanol : diethylamin = 40 : 40 : 6 : 2, *B401x14* Diethylether : diethylamin = 95 : 5, *B401x15* Ethylacetát : chloroform = 50 : 50, *B401x16* Ethylacetát : chloroform [A] = 50 : 50, *B401x17* Butanol : methanol = 40 : 60, *B401x18* Butanol : methanol *[A]* = 40 : 60, kde *[A]* značí, že byl užit 0.1M methanolát draselný.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
Atropine	20	16	29	23	62	33	04	02	13	47	25	42	18	12	00	00	05	08
...	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Procaine	64	60	70	65	82	73	8	5	16	66	24	54	37	50	1	11	29	53

Úloha B4.02 *Účinky neuroleptik při tlumení rozličných psychóz* (EDA, PCA, CLU)
Psychóza označuje mentální poruchy (nepřesné myšlení, nesprávné vnímání skutečnosti, nesprávné konání), jako je schizofrenie, mánie a paranoia[23]. Neuroleptika jsou léky, které sice neléčí příčiny psychózy, ale normalizují chování. Ze života pacientů odstraňují svěrací kazajku a pouta tím, že výrazně tlumí mánii, schizofrenii, třes a záchvaty. Psychóza je vyvolaná zvýšenou komunikací mezi buňkami mozku nadměrnou aktivitou dopaminu, serotoninu a adrenalinu. Při psychóze mozkové buňky uvolňují nadbytek dopaminu, což vyvolává zvýšený přenos signálů mezi mozkovými buňkami. Neuroleptika redukují nežádoucí účinky přebytečného dopaminu tím, že obsazují dopaminové receptory, a tak zabraňují působení nadbytku dopaminu, který zůstává volný. Tlumicí účinek neuroleptika se vyjadřuje ***mediánovou účinnou dávkou ED50*** **[mg/kg]**, která představuje mg léčiva na 1 kg hmotnosti organismu, utlumené poloviny celkového počtu pokusných zvířat. Nízká hodnota ED50 značí výkonnější léčivo, které působí již v malém množství. Častěji se však užívá ***převrácená hodnota 1/ED50*** **[kg/mg]**, která vyjadřuje přímou úměru mezi množstvím neuroleptika a tlumicí aktivitou. Neuroleptika se však liší ve svých účincích a vedlejších účincích: potlačují nervozitu, záchvaty, třes, ospalost, letargii, přibývání na hmotnosti, nejasné vidění, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění atd. Účelem je provést klasifikaci neuroleptik do shluků podobných účinků s ohledem na potlačení nervozity, potlačení stereotypního chování, potlačení záchvatu a třesu,

a konečně i velikost dávky smrticího účinku neuroleptika. Užijte také metodu hlavních komponent k jednoduššímu popisu účinku neuroleptik. Komentujte biologický smysl dosažených závěrů statistické klasifikace. K analýze užijte centrovaná data. Proveďte všechny kroky *Postupu analýzy vícerozměrných dat* a především komentujte proměnlivost v datech (ve sloupcích) v krabicovém grafu u všech proměnných. Vysvětlete korelační diagram 12 objektů a 13 proměnných. Dají se určit podobní pacienti z hvězdičkového grafu? V grafu komponentních vah nalezněte proměnné, které spolu silně korelují a současně si všimněte proměnných s vysokým stupněm důležitosti. Kolik shluků se dá obkroužit v rozptylovém diagramu komponentního skóre? Na 66 % hladině podobnosti určete v dendrogramu shluky objektů.

Data: Charakter proměnných (převrácená hodnota mediánové účinné dávky 1/ED50 [kg/mg]): *B402i* název neuroleptika, *B402x1* potlačení nervozity, *B402x2* potlačení stereotypního chování, *B402x3* potlačení záchvatu a třesu, a *B402x4* dávka smrtícího účinku.

i	*B402x1*	*B402x2*	*B402x3*	*B402x4*
1 Chlorphromazine	3.846	3.333	1.111	1.923
...	...	...	...	...
20 Molindone	7.692	7.692	0.140	0.006

Úloha B4.03 *Analýza psychosociálních vlivů na výskyt rakoviny prsu* (EDA, PCA, CLU)
Byly studovány psychosociální vlivy na výskyt rakoviny prsu u žen[24]. Vlivy označené jako proměnné naměřené a kategorizované jsou uvedeny v datech. Vyšetřované proměnné jsou jak kardinální, tak ordinální a nominální. Zobrazte standardizovaná vícerozměrná data a pokuste se proměnlivost v datech, tj. rozdílnost žen, vyjádřit menším počtem proměnných. Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: Charakter proměnných: *B403i* index pacientky, *B403x*1 binární proměnná stavu ženy před přechodem 0 a po přechodu 1, *B403x*2 nespojitá kvantitativní proměnná věku ženy, *B403x*3 a *B403x*4 nominální proměnná ke klasifikaci povahy, nálady a pocitů stupněm 0, 1, a 2, *B403x*5 až *B403x*9 pořadová kvalitativní proměnná, vyjadřující na stupnici 0 až 10 míru: *B403x*5 nepřátelství, *B403x*6 kritiku ostatních a okolí, *B403x*7 paranoidní nepřátelství, *B403x*8 sebekritiku a *B403x*9 vlastní vinu, *B403x*10 věk ženy, ve kterém se u ní objevila první menstruace, *B403x*11 a a *B403x*12 jsou binární proměnné, značící přítomnost či nepřítomnost *B403x*11 alergie a *B403x*12 štítné žlázy, *B403x*13 kvantitativní proměnná tělesné hmotnosti v kg.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1	1	49	0	2	3	7	0	6	2	15	1	1	52.46
..	..	..	..	..	..	..	..	..	..	..	..	..	..
12	1	50	1	1	3	8	4	2	0	16	1	1	50.22

Úloha B4.04 *Popis a třídění polétavých mšic* (EDA, CORA, PCA, FA, CLU)
Jeffers (1967)[25] studoval 40 jedinců polétavých mšic *(Alate adelges)* za pomoci světelné pasti. U mšic bylo změřeno 19 ukazatelů, sloužících k rozlišení druhů a typů tohoto hmyzu: 14 proměnných se týká délky nebo šířky, 4 proměnné se týkají počtu a 1 proměnná je binární, vyjadřující přítomnost či absenci. Mšice se totiž obtížně rozlišují dle běžných taxonometrických klíčů a obvykle je nutné detailní vyšetření dat, hledání společných rysů a následná klasifikace do větších celků. Užitím metody hlavních komponent PCA se pokuste snížit původní počet 19 proměnných na menší počet, který by vystihl co největší množství proměnlivosti mšic. Před užitím PCA je potřeba provést standardizaci dat, protože proměnné představují směs délek a počtů. Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: Charakter proměnných: *B404i* index jedince, *B404x*1 délka těla, *B404x*2 šířka těla, *B404x*3 délka předního křídla, *B404x*4 délka zadního křídla, *B404x*5 počet průduchů, *B404x*6 délka tykadla I, *B404x*7 délka tykadla II, *B404x*8 délka tykadla III, *B404x*9 délka tykadla IV, *B404x*10 délka tykadla V, *B404x*11 počet tykadlových ostnů, *B404x*12 délka posledního článku nohy, *B404x*13 délka holeně, tibia, *B404x*14 délka stehna, *B404x*15 délka sosáku, *B404x*16 délka kladélka, *B404x*17 počet kladélkových trnů, *B404x*18 řitní otvor, *B404x*19 počet háčků zadních křídel.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
1	21.2	11.0	7.5	4.8	5.0	2.0	2.0	2.8	2.8	3.3	3	4.4	4.5	3.6	7.0	4.0	8	0	3.0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
40	12.8	5.7	4.8	2.8	5.0	1.6	1.4	1.7	1.9	2.3	5	2.3	2.5	1.9	5.0	3.1	8	1	2.0

Úloha B4.05 *Odezva v obou očích na dva podněty u pacientů se sklerózou multiplex*
U 69 pacientů se sklerozou multiplex byla sledována odezva v obou očích, v levém L a v pravém R postupně na dva vizuální podněty S1 a S2, str. 40 v ref.[20]. Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: *B405i* je index pacienta, *B405x1* je věk pacienta [roky], *B405x2* celková odezva obou očí na první podnět S1 označená (S1L+S1R), *B405x3* rozdíl v odezvě levého a pravého oka na první podnět S1 označený *S1L-S1R*, *B405x4* celková odezva obou očí na druhý podnět S2 označená (S2L+S2R), *B405x5* rozdíl v odezvě levého a pravého oka na druhý podnět S2 označený *S2L-S2R*.

B405i	*B405x1*	*B405x2*	*B405x3*	*B405x4*	*B405x5*
1	18	152	1.6	198.4	0
...	...	...	...	...	...
98	59	199.8	4.6	250.2	1

Úloha B4.06 *Přehled radioterapeutického léčení u vybraných pacientů*
U 98 pacientů byl sledováno radioterapeutické léčení a byla zaznamenáma následující data, str. 40 v ref.[20]. Proveďte kroky *Postupu analýzy vícerozměrných dat* a soustřeďte se především na proměnné x_2 a x_3. Jsou v datech odlehlé objekty? Jak je vysvětlíte? Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Které objekty lze separovat v korelačním nebo ve hvězdičkovém a sluníčkovém grafu? Kolik latentních proměnných dostatečně popisuje objekty? Jsou nějaké původní proměnné, které jsou v silné korelaci? Proveďte klasifikaci pacientů a rozhodněte, do kolika shluků lze třídit 98 pacientů?

Data: *B406i* je index pacienta, *B406x1* počet symptomů jako je pálení žáhy, zvracení atd., *B406x2* objem provedených činností ve stupnici 1 až 5, *B406x3* objem spánku ve stupnici 1 až 5, *B406x4* množství zkonzumované stravy, *B406x5* apetit ve stupnici 1 až 5, *B406x6* podrážděnost kůže ve stupnici 0 až 3.

Pacient, i	*B406x1*	*B406x2*	*B406x3*	*B406x4*	*B406x5*	*B406x6*
1	0.889	1.389	1.555	2.222	1.945	1
...	...	...	...	...		...
98	0.889	1	1	2	1	2

Úloha B4.07 *Úbytek kostní hmoty starších žen po cvičeních a dietách* (EDA, CORA, PCA)
Bylo zkoumáno, zda cvičení nebo doplňky vhodné diety zpomalí úbytek kostní hmoty u žen, str. 43 v ref.[20]. K měření obsahu minerálů v kostech byla použita absorpční fotometrie, a to pro tři kosti na dominantní a na vedlejší straně. Proveďte kroky *Postupu analýzy*

vícerozměrných dat a soustřeďte se na redukci proměnných. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Data: *B407i* je index pacienta, *B407x1* poloměr u dominantní kosti, *B407x2* poloměr u vedlejší kosti, *B407x3* dominantní část kosti pažní, *B407x4* vedlejší část kosti pažní, *B407x5* dominantní část kosti loketní, *B407x6* vedlejší část kosti loketní.

B407i	*B407x1*	*B407x2*	*B407x3*	*B407x4*	*B407x5*	*B407x6*
1	1.103	1.052	2.139	2.238	0.873	0.872
...	...	...	...	...		...
25	0.915	0.936	1.971	1.869	0.869	0.868

Úloha B4.08 *Významnost rozdílu a struktura funkce plic při velké zátěži u mužů a žen*

U vzorku zdravé populace studentů byla sledována plicní funkce za okolností, že vybraní studenti-dobrovolníci pracovali do úplného vyčerpání, str. 368 v ref.[20]. Vydechnutý vzduch byl analyzován. Výsledky 4 spotřeb kyslíku pro 25 mužů a 25 žen jsou uvedeny v datech. Proveďte kroky *Postupu analýzy vícerozměrných dat* a sledujte především výsledky u všech pacientů, potom u mužů *vs.* u žen. Sestrojte interval spolehlivosti pro ozdíl středních hodnot mužů a středních hodnot žen. Které proměnné spolu silně korelují? Která proměnná má dle své proměnlivosti největší důležitost? Analýzou hvězdiček a sluníček nalezněte shluky podobných studentů. Kolik shluků studentů detekuje diagram komponentního skóre? Lze nějak vysvětlit interakci blízkého umístění studenta a proměnné na ploše dvojného grafu? Kolik zřetelných shluků lze odhalit dendrogramem optimálního shlukovacího postupu?

Data: *B408i* je index pacienta, *B408x1* zbývající objem kyslíku [kg/min], *B408x2* zbývající objem kyslíku [ml/kg/min], *B408x3* maximální objem kyslíku [kg/min], *B408x4* maximální objem kyslíku [ml/kg/min], *B408x5* pohlaví.

i	*B408x1*	*B408x2*	*B408x3*	*B408x4*	*B408x5*
1	0.34	3.71	2.87	30.87	male
...	...	...	...	...	...
50	0.35	5.37	2.25	35.07	female

Úloha B4.09 *Sledování složení potu u zdravých žen*

Bylo analyzováno složení potu u 20 zdravých žen a byly sledovány tři proměnné, str. 229 v ref.[20]. Proveďte kroky *Postupu analýzy vícerozměrných dat* a odpovězte především na otázky: Korelují nějaké proměnné? Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Ukazují hvězdičky, že některé ženy jsou si podobné? Lze v diagramu komponentního skóre nalézt odlehlé objekty nebo shluky podobných objektů? Lze nalézt interakci objektu a proměnné ve dvojném grafu? Kolik shluků lze nalézt na 50 % hladině podobnosti v dendrogramu objektů?

Data: *B409i* je index ženy, *B409x1* nasládlost, *B409x2* obsah sodíku, *B409x3* obsah draslíku.

i	*B409x1*	*B409x2*	*B409x3*
1	3.7	48.5	9.3
...	...	...	...
20	5.5	40.9	9.4

Úloha B4.10 *Korelace výšek a stáří muže a ženy u 169 manželských dvojic* (CORA)
Pokuste se v rozptylovém diagramu nebo v grafu korelační matice vyšetřit korelaci dvojic proměnných: x_1 věk manžela, x_2 výška manžela, x_3 věk manželky, x_4 výška manželky a x_5 věk manžela při první svatbě. Vložte do grafu přímku $y = x$ a vyšetřete procento bodů, odchylujících se od této přímky. Jeví se závislost jako nelineární, tzn. vystižená spíše křivkou? Dopňte osu x a osu y diagramem rozptýlení a sledujte také rozptýlení těchto hodnot. Potvrďte, že u většiny párů je muž vyšší a starší než žena, str. 10, ref.[30]. Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: *B410i* je index dvojice, *B410x1* věk manžela [roky], *B410x2* výška manžela [mm], *B410x3* věk manželky [roky], *B410x4* výška manželky [mm] a *B410x5* věk manžela při první svatbě [roky].

B410i	*B410x1*	*B410x2*	*B410x3*	*B410x4*	*B410x5*
1	49	1809	43	1590	25
...	...	...	...	...	...
169	59	1720	56	1530	24

Úloha B4.11 *Korelace počtů narozených a zemřelých v populaci 69 států*
Ukažte, zda lze v korelační matici grafu dvou proměnných najít nějaké shluky obdobných států, ref.[30], a proveďte kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Data: *B411i* název státu, *B411x1* promile narozených, tj. počet na 1000 obyvatel, *B411x2* promile zemřelých.

B411i	*B411x1*	*B411x2*
alg	36.4	14.6
...	...	...
nzl	25.5	8.8

Úloha B4.12 *Aplikace logistické diskriminační analýzy u rakoviny prostaty* (LDA)
Režim léčení je velmi závislý na velikosti rozšíření rakoviny na lymfatické uzliny, která se zjistí laparotomií. Rozhodující metodou vyšetření uzlin je laparotomie. Brownův postup následujícího vyšetření pěti proměnných u 53 pacientů by měl do jisté míry nahradit obtížnější laparotomické vyšetření specialistou. Nalezněte parametry logistické diskriminační funkce, včetně jejich směrodatných odchylek. Dosazením aktuálních dat každého pacienta obdržíme pacientův koeficient, který vyjadřuje míru nodálního rozšíření, str. 261, ref.[30]. Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: *B412i* index pacienta, *B412x1* věk pacienta, *B412x2* hladina sérové kyselé fosfatázy v Kingových-Armstrongových jednotkách, *B412x3* výsledek roentgenového vyšetření (0 = negativní, 1 = pozitivní), *B412x4* velikost tumoru rektálním vyšetřením (0 = malý, 1 = velký), *B412x5* závěr pathologického bodování z biopsie (0

= méně vážný, 1 = velmi vážný), *B412x6* výsledek laparotomického vyšetření (0 = negativní, 1 = pozitivní přítomnost nodálního rozšíření).

B412i	*B412x1*	*B412x2*	*B412x3*	*B412x4*	*B412x5*	*B412x6*
1	66	0.48	0	0	0	0
...	...	...	...	...	...	
53	68	1.26	1	1	1	1

Úloha B4.13 *Porovnání účinnosti léčení Alzheimerovy nemoci lecithinem s placebo efektem* (EDA, CORA, PCA). Pokuste se zobrazit primární data v exploratorních grafech, především časové závislosti bodového grafu, krabicového grafu a pacientovy profily (řádkové). Sestrojte korelační matici, ve které zobrazíte placebo znakem 0 a licithin znakem 1, str. 10, ref.[30]. Potom proveďte kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Data: *B413i* je index pacienta, *B413x1* značí placebo (=1) nebo lecithin (=2), *B413x2* počet slov opakovaný pacientem ze standardní testační předlohy, *B413x3* totéž po 2. měsíci, *B413x4* totéž po 3. měsí- ci, *B413x5* totéž po 4. měsíci, *B413x6* totéž po 5. měsíci.

B413i	*B413x1*	*B413x2*	*B413x3*	*B413x4*	*B413x5*	*B413x6*
1	1	20	15	14	13	13
...	...	...	...	...	...	...
47	2	10	11	13	17	21

Úloha B4.14 *Aglomerativní hierarchické shlukování při analýze lebek Tibeťanů* (EDA, PCA, CLU). Databáze nalezených lebek na pohřebištích v Tibetu svědčí o dvou skupinách lidí: prvních 13 bylo nalezeno v hrobech v Sikkimu a okolí, zatímco druhých 15 lebek na bojištích okolo Lhasy. Bylo odhadnuto, že tyto lebky patří domorodým bojovníkům z východní provincie Khams. Jedná se o lebky poněkud zvláštní, které se odlišují od lebek Mongolů a Indů, které je ale také na pohřebištích bojišť obklopovaly. Cílem je rozlišit a roztřídit lebky do několika možných shluků tak, jak byly vysloveny hypotézy o původu padlých bojovníků na rozličných bojištích Tibetu. Proveďte postupně kroky *Postupu analýzy vícerozměrných dat*.

Data: *B414i* index lebky, *B414x1* největší délka lebky [mm], *B414x2* největší horizontální šířka lebky [mm], *B414x3* výška lebky [mm], *B414x4* výška horní části obličeje [mm], *B414x5* šířka obličeje mezi body lícních kostí [mm].

B414i	*B414x1*	*B414x2*	*B414x3*	*B414x4*	*B414x5*
1	190.5	152.5	145	73.5	136.5
...	...	...	...	...	...
32	182.5	131	135	68.5	136

Úloha B4.15 *Vícenásobná korespondenční analýza ušních infekcí plavců* (CCA)
Cílem této analýzy bylo odhalit, zda plážoví plavci, kteří plavou především v mělčinách při pobřeží, jsou vystaveni většímu riziku ušní infekce než neplážoví plavci, kteří dávají přednost plavání na otevřeném moři. Kanonickou korelační analýzou je třeba vyšetřit, jaký vliv má v tomto případě i stáří a pohlaví plavce, str. 140, ref.[30] pomocí *Postupu analýzy vícerozměrných dat*.

Data: *B415i* index plavce, *B415x1* značí častý plavec v moři (=1 ano, =2 ne), *B415x2* značí plážový plavec (=1 ne, =4 ano), *B415x3* je věk (=2 je 15-19 let, =3 je 20-25 let, =4 je 26-29 let), *B415x4* značí pohlaví (=1 muž, =2 žena), *B415x5* značí počet samodiagnostikovaných ušních infekcí, které plavec oznámil.

B415i	*B415x1*	*B415x2*	*B415x3*	*B415x4*	*B415x5*
1	1	1	2	1	0
...	...	...	...	...	...
287	2	4	4	2	2

Úloha B4.16 *Faktorová analýza ordinálních proměnných na různých pacientkách* (FA, PCA, CLU). U 29 psychiatricky nemocných pacientek byly analyzovány ordinální proměnné (Everitt[30], str. 284). Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: *B416i* index pacienta, *B416x1* věk, *B416x2* IQ, *B416x3* úzkost (=1 značí žádnou, =2 mírnou, =3 střední, =4 silnou), *B416x4* je skleslost (=1 značí žádnou, =2 mírnou, =3 střední, =4 silnou), *B416x5* značí kvalitu spaní (=1 dobrá, =2 nespavost), *B416x6* značí zájem o sex (=1 ne, =2 ano), *B416x7* značí myšlenky na sebevraždu (=1 ne, =2 ano), *B416x8* parametr weight.

B416i	*B416x1*	*B416x2*	*B416x3*	*B416x4*	*B416x5*	*B416x6*	*B416x7*	*B416x8*
1	39	94	2	2	2	2	2	4.9
...	...	...	...	...	...	...	...	...
118	42	-99	3	2	2	2	2	4.9

Úloha B4.17 *Vícerozměrné škálování u analýzy příbuznosti 14 emocí* (MDS)

Vícerozměrným škálováním posuďte podobnost, příbuznost a vzájemný vztah 14 rozličných emocí u dvou osob, když byla psychology vypracována tabulka vzájemných vzdáleností způsobem porovnání "každé emoce s každou" a při dokonalé podobnosti, nerozlišitelnosti dvou emocí, byla přidělena vzdálenost 0, zatímco při naprosté nepodobnosti pak vzdálenost 10. Párové vzdálenosti byly zapsány do cely symetrické čtvercové matice. Z této matice se ve vstupních datech užije horní trojúhelníková část, tj. prvky nad diagonálou nul pro první osobu a prvky pod diagonálou pro druhou osobu. Aplikujte metodu klasického metrického škálování CMDS a porovnejte se závěry nemetrického škálování NNMDS, str. 121, ref.[30].

Data: prvky trojúhelníkové matice vyjadřují *vzdálenosti* (nepodobnosti, dissimilarities) objektů dvojice emocí: *x1* značí *B417x1* a znamená spokojený, *B417x2* vzrušující, *B417x3* překvapený, *B417x4* horlivý, *B417x5* šťastný, *B417x6* vášnivý, *B417x7* něžný, *B417x8* pohrdá, *B417x9* vylekaný, *B417x10* bázlivý, *B417x11* provinilý, *B417x12* smutný, *B417x13* zlostný, *B417x14* odmítnutý. Horní polovina na diagonálou se týká první osoby, druhá polovina pod diagonálou pak druhé osoby.

Obj	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*	*x14*
x1	-	3	4	6	1	2	3	9	8	9	8	8	8	9
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
x14	9	9	7	9	9	9	9	9	2	1	4	1	3	-

Úloha B4.18 *Sledování spotřeby proteinů v Evropě* (EDA, PCA, FA, CLU)

Byla sledována spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin. Proveďte kroky *Postupu analýzy vícerozměrných dat* a odpovězte na otázky: Existuje korelace mezi proměnnými? Budou data vyžadovat nějakou úpravu, standardizaci nebo centrování? Ukazuje graf komponentních vah na silně korelující proměnné? Dají se ve dvojném grafu odhalit interakce mezi jednotlivými proměnnými-druhy potravin a objekty-zeměmi?

Data: *B418i* značí index, *B418j* název země, *B418x1* červené maso, *B418x2* bílé maso, *B418x3* vejce, *B418x4* mléko, *B418x5* ryby, *B418x6* obilniny, *B418x7* škrob, *B418x8* ořechy, *B418x9* ovoce a zelenina,

i	*j*	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*
1	Albánie	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
..	...	...	...	...	...	...	...	...	...	...
25	Jugoslávie	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Úloha B4.19 *Struktura odpovědí pacientů s AIDS o spokojenosti s lékařem* (FA)
Van Servellen užil škály od Cope et al. (1986), str. 281, ref.[30]. Na 14 otázek odpovědělo několik set pacientů s nemocí AIDS a ohodnotilo svého ošetřujícího lékaře známkou od 1 do 5 (nejhorší). Posuzované položky byly: *1. Lékař mne léčí přátelsky. 2. Mám drobné pochybnosti o schopnostech mého lékaře. 3. Můj lékař je studený a neosobní. 4. Můj lékař činí to nejlepší, aby mne zbavil obav a strachu. 5. Můj lékař vyšetřuje šetrně, co jenom může. 6. Můj lékař by mne měl léčit s větším respektem. 7. Mám jisté pochybnosti o léčení navrženém mým lékařem. 8. Můj lékař se jeví jako velice kompetentní a zkušený. 9. Můj lékař projevuje nefalšovaný zájem o mne jako o osobu. 10. Můj lékař mne nechává v nejistotě s řadou nezodpovězených otázek. 11. Můj lékař užívá slova, kterým já nerozumím. 12. Mám velikou důvěru ve svého lékaře. 13. Cítím, že mohu říci mému lékaři o velmi osobních problémech. 14. Necítím se dobře, mám-li se zeptat svého lékaře na stav mé nemoci.* Provedete faktorovou analýzu s cílem nalézt faktorově čisté předměty po rotaci faktorů metodou Varimax. Navržená pojmenování pro faktory jsou, pro první faktor - důvěra v lékaře, pro druhý faktor - spolehnutí se na lékařovy schoponosti, a třetí faktor - důvěra v doporučené léčení. Při analýze postupujte podle *Postupu analýzy vícerozměrných dat*.

Data: *B419i* index, název předmětu, *B419x1* odpověd na 1. otázku, *B419x2* odpověd na 2. otázku, ..., *B419x14* odpověd na 14. otázku.

i	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*	*x14*
x1	1													
..	...	...	...	...	...	...	...	...	...	...	...	...	...	...
x14	0.62	0.42	0.33	0.47	0.38	0.58	0.51	0.49	0.53	0.56	0.23	0.51	0.56	1

Úloha B4.20 *Struktura kostry kuřat* (PCA, FA)
Dunn publikoval data, týkající se kostry a lebky kuřat bílé lenghornky, str. 558 v ref.[20]. Z 276 měření kostry byla vypočtena korelační matice ***R***. Užijte dvou či tří faktorů v analýze korelační matice ***R***. Navržená pojmenování pro faktory jsou, pro první faktor - velikost těla, pro druhý a třetí faktor týkající se velikosti lebky a jsou pojmenovány shodně s původními proměnnými. Při analýze postupujte podle kroků *Postupu analýzy vícerozměrných dat*.

Data: *B420i* index kuřete, *B420x1* délka lebky, *B420x2* šířka lebky, *B420x3* délka stehenní kosti, *B420x4* délka holení kosti, *B420x5* délka pažní kosti, *B420x6* délka loketní kosti.

B420i	*B420x1*	*B420x2*	*B420x3*	*B420x4*	*B420x5*	*B420x6*
B420x1	1					
..	...	...	...	...	...	...
B420x6	0.603	0.45	0.878	0.894	0.937	1

4.9.2 Analýza chemických a fyzikálních dat

Úloha C4.01 *Faktory ovlivňující koloristickou vydatnost versálové zeleně GL* (EDA, CORA, PCA, FA). Analyzujte vícerozměrná data, popisující koloristickou vydatnost versálové zeleně a zjistěte, který ze sledovaných ukazatelů ji významně ovlivňuje. Pokuste

se snížit počet proměnných aplikací metody hlavních komponent a faktorové analýzy. Postupujte podle kroků *Postupu analýzy vícerozměrných dat* a komentujte míry polohy, rozptýlení a tvaru. Které proměnné nevykazují Gaussovo rozdělení? V korelačním diagramu nalezněte proměnné silně korelující. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Tak se odpoví také na otázky, zda existují vzorky versálové zeleně (objekty), které jsou si podobné ve formě sluníček nebo hvězdiček. Které proměnné silně korelují a které nejsou důležité v grafu komponentních vah? Dají se nalézt v rozptylovém diagramu komponentního skóre nějaké odlehlé objekty? Pokuste se vysvětlit interakci objektů a proměnných, umístěných blízko sebe ve dvojném grafu. Kolik shluků na 60 % hladině podobnosti lze indikovat v dendrogramu podobnosti objektů při užití optimální shlukovací procedury?

Data: *C401i* index vzorku, *C401x1* obsah dusitanu, *C401x2* obsah modři, *C401x3* obsah žluti, *C401x4* pH, *C401x5* měření synthaminu, *C401x6* výtěžek, *C401x7* odstínové odchylky dR, *C401x8* dH, *C401x9* síla.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	35.5	18.52	30.73	3.47	6.5	43.2	0.7	-0.38	100
..	..	..	..	..	..	..	..	..	..
10	34.5	16.47	30.97	0.00	10.0	49.0	-1.4	0.72	105

Úloha C4.02 *Analýza kvality a znečištění různých šarží mleté síry* (EDA, CORA, PCA)
Při analýze 50 šarží mleté síry se zvýšeným obsahem oleje byly nalezeny hodnoty sledovaných parametrů: hmotnost, vlhkost v %, olej v %. Jsou uvedeny také indexy zkoumaných šarží. Pro srovnání jsou uvedeny rovněž limitní hodnoty sledovaných parametrů: vlhkost maximálně 0.3 %, olej minimálně 0.85 %. Metodou hlavních komponent a faktorové analýzy se pokuste snížit počet popisných proměnných a proveďte kroky *Postupu analýzy vícerozměrných dat*. Existují proměnné, které spolu korelují? Kolik shluků podobných objektů indikují hvězdičkové a sluníčkové graf? Které proměnné korelují v grafu komponentních vah? Odhalte odlehlé objekty v rozptylovém diagramu komponentního skóre? Lze v tomto diagramu indikovat nějaké shluky? Ve dvojném grafu vyšetřete interakci objektů a proměnných, umístěných na stejném místě. Jaké vysvětlení lze k tomuto jevu vyslovit? Kolik shluků odhalí nejvhodnější shlukovací metoda v dendrogramu objektů?

Data: *C402i* index vzorku, *C402x1* číslo šarže, *C402x2* hmotnost, *C402x3* vlhkost v %, *C402x4* olej v %.

C402i	*C402x1*	*C402x2*	*C402x3*	*C402x4*
1	5	2400	0.09	1.01
..	..	..	..	..
50	79	7500	0.00	0.87

Úloha C4.03 *Popis rozdílů polyesterových vláken* (EDA, CORA, PCA, FA)
Bylo hodnoceno 30 vzorků polyesterového vlákna a byly zapsány tři charakteristiky, pořadové číslo vlákna, množství aviváže a jemnost. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad

lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Sluníčkovými a hvězdicovými grafy zobrazte uvedená vícerozměrná data s ohledem na vybočující vlákna. Proveďte kroky *Postupu analýzy vícerozměrných dat*.
Data: *C403i* pořadové číslo vlákna, *C403x1* množství aviváže, *C403x2* jemnost.

C403i	*C403x1*	*C403x2*
1	0.9	66
..	..	..
30	0.9	60

Úloha C4.04 *Vliv podmínek na výtěžek redukce triazoloxidu na E-alkohol (EDA, CORA, PCA, FA).* Při výrobě E-alkoholu redukcí triazoloxidu byl sledován vliv proměnných jako jsou objem spotřebovaného 1M NaOH, sušina výtěžku redukce, objem použitého 56% hydrazinu v ml, a objem reakční směsi v ml. Vyšetřete statistickou významnost jednotlivých proměnných podle kroků *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).
Data: *C404i* index výrobku, *C401x1* objem NaOH o koncentraci 1 mol/l v ml, *C404x2* sušina výtěžku E - alkoholu v gramech, *C404x3* objem použitého 56 %ního hydrazinu v ml, *C404x4* objem směsi v ml.

C404i	*C404x1*	*C404x2*	*C404x3*	*C404x4*
1	39.0	259	25.6	230.0
...	..	..	..	..
12	52.0	23.5	21.4	350.0

Úloha C4.05 *Počet částic absorbujících světlo, faktorová analýza spekter* (FA)
Aplikací faktorové analýzy Kankareho metodou, kdy se ze 2. momentu $\boldsymbol{M}$ absorbanční matice $\boldsymbol{A}$, $\boldsymbol{M} = (1/M)\,\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, z vlastních čísel r_a matice $\boldsymbol{M}$ určí směrodatná odchylka absorbance

$$s_k(A) = \sqrt{\frac{\mathrm{tr}(\boldsymbol{M}) - \sum_{a=1}^{k} r_a}{N - k}},$$

(kde tr($\boldsymbol{M}$) je stopa matice $\boldsymbol{M}$ a k je určovaná hodnost matice (čili počet světlo absorbujících komponent), stanovte počet světlo absorbujících složek v rovnovážné směsi komplexů a oligomerů, které vzniknou reakcí sulfoazoxinu SNAZOXS s měďnatými ionty při pH okolo 5. Pro molární poměr $q_M = M/L$ SNAZOXSu L a měďnatých iontů M, proměňující se v intervalu od $q_M = 0.06$ do $q_M = 20.36$, byla naměřena absorbanční matice $\boldsymbol{A}$ 20 spekter (řádky) při 24 vlnových délkách (sloupce, proměnné) viditelné oblasti spektra od 380 do 610 nm v kyvetě 3 cm dlouhé. Diskutujte rozlišovací schopnost a i spolehlivost užité metody.
Data: Sledované proměnné: absorbanční matice u směsi SNAZOXS - Cu^{2+} naměřena: kyveta 2.996 cm, 25EC, octanový pufr pH 5, koncentrace c_{Pb} [mol/l] a molární poměr složek M:L pro spektra (řádky):

(1) 1.15E-06 0.060, (2) 2.31E-06 0.120, (3) 3.46E-06 0.180, (4) 4.62E-06 0.240,
(5) 5.77E-06 0.300, (6) 6.93E-06 0.360, (7) 8.08E-06 0.420, (8) 9.24E-06 0.480,
(9) 1.04E-05 0.540, (10) 1.15E-05 0.600, (11) 1.35E-05 0.700, (12) 1.54E-05 0.800,
(13) 1.73E-05 0.900, (14) 1.92E-05 1.000, (15) 2.12E-05 1.100, (16) 2.31E-05 1.200,
(17) 2.69E-05 1.400, (18) 3.27E-05 1.701, (19) 3.85E-05 2.002, (20) 3.85E-04 20.364,

Absorbance pro vlnové délky, které zde představují proměnné (sloupce) [nm]:

(Spektrum) x_1 380, x_2 390, x_3 400, x_4 410, x_5 420, x_6 430, x_7 440, x_8 450, x_9 460, x_{10} 470, x_{11} 480, x_{12} 490, x_{13} 500, x_{14} 510, x_{15} 520, x_{16} 530, x_{17} 540, x_{18} 550, x_{19} 560, x_{20} 570, x_{21} 580, x_{22} 590, x_{23} 600, x_{24} 610											
(1)	0.3760	0.3917	0.4183	0.4533	0.4989	0.5466	0.5827	0.6141	0.6380	0.6531	0.6619	0.6568
	0.6379	0.6015	0.5421	0.4667	0.3970	0.3363	0.2660	0.1785	0.0948	0.0420	0.0166	0.0077
	..	..	..	..	..	..	..	..	..	..	..	..
(20)	0.4760	0.5000	0.5518	0.6237	0.6988	0.7530	0.7792	0.7751	0.7289	0.6625	0.5903	0.5067
	0.4095	0.3075	0.2141	0.1409	0.0892	0.0564	0.0362	0.0238	0.0164	0.0115	0.0078	0.0073

Úloha C4.06 *Počet částic absorbujících světlo, faktorová analýza spekter* (FA)
Aplikací faktorové analýzy Kankareho metodou analogicky jako v **úloze C4.05** stanovte počet světlo absorbujících složek v rovnovážné směsi komplexů a oligomerů, které vzniknou reakcí sulfoazoxinu m-CAPAZOXS se zinečnatými ionty při pH okolo 5. Pro molární poměr $q_M = M/L$ m-CAPAZOXSu L a zinečnatých iontů M, měnící se v intervalu od $q_M = 0.05$ do 25.39, byla naměřena absorbanční matice $\boldsymbol{A}$ 20 spekter (řádky) při 20 vlnových délkách (sloupce, proměnné) viditelné oblasti spektra od 380 do 570 nm v kyvetě 3 cm dlouhé. Diskutujte rozlišovací schopnost a i spolehlivost užité metody.

Data: Sledované proměnné: absorbanční matice u směsi m-CAPAZOXS - Zn^{2+} naměřena: kyveta 2.996 cm, 25EC, octanový pufr pH 5, koncentrace c_{Zn} [mol/l] a molární poměr složek M:L pro spektra (řádky):

(1) 1.940E-06 0.051, (2) 3.879E-06 0.102, (3) 5.819E-06 0.152,
(4) 7.757E-06 0.203, (5) 1.163E-05 0.305, (6) 1.551E-05 0.406,
(7) 2.325E-05 0.609, (8) 3.486E-05 0.914, (9) 4.646E-05 1.219,
(10) 5.804E-05 1.523, (11) 6.960E-05 1.828, (12) 8.501E-05 2.234,
(13) 1.004E-04 2.641, (14) 1.157E-04 3.047, (15) 1.349E-04 3.555,
(16) 1.731E-04 4.570, (17) 2.680E-04 7.109, (18) 4.552E-04 12.188,
(19) 6.389E-04 17.266, (20) 9.259E-04 25.391

Absorbance pro vlnové délky, které zde představují proměnné (sloupce) [nm]:

(Spektrum) x_1 380, x_2 390, x_3 400, x_4 410, x_5 420, x_6 430, x_7 440, x_8 450, x_9 460, x_{10} 470, x_{11} 480, x_{12} 490, x_{13} 500, x_{14} 510, x_{15} 520, x_{16} 530, x_{17} 540, x_{18} 550, x_{19} 560, x_{20} 570,												
(1)	0.5160	0.4958	0.4938	0.5022	0.5262	0.5583	0.6049	0.6838	0.7683	0.8376	0.9159	0.9349
	0.8532	0.7873	0.7118	0.5091	0.2448	0.0916	0.0331	0.0139				
	..	..	..	..	..	..	..	..	..	..	..	..
(20)	0.5678	0.6660	0.7517	0.8106	0.8432	0.8307	0.8019	0.7793	0.7426	0.6689	0.5581	0.4253
	0.2937	0.1830	0.1059	0.0610	0.0362	0.0214	0.0140	0.0102				

Úloha C4.07 *Počet částic absorbujících světlo, faktorová analýza spekter* (FA)
Aplikací faktorové analýzy Kankareho metodou analogicky jako v **úloze C4.05** stanovte počet světlo absorbujících složek ve směsi tří spektrálních standardů, $K_2Cr_2O_7$ (proměřované koncentraci od 0.057 do 1.143×10^{-3} mol/l), $CoSO_4 \cdot 7H_2O$ (od 0.021 do 0.08786 mol/l), $CuSO_4 \cdot 5H_2O$ (od 0.0 do 0.02786 mol/l). Byla naměřena absorbanční matice $\boldsymbol{A}$ 19 spekter (řádky) při 43 vlnových délkách (sloupce, proměnné) viditelné oblasti spektra od 380 do 800 nm v kyvetě 3 cm dlouhé. Diskutujte rozlišovací schopnost a i spolehlivost užité metody.

Data: Sledované proměnné: absorbanční matice u směsi tří spektrálních standardů naměřena: kyveta 2.996 cm, 25EC, spektra (řádky), absorbance pro vlnové délky, která představují proměnné (sloupce) [nm]:

(Spektrum) x_1 380, x_2 390, x_3 400, x_4 410, x_5 420, x_6 430, x_7 440, x_8 450, x_9 460, x_{10} 470, x_{11} 480, x_{12} 490, x_{13} 500, x_{14} 510, x_{15} 520, x_{16} 530, x_{17} 540, x_{18} 550, x_{19} 560, x_{20} 570, x_{21} 580, x_{22} 590, x_{23} 600, x_{24} 610, x_{25} 620, x_{26} 630, x_{27} 640, x_{28} 650, x_{29} 660, x_{30} 670, x_{31} 680, x_{32} 690, x_{33} 700, x_{34} 710, x_{35} 720, x_{36} 730, x_{37} 740, x_{38} 750, x_{39} 760, x_{40} 770, x_{41} 780, x_{42} 790, x_{43} 800,

(01)	0.0371	0.0269	0.0199	0.0161	0.0155	0.0160	0.0164	0.0165	0.0159	0.0147	0.0136	
	0.0121	0.0114	0.0110	0.0108	0.0104	0.0101	0.0100	0.0097	0.0090	0.0077	0.0058	0.0038
	0.0037	0.0060	0.0066	0.0079	0.0086	0.0089	0.0092	0.0091	0.0090	0.0088	0.0090	0.0105
	0.0104	0.0106	0.0102	0.0100	0.0099	0.0098	0.0095	0.0095				
	...	...	...	...	...	...	...	...	...	...	...	...
(19)	0.4003	0.2888	0.2180	0.1898	0.2150	0.2786	0.3730	0.4984	0.6224	0.6965	0.7551	
	0.8063	0.8889	0.9355	0.8987	0.7822	0.6073	0.4334	0.2949	0.1987	0.1483	0.1311	0.1314
	0.1413	0.1608	0.1839	0.2119	0.2440	0.2790	0.3178	0.3602	0.4056	0.4531	0.5017	0.5516
	0.5980	0.6417	0.6809	0.7151	0.7431	0.7663	0.7817	0.7924				

Úloha C4.08 *Analýza komerčního granulátu polyethylenu (EDA, PCA, CLU)*
Při výrobě granulovaného polyethylenu byly v průběhu roku měřeny u jednotlivých várek produktu mechanické a fyzikálně-chemické parametry. Celkově se jedná o sadu 185 vzorků, u kterých bylo měřeno 9 veličin. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Aplikací metody hlavních komponent posuďte účelnost rozsahu škály měřených veličin, případně se metodou faktorové analýzy pokuste najít a pojmenovat jejich vnitřní vztah. Proveďte také kroky *Postupu analýzy vícerozměrných dat*. Využitím klasifikačních metod rozlište jednotlivé skupiny výrobků tj. várky a typy pomocí jejich vlastností. Které proměnné spolu korelují? Které proměnné mají největší proměnlivost a které největší důležitost? Dají se určit i redundandní proměnné a odlehlé objekty i v centrovaných datech? Kolik shluků lze nalézt na 75 % hladině podobnosti?

Data: *C408i* je index várky, *C408x1* index toku taveniny při 210EC a zatížení 49 N v g/10 min, *C408x2* index toku taveniny při 210EC a zatížení 21 N v g/10 min, *C408x3* hustota granulátu při 23EC v kg/m^3, *C408x4* je rázová houževnatost metodou Charpy při 23EC v kJ/m^2, *C408x5* je mez kluzu v Mpa, *C408x6* je napětí při přetržení v Mpa, *C408x7* tažnost na mezi kluzu v %, *C408x8* celkové protažení při přetržení v %, *C408x9* je tvrdost ve stupních Shore stupnice D.

Typ	*C408x1*	*C408x2*	*C408x3*	*C408x4*	*C408x5*	*C408x6*	*C408x7*	*C408x8*	*C408x9*
BB10	15.32	943.40	14.72	20.30	28.02	7.50	510.50	60.00	0.05
..	..	..	..	..	..	..	..	..	..
ZS70	5.62	942.50	23.07	21.12	27.77	10.00	430.00	63.00	0.02

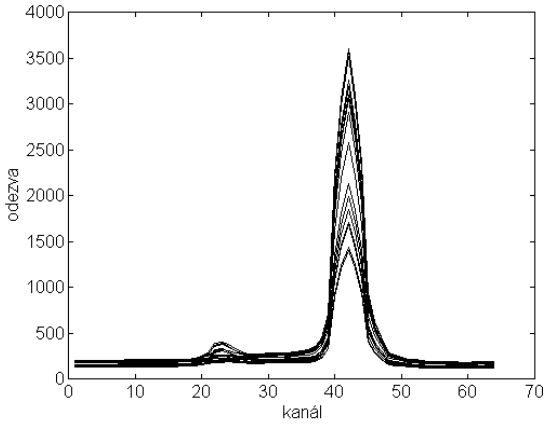
Úloha C4.09 *Analýza komerčního granulátu polypropylenu (EDA, PCA, CLU)*
Sortiment všech komerčních typů lineárního vysokohustotního polyethylenu Liten má hlavní aplikaci ve výrobě tlakových trubek, kterých lze využít pro rozvody vody nebo domovní rozvody plynu. Materiál Liten PL 10 byl podroben rozsáhlému testování a byly stanoveny rheologické a fyzikálně mechanické parametry materiálu, ze kterého byla vyrobena trubka. Aplikací metody hlavních komponent posuďte účelnost rozsahu škály měřených veličin, případně se metodou faktorové analýzy pokuste najít a pojmenovat jejich vnitřní vztah. Proveďte také kroky *Postupu analýzy vícerozměrných dat*. Využitím klasifikačních metod rozlište jednotlivé shluky výrobků tj. várky a typy pomocí jejich vlastností.

Data: C409i je index várky, *C409x1* je index toku taveniny při 210EC a zatížení 21 N v g/10 min (TI2), *C409x2* je index toku taveniny při 210EC a zatížení 49 N v g/10 min (TI5), *C409x3* je index toku taveniny při 210EC a zatížení 212 N v g/10 min, *C409x4* je poměr indexů toku při zatíženích 212 a 21 N (MFR) bezrozměrná veličina, *C409x5* hustota granulátu při 23EC v kg/m^3, *C409x6* pevnost na mezi skluzu (MK) v MPa, *C409x7* je napětí při přetržení (NPP) v MPa, *C409x8* tažnost na mezi kluzu v %, *C409x9* celková tažnost (CT) v %, *C409x10* doba do prasknutí při trubkovém testu (TrT) v hodinách.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
2	0.08	0.056	15.35	191.9	0.9551	21.4	34.7	9.4	573	572
..	..	..	..	..	..	..	..	..	..	..
745	0.08	0.42	11.11	144.3	0.952	20.6	31.7	6.4	537	2558

Úloha C4.10 *Vícerozměrná kalibrace bezolovnatých benzinů*
S pomocí dostupného softwaru sestrojte vícerozměrný kalibrační model pro 4 parametry kvality benzinů prezentované v kapitole 4.8. Interpretujte získané výsledky. Diskutujte aplikovatelnost metody pro rutinní analýzu.

Úloha C4.11 *Vícerozměrná kalibrace oceli pomocí ICP dat*
Sestrojte vícerozměrný kalibrační model umožňující stanovení koncentrace prvku 1 a prvku 2 v ušlechtilé oceli na základě ICP dat. Spektra jsou zobrazena na následujícím obrázku. Porovnejte modely dosažené pro obě proměnné a diskutujte jejich kvalitu a použitelnost.



NIR spektra

Úloha C4.12 *Vícerozměrná kalibrace polymeru pomocí NIR spekter*
Na základě NIR spekter sestrojte kalibrační model pro komplexní charakteristiky polymeru $\boldsymbol{c}_1$ a $\boldsymbol{c}_2$. Spektra jsou prezentována na obrázku. Diskutujte dosažené výsledky.

Úloha C4.13 *Závislost podélné a příčné síly papíru na hustotě (EDA, PCA, CLU)*
Papír je vyráběn ve spojitých rolích několik stop širokých, str. 17 v ref.[20]. Orientace vláken v papíru způsobuje, že papír má rozdílnou sílu v podélném a příčném směru, který je kolmý na podélný směr. Pro různé hustoty papíru byla naměřená síla ve směru podélném a příčném. Proveďte kroky *Postupu analýzy vícerozměrných dat* a odpovězte otázky: existuje nějaká významná korelace mezi proměnnými? Která proměnná má největší

proměnlivost a důležitost? Jsou některé proměnné blízko sebe v grafu komponentních vah? Kolik shluků lze detekovat v diagramu komponentního skóre? Která shlukovací metoda se jeví nejlepší při výstavbě dendrogramu? Kolik shluků na 80 % hladině podobnosti lze v dendrogramu detekovat?

Data: *C413i* index, *C413x1* hustota papíru [g/cm^3], *C413x2* síla měřená podélně [libra], *C413x3* síla měřená příčně [libra].

C413i	*C413x1*	*C413x2*	*C413x3*
1	0.801	121.41	70.42
...	...	...	...
41	0.758	113.8	52.42

4.9.3 Analýza environmetálních, potravinářských a zemědělských dat

Úloha E4.01 *Grafická prezentace a průzkumová analýza dat hornin z vrtů* (EDA, PCA, CLU)
Proveďte grafickou prezentaci 53 vzorků hornin, získaných z hlubiných vrtů 4500 stop pod zemí v horách Colorada, u kterých byl sledován obsah 12 rozličných kationtů a aniontů (proměnné). Proveďte kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).
Data: *E401i* index objektu, proměnné *E401x1* až *E401x12* představují obsah 12 kationtů a aniontů [100 mg/kg] v hornině z vrtu 4500 stp hlubokého v horách v Coloradu.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
1	320	105	057	050	001	001	001	060	020	250	210	370
..	..	..	..	..	..	..	..	..	..	..	..	
53	425	60	35	5	1	1	30	100	10	340	1	10

Úloha E4.02 *Sledování kvality říční vody v rozličných místech* (EDA, PCA, CLU)
Ze sledovaných 36 parametrů kvality říční vody bylo vybráno (a) 11 parametrů a (b) 7 parametrů, a oba výběry byly analyzovány postupy vícerozměrné statistické analýzy. Užijte standardizovaná data. Z 54 míst sledované kvality vody na šesti říčních profilech (označených MSH, OB, OLR, OLCT, OLV, OM) v období od ledna do září 1995 proveďte klasifikaci charakteristických vlastností, nejlépe popisujících variabilitu objektů a vzájemné porovnání. Proveďte také kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Především vytvořte shluky podobných říčních profilů.
Data:
(a) *E402Ai* index vzorku, *E402Ax1* vodivost, *E402Ax2* tvrdost, *E402Ax3* nerozpuštěné látky veškeré sušené při 105EC, *E402Ax4* nerozpuštěné látky žíhané při při 550EC, *E402Ax5* nerozpuštěné látky ztráta žíháním (dopočet) [mg/l], *E402Ax6* rozpuštěné látky sušené při 105EC, *E402Ax7* rozpuštěné látky žíhané při 550EC [mg/l], *E402Ax8* rozpuštěné látky ztráta žíháním (dopočet) [mg/l], *E402Ax9* pH, *E402Ax10* KNK4,5, *E402Ax11* procento nasycení kyslíkem [%]. Data obsahují řádek vzorku *i*, označení místa na profilu řeky a teplotu vody v EC: 1 MSH 1 0.2EC, 2 MSH 2 4.8EC, 3 MSH 3 1.9EC, 4 MSH 4 8.9EC, 5 MSH 5 9.7EC, 6 MSH 6 14.7EC, 7 MSH 7 16.5EC, 8 MSH 8 12.4EC, 9 MSH 9 11.9EC, 10 OB 1 1.9EC, 11 OB 2 5.0EC, 12 OB 3 3.5EC, 13 OB 4 8.5EC, 14 OB 5 9.7EC, 15 OB 6 14.1EC, 16 OB 7 19.1EC, 17 OB 8 20.7EC, 18 OB 9 18.1EC, 19 OLR 1 1.0EC, 20 OLCT 1 1.3EC, 21 OLV 1 1.4EC, 22 OLR 2 2.0EC, 23 OLCT 2 2.0EC, 24 OLV 2 3.0EC, 25 OLR 3 5.4EC, 26 OLCT 3 5.9EC, 27 OLV 3 6.2EC, 28 OLR 4 5.4EC, 29 OLCT 4 6.7EC, 30 OLV 4 8.0EC, 31 OLR 5 8.9EC, 32 OLCT 5 10.8EC, 33 OLV 5 12.4EC, 34 OLR 6 13.2EC, 35 OLCT 6 14.6EC, 36 OLV 6 16.4EC, 37 OLR 7 19.5EC, 38 OLCT 7 19.7EC, 39 OLV 7 20.9EC, 40 OLR 8 16.9EC, 41 OLCT 8 17.8EC, 42 OLV 8 18.0EC, 43 OLR 9 12.8EC, 44 OLCT 9 13.5EC, 45 OLV 9 14.6EC, 46 OM 1 3.6EC, 47 OM 2 8.1EC, 48 OM 3 7.0EC, 49 OM 4 7.5EC, 50 OM 5 10.5EC, 51 OM 6 18.3EC, 52 OM 7 20.4EC, 53 OM 8 20.6EC, 54 OM 9 14.4EC,

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	19.4	0.832	25	14	11	146	97	49	7.55	0.7	13.593

..	..	..	..	..	..	..	..	..	..	..	..
54	55.3	1.685	77	61	16	427	283	144	7.76	2.0	8.079

(b) *Data:* *E402Bi* index místa, *E402Bx1* koncentrace draslíku [mg/l], *E402Bx2* koncentrace železa (celkového) [mg/l], *E402Bx3* koncentrace manganu [mg/l], *E402Bx4* koncentrace zinku [mg/l], *E402Bx5* koncentrace mědi [mg/l], *E402Bx6* koncentrace niklu [mg/l], *E402Bx7* koncentrace chromu [mg/l]. Data obsahují řádek vzorku *i*, místo na profilu řeky a teplotu vody v EC:
1 MSH 1 0.2EC, 2 MSH 2 4.8EC, 3 MSH 3 1.9EC, 4 MSH 4 8.9EC, 5 MSH 5 9.7EC, 6 MSH 6 14.7EC, 7 MSH 7 16.5EC, 8 MSH 8 12.4EC, 9 MSH 9 11.9EC, 10 OB 1 1.9EC, 11 OB 2 5.0EC, 12 OB 3 3.5EC, 13 OB 4 8.5EC, 14 OB 5 9.7EC, 15 OB 6 14.1EC, 16 OB 7 19.1EC, 17 OB 8 20.7EC, 18 OB 9 18.1EC, 19 OLR 1 1.0EC, 20 OLCT 1 1.3EC, 21 OLV 1 1.4EC, 22 OLR 2 2.0EC, 23 OLCT 2 2.0EC, 24 OLV 2 3.0EC, 25 OLR 3 5.4EC, 26 OLCT 3 5.9EC, 27 OLV 3 6.2EC, 28 OLR 4 5.4EC, 29 OLCT 4 6.7EC, 30 OLV 4 8.0EC, 31 OLR 5 8.9EC, 32 OLCT 5 10.8EC, 33 OLV 5 12.4EC, 34 OLR 6 13.2EC, 35 OLCT 6 14.6EC, 36 OLV 6 16.4EC, 37 OLR 7 19.5EC, 38 OLCT 7 19.7EC, 39 OLV 7 20.9EC, 40 OLR 8 16.9EC, 41 OLCT 8 17.8EC, 42 OLV 8 18.0EC, 43 OLR 9 12.8EC, 44 OLCT 9 13.5EC, 45 OLV 9 14.6EC, 46 OM 1 3.6EC, 47 OM 2 8.1EC, 48 OM 3 7.0EC, 49 OM 4 7.5EC, 50 OM 5 10.5EC, 51 OM 6 18.3EC, 52 OM 7 20.4EC, 53 OM 8 20.6EC, 54 OM 9 14.4EC,

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	1.90	0.13	0.15	0.08	0.0025	0.005	0.005
..	..	..	..	..	..	..	..
54	19.00	1.53	0.20	0.36	0.0100	0.012	0.010

Úloha E4.03 *Shluková analýza vodárenských dat* (EDA, PCA, CLU)
Na 32 vzorcích pitné vody, u kterých bylo provedeno stanovení 17 proměnných kvality a výsledky byly rozděleny do tří výběrů: (a) výběr o prvních 9 proměnných, (b) výběr o zbylých 8 proměnných, (c) výběr všech 17 proměnných. Bylo odebráno celkem 28 vzorků na úpravnách vody a v přilehlé vodovodní síti 3 zdrojů, přičemž 4 vzorky jsou zakázky cizích organizací. Proveďte kroky *Postupu analýzy vícerozměrných dat* a odpovězte na otázky: které proměnné silně korelují? Je třeba data standardizovat nebo centrovat dle krabicového grafu všech proměnných? Ukazují sluníčka nebo hvězdičky na podobné vzorky vody (objekty)? Která původní proměnná nejvíce přispívá do první hlavní komponenty y_1 a která do druhé y_2? Která původní proměnná je nejdůležitější a která je nejméně důležitá - redundandní? Jsou v tomto grafu indikovány vůbec nějaké redundantní proměnné? Ukazuje graf komponentních vah na nějaké korelující proměnné? Pokuste se podle umístění původních proměnných ve dvojném grafu pojmenovat hlavní komponenty y_1 a y_2. Indikuje rozptylový diagram komponentního skóre nějaké odlehlé objekty? Změní se počet shluků v tomto diagramu při odstranění silně vybočujících objektů? Kolik shluků objektů odhaluje dendrogram podobnosti objektů za použití optimální shlukovací procedury? Shlukovou analýzou nalezněte vhodné třídění zdrojů vody. Lze nalézt i dominantní proměnné, které je třeba měřit v analytické laboratoři a redundantní proměnné, které je možné vypustit?
Data:
(a) *E403Ai* index vzorku, *E403Ax1* barva [mg Pt/l], *E403Ax2* zákal, *E403Ax3* vodivost [S/cm^2], *E403Ax4* pH, *E403Ax5* VL [mg/l], *E403Ax6* chemická spotřeba kyslíku CHSK [mg/l], *E403Ax7* tvrdost [mmol/l], *E403Ax8* alkalita [mmol/l], *E403Ax9* obsah železa [mg/l].

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	0.00	0.00	518	7.50	164	0.4	3.33	5.28	0.09
..	..	..	..	..	..	..	..	..	..
32	1.08	0.00	581	7.38	410	0.1	3.33	5.15	0.07

(b) *E403Bi* index vzorku, *E403Bx1* obsah manganu [mg/l], *E403Bx2* obsah amonných iontů [mg/l], *E403Bx3* obsah chloridů [mg/l], *E403Bx4* obsah síranů [mg/l], *E403Bx5* obsah dusičnanů [mg/l], *E403Bx6* obsah saponátů [mg/l], *E403Bx7* obsah huminových látek [mg/l], *E403Bx8* absorbance.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	0.02	0.06	16.3	109.5	10.3	0.06	0.00	0.01
..	..	..	..	..	..	..	..	..
32	0.06	0.05	13.8	91.2	9.59	0.05	0.00	0.00

Úloha E4.04 *Klasifikace zdrojů pitné vody* (EDA, PCA, CLU)
Na 62 vzorcích pitné vody, u kterých bylo provedeno stanovení 16 proměnných kvality, proveďte vícerozměrnou statistickou analýzu. Proveďte kroky *Postupu analýzy vícerozměrných dat* a ověřte, zda EDA ukazuje (např. krabicový graf), že je třeba proměnné standardizovat. Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Shlukovou analýzou proveďte klasifikaci zdrojů.

Data: *E404i* index vzorku, *E404x1* obsah dusičnanů [mg/l], *E404x2* obsah dusitanů [mg/l], *E404x3* obsah chloridů [mg/l], *E404x4* obsah celkového chloru [mg/l], *E404x5* obsah síranů [mg/l], *E404x6* obsah fosforečnanů [mg/l], *E404x7* obsah amonných solí [mg/l], *E404x8* obsah vápníku [mg/l], *E404x9* obsah hořčíku [mg/l], *E404x10* obsah železa (celkového) [mg/l], *E404x11* obsah manganu [mg/l], *E404x12* pH, *E404x13* KNK, *E404x14* ZNK, *E404x15* vodivost, *E404x16* nerozpuštěné látky [mg/l].

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
1	2.2	0.00	6.	6.	103.5	0.032	0.02	181	17	0.016	0.05	7.08	8.1	3.40	855	0.09
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
62	32.8	0.01	25.	25.	115.5	0.050	0.02	102	12	0.016	0.05	7.69	2.6	0.65	436	0.05

Úloha E4.05 *Porovnání a hodnocení mycích procesů v laboratoři* (EDA, PCA, CLU)
Při hodnocení mycích procesů byly testovány různé způsoby mycích procesů tzn. kombinace různě teplých vod s přípravky, umožňujícími snadnější mytí. Zde pak za přípravky chápeme HNO_3, NaOH, detergenty, přičemž písmeno u mycí metody odlišuje teplotu vody, použitou pro mytí, u metod se dvěma písmeny je teplota vody na dvou úrovních, vysoké nebo nízké. U ostatních metod jsou použity i jiné úrovně teploty mycího roztoku. Na základě zkušenosti obsluhy byly mycí procesy testovány na třech skupinách přípravků - snadno omyvatelné, středně omyvatelné, špatně omyvatelné. Pro hodnocení umytí zařízení byla použita pětistupňová škála: 1 značí čistě umyté, 2 značí mírně znečištěné, 3 značí středně znečištěné, 4 značí silně znečištěné, 5 značí neumyté.

Proveďte kroky *Postupu analýzy vícerozměrných dat* a rozhodněte o (a) správnosti expertního odhadu obsluhy přidělení přípravků do skupin, na kterých byly testovány mycí postupy, (b) přiřazení skupin přípravků nejvhodnějším mycím procesům. Které mycí procesy (proměnné) spolu korelují? Které hvězdičky nebo sluníčka jsou si podobná a tím svědčí o podobném přípravku (objektu)? Z grafu komponentích vah určete, které proměnné jsou v silné korelaci, a které dle své důležitosti je možné vypustit. Z dvojného grafu odhalte a vysvětlete interakci postupu (proměnné) a přípravku (objektu). Dendrogram odhaluje shluky silně podobných přípravků.

Porovnáním výsledků rozhodněte o použitelnosti jednotlivých mycích postupů pro jednotlivé skupiny přípravků. Rozhodování provádějte za předpokladu, že mycí režimy

jsou nastaveny tak, že množství použité oplachové vody je u všech cyklů stejné a náklady na jednotlivé mycí procesy jsou uspořádány podle následujícího pořadí: mytí H_2O < mytí pomocí HNO_3 < mytí pomocí NaOH < mytí pomocí detergentů, když časovou náročnost jednotlivých mycích postupů neuvažujeme; $x_2 < x_3 < x_4 < x_5 < x_6 < x_7 < x_8 < x_9 < x_{10} < x_{11} < x_{12}$.

Data: *E405i* index přípravku, *E406x1* až *E406x13* představují číslo přípravku, kód omyvatelnosti, a hodnocení umytí ve škále 1 až 5 pro všechny testované mycí postupy x_1 až x_{12}.

Přípravek *i*	Omyvat. x_1	H_2O				HNO_3		NaOH		DETERG.		
		x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
1	+	2	1	1	1	3	3	2	1	2	1	1
..	..	..	..	..	..	..	..	..	..	..	..	..
13	-	4	3	2	2	4	3	2	1	2	1	1

Úloha E4.06 *Průzkumová analýza a klasifikace vlastností rozličných druhů kávy* (EDA, PCA, CLU)

Byl získán výběr 43 vzorků kávy, pocházejících ze 30 zemí. U každého druhu kávy byly změřeny jeho chemické a fyzikální vlastnosti. Testujte tato data, zda-li splňují požadavky na homogenitu, nebo zda je možné indikovat dvě či více rozličných kategorií. Interpretujte data graficky. Proveďte kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu). Vytvořte dendrogram klasifikovaných druhů kávy.

Data: Soubor dat COFFEE obsahuje na 43 vzorcích kávy (řádky) popis pomocí 13 proměnných (sloupce) v pořadí: *E406i* index kávy, *E406j* je původ kávy, *E406x1* obsah vody, *E406x2* hmotnost zrn, *E406x3* extrakt, *E406x4* pH, *E406x5* volná acidita, *E406x6* obsah minerálů, *E406x7* tuky, *E406x8* kofein, *E406x9* trinonelin, *E406x10* kyselina chlorogeniková, *E406x11* kyselina neochlorogeniková, *E406x12* kyseliny isochlorogeniková, *E406x13* suma kyselin chlorogenikových.

i	*j*	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1.0	5.4	0.4	0.8	6.6
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
43	Hawai	9.7	191.2	35.1	5.6	34.6	4.2	14.2	1.1	0.9	0.7	0.5	0.3	6.5

Úloha E4.07 *Průzkumová analýza dat o znečištění ovzduší* (EDA, PCA, CLU, LDA)

Chromatograficky bylo stanoveno 10 sloučenin ve vzorcích ovzduší, odebraných z 20 míst ve městě. Ve městě existují 4 aktivní centra znečištění ovzduší a složení znečištění každého zdroje je rovněž uvedeno v datech. Vyšetřete, zda jsou ve městě skutečně pouze 4 uvedené zdroje znečištění či zda je zdrojů více. Proveďte kroky *Postupu analýzy vícerozměrných dat* a odpovězte na otázky: které proměnné silně korelují? Je třeba data standardizovat nebo centrovat dle krabicového grafu všech proměnných? Ukazují sluníčka nebo hvězdičky na podobné vzorky vody (objekty)? Která původní proměnná nejvíce přispívá do první hlavní komponenty y_1 a která do druhé y_2? Která původní proměnná je nejdůležitější a která je nejméně důležitá-redundandní? Jsou v tomto grafu indikovány nějaké redundantní proměnné? Ukazuje graf komponentních vah na nějaké korelující proměnné? Pokuste se podle umístění původních proměnných ve dvojném grafu pojmenovat hlavní

komponenty y_1 a y_2. Indikuje rozptylový diagram komponentího skóre nějaké odlehlé objekty? Změní se počet shluků v tomto diagramu při odstranění silně vybočujících objektů? Kolik shluků objektů odhaluje dendrogram podobnosti objektů za použití optimální shlukovací procedury? Shlukovou analýzou nalezněte vhodné třídění zdrojů vody. Lze nalézt i dominantní proměnné, které je třeba měřit v analytické laboratoři a redundantní proměnné, které je možné vypustit?

Data: *E407i* index místa, *E407x1* až *E407x10* představují 10 proměnných ke sledování znečištění, 20 vyšetřovaných míst ve městě (řádky).

Objekt i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	18.6	18.3	8.1	22.8	16.0	37.6	24.8	11.8	28.3	29.4
..	..	..	..	..	..	..	..	..	..	..
20	18.2	16.3	5.2	23.3	13.9	29.0	21.8	14.3	22.5	22.6
Zdroj										
A	35.0	40.0	60.0	75.0	40.0	30.0	55.0	50.0	80.0	40.0
B	40.0	20.0	20.0	70.0	20.0	80.0	40.0	50.0	30.0	60.0
C	20.0	60.0	55.0	80.0	10.0	45.0	75.0	30.0	70.0	20.0
D	30.0	65.0	10.0	20.0	45.0	75.0	65.0	15.0	85.0	65.0

Úloha E4.08 *Faktorová analýza při klasifikaci vzorků vín* (EDA, PCA, CLU)
Pro 90 vzorků italských vín bylo naměřeno 8 fyzikálně-chemických vlastností. Ve vínech jsou obsaženy tři kultury, a to Nebbiolo ve víně Barolo, Grignolino a Barbera ve vínech stejného jména, a to každá ve 30 vzorcích. Vzorky byly odebírány v průběhu několika let. Vyšetřete, kolik faktorů rozliší tři kategorie vín. Proveďte kroky *Postupu analýzy vícerozměrných dat* a rozhodněte, které proměnné jsou nejlepší co do efektivnosti a které původní proměnné jsou téměř nevýznamné - redundandní. Které proměnné jsou v korelaci? Do kolika shluků lze vína roztřídit? Souvisí počet shluků se zadanými druhy vín?

Data: *E408i* index vzorku vína, *E408j* jméno vzorku vína, *E408k* kategorie vzorku vína. Data obsahují 90 druhů vín (objekty, řádky) tří kategorií 1. Barolo, 2. Grignolino a 3. Barbera, popsaných 8 následujícími vlastnostmi (proměnné, sloupce): *E408x1* obsah alkoholu, *E408x2* necukerný extrakt, *E408x3* fosfáty, *E408x4* celkové fenoly, *E408x5* flavanoidy, *E408x6* poměr absorbancí při 280 a 315 nm pro naředěné víno, *E408x7* poměr absorbancí při 280 a 315 nm pro flavanoidy, *E408x8* obsah prolinu.

i	j	*k*	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	Olo0171	1	14.23	24.82	320	2.80	3.06	3.92	4.77	1065
..	..	..	..	..	..	..	..	..	..	..
90	Era2878	3	13.17	23.45	534	1.65	0.68	1.62	2.05	840

Úloha E4.09 *Klasifikace čistého mléka dle složení z různých komponent* (EDA, PCA, CLU). Z velkého souboru dat složení a parametrů mléka, určených metodou GC a HPLC, byl vybrán hypotetický výběr 17 rozličných zdrojů mléka, u kterých bylo sledováno procento tří mastných kyselin. Ve výběru je 7 vzorků kravského mléka ($p = 1$), 7 vzorků směsi 80 % kravského a 20 % kozího mléka ($p = 2$) a konečně 3 vzorky mléka, které nepatří ani do jednoho z obou výběrů. Proveďte kroky *Postupu analýzy vícerozměrných dat* a odpovězte na otázky: odhalte korelaci v datech. Ukazuje krabicový graf na nutnost standardizace nebo alespoň centrování? Ukazuje hvězdičkový graf na podobné vzorky mléka? Jsou původní proměnné důležité nebo redundandní v grafu komponentních vah? Ukazuje rozptylový diagram komponentího skóre na nějaké odlehlé objekty, které je třeba

odstranit? Odhaluje dendrogram objektů silně odlehlý objekt, který by bylo třeba z dat odstranit?

Data: *E409i* index vzorku tuku mléka, *E409p* je třída druhu mléka, procento tří mastných kyselin *E409x1* = FA1 [%], *E409x2* = FA2 [%], *E409x3* = FA3 [%] v 17 vzorcích mléčného tuku (řádky).

Vzorek *i*	Třída *p*	x_1	x_2	x_3
1	1	3.5	12.8	8.9
..	..	..	..	..
17	0	3.8	5.5	13.2

Úloha E4.10 *Faktory ovlivňující výnosnost petržele* (EDA, PCA, CLU)
Při biologickém sledování výnosnosti petržele na semeno byly sledovány následující charakteristiky odrůdy petržele, jako jsou průměr květu rostliny, výška rostliny, šířka rostliny a výnos semena na jednu rostlinu. Pokuste se o klasifikaci 14 druhů petržele a vyjádření závislosti výnosu na jednotlivých vlastnostech květu. Proveďte kroky *Postupu analýzy vícerozměrných dat* a rozhodněte analýzou korelační matice, která proměnná silně koreluje a které proměnné nekorelují. Pokuste se tento efekt biologicky vysvětlit. Lze najít ve hvězdičkovém grafu podobné druhy petržele? Určete nejlepší model hlavních komponent PCA a pokuste se odhalit vybočující objekty petržele. Kolik shluků petržele lze detekovat v dendrogramu?

Data: *E410i* je index druhu petržele, *E410x1* průměr květu rostliny, *E410x2* výška rostliny, *E410x3* šířka rostliny, *E410x4* výnos semena na jednu rostlinu.

i	x_1	x_2	x_3	x_4
1	4.5	7.0	15.0	0.190
..	..	..	..	..
14	5.5	9.5	15.0	0.480

Úloha E4.11 *Struktura a vazby proměnných při sledování kvality životního prostředí* (EDA, PCA, CLU). Po rozličné dny bylo zaznamenáváno 42 měření pro 7 sledovaných proměnných životního prostředí, a to vždy přesně ve 12 hodin v Los Angeles, str. 37 v ref.[20]. Proveďte kroky *Postupu analýzy vícerozměrných dat* a odpovězte na otázky: které proměnné silně korelují? Dle krabicového grafu všech proměnných je třeba data standardizovat nebo centrovat? Ukazují sluníčka nebo hvězdičky na podobné vzorky vody (objekty)? Která původní proměnná nejvíce přispívá do první hlavní komponenty y_1 a která do druhé y_2? Která původní proměnná je nejdůležitější a která je nejméně důležitá - redundandní? Ukazuje graf komponentních vah na nějaké korelující proměnné? Pokuste se podle umístění původních proměnných ve dvojném grafu pojmenovat hlavní komponenty y_1 a y_2. Indikuje rozptylový diagram komponentního skóre nějaké odlehlé objekty? Změní se počet shluků v tomto diagramu při odstranění silně vybočujících objektů? Kolik shluků objektů odhaluje dendrogram podobnosti objektů za použití optimální shlukovací procedury? Shlukovou analýzou nalezněte vhodné třídění. Lze nalézt i dominantní proměnné, které je třeba měřit v analytické laboratoři a redundantní proměnné, které je možné vypustit?

Data: *E411i* index dne, ve kterém bylo měřeno, *E411x1* rychlost větru, *E411x2* sluneční záření, *E411x3* obsah CO, *E411x4* obsah NO, *E411x5* obsah NO_2, *E411x6* obsah O_3, *E411x7* obsah CH.

E411i	*E411x1*	*E411x2*	*E411x3*	*E411x4*	*E411x5*	*E411x6*	*E411x7*
1	8	98	7	2	12	8	2

...	...	...	...			...	...
42	8	40	4	3	6	5	2

Úloha E4.12 *Podobnost vlastností křupavých lupínků od různých výrobců* (EDA, PCA, CLU). Tři americké firmy General Mills (G), Kellogg (K) a Quaker (Q) produkují křupavé obilné lupínky ke snídani. U řady produktů bylo sledováno 10 proměnných a vyšetřována struktura a vzájemné vazby mezi vlastnostmi jednotlivých produktů, ale i proměnných. Které produkty jsou si velice podobné? Proveďte kroky *Postupu analýzy vícerozměrných dat* a rozhodněte, které proměnné silně korelují? Dle krabicového grafu všech proměnných je třeba data standardizovat nebo centrovat? Ukazují sluníčka nebo hvězdičky na podobné vzorky (objekty)? Která původní proměnná nejvíce přispívá do první hlavní komponenty y_1 a která do druhé y_2? Která původní proměnná je nejdůležitější a která je nejméně důležitá - redundandní? Jsou v tomto grafu indikovány nějaké redundantní proměnné? Ukazuje graf komponentních vah na nějaké korelující proměnné? Pokuste se podle umístění původních proměnných ve dvojném grafu pojmenovat hlavní komponenty y_1 a y_2. Indikuje rozptylový diagram komponentního skóre nějaké odlehlé objekty? Změní se počet shluků v tomto diagramu při odstranění silně vybočujících objektů? Kolik shluků objektů odhaluje dendrogram podobnosti objektů za použití optimální shlukovací procedury? Shlukovou analýzou nalezněte vhodné třídění. Lze nalézt i dominantní proměnné a redundantní proměnné, které je možné vypustit?

Data: *E412i* index obilných lupínků, *E412x1* výrobce G, K či Q, *E412x2* kalorická hodnota [cal], *E412x3* bílkoviny, *E412x4* tuk, *E412x5* sodné kationty, *E412x6* vláknina, *E412x7* uhlovodíky, *E412x8* cukr, *E412x9* draselné kationty, *E412x10* skupina.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	ACCheerios	G	110	2	2	180	1.5	10.5	10	70	1
...	...	...	...	...	...	...	...	...	...	...	..
55	QuakerOatmeal	Q	100	5	2	0	2.7	1	1	110	3

Úloha E4.13 *Posouzení struktury, kvality a ceny hovězího masa mladých býků* (EDA, PCA, CLU). U 76 býků mladších dvou let byly sledovány vlastnosti, determinující výslednou kvalitu masa a dále cena, za kterou byli býci prodáni na jatka, str. 46 v ref.[20]. Soustřeďte se hlavně na *E413x1, E413x6* a *E413x7* a potom *E413x1, E413x4* a *E413x9*. Je možné vytvořit index "velikosti těla" nebo index "tvaru těla", uvažujíc sedm proměnných. Objasněte. Podaří se v grafech PCA nalézt tři skupiny plemene býků? Sestrojte *Q-Q* graf z první komponenty a vysvětlete.

Data: *E413i* index býka, *E413x1* plemeno (1 značí Angus, 5 značí Hereford, 8 značí Simental), *E413x2* prodejní cena [US $], *E413x3* výška v kohoutku u ročního býka [palce], *E413x4* hmotnost těla bez tuku [libry], *E413x5* procento hmoty masa bez tuku [%], *E413x6* velikost ve stupnici 1 (malý) až 8 (velký), *E413x7* hřbetní tuk [palce], *E413x8* výška v kohoutku při prodeji býka [palce], *E413x9* hmotnost býka [libry].

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	1	2200	51	1128	70.9	7	0.25	54.8	1720
...	...	...	...	...	...	...	...	...	...
76	8	1500	51.7	992	70.6	7	0.15	55.1	1458

Úloha E4.14 *Struktura a vazby mezi druhy a dobou odchycených ryb na 28 jezerech* (EDA, PCA, CLU). V průběhu pěti let byl v 90. letech dotázán vybraný vzorek rybářů na dobu k chycení určitého počtu ryb na 28 jezerech ve Wisconsinu. Šlo o ryby velice chutné,

a proto oblíbené v Kanadě. Doba na chycení 6 druhů ryb byla normována na 1 hodinu a databáze reakcí rybářů měla 120 řádků odpovědí. Z matice vstupních dat byla vypočtena korelační matice, která je nyní k dispozici. Lze očekávat, že ryby chycené jedním rybářem budou vždy ve stejném shluku.

Data: Korelační matice sledovaných proměnných: *E414x1* ryba Bluegill, *E414x2* ryba Black crappie, *E414x3* ryba Smallmouth bass, *E414x4* Largemouth bass, *E414x5* Walley, *E414x6* Northen pike.

1	0.4919	0.2636	0.4653	-0.2277	0.0652
0.4919	1	0.3127	0.3506	-0.1917	0.2045
0.2635	0.3127	1	0.4108	0.0647	0.2493
0.4653	0.3506	0.4108	1	-0.2249	0.2293
-0.2277	-0.1917	0.0647	-0.2249	1	-0.2144
0.0652	0.2045	0.2493	0.2293	-0.2144	1

Úloha E4.15 *Korelace mezi klimatickými a environmentálními proměnnými při vyšetřování ovzduší* (EDA, CORA, PCA, FA, CLU). Pokuste se v rozptylovém diagramu nebo v grafu korelační matice vyšetřit korelaci dvojic proměnných. V rozptylových diagramech (Casement Plot a Draftsman Plot) vyšetřete trendy či závislosti mezi klimatickými a environmentálními proměnnými a diskutujte míru korelace. Vložte do rozptylového grafu přímku $y = x$ a vyšetřete procento bodů, odchylujících se od této přímky. Jeví se závislost jako nelineární, tzn. vystižená spíše křivkou? Byl by zde užitečný i 3D obrázek? Doplňte osu x a osu y diagramem rozptýlení a sledujte také rozptýlení těchto hodnot. Testujte, zda vícerozměrné rozdělení je normálního charakteru pomocí Mahalanobisovy vzdálenosti každého objektu od jeho střední hodnoty. Pro tuto veličinu sestrojte *Q-Q* graf, resp. rankitový graf. Je možné sestrojit rankitový graf i pro každou proměnnou odděleně, str. 30, ref.[30]. Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: *E415i* index, *E415j* název města, *E415x1* koncentrace SO_2 [µg/m^3], *E415x2* roční průměrná teplota [EF], *E415x3* počet podniků s více než 20 zaměstnanci, *E415x4* počet obyvatelstva v tisících, *E415x5* průměrná rychlost větru [míle/hod], *E415x6* roční srážkový průměr [inch], *E415x7* počet deštivých dní v roce [dny].

i	*j*	*E415x1*	*E415x2*	*E415x3*	*E415x4*	*E415x5*	*E415x6*	*E415x7*
1	Phoenix	10	70.3	213	582	6	7.05	36
	...	...	...	...	...	...	...	...
41	Milwak	16	45.7	569	717	11.8	29.07	123

Úloha E4.16 *Vícerozměrné škálování u analýzy podobnosti 10 výrobků Coly* (MDS)
Vícerozměrným škálováním posuďte podobnost 10 výrobků Coly, testované dvěma nezávislými pracovišti A a B, a to vždy na základě výsledků ankety: 50 respondentů posoudilo a vzájemně porovnalo 10 výrobků Coly (objekty) způsobem "každý s každým" a při dokonalé podobnosti, nerozlišitelnosti byla přidělena nulová vzdálenost mezi dvěma objekty, zatímco při naprosté nepodobnosti vzdálenost 100. Z 50 hodnot párových vzdáleností byla vypočtena střední hodnota a zapsána do buňky vytvořené symetrické čtvercové matice. Z této matice se ve vstupních datech užije pouze horní trojúhelníková část, tj. prvky nad diagonálou nul. Aplikujte metodu klasického metrického škálování CMDS a porovnejte se závěry nemetrického škálování NNMDS, str. 95, ref.[30].

Data: prvky trojúhelníkové matice vyjadřují párové *vzdálenosti* (nepodobnosti, dissimilarities) dvojice výrobků Coly.

Značení: (a) Pracoviště A: *x1* značí sloupec dat *E416Ax1*, ...atd.,

Objekt	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*
x1	0									
x2	16	0								
..	..	..	..	..	..	..	..	..	..	
x10	16	92	90	83	79	44	24	18	98	0

(b) Pracoviště B: *x1* značí sloupec dat *E416Bx1*, ...atd.

Objekt	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*
x1	0									
x2	20	0								
..	..	..	..	..	..	..	..	..	..	
x10	12	90	96	89	75	40	27	14	90	0

Úloha E4.17 *Vícerozměrné škálování u podobnosti hryzců ze zemí Evropy* (MDS)
Vícerozměrným škálováním posuďte podobnost hryzců ze 14 různých hrabství Velké Británie a ostatních zemí Evropy způsobem porovnání jejich lebek, a to "každé s každou". Je dána tabulka vzájemných vzdáleností, a to při dokonalé podobnosti dvou lebek je přidělena vzdálenost 0, zatímco při naprosté nepodobnosti vzdálenost 1. Párové vzdálenosti jsou zapsány do cely symetrické čtvercové matice. Z této matice se ve vstupních datech užije pouze horní trojúhelníková část, tj. prvky nad diagonálou nul. Je třeba provést *dvojrozměrné škálování* a z výsledného grafu usoudit na podobné a nepodobné lebky hryzce. Aplikujte metodu klasického metrického škálování CMDS a porovnejte i se závěry nemetrického škálování NNMDS. K rozlišení užijte kritérium těsnosti proložení, vystižené koeficientem *stress,* str. 95, ref.[30].

Data: prvky trojúhelníkové matice vyjadřují *vzdálenosti* (nepodobnosti, dissimilarities) lebek hryzců z různých zemí Evropy: *x1* značí *E417x1* a znamená *Surrey, E417x2* Shropshire, *E417x3* Yorkshire, *E417x4* Perthshire, *E417x5* Aberdeen, E417x6 Eilean Gamhna, *E417x7* Alpy, *E417x8* Jugoslavie, *E417x9* Německo, *E417x10* Norsko, *E417x11* Pyreneje I, *E417x12* Pyreneje II, *E417x13* severní Španělsko, *E417x14* jižní Španělsko.

Obj	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*	*x14*
x1	0.000													
x2	0.099	0.000												
..	..	..	..	..	..	..	..	..	..	..	..	..	..	
x14	0.586	0.435	0.550	0.530	0.552	0.509	0.369	0.471	0.234	0.346	0.456	0.090	0.038	0.000

Úloha E4.18 *Fisherova úloha rozměrů okvětních lístků u 150 kosatců* (CLU)
Analyzujte předložený výběr kosatců, obsahujících čtvero popisných rozměrů okvětních lístků (čili diskriminátorů) u 150 květů kosatců (čili objektů), pocházejících ze tří základních tříd: (1) Iris setosa, (2) Iris versicolor, (3) Iris virginica. Z botaniky je známo, že druh Iris versicolor je hybridem zbývajících dvou druhů. Iris setosa je diploidní květ s 38 chromozomy, Iris virginica je tetraploidní a Iris versicolor je hexaploidní s 108 chromozomy. Květy kosatců jsou popsány čtyřmi diskriminátory: délkou kališních lístků v mm anglicky *lsepal,* šířkou *wsepal,* dále délkou korunních plátků v mm *lpetal* a šířkou *wpetal*. Budeme proto formulovat úlohu: jsou dána data o K třídách, např. $K = 3$, tři druhy čili třídy kosatců: *Setosa, Versicolor* a *Virginica* s N_k, $k = 1, ..., K,$ objekty v každé třídě, např. pro *Setosu* $k = 1$ $N_1 = 50$, pro *Versicolor* $k = 2$ $N_2 = 50$ a pro *Virginica* $k = 3$ $N_3 = 50$, N představuje celkový počet objektů, např. $N = N_1 + N_2 + N_3 = 150$. Každý objekt je popsán p diskriminátory, např. $p = 4$, a to *Sepal Length, Sepal Width, Petal Length, Petal*

Width. Takže každý *i*-tý objekt je prezentován prvkem x_{ki}. Nechť $\bar{\boldsymbol{x}}$ představuje vektor průměrů diskriminátorů ve všech třídách dohromady a $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$ je vektor průměrů objektů v *k*-té třídě. Cílem diskriminační analýzy je vyšetřit a ověřit botanické třídění a odpovědět na otázku, zda botanické třídění kosatců Iris do tří tříd je správné. Nelze zařadit 150 kosatců do jiného počtu tříd?

Data: použijeme data z úlohy **S2.18:** rozměry pro druhy lístků v mm, a to *lsepal (S218x1), wsepal (S218x2), lpetal (S218x3), wpetal (S218x4)* a druh kosatce *Iris*:

50	33	14	2	1,	64	28	56	22	3,	65	28	46	15	2,	67	31	56	24	3,
..	..	..	..	..,	..	..	..	..	..,	..	..	..	..	..,	..	..	..	..	..,
63	33	60	25	3,	53	37	15	2	1.										

4.9.4 Analýza hutnických a mineralogických dat

Úloha H4.01 *Popis uranového koncentrátu metodou hlavních komponent* (EDA, PCA, CLU)

Uranový koncentrát určený pro odbyt je analyzován na obsah uranu a ostatních příměsí, a to sodíku, draslíku, železa, síranů, uhličitanů a oxidu křemičitého. Analýzou hlavních komponent, resp. faktorovou analýzou se pokuste popsat matici vstupních dat co nejmenším počtem hlavních komponent či faktorů. Proveďte kroky *Postupu analýzy vícerozměrných dat* a rozhodněte, které proměnné korelují. Existují vzorky uranu (objekty), které vykazují podobné tvary sluníček a podobné tvary hvězdiček? Které proměnné jsou nejdůležitější dle grafu komponentních vah? Dají se v rozptylovém diagramu komponentního skóre odhalit odlehlé objekty a nalézt shluky podobných objektů? Kolik se dá najít shluků podobných objektů? Existují také podobné proměnné v dendrogramu podobnosti proměnných?

Data: *H401i* index vzorku, *H401x1* obsah kationtů a aniontů [%] v uranovém koncentrátu, tj. obsah uranu [%], *H401x2* obsah uhličitanů [%], *H401x3* obsah oxidu křemičitého [%], *H401x4* obsah síranů [%], *H401x5* obsah železa [%], *H401x6* obsah draslíku [%], *H401x7* obsah sodíku [%].

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	72.42	0.67	0.81	0.46	0.150	1.90	2.97
..	..	..	..	..	..	..	..
30	73.23	0.43	1.15	0.87	0.580	2.40	1.82

Úloha H4.02 *Klasifikace ropy dle podílu vanadu, železa a uhlovodíků* (EDA, PCA, CLU). Na 45 vzorcích ropy, pocházejících ze tří ložisek, byly sledovány hodnoty obsahu vanadu, železa, nasycených uhlovodíků a aromatických uhlovodíků. Aplikujte metodu hlavních komponent za účelem snížení počtu proměnných. Klasifikujte vzorky ropy do shluků a proveďte kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Data: *H402i* index vzorku ropy, proměnné: *H402x1* obsah vanadu a *H402x2* obsah železa vyjadřují obsah v popelu [promile], *H402x3* obsah nasycených uhlovodíků a *H402x4* obsah aromatických uhlovodíků vyjadřují obsah v desetinách promile, *H402x5* je číslo ložiska.

i	x_1	x_2	x_3	x_4	x_5
1	39	51	706	1219	1

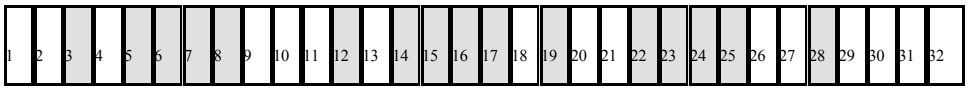
..	..	..	..	..	..
45	62	27	397	297	3

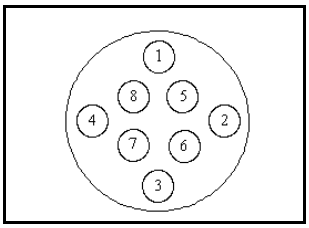
Úloha H4.03 *Celkové zhodnocení mezilaboratorního porovnávacího testu* (EDA, PCA, CLU). U 21 laboratoří (objekty) byla posuzována způsobilost k provádění kvantitativní analýzy vybraných prvků (proměnné) v CrNi ocelích metodou optické emisní spektroskopie s jiskrovým buzením. Na základě obdržených výsledků byly spočteny *Z*-skóre, jejichž absolutní hodnoty představují data pro vícerozměrné testování. Cílem hodnocení dat byla snaha o nalezení takových proměnných, které by nejlépe charakterizovaly porovnávané laboratoře v oblasti analyzování CrNi ocelí s ohledem k jejich použití. Proveďte kroky *Postupu analýzy vícerozměrných dat* a vyšetřete, kolik latentních proměnných popisuje alespoň 66%ní proměnlivost v datech? Z grafu komponentních vah vyšetřete korelaci proměnných, důležité proměnné, ale také redundandní proměnné. V rozptylovém diagramu komponentního skóre vyšetřete odlehlé objekty a označte shluky podobných objektů. Pokuste se odstranit nadbytečné proměnné a odlehlé objekty a aplikujte opětovně metodu hlavních komponent. Komentujte počet shluků v rozptylovém diagramu komponentního skóre a pokuste se nalézt interakci objektů a proměnných ve dvojném grafu. Pokuste se pojmenovat hlavní komponenty. Kolik shluků lze nalézt v dendrogramu podobnosti objektů nejlepší shlukovací procedurou? Komentujte vzniklé shluky dle jejich vlastností.

Data: *H403i* index vzorku, proměnné: *H403x1* obsah C, *H403x2* obsah Cu, *H403x3* obsah P, *H403x4* obsah S, *H403x5* obsah Si, *H403x6* obsah Ni, *H403x7* obsah Mo, *H403x8* obsah Cr, *H403x9* obsah Mn.

i	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*
1	0.50	0.39	1.09	0.18	0	0.36	0	0.02	0.06
..	...	...	...	...	...	...	...	...	...
21	1.13	0.27	1.36	0.18	0.31	0.93	0.56	0.79	0.60

Úloha H4.04 *Posouzení chemické homogenity v kruhové tyči CrNi oceli* (EDA, PCA, CLU). Kruhová tyč CrNi oceli o rozměrech 50 × 1000 mm, vyrobená v Poldi Kladno, byla rozřezána na 32 zkušebních vzorků o rozměrech 50 × 30 mm. Každý vzorek byl označen pořadovým číslem za současného zachování původní orientace v tyči. K testování homogenity bylo vybráno náhodně 16 vzorků. Metodou optické emisní spektroskopie s jiskrovým buzením byla na každém vzorku v předem určených místech provedena analýza. Na každém vzorku bylo provedeno 8 expozic. Cílem bylo vybrat pro vyhodnocení homogenity materiálu omezený počet proměnných a odhalit trendy v chemickém složení. Prvním číslem v matici dat je pořadí vzorku a číslicí po mezeře je pak umístění expozice. Proveďte kroky *Postupu analýzy vícerozměrných dat* a vyšetřete, kolik latentních proměnných popisuje alespoň 66% proměnlivost v datech? Z grafu komponentních vah vyšetřete korelaci proměnných, důležité proměnné, ale také redundandní proměnné. V rozptylovém diagramu komponentního skóre vyšetřete odlehlé objekty a označte shluky podobných objektů. Pokuste se odstranit nadbytečné proměnné a odlehlé objekty a aplikujte opětovně metodu hlavních komponent. Komentujte počet shluků v rozptylovém diagramu komponentního skóre a pokuste se nalézt interakci objektů a proměnných ve dvojném grafu. Pokuste se pojmenovat hlavní komponenty. Kolik shluků lze nalézt v dendrogramu podobnosti objektů nejlepší shlukovací procedurou? Komentujte vzniklé shluky dle jejich vlastností.





Data: Matice vstupních dat: *H404i* je index objektu, *H404x1* obsah C, *H404x2* obsah Mn, *H404x3* obsah Si, *H404x4* obsah P, *H404x5* obsah S, *H404x6* obsah Cu, *H404x7* obsah Cr, *H404x8* obsah Ni, *H404x9* obsah Al, *H404x10* obsah Mo, *H404x11* obsah Tl, *H404x12* obsah B.

i	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*
3 1	0.051	0.1531	1.4636	27.181	28.506	0.098	1.7085	0.728	1.9752	1.8461	2.5478	0.062
...	...	...	...	...	...	...	...	...	...	...	...	...
28 8	0.05	0.1531	1.4548	27.746	29.465	0.098	1.7127	0.7283	1.9348	1.8348	2.4939	0.062

Úloha H4.05 *Dispergace mletého kalcinátu titanové běloby rutilového typu* (EDA, PCA, CLU)

Materiál pigmentových vlastností titanové běloby se získává kalcinací hydratovaného gelu oxidu titaničitého při teplotách kolem 950°C. Výpad z kalcinační pece, kalcinát, se po zchlazení nejprve mele za sucha na běžných mlýnech pro sypké materiály. Byly stanoveny rheologické vlastnosti, konkrétně viskozita a doba výtoku Fordovým kelímkem. Smyslem bylo posoudit vzájemné vazby mezi parametry a najít charakteristiky, které umožní včasný odhad rheologických vlastností. Data je možné rozdělit do dvou skupin: (1) *První část dat*, z nichž většina se týká obsahu rutilové krystalické fáze a jednotlivých příměsí ze suroviny nebo přísad, byla získána pro provozně pomletý kalcinát. Jsou zde i data ze speciálních zkoušek, např. měrná vodivost a pH vodného výluhu z kalcinátu nebo spotřeba vody, bod tečení a bod smočení, které popisují dosažení určitého chování kalcinátu při přidání minimálního množství vody. (2) *Druhá část dat* je tvořena dobou výtoku Fordovým kelímkem a viskozitou a týká se rheologických vlastností suspenze připravené standardním postupem dispergace mletého kalcinátu.

Proveďte kroky *Postupu analýzy vícerozměrných dat* a vyšetřete, kolik latentních proměnných popisuje alespoň 66% proměnlivost v datech? Z grafu komponentních vah vyšetřete korelaci proměnných, důležité proměnné ale také redundandní proměnné. V rozptylovém diagramu komponentního skóre vyšetřete odlehlé objekty a označte shluky podobných objektů. Pokuste se odstranit nadbytečné proměnné a odlehlé objekty a aplikujte opětovně metodu hlavních komponent. Komentujte počet shluků v rozptylovém diagramu komponentního skóre a pokuste se nalézt interakci objektů a proměnných ve dvojném grafu. Pokuste se pojmenovat hlavní komponenty a komentujte vyluhovatelnost pigmentu v y_1 a hydrofilitu povrchu v y_2. Kolik shluků lze nalézt v dendrogramu podobnosti objektů nejlepší shlukovací procedurou? Komentujte vzniklé shluky dle jejich fyzikálně-chemických vlastností.

Data: *H405i* index objektu, *H405x1* značí Ford [s], *H405x2* značí viskozita [mPas], *H405x3* značí pH, *H405x4* značí vodivost [m S cm], H405x5 značí rutil [%], *H405x6* značí Fe [%], *H405x7* značí Sb_2O_3 [%], *H405x8* značí K_2O [%], *H405x9* značí obsah síry [%], *H405x10* značí P_2O_5 [%], *H405x11* značí SiO_2 [%], *H405x12* značí Nb_2O_5 [%], *H405x13* značí ZrO_2 [%], *H405x14* značí TiO_2 [%], *H405x15* značí Na_2O [%], *H405x16* značí specifická hmotnost vody [g/100 g], *H405x17* značí bod smočení [g/100 g], *H405x18* značí Sb_2O_3 [g/100 g].

i	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*	*x14*	*x15*	*x16*	*x17*	*x18*
1	13.1	573	8.06	526	99.7	0	0.004	0.206	0.049	0.172	0	0.056	0.035	98.74	0.021	21.54	20.53	88.17
..	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
28	13.3	883	7.91	585	99.3	0.0018	0.002	0.194	0.059	0.166	0.007	0.052	0.031	98.83	0.021	23.86	23.74	81.64

4.9.5 Analýza ekonomických a sociologických dat

Úloha S4.01 *Pevnost stavebního řeziva - dřevěných prken* (EDA, CORA, PCA)
U 30 náhodně vybraných vzorků stavebního řeziva - dřevěných prken[20] byla měřena pevnost v průhybu čtyřmi způsoby: na základě vyslané rázové vlny, při vibračním namáhání a pomocí dvou statických metod. Účelem je stanovit odhady parametrů polohy a rozptýlení klasickými i robustními metodami a testovat hypotézu, že každá metoda poskytuje nezávislé informace (korelační matice je jednotková, tj. H_0: $\boldsymbol{R} = \boldsymbol{E}$). Proveďte kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).
Data: *S401i* index vzorku řeziva, pevnost dřevěných desek (řádky) stanovená různými metodami (sloupce): *S401x1* pevnost v průhybu po rázové vlně, *S401x2* pevnost při vibračním namáhání, *S401x3* pevnost 1. statickou metodou, *S401x4* pevnost 2. statickou metodou. V tabulce dat lze vyčíslit i čtverec Mahalanobisovy vzdálenosti d^2: čím větší hodnota d^2, tím je bod vzdálenější od střední hodnoty (resp. těžiště), a tím je pevnost prkna odlišnější od ostatních.

i	x_1	x_2	x_3	x_4
1	1889	1651	1561	1778
..	..	..	..	..
30	1490	1187	1714	1284

Úloha S4.02 *Odolnost skel* (EDA, CORA, PCA, CLU)
U patnácti zkumavek byla měřena tvrdost skla pomocí Brinellovy metody x_1 a mez odporu proti lomu x_2. Stanovte odhad vektoru středních hodnot a zkonstruujte odpovídající 95 % oblast spolehlivosti. Proveďte kroky *Postupu analýzy vícerozměrných dat* a vyšetřete, kolik shluků podobných zkumavek je možné nalézt.
Data: *i* index vzorku skla, *S402x1* značí tvrdost skla podle Brinella a *S402x2* značí mez odporu proti lomu, cit.[21].

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	331	335	331	331	341	331	331	339	333	331	331	331	0	341	327
x_2	109	111.5	109.5	109.5	114	113	110.5	107.5	114.5	112	110	115	110	110.5	109.5

Úloha S4.03 *Pracnost žehlení obleků* (EDA, PCA)
U 76 pracovníků byly stanoveny časy potřebné k vykonání šesti operací, z nichž se skládá žehlení obleků. Účelem je zjistit, zda existuje závislost mezi jednotlivými operacemi (testování hypotézy H_0: $\boldsymbol{R} = \boldsymbol{E}$). Proveďte kroky *Postupu analýzy vícerozměrných dat*.
Data: Korelační matice $\boldsymbol{R}$ obsahuje v pořadí *S403x1* první operaci, *S403x2* druhou operaci, ..., *S403x6* šestou operaci.

$$\boldsymbol{R} \;=\; \begin{bmatrix} 1.000 & 0.088 & 0.334 & 0.191 & 0.173 & 0.123 \\ 0.088 & 1.000 & 0.186 & 0.384 & 0.262 & 0.040 \\ 0.334 & 0.186 & 1.000 & 0.343 & 0.144 & 0.080 \\ 0.191 & 0.384 & 0.343 & 1.000 & 0.375 & 0.142 \\ 0.173 & 0.262 & 0.144 & 0.375 & 1.000 & 0.334 \\ 0.123 & 0.040 & 0.80 & 0.142 & 0.334 & 1.000 \end{bmatrix}.$$

Úloha S4.04 *Shluková analýza původu ropy* (DA, CLU)
Na patnácti vzorcích ropy, pocházejících ze dvou ložisek (typ ložiska je kódován v proměnné u_i = 1, 2), byly stanoveny: obsah vanadu, obsah železa, obsah nasycených uhlovodíků, obsah aromatických uhlovodíků. Je podezření, že ložiska $u_1 = 1$ a $u_2 = 2$ poskytují ropu ze stejného geologického zdroje. Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: Obsahy různých látek v ropě získané ze dvou ložisek: *S404i* index ložiska, *S404j* druh ložiska, *S404x1* obsah vanadu, *S404x2* obsah železa, *S404x3* obsah nasycených uhlovodíků, *S404x4* obsah aromatických uhlovodíků.

i	*j*	x_1	x_2	x_3	x_4
1	1	39	51	706	1219
..	..	..	..	..	..
15	2	44	46	754	576

Úloha S4.05 *Sociologický průzkum názorů sedmi etnik* (EDA, PCA, CLU)
V rámci sociologického průzkumu bylo zástupcům několika etnik nastíněno několik životních situací a možných řešení. Tabulka obsahuje relativní četnosti souhlasných reakcí daných etnik v procentech: (a) pokuste se vyjádřit míru vzájemné podobnosti názorů jednotlivých etnik, (b) posuďte, zda pro uvedený účel je nutný daný počet otázek. Proveďte také kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Data: *S405i* index zástupce etnika, *S405x1, S405x2, S405x3, S405x4, S405x5, S405x6, S405x7* souhlasné reakce na 10 otázek (řádky) zástupcům 7 etnik (sloupce).

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	53	26	25	39	53	22	9
..	..	..	..	..	..	..	..
10	86	87	85	80	81	67	44

Úloha S4.06 *Rozbor klasifikace žáků osmých tříd* (EDA, PCA, CLU)
Na souboru dat známek žáků osmých tříd použijte metodu hlavních komponent nebo faktorovou analýzu, shlukovou analýzu a alespoň dvě grafické zobrazovací metody tak, že zredukujete nadměrný počet proměnných a matici popíšete co nejmenším počtem proměnných či faktorů. Které předměty spolu významně korelují a lze je eventuálně i vypustit? Proveďte kroky *Postupu analýzy vícerozměrných dat* a vyšetřete ve hvězdičkovém a sluníčkovém grafu, kolik žáků dosáhlo podobných výsledků. Které předměty korelují se shluky žáků v analýze dvojného grafu. Kolik shluků podobných žáků lze určit v dendrogramu objektů? Kterou shlukovací metodou sestrojíte dendrogram objektů?

Data: *S406i* index žáka, *S406x1* český jazyk, *S406x2* anglický nebo německý jazyk, *S406x3* dějepis, *S406x4* občanská výchova, *S406x5* zeměpis, *S406x6* matematika, *S406x7* přírodopis, *S406x8* fyzika, *S406x9* chemie, *S406x10* hudební výchova a zpěv, *S406x11* výtvarná výchova, *S406x12* tělesná výchova, *S406x13* rodinná výchova, *S406x14* volitelný předmět,

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
1	3	3	3	1	3	3	3	3	3	1	1	1	1	1
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
44	2	2	3	2	2	2	3	2	3	1	1	1	1	1

Úloha S4.07 *Klasifikace aut* (EDA, PCA, FA, CLU)
Databáze 155 aut byla klasifikována dle 11 parametrů dominantních vlastností. Aplikujte metodu hlavních komponent a faktorovou analýzu a pokuste se snížit počet proměnných. Vytvořte shluky podobných aut. Proveďte také kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Data: *S407i* index auta, *S407x1* spotřeba benzinu v počtu ujetých mil na 1 gallon (mpg), *S407x2* počet válců (cyl), *S407x3* vrtání (displ), *S407x4* výkon v koňských silách (horse), *S407x5* zrychlení (accel), *S407x6* poslední dvojčíslí roku výroby (year), *S407x7* hmotnost vozu (weight), *S407x8* země původu (origin: 1 USA, 2 Evropa, 3 Japonsko), *S407x9* výrobce (make), *S407x10* model, *S407x11* cena vozu v US$ v roce 1978 (price).

i	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*
1	43.1	4	90	48	21.5	78	1985	2	Volkswagen	Rabbit Dl	2400
..	..	..	..	..	..	..	..	..	..	..	..
155	31.0	4	119	82	19.4	82	2720	1	Chevrolet	S-10	4500

Úloha S4.08 *Sledování finačních nákladů nákladních vozů dle spotřeby benzinu a nafty* (EDA, PCA, CLU). Vyšetřete náklady na transport mléka při použití nákladního auta na benzin nebo na naftu, str. 365 v ref.[20]. Vyšetřete struktury sledovaných proměnných a strukturu sledovaných objektů. Proveďte také kroky *Postupu analýzy vícerozměrných dat*. Ukazuje EDA (např. krabicový graf), že je třeba proměnné standardizovat? Je třeba provést exploratorní (průzkumovou) analýzu dat, která (a) posoudí *podobnost objektů* pomocí rozptylových a symbolových grafů, (b) nalezne *vybočující objekty*, resp. jejich proměnné, (c) stanoví, zda lze použít předpoklad lineárních vazeb, a (d) ověří *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Data: Sledované proměnné jsou znormovány a vztaženy na 1 míli: *S408i* index auta, *S408x1* spotřeba pohonných hmot, *S408x2* náklady na opravy, *S408x3* pořizovací hodnota, *S408x4* druh pohonné hmoty.

S408i	*S408*x1	*S408*x2	*S408*x3	*S408x4*
1	16.44	12.43	11.23	gasoline
...	...	...	...	
59	12.03	9.22	23.09	diesel

Úloha S4.09 *Struktura a vazby burzovních akcií chemických firem* (EDA, PCA, CLU) Po 100 týdnů byl sledován pohyb akcií pěti velkých chemických firem na newyorské burze, str. 507 v ref.[20]. Proveďte také kroky *Postupu analýzy vícerozměrných dat* a hledejte strukturu a vnitřní vazby jak mezi proměnnými tak i mezi objekty (týdny).

Data: *S409i* index týdne, *S409x1* akcie firmy Allied Chemical, *S409x2* akcie firmy Du Pont, *S409x3* akcie firmy Union Carbide, *S409x4* akcie firmy Exxon, *S409x5* akcie firmy Texaco.

i	x_1	x_2	x_3	x_4	x_5
1	0	0	0	0.039473	0
...	...	...	...	...	...
100	0.019108	-0.033303	0.008362	0.033898	0.004566

Úloha S4.10 *Vícerozměrné škálování u příbuznosti míčových sportů* (MDS)
Vícerozměrným škálováním posuďte podobnost, příbuznost a vzájemný vztah 6 rozličných míčových a míčkových sportů, je-li dána tabulka vzájemných vzdáleností způsobem porovnání "každého s každým": při dokonalé podobnosti dvou sportů je přidělena vzdálenost 0, zatímco při naprosté nepodobnosti vzdálenost 10. Párové vzdálenosti jsou zapsány do cely symetrické čtvercové matice. Z této matice se ve vstupních užije pouze horní trojúhelníková část, tj. prvky nad diagonálou nul. Aplikujte metodu klasického metrického škálování CMDS a porovnejte i se závěry nemetrického škálování NNMDS.

Data: prvky trojúhelníkové matice vyjadřují vzdálenosti (nepodobnosti, dissimilarities) objektů dvojice sportů: *S410x1* hokej, *S410x2* fotbal, *S410x3* košíková, *S410x4* tenis, *S410x5* golf, *S410x6* kroket.

Obj	*S410x1*	*S410x2*	*S410x3*	*S410x4*	*S410x5*	*S410x6*
S410x1	0					
S410x2	2	0				
...	...	...	...	...	...	...
S410x6	5	5	6	3	2	0

Úloha S4.11 *Vícerozměrné škálování u podobnosti aktivit k osobní relaxaci* (MDS)
Vícerozměrným škálováním posuďte podobnost a vzájemný vztah 15 rozličných aktivit osobní relaxace, které tvoří 105 párů vzájemného porovnání "každé aktivity s každou". Je dána tabulka vzájemných vzdáleností, a to při dokonalé podobnosti dvou aktivit je přidělena vzdálenost 0, zatímco při naprosté nepodobnosti vzdálenost 25. Párové vzdálenosti jsou zapsány do symetrické čtvercové matice. Ve vstupních datech se užije pouze horní trojúhelníková část, tj. prvky nad diagonálou nul. Aplikujte metodu klasického metrického škálování CMDS a porovnejte i se závěry nemetrického škálování NNMDS. Jako rozhodčí kritérium těsnosti proložení užijte koeficient *stress*.

Data: prvky trojúhelníkové matice vyjadřují *vzdálenosti* (nepodobnosti, dissimilarities) objektů dvojice aktivit relaxace: *x1* značí *S411x1* a znamená koncert, *S411x2* muzeum, *S411x3* divadlo, *S411x4* kino, *S411x5* TV, *S411x6*

konference, *S411x7* četba, *S411x8* divák hokeje, *S411x9* balet, *S411x10* politika, *S411x11* móda, S411x12 dokumentaristika, *S411x13* výstavy, S411x14 nákupy, *S411x15* restaurace.

Obj	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*	*x14*	*x15*
x1	-														
x2	16	-													
x3	3	18	-												
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
x15	8	8	7	9	21	21	2	22	5	25	9	23	10	8	-

Úloha S4.12 *Skryté vazby mezi traťovými rekordy v lehké atletice žen* (PCA, CLU)
Byly zaznamenány národní traťové rekordy v lehké atletice žen, str. 44 v ref.[20]. Je třeba odhalit strukturu, skryté vazby či závislosti mezi jednotlivými běžeckými disciplinami. Všimněme si především párových korelačních koeficientů jednotlivých běžeckých disciplín. Proveďte kroky *Postupu analýzy vícerozměrných dat*.

Data: *S412i* index sportovce, *S412x1* běh na 100 m [s], *S412x2* běh na 200 m [s], *S412x3* běh na 400 m [s], *S412x4* běh na 800 m [min], *S412x5* běh na 1500 m [min], *S412x6* běh na 3000 m [min], *S412x7* maraton [min], *S412x8* země původu sportovce.

S412i	*S412x1*	*S412x2*	*S412x3*	*S412x4*	*S412x5*	*S412x6*	*S412x7*	*S412x8*
1	11.61	22.94	54.5	2.15	4.43	9.79	178.52	argentin
...	...	...	...	...	...			
55	12.74	25.85	58.73	2.33	5.81	13.04	306	wsamoa

Úloha S4.13 *Skryté vazby mezi traťovými rekordy v lehké atletice mužů* (PCA, CLU)
Byly zaznamenány národní traťové rekordy v lehké atletice mužů, str. 510 v ref.[20]. Je třeba odhalit strukturu, skryté vazby či závislosti mezi jednotlivými běžeckými disciplínami. Všimněme si především párových korelačních koeficientů jednotlivých běžeckých disciplín. Proveďte také kroky *Postupu analýzy vícerozměrných dat*.

Data: *S413i* index sportovce, *S413x1* běh na 100 m [s], *S413x2* běh na 200 m [s], *S413x3* běh na 400 m [s], *S413x4* běh na 800 m [min], *S413x5* běh na 1500 m [min], *S413x6* běh na 5000 m [min], *S413x7* běh na 10000 m [min], *S413x8* maraton [min], *S412x9* země původu sportovce.

i	*S413x1*	*S413x2*	*S413x3*	*S413x4*	*S413x5*	*S413x6*	*S413x7*	*S413x8*	*S413x9*
1	10.39	20.81	46.84	1.81	3.7	14.04	29.36	137.72	argentin
...	...	...	...	...	...				
55	10.82	21.86	49	2.02	4.24	16.28	34.71	161.83	wsamoa

Úloha S4.14 *Struktura a skryté vazby ve vlastnostech planetárních těles sluneční soustavy*
Sluneční soustava je mimo centrální hvězdy tvořena planetami a jejich měsíci, planetkami a kometami. Údaje o kometách jsou méně dostupné, přesněji vzato jsou v potřebném plném rozsahu k dispozici jen pro ojedinělá tělesa, jejichž hmotnost a často i tvar a rozměry se výrazně snižují každým přiblížením ke Slunci, kdy se velká část materiálu vypaří. Analyzovaná data byla získána z webovských stránek NASA: rozměr 1 značí průměr na rovníku planetárního objektu, který u rychle rotujících planet bývá nejvyšší vlivem snížení vlivu gravitace odstředivou silou. Rozměr 2 značí průměr tělesa na pólech. Rozměr 3 se týká průměru tělesa na pólech u planetek, které jsou vesměs tvořeny nepravidelnými kusy hmoty. Pojmem délka dne se rozumí doba jedné otáčky tělesa kolem své osy rotace, délka roku je doba oběhu tělesa kolem Slunce. Proveďte kroky *Postupu analýzy vícerozměrných dat* a soustřeďte se na otázky: které proměnné korelují v korelačním diagramu a v ostatních diagnostikách exploratorní analýzy dat? V grafu komponentních vah

vyšetřete, které proměnné korelují, které jsou důležité a které jsou redundandní: Vyskytují se v rozptylovém diagramu komponentního skóre nějaké odlehlé objekty? Které objekty jsou podobné a nacházejí se ve společném shluku? Dospěli jste k závěru v exploratorní analýze dat, že data je třeba standardizovat? Analýzou dvojného grafu se pokuste hlavní komponenty vysvětlit dle přiděleného fyzikálního smyslu: y_1 hmotnost a gravitace planety, y_2 poloha planety a její rotace. Do kolika shluků se podařilo planety roztřídit?

Data: *S414i* značí index planetového objektu, *j* název planetového objektu, *S414x1* rozměr 1 [km], *S414x2* rozměr 2 [km], *S414x3* rozměr 3 [km], *S414x4* hmotnost [teratuny], *S414x5* délka dne [hodiny], *S414x6* délka roku [dny], *S414x7* počet měsíců.

i	j	x1	x2	x3	x4	x5	x6	x7
1	Merkur	4878	4878	4878	330000000	1407.6	87.97	0
..	...	...	...	...	...	...	...	...
24	Toutatis	4.6	2.4	1.9	0.05	130	1453.7	0

Úloha S4.15 *Struktura ve známkách na vysvědčení žáků* (FA)

Pro šest předmětů na vysvědčení u 200 žáků byla vyčíslena korelační matice ***R***. Provedte faktorovou analýzu s cílem nalézt faktorově čisté předměty po rotaci faktorů. Označte jeden faktor za verbální myšlení a druhý faktor za logické myšlení, str. 279, ref.[30]. Při analýze postupujte podle kroků *Postupu analýzy vícerozměrných dat*.

Data: *S415i* index, název předmětu, *S415x1* francouzština, *S415x2* angličtina, *S415x3* dějepis, *S415x4* aritmetika, *S415x5* algebra, *S415x6* geometrie.

S415i	*S415x1*	*S415x2*	*S415x3*	*S415x4*	*S415x5*	*S415x6*
S415x1	1					
...	...	...	...	...	...	
S415x6	0.25	0.33	0.18	0.47	0.46	1

Úloha S4.16 *Shluky 12 superhvězd košíkové* (CLU)

Následující tabulka dat obsahuje informace o osmi hráčských vlastnostech a aktivitách 12 superhvězd košíkové v sezóně 1989. Cílem je najít shluky hráčů podobných vlastností a naopak odhalit jejich aktivity a vlastnost, ve které se hráč neshoduje s ostatními hráči.

Data: databáze hráčů košíkové obsahuje tyto proměnné: *S416i* značí index hráče, *S416j* značí jméno hráče, *S416x1* značí Height, výška hráče [palce], *S416x2* značí Weight, hmotnost [libry], *S416x3* značí FgPct, *S416x4* značí FtPct, *S416x5* značí Points, počet dosažených bodů, *S416x6* značí Rebounds, počet doskoků, *S416x7* značí počet Assist, počet asistencí, *S416x8* značí počet Fauls, počet faulů.

i	j	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*
1.	Jabbar K. A.	86	230	55.9	72.1	24.6	11.2	3.6	3
..	...	..	..	..	..	..	..	..	..
12.	West J.	74.5	180	47.4	81.4	27	5.8	6.7	2.6

Úloha S4.17 *Shluky jednotlivých barev na paletě (CLU)*
Je dáno 22 objektů barev, které vznikly podílem červené a modré. Je třeba provést klasifikaci objektů barev do shluků.
Data: *S417i* je index barvy, *S417x1* je podíl červené (Red), *S417x2* je podíl modré (Blue).

S417*i*	*S417x1*	*S417x2*
1	1	9
..	..	..
22	8	1

4.10 Doporučená literatura

[1] Siotani M., Hayakawa T., Fujikoshi Y.: Modern Multivariate Statistical Analysis, A Graduate Course and Handbook. American Science Press, Columbia 1985.

[2] Kendall M. G., Stuart A.: The Advanced Theory of Statistics, Vol. III. New York 1966.

[3] James W., Stein C.: Estimation with Quadratic Loss, Proceed. 4th Berkeley Symp. on Math. Statist., p. 361, 1961.

[4] Guanadeskian R., Kettenring J. R.: Biometrics **28**, 80 (1972).

[5] Campbell N. A.: Appl. Statist., 29, 231 (1980).

[6] Hu J., Skrabal P., Zollinger H.: Dyes and Pigments, **8**, 189 (1987).

[7] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: Graphical Methods for Data Analysis. Duxburg Press, Belmont, California 1983.

[8] Barnett V. (ed.): Interpreting Multivariate Data. Wiley, Chichester 1981, kap. 6.

[9] Jolliffe I. T.: Principal Component Analysis. Springer Verlag, New York 1986.

[10] Barnett V.,B. S.: Graphical Techniques for Multivariate Data. London 1978.

[12] Andrews D. F.: Biometrics, **28**, 125 (1972).

[13] Kulkarni S. R., Paranjape S. R.: Commun. Statist., **13**, 2511 (1984).

[14] Guanadeskian R.: Methods for Statistical Data Analysis of Multivariate Observations. Wiley, New York 1977.

[15] Kleiner B., Hartigan J. A., J. Amer. Statist. Assoc., **76**, 260 (1981).

[16] Kres H.: Statistical Tables for Multivariate Analysis. Springer, New York 1983.

[17] Seber G. A. F.: Multivariate Observations. Wiley, New York 1984.

[18] Stryjewska E., Rubel S., Henrion A., Henrion G.: Z. Anal. Chem., **327**, 679 (1987).

[19] Mudholkar G. S., Trivedi M. S., Lin T. C.: Technometrics, **24**, 139 (1982).

[20] Johnson R.A., Wichern D.W.: Applied Multivariate Statistical Analysis, Prentice Hall, 1998.

[21] Ajvjazin S., Bežajeva Z., Staroverov O.: Metody vícerozměrné analýzy, SNTL Praha 1981.

[22] Meloun M., Militký J. , Forina M.: Chemometrics for Analytical Chemistry, Volume 1. PC-Aided Statistical Data Analysis, Ellis Horwood, Chichester 1992.

[23] Brereton R. G. Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies, Elsevier 1992,

[24] Krzanowski W. J.: Principles of Multivariate Analysis, A User's Perspective, Oxford Science Publications, 1988.

[25] Jeffers J. N. R.: Applied Statistician, **16**, 225 (1967).

[26] Meloun M. , Militký J.: Statistické zpracování experimentálních dat, Plus Praha 1994.

[27] Martens H., Naes T.: Multivariate calibration, Wiley (1989) Chichester.

[28] Thomas E. V.: Anal. Chem., **66** (1994) 795A-804A.

[29] Malinowski F., Howery D.: Factor Analysis in Chemistry, Wiley (1980) New York.

[30] Everitt B. S., Dunn G.: Applied Multivariate Data Analysis, Arnold, London 2001.

5

ANALÝZA ROZPTYLU

Analýza rozptylu, ANOVA (z anglického **An**alysis **o**f **Va**riance), se v technické praxi používá buď jako samostatná technika, nebo jako postup, umožňující analýzu zdrojů variability u lineárních statistických modelů. Cílem je zjistit, které z kvalitativních nebo kvantitativních faktorů významně ovliňují sledované veličiny. Nejde přitom o to jak ovliňují, ale zda vůbec ovliňují. V technické praxi se ANOVA uplatňuje jako samostatná technika v úlohách:

(a) Určení vlivu způsobu přípravy vzorků na výsledek analýzy.

(b) Určení vlivu přístroje, lidského faktoru a obsluhy na výsledek měření.

(c) Zpracování různých mezilaboratorních experimentů.

(d) Zpracování plánovaných experimentů, u kterých se systematicky sleduje vliv rozličných faktorů (teploty, času, koncentrace a dalších) na výsledek reakce či analýzy.

Podstatou analýzy rozptylu je rozklad celkového rozptylu dat na *složky objasněné* (známé zdroje variability) a *složku neobjasněnou*, o níž se předpokládá, že je náhodná. Následně se testují hypotézy o významnosti jednotlivých zdrojů variability.

Prvním krokem analýzy rozptylu je určit, zda jde o model analýzy rozptylu s pevnými, náhodnými nebo smíšenými efekty. Vlastní postup analýzy rozptylu lze rozdělit do pěti kroků, jimiž jsou:

1. Odhad parametrů základního modelu ANOVA.
2. Testování jeho významnosti a konstrukce různých modelů.
3. Vyjádření složek rozptylů a testování jejich významnosti.
4. Ověření předpokladu normality a indikace silně vybočujících hodnot.
5. Interpretace výsledků s ohledem na zadání dat a jejich případné úpravy.

5.1 Jednofaktorová analýza rozptylu (ANOVA1)

Formulace modelu: sleduje se faktor A na K různých úrovních $A_1, ..., A_K$. Na každé úrovni A_i je provedeno n_i měření $\{y_{ij}\}, j = 1, ..., n_i$. Model analýzy rozptylu má tvar

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad j = 1, ..., n_i, \quad \text{kde} \quad \mu_{ij} = \mu + \alpha_i$$

a α_i je *efekt i-té úrovně*. Parametry μ_i, μ a α_i se odhadují pomocí odpovídajících výběrových

průměrů. *Celkový počet měření* je $N = \sum_{i=1}^{K} n_i$. *Sloupcový průměr* $\hat{\mu}_i$ představuje součet hodnot opakovaného měření y_{ij} pro úroveň faktoru A_i, dělený počtem opakování n_i,

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} .$$

Celkový průměr je součet všech hodnot y_{ij} dělený celkovým počtem dat N, který je roven průměru K sloupcových průměrů

$$\hat{\mu} = \frac{1}{K} \sum_{i=1}^{K} \hat{\mu}_i .$$

Pro výpočet odhadu i-tého efektu α_i lze použít vztah $\alpha_i = \hat{\mu}_i - \hat{\mu}$. Zavedením efektů vznikne přeurčený model, obsahující o jeden parametr více. Proto se při odhadu efektů používá omezující podmínka $\sum_{i=1}^{K} n_i \alpha_i = 0$ a pro vyvážené experimenty $\sum_{i=1}^{K} \alpha_i = 0$. Klasická jednofaktorová analýza rozptylu dat a Kruskalova-Wallisova jednofaktorová analýza rozptylu pořadí dat porovnává střední hodnoty dvou či více úrovní faktoru A čili sloupců v matici dat za účelem určit, zda alespoň jedna sloupcová střední hodnota se liší od ostatních. Statistická významnost je testována F-testem tak, že nulová hypotéza H_0 říká "*Všechny střední hodnoty jsou stejné*" proti alternativní H_A "*Alespoň jedna střední hodnota se odlišuje od ostatních*".

Základní předpoklady jednofaktorové analýzy rozptylu dat. Před použitím analýzy rozptylu musí být ověřeny následující předpoklady o výběru:

1. Data mají normální rozdělení: náhodné chyby ε_{ij} jsou náhodné veličiny s normálním rozdělením a střední hodnotou chyb rovnou nule $N(0, \sigma^2)$.
2. Rozptyly sloupcových výběrů σ^2 jsou stejné (homoskedasticita).
3. Každý sloupec je prostým náhodným výběrem ze svého souboru: každý prvek souboru má stejnou pravděpodobnost, že bude vybrán do výběru.

Základní předpoklady Kruskalova-Wallisova testu analýzy rozptylu dat. Před použitím analýzy rozptylu musí být ověřeny následující předpoklady o výběru:

1. Měrná stupnice je přinejmenším ordinální.
2. Rozdělení souborů musí být stejné, kromě míry polohy. Rozptyly jsou stejné (homoskedasticita).
3. Všechny sloupce představují náhodné výběry svých souborů.

Omezení: velikost výběru může být od několika jednotek až po několik stovek prvků.

Snad největší omezení v datech se týká náhodného výběru ze souboru: když totiž nebude výběr náhodný, hladiny významnosti budou nesprávné.

Testování: součet čtverců odchylek od celkového průměru $\hat{\mu}$, definovaný vztahem

$$S_c = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu})^2 ,$$

lze rozložit na dvě složky

$$S_c = \sum_{i=1}^{K} \sum_{j=1}^{n_i} [(y_{ij} - \hat{\mu}_i) + (\hat{\mu}_i - \hat{\mu})]^2 = S_A + S_R ,$$

kde S_A představuje *rozptyl mezi jednotlivými úrovněmi daného faktoru*

$$S_A = \sum_{i=1}^{K} n_i (\hat{\mu}_i - \hat{\mu})^2$$

a S_R je *rozptyl reziduální* tj. uvnitř jednotlivých úrovní,

$$S_R = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2 .$$

Nevychýleným odhadem rozptylu chyb σ_e^2 je *průměrný reziduální čtverec* MS_R dle

$$MS_R = \frac{S_R}{N - K} = \hat{\sigma}_e^2 .$$

Cílem jednofaktorové analýzy je především testování, zda jsou efekty α_i nulové, tedy zda jednotlivé úrovně daného faktoru jsou statisticky nevýznamně odlišné. Testuje se nulová hypotéza H_0: $\alpha_i = 0$, $i = 1, ..., K$, proti alternativní H_A: $\alpha_i \neq 0$, $i = 1, ..., K$. Při testování se využívá faktu, že veličina S_A / σ_e^2 má χ^2-rozdělení s $(K - 1)$ stupni volnosti a veličina S_R / σ_e^2 má nezávislé χ^2-rozdělení s $(N - K)$ stupni volnosti. Jejich podíl má pak F-rozdělení s $(K - 1)$ a $(N - K)$ stupni volnosti. Testační statistika F_e má tvar

$$F_e = \frac{MS_A}{MS_R} = \frac{S_A (N - K)}{S_R (K - 1)} .$$

Při platnosti nulové hypotézy H_0 má F_e statistika F-rozdělení s $(K - 1)$ a $(N - K)$ stupni volnosti. Vyjde-li F_e větší než kvantil $F_{1-\alpha}(K - 1, N - K)$, je nutné nulovou hypotézu H_0 na hladině významnosti α zamítnout a efekty považovat za nenulové, čili statisticky významné.

Vícenásobné porovnávání (**M**ultiple **C**omparison **P**rocedure, **MCP**). Když ANOVA určí, že faktor *A* je statisticky významný, je možné nalézt úrovně faktoru *A,* které se významně liší od ostatních.

Druhy porovnávacích metod MCP. Volba porovnávací metody je ovlivněna

odpovědí na následující dvě otázky: (1) Víme už v průběhu experimentu, co chceme porovnávat? (2) Zajímáme se o všechny nebo jenom o některá porovnání? Budeme přitom rozlišovat dva typy chyb, chyby typu metodického a chyby typu experimentálního.

Metodická chyba δ: každé porovnání dvou průměrů se bere jako jediný test, který se provedl a označuje se δ. Pojmovou jednotkou je proto *chyba pro jedno porovnání.* Ostatní testy na datech jsou pak ignorovány vzhledem k výpočtu hodnoty chyby.
Experimentální chyba δ_f : hodnota chyby vyšla ze skupiny nezávislých testů. Je to pravděpodobnost nabytí jedné či více chyb typu I ve skupině nezávislých porovnání. Označíme tuto *chybu u skupiny nezávislých testů* δ_f. Vztah mezi oběma typy chyb je

$$\delta_f = 1 - (1 - \delta)^c \quad ,$$

kde *c* je celkový počet porovnání, provedených v úloze.

Definice metod MCP. Všechny MCP metody předpokládají nezávislost mezi úrovněmi faktoru čili sloupcovými výběry, homoskedasticitu a normalitu (kromě Kruskalova-Wallisova testu). Budiž $\bar{y}_i$ sloupcový průměr a n_i velikost sloupcového výběru *i*-tého sloupce, s^2 představuje průměrný čtverec chyb, způsobených $N - K$ stupni volnosti a při uvažování K úrovní faktoru A. Vícenásobné porovnávání může být provedeno (1) *automaticky*, kdy každý sloupcový průměr je porovnáván s každým, nebo (2) *plánovaně* dle zadaných váhových koeficientů C_i , $i = 1, ..., K,$ kdy budeme porovnávat určité vybrané sloupce s jinými vybranými sloupci. Jestliže všechny koeficienty C_i vykazují součet rovný nule, porovnání se nazývá *kontrast*. Koeficienty C_i se zadávají následujícím způsobem: chceme, například, pro 6 sloupců porovnat pouze první dva sloupcové průměry s posledními dvěma, tzn. vyčíslit významnost rozdílů $\bar{y}_5 - \bar{y}_1$, $\bar{y}_5 - \bar{y}_2$, $\bar{y}_6 - \bar{y}_1$, $\bar{y}_6 - \bar{y}_2$, což zapíšeme pomocí vahových koeficientů: $C_1 = -1$, $C_2 = -1$, $C_3 = 0$, $C_4 = 0$, $C_5 = 1$, $C_6 = 1$. Všimněte si, že suma vahových koeficientů musí dávat nulu a porovnání je proto kontrastem.

Bonferroniho porovnání všech párů. Test má odhalit, které páry se liší. Zvolí se metodická chyba δ tak, že bude korigovat požadovanou experimentální chybu δ_f. Je-li K sloupcových průměrů a zájem vyšetřit všechny možné kombinace párů sloupcových průměrů, metodická chyba δ je definována vztahem:

$$\delta = \frac{\delta_f}{K.K - 1}$$

a testační kritérium statistické významnosti testovaných párů je pro $\nu = N - K$ stupňů volnosti a $\gamma = \alpha$ dáno vztahem

$$\frac{*\bar{y}_i - \bar{y}_j*}{\sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \geq t_{\gamma,\nu} \quad .$$

Bonferroniho porovnání sloupců vůči kontrolnímu. Jestliže jeden sloupec bude

představovat *kontrolní sloupec* a všechny ostatní sloupce budeme porovnávat s kontrolním, půjde o $\underline{\nu} = K$-1 porovnání. Zvolíme metodickou chybu δ dle vzorce[17]

$$\delta ' \frac{\delta_f}{2(K \& 1)} .$$

Standardní porovnání. Plánovaný test významnosti určitého zvoleného porovnání, který se týká metodické chyby. Zadává se jedna z porovnávacích voleb: (a) standardního porovnání, (b) porovnání ortogonálními polynomy, (c) porovnání každého sloupce s prvním, (d) porovnání každého sloupce s posledním, (e) porovnání při více než třech uživatelských kontrastech a další volby. Nastavíme hladinu metodické chyby tak, že dosáhneme specifické hodnoty celkové chyby. Studentovo testační kritérium významnosti sloupcových průměrů testovaných párů je pro $\nu = N - K$ stupňů volnosti a $\gamma = \alpha/2$ dáno vztahem

$$\frac{\left| \sum_{j=1}^{K} C_j \bar{y}_j \right|}{s \sqrt{\sum_{j=1}^{K} \frac{C_j^2}{n_j}}} \$ t_{\gamma,\nu} .$$

Uvedeme ukázku zadání způsobu (a) "standardního porovnání": pro například $K = 4$ úrovně faktoru A máme k dispozici tři volby porovnání sloupcových průměrů. Všimněte si, že suma vahových koeficientů dává vždy nulu čili jde o kontrast.

Váhové koeficienty C_j	Provádí porovnání sloupcových průměrů:
-3, 1, 1, 1	Porovnává průměr 1. sloupce s průměry všech vyšších ostatních, tj. s průměry 2., 3. a 4. sloupce.
0, -2, 1, 1	Porovnává průměr 2. sloupce s průměry všech vyšších ostatních, tj. s průměry 3. a 4. sloupce.
0, 0, -1, 1	Porovnává průměr 3. sloupce s průměry všech vyšších ostatních, tj. s průměry 4. sloupce.

Kruskalovo-Wallisovo porovnání Z-skóre. Hodnoty Z-skóre (*standardizovaná veličina*, kdy od prvků sloupce je odečten sloupcový průměr a pak jsou poděleny směrodatnou odchylkou) jsou zde využity k porovnání sloupcových mediánů v páru při nesplnění předpokladů výběrové normality. Test však vyžaduje výběr o minimální četnosti $n_i = 5$ (a lépe ještě vyšší) v každé úrovni faktoru A. Hladina chyby je nastavena na základě metodické chyby tak, aby poskytla experimentální hladinu chyby δ_f. Za střední hodnoty může test užít vedle mediánu také průměr pořadí, jak je zřejmé ze vzorce pro $\delta ' \delta_f /K(K \& 1)$, porovnávající sloupec i se sloupcem j

$$\frac{*\bar{R}_i \;\&\; \bar{R}_j*}{\sqrt{\frac{N(N \,\%\, 1)}{12}\left(\frac{1}{n_i} \,\%\, \frac{1}{n_j}\right)}} \;\$\; z_\alpha \;,$$

kde N je celkový počet prvků, n_i je počet prvků v i-tém sloupci, R_i je suma pořadí v i-tém sloupci. Rozdělení z_{ij} je normální se střední hodnotou nula a rozptylem jedna. Je-li vypočtená hodnota z_{ij} pro dva sloupce (i a j) větší než kritická hodnota, pak se sloupcové průměry významně liší.

Scheffeho porovnání. Vyšetřuje všechna možná porovnání K sloupcových průměrů. Testační kritérium významnosti pro páry sloupcových průměrů je pro K-1 a pro $\nu = N - K$ stupňů volnosti rovno nebo větší než $\sqrt{(K \;\&\; 1)\, F_{\alpha,K\&1,\nu}}$. Je stejné jako Bonferroniho porovnání.

Postup jednofaktorové analýzy rozptylu (ANOVA1)

Vstupem je tabulka dat, obsahující pro jednotlivé sloupce čili úrovně A_1, ..., A_K faktoru A vždy n_i pozorování $\{y_{ij}\}$, i = 1, ..., K a j = 1, ..., n_i. Pro všechny testy je obvykle uvažována hladina významnosti $\alpha = 0.05$. Postup obsahuje kroky:

1. Přípravu dat: už přípravou dat lze zajistit větší věrohodnost dosažených výsledků.

(a) *Velikost výběru* je počet plných řádků. ANOVA byla původně odvozena za předpokladu, že četnosti ve sloupcích jsou shodné. V praxi je však tento předpoklad zřídka splněn. Stejně však platí, že čím těsněji je toto pravidlo splněno, tím věrohodnější jsou výsledky. Lze analyzovat i malé výběry, 4 až 5 hodnot ve sloupci. Máme-li testovat všechny výběrové předpoklady, je třeba prvků ve sloupci více, ze statistického hlediska nejlépe 30 a více.

(b) *Chybějící hodnoty* mohou způsobit vychýlení výsledků. V každém případě je poněkud nebezpečné analyzovat výběr s řadou chybějících hodnot.

(c) *Typ dat*: matematické pozadí F-testu požaduje, aby data byla *spojitá.* Kvůli zaokrouhlo-vání při zápisu dat, jsou všechna data vlastně technicky vzato diskrétní. Požadavek spojitosti je proto na místě, jsou-li data hodně zaokrouhlovaná.

(d) *Odlehlé hodnoty* obecně způsobují zborcení F-testů. Je třeba prozkoumat data v grafech exploratorní analýzy dat EDA, často se užívá krabicový graf. Pak následuje vyšetření, zda se odlehlé hodnoty vyskytují pouze v jednom sloupci nebo i v ostatních. Je-li odlehlá hodnota v datech pouze jednou, je třeba ji odstranit. Pakliže ji v datech ponecháme, je třeba dát přednost neparametrickému testování, F-test by totiž mohl selhat.

2. Ověření výběrových předpokladů: nestačí se soustředit na tabulku výsledků testování ANOVA. Je třeba pečlivě ověřit splnění základních předpokladů o výběru. Často data nemají ve všech sloupcích normální rozdělení a je třeba použít mocninnou (nebo logaritmickou) transformaci dat. Po transformaci pak data již vykazují normální rozdělení. I když je pouze jediný sloupec s nenormální rozdělením, transformace celého výběru přinese zlepšení výsledků.

(a) *Náhodnost:* metoda odběru vzorku by měla zajistit, že každý prvek souboru má stejnou pravděpodobnost být vybrán do výběru.

(b) *Nezávislost:* aplikací von Neumannova testu ověříme nezávislost prvků výběru. Budeme-li, například, porovnávat levé a pravé pneumatiky u výběru určitého množství aut, nezávislost nebude

zaručena.

(c) *Normalita:* nejlépe je začít vyšetřením rankitového grafu odchylek od totálního průměru. Pak následuje řada testů normality. Síla těchto testů se zvyšuje s velikostí výběru. I když byla normalita potvrzena, prověříme velikost výběru, zda je možné brát výsledky testu za věrohodné.

(d) *Homoskedasticita:* aby bylo možné užít řadu statistických testů, je třeba ověřit, zda rozptyly sloupců jsou shodné (homoskedasticita). V krabicových grafech je sledována šířka krabic, zda je u všech sloupců stejná. Numericky lze ověřit homoskedasticitu pomoci modifikovaného Levenova testu[17].

3. Průměry a efekty úrovní: je proveden výpočet parametrů $\hat{\mu}_i$, $\hat{\mu}$, $\hat{\alpha}_i$, reziduí $\hat{e}_{ij}$, Jackknife reziduí $\hat{e}_{Ji}$ a diagonálních prvků H_{ii} projekční matice $\boldsymbol{H}$. Jsou identifikovány vlivné body, pro které je $\hat{e}_{Ji} > 2$ a $H_{ii} > 2K/\sum_{i=1}^{K} n_i$ (viz 6. kapitola).

4. Volbu statistických testů významnosti faktoru A v tabulce ANOVA: Je sestavena tabulka ANOVA a proveden *F*-test významnosti efektů faktoru *A*. Předem je třeba ověřit výběrové předpoklady a zvolit správný test:

(a) *Normalita a homoskedasticita dat:* aplikujeme *F*-test.

(b) *Normalita a heteroskedasticita dat:* pokusíme se stabilizovat rozptyl mocninnou transformací (nebo logaritmickou). Pak užijeme test shodnosti středních hodnot u dvou výběrů při nehomogenitě rozptylů. Nelze užít ani Kruskalův-Wallisův test, protože tento test také předpokládá shodné rozptyly obou výběrů.

(c) *Nenormalita a homoskedasticita dat:* užijeme Kruskalův-Wallisův test.

(d) *Nenormalita a heteroskedasticita dat:* když nejde data transformovat za účelem stabilizace rozptylu a zajištění normality, užijeme Kolmogorův-Smirnovův test (viz cit.[17]), který testuje obojí, průměry i rozptyly současně. Jelikož však už víme z Levenova testu (viz cit.[17]), že rozptyly nejsou stejné, je otázkou, zda Kolmogorův-Smirnovův test přinese něco nového.

Testování hypotéz: výklad analýzy rozptylu je snadný. Jednoduše sledujeme *F*-test. Je-li hodnota spočtené hladiny významnosti menší než předvolená hladina významnosti α (obyčejně 0.05), můžeme potvrdit, že přinejmenším dva sloupcové průměry jsou odlišné.

5. Vícenásobné porovnávání sloupcových průměrů MCP: postup předpokládá normalitu a homoskedasticitu výběrových sloupců. Není-li splněna normalita pro každý sloupec, je třeba užít Kruskalův-Wallisův test vícenásobného porovnávání MCP:

(a) *Plánované všechny možné páry:* víme-li dopředu, že budeme vyšetřovat všechny páry, užijeme Bonferroniho porovnávání párů.

(b) *Neplánované všechny možné páry:* užijeme Scheffeho porovnávání.

(c) *Každý versus kontrolní sloupec:* užijeme Bonferroniho porovnání všech sloupců vůči kontrolnímu.

(d) *Vybrané a plánované sloupce:* užijeme Standardní porovnávání a nastavíme hladinu α.

6. Grafy a diagramy: je konstruován rankitový graf Jackknife reziduí a transformační graf závislosti s_i vs. $\hat{\mu}_i$. Pokud jsou všechna data kladná a tato závislost je přibližně lineární, lze zvolit logaritmickou transformaci.

Vzorová úloha 5.1 *Zkrácený postup jednofaktorové analýzy rozptylu*

Na úloze **B5.02** *Porovnání nové metody v sedmi laboratořích* ukážeme postup

jednofaktorové analýzy rozptylu. Kirchhoefer[16] navrhl poloautomatickou metodu na stanovení maleátu chlorfeniraminu v tabletách. Sedm laboratoří A1 až A7 (faktor *A*) opakovalo analýzu tablety o obsahu 4 mg této látky celkem 10krát. Cílem je posoudit vliv laboratoří na výsledek analýzy.

Řešení:

1. Průměry a efekty úrovní: je proveden výpočet parametrů sloupcových průměrů $\hat{\mu}_i$, celkového průměru $\hat{\mu}$, sloupcových efektů α_i, reziduí $\bar{e}_{ij}$, (ADSTAT).

Počet sloupců, *K*	= 7	
Počet dat, *N*	= 70	
Celkový průměr $\hat{\mu}$	= 3.9846	
Reziduální rozptyl s^2 =	3.6730	
Úroveň faktoru *A*	**Sloupcový průměr** $\hat{\mu}$	**Efekt** α_i
1	4.0620	0.077429
2	3.9970	0.012429
3	4.0030	0.018429
4	3.9200	-0.04571
5	3.9570	-0.07571
6	3.9550	-0.09571
7	3.9980	0.013429

2. Tabulka ANOVA: je sestavena tabulka ANOVA (ADSTAT) a proveden *F*-test významnosti faktoru *A*. Jelikož Fisherovo-Snedecorovo testační kritérium F_e = 5.660 nabývá vyšší hodnoty než kvantil $F_{1-0.05}(7\text{-}1, 70\text{-}7) = 2.246$, je nulová hypotéza H_0: *Efekty faktoru A jsou nulové* zamítnuta a faktor *A* je statisticky významný.

H_0: Efekty faktoru *A* jsou nulové, **H_A:** ... nejsou nulové
Kvantil *F*(1-alfa, *K-1, N-K*) = 2.246

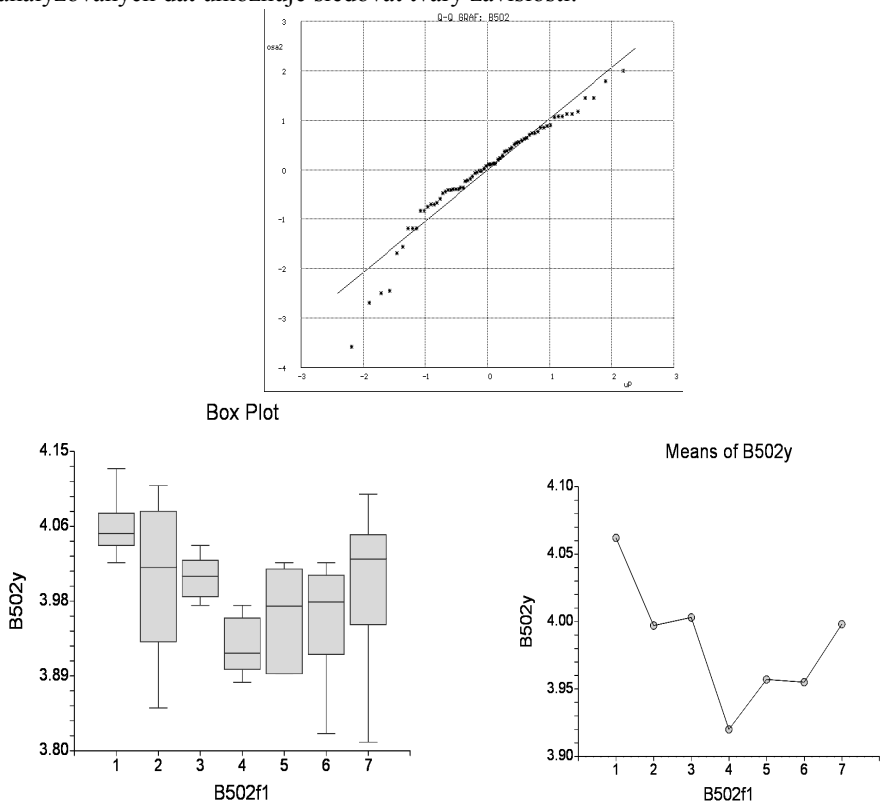
Zdroj rozptylu	**Stupně volnosti**	**Součet čtverců**	**Průměrný čtverec**	**Testační kritérium**	**Závěr H_0 je**	**Spočtená hlad. výz.**
Mezi úrovněmi *K*-1 = 6		0.12474	0.020790	5.660	Zamítnuta	0.000
Rezidua *N-K*	= 63	0.23140	0.0036730			
Celkový *N*-1	= 69	0.35614	0.0051614			

3. Vícenásobné porovnávání Schéffeho procedurou (ADSTAT): jsou testovány lineární kontrasty pro zadané kombinace úrovní, H_0: $\mu_i - \mu_j = 0$, Schefféhо metodou mnohonásobného porovnávání. Tři dvojice úrovní faktoru *A*, a to P1=P4, P1=P5, P1=P6, vycházejí odlišně od ostatních, totiž nerovnají se sobě. Ostatní úrovně jsou, co do hodnoty, považovány za shodné.

Hypotéza H_0	**Průměrný párový rozdíl**	**Meze konfidenčního intervalu dolní**	**horní**	**Hypotéza H_0 je**
P1 = P2	0.065	-0.035	0.165	Akceptována
P1 = P3	0.059	-0.041	0.159	Akceptována
P1 = P4	0.142	0.042	0.242	Zamítnuta
P1 = P5	0.105	0.005	0.205	Zamítnuta
P1 = P6	0.107	0.007	0.207	Zamítnuta
P1 = P7	0.064	-0.036	0.164	Akceptována
P2 = P3	-0.006	-0.106	0.094	Akceptována
P2 = P4	0.077	-0.023	0.177	Akceptována
P2 = P5	0.040	-0.060	0.140	Akceptována
P2 = P6	0.042	-0.058	0.142	Akceptována
P2 = P7	-0.001	-0.101	0.099	Akceptována
P3 = P4	0.083	-0.017	0.183	Akceptována

P3 = P5	0.046	-0.054	0.146	Akceptována
P3 = P6	0.048	-0.052	0.148	Akceptována
P3 = P7	0.005	-0.095	0.105	Akceptována
P4 = P5	-0.037	-0.137	0.063	Akceptována
P4 = P6	-0.035	-0.135	0.065	Akceptována
P4 = P7	-0.078	-0.178	0.022	Akceptována
P5 = P6	0.002	-0.098	0.102	Akceptována
P5 = P7	-0.041	-0.141	0.059	Akceptována
P6 = P7	-0.043	-0.143	0.057	Akceptována

4. Grafy a diagramy: je konstruován rankitový graf Jackknife reziduí. Rankitový graf dokazuje, že většina reziduí odpovídá předpokladu normality. V dolní části grafu je několik odlehlých hodnot, zbytek dobře splňuje lineární závislost. Zobrazení průměru analyzovaných dat umožňuje sledovat tvary závislosti.



Obr. 5.1 Jednofaktorová analýza rozptylu v úloze B502: rankitový graf reziduí, *ADSTAT* (nahoře), krabicový graf úrovní faktoru *A* (vlevo dole), diagram sloupcových průměrů *NCSS2000* (vpravo dole).

Vzorová úloha 5.2 *Podrobný postup v jednofaktorové analýze rozptylu*

Na vzorové úloze **H5.11** *Vliv tavby na obsah mědi v bronzu* ukážeme podrobný postup jednofaktorové analýzy rozptylu a především rozličné metody vícenásobného porovnání faktoru *A*. Bylo zkoumáno, zda se obsah mědi v bronzu mění od tavby k tavbě. U každé tavby byly odebrány 4 vzorky a stanoven procentuální obsah mědi v bronzu. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda existuje vliv tavby (faktor *A*) na obsah mědi v bronzu. Ověřte rovněž výběrové předpoklady vhodnou grafickou a numerickou metodou.

Řešení:

1. Průměry a efekty úrovní: je proveden výpočet odhadů sloupcových průměrů $\hat{\mu}_i$, celkového průměru $\bar{\mu}$, sloupcových efektů α_i a reziduí $\bar{e}_{ij}$, (NCSS2000).

Zdroj	**Počet**	**Průměr**	**Efekt**
Vše	20	87	
A: H511f1			
1	4	81.25	-5.75
2	4	88.5	1.50
3	4	86.25	-0.75
4	4	90.25	3.25
5	4	88.75	1.75

Zdroj: návěští řádku. **Počet:** četnost pozorování pro výpočet průměru. **Průměr:** hodnota aritmetického průměru. **Efekt:** velikost komponenty, kterou tento člen přispívá do průměru. **Graf průměrů:** vizuálně lze vyšetřit jednotlivé sloupce krabicovým grafem, který vyšetří symetrii, přítomnost odlehlých bodů, obecnou shodu středních hodnot a shodu rozptylů (obr. 5.2b).

2. Testy předpokladů o výběru: tři testy umožnují otestovat šikmost, špičatost a celkovou normalitu dat. Jestliže některý z nich zamítne hypotézu o normalitě, data nemohou být považována za normální, gaussovská.

Testy výběrových předpokladů:	**Testační**	**Spočtená**	**Závěr testu (0.05)**
Předpoklad	**kritérium**	**hlad. význ.**	**H_0 je**
Test šikmosti reziduí	-1.1439	0.252672	Přijata
Test špičatosti reziduí	-0.9800	0.327102	Přijata
Omnibus test reziduí	2.2688	0.321614	Přijata
Modifikovaný Levenův test stejných rozptylů	0.0765	0.988310	Přijata

Výsledky tří testů normality a jednoho testu homoskedasticity. ANOVA totiž předpokládá, že rezidua, odchylky od skupinových průměrů vykazují normální rozdělení. Testy nezávislosti prvků ve výběru a náhodnosti prvků ve výběru se zde neprovádějí. Obě vlastnosti by měly být ošetřeny v experimentálním postupu. Testy normality jsou provedeny na hladině významnosti $\alpha = 0.05$. Protože žádný z testů *nezamítl* nulovou hypotézu H_0 o normalitě, můžeme být přesvědčeni, že normalita je prokázána. Síla testu je ovlivněna velikostí výběru: malé výběry vykazují normalitu zřídka. Jsou vyšetřována dvě kritéria normality, šikmost a špičatost. Jestliže se normalita reziduí neprokáže kvůli šikmosti, mohla by se užít mocninová nebo logaritmická transformace k normování dat. V případě nenormality se použije Kruskalův-Wallisův neparametrický test. Základní předpoklady o nezávislosti výběru, spojité náhodné proměnné o měřicí stupnici se musí ovšem dodržet. Tento test má další předpoklady, že rozdělení sloupců jsou stejná (i když nemusí být normální) co do formy a tvaru a mohou se lišit pouze v parametru místa. **Test homoskedasticity** (modifikovaný Levenův test[17]): test byl shledán jedním z nejlepších testů shody rozptylů-homoskedasticity. Test ukazuje, že sloupce mají shodné rozptyly.

3. ANOVA tabulka: je proveden *F*-test významnosti efektů faktoru *A*. Jelikož nabývá Fisherovo-Snedecorovo testační kritérium $F_e = 5.988$ vyšší hodnoty než kvantil $F(1\text{-}0.05, 7\text{-}1, 70\text{-}7) = 3.056$, je nulová hypotéza H_0: *"Efekty faktoru A jsou nulové"* zamítnuta

a faktor *A* se uvažuje jako statisticky významný.

NCSS2000:						
Zdroj rozptylu	**SV**	**Suma čtverců**	**Průměrný čtverec**	**F-test**	**Spočtená hlad. význ.**	**Síla testu (α=0.05)**
A: H511f1	4	198.00	49.500	5.99	0.004374*	0.93514
S(A)	15	124.00	8.266666			
Total (Adjust.)	19	322.00				
* Faktor je významný při α = 0.05						

Zdroj rozptylu: obsahuje sloupce zdroje proměnlivosti v ANOVA modelu. **Suma čtverců:** suma čtverců odchylek se uvádí spíše pro úplnost této tabulky než pro přímé využití s vysvětlením. **Průměrný čtverec:** průměrný čtverec představuje sumu čtverců odchylek dělenou počtem stupnů volnosti. ***F*-test:** Fisherovo-Snedecorovo testační kritérium $F_e = 5.99$ je vyšší než kvantil $F(1\text{-}0.05, 7\text{-}1, 70\text{-}7) = 3.056$, a proto je nulová hypotéza H_0: "*Efekty faktoru A jsou nulové"* zamítnuta a faktor *A* se považuje za statisticky významný. **Spočtená hladina významnosti:** pro *F*-kritérium 0.004374 je menší než $\alpha = 0.05$, a proto je nulová hypotéza zamítnuta a *F*-kritérium, a tím pádem i faktor *A*, statisticky významné. **Síla testu** je pravděpodobnost zamítnutí hypotézy "o stejných průměrech", když průměry ve skutečnosti nejsou stejné. Je rovna 1 minus pravděpodobnost chyby typu II, čili β. Síla testu závisí na velikosti výběru, velikosti rozptylu, hladině významnosti α a skutečném rozdílu průměrů. Vysoká hodnota síly testu je žádoucí. Vysoká hodnota znamená vysokou pravděpodobnost zamítnutí nulové hypotézy, když nulová hypotéza je nesprávná. Je to vlastně kritická míra přesnosti testování hypotézy. Obecně bychom měli uvažovat sílu testu vždy, když přijmeme nulovou hypotézu. Když přijmeme nulovou hypotézu s velkou silou testu, není dále co řešit. Přinejmenším víme, že průměry *nejsou* různé. Když však přijmeme nulovou hypotézu s malou silou testu, je před námi jedna z následujících možných voleb:

1. Zvýšíme hladinu významnosti α. Např. zvýšíme α z hodnoty 0.01 na 0.05 a toto zvýšení způsobí i nárůst síly testu.
2. Zvýšíme velikost výběru, což opět způsobí nárůst síly testu. Je-li však síla testu veliká, zvětšení výběru má malý vliv na sílu testu.
3. Snížíme velikost rozptylu. Můžeme např. předělat strategii experimentu a získat tak přesnější měření, hodnoty bez okrajových hodnot a s menším rozptylem.

4. Rozličné metody vícenásobné porovnávání: jsou testovány lineární kontrasty pro zadané kombinace úrovní, H_0: $\mu_i - \mu_j = 0$.

Následující metody vícenásobného porovnávání mají stejný výstup: **Alfa:** značí zvolenou hladinu významnosti. **ST:** stupně volnosti. **MSE:** hodnota průměrného čtverce chyb. **Kritická hodnota:** tabulkový kvantil pro dané stuně volnosti a hladinu významnosti. **Sloupec** čili úrovně faktoru *A*. **Počet:** četnost pozorování v průměru. **Průměr:** hodnota aritmentického průměru. **Liší se od sloupců:** seznam sloupců, které se liší od dotyčného testovaného sloupce, nadepsaného v tomto řádku. Všechny sloupce zde nevyjmenované jsou statisticky nevýznamně odlišné od dotyčného sloupce, nadepsaného v řádku.

A. Bonferroniho porovnání všech párů: test má odhalit, které páry se liší.

Odezva: H511y, **Faktor *A*:** H511f1, **Alfa**=0.050, **SV**=15 **MSE**=8.267, **Kritická hodnota**=3.286,			
Sloupec	**Počet**	**Průměr**	**Liší se od sloupců**
1	4	81.25	2, 5, 4
3	4	86.25	
2	4	88.5	1
5	4	88.75	1
4	4	90.25	1

Test ukazuje, že tři páry úrovní faktoru *A*, a to 1. sloupec a 2. sloupec, 1. sloupec a 4. sloupec, a 1. sloupec a 5. sloupec vycházejí statisticky významně odlišné. Ostatní sloupce jsou považovány za shodné.

B. Bonferroniho porovnání sloupců vůči kontrolnímu: jestliže jeden sloupec bude představovat *kontrolní sloupec* a všechny ostatní sloupce budeme porovnávat s tímto kontrolním, půjde o *K*-1 porovnání.

Odezva: H511y, **Faktor *A*:** H511f1, **Alfa**=0.050, **SV**=15 **MSE**=8.267, **Kritická hodnota**=2.837,

Sloupec	Počet	Průměr	Liší se od sloupců
1	4	81.25	2, 5, 4
3	4	86.25	
2	4	88.5	1
5	4	88.75	1
4	4	90.25	1

Bylo dosaženo podobného závěru jako u předešlého testu, tři páry úrovní faktoru *A*, a to 1. a 2. sloupec, 1. a 4., a 1. a 5. sloupec vycházejí statisticky významně odlišné.

C. Standardní porovnání: test významnosti plánovaného porovnání:

(a) Porovnání 1. sloupce se všemi ostatními vyššími, tj. s 2., 3., 4. a 5. sloupcem:

Plánované porovnání: A1 (-4, 1, 1, 1, 1,)
Odezva: H511y, **Faktor *A*:** H511f1, **Alfa**=0.050, **SV**=15, **MSE**=8.267, **Porovnávaná hodnota**=28.75, **t-test**=4.472, **Prob>|t|**=0.00045, **Závěr testu(0.05)**= Zamítnuto, **Standardní chyba porovnávané hodnoty**=6.429,

Sloupec	Váhový koeficient	Počet	Průměr
1	-4	4	81.25
2	1	4	88.5
3	1	4	86.25
4	1	4	90.25
5	1	4	88.75

Vedle vysvětlených pojmů jsou zde ještě: **Porovnávaná hodnota:** vznikne násobením porovnávacího koeficientu (Váhový koeficient) svým sloupcovým průměrem a následujícím součtem přes všechny sloupce. ***t*-test:** testuje, zda se porovnávaná hodnota významně liší od nuly. **Prob , *t*** je spočtená hladina významnosti **Prob**, která by měla být nad kritickou hodnotou: je-li **Prob** menší než zadaná hladina významnosti α, je porovnávaná hodnota statisticky významná. **Závěr testu (0.05):** rozhodnutí na základě zadané hodnoty α. **Standardní chyba porovnávané hodnoty:** standardní chyby vyčíslované porovnávané hodnoty. Tvoří jmenovatele v testované statistice *t*-testu. **Váhový koeficient:** váhový koeficient pro tento sloupec. Porovnáním 1. sloupcového průměru se všemi ostatními vyššími 2., 3., 4., a 5. byla zamítnuta nulová hypotéza o shodnosti těchto sloupcových průměrů, protože **Závěr testu(0.05)** přináší závěr - Zamítnuto.

(b) Porovnání 2. sloupce se všemi ostatními vyššími, tj. s 3., 4. a 5. sloupcem:

Plánované porovnání A2 (0, -3, 1, 1, 1,)
Odezva: H511y, **Faktor *A*:** H511f1, **Alfa**=0.050, **SV**=15 **MSE**=8.267, **Porovnávaná hodnota**=-0.25, **t-test**=0.0502, **Prob>|t|**=0.961, **Závěr testu(0.05)**=Přijato, **Standardní chyba porovnávané hodnoty**=4.980,

Sloupec	Váhový koeficient	Počet	Průměr
1	0	4	81.25
2	-3	4	88.5
3	1	4	86.25
4	1	4	90.25
5	1	4	88.75

Porovnáním 2. sloupcového průměru se všemi ostatními vyššími 3., 4., a 5. byla přijata nulová hypotéza o shodnosti těchto sloupcových průměrů, protože **Závěr testu(0.05)** přináší závěr - Přijato.

(c) Porovnání 3. sloupce se všemi ostatními vyššími, tj. s 4. a 5. sloupcem:

Plánované porovnávání: A3 (0, 0, -2, 1, 1)
Odezva: H511y, **Faktor *A*:** H511f1, **Alfa**=0.050, **SV**=15, **MSE**=8.267, **Porovnávaná hodnota**=6.5,

t-test=1.846, **Prob>|t|**=0.0847, **Závěr testu(0.05)**=Přijato, **Standardní chyba porovnávané hodnoty**=3.521,

Sloupec	Váhový koeficient	Počet	Průměr
1	0	4	81.25
2	0	4	88.5
3	-2	4	86.25
4	1	4	90.25
5	1	4	88.75

Porovnáním 3. sloupcového průměru se všemi ostatními vyššími 4., a 5. byla přijata nulová hypotéza o shodnosti těchto sloupcových průměrů, protože **Závěr testu(0.05)** přináší závěr - Přijato.

(d) Porovnání 4. sloupce se všemi ostatními vyššími, tj. s 5. sloupcem:

Plánované porovnání: A4 (0, 0, 0, -1, 1)
Odezva: H511y, **Faktor *A*:** H511f1, **Alfa**=0.050, **SV**=15, **MSE**=8.267, **Porovnávaná hodnota**=-1.5, **t-test**=0.738, **Prob>|t|**=0.472, **Závěr testu(0.05)**=Přijato, **Standardní chyba porovnávané hodnoty**=2.033,

Sloupec	Váhový koeficient	Počet	Průměr
1	0	4	81.25
2	0	4	88.5
3	0	4	86.25
4	-1	4	90.25
5	1	4	88.75

Porovnáním 4. sloupcového průměru se všemi ostatními vyššími, tj. s 5., byla přijata nulová hypotéza o shodnosti těchto sloupcových průměrů, protože **Závěr testu(0.05)** přináší závěr - Přijato.

D. Vícenásobné porovnávání Schéffeho procedurou: jsou testovány lineární kontrasty pro zadané kombinace úrovní, H_0: $\mu_i - \mu_j = 0$, Scheffého metodou mnohonásobného porovnávání (ADSTAT).

NCSS2000: Scheffeho vícenásobné porovnávání
Odezva: H511y, **Faktor *A*:** H511f1, **Alfa**=0.050, **SV**=15, **MSE**=8.267, **Kritická hodnota**=3.496,

Sloupec	Počet	Průměr	Liší se od sloupců
1	4	81.25	2, 5, 4
3	4	86.25	
2	4	88.5	1
5	4	88.75	1
4	4	90.25	1

Bylo dosaženo podobného závěru jako u Bonferroniho testu, tři páry úrovní faktoru *A* dává 1. a 2. sloupec, 1. a 4., a 1. a 5. sloupec vycházejí statisticky významně odlišné.

ADSTAT: Scheffeho vícenásobné porovnávání

Hypotéza H_0	Průměrný párový rozdíl	Meze konfidenčního intervalu dolní	horní	Závěr
P1=P2	-7.250	-14.358	-0.142	Zamítnuta
P1=P3	-5.000	-12.108	2.108	Akceptována
P1=P4	-9.000	-16.108	-1.892	Zamítnuta
P1=P5	-7.500	-14.608	-0.392	Zamítnuta
P2=P3	2.250	-4.858	9.358	Akceptována
P2=P4	-1.750	-8.858	5.358	Akceptována
P2=P5	-0.250	-7.358	6.858	Akceptována
P3=P4	-4.000	-11.108	3.108	Akceptována

P3=P5	-2.500	-9.608	4.608	Akceptována
P4=P5	1.500	-5.608	8.608	Akceptována

Tři páry úrovní faktoru *A* P1=P2, P1=P4, P1=P5 vycházejí odlišně od ostatních, jsou totiž statisticky významně odlišné. Ostatní sloupce jsou považovány za shodné.

E. Kruskalovo-Wallisovo vícenásobné porovnávání pořadí: test je neparametrickou náhražkou jednofaktorové analýzy rozptylu, když předpoklad normality není splněn. Jsou-li navíc vnitřní vazby v datech, musíme užít korigovanou verzi tohoto testu.

Kruskalovo-Wallisovo vícenásobné porovnávání pořadí:
Hypotézy: **H_0:** Všechny mediány jsou stejné.
H_A: Přinejmenším dva mediány jsou vzájemně odlišné.

Metoda	**SV**	**χ^2 (H)**	**Spočtená hlad. významnosti**	**Závěr testu(0.05)**
Nekorigované na vazby	4	10.45357	0.033443	Zamítnuta H_0
Korigované na vazby	4	10.54075	0.032240	Zamítnuta H_0
Počet vázaných hodnot	5			
Korekční faktor	66			

Sloupec	**Počet**	**Suma pořadí**	**Průměrné pořadí**	**Z-skóre**	**Medián**
1	4	12.00	3.00	-2.8347	82
2	4	50.50	12.63	0.8032	89
3	4	35.00	8.75	-0.6614	87
4	4	60.50	15.13	1.7481	90.5
5	4	52.00	13.00	0.9449	89

Hypotézy: jsou naformulovány dvě hypotézy. Nulová hypotéza říká H_0: "*Mediány všech sloupců jsou stejné"* proti alternativní H_A: "*Alespoň jeden medián se liší od ostatních"*. **Metoda:** jsou prezentovány výsledky dvou testů: *první* je Kruskalův-Wallisův test bez korekcí na vnitřní vazby, čili stejná pořadí, a *druhý* je s korekcí na vnitřní vazby. Nejsou-li žádné vazby, žádná stejná pořadí jsou výsledky obou testů stejné. **SV:** stupně volnosti velkého výběru χ^2-aproximace rozdělením Kruskalova-Wallisova testu. Všimněme si, že počet stupňů volnosti je roven počtu sloupců (úrovní faktoru) minus 1. **χ^2(H):** hodnota testačního kritéria *H*, nekorigovaného Kruskalova-Wallisova testu se vyčíslí dle vzorce

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(N+1)$$

a týž test, korigovaný na případ stejných pořadí, se vyčíslí dle téhož *H*, děleného korekčním faktorem dle

$$H_C = H \Big/ \left[1 - \frac{\sum_{i=1}^{K} t_i(t_i^2 - 1)}{N(N^2 - 1)}\right] .$$

V obou vzorcích je *N* celkový počet prvků, n_i je počet prvků v *i*-tém sloupci, *K* je počet sloupců (úrovní faktoru *A*), R_i je suma pořadí *i*-tého sloupce a *t* je počet odpovídajících stejných pořadí. **Spočtená hladina významnosti:** o tom, že kritérium *H* nabývá χ^2 rozdělení. Pravděpodobnost kritéria *H* je větší než pravděpo-dobnost získaná touto analýzou, např. testovat na hladině významnosti $\alpha = 0.05$ znamená, že tato pravděpodobnost by měla být menší než 0.05, aby bylo *H* významné. **Závěr testu (0.05):** rozhodnutí o nulové hypotéze na bázi tohoto testu. **Počet souborů vázaných hodnot:** nejsou-li žádné vazby, čili jsou stejná pořadí, je toto číslo rovno nule. **Korekční faktor:** jde o část korekčního faktoru, který je roven *H* u kritéria $\sum_{i=1}^{k} t_i(t_i^2 - 1)$.

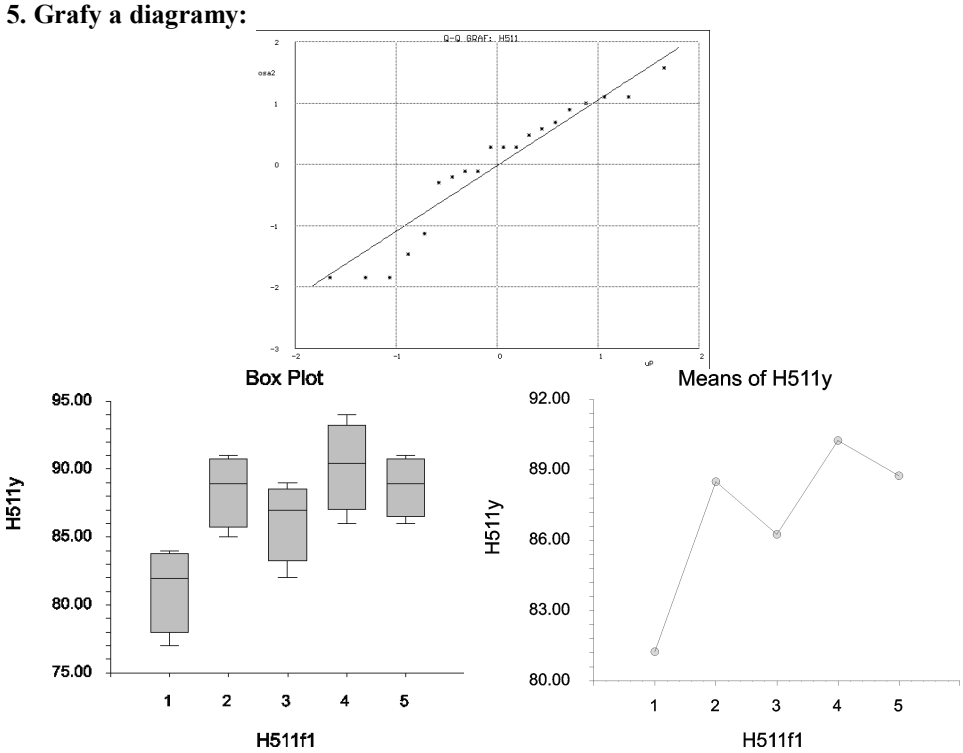
F. Kruskalův-Wallisův test vícenásobného porovnání pomocí *Z*-skóre: hodnoty *Z*-skóre (*standardizovaná veličina*, kdy od prvků sloupce je odečten sloupcový průměr a pak jsou vyděleny směrodatnou odchylkou) jsou testovány, zda mediány libovolných dvou sloupců jsou významně rozdílné: jde o *Z* skóre, které bylo upraveno pro vícenásobný

test tak, že α/2 je dělena hodnotou $K(K\text{-}1)/2$, a výsledkem je $z_{\alpha/K(K-1)}$. Zde $K(K+1)$ představuje počet možných párů všech K sloupců. Když ale provedeme omezený počet testování m, menší než je celkový počet sloupců, vydělíme α/2 pouze číslem m.

Kruskalův-Wallisův test vícenásobného porovnání pomocí *Z*-skóre					
H511y	1	2	3	4	5
1	0.0000	2.3104	1.3802	2.9105	2.4004
2	2.3104	0.0000	0.9302	0.6001	0.0900
3	1.3802	0.9302	0.0000	1.5303	1.0202
4	2.9105	0.6001	1.5303	0.0000	0.5101
5	2.4004	0.0900	1.0202	0.5101	0.0000
Vlastní test: mediány se významně liší, je-li *Z*-skóre > 1.9600					
Bonferroniho test: mediány se významně liší, je-li *Z*-skóre > 2.8070					

Porovnáním každého s každým plyne, že hodnoty testačního kritéria z_{ij} vyšší než kritická hodnota značí, že sloupcové průměry jsou statisticky významně odlišné.

5. Grafy a diagramy:



Obr. 5.2 Jednofaktorová analýza rozptylu v úloze H511: rankitový graf reziduí, *ADSTAT* (nahoře), krabicový graf úrovní faktoru A (vlevo dole), diagram sloupcových průměrů *NCSS2000* (vpravo dole).

5.2 Dvoufaktorová analýza rozptylu bez opakování v cele

Při dvoufaktorové analýze rozptylu se provádí experimenty na různých úrovních dvou faktorů A a B. Kombinace úrovní faktorů tvoří typickou mřížkovou strukturu, jejímž elementem je tzv. ***cela***. Platí, že (i, j)-tá cela odpovídá kombinaci úrovně A_i faktoru A a B_j

faktoru *B*. V každé cele je obecně n_{ij} pozorování. Často se však setkáváme s případem ***bez opakování***, kdy v každé cele je pouze jediné pozorování, $n_{ij} = 1$. Kromě řádkových α_i a sloupcových β_j efektů se zde vyskytuje také interakční člen τ_{ij}. Tento člen je důsledkem různých kombinací sloupcových a řádkových efektů.

	B_1	B_2	...	B_M
A_1	.	.	...	.
A_2	.		...	.
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
A_N	.	.	...	.

cela A_2 B_2

Obvykle se užívá ***Tukeyův model interakce***, vyjádřený tvarem $\tau_{ij} = C\,\alpha_i\,\beta_j$, kde *C* je konstanta. U těchto modelů obsahuje každá cela právě jednu hodnotu y_{ij}. O chybách ε_{ij} se předpokládá, že jsou to nezávislé a stejně rozdělené náhodné veličiny s nulovou střední hodnotou a konstantním rozptylem. K testování se navíc předpokládá, že rozdělení chyb je normální. Definují se omezující podmínky

$$\sum_{i=1}^{N} \alpha_i = 0; \quad \sum_{j=1}^{M} \beta_j = 0; \quad \sum_{i=1}^{N} \tau_{ij} = 0; \quad \sum_{j=1}^{M} \tau_{ij} = 0 .$$

V případě pouze aditivního působení jednotlivých faktorů je $\tau_{ij} = 0$ pro všechna $i = 1, ..., N$ a $j = 1, ..., M$. Odhady parametrů μ, α_i, β_j lze pak určit ze vztahů

$$\hat{\mu} = \frac{1}{N M} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}, \quad \alpha_i = \frac{1}{M} \sum_{j=1}^{M} y_{ij} - \hat{\mu}, \quad \hat{\beta}_j = \frac{1}{N} \sum_{i=1}^{N} y_{ij} - \hat{\mu} .$$

Pro rezidua $\hat{e}_{ij}$ platí $e_{ij} = y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. K určení interakcí můžeme využít skutečnosti, že $\tau_{ij} = E(y_{ij}) - \mu - \alpha_i - \beta_j$ a pro odhad interakcí platí přibližně $\hat{\tau}_{ij} \doteq \hat{e}_{ij}$.

Pak lze snadno identifikovat ***Tukeyův model*** interakce. Platí-li tento model, vyjde na grafu $\hat{e}_{ij}$ vs. $\hat{\alpha}_i \hat{\beta}_j$ lineární trend. Ze směrnice odpovídající regresní přímky se odhadne parametr *C*. Platí pro něj výraz:

$$C = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \hat{e}_{ij} \hat{\alpha}_i \hat{\beta}_j}{\sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\alpha}_i^2 \hat{\beta}_j^2} .$$

Graf e_{ij} vs. $\alpha_i\,\beta_j / \hat{\mu}$ se označuje jako graf *neaditivity*. Pokud vyjde v tomto grafu

nenáhodný trend, znamená to, že je třeba uvažovat interakce.

Analýza rozptylu pro dvojné třídění s interakcí Tukeyova typu

Součet čtverců pro	**Stupně volnosti**	**Průměrný čtverec**	**Kritérium *F***
Faktor *A*, $S_A = M \sum_{i=1}^{N} \alpha_i^2$	$N - 1$	$M_A = S_A/(N-1)$	$F_A = M_A/M_{AB}$
Faktor *B*, $S_B = N \sum_{j=1}^{M} \beta_j^2$	$M - 1$	$M_B = S_B/(M-1)$	$F_B = M_B/M_{AB}$
Interakce (Tukey)	1	$M_T = S_T$	$F_T = M_T/M_E$
Reziduální, $S_R = S_{AB}-S_T$	$NM - N - M$	$M_E = S_R/(NM-N-M)$	-
Celkový, $S_C = \sum_{(i)} \sum_{(j)} (\hat{\mu} - y_{ij})^2$	$NM - 1$	-	-

V tabulce představuje S_T součet čtverců odchylek odpovídající Tukeyově interakci

$$S_T = \frac{\left(\sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \hat{\alpha}_i \hat{\beta}_j \right)^2}{\sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\alpha}_i^2 \beta_j^2} .$$

Symbol S_{AB} označuje reziduální součet čtverců pro případ bez interakcí

$$S_{AB} = \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 .$$

Odpovídající průměrný čtverec je $M_{AB} = \frac{S_{AB}}{(N - 1)(M - 1)}$. Hodnota M_{AB} je nevychýleným odhadem rozptylu σ^2. Pomocí *F*-kritéria lze opět provádět statistické testy. Začíná se testováním nulové hypotézy H_0: "*Tukeyova interakce je nevýznamná*", pro kterou lze použít testační statistiku F_T. Za předpokladu platnosti nulové hypotézy H_0 má tato testační statistika *F*-rozdělení s 1 a ($N\,M - N - M$) stupni volnosti. Pokud nelze tuto hypotézu zamítnout, testuje se nulová hypotéza H_0: $\alpha_i = 0$, $i = 1, ..., N,$ (efekty řádků čili faktoru *A* jsou nevýznamné) pomocí statistiky F_A nebo nulová hypotéza H_0: $\beta_j = 0$, $j = 1, ..., M$, (efekty sloupců čili faktoru *B* jsou nevýznamné) pomocí statistiky F_B. Obě tyto testační statistiky jsou uvedeny v tabulce. Za předpokladu platnosti hypotézy H_0 má statistika F_A Fisherovo-Snedecorovo *F*-rozdělení s $(N - 1)$ a $(N - 1)\,(M - 1)$ stupni volnosti a statistika F_B také *F*-rozdělení s $(M - 1)$ a $(N - 1)(M - 1)$ stupni volnosti. Pokud však vyjde F_T vyšší než odpovídající kvantil *F*-rozdělení, je efekt Tukeyovy interakce významný.

Friedmanův pořadový test. V případě nenormality a heteroskedasticity se aplikuje tento neparametrický test, kdy původní data jsou nahrazena svými pořadími. V experimentu s *N* úrovněmi faktoru *A* v *N* řádcích a *M* úrovněmi faktoru B v *M* sloupcích matice

o rozměru $N \times M$ se užije *Friedmanovo testační kritérium Q* dle vzorce

$$Q = (M-1)\frac{12\sum_{i=1}^{M} R_i^2 - 3N^2 M(M+1)^2}{NM(M^2-1)^2 - \sum_{i=1}^{M}(t^3 - t)},$$

kde data v každém ze N řádků jsou nahražena pořadím. Pořadí jsou sečtena pro každý ze M sloupců. Tato suma pořadí se značí R_i. Faktor t ve jmenovateli testačního kritéria Q představuje počet opakující se jedné hodnoty v průběhu řádku. Když je tento člen nulový, tak se vynechá. Testační kritérium Q má přibližně χ^2 rozdělení s M-1 stupni volnosti. Toto kritérium je blízké *Kendalově koeficientu dobré shody*. U těchto testů *musí* být faktor *A vždy náhodným faktorem* a faktor *B vždy pevným faktorem.*

Vícenásobné porovnávání MCP. Dá se použít jenom pro pevné faktory. Plánovaná porovnávání se formulují v pojmech sloupcových průměrů dle vzorce

$$C_i = \sum_{j=1}^{M} w_{ij}\, m_j ,$$

kde M značí počet úrovní faktoru, m_j jsou průměry pro každou hladinu faktoru a w_{ij} představuje soubor M vah pro i-té porovnání. Porovnávací hodnota C_i se testuje pomocí t-testu. Všimněte si, že jestliže váha w_{ij} nabývá nuly při sumaci přes všechna j, porovnání budeme nazývat *kontrast průměrů.*

Porovnání může být zadáno jednoduše pomocí vah. Například, uvažujme faktor o 3 úrovních, kde 1. sloupec představuje *úroveň kontrolní*, druhý a třetí sloupec obsahují 2. a 3. úroveň faktoru. Porovnání zadáme pomocí vah: porovnání kontrolního 1. sloupce s 2. úrovní faktoru: -1, 1, 0. Porovnání kontrolního 1. sloupce se 3. úrovní faktoru: -1, 0, 1. Porovnání 2. úrovně s 3. úrovní: 0, -1, 1. Porovnání kontrolního 1. sloupce s průměrem 2. a 3. úrovně: -2, 1, 1.

Postup dvoufaktorové analýzy rozptylu bez opakování (ANOVA2P)

Pro dvoufaktorovou analýzu rozptylu a modelu s pevnými efekty v případě $n_{ij} = 1$, tedy bez opakování v cele, se předpokládá možnost interakce Tukeyova typu. Provádí se odhady parametrů, testy významnosti a ověření interakce, resp. transformace, vedoucí k aditivitě efektů. Vstupem je obdélníková tabulka dat pro $A_1, ..., A_N$ úrovní faktoru A (řádky) a $B_1, ..., B_M$ úrovní faktoru B (sloupce), $\{y_{ij}\}$, $j = 1, ..., N$ a $j = 1, ..., M$. Pro všechny testy je standardně uvažována hladina významnosti $\alpha = 0.05$. Postup obsahuje stejné kroky jako postup jednofaktorové analýzy rozptylu:

1. Příprava dat:

(a) Velikost výběru.

(b) Chybějící hodnoty.

(c) Typ dat.

(d) Odlehlé hodnoty.

2. Průměry a efekty úrovní: jsou vypočteny odhady parametrů: celkový průměr $\hat{\mu}$, řádkové efekty α_i, sloupcové efekty β_j, interakční člen τ_{ij} a Tukeyho konstanta C.

3. ANOVA tabulka: je sestavena tabulka ANOVA a provedeny testy významnosti faktorů *A*, *B* a *AB*.

(a) Za předpokladu normality a homoskedasticity: *F*-testy významnosti faktorů, resp. interakcí, včetně kombinovaných testů pro ověření celkové významnosti faktorů *A*, *B*.

(b) Za předpokladu nenormality nebo heteroskedasticity: Friedmanův pořadový neparametrický test.

4. Graf neaditivity: je kreslen graf neaditivity včetně určení optimální mocninné transformace $\hat{\lambda}$ pro zajištění aditivity. Lze zadat provedení analýzy pro transformovaná data, pokud jsou kladná.

Vzorová úloha 5.3 *Dvoufaktorová analýza rozptylu bez opakovaní*

Úloha **H5.02** *Vliv druhu svářecího kovu na pevnost svaru.* Vazebným pojítkem svaru zirkoniové slitiny bývá nikl, železo a měď. Byly vytvořeny svary se třemi typy svářecích drátů a cílem je vyšetřit pevnost svaru, tzn. největší tlak v tisících liber na čtvereční palec, nutný k přerušení svaru. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda záleží na druhu kovu (faktor *A* značený *H502F1*) ve svářecím drátu nebo na sedmi rozličných svárech (faktor *B* značený *H502F2*), zda tlaky k přerušení rozličných svarů jsou u všech drátů stejné. Prozkoumejte, zda je třeba provést transformaci vedoucí ke stabilizaci rozptylu.

Řešení:

1. Průměry a efekty úrovní: jsou vypočteny odhady celkového průměru $\hat{\mu}$, řádkových efektů $\hat{\alpha}_i$, sloupcových efektů $\hat{\beta}_j$, interakčního členu $\hat{\tau}_{ij}$ a konstanty *C*.

Průměry a efekty (NCSS2000):

Zdroj	**Počet**	**Průměr**	**Efekt**
Vše	21	72.39524	72.39524
***A*:** *H502f1*			
1	7	71.1	-1.295238
2	7	75.9	3.504762
3	7	70.18571	-2.209524
B: *H502f2*			
1	3	70.36667	-2.028571
2	3	67.56667	-4.828571
3	3	77.7	5.304762
4	3	72.7	0.3047619
5	3	73.5	1.104762
6	3	68.43333	-3.961905
7	3	76.5	4.104762
***AB*:** *H502f1, H502f2*			
1,1	1	67	-2.071429
1,2	1	67.5	1.228571
1,3	1	76	-0.4047619
1,4	1	72.7	1.295238
1,5	1	73.1	0.8952381
1,6	1	65.8	-1.338095
1,7	1	75.6	0.3952381
2,1	1	71.9	-1.971429
2,2	1	68.8	-2.271429
2,3	1	82.6	1.395238

2,4	1	78.1	1.895238
2,5	1	74.2	-2.804762
2,6	1	70.8	-1.138095
2,7	1	84.9	4.895238
3,1	1	72.2	4.042857
3,2	1	66.4	1.042857
3,3	1	74.5	-0.9904762
3,4	1	67.3	-3.190476
3,5	1	73.2	1.909524
3,6	1	68.7	2.476191
3,7	1	69	-5.290476

2. Tabulka ANOVA: je sestavena tabulka ANOVA a provedeny *F*-testy významnosti faktorů, resp. interakcí, včetně kombinovaných testů k ověření celkové významnosti faktorů *A* a *B*.

ANOVA MODEL I: Očekávaná suma čtverců pro vyvážená data (NCSS2000):

Zdroj rozptylu	**SV**	**Faktor pevný?**	**Faktor ve jmenovateli**	**Očekávaná suma čtverců**
A: *H502f1*	2	Ne	S	S+bsA
B: *H502f2*	6	Ano	*AB*	S+sAB+asB
AB	12	Ne	S	S+sAB
S	0	Ne		S

ANOVA tabulka:

Zdroj rozptylu	**SV**	**Suma čtverců**	**Průměrný čtverec**	**F-test**	**Spočtená hladina významnosti**	**Síla (α = 0.05)**
A: *H502f1*	2	131.901	65.95048			
B: *H502f2*	6	268.2895	44.71492	4.31	0.015087*	0.822970
AB	12	124.459	10.37159			
S	0	0				
Total (Adjust.)	20	524.6495				
Total	21					

* Faktor je významný při α = 0.05

ANOVA model I: jelikož hodnota testačního kritéria 4.31 je vyšší než kvantil Fisherova-Snedecorova rozdělení 3.095, je nulová hypotéza o nevýznamnosti faktoru *A* (druh sváru) zamítnuta a druh svaru je statisticky významným faktorem.

ANOVA MODEL II: Očekávaná suma čtverců pro vyvážená data (NCSS2000)

Zdroj rozptylu	**SV**	**Faktor pevný?**	**Faktor ve jmenovateli**	**Očekávaná suma čtverců**
A: *H502f1*	2	Ne	*AB*	*S+sAB+bsA*
B: *H502f2*	6	Ne	*AB*	*S+sAB+asB*
AB	12	Ne	*S*	*S+sAB*
S	0	Ne	*S*	

ANOVA tabulka:

Zdroj rozptylu	**SV**	**Suma čtverců**	**Průměrný čtverec**	***F*-test**	**Spočtená hladina významnosti**	**Síla (α = 0.05)**
A: *H502f1*	2	131.901	65.95048	6.36	0.013094*	
B: *H502f2*	6	268.2895	44.71492	4.31	0.015087*	
AB	12	124.459	10.37159	3.49		
S	0	0				

Total (Adjust.)	20	524.6495
Total	21	
*Faktor je významný při α = 0.05		

ANOVA model II: jelikož je hodnota druhého testačního kritéria 6.36 vyšší než kvantil Fisherova-Snedecorova rozdělení 3.982, je nulová hypotéza o nevýznamnosti faktoru *B* (druh kovu svářecího drátu) zamítnuta a tento kov je statisticky významným faktorem. Interakce má fyzikální význam, a proto ji budeme testovat: jelikož hodnota testačního kritéria 3.493 je nižší než kvantil Fisherova-Snedecorova rozdělení 4.844, je nulová hypotéza o nevýznamnosti interakčního členu *AB* (interakce druhu sváru s kovem svářecího drátu) přijata a interakce je statisticky nevýznamná. Jelikož odhad mocninné transformace -10.086 leží v akceptovatelném intervalu -15.766 až -4.4066, není třeba data transformovat mocninnou nebo Boxovu-Coxovou transformací.

Při porušení předpokladu o normalitě užijeme pořadový Friedmanův test.

Pořadí úrovní:

Úroveň faktoru *H502f2*	**Počet bloků**	**Medián**	**Průměr pořadí**	**Suma pořadí**
1	3	71.9	3.333333	10
2	3	67.5	1.666667	5
3	3	76	6.666667	20
4	3	72.7	3.666667	11
5	3	73.2	5	15
6	3	68.7	2	6
7	3	75.6	5.666667	17

Friedmanův test:

Vazby	**Friedmanův test. kritérium (Q)**	**SV**	**Spočtená hladina významnosti α**	**Test dobré shody (W)**
Ignorovány	13.428571	6	0.036713	0.746032
Uvažovány	13.428571	6	0.036713	0.746032
Multiplicita	0			

Poznámka: i když vyšel vliv faktoru *A* jako statisticky významný, demonstrujeme si další postup za předpokladu jeho znáhodnění.

Úroveň faktoru *B*: drženého jako pevný nebo náhodný faktor. **Počet bloků:** počet řádků náhodného faktoru *A*. **Medián:** hodnota mediánu pro odpovídající řádek. **Průměr pořadí:** průměr pořadí pro tuto hladinu faktoru. **Suma pořadí:** suma pořadí pro tuto hladinu pořadí. **Vazby:** vazby. **Ignorovány:** statistiky řádku pro neuvažované vazby. **Uvažovány:** statistiky řádku pro uvažované vazby. **Friedmanův test (Q):** hodnota Friedmanovy testační statistiky *Q*. **SV:** *K*-1 stupňů volnosti. **Spočtená hladina významnosti α:** pro testační kritérium *Q*. Je-li hodnota *menší* než zadané α = 0.05, nulová hypotéza o stejných mediánech je zamítnuta. **Test dobré shody (W):** hodnota Kendallova koeficientu dobré shody je mírou shody mezi prvky výběru. Nabývá hodnot mezi 0 a 1. Je-li blízká 1, indikuje perfektní shodu. Nula indikuje naprostou neshodu. **Korekční faktor:** pro vazebné podmínky.

3. Rozličné metody vícenásobné porovnávání MCP: jsou testovány lineární kontrasty pro zadané kombinace úrovní, H_0: $\mu_i - \mu_j = 0$ pro ANOVA MODEL I.

Následující metody vícenásobného porovnávání mají vesměs stejný výstup: **Alfa:** značí zvolenou hladinu významnosti. **DF:** stupně volnosti. **MSE:** hodnota průměrného čtverce chyb. **Kritická hodnota:** tabulkový kvantil pro dané stuně volnosti a hladinu významnosti. **Sloupec:** sloupec čili úroveň faktoru *B* (značeného *H502F2*). **Počet:** počet pozorování ve sloupcovém průměru. **Průměr:** hodnota aritmetického průměru. **Liší se od sloupců:** seznam sloupců, které se liší od sloupce citovaného v dotyčném řádku. Všechny sloupce zde nevyjmenované jsou statisticky nevýznamně odlišné od dotyčného sloupce.

A. Bonferroniho porovnání všech párů: test má odhalit, které páry se liší.

Bonferroniho porovnání všech párů

Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Kritická hodnota**=3.833,

Sloupec	Počet	Průměr	Liší se od sloupců
2	3	67.56667	3
6	3	68.43333	
1	3	70.36667	
4	3	72.7	
5	3	73.5	
7	3	76.5	
3	3	77.7	2

Test indikuje, že 2. a 3. sloupec faktoru *B* se statisticky významně odlišují.

B. Bonferroniho porovnání sloupců vůči kontrolnímu: jestliže jeden sloupec bude představovat *kontrolní sloupec* a všechny ostatní sloupce budeme porovnávat s kontrolním, půjde o *K*-1 porovnání.

Bonferroniho porovnání sloupců vůči kontrolnímu
Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Kritická hodnota**=3.153,

Sloupec	Počet	Průměr	Liší se od sloupců
2	3	67.56667	7, 3
6	3	68.43333	3
1	3	70.36667	
4	3	72.7	
5	3	73.5	
7	3	76.5	2
3	3	77.7	2, 6

Test indikuje, že od 2. se liší 3. a 7. sloupec, od 6. se liší 3. sloupec.

C. Scheffeho vícenásobné porovnání:

Scheffeho vícenásobné porovnání
Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Kritická hodnota**=4.240,

Sloupec	Počet	Průměr	Liší se od sloupců
2	3	67.56667	
6	3	68.43333	
1	3	70.36667	
4	3	72.7	
5	3	73.5	
7	3	76.5	
3	3	77.7	

Test neindikuje odlišné sloupce.

D. Plánované standardní porovnání: při plánovaném porovnání **B1** až **B6** se při standardní volbě porovnává vždy jeden konkrétní sloupec se všemi ostatními vyššími. U volby **B1** se porovnává pár 1. sloupce s 2., 3., 4., 5., 6. a 7. sloupcem, u volby **B2** pak pár 2. sloupce s 3., 4., 5., 6. a 7. sloupcem atd. Testování těchto všech párů sloupců pak tvoří celek nulové hypotézy, který může být přijat nebo zamítnut v návěští **Závěr testu(0.05).**

Plánované porovnání: B1
Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Porovnávaná hodnota**=14.2, **t-test**=1.178, **Prob>|t|**=0.261, **Závěr testu(0.05)**=Přijato, **Standardní chyba porovnávané hodnoty**=12.050,

Sloupec	Váhový koeficient	Počet	Průměr
1	-6	3	70.36667
2	1	3	67.56667
3	1	3	77.7
4	1	3	72.7
5	1	3	73.5
6	1	3	68.43333
7	1	3	76.5

U volby **B1** je nulová hypotéza o shodnosti testovaných párů sloupcových průměrů *přijata*.

Plánované porovnání: B2
Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Porovnávaná hodnota**=31, **t-test**=3.044, **Prob>|t|**=0.0102, **Závěr testu(0.05)**=Zamítnuto, **Standardní chyba porovnávané hodnoty**=10.184,

Sloupec	Váhový koeficient	Počet	Průměr
1	0	3	70.36667
2	-5	3	67.56667
3	1	3	77.7
4	1	3	72.7
5	1	3	73.5
6	1	3	68.43333
7	1	3	76.5

U volby **B2** je nulová hypotéza o shodnosti testovaných párů sloupcových průměrů *zamítnuta*.

Plánované porovnání: B3
Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Porovnávaná hodnota**=-19.667, **t-test**=2.365, **Prob>|t|**=0.0357, **Závěr testu(0.05)**= Zamítnuto, **Standardní chyba porovnávané hodnoty**=8.315,

Sloupec	Váhový koeficient	Počet	Průměr
1	0	3	70.36667
2	0	3	67.56667
3	-4	3	77.7
4	1	3	72.7
5	1	3	73.5
6	1	3	68.43333
7	1	3	76.5

U volby **B3** je nulová hypotéza o shodnosti testovaných párů sloupcových průměrů *zamítnuta*.

Plánované porovnání: B4
Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Porovnávaná hodnota**=0.333, **t-test**=0.0517, **Prob>|t|**=0.9596, **Závěr testu(0.05)**=Přijato, **Standardní chyba porovnávané hodnoty**=6.441,

Sloupec	Váhový koeficient	Počet	Průměr
1	0	3	70.36667

2	0	3	67.56667
3	0	3	77.7
4	-3	3	72.7
5	1	3	73.5
6	1	3	68.43333
7	1	3	76.5

U volby **B4** je nulová hypotéza o shodnosti testovaných párů sloupcových průměrů *přijata*.

Plánované porovnání: B5
Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Porovnávaná hodnota**=-2.067, **t-test**=0.4537, **Prob>|t|**=0.658, **Závěr testu(0.05)**=Přijato, **Standardní chyba porovnávané hodnoty**=4.554,

Sloupec	**Váhový koeficient**	**Počet**	**Průměr**
1	0	3	70.36667
2	0	3	67.56667
3	0	3	77.7
4	0	3	72.7
5	-2	3	73.5
6	1	3	68.43333
7	1	3	76.5

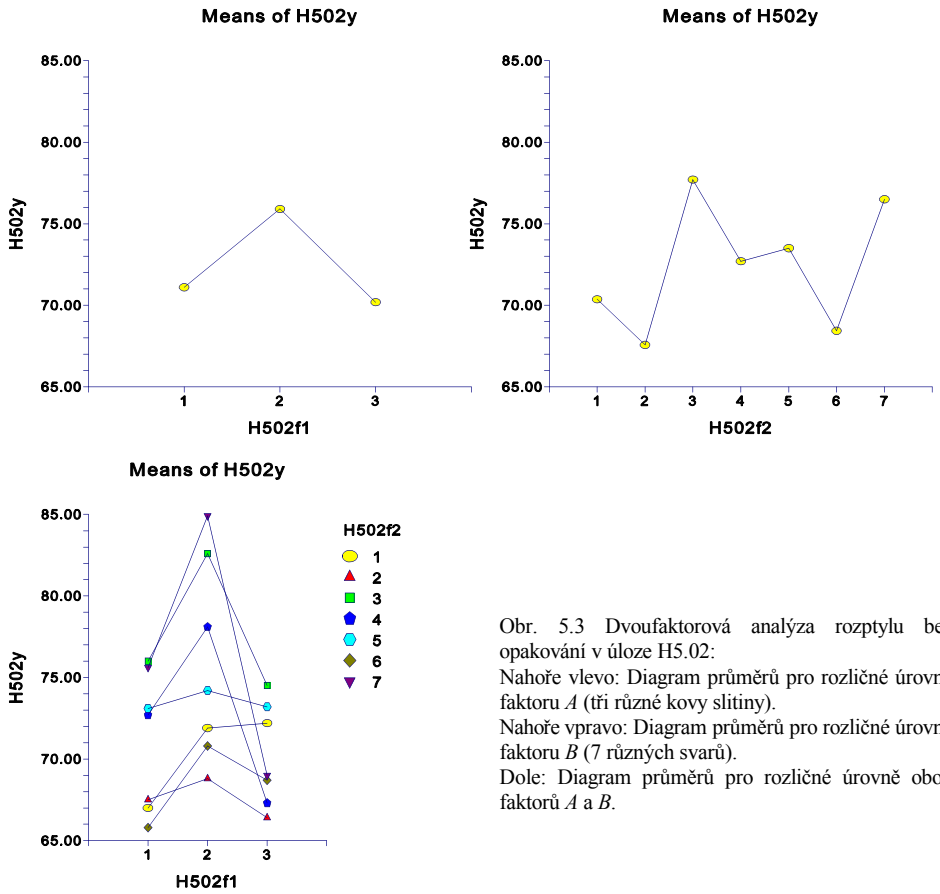
U volby **B5** je nulová hypotéza o shodnosti testovaných párů sloupcových průměrů *přijata.*

Plánované porovnání: B6
Odezva: *H502y*, **Faktor *B*:** *H502f2*, **Alfa**=0.050, **SV**=12, **MSE**=10.372, **Porovnávaná hodnota**=8.067, **t-test**=3.067, **Prob>|t|**=0.0098, **Závěr testu(0.05)**=Zamítnuto, **Standardní chyba porovnávané hodnoty**=2.630,

Sloupec	**Váhový koeficient**	**Počet**	**Průměr**
1	0	3	70.36667
2	0	3	67.56667
3	0	3	77.7
4	0	3	72.7
5	0	3	73.5
6	-1	3	68.43333
7	1	3	76.5

U volby **B6** je nulová hypotéza o shodnosti testovaných párů sloupcových průměrů *zamítnuta*.

4. Grafy a diagramy:



Obr. 5.3 Dvoufaktorová analýza rozptylu bez opakování v úloze H5.02:
Nahoře vlevo: Diagram průměrů pro rozličné úrovně faktoru *A* (tři různé kovy slitiny).
Nahoře vpravo: Diagram průměrů pro rozličné úrovně faktoru *B* (7 různých svarů).
Dole: Diagram průměrů pro rozličné úrovně obou faktorů *A* a *B*.

5.3 Vyvážená dvoufaktorová analýza rozptylu

Slouží ke dvoufaktorové analýze rozptylu u vyvážených experimentů $n_{ij} = n$ a modelů s pevnými efekty. Je hledán optimální model ANOVA, odhadnuty jeho parametry a provedeny testy významnosti. Vstupem jsou pro úrovně $A_1, ..., A_N$ faktoru A a úrovně $B_1, ..., B_M$ faktoru B hodnoty $\{y_{ijk}\}$, $i = 1, ..., N$, $j = 1, ..., M$ a $k = 1, ..., n$. Pro všechny testy je standardně uvažována hladina významnosti $\alpha = 0.05$. Pro výpočet se užívá ANOVA2B (ADSTAT).

Pro tyto modely platí, že v každé cele je $n_{ij} = n$ pozorování. Odhadem μ_{ij} jsou aritmetické průměry

$$\hat{\mu}_{ij} = \frac{1}{n} \sum_{k=1}^{n} y_{ijk} .$$

Pro odhady ostatních parametrů se použijí vztahy

$$\hat{\mu} = \frac{1}{N M} \sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\mu}_{ij}, \quad \alpha_i = \frac{1}{M} \sum_{j=1}^{M} \hat{\mu}_{ij} - \hat{\mu}, \quad \hat{\beta}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_{ij} - \hat{\mu},$$

Rezidua vyjádříme vztahem $\hat{e}_{ijk} = y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. Podobně lze i v tomto případě definovat odhad interakcí $\tau_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. Povšimněme si, že tento vztah se liší od předešlé rovnice jen tím, že se místo veličin y_{ij} používá průměrů $\hat{\mu}_{ij}$. Pro ověření Tukeyova modelu interakce neaditivity lze vynášet graf τ_{ij} vs. $\hat{\alpha}_i \hat{\beta}_j$. Náhodný obrazec zde svědčí o aditivním působení obou faktorů. Součty čtverců modelu analýzy rozptylu pro obecný případ interakcí jsou uvedeny v tabulce. Odpovídající střední hodnoty (očekávané hodnoty) průměrných čtverců jsou

$$E(M_A) = \sigma^2 + \frac{n M \sum_{i=1}^{N} \alpha_i^2}{(N - 1)\sigma^2} = \sigma^2 + n M \sigma_A^2,$$

$$E(M_B) = \sigma^2 + \frac{n N \sum_{j=1}^{M} \beta_j^2}{(M - 1)\sigma^2} = \sigma^2 + n N \sigma_B^2$$

a

$$E(M_{AB}) = \sigma^2 + \frac{n \sum_{i=1}^{N} \sum_{j=1}^{M} \tau_{ij}^2}{(N - 1)(M - 1)\sigma^2} = \sigma^2 + n \sigma_{AB}^2 .$$

Očekávaná hodnota $E(M_R) = \sigma^2$ ukazuje, že roztyl M_R je nevychýleným odhadem σ^2 rozptylu chyb. Rozptyly σ_A^2, σ_B^2 a σ_{AB}^2 odpovídají efektům řádků, sloupců a interakcí. Těchto vztahů lze využít i v případech, kdy se hledají odhady rozptylů příslušející faktorům a interakcím. Pak se místo středních hodnot *E(.)* dosazují přímo průměrné čtverce a místo rozptylu σ^2 reziduální rozptyl σ^2. Důležité je, že průměrné čtverce nejsou přímo odhady odpovídajících rozptylů.

Také v případě analýzy rozptylu, definované ANOVA tabulkou se využitím statistik F_{AB}, F_B a F_A testuje, zda je možné považovat sloupcové a řádkové efekty, resp. interakce, za nevýznamné. Pro test nulové hypotézy H_0: $\tau_{ij} = 0$, $i = 1, ..., N$ a $j = 1, ..., M$, lze použít testační statistiku F_{AB}, která má za předpokladu platnosti hypotézy H_0 F-rozdělení s $\{(N - 1)(M - 1)\}$ a $\{M N (n - 1)\}$ stupni volnosti. Při testování významnosti řádkových efektů faktoru A je H_0: $\alpha_i = 0$, $i = 1, ..., N$. Pokud nulová hypotéza platí, má testační F_A statistika F-rozdělení s $(N - 1)$ a $\{M N(n - 1)\}$ stupni volnosti. Analogicky při testování významnosti sloupcových efektů faktoru B je H_0: $\beta_j = 0$, $j = 1, ..., M$. Pokud nulová hypotéza platí, má testační F_B statistika F-rozdělení s $(M - 1)$ a $\{M N (n - 1)\}$ stupni volnosti. Nevychýleným odhadem rozptylu je zde M_R.

Analýza rozptylu pro dvojné třídění a vyvážený experiment

Součet čtverců pro	Stupně volnosti	Průměrný čtverec	Kritérium F
Faktor *A*			
$S_A = nM\sum_{i=1}^{N}\hat{\alpha}_i^2$	$N-1$	$M_A = \frac{S_A}{N-1}$	$F_A = \frac{M_A}{M_R}$
Faktor *B*			
$S_B = nN\sum_{j=1}^{M}\hat{\beta}_j^2$	$M-1$	$M_B = \frac{S_B}{M-1}$	$F_B = \frac{M_B}{M_R}$
Interakce *AB*			
$S_{AB} = n\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\tau}_{ij}^2$	$(N-1)(M-1)$	$M_{AB} = \frac{S_{AB}}{(N-1)(M-1)}$	$F_{AB} = \frac{M_{AB}}{M_R}$
Reziduální			
$S_R = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{n}(y_{ijk}-\hat{\mu}_{ij})^2$	$MN(n-1)$	$M_R = \frac{S_R}{MN(n-1)}$	-
Celkový			
$S_C = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{n}(y_{ijk}-\hat{\mu})^2$	$MNn-1$	-	-

Výhodou vyvážených experimentů je to, že jednotlivé složky modelů analýzy rozptylu jsou vzájemně nezávislé.

Postup vyvážené dvoufaktorové analýzy rozptylu (ANOVA2B)

Slouží ke dvoustupňové analýze rozptylu u vyvážených experimentů $n_{ij} = n$ a modelů s pevnými efekty. Je hledán optimální model ANOVA, odhadnuty jeho parametry a provedeny testy významnosti. Vstupem jsou pro úrovně A_1, ..., A_N faktoru A a úrovně B_1, ..., B_M faktoru B hodnoty $\{y_{ijk}\}$, $i = 1, ..., N$, $j = 1, ..., M$ a $k = 1, ..., n$. Pro všechny testy je standardně uvažována hladina významnosti $\alpha = 0.05$.

Postup obsahuje stejné kroky jako postup jednofaktorové analýzy rozptylu:

1. Příprava dat:

(a) Velikost výběru.

(b) Chybějící hodnoty.

(c) Typ dat.

(d) Odlehlé hodnoty.

2. Ověření výběrových předpokladů: z opakování v celách

(a) Náhodnost.

(b) Nezávislost.

(c) Normalita.

(d) Homoskedasticita.

3. Průměry a efekty úrovní: jsou vypočteny odhady: celkový průměr $\hat{\mu}$, řádkové efekty

α_i, sloupcové efekty $\bar{\beta}_j$, interakční člen $\bar{\tau}_{ij}$ a Tukeyho konstanta *C*.

4. ANOVA tabulka: je sestavena tabulka ANOVA a provedeny testy významnosti faktorů *A*, *B* a *AB*.

(a) Za předpokladu normality a homoskedasticity: *F*-testy významnosti faktorů, resp. interakcí, včetně kombinovaných testů pro ověření celkové významnosti faktorů *A*, *B*.

(b) Za předpokladu nenormality nebo heteroskedasticity: Friedmanův pořadový test.

5. Grafy a diagramy: je kreslena závislost výběrových směrodatných odchylek s_{ij} v celách na průměrech $\bar{\mu}_{ij}$. Pokud je nalezena monotónní závislost, lze zadat vhodnou transformaci, ve které se provede opakovaná analýza. Je konstruován rankitový graf pro rezidua $\bar{e}_{ijk}$.

Vzorová úloha 5.4 *Vyvážená dvoufaktorová analýza rozptylu*

Na úloze **H5.21** *Vliv teploty výpalu a druhu keramické suroviny na ztrátu hmotnosti pálením* ukážeme aplikaci vyvážené dvoufaktorové analýzy rozptylu s pevnými efekty se stejným počtem pozorování. Při čtyřech teplotách A1 až A4 tj. 950 EC, 1000 EC, 1050 EC a 1100 E C (faktor *A*) byly vypalovány tři typy surovin B1 až B3 (faktor *B*). Na výpalcích bylo provedeno zjištění ztráty hmotnosti pálením v procentech. Stanovení ztráty hmotnosti bylo provedeno u všech tří surovin pro každou teplotu, a to třikrát $n = 3$. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda má teplota výpalu A1 až A4 a typ suroviny B1 až B3 vliv na ztrátu hmotnosti pálením.

Řešení: Použijeme vyvážené dvoufaktorové analýzy rozptylu s pevnými efekty se stejným počtem pozorování ANOVA2B program ADSTAT pro hladinu významnosti $\alpha = 0.050$, počet úrovní faktoru *A*, $N = 4$, počet úrovní faktoru *B*, $M = 3$, počet opakování v jedné buňce $n = 3$.

1. Průměry a úrovně efektů:

Celkový průměr = 9.8833

Faktor *A*: *H521F1*			**Faktor *B*:** *H521F2*		
Úroveň	**Průměr**	**Efekt**	**Úroveň**	**Průměr**	**Efekt**
1	9.656	-0.2278	1	6.9500	-2.9333
2	9.822	-0.0611	2	10.450	0.5667
3	10.11	0.12778	3	12.250	2.3667
4	10.44	0.16111			

2. ANOVA tabulka pro model s interakcemi faktorů *A*, *B*:

H_0: Efekty faktoru *A* jsou nulové, **H_A:** ... nejsou nulové

Kvantil F(1-alfa, *N*-1, *M N*(*n*-1) = 3.009

H_0: Efekty faktoru *B* jsou nulové, **H_A:** ... nejsou nulové

Kvantil F(1-alfa, *M*-1, *M N*(*n*-1) = 3.403

H_0: Interakce I je nulová, **H_A:** ... není nulová, (zde I znamená efekty interakcí *A* a *B* dohromady)

Kvantil F(1-alfa, (*N*-1)(*M*-1), *N M*(*n*-1)) = 2.508

Zdroj rozptylu	**Stupně volnosti**	**Součet čtverců**	**Průměrný čtverec**	**Testační kritérium**	**Závěr H_0 je**	**Spočtená hlad. výz.**
Mezi úrovněmi *A*, *N*-1	= 3	0.8811	0.2937	0.405	Přijata	0.751
Mezi úrovněmi *B*, *M*-1	= 2	174.32	87.160	120.175	Zamítnuta	0.000
Interakce (*N*-1)(*M*-1)	= 6	0.2222	0.0370	0.051	Přijata	0.999

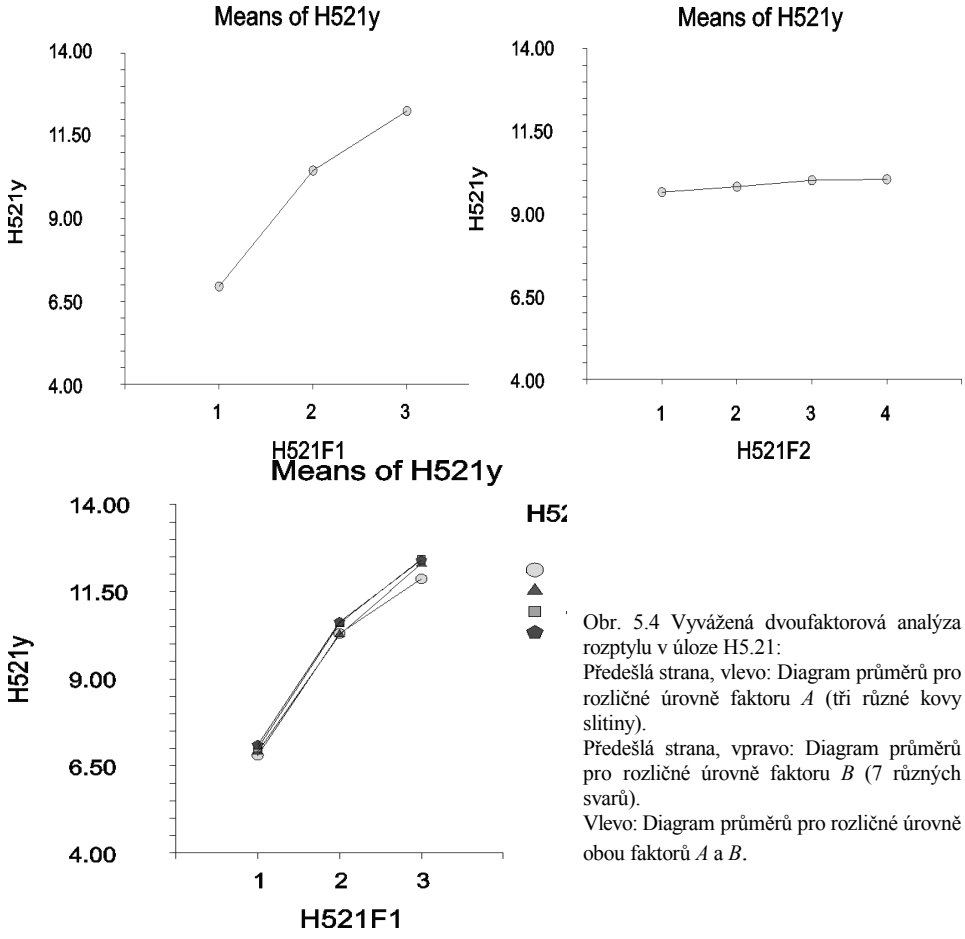
Rezidua $M\,N(n\text{-}1)$	= 18	17.407	0.7253
Celkový $(M\,N\,n\text{-}1)$	= 35	192.83	5.5094

Protože F_e = 0.405 je menší než $F_{1\text{-}0.05}(3, 18)$ = 3.009, je nulová hypotéza přijata a faktor *A* je statisticky nevýznamný. Protože F_e = 120.175 je větší než $F_{1\text{-}0.05}(2, 18)$ = 3.403, je nulová hypotéza zamítnuta a faktor *B* je statisticky významný. Protože F_e = 0.051 je menší než $F_{1\text{-}0.05}(3\ .\ 2, 18)$ = 2.508, je nulová hypotéza přijata a interakce faktor *A* a *B* je statisticky nevýznamná.

3. Zkouška transformace: korelační koeficient, R = 0.491, protože není blízký nule není transformace nutná.

4. Závěr: Na ztrátu hmotnosti pálením má významný vliv pouze typ keramické suroviny.

5. Grafy a diagramy:



Obr. 5.4 Vyvážená dvoufaktorová analýza rozptylu v úloze H5.21:
Předešlá strana, vlevo: Diagram průměrů pro rozličné úrovně faktoru *A* (tři různé kovy slitiny).
Předešlá strana, vpravo: Diagram průměrů pro rozličné úrovně faktoru *B* (7 různých svarů).
Vlevo: Diagram průměrů pro rozličné úrovně obou faktorů *A* a *B*.

5.4 Nevyvážená dvoufaktorová analýza rozptylu

Pro nevyvážené modely platí, že v (i, j)-té cele je n_{ij} pozorování. Pokud je experiment velmi špatně vyvážený, což znamená, že rozdíly mezi jednotlivými hodnotami n_{ij} jsou řádově v desítkách, je analýza rozptylu komplikovanější. Analýza rozptylu se pak provádí s využitím programů pro lineární regresi, kdy se modely ANOVA uvažují jako speciální regresní modely s vysvětlujícími proměnnými, které nabývají pouze hodnot 0 nebo 1.

Pro praktické účely se osvědčuje použití *přibližného rozkladu celkového součtu čtverců*. Začíná se výpočtem průměrů

$$\hat{\mu}_{ij} = \frac{1}{n_k} \sum_{k=1}^{n_k} y_{ijk}$$

pro všechny cely. Z těchto hodnot se dá odhadnout reziduální součet čtverců

$$S_R = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{n_k} (y_{ijk} - \hat{\mu}_{ij})^2 \quad .$$

Pro výpočet dalších složek rozkladu celkového součtu čtverců se používá $\hat{\mu}_{ij}$, o kterých se uvažuje, že jsou určeny z ekvivalentního počtu pozorování n^*, definovaného vztahem

$$n^* = \left[\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{n_{ij}} \right]^{-1} \quad .$$

Analýza rozptylu se pak provádí stejně jako u vyvážených experimentů s tím, že jsou jednotlivé součty čtverců definovány vztahy

$$S_A = n^* M \sum_{i=1}^{N} (\hat{\mu}_i - \hat{\mu})^2 \quad \text{s } (N - 1) \text{ stupni volnosti,}$$

$$S_B = n^* N \sum_{j=1}^{M} (\hat{\mu}_j - \hat{\mu})^2 \quad \text{s } (M - 1) \text{ stupni volnosti}$$

a $S_{AB} = n^* \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{\mu}_{ij} - \hat{\mu}_i - \hat{\mu}_j + \hat{\mu})^2$ s $(N - 1)(M - 1)$ stupni volnosti.

V těchto vztazích je použito označení

$$\hat{\mu}_i = \frac{1}{M} \sum_{j=1}^{M} \hat{\mu}_{ij}, \qquad \hat{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_{ij}, \qquad \hat{\mu} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\mu}_{ij}.$$

Součet $S_A + S_B + S_{AB} + S_R$ zde již není přesně roven S_C, ale rozdíly jsou poměrně malé. Testování hypotéz o řádkových a sloupcových efektech nebo interakcích se provádí stejně jako u vyvážených experimentů.

V případě více opakování v jednotlivých celách lze pro každou z nich určit výběrový rozptyl s_{ij}^2 a pomocí grafu s_{ij}^2 vs. $\hat{\mu}_{ij}$ testovat případnou závislost rozptylu na střední hodnotě (heteroskedasticitu).

Postup nevyvážené dvoufaktorové analýzy rozptylu (ANOVA2U)

Slouží ke dvoufaktorové analýze rozptylu u nevyvážených experimentů n_{ij} a modelů s pevnými efekty. Je hledán optimální model ANOVA, odhadnuty jeho parametry a provedeny testy významnosti. Vstupem jsou pro úrovně $A_1, ..., A_N$ faktoru A a úrovně $B_1, ..., B_M$ faktoru B hodnoty $\{y_{ijk}\}$, $i = 1, ..., N$, $j = 1, ..., M$ a $o = 1, ..., O$. Pro všechny testy je standardně uvažována hladina významnosti $\alpha = 0.05$.

Postup obsahuje stejné kroky jako postup jednofaktorové analýzy rozptylu:

1. Příprava dat:

(a) Velikost výběru.

(b) Chybějící hodnoty.

(c) Typ dat.

(d) Odlehlé hodnoty.

2. Ověření výběrových předpokladů: z opakování v celách

(a) Náhodnost.

(b) Nezávislost.

(c) Normalita.

(d) Homoskedasticita.

3. Průměry a efekty úrovní: jsou vypočteny odhady: celkový průměr $\hat{\mu}$, řádkové efekty α_i, sloupcové efekty β_j, interakční člen τ_{ij} a Tukeyho konstanta C.

4. Tabulka ANOVA: je sestavena tabulka ANOVA a provedeny testy významnosti faktorů A, B a AB.

(a) Za předpokladu normality a homoskedasticity: F-testy významnosti faktorů, resp. interakcí, včetně kombinovaných testů pro ověření celkové významnosti faktorů A, B.

(b) Za předpokladu nenormality nebo heteroskedasticity: Friedmanův pořadový neparametrický test.

5. Grafy a diagramy: je kreslena závislost výběrových směrodatných odchylek s_{ij} v celách na průměrech $\bar{\mu}_{ij}$. Pokud je nalezena monotónní závislost, lze zadat vhodnou transformaci, ve které se provede opakovaná analýza. Je konstruován rankitový graf pro rezidua $\bar{e}_{ijk}$.

Vzorová úloha 5.5 *Nevyvážená dvoufaktorová analýza rozptylu*

Na úloze **E5.05** *Porovnání stanovení arzeniku v pěti laboratořích* dvěma metodami ukážeme aplikaci nevyvážené dvoufaktorové analýzy rozptylu s pevnými efekty a s nestejným počtem pozorování. Arzenik lze v potravě stanovit reakcí s molybdenovou solucí A1 a reakcí diethyldithiokarbamátem stříbrným dle Vašáka a Šedivce A2. Ke vzorku potravy bylo přidáno 15 g arzeniku a vzorek byl analyzován oběma metodami (faktor A) v pěti laboratořích B1 až B5 (faktor B) s vícenásobnou reprodukovatelností. Vedou obě

metody ve všech laboratořích ke stejným výsledkům? Existuje statisticky významná interakce mezi analytickou metodou a laboratoří?

Řešení: Použijeme nevyvážené dvoufaktorové analýzy rozptylu s pevnými efekty s nestejným počtem pozorování ANOVA2U programem ADSTAT pro hladinu významnosti $\alpha = 0.050$. Počet úrovní faktoru A, $i = 1, ..., 5$. Počet úrovní faktoru B, $j = 1, ..., 2$. Nestejný počtem opakování o v každé jedné buňce.

1. Data na různých úrovních faktorů *A* a *B*:

Úrovně faktoru *A* (řádky) index *i*, úrovně faktoru *B* (sloupce) index *j*, opakování v cele index *o*.

i	*j*	*o*	Stanovený obsah arzeniku v gramech:					
1	1	6	12.90	13.20	12.90	12.90	13.10	13.00
1	2	5	13.40	13.00	13.00	17.00	16.60	
2	1	6	14.60	16.20	14.00	15.00	15.50	13.70
2	2	4	14.80	15.20	14.60	15.00		
3	1	6	13.40	13.00	13.20	13.20	13.10	13.20
3	2	5	14.80	14.80	15.00	14.90	14.80	
4	1	6	13.30	13.80	12.50	13.50	13.60	12.80
4	2	6	14.80	14.80	15.00	14.50	15.40	15.20
5	1	6	15.90	14.80	15.30	15.60	14.90	15.20
5	2	5	13.80	14.10	13.80	13.90	14.00	

2. Průměry a úrovně efektů:

Celkový průměr = 1.4278E+01
Reziduální rozptyl = 5.4025E-01

Faktor *A*: *E505F1*

Úroveň	Průměr	Efekt
1	13.80	-0.4780
2	14.86	0.5887
3	14.02	-0.2563
4	14.10	-0.1780
5	14.60	0.3237

Faktor *B*: *E505F2*

Úroveň	Průměr	Efekt
1	13.91	-0.3680
2	14.64	0.36800

3. Tabulka ANOVA s interakcemi faktorů *A*, *B*:

H_0: Efekty faktoru A jsou nulové, **H_A:** ... nejsou nulové
Kvantil F(1-alfa, *K*-1, *M K*(*O*-1) = 2.584

H_0: Efekty faktoru B jsou nulové, **H_A:** ... nejsou nulové
Kvantil F(1-alfa, *M*-1, *M K*(*O*-1) = 4.062

H_0: Interakce I je nulová, **H_A:** ... není nulová, (zde I znamená efekty interakcí A a B dohromady)
Kvantil F(1-alfa, (*K*-1)(*M*-1), *K M*(*O*-1)) = 2.584

Zdroj rozptylu	Stupně volnosti	Součet čtverců hlad.výz.	Průměrný čtverec	Testační kritérium	Závěr H_0 je	Spočtená
Mezi úrovněmi *A* *N*-1	= 4	8.4018	2.1004	3.888	Zamítnuta	0.009
Mezi úrovněmi *B* *M*-1	= 1	7.3202	7.3202	13.550	Zamítnuta	0.001
Interakce (*N*-1)(*M*-1)	= 4	20.04	5.0107	9.275	Zamítnuta	0.000
Rezidua *M N* (*O*-1)	= 44	23.80	0. 54025			
Celkový *M N O*-1	= 53	62.15	1.1715			

Protože $F_e = 3.888$ je větší než $F_{1-0.05}(4, 44) = 2.584$, je nulová hypotéza zamítnuta a faktor A (vliv analytické metody) je statisticky významný. Protože $F_e = 13.550$ je větší než $F_{1-0.05}(1, 44) = 4.062$, je nulová hypotéza zamítnuta a faktor B (vliv laboratoře) je statisticky významný. Protože $F_e = 9.275$ je větší než $F_{1-0.05}(4 . 1, 44) = 2.584$, je nulová hypotéza zamítnuta a interakce faktor A a B je statisticky významná.

4. Závěr: Vliv analytické metody, laboratoře a interakce metody a laboratoře pro stanovení obsahu arzeniku v potravě jsou statisticky významné.

5.5 Opakovatelnost a reprodukovatelnost (O&R analýza)

Populárně zvané *cejchování* se týká ověření přesnosti, zda dotyčná technika měření je co do přesnosti vhodně zvolena, a tím pro experimentální proces přiměřená. Je-li proměnlivost měření malá ve srovnání s proměnlivostí experimentálního procesu říkáme, že postup měření je adekvátní nebo odpovídající. Není-li, je třeba techniku měření zlepšit tak, aby vůbec mohla uspokojivě monitorovat experimentální proces. Jsou-li, například, míry opracovaného výrobku uváděny v toleranci milimetrů, nelze použít techniku měřením měřidlem, které má čtení jenom v centimetrech.

O&R analýza rozděluje celkovou proměnlivost do dvou složek: 1. *Složky měřicí techniky* a 2. *složky procesní,* týkající se vlastního experimentálního procesu. Proměn-livost složky měření je pak dále rozdělena: 1. Do *složky operátora O*, což vlastně představuje *reprodukovatelnost*, a 2. *složky měřicí techniky V*, což je *opakovatelnost*. Je důležité zdůraznit, že O&R analýza se týká jenom přesnosti složky měření. Data pro tuto analýzu pocházejí z experimentu, zvláště postaveného jenom a jenom k tomuto účelu. Není proto možné kombinovat O&R analýzu s ostatními experimenty v laboratoři. Platí *obecné pravidlo:* náhodné chyby měření by neměly být větší než jedna desetina rozptylu procesu. O&R analýza zjišťuje, jaká část pozorovaného rozptylu procesu náleží rozptylu měřicího systému. ANOVA rozděluje ještě reprodukova-telnost na vliv operátora a interakci operátor-vzorek.

Analýzu rozptylu vyšetřovaného experimentálního plánu vystihuje model ANOVA

$$y_{ijk} = \mu + V_i + O_j + (VO)_{ij} + \gamma_{ijk} \quad ,$$

kde $i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$ a V_i, O_j, $(VO)_{ij}$, γ_{ijk} jsou nezávislé, normální, náhodné proměnné se střední hodnotou nula a rozptyly σ_V^2, σ_O^2, σ_{VO}^2 a σ_γ^2. Tyto rozptyly jsou často označovány jako *složky rozptylu* nazývané také *rozptylové komponenty*. V tomto modelu ANOVA shodném s modelem ANOVA pro vyváženou dvoufaktorovou analýzu rozptylu značí písmeno *V náhodný vzorek*, písmeno *O* udává *operátora* a písmeno γ *náhodnou chybu.* Dále v modelu označíme složku rozptylu σ_γ^2 za *opakovatelnost,* dále složku rozptylu $\gamma_1 = \sigma_O^2 + \sigma_{VO}^2$ za *reprodukovatelnost*, složku rozptylu $\gamma_2 = \sigma_O^2 + \sigma_{VO}^2 + \sigma_\gamma^2$ za *celkovou proměnlivost měření*, která se také někdy nazývá O&R hodnota. *Proměnlivost procesu od vzorku k vzorku* představuje další složku rozptylu σ_V^2. Poměr, který porovnává dvě složky rozptylu, a sice proměnlivost experimentálního procesu vůči proměnlivosti samotného měření je dán vzorcem

$$\delta = \frac{\sigma_V^2}{\sigma_O^2 + \sigma_{VO}^2 + \sigma_\gamma^2} \quad .$$

V literatuře je popsána řada kritérií k posouzení O&R hodnot. V automobilovém průmyslu

se užívá kritérium *SNR* (**S**ignal-to-**N**oise **R**atio) *poměr signálu vůči šumu*, vyčíslené vzorcem *SNR* = %δ a dále *rozhodčí kategorie RK* = %(2δ). Existují dvě populární míry k porovnání rozptylu vůči toleranci, v nichž se za toleranci bere rozdíl horní a dolní toleranční meze *HSL - DSL*. Jsou to jednak *chyba měření M* dle vzorce

$$M \text{ ' } 3 \frac{\sqrt{\sigma_O^2 \% \sigma_{VO}^2 \% \sigma_g^2}}{HSL \& DSL} 100\% ,$$

a dále *poměr přesnosti vůči toleranci PT = 2M*. Obě kritéria jsou obvykle vyčíslována spolu se svými intervaly spolehlivosti. Cílem analýzy je vyčíslit tyto hodnoty a rozhodnout, zda padnou do předem určeného intervalu.

Vzorová úloha 5.6 *Schéma O&R analýzy*

Burdik a Larsen vyšetřovali způsobilost monitorování koncentrace kyseliny ve velké nádrži. Z nádrže bylo odebráno 10 vzorků kyseliny, $I = 10$. Náhodným výběrem byli dále vybráni 3 operátoři, $J = 3$, a každý operátor změřil koncentraci kyseliny u každého z 10 vzorků 3krát, $K = 3$, a to za použití vždy stejného laboratorního zařízení. Měření operátorů měla náhodné pořadí. Předpokládá se, že operátoři byli dostatečně zkušení chemici. Vstupní data O&R analýzy byla tvořena 90 hodnotami koncentrace kyseliny. Každý vzorek představuje tři řádky v matici dat, tři sloupce představují tři operátory. Chybějící hodnoty v této analýze nejsou dovoleny.

Data: Data jsou uvedena ve čtveřicích v pořadí: číslo vzorku, 1. operátor, 2. operátor, a 3. operátor.
1 67 66 69, 1 68 68 67, 1 68 68 68, 2 67 67 67, 2 66 67 66, 2 66 68 66, 3 68 70 68, 3 68 70 68,
3 67 68 68, 4 67 70 67, 4 67 68 68, 4 67 70 68, 5 68 70 69, 5 68 70 68, 5 68 70 69, 6 69 71 70,
6 68 70 70, 6 69 70 70, 7 67 68 68, 7 67 68 68, 7 67 68 69, 8 75 75 75, 8 74 75 75, 8 74 75 75,
9 67 69 68, 9 67 68 68, 9 67 69 68, 10 66 68 66, 10 66 66 66, 10 66 66 66.

Řešení: Software NCSS2000

1. Přehled dat: aktuální počet musí být roven očekávanému počtu, jinak by v datech byly díry, což je pro tento druh analýzy nepřípustné.

Položka	**Skutečná četnost**	**Očekávaná četnost**
Celkový počet	90	90
Počet vzorků	10	
Počet operátorů 3		
Počet opakování 3		

Počet vzorků: četnost vzorků. **Počet operátorů**. **Počet opakování:** počet opakování měření operátorem.

2. Očekávaný průměrný čtverec a složky rozptylu: jsou vyčísleny bodový a intervalový odhad od každé složky rozptylu. Jde o vyváženou dvoufaktorovou analýzu rozptylu.

Složka, zdroj rozptylu	**SV**	**Očekávaný průměrný čtverec**	**Složka rozptylu**	**Dolní mez 90% int. spolehlivosti rozptylu komp.**	**Horní mez**
Vzorky (*V*)	9	R+3(VO)+9(V)	5.615638	2.948817	15.33656
Operátoři (*O*)	2	R+3(VO)+30(O)	0.3563786	0.1016096	7.389713
Interakce (*VO*)	18	R+3(PO)	0.1251029	0.0238500	0.3455315
Opakování (*R*)	60	R	0.3444445	0.2613323	0.4785284

Složka, zdroj rozptylu: zdroj rozptylu v datech. **SV**: za stupně volnosti. **Očekávaný průměrný čtverec:** symbolická hodnota pro průměrný čtverec, když se v ANOVA modelu předpokládají vyvážená data. Zde *V* představuje σ_V^2, *O* představuje σ_O^2, *VO* představuje σ_{VO}^2 a *R* představuje σ_g^2. **Složky rozptylu:** σ_V^2, σ_O^2, σ_{VO}^2 a σ_g^2. **Dolní mez** 90% intervalu spolehlivosti dotyčné složky rozptylu; **Horní mez** 90% intervalu spolehlivosti dotyčné

složky rozptylu.

3. ANOVA tabulka: testuje statistickou významnost jednotlivých složek rozptylu.

Složka, zdroj rozptylu	SV	Suma čtverců	Průměrný čtverec	*F*-test	Spočtená hladiná významnosti
Vzorky	9	461.3445	51.26049	71.22	0.000000
Operátoři	2	22.82222	11.41111	15.85	0.000107
Interakce	18	12.95556	0.7197531	2.09	0.017450
opakování	60	20.66667	0.3444445		
Total (Adjust.)	89	517.7889			
Total	90				

Složka, zdroj rozptylu: zdroj rozptylu v modelu. **SV:** stupně volnosti. **Suma čtverců:** pro úplnost se zde uvádí suma čtverců. **Průměrný čtverec:** odhad rozptylu této složky. Jde o sumu čtverců odchylek dělenou patřičným počtem stupňů volnosti. ***F*-test:** poměr průměrného čtverce tohoto zdroje a průměrného čtverce jeho odpovídajícího chybového zdroje. **Spočtená hladina významnosti:** vypočtená α pro *F*-test. Je-li spočtená α menší než 0.05, *F*-test je statisticky významný. Hvězdička u hodnoty *F*-testu pak značí statistickou významnost.

4. Přehled složek, zdrojů rozptylu:

Složka, zdroj rozptylu	Rozptyl	Procento rozptylu	Směrodatná odchylka	Dolní mez 90%ní intervalu spolehliv.	Horní mez spolehliv.	Procento celk. směr. odchylky
Vzorky	5.615638	87.1782	2.3697	1.7172	3.9162	93.3693
Operátoři	0.356379	5.5325	0.5970	0.3188	2.7184	23.5212
Interakce	0.125103	1.9421	0.3537	0.1544	0.5878	13.9360
Reprodukovatelnost	0.481481	7.4746	0.6939	0.4349	2.7415	27.3397
Opakovatelnost	0.344444	5.3472	0.5869	0.5112	0.6918	23.1241
O a *R*	0.825926	12.8218	0.9088	0.7443	2.8044	35.8076
Celková proměnliv.	6.441564	100.0000	2.5380	1.9394	4.2947	100.0000

Složka, zdroj rozptylu: jména složek, zdrojů rozptylů, které se vyčíslují. První čtyři řádky byly vysvětleny předešle. *Vzorek* značí σ_V^2 rozptyl mezi vzorky. *Operátoři* značí σ_O^2 rozptyl mezi operátory. *Interakce* značí σ_{VO}^2 interakční rozptyl. *Opakovatelnost* značí σ_g^2 rozptyl, která se objeví když jeden operátor měří stejný vzorek neustále opakovaně. *Reprodukovatelnost* se týká rozptylu mezi operátory $\gamma_1 = \sigma_O^2 + \sigma_{VO}^2$. R&O analýza se týká *sumy reprodukovatelnosti a opakovatelnosti* $\gamma_2 = \sigma_O^2 + \sigma_{VO}^2 + \sigma_g^2$. *Celkový rozptyl* je přitom tvořena sumou všech čtyř zdrojů rozptylu $\sigma_T^2 = \sigma_V^2 + \sigma_O^2 + \sigma_{VO}^2 + \sigma_g^2$. **Rozptyly:** odhady všech složek rozptylu σ_V^2, σ_O^2, σ_{VO}^2, σ_g^2, γ_1, γ_2, σ_T^2.

Procento z celkového rozptylu: ukazuje na procentuální zastoupení dotyčné složky rozptylu v celkovém rozptylu. **Směrodatná odchylka:** je odmocnina z dotyčné složky rozptylu. **Dolní (a horní) mez 90% intervalu spolehlivosti směrodatné odchylky; Procento z celkové směrodatné odchylky:** ukazuje na procentuální zastoupení dotyčné složky směrodatné odchylky v celkové směrodatné odchylce. Suma jednotlivých složek směrodatné odchylky nemusí nutně dát 100, protože toto pravidlo platí pro rozptyly.

5. Procento procesního rozptylu: dává složky procesního rozptylu násobené hodnotou sigma koeficientu (který má předvolenou hodnotu 5.15). Toto vynásobení převede všechny hodnoty do stejné metriky jako jsou regulační limity, takže mohou být porovnávány přímo. Například, rozptyl, která se objeví, když stejný operátor měří stejný vzorek dvakrát přidá mezi 2.6327 a 3.5626 ke směrodatné odchylce.

Složka, zdroj rozptylu	Dolní mez 90% int. spol.	5.15 směr.odch.	Horní mez 90% int. spol.	Procento rozptylu	Procento příspěvku
Vzorky	8.8436	12.2041	20.1684	93.3693	87.1782
Operátoři	1.6416	3.0744	13.9998	23.5212	5.5325
Interakce	0.7953	1.8215	3.0273	13.9360	1.9421

Reprodukovatelnost	2.2395	3.5735	14.1187	27.3397	7.4746
Opakovatelnost	2.6327	3.0225	3.5626	23.1241	5.3472
O a *R*	3.8332	4.6803	14.4425	35.8076	12.8218
Celková proměnliv.	9.9877	13.0708	22.1175	100.0000	100.0000
Je-li procento *O&R* větší než 30%, měřicí systém je nepřijatelný.					

Reprodukovatelnost(2) značí σ_O^2 rozptyl od operátorů. **Dolní (a horní) mez 90% intervalu spolehlivosti směrodatné odchylky** ve dvou sloupcích. Hodnoty jsou násobeny sigma koeficientem, jak bylo uvedeno výše. **5.15 směrodatná odchylka:** je to odmocnina každé složky rozptylu, násobená sigma koeficientem (předvoleno 5.15). **Procento celkového rozptyl:** je to 100násobek poměru směrodatné odchylky tohoto zdroje ku celkové směrodatné odchylce v posledním řádku. **Procento příspěvku:** je 100násobkem poměru tohoto zdroje ku rozptylu celkovému v posledním řádku.

6. Procento tolerance: je podobné předchozí tabulce, kromě toho, že za jmenovatele v posledních dvou sloupcích je užita *tolerance* místo celkového rozptylu.

Složka, zdroj rozptylu	**Dolní mez 90% int. spol.**	**5.15 směr.odch.**	**Horní mez 90% int. spol.**	**Procento tolerance**
Vzorky	8.8436	12.2041	20.1684	30.5103
Operátoři	1.6416	3.0744	13.9998	7.6860
Interakce	0.7953	1.8215	3.0273	4.5539
Reprodukovatelnost	2.2395	3.5735	14.1187	8.9338
Opakovatelnost	2.6327	3.0225	3.5626	7.5563
O a *R*	3.8332	4.6803	14.4425	11.7009
Celková proměnliv.	9.9877	13.0708	22.1175	32.6771
Horní spec. mez	88			
Dolní spec. mez	48			
Tolerance	40			
Je-li procento *O&R* hodnoty mezi 10% a 20%, měřicí systém je přijatelný.				

7. Testační rozhodčí kritéria: tabulka přináší hodnoty intervalu spolehlivosti čtyř kritérií, užitečných pro *O&R* analýzu. Uživatel rozhodne, zda užije bodový nebo intervalový odhad k vyhodnocení dat. Kritéria jsou založena na hodnotě poměru δ. Jsou to *RK, SNR, M* a *PT*.

Index	**Dolní mez 90% int. spol.**	**Hodnota**	**Horní mez 90% int. spol.**
Rozhodčí kategorie *RK*	1.1924	3.6876	6.2979
SNR	0.8431	2.6075	4.4533
Chyba měření *M* 5.5823	6.8160	21.0328	
Poměr *PT*	11.1647	13.6321	42.0655
Je-li dolní mez intervalu spolehlivosti rozhodčí kategorie *RK* měnší než 3, měřicí proces je neadekvátní.			
Je-li horní mez intervalu spolehlivosti chyby měření *M* menší než 25%, chyba měření může být zanedbána.			

SNR* (poměr signál ku šumu):** *SNR* = %δ vyjadřuje poměr směrodatné odchylky od vzorku ke vzorku vůči rozptylu měření. Jako výrobce se především zajímáme o rozptyl od vzorku ke vzorku. Směrodatná odchylka měření odhaluje šum, který je přidán k rozptylu od vzorku ke vzorku díky přibližné povaze měřicího systému. ***RK* (rozhodčí kategorie)**: *RK* = %(2δ) vyjadřuje počet rozhodčích kategorií produktu, které mohou být spolehlivě rozlišeny měřicím postupem. **Chyba měření *M: porovnává směrodatnou odchylku měření vůči toleranci. Za toleranci se bere rozdíl mezi horní a dolní toleranční mezí. Platí pravidlo: tato hodnota by měla být menší než 25 %, aby byl měřicí systém uznán za přiměřený. **Poměr přesnosti vůči toleranci *PT***: je mírně odlišnou verzí kritéria *M*.

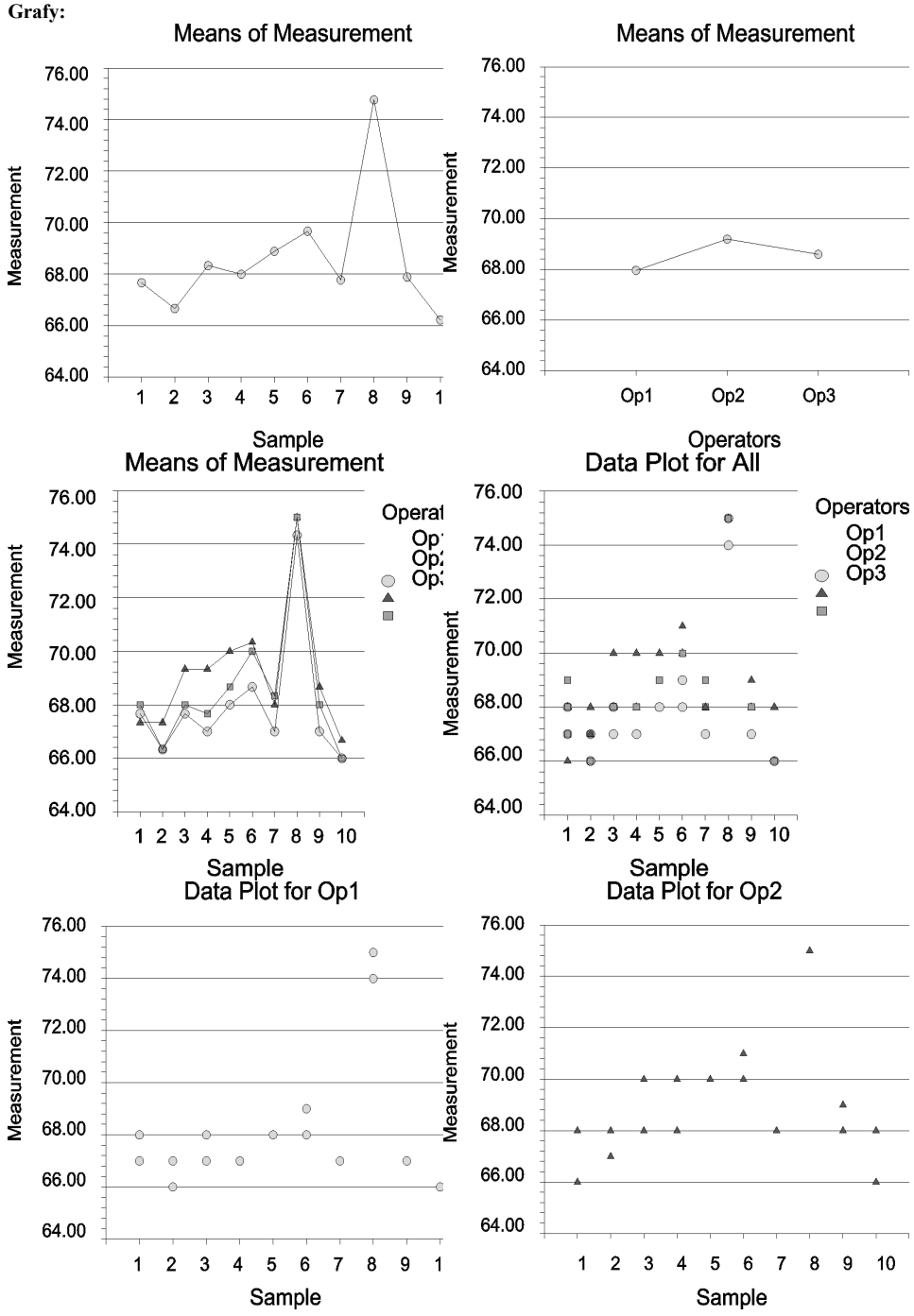
8. Průměry a vychýlení: umožňuje snadno nalézt odlehlé hodnoty.

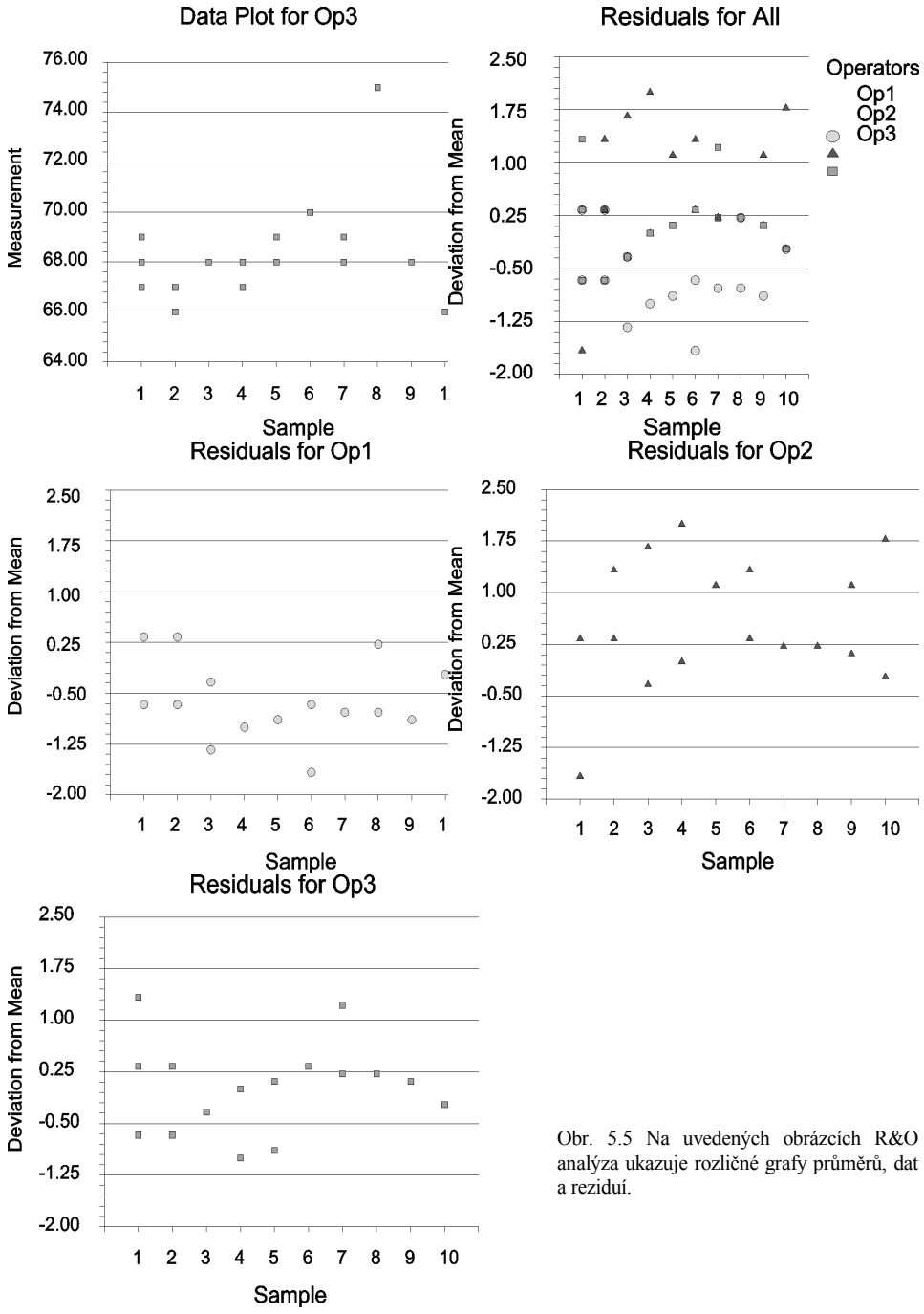
Zdroj rozptylu	**Počet**	**Průměr**	**Odchylka od cíle**

Celkově	90	68.589	0.589
Vzorek			
1	9	67.667	-0.333
2	9	66.667	-1.333
Vzorek			
3	9	68.333	0.333
4	9	68.000	0.000
5	9	68.889	0.889
6	9	69.667	1.667
Vzorek			
7	9	67.778	-0.222
8	9	74.778	6.778
9	9	67.889	-0.111
10	9	66.222	-1.778
Operátoři			
Op1	30	67.967	-0.033
Op2	30	69.200	1.200
Op3	30	68.600	0.600
Vzorek, Operátoři			
1,Op1	3	67.667	-0.333
1,Op2	3	67.333	-0.667
1,Op3	3	68.000	0.000
Vzorek, Operátoři			
2,Op1	3	66.333	-1.667
2,Op2	3	67.333	-0.667
2,Op3	3	66.333	-1.667
3,Op1	3	67.667	-0.333
Vzorek, Operátoři			
3,Op2	3	69.333	1.333
3,Op3	3	68.000	0.000
4,Op1	3	67.000	-1.000
4,Op2	3	69.333	1.333
Vzorek, Operátoři			
4,Op3	3	67.667	-0.333
5,Op1	3	68.000	0.000
5,Op2	3	70.000	2.000
5,Op3	3	68.667	0.667
Vzorek, Operátoři			
6,Op1	3	68.667	0.667
6,Op2	3	70.333	2.333
6,Op3	3	70.000	2.000
7,Op1	3	67.000	-1.000
Vzorek, Operátoři			
7,Op2	3	68.000	0.000
7,Op3	3	68.333	0.333
8,Op1	3	74.333	6.333
8,Op2	3	75.000	7.000
Vzorek, Operátoři			
8,Op3	3	75.000	7.000
9,Op1	3	67.000	-1.000
9,Op2	3	68.667	0.667

9,Op3	3	68.000	0.000

Grafy:





Obr. 5.5 Na uvedených obrázcích R&O analýza ukazuje rozličné grafy průměrů, dat a reziduí.

5.6 Úlohy

Úlohy jsou rozděleny do pěti kapitol: B5 (farmakologická a biochemická data), C5 (chemická a fyzikální data), E5 (environmentální, potravinářská a zemědělská data), H5 (hutní a mineralogická data) a S5 (ekonomická a sociologická data). Vysvětlete jednotlivé diagnostiky a učiňte své závěry o výsledcích analýzy rozptylu.

5.6.1 Analýza farmakologických a biochemických dat

Úloha B5.01 *Vliv analytické laboratoře na stanovení albuminu v lidském séru (ANOVA1)*. Standardní vzorek lidského séra obsahuje 42.0 g albuminu v 1 litru. Pět laboratoří A1 až A5 (faktor *A*) provedlo analytické stanovení při šesti opakovaných analýzách v jednom dni (str. 31 v cit.[14]). Posuďte, zda se výsledky z laboratoří významně liší. Dosáhla některá laboratoř odlehlých výsledků? Komentujte také správnost stanovení v jednotlivých laboratořích pomocí intervalového odhadu.

Data: Obsah určeného albuminu [g. l^{-1}] v pěti laboratořích, A1 až A5.

A1	42.5	41.6	42.1	41.9	41.1	42.2
...	...	...	...	...	...	...
A5	42.2	41.6	42.0	41.8	42.6	39.0

Úloha B5.02 *Porovnání nové metody v sedmi laboratořích (ANOVA1)*

Kirchhoefer[16] navrhl poloautomatickou metodu na stanovení maleátu chlorfeniraminu v tabletách. Sedm analytických laboratoří A1 až A7 (faktor *A*) opakovalo analýzu tablety o deklarovaném obsahu 4 mg celkem 10krát. Cílem je posoudit vliv laboratoří na výsledek analýzy. Dosáhla některá laboratoř silně vybočujících výsledků? Testujte také správnost stanovení v jednotlivých laboratořích pomocí intervalového odhadu.

Data: Obsah určeného maleátu [mg] v sedmi laboratořích, A1 až A7.

A1	A2	A3	A4	A5	A6	A7
4.13	3.86	4.00	3.88	4.02	4.02	4.00
...	...	...	...	...	...	...
4.04	3.95	3.98	3.90	4.00	3.93	4.06

Úloha B5.03 *Vliv doby skladování na stanovený obsah riboflavinu (ANOVA1)*

V rámci stabilitních testů u dražé B-komplexu bylo provedeno stanovení obsahu riboflavinu u jedné výrobní šarže ve třech různých časech (faktor *A*), a to A1 finální výstup po ukončení výroby, dále A2 po jednom měsíci a konečně A3 po půl roce při skladovacích teplotách do 25 EC. Stanovení bylo provedeno vždy stejným postupem metodou HPLC, a to na stejném zařízení. Ověřte, zda doba skladování měla významný vliv na obsah sledovaného riboflavinu. Liší se významně obsah riboflavinu v určitém čase od ostatních? Považujte finální výstup A1 za kontrolní a testujte, zda se zbývající dva časové úseky A2 a A3 liší od kontrolního A1.

Data: Obsah stanoveného riboflavinu [mg/drg] u jedné šarže v různých časových obdobích: A1 Finální výstup, A2 po 1 měsíci, A3 po 6 měsících.

A1	A2	A3
14.80	15.97	14.16
...	...	...
15.42	14.40	15.25

Úloha B5.04 *Vliv druhu biologického materiálu na obsah vanadu (ANOVA1)*

Vanad byl shledán důležitým biogenním prvkem. Byl sledován v biologickém materiálu metodou izotopového zřeďování hmotovou spektrometrií. Data přináší obsah vanadu v ng/g vysušeného biologického materiálu (faktor *A*), a to A1 v tkáni ústřice, A2 v listech citrusů,

A3 v hovězích játrech a A4 v lidském séru. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda se významně liší obsah vanadu dle druhu biologického materiálu nebo zda je všude stejný? Jaký je rozdíl mezi střední hodnotou obsahu vanadu v tkání ústřice a citrusovými listy? Jaká je střední hodnota obsahu vanadu v lidském séru?

Data: Obsah vanadu [ng/g] v rozličném biologickém materiálu: A1 tkáň ústřice, A2 listy citrusů, A3 hovězí játra, A4 lidské sérum.

A1	A2	A3	A4
2.35	2.32	0.39	0.1
...	...	...	...
-	-	-	0.16

Úloha B5.05 *Vliv hladiny inteligence na třídění dětí (ANOVA1)*

Psycholog roztřídil velkou skupinu dětí dle jejich chování, šikovnosti a úrovně inteligenčního kvocientu IQ do sedmi klasifikačních skupin A1 až A7. Otázkou je, zda navržené třídění není příliš zjednodušeno. Vyšetřete, zda toto kritérium (faktor *A*) dostatečně spolehlivě třídí děti a zbývající variabilita v datech je náhodného charakteru. Existuje skupina se silně odlehlými výsledky od ostatních skupin?

Data: IQ [skóre] dětí, roztříděných do sedmi skupin A1 až A7.

A1	A2	A3	A4	A5	A6	A7
105	115	103	124	115	85	79
...	...	...	...	...	...	...
-	-	112	-	-	-	-

Úloha B5.06 *Vliv pěti druhů diety na tělesnou hmotnost mužů (ANOVA1)*

Pět skupin po 4 mužích bylo vystaveno rozličné dietě A1 až A5. Na konci týdne byly vyčísleny kladné a záporné diference hmotnosti mužů po aplikaci týdenní diety. Porovnejte čtyři diety vůči dietě A1, kterou budeme chápat jako dietu kontrolní.

Data: kladné a záporné diference hmotnosti mužů A1 až A5.

A1	A2	A3	A4	A5
3	2	4	3	1
...	...	...	...	...
-2	1	2	1	-1

Úloha B5.07 *Vliv způsobu přípravy pomerančového džusu na obsah vitaminu C (ANOVA1)*. Je porovnáván obsah vitaminu C v mg u tří rozličných způsobů přípravy pomerančového džusu. Z každého způsobu přípravy A1 až A3 (faktor *A*) bylo provedeno pět opakovaných měření obsahu vitaminu. Je obsah vitaminu C závislý na způsobu přípravy? Přináší některý způsob přípravy silně odlehlé výsledky od ostatních?

Data: Obsah stanoveného vitaminu C [mg] u tří způsobů přípravy džusu A1 až A3.

A1	A2	A3
96	123	76
...	...	...
90	122	80

Úloha B5.08 *Vliv lokality lesa na hmotnost chytaných králíků (ANOVA1)*

V různých lokalitách australského lesa A1 až A5 (faktor *A*) bylo v nastražených pastích odchytnuto několik divokých králíků. Králíci dosahovali rozličné hmotnosti, uvedené v librách. Testujte, zda lokalita lesa má vliv na hmotnost chytaných králíků. Liší se významně nějaká lokalita od ostatních? Dosahuje silně odlehlých výsledků?

Data: Hmotnost králíků [libra] v pěti lokalitách lesa A1 až A5.

A1	A2	A3	A4	A5
37	29	49	40	50
...	...	...	...	...
-	-	-	41	-

Úloha B5.09 *Vliv sledované rodiny na nadváhu bratří (ANOVA1)*

Vyšetřete, zda nadváha bratří v 10 sledovaných rodinách A1 až A10 (faktor *A*) je v normě a nepřekračuje významně průměr, či zda existují výjimečné rodiny otylců. Nadváha je uvedena v librách. Existuje rodina, ve které je nadváha silně odlišná od ostatních?

Data: Nadváha bratrů [v librách] v rodině u 10 sledovaných rodin: A1 až A10.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
59	70	67	83	82	96	101	79	85	84
66	87	83	78	95	75	78	79	72	79
83	90	92	77	90	78	66	84	89	83

Úloha B5.10 *Vliv alkoholu a věku na reakční čas řidiče (ANOVA2B)*

Vyšetřete vliv věku řidiče A1 až A3 (faktor *A*) a množství vypitého alkoholu B1 až B3 (faktor *B*) na reakční čas řidiče v sekundách, když každé měření bylo 3× opakováno. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda oba faktory mají významný vliv na reakční čas řidiče a zda existuje významná interakce mezi věkem řidiče a vlivem alkoholu. Vezměte první sloupec reakčního času bez alkoholu B1 za kontrolní a porovnejte s ním zbývající dva sloupce B2 a B3. Jsou zde statisticky významné rozdíly?

Data: Reakční čas [s] v závislosti na věku řidiče A1 až A3 a vlivu alkoholu B1 až B3.

Věk	B1 (Žádný alkohol)	B2 (1 sklenička)	B3 (2 skleničky)
A1 (20 - 39 let)	0.42 0.43 0.41	0.47 0.46 0.46	0.65 0.66 0.68
A2 (40 - 59 let)	0.51 0.53 0.52	0.62 0.63 0.62	0.66 0.68 0.66
A3 (60 a více let)	0.57 0.58 0.57	0.73 0.73 0.72	0.79 0.80 0.80

Úloha B5.11 *Vliv tetrachlormethanu na počet škrkavek v kryse (ANOVA1)*

U 4 experimentálních skupin pokusných krys byl tetrachlormethan CCl_4 užit jako smrtící prostředek na škrkavky. Deset krys bylo infikováno larvami škrkavek a po 8 dnech byl pěti krysám aplikován tetrachlormethan (skupina A1) a zbylých pět bylo ponecháno jako kontrolní vzorek (skupina A2). Po 2 dnech byly krysy usmrceny a určen počet škrkavek. Celý pokus byl 1× opakován, a tak vůči skupině A1 vznikla skupina A3 a vůči skupině A2 skupina A4. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda je mezi čtyřmi skupinami (faktor *A*) významný rozdíl v počtu škrkavek a zda tetrachlormethan je účinný. Porovnejte skupiny mezi sebou.

Data: Počet škrkavek v jedné kryse v závislosti na čtyřech skupinách A1 až A4.

A1	A2	A3	A4
279	378	172	381
...	...	...	...

303	286	250	318

Úloha B5.12 *Vliv LRF faktoru na tvorbu lutenizačního hormonu LH (ANOVA2B)*
Ve studii tvorby lutenizačního hormonu LH u krys byli samci i samice krys drženi při konstantním světle a část samců a samic byli 14 hodin na světle a 10 hodin ve tmě. Krysám bylo podáno rozličné množství stimulátoru tvorby lutenizačního hormonu LRF (faktor *A*), který podporuje uvolnění lutenizačního hormonu. Obsah uvolněného hormonu LH v ng/ml séra byl stanoven po odběru krve. Na hladině významnosti $\alpha = 0.05$ vyšetřete zda (a) množství stimulátoru (faktor *A*) a (b) světlo (faktor *B*) ovlivňuje obsah uvolněného lutenizačního hormonu u všech krys, nebo pouze u samců anebo pouze u samic. Porovnejte obsah uvolněného hormonu LH u samců versus samice a dále denní versus konstantní světlo vždy u obou pohlaví zvlášť.
Data: Obsah uvolněného lutenizačního hormonu [ng/ml].

Množství stimulátoru	B1 (Samci, denní světlo)	B2 (Samci, konst. světlo)	B3 (Samice, denní světlo)	B4 (Samice, konst. světlo)
A1 (LRF, 0 ng)	212 27 68 72 130 153	72 64 78 20 56 70	71 159 208 161 187 51	197 115 28 48 229 424
...	...	...	...	...
A5 (LRF, 1250 ng)	296 545 630 418 396 227	137 426 178 208 196 251	2693 1719 2758 2040 3199 561	1482 1646 1646 1289 1982 1780

Úloha B5.13 *Vliv chleba, otrub a práce laboratoře na obsah vitaminu PP v obilninách (ANOVA3B)*. Při studiu homogenity obsahu vitaminu PP v obilninách byly homogenizované vzorky chleba a otrub obohaceny o 0, 2, 4 a 8 mg vitaminu PP na 100 g. Vzorky byly rozeslány 12 laboratořím, ve kterých byl vitamin PP v mg/100 g stanoven specifickou metodou ve třech po sobě jdoucích dnech. Na hladině významnosti $\alpha = 0.05$ proveďte analýzu rozptylu a vyšetřete, zda existuje vliv laboratoře (faktor *A*), zda existuje rozdíl mezi obsahem vitaminu PP ve chlebě a otrubách (faktor *B*) a dále zda existuje vliv obohacení o rozličná množství vitaminu PP (faktor C). Dá se uvažovat interakce mezi faktory? Je třeba data upravovat odečtením hodnoty obohacení před aplikací analýzy rozptylu?
Data: Stanovený obsah vitaminu PP [mg/100 g] se standardním přídavkem ve 12 laboratořích ve chlebě a v otrubách.

	B1 (Obsah vitaminu PP ve chlebě)			B2 (Obsah vitaminu PP v otrubách)		
Lab.	C1 (+0 mg)	C2 (+2 mg)	C3 (+4 mg)	C1 (+0 mg)	C2 (+4 mg)	C3 (+8 mg)
A1	3.42 3.66 3.26	5.25 5.63 5.25	7.17 7.50 7.25	7.58 7.87 7.71	11.63 11.87 11.40	15.00 15.92 15.58
...	...	...	...	...	...	...
A12	3.76 3.68 3.80	6.06 5.60 6.05	7.60 7.50 7.67	8.32 8.25 8.57	12.00 12.40 12.30	16.80 16.60 16.30

Úloha B5.14 *Vliv druhu anesteze a typu psa na koncentraci adrenalinu v krvi (ANOVA2P)*.
U 10 psů B1 až B10 byla měřena koncentrace adrenalinu pod anestezí plynnými anestetiky A1 isofluoranem, A2 halothanem a A3 cyklopropanem. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda druh anestetika (faktor *A*) ovlivní výslednou koncentraci adrenalinu v ng/ml krve a zda tuto koncentraci ovlivní i typ vyšetřovaného psa (faktor *B*). Lze uvažovat také interakci obou faktorů? Porovnejte koncentraci adrenalinu u jednotlivých psů a nalezněte

odlehlého jedince. Lze nalézt i anestetikum odlehlých vlastností?

Data: Koncentrace adrenalinu v krvi [ng/ml] po 3 anestetikách A1 až A3 pro 10 druhů psů B1 až B10.

	Druh psa									
Anestetikum	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
A1 (Isofluoran)	0.28	0.51	1.00	0.39	0.29	0.36	0.32	0.69	0.17	0.33
A2 (halothan)	0.30	0.39	0.63	0.68	0.38	0.21	0.88	0.39	0.51	0.32
A3 (Cyklopropan)	1.07	1.35	0.69	0.28	1.24	1.53	0.49	0.56	1.02	0.30

Úloha B5.15 *Vliv druhu myši na počet překonaných čtverečků v bludišti (ANOVA1)*
Tři druhy myší A1, A2, A3 byly testovány na "agresivitu" dle svého chování v bludišti. Každá myš byla umístěna do středního čtverečku bludiště, na dno krabice 1×1 m, kde základna krabice byla členěna na 49 stejných čtverců. Myš se pokoušela o únik z tohoto bludiště a musela překonat určitý počet čtverečků za čas 5 minut. Na hladině významnosti $\alpha = 0.05$ vyšetřete, je-li významný rozdíl mezi počtem prošlých čtverečků u tří druhů myší A1 až A3 (faktor *A*). Existuje druh myši silně odlehlý od ostatních?

Data: Počet prošlých čtverečků u tří druhů myší A1 až A3.

A1	309 229 182 228 326 289 231 225 307 281 316 290 318 273 328 325 191 219 216 221 198 181 110 256 240 122 290 253 164 211 215 211 152 178 194 144 95 157 240 146 106 252 266 284 274 285 366 360 237 270 114 176 224
A2	37 90 39 104 43 62 17 19 21 9 16 65 187 17 79 77 60 8 81 39 133 102 36 19 53 59 29 47 22 140 41 122 10 41 61 19 62 86 66 64 53 79 46 89 74 44 39 59 29 13 11 23 40
A3	140 218 215 109 151 154 93 103 90 184 7 46 9 41 241 118 15 156 111 120 163 101 170 225 177 72 288 129

Úloha B5.16 *Vliv druhu jedu a způsobu jeho aplikace na dobu přežití zvířete (ANOVA2B).*
Byla sledována doba přežití zvířete od okamžiku aplikace jedu. Data obsahují čas přežití v hodinách pro tři rozličné jedy A1 až A3 (faktor *A*) a čtyři způsoby jeho aplikace B1 až B4 (faktor *B*). V každé buňce jsou čtyři pozorování. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda druh jedu a způsob jeho aplikace mají významný vliv na délku přežití a zda existuje významná interakce mezi oběma faktory, jedem a dobou aplikace.

Data: Doba přežití [hodin] po aplikaci tří druhů jedu A1 až A3 a pro čtyři způsoby aplikace B1 až B4.

Jed	Aplikace B1	Aplikace B2	Aplikace B3	Aplikace B4
A1	3.1 4.5 4.6 4.3	8.2 11.0 8.8 7.2	4.3 4.5 6.3 7.6	4.5 7.1 6.6 6.2
A2	3.6 2.9 4.0 2.3	9.2 6.1 4.9 12.4	4.4 3.5 3.1 4.0	5.6 10.0 7.1 3.8
A3	2.2 2.1 1.8 2.3	3.0 3.7 3.8 2.9	2.5 3.0 2.0 3.1	3.0 3.6 3.1 3.3

Úloha B5.17 *Vliv dvou faktorů zařazení ve společnosti na krevní tlak jedince (ANOVA2B).*
Uvažujme hypotetickou studii, týkající se vlivu rychlých kulturních změn domorodců na ostrově v Mikronésii, na hladinu jejich systolického krevního tlaku. Krevní tlak byl měřen náhodnému vzorku 30 mužů starších 40 let, kteří pracují v hlavním městě. Těmto osobám byl předložen sociologický dotazník, ze kterého vyšlo jejich sociální zařazení do prozápadní společnosti (faktor *A*), ale také cítění dle svých tradic a vlastní kultury (faktor *B*). Jedinci byli charakterizováni hodnotou svého krevního tlaku. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda jsou oba faktory statisticky významné ve vlivu na systolický krevní tlak. Pokuste se vysvětlit ANOVA tabulku.

Data: Hodnota systolického krevního tlaku domorodců [mm Hg sloupce] dle jejich sociálního zařazení A1 až A3 a dle jejich kultury a tradice B1 až B3.

Hodnocení dle jejich kultury	Hodnocení dle jejich tradice		
	B1 (Vysoké)	B2 (Střední)	B3 (Nízké)
A1 (Vysoké)	130 140 135	150 145	175 160 170 165 155
A2 (Střední)	145 140 150	150 160 155	165 155 165 170 160
A3 (Nízké)	180 160 145	155 140 135	125 130 110

Úloha B5.18 *Vliv injekce estrogenu na změnu pulzu samce a samice šimpanze (ANOVA2B).* Způsobí injekce estrogenu A1 a A2 (faktor *A*) dospívajícímu šimpanzovi významnou změnu pulzu? Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda injekce má významný vliv na samce a samice šimpanze B1 a B2 (faktor *B*). Má smysl uvažovat interakci mezi oběma faktory?

Data: Změna pulzu šimpanze, (+ zvýšení, - je snížení) po injekci estrogenu A1 a A2 a pohlaví B1 a B2.

Injekce	B1 (Samci)	B2 (Samice)
A1 (Bez estrogenu)	5.1 -2.3 4.2 3.8 3.2 -1.5 6.1 -2.5	-2.3 -5.8 -1.5 3.8 5.5 1.6 -2.4 1.9
A2 (S estrogenem)	15.0 6.2 4.1 2.3 7.6 14.8 12.3 13.1	7.3 2.4 6.5 8.1 10.3 2.2 12.7 6.3

Úloha B5.19 *Vliv teploty vzduchu a pilulky tepelné regulace na tělesnou teplotu (ANOVA2B)*. Data jsou z hypotetické studie o ovlivnění tělesné teploty člověka pilulkou farmaka s pyrogenem (v mg na 1 kg tělesné váhy), zvyšujícího tepelnou regulaci člověka A1 až A3 (faktor *A*) anebo teplotou vzduchu v místnosti B1 až B4 (faktor *B*). Bylo testováno 36 lehkých atletů okamžitě po závodě v klimatizované místnosti. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda teplota místnosti a pilulka, zvyšující tepelnou regulaci, významně ovlivní tělesnou teplotu. Posuďte významnost interakce mezi teplotou vzduchu a požitím pilulky. Existuje silně vybočující výběr dle faktoru *A* nebo dle faktoru *B*?

Data: Dosažená tělesná teplota [^{E}C] po aplikovaném dražé o různé hmotnosti A1 až A3 a při různých teplotách místnosti B1 až B4.

Dražé [mg/kg]	B1 (21 ^{E}C)	B2 (25 ^{E}C)	B3 (29 ^{E}C)	B4 (33 ^{E}C)
A1 (0.0)	37.2 37.2 36.8	36.9 37.0 37.1	36.9 37.0 36.8	37.1 37.3 36.7
A2 (0.05)	37.1 36.9 36.8	37.1 36.7 37.0	36.9 37.0 36.9	36.9 37.0 37.0
A3 (0.10)	37.1 37.1 37.1	36.9 37.0 37.3	36.9 37.0 37.2	36.9 36.8 37.2

Úloha B5.20 *Test hladiny kyseliny močové v krvi u mongoloidních lidí (ANOVA2B)* Mongoloidní osoby věku 21 až 25 let byly testovány na hladinu kyseliny močové v krvi. U těchto pacietnů bývá totiž tato hladina mírně zvýšena. V datech jsou hodnoty A1 mongoloidních a A2 nemongoloidních, ale mentálně retardovaných pacientů (faktor *A*). Vyšetřovaní byli B1 mužského i B2 ženského pohlaví (faktor *B*). Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda mongoloidní pacienti mají vyšší hladinu kyseliny močové v krvi a zda je tato hladina ovlivněna pohlavím pacienta.

Data: Hladina kyseliny močové v krvi.

Pacienti	B1 (Muži)	B2 (Ženy)
A1 (Mongoloidní)	5.84, 6.30, 6.95, 5.92, 7.94	4.90, 6.95, 6.73, 5.32, 4.81
A2 (Nemongoloidní)	5.50, 6.08, 5.12, 7.58, 6.78	4.94, 7.20, 5.22, 4.60, 3.88

Úloha B5.21 *Vliv drogy a druhu zvířete na jeho stres (ANOVA2P)*
Je třeba vyšetřit, zda droga levorfanol snižuje stres živočichů, který se odráží v úrovni hormonů hypofýzy a nadledvinek, tj. kortikosteroidů. U pěti zvířat A1 až A5 (faktor *A*) byla sledována hladina kortikosteroidů, a to B1 bez drogy, B2 za působení samotného levorfanolu, B3 samotného adrenalinu a konečně B4 obou dohromady (faktor *B*). Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda dotyčné látky významně ovlivňují stres a konečně, zda u všech zvířat stejně.
Data: Hladina kortikosteroidů u pěti zvířat při různých hladinách aplikace kortikosteroidu B1 až B4.

Zvíře	B1 (Bez drogy)	B2 (Pouze levorfanol)	B3 (Pouze adrenalin)	B4 (Oba dohromady)
A1	1.90	0.82	5.33	3.08
...	...	...	...	...
A5	1.89	1.21	6.07	2.57

Úloha B5.22 *Vliv času na hladinu alkoholu v krvi (ANOVA1)*
U tří skupin mužů by sledován vliv času na hladinu alkoholu v krvi (faktor *A*) po vypití pěti skleniček alkoholu. U skupiny A1 byl alkohol v krvi měřen 1 hodinu po požití, u skupiny A2 pak 2 hodiny a u skupiny A3 4 hodiny. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda má čas po požití významný vliv na hladinu alkoholu v krvi.
Data: Obsah alkoholu v krvi [%] v závislosti na čase po požití A1 až A3.

A1 (po 1 hodině)	0.11 0.10 0.09 0.09 0.10 0.12 0.11
A2 (po 2 hodinách)	0.08 0.09 0.07 0.07 0.06 0.07
A3 (po 4 hodinách)	0.04 0.04 0.05 0.05 0.06

Úloha B5.23 *Vliv způsobu léčby na čas potřebný k uzdravení králíků (ANOVA1)*
Veterinární věda užívá tří rozličných způsobů léčby chorých králíků A1 až A3. Doba potřebný k jejich uzdravení je pak u každého způsobu ale i u každého králíka, jiný. Na hladině významnosti $\alpha = 0.01$ testujte, zda má způsob léčby (faktor *A*) významný vliv na dobu k uzdravení králíků. Posuďte, zda existuje statisticky významný rozdíl mezi způsobem léčby A1 a A3.
Data: Počet dní k uzdravení králíků u tří způsobů léčby A1 až A3.

Způsob léčby A1	6 8 12 9 7 2
Způsob léčby A2	9 8 7 6 9
Způsob léčby A3	11 9 10 8 11 9 12 14

Úloha B5.24 *Vliv způsobu aplikace zubařské výplně na dosaženou pevnost (ANOVA1)*
Zubařský tým testoval novou hmotu u zubních výplní za současného užití čtyř způsobů aplikace A1 až A4 a sledoval požadovanou pevnost a tvrdost. Na hladině významnosti $\alpha = 0.01$ testujte, zda všechny způsoby (faktor *A*) vedou ke stejnému indexu pevnosti.

Data: Index pevnosti zubařské výplně v zubu u čtyř způsobů aplikace A1 až A4.

A1	8.2 7.9 8.4 8.0 8.0
...	...
A4	8.0 7.2 7.3 7.1 7.9 7.3 7.1 7.4

Úloha B5.25 *Vliv zapomínání a hladiny IQ na výsledek psychologického testu (ANOVA2B)*
Psycholog vyšetřuje, zda zapomínání u člověka souvisí s jeho inteligenčním kvocientem *IQ* a zda ovlivní výsledek psychologického testu. Na hladině významnosti $\alpha = 0.05$ testujte, zda hladina inteligenčního kvocientu A1 až A3 (faktor *A*) a velikost zapomínání B1 až B3 (faktor *B*) mají významný vliv na výsledek psychologického testu.

Data: Výsledek testu [skóre] u lidí rozličného *IQ* A1 až A3 a rozličného zapomínání B1 až B3.

Úroveň IQ	B1 (Zapomíná zřídka)	B2 (Zapomíná občas)	B3 (Zapomíná často)
A1 (Nízké)	15 14 16	10 11 11	5 4 4
A2 (Střední)	20 21 22	30 31 34	10 9 9
A3 (Vysoké)	15 13 14	25 27 26	15 16 13

Úloha B5.26 *Vliv pacienta a diety na ztrátu hmotnosti při nadváze (ANOVA2P)*
Při testování účinnosti čtyř diet u pacientů s nadváhou byla sledována ztráta hmotnosti pacienta v určitém časovém období. Na hladině významnosti $\alpha = 0.05$ testujte, zda pacient A1 až A5 (faktor *A*) ovlivňuje ztrátu hmotnosti a dále, zda má zvolená dieta B1 až B4 (faktor *B*) významný vliv. Existuje pacient nebo dieta, které mají výjimečně odlišné postavení vzhledem k ostatním?

Data: Ztráta hmotnosti [kg] u pěti pacientů A1 až A5 při aplikaci čtyř různých diet B1 až B4.

Pacient	Dieta B1	Dieta B2	Dieta B3	Dieta B4
A1	8	10	6	21
...	...	...	...	...
A5	3	5	10	27

Úloha B5.27 *Vliv teploty a člověka na stanovení obsahu uronových kyselin (ANOVA2B)*
Byl sledován vliv teploty chemické reakce (faktor *A*) a vliv kvality práce člověka (faktor *B*) na stanovení obsahu uronových kyselin. Stanovení prováděli tři laboranti B1 až B3 při čtyřech teplotách A1 až A4 a každé stanovení bylo provedeno dvakrát. Vyšetřete, zda výsledek stanovení je ovlivněn teplotou reakce, kvalitou práce laborantů nebo obojím. Dosáhl některý z laborantů silně odlišných hodnot vzhledem k ostatním dvěma?

Data: Obsah uronových kyselin [hm.%] při čtyřech teplotách A1 až A4 třemi laboranty B1 až B3.

Teplota	Laborant B1	Laborant B2	Laborant B3
A1	97.6 97.9	96.8 96.6	95.8 96.6
...	...	...	...
A4	96.6 96.3	96.8 96.8	97.0 96.8

Úloha B5.28 *Vliv kadmia a kyseliny fytinové na koncentraci železa v játrech potkanů (ANOVA2B)*. Laboratorním potkanům bylo v potravě aplikováno kadmium a kyselina fytinová. Kadmium bylo podáváno v dietě ve dvou úrovních: A1 bez kadmia a A2 s přídavkem kadmia (faktor *A*). Kyselina fytinová byla aplikována ve třech úrovních: B1 nízké, B2 střední a B3 vysoké (faktor *B*). Protože došlo v průběhu experimentu k úhynu

6 potkanů, nejsou počty měření pro jednotlivé kombinace úrovní stejné. Hodnoty naměřené koncentrace železa v játrech [mg/kg] jsou v datech.

Data: Obsah železa v játrech [mg/kg] při hladinách kadmia A1 a A2 a kyseliny fytinové B1 až B3.

	Hladina kyseliny fytinové v potravě		
Kadmium v potravě	B1 (Nízká)	B2 (Střední)	B3 (Vysoká)
A1 (0)	102.34 92.90 92.15 91.61 123.88 117.71 95.26 96.94	83.19 92.63 79.12 112.50 122.21 58.53 86.58 79.50	108.25 84.31 88.67 130.02 104.91 104.65 117.09 100.50
A2 (1)	68.62 88.96 88.12 104.98 105.38 96.47 70.01 107.20	87.98 72.55 70.44 92.25 74.34 106.93 55.00 80.20	107.40 92.57 77.10 86.45 98.22 126.84 95.05 98.98

5.6.2 Analýza chemických a fyzikálních dat

Úloha C5.01 *Vliv laboratoře na výsledek analytického stanovení (ANOVA2P)*

Ve čtyřech rozličných laboratořích byl stanoven procentuální obsah ethylacetátu v pěti sudech A1 až A5 (str. 132 v cit. [13]). Posuďte, zda byl ethylacetát ve všech sudech A1 až A5 (faktor *A*) stejný a homogenní a zda všechny čtyři laboratoře B1 až B4 (faktor *B*) dospěly ke stejným výsledkům. Dosáhla některá laboratoř silně odlišných výsledků?

Data: Obsah ethylacetátu [%] v pěti sudech A1 až A5 a stanovený ve čtyřech laboratořích B1 až B4.

Sud	Lab. B1	Lab. B2	Lab. B3	Lab. B4
A1	73	74	68	71
...	...	...	...	...
A5	73	74	69	73

Úloha C5.02 *Vliv člověka na výsledek analýzy (ANOVA1)*

Pět analytických chemiků provedlo porovnávací stanovení s výsledky, uvedenými v datech (str. 136 v cit. [13]). Vyšetřete, zda všichni analytici A1 až A5 (faktor *A*) dospěli ke stejnému výsledku, když každý provedl jiný počet opakovaných analýz. Stanovte případně vybočující hodnoty. Který analytik dosáhl silně vybočujících výsledků?

Data: Obsah neznámé látky [mg], určený pěti chemiky A1 až A5.

A1	A2	A3	A4	A5
30.0	29.3	29.6	32.5	31.0
...	...	...	...	...
---	30.0	---	---	---

Úloha C5.03 *Vliv člověka a přístroje na stanovení chloridů (ANOVA2B)*

V laboratořích pracovalo 20 analytiků (faktor *A*) na 12 fotometrech (faktor *B*). Je třeba rozhodnout, zda člověk a přístroj významně ovlivňují stanovení obsahu chloridů [ppm] (str. 253 v cit.[13]). Každý ze tří náhodně vybraných analytiků A1 až A3 provedl tři opakovaná měření na třech náhodně vybraných fotometrech B1 až B3. Vypočtěte také odhad směrodatné odchylky každého fotometru.

Data: Obsah chloridů [ppm], určený analytiky A1 až A3 na fotometrech B1 až B3.

	B1 (1. fotometr)			B2 (2. fotometr)			B3 (3. fotometr)		
A1 analytik	2.3	3.4	3.5	3.7	2.8	3.7	3.1	3.2	3.5

A2 analytik	3.5	2.6	3.6	3.9	3.9	3.4	3.3	3.4	3.5
A3 analytik	2.4	2.7	2.8	3.5	3.2	3.5	2.6	2.6	2.5

Úloha C5.04 *Barvení textilií užitím různých šarží barviva (ANOVA1)*
Při sledování kvality barviva byla barvena standardní textilie jednotkové plochy pomocí šarže barviva o standardní koncentraci za přísně konstantních podmínek (str. 426 v cit.[16]). Barvení každou šarží A1 až A6 bylo celkem 6krát opakováno a vždy byl stanoven stupeň využití barviva. Vyšetřete, zdali se významně projevuje vliv dané výrobní šarže barviva (faktor *A*) na stupeň jeho využití.

Data: Stupeň využití barviva [%] v šesti výrobních šaržích A1 až A6.

A1 šarže	94.5	93.0	91.0	89.0	96.5	88.0
...	...	...	...	...	...	...
A6 šarže	98.5	100.0	98.0	100.0	96.5	98.0

Úloha C5.05 *Sledování homogenity produktu výrobní linky ve směnách (ANOVA2B)*
Při výrobě vápenaté soli kyseliny sorbové byly odebírány vzorky B1 ze začátku, B2 středu a B3 z konce výrobní linky (faktor *B*) a ve směnách (faktor *A*). V každém vzorku byl dvakrát stanovován obsah vápníku v procentech. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zdali výroba poskytuje produkty se stejným obsahem vápníku (a) ve všech čtyřech směnách A1 až A4 a (b) ve všech částech výrobní linky B1 až B3.

Data: Obsah vápníku [%], určený ve 4 směnách A1 až A4 a v různých částech výrobní linky B1 až B3.

Směna	B1 (Začátek linky)		B2 (Střed linky)		B3 (Konec linky)	
A1	3.28	3.09	3.52	3.48	2.88	2.80
...	...	...	...	...	...	...
A4	3.78	3.87	4.07	4.12	3.31	3.31

Úloha C5.06 *Vliv přístroje a laboranta na výsledek chemické analýzy*
V laboratoři pracují dva laboranti A1 a A2 (faktor *A*) na dvou různých chromatografech B1 a B2 (faktor *B*) a provádějí stanovení obsahu látky SIC. Na každém přístroji provedl každý laborant dvě opakovaná měření. Ověřte, zda na výsledek analýzy má vliv laborant či přístroj.

Data: Obsah látky SIC [%], určený dvěma laboranty A1 a A2 na dvou přístrojích B1 a B2.

	B1	B2
A1	65.3	65.7
	66.7	65.3
A2	63.7	63.5
	65.7	68.1

Úloha C5.07 *Vliv odměrného skla a analytu na výsledek stanovení (ANOVA2P)*
Zjistěte, je-li vliv použitého odměrného nádobí A1 až A5 (faktor *A*) a koncentrace stanovovaných nerozpuštěných látek B1 až B3 (faktor *B*) statisticky významný na výsledek stanovení.

Data: Výsledek stanovení pro pět druhů odměrného nádobí A1 až A5 a tři různé koncentrace nerozpuštěných látek B1 až B3.

	B1	B2	B3
A1	4.3050	4.1139	4.3283
...	...	...	...
A5	4.4579	4.3096	4.7310

Úloha C5.08 *Vliv tvrdidla a teploty tvrzení na pevnost spoje (ANOVA2P)*
Byl sledován vliv množství přidaného tvrdidla (faktor *A*) a teploty vytvrzování (faktor *B*) na pevnost lepeného spoje. Na hladině významnosti $\alpha = 0.05$ určete, zda (a) různá množství tvrdidla ovlivní mez pevnosti, (b) různé teploty mají vliv na mez pevnosti, a (c) zda-li existuje interakce obou vlivů na mez pevnosti. Prověřte, zda je možné použít mocninné transformace pro zlepšení aditivity působení tvrdidla a teploty.
Data: Mez pevnosti při různém množství tvrdidla A1 až A8 a teplotách vytvrzování [EC] B1 až B7.

Obsah tvrdidla	B1 (20E)	B2 (40E)	B3 (60E)	B4 (80E)	B5 (100E)	B6 (120E)	B7 (140E)
A1 (0.5%)	11	12.5	12.5	14.0	15.0	15.0	15.0
...	...	...	...	...	...	...	...
A8 (4.0%)	19.0	19.0	18.5	18.0	16.5	15.5	15.0

Úloha C5.09 *Vliv laboratoře na hodnotu v kruhovém testu (ANOVA1)*
Při kruhovém testu byla prověřována hodnota výsledku analýzy v deseti radiologických laboratořích. Vyšetřete, zda laboratoř A1 až A10 (faktor *A*) významně ovlivňuje výsledek stanovení. Stanovte případné vybočující hodnoty. Rozhodněte, zda významně ovlivňují výsledek analýzy.
Data: Obsah neznámé látky [%], stanovené v deseti radiologických laboratořích A1 až A10.

Označení posuzované laboratoře									
A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
2.94	2.39	3.24	2.28	2.40	1.82	2.53	2.17	1.76	3.03
...	...	...	...	...	...	...	...	...	...
-	-	2.88	2.81	-	2.51	-	-	-	-

Úloha C5.10 *Vliv místa v sušicím zařízení na změnu teploty (ANOVA1)*
V prostoru sušicího zařízení byla sledována teplota na šesti místech A1 až A6. Posuďte, zda rozložení teploty (faktor *A*) uvnitř sušicího zařízení lze považovat za rovnoměrné a pocházející z normálního rozdělení. Je splněn předpoklad výběrové homoskedasticity?
Data: Teplota v sušicím zařízení [E C] v různých sledovaných místech zařízení A1 až A6.

A1	A2	A3	A4	A5	A6
22.1	22.4	22.3	22.4	22.3	22.6
...	...	...	...	...	...
22.4	22.4	22.6	22.7	22.3	22.7

Úloha C5.11 *Vliv aditiva Irganox 1010 v polyethylenu Mosten (ANOVA1)*
Vzorek granulátu vysokohustotního polyethylenu Mosten VB33 z výrobní linky byl vylisován v tandemu horkého a studeného hydraulického lisu do tvaru fólie tloušťky 0.7 mm o rozměrech 100 × 100 mm. V pásku této fólie byl na infračerveném

spektrofotometru stanoven obsah aditiva Irganox 1010. Analýzy obsahu aditiva byly provedeny na 4 fóliích a v laboratořích dvou firem. Vyšetřete, zda má laboratoř A1 a A2 (faktor *A*) významný vliv na obsah stanoveného aditiva v polyethylenu. Proveďte porovnání výsledků také párovým *t*-testem.

Data: Obsah aditiva [mg], stanovený laboratořemi A1 a A2.

A1	A2
0.017	0.026
...	...
0.017	0.036

Úloha C5.12 *Test homogenity vzorkovnic (ANOVA1)*

Testujte homogenitu sedmi vzorkovnic, sloužících k přípravě uranového koncentrátu, když byl v 7 vzorkovnicích A1 až A7 (faktor *A*) opakovaně stanovován obsah železitých iontů Fe^{3+} [mg] fotometricky. Je splněn předpoklad výběrové normality a homoskedasticity? Přinesla některá vzorkovnice silně odlehlé výsledky od ostatních?

Data: Obsah železitých iontů Fe^{3+} [mg] v sedmi vzorkovnicích.

A1. vzorkovnice: 638, 653, 648, 625,
A2. vzorkovnice: 650, 640, 638, 652
A3. vzorkovnice: 633, 645, 618, 659,
A4. vzorkovnice: 640, 664, 650, 638,
A5. vzorkovnice: 636, 644, 660, 655,
A6. vzorkovnice: 651, 644, 644, 639,
A7. vzorkovnice: 651, 644, 644, 639.

Úloha C5.13 *Vliv vlhkosti vzduchu a teploty plamene na rychlost hoření grafitu, (ANOVA2P)*. Byla vyšetřována rychlost hoření umělého grafitu v proudu vzduchu, obohaceného rozličným obsahem vodní páry. Rychlost hoření byla měřena pro rozličné teploty ve stupních Kelvina A1 až A9 (faktor *A*) a pro rozličné molární zlomky vody B1 až B3 (faktor *B*). Na hladině významnosti $\alpha = 0.05$ vyšetřete statistickou významnost obou faktorů. Dá se prokázat vzájemná interakce faktorů ?

Data: Rychlost hoření pro rozličné teploty a molární zlomky vody x_{voda}.

Teplota [K]	B1 ($x_{voda} = 0.0022$)	B2 ($x_{voda} = 0.017$)	B3 ($x_{voda} = 0.080$)
A1 (1000)	1.68	1.69	1.72
...	...	...	...
A9 (1800)	4.63	4.64	4.71

Úloha C5.14 *Vliv katalyzátoru na pevnost betonu (ANOVA1)*

Koncentrace katalyzátoru v zálivkové cementové kaši ovlivňuje pevnost výsledného betonu. Byl vyšetřován vliv tří rozličných koncentrací katalyzátoru A1 až A3 (faktor *A*) na pevnost betonu. Pevnost byla měřena tlakem v librách na čtvereční palec tak, že blok betonu byl zatěžován v lisu až do okamžiku prasknutí. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda koncentrace katalyzátoru významně ovlivňuje pevnost betonu.

Data: Pevnost betonu [libra/palec^2] pro tři různé koncentrace katalyzátoru A1 až A3.

A1 (35 %)	A2 (40 %)	A3 (45 %)
5.9	6.8	9.9
...	...	...
7.7	8.2	8.7

Úloha C5.15 *Vliv teploty a času na množství odstraněné vody z papíru (ANOVA2B)*

Množství odstraněné vody z papíru závisí na rychlosti papíru čili expozičním čase na válci A1 až A3 (faktor *A*) a na teplotě sušicího válce B1 až B3 (faktor *B*). Vyšetřete, zda oba faktory jsou statisticky významné a ovlivňují množství odstraněné vody. Vyšetření provedete na hladině významnosti $\alpha = 0.05$. Jsou oba faktory v interakci?

Data: Množství odstraněné vody z papíru [%] pro expoziční časy A1 až A3 a různé teploty B1 až B3.

Čas [s]	B1 (100 °F)	B2 (120 °F)	B3 (140 °F)
A1 (10)	24 26 21 25	33 33 36 32	45 49 44 45
A2 (20)	39 34 37 40	51 50 47 52	67 64 68 65
A3 (30)	58 55 56 53	75 71 70 73	89 87 86 83

Úloha C5.16 *Test homogenity papíru na pórovitost u různých rolí papíru (ANOVA1)* Pórovitost papíru, vycházejícího z výrobního válce, byla testována na homogenitu a normalitu rozdělení. Z každé vyrobené role papíru byl odstřižen proužek z konce a provedeno čtvero měření pórovitosti papíru. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda všechny role A1 až A10 (faktor *A*) mají stejnou pórovitost, čili zda je vyráběný papír homogenní. Dosahuje některá role extrémně odlišných výslkedků?

Data: Pórovitost u 10 náhodně vybraných rolí A1 až A10.

Role	Pórovitost papíru
A1	974 978 976 975
...	....
A10	999 1002 998 1003

Úloha C5.17 *Pevnost papíru v závislosti na dnech a počtu vyrobených rolí (ANOVA2B)* Pevnost papíru odvisí od délky celulózových vláken ve dřevě a dalších vlastností dřeva. Jelikož se dodávky celulózy kvalitou dost mění, mění se také kvalita a pevnost vyráběného papíru. Náhodně bylo vybráno 6 výrobních dní B1 až B6 (faktor *B*) v rozmezí 4 měsíců, ve kterých byl odstřižen proužek papíru z konce role. Byl rozlišen i počet rolí A1 až A3 (faktor *A*), vyrobených za den. Každá zkouška na pevnost byla 1krát reprodukována. Celkem bylo testováno 18 proužků papíru. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda po dobu 4 měsíců byla výroba co do pevnosti papíru homogenní a zda je ovlivněna počtem rolí vyrobených za den. Byla pevnost papíru, vyrobeného některý den silně odlišná od ostatních dní?

Data: Pevnost papíru [libra/palec^2] ve 3 rolích A1 až A3 v šesti dnech B1 až B6 ze 4 měsíců.

Denně	B1 (1. den)	B2 (2. den)	B3 (3. den)	B4 (4. den)	B5 (5. den)	B6 (6. den)
A1 (1 role)	20.7 19.3	22.1 20.4	19.0 19.9	20.6 18.9	23.2 22.5	20.7 18.5
A2 (2 role)	21.2 20.1	21.6 22.5	18.8 19.3	19.8 20.1	24.2 22.9	19.6 21.3
A3 (3 role)	19.9 20.5	20.9 22.1	20.2 19.4	20.7 19.2	23.4 24.6	20.0 18.6

Úloha C5.18 *Vliv stáří činidla a doby reakce na stanovení obsahu hydroxyprolinu (ANOVA2P)* Byl sledován vliv stáří chloraminu T (faktor *A*) a doby chemické reakce (faktor *B*) na stanovení procentuálního obsahu hydroxyprolinu v elastinu. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda stáří činidla A1 až A3 nebo reakční doba B1 až B4 ovlivňují stanovení obsahu hydroxyprolinu a určete, zda existuje interakce vlivu obou vyšetřovaných faktorů na obsah hydroxyprolinu v elastinu.

Data: Obsah hydroxyprolinu [hmotnostní %] v elastinu pro různé stáří chloraminu T A1 až A3 a pro různé doby

B1 až B4 chemické reakce.

	Doba reakce			
Stáří chloraminu T	B1	B2	B3	B4
A1	0.45	0.46	0.45	0.42
A2	0.42	0.45	0.46	0.42
A3	0.37	0.43	0.43	0.36

Úloha C5.19 *Vliv obsahu dimethyltereftalátu (DMT) na reprodukovatelnost stanovení (ANOVA1)*. Byla zkoumána oxidace p-xylenu vzduchem za katalýzy různými katalyzátory. Výsledná reakční směs byla esterifikována diazomethanem a produkty esterifikace byly analyzovány plynovou chromatografií. K ověření analytické metody bylo naváženo několik modelových směsí s různým obsahem dimethyltereftalátu (DMT) a vedlejších látek. Bylo proměřováno pět vzorků o koncentracích A1 3.3, A2 1.9, A3 4.0, A4 5.1 a A5 25.1 hmotnostních % DMT, každý třikrát až osmkrát. Cílem bylo zjistit, zda skutečný obsah DMT ve vzorku A1 až A5 (faktor *A*) má vliv na reprodukovatelnost stanovení.

Data: Obsah DMT ve vzorcích [hmotn. %], vzniklých katalýzou různých katalyzátorů A1 až A5.

A1	A2	A3	A4	A5
18.48	23.68	18.75	16.86	11.87
...	...	...	...	...
24.24	-	-	-	-

Úloha C5.20 *Vliv času a nadbytku amoniaku na konverzi diketenu na acetoacetamid (ANOVA2P)*. Byl zkoumán vliv dvou faktorů na konverzi diketenu na acetoacetamid. Experiment byl realizován tak, že do připraveného roztoku amoniaku byl dávkován diketen za chlazení na teplotu 8EC. Vzorky reakční směsi byly odebírány při 4 časových úrovních (faktor *B*), a to v časech B1 0, B2 15, B3 30, B4 45 min po skončení dávkování diketenu, přeneseny do přebytku odměrného roztoku HCl a retitrací byla stanovena koncentrace nespotřebovaného amoniaku. Faktorem *A* (řádky) byl molární poměr amoniak: diketen s pěti úrovněmi, a to A1 1, A2 1.05, A3 1.1, A4 1.15, A5 1.2.

Data: Procento konverze diketenu [%] pro pět hodnot molárních poměrů A1 až A5 amoniak : diketen a čtyři časové úrovně B1 až B4.

Mol. poměr	B1 (0 min)	B2 (15 min)	B3 (30 min)	B4 (45 min)
A1 (1.00)	94.20	95.50	95.70	95.50
...	...	...	...	...
A5 (1.20)	96.70	97.00	96.60	96.9

Úloha C5.21 *Vliv druhu operace a změkčovadla na prodloužení kaučuku (ANOVA2B)* U 48 vzorků surového kaučuku byl měřen modul prodloužení. Pro každý vzorek byla provedena tři měření. Vzorky pocházely ze 4 výrobních operací A1 až A4 (faktor *A*). Do každého vzorku byly přimíchány stanoveným způsobem 4 druhy změkčovadel B1 až B4 (faktor *B*). Cílem bylo zjistit, zda faktory *A* a *B* mají vliv na modul 700%ního prodloužení vzorku kaučuku.

Data: 700%ní prodloužení vzorku kaučuku [lb/sq inch], vzniklého ve 4 operacích A1 až A4 a za působení 4 druhů změkčovadla B1 až B4.

Operace	B1	B2	B3	B4

A1	211.0	196.0	200.0	323.0
	215.0	186.0	221.0	279.0
	197.0	190.0	198.0	251.0
...	...	...	...	...
A4	229.0	198.0	196.0	273.0
	250.0	209.0	197.0	241.0
	238.0	221.0	186.0	221.0

Úloha C5.22 *Vliv způsobu promytí a teploty sušení na koloristické vlastnosti pigmentu (ANOVA2P)*. Při syntéze pigmentu při běžném promytí vodou vznikají po usušení tvrdé kusy pigmentu, které je obtížné jemně umlít a při koloristickém hodnocení pak dosahují horší barevné síly a zvýšené kalnosti odstínu. Proto je třeba nejprve určit vliv způsobu promytí a teploty sušení na kvalitu pigmentu. Zjištění vlivu teploty sušení (faktor *A*) a podmínek promytí (faktor *B*) lze řešit dvoufaktorovou analýzou rozptylu bez opakování. Za teploty sušení (faktor *A*) lze vybrat úrovně A1 = 50 EC, A2 = 55EC, A3 = 65 EC a A4 = 80 EC. Za způsoby promytí (faktor *B*) lze zvolit 4 úrovně jako pevné efekty, a to B1 promytí 250 ml vody 95 EC teplé a 250 ml ethanolu 70 EC teplého, B2 promytí 250 ml vody 95 EC teplé, 250 ml ethanolu 70 EC teplého a rozmíchání pasty pigmentu v 500 ml ethanolu a filtrace, B3 promytí 250 ml vody 95 EC teplé a 250 ml vody s 3 g Slovafolu 915, B4 promytí 500 ml vody 95 EC teplé. Test se provede tak, že pigment se připraví dvakrát a pokaždé se reakční suspenze rozdělí na 4 objemové díly a každý díl je filtrován zvlášť a promyt podle některého z uvedených způsobů promytí. Získané pasty pigmentu se rozdělí vždy na dva díly a jsou sušeny v prvním pokusu při 50, resp. 80 EC, v druhém pokusu při 55 EC, resp. 65 EC. Celkem se koloristicky vyhodnotí 16 sušin pigmentu. Vyšetřete nyní vlivy způsobu promytí a teplot sušení na koloristické vlastnosti pigmentu.

Data: Koloristická vlastnost sušiny pigmentu při čtyřech různých teplotách sušení A1 až A4 a čtyřech způsobech promytí B1 až B4.

Teplota sušení	B1	B2	B3	B4
A1 (50 EC)	125.5	144.9	135.6	133.1
...	...	...	...	...
A4 (80 EC)	153.1	159.3	151.2	119,2

5.6.3 Analýza environmetálních, potravinářských a zemědělských dat

Úloha E5.01 *Vliv umělého hnojiva na výnosy plodiny (ANOVA1)*

Průměrné výnosy určité zemědělské plodiny při použití umělého hnojiva A1 až A3 jsou uvedeny v datech (str. 213 v cit.[12]). Na hladině významnosti $\alpha = 0.05$ ověřte, zda průměrné výnosy (faktor *A*) jsou pro tato hnojiva stejné. Stanovte, zda je významný rozdíl mezi hnojivem A1 a A2. Existuje mezi třemi testovanými významně odlišné hnojivo?

Data: Výnos plodiny [q/ha] pro různá hnojiva A1 až A3.

Hnojivo	Výnos [q. ha^{-1}]
A1	40 42 45 40 44 47
A2	76 75 82
A3	60 58 62 64

Úloha E5.02 *Vliv druhu krmiva na přírůstky hmotnosti dobytka (ANOVA1)*
Byl sledován vliv čtyř druhů krmiva A1 až A4 (faktor *A*) na přírůstky hmotnosti šesti náhodně vybraných kusů dobytka (str. 214 v cit.[12]). Na hladině významnosti $\alpha = 0.05$ rozhodněte, zda přírůstky hmotnosti jsou pro různé druhy krmiva stejné. Vyčíslete také intervaly spolehlivosti středních hodnot. Existuje krmivo, které se svými vlastnostmi silně odlišuje od ostatních?

Data: Přírůstek hmotnosti [kg/ks] pro různá krmiva A1 až A4.

Krmivo	Přírůstky hmotnosti [kg/ks]
A1	22.8 25.2 27.6 28.1 27.3 28.6
...	...
A4	19.9 21.4 20.6 20.9 21.1 20.6

Úloha E5.03 *Vliv konzervačního činidla na trvanlivost potraviny (ANOVA1)*
Byl sledován vliv druhu konzervačního činidla A1 až A4 (faktor *A*) na trvanlivost potravinářského výrobku (str. 214 v cit.[12]). Vyšetřete, zda na hladině významnosti $\alpha = 0.05$ závisí trvanlivost potravin na druhu konzervačního činidla. Který druh konzervačního činidla dosahuje silně odlišných výsledků vůči ostatním druhům?
Data: Trvanlivost [počet dnů] pro různé druhy konzervačního činidla A1 až A4.

Druh A1	12 14 10 17 21 19 16
...	...
Druh A4	7 7 7 8 9 8 9

Úloha E5.04 *Vliv druhu pšenice a lokality orné půdy na výnos pšenice (ANOVA2P)*
Byl sledován vliv šesti odrůd pšenice A1 až A6 (faktor *A*) a šesti lokalit orné půdy B1 až B6 (faktor *B*) na hektarový výnos pšenice (str. 214 v cit.[12]). Na hladině významnosti $\alpha = 0.05$ a 0.01 vyšetřete, zda (a) různé odrůdy pšenice dávají stejný hektarový výnos, (b) různé lokality orné půdy ovlivňují hektarový výnos, (c) existuje interakce vlivu obou vyšetřovaných faktorů na hektarový výnos, (d) která lokalita přináší silně odlišné výnosy, (e) který druh pšenice vzhledem ke všem ostatním přináší silně odlišné výnosy?
Data: Hektarový výnos [q. ha^{-1}] pro různé odrůdy A1 až A6 a různé lokality B1 až B6.

Odrůda	B1	B2	B3	B4	B5	B6
Odrůda A1	30	27	19	25	20	26
...	...	...	...	...	...	...
Odrůda A6	17	17	19	20	17	18

Úloha E5.05 *Porovnání stanovení arzeniku v pěti laboratořích (ANOVA2U)*
Arzenik lze v potravě stanovit reakcí s molybdenovou solucí (B1) a reakcí s diethyldithiokarbamátem stříbrným dle Vašáka a Šedivce (B2). Ke vzorku potravy bylo přidáno 15 g arzeniku a vzorek byl analyzován oběma metodami B1 a B2 (faktor *B*) v pěti laboratořích A1 až A5 (faktor *A*) s vícenásobnou reprodukovatelností. Vedou obě metody ve všech laboratořích ke stejným výsledkům? Existuje statisticky významná interakce mezi analytickou metodou a laboratoří?

Data: Obsah arzeniku [g], stanovený v různých laboratořích A1 až A5 dvěma metodami B1 a B2.

1. laboratoř, metoda I: A1, B1	12.9	13.2	12.9	12.9	13.1	13.0
...	...	...	...	...	...	...
5. laboratoř, metoda II: A5, B2	13.8	14.1	13.8	13.9	14.0	

Úloha E5.06 *Vliv přístroje a člověka na stanovení mědi v půdě (ANOVA2B)*
Metodou atomové absorpční spektrofotometrie AAS byl stanoven obsah mědi v půdě. K rozkladu vzorku byly použity tři techniky A1 až A3 (faktor *A*), prováděné dvěma laborantkami B1 a B2 (faktor *B*). Rozklady vzorků byly prováděny a) směsí kyselin H_2SO_4 a HNO_3, b) směsí kyselin HNO_3a HCl, c) mikrovlnným rozkladem. Rozklady prováděly dvě laborantky tak, že každou technikou připravily dva vzorky. Rozhodněte, zda použitá technika, či lidský faktor mají vliv na stanovení. Existuje snad další faktor, který stanovení ovlivňuje?

Data: Obsah mědi [ppm], určené 3 technikami rozkladu A1 až A3 a dvěma laborantkami B1 a B2.

Technika rozkladu	B1 (Labor. 1)	B2 (Labor. 2)
A1 (Směs H_2SO_4 a HNO_3)	22.60	25.60
	23.10	23.90
...	...	...
A3 (Mikrovlnný rozklad)	40.00	47.10
	42.80	45.80

Úloha E5.07 *Vliv nadmořské výšky na koncentraci oxidu siřičitého v ovzduší (ANOVA2P)*.
Od 1. do 5. 2. 1993 se nad územím okresu jednoho města udržovala mohutná inverze, kdy panovalo bezvětří a rozptylové podmínky byly nepříznivé. V několika stanicích byly naměřeny denní průměrné hodnoty koncentrací oxidu siřičitého SO_2 [$\mu g/m^3$] v různých nadmořských výškách měřicích stanic. Vyšetřete, zda faktor nadmořské výšky A1 až A4 (faktor *A*) nebo časový faktor *B*1 až B5 (faktor *B*) ovlivňuje naměřené hodnoty. Pomocí grafu neaditivity nalezněte možnou transformaci dat. Existuje statisticky významná interakce obou faktorů?

Data: Hodnota SO_2 [$\mu g/m^3$] v různých nadmořských výškách A1 až A4 a v různém čase B1 až B5.

Stanice/Datum	B1 (1. 2.)	B2 (2. 2.)	B3 (3. 2.)	B4 (4. 2.)	B5 (5. 2.)
A1 (Město 1, OHS)	200	312	316	284	300
...	...	...	...	...	...
A4 (Město 3, OHS)	598	591	420	488	344

Úloha E5.08 *Vliv výšky petrklíče na výnos jeho osiva (ANOVA1)*
Při sledování výnosnosti petrklíčů Picotee byl posuzován vliv velikosti rostlin (faktor *A*) na výnos osiva v g/rostlinu. Rostliny byly sledovány po dobu jedné sezóny v průběhu prvního roku své vegetace. Rostliny byly rozděleny na tři výškové typy, a to typ A1 (nižší než 7 cm), typ A2 (7.5 - 9 cm) a typ A3 (vyšší než 9.5 cm). Od každého typu bylo odebráno a vyšetřováno 10 vzorků. Vyšetřete, zda je výnos osiva ovlivněn výškou rostliny.

Data: Výnos osiva petrklíče [g/rostlinu] v závislosti na výšce petrklíče A1 až A3.

A1 (< 7 cm)	A2 (7.5 - 9 cm)	A3 (> 9.5 cm)
0.19	0.347	0.343
...	...	...

0.039	0.184	0.383

Úloha E5.09 *Vliv průmyslové lokality na koncentraci ozónu v ovzduší*
Nadhraniční množství ozónu v ovzduší je kritériem znečištění ovzduší. Bylo shromážděno 6 vzorků vzduchu ze čtyř lokalit A1 až A4 průmyslové oblasti Středozápadu USA (faktor *A*) a byla stanovena koncentrace ozónu v ppm. Má lokalita významný vliv na množství ozonu v ovzduší? Existuje lokalita, která vykazuje silně odlišné výsledky od ostatních?
Data: Koncentrace ozónu ve vzduchu [ppm] na čtyřech lokalitách A1 až A4.

A1	A2	A3	A4
0.08	0.15	0.13	0.05
...	...	...	...
0.06	0.13	0.17	0,08

Úloha E5.10 *Vliv druhu řeky na koncentraci PCB v rybách (ANOVA1)*
PCB, nebezpečné karcenogenní látky, vznikají při výrobě elektrických transformátorů a kondenzátorů. Vzorky ryb z pěti řek A1 až A5 (faktor *A*) byly analyzovány na koncentraci PCB v ppm. Má druh řeky vliv na koncentraci PCB v rybách? Proveďte analýzu vlivných bodů a přepočtěte výsledky s vyloučením případných extrémních odchylek. Existuje řeka, která vykazuje silně odlišné výsledky od ostatních řek?
Data: Koncentrace PCB v rybách [ppm] rozličných řek A1 až A5.

A1	A2	A3	A4	A5
2	4	12	7	13
...	...	...	...	...
-	-	-	-	7

Úloha E5.11 *Vliv teploty a reakčního času na obsah uhlovodíků (ANOVA2P)*
Zahřívání rašeliny uvolňuje řadu fermentovatelných uhlovodíků, jež mají mnoho důležitých průmyslových využití. Byly studovány experimentální podmínky, za kterých dochází k uvolnění těchto uhlovodíků. Proces a obsah uvolněných rozpustných uhlovodíků byl sledován při třech teplotách A1 až A3 (faktor *A*) pro pět časů chemické reakce B1 až B5 (faktor *B*). Který ze dvou vyšetřovaných faktorů má významný vliv na obsah uhlovodíků? Existuje teplota, při které dochází k významně odlišnému uvolnění uhlovodíků?
Data: Obsah rozpustných uhlovodíků [%] pro různé teploty [EC] A1 až A3 a čas reakce [min] B1 až B5.

	B1 (0.5 min)	B2 (1 min)	B3 (2 min)	B4 (3 min)	B5 (5 min)
A1 (170 EC)	1.3	1.8	3.2	4.9	11.7
A2 (200 EC)	9.2	17.3	18.1	18.1	18.8
A3 (215 EC)	12.4	20.4	17.3	16	15.3

Úloha E5.12 *Vliv dne, pH kyselého deště a hloubky zeminy na kyselost půdy (ANOVA3P)*.
"Kyselé deště" jsou považovány za jedno z nebezpečných znečištění životního prostředí. Vznikají reakcí vodní páry v mracích a oxidu dusíku a oxidu siřičitého, které jsou uvolňovány hořením uhlí. Kyselé deště také ovlivňují aciditu půdy. Acidita půdy v oblasti Gainesville, Florida byla měřena ve třech rozličných hloubkách A1 až A3 (faktor *A*) a ve třech kritických dnech roku 1981 po dešti, jehož pH bylo právě změřeno. Vyšetřete, (a) zda hloubka půdy A1 až A3 a kyselost deště C1 a C2 (faktor *C*) bez ohledu na datum v roce

mají vliv na kyselost půdy, (b) zda kyselost deště C1 a C2 a datum v roce B1 až B3 (faktor *B*) mají vliv na kyselost půdy bez ohledu na její hloubku A1 až A3. Lze prohlásit, že jejich působení je aditivní? Existuje interakce faktorů a má logický smysl?

Data: Vliv hloubky půdy A1 až A3, datumu v roce B1 až B3 a kyselosti deště C1 až C2 na kyselost půdy v jednotkách pH.

Datum kyselého deště: 6	B1 (3. 4. 1981)		B2 (16. 6. 1981)		B3 (30. 6. 1981)	
Hloubka zeminy: 9	C1 (pH 3.7)	C2 (pH 4.5)	C1 (pH 3.7)	C2 (pH 4.5)	C1 (pH 3.7)	C2 (pH 4.5)
A1 (0 - 15 cm)	5.33	5.33	5.47	5.47	5.2	5.13
A2 (15 - 30 cm)	5.27	5.03	5.5	5.53	5.33	5.2
A3 (30 - 46 cm)	5.37	5.4	5.8	5.6	5.33	5.17

Úloha E5.13 *Vliv vyšlechtěných hybridů a lokalit území na výnos kukuřice (ANOVA2P)*
Byly sledovány čtyři vyšlechtěné hybridy kukuřice A1 až A4 (faktor *A*), jež odolávají houbovým infekcím. Na pěti lokalitách státu v USA B1 až B5 (faktor *B*) byl sledován vliv vyšlechtěného hybridu a vliv lokality území na výnos kukuřice. Je některý z těchto faktorů statisticky významný ? Existuje interakce obou faktorů a má logický smysl?

Data: Výnos [q] čtyř hybridů kukuřice A1 až A4 v pěti lokalitách B1 až B5.

	B1 (NW)	B2 (NE)	B3 (C)	B4 (SE)	B5 (SW)
A1 (FR-11)	62.3	64.0	64.3	65.0	66.4
...	...	...	...	...	...
A4 (RC-3)	55.4	56.0	59.8	58.0	58.8

Úloha E5.14 *Vliv druhu borovice a lokality výskytu na průměr kmene (ANOVA2B)*
Průměry kmene tří druhů borovice A1 až A3 (faktor *A*) byly porovnány na čtyřech lokalitách B1 až B4 výskytu borovice (faktor *B*), a to pomocí 5 náhodně vybraných stromů od každého druhu v každé lokalitě. Na hladině významnoti $\alpha = 0.05$ vyšetřete, zda oba faktory významně ovlivnují průměr borovice. Existuje interakce obou faktorů a má fyzikální smysl?

Data: Průměr borovice [palec] pro 3 druhy A1 až A3 a 4 lokality B1 až B4.

Druh borovice	Lokalita B1	Lokalita B2	Lokalita B3	Lokalita B4
A1	23 15 26 13 21	25 20 21 16 18	21 17 16 24 27	14 17 19 20 24
A2	28 22 25 19 26	30 26 26 20 28	19 24 19 25 29	17 21 18 26 23
A3	18 10 12 22 13	15 21 22 14 12	23 25 19 13 22	18 12 23 22 19

Úloha E5.15 *Vliv místa a měsíce na obsah oxidu siřičitého v ovzduší (ANOVA2P)*
Byl sledován obsah oxidu siřičitého v ovzduší během roku 1993, a to průměrné měsíční koncentrace na různých místech okresu Ústí nad Orlicí: A1-Česká Třebová, A2-Dolní Lipka, A3-Kameničná, A4-Podlesí, A5-Žichlínek, A6-Vraclav. Ověřte, zda je významný vliv místa A1 až A6 (faktor *A*) a měsíce B1 až B12 (faktor *B*) na obsah oxidu siřičitého v ovzduší. Existuje statisticky významná interakce obou faktorů a má fyzikální smysl?

Data: Obsah SO_2 [$mg.m^{-3}$] v ovzduší na 6 místech A1 až A6 v průběhu 12 měsíců B1 až B12.

Místo	B1 (leden)	B2 (únor)	B3 (březen)	B4 (duben)	B5 (květen)	B6 (červen)
A1	44	103	47	7	2	5
..	..	...	...	...	...	...
A6	9	6	10	21	38	13

Úloha E5.16 *Vliv instrumentální metody a laboranta na obsah dusičnanů v salátu (ANOVA2B)*. Dusičnany v zeleninovém salátu byly stanovovány třemi laboranty A1 až A3 (faktor *A*) a třemi instrumentálními metodami (faktor *B*), a to B1 izotachoforeticky ITP, B2 fotometricky s 3,4-xylenolem XYL a B3 kolorimetricky po redukci dusičnanů na kadmiové koloně. Každý laborant přitom provedl 3 paralelní stanovení. Který z testovaných faktorů má statisticky významný vliv na obsah dusičnanů v zeleninovém salátu?

Data: Obsah dusičnanů v zeleninovém salátu [mg $NaNO_3$/kg salátu] třemi laboranty A1 až A3 a třemi metodami B1 (ITP), B2 (XYL) a B3 (NEDA).

Laborant	B1 (ITP)	B2 (XYL)	B3 (NEDA)
A1	211.3, 211.3, 220.5	221.2, 216.8, 201.9	219.9, 210.1, 204.8
A2	206.9, 218.7, 221.2	210.3, 203.9, 205.0	215.7, 210.9, 212.6
A3	223.0, 216.4, 220.0	216.8, 210.3, 226.8	217.2, 205.6, 214.8

5.6.4 Analýza hutnických a mineralogických dat

Úloha H5.01 *Vliv obsahu kovu na intenzitu spektrální čáry Mn a Co (ANOVA1)*

V emisní spektrální analýze oceli byl sledován obsah manganu a kobaltu dle intenzity zčernání páru čar S(Mn) u Λ(Mn) = 2576.1 Å a S(Co) u Λ(Co) = 2583.1 Å (str. 60 v cit.[15]). Na hladině významnosti α = 0.05 vyšetřete, zda rozdíl v intenzitě zčernání těchto čar ΔS = S(Mn) - S(Co) je u šesti různých ocelí A1 až A6 (faktor *A*) ovlivněn rozličným obsahem železa, chromu a niklu. Každá ze šesti ocelí byla na rychlofotometru proměřena šestkrát. Existuje druh oceli se silně odlišnými hodnotami zčernání vzhledem k ostatním?

Data: Rozdíl v intenzitě zčernání spektrálních čar ΔS u šesti ocelí A1 až A6.

	ΔS					
A1. ocel	0.14	0.17	0.13	0.15	0.15	0.17
...	...	...	...	...	...	...
A6. ocel	0.17	0.17	0.17	0.14	0.17	0.16

Úloha H5.02 *Vliv druhu svářecího kovu na pevnost sváru (ANOVA2P)*

Vazebným pojítkem svaru zirkoniové slitiny bývá nikl, železo a měď. Byly vytvořeny svary o rozličném složení těchto svářecích komponent a cílem je vyšetřit pevnost svaru, tzn. největší tlak v tisících liber na čtvereční palec k přerušení svaru. Na sedmi svarech A1 až A7 (faktor *A*) a hladině významnosti α = 0.05 vyšetřete, zda záleží na druhu kovu B1 až B3 (faktor *B*), užitého ve svářecím drátu, zda tlaky k přerušení svaru jsou u všech drátů stejné. Prozkoumejte, zda je třeba provést transformaci vedoucí ke stabilizaci rozptylu. Ovlivnil některý kov odlišně pevnost svaru ve srovnání s ostatními kovy? Existuje statisticky významná interakce obou faktorů a má logický smysl?

Data: Tlak k roztržení svaru zirkonové slitiny [10^3 liber/palec^2] pro sedm svarů A1 až A7 a tři druhy svářecích drátů B1 až B3.

Svár	B1 (Nikl)	B2 (Železo)	B3 (Měď)
A1	67.0	71.9	72.2
...	...	...	...
A7	75.6	84.9	69.0

Úloha H5.03 *Vliv laboratoře a metody na stanovení obsahu síry v uhlí (ANOVA2B)*
Byl vyšetřován obsah síry v uhlí dvěma nezávislými analytickými metodami v sedmi laboratořích. Každé měření bylo 2× opakováno. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda je obsah síry v uhlí ovlivněn analytickou metodou A1 a A2 (faktor *A*) nebo některou ze sedmi laboratoří B1 až B7 (faktor *B*), kde byla analýza provedena. Jaký je praktický výklad interakce užité metody a laboratoře? Lze eliminovat interakci mocninnou transformací?

Data: Obsah síry v uhlí [%] metodami A1 a A2 a v laboratořích B1 až B7.

Metoda 9	B1	B2	B3	B4	B5	B6	B7
A1	0.107	0.127	0.115	0.108	0.097	0.114	0.155
	0.105	0.122	0.112	0.108	0.096	0.119	0.145
A2	0.105	0.127	0.109	0.117	0.110	0.116	0.164
	0.103	0.124	0.111	0.115	0.097	0.122	0.160

Úloha H5.04 *Vliv obsahu antimonu a způsobu chlazení na pevnost sváru (ANOVA2B)*
Antimon často nahrazuje v pájení kompozicí olovo-cín právě dražší cín. Je třeba vyšetřit, zda zvyšující se obsah antimonu a způsob ochlazení pájené slitiny významně ovlivňují pevnost svaru. Byly užity čtyři obsahy antimonu v pájce A1 0%, A2 3%, A3 5%, A4 10% (faktor *A*) a čtyři způsoby chlazení (faktor *B*): B1 vodním kalením, B2 olejovým kalením, B3 proudem vzduchu, B4 chlazení plamenem. Každé měření bylo 3× opakováno. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda obsah antimonu a způsob chlazení významně ovlivňují pevnost svaru, tj. tlak nutný k přerušení svaru v MPa. Existuje významná interakce mezi obsahem antimonu a způsobem chlazení svaru?

Data: Tlak k přerušení svaru [MPa] pro 4 obsahy Sb [hm.%] A1 až A4 a 4 způsoby chlazení B1 až B4.

Obsah Sb	B1	B2	B3	B4
A1 (0 %)	17.6 19.5 18.3	20.0 24.3 21.9	18.3 19.8 22.9	19.4 19.8 20.3
...	...	...	...	...
A4 (10 %)	15.2 17.1 16.6	16.4 19.0 18.1	15.8 17.3 17.1	16.4 17.6 17.6

Úloha H5.05 *Vliv času a druhu kovů na tlakové pnutí v sintrované oceli (ANOVA2B)*
Byl vyšetřován vliv dvou druhů kovu A1 a A2 (faktor *A*) a vliv času sintrování na dvou úrovních B1 a B2 (faktor *B*) na vnitřní tlakové pnutí (v tisících liber na čtvereční palec) sintrovaných kovů. Pět vzorků bylo sintrováno při dvou rozličných časech a se dvěma rozličnými kovy. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda má doba sintrování a druh sintrovaného kovu statisticky významný vliv na tlakové pnutí. Existuje interakce mezi oběma faktory? Pokuste se tuto interakci fyzikálně vysvětlit. Určete vhodnou mocninnou transformaci, eliminující interakci.

Data: Tlakové pnutí kovů [10^3 liber/palec^2] pro různé druhy kovu A1 aA2 a různé časy sintrování B1 a B2 [minut].

	B1 (Sintrování 100 minut)	B2 (Sintrování 200 minut)
A1 (Kov 1)	17.1 16.5 14.9 15.2 16.7	19.4 18.9 20.1 17.2 20.7

A2 (Kov 2)	12.3 13.8 10.8 11.6 12.1	15.6 17.2 16.7 16.1 18.3

Úloha H5.06 *Vliv tří faktorů na množství popela v uhlí (ANOVA3B)*
Tři faktory mají vliv na množství popela v uhlí, tj. maximální velikost částeček uhlí A1 až A4 (faktor *A*), druh uhlí B1 až B4 (faktor *B*), a velikost sledované navážky C1 až C4 (faktor *C*). Každý experiment byl 3× opakován. Sestrojte tabulku třífaktorové analýzy rozptylu. Na hladině významnosti $\alpha = 0.05$ testujte statistickou významnost jednotlivých faktorů. Existuje nějaká interakce mezi faktory?
Data: Procento popela v uhlí pro maximální velikost částeček A1 až A4, pro různé druhy uhlí B1 až B4 a velikost navážky C1 až C4.

		B1 (uhlí Mojiri)	B2 (uhlí Michel)	B3 (uhlí Kairan)	B4 (metalurgické uhlí)
A1 (246 μm)	C1(1 g)	7.30 7.35 7.42	10.69 10.58 10.72	12.20 12.27 12.23	9.99 10.02 9.95
	C2 (100 mg)	6.84 6.07 6.91	10.26 10.35 10.42	11.85 11.85 12.05	9.45 9.86 9.78
	C3 (20 mg)	7.05 6.49 7.24	10.61 10.08 10.31	12.34 11.74 11.44	9.76 9.79 9.77
	C4 (5 mg)	6.75 5.62 7.24	10.66 10.61 10.01	12.22 11.68 12.09	9.92 10.17 10.50
...	...	...	...	...	...
A4 (48 μm)	C1 (1 g)	7.45 7.49 7.47	10.85 10.89 10.85	12.23 12.30 12.17	10.06 10.07 10.11
	C2 (100 mg)	7.15 7.68 7.18	10.37 10.79 10.71	11.52 12.17 11.82	9.71 9.86 9.78
	C3 (20 mg)	7.60 7.55 6.61	10.82 10.82 10.88	12.40 11.99 12.17	10.13 9.93 10.01
	C4 (5 mg)	8.06 7.05 7.57	11.26 10.56 10.31	11.96 11.87 12.06	10.01 9.98 9.84

Úloha H5.07 *Vliv šarže míchaného betonu na tlakové pnutí (ANOVA1)*
Tlakové pnutí betonu je závislé na poměru vody a cementu, na době promíchání, důkladnosti promíchání atd. I když zkušený pracovník se snaží dodržet všechny faktory na konstantních hodnotách, vyskytnou se vždy jemné odchylky od šarže k šarži A1 až A6 (faktor *A*). Existuje šarže s velmi odlehlými hodnotami tlakového pnutí od ostatních šarží? Vezměte 1. šarži za kontrolní a ostatní porovnejte.
Data: tlakové pnutí pro náhodnou šarži betonu A1 až A6.

A1	A2	A3	A4	A5	A6
5.01	4.74	4.99	5.64	5.07	5.90
...	...	...	...	...	...
5.37	4.80	4.77	5.17	5.48	5.39

Úloha H5.08 *Vliv lokality na hustotu horniny v dole (ANOVA1)*
Cílem studie mineralogických vlastností černého jílu a jemu podobného bláta bylo odhalit homogenitu této horniny v dole a předejít katastrofám, skluzům jílovité horniny a závalům. Vzorky horniny byly odebrány ze tří rozličných kritických lokalit dolu A1 až A3 (faktor *A*) a byla stanovena hustota horniny v kg/m^3. Na hladině významnosti $\alpha = 0.05$ vyšetřete homogenitu horniny v dole. Dosahuje některá lokalita odlišných hodnot než zbývající dvě?
Data: Hustota horniny [kg/m^3] ve třech lokalitách A1 až A3.

Lokalita A1	Lokalita A2	Lokalita A3
2.06	2.09	2.07
...	...	...

2.00	2.41	2.64

Úloha H5.09 *Vliv sady na velikost měrného odporu krystalu (ANOVA1)*
Měrný odpor křemíkových monokrystalů byl sledován u 8 náhodně vybraných krystalů, jež pocházely z 5 rozličných sad. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda sada A1 až A5 (faktor *A*) ovlivňuje velikost měrného odporu krystalu. Stanovte, zda existuje významný rozdíl mezi sadou č. 1 a č. 4. Existuje nějaká silně odlišná sada?
Data: Velikost měrného odporu krystalu [S cm^{-1}] pro pět sad odporů A1 až A5.

Sada	Reprodukovaná měření							
A1	2.8	2.7	2.3	2.6	2.7	2.3	2.7	2.7
...	...	...	...	...	...	...	...	...
A5	3.1	3.3	2.9	2.5	2.5	3.1	2.5	3.0

Úloha H5.10 *Vliv volby dne a užitého nákladního auta na obsah síry v uhlí (ANOVA2B)*
Byl sledován obsah síry v uhlí. Z pěti dodávek nákladními auty A1 až A5 (faktor *A*) se vzaly 2 náhodné kousky uhlí vždy po dobu pěti náhodně vybraných dnů B1 až B5 (faktor *B*) a ve vzorcích byl opakovaně stanoven obsah síry v procentech. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda je obsah síry ovlivněn volbou nákladního auta nebo volbou dne. Lze vysvětlit a dokázat interakci obou faktorů?
Data: Obsah síry [%] z pěti nákladních aut A1 až A5 a po dobu pěti dnů B1 až B5.

	B1 (1. den)	B2 (2. den)	B3 (3. den)	B4 (4. den)	B5 (5. den)
A1 (1. auto)	0.107 0.105	0.091 0.089	0.110 0.113	0.088 0.092	0.089 0.088
...	...	...	...	...	...
A5 (5. auto)	0.108 0.104	0.092 0.090	0.106 0.109	0.091 0.088	0.086 0.089

Úloha H5.11 *Vliv tavby na obsah mědi v bronzu (ANOVA1)*
Bylo zkoumáno, zda se obsah mědi v bronzu mění od tavby k tavbě. U každé tavby byly odebrány 4 vzorky a stanoven procentuální obsah mědi v bronzu. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda existuje vliv tavby A1 až A5 (faktor *A*) na obsah mědi v bronzu. Ověřte rovněž normalitu a homoskedasticitu výběrového rozdělení vhodnou grafickou metodou.
Data: Obsah mědi v bronzu [%] pro různé tavby A1 až A5.

A1	A2	A3	A4	A5
81	85	87	94	88
...	...	...	...	...
84	90	87	91	90

Úloha H5.12 *Vliv tavby na pórovitost měděné slitiny (ANOVA1)*
"Sintrování" je proces, kdy se práškový kov za vyšší teploty převede do stavu pórovité slitiny. Byla sledována pórovitost taveb sintrované mědi tak, že u 3 vzorků bylo měřeno procento prázdného prostoru pórů slitiny. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda má tavba A1 až A6 (faktor *A*) vliv na pórovitost měděné slitiny. Dosáhla některá tavba silně odlišných hodnot od ostatních taveb?
Data: Prázdný prostor v pórech měděné slitiny [%] pro 6 taveb A1 až A6.

A1	A2	A3	A4	A5	A6

21	24	19	27	22	27
23	25	22	24	20	26
21	23	20	25	20	23

Úloha H5.13 *Vliv oblasti jezera a vzorku na koncentraci fosforu ve vodě (ANOVA2B)*
Byla stanovována koncentrace fosforu v mg/l ve vodě velkého jezera na severu USA. Jezero bylo rozděleno na pět oblastí B1 až B5 (faktor *B*) a z každé oblasti odebrány tři vzorky A1 až A3 (faktor *A*). U každého vzorku byla provedena dvě opakovaná stanovení koncentrace fosforu. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda oblasti a odebrané vzorky mají významný vliv na koncentraci fosforu ve vodě jezera. Dochází k vzájemnému ovlivnění těchto faktorů?

Data: Koncentrace fosforu ve vodě jezera [mg/l] v oblastech B1 až B5 a pro vzorky A1 až A3.

	Oblast B1	Oblast B2	Oblast B3	Oblast B4	Oblast B5
A1	0.010 0.008	0.013 0.017	0.009 0.015	0.011 0.015	0.014 0.006
A2	0.009 0.012	0.008 0.010	0.010 0.014	0.008 0.013	0.018 0.010
A3	0.011 0.006	0.012 0.011	0.017 0.011	0.010 0.014	0.005 0.013

Úloha H5.14 *Vliv hloubky moře na hloubku proniknutí kyslíku do sedimentu (ANOVA1)*
Dodávka kyslíku pobřežních mořských sedimentů je velmi důležitá pro existenci živočichů v moři. Cílem studie bylo vyšetření průměrné hloubky při infiltrování kyslíku do pobřežních mořských sedimentů v mm, měřených v pěti rozličných hloubkách mořské vody [m]: A1 = 5, A2 = 10, A3 = 15, A4 = 20 a A5 = 40 m (faktor *A*). Kyslíkovou elektrodou byl stanoven obsah kyslíku v mořských sedimentech s pětinásobnou reprodukovatelností. Existuje důkaz rozdílnosti proniknutí kyslíku do mořských sedimentů v pěti hloubkách? Liší se významně výsledky pro hloubku moře A1 = 5 m a A5 = 40 m? Vykazují výběry normalitu rozdělení?

Data: Hloubka proniknutí kyslíku do mořského sedimentu [mm] v pěti hloubkách moře A1 až A5.

A1 (5 m)	A2 (10 m)	A3 (15 m)	A4 (20 m)	A5 (40 m)
1.2	3.7	1.7	2.8	4.4
...	...	...	...	...
2.4	3.0	2.0	3.1	3.9

Úloha H5.15 *Test homogenity hořčíkové slitiny (ANOVA2B)*
K vyšetření homogenity hořčíkové slitiny byl proveden následující experiment: ingot byl vytažen na 100 m dlouhou tyč čtvercového průměru o straně asi 4.5 cm. Tyč byla rozřezána na 100 sloupků o délce 1 m. Pět náhodně vybraných sloupků B1 až B5 (faktor *B*) bylo podrobeno analýze tak, že z každého sloupku byl uříznut testační proužek o délce 1.2 cm. Na testačním proužku bylo naznačeno 9 bodů k testování A1 až A9 (faktor *A*) a v každém bodě byla provedena dvě stanovení obsahu hořčíku v procentech. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda je materiál dokonale homogenní či zda záleží na poloze na 100 m tyči, odkud byly vzorky odebrány.

Data: Obsah hořčíku ve slitině [1000 .%] v pěti sloupcích B1 až B5 a v 9 polohách na sloupku A1 až A9.

Poloha	B1 (Sloupek 1)	B2 (Sloupek 5)	B3 (Sloupek 20)	B4 (Sloupek 50)	B5 (Sloupek 85)
A1	76 67	69 66	73 70	73 63	70 69
...	...	...	...	...	...

A9	66 63	70 65	67 73	69 66	69 68

Úloha H5.16 *Shodnost obsahu tantalu třemi instrumentálními metodami (ANOVA1)*
Ke stanovení tantalu bylo použito tří metod (faktor *A*): A1 nukleární metody, A2 hmotnostní spektroskopie a A3 rentgenová fluoroscenční spektrometrie. Na hladině významnosti $\alpha = 0.05$ testujte, zda všechny tři metody vedou ke stejnému výsledku. Je splněn předpoklad výběrové normality a homoskedasticity? Existují statisticky významné rozdíly mezi metodou A1 a A3?

Data: Obsah tantalu ve vzorku [ppm] třemi analytickými metodami A1 až A3.

A1	68.00 78.72 79.79 80.90 81.43 85.30 86.50 89.00 92.70 100.00
A2	32.50 77.16 82.00 83.00 85.63 90.59 93.00
A3	56.00 65.00 74.00 82.00 82.33 86.17

Úloha H5.17 *Shodnost obsahu lanthanu v cinvalditu čtyřmi metodami (ANOVA1)*
Pro stanovení lanthanu v cinvalditu bylo použito čtyř metod (faktor *A*): A1 hmotnostní spektroskopie, A2 rentgenová fluoroscenční spektrometrie, A3 emisní spektroskopie a A4 nukleární metody. Na hladině významnosti $\alpha = 0.05$ testujte, zda všechny čtyři metody vedou ke stejnému výsledku. Ověřte normalitu a homoske-dasticitu rozdělení výběru. Dosahuje některá metoda silně odlišných výsledků?

Data: Obsah lanthanu v cinvalditu [ppm] čtyřmi analytickými metodami A1 až A4.

A1	27.10 27.59 27.70 29.00 29.35 30.00 30.00 30.40 32.00 33.10 33.28 33.60 33.70 35.40
...	...
A4	24.25 25.45 26.35 27.34 28.00 28.00 29.30 29.70 29.90 30.00 33.00

Úloha H5.18 *Shodnot obsahu ceru čtyřmi metodami (ANOVA1)*
Pro stanovení ceru v neznámém vzorku horniny bylo použito čtyř metod (faktor *A*): A1 hmotnostní spektroskopie, A2 rentgenová fluoroscenční spektrometrie, A3 emisní spektroskopie a A4 nukleární metody. Na hladině významnosti $\alpha = 0.05$ testujte, zda všechny čtyři metody A1 až A4 vedou ke stejnému výsledku. Povede použití mocninné transformace k homogenitě rozptylů? Ověřte předpoklad výběrové normality a homoskedasticity. Která metoda dosáhla silně odlišných výsledků od ostatních metod?

Data: Obsah ceru v neznámém vzorku [ppm] čtyřmi analytickými metodami A1 až A4.

A1	92.90 93.00 96.00 96.00 96.50 97.06 97.29 97.50 102.10 102.70 105.00 105.20 107.00 116.40
...	...
A4	72.00 75.00 84.42 89.11 96.50 97.45 104.40 106.00 106.00 109.00

Úloha H5.19 *Test stálosti absorbance barevného roztoku (ANOVA1)*
Vyšší obsahy oxidu železitého lze stanovit měřením absorbance fialového komplexu železitých iontů s EDTA a H_2O_2 v amoniakálním prostředí. Po určité době se peroxid vodíku H_2O_2 rozkládá; bublinky kyslíku zvyšují absorbanci roztoku. Absorbance byla proto 3× opakovaně změřena po různé době od přípravy roztoku. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda lze považovat absorbanci za konstantní a nezávislou na čase A1

až A4 (faktor *A*). Zjistěte přípustný čas k proměření absorbance. Liší se významně výsledky pro stanovení ihned A1 a po 4 hodinách A4?

Data: Absorbance fialového komplexu v závislosti na čase A1 až A4.

A1 (Ihned)	0.423	0.429	0.415
...	...	...	...
A4 (Po 4 hodinách)	0.522	0.677	0.653

Úloha H5.20 *Vliv teploty výpalu a keramické suroviny na ztrátu hmotnosti pálením (ANOVA2B)*. Při čtyřech teplotách A1 950 EC, A2 1000 EC, A3 1050 EC a A4 1100 EC (faktor *A*) byly vypalovány tři suroviny B1 až B3 (faktor *B*). Na výpalcích bylo provedeno zjištění ztráty hmotnosti pálením v procentech. Stanovení ztráty hmotnosti bylo provedeno u všech tří surovin pro každou teplotu třikrát. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda má teplota výpalu a surovina vliv na ztrátu hmotnosti pálením. Lze vysvětlit a dokázat interakci obou faktorů?

Data: Ztráta hmotnosti pálením [%] při 4 teplotách A1 až A4 pro 3 různé typy suroviny B1 až B3.

Teplota	Surovina B1			Surovina B2			Surovina B3		
A1 (950E)	7.1	6	7.3	10.8	10.6	9.5	11.3	11.7	13
...	...	...	...	...	...	...	...	...	...
A4 (1100E)	7.5	6.2	7.6	11.3	10.9	9.7	11.6	12	13.6

Úloha H5.21 *Vliv šarže vápence a analytické metody na obsah CaO (ANOVA2P)*
Z technologického procesu úpravy horniny, vápence, bylo odebráno 5 šarží v průběhu jednoho týdne A1 až A5 (faktor *A*), které byly homogenizovány a analyzovány z hlediska obsahu oxidu vápenatého, CaO (hm.%), a to čtyřmi metodami (faktor *B*): B1 gravimetrickou, B2 titrační, B3 rentgenová fluorescenční a B4 plamenové atomové absorpční spektrometrie. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda šarže vápence a zvolená analytická metoda mají významný vliv na obsah CaO ve vápenci. Vede některá analytická metoda k silně odlišným výsledků?

Data: Obsah CaO [hm.%] v pěti šaržích vápence A1 až A5 a čtyřmi metodami B1 až B4.

Šarže	B1	B2	B3	B4
A1	50.9	49.5	47.4	50.3
...	...	...	...	...
A5	53.0	47.2	48.9	52.2

5.6.5 Analýza ekonomických a sociologických dat

Úloha S5.01 *Vliv člověka na výsledek analýzy (ANOVA1)*
Pět chemiků A1 až A5 (faktor *A*) provedlo analytické stanovení s rozličnými výsledky. Každy chemik provedl jiný počet opakovaných pokusů. Vyšetřete, zda všichni dosáhli stejného výsledku. Existuje chemik, který dosáhl silně odlišných výsledků?

Data: Obsah neznámé látky [%], určený pěti chemiky A1 až A5.

A1	A2	A3	A4	A5
30.0	29.3	29.6	32.5	31.0
...	...	...	...	...
-	30.0	-	-	-

Úloha S5.02 *Vliv odhadce a druhu odhadu na výši odhadu (ANOVA2P)*
Firma zaměstnává tři inženýry - odhadce (faktor *B*), kteří jsou v práci vzájemně nahraditelní. Aby se otestovala jejich opravdová nahraditelnost při plnění téhož úkolu, byli vyzváni aby provedli šest typů odhadu (faktor *A*) v tisících US $. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda odhad je ovlivněn zvoleným inženýrem B1 až B3, nebo zda je ovlivněn typem odhadu A1 až A6. Ověřte vhodnost užití Tukeyova modelu interakce. Dosáhl některý odhadce silně odlehlých výsledků od ostatních dvou?
Data: Odhad [US $] pro různé typy odhadu A1 až A6 a provedený třemi odhadci B1 až B3.

Typ odhadu	Odhadce B1	Odhadce B2	Odhadce B3
A1	27.3	26.5	28.2
...	...	...	...
A6	58.7	59.2	60.1

Úloha S5.03 *Vliv těsnění a tiskařského stroje na počet vyrobených kusů (ANOVA2B)*
Produkce určitého výrobku je ovlivněna tiskářským strojem A1 a A2 (faktor *A*) a těsnícím materiálem B1 až B3 (faktor *B*). Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda počet výrobků je ovlivněn druhem těsnění nebo tiskařským strojem. Existuje statisticky významná interakce obou faktorů a má logický smysl?
Data: Počet výrobků [tisíce ks] na dvou tiskařských strojích A1 a A2 při užití 3 druhů těsnění B1 až B3.

	B1 (Korkové těsnění)	B2 (Gumové těsnění)	B3 (Plastické těsnění)
Tiskařský stroj A1	4.31 4.27 4.40	3.36 3.42 3.48	4.01 3.94 3.89
Tiskařský stroj A2	3.94 3.81 3.99	3.91 3.80 3.85	3.48 3.53 3.42

Úloha S5.04 *Vliv typu managementu na hodinovou mzdu zaměstnance (ANOVA1)*
V USA jsou dva typy managementu: u prvního A1 šéfové věří, že pracující jsou v podstatě leniví a nedůvěryhodní, u druhého A2 věří svým pracovníkům, považují je za pilné a pracovité, závislé na silných individualitách. Japonci A3 však prosazují třetí směr: širokospektré plánování, konsenzus rozhodování-výroba, oboustrannou loajalitu zaměstnanec-zaměstnavatel. Cílem je porovnat hodinové mzdy podobných inženýr-ských firem se třemi typy managementu (faktor *A*), když od každého typu bylo náhodně vybráno 6 zaměstnanců. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda lze považovat mzdy u tří managementů za stejné.
Data: Hodinová mzda zaměstnance u tří typů managementu [US $].

	Hodinová mzda [US $]
Management A1	5.20 5.20 6.10 6.00 5.75 5.60
Management A2	6.25 6.80 6.87 7.10 6.30 6.35
Management A3	5.50 5.75 4.60 5.36 5.85 5.90

Úloha S5.05 *Vliv typu auta na počet ujetých mil na 1 galon benzinu (ANOVA2B)*

Taxikářská firma se rozhoduje o větším nákupu osobních aut, vhodných pro taxi službu. Vybírá mezi pěti značkami aut, jež jsou srovnatelné co do pořizovací ceny a co do měsíční údržby. Bylo proto testováno několik vozů, dva A1 a A2 (faktor *A*) od každé značky. Rozhodnutí o nákupu značky B1 až B5 (faktor *B*) padne až podle spotřeby benzinu, tzn. počtu mil ujetých na 1 galon benzinu. Každé auto bylo testováno 3×. Vyšetřete na hladině významnosti $\alpha = 0.05$, zda jsou oba vozy od dané značky stejné a zda všechny typy aut jsou stejné co do počtu ujetých mil na 1 galon benzinu. Stanovte, zda je významný rozdíl mezi auty různých značek za předpokladu, že v rámci téže značky považujeme auta za stejná.

Data: Počet mil na 1 galon benzinu pro dvě auta A1 a A2 a pro pět značek aut B1 až B5.

	B1	B2	B3	B4	B5
A1. auto:	15.8 15.6 16.0	18.5 18.0 18.4	12.3 13.0 12.7	19.5 17.5 19.1	16.0 15.7 16.1
A2. auto:	13.9 14.2 13.5	17.9 18.1 17.4	14.0 13.1 13.5	18.7 19.0 18.8	15.8 15.6 16.3

Úloha S5.06 *Vliv typu auta a města na počet mil na 1 galon benzinu (ANOVA2P)*
Firma zamýšlí koupit svým zaměstnancům auta pro služební cesty. Je testováno 5 konkrétních modelů A1 až A5 (faktor *A*) v 5 náhodně vybraných městech B1 až B5 (faktor *B*) co do počtu ujetých mil na 1 galon benzinu. Je vliv aut charakteru náhodného efektu nebo pevného efektu? Je vliv měst charakteru náhodného efektu nebo pevného efektu? Jak je formulována nulová hypotéza, která zajímá hodnotitele? Jsou rozdíly ve vzdálenosti mil, ujetých jednotlivými auty? Testování proveďte na hladině významnosti $\alpha = 0.05$. Ověřte graficky aditivitu obou faktorů.

Data: Počet mil na 1 galon benzinu pro 5 typů aut A1 až A5 v 5 městech B1 až B5.

	Město B1	Město B2	Město B3	Město B4	Město B5
Auto A1	15.83	17.56	21.11	20.48	26.04
...	...	...	...	...	...
Auto A5	21.24	21.29	20.34	19.43	25.05

Úloha S5.07 *Vliv času a stánku na počet prodaných kusů románu (ANOVA2P)*
Marketingový expert chce změřit čtenářskou oblibu vázaného a barevně ilustrovaného románu. Bylo vybráno pět novinových stánků A1 až A5 (faktor *A*) a román zde byl vystaven po dobu pěti týdnů B1 až B5 (faktor *B*). Období týdne se totiž jevilo postačující ke zjištění potencionální poptávky. Na hladině významnosti $\alpha = 0.05$ zjištěte vliv času a vliv novinového stánku na prodej vybraného románu. Existuje interakce mezi oběma faktory a má logické vysvětlení? Ověřte i graficky.

Data: Počet prodaných knih vybraného románu [ks] v 5 stáncích A1 až A5 po dobu 5 týdnů B1 až B5.

	B1 (1. týden)	B2 (2. týden)	B3 (3. týden)	B4 (4. týden)	B5 (5. týden)
Stánek A1	200	290	280	230	265
...	...	...	...	...	...
Stánek A5	340	335	265	270	230

Úloha S5.08 Vliv různých způsobů cesty do práce na potřebnou dobu cestování *(ANOVA1)*.
Mezi místy *A* a *B* jezdí spoj A1 (jedna tramvajová linka) a spoj A2 (jedna autobusová linka), doprava je též možná spojem A3 (metro s přestupem na tramvaj). V době ranní

špičky byl při cestě do práce šestkrát použit A1 ($n_1 = 6$), pětkrát A2 ($n_2 = 5$), a sedmkrát A3 ($n_3 = 7$). Naměřená doba cestování včetně čekání na příslušný spoj je v minutách. Je třeba vyšetřit, zda je doba ranního cestování mezi uvedenými místy u všech spojů (faktor *A*) stejná, tzn. za daných podmínek tato doba nezávisí na použitém spoji. Liší se významně výsledky spojů A1 a A3?

Data: Doba cesty do práce [minut] různými dopravními prostředky A1 a A3.

Spoj A1	32, 39, 42, 37, 34, 38
Spoj A2	30, 34, 28, 26, 32
Spoj A3	40, 37, 31, 39, 38, 33, 34

Úloha S5.09 *Vliv druhů osvětlení a hlučnosti na potřebný čas k výpočtu (ANOVA2B)*
Zajímá nás vliv hlučnosti (faktor *A*) na úrovních absolutní ticho A1, hluk z ulice A2, hlasitá reprodukce hudby A3 a dále vliv osvětlení (faktor *B*) na úrovních přímé denní světlo B1, osvětlení stolní lampou B2 a stropní osvětlení B3 na čas, potřebný k provedení určitého příkladu elektronickou kalkulačkou. Bylo vybráno 18 pracovnic výpočetního centra a každá z nich nezávisle na ostatních řešila stejnou výpočetní úlohu. Pracovnice byly náhodně rozděleny mezi kombinace úrovní sledovaných faktorů tak, že každá kombinace byla přidělena vždy dvěma z nich. Doba v minutách k vyřešení úlohy je v datech. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda uvedené faktory mají významný vliv na sledovanou dobu výpočtu.

Data: Doba výpočtu [minuty] pro 3 úrovně hlučnosti A1, A2, A3 a 3 typy osvětlení B1, B2, B3.

	B1	B2	B3
A1	3 4	5 3	4 6
A2	3 1	3 5	5 5
A3	2 4	6 5	5 4

Úloha S5.10 *Vliv obalu a vzorku krekeru na jeho zvlhnutí (ANOVA2B)*
Sušenky a krekery ztrácejí svou charakteristickou křupavost, když špatným skladováním zvlhnou. Vedle starého balení v tvrdém kartonu (obal B1) byly testovány i nově navržené obaly, a to krabice: z voskovaného papíru (obal B2), s kovovou fólií (obal B3), plastická krabice (obal B4) a kombinovaná plastická krabice s kovovou fólií (obal B5). Zboží bylo vystaveno 24 hodinovému působení 80 % vlhkosti a pak byly z každé krabice náhodně odebrány čtyři krekry a stanoven obsah vody v mg, který za 24 hodin adsorbovaly. Měření bylo 3× opakováno. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda záleží na druhu krekeru A1 až A4 (faktor *A*) a zda se funkce obalu krabic B1 až B5 (faktor *B*) liší v izolaci vůči vlhkosti. Existuje obal, který dosahuje silně odlišných výsledků od ostatních obalů?

Data: Obsah vody [mg/ks] u 4 druhů kreker A1 až A4 a pro 5 různých obalů B1 až B5.

	Obal B1	Obal B2	Obal B3	Obal B4	Obal B5
A1. vzorek	73 75 77	60 61 63	46 48 46	60 53 60	38 37 38
...	...	...	...	...	...
A4. vzorek	67 69 62	53 50 55	58 53 53	52 55 49	48 47 46

Úloha S5.11 *Vliv denní doby na koncentraci a výkon pracovníka (ANOVA1)*
Pro posouzení výkonnosti pracovníka během dne byl proveden experiment, při němž byla

v různou denní dobu A1 až A5 (faktor *A*) u 6 pokusných osob měřena schopnost koncentrace, tj. počet správně provedených úkonů za standardních podmínek. Proveďte analýzu rozptylu k ověření hypotézy, že podmínky pro koncentraci pracovníka nemají souvislost s denní dobou. Porovnejte výsledky z rána A1 a z noci A5. Ve kterou denní dobu je dosaženo nejvyšších výkonů a ve kterou silně odlišných výkonů od ostatní denní doby?

Data: Výkon [počet správně provedených úkonů] v různých dobách pracovního dne A1 až A5.

Část dne	Výkon
A1 (Ráno)	162 162 150 151 164 155 155
...	...
A5 (V noci)	148 150 160 156 159 156 163

Úloha S5.12 *Vliv druhu osvětlení na směnový výkon dělníků (ANOVA1)*

Pokuste se prokázat, že směnový výkon dělníků závisí na osvětlení pracoviště A1 až A3 (faktor *A*), za předpokladu normality a shody rozptylů. Výsledky jsou zaznamenány u 16 náhodně vybraných osob. Analýzu rozptylu proveďte na hladině významnosti $\alpha = 0.05$.

Data: Směnový výkon dělníků za rozličného osvětlení A1 až A3.

Osvětlení	Směnový výkon dělníků
A1 (Přímé)	64 54 60 50
A2 (Kombinované)	59 67 72 69 74 67
A3 (Nepřímé)	60 63 57 66 62 64

Úloha S5.13 *Vliv textu a čtenáře na hodnotu indexu čitelnosti textu (ANOVA2P)*

Text lze posuzovat dle indexu čitelnosti, přijatelnosti a pochopitelnosti pro čtenáře. Pomocí pěti čtenářů A1 až A5 (faktor *A*) byly testovány čtyři rozličné texty B1 až B5 (faktor *B*). Na hladině významnosti $\alpha = 0.05$ testujte, zda všechny čtyři texty vedou ke stejnému indexu čitelnosti a zda všech pět čtenářů má stejný názor. Povede mocninná transformace ke zlepšení aditivity těchto faktorů? Existuje statisticky významná interakce obou faktorů a má logický smysl?

Data: Index čitelnosti u 5 čtenářů A1 až A5 na čtyřech rozličných textech B1 až B4.

	B1	B2	B3	B4
Student A1	50	59	48	60
...	...	...	...	...
Student A5	53	61	50	70

Úloha S5.14 *Vliv kvality předešlé školy na výsledek zkoušky z aritmetiky (ANOVA1)*

Test znalosti žáků z aritmetiky je do značné míry ovlivněn úrovní znalostí z předešlé školy. Na hladině významnosti $\alpha = 0.05$ je třeba testovat, je-li vliv dosažených znalostí předešlé školy A1 až A3 (faktor *A*) opravdu významný. Liší se významně vysoká A1 a nízká A3 kvalita předešlé školy? Je splněn předpoklad normality rozdělení každého výběru?

Data: Dosažené skóre u zkoušky z aritmetiky pro rozličné znalosti z předešlé školy A1 až A3.

Kvalita předešlé školy	Dosažené skóre
A1 (Vysoká)	90 86 88 93 80
A2 (Střední)	80 70 61 52 73 65 83
A3 (Nízká)	60 60 55 62 50 70

Úloha S5.15 *Vliv počtu let studií na celoživotní příjem v dolarech (ANOVA1)*
Dosažené vzdělání značně ovlivňuje finanční příjem v zaměstnání. Byl proveden náhodný výběr zaměstnanců a data přinášejí souvislost mezi dosažených vzděláním či počtem let strávených na studiích A1 až A5 (faktor *A*) a celoživotním příjmem v tisících US $. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda vzdělání opravdu významně ovlivňuje příjem. Pro jaký počet let studií bylo dosaženo silně odlišného finačního příjmu?
Data: Celoživotní příjem [1000 dolarů] v závislosti na počtu let na studiích A1 až A5.

A1 (8 let a méně)	A2 (9 - 11 let)	A3 (12 let)	A4 (13 - 15 let)	A5 (16 let a více)
300	270	400	420	570
-	-	-	-	660

Úloha S5.16 *Vliv programátora na počet chyb v programu (ANOVA1)*
Byla sledována chybovost počítačového programátora. V náhodně vybraných dnech byl počítán počet chyb při sestavování programu u čtyř testovaných programátorů A1 až A4 (faktor *A*). Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda chybovost testovaných programátorů je shodná, či zda se liší. Lze přijmout předpoklad stejné přesnosti programátorů, vyjádřené rozptylem? Který programátor dosáhl silně odlišných výsledků od ostatních programátorů. Jsou splněny výběrové předpoklady?
Data: Počet chyb v programu pro jednotlivé programátory A1 až A4.

Programátor A1	14 16 18 14 22 9
Programátor A4	16 18 20 17 21

Úloha S5.17 *Vliv rozvodovosti na sociální postavení rodiny (ANOVA1)*
V sociologickém průzkumu bylo vyšetřováno sociální postavení, které bylo vyjádřeno v bodech. Rodiny byly rozděleny na stabilní, rozvedené a rodiny v přechodu A1 až A3 (faktor *A*). Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda všechny rodiny pocházejí ze statistického hlediska svého sociálního postavení z jednoho souboru. Existují významné rozdíly mezi stabilními a rozvedenými rodinami?
Data: Vyjádření sociálního postavení [body] v závislosti na typu postavení rodiny A1 až A3.

A1 (Stabilní)	A2 (Rozvedené)	A3 (V přechodu)
108	113	92
...	...	...
118	114	105

Úloha S5.18 *Vliv reklamy a dne v týdnu na počet kusů prodaného zboží (ANOVA2P)*
Velkosklad sledoval počet kusů prodaného zboží ve dnech v týdnu A1 až A5 (faktor *A*) a v závislosti na působení reklamy na zboží v novinách, v rádiu a konečně bez reklamy B1 až B3 (faktor *B*). Na hladině významnosti $\alpha = 0.05$ testujte, zda je typ reklamy účinný a zvýší počet kusů prodaného zboží a zda je vliv dnů v týdnu významný. Je typ reklamy ovlivněn dnem v týdnu?
Data: Počet kusů prodaného zboží ve dnech týdne A1 až A5 a dle uveřejněné reklamy B1 až B3.

	B1 (V novinách)	B2 (V rádiu)	B3 (Bez reklamy)
A1 (Pondělí)	845	811	612
...	...	...	...

A5 (Pátek)	806	872	666

Úloha S5.19 *Vliv pěti pedagogických metod na studijní výsledky studentů (ANOVA1)*
Studenti byli vyučováni předmětu za využití pěti pedagogických metod (faktor *A*): A1 tradiční způsob, A2 programová výuka, A3 audio technika, A4 audiovizuální technika a A5 vizuální technika. Z každé skupiny byl vybrán náhodný vzorek studentů a všichni byli podrobeni stejnému písemnému testu. Jsou znalosti všech studentů stejné, nezávislé na užité pedagogické metodě? Testujte na hladině významnosti $\alpha = 0.05$. Je variabilita v datech dokonale popsána jediným faktorem? Lze přijmout předpoklad stejné variability metod, vyjádřené rozptylem? Vykazují výběry normální rozdělení?
Data: Bodové hodnocení studentů [body] vyučovaných pěti pedagogických metod A1 až A5.

A1 (Tradiční)	A2 (Programová)	A3 (Audio)	A4 (Audiovizuální)	A5 (Vizuální)
76.2	85.2	67.3	75.8	50.5
...	...	...	...	...
-	60.4	-	-	-

Úloha S5.20 *Vliv typu stroje na hodinový výkon (ANOVA1)*
Při normování určitého druhu strojové operace se má rozhodnout, zda je třeba rozlišovat zpracování na jednotlivých typech strojů A1 až A4 (faktor *A*). Hodinový výkon těchto strojů značně kolísá vlivem lidské obsluhy a vlivem dalších neodstranitelných příčin. Na náhodně vybraných strojích typů A1 až A4 byly zjištěny hodinové výkony [ks]. Na hladině významnosti $\alpha = 0.05$ vyšetřete, zda se hodinové výkony na jednotlivých typech strojů významně liší. Bylo dosaženo silně odlišného hodinového výkonu u některého stroje?
Data: Hodinový výkon stroje [ks] pro čtyři typy stroje A1 až A4.

Typ stroje	Hodinový výkon [ks]
A1	515 860 710 720 655 670 645 570 685
...	...
A4	870 785 765 890 715

Úloha S5.21 *Vliv zkušebny na naměřené velikosti pevnosti drátu v tahu (ANOVA1)*
Pro porovnání práce tří zkušeben A1 až A3 (faktor *A*) byl proveden experiment: z drátu o průměru 2.4 mm byly vzaty vzorky a každé ze tří zkušeben bylo dáno 10 vzorků. V každé zkušebně byly na jednom stroji provedeny zkoušky. V datech jsou uvedeny hodnoty pevnosti v tahu [kg /mm^2]. Zjistěte, zda existují významné rozdíly mezi výsledky jednotlivých zkušeben. Vykazují výběry normální rozdělení?
Data: Hodnoty pevnosti v tahu [kg /mm^2] dosažené ve třech zkušebnách A1 až A3.

A1	A2	A3
194.6	190.2	194.5
...	...	...
194.6	191.3	193.4

Úloha S5.22 *Vliv výšky regálu na maximální denní teplotu uskladněného materiálu*

(ANOVA2P). Ve skladu obalového materiálu je materiál uložen na 4 patrových regálových konstrukcích. Během měření byla v kritických místech skladu měřena teplota pomocí min.-max. teploměrů. Po dobu 4 dnů byly rozmístěny na jednom patře regálové konstrukce, poté byly přemístěny do dalšího patra. Každý den byla odečtena maximální teplota ve třech měřicích místech. Aplikujte na naměřené hodnoty test dvourozměrné analýzy rozptylu a vyšetřete vliv regálu A1 až A4 (faktor *A*) a vliv měřicího místa B1 až B3 (faktor *B*).

Data: Teplota uskladněného materiálu [EC] ve čtyřech regálech A1 až A4 a na třech místech B1 až B3.

A1. regál	B1: 19 19 19 20
	B2: 19 19 20 20
	B3: 20 20 20 20
...	...
A4. regál	B1: 20 22 22 21
	B2: 21 21 25 21
	B3: 20 20 20 20

Úloha S5.23 *Vliv druhu oleje a typu lokomotivy na počet ujetých kilometrů (ANOVA2U)* V lokomotivním depu jsou na 6 různých typech lokomotiv A1 až A6 (faktor *A*) používány dva druhy oleje B1 a B2 (faktor *B*). Úkolem je zjistit, zda typ lokomotiv a druh oleje mají vliv na počet ujetých kilometrů, které lokomotiva ujede mezi výměnami olejové náplně motoru. Olej se vyměňuje, jakmile viskozita překročí určitou hranici. *Lokomotivy* (faktor *A*, řádky, index *i, N* = 6): byly vybrány náhodně z parku depa. *Oleje* (faktor *B*, sloupce, index *j, M* = 2): 1. M6ADV, 2. M6ADSII+. Existuje statisticky významná interakce obou faktorů a má logický smysl?

Data: Počet ujetých kilometrů pro 6 lokomotiv A1 až A6 a 2 druhy olejů B1 a B2 a opakování *o*.

i	*j*	*o*	Počet ujetých km			
1	1	3	19300.0	54260.0	53220.0	
1	2	4	11600.0	95480.0	115660.0	142370.0
...	...	...	...	...	...	
6	2	3	65730.0	14000.0	13200.0	

5.7 Kontrolní hodnoty (ADSTAT, NCSS2000)

5.7.1 Analýza farmakologických a biochemických dat

B5.01 *ANOVA#1, F = 4.319, H_0 zamítnuta*
B5.02 *ANOVA#1, F = 5.660, H_0 zamítnuta*
B5.03 *ANOVA#1, F = 1.932, H_0 akceptována*
B5.04 *ANOVA#1, F = 17.640, H_0 zamítnuta*
B5.05 *ANOVA#1, F = 8.190, H_0 zamítnuta*
B5.06 *ANOVA#1, F = 1.545, H_0 akceptována*
B5.07 *ANOVA#1, F = 216.0, H_0 zamítnuta*
B5.08 *ANOVA#1, F = 24.325, H_0 zamítnuta*
B5.09 *ANOVA#1, F = 0.689, H_0 akceptována*
B5.10 *ANOVA#2B, F_A = 928, H_0 zamítnuta, F_B = 1154, H_0 zamítnuta, F_{AB} = 77, H_0 zamítnuta*
B5.11 *ANOVA#1, F = 2.271, H_0 akceptována*
B5.12 *ANOVA#2B, (a) F_A = 28.646, H_0 zamítnuta, F_B = 38.882, H_0 zamítnuta, F_{AB} = 6.473, H_0 zamítnuta, (b) F_A = 22.647, H_0 zamítnuta, F_B = 40.247, H_0 zamítnuta, F_{AB} = 2.285, H_0 akcept., (c) F_A = 22.283, H_0 zamítnuta, F_B = 0.004, H_0 akcept., F_{AB} = 1.728, H_0 akcept.*
B5.13 *ANOVA#2B, F_A = 18.38, H_0 zamítnuta, F_B = 6079.7, H_0 zamítnuta, F_{AB} = 2.07, H_0 zamítnuta*
B5.14 *ANOVA#2P, F_A = 4.161, H_0 zamítnuta, F_B = 0.442, H_0 akcept., F_{AB} = 2.791, H_0 akcept.*
B5.15 *ANOVA#1, F = 122.6, H_0 zamítnuta*
B5.16 *ANOVA#2B, F_A = 22.267, H_0 zamítnuta, F_B = 13.588, H_0 zamítnuta, F_{AB} = 1.965, H_0 akcept.*
B5.17 *ANOVA#2U, F_A = 4.088, H_0 zamítnuta, F_B = 0.197, H_0 akcept., F_{AB} = 12.986, H_0 zamítnuta*
B5.18 *ANOVA#2B, F_A = 23.968, H_0 zamítnuta, F_B = 2.108, H_0 zamítnuta, F_{AB} = 0.023, H_0 akcept.*
B5.19 *ANOVA#2B, F_A = 1.038, H_0 akcept., F_B = 0.317, H_0 akcept. F_{AB} = 0.289, H_0 akcept.*
B5.20 *ANOVA#2B, F_A = 1.047, H_0 akcept., F_B = 4.135, H_0 akcept. F_{AB} = 0.044, H_0 akcept.*
B5.21 *ANOVA#2P, F_A = 0.088, H_0 akcept., F_B = 8.724, H_0 zamítnuta F_{AB} = 0.026, H_0 akcept.*
B5.22 *ANOVA#1, F = 43.964, H_0 zamítnuta*
B5.23 *ANOVA#1, F = 3.827, H_0 zamítnuta*
B5.24 *ANOVA#1, F = 21.351, H_0 = zamítnuta*
B5.25 *ANOVA#2B, F_A = 216.8, H_0 zamítnuta, F_B = 309.3, H_0 zamítnuta F_{AB} = 75.671, H_0 zamítnuta*
B5.26 *ANOVA#2P, F_A = 1.246, H_0 akcept., F_B = 3.563, H_0 akcept. F_{AB} = 2.674, H_0 akcept.*
B5.27 *ANOVA#2B, F_A = 0.357, H_0 akcept., F_B = 2.378, H_0 akcept. F_{AB} = 3.952, H_0 zamítnuta*
B5.28 *ANOVA#2B, F_A = 3.860, H_0 akcept., F_B = 4.790, H_0 zamítnuta F_{AB} = 0.049, H_0 akcept.*

5.7.2 Analýza chemických a fyzikálních dat

C5.01 *ANOVA#2P, F_A = 0.964, H_0 akcept., F_B = 40.381, H_0 zamítnuta F_{AB} = 0.231, H_0 akcept.*
C5.02 *ANOVA#1, F = 3.678, H_0 zamítnuta*
C5.03 *ANOVA#2B, F_A = 5.862, H_0 akcept., F_B = 5.293, H_0 zamítnuta F_{AB} = 0.728, H_0 akcept.*
C5.04 *ANOVA#1, F = 18.592, H_0 zamítnuta*
C5.05 *ANOVA#2B, F_A = 378.8, H_0 zamítnuta, F_B = 91.0, H_0 zamítnuta F_{AB} = 35.545, H_0 zamítnuta*
C5.06 *ANOVA#2B, F_A = 0.147, H_0 akcept., F_B = 0.053, H_0 akcept. F_{AB} = 0.375, H_0 akcept.*
C5.07 *ANOVA#2P, F_A = 2.375, H_0 akcept., F_B = 2.960, H_0 akcept.F_{AB} = 0.507, H_0 akcept.*
C5.08 *ANOVA#2P, F_A = 5.267, H_0 zamítnuta, F_B = 2.523, H_0 zamítnuta F_{AB} = 4.318, H_0 zamítnuta*
C5.09 *ANOVA#1, F = 8.161, H_0 zamítnuta*
C5.10 *ANOVA#1, F = 38.613, H_0 zamítnuta*
C5.11 *Porovnání dvou výběrů, F = 12.887, H_0 zamítnuta*
C5.12 *ANOVA#1, F = 0.488, H_0 akcept.*
C5.13 *ANOVA#2P, F_A = 39050, H_0 zamítnuta, F_B = 119.9, H_0 zamítnuta, F_{AB} = 13.662, H_0 zamítnuta*
C5.14 *ANOVA#1, F = 6.632, H_0 zamítnuta*
C5.15 *ANOVA#2B, F_A = 856.5, H_0 zamítnuta, F_B = 457.1, H_0 zamítnuta F_{AB} = 5.394, H_0 zamítnuta*
C5.16 *ANOVA#1, F = 78.994, H_0 zamítnuta*
C5.17 *ANOVA#2B, F_A = 0.842, H_0 akcept., F_B = 17.542, H_0 zamítnuta F_{AB} = 0.392, H_0 akcept.*
C5.18 *ANOVA#2P, F_A = 9.428, H_0 zamítnuta, F_B = 6.104, H_0 zamítnuta, F_{AB} = 3.217, H_0 akcept.*
C5.19 *ANOVA#1, F = 18.190, H_0 zamítnuta*
C5.20 *ANOVA#2P, F_A = 16.334, H_0 zamítnuta, F_B = 1.565, H_0 akcept. F_{AB} = 5.755, H_0 zamítnuta*

C5.21 *ANOVA#2B, F_A = 1.706, H_0 akcept., F_B = 30.777, H_0 zamítnuta F_{AB} = 2.599, H_0 zamítnuta*
C5.22 *ANOVA#2P, F_A = 2.427, H_0 akcept., F_B = 0.240, H_0 akcept. F_{AB} = 4.148, H_0 akcept.*

5.7.3 Analýza environmetálních, potravinářských a zemědělských dat

E5.01 *ANOVA#1, F = 153.8, H_0 zamítnuta*
E5.02 *ANOVA#1, F = 18.813, H_0 zamítnuta*
E5.03 *ANOVA#1, F = 19.513, H_0 zamítnuta*
E5.04 *ANOVA#2P, F_A = 7.324, H_0 zamítnuta, F_B = 2.436, H_0 akcept.F_{AB} = 14.643, H_0 zamítnuta*
E5.05 *ANOVA#2U, F_A = 3.888, H_0 zamítnuta, F_B = 13.550, H_0 zamítnuta F_{AB} = 9.275, H_0 zamítnuta*
E5.06 *ANOVA#2B, F_A = 599.2, H_0 zamítnuta, F_B = 4.493, H_0 akcept. F_{AB} = 14.217, H_0 zamítnuta*
E5.07 *ANOVA#2P, F_A = 20.138, H_0 zamítnuta, F_B = 1.380, H_0 akcept. F_{AB} = 1.317, H_0 zamítnuta*
E5.08 *ANOVA#1, F = 0.306, H_0 akcept.*
E5.09 *ANOVA#1, F = 3.499, H_0 zamítnuta*
E5.10 *ANOVA#1, F = 12.13, H_0 zamítnuta*
E5.11 *ANOVA#2P, F_A = 21.877, H_0 zamítnuta, F_B = 2.305, H_0 akcept. F_{AB} = 0.008, H_0 akcept.*
E5.12 *ANOVA#2P, F_A = 2.906, H_0 akcept., F_B = 7.188, H_0 zamítnuta F_{AB} = 1.344, H_0 akcept.*
E5.13 *ANOVA#2P, F_A = 35.724, H_0 zamítnuta, F_B = 5.332, H_0 zamítnuta F_{AB} = 0.476, H_0 akcept.*
E5.14 *ANOVA#2B, F_A = 9.455, H_0 zamítnuta, F_B = 0.841, H_0 akcept. F_{AB} = 1.038, H_0 akcept.*
E5.15 *ANOVA#2P, F_A = 3.037, H_0 zamítnuta, F_B = 7.140, H_0 zamítnuta F_{AB} = 25.072, H_0 zamítnuta*
E5.16 *ANOVA#2B, F_A = 1.456, H_0 akcept., F_B = 1.194, H_0 akcept. F_{AB} = 0.772, H_0 akcept.*

5.7.4 Analýza hutnických a mineralogických dat

H5.01 *ANOVA#1, F = 0.926, H_0 akceptována*
H5.02 *ANOVA#2P, F_A = 3.952, H_0 zamítnuta, F_B = 5.829, H_0 zamítnuta F_{AB} = 3.493, H_0 akcept.*
H5.03 *ANOVA#2B, F_A = 6.158, H_0 zamítnuta, F_B = 94.750, H_0 zamítnuta, F_{AB} = 2.198, H_0 akcept.*
H5.04 *ANOVA#2B, F_A = 20.117, H_0 zamítnuta, F_B = 5.527, H_0 zamítnuta F_{AB} = 1.617, H_0 akcept.*
H5.05 *ANOVA#2B, F_A = 41.153, H_0 zamítnuta, F_B = 60.990, H_0 zamítnuta, F_{AB} = 2.173, H_0 akcept.*
H5.06 *ANOVA#3, zjednodušeně v ADSTAT: ANOVA#2B, F_A = 5.70H_0 zamítnuta, F_B = 3478.5, H_0 zamítnuta, F_{AB} = 1.77, H_0 zamítnuta*
H5.07 *ANOVA#1, F = 5.208, H_0 zamítnuta*
H5.08 *ANOVA#1, F = 0.023, H_0 akceptována*
H5.09 *ANOVA#1, F = 6.337, H_0 zamítnuta*
H5.10 *ANOVA#2B, F_A = 0.522, H_0 akcept., F_B = 309.6, H_0 zamítnuta F_{AB} = 2.73, H_0 zamítnuta.*
H5.11 *ANOVA#1, F = 5.988, H_0 zamítnuta*
H5.12 *ANOVA#1, F = 7.432, H_0 zamítnuta*
H5.13 *ANOVA#2B, F_A = 0.131, H_0 akcept., F_B = 0.726, H_0 akcept. F_{AB} = 0.728, H_0 akcept.*
H5.14 *ANOVA#1, F = 9.789, H_0 zamítnuta*
H5.15 *ANOVA#2B, F_A = 1.040, H_0 akcept., F_B = 1.285, H_0 akcept. F_{AB} = 1.037, H_0 akcept.*
H5.16 *ANOVA#1, F = 1.099, H_0 = akcept.*
H5.17 *ANOVA#1, F = 7.266, H_0 zamítnuta*
H5.18 *ANOVA#1, F = 1.510, H_0 akcept.*
H5.19 *ANOVA#1, F = 13.941, H_0 zamítnuta*
H5.20 *ANOVA#1, F = 2.648, H_0 akcept.*
H5.21 *ANOVA#2B, F_A = 0.334, H_0 akcept., F_B = 123.9, H_0 zamítnuta F_{AB} = 0.029, H_0 akcept.*
H5.22 *ANOVA#2P, F_A = 0.433, H_0 akcept., F_B = 9.807, H_0 zamítnuta F_{AB} = 0.007, H_0 akcept.*

5.7.5 Analýza ekonomických a sociologických dat

S5.01 *ANOVA#1, F = 3.678, H_0 zamítnuta*
S5.02 *ANOVA#2P, F_A = 1324.2, H_0 zamítnuta, F_B = 0.304, H_0 akcept. F_{AB} = 2.328, H_0 akcept.*
S5.03 *ANOVA#2P, F_A = 23.040, H_0 zamítnuta, F_B = 92.383, H_0 akcept. F_{AB} = 87.387, H_0 akcept.*
S5.04 *ANOVA#1, F = 13.003, H_0 zamítnuta*
S5.05 *ANOVA#1, ANOVA#2B, F_A = 3.216, H_0 zamítnuta, F_B = 161.1H_0 zamítnuta, F_{AB} = 8.087, H_0 zamítnuta*
S5.06 *ANOVA#2P, F_A = 3.365, H_0 zamítnuta, F_B = 2.429, H_0 akcept. F_{AB} = 1.304, H_0 akcept.*
S5.07 *ANOVA#2P, F_A = 2.779, H_0 akcept., F_B = 1.226, H_0 akcept. F_{AB} = 0.805, H_0 akcept.*
S5.08 *ANOVA#1, F = 6.802, H_0 zamítnuta*

S5.09 *ANOVA#2B,* $F_A = 1.305$, H_0 *akcept.,* $F_B = 3.707$, H_0 *akcept.* $F_{AB} = 0.600$, H_0 *akcept.*
S5.10 *ANOVA#2B,* $F_A = 3.228$, H_0 *zamítnuta,* $F_B = 122.3$, H_0 *zamítnuta* $F_{AB} = 14.470$, H_0 *zamítnuta*
S5.11 *ANOVA#1,* $F = 1.548$, H_0 *akceptována*
S5.12 *ANOVA#1,* $F = 6.610$, H_0 *zamítnuta*
S5.13 *ANOVA#2P,* $F_A = 2.096$, H_0 *akcept.,* $F_B = 37.692$, H_0 *zamítnuta* $F_{AB} = 1.911$, H_0 *akcept.*
S5.14 *ANOVA#1,* $F = 15.950$, H_0 *zamítnuta*
S5.15 *ANOVA#1,* $F = 50.704$, H_0 *zamítnuta*
S5.16 *ANOVA#1,* $F = 12.083$, H_0 *zamítnuta*
S5.17 *ANOVA#1,* $F = 0.664$, H_0 *akcept.*
S5.18 *ANOVA#2P,* $F_A = 3.409$, H_0 *akcept.,* $F_B = 22.682$, H_0 *zamítnuta* $F_{AB} = 0.522$, H_0 *akcept.*
S5.19 *ANOVA#1,* $F = 1.638$, H_0 *akcept.*
S5.20 *ANOVA#1,* $F = 6.915$, H_0 *zamítnuta*
S5.21 *ANOVA#1,* $F = 48.190$, H_0 *zamítnuta*
S5.22 *ANOVA#2B,* $F_A = 5.642$, H_0 *zamítnuta,* $F_B = 1.180$, H_0 *akcept.*
S5.23 *ANOVA#2U,* $F_A = 4.110$, H_0 *zamítnuta.,* $F_B = 2.175$, H_0 *akcept.* $F_{AB} = 1.762$, H_0 *akcept.*

5.8 Doporučená literatura

[1] Searle S. R.: *Biometrics* **27**, 1 (1971).
[2] Bartlett M. S., Kendall D. G.: *J. Roy. Stat. Soc.* **B8**, 128 (1946).
[3] Schéffe H.: *The Analysis of Variance.* J. Wiley, New York 1959.
[4] Searle S. R.: *Linear Models.* J. Wiley, New York 1971.
[5] Miller P. G.: *Beyond ANOVA, Basics of Applied Statistics.* J. Wiley, New York 1986.
[6] Speed T. P.: *Annals of Statist.* **15**, 885 (1987).
[7] Emerson J. D., Hoaglin D. C., Kempthorne P. I.: *J. Amer. Statist. Assoc.* **79**, 329 (1984).
[8] Bradu D., Hawkins D. M.: *Technometrics* **24**, 103 (1982).
[9] Bloomfield P., Steiger W.: *Least Absolute Deviations: Theory, Applications and Algorithms.* Birkhäuser, Boston 1983.
[10] Gabriel K. R.: *J. R. Stat. Soc.* **B40**, 186 (1978).
[11] Cressie N. A. C.: *Biometrics* **34**, 505 (1978).
[12] Potocký R a kol.: *Zbierka úloh z pravdepodobnosti a matematickej štatistiky.* ALFA-SNTL, Bratislava 1986.
[13] Anderson R. L.: *Practical Statistics for Analytical Chemists.* van Nostrand Reinhold Comp., New York 1987.
[14] Miller J. C., Miller J. N.: *Statistics for Analytical Chemistry*. Ellis Horwood, Chichester 1984.
[15] Liteanu C., Rica I.: *Statistical Theory and Methodology of Trace Analysis*. Ellis Horwood, Chichester 1980.
[16] Rice J. A.: *Mathematical Statistics and Data Analysis*. Wadsworth & Brooks, California 1988, s. 397.
[17] Hintze J.: *Number Cruncher Statistical Systems 2000.* Manuál, Kaysville, Utah, October 1998.
[18] Hintze J.: *User's Guide NCSS2000.* Statistical System for Windows, Kaysville, Utah 1999.

6

LINEÁRNÍ REGRESNÍ MODELY

Při budování regresních modelů se běžně užívá metody nejmenších čtverců. Metoda nejmenších čtverců poskytuje postačující odhady parametrů jenom při současném splnění všech předpokladů o datech a o regresním modelu. Pokud tyto předpoklady nejsou splněny, ztrácí výsledky metodou nejmenších čtverců své vlastnosti. Užití lineární regresní analýzy se týkají následujících možností:

1. *Popis dat:* hledáme vztah, lineární regresní model, který sumarizuje soubor dat.

2. *Určení parametrů:* nejběžnějším cílem regresní analýzy je vyčíslení nejlepších odhadů neznámých parametrů regresního modelu. Uživatel navrhne regresní model a regresní analýzou se snaží model prokázat. Často tento cíl překrývá i ostaní záměry regresní analýzy.

3. *Predikce:* nejdůležitějším cílem regresní analýzy je predikce, vyčíslení hodnot závisle proměnných. Bývá to často cena, dodací lhůta, účinnost, obložnost v nemocnici, výtěžek reakce, síla kovu, atd. Predikce jsou důležité i v plánování, monitoringu, vyhodnocování chemických procesů atd. Je však řada předpokladů a kvalifikací, které se musí respektovat v regresním modelu a datech. Často se například nesmí extrapolovat mimo rozsah dat. Intervalové odhady vyžadují dodržení předpokladu normality. Metoda nejmenších čtverců (MNČ) má svých sedm důležitých předpokladů, které je třeba respektovat a dodržet.

4. *Řízení:* regresní modely lze využít také k monitoringu a řízení systémů, například ke kalibraci měřicího systému. Když využijeme regresní model k řídicím účelům, nezávisle proměnné musí být vztaženy k závisle proměnné kauzálním způsobem.

5. *Výběr (volba) proměnných:* volba proměnných sleduje ty nezávisle proměnné, které vysvětlují významný objem proměnlivosti závisle proměnné. V řadě aplikací nejde o jednorázový proces, ale o spojitý proces výstavby modelu.

Základní předpoklady metody nejmenších čtverců: statistické vlastnosti odhadů $\hat{\boldsymbol{y}}_P, \hat{\boldsymbol{e}}$, $\boldsymbol{b}$ závisí na splnění jistých základních předpokladů.

Předpoklady metody nejmenších čtverců:

I. *Regresní parametry β mohou nabývat libovolných hodnot.* V praxi však existují často omezení parametrů, která vycházejí z jejich fyzikálního smyslu.

II. *Regresní model je lineární v parametrech a platí aditivní model měření.*

III. *Matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných X má hodnost rovnou právě m*. To znamená, že žádné její dva sloupce $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ nejsou *kolineární*, tj. rovnoběžné vektory. Tomu odpovídá i formulace, že matice $\boldsymbol{X}^T\boldsymbol{X}$ je symetrická regulární matice, ke které existuje inverzní matice a jejíž determinant je větší než nula.

IV. *Náhodné chyby g_i mají nulovou střední hodnotu $E(g_i) = 0$.* To musí u kore-lačních modelů platit vždy. U regresních modelů se může stát, že $E(g_i) = K$, $i = 1, ..., n$, což znamená, že model neobsahuje absolutní člen. Po jeho zavedení bude $E(\hat{g}_i) = 0$, kde $\hat{g}_i = y_i - \hat{y}_{P,i} - K$.

V. *Náhodné chyby g_i mají konstantní a konečný rozptyl $E(g_i^2) = \sigma^2$*. Také podmíněný rozptyl $D(y/\boldsymbol{x}) = \sigma^2$ je konstantní a jde o *homoskedastický* případ.

VI. *Náhodné chyby g_i jsou vzájemně nekorelované a platí* $\mathrm{cov}(g_i\, g_j) = E(g_i\, g_j) = 0$. Pokud mají chyby normální rozdělení, jsou nezávislé. Tento požadavek odpovídá požadavku nezávislosti měřených veličin *y*.

VII. *Chyby g_i mají normální rozdělení $N(0, \sigma^2)$.* Vektor $\boldsymbol{y}$ má pak vícerozměrné normální rozdělení se střední hodnotou $\boldsymbol{X}\boldsymbol{\beta}$ a kovarianční maticí $\sigma^2\boldsymbol{E}$, kde $\boldsymbol{E}$ je jednotková matice.

Pokud platí předpoklady I až VI, jsou odhady $\boldsymbol{b}$ parametrů $\boldsymbol{\beta}$ **n**ejlepší, **n**estranné a **l**ineární (NNLO). Navíc mají asymptoticky normální rozdělení. Pokud platí ještě předpoklad VII, mají odhady $\boldsymbol{b}$ normální rozdělení i pro konečné výběry.

Regresní diagnostika: metoda nejmenších čtverců nezajišťuje obecně nalezení přijatelného modelu, a to jak ze statistického, tak i z fyzikálního hlediska. Musí být splněny podmínky, odpovídající složkám tzv. *regresního tripletu* (kritika dat, kritika modelu a kritika metody odhadu). Regresní diagnostika obsahuje postupy k identifikaci

a) vhodnosti dat pro navržený regresní model (složka *data*),

b) vhodnosti modelu pro daná data (složka *model*),

c) splnění základních předpokladů MNČ (složka *metoda*).

Základní rozdíl mezi regresní diagnostikou a klasickými testy spočívá v tom, že u regresní diagnostiky není třeba přesně formulovat alternativní hypotézu H_A. Tímto pojetím se regresní diagnostika blíží spíše k *exploratorní regresní analýze*, která vychází z faktu, že "uživatel ví o analyzovaných datech přece jenom více než počítač". Počítač zde slouží pouze jako nástroj analýzy dat, modelu a metody odhadu. Model je navrhován v interakci uživatele s programem. Tím by měl být omezen vznik formálních regresních modelů, které nemají fyzikální smysl a jsou v technické praxi obyčejně jen omezeně použitelné.

1. Data: mezi základní techniky regresní diagnostiky patří stanovení rozmezí dat, jejich variability a přítomnosti vybočujících pozorování. K tomu lze využít grafů rozptýlení

s kvantily a řady postupů průzkumové analýzy jednorozměrných dat z kap. 2. Přes svoji jednoduchost umožňuje regresní diagnostika identifikovat ještě před vlastní regresní analýzou

a) *nevhodnost dat* (malé rozmezí nebo přítomnost vybočujících bodů),

b) *nesprávnost navrženého modelu* (skryté proměnné),

c) *multikolinearitu*,

d) *nenormalitu* v případě, kdy jsou vysvětlující proměnné náhodnými veličinami.

Kvalita dat úzce souvisí s užitým regresním modelem. Při posuzování se sleduje především výskyt *vlivných bodů*, které mohou být hlavním zdrojem řady problémů, jako je zkreslení odhadů a růst rozptylů až k naprosté nepoužitelnosti regresních modelů. Vlivné body lze rozdělit do tří skupin:

a) *Hrubé chyby*, které jsou způsobeny měřenou veličinou (*vybočující pozorování*) nebo nevhodným nastavením vysvětlujících proměnných (*extrémy*). Hrubé chyby jsou obyčejně důsledkem chyb při manipulaci s daty.

b) *Body s vysokým vlivem* (tzv. golden points) jsou speciálně vybrané body, které byly přesně změřeny, a které obvykle rozšiřují predikční schopnosti modelu.

c) *Zdánlivě vlivné body* vznikají jako důsledek nesprávně navrženého regresního modelu.

Podle toho, kde se vlivné body vyskytují, lze provést dělení na

1. *Vybočující pozorování* (outliers), které se liší v hodnotách vysvětlované (závisle) proměnné y od ostatních, a

2. *extrémy* (high leverage points), které se liší v hodnotách vysvětlujících (nezávisle) proměnných $\boldsymbol{x}$ nebo v jejich kombinaci (v případě multikolinearity) od ostatních bodů.

Vyskytují se však i body, které jsou jak vybočující, tak i extrémní. O jejich výsledném vlivu však především rozhoduje to, že jsou extrémy. K identifikaci vlivných bodů typu vybočujícího pozorování se využívá zejména různých typů reziduí a k identifikaci extrémů pak diagonálních prvků H_{ii} projekční matice $\boldsymbol{H}$ (detaily v učebnici[72]).

2. Model: kvalitu regresního modelu lze posoudit v případě jedné vysvětlující proměnné x přímo z rozptylového grafu závislosti y na x. V případě více vysvětlujících proměnných a multikolinearity mohou však rozptylové grafy *mylně indikovat* nelineární trend i u lineárního modelu. Z řady různých grafů k posouzení vztahu y a x_j se omezíme na a) parciální regresní grafy, a b) parciální reziduální grafy.

Parciální regresní grafy byly Belseyem zařazeny mezi základní nástroje počítačové interaktivní analýzy regresních modelů. Umožňují nejenom posouzení kvality navrženého regresního modelu, ale indikují i přítomnost vlivných bodů a nesplnění předpokladů klasické metody nejmenších čtverců. Parciální regresní graf pro posouzení vztahu mezi y a i-tou vysvětlující proměnnou x_i je závislost *reziduí* $\boldsymbol{v}$ regrese y na sloupcích matice $\boldsymbol{X_{(i)}}$ a reziduí u regrese x_i na sloupcích matice $\boldsymbol{X_{(i)}}$. Přitom matice $\boldsymbol{X_{(i)}}$ vznikne z matice $\boldsymbol{X}$ vynecháním i-tého sloupce $\boldsymbol{x_i}$, odpovídajícího i-té vysvětlující proměnné. Parciální regresní grafy mají tyto vlastnosti:

a) Směrnice přímky v parciálním regresním grafu je stejná jako odhad b_j v neděleném modelu a úsek je roven nule. Tato lineární závislost platí pouze v případě, že navržený model je správný.

b) Korelační koeficient mezi oběma proměnnými parciálního regresního grafu odpovídá parciálnímu korelačnímu koeficientu $R_{yx_j(\boldsymbol{x})}$.

c) Rezidua v parciálním regresním grafu jsou shodná s klasickými rezidui e_i pro nedělený model.

d) V grafu jsou indikovány vlivné body a i některá porušení předpokladů metody nejmenších čtverců (heteroskedasticita).

Parciální reziduální grafy se označují také jako grafy *"komponenta + reziduum"*. Parciální reziduální grafy však poskytují poněkud odlišné informace než parciální regresní grafy:

a) Směrnice lineární závislosti je rovna b_j a úsek je nulový. Lineární závislost pak ukazuje na vhodnost navržené proměnné x_j v modelu.

b) Rezidua regresní přímky jsou přímo rezidua $\hat{e}_i$ pro nedělený model.

c) Pokud je úhel mezi $\boldsymbol{x}_j$ a některými sloupci matice $\boldsymbol{X}_{(j)}$ malý (*multikolinearita*), ukazuje parciální reziduální graf nesprávně malý rozptyl kolem regresní přímky $b_j \boldsymbol{x}_j$ a dochází navíc i k potlačení efektu vlivných bodů.

Parciální reziduální grafy se doporučují především k indikaci rozličných typů nelinearity v případě nesprávně navrženého regresního modelu.

3. Metoda: v praxi bývají některé předpoklady MNČ porušeny, což vede k použití jiných kritérií. K porušení předpokladů dochází v těchto základních případech:

a) Na parametry jsou kladena omezení, což vede na užití *metody podmínkových nejmenších čtverců (MPNČ).*

b) Kovarianční matice chyb není diagonální (autokorelace), popř. data nemají stejný rozptyl (heteroskedasticita), což vede na užití *metody zobecněných nejmenších čtverců (MZNČ)*, resp. *metody vážených nejmenších čtverců (MVNČ).*

c) Rozdělení dat nelze považovat za normální nebo se v datech vyskytují vlivné body. V takovém případě se místo kritéria metody nejmenších čtverců užije *robustního* kritéria, které je na porušení předpokladu o rozdělení chyb a na vlivné body málo citlivé. Z robustních kritérií jsou nejznámější *M-odhady*. Jedná se o maximálně věrohodné odhady pro vhodnou hustotu pravděpodobnosti chyb. Pro odhad parametrů $\boldsymbol{b}$ se užívá *iterační metody vážených nejmenších čtverců* (*IVNČ*).

d) Také proměnné x mohou být zatíženy náhodnými chybami, což vede na užití *metody rozšířených nejmenších čtverců (MRNČ).* Pro případ regresní přímky je použití metody rozšířených nejmenších čtverců velmi jednoduché. Postačuje znalost poměru rozptylu σ_y^2 (vysvětlovaná proměnná) a σ_x^2 (vysvětlující proměnné), $K = \sigma_y^2/\sigma_x^2$. Pro odhad směrnice regresní přímky $y = a\,x + b$ pak platí

$$a = L + \operatorname{sign}(S_{yx}) \sqrt{K + L^2} \quad ,$$

kde

$$L = \frac{S_{yx} - K\,S_x}{2S_x}$$

a sign S_{yx} je znaménková funkce. Symboly S označují součty čtverců, odpovídajících proměnných

$$S_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad , \quad S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad ,$$

$$S_{yx} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad .$$

Při znalosti odhadu směrnice $\hat{a}$ se snadno určí odhad úseku $\hat{b}$ ze vztahu

$$\hat{b} = \bar{y} - \hat{a}\,\bar{x} \quad .$$

Pro případ stejných rozptylů, tj. $K = 1$, vede dosazení do uvedených vztahů k odhadům minimalizujícím kolmé vzdálenosti (*ortogonální regrese).* Pro odhady rozptylů odhadů $\hat{a}$, $\hat{b}$ se pak používá speciálních vztahů.

e) Pro špatně podmíněné matice $\boldsymbol{X^T X}$ se používá *metoda racionálních hodností*, (**G**eneral **P**rincipal **C**omponent **R**egression) vedoucí k systému vychýlených odhadů, kde vychýlení je řízeno jedním parametrem.

Postup výstavby lineárního regresního modelu (ADSTAT)

(definice regresních diagnostik a ostatních statistických pojmů jsou v učebnici[72]):

1. Návrh modelu: začíná se vždy od nejjednoduššího modelu, u kterého vystupují jednotlivé vysvětlující proměnné v prvních mocninách a nevyskytují se žádné interakční členy typu $x_j\,x_k$. Pouze v případech, u kterých je předem známo, že model má obsahovat funkce vysvětlujících proměnných, může být výchozí model dle těchto požadavků upraven.

2. Předběžná analýza dat: sleduje se proměnlivost jednotlivých proměnných a možné párové vztahy. Užívá se proto rozptylových diagramů závislosti x_j na x_k nebo indexových grafů závislosti x_j na j. Posuzuje se významnost proměnných s ohledem na jejich proměnlivost a přítomnost multikolinearity. Přibližně lineární vztah mezi proměnnými v rozptylových grafech závislosti x_j na x_k indikuje multikolinearitu. Lze rovněž odhalit i vlivné body, které způsobují multikolinearitu.

Podle volby uživatele se provedou požadované transformace původních proměnných. Zadává se, zda model obsahuje absolutní člen. Uživatel může volit polynomickou transformaci zadáním stupně polynomu, Taylorův rozvoj do 2. stupně a lineární model s interakcemi. Uživatel může zadat libovolnou mocninu původních proměnných včetně logaritmů. Ostatní typy transformací se provádějí při přípravě dat k výpočtu v datovém editoru. K odstranění případné heteroskedasticity, vzniklé nelineární transformací proměnné y, je možné zadat nestatistické váhy, jež odpovídají kvazilinearizaci.

Provádí se sestavení korelační matice $\boldsymbol{R}$ a její rozklad na vlastní čísla a vlastní vektory. Jsou vypočteny faktory *VIF* (variation inflation factor) k indikaci multikolinearity a dále jsou vyčíslena setříděná vlastní čísla. K určení inverzní matice $\boldsymbol{R}^{-1}$ se užívá metoda racionálních hodností GPCR pro standardně zadávané vychýlení $P = 10^{-15}$. Uživatel může zadat jinou hodnotu parametru vychýlení P, což však vede pro vyšší hodnoty P k vychýleným odhadům. Bývá proto vhodné volit P z intervalu $10^{-5} \# P \# 10^{-3}$.

3. Odhadování parametrů: odhadování parametrů modelu se provádí metodou racionálních hodností GPCR s volbou $P = 10^{-5}$, což je vlastně MNČ. Ze zobecněné inverzní matice $\boldsymbol{R}^{-1}$ jsou určovány odhady parametrů $\boldsymbol{b}$, jejich směrodatné odchylky $\sqrt{D(b_j)}$ a velikosti testačních statistik Studentova *t*-testu významnosti pro $\beta_j = 0$. Dále jsou provedeny testy významnosti odhadů b_j, vícenásobného korelačního koeficientu R a koeficientu determinace D. Je vhodné sledovat souhrnné charakteristiky regrese jako je střední kvadratická chyba predikce *MEP* a Akaikovo informační kritérium *AIC*, popř. posoudit linearitu modelu.

4. Regresní diagnostika: identifikace vlivných bodů je prováděna využitím pěti rozličných grafů, a to *grafů Wiliamsova, Pregibonova, McCullohova-Meeterova, L-R,* a *grafu predikovaných reziduí*. Dále musíme ověřit splnění předpokladů metody nejmenších čtverců, jako je homoskedasticita, nepřítomnost autokorelace a normalita rozdělení chyb. Pokud dojde k úpravě dat, je třeba provést znovu regresní diagnostiku se zaměřením na porušení předpokladů metody nejmenších čtverců a posouzení vlivu multikolinearity. V případě více vysvětlujících proměnných se posoudí vhodnost jednotlivých proměnných a jejich funkcí využitím parciálních regresních grafů nebo grafů "komponenta + reziduum". Obvykle jsou využívány následující tabulky:

Tabulka výsledků obsahuje hodnoty predikce $\hat{y}_i$, rozptylů predikce $D(\hat{y}_i)$ a relativní odchylky predikce od experimentálních dat. Je uvedena i průměrná absolutní, resp. relativní odchylka a reziduální suma čtverců *RSC*. Následuje statistická analýza klasických reziduí.

Tabulka reziduí obsahuje klasická rezidua $\hat{e}_i$, normovaná rezidua $\hat{e}_{Ni}$, standardizovaná rezidua $\hat{e}_{Si}$ a Jackknife rezidua $\hat{e}_{Ji}$. Je uveden odhad autokorelačního koeficientu reziduí prvního řádu $\hat{\kappa}_1$.

Tabulka vlivných bodů obsahuje veličiny H_{ii}, $H_{ii}^{(}$, D_i, A_i, DF_i, $LD_i(\boldsymbol{b})$, $LD_i(\hat{\sigma}^2)$ a $LD_i(\boldsymbol{b}, \sigma^2)$. Hvězdičkou bývají označeny hodnoty silně vlivných bodů.

5. Konstrukce zpřesněného modelu: při využití

a) *metody vážených nejmenších čtverců (MVNČ)* při nekonstantnosti rozptylů,

b) *metody zobecněných nejmenších čtverců (MZNČ)* při autokorelaci,

c) *metody podmínkových nejmenších čtverců (MPNČ)* při omezeních, kladených na parametry,

d) *metody racionálních hodností GPCR* u multikolinearity,

e) *metody rozšířených nejmenších čtverců (MRNČ)* pro případ, že všechny proměnné jsou zatížené náhodnými chybami,

f) *robustní metody* - parametry zpřesněného modelu jsou odhadovány pro jiná rozdělení dat než normální a data s vybočujícími hodnotami a extrémy.

6. Zhodnocení kvality modelu: provede se s využitím klasických testů, postupů regresní diagnostiky a doplňkových informací o modelované soustavě posouzení kvality navrženého lineárního regresního modelu.

7. Kalibrační modely: u kalibračních modelů se pro daný signál y^* vypočte hodnota x^* spolu se svým konfidenčním intervalem. Před vlastním užitím kalibračního modelu

je vhodné určit limitu detekce a limitu stanovení, které určují použitelnou dolní hranici kalibračního modelu nebo odpovídající metody. Postup obsahuje

(a) Návrh modelu.

(b) Statistickou analýzu reziduí.

(c) Výpočet derivací a integrálů.

(d) Určení kalibračních mezí.

(e) Sestavení kalibrační tabulky.

8. Testování různých hypotéz: ve zvláštních případech, jako je porovnání několika přímek atd., se provádí testování pomocí dalších testů k ověřování rozličných typů hypotéz.

Uživatel může při interaktivní práci s počítačem některé tabulky nebo grafy vynechat. Na základě analýzy vlivných bodů a reziduí lze provést i vypuštění některých bodů a výpočty pak zopakovat. Podrobný popis, vzorce a statistické testy najde čtenář v doporučované učebnici[72]. Úlohy jsou v 6. kapitole rozděleny do pěti skupin: J6 jednorozměrné lineární regresní modely, V6 validace nové metody, K6 lineární a nelineární kalibrace, L6 polynomické regresní modely, M6 vícerozměrné lineární regresní modely. Ve výsledcích značí *r* korelační koeficient, *D* koeficient determinace v procentech, regresní model $y = \beta_1(s_1 , A \text{ či } Z) + ... + \beta_m(s_m , A \text{ či } Z)\, x$, kde *A* značí, že hypotéza H_0 o nevýznamnosti úseku je akceptována a *Z* zamítnuta, dále je uveden počet odlehlých bodů *o*, počet extrémů *e*.

6.1 Jednorozměrné lineární regresní modely

Vzorová úloha 6.1 *Postup výstavby modelu a regresní diagnostika*

Na úloze **J6.01** *Model teplotní závislosti přechodového tlaku bismutu* ukážeme postup analýzy jednorozměrného lineárního regresního modelu. Houck studoval přechodový tlak bismutu I - II *p* jako funkci teploty *t*, str. 501 v cit[62]. Nalezněte lineární regresní model, který bude adekvátní daným datům. Vyšetřete regresní triplet a indikujte vlivné body. Vykazují klasická rezidua normalitu? Vyšetřete rankitový a kvantilový graf.

Řešení:

1. Návrh modelu: na začátku analýzy vždy zařadíme absolutní člen β_0 a navržený regresní model (přímky) bude mít tvar $y = \beta_0 + \beta_1 x$.

2. Předběžná analýza dat: poloha a proměnlivost proměnných *y, x* se posuzuje na základě *průměru* a *směrodatné odchylky* hodnot každé proměnné. *Pearsonův párový korelační koeficient r* ukazuje na vysokou korelaci proměnných *y* a *x*.

Proměnná	Průměr	Směrodatná odchylka	Párový korelační koeficient	Spočtená hladina významnosti
y	24859.0	427.41	1.0000	-----
x	30.313	10.701	-0.9983	0.000

3. Odhadování parametrů: klasickou metodou nejmenších čtverců MNČ byly nalezeny nejlepší odhady úseku β_0 a směrnice β_1. Studentův *t*-test ukázal při porovnání *t*-kritéria

s kritickou hodnotou $t_{0.95}(23\text{-}2) = 2.080$, že úsek (absolutní člen) β_0 je statisticky významný a směrnice β_1 je rovněž statisticky významná.

Parametr	**Odhad**	**Směrodatná odchylka**	**Test H_0: b_j vs. H_A: b_j ... 0 t-kritérium**	**hypotéza H_0 je**	**Hlad. význam.**
b_0	26068.0	1.6169E+01	1612.2	Zamítnuta	0.000
b_1	-39.874	5.0419E-01	-79.084	Zamítnuta	0.000

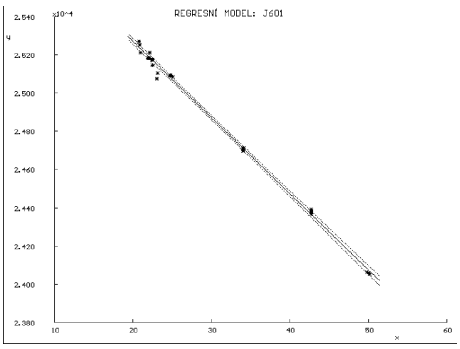
4. Základní statistické charakteristiky: absolutní hodnota *párového korelačního koeficientu r* ukazuje, že navržený lineární regresní model je statisticky významný. Vysoká hodnota *koeficientu determinace D* = 100 % r^2 (99.67 %), představuje procen-to variability, vysvětlené modelem. *Predikovaný koeficient determinace* R^2_p ukazuje na predikční schopnost modelu, je však vyčíslen jinak než r^2, místo *RSC* se ve vztahu užije *MEP*. *Střední kvadratická chyba predikce MEP* a *Akaikovo informační kritérium AIC* se užívají k rozlišení mezi několika navrženými modely. Za optimální se považuje model, pro který dosahuje *MEP* a *AIC* minimální hodnotu.

Vícenásobný korelační koeficient, *r*	: 0.99833
Koeficient determinace, *D* [%]	: 99.665
Predikovaný koeficient determinace, R^2_P	: 99.804
Střední kvadratická chyba predikce, *MEP*	: 685.46
Akaikovo informační kritérium, *AIC*	: 150.54

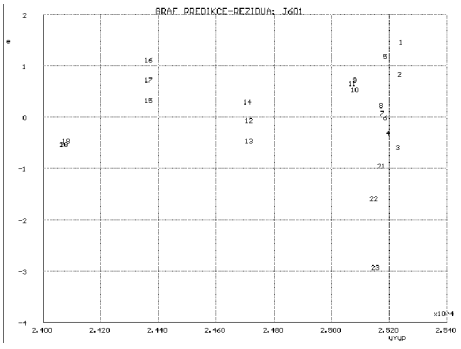
5. Regresní diagnostika: obsahuje pomůcky a postupy pro interaktivní analýzu (a) dat, (b) modelu, (c) metody, což jsou složky tzv. *regresního tripletu.*

Data: Regresní diagnostika se skládá z analýzy několika druhů grafických diagnostik (obr. 6.1-1a, b) a tabulek různých druhů reziduí.

(a) *Analýza klasických reziduí* $\hat{e}$ není příliš spolehlivá, protože klasická rezidua $\hat{e}$ jsou korelovaná, s nekonstantním rozptylem, jeví se normálnější než náhodné chyby (*efekt supernormality)* a nemusí indikovat silně odlehlé hodnoty. Grafická analýza $\hat{e}$ vs. $\hat{y}_P$ (obr. 6.1-1b) je schopna indikovat pouze podezřelé body, trend a nekon-stantnost rozptylu, tzv. heteroskedasticitu. Míry rozptýlení klasických reziduí by měly dosahovat hodnot blízkých experimentálnímu šumu. *Odhad směrodatné odchylky s(e)* by se měl blížit svou velikostí experimentální chybě, kterou je zatížena závisle proměnná *y*. Odhady šikmosti a špičatosti by měly indikovat Gaussovo normální rozdělení reziduí.



Obr. 6.1-1a Graf regresního modelu, ***ADSTAT.***



Obr. 6.1-1b Analýza klasických reziduí, ***ADSTAT.***

Bod	**Měřená hodnota**	**Predikovaná hodnota**	**Směrodatná odchylka**	**Klasické reziduum**	**Relativní reziduum**
i	$y_{exp,\,i}$	$y_{vyp,\,i}$	$s(y_{vyp,\,i})$	e_i	$e_{r,\,i}$
1	2.5276E+04	2.5239E+04	7.1310E+00	3.7376E+01	1.4787E-01
2	2.5256E+04	2.5235E+04	7.0971E+00	2.1363E+01	8.4587E-02
3	2.5216E+04	2.5231E+04	7.0635E+00	-1.4649E+01	-5.8095E-02
4	2.5187E+04	2.5195E+04	6.7704E+00	-7.7630E+00	-3.0821E-02
5	2.5217E+04	2.5187E+04	6.7076E+00	3.0212E+01	1.1981E-01
6	2.5187E+04	2.5187E+04	6.7076E+00	2.1178E-01	8.4085E-04
7	2.5177E+04	2.5175E+04	6.6153E+00	2.1738E+00	8.6342E-03
8	2.5177E+04	2.5171E+04	6.5851E+00	6.1612E+00	2.4472E-02
9	2.5098E+04	2.5079E+04	5.9642E+00	1.8871E+01	7.5187E-02
10	2.5093E+04	2.5079E+04	5.9642E+00	1.3871E+01	5.5277E-02
11	2.5088E+04	2.5071E+04	5.9178E+00	1.6845E+01	6.7145E-02
12	2.4711E+04	2.4712E+04	5.5947E+00	-1.2920E+00	-5.2284E-03
13	2.4701E+04	2.4712E+04	5.5947E+00	-1.1292E+01	-4.5715E-02
14	2.4716E+04	2.4708E+04	5.6116E+00	7.6953E+00	3.1135E-02
15	2.4374E+04	2.4365E+04	8.1762E+00	8.6087E+00	3.5319E-02
16	2.4394E+04	2.4365E+04	8.1762E+00	2.8609E+01	1.1728E-01
17	2.4384E+04	2.4365E+04	8.1762E+00	1.8609E+01	7.6315E-02
18	2.4067E+04	2.4078E+04	1.1197E+01	-1.1301E+01	-4.6957E-02
19	2.4057E+04	2.4070E+04	1.1286E+01	-1.3327E+01	-5.5396E-02
20	2.4057E+04	2.4070E+04	1.1286E+01	-1.3327E+01	-5.5396E-02
21	2.5147E+04	2.5171E+04	6.5851E+00	-2.3839E+01	-9.4798E-02
22	2.5107E+04	2.5147E+04	6.4087E+00	-3.9915E+01	-1.5898E-01
23	2.5077E+04	2.5151E+04	6.4374E+00	-7.3902E+01	-2.9470E-01

Reziduální součet čtverců, *RSC*	: 1.3449E+04
Průměr absolutních hodnot reziduí, M_e	: 1.8314E+01
Průměr relativních reziduí, $M_{e,\,r}$	: 7.3476E-02
Odhad reziduálního rozptylu, $s^2(e)$	: 6.4043E+02
Odhad směrodatné odchylky reziduí, *s(e)*	: 2.5307E+01
Odhad šikmosti reziduí, $g_1(e)$	:-1.098
Odhad špičatosti reziduí, $g_2(e)$	: 4.689

(b) ***Analýza ostatních reziduí:*** Jackknife rezidua indikují odlehlé body, z diagonálních prvků H_{ii} projekční matice $\boldsymbol{H}$ a diagonálních prvků H_{mii} zobecněné projekční matice $\boldsymbol{H_m}$ pouze extrémy. Ostatní druhy reziduí a kritéria v tabulce pak obojí (značeno hvězdičkou

u dotyčné hodnoty). Jackknife rezidua $e_{J,i}$ ukazují, že bod č. 23 je odlehlý, stejně tak i Cookova vzdálenost D_i , Atkinsova vzdálenost A_i ukazují na č. 23 a kritérium DF_i pak na č. 23 a věrohodnostní vzdálenosti $LD(b)_i$ a $LD(s^2)_i$ na č. 23 a $LD(b, s^2)_i$ na č. 23. Diagonální prvky H_{ii} projekční matice $\boldsymbol{H}$ ukazují na extrémy č. 18, 19, 20 a 23, diagonální prvky zobecněné H_{mii} projekční matice $\boldsymbol{H_m}$ na extrémy 23.

Indikace vlivných bodů: (* indikuje odlehlý nebo vlivný bod)

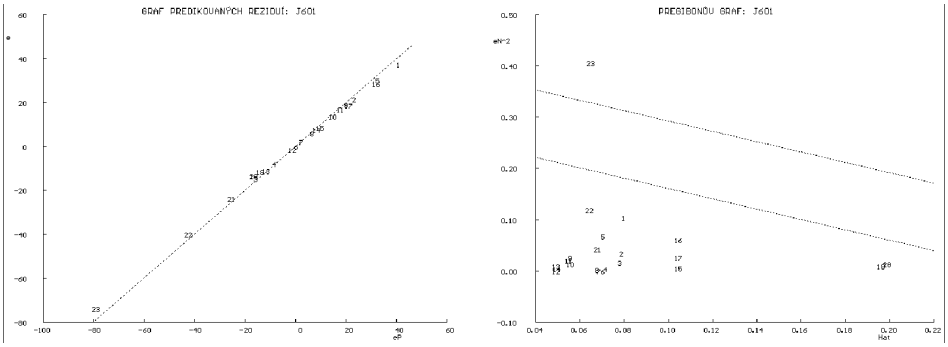
Bod	**Standardizované reziduum**	**Jackknife reziduum**	**Predikované reziduum**	**Diagonální prvky**
i	e_{Si}	e_{Ji}	e_{Pi}	H_{ii}
1	1.5393E+00	1.5949E+00	4.0600E+01	7.9400E-02
2	8.7947E-01	8.7453E-01	2.3187E+01	7.8649E-02
3	-6.0282E-01	-5.9345E-01	-1.5887E+01	7.7906E-02
4	-3.1836E-01	-3.1144E-01	-8.3614E+00	7.1573E-02
5	1.2381E+00	1.2549E+00	3.2495E+01	7.0253E-02
6	8.6791E-03	8.4700E-03	2.2779E-01	7.0253E-02
7	8.8994E-02	8.6866E-02	2.3333E+00	6.8333E-02
8	2.5215E-01	2.4644E-01	6.6087E+00	6.7709E-02
9	7.6729E-01	7.5952E-01	1.9980E+01	5.5543E-02
10	5.6398E-01	5.5461E-01	1.4686E+01	5.5543E-02
11	6.8463E-01	6.7571E-01	1.7820E+01	5.4683E-02
12	-5.2348E-02	-5.1090E-02	-1.3584E+00	4.8874E-02
13	-4.5753E-01	-4.4874E-01	-1.1872E+01	4.8874E-02
14	3.1184E-01	3.0504E-01	8.0933E+00	4.9171E-02
15	3.5945E-01	3.5187E-01	9.6120E+00	1.0438E-01
16	1.1945E+00	1.2075E+00	3.1943E+01	1.0438E-01
17	7.7699E-01	7.6941E-01	2.0777E+01	1.0438E-01
18	-4.9796E-01	-4.8885E-01	-1.4052E+01	1.9576E-01*
19	-5.8835E-01	-5.7896E-01	-1.6635E+01	1.9889E-01*
20	-5.8835E-01	-5.7896E-01	-1.6635E+01	1.9889E-01*
21	-9.7560E-01	-9.7443E-01	-2.5570E+01	6.7709E-02
22	-1.6304E+00	-1.7025E+00	-4.2650E+01	6.4130E-02
23	-3.0196E+00	-3.9175E+00*	-7.9015E+01	6.4707E-02

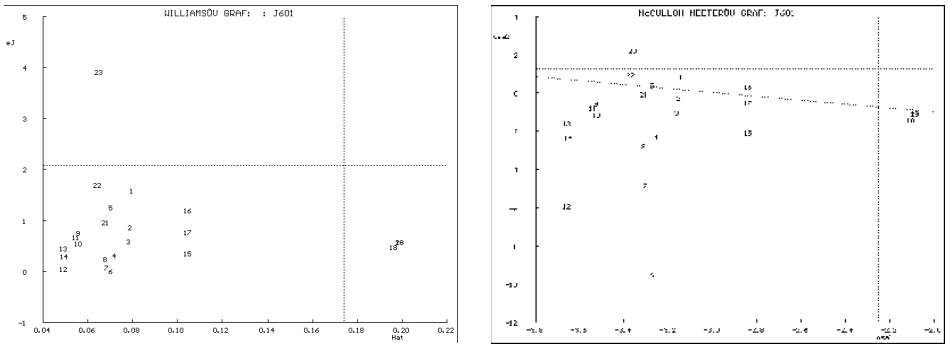
Bod	**Zobecněné diag. prvky**	**Cookova vzdálenost**	**Atkinsonova vzdálenost**	**Vliv na predikci**
i	H_{mii}	D_i	A_i	DF_i
1	1.8327E-01	1.0218E-01	1.5177E+00	4.6838E-01
2	1.1258E-01	3.3013E-02	8.2795E-01	2.5551E-01
3	9.3862E-02	1.5351E-02	5.5896E-01	-1.7250E-01
4	7.6054E-02	3.9067E-03	2.8020E-01	-8.6472E-02
5	1.3812E-01	5.7915E-02	1.1178E+00	3.4496E-01
6	7.0257E-02	2.8459E-06	7.5444E-03	2.3283E-03
7	6.8684E-02	2.9044E-04	7.6230E-02	2.3525E-02
8	7.0531E-02	2.3087E-03	2.1521E-01	6.6415E-02
9	8.2020E-02	1.7311E-02	5.9683E-01	1.8419E-01
10	6.9848E-02	9.3529E-03	4.3582E-01	1.3450E-01
11	7.5782E-02	1.3557E-02	5.2662E-01	1.6252E-01
12	4.8998E-02	7.0407E-05	3.7528E-02	-1.1581E-02
13	5.8355E-02	5.3782E-03	3.2962E-01	-1.0172E-01

14	5.3574E-02	2.5145E-03	2.2478E-01	6.9367E-02
15	1.0989E-01	7.5293E-03	3.8925E-01	1.2013E-01
16	1.6524E-01	8.3153E-02	1.3358E+00	4.1223E-01
17	1.3013E-01	3.5181E-02	8.5115E-01	2.6267E-01
18	2.0526E-01	3.0179E-02	7.8153E-01	-2.4119E-01
19	2.1209E-01	4.2969E-02	9.3477E-01	-2.8847E-01
20	2.1209E-01	4.2969E-02	9.3477E-01	-2.8847E-01
21	1.0996E-01	3.4563E-02	8.5092E-01	-2.6260E-01
22	1.8259E-01	9.1073E-02	1.4441E+00	-4.4566E-01
23	4.7079E-01*	3.1540E-01*	3.3389E+00*	-1.0304E+00*

Bod	**Věrohodnostní vzdálenosti**		
i	$LD(b)_i$	$LD(s^2)_i$	$LD(b,s^2)_i$
1	2.2274E-01	6.6813E-02	3.0813E-01
2	7.2200E-02	5.5015E-04	7.2365E-02
3	3.3602E-02	8.3059E-03	4.1037E-02
4	8.5559E-03	1.7810E-02	2.6036E-02
5	1.2651E-01	1.1418E-02	1.4232E-01
6	6.2339E-06	2.2387E-02	2.2393E-02
7	6.3620E-04	2.2015E-02	2.2624E-02
8	5.0567E-03	1.9460E-02	2.4312E-02
9	3.7889E-02	2.9345E-03	4.0252E-02
10	2.0478E-02	9.7036E-03	2.9602E-02
11	2.9676E-02	5.4651E-03	3.4518E-02
12	1.5422E-04	2.2260E-02	2.2408E-02
13	1.1778E-02	1.3480E-02	2.4862E-02
14	5.5073E-03	1.7986E-02	2.3279E-02
15	1.6487E-02	1.6639E-02	3.2512E-02
16	1.8143E-01	7.7907E-03	1.9472E-01
17	7.6935E-02	2.6720E-03	7.8567E-02
18	6.6012E-02	1.2071E-02	7.6058E-02
19	9.3931E-02	8.8254E-03	1.0036E-01
20	9.3931E-02	8.8254E-03	1.0036E-01
21	7.5584E-02	4.2910E-05	7.5898E-02
22	1.9863E-01	9.7932E-02	3.1641E-01
23	6.8070E-01*	3.8060E+00*	4.9740E+00*

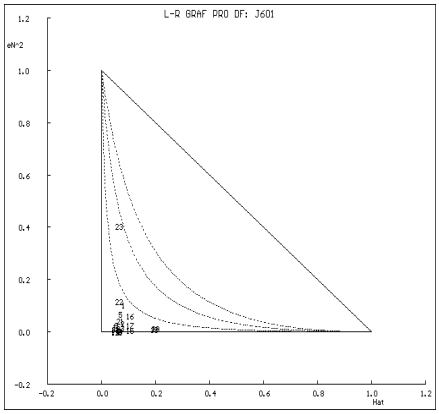
(c) ***Grafy vlivných bodů*** (obr. 6.1-2) jsou schopny indikovat přítomnost odlehlých hodnot a extrémů. *Graf predikovaných reziduí* ukazuje na odlehlý bod č. 23. *Pregibonův graf* ukazuje na silně vlivný bod č. 23. *Williamsův graf* indikuje č. 23 jako odlehlý bod a extrémy č. 18, 19, 20. *McCullohův-Meeterův graf* dokazuje odlehlý bod č. 23, extrémy č. 18, 19 a 20. Konečně *L-R graf* dokazuje odlehlý bod č. 23 a současně extrémy č. 18, 19, 20. Lze uzavřít, že bod č. 23 je většinou diagnostik prokázán za odlehlý, a proto je vhodné ho dále analyzovat resp. z výběru vyloučit.





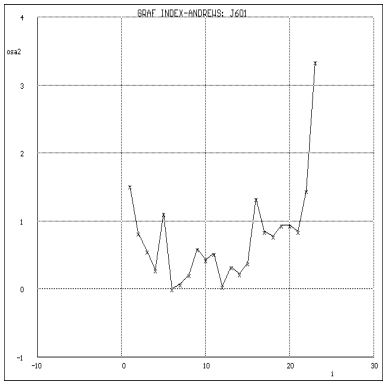
Obr. 6.1-2 Grafy vlivných bodů, vlevo graf predikovaných reziduí, a vpravo Pregibonův graf, ***ADSTAT.***

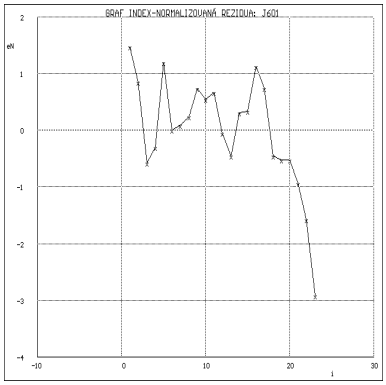
Obr. 6.1-2 Grafy vlivných bodů, vlevo Williamsův graf, a vpravo McCullohův-Meeterův graf, ***ADSTAT.***



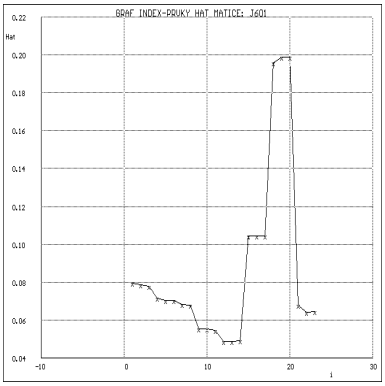
Obr. 6.1-2 Grafy vlivných bodů, L-R graf, ***ADSTAT.***

(d) ***Indexové grafy*** (obr. 6.1-3) upozorňují na *podezřelé body*, a kvantifikují velikost vlivu. *Andrewsův indexový graf* a *graf normovaných reziduí* ukazují na podezřelé body č. 1 a 23. *Indexový graf prvků* ***H*** *projekční matice* pak na extrémy č. 18, 19, 20.



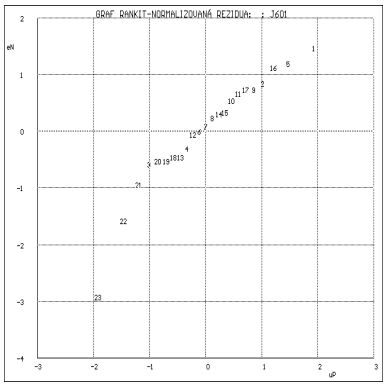


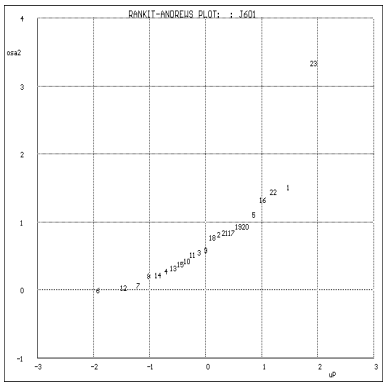
Obr. 6.1-3 Indexové grafy, vlevo, Andrewsův graf, a vpravo graf normovaných reziduí, ***ADSTAT.***

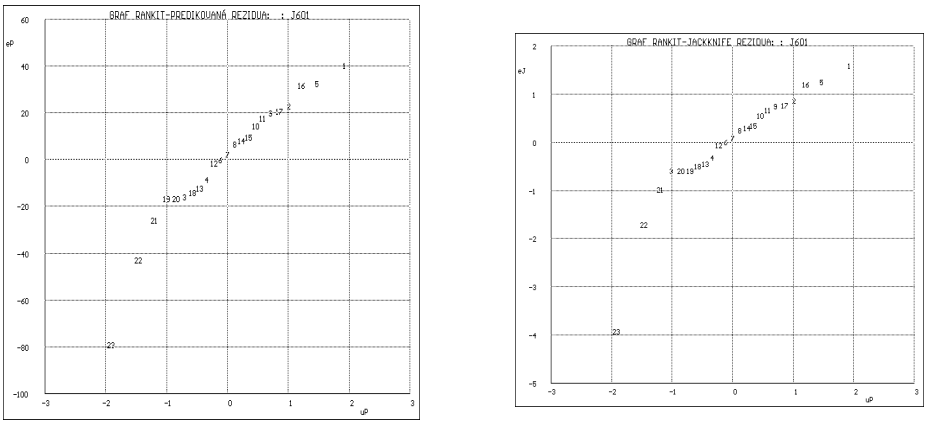


Obr. 6.1-3 Graf prvků H-projekční matice, ***ADSTAT.***

(e) ***Rankitové grafy*** (obr. 6.1-4) ukazují vedle normality rozdělení dotyčných reziduí i na vlivné (zde odlehlé) body. *Graf normovaných reziduí* ukazuje na začátku č. 23 a na konci č. 1 jako odlehlé body. *Andrewsův graf* ukazuje bod č. 23 jako odlehlý. *Graf Jackknife reziduí* ukazuje č. 23 jako odlehlý.







Obr. 6.1-4 Rankitové grafy, vlevo, graf predikovaných reziduí, a vpravo, graf Jackknife reziduí, ***ADSTAT.***

Model: *Parciální regresní grafy* a *parciální reziduální grafy* jsou určeny pro vícerozměrné lineární regresní modely a nemají proto u jednorozměrného regresního modelu smysl. Vhodnost modelu se posuzuje přímo v grafu, obsahujícím data a průběh modelové funkce. Je patrné, že v tomto případě je přímka akceptovatelná a data nevykazují nelineární průběh.

Metoda: do této části patří vyšetření splnění základních předpokladů metody nejmenších čtverců MNČ čili statistické testy předpokladů, za kterých by měla metoda vést k nejlepším nestranným lineárním odhadům regresních parametrů:

Fisherův-Snedecorův test významnosti regrese potvrdil, že navržený model je přijat jako významný, jinými slovy: závisle proměnná *y* a nezávisle proměnná *x* jsou silně lineárně závislé.

Scottovo kritérium multikolinearity nemá u jednorozměrného regresního modelu smysl.

Cookův-Weisbergův test heteroskedasticity ukazuje, že rezidua vykazují heteroskedasticitu (nekonstantnost rozptylu).

Jarqueův-Berraův test normality reziduí ukazuje, že klasická rezidua nevykazují Gaussovo, normální rozdělení.

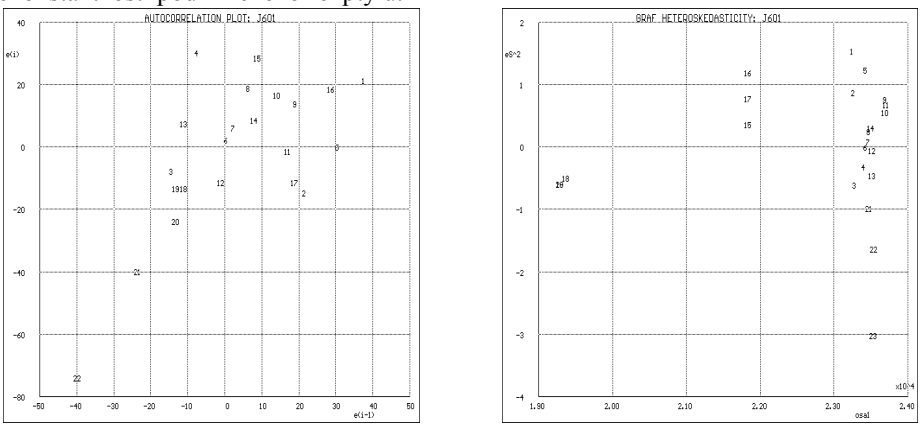
Waldův test autokorelace ukazuje, že klasická rezidua jsou autokorelována. To je vážné upozornění ke zhodnocení provedeného experimentu a dokazuje, že došlo k narušení podmínek, např. objevil se vliv nežádoucího konkurenčního děje (zanedbané proměnné), nevhodných podmínek měření na téže soustavě atd.

Znaménkový test prokazuje, že znaménko klasických reziduí se dostatečně střídá, a proto rezidua nevykazují žádný trend.

TESTOVÁNÍ REGRESNÍHO TRIPLETU (DATA + MODEL + METODA):	
Fisherův-Snedocorův test významnosti regrese, F_{exp}	: 6.2543E+03
Tabulkový kvantil, $F_{1-\alpha}(m\text{-}1, n\text{-}m)$	: 4.3248E+00
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	: 0.000
Scottovo kritérium multikolinearity, *M*	: -4.07117E-15

Závěr: Navržený model je korektní.	
Cookův-Weisbergův test heteroskedasticity, S_f	: 1.1084E+02
Tabulkový kvantil, $\chi^2_{1-\alpha}(1)$	: 3.8415E+00
Závěr: Rezidua vykazují heteroskedasticitu.	
Spočtená hladina významnosti	: 0.000
Jarqueův-Berraův test normality reziduí, $L(e)$	: 7.3537E+00
Tabulkový kvantil, $\chi^2_{1-\alpha}(2)$	: 5.9915E+00
Závěr: Normalita není přijata.	
Spočtená hladina významnosti	: 0.025
Waldův test autokorelace, W_a	: 3.1791E+01
Tabulkový kvantil, $\chi^2_{1-\alpha}(1)$	: 3.8415E+00
Závěr: Rezidua jsou autokorelována.	
Spočtená hladina významnosti	: 0.000
Znamékový test, D_t	:-2.5225E+00
Tabulkový kvantil, $N_{1-\alpha/2}$	: 1.6449E+00
Závěr: Rezidua nevykazují trend.	
Spočtená hladina významnosti	: 0.006

Graf autokorelace (obr. 6.1-5) vykazuje přibližně mrak bodů bez výrazné orientace.
Graf heteroskedasticity (obr. 6.1-5) vykazuje klín, což odpovídá heteroskedasticitě, nekonstantnosti podmíněného rozptylu.



Obr. 6.1-5 Vlevo: graf autokorelace, a vpravo: graf heteroskedasticity, ***ADSTAT.***

6. Konstrukce zpřesněného modelu:

(a) Po odstranění bodů č. 23 (*kritika dat*) byly nalezeny nové odhady parametrů zpřesněného modelu, když $t_{0.95}(22\text{-}2) = 2.086$.

Parametr	**Odhad**	**Směrodatná odchylka**	**Test H_0: $b_j = 0$ vs. H_A: b_j ... 0** ***t*-kritérium**	**hypotéza H_0 je**	**Hlad. význam.**
b_0	26078.0	12.742	2046.7	Zamítnuta	0.000
b_1	-40.103	0.3930	-102.04	Zamítnuta	0.000

Zpřesněný model (v závorce je uveden vždy odhad směrodatné odchylky parametru)

$$y = 26\,078\ (13) - 40.1\ (0.4)\ x$$

je doložen statistickými charakteristikami: *Pearsonův párový korelační koeficient r, koeficient determinace D a predikovaný korelační koeficient* R_P dosáhly vesměs vysokých hodnot. *Střední kvadratická chyba predikce MEP* a *Akaikovo informační kritérium AIC* dosáhly nižších hodnot než u předešlého modelu, což dokazuje, že zpřesněný model je lepší.

Vícenásobný korelační koeficient, *r*	: 0.9990
Koeficient determinace, 100 *% D* [%]	: 99.808
Predikovaný koeficient determinace, R^2_P	: 0.9989
Střední kvadratická chyba predikce, *MEP*	: 414.22
Akaikovo informační kritérium, *AIC*	: 132.62

Rezidua nyní vykazují normální rozdělení a nevykazují trend, stále však vykazují heteroskedasticitu, a proto lze doporučit použití metody vážených nejmenších čtverců.

(b) Užitím statistické váhy ($w_i = 1/y_i^2$) kompenzujeme heteroskedasticitu v datech. Obdržíme nové odhady parametrů, když $t_{0.95}(22\text{-}2) = 2.086$.

Parametr	**Odhad**	**Směrodatná odchylka**	**H_0: $b_j = 0$ vs. H_A: b_j ... 0** *t*-**kritérium**	**hypotéza H_0 je**	**Hlad. význam.**
b_0	26079.0	12.666	2059.0	Zamítnuta	0.000
b_1	-40.110	0.3860	-103.92	Zamítnuta	0.000

Opravený model má tvar (v závorce je odhad směrodatné odchylky parametru)

$$y = 26\,079\ (13) - 40.1\ (0.4)\ x.$$

Jelikož došlo ke snížení rozhodujících kritérií, tj. střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*, lze považovat tyto odhady za lepší než předešlé. Pearsonův korelační koeficient *r,* a tím pádem i koeficient determinace *D* vychází nepatrně lepší nebo stejný než u předešlého odhadu bez statistické váhy.

Vícenásobný korelační koeficient, *r*	: 0.9991
Koeficient determinace, 100 *% D* [%]	: 99.815
Predikovaný koeficient determinace, R^2_P	: 0.9989
Střední kvadratická chyba predikce, *MEP*	: 410.29
Akaikovo informační kritérium, *AIC*	: 132.39

7. Zhodnocení kvality modelu: porovnáním hodnot regresní diagnostiky lze snadno provést zhodnocení *regresního tripletu* získaného lineárního regresního modelu pro upravená data, zbavená odlehlých hodnot, metodou vážených nejmenších čtverců. Nalezený a prokázaný model teplotní závislosti přechodového tlaku bismutu má tvar, (v závorce je vždy uveden odhad směrodatné odchylky parametru)

$$y = 26\,079\ (13) - 40.1\ (0.4)\ x.$$

6.1.1 Úlohy na jednorozměrné lineární regresní modely

Vzorce, postupy a statistické testy jsou podrobně popsány v doporučené učebnici[72]. Dotyčnou stránku, na kterou je v jednotlivých úlohách uveden odkaz uvádíme u každé úlohy.

Úloha J6.01 *Model teplotní závislosti přechodového tlaku bismutu*

Houck studoval přechodový tlak bismutu I - II *y* jako funkci teploty *x*, str. 501 v cit[62]. (1) Nalezněte lineární regresní model, který bude adekvátní daným datům. (2) Vyšetřete regresní triplet a indikujte vlivné body. (3) Jaký lze očekávat tlak bismutu pro teplotu 40 EC a 50 EC? (4) Uveďte také intervalový odhad tlaku při těchto teplotách pro hladinu významnosti $\alpha = 0.05$. (5) Vykazují klasická rezidua normalitu? Vyšetřete proto rankitový a kvantilový graf.

Data: Teplota *x* [EC], tlak *y* [bar].

20.8	25276,	20.9	25256,	21.0	25216,	21.9	25187,	22.1	25217,
...	...	...	...	...	...	...	...	...	...
22.5	25147,	23.1	25107,	23.0	25077,				

Úloha J6.02 *Ověření správnosti metody po výměně lampy ve spektrofotometru*

Po výměně originální lampy UV-VIS spektrofotometru byly srovnávány dosahované výsledky stanovení mastných kyselin v kalech s původními hodnotami. (1) Vyšetřete regresní model a testujte vliv nové lampy. (2) Jsou v datech vlivné body? (3) Porovnejte závěry o indikaci vlivných bodů z tabulek a z grafů. (4) Ověřte, zda regresní přímka prochází počátkem. (5) Obsahuje 95 % interval spolehlivosti směr-nice číslo $\beta_1 = 1$? (6) Co ukazují grafy klasických reziduí proti predikci a proti *x*, jde o stejnou informaci? (7) Testem složené hypotézy proveďte test úseku (má být $\beta_0 = 0$) a směrnice (má být $\beta_1 = 1$) dle postupu na str. 365 v učebnici[72].

Data: Obsah mastných kyselin [mg/l] před *x* a po *y* výměně lampy.

0.082	0.082,	0.159	0.160,	0.240	0.242,	0.312	0.329,	0.376	0.374,
0.475	0.469,	0.548	0.545,	0.637	0.630,				

Úloha J6.03 *Závislost obsahu křemíku v surovém železe na teplotě strusky*

Procentuální obsah křemíku v surovém železe *y* závisí na teplotě strusky *x*, str. 210 v cit[64]. (1) Pomocí regresní diagnostiky navrhněte lineární regresní model. (2) Vyšetřete regresní triplet a odstraňte vybočující hodnoty. (3) Komentujte velikost korelačního koeficientu *r* a koeficientu determinace *D* mezi proměnnými *x* a *y*. (4) Odhadněte procentuální obsah křemíku v surovém železe pro teplotu strusky 1350EC a 1450EC.

Data: Teplota strusky *x* [EC], obsah křemíku *y* [%].

1300	0.3	1368	0.341	1380	0.38	1444	0.555	1460	0.62
1472	0.643	1320	0.29	1400	0.42	1480	0.680	1340	0.35
1420	0.47	1500	0.7	1360	0.28	1440	0.510		

Úloha J6.04 *Užití titrační a spektrofotometrické metody pro dva typy vázaného dusíku*

Při výrobě dusíkatých hnojiv se množství čpavkového dusíku a amidického dusíku ve výrobku stanovuje titrační a spektrofotometrickou metodou. (1) Určete regresní přímky závislosti naměřeného dusíku metodou spektrofotometrickou na metodě titrační. (2) Rozhodněte, zda regresní přímky, stanovené pro nízké hodnoty obsahu dusíku čpavkového a vyšší hodnoty dusíku amidického, jsou totožné. (3) Na oba druhy dusíku aplikujte Chowův test shody dvou lineárních modelů za předešlého vyšetření homoskedasticity.

Data: Čpavkový dusík [%], metoda titrační *x*, kolorimetrická *y*.

7.62	7.51,	7.56	7.52,	7.56	7.51,	7.58	7.51,	6.24	6.36,	6.48	6.50,	7.53	7.50,	7.40	7.43,
6.66	6.71,														

Amidický dusík [%], metoda titrační x, kolorimetrická y.

15.49	15.55,	15.56	15.52,	15.52	15.57,	15.50	15.44,	16.73	16.73,
...	...	...	...	...	...	...	...	...	...
14.61	14.72,	15.49	15.46,	15.27	15.26,	14.92	14.97,	14.42	14.41,

Úloha J6.05 *Stanovení tří redukujících cukrů podle Luffa-Schoorla*

Redukující cukry glukóza, fruktóza, laktóza a maltóza redukují za varu v alkalickém prostředí měďnatou sůl na oxid měďný, str. 239 v cit[59]. Nezreagovaný přebytek měďnaté soli zoxiduje jodid draselný na jod, a ten se pak stanoví thiosíranem. (1) Určete model závislosti spotřeby tiosíranu x na obsahu dotyčného redukujícího cukru y využitím testu linearity dle Uttsové a dle kritérií *MEP* a *AIC*, str. 372 v cit[72]. Použijte také graf. (2) Testujte, zda jsou všechny tři přímky shodné, tj. zda mají stejné směrnice a společný úsek? dle str. 380 v cit[72].

Data: Spotřeba 0.05 M tiosíranu x [ml], obsah redukujícího cukru y [mg].

x	y pro glukózu	y pro laktózu	y pro maltózu
1	2.4	3.6	3.9
...	...	...	...
23	62.2	88.0	94.6

Úloha J6.06 *Model teplotní závislosti disociačního tlaku nitridu barnatého*

Orcutt studoval závislost disociačního tlaku y reakce nitridu barnatého na teplotě x a navrhl lineární regresní model $\ln y = A + B/x$, kde x je absolutní teplota, str. 501 v cit[62]. (1) Ověřte navržený model s využitím kritérií *MEP* a *AIC* a dále vyšetřete regresní triplet. (2) Jsou splněny požadavky metody nejmenších čtverců MNČ? (3) Jsou v datech nějaké odlehlé body, které je třeba vyloučit? (4) Jakou informaci získáte z diagramu rozptýlení a krabicového grafu exploratorní analýzy nezávisle proměnné $1/x$ a závisle proměnné $\ln y$? (5) Odhadněte disociační tlak nitridu barnatého pro teplotu x = 1090 K a 1140 K.

Data: Teplota x [stupně Kelvina], tlak y [torr].

738	0.0000211,	748	0.0000480,	764	0.0000595,	770	0.0000920,	792	0.0002060,
...	...	...	...	...	...	...	...	...	...
1135	0.6220000,	1150	0.7240000.						

Úloha J6.07 *Porovnání dvou regresních přímek obsahu meziproduktu*

Při výrobě ostazinové modře byla sledována závislost množství meziproduktu y na množství výchozí reakční komponenty x za jinak stejných provozních podmínek. Obsah meziproduktu v reakční směsi byl stanovován po skončení reakce jednak dusitanem sodným y_1, jednak spektrofotometricky y_2. Aplikujte Chowův test shody dvou lineárních modelů, str. 372 v cit.[72].

Data: Množství výchozí komponenty x [kg], množství meziproduktu pomocí $NaNO_3$ y_1 [kg].

9.00	24.30,	10.20	27.42,	11.60	31.40,	12.00	32.36,	5.500	13.25,	13.00	34.05,
10.50	25.79,	10.50	25.90,	11.50	30.81,	9.00	23.93				

Množství výchozí komponenty x [kg], množství meziproduktu spektrofotometricky y_2 [kg].

9.00	24.55,	10.20	27.93,	11.60	31.80,	12.00	31.65,	5.500	12.83,	13.00	35.24,
10.50	28.20,	10.50	27.39,	11.50	32.53,	9.00	25.48,				

Úloha J6.08 *Modely cen aut v závislosti na jejich stáří a na ujetých kilometrech*

U dvaceti prodaných ojetých automobilů určité značky byla zjištěna cena y [tisíce Kč], stáří auta x_1 [roky] a počet ujetých kilometrů x_2 [tisíce km]: (1) Popište regresním modelem závislost ceny aut y na jejich stáří x_1. (2) Stejně tak popište regresním modelem i závislost ceny aut y na počtu ujetých kilometrů x_2 a obě přímky porovnejte. (3) Jaký je Pearsonův korelační koeficient r mezi závisle proměnnou a oběma nezávisle proměnnými?

Data: (a) Stáří auta x_1 [roky], cena auta y [tisíce Kč].

0.60	55.0,	1.00	54.6,	1.10	50.6,	2.00	51.1,	2.30	47.0,	2.50	50.0,
...	...	...	...	...	...	...	...	...	...	...	...
6.80	27.0,	7.50	17.6,								

(b) Počet ujetých km x_2 [tisíce km], cena auta y [tisíce Kč].

1.10	55.0,	2.50	54.6,	10.4	50.6,	4.50	51.1,	31.4	47.0,	8.60	50.0,
...	...	...	...	...	...	...	...	...	...	...	...
78.7	27.0,	90.2	17.6,								

Úloha J6.09 *Závislost kvality pěti výrobků na čase během dvou směn*

Ve dvou po sobě následujících dnech byla sledována kvalita různých výrobků záznamem řady faktorů; jedním z nich byla i hmotnost y. Hmotnosti pěti různých výrobků y_1 , y_2 , y_3, y_4 , y_5 v gramech byly stanoveny v každou celou hodinu x [h]. (1) Zjistěte, je-li hmotnost výrobků konstantní, či zda je v ní nějaký trend. (2) Vyšetřete, zda se případný trend v obou dnech dá vystihnout stejnou regresní závislostí. (3) Chowovým testem shody dvou lineárních modelů vyšetřete, zda výrobek ze dvou směn má stejné hmotnosti y v čase x, str. 372 v cit.[72]. (4) Jsou nějaké rozdíly mezi výrobky v průběhu jedné směny?

Data: Čas průběhu směny x [h], hmotnost výrobku [g] y.

1. den.

x	y_1	y_2	y_3	y_4	y_5
7	18.70	18.70	18.68	18.66	18.66
...	...	...	...	...	...
16	18.28	18.28	18.28	18.27	18.26

2. den.

x	y_1	y_2	y_3	y_4	y_5
7	18.34	18.35	18.35	18.31	18.34
...	...	...	...	...	...
18	18.66	18.68	18.68	18.67	18.67

Úloha J6.10 *Závislost koncentrace nositelů náboje v Sb_2Te_3 na koncentraci titanu*

Na základě měření byla stanovena závislost koncentrace volných nositelů náboje y_1 a y_2 na koncentraci titanu Ti x v teluridu antimonitém Sb_2Te_3. Koncentrace volných nositelů náboje byla zjištěna (a) z rezonanční frekvence plasmatu metodou y_1, (b) z Hallovy konstanty metodou y_2. (1) Rozhodněte, zda jsou obě metody srovnatelné a poskytují shodné výsledky? (2) Mají obě přímky shodný úsek β_0 a stejnou směrnici β_1? str. 380 v cit.[72].

Data: (a) Koncentrace Ti x [%], koncentrace nositelů náboje [10^{19} cm^{-3}] metodou y_1.

0.0	10.0,	0.2	9.10,	0.4	8.00,	0.6	7.60,	0.8	6.90,	1.0	5.90,
1.20	4.40,	1.40	3.60,	1.60	3.00,	1.80	1.70,	2.00	0.90		

(b) Koncentrace Ti x [%], koncentrace nositelů náboje [10^{19} cm^{-3}] metodou y_2.

0.0	9.60,	0.20	8.20,	0.40	7.60,	0.60	7.00,	0.80	4.40,	1.00	5.20,
1.20	3.00,	1.40	3.30,	1.60	1.00,	1.80	1.00,	2.00	0.90		

Úloha J6.11 *Závislost hustoty kalibrační kuličky na hloubce ponoru*
Hustota lineárního vysokohustotního polyethylenu je měřena metodou hustotních gradientních kolon. Skleněný odměrný válec, naplňovaný vodným roztokem isopropylalkoholu s plynule proměnnou hustotou, je ponořen do vodní lázně temperované na (23.0 ± 0.5)EC. Do kolony jsou vnořeny kalibrační kuličky o známé, pyknometricky stanovené hustotě. Výška roztoku v kalibrované části válce je 800 mm. Měření hustoty se provádí tak, že se dvě kuličky vloží do kolony, odečte se výška ponoru kalibrační kuličky x a z kalibračního grafu příslušné kolony se odečte hustota y s přesností na 4 desetinná místa. (1) Sestavte kalibrační graf a předem sestavte regresní model. (2) Vyšetřete regresní triplet, vlivné body a odstraňte z dat odlehlé hodnoty.
Data: Hloubka ponoru kalibrační kuličky x [mm], hustota kalibrační kuličky y [g/cm^3].

6.00	0.938,	102.0	0.939,	198.0	0.940,	301.0	0.941,	403.0	0.942,	499.0	0.943,
602.0	0.944,	698.0	0.945,								

Úloha J6.12 *Závislost indexu toku vysokohustotního polyethylenu při dvou zatíženích*
Při výrobě lineárního vysokohustotního polyethylenu je jedním z rozhodujících parametrů pro řízení technologického režimu hodnota indexu toku vyráběného polymeru, měřená při teplotě 190 EC a zatížení 21 N. U polyethylenu pro tlakové trubky je hodnota indexu toku při zatížení 21 N velmi nízká a její měření na plastometru je zatíženo vysokou relativní chybou. Pro snížení značného rozptylu je výhodné provést měření indexu toku při zatížení 49 N, kde je přesnost měření mnohem vyšší. Byla změřena odpovídající hodnota indexu toku při zatíženích 21 N a 49 N v celém rozsahu indexů toku trubkového typu polyethylenu. (1) Nalezněte regresní model, kterým ze změřeného indexu při 49 N dostatečně přesně odhadneme index polyethylenu při 21 N. (2) Jakou informaci přináší diagram rozptýlení a krabicový graf v exploratorní analýze obou proměnných x a y? (3) Vykazují klasická rezidua v rankitovém grafu normalitu? (4) Jaký je Pearsonův korelační koeficient r mezi proměnnými x a y? Odhadněte také jeho statistickou významnost. (5) Proveďte testy vhodnosti lineárního regresního modelu.
Data: Index toku x při 21 N, index toku y při 49 N.

0.0500	0.340,	0.0600	0.385,	0.0620	0.335,	0.0630	0.368,	0.0640	0.388,	0.0650	0.407,
...	...	...	...	...	...	...	...	...	...	...	...
0.192	0.927,										

Úloha J6.13 *Shodnost tří regresních přímek stanovení složky polyolefinu*
Ke stanovení charakteristické složky stabilizačního systému komerčního typu polyolefinu se vzorek granulátu po rozemletí extrahuje 3 hodiny dvěma různými rozpouštědly. Extrakt po druhé extrakci je po ochlazení měřen spektrofotometricky a z plochy píku je stanoven obsah charakteristické složky stabilizačního systému. Byly provedeny tři varianty úprav klasického systému A, B, C a je třeba posoudit, zda všechny vedou ke shodným výsledkům. Dále je třeba prověřit shodnost tří regresních přímek A, B, C, str. 384 v cit.[72].
Data: (a) Klasický postup x [mm^2], varianta A y_1 [mm^2].

34.0	32.0,	40.7	39.5,	48.5	45.5,	55.5	53.5,	62.5	60.3,	70.0	68.2,
77.5	75.0,	84.2	82.0,	91.5	89 .5,	97.2	96.5				

(b) Klasický postup x [mm^2], varianta B y_2 [mm^2].

34.0	36.5,	40.7	43.5,	48.5	50.5,	55.5	58.5,	62.5	65.0,	70.0	72.0,
77.5	79.0,	84.2	85.5,	91.5	93.5,	97.2	101.0				

(c) Klasický postup x [mm^2], varianta C y_3 [mm^2].

34.0	37.5,	40.7	45.0,	48.5	51.5,	55.5	59.5,	62.5	66.0,	70.0	73.5,
77.5	80.5,	84.2	87.5,	91.5	95.0,	97.2	102.5				

Úloha J6.14 *Závislost bodu vzplanutí na viskozitě oleje*

Byla sledována závislost bodu vzplanutí oleje y na viskozitě oleje x. (1) Nalezněte lineární regresní model a vyšetřete vlivné body. (2) V rámci exploratorní analýzy komentujte diagram rozptýlení a krabicový graf pro proměnnou x a y. (3) Jaký je Pearsonův korelační koeficient r mezi x a y? (4) Co ukazují grafy klasických reziduí $\hat{e}$ v závislosti na predikci $\hat{y}_P$ a na nezávisle proměnné x. (5) Jakou informaci přináší rankitové grafy rozličných druhů reziduí?

Data: Viskozita x [mm^2.s^{-1}], bod vzplanutí y [EC].

82.9	238,	90.14	222,
...	...	...	...
86.91	215,	92.6	210,

Úloha J6.15 *Porovnání dvou regresních kalibračních přímek*

UV spektrum účinné substance tetryzolinu vykazuje 2 maxima při vlnových délkách 214 a 248 nm. Pro určování koncentrace tetryzolinu v očních kapkách byly naměřeny kalibrační přímky při obou vlnových délkách. (1) Porovnejte tyto kalibrační přímky včetně testování jejich úseku a směrnice, s vyšetřením vlivných bodů a posouzením míry spolehlivosti navrženého modelu dle str. 380 v cit.[72]. (2) Proveďte test shodnosti dvou přímek, test jejich rovnoběžnosti a test společného úseku. (3) Jaké jsou míry přesnosti obou kalibračních přímek, jsou shodné? (4) Nalezněte koncentraci tetryzolinu v očních kapkách pro absorbanci $A^* = 0.100$, 0.200 a 0.800 z obou kalibrací. (5) Jaký bude 95 % intervalový odhad?

Data: Koncentrace x [mg/l], absorbance A_1 při 214 nm, absorbance A_2 při 248 nm.

1.0	0.102	0.265,	1.5	0.213	0.283,	2.0	0.385	0.344,	2.5	0.561	0.702,
4.5	0.646	1.113,	5.0	0.924	1.135,	7.0	1.133	1.417,	8.0	1.407	1.833,
9.0	1.598	2.106,	10.5	1.812	2.327,						

Úloha J6.16 *Porovnání dvou metod chemické analýzy meziproduktu ostazinové modře*

Ve výrobě ostazinové modře byla sledována závislost množství meziproduktu y na množství výchozí reakční komponenty x za jinak stejných provozních podmínek. Obsah meziproduktu v reakční směsi byl stanovován po skončení reakce jednak titrací dusitanem sodným y_1, jednak spektrofotometricky y_2. Očekávalo se, že bude-li potvrzena shoda lineárních závislostí, bude možné tvrdit, že obě metody analýzy jsou rovnocenné. (1) Testujte rovnoběžnost nebo shodnost obou přímek dle str. 383 v cit.[72]. (2) Užijte také Chowův test shody dvou lineárních modelů. (3) Testujte rovněž vhodnost lineárního modelu dle střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*.

Data: Množství výchozí reakční komponenty x [kg], obsah meziproduktu ostazinové modře titračně y_1 [kg] a spektrofotometricky y_2 [kg].

x:	9.0	10.2	11.6	12.0	5.5	13.0	10.5	10.5	11.5	9.0
y_1:	24.30	27.42	31.40	32.36	13.25	34.05	25.79	25.90	30.81	23.93
y_2:	24.55	27.93	31.80	31.65	12.83	35.24	28.20	27.39	32.53	25.48

Úloha J6.17 *Porovnání hospodaření dvou typů domácností*

V rámci průzkumu ekonomické situace rodin byl pořízen náhodný výběr 17 domácností "dělnického typu" D a 17 domácností "zemědělského typu" Z. Údaje se týkají měsíčních výdajů domácností za potraviny y a celkového měsíčního příjmu x těchto domácností v tisících Kč. Určete regresní přímky závislosti výdajů za potraviny y na příjmech x u domácností obou typů, a rozhodněte, zda lze obě regresní přímky považovat za shodné, tj. zda neexistuje statisticky významný rozdíl mezi rodinami typů D a Z ve výdajích za potraviny v závislosti na jejich příjmech.

Data: Příjem domácnosti x_{ij}, výdej domácnosti za potraviny y_{ij}, j = 1, 2.

Domácnosti typu D				Domácnosti typu Z			
x_{i1}	y_{i1}	x_{i1}	y_{i1}	x_{i2}	y_{i2}	x_{i2}	y_{i2}
14.29	5.18	11.45	4.67	19.73	7.13	13.74	2.45
...	...	...	...	...	...	...	...
7.91	4.58			24.23	5.47		

Úloha J6.18 *Porovnání vlivu dusíku na Ullmannovu reakci*

Při řešení technologie výroby modré báze MB H-3R byly provedeny kinetické pokusy, ke zjištění rozdílu mezi Ullmannovou reakcí, provedenou v atmosféře dusíku x a za přístupu vzduchu y. Byly provedeny tři reakce pod dusíkem, data A, B, C (na ose x) a tři na vzduchu, data D, E, F (na ose y). Bylo provedeno porovnání regresních přímek, aby se zjistilo, zda jsou opakovaná měření shodná. Zjistěte paralelnost nebo přímo shodnost tří přímek A-D, B-E, C-F dle str. 380 v cit.[72].

Data: x, y.

A:	0.00	0.04	2.83	0.13	5.53	0.23
	8.75	0.27	13.08	0.40	21.67	0.63
	24.33	0.68	29.17	0.80	39.42	1.03
	46.92	1.23				
	...	...	...	...	...	...
F:	0.00	0.01	3.33	0.16	5.55	0.23
	8.20	0.31	10.88	0.38	14.28	0.48
	17.17	0.56	20.50	0.65	22.87	0.73
	25.33	0.80				

Úloha J6.19 *Vliv dvou preparátů s mesalazinem na exkreci N-acetyl-beta-D-glukozaminidázy (NAG)*

Pacienti s idiopatickými střevními záněty jsou obvykle léčeni potenciálně nefroto-xickými preparáty s účinnou látkou mesalazinem, tj. 5-aminosalycilovou kyselinou. Každý ze

sledovaných souborů pacientů byl léčen jiným typem preparátu v rozdílném dávkování. Byla sledována exkrece N-acetyl-beta-D-glukozaminidázy (NAG) jako citlivého ukazatele tubulárního poškození. Je třeba vyšetřit, zda se liší nefrotoxicita obou preparátů, tj. zda jsou obě přímky shodné, či alespoň rovnoběžné dle str. 380 v cit.[72].

Data: Denní dávka x [g/den], NAG y [μg/l].

Preparát č. 1:		0.50	1.00,	0.75	6.20,	1.00	5.90,	1.25	7.8,
1.50	8.40,	1.75	7.10,	2.00	11.4,	2.25	15.6,	2.50	17.4,
Preparát č. 2:		0.50	3.90,	0.75	6.50,	1.00	7.90,	1.25	10.9,
1.50	13.2,	1.75	15.1,	2.00	17.3,	2.25	20.6,	2.50	27.8,

Úloha J6.20 *Shodnost dvou preparátů s mesalazinem na exkreci beta-2-mikroglobulinu*
Pacienti s idiopatickými střevními záněty jsou obvykle léčeni potenciálně nefrotoxickými preparáty s účinnou látkou mesalazinem. Každý ze sledovaných souborů pacientů byl léčen jiným typem preparátu v rozdílném dávkování. Sledovali jsme exkreci beta-2-mikroglobulinu jako citlivého ukazatele tubulárního poškození. Je třeba vyšetřit, zda se liší nefrotoxicita obou preparátů, tj. zda obě přímky jsou shodné či alespoň rovnoběžné dle str. 380 v cit.[72].

Data: Denní dávka x [g/den], beta-2-mikroglobulin y [μg/l].

Preparát č.1:		0.50	28,	0.75	41,	1.00	60,	1.25	88,
1.50	78,	1.75	128,	2.00	117,	2.25	159,	2.50	189,
Preparát č. 2:		0.50	60,	0.75	85,	1.00	115,	1.25	149,
1.50	162,	1.75	181,	2.00	178,	2.25	199,	2.50	201,

Úloha J6.21 *Porovnání závislosti příkonu na napětí u dvou žárovek*
Byla změřena závislost činného příkonu žárovky y na napětí x. Pro měření byl použit elektrodynamický wattmetr. Byly použity dvě žárovky, 100 W a 200 W. Úkolem je zjistit, zda zjištěné závislosti u obou žárovek jsou stejné, shodné či alespoň paralelní. Rozhodněte, zda lze závislosti aproximovat přímkami?

Data: Napětí x [V], příkon y_1 [W] pro žárovku 200 W a y_2 [W] pro žárovku 100 W.

0	0	0,	20	4	2,	40	14	6.5,	60	26	12,	80	42	18.5,	100	58	26,
120	78	35,	140	100	44,	160	120	54,	180	146	64,	200	172	76,	220	200	88.

Úloha J6.22 *Porovnání dvou titračních stanovení oxidu boritého*
Stanovení oxidu boritého za použití normované metody je založeno na jeho titraci odměrným roztokem hydroxidu sodného, c(NaOH) = 0.06 mol/l v přítomnosti mannitu. Při titraci známého předloženého množství oxidu boritého x v mg ve formě standardní látky, tj. kyseliny borité nebo dekahydrátu tetraboritanu sodného hydroxidem sodným y [ml] byla získána experimentální data. (1) Testujte lineární regresní model a existenci vlivných bodů. (2) Testujte statistickou významnost úseku β_0 a směrnice β_1. (3) Co ukazuje graf klasických reziduí e *vs.* x a co $\hat{e}$ *vs.* $\hat{y}_P$? (4) Komentujte vyšetření regresního tripletu s využitím regresních diagnostik.

Data: Spotřeba hydroxidu sodného y [ml] v závislosti na hmotnosti oxidu boritého x [mg].

x:	39.5	22.5	25.0	26.4	23.8
y:	12.9	7.5	8.5	8.9	8.0
..	...	...	...	...	...
x:	33.8	25.8	28.0		
y:	11.2	8.7	9.3		

Úloha J6.23 *Závislost systolického krevního tlaku na stáří jedince*
U výběru 24 jedinců ve věku 21 až 70 let, náhodně vybraných ze stejné etnické skupiny, byl po dobu 2 týdnů denně vždy v 8 hodin ráno měřen systolický krevní tlak, str. 167 v cit.[70]. Nalezněte regresní model závislosti krevního tlaku y [mm] na stáří jedince x [roky] a odpovězte na následující otázky: (1) Jaký je průměrný věk jedinců celé etnické skupiny? (2) Jaký je průměrný krevní tlak jedinců celé etnické skupiny? (3) Jaký bude krevní tlak jedinců starých 65 let? Jak spolehlivě můžeme tento krevní tlak odhadnout, pracujeme-li s 95% statistickou jistotou? (4) O kolik mm se zvýší krevní tlak jedince každým rokem? V jakém rozmezí bude tato hodnota, odhadujeme-li tento krevní tlak s 95% statistickou jistotou? (5) Na základě dat tohoto výběru odhadněte krevní tlak novorozence. V jakém intervalu bude tato hodnota, odhadujeme-li krevní tlak novorozence s 95% statistickou jistotou? (6) Vypočtěte 95% oboustranný intervalový odhad predikovaného krevního tlaku pro jedince staré 60 a 55 let. (7) Analýzou regresního tripletu s využitím regresní diagnostiky potvrďte navržený regresní model a odhalte i vlivné body.
Data: Věk x [roky], systolický krevní tlak y [mm Hg sloupce].

34	116,	26	112,	51	151,	58	161,	34	122,
...	...	...	...	...	...	...	...	...	...
57	159,	66	177,	42	135,	53	149,		

Úloha J6.24 *Závislost tělesného tuku atletů-běžců na obsahu tuku ve stravě*
Cílem studie bylo nalézt závislost mezi tělesným tukem lehkých atletů-běžců y, kteří týdně trénují asi 12 hodin, a zkonzumovaným tukem v jejich každodenní stravě x (str. 202 v cit[70]). U náhodného vzorku 18 běžců byl měřen jejich tělesný podkožní tuk y [%] a sledován v závislosti na zkonzumovaném tuku ve stravě x [%]. Ověřte, zda lze uvedenou závislost popsat jednoduchým lineárním regresním modelem $y = \beta_0 + \beta_1 x$. (1) Jaký lze očekávat tělesný tuk u běžce, který spotřeboval ve stravě 25 % tuku? (2) Jaké procento tuku ve stravě očekává běžec, který má tělesný tuk 25 %? Uveďte i rozmezí této hodnoty, a to s 95 % statistickou jistotou. (3) Analýzou regresního tripletu s využitím regresní diagnostiky potvrďte navržený regresní model a odhalte také vlivné body. (4) Komentujte rankitové grafy rozličných druhů reziduí.
Data: Spotřebovaný tuk ve stravě x [%], tělesný podkožní tuk y [%].

22	9.8,	22	11.7,	14	8.0,	21	9.7,	32	10.9,	26	7.8,
...	...	...	...	...	...	...	...	...	...	...	...
24	12.0,	36	11.6,	20	10.4,	37	10.8,	35	11.5,	14	7.9,

Úloha J6.25 *Závislost celkového cholesterolu v krvi na denní spotřebě tuku*
U náhodného vzorku 20 Američanů byla provedena analýza krve a sledována denní spotřeba tuku ve stravě x v gramech a hodnota celkového cholesterolu y v mg na 100 ml krve, str. 215 v cit[70]. Pro tuto závislost byl navržen jednoduchý lineární regresní model $y = \beta_0 + \beta_1 x$. Na základě analýzy regresního tripletu s využitím regresní diagnostiky dokažte (1) platnost navrženého regresního modelu a existenci vlivných bodů. (2) Testujte statistickou významnost obou parametrů, úseku β_0 a směrnice β_1. (3) Sestrojte 95%ní jednostranný interval spolehlivosti úseku β_0, a dále vysvětlete fakt, že $\beta_0 = 0$. (4) Sestrojte 95 % oboustranný interval spolehlivosti směrnice β_1. (5) Sestrojte také 99.5 % interval spolehlivosti směrnice β_1 a vysvětlete oba konfidenční pásy. (6) Nalezněte 95 % interval

spolehlivosti celkového cholesterolu u lidí, kteří denně spotřebují 50 g tuku. (7) Jaký je Pearsonův korelační koeficient mezi celkovým cholesterolem v krvi *y* a denní spotřebou tuku *x* u sledovaných jedinců? (8) Testujte nulovou hypotézu H_0: $\beta_1 \# 2$ vs. H_A: $\beta_1 > 2$ a komentujte výsledek testování.

Data: Denní spotřeba tuku *x* [g], obsah celkového cholesterolu v krvi *y* [mg/100 ml].

21	130,	29	163,	43	169,	52	136,	56	187,	64	193,
...	...	...	...	...	...	...	...	...	...	...	...
148	297,	157	316,								

6.2 Validace nové analytické metody

Vzorová úloha 6.2 *Postup validace a regresní diagnostika*

Na úloze **V6.14** *Validace stanovení amonných iontů v pitných vodách* proveďte ověření časově nenáročné metody stanovení obsahu amonných iontů *y* soupravou Spektroquant, a to srovnáním se standardní metodou *x* stanovení amoniaku podle ČSN - ISO 7150-1, která je však náročná na provedení. Pro účely vyhodnocení se předpokládá, že rozptyl obsahu u standardní metody je zanedbatelný. (1) Vyšetřete statistickou významnost úseku b_0 (má být $\beta_0 = 0$). (2) Odstraňte z dat odlehlé hodnoty. (3) K jakým závěrům vede kombinovaný test úseku a směrnice?

Řešení:

1. Návrh modelu: navrhneme regresní model (přímky) $y = \beta_0 + \beta_1 x$, u kterého budeme testovat nulovou hypotézu H_0: $\beta_0 = 0$, $\beta_1 = 1$.

2. Předběžná analýza dat: poloha a proměnlivost proměnných *y, x* se posuzuje na základě *průměru* a *směrodatné odchylky* hodnot každé proměnné. *Pearsonův párový korelační koeficient* ukazuje vysokou korelaci proměnných *y* a *x*.

Proměnná	Průměr	Směrodatná odchylka	Párový korelační koeficient	Spočtená hladina významnosti
y	2.0443E-01	1.9296E-01	1.0000	-----
x	2.1048E-01	2.0848E-01	0.9958	0.000

3. Odhadování parametrů: klasickou metodou nejmenších čtverců (MNČ) byly nalezeny odhady parametrů, úseku β_0 a směrnice β_1. Studentův *t*-test ukázal, že úsek (absolutní člen) β_0 je statisticky nevýznamný, zatímco směrnice β_1 je statisticky významná, když $t_{0.95}(20\text{-}2) = 2.101$.

Parametr	Odhad	Směrodatná odchylka	H_0: $b_j = 0$ vs. H_A: $b_j \neq 0$ *t*-kritérium	hypotéza H_0 je	Spočtená hladina významnost.
b_0	0.010432	0.00566	1.8436	Akceptována	0.081
b_1	0.92170	0.01933	47.681	Zamítnuta	0.000

4. Základní statistické charakteristiky: *párový korelační koeficient r* ukazuje, že navržený lineární regresní model je statisticky významný. Vysoká hodnota *koeficientu determinace D* (= 99.17 %), představující procento bodů vyhovujících regresnímu modelu, ukazuje, že všechny body výtečně korespondují s modelem přímky. *Střední kvadratická*

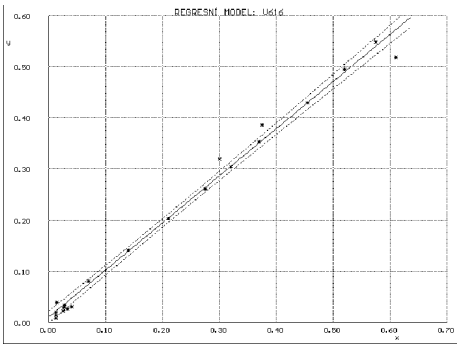
chyba predikce MEP a *Akaikovo informační kritérium AIC* se užívají k rozlišení mezi několika navrženými modely. Za optimální se považuje model, pro který dosahuje *MEP* a *AIC* minimální hodnotu.

Vícenásobný korelační koeficient, *r*	: 0.99585
Koeficient determinace, *D*[%]	: 99.171
Predikovaný koeficient determinace, R^2_P	: 0.99414
Střední kvadratická chyba predikce, *MEP*	: 4.1406E-04
Akaikovo informační kritérium, *AIC*	: -166.78

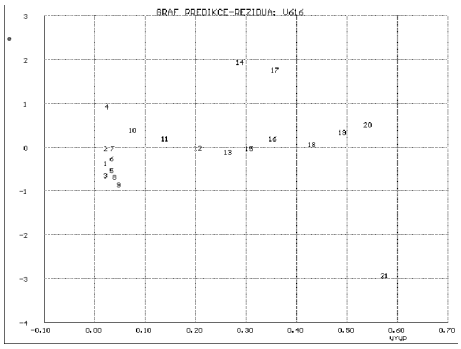
5. Regresní diagnostika: obsahuje pomůcky a postupy pro interaktivní analýzu (a) dat, (b) modelu, (c) metody, což jsou složky tzv. *regresního tripletu.*

Kritika dat: věrohodnost nalezených odhadů parametrů β_0, β_1 lze posoudit na základě grafu regresního modelu (obr. 6.2-1a).

(a) *Analýza klasických reziduí* není příliš spolehlivá a nemusí indikovat silně odlehlé hodnoty. Grafická analýza $\hat{e}$ vs. $\hat{y}_P$ (obr. 6.2-1b) je schopna indikovat podezřelé body, trend a heteroskedasticitu. Míry polohy a rozptýlení klasických reziduí by měly dosahovat hodnot, blízkých experimentálnímu šumu. *Odhad směrodatné odchylky s(e)* se totiž blíží svou velikostí experimentální chybě, kterou je zatížena závisle proměnná. Odhad šikmosti a špičatosti nedokazují Gaussovo normální rozdělení reziduí, normalitu.



Obr. 6.2-1a Graf regresního modelu**,** ***ADSTAT.***



Obr. 6.2-1b Analýza klasických reziduí, ***ADSTAT.***

Bod	**Měřená hodnota**	**Predikovaná hodnota**	**Směrodatná odchylka**	**Klasické reziduum**	**Relativní reziduum**
i	$y_{exp,\,i}$	$y_{vyp,\,i}$	$s(y_{vyp,\,i})$	e_i	$e_{r,\,i}$
1	1.5000E-02	2.1493E-02	5.4943E-03	-6.4929E-03	-4.3286E+01
2	2.1000E-02	2.1493E-02	5.4943E-03	-4.9287E-04	-2.3470E+00
3	1.0000E-02	2.1493E-02	5.4943E-03	-1.1493E-02	-1.1493E+02
4	4.0000E-02	2.3336E-02	5.4674E-03	1.6664E-02	4.1659E+01
5	2.4000E-02	3.3475E-02	5.3219E-03	-9.4750E-03	-3.9479E+01
6	3.0000E-02	3.4397E-02	5.3089E-03	-4.3967E-03	-1.4656E+01
7	3.5000E-02	3.5318E-02	5.2959E-03	-3.1838E-04	-9.0966E-01
8	2.8000E-02	3.9927E-02	5.2317E-03	-1.1927E-02	-4.2596E+01
9	3.2000E-02	4.7300E-02	5.1310E-03	-1.5300E-02	-4.7814E+01

10	8.2000E-02	7.4952E-02	4.7793E-03	7.0485E-03	8.5957E+00
11	1.4300E-01	1.3947E-01	4.1622E-03	3.5294E-03	2.4681E+00
12	2.0400E-01	2.0399E-01	3.9329E-03	1.0337E-05	5.0673E-03
13	2.6200E-01	2.6390E-01	4.1260E-03	-1.9002E-03	-7.2528E-01
14	3.2200E-01	2.8694E-01	4.2968E-03	3.5057E-02	1.0887E+01
15	3.0500E-01	3.0538E-01	4.4666E-03	-3.7676E-04	-1.2353E-01
16	3.5500E-01	3.5146E-01	4.9977E-03	3.5382E-03	9.9666E-01
17	3.8800E-01	3.5607E-01	5.0579E-03	3.1930E-02	8.2293E+00
18	4.3100E-01	4.2981E-01	6.1490E-03	1.1936E-03	2.7693E-01
19	4.9600E-01	4.8972E-01	7.1601E-03	6.2830E-03	1.2667E+00
20	5.5000E-01	5.4041E-01	8.0697E-03	9.5895E-03	1.7435E+00
21	5.2000E-01	5.7267E-01	8.6668E-03	-5.2670E-02	-1.0129E+01

Reziduální součet čtverců, *RSC*	: 6.1717E-03
Průměr absolutních hodnot reziduí, M_e	: 1.0937E-02
Průměr relativních reziduí, M_{erel}	: 1.8720E+01
Odhad reziduálního rozptylu, $s^2(e)$	: 3.2482E-04
Odhad směrodatné odchylky reziduí, $s(e)$	: 1.8023E-02
Odhad šikmosti reziduí, $g_1(e)$	: -6.8328E-01
Odhad špičatosti reziduí, $g_2(e)$	: 5.7550E+00

(b) ***Analýza ostatních reziduí:*** Jackknife rezidua indikují odlehlé body, z diagonál-ních prvků H_{ii} projekční matice $\boldsymbol{H}$ a diagonálních prvků H_{mii} zobecněné projekční matice $\boldsymbol{H_m}$ pouze extrémy. Ostatní druhy reziduí a kritéria v tabulce indikují obecně vlivné body (značeno hvězdičkou u hodnoty). Jackknife rezidua $e_{J,i}$ ukazují, že body č. 14 a 21 jsou odlehlé, stejně tak i Cookova vzdálenost D_i; Atkinsonova vzdálenost A_i na č. 14, 17, 21; kritérium DF_i na č. 21, věrohodnostní vzdálenosti $LD(b)_i$, $LD(s^2)_i$ na č. 21 a $LD(b, s^2)_i$. Diagonální prvky H_{ii} projekční matice $\boldsymbol{H}$ ukazují na extrémy č. 20, 21, a diagonální prvky zobecněné H_{mii} projekční matice $\boldsymbol{H_m}$ pak na extrémy č. 21.

INDIKACE VLIVNÝCH BODU: (* indikuje odlehlý nebo vlivný bod)

Bod	**Standardizované reziduum**	**Jackknife reziduum**	**Predikované reziduum**	**Diagonální prvky**
i	e_{Si}	e_{Ji}	e_{Pi}	H_{ii}
1	-3.7826E-01	-3.6957E-01	-7.1581E-03	9.2936E-02
2	-2.8713E-02	-2.7948E-02	-5.4336E-04	9.2936E-02
3	-6.6955E-01	-6.5952E-01	-1.2670E-02	9.2936E-02
4	9.7031E-01	9.6874E-01	1.8353E-02	9.2027E-02
5	-5.5026E-01	-5.3990E-01	-1.0380E-02	8.7194E-02
6	-2.5528E-01	-2.4889E-01	-4.8144E-03	8.6768E-02
7	-1.8481E-02	-1.7989E-02	-3.4847E-04	8.6345E-02
8	-6.9154E-01	-6.8173E-01	-1.3024E-02	8.4263E-02
9	-8.8560E-01	-8.8034E-01	-1.6650E-02	8.1052E-02
10	4.0561E-01	3.9651E-01	7.5816E-03	7.0320E-02
11	2.0127E-01	1.9611E-01	3.7282E-03	5.3333E-02
12	5.8773E-04	5.7205E-04	1.0854E-05	4.7619E-02
13	-1.0831E-01	-1.0546E-01	-2.0053E-03	5.2408E-02
14	2.0029E+00	2.1949E+00*	3.7170E-02	5.6839E-02
15	-2.1578E-02	-2.1003E-02	-4.0142E-04	6.1418E-02
16	2.0433E-01	1.9910E-01	3.8329E-03	7.6894E-02
17	1.8458E+00	1.9831E+00	3.4659E-02	7.8758E-02
18	7.0453E-02	6.8582E-02	1.3508E-03	1.1640E-01
19	3.7988E-01	3.7116E-01	7.4605E-03	1.5783E-01
20	5.9505E-01	5.8466E-01	1.1994E-02	2.0048E-01*

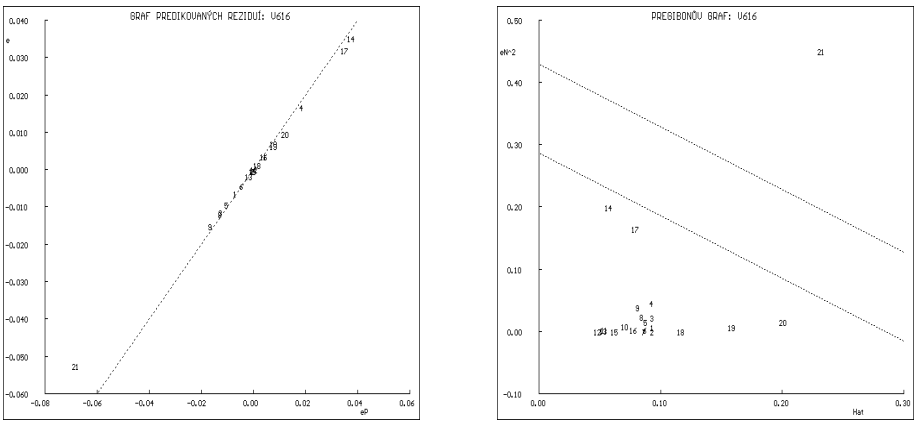
21	-3.3331E+00	-5.0342E+00*	-6.8513E-02	2.3124E-01*

Bod	**Zobecněné diag. prvky**	**Cookova vzdálenost**	**Atkinsonova vzdálenost**	**Vliv na predikci**
i	H_{mii}	D_i	A_i	DF_i
1	9.9767E-02	7.3300E-03	3.6461E-01	-1.1830E-01
2	9.2975E-02	4.2236E-05	2.7573E-02	-8.9460E-03
3	1.1434E-01	2.2966E-02	6.5067E-01	-2.1111E-01
4	1.3702E-01	4.7713E-02	9.5058E-01	3.0841E-01
5	1.0174E-01	1.4461E-02	5.1431E-01	-1.6687E-01
6	8.9900E-02	3.0958E-03	2.3647E-01	-7.6720E-02
7	8.6361E-02	1.6139E-05	1.7045E-02	-5.5300E-03
8	1.0731E-01	2.2002E-02	6.3739E-01	-2.0680E-01
9	1.1898E-01	3.4587E-02	8.0583E-01	-2.6145E-01
10	7.8370E-02	6.2219E-03	3.3611E-01	1.0905E-01
11	5.5351E-02	1.1411E-03	1.4347E-01	4.6548E-02
12	4.7619E-02	8.6356E-09	3.9426E-04	1.2791E-04
13	5.2994E-02	3.2441E-04	7.6440E-02	-2.4800E-02
14	2.5598E-01	1.2088E-01*	1.6608E+00	5.3883E-01
15	6.1441E-02	1.5234E-05	1.6559E-02	-5.3726E-03
16	7.8922E-02	1.7389E-03	1.7711E-01	5.7462E-02
17	2.4395E-01	1.4563E-01*	1.7872E+00	5.7985E-01
18	1.1663E-01	3.2694E-04	7.6724E-02	2.4892E-02
19	1.6423E-01	1.3522E-02	4.9524E-01	1.6068E-01
20	2.1538E-01	4.4394E-02	9.0237E-01	2.9277E-01
21	6.8074E-01*	1.6709E+00*	8.5100E+00*	-2.7610E+00*

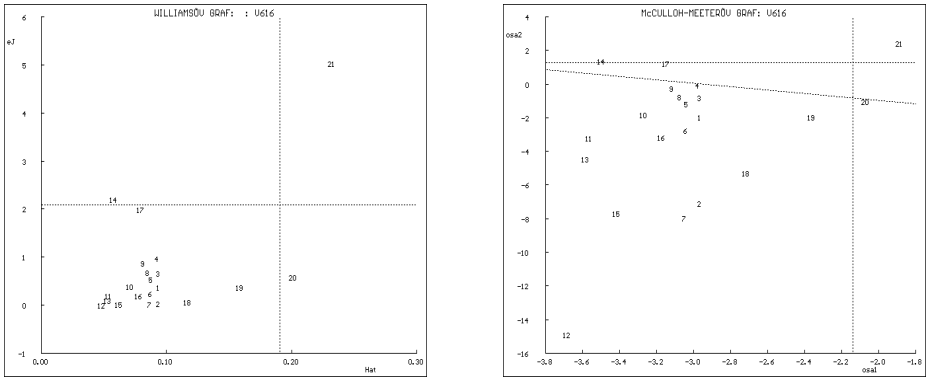
Bod	**Věrohodnostní vzdálenosti**		
i	$LD(b)_i$	$LD(s^2)_i$	$LD(b,s^2)_i$
1	1.6197E-02	1.7607E-02	3.3156E-02
2	9.3364E-05	2.4550E-02	2.4639E-02
3	5.0706E-02	6.4629E-03	5.5981E-02
4	1.0521E-01	4.3416E-05	1.0573E-01
5	3.1943E-02	1.1124E-02	4.2062E-02
6	6.8422E-03	2.1276E-02	2.7816E-02
7	3.5677E-05	2.4575E-02	2.4609E-02
8	4.8581E-02	5.6554E-03	5.3172E-02
9	7.6317E-02	4.6137E-04	7.6411E-02
10	1.3749E-02	1.6655E-02	2.9869E-02
11	2.5223E-03	2.2505E-02	2.4912E-02
12	1.9089E-08	2.4593E-02	2.4593E-02
13	7.1711E-04	2.3980E-02	2.4663E-02
14	2.6552E-01	3.9713E-01	7.1973E-01
15	3.3675E-05	2.4569E-02	2.4601E-02
16	3.8434E-03	2.2442E-02	2.6111E-02
17	3.1948E-01	2.4454E-01	6.1812E-01
18	7.2270E-04	2.4333E-02	2.5021E-02
19	2.9870E-02	1.7552E-02	4.6238E-02
20	9.7905E-02	9.3426E-03	1.0458E-01
21	3.4023E+00	9.7291E+00*	1.8199E+01*

(c) ***Grafy vlivných bodů*** (obr. 6.2-2) jsou schopny indikovat přítomnost odlehlých hodnot a extrémů. *Graf predikovaných reziduí* ukazuje na odlehlé body č. 21, 14, 17. *Pregibonův graf* ukazuje na silně vlivný bod č. 21. *Williamsův graf* indikuje č. 14 a 21 jako odlehlé body a jako extrémy č. 20, 21. *McCullohův-Meeterův graf* dokazuje odlehlé body

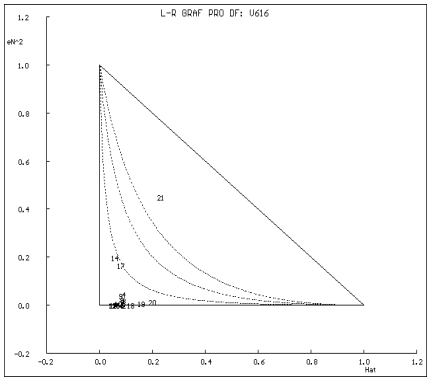
č. 14, 17, 21 a extrémy č. 20, 21. Konečně *L-R graf* dokazuje odlehlé body č. 14, 17, 21 a současně extrém č. 20. Lze uzavřít, že body č. 14, 21 jsou většinou diagnostik indikovány jako odlehlé.



Obr. 6.2-2 Grafy vlivných bodů, vlevo, graf predikovaných reziduí, a vpravo, Pregibonův graf, ***ADSTAT.***

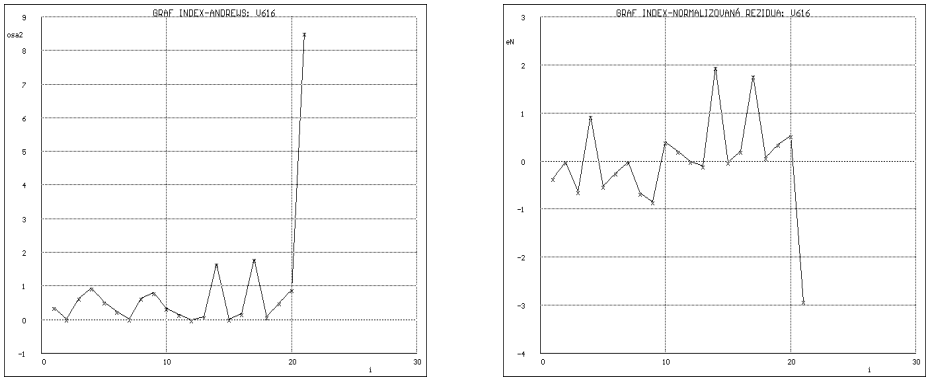


Obr. 6.2-2 Grafy vlivných bodů, vlevo, Williamsův graf, a vpravo, McCullohův-Meeterův graf, ***ADSTAT.***

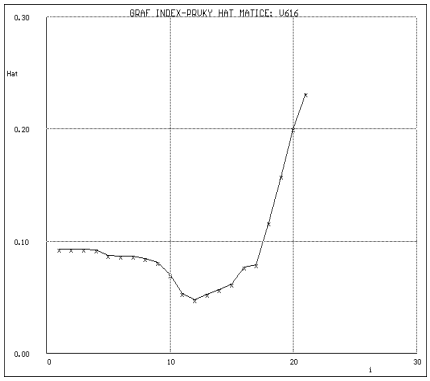


Obr. 6.2-2 Grafy vlivných bodů, L-R graf, ***ADSTAT.***

(d) ***Indexové grafy*** (obr. 6.2-3) upozorňují pouze na podezřelé body. *Andrewsův indexový graf* a *graf normovaných reziduí* ukazují na podezřelé body č. 14, 17 a 21. *Indexový graf prvků* ***H*** *projekční matice* pak na podezřelé extrémy č. 21.

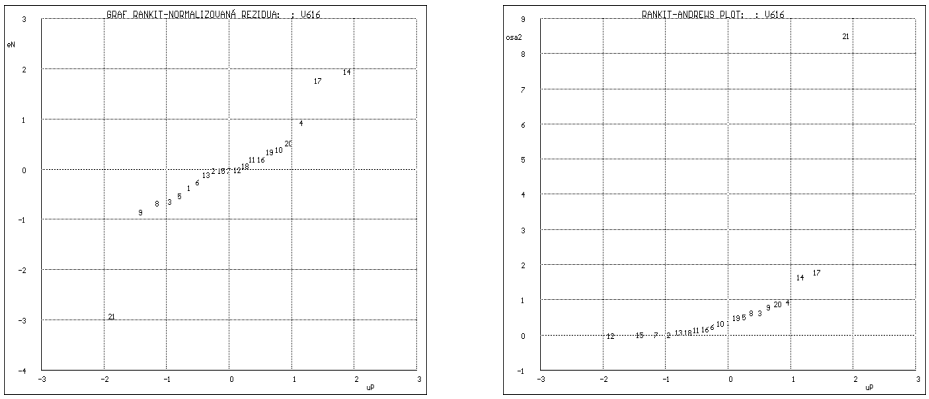


Obr. 6.2-3 Indexové grafy, vlevo: Andrewsův graf, a vpravo: graf normovaných reziduí, ***ADSTAT.***



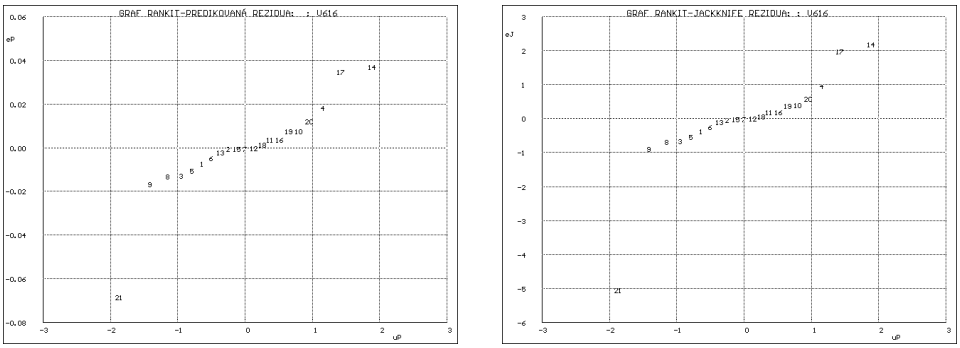
Obr. 6.2-3 Graf prvků H-projekční matice, ***ADSTAT.***

(e) ***Rankitové grafy*** (obr. 6.2-4) ukazují vedle normality rozdělení dotyčných reziduí i na vlivné (zde odlehlé) body.



Obr. 6.2-4 Rankitové grafy, vlevo, graf normovaných reziduí, a vpravo, Andrewsův graf, ***ADSTAT.***

Graf normovaných reziduí ukazuje na č. 21 a na č. 17 a 14 jako na odlehlé body. *Andrewsův graf* představuje č. 21 jako odlehlý bod. *Graf predikovaných reziduí* a *graf Jackknife reziduí* č. 21, 14, 17 jako odlehlé body.



Obr. 6.2-4 Rankitové grafy, vlevo, graf predikovaných reziduí, a vpravo, graf Jackknife reziduí, ***ADSTAT.***

Model: *Parciální regresní grafy* a *parciální reziduální grafy* jsou určeny pro vícerozměrné lineární regresní modely a nemají proto smysl u jednorozměrného regresního modelu. Vhodnost modelu se posuzuje přímo v grafu obsahujícím data a průběh modelové funkce. Je patrné, že v tomto případě je přímka akceptovatelná a data nevykazují nelineární průběh.

Metoda: do této části patří vyšetření splnění základních předpokladů metody nejmenších čtverců (MNČ), za kterých by měla metoda vést k nejlepším lineárním nestranným odhadům regresních parametrů:

Fisherův-Snedecorův test významnosti regrese potvrdil, že navržený model je přijat jako významný.

Scottovo kritérium multikolinearity nemá smysl u jednorozměrného regresního modelu.

Cookův-Weisbergův test heteroskedasticity dokazuje, že rezidua vykazují heteroskedasticitu (nekonstantnost rozptylu).

Jarqueův-Berraův test normality reziduí ukazuje, že klasická rezidua nevykazují Gaussovo rozdělení.

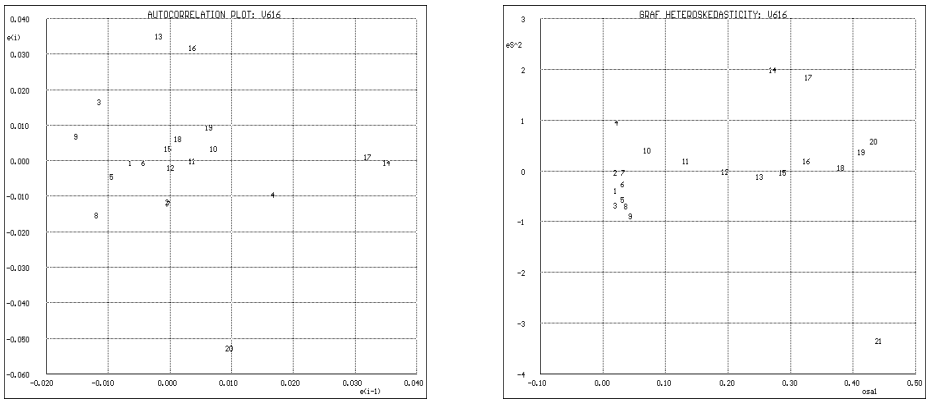
Waldův test autokorelace ukazuje, že klasická rezidua nejsou autokorelována. To by bylo totiž vážným upozorněním ke zhodnocení provedeného experimentu, že došlo k narušení podmínek. Mnohdy však může způsobit heteroskedasticitu i jeden odlehlý bod.

Znaménkový test prokazuje, že znaménko klasických reziduí se dostatečně střídá, a proto rezidua nevykazují žádný trend.

TESTOVÁNÍ REGRESNÍHO TRIPLETU (DATA + MODEL + METODA):	
Fisherův-Snedocorův test významnosti regrese, F_{exp}	: 2273.5
Tabulkový kvantil, $F_{1-\alpha}(m\text{-}1, n\text{-}m)$	: 4.3807
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	: 0.000
Scottovo kritérium multikolinearity, M	: 3.0004E-15
Závěr: Navržený model je korektní.	
Cookův-Weisbergův test heteroskedasticity, S_f	: 902.50
Tabulkový kvantil, $\chi^2_{1-\alpha}(1)$	: 3.8415
Závěr: Rezidua vykazují heteroskedasticitu.	
Spočtená hladina významnosti	: 0.000
Jarqueův-Berraův test normality reziduí, $L(e)$	: 8.2754
Tabulkový kvantil, $\chi^2_{1-\alpha}(2)$	: 5.9915
Závěr: Normalita není přijata.	
Spočtená hladina významnosti	: 0.016
Waldův test autokorelace, W_a	: 0.5898
Tabulkový kvantil, $\chi^2_{1-\alpha}(2)$	3.8415
Závěr: Rezidua nejsou autokorelována.	
Spočtená hladina významnosti	: 0.000
Znamékový test, D_t	:-0.8870
Tabulkový kvantil, $N_{1-\alpha/2}$	: 1.6449
Závěr: Rezidua nevykazují trend.	
Spočtená hladina významnosti	: 0.188

Graf autokorelace (obr. 6.2-5) vykazuje náhodný mrak bodů reziduí.

Graf heteroskedasticity (obr. 6.2-5) vykazuje trend, klín, což odpovídá heteroskedasticitě, nekonstantnosti rozptylu.



Obr. 6.2-5 Vlevo, graf autokorelace, a vpravo, graf heteroskedasticity, ***ADSTAT.***

6. Konstrukce zpřesněného modelu: (a) Po odstranění bodů č. 14, 17, 21 byly nalezeny nové odhady parametrů zpřesněného modelu.

Parametr	**Odhad**	**Směrodatná odchylka**	**H_0: $b_j = 0$ vs. H_A: $b_j \ldots 0$** *t*-**kritérium**	**hypotéza H_0 je**	**Spočtená hlad. význam.**
b_0	0.00639	0.002412	2.6490	Zamítnuta	0.018
b_1	0.94034	0.009356	100.51	Zamítnuta	0.000

Zpřesněný model (v závorce je uveden odhad směrodatné odchylky parametru)

$$y = 0.00639\ (0.00241) + 0.9403\ (0.0094)\ x$$

je doložen statistickými charakteristikami: *střední kvadratická chyba predikce MEP* a *Akaikovo informační kritérium AIC* dosáhly nižších hodnot, čímž dokazují kvalitnější model než předešlý.

Vícenásobný korelační koeficient, *r*	: 0.99921
Koeficient determinace, 100 *% D*	: 99.842
Predikovaný koeficient determinace, R^2_P	: 0.9990
Střední kvadratická chyba predikce, *MEP*	: 6.1534E-05
Akaikovo informační kritérium, *AIC*	: -174.03

Rezidua nyní vykazují normální rozdělení a nevykazují trend, stále však vykazují heteroskedasticitu, a proto lze doporučit užití metody vážených nejmenších čtverců.

(b) Užitím statistické váhy ($w_i = 1/y_i^2$) kompenzujeme heteroskedasticitu v datech. Obdržíme nové správnější odhady parametrů.

Parametr	**Odhad**	**Směrodatná odchylka**	**H_0: $b_j = 0$ vs. H_A: $b_j \ldots 0$** *t*-**kritérium**	**hypotéza H_0 je**	**Spočtená hlad. význam.**
b_0	0.002129	0.001974	1.0782	Akceptována	0.297
b_1	0.94617	0.073078	12.947	Zamítnuta	0.000

Opravený model má tvar, (v závorce je vždy uveden odhad směrodatné odchylky parametru):

$$y = 0.00213\ (0.00197) + 0.9462\ (0.0731)\ x.$$

Jelikož došlo ke snížení rozhodujících kritérií, *střední kvadratické chyby predikce MEP* a *Akaikova informačního kritéria AIC,* lze považovat tyto odhady za lepší než předešlé. Pearsonův korelační koeficient *r*, a tím pádem i koeficient determinace *D* vychází nepatrně horší než u předešlého odhadu bez statistické váhy.

Vícenásobný korelační koeficient, *r*	: 0.95544
Koeficient determinace, 100 *% D*	: 91.287
Predikovaný koeficient determinace, R^2_P	: 0.93371
Střední kvadratická chyba predikce, *MEP*	: 5.1179E-05
Akaikovo informační kritérium, *AIC*	: -181.63

7. Zhodnocení kvality modelu: nalezený model má tvar (v závorce je vždy uveden odhad směrodatné odchylky parametru)

$$y = 0.00213\ (0.00197) + 0.9462\ (0.0731)\ x$$

a intervalový odhad parametrů úseku β_0 a směrnice β_1 bude

$$b_0\ \&\ t_{1\&\alpha/2}(18)\sqrt{D(b_0)}\ \ \#\ \ \beta_0\ \ \#\ \ b_0\ \%\ t_{1\&\alpha/2}(18)\sqrt{D(b_0)}$$

a po dosazení

$$0.00213 - 2.12 \times 0.00197 \# \beta_0 \# 0.00213 + 2.12 \times 0.00197$$

vyjde

$$-0.00205 \# \beta_0 \# 0.00630.$$

Tento interval spolehlivosti úseku regresní přímky zahrnuje nulu, takže lze úsek β_0 považovat za nulový.

Analogicky dosazením do intervalu spolehlivosti směrnice obdržíme nerovnost

$$0.9462 - 2.12 \times 0.0731 \# \beta_1 \# 0.9462 + 2.12 \times 0.0731$$

a po vyčíslení

$$0.7912 \# \beta_1 \# 1.1012.$$

Jelikož *tento interval obsahuje jedničku, lze považovat směrnici β_1 za jednotkovou.*

Lze uzavřít, že úsek regresní přímky lze považovat za nulový $\beta_2 = 0$ a směrnice β_1 není významně odlišná od jedničky. Výsledky nové metody se proto statisticky významně neliší od metody standardní.

6.2.1 Úlohy na validaci nové analytické metody

Úloha V6.01 *Validace stanovení molybdenu rentgenově-fluorescenční metodou*
Správnost a přesnost metody stanovení obsahu molybdenu lze určit porovnáním výsledků z rentgenově-fluorescenční metody *y* s deklarovanými hodnotami obsahu u standardů ocelí *x*, str. 80 v cit[58]. (1) Určete velikost systematické chyby metody, která se rovná velikosti úseku β_0 v modelu $y = \beta_0 + \beta_1 x$, a dále správnost metody, pro kterou by směrnice měla být 1. (2) U řady kontrolních standardů se pokuste vyjádřit i přesnost metody. (3) Jsou v datech vlivné a vybočující body?

Data: Obsah molybdenu, dáno *x* [%], stanoveno *y* [%].

0.018	0.019,	0.068	0.069,	0.045	0.045,	0.052	0.051,	0.061	0.060,	0.075	0.075,
0.035	0.034,	0.025	0.025,	0.039	0.037,	0.085	0.083,	0.011	0.012,	0.014	0.016,

Úloha V6.02 *Bichromátometrická metoda stanovení železitých iontů*
Kraft a Dosch[60] navrhli titrační stanovení železa v odpadních vodách. Železité ionty Fe^{3+} v oxidu železitém Fe_2O_3 se redukují titanitou solí v přebytku a vzniklé železnaté ionty Fe^{2+} se pak určí bichromátometricky. Metoda byla validována na navážku *x* oxidu železitého.

(1) Vede titrační stanovení ke správným výsledkům? (2) Proveďte Studentův *t*-test úseku b_0 (má být $\beta_0 = 0$) a směrnice b_1 (má být $\beta_1 = 1$), dále kombinovaný test obou parametrů v modelu y ' β_0 % β_1x dle str. 352 v cit.[72].

Data: Obsah Fe_2O_3 [mg], dáno *x*, nalezeno *y*.

515	514.42,	333.07	333.80,	533.17	532.93,	499.07	498.47,	543.61	543.78,
...	...	...	...	...	...	...	...	...	...
400.10	400.20,	350.50	350.30.						

Úloha V6.03 *Metoda iontové chromatografie stanovení bromidů a jodidů*

Ke srovnání metody iontové chromatografie IC stanovení bromidů a jodidů *y* s extrakční fotometrickou metodou EX *x* (cit.[61]) bylo změřeno 13 vzorků povrchových vod, znečištěných salinními odpadními vodami báňského průmyslu. Předpokládejte, že fotometrická metoda EX poskytuje správné informace *x* se zanedbatelnou chybou. (1) Testujte shodnost obou metod a vyhodnoťte případnou systematickou chybu metody IC. (2) Jaká je přesnost ověřované metody IC? (3) Vyskytují se v datech odlehlé body?

Data: (a) Koncentrace bromidů [mg. l^{-1}], *x* EX, *y* IC.

7.10	6.30,	1.56	1.29,	8.10	7.80,	14.30	14.10,	11.66	11.43,
...	...	...	...	...	...	...	...	...	...
1.43	1.92,	5.40	5.64,	9.97	7.59,	4.56	4.10,	6.39	6.53,

(b) Koncentrace jodidů [mg. l^{-1}], *x* EX, *y* IC.

0.02	0.03,	0.02	0.14,	0.21	0.35,	3.90	3.04,	2.12	1.71,	0.60	0.73,
...	...	...	...	...	...	...	...	...	...	...	...
3.10	2.50,	1.74	1.55,	0.88	0.90,						

Úloha V6.04 *Stanovení kyseliny ftalové tenkovrstvou chromatografií*

Obsah kyseliny ftalové byl stanoven tenkovrstvou chromatografií a chromatogram byl vyhodnocován remisním fotometrem[66]. (1) Stanovte parametry lineárního regresního modelu a testujte, zda je úsek nulový a směrnice jednotková. (2) Jsou v datech vybočující hodnoty? (3) Lze dospět k závěru, že stanovení je správné?

Data: Obsah kyseliny ftalové [µg], dáno *x*, nalezeno *y* (opakovaně).

0.50	0.48,	0.49	0.51,	0.50	0.47,	0.99	1.00,	0.99	0.98,
...	...	...	...	...	...	...	...	...	...
5.23	5.02,								

Úloha V6.05 *Validace stanovení benfluralinu metodou GC vůči HPLC*

Pro srovnání standardní metody plynové chromatografie GC při stanovení obsahu účinné látky benfluralinu *x* v přípravku Balan s metodou HPLC *y* bylo změřeno 13 různých šarží vzorků přípravku Balan. (1) Použijte metodu ortogonální regrese za předpokladu, že rozptyl obou metod je shodný. (2) Je testovaná metoda GC zatížena vůči standardní HPLC nějakou systematickou odchylkou?

Data: Obsah benfluralinu [mg/100 ml], GC *x*, HPLC *y*.

x:	17.41	17.03	16.41	16.53	17.04	17.37	16.81	16.94	16.76	17.09
	17.42	17.33	16.91	16.61	17.18	16.45				

y:	17.23	16.96	16.15	16.49	16.83	16.92	16.56	16.74	16.54	16.88
	17.14	17.06	16.78	16.45	16.95	16.31,				

Úloha V6.06 *Ověření stanovení železa spektrofotometrickou metodou*
Prove(te validaci stanovení obsahu železa *y* v $CoSO_4$ spektrofotometrickou metodou SFM *y* porovnáním výsledků standardního stanovení obsahu *x* metodou AAS, u které je zde předpokládána zanedbatelná náhodná chyba. (1) Vedou pak obě metody ke stejným výsledkům? (2) Jsou v datech odlehlé hodnoty?

Data: Obsah železa v $CoSO_4$ [%].

SFM:	0.025	0.048	0.068	0.092	0.103	0.110	0.119	0.130	0.145	0152
0.010	0.036									
AAS:	0.023	0.049	0.071	0.090	0.099	0.111	0.120	0.132	0.140	0.149
0.011	0.036									

Úloha V6.07 *Stanovení dusičnanů v pitné a povrchové vodě*
V chemických laboratořích geochemické firmy se nedávno zavedla nová metoda stanovení obsahu dusičnanů *y* v pitných a povrchových vodách pomocí iontově párové chromatografie. (1) Je třeba provést validaci nové metody v porovnání s deklarovanými hodnotami obsahu dusičnanů *x* uměle připravených vzorků. Použijte pro odhad parametrů metodu ortogonální regrese. (2) Vede nová metoda ke správným výsledkům? (3) Proveďte simultánní test úseku, zda je roven nule a směrnice, zda je rovna jedné, str. 352 v cit[72].

Data: Obsah dusičnanů NO_3^- [mg/l]: dáno *x*, nalezeno *y*.

5.2	4.2,	2.1	2.2,	20.0	18.7,	50.0	54.1,	19.0	17.4,	22.3	19.7,
200	195,	80.0	78.2,	116.0	113.6,	164.0	160.6,				

Úloha V6.08 *Nová metoda stanovení organických látek v trhavinách*
Prove(te posouzení nové analytické metody HPLC stanovení obsahu organických látek v Perunitu *y* porovnáním výsledků s výsledky standardní extrakční metody měření *x* se zanedbatelným rozptylem. Dvojice *x*, *y* představují obsah organických látek v Perunitu v procentech. Pro odhad parametrů použijte metodu ortogonální regrese. Testujte úsek b_0 (má být $\beta_0 = 0$) a směrnici b_1 (má být $\beta_1 = 1$) individuálními *t*-testy parametrů i kombinovaným testem obou, a konečně elipsou spolehlivosti modelu $y = \beta_0 + \beta_1 x$ dle str. 352 v cit[72].

Data: Obsah organických látek [%], standardní metoda *x* a HPCL metoda *y*.

x:	32.96	32.99	32.41	32.09	32.28	32.64	33.15	31.83	32.85	32.51
	32.20	31.16	32.84	32.79	32.18	31.85	32.46			
y:	32.05	31.50	31.89	31.55	32.07	32.00	33.07	33.34	33.19	32.95
	31.83	31.26	32.89	33.30	32.00	31.46	32.00			

Úloha V6.09 *Validace stanovení dusičnanů selektivní elektrodou v pitné vodě*
Rozhodčí metodou pro stanovení obsahu dusičnanů v pitné vodě je podle ČSN 83 0520 spektrofotometrická metoda se salicylanem sodným *x*. Je testováno stanovení dusičnanů selektivní elektrodou *y*. Obě metody poskytují výsledky se shodným rozptylem. (1) Úkolem je zjistit, zda obě metody poskytují stejné výsledky. Použijte pro odhad parametrů metodu ortogonální regrese. (2) Není metoda selektivní elektrody zatížena systematickou chybou?

(3) Obsahují data vlivné body?

Data: Obsah dusičnanů [mg/l] rozhodčí metodou *x* a selektivní elektrodou *y*.

x:	68.9	54.6	54.2	50.8	43.7	39.2	33.7	19.1	24.6	29.0
	32.7	36.1	41.5	43.3	19.4	11.1	16.7	76.7,		
y:	70.6	58.7	54.7	47.1	43.5	38.7	33.7	19.4	26.2	29.7
	34.4	37.6	42.9	45.6	20.3	13.0	15.7	88.1,		

Úloha V6.10 *Validace stanovení obsahu pyridinových a chinolinových zásad*

Byly srovnávány dva způsoby stanovení obsahu pyridinových a chinolinových zásad v destilačních frakcích kyseliny karbolové, metoda chromatografické GLC analýzy *x* a titrační metoda *y*. Obě metody poskytují výsledky se shodným rozptylem. Použijte pro odhad parametrů metodu ortogonální regrese. (1) Vedou obě metody ke stejným výsledkům? Jsou v datech vlivné body? (2) Proveďte test úseku b_0 (má být $\beta_0 = 0$) a směrnice b_1 (má být $\beta_1 = 1$), dále aplikujte kombinovaný test obou parametrů a test elipsou spolehlivosti modelu $y = \beta_0 + \beta_1 x$ dle str. 352 v cit[72].

Data: Obsah zásad [%] GLC metodou *x* a titrační metodou *y*.

x:	0.40	0.50	0.17	1.49	3.50	0.39	0.25	0.52	1.92	2.60
	0.50	0.36	0.18	1.30	1.62	2.66	3.02	2.30	0.37,	
y:	0.35	0.39	0.17	1.92	3.37	0.37	0.28	0.45	1.78	2.25
	0.08	0.04	0.20	1.30	1.60	1.91	2.75	2.09	0.31,	

Úloha V6.11 *Validace stanovení kyseliny chromotropové*

Obsah kyseliny chromotropové se stanovuje titrací s 4-nitrobenzendiazoniovou solí. Konec titrace se sleduje reakcí první nadbytečné kapky diazo-p-nitroanilinu s H-kyselinou za vzniku azobarviva. Standardně prováděná vizuální kopulace byla převedena na automatický titrátor. Úkolem je provést validaci obsahu, stanoveného novou analytickou metodou *y* v porovnání se standardní metodou *x*, a to na základě několika paralelních stanovení. Obě metody poskytují výsledky se shodným rozptylem. (1) Dosahují obě metody stejných výsledků? (2) Proveďte Studentův *t*-test úseku b_0 (má být $\beta_0 = 0$) a směrnice b_1 (má být $\beta_1 = 1$), dále kombinovaný test obou parametrů modelu $y = \beta_0 + \beta_1 x$ dle str. 352 v cit[72].

Data: Obsah kyseliny chromotropové [%] vizuálně *x* a titračně *y*.

74.8	75.2,	76.4	76.7,	73.2	73.4,	74.4	74.8,	74.5	74.6,
...	...	...	...	...	...	...	...	...	...
76.1	76.2,								

Úloha V6.12 *Validace analytické metody stanovení dusíku*

Při výrobě dusíkatých hnojiv se množství dusíku ve výrobku stanovuje titrační *x* a spektrofotometrickou metodou *y*. Obě metody poskytují výsledky se shodným rozptylem. (1) Na naměřených datech určete regresní přímku závislosti dusíku stanoveného metodou spektrofotometrickou *y* a metodou titrační *x* a rozhodněte, zda obě metody lze považovat za shodné. Použijte pro odhad parametrů metodu ortogonální regrese. (2) Dosahují obě metody stejných výsledků? (3) Proveďte Studentův *t*-test úseku b_0, (má být $\beta_0 = 0$) a směrnice b_1, (má být $\beta_1 = 1$).

Data: Obsah stanovený titračně *x* a spektrofotometricky *y*.

30.73	30.52,	30.69	30.51,	30.64	30.54,	30.62	30.41,	29.21	29.35,
...	...	...	...	...	...	...	...	...	...

30.01	30.06,	30.59	30.49,	30.15	30.30,	30.24	30.27,	29.44	29.39,

Úloha V6.13 *Validace analytické metody stanovení lipázy*
Proveďte ověření nově navržené analytické metody pro stanovení obsahu lipázy y diagnostickou soupravou Sentinel porovnáním s výsledky x dosud používané soupravy Wako, a to na základě paralelních stanovení. Vyjděte z předpokladu, že obě metody poskytují výsledky se shodným rozptylem. (1) Testujte významnost vlivných bodů. (2) Je nová metoda zatížena systematickou chybou? (3) Dosahují obě metody stejných výsledků? (4) Proveďte Studentův t-test úseku b_0 (má být $\beta_0 = 0$) a směrnice b_1 (má být $\beta_1 = 1$) a dále i kombinovaný test úseku a směrnice přímky $y = \beta_0 + \beta_1 x$.

Data: Obsah lipázy [μ/l kat] Wako x, Sentinel y.

1.60	0.89,	3.92	1.72,	32.46	10.45,	2.87	1.34,	2.55	1.26,
...	...	...	...	...	...	...	...	...	...
18.80	6.26,	23.60	7.76,						

Úloha V6.14 *Validace stanovení amonných iontů v pitných vodách*
Proveďte ověření časově nenáročné metody stanovení obsahu amonných iontů y soupravou Spektroquant srovnáním s obsahem x určeným standardní metodou stanovení amoniaku podle ČSN - ISO 7150-1, která je však náročná na provedení. Pro účely vyhodnocení se předpokládá, že rozptyl standardní metody je zanedbatelný. (1) Vyšetřete statistickou významnost úseku b_0 (má být $\beta_0 = 0$). (2) Odstraňte z dat odlehlé hodnoty. (3) K jakým závěrům vede kombinovaný test úseku a směrnice?

Data: Obsah amonných iontů [mg/l] standardní metodou x a metodou Spektroquant y.

x: 0.012	0.012	0.012	0.014	0.025	0.026	0.032	0.040	0.070	0.140	
y: 0.010	0.015	0.021	0.040	0.024	0.030	0.028	0.032	0.082	0.143	
x: 0.210	0.275	0.300	0.320	0.370	0.375	0.455	0.575	0.610	0.520	0.035
y: 0.204	0.262	0.322	0.305	0.355	0.388	0.431	0.550	0.520	0.496	0.035

Úloha V6.15 *Validace metody plamenné fotometrie AAS a ICP pro stanovení mědi*
V deseti vzorcích mědnatého katalyzátoru byl standardní elektrogravimetrickou metodou stanoven obsah x oxidu měďnatého CuO. Ve stejných vzorcích byl stanoven obsah CuO také (a) plamenovou metodou AAS, proměnná y_1, a (b) metodou ICP, proměnná y_2. (1) Pro účely vyhodnocení předpokládejte zanedbatelný rozptyl elektrogravimetrické metody. (2) Je možné oběma navrženými metodami nahradit metodu standardní? (3) Aplikujte Studentův t-test úseku b_0, (má být $\beta_0 = 0$) a směrnice b_1 (má být $\beta_1 = 1$), a dále i kombinovaný test obou parametrů přímky $y = \beta_0 + \beta_1 x$. (4) Vyšetřete, zda jsou obě přímky shodné či paralelní? (5) Mají společný úsek? Postupujte dle návodu na str. 381 v cit[72].

Data: Obsah CuO [%] metodou standardní x, plamenné fotometrie y_1, a ICP y_2.

37.8	37.0	37.2,	38.2	38.3	37.9,	40.2	40.2	39.6,	39.5	38.6	39.1,
...	...	...	...	...	...	...	...	...	...	...	...
41.0	40.2	39.9,	37.5	34.3	38.4,						

Úloha V6.16 *Validace analytické metody stanovení formaldehydu*
Obsah formaldehydu x ve vzorcích fenolových vod je stanoven polarografickou metodou x. Laboratoř fenoplastů navrhla používat jednodušší metodu y, redox-titraci. Rozptyl obou

metod je v zásadě stejný. (1) Rozhodněte, zda tato metoda bude poskytovat správné a reprodukovatelné výsledky. (2) Jsou v datech odlehlé hodnoty? (3) Jsou výsledky nové metody zatíženy systematickou chybou? (4) Aplikujte Studentův *t*-test úseku b_0 (má být $\beta_0 = 0$) a směrnice b_1 (má být $\beta_1 = 1$).

Data: Obsah formaldehydu ve vodě [mg/l] polarograficky *x*, redox-titrací *y*.

x:	76.8	117	129.1	160.1	236.4	258	284.2	303.2	386.2
	474.3	532.4	937.6	2654.3					
y:	80.5	112.6	128	152.2	239.4	250	287	307.8	391.7
	480.2	530.8	934.2	2647.2					

Úloha V6.17 *Ověření chromatografické metody vůči redukční*

V chemické laboratoři se obsah látky stanovoval redukční metodou, později se přešlo na chromatografické stanovení. Vyjděte z předpokladu, že obě metody poskytují výsledky se shodným rozptylem. Použijte pro odhad parametrů metodu ortogonální regrese. Testujte, zda obě metody poskytují stejné výsledky.

Data: Výsledek stanovení obsahu látky redukcí *x* [%], chromatograficky *y* [%].

27.48	27.40,	27.68	27.30,	25.88	26.05,	27.20	26.88,	27.10	26.47,
...	...	...	...	...	...	...	...	...	...
28.40	27.88,	26.86	26.77,						

Úloha V6.18 *Stanovení obsahu kyseliny sírové v nitrační směsi dvěma metodami*

Pro stanovení obsahu kyseliny sírové v nitrační směsi byla připravena série umělých vzorků. Vzorky byly analyzovány jednak použitím titrátoru DL 40 RC *x*, jednak ruční titrací *y*. Předpokládá se, že obě metody poskytují výsledky se shodným rozptylem. Použijte pro odhad parametrů metodu ortogonální regrese. Rozhodněte, zda obě metody poskytují srovnatelné výsledky.

Data: Stanovení obsahu titrátorem *x* [%] a ruční titrací *y* [%].

75.073	75.072,	74.349	74.418,	74.527	74.497,	74.866	74.898,	75.499	75.407,
...	...	...	...	...	...	...	...	...	...
75.954	75.977,	76.131	76.103,						

Úloha V6.19 *Porovnání výsledků dvou laboratoří při stanovení MgO v hnojivech*

Dvě laboratoře stanovily obsah oxidu hořečnatého MgO v hnojivu, první *x* a druhá *y*. Mezi výsledky obou laboratoří předpokládejte jednoduchý lineární regresní vztah. Pro odhad parametrů použijte metodu ortogonální regrese: (1) Stanovte, jsou-li výsledky obou laboratoří statisticky významně odlišné. (2) Výsledky testování ověřte též párovým testem pro dva výběry.

Data: Obsah oxidu hořečnatého [%] v první laboratoři *x* a v druhé laboratoři *y* .

x:	10.36	8.22	7.36	9.52	7.8	8.47	6.86	7.56	16.28	19.00
y:	10.18	8.33	8.27	9.18	7.75	9.05	7.65	7.67	18.85	18.79

Úloha V6.20 *Validace nové metody stanovení arsenu v odpadní vodě*

Byla navržena jednodušší metoda stanovení arsenu ve vodě a bylo třeba validovat na standardech, zda přináší správné výsledky[70]. K vyšetření vztahu mezi naměřenou koncentrací arsenu *y* a známou koncentrací arsenu ve standardu *x*, udávanou v µg/ml, je předpokládán jednoduchý lineární regresní model $y = \beta_0 + \beta_1 x$. (1) Ověřte, zda nová

metoda přináší správné výsledky. Použijte pro odhad parametrů metodu ortogonální regrese. (2) K jakému výsledku dospěje nová metoda, když standard žádný arsen neobsahuje, čili absolutní člen je nulový, $\beta_0 = 0$? (3) Vyšetřete zda při stanovení arsenu nová metoda nadhodnocuje či podhodnocuje?

Data: Koncentrace standardu arsenu x [μg. cm^{-3}], nalezená koncentrace arsenu y [μg. cm^{-3}].

0	0.17,	0	0.25,	0	0.01,	0	0.12,	1	1.25,	1	0.86,	1	1.25,
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7	6.67,	7	7.02,	7	7.14,	7	7.30,						

Úloha V6.21 *Validace HPLC metody při stanovení kyantetralinu*

Porovnejte novou metodu stanovení obsahu kyantetralinu y v surovém vzorku kapalinovou chromatografií HPLC s jeho stanovením plynovou chromatografií GC x. (1) Vyšetřením regresního tripletu testujte a diskutujte statistickou významnost jednotlivých parametrů v modelu stejně jako i jejich fyzikální smysl, zdůvodněte všechny užité diagnostiky. (2) Jsou výsledky nové metody zatíženy systematickou chybou?

Data: Koncentrace kyantetralinu [%] metodou GC x a metodou HPLC y.

x:	65.5	67.3	62.2	*y:*	63.8	68.1	63.1
	...	...	...		...	...	...
	67.7	66.8					

Úloha V6.22 *Validace titrační metody při stanovení modré báze MB H-3R*

Při pokusné výrobě modré báze MB H-3R byl stanovován její obsah v pastě z kalolisu titračně dusitanem v kyselém prostředí y a spektrofotometricky x. Obě metody v sobě zahrnují chybu danou obsahem vedlejších látek. Při titračním stanovení se však vedle modré báze stanoví i diaminomesitylensulfonová kyselina přítomná v malém množství v pastě, a dále se ztitrují i vedlejší modře bez aminoskupiny. Při spektrofotometrickém stanovení mohou absorbanci zvýšit i vedlejší modře bez aminoskupiny, které barvivo netvoří. Za základ byla vzata titrační metoda, která je na provoze považována za směrodatnou pro výpočet výtěžku a ekonomii výroby. Rozptyl této metody se považuje za zanedbatelný vůči rozptylu spektrofotometrické metody. Při validaci s ní byla porovnávána spektrofotometrická metoda.

Data: Koncentrace modré báze spektrofotometrickou metodou x a titrační metodou y.

52.0	50.3	45.3	42.5	54.3	50.8
...	...	...	...	...	...
64.1	60.1	58.1	55.2	69.2	60.1

Úloha V6.23 *Validace metody Cheminova při stanovení alaninaminotransferázy*

Proveďte ověření nově navržené analytické metody pro stanovení obsahu y alanin-aminotransferázy ALT diagnostickou soupravou Cheminova y porovnáním s výsledky obsahu x dosud používané soupravy Lachema. Použijte předpokladu stejných rozptylů obou metod. Aplikujte t-test úseku b_0 (má být $\beta_0 = 0$) a směrnice b_1 (má být $\beta_1 = 1$), a dále i kombinovaný test parametrů β_0 a β_1 v regresním modelu $y = \beta_0 + \beta_1 x$.

Data: Aktivita ALT v krevním séru [μkat/l], stanovená metodou Lachema x a Cheminova y.

Lachema *x*:		0.46	0.38	0.19	0.11	1.28	1.37	1.62	1.95
2.10	2.97	0.32	0.77	3.28	1.65	4.15	0.95	3.99	2.41
2.99									

Cheminova *y*:		0.43	0.35	0.20	0.09	1.26	1.33	1.55	2.00
2.13	2.93	0.29	0.74	3.14	1.62	4.00	1.15	3.52	2.35
3.18									

Úloha V6.24 *Validace metody WAKO při stanovení mikroalbuminurie MAU v moči*
Proveďte ověření nově navržené analytické metody pro stanovení obsahu mikroalbu-minurie MAU *y* v moči diagnostickou soupravou WAKO porovnáním vůči výsledkům dosud používané soupravy Boehringer *x*. Použijte předpokladu stejných rozptylů obou metod. (1) Testujte statistickou významnost odhadovaných parametrů $\beta_0 = 0$ a $\beta_1 = 1$. (2) Užijte také kombinovaný test obou parametrů.

Data: Koncentrace albuminu v moči [mg/l], stanovená metodou Boehringa *x* a WAKO *y*.

Boehringer *x*:	3.9	11.2	5.2	7.6	15.4	21.0	30.2	47.7
	55.9	57.4	78.0	124.7	183.4	256.6	279.4	300.2
Wako *y*:	4.2	10.8	5.4	7.9	15.0	20.4	29.7	48.4
	55.2	58	77.2	125.6	182.3	251.5	278.3	299.4

Úloha V6.25 *Validace metody ITP pro stanovení dusičnanů*
Proveďte validaci metody ITP *y* stanovení obsahu dusičnanů ve vodách porovnáním výsledků obsahu s výsledky standardní fotometrické metody se sulfosalicylanem v semimikroprovedení *x*. Použijte předpokladu stejných rozptylů obou metod. (1) Dosahují obě metody stejných výsledků? (2) Proveďte Studentův *t*-test úseku b_0 (má být $\beta_0 = 0$) a směrnice b_1 (má být $\beta_1 = 1$).

Data: Koncentrace dusičnanů [mg/l], stanovené fotometricky *x* a metodou ITP *y*.

x:	15.8	6.3	10.2	28.4	16.9	25.7	25.4	76.2
y:	16.8	6.9	10.4	29.3	16.4	27.9	23.4	73.6

6.3 Lineární a nelineární kalibrace

Pro nelineární kalibrační úlohy se s výhodou používá úsekové (spline) regrese (viz kap. 9). Jsou přitom vypočteny také odhady kritické úrovně y_c, limita detekce y_d a limita stanovení y_s, zvaná též limita kvantifikace y_q. Pro zadaný soubor *M* opakování signálu $\{y_j^{(}\}$, $j = 1, ..., M$, je vyčíslena průměrná hodnota $\bar{y}^{(}$, pro kterou jsou pak stanoveny bodové odhady $\hat{x}$ a odpovídající 95 % konfidenční intervaly.

Postup výpočtu:

1. Zadání experimentálních dat.
2. Opakované zadávání počtu *M* a hodnot $y_j^{(}$, $j = 1, ..., M$.
3. Po zadání experimentálních dat:
   a) Jsou vypočteny průměry, rozptyly, součty čtverců a kovariance;
   b) Je určen odhad směrnice a úseku spolu s odpovídajícími rozptyly. Je vyčíslen i Pearsonův korelační koeficient a tabulka výsledků, obsahující predikce $\hat{y}_i$ spolu s absolutními a relativními odchylkami;
   c) Je vypočtena kritická úroveň (x_c , y_c), limita detekce (x_d , y_d) a limita kvantifikace čili limita stanovení (x_q , y_q).

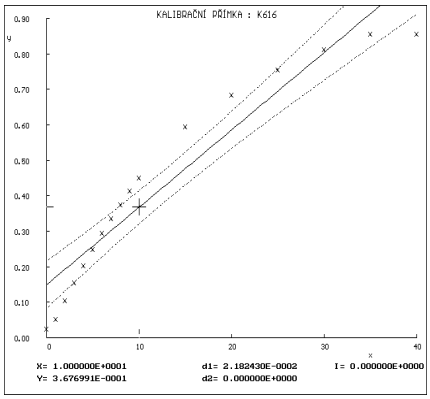
d) Je možné kreslit i graf kalibrační přímky spolu s experimentálními body.
e) Pro zadané hodnoty signálu $\{y_j^{(}\}$, j = 1, ..., M, jsou určeny jak bodové, tak i intervalové odhady pro x^*. Zadávání hodnot signálů lze opakovat.

Vzorová úloha 6.3 *Postup nelineární kalibrace spline funkcí*

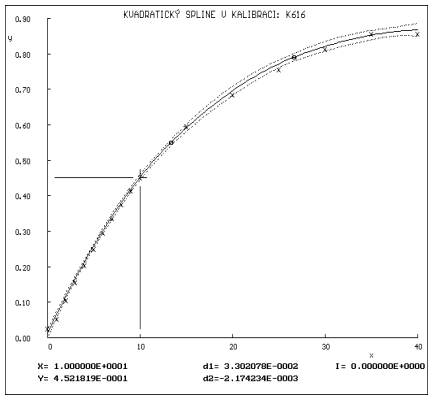
Na úloze **K6.16** *Kalibrační model chromu v mineralizátech kalů a míra detekce* ukážeme postup kalibrace: sestrojte kalibrační model pro stanovení obsahu chromu v mineralizátech kalů metodou AAS. Pro kalibraci bylo použito standardů, které mají obsah 10 % Lefortovy lučavky, shodný s mineralizáty kalů. Čára chromu byla 357.9 nm, šířka spektrálního intervalu 0.2 nm a plamen acetylen-vzduchový. (1) Vyšetřete linearitu kalibračního modelu, určete jeho parametry a míry přesnosti kalibrace. (2) Jaká je koncentrace chromu u neznámých vzorků, když naměřené absorbance vzorků byly y^* = 0.272, 0.464 a 0.535? (3) Jsou v kalibračních datech odlehlé hodnoty? (4) Je nejnižší koncentrace neznámého vzorku ještě nad limitou detekce?

Řešení:

1. Návrh modelu: podle typu kalibrace nebo charakteru dat se hledá návrh regresního modelu tak, že se daty prokládá *lineární model*, tj. přímka, nebo *nelineární model*, tj. křivka. Jako "univerzální" křivky jsou voleny lineární spline, kvadratický spline, kubický spline, tj. regrese s různými alternativami zadávání uzlových bodů. Kvalitou proloženého kalibračního grafu je jak grafická, tak i statistická analýza reziduí.



Obr. 6.3-1 Lineární (přímkový) kalibrační graf.



4Obr. 6.3-2 Nelineární kalibrační graf, ***ADSTAT.***

2. Statistická analýza reziduí: těsnost proložení, čili velikost klasických reziduí, střídání znaménka a statistické charakteristiky jsou mírou vhodnosti navrženého kalibračního grafu.

Bod	**Měřená hodnota**	**Predikovaná hodnota**	**Absolutní reziduum**	**Relativní reziduum**
i	$y_{exp,\,i}$	$y_{vyp,\,i}$	e_i	$e_{r,\,i}$
1	3.2000E-02	1.3262E-02	-1.8738E-02	-1.4128E+02
2	5.8900E-02	6.6938E-02	8.0383E-03	1.2009E+01
3	1.1120E-01	1.1844E-01	7.2401E-03	6.1129E+00
4	1.6190E-01	1.6777E-01	5.8677E-03	3.4975E+00
5	2.1040E-01	2.1492E-01	4.5210E-03	2.1035E+00
6	2.5670E-01	2.5990E-01	3.2000E-03	1.2313E+00
7	3.0120E-01	3.0270E-01	1.5049E-03	4.9714E-01

8	3.4340E-01	3.4334E-01	-6.4528E-05	-1.8795E-02
9	3.8250E-01	3.8179E-01	-7.0815E-04	-1.8548E-01
10	4.2150E-01	4.1807E-01	-3.4260E-03	-8.1948E-01
11	4.5740E-01	4.5218E-01	-5.2181E-03	-1.1540E+00
12	6.0300E-01	5.9152E-01	-1.1480E-02	-1.9408E+00
13	6.9180E-01	6.9627E-01	4.4725E-03	6.4235E-01
14	7.6350E-01	7.7209E-01	8.5880E-03	1.1123E+00
15	8.1930E-01	8.2159E-01	2.2950E-03	2.7933E-01
16	8.6310E-01	8.5334E-01	-9.7633E-03	-1.1441E+00
17	8.6430E-01	8.6797E-01	3.6703E-03	4.2287E-01
Reziduální součet čtverců, *RSC*			: 9.1459E-04	
Průměr absolutních hodnot reziduí, M_e			: 5.8115E-03	
Průměr relativních reziduí, M_{er} [%]			: 10.262	
Odhad reziduálního rozptylu, $s^2(e)$			: 7.6215E-05	
Odhad směrodatné odchylky reziduí, *s(e)*			: 8.7301E-03	

3. Analýza derivací a integrálů kalibračního grafu:

Bod	**Predikovaná hodnota**	**První derivace**	**Druhá derivace**	**Integrál**
i	$y_{vyp,i}$	$der1_i$	$der2_i$	int_i
1	1.3262E-02	5.4763E-02	-2.1742E-03	0.0000E+00
2	6.6938E-02	5.2589E-02	-2.1742E-03	4.0282E-02
3	1.1844E-01	5.0415E-02	-2.1742E-03	1.3315E-01
4	1.6777E-01	4.8240E-02	-2.1742E-03	2.7644E-01
5	2.1492E-01	4.6066E-02	-2.1742E-03	4.6796E-01
6	2.5990E-01	4.3892E-02	-2.1742E-03	7.0555E-01
7	3.0270E-01	4.1718E-02	-2.1742E-03	9.8704E-01
8	3.4334E-01	3.9543E-02	-2.1742E-03	1.3102E+00
9	3.8179E-01	3.7369E-02	-2.1742E-03	1.6730E+00
10	4.1807E-01	3.5195E-02	-2.1742E-03	2.0731E+00
11	4.5218E-01	3.3021E-02	-2.1742E-03	2.5084E+00
12	5.9152E-01	2.3844E-02	-1.1575E-03	5.1376E+00
13	6.9627E-01	1.8057E-02	-1.1575E-03	8.3691E+00
14	7.7209E-01	1.2269E-02	-1.1575E-03	1.2052E+01
15	8.2159E-01	8.0592E-03	-6.8432E-04	1.6045E+01
16	8.5334E-01	4.6375E-03	-6.8432E-04	2.0239E+01
17	8.6797E-01	1.2159E-03	-6.8432E-04	2.4550E+01

4. Kalibrační meze přesnosti:

Kritická úroveň,	y_c:	0.02586	x_c:	0.2293
Limita detekce,	y_d:	0.0370	x_d:	0.4371

5. Kalibrační tabulka: obsahuje výsledky neznámých koncentrací či obsahů. Vedle bodového odhadu je tištěn i intervalový odhad neznámé koncentrace či neznámého obsahu (ADSTAT).

Měřený signál	**Inverzní odhad (bodový)**	**Interval spolehlivosti koncentrace (obsahu)**	
		dolní mez	**horní mez**
$y_{exp,i}$	$x_{vyp,i}$	$LDx_{vyp,i}$	$LHx_{vyp,i}$
0.2720	5.278	5.128	5.431
0.4640	10.362	10.116	10.613

0.5350	12.758	12.454	13.083

Lze uzavřít, že každý výsledek kvantitativního stanovení z kalibrační křivky se v analytické chemii uvádí intervalovým odhadem, jímž se vystihuje interval, ve kterém se s 95% statistickou jistotou nachází stanovovaný obsah. Ten bývá dále doplněn limitou detekce, výjimečně také limitou kvantifikace. Kritická úroveň (*slepý pokus*) se uvádí zřídka. Limita detekce dokresluje úroveň práce analytické laboratoře a třídu přesnosti přístroje, takže vyjadřuje všechny nejistoty analytických operací, chemikálií, laboratorního skla atd.

6.3.1 Úlohy na lineární a nelineární kalibraci

Úloha K6.01 *Kalibrace nefelometru*
Nefelometr je kalibrován na obsah pevné fáze dispergované v destilované vodě, str. 95 v cit[67]. Pro standardní suspenzi jsou změřena kalibrační data. Zjistěte míry přesnosti kalibrace a obsah neznámých vzorků, jež vykazovaly na stupnici hodnoty $y^* = 39$, 46, 66 a 80 dílků. Jsou v kalibračních datech nějaké odlehlé hodnoty? Jsou splněny předpoklady metody nejmenších čtverců? Jde o lineární nebo nelineární kalibraci? Je rozdíl mezi hodnotou limity detekce lineární a nelineární kalibrace?

Data: Koncentrace pevné fáze x [ppm], velikost signálu y [dílky].

0.15	23,	0.30	38,	0.40	45,	0.50	61,	0.60	76,	0.70	82,
0.23	31.6,	0.086	14.1,	0.25	34.2,	0.45	52,	0.55	69.2,	0.044	5.94,

Úloha K6.02 *Kalibrační model odezvy GC detektoru na koncentraci*
Metodou plynové chromatografie se na různých kalibračních roztocích n-nonanu v dec-1-enu posuzuje závislost odezvy plameno-ionizačního detektoru na koncentraci v rozmezí 0.05 až 0.4 hmotnostních procent. (1) Určete kalibrační model pro hladinu významnosti $\alpha = 0.05$. (2) Určete obsah neznámého roztoku pro plochu píku $y^* = 2100$, 3100 a 9500. (3) Vyčíslete také míry přesnosti kalibrace?

Data:

Obsah x [hmot%]:	0.050	0.075	0.100	0.150	0.200	0.300	0.400
Plocha píku y:	2065	3117	4173	6132	8232	12405	16592

Úloha K6.03 *Kalibrační závislost absorbance na koncentraci zinku*
Při stanovení zinku metodou atomové absorbční spektroskopie byla proměřována kalibrační závislost absorbance y na koncentraci x. Úkolem je (1) nalézt lineární či nelineární kalibrační model, který nejlépe popisuje průběh naměřené závislosti a stanovit koncentraci zinku pro absorbanci $y^* = 0.005$, 0.155, 0.355 a 0.555. (2) Jaké jsou míry přesnosti kalibrace? (3) Je vzorek o nejnižší absorbanci ještě nad limitou stanovení?

Data: Koncentrace Zn x [mg.l^{-1}], absorbance y.

c:	0.02	0.05	0.1	0.25	0.5	1.0	1.5	2.0	2.5
A:	0.012	0.032	0.057	0.149	0.270	0.460	0.568	0.670	0.730

Úloha K6.04 *Kalibrační model mědi v oceli rentgenově-fluorescenční metodou*
Rentgenově-fluorescenční metodou byl stanovován obsah mědi v rozsahu 0.10 až 0.47% ve standardních vzorcích oceli, str. 158 v cit[58]. Neznámé vzorky vykazovaly $y^* = 555$, 1005

a 1505 impulzů signálu. (1) Ověřte homoskedasticitu, nalezněte parametry kalibrační funkce, rozptyl měřeného signálu a vyšetřete vlivné body. (2) Sestrojte konfidenční interval kalibrační křivky a stanovte míry přesnosti kalibrace.

Data: Obsah mědi *x* [%], velikost signálu *y* [impulzy].

x:	0.10	*y*:	694	685	685	697	684
	....		...	...	...	...	...
	0.47		2175	2172	2186	2175	2171,

Úloha K6.05 *Kalibrační model obsahu fenolu v odpadních vodách*

Spektrofotometrické stanovení fenolů v průmyslových a odpadních vodách se provádí fotometricky, a to vícenásobným standardním přídavkem p-nitroanilinu k neznámému vzorku, str. 165 v cit[58]. Každé měření se koriguje na pozadí. Neznámé vzorky vykazovaly absorbance y^* = 0.105, 0.205, 0.505 a 0.905. (1) Sestrojte kalibrační model a určete charakteristiky přesnosti kalibrace. (2) Vyšetřete vlivné body a eventuální vybočující hodnoty z dat odstraňte.

Data: Obsah přidaného p-nitroanilinu *x* [mg], absorbance *y*.

0.00	0.140,	0.10	0.310,	0.15	0.385,	0.20	0.460,	0.30	0.640,	0.40	0.820,
0.50	0.980,										

Úloha K6.06 *Nelineární kalibrace u stanovení mědi v oceli*

Při stanovení obsahu mědi v oceli *x* emisní spektrální analýzou u 11 vzorků, bylo každé měření signálu *y* bylo 5krát opakováno, viz str. 173 v cit[58]. Neznámé vzorky vykazovaly y^* = 50, 105, 157 a 205 impulzy signálu. (1) Nalezněte a sestrojte kalibrační model a určete míry přesnosti kalibrace. (2) Jsou v datech vlivné body? (3) Je třeba odstranit nějaké vybočující hodnoty?

Data: Koncentrace mědi *x* [%], velikost signálu *y* [impulzy].

c:	0.019	*y:*	26	28	27	27	30,
	...		...	...	...	...	...
	0.7		366	380	400	375	387,

Úloha K6.07 *Kalibrační model bromidů a jodidů iontovou chromatografií*

Proměřením řady obsahů standardů bromidů a jodidů byl metodou iontové chromatografie určen kalibrační graf závislosti plochy pod píkem *y* na obsahu halogenidu *x*, str. 52 v cit[61]. (1) Určete limity detekce a kvantifikace bromidů a jodidů ve vodách a vyšetřete linearitu obou kalibračních modelů. (2) Jde skutečně o lineární model? (3) Neznámé vzorky bromidů vykazovaly plochu pod píkem y^* = 505, 1005 a 1505 jednotek a u jodidů y^* = 605, 1105 a 2505 jednotek. (4) Existují v datech odlehlé body, které je třeba odstranit? (5) Jsou splněny předpoklady metody nejmenších čtverců MNČ?

Data: Koncentrace bromidů *x* [mg. l^{-1}], plocha pod píkem *y* [jednotky].

5.0	157,	10.0	279,	15.0	428,	20.0	533,	25.0	723,	30.0	778,
40.0	1008,	50.0	1251,	60.0	1445,	70.0	1653,	90.0	2009,		

Koncentrace jodidů *x* [mg. l^{-1}], plocha pod píkem *y* [jednotky].

10.0	406,	20.0	777,	30.0	1180,	40.0	1567,	50.0	1940,	60.0	2360,
70.0	2727,	80.0	3005,	90.0	3644,	100.0	3803,				

Úloha K6.08 *Kalibrace obsahu arsenu v oceli a míry přesnosti kalibrace*

Lineární kalibrační model obsahu arsenu x ve standardních vzorcích oceli byl sestrojen na základě opakovaných měření rentgenovou fluorescenční metodou signálu y, str. 173 v cit[58]. Neznámé vzorky vykazovaly odezvu $y^* = 505$, 1005 a 1205 jednotek signálu. (1) Určete míry přesnosti kalibrace a koncentraci zadaných vzorků. (2) Je třeba odstranit vybočující hodnoty?

Data: Obsah arsenu x [%], velikost signálu y [jednotky].

x:	0.003	*y:*	916	916	916	902	911,
	...		...	...	...	...	...
	0.015		997	1020	984	999	1004,

Úloha K6.09 *Kalibrace wolframu v oceli a míry přesnosti kalibrace*

Lineární kalibrační model obsahu wolframu x ve standardních vzorcích oceli byl sestrojen na základě opakovaných měření velikosti signálu y rentgenovou fluorescenční metodou, str. 176 v cit[58]. Neznámé vzorky vykazovaly odezvu $y^* = 305$, 505 a 705 jednotek signálu. (1) Určete míry přesnosti kalibrace a koncentraci wolframu neznámých vzorků. (2) Je třeba odstranit vybočující hodnoty?

Data: Koncentrace wolframu x [%], opakovaná měření velikosti signálu y [jednotky].

x:	0.003,	*y:*	133	129	132	134	128
	...		...	...	...	...	...
	0.015		192	181	180	176	181

Úloha K6.10 *Míry přesnosti kalibrace u fotometrického stanovení fosforu*

Gottschalk[68] publikoval fotometrické stanovení fosforu $y = f(x)$ v komplexní formě v koncentračním rozmezí x od 10 do 100 μmolů fosforu (3.1 až 31 $\mu g.ml^{-1}$). (1) Určete míry přesnosti kalibrace, parametry lineárního kalibračního modelu a vlivné body. (2) Jaká je koncentrace fosforu ve vzorcích, jež vykazovaly absorbance $y^* = 0.250$, 0.450, 0.750 a 0.950. (3) Je třeba odstranit vybočující hodnoty?

Data: Koncentrace fosforu x [μmol], absorbance y.

x:	10	*y:*	0.1141	0.1141	0.1151	0.1151,
	...		...	...	...	...
	100		1.1468	1.1498	1.1478	1.1468,

Úloha K6.11 *Kalibrační model stanovení koncentrace glukózy fotometricky*

Sestrojte kalibrační model koncentrace glukózy z fotometrických měření standardních roztoků glukózy $y = \beta_0 + \beta_1 x$, str. 116 v cit[69]. (1) Určete parametry kalibračního modelu a míry přesnosti kalibrace. (2) Jaká je koncentrace glukózy u vzorků, jež vykazovaly absorbance $y^* = 0.205$, 0.450 a 0.650? (3) Jsou v datech nějaké odlehlé hodnoty?

Data: Koncentrace glukózy x [$mmol.dm^{-3}$], absorbance y.

0	0.002,	2	0.150,	4	0.294,	6	0.434,	8	0.570,	10	0.704,

Úloha K6.12 *Kalibrační model obsahu stříbra metodou AAS a limita detekce*

Sestrojte kalibrační model obsahu stříbrných iontů x v roztoku proměřením standardních roztoků stříbra metodou AAS $y = \beta_0 + \beta_1 x$, str. 117 v cit[69]. (1) Určete parametry kalibračního modelu a míry přesnosti kalibrace. (2) Jaká je koncentrace stříbra ve vzorcích, které vykazovaly absorbanci $y^* = 0.456$ a jeden vzorek opakovanou absorbanci $y^* = 0.308$, 0.314, 0.347 a 0.312? (3) Jsou v datech nějaké odlehlé hodnoty?

Data: Koncentrace stříbra x [ng/ml], absorbance y.

0	0.003,	5	0.127,	10	0.251,	15	0.390,	20	0.498,	25	0.625,	30	0.763,

Úloha K6.13 *Kalibrace koncentrace zlata v mořské vodě metodou AAS a míry přesnosti*
Sestrojte kalibrační model koncentrace zlata v mořské vodě x metodou AAS proměřením absorbance y u standardních vzorků zlata technikou vícenásobných standardních přídavků $y = \beta_0 + \beta_1 x$, str. 117 v cit[69]. (1) Určete parametry kalibračního modelu a míry přesnosti kalibrace. (2) Jaká je koncentrace glukózy v neznámých vzorcích, jež vykazovaly absorbanci $y^* = 0.280$, 0.385, 0.490 a 0.565? (3) Je nejnižší koncentrace neznámého vzorku ještě nad limitou detekce?
Data: Koncentrace zlata x [ng. ml^{-1}], absorbance y.

0	0.257,	10	0.314,	20	0.364,	30	0.413,	40	0.468,	50	0.528,	60	0.574,	70	0.635,

Úloha K6.14 *Kalibrační graf chininu fluorescenční metodou a limita detekce*
Sestrojte kalibrační model pro stanovení koncentrace chininu fluorescenční metodou signálu y dle modelu $y = \beta_0 + \beta_1 x$, str. 117 v cit[69]. (1) Určete parametry kalibračního modelu, jednak metodou vážených nejmenších čtverců a jednak metodou prostých nevážených nejmenších čtverců. (2) Vypočítejte míry přesnosti kalibrace. (3) Jaká je koncentrace chininu ve čtyřech vzorcích osvěžujícího nápoje Tonik, když byly naměřeny následující hodnoty fluorescence y^* = 15, 38, 72, 99 jednotek? (4) Jsou v datech nějaké vlivné body? (5) Kterému modelu dáváte přednost, lineárnímu či nelineárnímu? (6) Projeví se volba lineárního či nelineárního modelu na hodnotě limity detekce?
Data: Koncentrace chininu x [ng. ml^{-1}], signál fluorescence y [jednotek].

x:	0	*y:*	4	3	4	5	4,
	...		...	...	...	...	...
	50		104	109	107	101	105,

Úloha K6.15 *Kalibrační model obsahu olova metodu AAS a míry přesnosti kalibrace*
Sestrojte kalibrační model koncentrace olova, získaný proměřením standardních vzorků olova metodou elektrotermické atomové absorpce s grafitovou kyvetou $y = \beta_0 + \beta_1 x$, str. 117 v cit[69]. (1) Vyšetřete linearitu kalibračního grafu, určete parametry kalibračního grafu β_0, β_1 a kalibrační limity. (2) Jaká je koncentrace neznámých vzorků olova, když naměřené absorbance vzorků byly y^* = 0.282, 0.444, 0.725 a 0.995? (3) Jsou v datech vybočující hodnoty?
Data: Koncentrace olova x [ng.ml^{-1}], absorbance y.

10	0.050,	25	0.170,	50	0.320,	70	0.437,	100	0.600,	136	0.927,	200	1.070,
250	1.250,	280	1.344,	300	1.400,								

Úloha K6.16 *Kalibrační model obsahu chromu v mineralizátech kalů a míra detekce*
Sestrojte kalibrační model $y = f(x)$ pro stanovení koncentrace chromu x v mineralizátech kalů metodou AAS. Pro kalibraci bylo použito standardů, které jsou obsaženy v mineralizátech kalů. Čára chromu byla 357.9 nm, šířka spektrálního intervalu 0.2 nm a plamen acetylen-vzduchový. (1) Vyšetřete linearitu kalibračního modelu, určete jeho parametry a míry přesnosti kalibrace. (2) Jaká je koncentrace chromu u neznámých vzorků, když naměřené absorbance vzorků byla y^* = 0.272, 0.464, 0.535 a 0.785? (3) Jsou v kalibračních datech odlehlé hodnoty? (4) Je nejnižší koncentrace neznámého vzorku ještě

nad limitou detekce?

Data: Obsah chromu *x* [ppm], absorbance *y*.

0.0	0.3200,	1.0	0.5890,	2.0	0.1112,	3.0	0.1619,	4.0	0.2104,	5.0	0.2567,
...	...	...	...	...	...	...	...	...	...	...	...
20.0	0.6918,	25.0	0.7635,	30.0	0.8193,	35.0	0.8631,	40.0	0.8643,		

Úloha K6.17 *Kalibrační model koncentrace zinku v mléce metodou plamenové AAS*
Sestrojte nelineární kalibrační model $y = f(x)$ koncentrace zinku x v mléce metodou plamenné fotometrie y, když byla použita spektrální čára 213.9 nm, šířka spektrálního intervalu 0.2 nm a plamen acetylen-vzduchový. (1) Vyšetřete parametry kalibračního modelu, míry přesnosti kalibrace a stanovte koncentraci zinku u neznámých vzorků, jež vykazovaly absorbance $y^* = 0.105$, 0.205, 0.315 a 0.445. (2) Jsou v kalibračních datech odlehlé hodnoty? (3) Je nejnižší koncentrace neznámého vzorku ještě nad limitou detekce a limitou kvantifikace?

Data: Koncentrace zinku *x* [ppm], absorbance *y*.

0.050	0.031,	0.100	0.062,	0.150	0.095,	0.200	0.128,	0.250	0.161,	0.300	0.195,
...	...	...	...	...	...	...	...	...	...	...	...
1.500	0.844,	1.600	0.867,								

Úloha K6.18 *Kalibrace obsahu draslíku v kamenci plamenovou fotometrií a limita detekce.*
Sestrojte kalibrační model $y = f(x)$ a stanovte stopovou koncentraci draslíku x v kamenci rubidnocesném. (1) Určete míry přesnosti kalibrace, parametry kalibračního modelu a koncentraci draslíku u neznámých vzorků, jež vykazovaly výchylku $y^* = 2050$, 2160, 2310 a 2450 jednotek. (2) Jsou v kalibračních datech odlehlé hodnoty? (3) Je nejnižší koncentrace neznámého obsahu ještě nad limitou detekce?

Data: Koncentrace draslíku *c* [mg/ml], signál *y* [jednotky]}.

0.0010	1937,	0.0015	1961,	0.0016	1969,	0.0020	1997,	0.0025	2044,	0.0030	2100,
...	...	...	...	...	...	...	...	...	...	...	...
0.0060	2412,	0.0064	2446,	0.0065	2458,	0.0070	2499,	0.0075	2538,	0.0080	2574,

Úloha K6.19 *Kalibrace obsahu sodíku v jehličí plamenovou fotometrií a limita detekce*
Sestrojte kalibrační model $y = f(x)$ a stanovte stopový obsah sodíku v jehličí x metodou plamenové fotometrie y. (1) Určete limitu detekce, parametry kalibračního modelu a koncentraci sodíku u neznámých vzorků, jež vykazovaly výchylky $y^* = 2250$, 2360, 2410 a 2650 jednotek. (2) Vyšetřete regresní triplet a vlivné body. (3) Jsou v kalibračních datech nějaké odlehlé hodnoty? (4) Je nejnižší koncentrace neznámého vzorku "ještě měřitelná"?

Data: Koncentrace sodíku *x* [mg/ml], signál *y* [jednotky].

0.000	112,	0.001	166,	0.0015	193,	0.0025	220,	0.005	380,	0.010	655,
...	...	...	...	...	...	...	...	...	...	...	...
0.035	1890,	0.040	2060,	0.043	2190,	0.048	2350,	0.050	2420,		

Úloha K6.20 *Lineární kalibrační model DHTTSK v primulinsulfokyselině*
Stanovení obsahu dehydrothiotoluidinsulfokyseliny (DHTTSK) x v primulinsulfo-kyselině se provádí fluorimetricky y. (1) Vyšetřete kalibrační přímku a určete koncentraci tří obsahů DHTTSK, jež vykazovaly hodnoty $y^* = 10.1$, 25.5 a 60.3 jednotek. (2) Jsou v kalibračních

datech odlehlé hodnoty? (3) Je nejnižší koncentrace neznámého vzorku ještě nad limitou detekce?

Data:

Koncentrace DHTTSK x [mg/l]:	699.9	1399.8	2099.7	2799.6	3499.5	4199.4
Hodnoty signálu y [jednotky]:	14.1	28.2	40.5	54.5	66.0	81.5

Úloha K6.21 *Kalibrační model stanovení fosforu fotometricky a míry přesnosti*

Stanovte kalibrační model $y = f(x)$ a míry přesnosti kalibrace spektrofotometrického stanovení koncentrace celkového fosforu. (1) Je třeba provést stanovení intervalového odhadu neznámých koncentrací celkového fosforu o naměřené absorbanci $y^* = 0.05$ a 0.305. Jsou v kalibračních datech odlehlé hodnoty? (2) Je nejnižší koncentrace určovaného vzorku ještě měřitelná?

Data: Koncentrace x [mg/l], absorbance y.

x:	0.05	0.078	0.10	0.15	0.20	0.25	0.40	0.50	0.60	0.80	1.00
y:	0.019	0.023	0.033	0.044	0.069	0.084	0.125	0.158	0.183	0.247	0.332

Úloha K6.22 *Odhady koncentrace z kalibrační přímky "koncentrace-hustota"*

Byla změřena hustota vodných roztoků dichromanu sodného $Na_2Cr_2O_7$ y s přesnou koncentrací soli x. (1) Určete kalibrační přímku a z ní pak odhadněte koncentraci roztoku dichromanu $Na_2Cr_2O_7$ pro naměřené hustoty tří neznámých vzorků $y^* = 1.6701$ g/cm^3, 1.6812 g/cm^3, 1.6860 g/cm^3. (2) Jsou v kalibračních datech odlehlé hodnoty? (3) Je nejnižší koncentrace neznámého vzorku ještě nad limitou detekce?

Data: Koncentrace roztoku x [hm.%], hustota y [g/cm^3].

x:	58.0	59.0	60.0	60.3	60.8	61.2	62.0	63.0
y:	1.6424	1.6617	1.6792	1.6857	1.6956	1.7042	1.7145	1.7345

Úloha K6.23 *Kalibrace turbidimetrického stanovení formazinu*

Proměřením zákalu rozličné koncentrace x standardních roztoků formazinu byla získána data pro sestrojení nelineární kalibrační křivky y. (1) Nalezněte kalibrační model $y = f(x)$, vyšetřete vlivné body, posuďte míru spolehlivosti navrženého modelu, míry přesnosti kalibrace a intervalový odhad koncentrace formazinu pro změřený zákal o hodnotách $y^* =$ 55.1 a 55.3. (2) Jsou v datech nějaké odlehlé body?

Data:

x [mg/l]:	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.85
Turbidance y:	0.15	1.20	4.05	9.60	18.8	32.0	51.5	76.8	92.1

Úloha K6.24 *Kalibrace obsahu dusitanů a míry přesnosti kalibrace*

Proměřením absorbance standardních roztoků dusitanu sodného byla získána data pro sestrojení kalibračního grafu závislosti absorbance y na koncentraci x. (1) Vyšetřete vlivné body a vylučte vybočující hodnoty. (2) Posuďte míru spolehlivosti navrženého modelu, míry přesnosti kalibrace. (3) Na hladině významnosti $\alpha = 0.05$ vyčíslete intervalový odhad neznámé koncentrace pro absorbanci $y^* = 0.40$ a 0.45.

Data: Koncentrace dusitanů x [mol/l], absorbance y.

x:	0.000	0.020	0.020	0.100	0.100	0.500	0.500	0.200	0.300	0.400
y:	0.022	0.043	0.043	0.152	0.158	0.612	0.620	0.263	0.381	0.499

Úloha K6.25 *Stanovení chromu metodou AAS a míry přesnosti kalibrace*
Při stanovení chromu metodou atomové absorbční spektrometrie byla proměřována kalibrační závislost absorbance y na koncentraci chromu x v roztoku. (1) Úkolem je navrhnout kalibrační model, včetně testování významnosti parametrů modelu s vyšetřením vlivných bodů, míry přesnosti kalibrace. (2) Určete intervalové odhady koncentrace pro naměřené hodnoty absorbance $y^* = 0.018, 0.026, 0.057, 0.093, 0.124, 0.183$ na hladině významnosti $\alpha = 0.05$.

Data: Koncentrace chromu x [mg/l], absorbance y.

x:	0.1	0.25	0.5	1.0	2.5	5.0	1.7	3.0	4.0
y:	0.004	0.010	0.021	0.041	0.103	0.206	0.070	0.124	0.165

Úloha K6.26 *Stanovení dusičnanů spektrofotometricky*
Pro vodné roztoky s různou koncentrací dusičnanů byla proměřována kalibrační závislost absorbance v UV oblasti (vlnová délka 220 nm) na obsahu dusičnanových iontů v roztoku. Úkolem je (1) navrhnout kalibrační model včetně testování významnosti parametrů modelu, (2) posoudit míru spolehlivosti navrženého modelu, míru přesnosti kalibrace a (3) určit intervalové odhady koncentrace pro naměřené hodnoty absorbance $y^* = 0.117, 0.491, 0.710, 0.969, 1.014, 1.255, 1.488, 1.750, 2.016$ a 2.106 na hladině významnosti $\alpha = 0.05$.

Data: Koncentrace dusičnanů x [mg.l^{-1}], absorbance y.

c:	1	5	10	15	20	30	50	25	33	40	45
y:	0.065	0.295	0.559	0.816	1.062	1.512	2.223	1.293	1.638	1.908	2.076

Úloha K6.27 *Kalibrace plynového chromatografu pomocí kalibračního plynu*
Naměřené kalibrační hodnoty jsou uvedeny tak, že hodnoty x představují známé obsahy kalibrační látky a hodnoty y počet integračních jednotekkrát 10^{-4} v neznámém kalibračním modelu $y = f(x)$. (1) Jaká je koncentrace neznámé látky pro $y^* = 0.00015$ a 0.0080? (2) Jaké jsou míry přesnosti kalibrace?

Data: koncentrace x, signál y [jednotek].

x:	1.5	8.0	36	56	119	238	315	397
y:	0.0001	0.0010	0.0050	0.0070	0.0150	0.0300	0.0400	0.0500

Úloha K6.28 *Kalibrace benzenu na plynovém chromatografu*
Na chromatografickém přístroji TDAS 5000 byla provedena kalibrace benzenu signály pro koncentraci x. (1) Nalezněte vhodný kalibrační model $y = f(x)$, vyšetřete vlivné body, posuďte spolehlivost navrženého modelu a míry přesnosti kalibrace. (2) Jaká je koncentrace benzenu pro velikost signálu $y^* = 333, 444$ a 555?

Data: Koncentrace x, signál y.

x:	200.0	200.0	200.0	400.0	400.0	600.0	600.0	800.0	800.0	800.0	800.0	1000.0
	1000.0	2000.0	2000.0	4000.0	6000.0	6000.0	8000.0	10000	9000	6800	4800	2800
y:	336.65	337.40	309.91	609.41	544.90	785.93	808.44	1027.8	1060.1	991.51	1010.6	1189.4
	1232.4	2399.9	2438.5	4524.6	5924.8	5972.9	6804.6	7812.6	7416	6371	5052	3294

Úloha K6.29 *Stanovení koncentrace manganu a míry přesnosti kalibrace*
Stanovení manganu ve vodách se provádí oxidací jodistanem v kyselém prostředí až na manganistan. (1) Sestrojte kalibrační model $y = f(x)$ pro naměřená data koncentrace x a absorbanci y a určete míry přesnosti kalibrace. (2) Neznámé roztoky vykazovaly absorbance $y^* = 0.005, 0.020, 0.115$. (3) Jsou v datech vybočující hodnoty?

Data: Koncentrace x [μg/ml Mn], absorbance y.

2.5	0.026,	5.0	0.052,	8.0	0.082,	10.0	0.102,	13.0	0.132,	15.0	0.150,
17.0	0.173,	20.0	0.200,								

Úloha K6.30 *Turbidimetrické stanovení síranů ve vodách*

Rozhodněte o lineární či nelineární kalibrační závislosti $y = f(x)$ pro stanovení sraženiny síranu barnatého ve vodě turbidimetricky při $\lambda_{max} = 680$ nm. Neznámé roztoky vykazovaly turbidanci y v dílcích stupnice $y^* = 55.0, 100.0, 120.0$.

Data: Koncentrace x [mg/l], turbidance y [mm].

88.8	52.0,	150.0	75.0,	177.61	84.0,	230.0	98.0,	266.41	106.0,	330.0	118.0,
355.22	110.0,	444.0	134.0								

Úloha K6.31 *Kalibrace obsahu látky MCPA metodou plynové chromatografie*

Proměřením standardních vzorků MCPA byla metodou plynové chromatografie GC určena kalibrační přímka $y = f(x)$. Neznámé vzorky vykazovaly plochu pod píkem $y^* = 985$ a 1356. Vyšetřete vlivné body, parametry kalibrace a intervalový odhad neznámé koncentrace na hladině významnosti $\alpha = 0.05$.

Data:

Koncentrace c [mg/l]:	25.0	31.6	37.5	43.1	49.0	83.7	7.0	13.0	17.8
y [jednotky]:	710	940	1140	1280	1510	2590	4963	7546	8716

Úloha K6.32 *Kalibrace zbytkového obsahu lindanu chromatograficky*

Sestrojte kalibrační model pro stanovení zbytkového obsahu lindanu v půdě x metodou plynové chromatografie s EC detektorem y. (1) Vyšetřete linearitu kalibračního grafu a míry přesnosti kalibrace. (2) Určete množství lindanu ve dvou vzorcích půdy, jež vykazovaly plochu píku $y^* = 1200$ a 2500 jednotek. (3) Není hodnota 1200 pod limitou detekce lindanu?

Data: Obsah lindanu x [ppb], y plocha píku [jednotky].

x [ppb]:	5.0	10.0	15.0	20.0	25.0	30.0
y [jednotky]:	3743.0	6662.0	7913.0	9026.0	9835.0	10081

Úloha K6.33 *Kalibrace obsahu fenolu v odpadních vodách fotometricky*

Spektrofotometrické stanovení fenolů v odpadních vodách se provádí fotometricky, a to vícenásobným standardním přídavkem p-nitroanilinu k neznámému vzorku $y = \beta_0 + \beta_1 x$. Každé měření se koriguje na pozadí. (1) Určete obsahy fenolů v odpadních vodách, když vzorky vykazovaly absorbance $y^* = 0.105, 0.205, 0.505$ a 0.905. (2) Je nejnižší obsah fenolu ve vzorcích nad limitou "ještě stanovitelného obsahu"?

Data: Obsah p-nitroanilinu x [mg], absorbance y.

0.00	0.140,	0.10	0.310,	0.15	0.385,	0.20	0.460,	0.30	0.640,	0.40	0.820,
0.50	0.980										

Úloha K6.34 *Nelineární kalibrace koncentrace kyseliny kyanurové polarograficky*

Metodou diferenční pulzní polarografie byla zjišťována výšky vlny kyseliny kyanurové y na její koncentraci x. Určete nelineární kalibrační model a vyšetřete vlivné body. Neznámé roztoky vykazovaly výšky vln $y^* = 185, 205, 250$ mm. Je nejnižší koncentrace neznámého vzorku pod či nad ní?

Data:

Koncentrace x [ppb]:	5	7	9	12	14	16	18	20	22
Výška vlny y [mm]:	178	181	186	199	211	227	242	264	285

Úloha K6.35 *Fotometrické určení chemické spotřeby kyslíku CHSK*
Na výběru dat kalibrační přímky $y = \beta_0 + \beta_1 x$ fotometrického stanovení y chemické spotřeby kyslíku x je třeba provést testování úseku a směrnice, vyšetření vlivných bodů a jejich event. odstranění, posouzení míry spolehlivosti navrženého modelu, míry přesnosti kalibrace, intervalový odhad neznámé koncentrace s důrazem na vysvětlení tzv. "ještě stanovitelné koncentrace" na hladině významnosti $\alpha = 0.05$. Neznámé vzorky měly absorbanci $y^* =$ 0.092, 0.127, 0.214 a 1.153.

Data: Koncentrace CHSK x [mg/l], absorbance y.

0.0	0.002,	20.0	0.041,	50.0	0.086,	100.0	0.157,	200.0	0.329,	500.0	0.790,
800.0	1.262,	900.0	1.406,								

Úloha K6.36 *Oxidovatelnost odpadní vody dichromanem fotometricky*
Stanovení oxidovatelnosti dichromanem semimikrometodou s fotometrickou koncovkou vyniká širokým rozsahem. (1) Na souboru kalibračních dat závislosti absorbance y při 600 nm na koncentraci x je třeba sestavit kalibrační model $y = f(x)$ s vyšetřením vlivných bodů, mírami přesnosti kalibrace a intervalovým odhadem neznámé koncentrace. (2) Neznámé vzorky měly absorbanci $y^* = 0.112$, 0.213 a 0.982. (3) Je nejnižší koncentrace již pod limitou detekce? (4) Jsou v kalibračních datech odlehlá měření? Hladina významnosti je $\alpha = 0.05$.

Data: Koncentrace x [mg/l], absorbance y (3× opakovaná).

20	0.042	0.031	0.035,	40	0.080	0.062	0.056,	80	0.148	0.134	0.125,
...	...	...	...	...	...	...	...	...	...	...	...
600	0.918	0.926	0.921,	800	1.200	1.222	1.238,	1000	1.448	1.437	1.441,

Úloha K6.37 *Kalibrační model obsahu Centralitu I v nitrocelulózovém prachu*
Centralit I, přidávaný do nitrocelulózového prachu jako stabilizátor, se stanovuje rozkladem prachu roztokem hydroxidu draselného a následnou analýzou na plynovém chromatografu metodou ISTD. Při kalibraci byla měřena plocha pod chromatografickým píkem pro přesně určené koncentrace Centralitu I. Do nelineárního kalibračního grafu $y = f(x)$ se vynáší poměr ploch Centralitu I a vnitřního standardu jako hodnoty y vůči koncentraci Centralitu I jako hodnoty x. (1) Zjistěte kalibrační meze a obsah Centralitu I v neznámých vzorcích, které vykazovaly následující hodnoty poměru ploch Centralitu I a ISTD $y^* = 0.9927$, 1.5088, 2.0628. (2) Sestrojte kalibrační model a určete charakteristiky přesnosti kalibrace na hladině významnosti $\alpha = 0.05$. (3) Jsou v kalibračním grafu nějaké vlivné body?

Data: Koncentrace Centralitu I x [%], poměr ploch Centralitu I vůči vnitřnímu standardu y.

x:	1.03	1.20	1.93	2.99	3.89	4.96	0.50	1.50	2.50	3.5
y:	0.9467	1.0113	1.3248	1.6099	1.8316	2.0693	0.675	1.149	1.497	1.735

Úloha K6.38 *Stanovení fenolů fotometricky*
Byly změřeny absorbance y standardních roztoků jednomocných fenolů spektrofotometrickou metodou. (1) Proveďte stanovení intervalového odhadu neznámé koncentrace vzorků jednomocných fenolů, jež vykazují absorbanci $y^* = 0.120$, 0.320, 0.420. (2) Jsou v kalibračních datech odlehlé body, které je třeba vyloučit?

Data: Koncentrace *x* [mg/l], absorbance *y*.

x:	0.08	0.159	0.227	0.315	0.398	0.495	0.598	0.707	0.836
y:	0.050	0.100	0.150	0.200	0.250	0.300	0.350	0.400	0.450

Úloha K6.39 *Kalibrační model pro stanovení amino-R-kyseliny*

Stanovení amino-R-kyseliny v amino-G-kyselině se provádí kapilární elektroforézou metodou vnějšího standardu. Proměřením pěti kalibračních roztoků byla získána závislost signálu *y* v jednotkách mAU/s na koncentraci amino-R-kyseliny *x*. Úkolem bylo zjistit obsah nečistoty ve třech vzorcích, které poskytly signál $y^* = 39.00$, 46.82 a 53.36 mAU/s.

Data: Koncentrace amino-R-kyseliny *x* [%], velikost signálu *y* [mAU/s].

0.0831 11.13, 0.4000 59.89, 0.1385 18.13, 0.2771 39.79, 0.5541 84.39, 0.9995 154.23, 0.7500 114.9,

Úloha K6.40 *Stanovení koncentrace p-dichlorbenzenu v chlorbenzenu*

Stanovení koncentrace p-dichlorbenzenu v chlorbenzenu se provádí metodou plynové chromatografie na vnitřní standard toluen. Proměřením kalibračních roztoků byla získána nelineární závislost poměru ploch p-dichlorbenzenu a vnitřního standardu *y* na procentní koncentraci p-dichlorbenzenu *x*. Proměřením dvou vzorků chlorben-zenu byly získány odezvy $y^* = 0.024$ a 0.048. Úkolem je zjistit (1) koncentraci p-dichlorbenzenu v obou vzorcích a (2) míry přesnosti kalibrace.

Data: koncentrace p-dichlorbenzenu *x* [%], poměr ploch p-dichlorbenzenu a vnitřního standardu *y*.

0.01 0.013,	0.02 0.027,	0.03 0.040,	0.04 0.048,	0.05 0.053,	0.08 0.070,
0.10 0.090,	0.15 0.120,	0.20 0.150,	0.30 0.220,	0.40 0.270,	0.50 0.320,

Úloha K6.41 *Stanovení koncentrace chloridů fotometricky*

Ve vzorku jodistanu draselného se stanovuje koncentrace chloridů a chlorečnanů jako celkové chloridy. Chlorečnany se redukují na chloridy jodidem draselným a jodistan se redukuje na jod. Po odstranění jodu se chloridy srážejí dusičnanem stříbrným a vzniklý zákal se měří fotometricky. (1) Ze změřených absorbancí kalibračních roztoků o různé koncentraci chloridů určete kalibrační přímku a z ní odhadněte koncentraci chloridů pro naměřené absorbance vzorků $y^* = 0.0152$, 0.0199. (2) Vyšetřete, zda je místo lineárního kalibračního modelu vhodnější použít kvadratický kalibrační model. Pracujte na hladině významnosti $\alpha = 0.05$.

Data: Koncentrace *x* [mg/ml Cl^-], absorbance *y*.

x:	0.005	0.010	0.015	0.020	0.025	0.030
y:	0.0046	0.0113	0.0172	0.0227	0.0288	0.0339

Úloha K6.42 *Kalibrační model fotometrického stanovení*

Metodou AAS byla změřena absorbance *y* roztoků o různé koncentraci *x* stanovované složky B [ppm]. (1) Určete kalibrační přímku a z ní odhadněte koncentraci složky B pro naměřené absorbance neznámého vzorku $y^* = 0.149$, 0.237. (2) Jsou v kalibračních datech odlehlé hodnoty? (3) Vyčíslete Naszodiho modifikovaný odhad neznámé koncentrace uvedený na str. 497 v cit.[72].

Data: Koncentrace *x* [ppm], absorbance *y*.

0 0.021, 30 0.077, 60 0.130, 100 0.200, 150 0.291, 200 0.380,

Úloha K6.43 *Kalibrační model stanovení AAS a limity přesnosti dle Ebela a Kamma*

Metodou AAS byla změřena absorbance y roztoků o různém obsahu x složky B. Určete kalibrační model a odhadněte obsah složky B v ppm pro naměřené absorbance y neznámého vzorku $y^* = 0.060, 0.152$. Vyčíslete i limitu detekce dle Ebela a Kamma, viz str. 506 v cit.[72].

Data: Obsah x [ppm], absorbance y.

0	0.025,	20	0.028,	30	0.033,	40	0.047,	50	0.057,	60	0.064,	80	0.080,	100	0.089,	
120	0.112,	140	0.118,	150	0.137,	180	0.141,	200	0.186,							

Úloha K6.44 *Rozlišení mezi lineárním a nelineárním kalibračním modelem*

Metodou AAS byl stanovena koncentrace složky B v daném vzorku. (1) Rozhodněte, zda se jedná o lineární nebo nelineární model kalibrace. (2) Nalezněte odpovídající model $y = f(x)$ a stanovte koncentraci tří neznámých vzorků. (3) Určete koncentrace složky B, je-li absorbance $y^* = 0.250, 0.780$. (4) Vyčíslete Naszodiho modifikovaný odhad neznámé koncentrace dle str. 497 v cit.[72].

Data: Koncentrace x [mg/l], absorbance y.

0	0.007,	0.5	0.238,	1.0	0.367,	1.5	0.590,	2.0	0.715,	2.5	1.006,	

Úloha K6.45 *Kalibrační model dusičnanů v pitné vodě fotometricky*

Proměřením řady standardních roztoků dusičnanů spektrometrickou metodou v UV oblasti bylo zjištěno, že závislost absorbance y na koncentraci x vykazuje od koncentrace nad 10 mg/l výrazné odchylky od linearity. (1) Vytvořte vhodný kalibrační model $y = f(x)$. (2) Vyšetřete data, zda-li v nich nejsou vlivné body, případně proveďte jejich eliminaci. (3) Posuďte míru spolehlivosti navrženého modelu, míru přesnosti kalibrace a intervalové odhady neznámých koncentrací vzorků, u kterých byla naměřena absorbance $y^* = 1.305$, 2.478 a 2.878. Pracujte na hladině významnosti $\alpha = 0.05$.

Data: Koncentrace dusičnanů x [mg/l], absorbance y.

x:	0.00	5.00	10.00	15.00	20.00	25.00	30.00	40.00	50.00
y:	0.127	0.662	1.185	1.690	2.186	2.510	2.791	2.983	3.178

Úloha K6.46 *Kalibrační modely při stanovení hliníku ve vodě*

Na základě vyšetření regresního tripletu technikou regresních diagnostik a těsnosti proložení kalibračních dat rozhodněte, zda u kalibrační závislosti absorbance y na koncentraci hliníku x v pitné vodě jde o lineární či nelineární model. Pomocí vhodného, předem zjištěného modelu určete koncentraci tří neznámých vzorků, pokud naměřené absorbance jsou $y^* = 0.750, 1.350, 2.180$.

Data: Koncentrace hliníku x [mg/l], absorbance y.

x:	0.00	0.05	0.10	0.15	0.25	0.35	0.50
y:	0.717	0.991	1.231	1.487	1.777	2.029	2.447

Úloha K6.47 *Kalibrace nefelometru*

Nefelometr byl kalibrován na obsah pevné fáze dispergované v destilované vodě. Pro standartní suspenzi byla změřena kalibrační data $y = f(x)$. (1) Zjistěte lineární kalibrační model, míru spolehlivosti modelu, míry přesnosti kalibrace a intervalový odhad neznámé koncentrace pro turbidance $y^* = 15$, 25 a 55. (2) Vyčíslete i modifikovanou limitu kvantifikace.

Data: Obsah x [ppm], turbidance y [dílky].

x:	0.00	0.60	0.15	0.70	0.30	0.40	0.50
y:	0.00	76.00	23.00	82.00	38.00	45.00	61.00

Úloha K6.48 *Určení stopové koncentrace draslíku v kamenci rubidnocesném*
Sestrojte kalibrační model $y = f(x)$ a stanovte stopovou koncentraci draslíku v kamenci rubidnocesném. Určete (1) kalibrační limity, (2) parametry kalibračního modelu a (3) koncentraci draslíku u neznámých vzorků, jež vykazovaly odchylku $y^* = 2100$, 2310 a 2450 jednotek. (4) Jsou v datech vybočující hodnoty?

Data: Koncentace draslíku *x* [mg/l], hodnota signálu *y* [jednotky].

c:	1.0	1.6	2.5	3.2	4.0	4.8	6.0	6.5	7.5	1.5	2.0	3.0
	3.5	4.5	5.0	6.4	7.0	8						
y:	1937	1967	2044	2118	2210	2298	2412	2458	2538	1961	1997	2100
	2155	2263	2315	2446	2499	2574						

Úloha K6.49 *Fotometrické stanovení CHSK-Mn v pitné vodě*
Pomocí 40 dat naměřených v prosinci 1994 a v lednu 1995 v laboratoři pitných vod zjistěte, zda-li existuje lineární závislost mezi absorbancí vzorku *y*, měřenou při vlnové délce 254 nm a chemickou spotřebou kyslíku manganometricky CHSK-Mn *x* v mg/l. (1) Proveďte vyšetření vlivných bodů, případně jejich odstranění a posouzení míry spolehlivosti navrženého modelu. (2) Na základě odděleného vyšetření dat z prosince 1994 a ledna 1995 porovnejte obě přímky, a to pomocí (a) testu shody rozptylů, (b) testu homogenity úseků a homogenity směrnic a (c) testu shody regresních přímek. (3) Jaké jsou míry přesnosti kalibrace?

Data: Obsah CHSK-Mn *x* [mg/l], absorbance *y*.

Prosinec 1994:

x:	0.3	0.5	0.3	0.4	3.0	3.0	2.4	0.3	4.2	0.8	0.4	1.3
	2.6	2.7	0.6	3.0	1.8	3.1	2.5	2.2				
y:	0.009	0.013	0.008	0.000	0.107	0.087	0.072	0.004	0.086	0.042	0.024	0.006
	0.073	0.091	0.050	0.091	0.074	0.072	0.072	0.050				

Leden 1995:

x:	0.4	2.2	0.3	0.4	1.1	3.2	0.5	0.4	0.6	2.0	2.6	1.3
	0.5	0.2	0.8	1.1	0.9	1.2	1.1	1.3				
y:	0.010	0.110	0.020	0.020	0.040	0.100	0.020	0.010	0.010	0.060	0.080	0.050
	0.020	0.020	0.040	0.030	0.030	0.020	0.030	0.040				

Úloha K6.50 *Fotometrické stanovení koncentrace fenolů v pitné vodě*
Proměřením řady standardů fenolů byla fotometrickou metodou určena kalibrační přímka závislosti absorbance *y* na koncentraci *x*. (1) Zjistěte parametry kalibrační přímky, včetně testování jejího úseku (má být roven nule) a směrnice (má být rovna jedné). (2) Vyšetřete, zda-li v naměřených datech existují odlehlé hodnoty, případně proveďte jejich eliminaci. (3) Posuďte míru přesnosti kalibrace a intervalový odhad neznámé koncentrace vzorku na hladině významnosti $\alpha = 0.05$, když byla změřena absorbance $y^* = 0.188$.

Data: Koncentrace *x* [mg/l], absorbance *y*.

x:	0.00	0.02	0.05	0.10	0.20	0.30	0.50
y:	0.095	0.127	0.177	0.235	0.380	0.503	0.792

Úloha K6.51 *Fotometrické stanovení oxidů dusíku v topném oleji*
Postupem dle normy ČSN 38 5537 byla třemi laboranty naměřena kalibrační křivka fotometrického stanovení obsahu oxidů dusíku NO v topných plynech za užití Griessova činidla. Zjistěte přesnost práce laborantů, limity přesnosti kalibrace a intervalový odhad obsahu oxidů dusíku NO ve dvou neznámých vzorcích pro absorbance $y^* = 0.140, 0.040$ na hladině významnosti $\alpha = 0.05$.

Data: Obsah NO *x*, absorbance *y*.

(a) 1. laborant:	0.14	0.0046,	0.42	0.011,	0.78	0.017,	0.98	0.026,	1.26	0.030,
	1.40	0.036	4.2	0.104,	7	0.172,	11.2	0.278,	14	0.34
(b) 2. laborant:	0.14	0.0042,	0.42	0.009,	0.78	0.012,	0.98	0.020,	1.26	0.025,
	1.40	0.03	4.2	0.096,	7	0.154,	11.2	0.237,	14	0.31
(c) 3. laborant:	0.14	0.0050,	0.42	0.014,	0.78	0.021,	0.98	0.032,	1.26	0.034,
	1.40	0.043	4.2	0.121,	7	0.196,	11.2	0.301,	14	0.370,

Úloha K6.52 *Stanovení arsenu hybridovou technikou AAS*
Aproximujte kalibrační závislost $y = f(x)$ absorbance *y* na koncentraci *x* při stanovení arsenu přímkou a polynomem a rozhodněte, který model je vhodnější. (1) Rozlište mezi lineárním a kvadratickým modelem. (2) Stanovte koncentrace tří vzorků se signály absorbance $y^* = 0.029, 0.091$ a 0.122. (3) Jaké jsou míry přesnosti kalibrace?

Data: Koncentrace *x* [µg/l], absorbance *y*.

x:	0.5	1.0	1.5	2.0	3.0	4.0
y:	0.019	0.040	0.060	0.079	0.116	0.151

Úloha K6.53 *Stanovení bromidů ve vodách metodou iontové chromatografie*
Sestrojte lineární kalibrační model stanovení bromidů ve vodách metodou iontové chromatografie. (1) Metodou regresní analýzy vyšetřete linearitu kalibračního modelu, určete parametry kalibračního modelu, kalibrační limity a proveďte odhady koncentrace neznámých vzorků bromidů, které vykazovaly plochu $y^* = 505, 1005$ a 1505. (2) Jsou tyto koncentrace ještě detekovatelné? (3) Obsahuje kalibrační graf vlivné body?

Data: Koncentrace bromidů *x* [mg. l^{-1}], plocha pod píkem *y* [jednotky].

x:	5.0	10.0	15.0	20.0	25.0	30.0	40.0	50.0	60.0	70.0	90.0
y:	157	279	428	533	723	778	1008	1251	1445	1653	2009

Úloha K6.54 *Stanovení koncentrace zinku v mléce metodou plamenné fotometrie*
Sestrojte nelineární kalibrační model stanovení obsahu zinku v mléce metodou plamenné fotometrie $y = f(x)$. Pro 20 úrovní obsahu zinku byl získán naměřením hodnot absorbance kalibrační graf. Pro navržený model stanovte parametry, kalibrační limity a proveďte odhady koncentrace zinku u neznámých vzorků, jež vykazovaly absorbance $y^* = 0.105, 0.205, 0.315$ a 0.445.

Data: Koncentrace zinku *x* [ppm], absorbance *y*.

0.050	0.031,	0.300	0.195,	0.700	0.453,	1.200	0.718,
...	...	...	...	...	...	...	...
0.250	0.161,	0.600	0.396,	1.100	0.668,	1.600	0.867,

Úloha K6.55 *Stanovení obsahu mědi v oceli emisní spektrální analýzou*
Najděte kalibrační model obsahu $y = f(x)$ mědi v oceli emisní spektrální analýzou u 11 vzorků. Určete (1) parametry kalibračního modelu, (2) charakteristiky přesnosti kalibrace a (3) proveďte odhady obsahu mědi u neznámých vzorků, které vykazovaly signál $y^* = 55$, 105, 155 a 205 jednotek.

Data: Obsah mědi x [%], hodnota signálu y [jednotky].

x:	0.019	0.030	0.047	0.067	0.072	0.084	0.16	0.26	0.36	0.47	0.70
y:	28	33	48	72	68	86	119	166	216	284	380

Úloha K6.56 *Kalibrační model polarografického stanovení thalných iontů*
Thalné ionty lze nakoncentrovat do visící rtuťové kapky, jejíž potenciál je vyšší, než je vylučovací potenciál thalných iontů. Po ukončení míchací fáze bylo "stripování" dokončeno uklidněním roztoku a při současném snižování potenciálu rtuťové elektrody byla zaznamenávána polarografická křivka "rozpouštění" thalia zpět do roztoku. Byla měřena výška píků, odpovídající thalným iontům s přepočtem na korekci objemu. (1) Sestrojte kalibrační model $y = f(x)$ stanovení thalných iontů včetně testování statistické významnosti úseku a směrnice. (2) Proveďte také vyšetření vlivných bodů a (3) posouzení míry spolehlivosti navrženého modelu, (4) vypočtěte míry přesnosti kalibrace a intervalový odhad neznámé koncentrace. (5) Zdůrazněte "ještě stanovitelnou koncentraci". Výška píku neznámých vzorků byla $y^* = 15.2$, 75.5 a 111.1 mm.

Data: Množství thalia x [μg], výška píků thalných iontů y [mm] po korekci na objem.

x:	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
y:	16.06	32.76	51.51	70.71	88.74	105.37	123.87	139.22	159.44

Úloha K6.57 *Nelineární kalibrační model koncentrace měďnatých iontů*
Na základě výsledků regresní diagnostiky rozhodněte o druhu kalibračního modelu pro případ titrace měďnaté soli roztokem chelatotvorné organické látky DODA-DTPA. Měřena byla výška píků odpovídající koncentraci komplexu Cu-DODA-DTPA. Sestrojte kalibrační model $y = f(x)$ s vyšetřením vlivných bodů a s posouzením míry spolehlivosti navrženého modelu, míry přesnosti kalibrace, intervalového odhadu neznámé koncentrace. Výška píku neznámých vzorků byla $y^* = 35$, 175 a 250 mm.

Data: Koncentrace měďnatého komplexu x [10^{-5} mol. l^{-1}], výška píků y [mm].

x:	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
y:	37	80	117	136	168	189	217	328

Úloha K6.58 *Rozlišení u nepřímého polarografického stanovení dusičnanů*
Dusičnanové a dusitanové anionty lze stanovit nepřímo polarograficky ze zvýšení píku redukce uranylového kationtu. (1) Proveďte rozlišení mezi lineárním a nelineárním kalibračním modelem $y = f(x)$ u nepřímého polarografického stanovení dusičnanů. (2) Vyšetřením regresního tripletu rozhodněte, zda jde u dané kalibrační závislosti o lineární či nelineární kalibrační model a určete koncentraci tří neznámých vzorků s výškou píku $y^* = 30$, 70 a 90 mm.

Data: Koncentrace dusičnanů x [mg/l], výška píků y [mm] po korekci objemu.

x:	62	124	186	248	310	372	434
y:	19.9	40.5	59.9	75.1	89.0	100.2	111.7

Úloha K6.59 *Kalibrační model CRP imunoturbidimetrickou metodou*

Sestrojte kalibrační model C-reaktivního proteinu CRP z imunoturbidimetrických měření standardních roztoků, určete parametry kalibračního modelu a míry přesnosti kalibrace. (1) Jsou obě kalibrační přímky rovnoběžné nebo zcela totožné? (2) Jaká je koncentrace CRP u vzorků, jež vykazovaly absorbance $y^* = 0.0638, 0.1447, 0.3573$ a 0.9008?

Data: Koncentrace CRP x [mg/l], dvě opakovaná měření absorbance y_1 a y_2.

6.25	0.0248	0.0262,	10.0	0.0348	0.0358,	12.5	0.0410	0.0485,
...	...	...	...	...	...	...	...	...
200.0	0.9330	0.9219,						

Úloha K6.60 *Kalibrační model haptoglobinu imunoturbidimetrickou metodou*

Nalezněte (1) vhodný kalibrační model $y = f(x)$ pro imunoturbidimetrické stanovení haptoglobinu Hpl v krevním séru. (2) Určete parametry kalibračního modelu a míry přesnosti kalibrace. (3) Jsou v datech nějaké vlivné body? (4) Prochází kalibrační přímka počátkem? (5) Jaká je koncentrace Hpl u vzorků, jež vykazovaly absorbance $y^* = 0.095$, 0.392 a 0.598?

Data:

Koncentrace Hpl x [g/l]:	0.00	0.30	0.59	0.70	0.89	1.77	2.35
	3.54	4.15	5.60	7.08			
Absorbance y:	0.022	0.050	0.094	0.115	0.150	0.290	0.392
	0.572	0.688	0.917	1.137			

Úloha K6.61 *Stanovení hořečnatých iontů enzymatickou fotometrickou metodou*

Sestrojte kalibrační model $y = f(x)$ enzymatického stanovení hořečnatých iontů Mg^{2+} z naměřených absorbancí standardních roztoků, určete parametry kalibračního modelu a míry přesnosti kalibrace. Jaká je koncentrace Mg^{2+} u vzorků s absorbancemi $y^* = 0.405$, 0.652, 0.154?

Data: Koncentrace Mg^{2+} x [mmol/l], absorbance y.

x:	0.20	0.40	0.50	0.60	0.80	1.00	1.10	1.20	1.30
	1.40	1.6							
y:	0.122	0.247	0.316	0.370	0.498	0.627	0.688	0.751	0.813
	0.878	0.999							

Úloha K6.62 *Vyšetření kalibračního grafu kyseliny hippurové v moči*

Měřením absorbance v UV oblasti lze stanovit kyselinu hippurovou v moči, a to jako psychotropní metabolit toluenu. Během tří let byly naměřeny čtyři kalibrační křivky, které se mají vyšetřit testem shodnosti přímek, testem paralelnosti a společného úseku. Jaká je koncentrace kyseliny hippurové, když neznámé roztoky vykazovaly absorbanci $y^* = 0.050$, 0.420, 0.650 a 0.811?

Data: Koncentrace kyseliny hippurové x [mmol/l], absorbance y_A, y_B, y_C, y_D.

x	y_A	y_B	$\boldsymbol{y_C}$	$\boldsymbol{y_D}$
2.8	0.085	0.110	0.115	0.096
...	...	...	...	...
27.9	0.889	0.968	0.900	0.897

Úloha K6.63 *Kalibrační přímka stanovení amonných iontů*

Amonné ionty se stanovují spektrofotometricky dle ČSN ISO 7150-1. (1) Vyšetřete vlivné body, určete parametry přesnosti kalibrace a intervalový odhad neznámé koncentrace o absorbanci $y^* = 0.0105$, 0.1800 a 0.3795. (2) Testujte také statistickou významnost úseku (má být $\beta_0 = 0$) a směrnice (má být $\beta_1 = 1$) kalibrační přímky.

Data: Koncentrace x [mg NH_4^+/l], absorbance y.

0.025	0.0254,	0.050	0.0413,	0.100	0.0798,	0.125	0.0929,	0.250	0.1872,
0.500	0.3569,	0.750	0.5381,	1.00	0.7111,	1.25	0.8734		

Úloha K6.64 *Nelineární kalibrace Zn metodou AAS*

Stanovení zinku bylo provedeno metodou AAS v rozmezí od 0.5 mg/l do 50 mg/l. (1) Určete kalibrační model a vyšetřete vlivné body. (2) Neznámé vzorky vykazovaly absorbanci $y^* =$ 0.021, 0.305 a 0.695. (3) Určete také intervalový odhad neznámé koncentrace. (4) Jsou neznámé koncentrace v oblasti "ještě detekovatelné" absorbance?

Data: Koncentrace Zn x [mg/l], absorbance y.

x:	0.5	1.0	2.0	15.0	25.0	30.0	35.0	40.0	50.0
y:	0.017	0.033	0.064	0.417	0.591	0.630	0.677	0.706	0.742

Úloha K6.65 *Stanovení NEL metodou UV spektrofotometrie*

Nepolární extrahovatelné látky NEL jsou stanovovány spektrofotometrickou metodou. Pro oblast UV spektrofotmetrie byla zhotovena kalibrační křivka $y = f(x)$ nástřikem trafo-oleje do destilované vody. (1) Určete, zda je lépe použít k proložení bodů přímky nebo kvadratický spline. (2) Neznámé vzorky vykazovaly absorbanci $y^* = 0.050$, 0.401 a 0.750. (3) Určete také intervalový odhad neznámé koncentrace.

Data: Nástřik x [µl LTO], absorbance y.

10	0.063,	20	0.185,	30	0.302,	40	0.437,	50	0.554,	70	0.733,
80	0.798,	90	0.828,	100	0.886						

Úloha K6.66 *Lineární kalibrace výkonu žárovky 200 W*

Byla změřena závislost výkonu žárovky y na napětí x. Je třeba provést lineární kalibraci a zjistit hodnoty napětí při úrovních výkonu $y^* = 50$ W, 90 W, 150 W. Prochází kalibrační přímka počátkem?

Data: Napětí x [V], výkon y [W].

x:	60	80	100	120	140	160	180	200	220
y:	26	42	58	78	100	120	146	172	200

Úloha K6.67 *Nelineární kalibrace výkonu zářivky 100 W*

Byla změřena závislost výkonu zářivky y na napětí x. Je třeba určit typ nelineární kalibrace a zjistit hodnoty predikce napětí při výkonu $y^* = 10$ W, 50 W, 75 W.

Data: Napětí x [V], výkon y [W].

x:	0	20	40	60	80	100
y:	0	2	6.5	12	18.5	26
x:	120	140	160	180	200	220
y:	35	44	54	64	76	88

6.4 Polynomické regresní modely

Při hledání optimálního stupně *m* polynomu (str. 372 v cit.[72])

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_m x^m \quad ,$$

kde β_0, β_1, β_2, β_3, ..., β_m, jsou odhadované parametry, patří mezi často užívaná rozhodčí kritéria *střední kvadratická chyba predikce*

$$MEP = \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i}^{\mathrm{T}} \boldsymbol{b_{(i)}})^2 \quad ,$$

kde $\boldsymbol{b_{(i)}}$ je odhad parametrů regresního modelu určený ze všech bodů kromě *i*-tého a $\boldsymbol{x}_i$ je *i*-tý řádek matice $\boldsymbol{X}$. Statistika *MEP* využívá predikce $y_{P,i}$ z odhadu, při jehož konstrukci byla informace o *i*-tém bodu vypuštěna. Dá se snadno ověřit, že pro *MEP* platí vztah

$$MEP = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{e}_i^2}{(1 - H_{ii})^2} \quad ,$$

Pro velké rozsahy dat *n* jsou prvky $H_{ii} \doteq 0$ a pak *MEP* = *RSC*/*n*, dle str. 374 v cit[72]. Užije-li se charakteristiky *MEP* místo *RSC* ve výpočtu koeficientu determinace, bude výsledkem *predikovaný koeficient determinace* D_P značený v literatuře také R^2_P

$$D_P = 1 - \frac{n\ MEP}{\sum_{i=1}^{n} y_i^2 - n\bar{y}} = \hat{R}_P^2 \quad ,$$

Univerzální použití mají také rozličná kritéria vycházející z teorie informace a entropie. Mezi nejznámější patří *Akaikovo informační kritérium AIC*

$$AIC = n \ln\left(\frac{RSC}{n}\right) + 2m \quad .$$

Při hledání stupně polynomu je za nejvhodnější považován takový model, pro který je *AIC*, ale také *MEP* minimální, zatímco D_P maximální.

Vzorová úloha 6.4 *Optimální stupeň polynomu a snížení multikolinearity*
Na úloze **L6.02** *Závislost výšky píku kyseliny kyanurové na koncentraci želatiny* ukážeme hledání optimálního stupně polynomu a snížení multikolinearity aplikací metody racionálních hodností (anglicky **G**eneralized **P**rincipal **C**omponent **R**egression, GPCR): Při stanovení kyseliny kyanurové metodou diferenční pulsní polarografie byl sledován vliv přítomnosti povrchově aktivních látek. (1) Určete stupeň polynomu *m* závislosti výšky píku kyseliny kyanurové *y* na koncentraci želatiny *x*. (2) Které z kritérií, *MEP* nebo *AIC*, má lepší rozlišovací schopnost při určení stupně polynomu? (3) Pokuste se snížit multikolinearitu.
Řešení:
1. Návrh modelu:
(a) *Určení stupně polynomu metodou nejmenších čtverců MNČ:*
K určení optimálního stupně polynomu budeme sledovat následující statistické charakteristiky pro rozličné stupně polynomu: *vícenásobný korelační koeficient r* a *koeficient determinace D* nejsou pro hledání optimálního stupně polynomu *m* vhodné.

Stupeň polynomu $m = 1$

Vícenásobný korelační koeficient r	: 0.97029
Koeficient determinace 100 % D	: 94.145
Predikovaný koeficient determinace R^2_P	: 0.94633
Střední kvadratická chyba predikce MEP	: 109.95
Akaikovo informační kritérium AIC	: 45.211

Stupeň polynomu $m = 2$

Vícenásobný korelační koeficient r	: 0.99800
Koeficient determinace 100 % D	: 99.600
Predikovaný koeficient determinace R^2_P	: 0.99481
Střední kvadratická chyba predikce MEP	: 10.894
Akaikovo informační kritérium AIC	: 20.376

Stupeň polynomu $m = 3$

Vícenásobný korelační koeficient r	: 0.99913
Koeficient determinace 100 % D	: 99.827
Predikovaný koeficient determinace R^2_P	: 0.99564
Střední kvadratická chyba predikce MEP	: 9.166
Akaikovo informační kritérium AIC	: 14.022

Stupeň polynomu $m = 4$

Vícenásobný korelační koeficient r	: 0.99965
Koeficient determinace 100 % D	: 99.929
Predikovaný koeficient determinace R^2_P	: 0.99398
Střední kvadratická chyba predikce MEP	: 12.639
Akaikovo informační kritérium AIC	: 7.080

Stupeň polynomu $m = 5$

Vícenásobný korelační koeficient r	: 0.99980
Koeficient determinace 100 % D	: 99.959
Predikovaný koeficient determinace R^2_P	: 0.98033
Střední kvadratická chyba predikce MEP	: 41.006
Akaikovo informační kritérium AIC	: 3.502

Stupeň polynomu $m = 6$

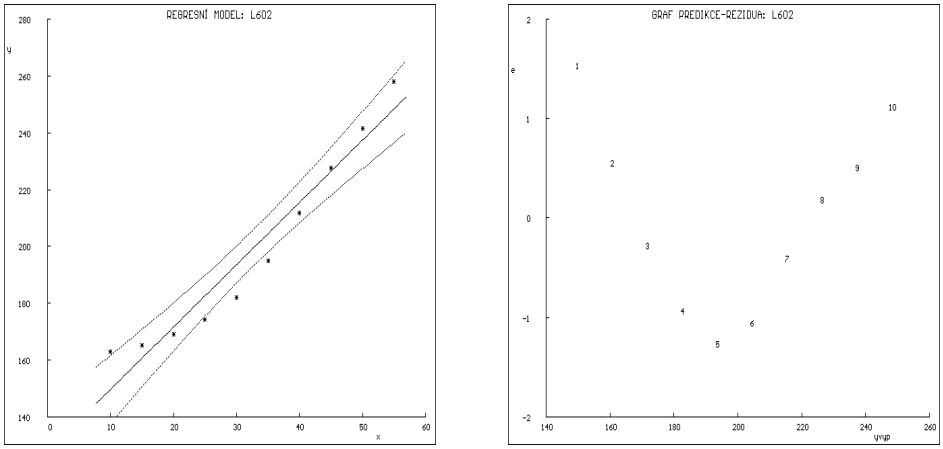
Vícenásobný korelační koeficient r	: 0.99999
Koeficient determinace 100 % D	: 99.999
Predikovaný koeficient determinace R^2_P	: 0.99863
Střední kvadratická chyba predikce MEP	: 2.883
Akaikovo informační kritérium AIC	:-29.367

Stupeň polynomu $m = 7$

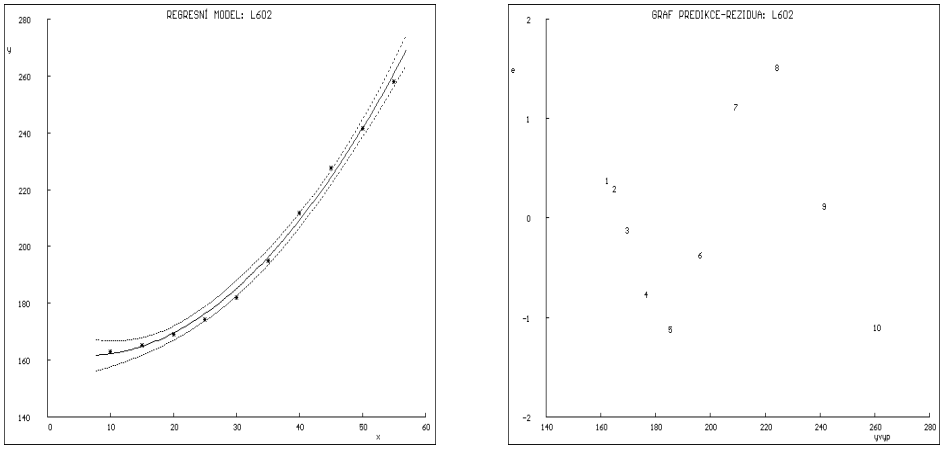
Vícenásobný korelační koeficient r	: 1.0000
Koeficient determinace 100 % D	: 99.999
Predikovaný koeficient determinace R^2_P	: 0.97561
Střední kvadratická chyba predikce MEP	: 50.729

Akaikovo informační kritérium *AIC*	:-30.950

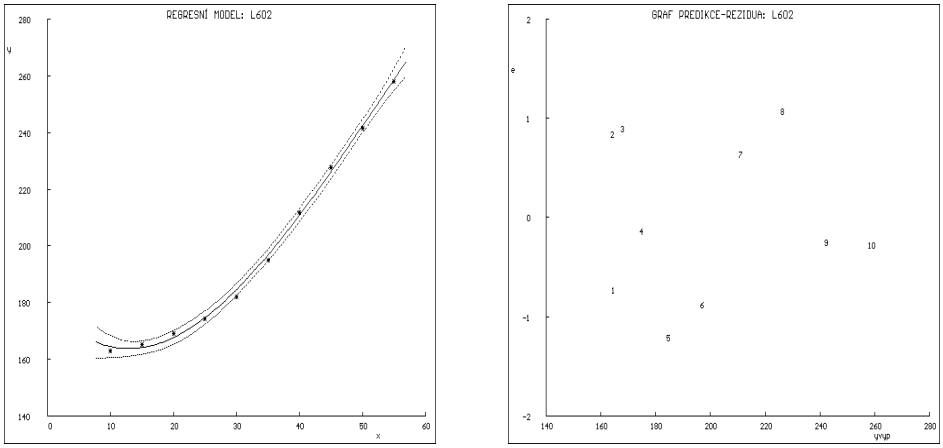
Grafická analýza klasických reziduí a rozptylový graf jsou rovněž užitečnou a názornou pomůckou při hledání stupně polynomu *m*. Vynesením *MEP* nebo *AIC* proti stupni polynomu *m* vidíme, že se nabízí dvě řešení, a to $m = 3$ a $m = 6$. S ohledem na maximální jednoduchost modelu lze volit $m = 3$. V tomto případě je účelem maximální těsnost proložení, a proto dáme přednost $m = 6$.



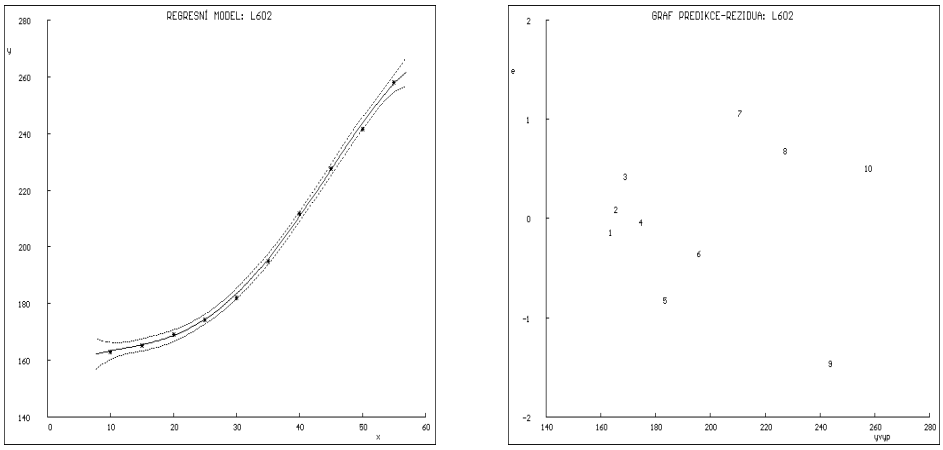
Obr. 6.4-1 Polynom 1. stupně, vlevo: graf těsnosti proložení, a vpravo: analýza reziduí, ***ADSTAT.***



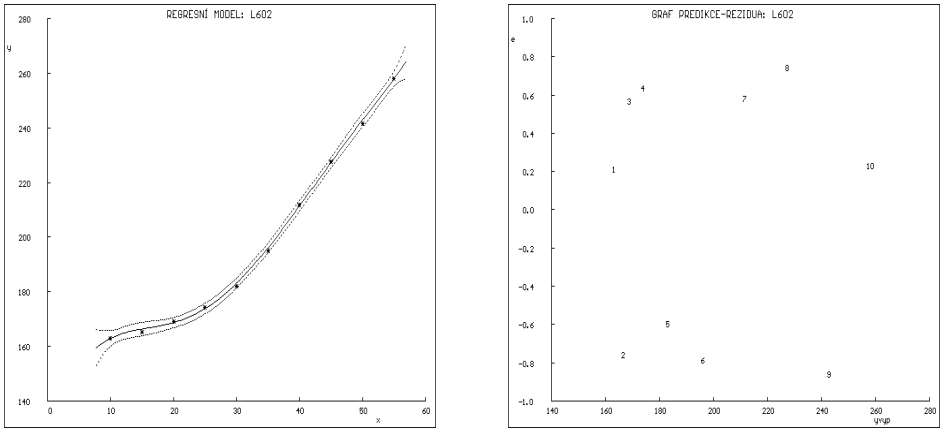
Obr. 6.4-2 Polynom 2. stupně, vlevo: graf těsnosti proložení, a vpravo: analýza reziduí, ***ADSTAT.***



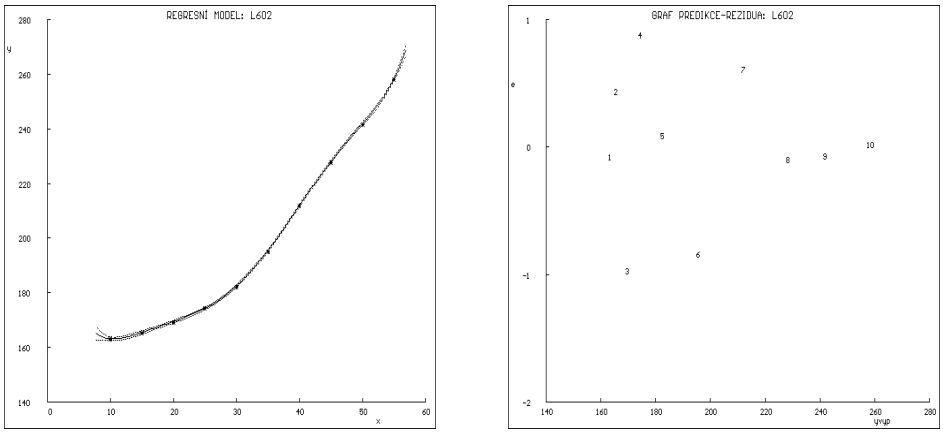
Obr. 6.4-3 Polynom 3. stupně, vlevo: graf těsnosti proložení, a vpravo: analýza reziduí, ***ADSTAT.***



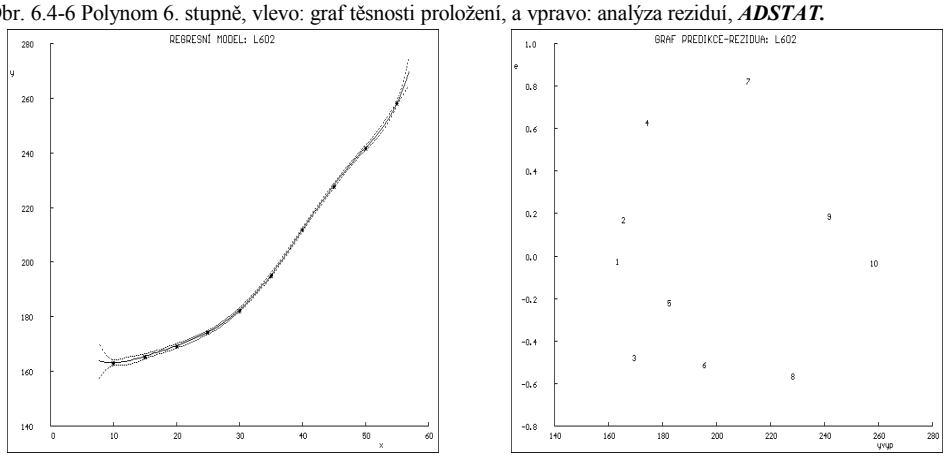
Obr. 6.4-4 Polynom 4. stupně, vlevo: graf těsnosti proložení, a vpravo: analýza reziduí, ***ADSTAT.***



Obr. 6.4-5 Polynom 5. stupně, vlevo: graf těsnosti proložení, a vpravo: analýza reziduí, ***ADSTAT.***



Obr. 6.4-6 Polynom 6. stupně, vlevo: graf těsnosti proložení, a vpravo: analýza reziduí, ***ADSTAT.***



Obr. 6.4-7 Polynom 7. stupně, vlevo: graf těsnosti proložení, a vpravo: analýza reziduí, ***ADSTAT.***

(b) *Určení odhadů parametrů metodou racionálních hodnostní GPCR:*
Omezení P na vlastní čísla matice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$: Jelikož všechny odhady parametrů v bloku 3 jsou statisticky významné, není třeba hledat novou hodnotu omezení na vlastní čísla P a je možné užít přímo tyto odhady metodou nejmenších čtverců, ($P = 10^{-34}$).

2. Předběžná analýza dat: polohu a proměnlivost proměnných *y, x* charakterizuje *průměr* a *směrodatná odchylka* hodnot každé proměnné. *Pearsonův párový korelační koeficient* všech párů *y* vs. *x, y* vs. x^2, *y* vs. x^3, *y* vs. x^4, *y* vs. x^5, *y* vs. x^6 ukazuje vysokou korelaci. Velká korelace je indikována i mezi jednotlivými nezávisle proměnnými, což je charakteristickým rysem polynomů, protože mezi x^j a x^k je zřejmá vazba.

Proměnná	**Průměr**	**Směrodatná odchylka**	**Párový korelační koeficient**	**Spočtená hlad. význam.**
y	1.9906E+02	3.4199E+01	1.0000	-----
x^1	3.2500E+01	1.5138E+01	0.9703	0.000
x^2	1.2625E+03	1.0024E+03	0.9970	0.000
x^3	5.4438E+04	5.6724E+04	0.9903	0.000

x^4	2.4983E+06	3.1062E+06	0.9682	0.000
x^5	1.1934E+08	1.6883E+08	0.9405	0.000
x^6	5.8593E+09	9.1697E+09	0.9117	0.000

Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných		**Spočtená hladina významnosti**
x^1 versus x^2:	9.8159E-01	0.000
x^1 versus x^3:	9.4340E-01	0.000
x^1 versus x^4:	9.0123E-01	0.000
x^1 versus x^5:	8.6118E-01	0.001
x^1 versus x^6:	8.2500E-01	0.003
x^2 versus x^3:	9.8890E-01	0.000
x^2 versus x^4:	9.6530E-01	0.000
x^2 versus x^5:	9.3774E-01	0.000
x^2 versus x^6:	9.0996E-01	0.000
x^3 versus x^4:	9.9316E-01	0.000
x^3 versus x^5:	9.7795E-01	0.000
x^3 versus x^6:	9.5925E-01	0.000
x^4 versus x^5:	9.9554E-01	0.000
x^4 versus x^6:	9.8522E-01	0.000
x^5 versus x^6:	9.9693E-01	0.000

INDIKACE MULTIKOLINEARITY:

Č	**Vlastní čísla**	**Čísla podmí-**	**Variance inflation**	**Vícenás. korel.**
j	**korel. matice λ_j**	**nění K_j**	**factor VIF_j**	**koef pro X_j**
1	1.8216E-09	3.1529E+09	4.3737E+05	1.0000
2	7.4534E-07	7.7056E+06	1.8050E+07	1.0000
3	1.0269E-04	5.5929E+04	1.3400E+08	1.0000
4	6.7786E-03	8.4726E+02	2.6046E+08	1.0000
5	2.4987E-01	2.2985E+01	1.2751E+08	1.0000
6	5.7432E+00	1.0000E+00	9.8657E+06	1.0000

Maximální číslo podmíněnosti *K* : 3.1529E+09

Nápověda: *K[j], K > 1000 indikuje silnou multikolinearitu, VIF[j] > 10 indikuje silnou multikolinearitu.*

Maximální číslo podmíněnosti K = 3.1529E+09 vysoko převyšuje hodnotu 1000, a proto je v datech indikována silná multikolinearita. Rovněž i kritérium *VIF* dosahuje hodnot vyšších než 10, což potvrzuje v datech silnou multikolinearitu.

3. Odhadování parametrů: klasickou metodou nejmenších čtverců (MNČ) byly nalezeny nejlepší odhady sedmi parametrů, β_0 až β_6. Studentův *t*-test ukázal, že všechny parametry jsou statisticky významné, když $t_{0.95}(10\text{-}7) = 3.182$.

Parametr	**Odhad**	**Směrodatná odchylka**	**H_0: $b_j = 0$ vs. H_A: b_j ... 0 t-kritérium**	**hypotéza H_0 je**	**Spočtená hlad. význ.**
b_0	2.2938E+02	1.2016E+01	19.090	Zamítnuta	0.000
b_1	-1.9126E+01	3.0407E+00	-6.2900	Zamítnuta	0.008
b_2	2.0945E+00	2.9499E-01	7.1002	Zamítnuta	0.006
b_3	-1.1157E-01	1.4204E-02	-7.8552	Zamítnuta	0.004
b_4	3.1315E-03	3.6162E-04	8.6597	Zamítnuta	0.003
b_5	-4.3280E-05	4.6552E-06	-9.2970	Zamítnuta	0.003
b_6	2.3244E-07	2.3841E-08	9.7496	Zamítnuta	0.002

4. Základní statistické charakteristiky: *Střední kvadratická chyba predikce MEP* a *Akaikovo informační kritérium AIC* se užívají k rozlišení mezi několika navrženými

modely, zde stupni polynomu *m*. Pro nejlepší stupeň polynomu *m* dosahují obě kritéria svých minimálních hodnot.

Vícenásobný korelační koeficient *r*	: 0.99999
Koeficient determinace 100 % *D*	: 99.999
Predikovaný koeficient determinace R^2_P	: 0.99863
Střední kvadratická chyba predikce *MEP*	: 2.8832
Akaikovo informační kritérium *AIC*	:-29.367

5. Regresní diagnostika: obsahuje pomůcky a postupy pro interaktivní analýzu (a) dat, (b) modelu, (c) metody, což jsou složky tzv. *regresního tripletu.*

A. Kritika dat: Skládá se z analýzy několika druhů grafických diagnostik a tabulek různých druhů reziduí. Je třeba si uvědomit, že o vhodnosti modelu rozhoduje především statistická analýza reziduí (obr. 6.4-1a až 6.4-7a).

(a) *Analýza klasických reziduí* není příliš spolehlivá, protože klasická rezidua jsou korelovaná, s nekonstantním rozptylem, jeví se normálnější než náhodné chyby (*efekt supernormality)* a nemusí indikovat silně odlehlé hodnoty. Grafická analýza $\hat{e}$ vs. $\hat{y}_P$ (obr. 6.4-1b až 6.4-7b) je schopna indikovat podezřelé body, trend a nekonstantnost rozptylu, tzv. heteroskedasticitu. Míry polohy a rozptýlení klasických reziduí by měly dosahovat hodnot blízkých experimentálnímu šumu. *Odhad směrodatné odchylky s(e)* by se měl blížit svou velikostí experimentální chybě, kterou je zatížena závisle proměnná. Odhady šikmosti a špičatosti by měly dokazovat normální rozdělení reziduí, normalitu.

Bod	**Měřená hodnota**	**Predikovaná hodnota**	**Směrodatná odchylka**	**Klasické reziduum**	**Relativní reziduum**
i	$y_{exp,i}$	$y_{vyp,i}$	$s(y_{vyp,i})$	e_i	$e_{r,i}$
1	1.6320E+02	1.6322E+02	2.0848E-01	-1.5771E-02	-9.6636E-03
2	1.6560E+02	1.6551E+02	1.9869E-01	9.1293E-02	5.5129E-02
3	1.6930E+02	1.6950E+02	1.5902E-01	-2.0217E-01	-1.1942E-01
4	1.7450E+02	1.7432E+02	1.5061E-01	1.8356E-01	1.0519E-01
5	1.8250E+02	1.8248E+02	1.4727E-01	1.8601E-02	1.0192E-02
6	1.9530E+02	1.9547E+02	1.4727E-01	-1.7427E-01	-8.9230E-02
7	2.1200E+02	2.1187E+02	1.5061E-01	1.2776E-01	6.0263E-02
8	2.2810E+02	2.2812E+02	1.5902E-01	-2.0205E-02	-8.8578E-03
9	2.4190E+02	2.4191E+02	1.9869E-01	-1.3461E-02	-5.5646E-03
10	2.5820E+02	2.5820E+02	2.0848E-01	4.6563E-03	1.8034E-03

Reziduální součet čtverců *RSC*	: 1.3080E-01
Průměr absolutních hodnot reziduí *Me*	: 8.5175E-02
Průměr relativních reziduí Me_r	: 4.6531E-02
Odhad reziduálního rozptylu $s^2(e)$	: 4.3600E-02
Odhad směrodatné odchylky reziduí *s(e)*	: 2.0881E-01
Odhad šikmosti reziduí $g_1(e)$	:-0.303
Odhad špičatosti reziduí $g_2(e)$	: 2.376

(b) ***Analýza ostatních reziduí:*** Jackknife rezidua indikují odlehlé body, diagonální prvky H_{ii} od projekční matice $\boldsymbol{H}$ a diagonální prvky $H_{mi,i}$ od zobecněné projekční matice $\boldsymbol{H_m}$, pouze extrémy. Ostatní druhy reziduí a kritéria v tabulce pak obojí (značeno hvězdičkou u hodnoty). Jackknife rezidua $e_{J,i}$ ukazují, že žádný bod není odlehlý. Cookova vzdálenost D_i a Atkinsova vzdálenost A_i ukazují na vlivné body č. 1, 2, 10 a kritérium DF_i na č. 1, 2, 3, 10 a věrohodnostní vzdálenosti $LD(b)_i$ na č. 1, 2, 10 a $LD(s^2)_i$ na č. 1, 2, 3 a $LD(b, s^2)_i$ na č. 1,

2, 3, 10. Diagonální prvky H_{ii} projekční matice $\boldsymbol{H}$ ukazují, že nejsou extrémy, a diagonální prvky $H_{mi,\,i}$ zobecněné projekční matice $\boldsymbol{H_m}$ pak opět, že nejsou extrémy.

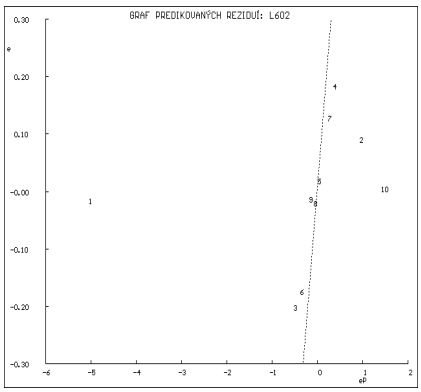
INDIKACE VLIVNÝCH BODU: (* indikuje odlehlý nebo vlivný bod)

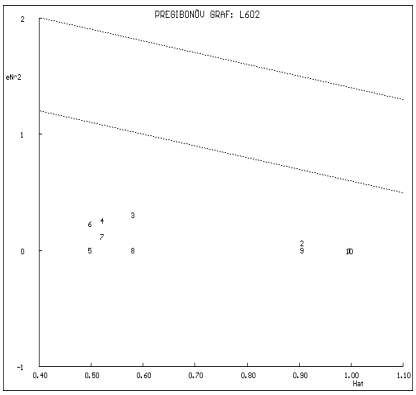
Bod	**Standardiz. reziduum**	**Jackknife reziduum**	**Predikované reziduum**	**Diagonální prvky**
i	$e_{S,i}$	$e_{J,i}$	$e_{P,i}$	$H_{i,i}$
1	-1.3464E+00	-1.7476E+00	-5.0116E+00	9.9685E-01
2	1.4221E+00	2.0340E+00	9.6584E-01	9.0548E-01
3	-1.4939E+00	-2.4105E+00	-4.8131E-01	5.7995E-01
4	1.2693E+00	1.5231E+00	3.8265E-01	5.2028E-01
5	1.2566E-01	1.0287E-01	3.7013E-02	4.9744E-01
6	-1.1773E+00	-1.3105E+00	-3.4675E-01	4.9744E-01
7	8.8338E-01	8.3854E-01	2.6632E-01	5.2028E-01
8	-1.4930E-01	-1.2236E-01	-4.8101E-02	5.7995E-01
9	-2.0968E-01	-1.7247E-01	-1.4241E-01	9.0548E-01
10	3.9752E-01	3.3347E-01	1.4796E+00	9.9685E-01

Bod	**Zobecněné diag. prvky**	**Cookova vzdálenost**	**Atkinsonova vzdálenost**	**Vliv na predikci**
i	H_{mii}	D_i	A_i	DF_i
1	9.9875E-01	8.2037E+01*	2.0362E+01*	-3.1103E+01*
2	9.6920E-01	2.7676E+00*	4.1213E+00*	6.2954E+00*
3	8.9244E-01	4.4020E-01*	1.8542E+00*	-2.8324E+00*
4	7.7790E-01	2.4961E-01	1.0384E+00	1.5862E+00
5	5.0008E-01	2.2328E-03	6.7002E-02	1.0235E-01
6	7.2961E-01	1.9597E-01	8.5352E-01	-1.3038E+00
7	6.4507E-01	1.2091E-01	5.7169E-01	8.7327E-01
8	5.8307E-01	4.3966E-03	9.4123E-02	-1.4378E-01
9	9.0686E-01	6.0168E-02	3.4946E-01	-5.3382E-01
10	9.9702E-01	7.1509E+00*	3.8855E+00*	5.9352E+00*

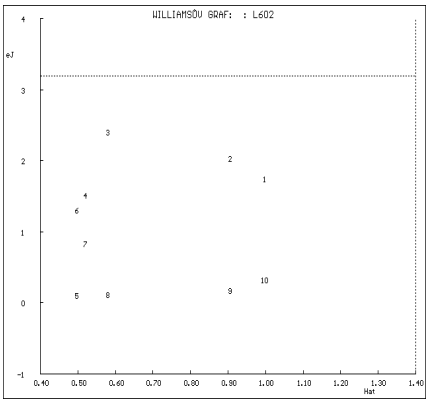
Bod	**V ě r o h o d n o s t n í v z d á l e n o s t i**		
i	$LD(b)_i$	$LD(s^2)_i$	$LD(b, s^2)_i$
1	5.2597E+01*	4.5261E+00*	4.3579E+03*
2	2.0093E+01*	7.4589E+00*	1.8581E+02*
3	7.0662E+00	1.2577E+01*	4.8678E+01*
4	4.5895E+00	2.7920E+00	1.4114E+01
5	5.1964E-02	4.8453E-02	9.5590E-02
6	3.7656E+00	1.5830E+00	9.2322E+00
7	2.4851E+00	2.0507E-01	3.6368E+00
8	1.0207E-01	4.6398E-02	1.3942E-01
9	1.3137E+00	3.9827E-02	1.3221E+00
10	2.8727E+01*	1.2911E-02	1.5853E+02*

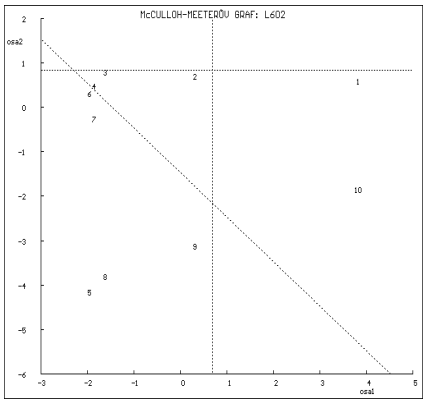
(c) ***Grafy vlivných bodů*** (obr. 6.4-8) jsou schopny indikovat a současně i testovat, dokazovat přítomnost odlehlých hodnot a extrémů.



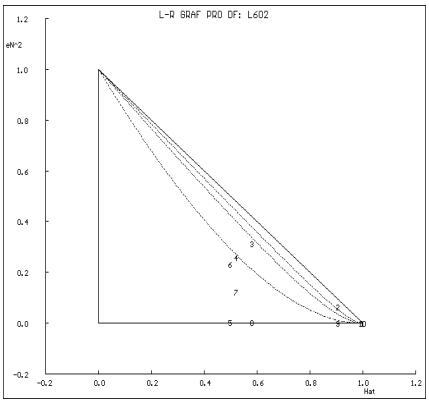


Obr. 6.4-8 Grafy vlivných bodů pro m = 6, vlevo: graf predikovaných reziduí, a vpravo: Pregibonův graf, ***ADSTAT.***





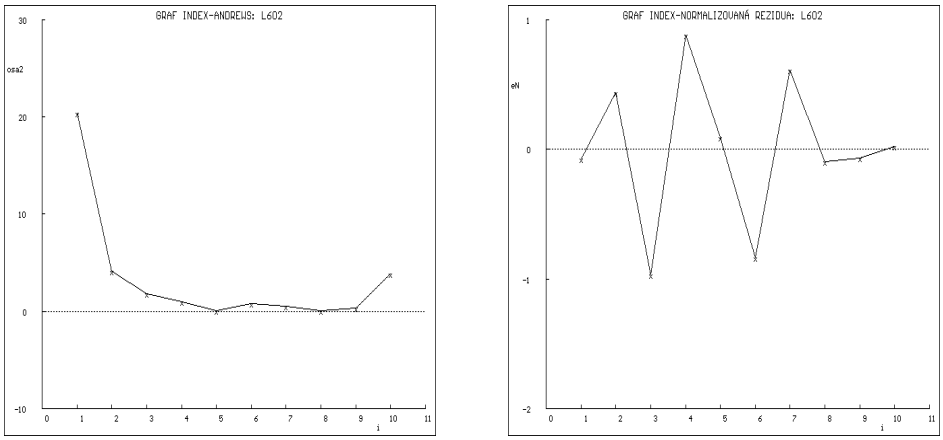
Obr. 6.4-8 Grafy vlivných bodů pro m = 6, vlevo: Williamsův graf, a vpravo: McCullohův-Meeterův graf, ***ADSTAT.***



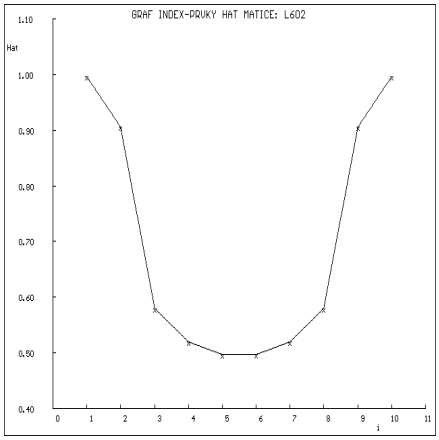
Obr. 6.4-8 Grafy vlivných bodů pro m = 6, L-R graf, ***ADSTAT.***

Grafy predikovaných reziduí neindikují žádný odlehlý bod, ale extrém č. 1, 2, 10. *Pregibonův graf* neukazuje na žádný vlivný bod. *Williamsův graf* neindikuje nějaký odlehlý bod a extrém. *McCullohův-Meeterův graf* nedokazuje odlehlý bod avšak extrémy č. 1 a 10. Konečně *L-R graf* dokazuje odlehlý bod č. 3 a současně extrémy č. 2, 9. Lze uzavřít, že žádný bod není většinou diagnostik prokázán za odlehlý.

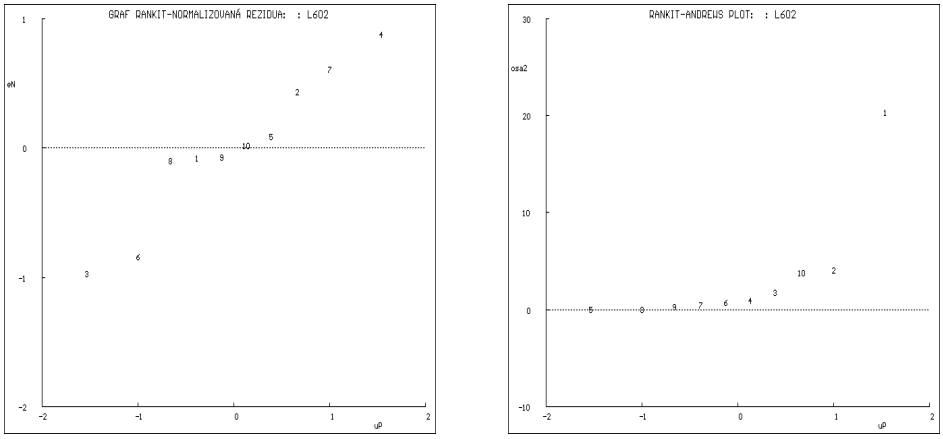
(d) ***Indexové grafy*** (obr. 6.4-9) upozorňují pouze na *podezřelé body*, nejsou schopny testovat odlehlé body a extrémy. *Andrewsův indexový graf* ukazuje na podezřelé body č. 1 a 10. *Indexový graf prvků projekční matice* ***H*** pak na podezřelé extrémy č. 1, 2, 9, 10.



Obr. 6.4-9 Indexové grafy pro $m = 6$, vlevo: Andrewsův graf, a vpravo: graf normovaných reziduí, ***ADSTAT.***

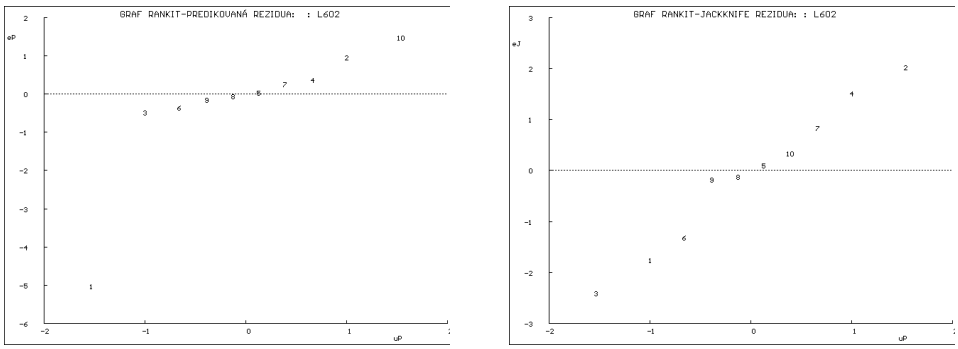


Obr. 6.4-9 Graf prvků projekční matice ***H***, ***ADSTAT.***



Obr. 6.4-10 Rankitové grafy pro $m = 6$, vlevo: graf normovaných reziduí, a vpravo: Andrewsův graf, ***ADSTAT.***

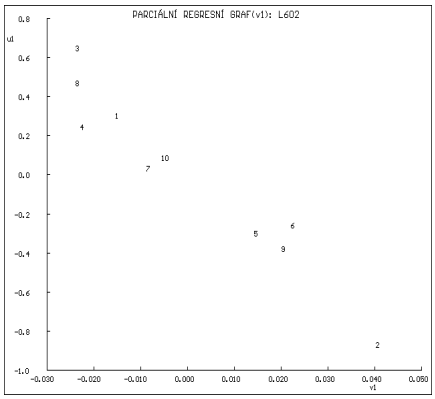
(e) ***Rankitové grafy*** (obr. 6.4-10) ukazují vedle normality rozdělení dotyčných reziduí i na vlivné (zde odlehlé) body. *Graf normovaných reziduí* neukazuje na odlehlé body. *Andrewsův graf* ukazuje na č. 1 jako na odlehlý bod. *Rankitový graf predikovaných reziduí* ukazuje na jeden odlehlý bod, a to na č. 1 a *graf Jackknife reziduí* pak na č. 1 jako na odlehlý bod.



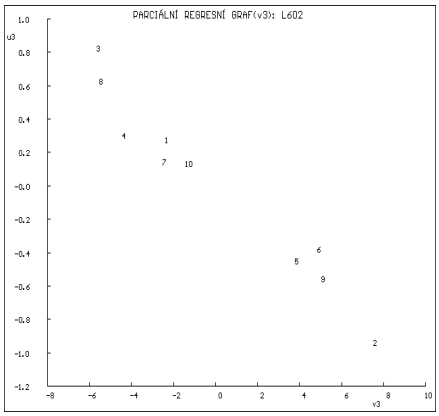
Obr. 6.4-10 Rankitové grafy, vlevo: graf predikovaných reziduí, a vpravo: graf Jackknife reziduí, ***ADSTAT.***

B. Model: *Parciální regresní grafy* (obr. 6.4-11) ale především *parciální reziduální grafy* (obr. 6.4-12) ukazují na lineární závislost všech nezávisle proměnných. Navržený model je formulován ve tvaru

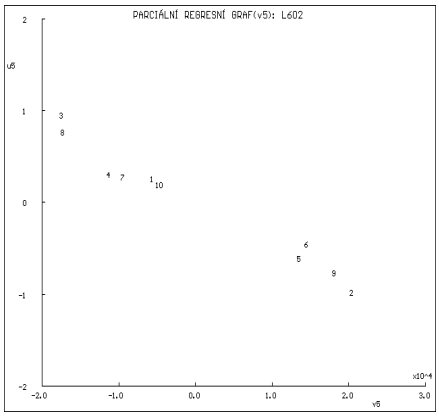
$$\beta_0 \,\%\, \beta_1 x \,\%\, \beta_2 x^2 \,\%\, \beta_3 x^3 \,\%\, \beta_4 x^4 \,\%\, \beta_5 x^5 \,\%\, \beta_6 x^6 .$$



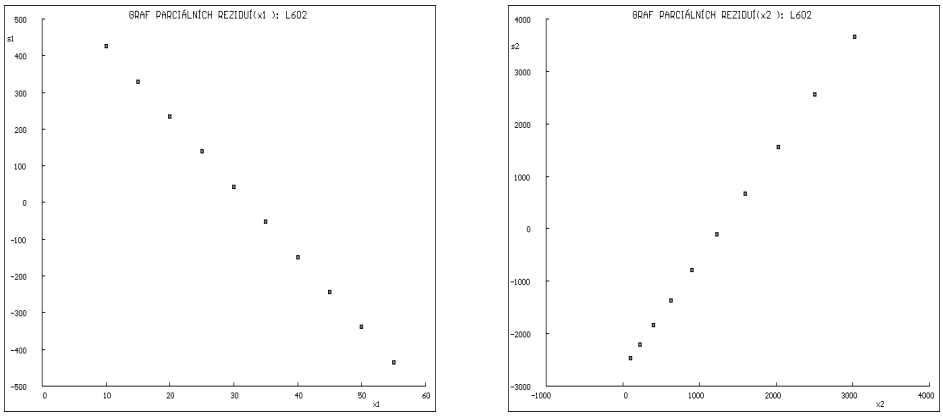
Obr. 6.4-11 Parciální regresní grafy, vlevo: pro proměnnou x, vpravo: pro proměnnou x^2.



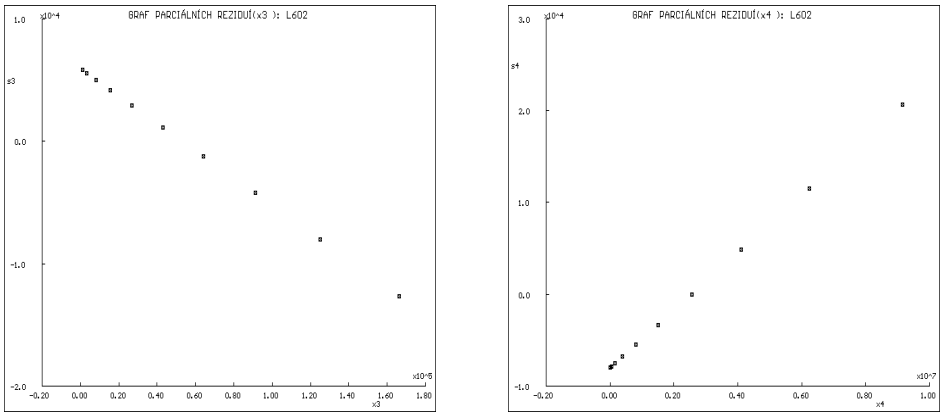
Obr. 6.4-11 Parciální regresní grafy, vlevo: pro proměnnou x^3, vpravo: pro proměnnou x^4.



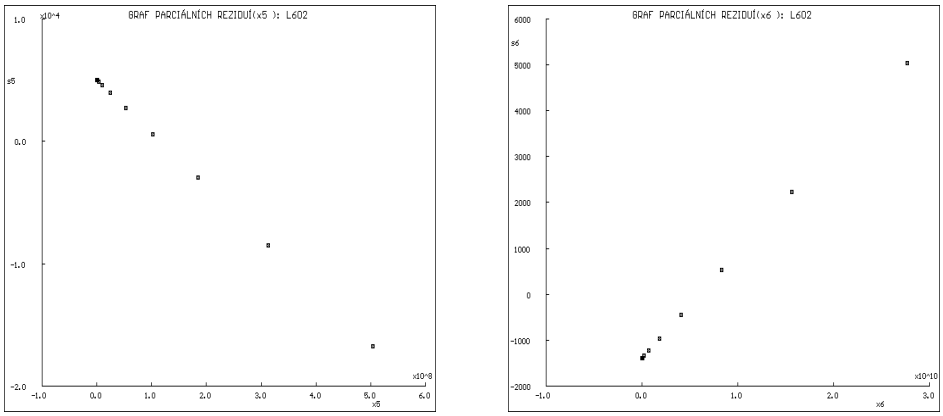
Obr. 6.4-11 Parciální regresní grafy, vlevo: pro proměnnou x^5, vpravo: pro proměnnou x^6.



Obr. 6.4-12 Parciální reziduální grafy, vlevo: pro proměnnou x, vpravo: pro proměnnou x^2.



Obr. 6.4-12 Parciální reziduální grafy, vlevo: pro proměnnou x^3, vpravo: pro proměnnou x^4.



Obr. 6.4-12 Parciální reziduální grafy, vlevo: pro proměnnou x^5, vpravo: pro proměnnou x^6.

C. Metoda: Do této části patří vyšetření splnění základních předpokladů metody

nejmenších čtverců (MNČ), za kterých by měla metoda vést k nejlepším nestranným odhadům lineárních regresních parametrů:

Fisherův-Snedecorův test významnosti regrese potvrdil, že navržený model je přijat jako významný, jinými slovy: závisle proměnná *y* a všechny nezávisle proměnné jsou v lineární závislosti.

Scottovo kritérium multikolinearity ukazuje, že navržený model je korektní.

Cookův-Weisbergův test heteroskedasticity dokazuje, že rezidua vykazují heteroskedasticitu (nekonstantnost rozptylu).

Jarqueův-Berraův test normality reziduí ukazuje, že klasická rezidua nevykazují Gaussovo rozdělení.

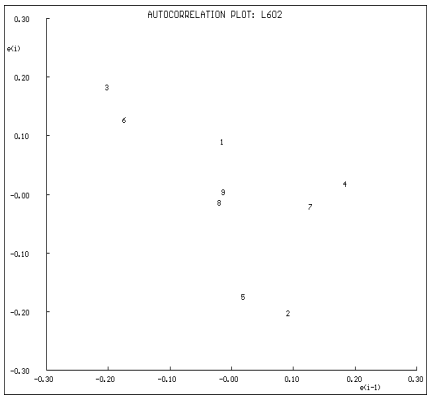
Waldův test autokorelace ukazuje, že klasická rezidua jsou autokorelována.

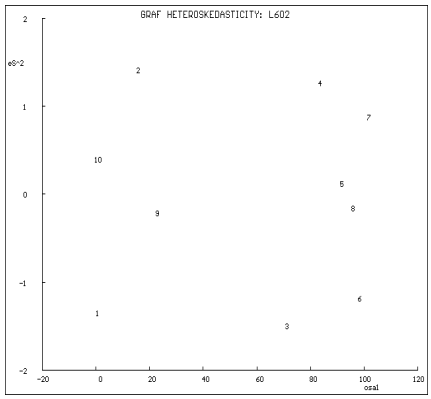
Znaménkový test prokazuje, že znaménko klasických reziduí se dostatečně střídá, a proto rezidua nevykazují žádný trend.

TESTOVÁNÍ REGRESNÍHO TRIPLETU (DATA + MODEL + METODA):	
Fisherův-Snedocorův test významnosti regrese F_{exp}	: 40238.0
Tabulkový kvantil $F_{1-\alpha}(m\text{-}1,\ n\text{-}m)$	: 8.9406
Závěr: Navržený model je přijat jako významný.	
Spočtená hladina významnosti	: 0.000
Scottovo kritérium multikolinearity M	: 0.9966
Závěr: Navržený model je korektní.	
Cookův-Weisbergův test heteroskedasticity S_f	: 0.3053
Tabulkový kvantil $\chi^2_{1-\alpha}(1)$	: 3.8415
Závěr: Rezidua vykazují heteroskedasticitu.	
Spočtená hladina významnosti	: 0.000
Jarqueův-Berraův test normality reziduí $L(e)$	: 0.3154
Tabulkový kvantil $\chi^2_{1-\alpha}(2)$	: 5.9915
Závěr: Normalita není přijata.	
Spočtená hladina významnosti	: 0.854
Waldův test autokorelace W_a	: 6.3423
Tabulkový kvantil $\chi^2_{1-\alpha}(1)$	: 3.8415
Závěr: Rezidua jsou autokorelována.	
Spočtená hladina významnosti	: 0.012
Znamékový test D_t	: 1.0062
Tabulkový kvantil $N_{1-\alpha/2}$	: 1.6449
Závěr: Rezidua nevykazují trend.	
Spočtená hladina významnosti	: 0.157

Graf autokorelace (obr. 6.4-13) vykazuje náhodný mrak bodů, což znamená, že rezidua autokorelaci nevykazují.

Graf heteroskedasticity (obr. 6.4-13) nevykazuje klín nýbrž elipsovitý mrak, a proto rezidua nevykazují heteroskedasticitu, nekonstantnost rozptylu.





Obr. 6.4-13 Vlevo: Graf autokorelace, a vpravo: graf heteroskedasticity, ***ADSTAT.***

7. Zhodnocení kvality modelu: porovnáním hodnot regresní diagnostiky lze snadno provést zhodnocení *regresního tripletu* dosaženého lineárního regresního modelu pro daná data. Nalezený model polynomické závislosti má tvar (v závorce je vždy uveden odhad směrodatné odchylky parametru):

$$y = 229.4\ (12.0) - 19.1\ (3.0)\ x + 2.09\ (0.29)\ x^2 - 0.11\ (0.01)\ x^3 + \\ + 3.13\ (0.36) \times 10^{-3}\ x^4 - 4.33\ (0.47) \times 10^{-5}\ x^5 + 2.32\ (0.24) \times 10^{-7}\ x^6.$$

6.4.1 Úlohy na polynomické regresní modely

Úloha L6.01 *Závislost hektarového výnosu obiloviny na množství hnojiva*
Zemědělský ústav zkoumal závislost hektarového výnosu určité obiloviny y v [t/ha] na množství hnojiva x, a to ledku amonného v [kg/ha], str. 52 v cit[63]. Bylo zjištěno, že regresní model je tvořen polynomem m-tého stupně. (1) Stanovte stupeň polynomu m. (2) Věnujte zvláštní pozornost multikolinearitě a pokuste se o její snížení. (3) Jsou parametry polynomu, určené metodou nejmenších čtverců MNČ a racionálních hodností GPCR, statisticky významné? (4) Jaký hektarový výnos lze očekávat při hnojení 85 kg/ha a 115 kg/ha?

Data: Množství hnojiva x [kg/ha], hektarový výnos y [t/ha].

40	1.9,	50	2.5,	60	2.9,	65	3.1,	70	3.1,	75	3.3,	80	3.3,		
85	3.5,	90	3.5,	100	3.4,										

Úloha L6.02 *Závislost výšky píku kyseliny kyanurové na koncentraci želatiny*
Při stanovení kyseliny kyanurové metodou diferenční pulsní polarografie byl sledován vliv přítomnosti povrchově aktivních látek. (1) Určete stupeň polynomu m závislosti výšky píku kyseliny kyanurové y na koncentraci želatiny x. (2) Které z kritérií, *MEP* nebo *AIC*, má lepší rozlišovací schopnost při určení stupně polynomu? (3) Pokuste se snížit multikolinearitu.

Data: Koncentrace x [mg/l], výška píku y [mm].

c:	10	15	20	25	30	35	40	45	50	55
h:	163.2	165.6	169.3	174.5	182.5	195.3	212.0	228.1	241.9	258.2

Úloha L6.03 *Závislost rozpustnosti nitrocelulózy na viskozitě*
Byla změřena závislost rozpustnosti nitrocelulózy *y* ve směsi etylalkoholu s éterem na její viskozitě *x*. (1) Určete stupeň polynomu *m* pro popis této závislosti. (2) Je rozdíl v odhadech parametrů, určených metodou nejmenších čtverců MNČ a racionálních hodností GPCR? (3) Jsou v datech vlivné body? (4) Jaká bude rozpustnost pro viskozitu 5.50 a 7.50 [mPa.s]?
Data: Viskozita nitrocelulózy *x* [mPa.s], rozpustnost nitrocelulózy *y* [hmot%].

x:	6.15	5.70	5.20	6.15	5.95	6.40	7.65	5.10	5.00	5.00	5.40	5.70	5.85
	8.60	9.10											
y:	12.15	12.35	11.70	12.10	13.95	11.15	12.50	13.80	15.40	15.00	15.50	15.10	15.05
	10.80	16.95											

Úloha L6.04 *Závislost biologického rozkladu kalu na teplotě*
Je sledována závislost biologického rozkladu kalu, vyjádřená procentem organických látek *y* na teplotě procesu *t* při anaerobní stabilizaci. (1) Určete polynom k popisu této závislosti, vyšetřete regresní triplet. (2) Soustřeďte se především na snížení vlivu multikolinearity. (3) Které z rozlišovacích kritérií k určení stupně polynomu a filtru multikolinearity je citlivější, střední kvadratická chyba predikce *MEP* nebo Akaikovo informační kritérium *AIC*?
Data: Teplota *x* [EC], koncentrace organických látek *y* [%].

t:	20	25	30	35	40	45	50	55
y:	40.19	40.00	39.73	39.63	39.33	39.14	38.75	38.45

Úloha L6.05 *Závislost relativní síly na relativním stlačení ovoce*
Při experimentech se stlačováním ovoce byly naměřeny následující hodnoty relativní síly *y* při relativním stlačení *x*. (1) Určete stupeň polynomu křivkové závislosti. (2) Jsou parametry, určené metodou MNČ nebo GPCR, statisticky významné? (3) Lze vedle kritérií *MEP* a *AIC* užít také predikovaného koeficientu determinace R^2_P?
Data: Relativní stlačení x [%], relativní síla *y* [%].

0.0	0.0,	0.5	0.9,	1.0	1.3,	1.5	2.4,	2.0	3.9,	2.5	5.8,	3.0	8.2,	3.5	11.7,
4.0	17.1,	4.5	22.6,												

Úloha L6.06 *Závislost signálu metody HPLC na koncentraci phenmediphamu*
Využitím střední kvadratické chyby predikce *MEP* určete (1) stupeň polynomu *m* experimentální závislosti signálu HPLC *y* na koncentraci *x* účinné látky phenme-diphamu v přípravku Synbetan P. (2) Jsou všechny parametry modelu statisticky významné? (3) Pokuste se snížit multikolinearitu v modelu.
Data: Koncentrace *x* [mg Ph/skvrna], plocha *y* [jednotky].

0.15537	48528.0,	0.31076	73825.0,	0.46613	85200.0,	0.62150	102660.0,
0.77688	116050.0,	0.93226	126090.0,	1.08760	135590.0,	1.24300	144090.0,

Úloha L6.07 *Závislost vodivostního koeficientu na teplotě*
Je studována aproximace konvexně klesající závislosti vodivostního korekčního koeficientu *y* na teplotě *x* polynomem. Využitím kritérií střední kvadratické chyby *MEP* a Akaikova informačního kritéria *AIC* (1) určete vhodný stupeň polynomu této závislosti, (2) metodou nejmenších čtverců MNČ a metodou racionálních hodností GPCR zjistěte tvar optimálního

regresního modelu. (3) Proveďte testování modelu, test statistické významnosti nalezených parametrů, (4) Proveďte také vyšetření regresního tripletu a multikolinearity. (5) Zdůvodněte užité diagnostiky a statistiky.

Data: Teplota *x* [EC], vodivostní korekční koeficient *y*.

x:	16.0	16.5	17.0	17.5	18.0	18.5	19.0	19.5	20.0	20.5	21.5	22.0
	22.5	23.0	23.5	24.0	24.5	25.0	25.5					
y:	1.101	1.088	1.074	1.061	1.048	1.036	1.024	1.012	1.000	0.989	0.978	0.967
	0.956	0.946	0.936	0.926	0.916	0.906	0.897	0.888				

Úloha L6.08 *Závislost změny analytického signálu na koncentraci HNO_3*

Při stanovení arsenu metodou generování hydridů je zařazován redukční krok As^{5+} W As^{3+}. K této předredukci roztoků je používán L-cystein v prostředí HNO_3. Byla sledována změna velikosti analytického signálu *y* při redukci roztoku obsahujícího 20 mg As^{5+}/l v závislosti na koncentraci HNO_3 *x*. (1) Určete stupeň polynomu této závislosti metodou MNČ a GPCR. (2) Pokuste se snížit multikolinearitu v modelu.

Data: Koncentrace HNO_3 *x* [ml HNO_3/100ml], signál *y*.

x:	0.25	0.5	1	2	4	6	8	10	20	30
y:	0.01	0.019	0.038	0.087	0.150	0.171	0.196	0.232	0.290	0.310

Úloha L6.09 *Fotometrická závislost koncentrace těhotenského hormonu (HCG)*

Pokuste se (1) aproximovat kalibrační závislost absorbance *y* na koncentraci těhotenského hormonu HCG *x* [mIU/l] ve standardních sérech při enzymoimunologickém stanovení HCG polynomem a (2) stanovte jeho stupeň *m*. (3) Jsou v datech nějaké odlehlé body? (4) Pokuste se snížit multikolinearitu v datech. (5) Porovnejte statistickou významnost odhadů parametrů, získaných metodou MNČ a GPCR.

Data: Koncentrace HCG *x* [mIU/l], absorbance *y*.

x:	0.100	0.098	0.099	0.113	13.60	13.60	13.68	200.0	200.0	200.0	0.401	0.401
	0.401	0.426	0.401	0.571	0.571	0.571	0.571	0.571				
y:	0.000	0.000	0.000	0.000	0.018	0.017	0.017	0.101	0.101	0.104	0.826	0.850
	0.742	0.819	0.845	1.252	1.303	1.135	1.297	1.237				

Úloha L6.10 *Závislost teploty tuhnutí nitrobenzenu na obsahu vody*

Byla změřena závislost bodu tuhnutí nitrobenzenu *y* na obsahu vody *x*. (1) Rozhodněte, zda je daná závislost lépe vystižena lineárním nebo kvadratickým modelem. (2) Užijte testu kvadratického členu. (3) Mají nalezené odhady parametrů statistický význam?

Data: Obsah vody *x* [hm%], bod tuhnutí *y* [EC].

0.0410	5.57,	0.0530	5.56,	0.0730	5.50,	0.158	5.39,	0.310	5.31,	0.360	5.28,	0.380	5.25,

Úloha L6.11 *Časová závislost zbarvení anilinu*

Byla naměřena data pro závislost zbarvení anilinu *y* na čase *x*. (1) Určete nejlepší stupeň polynomu *m* této závislosti metodou MNČ, resp. GPCR. (2) Testujte statistickou významnost jednotlivých parametrů polynomu a pokuste se snížit multikolinearitu v modelu.

Data: Čas *x* [h], zbarvení *y* [jednotky Hasena].

0	42,	17.5	51,	22	56,	65	54,	89	95,	113	123,	137	121,	163	183,	

233	355,	259	444,								

Úloha L6.12 *Rozlišení mezi dvěma modely u časové závislosti zbarvení anilinu*
Proveďte (1) rozlišení mezi lineárním a kvadratickým modelem pro časovou závislost zbarvení anilinu $y = f(x)$ při jinak konstantních fyzikálních podmínkách. (2) Vyšetřete, zda naměřená závislost je lépe vystižena lineárním nebo kvadratických modelem.
Data: Čas x [s], zbarvení y [jednotky Hasena].

0.0	42.0,	17.5	51.0,	22.0	56.0,	65.0	54.0,	89.0	95.0,	113.0	123.0,
137.0	121.0,	163.0	183.0,	233.0	355.0,	259.0	444.0,				

Úloha L6.13 *Závislost obsahu berylia na obsahu popela v uhlí*
Při studiu geochemické pozice beryllia v uhlí byla zjištěna závislost obsahu beryllia y na obsahu popela x v uhlí v procentech. (1) Metodou nejmenších čtverců MNČ a metodou racionálních hodností GPCR najděte parametry optimálního polynomického modelu pro ln y v závislosti na obsahu popela x. (2) Vyšetřete velikost multikolinearity a případně ji i odstraňte.
Data: Obsah popela x [%], obsah beryllia y [%].

4.58	2.22,	3.35	2.53,	2.50	2.80,	3.24	2.51,	5.13	2.08,	8.18	1.61,
5.57	2.02,	29.6	0.33,	17.4	0.83,	14.5	0.72,	3.78	2.36,	6.67	1.84,
8.73	1.55,	2.19	2.90,	1.83	3.22,	2.30	2.86,	21.39	0.65,	5.30	2.14,
1.58	3.35										

Úloha L6.14 *Závislost koncentrace nukleových kyselin v séru*
Při vypracovávání HPLC metody stanovení nukleových kyselin v séru byly propočítány kapacitní faktory standardních látek, třeba inosinu y v závislosti na různém složení mobilní fáze, např. na procentu methanolu v mobilní fázi x při pH 5.5. (1) Metodou nejmenších čtverců MNČ a metodou racionálních hodností GPCR najděte vhodný stupeň polynomu m. (2) Vyšetřete i statistickou významnost jednotlivých parametrů.
Data: Koncentrace methanolu v mobilní fázi x [%], kapacitní faktor inosinu y.

50.0	9.02,	45.0	12.3,	40.0	19.0,	35.0	36.6,	30.0	58.7,	25.0	110.0,
20.0	180.0										

Úloha L6.15 *Závislost výtěžku epoxidace alkenů na výtěžku syntézy katalyzátoru*
Při výzkumu epoxidace alkenů peroxidem vodíku pomocí katalyzátorů titanových silikátů byl vyšetřován katalytický účinek na epoxidaci alkylchloridu peroxidem vodíku. Byly hledány závislosti aktivity epoxidačních katalyzátorů. Nezávisle proměnnou x je množství katalyzátoru v hmotnostních procentech a závisle proměnnou y je výtěžek epoxidační reakce. (1) Lze přijmout předpoklad, že hodnoty nezávisle proměnné byly podstatně přesnější než hodnoty závisle proměnné, při jejichž získávání se projevovalo více náhodných vlivů. (2) Stanovte stupeň polynomu m a odstraňte případnou multikolinearitu.
Data: Množství x [hm. %], výtěžek epoxidační reakce y [hm. %].

87.89	80.7,	28.13	35.6,	45.81	20.1,	74.67	68.4,	109.01	82.5,	106.06	81.0,
74.25	66.86,	113.91	82.1,	97.1	92.1,	122.46	90.0,	91.35	81.0,	116.1	81.9,
61.05	83.2										

Úloha L6.16 *Závislost přírustku investic v USA v průběhu let 1920 - 1941*
Pro data představující ukazatele "přírůstku investic v USA v miliardách US $ *y* v cenové hladině roku 1934" v závislosti na letech *x* byl v literatuře původně navržen kvadratický model. (1) Regresní analýzou vyšetřete, zda by datům lépe vyhovoval polynom vyššího stupně. (2) Porovnejte statistickou významnost odhadů parametrů, získaných metodou nejmenších čtverců MNČ a metodou racionálních hodností GPCR.
Data: Roky *x* [roky], přírůstek investic *y* [miliardy US $].

1920	2.70,	1921	-0.20,	1922	1.90,	1923	5.20,	1924	3.00,	1925	5.10,
...	...	...	...	...	...	...	...	...	...	...	...
1938	-1.90,	1939	1.30,	1940	3.30,	1941	4.90,				

Úloha L6.17 *Závislost relativní fluorescence na obsahu teofylinu*
Při stanovení teofylinu pomocí soupravy TDA Ames byly při měření kalibrátorů *x* zjištěny hodnoty relativní fluorescence *y*. (1) Stanovte stupeň polynomu *m* této závislosti. (2) Pokuste se snížit vliv multikolinearity. (3) Porovnejte statistickou významnost odhadů parametrů, získaných metodou nejmenších čtverců MNČ a metodou racionálních hodností GPCR.
Data: Koncentrace kalibrátorů *x* [mg/l], relativní fluoroscence *y* [%].

90.0	124.7,	80.0	116.3,	70.0	109.2,	60.0	101.3,	50.0	90.5,	40.0	83.0,
30.0	68.8,	20.0	51.8,	10.0	25.1,	0.0	0.0,				

Úloha L6.18 *Závislost elektrochemické účinnosti akumulátorové hmoty NICOS*
U akumulátorové hmoty NICOS byla sledovávána elektrochemická účinnost *y* v závislosti na obsahu kobaltu ve hmotě *x*. (1) Stanovte stupeň polynomu *m* této závislosti. (2) Pokuste se snížit vliv multikolinearity. (3) Porovnejte statistickou významnost odhadů parametrů, získaných metodou nejmenších čtverců MNČ a metodou racionálních hodností GPCR.
Data: Obsah kobaltu *x* [ppm], elektrochemická účinnost *y* [%].

182.1	1.97,	182.6	1.98,	179.8	2.02,	180.2	2.07,	180.1	2.04,	179.3	2.11,
178.6	2.13,	177.4	2.03,	176.8	2.15,	177.1	2.09,	176.6	2.08,	176.2	2.08,
175.0	2.20,	179.7	2.10,								

Úloha L6.19 *Závislost hmotnosti výrobků na hodině probíhající směny*
V průběhu jedné směny byla sledována kvalita výrobku záznamem řady faktorů, jedním z nich byla i hmotnost *y*. Vzorky pěti výrobků byly odebírány v každou celou hodinu a jejich hmotnost *y* sledována v závislosti na čase *x*. (1) Zjistěte, je-li hmotnost výrobků konstantní, či zda je v ní nějaký trend. (2) Dá se případný trend vystihnout nějakou závislostí, např. polynomickou? (3) Stanovte stupeň polynomu *m* této závislosti. (4) Pokuste se snížit vliv multikolinearity. (5) Porovnejte statistickou významnost odhadů parametrů, získaných metodou nejmenších čtverců MNČ a metodou racionálních hodností GPCR.
Data: Čas v průběhu směny *x* [h], hmotnost výrobku [g] *y*.

7.00	18.70,	8.00	18.88,	9.00	18.79,	10.0	18.46,	11.0	18.85,	12.0	18.45,
13.0	18.41,	14.0	17.61,	15.0	18.26,	16.0	18.28				

Úloha L6.20 *Závislost měrné vodivosti telluridu bismutitého na teplotě*
Nalezněte polynomický model pro popis závislosti měrné vodivosti telluridu bismutitého

Bi_2Te_3 y [Ω^{-1}. cm^{-1}] na teplotě *x* [K]. (1) Určete stupeň polynomu *m* metodami MNČ a GPCR. (2) Porovnejte statistickou významnost odhadů parametrů, získaných oběma metodami.

Data: Teplota *x* [K], měrná vodivost Bi_2Te_3 *y* [Ω^{-1}. cm^{-1}].

5.00	157.0,	10.0	279.0,	15.0	428.0,	20.0	533.0,	25.0	723.0,
30.0	778.0,	40.0	1008.0,	50.0	1251.0,	60.0	1445.0,	70.0	1653.0,
90.0	2009.0								

Úloha L6.21 *Závislost koncentrace makroglobulinu v krvi zdravých žen na věku*

Nalezněte polynomický regresní model závislosti koncentrace α-2-makro-globulinu A2M *y* v krvi zdravých žen na jejich věku *x*, a to od 11. do 78. roku. (1) Testujte statistickou významnost jednotlivých parametrů, určených jak metodou nejmenších čtverců MNČ, tak i metodou racionálních hodností GPCR. (2) Jsou v souboru dat vlivné body a odlehlé hodnoty?

Data: Věk *x* [rok], koncentrace α-2-makro-globulinu A2M v krvi zdravých žen *y*.

11.0	3.25,	12.0	3.10,	13.0	3.25,	14.0	3.10,	15.0	2.92,	16.0	2.48,
...	...	...	...	...	...	...	...	...	...	...	...
71.0	2.65,	72.0	2.04,	73.0	2.70,	78.0	2.20,				

Úloha L6.22 *Korekce hustoty vody na neideální chování*

Při měření hustoty *y* směsi isopropylamin + diisopropylamin + isopropylalkohol + voda je nutno provést, vzhledem k neideálnímu chování vody, korekci na molární zlomek vody *x*. (1) Určete stupeň polynomu *m* metodou nejmenších čtverců a racionálních hodností. (2) Vyšetřete statistickou významnost jednotlivých parametrů.

Data: Molární zlomek vody *x* a korekce hustoty *y*.

0.01420	8.70,	0.05340	21.70,	0.09540	27.80,	0.1570	34.20,	0.1874	36.10,
0.3177	46.10,	0.5430	49.30,						

Úloha L6.23 *Teplotní závislost křivky rázové houževnatosti polyolefinu*

Pro automobilku Škoda-Volkswagen je dodáván polyolefin, u kterého je požadována vysoká hodnota rázové houževnatosti v širokém rozsahu teplot. Při vývoji tohoto materiálu byla stanovena tzv. křivka rázové houževnatosti *y* měřením těchto hodnot v rozsahu teplot *x* od -60E do +23EC.

Data: Teplota *x* [EC], rázová houževnatost *y* [kJ/m^2].

23.0	60.8,	20.0	33.7,	10.0	14.5,	0.0	10.1,	-10.0	7.50,	-20.0	5.80,
-30.0	4.80,	-40.0	3.80,	-50.0	3.10,	-60.0	2.80,				

Úloha L6.24 *Určení stupně polynomu závislosti podílu vadných výrobků*

Při výrobě jistého elektronického výrobku se od pracovníků vyžaduje rychlost a vysoká přesnost. Výstupní kontrola zjistila vysoký podíl vadných výrobků. Byl proveden průzkum, jehož cílem bylo popsat průběh závislosti procenta vadných výrobků *y* na výkonu za směnu *x*. Tabulka obsahuje údaje o výkonu za směnu *x* a procentu vadných výrobků *y* u 20 náhodně vybraných pracovníků. (1) Určete optimální stupeň polynomu *m* a s ohledem na predikční schopnost modelu a metodou racionálních hodností proveďte odhad parametrů.

(2) Testujte statistický význam odhadů parametrů.

Data: Výkon za směnu *x* [ks], podíl vadných výrobků *y* [%].

84	2.8,	86	2.2,	68	4.5,	50	6.0,	75	3.5,	88	1.8,	142	3.2,	132	2.4,	123	2.8,	93	2.3,
107	1.3,	114	2.2,	138	2.8,	98	1.7,	56	6.7,	104	1.7,	79	3.5,	126	1.9,	57	4.9,	130	2.2,

Úloha L6.25 *Polynomická závislost koncentrace beryllia na obsahu popela*

Při studiu geochemické pozice beryllia v uhlí byla zjištěna závislost koncentrace beryllia *y* na obsahu popela v uhlí *x*. (1) Určením stupně polynomu MNČ a metodou GPCR najděte regresní model závislosti ln [Be] na obsahu popela *x*. Které kritérium má lepší rozlišovací schopnost, *MEP* nebo *AIC*?

Data: Obsah popela *x* [%], a koncentrace beryllia představuje ln [Be] je *y*.

x:	4.6	3.4	2.5	3.2	5.1
y:	2.2	2.5	2.8	2.5	2.1
...	...	...	...	...	...
x:	2.3	21.4	5.3	1.6	
y:	2.9	0.7	2.1	3.4	

Úloha L6.26 *Určení stupně polynomu u kinetické závislosti*

Při kinetickém pokusu reakce kyseliny bromaminové s anilinem za podmínek pseudoprvního řádu byl očekáván přibližně přímkový průběh závislosti koncentrace reakčního produktu *y* na čase *x*. Po provedení experimentu byl však zjištěn nelineární průběh, u něhož byl učiněn pokus o aproximaci polynomem vhodného stupně. (1) Určete stupeň polynomu *m*. (2) Kterou modifikaci metody MNČ užijete k odhadu parametrů polynomu?

Data: Čas *x* [s], koncentrace reakčního produktu *y* [mg/l].

0.0	2.5	285.0	2.6	568.0	2.8
...	...	...	...	...	...
3623.0	3.6				

Úloha L6.27 *Určení stupně polynomu při popisu spektra*

Pro stanovení amonných iontů ve vzorcích vod se používá spektrofotometrická metoda, využívající reakce amoniaku se salicylanem sodným a chlorem uvolňovaným z dichloroizokyanaturanu sodného za vzniku indofenolového barviva. Pro zjištění tvaru regresního modelu, popisujícího záznam spektra v oblasti spektrofotometrického maxima, tj. závislosti absorbance *y* na vlnové délce *x* [nm], bylo použito polynomu. (1) Určete stupeň polynomu *m* a (2) Použijte metodu racionálních hodnostní a určete optimální parametr *P*. Porovnejte odhady parametrů především co do jejich statistické významnosti.

Data: vlnová délka *x* [nm], absorbance *y*.

x:	600	603	606	609	612
y:	0.476	0.500	0.524	0.550	0.576
	...	...	...	...	...
x:	690	693	696	699	
y:	0.845	0.831	0.815	0.800	

Úloha L6.28 *Závislost koncentrace kreatininu v krevním séru na věku dárců*

Nalezněte polynomický regresní model závislosti koncentrace kreatininu v krevním séru *y* na věku mužů, dárců krve *x*, a to od 20. do 75. roku života. (1) Určete stupeň polynomu *m* a vyšetřete multikolinearitu. (2) Které kritérium má lepší rozlišovací schopnost, *MEP* či *AIC*? (3) Jsou odhady parametrů statisticky významné?

Data: Věk *x* [roky], koncentrace kreatininu *y* [μmol/l].

20	66,	50	91,	24	68,	52	93,	26	69,	54	97,	30	72,	57	99,
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
42	79,	70	108,	45	85,	72	115,	47	88,	75	119,				

Úloha L6.29 *Závislost koncentrace cholesterolu v krevním séru na věku dárců*

Nalezněte polynomický regresní model závislosti koncentrace cholesterolu v krevním séru *y* na věku mužů, dárců krve *x*, a to od 20. do 75. roku života. (1) Určete stupeň polynomu *m*. (2) Které kritérium má lepší rozlišovací schopnost, *MEP* či *AIC*? (3) Jsou odhady parametrů statisticky významné?

Data: Věk *x* [roky], koncentrace cholesterolu *y* [mmol/l].

20	3.80,	24	4.14,	26	4.55,	30	5.00,	33	5.60,	35	5.45,	38	5.90,
...	..	...	...	...	...	...	...	...	...	...	...	...	...
75	7.90,												

Úloha L6.30 *Aproximace koncentrační závislosti zinku (AAS) polynomem*

Metodou AAS byly proměřeny hodnoty absorbance *y* pro koncentrace zinku *x* od 0.5 mg/l do 50 mg/l. (1) Určete stupeň polynomu *m* proložením této závislosti metodou MNČ a GPCR, (2) Pokuste se o odstranění multikolinearity. (3) Proveďte testování statistické významnosti nalezených parametrů.

Data: Koncentrace Zn *x* [mg/l], absorbance *y*.

x:	0.5	1.0	2.0	15.0	25.0	30.0	35.0	40.0	50.0
y:	0.017	0.033	0.064	0.417	0.591	0.630	0.677	0.706	0.742

Úloha L6.31 *Určení stupně polynomu útlumové charakteristiky*

Na generátoru s vnitřní impedancí 600 ohmů byly změřeny útlumové charakteristiky dvou propustí zapojených vedle sebe. Byl změřen útlum *y* [dB] v závislosti na kmitočtu *x* [Hz]. (1) Stanovte optimální stupeň polynomu *m*. (2) Pokuste se snížit multikolinearitu. (3) Porovnejte odhady parametrů, získaných metodami MNČ a GPCR.

Data: Kmitočet *x* [Hz], útlum *y* [dB].

x:	250	300	500	1000	1500	2000	2500	2600
y:	0.4	0.6	1.1	1.2	0.6	1.8	1.9	2.6
x:	2700	2750	2800	2900	3000	3200	3500	
y:	3	3.6	4.7	12	21.6	43.5	57.5	

Úloha L6.32 *Stupeň polynomu závislosti retence na podílu pufru v mobilní fázi*

Při optimalizaci metody HPLC, určující čistotu substance S, byla zjištěna silná závislost retence analytu *y* na podílu pufru v mobilní fázi *x*. (1) Nalezněte polynom, popisující tuto závislost. (2) Stupeň polynomu *m* určete pomocí metody nejmenších čtverců a metody racionálních hodností GPCR. (3) Testujte statistickou významnost jednotlivých parametrů.

Data: Podíl pufru v mobilní fázi x [%], kapacitní poměr pro substanci y.

x:	78.0	79.0	80.0	80.5	81.0	81.5	82.2	82.5
y:	7.47	10.63	15.19	18.45	22.00	27.18	33.57	42.25

6.5 Vícerozměrné lineární regresní modely

Vícenásobná lineární regrese se týká skupiny technik sloužících ke studiu lineární závislosti mezi dvěma či více proměnnými. Určuje odhady parametrů β v regresním modelu

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_m x_{i,m} + \varepsilon_i \ ,$$

kde $\boldsymbol{x}$ jsou nezávisle proměnné a $\boldsymbol{y}$ je závisle proměnná. Index i značí pořadové číslo měření a $\boldsymbol{\beta}$ jsou neznámé regresní parametry a $\boldsymbol{b}$ jejich odhady o počtu m. Pro $m = 1$ se regresní model zjednoduší na jednoduchou lineární regresi. Absolutní člen β_0 je průsečíkem regresní nadroviny s osou y. Odhady b_j jsou směrnice regresní nadroviny ze směru x_j a jsou nazvány parciálními regresními parametry (nebo parciálními regresními koeficienty). Každý takový parciální regresní parametr představuje síť efektů j-té proměnné působící na závisle proměnnou, když ostatní x jsou v regresním modelu drženy na konstantních hodnotách.

Lineární regresní model je schopen změřit pouze lineární, přímkový vztah. Jsou-li body v kruhu, regresní analýza nebude detekovat lineární vztah. Doporučuje se vynášet parciální regresní grafy pro všechny proměnné a vyšetřovat přímkový charakter grafu. Tak se odhalí nepřímkový tvar, vybočující body, míra rozptýlení bodů okolo přímky a řada dalších anomálií přímkového grafu. Nejdůležitějším kritériem linearity je *Pearsonův korelační koeficient r*. Blíží-li se jeho hodnota +1 nebo -1, jde o přímkový vztah; blíží-li se však nule, nejde o lineární (přímkový) vztah. Perfektní přímka má r rovno +1 (vzestupná přímka) nebo r rovno -1 (sestupná přímka).

Vzorová úloha 6.5 *Regresní triplet u vícerozměrného lineárního regresního modelu*

Na úloze **M6.19** *Na vlivu tří parametrů na obsah kadmia v potravinářské pšenici* ukážeme postup analýzy vícerozměrného lineárního regresního modelu

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \ ,$$

kde β_0, β_1, β_2, ..., β_m jsou odhadované parametry: u vzorků potravinářské pšenice byl zjišťován obsah kadmia v zrnu y v závislosti na obsahu kadmia v otrubách x_1, ve stonku s listy x_2 a v kořenovém systému x_3. (1) Vyšetřením regresního tripletu nalezněte nejlepší vícerozměrný regresní model. (2) Využijte k tomu regresní diagnostiku a pomocí parciálních regresních a parciálních reziduálních grafů diskutujte významnost jednotlivých parametrů v modelu stejně jako i jejich fyzikální smysl.

Řešení:

1. Návrh modelu: Začíná se vždy od nejjednoduššího modelu, u kterého vystupují x_1, x_2, x_3 v prvních mocninách a nevyskytují se žádné interakční členy. Na začátku analýzy vždy zařadíme i absolutní člen β_0, takže pro daná data bude navržený regresní model tvaru:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \, .$$

Výběr nezávisle proměnných hledá nejmenší počet smysluplných proměnných. Užívá se k tomu několik technik: kroková regrese, algoritmus "všech možných regresí" a vícerozměrný

výběr proměnných. Každá z těchto technik má své výhody a nevýhody.
Je-li méně než 15 nezávisle proměnných, užijeme především techniku algoritmu "všech možných regresí". Pro více než 15 nezávisle proměnných dáme přednost vícerozměrnému výběru proměnných. Tyto techniky se zde uplatní také v 6. kroku: konstrukci zpřesněného regresního modelu.

2. Předběžná analýza dat: Polohu a proměnlivost proměnných y, x_1, x_2, x_3 charakterizuje *průměr* a *směrodatná odchylka* hodnot každé proměnné.

Proměnná	**Průměr**	**Směrodatná odchylka**	**Párový korelační koeficient**	**Spočtená hladina významnosti**
y	6.0125	4.8734	1.0000	-----
x_1	4.8937	3.5692	0.9837	0.000
x_2	5.7813	4.5296	0.9935	0.000
x_3	5.0813	3.8782	0.9948	0.000

Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných		**Spočtená hladina významnosti**
x_1 versus x_2:	0.99344	0.000
x_1 versus x_3:	0.98693	0.000
x_2 versus x_3:	0.98847	0.000

Pearsonův párový korelační koeficient y* vs. x_1, *y* vs. x_2, *y* vs. x_3** ukazuje na vysokou korelaci, všechny tři nezávislé proměnné x_1, x_2, x_3 jsou se závisle proměnnou y spjaty silnou lineární závislostí. ***Párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných ukazují na silnou korelaci i mezi nezávisle proměnnými. Nejsilnější lineární vztah existuje mezi x_1 vs. x_2 a u x_1 vs. x_3 a x_2 vs. x_3 je rovněž silná korelace.

3. Odhadování parametrů: klasickou metodou nejmenších čtverců (MNČ) byly nalezeny nejlepší odhady čtyř parametrů β_0, β_1, β_2, β_3.

Parametr	**Odhad**	**Směrodatná odchylka**	**H_0: $b_j = 0$ vs. H_A: b_j ... 0** *t*-kritérium	**hypotéza H_0**	**Spočtená hlad. význam.**
b_0	-0.072666	0.13791	-0.52692	Akceptována	0.608
b_1	-0.68505	0.19165	-3.5746	Zamítnuta	0.004
b_2	0.89619	0.16072	5.5761	Zamítnuta	0.000
b_3	0.83769	0.13322	6.2879	Zamítnuta	0.000

Studentův *t*-test statistické významnosti jednotlivých parametrů t_{exp} ' $(b_j$ & $0)/s(b_j)$ ukázal při $t_{0.95}(16\text{-}4) = 2.179$, že absolutní člen β_0 je statisticky nevýznamný, zatímco ostatní parametry statisticky významné jsou. Je-li spočtená hladina významnosti větší než $\alpha = 0.05$, dotyčný parametr je statisticky nevýznamný. Opět vychází, že pouze β_0 je statisticky nevýznamný. To je v souladu i s biologickou interpretací: β_0 se týká zbytkového obsahu kadmia v zrnu, když je obsah kadmia v otrubách nulový ($x_1 = 0$), ve stonku ($x_2 = 0$) a v kořenovém systému rovněž nulový ($x_3 = 0$). Je zřejmé, že když v celé rostlince bude obsah kadmia nulový, musí být nulový obsah i v zrnu a zbytkový obsah proto nemá ani biologický smysl. Absolutní člen β_0 je nutno ve zpřesněném modelu vynechat.

Intervalové odhady regresních parametrů (i. s.): při užití $t_{0.95}(16\text{-}4) = 2.1788$.

Parametr	**Odhad parametru**	**Směrodatná odchylka**	**Dolní mez 95 % i. s.**	**Horní mez 95 % i. s.**	**Odhad standar. parametru**
Úsek	-7.266543E-02	0.137906	-0.3731368	0.2278059	0.0000
x_1	-0.6850502	0.1916463	-1.102612	-0.2674888	-0.5017
x_2	0.8961921	0.1607203	0.5460125	1.246372	0.8330

x_3	0.8376914	0.1332221	0.5474254	1.127957	0.6666

Vyčíslení dolní a horní meze intervalového odhadu určovaných parametrů je založeno na Studentově *t*-rozdělení s *n*-*m*-1 stupni volnosti vztahem $b_j \pm t_{1\&\alpha/2}(n \,\&\, m \,\&\, 1).s(b_j)$. Tento interval předpokládá, že rezidua regresního modelu jsou normálně rozdělena. Intervalový odhad lze využít i ke statistickému testování významnosti parametru β_j. Leží-li nula v intervalu spolehlivosti parametru (i. s.), je parametr statisticky nevýznamný. **Standardizovaný odhad parametrů:** vyčíslují se ze standardizovaných dat ***y*** a ***X***, když totiž od každého prvku vektoru ***y*** či $\boldsymbol{x}_j$ odečteme jeho vektorový průměr a podělíme jeho vektorovou směrodatnou odchylkou. Často jsou proměnné v rozličných jednotkách a numericky se liší o mnoho řádů. Je proto užitečné předejít standardizací dat jednak numerickým obtížím zaokrouhlování a jednak docílit i vzájemného porovnání parametrů. Existuje vztah mezi standardizovaným $b_{std,\,j}$ a nestandardizovaným odhadem b_j dle $b_{\mathrm{std},j} = b_j\, s(x_j) / s(y)$, kde $s(y)$ a $s(x_j)$ jsou vektorové směrodatné odchylky závisle proměnné *y* a *j*-té nezávisle proměnné x_j.

4. Základní statistické charakteristiky: slouží zde především jako kritéria rozlišení mezi rozličnými návrhy regresního modelu.

Vícenásobný korelační koeficient *r*	: 0.99858
Koeficient determinace 100 % *D*	: 99.716
Predikovaný koeficient determinace R^2_P	: 0.99527
Střední kvadratická chyba predikce *MEP*	: 0.2101
Akaikova informační kritérium *AIC*	:-36.180

Vícenásobný korelační koeficient r ukazuje, že navržený lineární regresní model je statisticky významný. Vysoká hodnota ***koeficientu determinace*** *D* ukazuje, že všechny body výtečně korespondují s modelem. ***Predikovaný koeficient determinace*** D_P má podobný význam jako koeficient determinace, je však vyčíslen jinak, místo *RSC* se ve vztahu užije *MEP*. ***Střední kvadratická chyba predikce*** *MEP* a ***Akaikovo informační kritérium*** *AIC* se užívají k rozlišení mezi několika navrženými modely. Za optimální se považuje model, pro který dosahuje *MEP* i *AIC* minimální hodnotu.

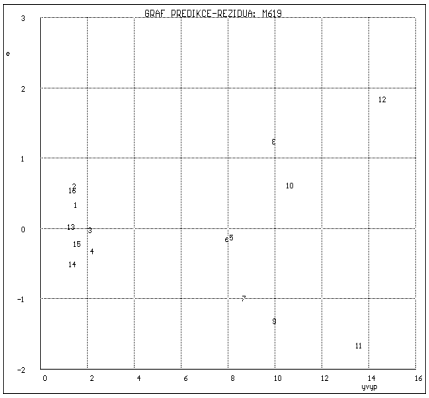
5. Regresní diagnostika: obsahuje pomůcky a postupy pro interaktivní analýzu (a) dat, (b) modelu, (c) metody, což jsou složky tzv. *regresního tripletu.*

A. Data: skládá se z analýzy několika druhů grafických diagnostik a tabulek různých druhů reziduí.

(a) ***Analýza klasických reziduí*** není příliš spolehlivá, protože klasická rezidua jsou korelovaná, s nekonstantním rozptylem, jeví se normálnější než náhodné chyby (*efekt supernormality)* a nemusí indikovat silně odlehlé hodnoty. Grafická analýza $\hat{e}$ vs. $\hat{y}_P$ (obr. 6.5-1) je však schopna indikovat podezřelé body, trend, a nekonstantnost podmíněného rozptylu, tj. heteroskedasticitu. Míry polohy a rozptýlení klasických reziduí by měly dosahovat hodnot blízkých experimentálnímu šumu. *Odhad směrodatné odchylky s(e)* by se měl blížit svou velikostí experimentální chybě, kterou je zatížena závisle proměnná.

Kritérium testu	**Hodnota kritéria**	**Spočtená hladina významnosti α**	**Závěr testování(5%): H_0 o normalitě je**
Šikmost	0.1918	0.847934	Přijata
Špičatost	0.5255	0.599260	Přijata
Omnibus	0.3129	0.855181	Přijata

Je možné užít také testů na šikmost, špičatost a omnibus testu, založeného na šikmosti a špičatosti a obojím. **Testy normality:** normalita reziduí by se měla vyšetřit vizuálně krabicovým grafem reziduí, rankitovým grafem a grafem hustoty pravděpodobnosti.



Obr. 6.5-1 Analýza klasických reziduí, ***ADSTAT.***

Bod	**Měřená hodnota**	**Predikovaná hodnota**	**Směrodatná odchylka**	**Klasické reziduum**	**Relativní reziduum**
i	$y_{exp,i}$	$y_{vyp,i}$	$s(y_{vyp,i})$	e_i	$e_{r,i}$
1	1.6000E+00	1.5006E+00	1.0203E-01	9.9415E-02	6.2135E+00
2	1.6000E+00	1.4227E+00	1.0587E-01	1.7733E-01	1.1083E+01
3	2.1000E+00	2.1029E+00	1.1045E-01	-2.9207E-03	-1.3908E-01
4	2.1000E+00	2.1925E+00	1.0412E-01	-9.2540E-02	-4.4067E+00
5	8.1000E+00	8.1354E+00	2.0871E-01	-3.5420E-02	-4.3729E-01
6	7.9000E+00	7.9410E+00	1.3352E-01	-4.1016E-02	-5.1919E-01
7	8.4000E+00	8.6869E+00	1.6249E-01	-2.8690E-01	-3.4155E+00
8	1.0300E+01	9.9377E+00	1.7913E-01	3.6232E-01	3.5176E+00
9	9.6000E+00	9.9805E+00	1.3555E-01	-3.8050E-01	-3.9636E+00
10	1.0800E+01	1.0621E+01	1.3061E-01	1.7918E-01	1.6591E+00
11	1.3100E+01	1.3581E+01	2.2832E-01	-4.8080E-01	-3.6703E+00
12	1.5100E+01	1.4565E+01	1.8609E-01	5.3530E-01	3.5450E+00
13	1.3000E+00	1.2908E+00	1.0514E-01	9.1821E-03	7.0632E-01
14	1.2000E+00	1.3441E+00	1.1635E-01	-1.4406E-01	-1.2005E+01
15	1.5000E+00	1.5597E+00	1.1723E-01	-5.9675E-02	-3.9784E+00
16	1.5000E+00	1.3389E+00	1.0987E-01	1.6110E-01	1.0740E+01

Rezidualní součet čtverců *RSC*	: 1.0114
Průměr absolutních hodnot reziduí *Me*	: 0.19048
Průměr relativních reziduí Me_r	: 4.3750
Odhad reziduálního rozptylu $s^2(e)$	: 8.4282E-02
Odhad směrodatné odchylky reziduí *s(e)*	: 0.29031
Odhad šikmosti reziduí $g_1(e)$	: 0.0923
Odhad špičatosti reziduí $g_2(e)$	: 2.8755

(b) ***Analýza ostatních reziduí:*** Jackknife rezidua indikují odlehlé body, diagonální prvky H_{ii} projekční matice $\boldsymbol{H}$ a diagonální prvky H_{mii} zobecněné projekční matice $\boldsymbol{H_m}$ pouze extrémy. Ostatní druhy reziduí a kritéria v tabulce pak obojí (značeno hvězdičkou u hodnoty). Jackknife rezidua $e_{J,i}$ ukazují, že bod č. 11 a 12 je odlehlý, stejně tak i Cookova vzdálenost D_i a Atkinsova vzdálenost A_i ukazují na odlehlost č. 8, 11, 12, kritérium DF_i na č. 8, 11, 12 a věrohodnostní vzdálenosti $LD(b)_i$ a $LD(s^2)_i$ na č. 11 a $LD(b, s^2)_i$ na č. 11 a 12. Diagonální prvky $H_{i,i}$ projekční matice $\boldsymbol{H}$ ukazují na extrémy č. 5 a 11, a diagonální prvky $H_{mi,i}$ zobecněné projekční matice $\boldsymbol{H_m}$ pak na extrémy č. 11 a 12.

Indikace vlivných bodů: (* indikuje odlehlý nebo vlivný bod)

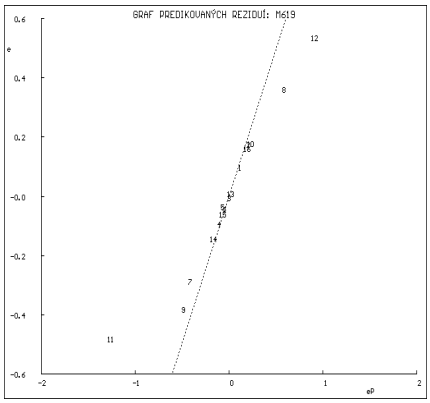
Bod	Standardizované reziduum	Jackknife reziduum	Predikované reziduum	Diagonální prvky
i	$e_{S,i}$	$e_{J,i}$	$e_{P,i}$	$H_{i,i}$
1	3.6577E-01	3.5217E-01	1.1343E-01	1.2352E-01
2	6.5601E-01	6.3966E-01	2.0453E-01	1.3298E-01
3	-1.0879E-02	-1.0415E-02	-3.4150E-03	1.4474E-01
4	-3.4147E-01	-3.2854E-01	-1.0620E-01	1.2863E-01
5	-1.7552E-01	-1.6827E-01	-7.3308E-02	5.1683E-01*
6	-1.5911E-01	-1.5249E-01	-5.2019E-02	2.1152E-01
7	-1.1925E+00	-1.2161E+00	-4.1777E-01	3.1326E-01
8	1.5859E+00	1.7078E+00	5.8504E-01	3.8069E-01
9	-1.4821E+00	-1.5700E+00	-4.8658E-01	2.1800E-01
10	6.9110E-01	6.7525E-01	2.2465E-01	2.0240E-01
11	-2.6814E+00	-4.0550E+00*	-1.2604E+00	6.1852E-01*
12	2.4023E+00	3.1923E+00*	9.0863E-01	4.1087E-01
13	3.3931E-02	3.2488E-02	1.0568E-02	1.3115E-01
14	-5.4162E-01	-5.2502E-01	-1.7163E-01	1.6062E-01
15	-2.2469E-01	-2.1557E-01	-7.1301E-02	1.6305E-01
16	5.9952E-01	5.8279E-01	1.8803E-01	1.4322E-01

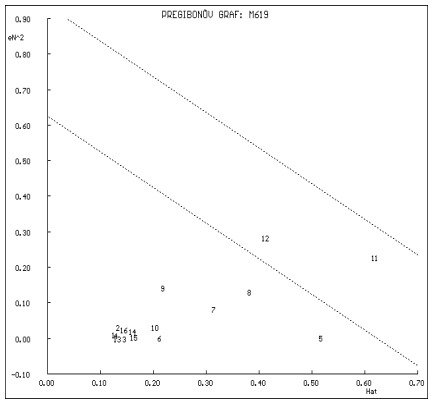
Bod	Zobecněné diag. prvky	Cookova vzdálenost	Atkinsonova vzdálenost	Vliv na predikci
i	$H_{mi,i}$	D_i	A_i	Df_i
1	1.3329E-01	4.7136E-03	2.2899E-01	1.3221E-01
2	1.6407E-01	1.6501E-02	4.3390E-01	2.5051E-01
3	1.4475E-01	5.0070E-06	7.4214E-03	-4.2848E-03
4	1.3709E-01	4.3030E-03	2.1863E-01	-1.2622E-01
5	5.1807E-01	8.2386E-03	3.0143E-01	-1.7403E-01
6	2.1318E-01	1.6978E-03	1.3680E-01	-7.8983E-02
7	3.9465E-01	1.6218E-01	1.4226E+00	-8.2134E-01
8	5.1049E-01	3.8650E-01*	2.3192E+00*	1.3390E+00*
9	3.6116E-01	1.5310E-01	1.4358E+00	-8.2895E-01
10	2.3414E-01	3.0299E-02	5.8916E-01	3.4015E-01
11	8.4709E-01*	2.9145E+00*	8.9433E+00*	-5.1634E+00*
12	6.9419E-01*	1.0062E+00*	4.6176E+00*	2.6660E+00*
13	1.3124E-01	4.3449E-05	2.1863E-02	1.2623E-02
14	1.8114E-01	1.4034E-02	3.9779E-01	-2.2966E-01
15	1.6657E-01	2.4587E-03	1.6480E-01	-9.5150E-02
16	1.6888E-01	1.5020E-02	4.1270E-01	2.3827E-01

Bod	Věrohodnostní vzdálenosti		
i	$LD(b)_i$	$LD(s^2)_i$	$LD(b,s^2)_i$
1	2.5120E-02	2.2351E-02	4.6185E-02
2	8.7766E-02	6.2215E-03	9.1797E-02
3	2.6704E-05	3.2606E-02	3.2632E-02
4	2.2933E-02	2.3569E-02	4.5296E-02
5	4.3879E-02	3.0095E-02	7.1394E-02
6	9.0523E-03	3.0538E-02	3.9045E-02
7	8.4238E-01	3.0992E-02	9.5089E-01
8	1.9389E+00	2.4682E-01	2.6917E+00
9	7.9637E-01	1.5878E-01	1.0958E+00

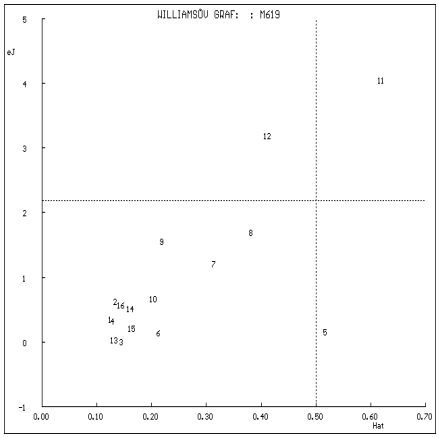
10	1.6079E-01	4.5425E-03	1.6232E-01
11	1.0861E+01*	7.8276E+00*	4.4183E+01*
12	4.6276E+00	3.4386E+00	1.3130E+01*
13	2.3173E-04	3.2520E-02	3.2738E-02
14	7.4671E-02	1.2499E-02	8.4425E-02
15	1.3108E-02	2.8534E-02	4.0880E-02
16	7.9906E-02	9.2132E-03	8.6631E-02

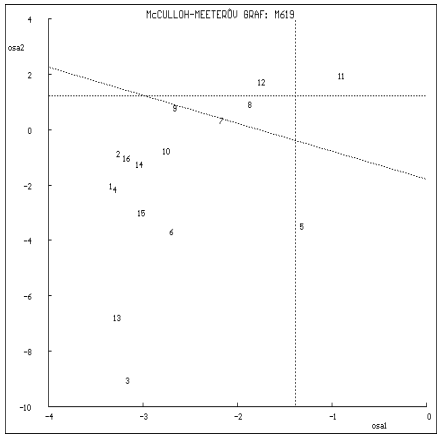
(c) ***Grafy vlivných bodů*** (obr. 6.5-2) jsou schopny indikovat a současně i testovat, dokazovat přítomnost odlehlých hodnot a extrémů. *Graf predikovaných reziduí* ukazuje na odlehlé body č. 8, 11, 12 a částečně na extrémy č. 11 a 12. *Pregibonův graf* ukazuje na středně vlivné body č. 11 a 12. *Williamsův graf* indikuje č. 11 a 12 jako odlehlé body a č. 11 a 12 jako extrémy. *McCullohův-Meeterův graf* dokazuje odlehlé body č. 11 a 12, extrémy č. 8 a 11. Konečně *L-R graf* dokazuje odlehlé body č. 8, 11 a 12 a extrémy č. 11, 12 a 9. Lze uzavřít, že body č. 11 a 12 jsou většinou diagnostik prokázány jako odlehlé, a proto je třeba je dále prověřit nebo z výběru vyloučit**.**



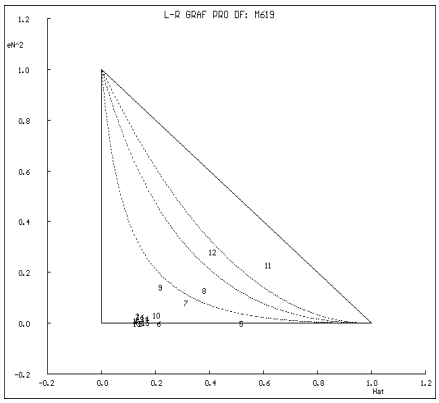


Obr. 6.5-2 Grafy vlivných bodů, vlevo, graf predikovaných reziduí, a vpravo, Pregibonův graf, ***ADSTAT.***



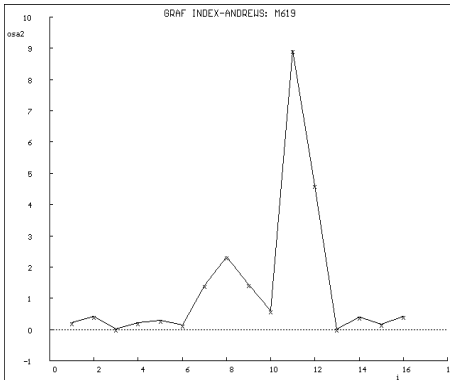


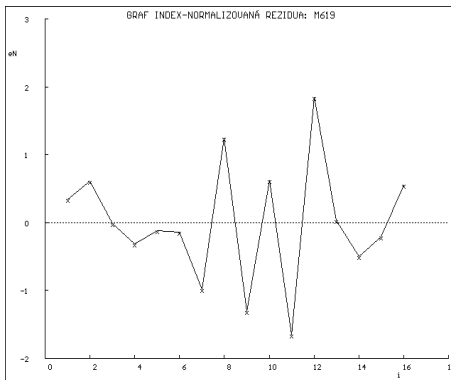
Obr. 6.5-2 Grafy vlivných bodů, vlevo, Williamsův graf, a vpravo, McCullohův-Meeterův graf, ***ADSTAT.***



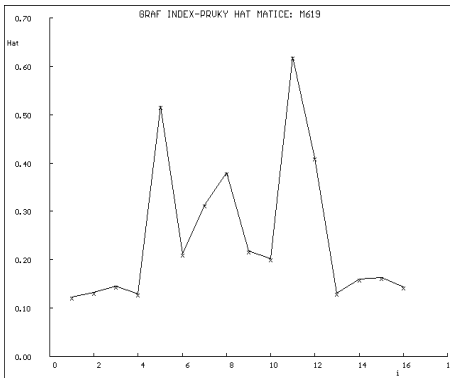
Obr. 6.5-2 Grafy vlivných bodů, L-R graf, ***ADSTAT.***

(d) ***Indexové grafy*** (obr. 6.5-3) upozorňují na podezřelé body. *Andrewsův indexový graf* a *graf normovaných reziduí* ukazují na podezřelé body č. 5, 8, 11, 12. *Indexový graf prvků projekční matice* ***H*** pak na podezřelé extrémy č. 5, 8, 11.



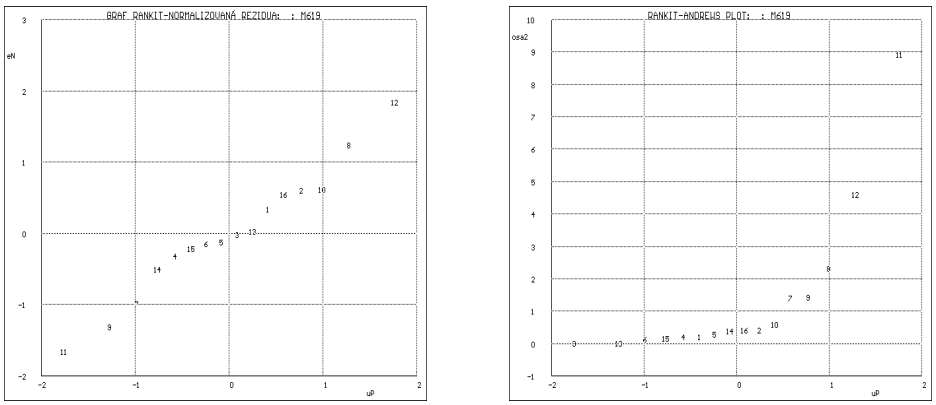


Obr. 6.5-3 Indexové grafy, vlevo, Andrewsův graf, a vpravo, graf normovaných reziduí, ***ADSTAT.***

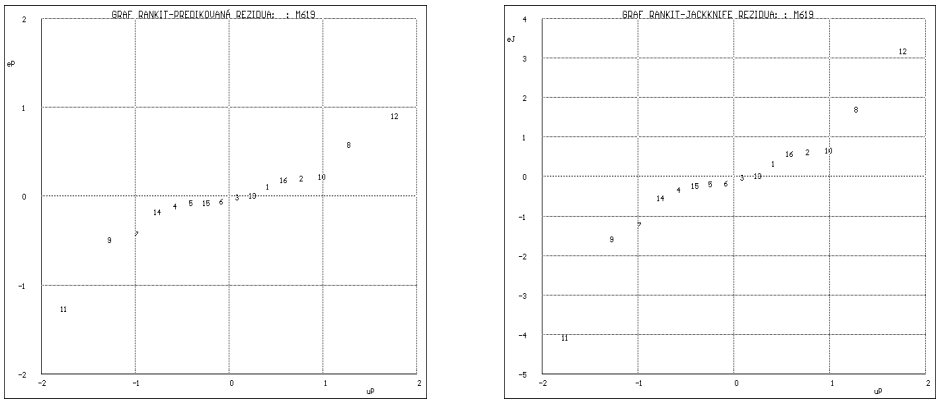


Obr. 6.5-3 Graf prvků *H*-projekční matice, ***ADSTAT.***

(e) ***Rankitové grafy*** (obr. 6.5-4) ukazují vedle normality rozdělení dotyčných reziduí i na vlivné (zde odlehlé) body. *Graf normovaných reziduí* ukazuje na začátku č. 11 a na konci č. 12 jako odlehlé body. *Andrewsův graf* ukazuje na č. 8, 11 a 12 jako na odlehlé, *graf Jackknife reziduí* č. 11 a 12 jako odlehlé body. Po odstranění dvou odlehlých bodů č. 11 a 12 lze konstatovat, že zbytek dat nevykazuje odchylky od normality.

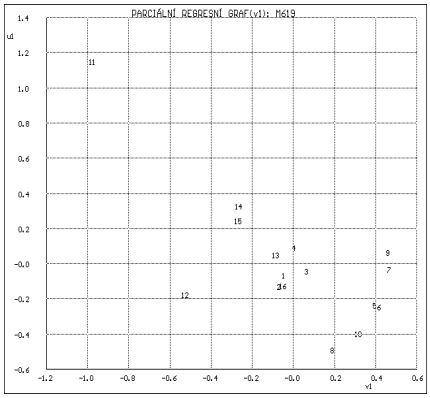


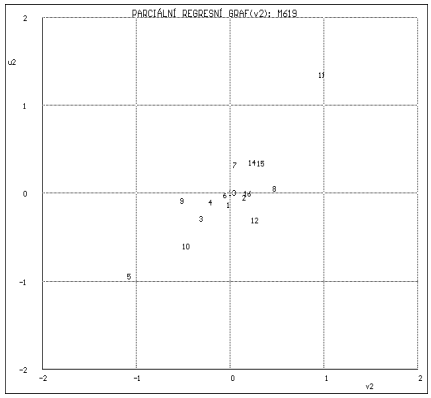
Obr. 6.5-4 Rankitové grafy, vlevo, graf normovaných reziduí, a vpravo, Andrewsův graf, ***ADSTAT.***



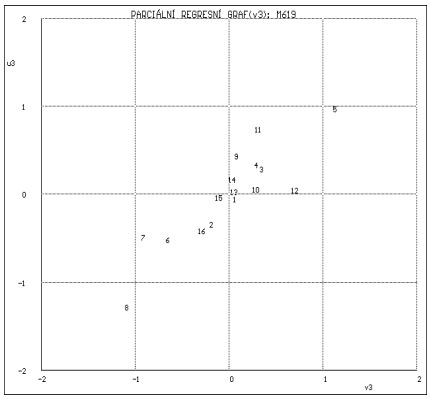
Obr. 6.5-4 Rankitové grafy, vlevo, graf predikovaných reziduí, a vpravo, graf Jackknife reziduí, ***ADSTAT.***

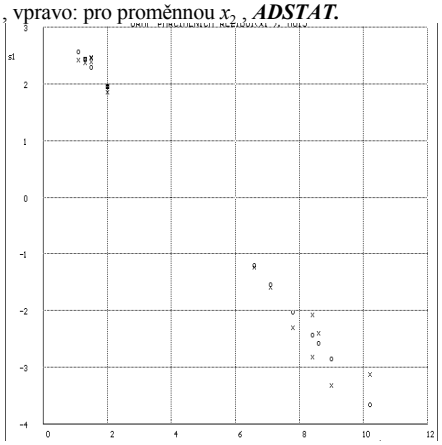
B. Model: *Parciální regresní grafy* (obr. 6.5-5), ale především *parciální reziduální grafy* (obr. 6.5-6), ukazují na čisté lineární závislosti jednotlivých nezávisle proměnných. Vedle posouzení závislosti navrženého regresního modelu umožňují také indikovat vlivné body, a to č. 5, 8, 11 a 12. Navržený model se jeví stran členů $\beta_1 x_1$ % $\beta_2 x_2$ % $\beta_3 x_3$ správný, pouze β_0 je nadbytečné.



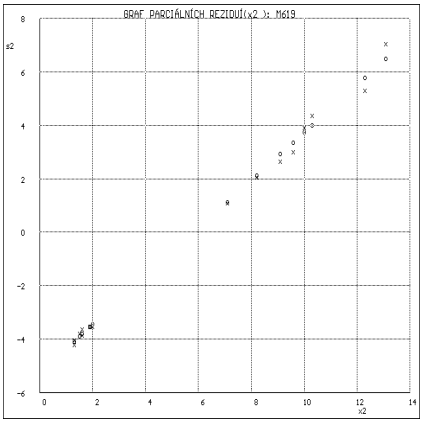


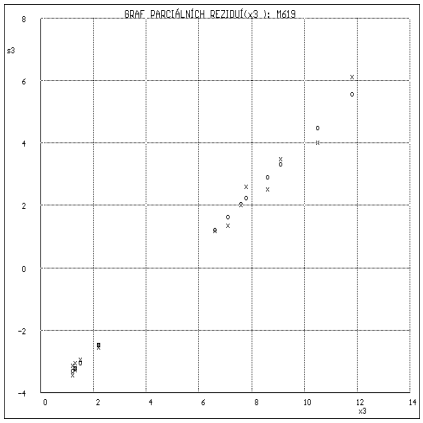
Obr. 6.5-5 Parciální regresní grafy, vlevo pro proměnnou x_1 , vpravo: pro proměnnou x_2 , ***ADSTAT.***





Obr. 6.5-5 Parciální regresní grafy, vlevo: pro proměnnou x_3, vpravo: parciální reziduální graf pro proměnnou x_1, ***ADSTAT***.





Obr. 6.5-6 Parciální reziduální grafy, vlevo: pro proměnnou x_2, vpravo: pro proměnnou x_3, ***ADSTAT***.

Koeficient determinace *D:* vystihuje procento proměnlivosti závisle proměnné *y*, objasněné nezávisle proměnnými v regresním modelu. Různé typy koeficientu determinace vnáší světlo do proměnlivosti *y*, když jsou postupně přidávány jednotlivé proměnné v pořadí x_1, x_2, ..., x_m nebo nezávisle proměnnými odebranými z modelu.

Nezávisle proměnná	Kumulativní přidání	Izolované přidání	Přidaná poslední	Přidaná sama	Parciální (pro zbytek)
x_1	0.967654	0.967654	0.003023	0.967654	0.515689
x_2	0.987807	0.020153	0.007356	0.986992	0.721532
x_3	0.997161	0.009354	0.009354	0.989671	0.767163

Kumulativní přidání: hodnota *D* pro tuto nezávisle proměnnou a všechny v modelu nad ní. Nezávisle proměnné pod ní jsou ignorovány. **Izolované přidání:** příspěvek do *D*, který způsobí tato nezávisle proměnná, přidaná do modelu, který již obsahuje nezávisle proměnné umístěné pod ní. **Přidaná poslední:** příspěvek do *D*, o který bude *D* sníženo, když se tato proměnná z modelu odebere. Velké hodnoty ukazují na důležitou nezávisle proměnnou, zatímco malé hodnoty na nevýznamnou proměnnou. **Přidaná sama:** je výsledná hodnota *D*, když závisle proměnná *y* je v modelu funkcí pouze této jediné nezávisle proměnné. Samozřejmě, vysoká hodnota *D* značí vysokou důležitost této nezávisle proměnné. **Parciální (nastavena pro zbytek):** čtverec parciálního korelačního koeficientu D_i vyjadřuje procento proměnlivosti v závisle proměnné *y*, vysvětlené pouze jedinou nezávisle proměnnou, řídící efekty zbytku nezávisle proměnných.

Odstranění nezávisle proměnné z modelu: jeden ze způsobů vyhodnocení důležitosti nezávisle proměnné x_j je vyšetření vlivu odstranění této proměnné z modelu na různé statistiky těsnosti proložení.

Nezávisle proměnná	*D*	*MEP*	Mallowovo *Cp*	Spočtená hladina významnosti	*D* vs ostatní *X*
			při vypuštění této proměnné		
Celý model	0.997161	0.0842824			
x_1	0.994138	0.1606387	14.777442	0.003820	1.000000
x_2	0.989805	0.2793823	33.092846	0.000121	1.000000
x_3	0.987807	0.3341351	41.538095	0.000040	1.000000

Nezávisle proměnná: název nezávisle proměnné, která je z modelu vypuštěna. Celý model značí, že žádná proměnná nebyla vypuštěna. **Hodnota *D* při vypuštění této proměnné:** *D* pro vícerozměrný regresní model, když tato nezávisle proměnná je vypuštěna a zbývající proměnné jsou užity. Je-li toto *D* blízké hodnotě pro *D* celého modelu, není tato proměnná velmi důležitá. Naopak, je-li toto *D* mnohem menší než *D* celého modelu, je tato nezávisle proměnná důležitá. ***MEP* při vypuštění této proměnné:** střední kvadratická chyba predikce, když tato nezávisle proměnná je vypuštěna a zbývající proměnné jsou užity. **Mallowovo *Cp* při vypuštění této proměnné:** optimální model bude mít *Cp* statistiku blízkou hodnotě (*m*+1). Je-li *Cp* větší než (*m*+1), znamená to, že regresní model je přeurčen a obsahuje příliš mnoho proměnných a hrozí nebezpečí multikolinearity. Je-li *Cp* menší než (*m*+1) ukazuje to, že regresní model je podceněn a minimálně jedna důležitá nezávisle proměnná byla vypuštěna. Vzorec pro výpočet *Cp* má tvar

$$Cp \;=\; \frac{MEP_m}{MEP_k}\,(n \,\&\, m \,\&\, 1) \,\&\, (n \,\&\, 2m \,\&\, 2) \quad ,$$

kde *k* značí maximální počet nezávisle proměnných. **Spočtená hladina významnosti α:** jde o oboustranný test spočtené hladiny významnosti parciálního regresního koeficientu. Nejspíše budou důležité nezávisle proměnné s malou hodnotou spočteného α. Kolinearita však může způsobit zvláště veliké hodnoty spočteného α, proto je třeba nejprve vyšetřovat multikolinearitu v modelu. ***D* vs ostatní *X*:** toto *D* vznikne, jestliže se tato nezávisle proměnná bude regresovat v závislosti na zbývajících nezávisle proměnných. Vysoká hodnota značí nadbytečnost této proměnné. Nezávisle proměnné s vysokou hodnotou vyšší než 90% je třeba z modelu odstranit.

Krokové přidávání proměnných do regresního modelu: vyšetřuje krok po kroku přidávání nezávisle proměnných do regresního modelu.

Proměnná	*D* pro zařazené proměnné	*D* pro nezařazené proměnné	*F*-test	Spočtená α	*F*-test	Spočtená α
			pro zařazené proměnné		pro nezařazené proměnné	
x_1	0.967654	0.029507	418.82	0.00000	62.36	0.000000
x_2	0.987807	0.009354	526.60	0.00000	39.54	0.000040
x_3	0.997161	0.000000	1404.98	0.00000		

***D* pro zařazené proměnné:** velikost *D*, když bude užita pouze tato nezávisle proměnná a všechny ostatní nad ní. **D pro nezařazené proměnné:** velikost *D* pro celý model minus část "zařazené proměnné". Jde o část *D* vysvětlenou všemi nezávisle proměnnými pod tímto řádkem. Velké hodnoty ukazují, že důležitější jsou právě nezávisle proměnné pod tímto řádkem. ***F*-test pro zařazené proměnné:** *F*-testační kritérium k testování nulové hypotézy H_0: "*všechny β pro tento řádek a řádky nad ním jsou nulové*". Alternativní H_A říká, že alespoň jeden parametr β_j je nenulový. **Spočtená hladina α pro zařazené proměnné:** spočtená hladina významnosti α pro dotyčné *F*-kritérium. ***F*-test pro nezařazené proměnné:** *F*-testační kritérium k testování nulové hypotézy H_0: "*všechny β pro řádky nižší než tento jsou nulové*". Alternativní H_A říká, že alespoň jeden β_j je nenulový. **Spočtená hladina α pro nezařazené proměnné:** spočtená hladina významnosti α pro dotyčné *F*-kritérium.

Predikce a intervalový odhad středních hodnot signálu: intervaly spolehlivosti střední hodnoty signálu $\bar{y}_i$ v *i*-tém řádku jsou vypočteny pro zadané hodnoty nezávisle proměnné *X*. Je třeba si povšimnout, že porušení předpokladů MNČ znekvalitní tento intervalový odhad.

Řádek	Exper.	Predikce	Směr. odchylka predikce	Dolní mez 95% i. s.	Horní mez 95% i. s.
1	1.6	1.500584	0.1020315	1.278277	1.722892
2	1.6	1.422665	0.1058675	1.192	1.653331
3	2.1	2.10292	0.1104498	1.862271	2.34357
4	2.1	2.192539	0.1041193	1.965683	2.419396
5	8.1	8.135422	0.2087092	7.680683	8.590159
6	7.9	7.941016	0.1335198	7.650102	8.231931
7	8.4	8.6869	0.1624884	8.332868	9.040932
8	10.3	9.937684	0.1791252	9.547404	10.32796
9	9.6	9.980503	0.1355499	9.685164	10.27584
10	10.8	10.62082	0.1306079	10.33625	10.90539
11	13.1	13.5808	0.2283214	13.08333	14.07827
12	15.1	14.5647	0.1860882	14.15925	14.97015
13	1.3	1.290818	0.105138	1.061742	1.519894
14	1.2	1.344059	0.1163509	1.090552	1.597566
15	1.5	1.559675	0.1172271	1.30426	1.815091
16	1.5	1.338896	0.1098664	1.099518	1.578274

Exper.: naměřená experimentální hodnota signálu y_i pro *i*-tý řádek. **Predikce**: vypočtená hodnota signálu $\hat{y}_i$ pro *i*-tý řádek je vypočtena na základě zadaných hodnot nezávisle proměnných pro tento řádek. **Směrodatná odchylka predikce:** směrodatná odchylka střední hodnoty y_i v *i*-tém řádku pro zadaná *x*. Všimněte si, že tato hodnota nemusí být konstantní pro všechny ostatní řádky. **Dolní mez 95% intervalu spolehlivosti střední hodnoty** $\bar{y}_i$: je vypočtena pro nezávisle proměnné v tomto *i*-tém řádku. **Horní mez 95% intervalu spolehlivosti střední hodnoty** $\bar{y}_i$**:** je vypočtena pro nezávisle proměnné v tomto *i*-tém řádku. Je třeba pouze zadat hladinu významnosti α.

Predikce a intervalový odhad jednotlivých hodnot signálu:

Řádek	Exper.	Predikce	Směr. odchylka predikce	Dolní mez 95% i. s.	Horní mez 95% i. s.
1	1.6	1.500584	0.3077221	0.8301157	2.171053
2	1.6	1.422665	0.3090151	0.7493793	2.095952
3	2.1	2.10292	0.3106149	1.426149	2.779692
4	2.1	2.192539	0.3084206	1.520549	2.86453
5	8.1	8.135422	0.3575499	7.356387	8.914456
6	7.9	7.941016	0.3195465	7.244784	8.637248
7	8.4	8.6869	0.3326934	7.962023	9.411777
8	10.3	9.937684	0.3411279	9.19443	10.68094
9	9.6	9.980503	0.3204001	9.282411	10.67859
10	10.8	10.62082	0.3183408	9.927211	11.31442
11	13.1	13.5808	0.3693415	12.77608	14.38553
12	15.1	14.5647	0.3448351	13.81337	15.31603
13	1.3	1.290818	0.308766	0.6180745	1.963561
14	1.2	1.344059	0.3127618	0.6626092	2.025508
15	1.5	1.559675	0.3130888	0.8775135	2.241838
16	1.5	1.338896	0.3104079	0.6625756	2.015217

Jsou vyčísleny intervaly spolehlivosti i. s. jednotlivých hodnot signálu y_i pro zadané hodnoty nezávisle proměnných v tomto řádku. Výklad tabulky je obdobný jako předešle.

C. Metoda: do této části patří vyšetření splnění základních předpokladů metody nejmenších čtverců (MNČ), za kterých by měla vést k nejlepším nestranným lineárním odhadům regresních parametrů:

(a) *Statistická významnost navrženého regresního modelu:* Fisherův-Snedecorův test významnosti regresního modelu potvrdil, že navržený model je přijat jako významný, jinými slovy: závisle proměnná y a nezávisle proměnné x_1, x_2, x_3 jsou v lineární závislosti. ANOVA přináší detailní informace o zdrojích proměnlivosti v datech a o tomto testu.

Zdroj	*SV*	Suma čtverců	Průměrný čtverec	*F*-test	Spočtená hlad. význam.	Síla (5%)
Úsek	1	578.4025	578.4025			
Model	3	355.2461	118.4154	1404.9828	0.0000	1.0000
Chyba	12	1.011389	0.084282			
Total (Adjustov.)	15	356.2575	23.7505			

Tabulkový kvantil, $F_{1-\alpha}(m\text{-}1, n\text{-}m)$: 3.4903

Závěr: Navržený regresní model je přijat jako významný.

Odmocnina průměr. kvadrat. chyby: 0.2903144 **Koeficient determinace *D***: 0.9972

Průměr hodnot závisle proměnné : 6.0125 **Adjustovaný koeficient determinace *D***: 0.9965

Variační koeficient : 0.04828 ***PRESS***: 3.362263

Suma abs. hodnot reziduí VK : 4.877476 **Predikovaný koeficient determinace D_P** : 0.9906

Zdroj: přináší názvy zdrojů proměnlivosti v datech, jsou čtyři: absolutní člen, tj. úsek β_0; model; chyba a celkový zdroj. **SV:** stupně volnosti představují počet rozměrů, spojených s dotyčným zdrojem proměnlivosti. Všimněte si, že každé pozorování může být interpretováno jako bod v *n*-rozměrném prostoru. Absolutní člen β_0 má 1 stupeň volnosti, model má *m* stupňů volnosti, chyba má *(n-m-*1*)* stupňů volnosti a celkový zdroj má *(n-*1*)* stupňů volnosti. **Suma čtverců:** je několik sum čtverců, spojených se zdrojem proměnlivosti v proměnné *y*. Všechny budou ve výrazech závisle proměnné *y* dle vzorců:

$$SS_{\text{úsek}} = n\,\bar{y}^2, \quad SS_{\text{model}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \quad SS_{\text{chyba}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad SS_{\text{Total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

Průměrný čtverec: čili rozptyl je suma čtverců dělená počtem stupňů volnosti. Představuje odhadovaný rozptyl dotyčného zdroje proměnlivosti. ***F*-test:** testování nulové hypotézy H_0: všechny parametry jsou nulové, $\beta_j = 0$. Toto kritérium má *m* stupňů volnosti pro čitatele a (*n-m-*1*)* stupňů volnosti pro jmenovatele. **Spočtená hladina významnosti *α*:** spočtená hladina statistické významnosti α pro *F*-test. Je-li spočtená α menší než zadaná α = 0.05, nulová hypotéza H_0 o nulovosti všech parametrů β je zamítnuta. Je-li spočtená α větší než α = 0.05, je H_0 o nulovosti všech parametrů β_j přijata. **Síla(5%):** síla testu je pravděpodobnost zamítnutí nulové hypotézy, že všechny regresní parametry β_j jsou nulové, když alespoň jeden není nulový. **Odmocnina průměrné kvadratické chyby:** představuje odhad směrodatné odchylky reziduí e_i. **Průměr hodnot závisle proměnné:** aritmetický průměr závisle proměnné. **VK:** variační koeficient čili relativní směrodatná odchylka je mírou rozptýlení. Vypočte se dle vzorce $VK = \sqrt{MSE}\,/\bar{y}$. **Koeficient determinace *D*:** je definovaný $D = SS_{\text{Model}}/SS_{\text{Total}}$ a představuje nejpopulárnější míru těsnosti proložení dat regresním modelem. Udává se spíše v procentech, takže vyjadřuje, jak velké procento bodů vyhovuje navrženému regresnímu modelu. Hodnota blízká nule značí, že lineární vztah mezi $\bar{y}$ a ***X*** vlastně neexistuje, zatímco hodnota blízká 100% značí perfektní lineární proložení. **PRESS:** predikovaná suma čtverců se užívá k validaci regresního modelu v predikční schopnosti. Při jeho výpočtu se postupně vynechávají jednotlivá pozorování. Zbývajících *(n-*1) pozorování slouží k regresi. Proces se *n*krát opakuje, jednou pro každé pozorování. Rozdíl mezi experimentálním y_i a vypočteným $\hat{y}_{i,\,\&i}$ za vynechání *i*-tého bodu se nazývá predikční chyba, $PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i,\&i})^2$. Suma čtverců predikčních chyb je *PRESS*. Čím menší *PRESS*, tím lepší predikční schopnost modelu. *PRESS* se užívá především při výběru proměnných, kde slouží jako kritérium k porovnání mezi regresními modely. **Predikovaný koeficient determinace D_P, R^2_P:** má obdobné využití jako *D*. Vystihuje predikční schopnost modelu bez rozdělování dat na dvě poloviny. Je zcela běžné mít vysoké *D* a velmi nízké D_P. Znamená to totiž, že prokládaný regresní model je silně závislý na datech. D_P leží v rozmezí 0 do 1 či od 0 % do 100 %. Vztah vůči *PRESS* vystihuje vzorec $D_P = 1 - (PRESS/SS_{\text{Total}})$. **Suma absolutních hodnot *PRESS* reziduí:** Je-li příliš veliká v důsledku jednoho či několika PRESS reziduí, je tato statistika jedna z nejvhodnějších k vyjádření predikční schopnosti modelu. Vypočte se vztahem $\sum_{i=1}^{n} |y_i - \hat{y}_{i,\&i}|$.

(b) *Multikolinearita:* blok Indikace multikolinearity zde vykazuje ve všech kritériích multikolinearitu, protože korelace mezi x_i a x_j je značná.

Indikace multikolinearity (ADSTAT):

Čj	**Vlastní čísla korel. matice λ_j**	**Číslo podmíněnosti K_j**	**Variance inflation faktor VIF_j**	**Vícenás. korel. koef. pro x_j**
1	0.0064568	461.41	83.272	0.9940
2	0.014307	208.23	94.324	0.9947
3	2.9792	1.0000	47.508	0.9894
Maximální číslo podmíněnosti *K*			: 461.41	
Scottovo kritérium multikolinearity, *M*			: 0.96119	

Závěr: Navržený model není korektní.

Vlastní čísla korel. matice λ_j. Suma vlastních čísel je rovna počtu nezávisle proměnných. Vlastní čísla blízká nule znamenají, že v datech je kolinearita. **Číslo podmíněnosti *K*:** představuje podíl největšího vlastního čísla, poděleného dotyčným vlastním číslem. Protože jsou vlastní čísla vlastně rozptyly, číslo podmíněnosti je podíl rozptylů. Je-li maximální číslo podmíněnosti $K > 1000$, pak indikuje silnou multikolinearitu. **Variance inflation factor VIF_j:** Je mírou multikolinearity a roven $1/(1-D_j)$, kde D_j se získá, když *j*-tá nezávisle proměnná je regresována vůči zbytku nezávisle proměnných. Je-li $VIF_j > 10$, pak indikuje rovněž silnou multikolinearitu.

Proměnné s vysokou *VIF* jsou předurčeny k vyloučení z modelu. **Scottovo kritérium multikolinearity:** ukazuje, že navržený model není korektní s ohledem na vazby mezi proměnnými.

(c) *Heteroskedasticita reziduí:* Cookův-Weisbergův test heteroskedasticity reziduí dokazuje, že rezidua vykazují heteroskedasticitu (nekonstantnost rozptylu), protože Cookovo-Weisbergovo testační kritérium S_f = 19.926 je větší než tabulkový kvantil $\chi^2_{0.95}(1) = 3.8415$.

Cookův-Weisbergův test heteroskedasticity S_f	: 19.926
Tabulkový kvantil, $\chi^2_{0.95}(1)$	: 3.8415
Spočtená hladina významnosti	: 0.000
Závěr: Rezidua vykazují heteroskedasticitu.	

(d) *Normalita reziduí:* Jarqueův-Berraův test normality reziduí ukazuje, že klasická rezidua vykazují Gaussovo normální rozdělení, protože Jarqueovo-Barraovo testační kritérium *L(e)* = 0.03309 je menší než tabulkový kvantil $\chi^2_{0.95}(2) = 5.9915$.

Jarqueův-Berraův test normality reziduí *L(e)*	: 0.03309
Tabulkový kvantil, $\chi^2_{0.95}(2)$	: 5.9915
Spočtená hladina významnosti	: 0.984
Závěr: Normalita je přijata.	

(e) *Autokorelace reziduí:* Waldův test autokorelace ukazuje, že klasická rezidua jsou autokorelována, protože Waldovo testační kritérium W_a = 10.320 je větší než tabulkový kvantil $\chi^2_{0.95}(1) = 3.8415$.

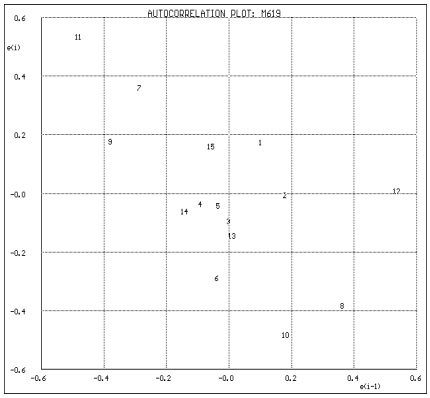
Waldův test autokorelace W_a	: 10.320
Tabulkový kvantil, $\chi^2_{0.95}(1)$	: 3.8415
Spočtená hladina významnosti	: 0.001
Závěr: Rezidua jsou autokorelována.	

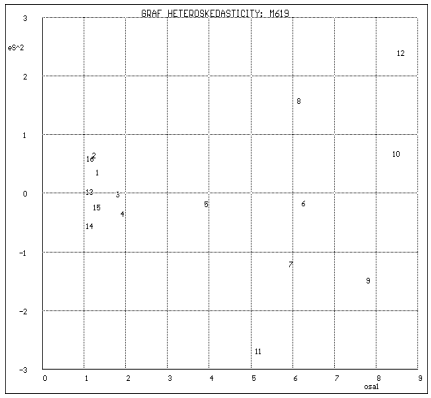
(f) *Trend v reziduích:* znaménkový test prokazuje, že znaménko klasických reziduí se dostatečně střídá, a proto rezidua nevykazují žádný trend. Absolutní hodnota testačního kritéria znaménkového testu D_t = -0.19739 je menší než tabulkový kvantil $N_{1-\alpha/2}$ = 1.6449, a proto je H_0 o nevýznamnosti trendu reziduí přijata.

Znamékový test D_t	:-0.19739
Tabulkový kvantil, $N_{1-\alpha/2}$	: 1.6449
Spočtená hladina významnosti	: 0.422
Závěr: Rezidua nevykazují trend.	

Graf autokorelace (obr. 6.5-7a) vykazuje přibližně mrak bodů reziduí.

Graf heteroskedasticity (obr. 6.5-7b) vykazuje klín, a proto rezidua vykazují heteroskedasticitu, nekonstantnost rozptylu.





Obr. 6.5-7 Vlevo: graf autokorelace, a vpravo: graf heteroskedasticity, ***ADSTAT.***

6. Konstrukce zpřesněného modelu:

(a) Po odstranění bodů č. 11 a 12 a absolutního členu β_0 byly nalezeny nové odhady parametrů zpřesněného modelu.

Par.	**Odhad**	**Směrodatná odchylka**	**H_0: $b_j = 0$ vs. H_A: b_j ... 0 t-kritérium**	**hypotéza H_0 je**	**Spočtená hlad. význam.**
b_0	0.0000	-------	-------	--------	-----
b_1	-1.1808	0.38271	-3.0854	Zamítnuta	0.010
b_2	1.2454	0.23610	5.2751	Zamítnuta	0.000
b_3	0.91667	0.15049	6.0910	Zamítnuta	0.000

Zpřesněný model (v závorce je uveden odhad směrodatné odchylky parametru)

$$y = -1.18\ (0.38)\ x_1 + 1.25\ (0.24)\ x_2 + 0.92\ (0.15)\ x_3$$

je doložen statistickými charakteristikami: *vícenásobný korelační koeficient r, koeficient determinace D a predikovaný koeficient determinace* $D_P = R^2_P$ dosáhly vesměs vysokých hodnot. *Střední kvadratická chyba predikce MEP* a *Akaikovo informační kritérium AIC* dosáhly nižších hodnot, což dokazuje lepší model než předešlý.

Vícenásobný korelační koeficient *r*	: 0.99900
Koeficient determinace 100 % *D*	: 99.800
Predikovaný koeficient determinace R^2_P	: 0.99792
Střední kvadratická chyba predikce *MEP*	: 0.0608
Akaikovo informač ní kritérium *AIC*	: -43.472

(b) Užitím statistické váhy ($w_i = 1/y_i^2$) kompenzujeme heteroskedasticitu v datech. Obdržíme nové odhady parametrů, v nichž však parametr β_1 vychází jako statisticky nevýznamný.

Par.	**Odhad**	**Směrodatná odchylka**	**H_0: $b_j = 0$ vs. H_A: b_j ... 0 *t*-kritérium**	**hypotéza H_0 je**	**Spočtená hlad. význam.**
b_0	0.0000E+00	-------	-------	--------	-----
b_1	0.05644	0.26023	0.21689	Akceptována	0.832
b_2	0.62215	0.16877	3.6863	Zamítnuta	0.004
b_3	0.36328	0.15078	2.4094	Zamítnuta	0.035

Opravený model má tvar, (v závorce je vždy uveden odhad směrodatné odchylky parametru): $y = 0.62\ (0.17)\ x_2 + 0.36\ (0.15)\ x_3$.

Jelikož došlo k významnému snížení hodnot rozhodujících kritérií, *střední kvadratické chyby predikce MEP* a *Akaikovo informačního kritéria AIC*, lze považovat tyto odhady za lepší než předešlé. *Pearsonův korelační koeficient r,* a tím pádem i *koeficient determinace D* vychází nepatrně horší než u předešlého odhadu bez statistické váhy.

Vícenásobný korelační koeficient *r*	: 0.99689
Koeficient determinace 100 % *D*	: 99.379
Predikovaný koeficient determinace R^2_P	: 0.99505
Střední kvadratická chyba predikce *MEP*	: 0.01230
Akaikovo informační kritérium *AIC*	: -62.092

7. Zhodnocení kvality modelu: porovnáním hodnot regresní diagnostiky lze snadno provést zhodnocení *regresního tripletu* dosaženého lineárního regresního modelu pro upravená data, zbavená odlehlých hodnot a upravený regresní model bez absolutního členu a metodou vážených nejmenších čtverců. Nalezený model má tvar (v závorce je vždy uveden odhad směrodatné odchylky parametru)

$$y = 0.62\ (0.17)\ x_2 + 0.36\ (0.15)\ x_3,$$

čili obsah kadmia v zrnu potravinářské pšenice je funkcí pouze obsahu kadmia ve stonku a v kořenovém systému a není funkcí obsahu kadmia v otrubách a dále nemá smysl uvádět ani zbytkový obsah kadmia v zrnu absolutního členu při nulovém obsahu kadmia ve zbytku rostlinky.

6.5.1 Úlohy na vícerozměrné lineární regresní modely

Úloha M6.01 *Vliv výchozích sloučenin na vznik fosfomolybdenové modře*

Nalezněte vyhovující vícerozměrný lineární regresní model a určete, které výchozí sloučeniny mají největší vliv na vznik fosfomolybdenové modře. Koncentrace fosforečnanů je při všech reakcích stejná. Jde o následující proměnné vlivu na koncentraci fosfomolybdenové modři *y*: přídavek molybdenanu x_1, přídavek kyseliny sírové x_2, přídavek chloridu cínatého x_3.

Data: Proměnné: přídavek x_1 [ml 5% molybdenanu], přídavek x_2 [ml koncentrované H_2SO_4], přídavek x_3 [ml 10% chloridu cínatého], koncentrace *y* [mg/l fosfomolybdenové modře].

1	1	1	30,	0.2	2	1.5	33,	1	5	0.5	28,	0.5	3	0.3	22,	0.6	0.5	0.1	17,
0.5	0.5	0.15	18,	2.5	4	0.2	31,	4.5	5	0.3	44,	5	5	3	53,	0.4	5	0.4	23,

Úloha M6.02 *Vznik dusitanů z dusičnanů v závislosti na množství komponent*

Na redukci dusičnanů mají vliv tři reakční komponenty, zinek Zn x_1, mangan Mn x_2 a glukóza x_3. (1) Nalezněte vícerozměrný lineární regresní model a vyšetřete regresní triplet. (2) Určete, která komponenta má na koncentraci vznikajících dusitanů *y* největší vliv. Koncentrace dusičnanů je při všech reakcích konstantní. (3) Jsou v datech nějaké vlivné body? (4) K jakým závěrům dospěly rankitové grafy?

Data: Obsah zinku x_1 [g], obsah manganu x_2 [g], obsah glukózy x_3 [g], koncentrace vznikajících dusitanů *y* [mg/l].

5	4	400	80,	4	3.4	250	120,	2	3.2	250	300,	4	5	350	600,	2	2.1	120	200,
3	3	200	150,	6	3.8	250	50,	4	4.2	180	300,	3	3.8	150	450,	5	4.5	150	150,

Úloha M6.03 *Vliv tří faktorů na úbytek reaktivního podílu barviva*

Byl vyšetřován vliv pH x_1, přídavku chloridu amonného x_2 a přídavku hydrogen-uhličitanu sodného x_3 na úbytek reaktivního podílu barviva y u ostazinové černi HN. Reaktivní podíl barviva je stanovován metodou HPLC. (1) Vyšetřením regresního tripletu nalezněte nejlepší model. (2) Diskutujte významnost jednotlivých parametrů v modelu. (3) Vyšetřete také vlivné body. (4) Odstraňte odlehlé hodnoty.

Data: pH x_1, přídavek NH_4Cl x_2 [ml], přídavek $NaHCO_3$ x_3 [ml], úbytek reaktivního podílu barviva y [%].

x_1	4.0	7.7	4.0	8.3	4.0	8.2	4.3	8.3	8.4	4.1	4.2	7.8
x_2	0.0	0.0	7.0	7.0	14.0	14.0	20.0	20.0	7.0	7.0	20.0	1.0
x_3	0.0	4.0	0.0	15.0	0.0	26.0	0.0	35.4	16.0	1.0	2.0	4.0
y	8.6	5.3	10.2	6.9	9.4	10.9	10.4	12.5	6.9	10.5	10.1	5.4

Úloha M6.04 *Vliv čtyř faktorů stanovení hořčíku na absorbanci AAS*

Při stanovení hořčíku ve vodě metodou plamenné AAS byla sledována závislost intenzity absorbance y při vlnové délce $\lambda = 202.5$ nm na výšce pozorování (hořáku) x_1, nasávaném množství vzorku x_2, průtoku vzduchu x_3 a průtoku acetylenu x_4. Za předpokladu platnosti vícerozměrného lineárního regresního modelu ve tvaru

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

(1) vyšetřete regresní triplet, (2) regresní diagnostiky a (3) pomocí parciálních regresních a u parciálních reziduálních grafů diskutujte významnost jednotlivých parametrů modelu, stejně jako i jejich fyzikální smysl.

Data: Výška hořáku x_1, množství vzorku x_2, průtok vzduchu x_3, průtok acetylenu x_4, měřená absorbance y.

10.4	4.4	12.3	37	0.25		10.5	4.5	7.6	38.5	0.229
...	...	...	...	...	...	...	...	...	...	...
10.1	4.1	8.3	30.6	0.264		9.6	3.9	11.2	31.9	0.275

Úloha M6.05 *Vliv tří faktorů na stabilitu průmyslové nitrocelulózy*

Byla sledována závislost stability y průmyslové nitrocelulózy podle Bergmanna a Junka na viskozitě x_1, obsahu dusíku x_2 a rozpustnosti průmyslové nitrocelulózy v etanolu x_3. Stanovte vliv těchto tří faktorů x_1, x_2, x_3 na y za předpokladu platnosti vícerozměrného lineárního regresního modelu a vyšetřete vlivné body.

Data: Viskozita x_1 [mPa.s], obsah dusíku x_2 [hm.%], rozpustnost x_3 [hm.%], stabilita nitrocelulózy y [ml/g NO].

11.67	11.94	8.30	1.50	11.20	12.06	6.60	1.60	11.00	12.05	7.10	1.50
...	...	...	...	...	...	...	...	...	...	...	...
11.20	12.00	7.70	1.34	10.90	11.96	7.74	1.31	12.15	11.97	7.90	1.48
11.60	11.98	6.85	1.45								

Úloha M6.06 *Vliv čtyř faktorů na koncentraci amoniakálního dusíku*

Je vyšetřován vliv teploty x_1, pH x_2, koncentrace celkového dusíku x_3 a koncentrace rozpuštěného kyslíku x_4 na koncentraci amoniakálního dusíku y v odtoku z dosazovací nádrže. (1) Postavte vícerozměrný lineární regresní model a vyšetřete regresní triplet. (2) Pomocí parciálních regresních grafů vyšetřete statistickou významnost jednotlivých faktorů. (3) Jsou v datech vlivné body? (4) Je nutné odstranit vybočující hodnoty?

Data: Teplota x_1, pH x_2, koncentrace celkového dusíku x_3, koncentrace rozpuštěného kyslíku x_4, koncentrace amoniakálního dusíku y.

21.0	7.2	35.0	0.1	18.5,	21.2	7.2	35.5	0.1	19.0,	21.0	7.2	37.0	0.1	20.0,
...	...	...	...	...,	...	...	...	...	...	...	...	...	...	...,

22.0	7.3	38.0	0.1	22.8,	23.0	7.3	37.0	0.1	23.5,	24.0	7.4	41.0	0.1	23.9,	
21.5	7.5	41.1	0.1	23.1,	22.5	7.5	38.5	0.1	25.5.						

Úloha M6.07 *Závislost tvrdosti oceli na obsahu uhlíku a teplotě*
Melnik (str. 52 v cit.[63]) zkoumal závislost tvrdosti uhlíkové oceli y na obsahu uhlíku x_1 a na teplotě x_2. (1) Určete vícerozměrný lineární regresní model. (2) Jsou v datech vlivné body? (3) Kterým testem spolehlivě vyšetříte statistickou významnost jednotlivých parametrů? (4) Porovnejte užitečnost parciálních regresních a parciálních reziduálních grafů. Vedou ke stejným závěrům?
Data: Obsah uhlíku x_1 [%], teplota x_2 [EF], tvrdost oceli y.

0.06	1330	555,	0.10	1220	499,	0.23	1120	588,	0.24	1030	559,
...	...	...,	...	...	...,	...	...	...,	...	...	...,
0.83	400	724,	1.00	470	703,	1.16	310	749,			

Úloha M6.08 *Závislost obsahu lipoproteinu v krevním séru na třech faktorech*
Při kvantitativní analýze lidského krevního séra ovlivňují hodnotu obsahu vysokohustotního lipoproteinu y tři proměnné, a to obsah celkového cholesterolu x_1, obsah celkového triglyceridu x_2 a konečně tzv. pre-beta komponenty x_3, které jsou buď přítomné ($x_3 = 1$), nebo nepřítomné ($x_3 = 0$), str. 141 v cit[65]. (1) Navrhněte regresní model a rozhodněte, zda (2) x_1, x_2, x_3 *samostatně* ovlivňují v predikci model jako celek, (3) x_1, x_2, x_3 *společně* ovlivňují v predikci model jako celek, (4) testujte i členy x_1x_2 a x_2x_3. Testy proveďte na hladině významnosti $\alpha = 0.05$. (5) Jsou v datech vlivné body? Je třeba odstranit vybočující hodnoty?
Data: Obsah cholesterolu x_1, obsah triglyceridu x_2, přítomnost pre-beta komponenty x_3, obsah lipoproteinu y.

287	111	0	47,	236	135	0	38,	255	98	0	47,	135	63	0	39,
...	...	...	...,	...	...	...	...,	...	...	...	...,	...	...	...	...,
326	236	1	56,	248	92	1	40,	285	153	1	58,	361	126	1	43,
248	226	1	40,	280	176	1	46,								

Úloha M6.09 *Závislost obsahu cholesterolu na věku a váze pacienta*
Výběr 25 pacientů, nemocných s hyperliprotenemií byl vyšetřován na hladinu lipidů v plasmě totálního cholesterolu y s přihlédnutím k váze x_1 a věku x_2 pacienta, str. 116 v cit[65]: (1) Navrhněte regresní model a testujte statistickou významnost jednotlivých parametrů. (2) Vyšetřete regresní triplet a soustřeďte se na odhalení vlivných bodů. (3) Jaké závěry můžete učinit z parciálních regresních a parciálních reziduálních grafů?
Data: Váha x_1 [kg] a věk pacienta x_2 [roky], obsah cholesterolu y [mg/100 ml].

84	46	354,	73	20	190,	65	52	405,	70	30	263,	76	57	451,	69	25	302,		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		
63	30	244,																	

Úloha M6.10 *Závislost vlastnosti modulu pryže na dvou proměnných*
Wood studoval závislost modulu přírodní pryže y jako funkci koncentrace řetězícího činidla x_1 a jeho teploty x_2 a navrhl regresní model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, (str. 506 v cit.[62]). (1) Ověřte tento model a navrhněte jeho případné změny. (2) Vyšetřete regresní triplet a soustřeďte se na vlivné body. (3) Jsou v datech vybočující hodnoty?
Data: Koncentrace x_1 [%], teplota x_2 [EC], modul y.

0.48	-50	3.20,	0.95	-50	3.54,	1.90	-50	4.85,	2.86	-50	6.24,
...	...	...,	...	...	...,	...	...	...,	...	...	...,
9.52	100	22.5,	14.30	100	36.52,	19.00	100	49.45,	23.80	100	51.50,

Úloha M6.11 *Vliv tří rozličných chemikálií na zbarvení pitné vody*
Ve vzorku pitné vody byly na sobě nezávislými postupy měřeny koncentrace železa Fe x_1 v mg/l, manganu Mn x_2 v mg/l, huminových látek HL x_3 v mg/l a zbarvení y, způsobené obsahem platiny v mg Pt/l. Závislost zbarvení y byla sledována na koncentracích Fe x_1, Mn x_2 a HL x_3. (1) Vyšetřením regresního tripletu nalezněte nejlepší model. (2) Využijte regresní diagnostiku a pomocí parciálních regresních a parciálních reziduálních grafů diskutujte významnost jednotlivých parametrů v modelu, jakož i jejich fyzikální smysl. (3) Jsou v datech vlivné body, a to především vybočující?
Data: Koncentrace Fe x_1 [mg/l], koncentrace Mn x_2 [mg/l], koncentrace HL x_3 [mg/l], barva y [mg Pt/l].

0.03	0.01	0.68	4.86,	0.06	0.03	0.83	8.73,	0.68	0.15	1.30	11.16,
...	...	...	...,	...	...	...	...,	...	...	...	...
0.45	0.03	0.58	7.5,								

Úloha M6.12 *Vliv parametrů na výrobu kyseliny šťavelové*
Laboratorně byl sledován vliv teploty reakční směsi x_1, koncentrace HNO_3 x_2 a doby reakce x_3 na konečný výtěžek y při výrobě kyseliny šťavelové. (1) Navrhněte regresní model, diskutujte významnost jednotlivých parametrů v modelu. (2) Jsou v datech vlivné body? (3) Určete, který parametr je statisticky významný?
Data: Teplota x_1 [EC], koncentrace HNO_3 x_2 [%], doba reakce x_3 [hod], výtěžek y [g].

32.5	44	8	0,	34.5	44	8	0,	38.0	45	8	12.3,
...	...	...	...,	...	...	...	...,	...	...	...	...,
57.1	50	18	33.2,	57.1	50	18	33.2,	57.1	50	18	33.2,

Úloha M6.13 *Vliv šesti parametrů na výtěžek destilace cyklohexanolu*
Při studiu destilační kolony byly proměřovány jednotlivé fyzikálně-chemické veličiny, ovlivňující výtěžek destilace. Pomocí lineárního regresního modelu diskutujte vliv dále v datech uvedených šesti sledovaných veličin x_1 až x_6 na koncentraci cyklohexanolu y. Testujte statistickou významnost jednotlivých parametrů. Jsou v datech vlivné body?
Data: Koncetrace cyklohexanolu v surovině x_1 [ppm], teplota na hlavě kolony x_2[EC], tlak na hlavě kolony x_3 [atm], teplota na patě kolony x_4 [EC], reflux x_5 [kg/h], odtah x_6 [kg/h], koncetrace výsledného cyklo-hexanolu v produktu y [ppm].

80	155	0.53	165	10000	9000	29,	70	156	0.55	166	11000	9000	34,
...	...	...	...	...	...	...,	...	...	...	...	...	...	...,
2020	150	0.62	163	18000	4000	200,	1760	156	0.52	166	9200	8000	327,

Úloha M6.14 *Vliv tří rozličných chemikálií na obsah látky v původním vzorku*
V laboratoři byl sledován vliv tří parametrů regresního modelu x_1, x_2, x_3 na obsah látky v původním vzorku. (1) Určete regresní model, testujte statistickou významnost jednotlivých parametrů a vyšetřete regresní triplet. (2) Jsou v datech vlivné body?
Data: Obsah složky A x_1 [%], obsah složky B x_2 [%], obsah složky C x_3 [%], obsah látky y [%].

79.3	0.5	1.2	55.8,	75.7	0.5	1.0	52.3,	76.0	0.4	1.3	53.6,
...	...	...	...,	...	...	...	...,	...	...	...	...,
80.2	0.6	1.2	58.1,	78.9	0.6	1.4	55.8,				

Úloha M6.15 *Vliv obsahu PCB v kongenerech na PCB v mateřském mléce*
V tuku mateřského mléka 74 matek byl stanoven obsah polychlorovaných bifenylů, a to jednak jako celkový obsah, tzv. suma PCB y, a jednak jako obsah PCB jednotlivých kongenerů x_1, x_2, x_3. (1) Vyšetřete lineární regresní model mezi sumou PCB y a obsahem kongenerů PCB#138 x_1, PCB#153 x_2 a PCB#180 x_3 v mateřském mléce. (2) Posuďte predikční schopnost modelu a spolehlivost predikce celkového obsahu PCB na základě znalosti tří kongenerů. (3) Vyšetřete regresní triplet a testujte statistickou významnost jednotlivých kongenerů.

Data: Obsah PCB#138 x_1 [mg/kg], obsah PCB#153 x_2 [mg/kg], obsah PCB#180 x_3 [mg/kg], suma PCB y [mg/kg].

0.184	0.218	0.228	1.737,	0.310	0.239	0.209	2.090,
...	...	...	...,	...	...	...	...,
0.082	0.092	0.034	0.574,	0.107	0.186	0.081	1.031,

Úloha M6.16 *Vliv čtyř parametrů na retenční čas eluovaného píku u GC*
Vyšetřete statistickou významnost vlivu čtyř proměnných parametrů u plynové chromatografie GC, tj. nástupní teploty teplotně programovaného režimu eluce x_1, tlaku nosného plynu na kolonu x_2, nárůstu teploty x_3 a koncentrace propanolu x_4 na retenční čas eluovaného píku y. Teplota septa, detektoru a dávkování 5 µl byly drženy na konstantních hodnotách. (1) Který z parametrů se jeví jako statisticky nevýznamný? (2) Vyšetřete také regresní triplet a odhalte vlivné body. (3) Užijte parciální regresní a parciální reziduální grafy a porovnejte závěry z nich se závěry Studentova t-testu.

Data: Velikost složky x_1 [EC], velikost složky x_2 [mPa], velikost složky x_3 [EC], velikost složky x_4 [mg/l], retenční čas y [mm].

150	1.0E+05	5	10	8.37,	150	1.0E+05	5	30	8.32,
...	...	...	...	...,	...	...	...	...	...,
170	1.5E+05	10	10	4.11,	170	1.5E+05	10	30	4.00,

Úloha M6.17 *Vliv čtyř parametrů na retenční čas eluovaného píku u GC*
Vyšetřete (1) statistickou významnost vlivu čtyř proměnných parametrů u plynové chromatografie GC, tj. nástupní teploty kolony teplotně programovaného režimu eluce x_1, tlaku nosného plynu na kolonu x_2, nárůstu teploty x_3 a koncentrace etheru x_4 na retenční čas eluovaného píku y. Teplota septa, detektoru a dávkování byly drženy na konstantních hodnotách. (2) Který z parametrů je statisticky nevýznamný? (3) Vyšetřete také regresní triplet a odhalte vlivné body. (4) Využijte i parciálních regresních grafů.

Data: Velikost složky x_1 [EC], velikost složky x_2 [mPa], velikost složky x_3 [EC], velikost složky x_4 [mg/l], retenční čas y [mm].

150	1.8	50.0	0.10	1.75	150	1.8	50.0	0.050	1.73
...	...	...	...	...	...	...	...	...	...
200	1.88	20.0	0.10	1.59	200	1.88	20.0	0.050	1.57

Úloha M6.18 *Vliv tří parametrů na extrakci dalaponu z vody do éteru*
Dalapon (kyselinu 2,2-dichlorpropionovou) lze stanovit chromatograficky GLC po extrakci z vody do éteru. Výtěžek extrakce y však závisí na pH x_1 a koncentraci NaCl x_2 ve vodné fázi, a konečně i na koncentraci dalaponu v systému x_3. (1) Posuďte statistickou významnost

vlivu jednotlivých parametrů na výtěžek extrakce, je-li zachován konstantní poměr obou fází 1 : 1. (2) Využijte také parciálních regresních a parciálních reziduálních grafů a závěry porovnejte s výsledky Studentova *t*-testu.

Data: pH x_1, koncentrace NaCl x_2 [g/100 ml], koncentrace dalaponu x_3 [mg/100 ml], výtěžek y [%].

2.1	2.0	40.2	67.5	2.1	6.0	120.6	68.9	3.2	4.0	80.4	54.2
...	...	...	...	...	...	...	...	...	...	...	...
6.8	8.0	40.2	10.4								

Úloha M6.19 *Vliv tří parametrů na obsah kadmia v potravinářské pšenici*

U vzorků potravinářské pšenice byl zjišťován obsah kadmia v zrnu y v závislosti na obsahu kadmia v otrubách x_1, ve stonku s listy x_2 a v kořenovém systému x_3 (1) Vyšetřením regresního tripletu nalezněte nejlepší vícerozměrný regresní model. (2) Využijte také regresní diagnostiku a pomocí parciálních regresních a parciálních reziduálních grafů diskutujte významnost jednotlivých parametrů v modelu stejně jako i jejich fyzikální smysl.

Data: Obsah kadmia v otrubách x_1 [mg/l], obsah kadmia ve stonku s listy x_2 [mg/l] a obsah kadmia v kořenovém systému x_3 [mg/l], obsah kadmia v zrnu y [mg/l].

1.5	1.5	1.5	1.6	1.5	1.6	1.3	1.6	2.0	1.9	2.2	2.1
2.0	2.0	2.2	2.1	6.6	7.1	7.6	8.1	7.1	8.2	6.6	7.9
7.8	9.1	7.1	8.4	8.4	10.3	7.8	10.3	8.4	9.6	8.6	9.6
8.6	10.0	9.1	10.8	9.0	12.3	10.5	13.1	10.2	13.1	11.8	15.1
1.3	1.3	1.3	1.3	1.1	1.3	1.2	1.2	1.3	1.6	1.3	1.5
1.5	1.6	1.2	1.5								

Úloha M6.20 *Vliv tří parametrů na obsah zinku v potravinářské pšenici*

U vzorků potravinářské pšenice byl zjišťován obsah zinku v zrnu y v závislosti na obsahu zinku v kořenech x_1, v otrubách x_2 a v nadzemních částech rostliny, ve stonku s listy x_3. (1) Vyšetřením regresního tripletu nalezněte nejlepší vícerozměrný regresní model. (2) Využijte k tomu regresní diagnostiku a pomocí parciálních regresních a parciálních reziduálních grafů diskutujte významnost jednotlivých parametrů v modelu stejně jako i jejich fyzikální smysl.

Data: Obsah zinku v kořenech x_1 [mg/l], v otrubách x_2 [mg/l] a ve stonku s listy x_3 [mg/l], obsah zinku v zrnu y [mg/l].

164 198 162 175,	160 198 159 169,	158 211 164 175,	162 211 162 181,
... ...	... ...	... ...	... ...
806 946 834 903,	793 912 824 927,	820 919 807 889,	

Úloha M6.21 *Vliv pěti parametrů na obsah manganu ve vodárenské vodě*

Nalezněte závislost mezi obsahem rozpuštěného manganu y na úrovni posledního odběrného horizontu vodárenské nádrže od začátku letní stagnace a pěti parametry, a to teplotou vody x_1, obsahem kyslíku x_2, pH x_3, průtokem vody vypouštěné dnovými výpusťmi x_4 a časem x_5. Ve vyšetřovaném letním období dochází k izolaci spodních vrstev vody a k tradičnímu nárůstu obsahu manganu, který je z hlediska úpravy vody nežádoucí. V pravidelných týdenních intervalech byla získána následující vodohospo-dářská data. (1) Jsou všechny parametry statisticky významné? (2) Vyšetřete regresní triplet a (3) odhalte i vlivné body.

Data: Teplota vody x_1 [EC], obsah kyslíku x_2 [mg/l], pH x_3, průtok vody vypouštěné dnovými výpusťmi x_4 [%] a časem x_5 [týden], obsah rozpuštěného manganu y [µg/l].

3.4	9.9	7.7	16.26	17	10	3.5	9.1	7.7	4.26	18	60
...	...	...	...	...	...	...	...	...	...	...	...
4.9	5.0	6.9	6.55	49	410						

Úloha M6.22 *Vliv čtyř parametrů na výkon linky granulace*
Výkon linky granulace polyolefinového prášku y při konstantních otáčkách šneku extruderu závisí na teplotě extruderu x_1, otáčkách sekačky x_2, indexu toku zpracovávaného prášku x_3 a tlaku na hlavě extruderu x_4. (1) Jsou všechny parametry statisticky významné? (2) Vyšetřete regresní triplet a odhalte i vlivné body.
Data: Teplota extruderu x_1, otáčky sekačky x_2, index toku zpracovávaného prášku x_3, tlak na hlavě extruderu x_4, výkon linky granulace y.

270	1500	1.5	17.0	5.3	270	1500	1.5	19.0	5.0	280	1600	1.5	18.0	5.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
275	1600	18.0	18.0	5.4	270	1700	23.5	19.0	5.3	275	1600	28.0	18.0	5.4

Úloha M6.23 *Predikce obsahu rozpuštěných látek v říční vodě*
Nalezněte vztah $y = f(x_1, x_2, x_3)$, pomocí kterého by bylo možno predikovat obsah rozpuštěných látek y ze známé hodnoty chemické spotřeby kyslíku CHSK-Mn x_1, známé koncentrace síranů x_2 a chloridů x_3 u většiny říčních profilů daného povodí.
Data: x_1 CHSK-Mn [mg/l], x_2 koncentrace síranů [mg/l] a x_3 chloridů [mg/l], y obsah rozpuštěných látek [mg/l].

				292.00	9.90	69.00	54.00	362.00	9.40	83.00	97.00
390.00	9.60	84.00	82.00	190.00	3.12	41.00	14.50	204.00	2.64	47.00	20.60
...	...	...	...	...	...	...	...	...	...	...	...
536.00	5.00	94.60	134.70	2046.00	4.80	119.00	878.00	305.00	4.40	58.40	39.00

Úloha M6.24 *Vliv šesti parametrů destilační kolony na výtěžek nitrobenzenu*
Vyšetřete (1) vliv šesti parametrů destilační kolony, tj. koncentrace nitrobenzenu v surovině x_1, teploty na hlavě kolony x_2, tlaku na hlavě kolony x_3, teploty na patě kolony x_4, refluxu x_5 a odtahu x_6 na výtěžek, tj. koncentraci nitrobenzenu v produktu y. (2) Nalezněte vícerozměrný regresní model. (3) Testujte statistickou významnost jednotlivých parametrů. (4) Jsou v datech vlivné body?
Data: Koncentrace nitrobenzenu v surovině [ppm] x_1, teplota na hlavě kolony [EC] x_2, tlaku na hlavě kolony [atm] x_3, teploty na patě kolony [EC] x_4, refluxu [kg/h] x_5, a odtahu [kg/h] x_6 na koncentraci nitrobenzenu v produktu [ppm] y.

14.75	151.40	0.58	161.40	8125.00	6833.00	1.00
...	...	...	...	...	...	...
54000.00	156.00	0.52	166.00	9200.00	8000.00	2465.00

Úloha M6.25 *Závislost výskytu bakterií Escherichia coli v řece*
Na základě laboratorních měření byla zjišťována závislost výskytu bakterií Escherichia coli v řece y na průtoku x_1, teplotě vody x_2, teplotě ovzduší x_3 a koncentraci amonných iontů x_4. (1) Nalezněte nejlepší regresní model. (2) Diskutujte statistickou významnost jednotlivých parametrů na základě parciálních regresních a parciálních reziduálních grafů. (3) Vyšetřete regresní triplet a odstraňte odlehlé body.
Data: Průtok x_1, teplota vody x_2, teplota ovzduší x_3 a koncentrace NH_4^+ iontů x_4, výskyt bakterií y.

6.88	-1.00	-12.00	0.55	226.00	5.18	0.00	7.00	0.65	18.00
...	...	...	...	...	...	...	...	...	...
2.80	7.00	7.00	0.30	55.00	3.52	4.00	1.00	0.40	70.00

Úloha M6.26 *Vliv čtyř parametrů na energetickou hodnotu kukuřice*
U osmi hybridů kukuřice byl stanoven obsah dusíkatých látek x_1, poměr hmotnosti klasu k hmotnosti stvolu x_2, obsah sušiny x_3 a vlákniny x_4. Úkolem je (1) stanovit, jaký mají tyto parametry vliv na výživnou energetickou hodnotu kukuřice jako krmiva y, na tzv. metabolizovanou energii pro skot. (2) Který z parametrů je statisticky nejvýznamnější? (3) Určete regresní model a vyšetřete regresní triplet.
Data: Obsah dusíkatých látek x_1, poměr hmotnosti klasu k hmotnosti stvolu x_2, obsah sušiny x_3 a vlákniny x_4, metabolizovaná energie pro skot y.

8.14	0.702	24.47	20.61	4.54	8.49	0.562	25.75	19.87	4.76
...	...	...	...	...	...	...	...	...	...
8.38	0.613	25.57	21.54	4.74	8.91	0.673	24.99	20.93	4.67

Úloha M6.27 *Vliv šesti parametrů na jednotnou klasifikaci sedimentů JKS*
Pro hodnocení "rozpojitelnosti" je užívána jednotná klasifikace sedimentů JKS. Hodnotícím kritériem je *index JKS*, jehož hodnota je stanovena na základě výsledků laboratorních zkoušek: obsahu jílových materiálů x_1 v % a karbonátů x_2 v %, objemové hmotnosti x_3 v g/cm^3 a vlhkosti x_4 v % objemu, odporu v penetraci x_5 v N/cm^3, pevnosti v prostém tlaku x_6 v MPa. (1) Vyšetřete statistickou významnost jednotlivých parametrů a regresní triplet. (2) Jsou v datech odlehlé body? (3) Lze vyloučit bod č. 39?
Data: Obsah jílových materiálů [%] x_1 a karbonátů [%] x_2, objemová hmotnost [g/cm^3] x_3, vlhkost [% objemu] x_4, odpor v penetraci [N/cm^3] x_5, pevnost v prostém tlaku [MPa] x_6, index JKS y.

74.0	7.0	2.0	38.3	126.0	1.1	99.0	49.0	3.0	2.1	28.6	199.0	1.9	101.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
49.0	7.0	2.1	31.0	177.0	2.1	101.0							

Úloha M6.28 *Vliv čtyř parametrů na práci levé srdeční komory*
Byla sledována závislost práce levé srdeční komory LVSWI y na čtyřech parametrech, a to srdečním indexu CI x_1, pulmonární vaskulární rezistenci PVRI x_2, systémové vaskulární rezistenci SVRI x_3 a spotřebě kyslíku VO_2x_4. (1) Vyšetřete statistickou významnost jednotlivých parametrů a regresní triplet. (2) Jsou v datech vlivné body? (3) Který z vyšetřovaných parametrů vyšel jako statisticky nejvýznamnější?
Data: Srdeční index CI x_1, pulmonární vaskulární rezistence PVRI x_2, systémová vaskulární rezistence SVRI x_3, spotřeba kyslíku VO_2 x_4, práce levé srdeční komory LVSWI y.

4.6	211	880	367	20	5.4	194	701	385	22
...	...	...	...	...	...	...	...	...	...
6.1	210	852	523	41	6.2	206	852	555	43

Úloha M6.29 *Vliv pěti parametrů experimentálních podmínek na výtěžek syntézy*
Syntéza 1-fenyl-3-methylpyrazolonu (FMP) se provádí dvoustupňově: v prvním stupni reaguje diketen s amoniakem a ve druhém stupni vzniklý acetoacetamid s hydrochloridem fenylhydrazinu. V průběhu řady syntéz byly měněny parametry: molární poměr amoniaku a diketenu x_1, molární poměr acetoacetamidu AAA a fenyl-hydrazinu FH x_2, reakční teplota x_3, reakční doba x_4 a čistota diketenu v % x_5. Výsledkem experimentů byl výtěžek FMP y,

vyjádřený procentem vůči teoretickému výtěžku fenylhydrazinu. (1) Určete regresní model a testujte statistickou významnost jednotlivých parametrů. (2) Jsou v datech odlehlé body? (3) Je třeba odstranit nějaké odlehlé body?

Data: Molární poměr amoniaku a diketenu x_1, molární poměr acetoacetamidu AAA a fenylhydrazinu FH x_2, reakční teplota [E C] x_3, reakční doba [min] x_4 a čistota diketenu [%] x_5, výtěžek [%] y.

1.020	1.210	60	40	60.0	73.6	1.010	1.041	90	40	70.0	81.9
...	...	...	...	...	...	...	...	...	...	...	...
1.100	1.149	40	40	95.4	93.0	1.150	1.150	30	60	95.4	91.6

Úloha M6.30 *Vliv čtyř parametrů na spotřebu kyslíku v arteriální krvi člověka*

Na jednotce intenzivní péče byla u jednoho pacienta sledována spotřeba kyslíku y v ml/min na čtyřech sledovaných parametrech, a to na množství hemoglobinu v jeho krvi x_1, na saturaci kyslíku ve smíšené venósní krvi x_2, na saturaci kyslíku v arteriální krvi x_3 a na indexu pronikání kyslíku do arteriální krve x_4. (1) Určete regresní model a vyšetřete, který parametr nejvíce ovlivňuje spotřebu kyslíku. (2) Lze zde mluvit o vybočujících hodnotách?

Data: Množství hemoglobinu v jeho krvi x_1, saturace kyslíku ve smíšené venósní krvi x_2, saturace kyslíku v arteriální krvi x_3, index pronikání kyslíku do arteriální krve x_4, spotřeba kyslíku [ml/min] y.

1.6	417.1	252.7	203.0	484.1	517.7,	1.6	416.9	241.3	197.1	484.1	531.8,
...	...	...	...	...	...	...	...	...	...	...	...
0.1	25.0	12.1	13.7	29.7	30.7,						

Úloha M6.31 *Faktory ovlivňující čtvrletní výdaje rodiny*

U dvaceti vybraných domácností byly zjištěny údaje o čtvrtletních výdajích na potraviny a nápoje y, čtvrtletním příjmu domácnosti x_1, počtu dětí x_2, průměrném věku vydělečně činných členů domácnosti x_3 a počtu členů domácnosti x_4. Rozhodněte, které proměnné významně přispívají k vysvětlení variability hodnot čtvrtletních výdajů a zkontrolujte lineární regresní model s nejlepšími vysvětlujícími proměnnými. Jsou v datech odlehlé hodnoty?

Data: Příjem x_1 [Kč], počet dětí x_2, průměrný věk x_3 [roky], počet členů x_4, výdaje y [Kč].

11172	0	55.0	1	3464	8868	0	21.0	1	1982	17414	0	49.0	1	3228
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
24920	2	33.5	4	8584	40064	3	47.0	5	16950					

Úloha M6.32 *Vliv variability příměsí na čistotu chemikálie $ZnCl_2$*

U chloridu zinečnatého bezvodého $ZnCl_2$ bylo zkoumáno, zda obsah této chemikálie y je ovlivněn příměsenými nečistotami, a to solemi alkalických kovů a zemin x_1, oxichloridy x_2 a sloučeninami železa x_3. Nalezněte regresní model této závislosti a vyšetřete statistickou významnost jednotlivých parametrů.

Data: Obsah solí alkalických kovů a zemin x_1, obsah oxychloridů x_2, obsah sloučenin železa x_3, obsah $ZnCl_2$ y [%].

0.079	0.48	0.0010	98.78	0.086	0.50	0.0020	99.47	0.076	0.46	0.0018	99.91
...	...	...	...	...	...	...	...	...	...	...	...
0.010	0.57	0.0018	99.43	0.070	0.56	0.0015	99.69	0.090	0.52	0.0016	99.99

Úloha M6.33 *Vliv čtyř parametrů na tloušťku povlaku tablet a dražé*

Kvalita potahovaných tablet a dražé se posuzuje mimo jiné také na základě tloušťky povlaku. Tloušťka povlaku y v mm závisí na mnoha ovlivňujících faktorech, z nichž posoudíme čtyři: teplotu skladování ve stupních Celsia x_1, teplotu reakce ve stupních Celsia x_2, dobu reakce v minutách x_3 a pH barevného roztoku x_4. (1) Nalezněte regresní model a vyšetřete statistickou významnost jednotlivých faktorů. (2) Jsou v datech vlivné body?

Data: Teplota skladování ve EC x_1, teplota reakce ve EC x_2, doba reakce v minutách a pH barevného roztoku x_4, tloušťka povlaku tablet y v mm.

8.9 79.7 40 4.6 0.7097	5.3 69.0 70 6.0 1.3842	4.2 82.3 70 5.3 1.4593
... ...	... ...	... ...
8.6 71.1 50 4.5 1.2583	8.8 89.8 70 4.5 1.8073	

Úloha M6.34 *Vliv teploty a obsahu prvků C, Cr a V na pevnost oceli v tahu*

Vyšetřete (1) závislost pevnosti v tahu y oceli na obsahu uhlíku x_1, chromu x_2, vanadu x_3 a teplotě temperace x_4. (2) Navrhněte regresní model a vyšetřete i statistický význam jednotlivých faktorů. (3) Jsou v datech vlivné nebo odlehlé hodnoty?

Data: Obsah uhlíku x_1 [hm%], chromu x_2 [hm%], vanadu x_3 [hm%] a teploty temperace x_4 [EC], pevnost v tahu y [GPa].

0.40	0.00	0.00	500.00	1.58	0.40	0.40	0.00	500.00	1.32
...	...	...	...	...	...	...	...	...	...
1.60	0.00	2.00	0.00	3.00					

Úloha M6.35 *Predikce studijních výsledků po 1. roce studia na technické univerzitě*

Na základě náhodného výběru studentů 2. ročníku technické univerzity je třeba hledat závislost mezi dosaženým studijním průměrem předmětů v 1. ročníku y a studijními výsledky na střední škole, tj. skóre maturitního testu z matematiky x_1, z jazyků x_2 a průměrem z matematických předmětů x_3 a jazyků či verbálních předmětů x_4 středoškolského studia. Postavte regresní model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, popisující závislost mezi y a x_1, x_2, x_3, x_4, pokuste se predikovat studijní výsledky y nových studentů, kteří právě začali studium na univerzitě, z jejich dosavadních výsledků na střední škole x_1, x_2, x_3, x_4. Testujte statistickou významnost jednotlivých parametrů.

Data: Skóre maturitního testu z matematiky x_1, z jazyků x_2, průměr z matematických předmětů x_3 a z jazyků x_4, dosažený průměr známek po 1. ročníku na technické univerzitě y.

Student	x_1	x_2	x_3	x_4	y
1	321	247	2.30	2.63	1.97
2	718	436	3.80	3.57	2.74
...	...	...	...	...	...
20	653	606	3.69	3.52	3.20

Úloha M6.36 *Dědičné vlivy na výšku 18letých mladíků*

Byla vyšetřována výška dvaceti 18letých mladíků y a výška jejich rodičů a obou prarodičů, žijících izolovaně v horské vesnici po několik generací a hledána lineární závislost mezi y a proměnnými x_1 až x_7. (1) Postavte lineární regresní model a testujte statistickou významnost jednotlivých parametrů β_0, ..., β_7. (2) Rozhodněte mezi dvěma navrženými regresními modely A a B. *Model A:* $y = f(\beta_0, ..., \beta_7)$ a *model B:* $y = f(\beta_0, ..., \beta_3)$. (3) Vypočtěte oboustranný 95 % interval spolehlivosti všech odhadnutých parametrů a vysvětlete slovně význam intervalu spolehlivosti pro β_1. (4) Predikujte výšku 18letého mladíka z dat jeho

rodičů a prarodičů: x_1 = 50.8 cm, x_2 = 152.4 cm, x_3 = 182.9 cm, x_4 = 154.9 cm, x_5 = 180.3 cm, x_6 = 157.7 cm, x_7 = 177.8 cm. (5) Jaká je průměrná výška mladíka ve věku 18 let?

Data: x_1 porodní délka chlapce, x_2 výška matky v jejím věku 18 let, x_3 výška otce v jeho věku 18 let, x_4 výška babičky z matčiny strany v jejím věku 18 let, x_5 výška dědečka z matčiny strany v jeho věku 18 let, x_6 výška babičky z otcovy strany v jejím věku 18 let, x_5 výška dědečka z otcovy strany v jeho věku 18 let, výška 18letého chlapce y [cm].

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	50.0	153.7	178.6	166.9	176.0	166.9	170.9	170.7
...	...	...	...	...	...	...	...	...
20	48.8	158.0	170.9	161.5	180.1	161.5	169.4	167.9

Úloha M6.37 *Optimalizace výroby thiaminphenicol glycinát hydrochloridu*

V laboratoři byla optimalizována technologie výroby thiamphenicol glycinát hydro-chloridu. Byl sledován vliv přebytku draselné soli chráněného glycinu x_1 [% teoretického množství], reakční doby x_2 [hodiny], poměru isopropylalkoholu k thiamphenicolu x_3 a množství přidané koncentrované kyseliny chlorovodíkové x_4 [% teoretického množství] na výtěžek reakce y [% teorie]. Vyšetřením regresního tripletu (1) nalezněte nejlepší model, (2) využijte regresní diagnostiku a (3) pomocí parciálních regresních a parciálních reziduálních grafů diskutujte statistickou významnost jednotlivých parametrů v modelu stejně jako i jejich fyzikální smysl.

Data: Přebytek draselné soli x_1 [%], reakční doba x_2 [hodiny], poměr isopropylalkoholu k thiamphenicolu x_3 a množství přidané koncentrované HCl x_4 [%], výtěžek reakce y [%].

160	6.0	0	75.7	51.8,
...	...	...	...	...
173	6.5	2.8	86.0	77.5,

Úloha M6.38 *Vyšetření ekonomické situace domácnosti*

V rámci průzkumu ekonomické situace rodin byl pořízen náhodný výběr 34 domácností, ve kterých byly zjištěny údaje o měsíčních výdajích za potraviny v tisících Kč y, počtu členů domácnosti x_1, počtu dětí x_2, průměrném věku vydělávajících členů domácnosti x_3, měsíčním příjmu domácnosti v tisících Kč x_4 a typu domácnosti x_5, dělnické jsou označené $x_5 = 0$ a zemědělské $x_5 = 1$. Vyšetřete vliv faktorů x_1, x_2, x_3, x_4 a x_5 na proměnnou y a výsledky prakticky interpretujte.

Data: Počet členů domácnosti x_1, počet dětí x_2, průměrný věk vydělávajících členů domácnosti x_3, měsíční příjem domácnosti v tisících Kč x_4 a typ domácnosti x_5, dělnické $x_5 = 0$ a zemědělské $x_5 = 1$, měsíční výdaje za potraviny v tisících Kč y.

x_1	x_2	x_3	x_4	x_5	y
2	1	55	5.59	1	2.73
...	...	...	...	...	...
6	4	33.5	24.46	2	3.79,

Úloha M6.39 *Optimalizace výroby modrého barviva*

Pokusná výroba modrého barviva MB H-3R sestávala z několika technologických kroků, produkt byl získán vykyselením volné sulfokyseliny a filtrací na kalolisu. Závěrečná fáze výroby byla vyhodnocena jako závislost poměru výtěžku modrého barviva a výchozího modrého barviva v roztoku (tj. % získané z klerátu 2) na některých technologických

faktorech: x_1 představuje koncentraci modrého barviva v g/l v roztoku před vykyselením (klerátu 2), x_2 představuje pH roztoku modrého barviva před očkováním krystaly, x_3 je spotřeba 35 % kyseliny sírové v litrech při vykyselení, x_4 je pH před filtrací, x_5 je teplota suspenze před filtrací, x_6 je koncentrace modrých látek v g/l ve filtrátech, y je výtěžek modrého barviva v procentech výchozího modrého barviva v klerátu 2. (1) Nalezněte regresní model. (2) Testujte statistickou významnost jednotlivých parametrů. (3) Jsou v datech vybočující hodnoty?

Data: Koncentrace modrého barviva MB v g/l v roztoku před vykyselením (klerátu 2) x_1, pH roztoku MB před očkováním krystaly x_2, spotřeba 35 % kyseliny sírové v litrech při vykyselení x_3, pH před filtrací x_4, teplota suspenze před filtrací x_5, koncentrace modrých látek v g/l ve filtrátech x_6, výtěžek MB v procentech výchozího MB v klerátu 2 y.

x_1	x_2	x_3	x_4	x_5	x_6	y
57.60	2.73	242.00	0.70	24.50	34.70	84.00
...	...	...	...	...	...	...
79.00	2.75	303.00	0.51	25.00	23.30	291.00

Úloha M6.40 *Závislost koncentrace kreatininu na hmotnosti, věku a výšce mužů*

Zjistěte, zda existuje závislost celkové koncentrace kreatininu y u zdravých mužů na jejich hmotnosti x_1, věku x_2 a výšce x_3. Soubor je tvořen muži, u kterých nebylo prokázáno renální onemocnění. (1) Postavte regresní model, (2) vyšetřete statistickou významnost jednotlivých parametrů. (3) Užijte také parciální regresní grafy a komentujte fyzikální význam jednotlivých parametrů.

Data: Věk x_1 [roky], hmotnost x_2 [kg], výška x_3 [cm], koncentrace kreatininu y [μmol/l].

x_1:	20	24	26	30	33	35	38	40	42	45	47	50	52	54	57	60	63	65
	67	70	72	75														
x_2:	70	68	80	79	85	82	63	91	87	88	96	92	81	102	99	78	89	83
	76	72	75	84														
x_3:	180	185	178	182	190	172	168	180	178	183	188	172	179	198	181	184	188	170
	174	178	187	179														
y:	66	68	69	72	74	73	77	80	79	85	88	91	93	97	99	89	101	107
	110	108	115	119														

Úloha M6.41 *Závislost alfa-1-globulinu na koncentraci proteinů akutní fáze*

Zjistěte, zda existuje závislost celkové koncentrace alfa-1-globulinu y na koncentraci dvou proteinů akutní fáze s elektroforetickou pohyblivostí v oblasti alfa-1-globulinů x_1 a x_2. Soubor je tvořen pacienty s idiopatickými střevními záněty v různé fázi aktivity chorobného procesu. (1) Postavte regresní model. (2) Využijte regresních diagnostik k vyřešení regresního tripletu. (3) Diskutujte statistickou významnost jednotlivých parametrů v modelu na základě parciálních regresních grafů a na základě Studentova t-testu.

Data: Koncentrace alfa-1-antitrypsinu x_1 [g/l], koncentrace orosomukoidu x_2 [g/l], koncentrace alfa-1-globulinu y [g/l].

x_1	2.5	4.4	4.9	1.8	1.9	6.3	5.6	4.2	2.3	3.9	2.6	2.0	4.8	5.5	1.2	0.9	1.7	5.3
x_2	2.0	2.9	3.1	1.6	1.3	4.5	3.9	2.8	1.8	2.5	2.3	2.0	3.0	3.7	1.1	0.8	1.9	4.0
y	5.5	7.5	8.1	4.5	3.7	11.2	9.9	7.1	5.3	6.8	5.4	4.9	8.3	9.8	3.1	2.8	4.3	9.6

Úloha M6.42 *Vícerozměrný lineární model při sledování účinnosti ČOV*

Při sledování účinnosti čistíren odpadních vod ČOV se sleduje ve výtoku několik parametrů. (1) Vyšetřením regresního tripletu nalezněte regresní model a významnost jednotlivých parametrů při sledování závislosti BSK5 (biologické spotřeby kyslíku za 5 dní) y na CHSK-Cr (chemické spotřeby kyslíku) x_1, RL (rozpuštěných látek) x_2 a N-NH_4^+ (amoniakálním dusíku) x_3 za měsíc leden. (2) Jsou v datech odlehlé body? (3) Testujte statistickou významnost jednotlivých parametrů a komentujte její fyzikální význam.

Data: CHSK-Cr x_1 [mg/l], RL x_2 [mg/l], N-NH_4^+ x_3 [mg/l], BSK5 y [mg/l].

x_1	x_2	x_3	y
320	363	23	66,
...	...	...	...
270	136	24	74,

Úloha M6.43 *Závislost délky železničních tratí na třech parametrech*

Jsou známy údaje o délce železniční tratí, počtu obyvatel, rozloze a délce silnic v některých vybraných státech Evropy. Máme zjistit jak závisí délka železničních tratí y na počtu obyvatel x_1, rozloze x_2 a délce silnic x_3. Nalezněte regresní model a s využitím regresních diagnostik prokažte jeho platnost. Jsou v datech vybočující hodnoty?

Data: K datu 31.12.1995 byla rozloha země x_1 [tisíce km^2], počet obyvatelstva x_2 [miliony], délka silnic x_3 [tisíce km] a délka železničních tratí y [km].

	x_1	x_2	x_3	y
Německo	357	81.34	642.2	41719
Rakousko	83.9	8.04	106.3	5672
...	...	...	...	...
Rumunsko	238.4	22.73	72.8	11376

Úloha M6.44 *Závislost retence LVP na čtyřech faktorech HPLC*

Při vývoji HPLC metody stanovující čistotu a obsah substance LVP ve vzorku peptidu byla získána data, charakterizující závislost retence y na čtyřech základních faktorech x_1 až x_4. Navrhněte lineární regresní model, popisující tuto závislost a stanovte vliv uvedených faktorů na retenci substance LVP.

Data: Procentuální podíl fosfátového pufru v mobilní fázi x_1 [%], teplota kolony x_2 [°C], objemový průtok mobilní fáze x_3 [ml. min^{-1}], pH mobilní fáze x_4, retenční čas y.

x_1	x_2	x_3	x_4	y
88.5	35.0	1.5	6.5	27.8
...	...	...	...	...
90.0	40.0	1.3	3.0	20.9

6.6 Kontrolní hodnoty (ADSTAT, NCSS2000)

Kontrolní hodnoty v kapitole Lineární regresní modely nejsou v žádném případě konečné výsledky nalezeného regresního modelu. Protože vyšetřování regresního tripletu je tvořivá činnost, není zde čtenáři prozrazen cíl, ke kterému má dospět, ale je poskytnut pouze určitý mezivýsledek těsně před cílem. V každém případě by čtenář měl nalézt ještě lepší regresní

model s lepšími statistickými charakteristikami. Kontrolní hodnoty zde představují jistou míru nápovědy.

6.6.1 Jednorozměrné lineární regresní modely

J6.01 *R = 0.9983, D = 99.66 %, y = 26079 (13, Z) - 40.1 (0.4)× x, 1 o, 3 e*

J6.02 *R = 0.9994, D = 99.88 %, y = 0.0063 (0.0056, A) + 0.9828 (0.014)× x, 1 o, 0 e*

J6.03 *R = 0.9604, D = 92.23 %, c = -2.8(0.3, Z) + 2.304E-04(1.93E-04)× t, 2 o, 1 e*

J6.04 *(a) R = 0.9929, D = 98.59 %, y = 0.47 (0.17, Z) + 0.936 (0.023)× x, 3o, 3 e, (b) R = 0.9891, D = 97.83 %, y = 0.39 (0.53, A) + 0.977 (0.034)× x, 2o, 2 e*

J6.05 *(a) R = 0.9993, D = 99.87 %, y = -1.3 (0.3, Z) + 2.71 (0.03)× x, 1 o, 1 e, (b) R = 0.9999, D = 99.97 %, y = -0.7 (0.2, Z) + 3.82 (0.02)× x, 1o, 1 e, (c) R = 0.9998, D = 99.96 %, y = -0.9 (0.3, Z) + 4.10 (0.03)× x, 1 o, 1 e*

J6.06 *R = 0.7808, D = 60.96 %, y = -1.12(0.19, Z) + 1.32E-03(1.9E-04) × x, 1 o, 1 e*

J6.07 (a) R = 0.9858, D = 97.17 %, *y* = -1.9 (1.8, A) + 2.80 (0.17)× x, 0 o, 1 e, *(b) R = 0.9891, D = 97.83 %, y = -2.3 (1.6, A) + 2.92 (0.15)× x, 0 o, 1 e*

J6.08 *(a) R = 0.9626, D = 92.66 %, y = 60.0 (1.6, Z) - 5.16 (0.34)× x, 0 o, 1 e, (b) R = 0.9321, D = 86.88 %, y = 52.6 (1.6, Z) - 0.37 (0.03)× x, 0 o, 1 e*

J6.09 *1. den: (a) R = 0.7078, D = 50.09 %, y = 19.47 (0.37, Z) - 0.087 (0.031)× x, 1o, 0 e, (b) R = 0.8451, D = 71.42 %, y = 19.31 (0.18, Z) - 0.066 (0.015)× x, 1o, 0 e, (c) R = 0.8970, D = 80.46 %, y = 19.23 (0.13, Z) - 0.065 (0.011)× x, 0o, 0 e, (d) R = 0.8872, D = 78.72 %, y = 19.22 (0.14, Z) - 0.063 (0.012)× x, 0o, 0 e, (e) R = 0.8833, D = 78.03 %, y = 19.26 (0.14, Z) - 0.064 (0.012)× x, 0o, 0 e,*

2. den: (a) R = 0.8130, D = 66.09 %, y = 17.84 (0.18, Z) + 0.050 (0.013)× x, 1 e,(b) R = 0.7886, D = 62.19 %, y = 17.80 (0.20, Z) + 0.053 (0.015)× x, 1 e, (c) R = 0.7787, D = 60.64 %, y = 17.84 (0.20, Z) + 0.051 (0.015)× x, 1 e, (d) R = 0.7610, D = 57.92 %, y = 17.78 (0.23, Z) + 0.055 (0.017)× x, 1 e, (e) R = 0.8072, D = 65.15 %, y = 17.81 (0.18, Z) + 0.052 (0.014)× x, 1o, 1 e

J6.10 *(a) R = 0.9966, D = 99.33 %, y = 1.01 (0.15, Z) - 4.57 (0.13)× x, 0 o, 0 e, (b) R = 0.9771, D = 95.47 %, y = 9.24 (0.39, Z) - 4.59 (0.33)× x, 0 o, 0 e,*

J6.11 *R = 0.9999, D = 99.99 %, y = 0.938 (0.000, A) + 1.00E-05(3.90E-08)× x, 1 o, 0 e*

J6.12 *R = 0.9796, D = 95.96 %, y = 0.094(0.013, Z) + 4.459(0.133, Z)× x*

J6.13 *(a) R = 0.9996, D = 99.92 %, y = -2.43 (0.70, Z) + 1.01 (0.01)× x, 1o, (b) R = 0.9994, D = 99.88 %, y = 2.52 (0.84, Z) + 0.99 (0.01)× x, 1 o, (c) R = 0.9995, D = 99.90 %, y = 3.22 (0.77, Z) + 1.01 (0.01)× x, 1 o.*

J6.14 *R = 0.2702, D = 7.30%, y = 183.6 (43.9, Z) + 0.480 (0.494, A)× x, model není statisticky významný, 1o, 1e,*

J6.15 *(a) R = 0.9914, D = 98.29%, y = -0.011(0.049, A) + 0.174 (0.008, Z)× x, 0o, 0e, (b) R = 0.9924, D = 98.49%, y = 0.012(0.059, A) + 0.224 (0.010, Z)× x, 0o, 0e,*

J6.16 *(a) R = 0.9858, D = 97.17%, y = -1.85 (1.77, A) + 2.80 (0.17, Z)× x, 0o, 1e, (b) R = 0.9891, D = 97.83%, y = -2.34 (1.61, A) + 2.93 (0.15, Z)× x, 0o, 1e,*

J6.17 *(a) R = 0.5744, D = 32.99%, y = 1.64 (1.26, A) + 0.23 (0.09, Z)× x, 2o, 2e, (b) R = 0.6698, D = 44.86%, y = 0.81 (0.74, A) + 0.17 (0.05, Z)× x, 1o, 1e,*

J6.18 *(a) R = 0.9991, D = 99.81%, y = 0.07 (0.01, Z) + 0.025 (0.000, Z)× x, 0o, 1e, (b) R = 0.9979, D = 99.58%, y = 0.074 (0.016, Z) + 0.031 (0.000, Z)× x, 1o, 1e, (c) R = 0.9983, D = 99.67%, y = 0.06 (0.01, Z) + 0.025 (0.000, Z)× x, 1o, 0e, (d) R = 0.9969, D = 99.37%, y = 0.03 (0.01, Z) + 0.033 (0.000, Z)× x, 0o, 0e, (e) R = 0.9965, D = 99.30%, y = 0.06 (0.02, Z) + 0.034 (0.001, Z)× x, 1o, 0e, (f) R = 0.9979, D = 99.58%, y = 0.046 (0.010, Z) + 0.030 (0.000, Z)× x, 1o, 0e,*

J6.19 *(a) R = 0.9347, D = 87.37%, y = -1.43 (1.63, A) + 6.9 (1.0, Z)× x, 1o, 0e, (b) R = 0.9782, D = 95.68%, y = -2.4 (1.4, A) + 10.7 (0.9, Z)× x, 1o, 1e,*

J6.20 *(a) R = 0.9732, D = 94.72%, y = -16.5 (11.2, A) + 76.8 (6.9, Z)× x, 0o, 0e, (b) R = 0.9618, D = 92.50%, y = 41.4 (12.5, Z) + 70.9 (7.6, Z)× x, 0o, 0e,*

J6.21 *(a) R = 0.9852, D = 97.05%, y = -21.9 (6.6, Z) + 0.927 (0.051, Z)× x, 1o, 0e, (b) R = 0.9860, D = 97.22%, y = -9.4 (2.8, Z) + 0.41 (0.02, Z)× x, 1o, 0e,*

J6.22 *R = 0.9996, D = 99.91%, y = 0.45 (0.09, Z) + 0.32 (0.00, Z)× x, 0o, 1e,*

J6.23 *R = 0.9907, D = 98.14%, y = 66.8 (2.2, Z) + 1.6 (0.1, Z)× x, 1o, 2e,*

J6.24 *R = 0.6135, D = 37.64%, y = 7.3 (1.0, Z) + 0.12 (0.04, Z)× x, 1o, 0e,*

J6.25 *R = 0.8842, D = 78.18%, y = 90.7 (17.4, Z) + 1.4 (0.2, Z)× x, 1o, 2e,*

6.6.2 Validace nové analytické metody

V6.01 *R = 0.9990, D = 99.80 %, y = 0.0010 (0.0007, A) + 0.973 (0.014)×x, 0 o, 1 e*

V6.02 *R = 0.9991, D = 99.82 %, y = -1.13 (4.35, A) + 1.005 (0.010)× x, 1 o, 3 e*
V6.03 *(a) R = 0.9785, D = 95.75 %, y = -0.36 (0.50, A) + 0.988 (0.058)×x, 2 o, 1e, (b) R = 0.9772, D = 95.48 %, y = -0.32 (0.07, Z) + 0.708 (0.043)×x, 2 o, 1e*
V6.04 *R = 0.9990, D = 99.82 %, y = -0.011(0.019, A) + 1.006(0.006)× x, 4 o, 0 e*
V6.05 *R = 0.9625, D = 92.64 %, 1 o, 0 e*
V6.06 *R = 0.9987, D = 99.75 %, y = 0.0010 (0.0015, A) + 0.982 (0.016)× x, 0 o, 0 e*
V6.07 *R = 0.9996, D = 99.92 %, y = -0.14 (0.94, A) + 0.980 (0.010)× x, 1 o, 1 e*
V6.08 *R = 0.4475, D = 20.03 %, y = 12.27(10.31, A) + 0.616(0.318)× x, 1 o, 1 e*
V6.09 *R = 0.9909, D = 98.18 %, y = -1.71 (1.56, A) + 1.080 (0.037)× x, 2 o, 2 e*
V6.10 *R = 0.9770, D = 95.44 %, y = -0.030 (0.081, A) + 0.923 (0.049)× x, 2 o, 1 e*
V6.11 *R = 0.9763, D = 95.31 %, y = 3.5 (4.2, A) + 0.956 (0.057)× x, 1 o, 1 e*
V6.12 *R = 0.9735, D = 94.76 %, y = 1.22 (1.61, A) + 0.959 (0.053)× x, 1 o, 1 e*
V6.13 *R = 0.9978, D = 99.56 %, y = 0.388 (0.060, A) + 0.312 (0.004)× x, 2 o, 4 e*
V6.14 *R = 0.9959, D = 99.17 %, y = 0.010 (0.006, A) + 0.922 (0.020)× x, 3 o, 2 e*
V6.15 *(a) R = 0.8848, D = 78.28 %, y = -9.7 (8.9, A) + 1.227 (0.229)× x, 1 o, 1 e, (b) R = 0.8429, D = 71.05 %, y = 15.2 (5.2, A) + 0.592 (0.134)× x, 0 o, 1 e*
V6.16 *R = 0.9999, D = 99.99 %, y = 0.66 (1.43, A) + 0.9975 (0.001)× x, 2 o, 2 e*
V6.17 *R = 0.8208, D = 67.37 %, y = 7.84 (3.40, A) + 0.700 (0.126)× x, 1 o, 3 e*
V6.18 *R = 0.9718, D = 94.44 %, y = -3.59 (6.06, A) + 1.048 (0.080)× x, 1 o, 1 e*
V6.19 *R = 0.9816, D = 96.36 %, y = -0.07 (1.05, A) + 1.050 (0.072)× x, 1 o, 1 e*
V6.20 *R = 0.9718, D = 94.44%, y = -3.6 (6.1, A) + 1.05 (0.08, Z)× x, 1o, 1e,*
V6.21 *R = 0.9816, D = 96.36%, y = -0.07 (0.78, A) + 1.05 (0.07, Z)× x, 1o, 1e,*
V6.22 *R = 0.9676, D = 93.63%, y = -0.66 (3.14, A) + 0.96 (0.05, Z)× x, 2o, 2e,*
V6.23 *R = 0.9950, D = 99.01%, y = 0.048 (0.049, A) + 0.952 (0.023, Z)× x, 2o, 2e,*
V6.24 *R = 0.9999, D = 99.99%, y = 0.163 (0.397, A) + 0.993 (0.003, Z)× x, 1o, 2e,*
V6.25 *R = 0.9081, D = 99.63%, y = 1.1 (0.8, A) + 0.956 (0.024, Z)× x, 0o, 1e,*

6.6.3 Úlohy na lineární a nelineární kalibraci

K6.01 *nelin., x_D = 0.029, x^* = 0.34, 0.40, 0.53, 0.88; ⟨L_D, L_H⟩: ⟨0.32, 0.36⟩, ⟨0.39, 0.41⟩, ⟨0.52, 0.88⟩, ⟨0.64, --⟩,*
K6.02 *x_c = 0.0019, x_D nelze, x^* = 0.051, 0.075, 0.230; ⟨L_D, L_H⟩: ⟨0.048, 0.054⟩, ⟨0.072, 0.078⟩, ⟨0.227, 0.233⟩,*
K6.03 *x_D = 0.039, c^* = 0.011, 0.262, 0.698, 1.402; ⟨L_D, L_H⟩: ⟨0.0, 0.030⟩, ⟨0.24, 0.28⟩, ⟨0.66, 0.74⟩, ⟨1.33, 3.63⟩,*
K6.04 *x_c = 0.0017, x_D nelze, x^* = 0.067, 0.178, 0.301; ⟨L_D, L_H⟩: ⟨0.060, 0.074⟩, ⟨0.17, 0.18⟩, ⟨0.29, 0.31⟩,*
K6.05 *x_D = 0.016, c^* = -0.018, 0.041, 0.219, 0.455; ⟨L_D, L_H⟩: ⟨-0.033, -0.003⟩, ⟨0.027, 0.056⟩, ⟨0.206, 0.232⟩, ⟨0.441, 0.470⟩,*
K6.06 *nelin., x_D = 0.015, c^* = 0.047, 0.138, 0.229, 0.317; ⟨L_D, L_H⟩: ⟨0.042, 0.052⟩, ⟨0.132, 0.145⟩, ⟨0.220, 0.237⟩, ⟨0.309, 0.325⟩,*
K6.07 *(a) x_D = 4.9, c^* = 18.3, 41.0, 63.7; ⟨L_D, L_H⟩: ⟨13.2, 23.4⟩, ⟨36.0, 46.9⟩, ⟨58.6, 68.9⟩, (b) x_D = 5.5, c^* = 15.2, 28.2, 64.4; ⟨L_D, L_H⟩: ⟨10.4, 20.0⟩, ⟨23.6, 32.8⟩, ⟨60.0, 68.9⟩,*
K6.08 *x_c = 0.0061, x_D nelze, x^* = -0.054, 0.0086, 0.034; ⟨L_D, L_H⟩: ⟨-0.081, -0.028⟩, ⟨-0.013, 0.029⟩, ⟨0.013, 0.054⟩,*
K6.09 *x_c = 0.0027, x_D nelze, x^* = 0.038, 0.086, 0.133; ⟨L_D, L_H⟩: ⟨0.023, 0.053⟩, ⟨0.070, 0.100⟩, ⟨0.118, 0.148⟩,*
K6.10 *x_D = 0.2, c^* = 21.7, 39.1, 65.2, 82.7; ⟨L_D, L_H⟩: ⟨21.4, 21.9⟩, ⟨38.8, 39.3⟩, ⟨65.0, 65.5⟩, ⟨82.4, 82.9⟩,*
K6.11 *x_D = 0.3, c^* = 2.8, 6.3, 9.2; ⟨L_D, L_H⟩: ⟨2.6, 3.1⟩, ⟨6.1, 6.5⟩, ⟨8.9, 9.4⟩,*
K6.12 *x_D = 1.0, c^* = 18.0, 12.6; ⟨L_D, L_H⟩: ⟨17.3, 18.8⟩, ⟨11.9, 13.4⟩,*
K6.13 *x_D = 2.1, c^* = 4.3, 24.0, 43.6, 57.6; ⟨L_D, L_H⟩: ⟨2.4, 6.3⟩, ⟨22.1, 25.8⟩, ⟨41.8, 45.4⟩, ⟨55.7, 59.5⟩,*
K6.14 *x_D = 2.0, c^* = 6.1, 17.7, 34.9, 48.5; ⟨L_D, L_H⟩: ⟨2.9, 9.3⟩, ⟨14.5, 20.9⟩, ⟨31.7, 38.0⟩, ⟨45.2, 51.8⟩,*
K6.15 *nelin., 1 uzel, x_D = 3.8, c^* = 44.3, 71.2, 123.0, 181.9; ⟨L_D, L_H⟩: ⟨43.2, 45.3⟩, ⟨69.9, 72.4⟩, ⟨121.5, 124.5⟩, ⟨180.0, 183.9⟩,*
K6.16 *nelin., 2 uzly, x_D = 0.4, c^* = 5.3, 10.4, 12.8, 57.5; ⟨L_D, L_H⟩: ⟨5.1, 5.4⟩, ⟨10.1, 10.6⟩, ⟨12.5, 13.1⟩, ⟨25.2, 27.2⟩,*
K6.17 *x_D = 0.025, c^* = 0.162, 0.310, 0.459, 0.695; ⟨L_D, L_H⟩: ⟨0.156, 0.168⟩, ⟨0.303, 0.316⟩, ⟨0.452, 0.466⟩, ⟨0.687, 0.703⟩,*
K6.18 *x_c = 0.00014, x_D nelze, c^* = 0.0024, 0.0035, 0.0051, 0.0065; ⟨L_D, L_H⟩: ⟨0.0021, 0.0027⟩, ⟨0.0033, 0.0038⟩, ⟨0.0048, 0.0053⟩, ⟨0.0062, 0.0068⟩,*
K6.19 *x_D = 0.0231, c^* = 0.0439, 0.0462, 0.0472, 0.0523; ⟨L_D, L_H⟩: ⟨0.0407, 0.0470⟩, ⟨0.0430, 0.0493⟩, ⟨0.0440, 0.0504⟩, ⟨0.0490, 0.0550⟩,*
K6.20 *x_D = 24, c^* = 479, 1291, 3126; ⟨L_D, L_H⟩: ⟨309, 649⟩, ⟨1137, 1445⟩, ⟨2979, 3274⟩,*
K6.21 *x_D = 0.043, c^* = 0.158, 0.954; ⟨L_D, L_H⟩: ⟨0.11, 0.21⟩, ⟨0.90, 1.00⟩,*

K6.22 x_D = 5.2, c^* = 59.5, 60.1, 60.3; ⟨L_D, L_H,: ⟩59.3, 59.7,, ⟩59.9, 60.3,, ⟩60.1, 60.5,
K6.23 x_c = 0.26, x_D = -0.17, c^* = 0.717, 0.718; ⟨L_D, L_H,: ⟩0.715, 0.719,, ⟩0.716, 0.720,
K6.24 x_D = 0.016, c^* = 0.316, 0.358; ⟨L_D, L_H,: ⟩0.300, 0.331,, ⟩0.343, 0.374,
K6.25 x_D = 0.07, c^* = 0.43, 0.63, 1.38, 2.25, 3.00, 4.42; ⟨L_D, L_H,: ⟩0.36, 0.51,, ⟩0.55, 0.70,, ⟩1.30, 1.45,, ⟩2.18, 2.32,, ⟩2.92, 3.07,, ⟩4.34, 4.50,
K6.26 x_D = 0.3, c^* = 1.9, 8.7, 18.1, 35.8; ⟨L_D, L_H,: ⟩1.8, 2.0,, ⟩8.6, 8.8,, ⟩18.0, 18.2,, ⟩35.7, 36.0,
K6.27 x_D = 4.2, c^* = 0.7, 63.0; ⟨L_D, L_H,: ⟩-4.0, 5.3,, ⟩58.5, 67.4,
K6.28 *nelin., 2 uzly,* x_D = 80.1, c^* = 215.4, 307.0, 399.1; ⟨L_D, L_H,: ⟩181.8, 246.4,, ⟩277.1, 334.8,, ⟩372.3, 424.3,
K6.29 x_D = 0.43, c^* = 0.31, 1.81, 11.35; ⟨L_D, L_H,: ⟩-0.04, 0.66,, ⟩1.48, 2.15,, ⟩11.06, 11.64,
K6.30 x_D = 187.9, c^* = 96.3, 232.8, 393.0; ⟨L_D, L_H,: ⟩--, 119.8,, ⟩-45.7, 564.1,, ⟩319.0, 438.7,
K6.31 x_D = 2.0, c^* = 33.2, 44.8; ⟨L_D, L_H,: ⟩32.0, 34.3,, ⟩43.7, 45.9,
K6.32 x_D *nelze,* c^* = 1.9, 3.4; ⟨L_D, L_H,: ⟩0.99, --,, ⟩2.8, 3.9,
K6.33 x_D = 0.016, c^* = -0.018, 0.041, 0.219, 0.455; ⟨L_D, L_H,: ⟩-0.033, -0.003,, ⟩0.027, 0.056,, ⟩0.206, 0.232,, ⟩0.441, 0.470,
K6.34 x_D = 5.5, c^* = 8.3, 11.4, 18.5; ⟨L_D, L_H,: ⟩4.1, 12.3,,⟩7.5, 15.3,, ⟩14.5, 22.6,
K6.35 x_D = 9.6, c^* = 54.2, 76.6, 132.3, 733.5; ⟨L_D, L_H,: ⟩43.2, 65.1,, ⟩65.7, 87.4,, ⟩121.5, 143.0,, ⟩722.1, 744.9,
K6.36 x_D = 19.0, c^* = 59.7, 129.0, 656.1; ⟨L_D, L_H,: ⟩24.2, 9.5,, ⟩93.6, 164.3,, ⟩620.7, 691.7,
K6.37 x_D = 0.18, c^* = 1.13, 2.57, 4.93; ⟨L_D, L_H,: ⟩1.1, 1.2,, ⟩2.5, 2.6,, ⟩4.8, 5.1,
K6.38 x_D = 0.034, c^* = 0.185, 0.534, 0.758; ⟨L_D, L_H,: ⟩0.18, 0.19,, ⟩0.53, 0.54,, ⟩0.75, 0.77,
K6.39 x_D = 0.0146, c^* = 0.267, 0.317, 0.359; ⟨L_D, L_H,: ⟩0.255, 0.279,, ⟩0.305, 0.329,, ⟩0.347, 0.370,
K6.40 *nelin., 2 uzly,* x_D = 0.018, c^* = 0.016, 0.046; ⟨L_D, L_H,: ⟩0.0060, 0.0224,, ⟩0.0405, 0.0518,
K6.41 x_D = 0.0019, c^* = 0.0136, 0.0176; ⟨L_D, L_H,: ⟩0.0124, 0.0148,, ⟩0.0165, 0.0188,
K6.42 x_D = 2.1, c^* = 70.9, 120.1; ⟨L_D, L_H,: ⟩69.2, 72.7,, ⟩118.3, 121.8,
K6.43 x_D = 20.3, c^* = 56.6, 175.1; ⟨L_D, L_H,: ⟩33.3, 79.8,, ⟩150.7, 199.8,
K6.44 x_D = 0.40, c^* = 0.628, 2.019; ⟨L_D, L_H,: ⟩0.28, 0.97,, ⟩1.67, 2.37,
K6.45 x_D = 0.78, c^* = 11.2, 24.1, 34.2; ⟨L_D, L_H,: ⟩10.87, 11.49,, ⟩23.64, 24.58,, ⟩32.60, 36.36,
K6.46 x_D = 0.016, c^* = 0.0067, 0.122, 0.406; ⟨L_D, L_H,: ⟩-0.0020, 0.014,, ⟩0.114, 0.131,, ⟩0.388, 0.422,
K6.47 x_D = 0.092, c^* = 0.111, 0.196, 0.451; ⟨L_D, L_H,: ⟩0.031, 0.190,, ⟩0.119, 0.272,, ⟩0.378, 0.526,
K6.48 *nelin., 2 uzly,* x_D = 0.61, c^* = 3.04, 4.93, 6.43; ⟨L_D, L_H,: ⟩3.02, 3.07,, ⟩4.91, 4.96,, ⟩6.40, 6.46,
K6.49 *(a) Prosinec 1994:* x_D = 0.94, R = 0.8848, D = 78.29 %, y = 6.80E-03 (6.73E-03, *A*) + 0.0253 (0.0031)×x, *(b) Leden 1995:* x_D = 0.52, R = 0.9135, D = 83.45 %, y = 2.39E-03 (4.61E-03, *A*) + 0.0322 (0.0034)×x
K6.50 x_D = 0.10, c^* = 0.079; ⟨L_D, L_H,: ⟩-0.042, 0.197,
K6.51 *(a)* x_D = 0.16, c^* = 1.61, 5.70; ⟨L_D, L_H,: ⟩5.49, 5.90,, ⟩1.40, 1.82,, *(b)* x_D = 0.34, c^* = 1.87, 6.43; ⟨L_D, L_H,: ⟩1.43, 2.31,, ⟩5.99, 6.87,, *(c)* x_D = 0.30, c^* = 1.36, 5.13; ⟨L_D, L_H,: ⟩0.97, 1.75,, ⟩4.75, 5.52,
K6.52 x_D = 0.197, c^* = 0.710, 2.359, 3.184; ⟨L_D, L_H,: ⟩0.56, 0.86,, ⟩2.22, 2.50,, ⟩3.04, 3.33,
K6.53 x_D = 4.87, c^* = 18.27, 41.00, 63.72; ⟨L_D, L_H,: ⟩13.17, 23.35,, ⟩36.02, 45.98,, ⟩58.55, 68.92,
K6.54 *2 uzly,* x_D = 0.025, c^* = 0.162, 0.310, 0.481, 0.695; ⟨L_D, L_H,: ⟩0.156, 0.168,, ⟩0.303, 0.316,, ⟩0.474, 0.487,, ⟩0.687, 0.703,
K6.55 *1 uzel,* x_D = 0.038, y_D = 45.6, c^* = 0.053, 0.138, 0.159, 0.323; ⟨L_D, L_H,: ⟩0.039, 0.066,, ⟩0.121, 0.158,, ⟩-, -,, ⟩0.299, 0.345,
K6.56 x_D = 0.021, y_D = 1.83, c^* = 0.095, 0.433, 0.632; ⟨L_D, L_H,: ⟩0.079, 0.112,, ⟩0.418, 0.448,, ⟩0.617, 0.647,
K6.57 *bez 1o 1 uzel,* x_D = 0.21, y_D = 29.3, c^* = 0.238, 1.356, 2.033; ⟨L_D, L_H,: ⟩0.148, 0.292,, ⟩1.27, 1.44,, ⟩1.87,-,
K6.58 *1 uzel,* x_D = 18.44, y_D = 3.0, c^* = 90.9, 226.3, 316.7; ⟨L_D, L_H,: ⟩87.3, 94.3,, ⟩221.7, 230.8,, ⟩310.4, 323.1,
K6.59 *(a)* x_D = 8.3, y_D = 0.04, c* = 13.9, 30.7, 74.8, 187.6; ⟨L_D, L_H,: ⟩3.2, 24.5,, ⟩20.1, 41.2,, ⟩64.3, 85.3,, ⟩174.8, 200.5,, *(b)* x_D = 10.5, y_D = 0.05, c^* = 13.2, 30.1, 74.6, 188.3; ⟨L_D, L_H,: ⟩0, 26.8,, ⟩16.7, 43.5,, ⟩61.3, 88.0,, ⟩172.0, 204.7,
K6.60 x_D = 0.112, y_D = 0.026, c^* = 0.54, 2.39, 3.67; ⟨L_D, L_H,: ⟩0.40, 0.68,, ⟩2.25, 2.52,, ⟩3.53, 3.80,
K6.61 x_D = 0.013, y_D = 0.005, c^* = 0.65, 1.04, 0.25; ⟨L_D, L_H,: ⟩0.64, 0.66,, ⟩1.03, 1.05,, ⟩0.24, 0.26,
K6.62 *(a)* x_D = 0.91, y_D = 0.032, c^* = 1.47, 13.18, 20.46, 25.55; ⟨L_D, L_H,: ⟩0.73, 2.20,, ⟩12.53, 13.82,, ⟩19.79, 21.12,, ⟩24.84, 26.27,, *(b)* x_D = 2.11, y_D = 0.090, c^* = 0.95, 11.89, 18.69, 23.45; ⟨L_D, L_H,: ⟩0, 2.72,, ⟩10.33, 13.44,, ⟩17.12, 20.26,, ⟩21.80, 25.11,, *(c) 1 uzel,* x_D = 0.72, y_D = 0.045, c^* = 0.86, 11.72, 18.55, 24.18; ⟨L_D, L_H,: ⟩0.51, 1.15,, ⟩11.55, 11.90,, ⟩18.37, 18.73,, ⟩23.98, 24.38,, *(d)* x_D = 1.62, y_D = 0.052, c^* = 1.56, 12.93, 20.00, 24.95; ⟨L_D, L_H,: ⟩0.21, 2.89,, ⟩11.76, 14.11,, ⟩18.80, 21.21,, ⟩23.67, 26.24,,
K6.63 x_D = 0.015, y_D = 0.020, c^* = <x_D, 0.245, 0.531; ⟨L_D, L_H,: ⟩<x_D,, ⟩0.228, 0.262,, ⟩0.515, 0.548,
K6.64 *2 uzly,* x_D = 0.642, y_D = 0.020, c^* = 0.663, 10.12, 38.27; ⟨L_D, L_H,: ⟩0.34, 0.95,, ⟩9.40, 10.88,, ⟩36.26, 40.91,

K6.65 *1 uzel, x_D = 6.29, y_D = 0.006, c^* = 9.50, 37.23, 72.73; ⟨L_D, L_H,: ⟩7.0, 11.4,, ⟩35.6, 38.9,, ⟩69.8, 126.5,*
K6.66 *2 uzly, x_D = 51.6, y_D = 20.35, c^* = 90.22, 131.7, 183.5; ⟨L_D, L_H,: ⟩87.8, 92.7,, ⟩129.9, 133.5,, ⟩181.8, 185.2,*
K6.67 *2 uzly, x_D = -90.9, y_D = 6.9, x_c = 6.7, y_c = 0.530, c^* = 54.07, 152.5, 198.6; ⟨L_D, L_H,: ⟩0, 55.65,, ⟩151.7, 153.3,, ⟩197.9, 199.3,*

6.6.4 Úlohy na polynomické regresní modely

L6.01 *m = 2*
L6.02 *m = 3*
L6.03 *m = 2*
L6.04 *m = 2*
L6.05 *m = 3*
L6.06 *m = 2*
L6.07 *m = 2*
L6.08 *m = 2*
L6.09 *m = 1*
L6.10 *m = 3*
L6.11 *m = 3*
L6.12 *m = 3*
L6.13 *m = 3*
L6.14 *m = 3*
L6.15 *m = 1*
L6.16 *m = 2*
L6.17 *m = 3*
L6.18 *m = 1*
L6.19 *m = 1*
L6.20 *m = 2*
L6.21 *m = 2*
L6.22 *m = 2*
L6.23 *m = 3*
L6.24 *m = 3*
L6.25 *m = 3*
L6.26 *m = 2*
L6.27 *m = 2*
L6.28 *m = 3*
L6.29 *m = 4*
L6.30 *m = 3*
L6.31 *m = 4*
L6.32 *m = 3*

6.6.5 Vícerozměrné lineární regresní modely

M6.01 *R = 0.9622, D = 92.58 %, s(e) = 3.79, y = 17.11 (2.42, Z) + 3.79 (0.97, Z) x_1 + 0.86 (78, A) x_2 + 5.37 (1.60, Z) x_3, 3 o, 1 e*
M6.02 *R = 0.9006, D =81.11 %, s(e) = 93.0, y = -58.44 (144.0, A) - 137.2 (30.6, Z) x_1 + 223.6 (50.3, Z) x_2 - 329.1(39.0, A) x_3, 0 o, 0 e*
M6.03 *R = 0.9745, D = 94.96 %, s(e) = 0.61, y = 14.66 (1.07, Z) -1.35 (0.17, Z) x_1 + 0.055 (0.023, A) x_2 + 0.23 (0.032, Z) x_3, 0 o, 0 e*
M6.04 *R = 0.9498, D = 90.22 %, s(e) = 0.87, y = 119.4 (13.9, Z) -255.7 (29.4, Z) x_1 - 2.2 (0.8, Z) x_2 - 0.5 (0.7, A) x_3 + 0.5 (0.2, Z)x_4, 0 o*
M6.05 *R = 0.6070, D = 36.85 %, s(e) = 0.11, y = 9.69 (9.61, A) + 0.017 (0.060, A) x_1 - 0.64 (0.77, A) x_2 - 0.10 (0.05, A) x_3, 1 o, 1 e*
M6.06 *R = 0.9486, D = 90.00 %, s(e) = 0.76, y = -105.8 (15.9, Z) + 0.5 (0.3, A) x_1 + 15.7 (2.8, Z) x_2 - 0.03 (0.15, A) x_3 + 27.9 (14.0, A) x_4, 0 o, 0 e*
M6.07 *R = 0.9094, D = 82.71 %, s(e) = 35.0, y = 669.2 (103.6, Z) + 102.7 (78.6, A) x_1 - 0.1 (0.1, A) x_2, 1 o, 1 e*
M6.08 *R = 0.4215, D = 17.77 %, s(e) = 10.0, y = 48.9 (7.3, Z) - 0.03 (0.03, A) x_1 + 0.02 (0.03, A) x_2 + 8.2 (3.1, Z) x_3, 1 o, 1 e*

M6.09 *R = 0.8400, D = 70.56 %, s(e) = 44.1, y = 78.0 (52.4, A) + 0.42 (0.73, A) x_1 + 5.2 (0.8, Z) x_2, 1 o, 1 e*

M6.10 *R = 0.9955, D = 99.10 %, s(e) = 1.3, y = 2.0 (0.2, Z) + 1.7 (0.02, Z) x_1 - 0.003 (0.004, A) x_2 + 0.006 (0.0004, Z) x_3*

M6.11 *R = 0.9633, D = 92.80 %, s(e) = 2.0, y = 2.8 (0.9, Z) + 3.8 (0.3, Z) x_1 - 28.3 (5.1, Z) x_2 + 7.7 (1.4, Z) x_3, 2 o, 2 e*

M6.12 *R = 0.9931, D = 98.62 %, s(e) = 1.8, y = -117.0 (47.2, Z) + 1.6 (0.4, Z) x_1 + 1.8 (1.3, A) x_2 - 1.6 (0.5, Z) x_3, 1 o, 0 e*

M6.13 *R = 79.10, D = 62.57 %, s(e) = 39.9, y = 672.9 (277.5, Z) + 0.12 (0.02, Z) x_1 - 1.6 (1.0, A) x_2 - 77.8 (177.7, A) x_3 - 1.9 (1.1, A)x_4 - 3.7 (0.00, A) x_5 + 0.0014 (0.0054, A)x_6,, 4 o, 2 e*

M6.14 *R = 0.8540, D = 72.93 %, s(e) = 1.3, y = -35.6 (14.4, Z) + 1.1 (0.2, Z) x_1 + 2.4 (2.6, A) x_2 + 3.0 (1.3, Z) x_3, 1 o, 2 e*

M6.15 *R = 0.9587, D = 91.91 %, s(e) = 0.3, y = 0.28 (0.08, Z) + 2.2 (0.6, Z) x_1 - 1.7 (0.8, Z) x_2 + 7.2 (0.4, Z) x_3, 6 o, 5 e*

M6.16 *R = 0.9741, D = 94.89 %, s(e) = 0.4, y = 25.6 (1.5, Z) - 0.1 (0.01, Z) x_1 - 2.4E-05 (3.5E-06, Z) x_2 - 0.24 (0.04, Z) x_3 - 5.4E-03 (8.9E-03, A) x_4, 0 o, 0 e*

M6.17 *R = 0.9851, D = 97.04 %, s(e) = 0.01, y = 2.1 (0.7, Z) - 4.2E-03 (8.1E-04, Z) x_1 + 0.2 (0.5, A) x_2 - 1.3E-03 (3.8E-04, Z) x_3 + 0.4 (0.2, Z) x_4, 0 o, 0 e*

M6.18 *R = 1.0000, D = 99.99 %, s(e) = 0.06, y = 93.3 (0.1, Z) - 12.6 (0.01, Z) x_1 + 0.4 (0.00, Z) x_2 - 1.0E-03 (5.9E-04, A) x_3, 1 o, 0 e*

M6.19 *R = 0.9986, D = 99.72 %, s(e) = 0.3, y = -0.07 (0.1, A) - 0.7 (0.2, Z) x_1 + 0.9 (0.2, Z) x_2 + 0.8 (0.1, Z) x_3, 2 o, 2 e*

M6.20 *R = 0.9983, D = 99.65 %, s(e) = 18.1, y = -28.8 (10.6, Z) - 0.04 (0.3, A) x_1 + 0.81 (0.18, Z) x_2 + 0.28 (0.32, A) x_3, 2 o, 0 e*

M6.21 *R = 0.7653, D = 58.56 %, s(e) = 150.4, y = 1555 (566, Z) + 20.5 (60.1, A) x_1 - 34.3 (14.9, Z) x_2 - 170.9 (45.0, Z) x_3 - 9.5 (9.5, A)x_4 + 4.8 (1.3, Z)x_4,, 2 o, 1 e*

M6.22 *R = 0.9263, D = 85.81 %, s(e) = 0.08, y = 2.6 (1.7, A) + 0.01 (0.01, A) x_1 + 4.5E-04 (3.2E-04, A) x_2 + 3.8E-03 (2.8E-03, A) x_3 - 0.08 (0.03, Z)x_4, 2 o, 0 e*

M6.23 *R = 0.9734, D = 94.75 %, y = -55.8 (4.6, Z) + 0.5 (0.01, Z) x_1 - 0.6 (0.5, A) x_2 - 0.7 (0.01, Z) x_3, 16 o, 45 e*

M6.24 *R = 0.9820, D = 96.43 %, s(e) = 136.6, y = 2919 (7123, A) + 0.0424 (0.0030, Z) x_1 - 121.9 (71.8, A) x_2 - 3848 (3203, A) x_3 + 110.1 (69.19, A) x_4 - 0.012 (0.01, A) x_5 + 0.009 (0.029, A) x_6, 3 o, 1 e*

M6.25 *R = 0.2912, D = 8.48 %, s(e) = 238.6, y = 111.1 (191.0, A) - 0.4 (18.1, A) x_1 + 15.6 (17.0, A) x_2 - 10.4 (10.7, A) x_3 - 42.8 (25.3, A) x_4, 1 o, 2 e*

M6.26 *R = 0.9995, D = 99.92 %, s(e) = 0.02, y = -0.6 (0.4, A) + 0.03 (0.03, A) x_1 + 0.13 (0.08, A) x_2 + 0.19 (0.00, Z) x_3 + 0.0057 (0.0057, A) x_4, 0 o, 0 e*

M6.27 *R = 0.9948, D = 98.95 %, s(e) = 0.44, y = 75.9 (1.2, Z) + 0.06 (5.5E-03, Z) x_1 + 0.17 (7.0E-03, Z) x_2 + 9.8 (0.6, Z) x_3 - 0.12 (0.00, Z) x_4 + 0.02 (0.00, Z) x_5 + 0.28 (0.02, Z) x_6 , 4 o, 3 e*

M6.28 *R = 0.9431, D = 88.95 %, s(e) = 2.9, y = -62.6 (16.2, Z) + 14.0 (2.5, Z) x_1 - 0.05 (0.04, A) x_2 + 0.03 (0.01, Z) x_3 + 4.7E-03 (0.013, A) x_4, 2 o, 0 e*

M6.29 *R = 0.4540, D = 20.61 %, s(e) = 7.4, y = 0.6 (71.1, A) + 31.7 (53.2, A) x_1 + 29.9 (33.9, A) x_2 + 0.012 (0.032, A) x_3 - 0.04 (0.24, A) x_4 + 0.19(0.18, A)x_5, 0 o, 1 e*

M6.30 *R = 0.9534, D = 90.90 %, s(e) = 35.9, y = -1902 (717, Z) + 77.2 (13.1, Z) x_1 - 19.7 (4.1, Z) x_2 + 29.8 (8.1, Z) x_3 + 0.4 (0.4, A) x_4, 1 o, 1 e*

M6.31 *R = 0.8411, D = 70.74 %, s(e) = 2448, y = -4027 (2981, A) + 0.04 (0.10, A) x_1 - 0.0014 (0.002, A) x_2 + 84.2 (52.1, A) x_3 + 3353 (2068, A) x_4, 2 o, 0 e*

M6.32 *R = 0.7088, D = 50.24 %, s(e) = 0.22, y = 98.5 (0.6, Z) - 8.7 (5.0, A) x_1 + 0.9 (0.5, A) x_2 + 817.3 (253.6, Z) x_3, 2 o, 2 e*

M6.33 *R = 0.9641, D = 92.95 %, s(e) = 0.2, y = -0.45 (0.34, A) + 0.02 (0.03, A) x_1 - 0.002 (0.004, A) x_2 + 0.026 (0.002, Z) x_3 + 0.018 (0.050, A) x_4, 1 o, 0 e*

M6.34 *R = 0.8741, D = 76.41 %, s(e) = 0.25, y = 2.1 (0.2, Z) + 0.4 (0.1, Z) x_1 - 0.6 (0.1, Z) x_2 + 0.3 (0.1, Z) x_3 - 1.1E-03 (2.6E-04, Z) x_4, 1 o, 0 e*

M6.35 *R = 0.9235, D = 85.28%, y = 0.162 (0.438, A) + 0.00201 (0.00058, Z)× x_1 + 0.00125 (0.00055, Z) x_2 + 0.189 (0.092, A) x_3 + 0.0876 (0.1765, A) x_4, 1o, 1e*

M6.36 *R = 0.9569, D = 91.56%, y = -193.6 (68.7, Z) + 1.402 (0.530, Z)× x_1 + 0.772 (0.203, Z) x_2 + 1.048 (0.136, Z) x_3 - 0.124 (0.173, A) x_4 + 0.0719 (0.131, A)× x_5 + 0.0918 (0.163, A) x_6 - 0.1067 (0.156, A) x_7, 2o, 0e*

M6.37 *R = 0.7298, D = 53.26%, y = 3.718 (34.08, A) + 0.1402 (0.0615, Z)× x_1 - 2.712 (2.348, A) x_2 + 1.263 (0.590, A) x_3 + 0.591 (0.328, A) x_4, 1o, 0e*

M6.38 *nejde*

M6.39 *R = 0.7749, D = 60.05%, y = 1105.4 (715.3, A) + 4.006 (1.079, Z)× x_1 - 516.6 (242.5, Z) x_2 + 0.510 (0.235, Z) x_3 + 185.75 (168.10, A) x_4 + 10.526 (8.825, A)× x_5 - 9.20 (3.82, Z) x_6 , 1o, 1e*

M6.40 *R = 0.9777, D = 95.60%, y = 37.32 (20.58, A) + 0.960 (0.050, Z)× x_1 + 0.00183 (0.09054, A) x_2 + 0.0285 (0.124, A) x_3, 1o, 1e*

M6.41 *R = 0.9956, D = 99.12%, y = 0.983 (0.178, Z) + 0.722 (0.149, Z)× x_1 + 1.227 (0.244, Z) x_2, 0o, 0e,*

M6.42 *R = 0.8839, D = 78.13%, y = -83.77 (42.64, A) + 0.392 (0.092, Z)× x_1 - 0.138 (0.049, Z) x_2 + 2.768 (1.022, Z) x_3, 1o, 0e*

M6.43 *R = 0.9422, D = 88.77%, y = -983.7 (1228.9, A) + 8.683 (6.564, A)× x_1 + 277.23 (67.62, Z) x_2 + 13.83 (7.50, A) x_3, 3o, 2e,*

M6.44 *R = 0.9181, D = 84.30%, y = -975.34 (163.40, Z) + 11.65 (1.62, Z)× x_1 - 0.719 (0.629, A) x_2 - 45.35 (40.25, A) x_3 + 9.99 (1.53, Z) x_4, 1o, 2e*

6.7 Doporučená literatura

[1] Draper N. R. a Smith H.: *Applied Regression Analysis*. 2nd Ed., Wiley, New York 1981.

[2] Seber G. A. F.: *Linear Regression Analysis*. Wiley, New York 1977.

[3] Guttman I.: *Linear Models - An Introduction*. Wiley, New York 1982.

[4] Searle S. R.: *Linear Models*. Wiley, New York 1971.

[5] Anscombe F. J.: *Amer. Statist.* **27**, 17 (1973).

[6] Utts J.: *Commun. Statist.* **11**, 2801 (1982).

[7] Krämer W. a Sonnberger H.: *The Linear Regression Model under Tests.* Physica Verlag Heidelberg, 1986.

[8] Scott J. R.: *Appl. Statist.* **24**, 42 (1975).

[9] Cassela J.: *Amer. Statist.* **37**, 147 (1983).

[10] Suich R., Derringer G. C.: *Technometrics* **19**, 213 (1977).

[11] Kornilov A. N. a Smenina L. B.: *Žurnal fyzičeskoj chimiji* **44**, 1932 (1970).

[12] Militký J.: *Proc. Conf. ESC 87*, Praha září 1987.

[13] Phillip G. R., Harris J. M. a Eyring E. M.: *Anal. Chem.* **54**, 2053 (1982).

[14] Neil J. W. a Johnson D. E.: *Commun. Stat.* **13**, 485 (1984).

[15] Green J. R. a Margerison D.: *Statistical Treatment of Experimental Data*. Elsevier, Amsterdam 1978.

[16] Nash J. C.: *Compact Numerical Algorithms for Computer*. A. Hilger, Bristol, 1979.

[17] Antila A.-M. a Sikvonen M.-L.: *Fresenius Z. Anal. Chem.*, **327**, 799 (1987).

[18] Lawson Ch. a Hanson R.: *Solving Least-Squares Problems*. Englewood Cliffs, New Jersey, 1974.

[19] Dahlquist A. G. a Björck A.: *Numerical Methods*. Emglewood Cliffs, 1974.

[20] Marquardt D. M.: *Technometrics* **12**, 591 (1970).

[21] Belsey D. A., Kuh E. a Welsch R. E.: *Regression Diagnostics*. Wiley New York 1980.

[22] Atkinson A. C.: *Plot, Transformation, Regression*. Claredon Press, Oxford 1986.

[23] Weisberg S.: *Technometrics* **25**, 219 (1983).

[24] Cook R. D. a Weisberg S.: *Reziduals and Influence in Regression*. Chapman and Hall, New York 1982.

[25] Joiner B.: *Amer. Statist.* **35**, 227 (1981).

[26] Chattarjee S. a Hadi A. S.: *Statist. Sci.* **1**, 379 (1986).

[27] Gorman M. A. a Myers R. M.: *Commun. Statist.* **16**, 770 (1987).

[28] Gray J. B.: *Proc. Stat. Comput. Sect.*, p. 159, ASA Washington 1983.

[29] Mallows C. L.: *Technometrics* **28**, 313 (1986).
[30] Cook R. D. a Weisberg S.: *Biometrika* **70**, 1 (1983).
[31] Querry N.: *Technometrics* **6**, 225 (1964).
[32] Jarque C. M. a Bera A. K.: *Int. Stat. Rev.* **55**, 163 (1987).
[33] Judge G. G. a Bock M. E.: *Statistical Implications of Pre-test and Stein Rule Estimators in Econometrics*. North Holland, Amsterdam 1978.
[34] Leary J. J. a Messick E. B.: *Anal. Chem.* **57**, 956 (1985).
[35] Horn S. D., Horn R. A. a Duncan D. B.: *J. Amer. Statist. Assoc.* **70**, 380 (1975).
[36] Fuller W. A.: *Measurement Error Models*. Wiley, New York 1987.
[37] Hill R. W. a Holland D. W.: *J. Amer. Statist. Assoc.* **72**, 828 (1977).
[38] Huber D. J.: *Robust Statistics*. Wiley, New York 1981.
[39] Li G. in Hoaglin D. C. a kol. (eds), *Exploring Data Tables,Trends and Shapes*. Wiley, New York, 1985 (kap. 8).
[40] Phillip G. R. a Eyring E. M.: *Anal. Chem.* **55**, 1134 (1983).
[41] Nikuličev J. G., Kančenko G. a kol.: *Kolloidnyj Žurnal* **50**, 473 (1988).
[42] Malá J., Sláma I.: *Chem. Papers* **42**, 319 (1988).
[43] Krasker W. S. a Welsch R. E.: *J. Amer. Statist. Assoc.* **77**, 595 (1982).
[44] Hettmansperger T. P.: Austral. *J. Statist.* **29**, 1 (1987).
[45] Rousseuw P. J. a Leroy A. M.: *Robust Regression and Outliers Detection*. Wiley, New York 1987.
[46] Rosenblatt J. R. a Spiegelman C. H.: *Technometrics* **23**, 329 (1981).
[47] Ebel S. a Becht U.: *Fresenius Z. Anal. Chem.*, **158** (1987).
[48] Schwartz L. M.: *Anal. Chem.* **48**, 2287 (1976).
[49] Naszodi L. J.: *Technometrics* **20**, 201 (1978).
[50] Krutchkoff R. G.: *Technometrics* **9**, 425 (1967).
[51] Schwartz L. M.: *Anal. Chem.* **49**, 2062 (1977).
[52] Oppenhelmer L. a kol.: *Anal. Chem.*, **55**, 638 (1983).
[53] Schwartz L. M.: *Anal. Chem.* **55**, 1424 (1983).
[54] Ebel S., Kamm U.: *Fresenius Z. Anal. Chem.* **318**, 293 (1984).
[55] Ebel S.a Brockmeyer R.: *Fresenius Z. Anal. Chem.* **326**, 770 (1970).
[56] Himmelblau D.: *Process Analysis by Statistical Methods*. Wiley, New York 1969.
[57] Swed F., Eisenhart C.: *Annal of Math. Statist.* **14**, 66 (1943)
[58] Liteanu C., Rica I.: *Statistical Theory and Methodology of Trace Analysis.* Ellis Horwood, Chichester 1980.
[59] Davídek J. a kol.: *Laboratorní příručka analýzy potravin.* SNTL, Praha 1981.
[60] Kraft G., Dosch H.: Z. Anal. Chem. 271, 264 (1974).
[61] Truxová I.: *Diplomová práce,* VŠCHT Pardubice 1991.
[62] Rice J. A.: *Mathematical Statistics and Data Analysis.* Wadsworth & Brooks, California 1988.
[63] Cyhelský L. a kol.: *Úlohy k základům statistiky.* SNTL, Praha 1988.
[64] Potocký R. a kol.: *Zbierka úloh z pravdepodobnosti a matematickej štatistiky.* ALFA, Bratislava 1986.
[65] Kleinbaum D. G. a kol.: *Applied Regression Analysis and Other Multivariate Methods.* PWS-KENT Publishing Comp., Boston, 1988.
[66] Ebel S., G. Herold: *Z. Anal. Chem.* **270**, 20 (1974).

[67] Anderson R. L.: *Practical Statistics for Analytical Chemists.* van Nostrand Reinhold Company, New York 1987.
[68] Gottschalk G.: *Z. Anal. Chem.* **282,** 1 (1976).
[69] Miller J. C., Miller J. N.: *Statistics for Analytical Chemistry.* Ellis Horwood, Chichester, 1984.
[70] Graybill F. A., Iyer H. K.: *Regression Analysis: Concepts and Applications*. Duxbury Press, International Thomson Publishing 1994.
[71] Neter J., Kutner M. H., Nachtsheim CH.J., Wasserman W.: *Applied Linear Statistical Models*. RICHARD D. IRWIN, Chicago 1990.
[72] Meloun M., Militký J.: *Statistické zpracování experimentálních dat.* East Publishing, Praha 1996.

7

KORELACE

Pro vyjádření intenzity vztahů mezi složkami $\xi_1, ..., \xi_m$ m-rozměrného náhodného vektoru $\boldsymbol{\xi}$ se používá *korelačních koeficientů*. Data tvoří *náhodný výběr* z m-rozměrného rozdělení náhodného vektoru $\boldsymbol{\xi}$. Neuvažuje se obyčejně a priori, která složka ξ_j náhodného vektoru $\boldsymbol{\xi}$ je *vysvětlovaná* (u lineárního regresního modelu označovaná jako výstupní závisle proměnná) a které složky vektoru ξ jsou *vysvětlující* (u lineárního regresního modelu označované jako vstupní nezávisle proměnné). Náhodný výběr $\{x_{ij}\}$, $i = 1, ..., n$, $j = 1, ..., m$, velikosti n je tvořen ($n \times m$) rozměrným polem dat

$$\begin{bmatrix} x_{11} & . & . & . & x_{12} & . & . & . & . & x_{1m} \\ x_{21} & . & . & . & x_{22} & . & . & . & . & x_{2m} \\ . & & & & & & & & & \\ . & & & & & & & & & \\ . & & & & & & & & & \\ x_{n1} & . & . & . & x_{n2} & . & . & . & . & x_{nm} \end{bmatrix} .$$

Platí, že

a) počet řádků n (tj. počet m-rozměrných "bodů" $\boldsymbol{x}_i$) je výrazně větší, než počet sloupců m (tj. počet "proměnných" čili složek vektoru $\boldsymbol{x}$).

b) Všechny složky vektoru $\boldsymbol{x}_i$ jsou *náhodné* a předem *neovlivnitelné* experimentátorem.

c) Mezi složkami jsou pouze lineární vazby.

7.1 Druhy korelačních koeficientů

7.1.1 Párový korelační koeficient

Korelační koeficienty slouží jako míry pro vyjádření "těsnosti lineární stochastické vazby" mezi složkami náhodného vektoru $\boldsymbol{\xi}$. *Pearsonův párový korelační koeficient* $\rho(\xi_i, \xi_j) = r_{ij}$ vyjadřuje míru lineární stochastické vazby mezi náhodnou veličinou ξ_i a ξ_j. Označme *populační párový korelační koeficient* ρ a *výběrový párový korelační koeficient* r.

Nahradíme střední hodnoty μ_1 a μ_2 aritmetickými průměry $\bar{x}_1$ a $\bar{x}_2$, dále rozptyly σ_1^2 a σ_2^2 výběrovými rozptyly s_1^2 a s_2^2. Pro *výběrový korelační koeficient* platí výraz

$$r = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2}} .$$

K interpretaci korelačních koeficientů je třeba přistupovat velmi obezřetně. Platí pravidlo, že *významná párová korelace není důkazem příčinné souvislosti*. Někdy vznikají falešné korelace, kdy jak ξ_1, tak i ξ_2 silně korelují s neuvažovanou náhodnou veličinou ξ_3 a vysoká hodnota $\rho(\xi_1, \xi_2)$ je důsledek vysokých hodnot $\rho(\xi_1, \xi_3)$ a $\rho(\xi_2, \xi_3)$. Při interpretaci korelačních koeficientů je pak vhodné užít i parciální korelační koeficienty.

Při konstrukci testů významnosti se využívá testační statistiky

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} ,$$

která má pro případ $\rho = 0$ Studentovo rozdělení s $(n - 2)$ stupni volnosti. Toho lze využít k testování nekorelovanosti, resp. lineární nezávislosti dvojice náhodných veličin. Je-li jejich rozdělení dvourozměrné normální, je nekorelovanost totožná s nezávislostí. Testuje se hypotéza H_0: $\rho = 0$ proti různým alternativám H_A. Vyjde-li $*t*$ větší než odpovídající kvantil Studentova rozdělení, zamítá se H_0 a náhodné veličiny nejsou nekorelované. Uvedený test je silně nerobustní a platí pouze v případě dvourozměrné normality ξ_1, ξ_2. Pro urychlení konvergence $f(r)$ k normálnímu rozdělení se používá různých transformací. Jednoduchá *Rubenova transformace* má tvar

$$R(r) = \frac{\sqrt{n - 2.5}\, r}{\sqrt{1 - 0.5\, r^2}} .$$

Náhodná veličina $R(r)$ již má i pro menší výběry normované normální rozdělení N(0, 1).

7.1.2 Parciální korelační koeficient

V řadě případů je účelné sledovat vztah mezi dvěma složkami ξ_1 a ξ_j náhodného vektoru při zkonstantnění dalších složek vektoru $\boldsymbol{\xi}$. Pro vyjádření intenzity tohoto vztahu se používají parciální korelační koeficienty různých řádů. Nejjednodušší jsou parciální korelační koeficienty nultého řádu, které odpovídají párovým korelačním koeficientům.

Parciální korelační koeficienty prvního řádu $r_{1,3(2)}$ odpovídají párovému korelačnímu koeficientu mezi rezidui $\gamma_2 = \xi_1 - E(\xi_1/x_2)$

a rezidui $\iota_2 = \xi_3 - E(\xi_3/x_2)$

a mají tvar $r_{1,3(2)} = \dfrac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$.

Analogicky lze definovat i další parciální korelační koeficienty $r_{1i(j)}$ prvního řádu jako párové korelační koeficienty mezi rezidui $\gamma_j = \xi_1 - E(\xi_1/x_j)$

a rezidui $\iota_j = \xi_i - E(\xi_i/x_j)$,

pro které platí tvar $r_{1,i(j)} = \dfrac{r_{1i} - r_{1j}\, r_{ij}}{\sqrt{(1 - r_{1i}^2)(1 - r_{ij}^2)}}$.

Parciální korelační koeficienty druhého řádu $r_{1i(j,k)}$ jsou vlastně párové korelační koeficienty reziduí $\gamma_{j,k} = \xi_1 - E(\xi_1/(x_j, x_k))$

a reziduí $\iota_{j,k} = \xi_i - E(\xi_i/(x_j, x_k))$

a mají tvar $r_{1i(j,k)} = \dfrac{r_{1i(j)} - r_{1j(k)}\, r_{ij(k)}}{\sqrt{(1 - r_{1j(k)}^2)(1 - r_{ij(k)}^2)}}$.

Parciální korelační koeficient (m - 1). řádu $r_{1i(2,\,3,\ldots,m)}$ odpovídá jednoduchému korelačnímu koeficientu mezi rezidui $\gamma_{2,\ldots,m} = \xi_1 - E(\xi_1/\boldsymbol{x}^*)$

a rezidui $\iota_{2,\ldots,m} = \xi_i - E(\xi_i/\boldsymbol{x}^*)$,

kde vektor $\boldsymbol{x}^*$ obsahuje složky $x_2, x_3, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m$.

Obecně se počítají *parciální korelační koeficienty vyšších řádů* podle rekurentní formule $r_{1,j(2,3,\ldots,j-1)} = \dfrac{A - B\,C}{\sqrt{(1 - B^2)(1 - C^2)}}$,

kde $A = r_{1,j(2,3,\ldots,j-2)}$ $B = r_{1,j-1(2,3,\ldots,j-2)}$ $C = r_{j,j-1(2,3,\ldots,j-2)}$.

Pro statistické testování a konstrukci intervalů spolehlivosti se využívá pravidlo, že rozdělení parciálního korelačního koeficientu řádu (m - 1) je stejné jako rozdělení párového korelačního koeficientu pro rozsah výběru (n - m + 1).

7.1.3 Vícenásobný korelační koeficient

Vícenásobný korelační koeficient $R_{1(2,\dots,m)}$ definuje *míru lineární stochastické závislosti* mezi náhodnou veličinou ξ_1 a nejlepší lineární kombinací složek $\xi_2, \dots, \xi_m$ náhodného vektoru. Pro tento korelační koeficient platí, že

$$R_{1(2,\dots,m)} = \sqrt{1 - \frac{\det(\boldsymbol{R})}{\det(\boldsymbol{R}_{11})}},$$

kde det(.) označuje determinant a $\boldsymbol{R}_{ij}$ je matice vzniklá vypuštěním i-tého řádku a j-tého sloupce korelační matice $\boldsymbol{R}$.

Mezi základní vlastnosti vícenásobného korelačního koeficientu patří:

1. Platí nerovnost $0 \leq R_{1(2,\dots,m)} \leq 1$.

2. Pokud je $R_{1(2,\dots,m)} = 1$, znamená to, že náhodná veličina ξ_1 je přesně lineární kombinací veličin $\xi_2, \dots, \xi_m$.

3. Pokud je $R_{1(2,\dots,m)} = 0$, jsou také všechny odpovídající párové korelační koeficienty rovny nule $\rho(\xi_1, \xi_j) = 0, j = 2, \dots, m$.

4. Pro případ jedné vysvětlující proměnné je $R_{1(2)} = |\rho(\xi_1, \xi_2)|$, tj. vícenásobný korelační koeficient je totožný s absolutní hodnotou párového korelačního koeficientu.

5. Platí, že s růstem počtu vysvětlujících proměnných vícenásobný korelační koeficient nikdy neklesá

$$R^2_{1(2)} \leq R^2_{1(2,3)} \leq R^2_{1(2,3,4)} \leq \dots \leq R^2_{1(2,\dots,m)} .$$

Při znalosti jednotlivých parciálních korelačních koeficientů všech řádů je možné vyčíslit také vícenásobný korelační koeficient ze vztahu

$$R^2_{1(2,\dots,m)} = 1 - (1 - R^2_{1,2})(1 - R^2_{1,3(2)})(1 - R^2_{1,4(2,3)}) \dots$$

$$\dots (1 - R^2_{1,m(2,3,\dots,m-1)}).$$

Pro výpočet parciálních korelačních koeficientů je výhodné využít vztah

$$R_{1i(2,3,\dots,m)} = \frac{(-1)^i \det(\boldsymbol{R}_{1,i})}{\sqrt{\det(\boldsymbol{R}_{11})\ \det(\boldsymbol{R}_{i,i})}} ,$$

kde $\boldsymbol{R}$ je korelační matice odpovídající vektoru $\boldsymbol{\xi}$ a $\boldsymbol{R}_{i,j}$ je matice vzniklá vynecháním i-tého řádku a j-tého sloupce matice $\boldsymbol{R}$.

7.2 Pořadový korelační koeficient

V některých případech je výhodné nahradit klasický párový korelační koeficient pořadovým (neparametrickým) korelačním koeficientem podle Spearmana, který je málo citlivý na přítomnost vybočujících hodnot. Pořadí i-tého prvku výběru je rovno indexu odpovídající pořádkové statistiky. Označme pořadí prvků výběru vzhledem k proměnné ξ_1 jako x_{1si} a pořadí prvků výběru vzhledem k proměnné ξ_2 jako x_{2si}.

Pro *Spearmanův pořadový korelační koeficient* pak platí

$$\bar{\rho}_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (x_{1si} - x_{2si})^2.$$

Rozdělení veličiny $\hat{\rho}_s$ je symetrické se střední hodnotou $E(\hat{\rho}_s) = 0$ a rozptylem $D(\hat{\rho}_s) = 1/(n - 1)$. Pro $n > 10$ se často využívá toho, že veličina

$$t_s = \frac{|\hat{\rho}_s| \sqrt{n - 2}}{\sqrt{1 - \hat{\rho}_s^2}}$$

má asymptoticky Studentovo rozdělení s $(n - 2)$ stupni volnosti, pokud teoretický koeficient $\rho_s = 0$.

V praxi se stává, že pro několik prvků výběru vychází stejné pořadí. Pak se všem přiřadí průměr z pořadí, které by měly, pokud by nabývaly různých hodnot, a *Spearmanův korelační koeficient* se počítá dle upravené formule

$$\rho_s = \frac{\frac{n(n^2 - 1)}{6} - \sum_{i=1}^{n} (x_{1si} - x_{2si})^2 - a - b}{\sqrt{\left(\frac{n(n^2 - 1)}{6} - 2a\right)\left(\frac{n(n^2 - 1)}{6} - 2b\right)}},$$

kde a, b jsou opravné koeficienty na pořadí

$$a = \frac{1}{12} \sum_{(j)} (a_j^3 - a_j),$$

$$b = \sum_{(k)} (b_k^3 - b_k),$$

kde j označují čísla shluků stejných pořadí pro x_1 a a_j je počet hodnot se stejným pořadím v j-tém shluku. Analogicky je definováno také k a b_k.

Spearmanův pořadový korelační koeficient ρ_s leží v intervalu $-1 \leq \rho_s \leq 1$. Pokud výběr pochází z dvourozměrného normálního rozdělení a $n \geq 30$, platí vztah, že

$$\rho(\xi_1, \xi_2) = 2 \sin\left(\frac{\pi}{6} \rho_s\right).$$

Při použití pořadových korelačních koeficientů je třeba mít stále na paměti, že při přechodu z dat x_{1i}, x_{2i} na pořadí x_{1si}, x_{2si} dochází vždy ke ztrátě informace. Na druhé straně je však docíleno zrobustnění a snížení citlivosti na odchylky od normality.

7.3 Cronbachův korelační koeficient γ spolehlivosti výsledku

Spolehlivost výsledku, měření může být rozdělena na dvě kategorie: správnost a přesnost (viz 1. kapitola). *Správnost* se týká důkazu, zda naměřená hodnota je správná. *Přesnost* se týká důkazu, zda naměřené hodnoty jsou stejné při svém opakovaní. Přístroj může být správný při měření jedné veličiny, ale nemusí být správný při měření jiné. Bylo

navrženo několik metod na prokázání spolehlivosti přístroje. Zaměříme se nyní na ověření *vnitřní jednotnosti výsledku (konzistentnosti).*

Cronbachův korelační koeficient γ: představuje nejrozšířenější kritérium posouzení vnitřní jednotnosti výsledku a vypočte se dle vzorce

$$\gamma = \frac{m}{m-1}\left[1 - \frac{\sum_{i=1}^{m}\sigma_{ii}}{\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{ij}}\right],$$

kde m je počet proměnných a σ_{ij} je vypočtená kovariance mezi proměnnou i a j, σ_{ii} je rozptyl proměnné i. Jsou-li data předem standardizována (odečtením průměru a podělením směrodatnou odchylkou položky), dostaneme standardizovanou verzi Cronbachova koeficientu

$$\gamma = \frac{m\,\bar{\rho}}{1 + \bar{\rho}\,(m-1)},$$

kde $\bar{\rho}$ je průměr všech korelačních koeficientů mezi všemi m proměnnými.

Cronbachův koeficient γ má několik interpretací: rovná se průměru všech Cronbachových koeficientů, získaných pro všechny možné kombinace rozdělením $2m$ proměnných do dvou skupin, každé o m proměnných, a vypočtením dvou polovičních testů. Dále odhaduje očekávanou korelaci jednoho přístroje s alternativní formou jiného, obsahujícího stejný počet měřených proměnných. Může odhadovat také očekávanou korelaci mezi aktuálním testem a hypotetickým testem, který nikdy nebyl popsán. Protože jde o korelační koeficient, je Cronbachův koeficient γ definován v intervalu -1 až +1. Ve většině případů jde o kladné číslo. Existuje pravidlo, že γ by mělo pro většinu přístrojů dosáhnout hodnoty alespoň 0.8. Koeficient γ lze zlepšit či zvýšit zvětšením počtu měření nebo zvýšením průměrné korelace mezi proměnnými.

Postup analýzy korelace

1. Návrh modelu: zařadíme obvykle i absolutní člen β_0 a nejprve budeme uvažovat lineární regresní model ve tvaru $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$. Polohu a proměnlivost proměnných y, x_1, x_2, x_3 přináší *průměr* a *směrodatná odchylka* hodnot každé proměnné. Zatímco *Pearsonův vícenásobný korelační koeficient r* ukazuje, do jaké míry je navržený lineární regresní model statisticky významný, hodnota *koeficientu determinace* $D = r^2$ vyjadřuje kolik procent bodů dobře koresponduje s modelem. *Predikovaný koeficient determinace* D_p má podobný význam jako *koeficient determinace D*, je však vyčíslen jinak, místo sumy čtverců odchylek *RSC* se ve vztahu užije střední kvadratická chyba predikce *MEP*.

2. Korelační matice Pearsonovy a Spearmanovy pořadové: výpočet umožňuje likvidaci děravých cel párovým nebo řádkovým způsobem. Korelace jsou však silně ovlivněny odlehlými hodnotami, heteroskedasticitou, nenormalitou rozdělení a nelinearitami. Vhodným doplňkem Pearsonova korelačního koeficientu je Spearmanův pořadový korelační koeficient. Pořadová korelace se vyčíslí Pearsonovým korelačním vzorcem, aplikovaným na pořadové číslo dat ne na numerické hodnoty dat samotných.

V případě odlehlých hodnot se bude velice lišit parametrická a neparametrická míra korelace, tj. Pearsonův korelační koeficient a Spearmanův pořadový korelační koeficient. V případě kolinearity jsou vysoké hodnoty párových korelací první indikací kolinearity.

3. Matice rozdílů: Aby se umožnilo porovnat tyto dva typy korelačních matic, vypočte se také matice rozdílů. Tím se ukáže, která dvojice proměnných si žádá hlubšího vyšetření.

Vzorová úloha 7.1 *Postup vyšetření korelace*

Jako vzorovou použijeme **Úlohu B7.05** *Obsah dehtu, nikotinu a CO v cigaretách* se zadáním: Federální komise obchodu USA posuzuje domácí cigarety dle obsahu dehtu x_1 [mg], nikotinu x_2 [mg] a hmotnosti cigarety x_3 [g] a konečně i obsahu oxidu uhelnatého CO x_4 [mg] v uvolněném cigaretovém kouři. Hlavní hygienik USA totiž považuje faktory x_1, x_2 a x_4 za vysoce nebezpečné pro zdraví člověka. Poslední studie ukázaly, že zvyšující se obsah dehtu a nikotinu spolu nesou i zvýšení obsahu oxidu uhelnatého. Vyšetřete, zda existuje na hladině významnosti $\alpha = 0.05$ korelace mezi proměnnými (a) x_1 a x_4, dále (b) x_2 a x_4, a (c) x_3 a x_4.

Řešení:

1. Návrh modelu: zařadíme i absolutní člen β_0 a nejprve budeme uvažovat lineární regresní model ve tvaru $x_4 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

Proměnná	**Průměr**	**Směrodatná odchylka**	**Pearsonův pár. korel. koeficient**	**Spočtená hladina význ.**	**Spearmanův pár. korel. koeficient**	**Spočtená hladina význ.**
x_1	12.216	5.6658	0.9575	0.0000	0.9448	0.0000
x_2	0.8764	0.3541	0.9259	0.0000	0.8778	0.0000
x_3	0.9703	0.0877	0.4640	0.0195	0.2170	0.2975
x_4	12.528	4.7397	1.0000	-----	1.0000	-----
Vícenásobný korelační koeficient r				: 0.95843		
Koeficient determinace 100 % D				: 91.859		
Predikovaný koeficient determinace D_p				: 0.91326		

Polohu a proměnlivost proměnných x_4, x_1, x_2, x_3 přináší *průměr* a *směrodatná odchylka* hodnot každé proměnné. *Pearsonův párový korelační koeficient* x_4 vs. x_1, x_4 vs. x_2, x_4 vs. x_3 ukazuje na vysokou korelaci, dvě nezávislé proměnné x_1, x_2 jsou se závisle proměnnou x_4 spjaty silnou lineární závislostí, zatímco poslední nezávisle proměnná x_3 slabou. Analogické závěry vyplývají i ze Spearmanova pořadového korelačního koeficientu. *Pearsonův vícenásobný korelační koeficient r* ukazuje, že navržený lineární regresní model je statisticky významný a jeho druhá mocnina, nazvaná *koeficient determinace* $D = r^2$ ukazuje, že 91.86 % bodů dobře koresponduje s modelem. *Predikovaný koeficient determinace* D_p má podobný význam jako *koeficient determinace D*, je však vyčíslen jinak, místo *RSC* se ve vztahu užije *MEP*.

2. Pearsonova korelační matice: vyčíslí se Pearsonovy párové korelační koeficienty parametrického charakteru a Cronbachův korelační koeficient γ.

Pearsonovy párové korelační koeficienty mezi dvojicemi vysvětlujících proměnných		**Spočtená hladina významnosti**
x_1 versus x_2:	0.97661	0.0000
x_1 versus x_3:	0.49077	0.0127
x_1 versus x_4:	0.95749	0.0000
x_2 versus x_3:	0.50018	0.0109
x_2 versus x_4:	0.92595	0.0000
x_3 versus x_4:	0.46396	0.0195
Cronbachův korelační koeficient γ		: 0.6939
Standardizovaný Cronbachův korelační koeficient γ		: 0.9111

Pearsonův párový korelační koeficient x_4 vs. x_1, x_4 vs. x_2, x_4 vs. x_3 ukazuje na vysokou korelaci, dvě nezávislé proměnné x_1, x_2, jsou se závisle proměnnou x_4 spjaty silnou lineární závislostí, zatímco poslední nezávisle proměnná x_3 slabou korelací.

3. Spearmanova korelační matice: vyčíslí se Spearmanovy pořadové párové korelační koeficienty neparametrického charakteru.

Spearmanovy pořadové korelační koeficienty mezi dvojicemi vysvětlujících proměnných		**Spočtená hladina významnosti**
x_1 versus x_2:	0.92843	0.0000
x_1 versus x_3:	0.15539	0.4583
x_1 versus x_4:	0.94480	0.0000
x_2 versus x_3:	0.19623	0.34717
x_2 versus x_4:	0.87781	0.0000
x_3 versus x_4:	0.21697	0.29752

Analogicky jako u Pearsonova korelačního koeficientu i *Spearmanovy párové pořadové korelační koeficienty mezi dvojicemi nezávislých proměnných* ukazují na silnou korelaci mezi prvními dvěma nezávisle proměnnými x_1 vs. x_2. Daleko slabší lineární vztah existuje mezi x_1 vs. x_3 a x_2 vs. x_3.

4. **Matice rozdílů Pearsonových a Spearmanových korelačních koeficientů:** některý software nabízí porovnání těchto dvou typů korelačních matic tak, že se vypočte matice jejich rozdílů.

Matice rozdílů Pearsonových a Spearmanových korelačních koeficientů:					
x_1 versus x_2:	0.04817				
x_1 versus x_3:	0.33538	x_2 versus x_3:	0.30395		
x_1 versus x_4:	0.01269	x_2 versus x_4:	0.04813	x_3 versus x_4:	0.24699

Někteří autoři pak doporučují takto identifikovat, která dvojice proměnných si žádá hlubšího vyšetření. Nám se však ukázalo, že v takovém případě je daleko účinnějším pomocníkem v této knize hodně využívaná regresní diagnostika, která totiž v grafech vlivných bodů snadno a jednoznačně odhalí odlehlé hodnoty.

7.4 Úlohy na korelaci

Úlohy jsou rozděleny do pěti kapitol: B7 (farmakologická a biochemická data), C7 (chemická a fyzikální data), E7 (environmentální, potravinářská a zemědělská data), H7 (hutní a mineralogická data) a S7 (ekonomická a sociologická data). Vysvětlete jednotlivé regresní a korelační statistiky a učiňte své závěry o výběru dat.

7.4.1 Analýza farmakologických a biochemických dat

Úloha B7.01 *Vliv věku dítěte na frekvenci píku EEG*
Neurologové nalezli, že se frekvence píku elektroencefalogramu EEG u normálních dětí zvyšuje s věkem. Bylo vyšetřeno 287 dětí ve věku od 2 do 16 let, které měly držet na otevřené dlani předmět o hmotnosti 65 g po určitou nespecifikovanou dobu, načež byla pro každé dítě zaznamenána frekvence píku EEG v Hz. Data byla setříděna dle věku dítěte x_1, u každé třídy byla vypočtena střední hodnota frekvence píku EEG x_2 v Hz. Za předpokladu homoskedasticity (stejných rozptylů u všech věkových skupin) otestujte, zda lze přijmout předpoklad linearity mezi věkem a frekvenci EEG píku.

Data: Věk dítěte x_1 [roky], střední hodnota frekvence píku EEG x_2 [Hz].

2	5.33,	3	5.75,	4	5.80,	5	5.60,	6	6.00,	7	5.78,	8	5.90,		
9	6.23,	10	7.28,	11	7.06,	12	7.60,	13	7.45,	14	8.23,	15	8.50,		
16	9.38,														

Úloha B7.02 *Vliv úniku radioaktivního odpadu na růst úmrtnosti na rakovinu*
Při úniku radioaktivního odpadu ze skládky v Hanfordu do řeky Columbia bylo vystaveno radioaktivitě obyvatelstvo v 9 okresech. Byla sledována úmrtnost na rakovinu x_1 (úmrtí na 100000 lidí v letech 1959-64) v různých vzdálenostech od Hanfordu x_2, str. 395 v cit.[12]? Účelem je zjistit, zda existuje vztah mezi úmrtností a ozářením, vyjádřeným vzdáleností od skládky.

Data: Úmrtnost na rakovinu x_1 [počet], vzdálenost od radioaktivní skládky x_2 [km].

1.20	120,	2.50	150,	1.60	140,	8.30	210,	6.40	180,
3.40	130,	3.80	170,	2.60	130,	11.6	210,		

Úloha B7.03 *Spotřeba cigaret a úmrtí na rakovinu plic*
Z náhodného výběru v šesti státech USA byla zjištěna spotřeba cigaret na obyvatele x_1 a roční míra úmrtnosti na 100 000 lidí následkem rakoviny plic x_2, str. 520 v cit.[12]. Vyšetřete, zda existuje korelace mezi oběma proměnnými x_1 a x_2 na hladině významnosti $\alpha = 0.05$.

Data: Spotřeba cigaret x_1 [četnost], úmrtnost x_2 [četnost].

3400	24,	2600	20,	2200	17,	2400	19,	2900	26,	2100	20,

Úloha B7.04 *Závislost věku žen a koncentrace cholesterolu v krvi*
Z náhodného výběru 50 amerických žen byla zjištěna následující data o věku x_1 a koncentraci cholesterolu v krvi [g/l] x_2 u prvních pěti žen, str. 528 v cit.[12]. Vyšetřete míru korelace mezi oběma proměnnými x_1 a x_2.

Data: Věk žen x_1 [roky], koncentrace cholesterolu v krvi x_2 [g/l].

30	1.6,	60	2.5,	40	2.2,	20	1.4,	50	2.7,

Úloha B7.05 *Obsahu dehtu, nikotinu a CO v cigaretách*
Federální komise obchodu USA posuzuje domácí cigarety dle obsahu dehtu x_1 [mg], nikotinu x_2 [mg] a hmotnosti cigarety x_3[g] a konečně i obsahu oxidu uhelnatého CO x_4 [mg] v uvolněném cigaretovém kouři. Hlavní hygienik USA totiž považuje faktory x_1, x_2 a x_4 za vysoce nebezpečné pro zdraví člověka. Poslední studie ukázaly, že zvyšující se obsah dehtu a nikotinu spolu nesou i zvýšení obsahu oxidu uhelnatého. Vyšetřete, zda existuje na hladině výynamnosti $\alpha = 0.05$ korelace mezi proměnnými (a) x_1 a x_4, dále (b) x_2 a x_4, a (c) x_3 a x_4.

Data: Obsah dehtu x_1 [mg], obsah nikotinu x_2[mg], hmotnost cigarety x_3[g], obsah oxidu uhelnatého CO x_4 [mg].

Alpine	14.1	0.86	0.9853	13.6,
...	...	...	...	...
Winston L.	12.0	0.82	1.1184	14.9,

Úloha B7.06 *Vliv počtu vypitých skleniček rumu na obsahu alkoholu v krvi*
Vyšetřete, zda obsah alkoholu v krvi x_2 je přímo úměrný počtu vypitých skleniček rumu x_1.

Data: Počet vypitých skleniček rumu x_1, obsah alkoholu v krvi x_2.

2	0.05,	3	0.060,	4	0.11,	5	0.13,	8	0.22,

Úloha B7.07 *Vliv věku člověka na hladinu cholesterolu v krvi*
Vyšetřete, zda věk člověka x_1 ovlivňuje hladinu cholesterolu v krvi x_2 v mg/100 ml, jsou-li k dispozici následující údaje.

Data: Věk x_1 [roky], cholesterol v krvi x_2 [mg/100 ml].

19	217.0,	27	221.4,	21	191.3,	45	321.5,	46	196.0,	58	284.6,
37	286.8,	42	194.2,	30	247.7,						

Úloha B7.08 *Spotřeba cigaret v USA a procento pacientů psychiatrie*
Vyšetřete, zda existuje korelace spotřeby cigaret v USA x_1 v různých letech a počtu pacientů psychiatrických léčeben, vyjádřeným procentem z celkové populace x_2. Diskutujte příčiny případné korelace.

Data: Spotřeba cigaret x_1 [počet], procento populace na psychiatrii x_2 [%].

3522	0.20,	3597	0.22,	4171	0.23,	4258	0.29,
3993	0.31,	3971	0.33,	4042	0.33,	4053	0.32,

Úloha B7.09 *Koncentrace SO_2 a procento nemocných dětí*
Vyšetřete, zda koncentrace oxidu siřičitého SO_2 v ovzduší x_1, vyjádřená střední hodnotou za 2 týdny v jednotkách $\mu g/m^3$ vzduchu, významně ovlivňuje procento dětí, se symptomy nemocí horních cest dýchacích x_2.

Data: Koncentraci oxidu siřičitého v ovzduší x_1 [$\mu g/m^3$ vzduchu], procento nemocných x_2 [%].

69	17,	147	24,	59	22,	132	31,	33	14,
116	33,	67	14,	120	25,	58	19,	92	27,

Úloha B7.10 *IQ a hypnabilita člověka*
Vyšetřete, zda existuje pozitivní korelace mezi inteligenčním kvocientem IQ x_1

a hypnabilitou člověka x_2. Kontrolní hodnoty byly získány náhodným výběrem z velkého souborů respondentů.

Data: Inteligenční kvocient IQ x_1 [hodnota], hypnabilita x_2.

103	55,	113	55,	119	59,	107	64,	78	45,	153	72,	114	42,	101	63,
103	62,	111	46,	105	41,	82	49,	110	57,	90	52,	92	41,		

Úloha B7.11 *Změna krevního tlaku před požitím pilulky léku a po něm*

Vyšetřete, zda požití pilulky léku ovlivňuje hodnoty dolního systolického krevního tlaku. Data obsahují údaje o systolickém tlaku před požitím pilulky léku x_1 a po požití pilulky x_2.

Data: Krevní tlak před x_1 [mm Hg] a po požití léku x_2 [mm Hg].

120	118,	136	122,	160	143,	98	105,	115	98,
110	98,	180	180,	190	175,	138	105,	128	112,

Úloha B7.12 *Vliv množství prodaného benzinu na obsah olova v krvi novorozeňat*

Data přinášejí měsíční objem prodaného olovnatého benzinu x_1 v tunách ve státě Massachusetts a střední hodnotu koncentrace olova v krvi x_2 v µg/100 ml, určeného v pupeční šnůře dětí, narozených v hlavní Bostonské nemocnici v průběhu 14 měsíců roku 1980-1981.

Data: Datum, hmotnost olovnatého benzinu x_1 [tuny], střední hodnota koncentrace olova v krvi x_2 [µg/100 ml].

březen 1980	141	6.4,	duben 1980	166	6.1,	květen 1980	161	5.7,	
...	...	...	...	...	...	...	...	...	...
březen 1981	82	4.5,	duben 1981	75	6.0,				

Úloha B7.13 *Faktory ovlivňující chemickou otravu berylliem*

Předzvěstí chronické otravy berylliem je vyšší rychlost blastogenické lymfocitní transformace (tzv. LT poměr). Tato transformace se objevuje především u lidí exponovaných berylliem v povrchových dolech. Vyšetřete závislost LT poměru x_4 na stáří x_1, délce zaměstnání v dolech x_2 a délce exponování berylliem x_3. Testujte především statistickou významnost délky zaměstnání a délky expozice berylliem. Sestrojte také parciální regresní grafy všech tří parametrů a diskutujte co indikují.

Data: Stáří x_1 [roky], délka zaměstnání v dolech x_2 [roky], délka exponování berylliem x_3 [roky], LT poměr x_4.

42	13	1	5.4,	46	14	2	7.3,	43	4	4	3,
...	...	...	...	...	...	...	...	...	...	...	...
26	3	4	2.5,	28	9	4	0.8,	26	4	2	1.2,

Úloha B7.14 *Faktory ovlivňující produkci maniok*

Maniok, plodina s škrobovitým kořenem, je hlavním, ale také dietním produktem v mnoha tropických zemích, především ve středoafrických republikách. Existuje korelace produkce manioku x_3 a roku x_1 nebo produkce manioku x_3 a velikosti obdělané půdy x_2? Vedle korelačních koeficientů vyšetřete korelaci také v parciálním reziduálním grafu. Aplikujte Studentův *t*-test statistické významnosti jednotlivých parametrů.

Data: Rok sklizně x_1 [rok], obdělaná půda x_2 [ha], produkce manioku x_3 [tisíce tun].

1961	204	746,	1962	204	746,	1963	204	746,	1964	204	746,
..	..	..	..	..	..	..	..	..	..	..	..
1982	331	676,									

Úloha B7.15 *Vliv stáří pstruha na koncentrace polychlorovaných bifenylů v jeho těle*
Vyšetřete, zda existuje přímý lineární vztah mezi stářím pstruha x_1 a obsahem rakovinotvorných látek polychlorovaných bifenylů PCB v jeho těle x_2 v jednotkách ppm.
Data: Stáří pstruha x_1, koncentrace PCB v jeho těle x_2 [ppm].

1.0	0.6,	1.0	1.6,	1.0	0.5,	1.0	1.2,	2.0	2.0,	2.0	1.3,	2.0	2.5,	3.0	2.2,	3.0	2.4,
3.0	1.2,	4.0	3.5,	4.0	4.1,	4.0	5.1,	5.0	5.7,	6.0	3.4,	6.0	9.7,	6.0	8.6,	7.0	4.0,
7.0	5.5,	7.0	10.5,	17.5	13.4,	8.0	4.5,	9.0	30.4,	11	12.4,	12.0	13.4,	12	26.2,	12	7.4,

Úloha B7.16 *Přírůstek počtu hrochů v Zambii u řeky Luangwa v čase*
Vyšetřete, zda změny v předpisech pro ochranu hrochů vedou k významnému přírůstku jejich počtu. Data obsahují informace o čase x_1 [roky] a odhadovaném počtu hrochů x_2 v Zambii na území poblíž řeky Luangwa. Stanovte autokorelační koeficient prvního řádu a významnost lineárního, resp. logaritmického trendu. Odhadněte, jaký počet hrochů bude v roce 2000?
Data: Čas x_1 [kalendářní rok], odhad počtu hrochů x_2 [počet].

1970	2815,	1972	2919,	1975	2342,	1976	4501,	1977	5147,
1978	4765,	1979	5151,	1981	4884,	1982	6293,	1983	6544,

Úloha B7.17 *Znečištění ovzduší a počet zemřelých ve městě*
Vyšetřete, zda existuje korelace na hladině významnosti $\alpha = 0.05$ mezi znečištěním ovzduší nad 60 velkými městy x_1 a počtem zemřelých ve městě x_2.
Data: Znečištění ovzduší x_1, počet zemřelých x_2.

105	791,	20	1113,	648	862,	144	840,	43	1071,	4	824,
...	...	...	...	...	...	...	...	...	...	...	...
23	959,	65	968,	31	954,	4	923,	11	942,	1	892,

Úloha B7.18 *Vliv imobilizační dávky na vzdálenost útěku před padnutím zvířete*
Rozhodněte, zda mezi dávkou imobilizační, uspávací injekce x_1, vztažené na 1 kg hmotnosti zvířete, a vzdáleností, kterou ještě antilopa uběhne než padne x_2 existuje lineární závislost.
Data: Dávka uspávací injekce x_1 [ml/kg], vzdálenost x_2 [m].

0.020	850,	0.011	950,	0.022	750,	0.023	3000,	0.033	100,
0.020	350,	0.026	450,	0.021	1300,	0.017	1150,		

Úloha B7.19 *Vliv množství farmaka na dobu práce pacienta*
Byl sledován účinek množství podpůrného farmaka na organismus v době, ve které je pacient schopen provést standardní manuální výkon. Nalezněte empirický regresní model pro vyjádření doby manuální práce x_2 na množství farmaka x_1.
Data: Množství farmaka x_1 [mg], doba práce x_2 [min].

05	48,	20	46,	25	55,	30	54,	35	60,	40	58,	45	73,
50	74,	55	82,	60	90,	65	105,	70	130,	75	200,		

7.4.2 Analýza chemických a fyzikálních dat

Úloha C7.01 *Vliv teploty na odbarvování barviva*
Při studiu vlivu teploty na odbarvování organického barviva byla získána experimentální data, str. 43 v cit.[11] o teplotě [K] x_1 a absorbanci x_2. Vyšetřete na hladině významnosti $\alpha = 0.05$, zda data vystihuje lineární, resp. exponenciální funkce.
Data: Teplota x_1 [K], absorbance x_2.

460	0.3,	450	0.3,	440	0.4,	430	0.4,	420	0.6,	410	0.5,
450	0.6,	440	0.6,	430	0.6,	420	0.7,	410	0.6,	400	0.6,
420	0.6,	410	0.6,	400	0.6,						

Úloha C7.02 *Pevnost dřeva a jeho specifická hmotnost*
Vyšetřete, zda existuje korelace mezi pevností dřevěné tyčinky x_1 a specifickou hmotností dřeva x_2, a to analýzou následujících dat deseti náhodně vybraných tyčinek, str. 437 v cit.[12]. Diskutujte příčiny této korelace.
Data: Limita pevnosti dřevěné tyčinky x_1, specifická hmotnost dřeva x_2 [kg/dm^3].

11.14	0.499,	12.74	0.558,	13.13	0.604,	11.51	0.441,	12.38	0.550,
12.60	0.528,	11.13	0.418,	11.00	0.480,	11.02	0.406,	11.41	0.467,

Úloha C7.03 *Fyzikální vlastnosti planet Sluneční soustavy*
Vyšetřete, zda existují nějaké korelace mezi fyzikálními vlastnostmi planet Sluneční soustavy a pokuste se diskutovat jejich příčiny.
Data: Planeta, vzdálenost od Slunce x_1 [10^6 km], poloměr při rovníku x_2 [km], hmotnost x_3 [kg], střední hustota x_4 [g/m^3], počet známých měsíců x_5.

	x_1	x_2	x_3	x_4	x_5
Merkur	57.9	2439	3.3000e+23	5.42	0
...	...	...	...	...	...
Pluto	5900	1550	1.1000e+22	1.2	1

Úloha C7.04 *Vliv množství hnojiva na dosažený výnos*
Analýzou následujících dat vyšetřete, zda existuje lineární vztah mezi výnosem pšenice x_1 [bušel/akr] a množstvím použitého hnojiva x_2 [libra/akr], str. 388 v cit.[12]:
Data: Výnos pšenice x_1 [bušel/akr], množství hnojiva x_2 [libra/akr].

100	40,	200	50,	300	50,	400	70,	500	65,	600	65,	700	80,

Úloha C7.05 *Vliv difuze kyslíku vodní parou na teplotu hoření*
Vyšetřete, zda existuje korelace mezi teplotou x_1 hoření a difuzí kyslíku $x_2$ve vlhkém prostředí. Matsui, Tsuji a Makino sledovali rychlost hoření umělého grafitu v proudu vlhkého vzduchu za účelem vyšetření difuze kyslíku směsí vodní páry. Molární zlomek vody ve směsi kyslíku a dusíku byl 0.017. Experiment byl sledován při 9 teplotách.
Data: Teplota plamene x_1 [stupně Kelvina, x_2 [difuze kyslíku].

1000	1.69,	1100	1.99,	1200	2.31,	1300	2.65,	1400	3.01,
1500	3.39,	1600	3.79,	1700	4.21,	1800	4.64,		

Úloha C7.06 *Termogravimetrické a redox stanovení olejů*
Termogravimetrické váhy představují přístroj k vyšetření termochemického chování chemické sloučeniny. El Naga a Salem (1986) porovnali termogravimetrii TG se standardní metodou SM vyhodnocení termooxidační stability alkalických olejů a jejich aditiv, např. transformátorových olejů, olejů k turbíně nebo k transmisi. Pro každý vzorek bylo určeno množství oxidovatelné sloučeniny x_1 termogravimetricky a celkové množství zoxidovaných produktů pak standarní redox metodou x_2. Vyšetřete, zda obě metody poskytují stejné kontrolní hodnoty x_1 a x_2.
Data: TG je množství oxidovatelné sloučeniny x_1 [% hmotnosti], SM je množství zoxidovaných produktů x_2 [%].

25.4	2.3,	27.11	2.5,	28.0	2.65,	17.9	1.3,	18.9	1.45,
22.9	1.9,	30.8	3.3,	18.6	1.4,	24.4	2.1,	29.8	2.9,

Úloha C7.07 *Poměr nečistot v pevném heliu a teplota*
Při teplotách okolo absolutní nuly (-273 EC) vykazuje helium vlastnosti, jež odporují obvyklým fyzikálním zákonům. Pevné helium spolu s nečistotami o hmotnostním zlomku x_1 bylo umístěno v tepelně izolovaném boxu. Experiment byl proveden s heliem v pevném stavu při rozličných teplotách x_2, blízkých absolutní nule. Vyšetřete, zda existuje korelace mezi proměnnými x_1 a x_2.
Data: Hmotnostní zlomek x_1 [-], teplota x_2 [K].

0.315	-262,	0.202	-265,	0.204	-256,	0.620	-267,	0.715	-270,
0.935	-272,	0.957	-272,	0.906	-272,	0.985	-273,	0.987	-273,

Úloha C7.08 *Tlak par bromokomplexu v baterii v závislosti na její teplotě*
Bajpal ukázal, že operační charakteristiky baterie, obsahující zinek a jeho bromo-komplex, jsou částečně závislé na tlaku par tetramethylamoniumbromidu v baterii. Data obsahují tlak par tetramethylamoniumbromidu x_2 v závislosti na teplotě x_1. Vyšetřete, zda existuje lineární vztah mezi oběma proměnnými x_1 a x_2.
Data: Teplota x_1 [EC], tlak par x_2 [mm Hg].

00	0.7,	5.10	4.30,	10.4	11.0,	14.7	17.4,
25.5	28.7,	32.0	35.6,	39.5	49.0,	49.7	67.0,

Úloha C7.09 *Vliv teploty solární cely na množství uvolněné energie*
Solární systémy nabývají stále větší důležitosti a jejich pořizovací cena klesá. Laminované solární moduly obsahují vysoce kvalitní solární cely křemíkových krystalů, spojených do série, aby poskytly co největší výstupní elektrickou energii. Výzkum byl zaměřen k vyšetření závislosti mezi teplotou solární cely x_1 ve stupních Celsia a množstvím

elektrické energie uvolněném v megawatech na cm^2 x_2. Vyšetřete, zda existuje korelace mezi oběma proměnnými x_1 a x_2. Nalezněte empirický regresní model pro vyjádření vlivu teploty na množství uvolněné energie.

Data: Teplota solární cely x_1 [EC], množství energie x_2 [megawatt/cm^2].

9	25,	25	70,	20	50,	12	30,	15	45,	22	60,		

Úloha C7.10 *Vliv koncentrace inhibitoru NaH_2PO_4 na korozi železa*

Andrzejczek studoval závislost koroze železa na inhibičních vlastnostech dihydrofosforečnanu sodného. Data ukazují míru koroze Armco železa ve vodě x_2 v závislosti na koncentraci x_1 inhibitoru NaH_2PO_4. Vyšetřete, zda existuje nějaký typ závislosti mezi oběma proměnnými x_1 a x_2.

Data: Koncentrace inhibitoru NaH_2PO_4 x_1 [ppm], míra koroze x_2 [%].

2.50	3.08,	5.03	6.95,	7.60	6.30,	11.60	5.75,	13.00	5.01,
19.60	1.43,	26.20	0.93,	33.00	0.72,	40.00	0.68,	50.00	0.65,
55.00	0.56,								

Úloha C7.11 *Vliv obsahu asfaltu na permeabilitu asfaltového betonu*

Woelfl a kol. studovali vliv obsahu asfaltu na stabilitu a permeabilitu hrubozrného asfaltového betonu. Každý ze čtyř vzorků tohoto betonu obsahoval odlišné procento asfaltu x_1. Permeabilita betonu byla určena prouděním odvzdušněné vody vzorkem betonu a měřením ztráty vody. Hodnoty permeability x_2 v palcích za hodinu byly změřeny pro 24 vzorků o různém procentickém obsahu asfaltu x_1. Vyšetřete, zda existuje typ závislosti mezi oběma proměnnými x_1 a x_2.

Data: Obsah asfaltu x_1 [%], permeabilita x_2 [palec/h].

3	1189,	3	840,	3	1020,	3	980,	4	1440,	4	1227,	4	1022,	4	1293,
5	1227,	5	1180,	5	980,	5	1210,	6	707,	6	927,	6	1067,	6	822,
7	853,	7	900,	7	733,	7	585,	8	395,	8	270,	8	310,	8	208,

Úloha C7.12 *Vliv vlnočtu a tloušťky filmu na absorbanci PPFPO*

Pacansky a kol. studovali infračervené reflektanční spektra látky PPFPO, tj. polyperfluoropropylenoxidu, viskózní tekutiny, která se užívá v elektronickém průmyslu jako mazadlo. Absorbance x_3 byla zaznamenána pro různé vlnočty x_1 v cm^{-1} a tloušťky filmu x_2 v mm na spektrofotometru Perkin-Elmer 621. Vyšetřete, zda lze použít lineární regresní model pro popis absorbance

$$x_3 \text{ ' } \beta_0 \text{ \% } \beta_1 \, x_1 \text{ \% } \beta_2 \, x_2 .$$

Data: Vlnočet x_1 [cm^{-1}], tloušťka filmu x_2 [mm], absorbance x_3.

740	1.1	0.231,	740	0.62	0.107,	740	0.31	0.053,
...	...	...	...	...	...	...	...	...
1235	0.31	0.934,						

Úloha C7.13 *Deformace měděného drátu v závislosti na jeho zatížení*

Vyšetřete, zda existuje lineární vztah zátěže měděného drátku x_1 v newtonech a prodloužení jeho délky x_2 v cm.

Data: Zátěž x_1 [newton], prodloužení x_2 [cm].

00	0.00,	10	0.05,	20	0.10,	30	0.15,
40	0.20,	50	0.25,	60	0.30,	70	1.25,

7.4.3 Analýza environmetálních, potravinářských a zemědělských dat

Úloha E7.01 *Vliv stáří stromu na roční sklizeň jablek*

U 42 zákrsků jabloní bylo zaznamenáno stáří stromu x_1 a roční sklizeň jablek v kg x_2. Rozhodněte, zda v rámci sledovaného souboru jabloní existuje lineární vztah mezi sklizní plodů a stářím stromu.

Data: Stáří stromu x_1 [roky], roční sklizeň x_2 [kg].

3.0	5.0,	4.0	8.0,	6.0	8.0,	8.0	10.0,	5.0	9.0,	6.0	10.0,	7.0	7.0,
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6.0	10.0,	9.0	4.0,	4.0	5.0,	4.0	7.0,	7.0	8.0,	9.0	6.0,	9.0	5.0,

Úloha E7.02 *Vliv nadmořské výšky na hektarový výnos ječmene*

Jsou dány údaje o 27 vybraných pozemcích, na nichž zemědělské závody pěstují v určité oblasti ozimý ječmen. Nadmořskou výšku pozemku v metrech označíme x_1, hektarový výnos ječmene v t/ha x_2. Nalezněte regresní model, popisující vliv nadmořské výšky x_1 na výnos ječmene x_2.

Data: Nadmořská výška pozemku x_1 [m], hektarový výnos ječmene x_2 [t/ha].

215.0	6.30,	220.0	6.50,	228.0	5.90,	246.0	5.80,	256.0	5.50,	260.0	5.60,
...	...	...	...	...	...	...	...	...	...	...	...
468.0	3.60,	475.0	3.30,	489.0	3.40,						

Úloha E7.03 *Vliv množství ledku amonného na výnos obiloviny*

Zemědělský výzkumný ústav zkoumá korelaci výnosu určité obiloviny x_2 v t/ha a množství hnojiva ledku amonného x_1 v kg/ha. Je předpoklad, že vhodným regresním modelem je parabola druhého stupně. Ověřte tuto alternativu.

Data: Množství hnojiva x_1 [kg/ha], výnos obiloviny x_2 [t/ha].

1.90	40.0,	2.50	50.0,	2.90	60.0,	3.10	65.0,	3.10	70.0,
3.30	75.0,	3.30	80.0,	3.50	85.0,	3.50	90.0,	3.40	100.0,

Úloha E7.04 *Vliv množství dusíku v hnojivu na výnos plodiny*

Pro popis závislosti výnosu určité plodiny y na množství dusíku v použitém hnojivu x byla doporučena funkce $y = e^{b_1 + b_2 x} x^{b_3}$. Vyšetřete správnost tohoto modelu pro aditivní model (původní data) a multiplikativní model (logaritmická transformace obou stran).

Data: Množství dusíku v hnojivu x [kg/ha], výnos určité plodiny y [t/ha].

15.1	0.09,	57.3	0.32,	103.3	0.69,	174.6	1.51,	191.5	2.29,
193.2	3.06,	178.7	3.39,	172.3	3.63,	167.5	3.77,		

Úloha E7.05 *Vliv věku stromů určité odrůdy na velikost úrody*
Při zkoumání úrody stromů určité odrůdy x_2 a věku stromů x_1 od jejich přesazení zjistil ovocnářský ústav o náhodně vybraných stromech následující údaje (v tabulce dat). Pokuste se navrhnout vhodný regresní model.

Data: Stáří stromu x_1 [roky], roční sklizeň x_2 [kg].

2.00	2.00,	3.00	2.00,	2.00	3.00,	10.0	5.00,	5.00	4.00,
3.00	3.00,	1.00	2.00,	4.00	3.00,	7.00	4.00,	6.00	4.00,
3.00	3.00,	12.0	5.00						

Úloha E7.06 *Pevnost dřeva a měrná hmotnost*
Bylo zkoumáno, zda koreluje síla potřebná k prasknutí nosníku, tzv. pevnost dřevěného nosníku x_2, a měrná hmotnost dřeva x_1. Náhodně bylo vybráno deset dřevěných trámů, nosníků. Ty byly namáhány postupně zvyšující se silou k prasknutí. Vyšetřete, zda typ závislosti měrné hmotnosti dřeva x_1 a meze pevnosti dřevěného nosníku x_2. Odhadněte mez pevnosti nosníku pro měrnou hmotnost 0.590.

Data: Měrná hmotnost dřeva x_1, mez pevnosti nosníku x_2.

0.499	11.14,	0.558	12.74,	0.604	13.13,	0.441	11.51,	0.550	12.38,
0.528	12.60,	0.418	11.13,	0.480	11.70,	0.406	11.02,	0.467	11.41,

Úloha E7.07 *Vliv HC u vozidel na jejich emise CO*
Lorezen studoval vliv HC u rozličných vozidel x_1 na emisi CO x_2. Vyšetřete, zda existuje korelace mezi oběma proměnnými x_1 a x_2 na hladině významnosti $\alpha = 0.05$.

Data: HC vozidla x_1, emise CO x_2 [g/m^3].

0.65	14.7,	0.55	12.3,	0.72	14.6,	0.83	15.1,
0.57	5.0,	0.51	4.1,	0.43	3.8,	0.37	4.1,

Úloha E7.08 *Vliv nadmořské výšky na srážkovou činnost*
Byla sledována závislost srážkové činnosti, deštů v cm^3 na cm^2 plochy, na nadmořské výšce v některých městech USA a zda nadmořská výška místa x_1 a množství srážek, spadlých v tomto místě x_2 nějak souvisí.

Data: Nadmořská výška x_1 [m], množství srážek x_2 [cm^3. cm^{-2}].

Berlin	284	97.5,	Bradford	296	112.4,
...	...	...	...	...	...
Ticonderoga	50	86.4,			

Úloha E7.09 *Teplota vzduchu a tloušťky vrstvy sněhu na Zemi*
Existuje mnoho lokalit na Zemi, kde sníh je hlavní zásobárnou vody. Bariéry sněhu na Zemi byly sledovány z družice spolu s měřením teploty vzduchu v této lokalitě. Navrhněte typ závislosti mezi teplotou vzduchu x_1 a tloušťkou vrstvy sněhu x_2.

Data: Teplota vzduchu x_1 [EC], vrstva sněhu x_2 [cm].

-62	21,	-41	13,	-36	12,	-26	3,
-33	6,	-56	22,	-50	14,	-66	19,

Úloha E7.10 *Korelace pH a počtu druhů ryb v řece Millers River*
Průmyslové spalování olejů a uhlí způsobuje kyselé deště, které silně ničí faunu a floru jezer, řek a lesů. Byl sledován počet druhů ryb přežívajících v řece Millers River ve střední části státu Massachusetts. Kyselost vody v řece je ovlivněna 15 přítoky, jež se liší v hodnotě pH. Vyšetřete, zda existuje na hladině významnosti $\alpha = 0.05$ korelace mezi pH přítoku x_1 a počtem druhů ryb v řece x_2.
Data: Název přítoku, pH přítoku x_1, počet druhů ryb v řece x_2.

Moss	6.3	6,	Orcutt	6.3	9,	Ellinwood	6.3	6,
...	...	...	...	...	...	...	...	...
Lawrence	5.4	5,	Wilder	4.7	0,	Templeton	4.5	0,

Úloha E7.11 *Vliv velikosti obsazeného území na počet párů Racků mořských*
U 22 kolonií mořských racků na Shetlandských a Orkneyských ostrovech byla studována závislost počtu rodičovských párů Racka mořského x_2 na velikosti okupovaného území x_1, využívaného racky k získání základní potravy. Vyšetřete, zda existuje korelace mezi oběma proměnnými x_1 a x_2 na hladině významnosti $\alpha = 0.05$. *Data:* Kolonie, velikost území x_1 [km^2], počet rodičovských párů mořských racků x_2 [počet].

W. Unst	208	311,	Hermaness	1570	3872,
...	...	...	...	...	...
Gruney	565	1364,	Fair Isle	3957	17000,

Úloha E7.12 *Vztah mezi nadmořskou výškou a počtem uschlých lesních stromů*
Většina lesních porostů kanadských červených smrků ve výše položených oblastech Apalačských hor začíná vykazovat chorobu - usychání stromů. Přispívá k tomu především znečištění ovzduší. Největší vliv na zkázu lesů ve vyšších nadmořských výškách, kde se tyto stromy vyskytují, má totiž obsah těžkých kovů v ovzduší a působení kyselých deštů. Vyšetřete 64 lokalit Apalačských hor a testujte, zda existuje korelace nadmořské výšky x_1 a procentuálního počtu odumřelých, uschlých lesních stromů x_2 na hladině významnosti $\alpha = 0.05$.
Data: Nadmořská výška x_1 [m], procento odumřelých stromů x_2 [%].

1615	5,	1768	13,	1524	6,	1311	21,	1128	4,
...	...	...	...	...	...	...	...	...	...
950	35,	1000	42,	1060	58,	1120	24,		

Úloha E7.13 *Vliv vzdálenosti od koželužny na koncentraci chromitých iontů v řece*
Z koželužen na řece Cocheco River v New Hampshire je vypouštěn odpad, obsahující jedovaté ionty chromu. Koncentrace chromu x_2 byla měřena v rozličných vzdálenostech od koželužny x_1 a zkoumána hypotéza o postupném usazovaní chromitých kalů v řece. Vyšetřete, zda existuje korelace mezi oběma proměnnými x_1 a x_2 a pokuste se najít i vhodný regresní model.
Data: Vzdálenost od koželužny x_1 [km], koncentrace chromu x_2 [ppm].

2.6	221,	5.6	77,	4.4	72,	5.1	56,
7.2	44,	9.9	56,	10.8	47,		

Úloha E7.14 *Vztah mezi odlesněním krajiny, populačním koeficientem a velikostí hrubého*

národního produktu u 50 států světa. U 50 států byla zkoumána závislost procenta odlesnění krajiny x_3 na populačním koeficientu x_b, tj. počtu narozených dětí minus počet zemřelých obyvatel děleno 1000 obyvateli a hrubým národním důchodem x_2. Vyšetřete, zda existuje korelace mezi proměnnou x_3 a oběma proměnnými x_1 a x_2, a konečně i vzájemnou korelaci obou proměnných x_1 a x_2. Pokuste se nalézt regresní model.

Data: Země, populační koeficient x_1, hrubý národní důchod x_2 [US $], odlesnění krajiny x_3 [%].

Angola	25.1	940	1.0,	Benin	28.5	310	2.6,
...	...	...	...	...	...	...	...
Zaire	29.4	190	0.2,	Zambie	33.0	640	1.2,

Úloha E7.15 *Hladina PCB v přístavech v roce 1984 a 1985*

V 37 přístavech v USA byla sledována hladina PCB v roce 1984 x_1 a v roce 1985 x_2. Vyšetřete, zda existuje lineární vztah mezi oběma proměnnými x_1 a x_2, resp. mezi jejich logaritmy, tj. $\log x_1$ a $\log x_2$. Nalezněte vlivné body, odlehlé body a upozorněte na ně.

Data: Hladina PCB v 1984 x_1 [ppb], v 1985 x_2 [ppb].

95.28	77.55,	52.97	29.23,	533.58	403.10,	308.46	192.15,
...	...	...	...	...	...	...	...
6.60	5.08,						

7.4.4 Analýza hutnických a mineralogických dat

Úloha H7.01 *Destrukční tlak a axiální napětí břidlice*

V Caswellově publikaci byla testována kompaktní břidlice, zda je použitelná jako stavební materiál vhodný do základů. Byla vyšetřována závislost destrukčního tlaku x_2 na axiálním napětí v fragmentované břidlici x_1, které je potřebné k rozdrcení vzorku břidlice. Vyšetřete, zda existuje nějaká korelace mezi oběma proměnnými, eventuálně se pokuste navrhnout regresní model.

Data: Axiální napětí ve fragmentovaném materiálu x_1 [%], velikost destrukčního tlaku x_2 [kPa].

1.0	500,	2.8	2000,	4.3	2750,	6.0	3500,	7.5	4375,
9.0	4875,	10.5	5250,	13.5	6000,	16.7	6625,	19.8	7000,
23.0	7125,	26.0	7000,	27.5	7125,				

Úloha H7.02 *Vliv hloubky dolu na množství těženého metanu*

V uhelných dolech se často objevuje methan. Metan je důležitý plyn, který se vyplatí čerpat, těžit a využívat. Metan se vyskytuje více ve hlubších dolech a množství x_2 je závislé na hloubce dolu x_1. Nalezněte regresní model.

Data: Hloubka dolu x_1 [stopy], množství metanu x_2 [cm^3/g].

175	2.5,	252	5.7,	318	8.7,	356	10.8,	516	11.8,
553	13.1,	571	11.8,	561	11.5,	556	10.9,	823	15.5,
892	16.8,	1439	17.1,	1440	16.7,	489	10.9,	1296	17.5,

Úloha H7.03 *Roční průměr hodnot odtékající deštové vody, srážky a sediment v horách*

Nového Zélandu. Byl sledován roční průměr odtékající deštové vody x_1 [mm^3.mm^{-2}], roční průměr srážek x_2 [mm] a roční průměr sedimentu x_3 [tuna/km^2] na 19 místech v horách Severních Alp Nového Zélandu. Vyšetřete korelace jednotlivých dvojic proměnných a diskutujte je.

Data: Roční průměr odtékající deštové vody x_1 [mm^3.mm^{-2}], roční průměr srážek x_2 [mm], roční průměr sedimentu x_3 [tuna/km^2].

11300	8600	14900,	11500	10070	32600,	6800	7300	12500,
...	...	...	...	...	...	...	...	...
1480	1947	3010,						

Úloha H7.04 *Vliv vzdálenosti od pobřeží na hustotu korálu*

Risk a Sammarco nalezli, že skeletární hustota mořského korálu *Porites lobata* se zvyšuje se vzdáleností od australských břehů, a to v důsledku rozdílu přílivového a odlivového složení znečištěné mořské vody. Vyšetřete, zda existuje vhodný regresní model vzdálenosti od pobřeží x_1 v km a hustotou korálu x_2 v g/cm^3.

Data: Vzdálenost od pobřeží x_1 [km], hustota korálu x_2 [g/cm^3].

3.5	1.337,	3.5	1.216,	3.5	1.309,	14.3	1.053,	14.3	1.082,
...	...	...	...	...	...	...	...	...	...
74.5	1.589,	74.5	1.461,						

Úloha H7.05 *Doba hoření dřeva v kamnech, množství spotřebovaného dřeva a množství uvolněného oxidu uhelnatého CO.* Ceny topných olejů a uhlí neustále stoupají, řada lidí se vrací ke dřevu jako topnému materiálu. Kamna na dřevo znečišťují vytápěný prostor o oxid uhelnatý a uhličitý a spotřebovávají kyslík v místnosti. Vyšetřete významnost korelací doby hoření dřeva v kamnech x_1 v hodinách, množstvím spotřebovaného dřeva x_2 v kg a koncentrací uvolněného oxidu uhelnatého CO x_3. Pokuste se odhadnout regresní model vzniku oxidu uhelnatého CO na době hoření a množství spotřebovaného dřeva.

Data: Doba hoření dřeva v kamnech x_1 [hodiny], množství spotřebovaného dřeva x_2 [kg], koncentrace uvolněného oxidu uhelnatého CO x_3 [ppm].

14.8	37.3	2.8,	8.8	38.4	1.2,	8.8	38.4	1.2,
...	...	...	...	...	...	...	...	...
10.4	32.4	35.0,	5.4	23.2	43.0,	9.5	38.6	3.5,

7.4.5 Analýza ekonomických a sociologických dat

Úloha S7.01 *Vliv stáří stroje na náklady na opravy*

Dá se předpokládat, že čím starší je stroj, tím nákladnější je jeho údržba. Vyšetřete, zda mezi stářím stroje x_1 [roky] a cenou jeho opravy x_2 [$] existuje lineární závislost, str. 41 v cit.[11].

Data: Stáří stroje x_1 [roky], cena jeho opravy x_2 [$].

4.5	619,	4.5	1049,	4.5	1033,	4.0	495,	4.0	723,	4.0	681,
5.0	890,	5.0	1522,	5.5	987,	5.0	1194,	0.5	163,	0.5	182,
6.0	764,	6.0	1373,	1.0	978,	1.0	466,	1.0	549,		

Úloha S7.02 *Matematické a verbální výsledky studentů*
U 8 studentů byly porovnávány bodové výsledky zkoušek matematických x_1 a verbálních x_2 předmětů, str. 514 v cit.[12]. Vyšetřete, zda existuje nějaká korelace na hladině významnosti $\alpha = 0.05$.

Data: Body ze zkoušek matematických x_1 a verbálních x_2 předmětů.

80	65,	50	60,	36	35,	58	39,	72	48,	60	44,	56	48,	68	61,

Úloha S7.03 *Vliv počtu obyvatel bytu na počet telefonních hovorů*
Ze 36 vybraných bytových telefonních stanic je třeba vyšetřit korelaci znaku x_1, tj. počtu obyvatel bytu starších 10 let, a znaku x_2, tj. počtu místních telefonních hovorů za měsíc. Vztah mezi oběma proměnnými znázorněte bodovým diagramem. Co lze usoudit o možné korelaci? Lze navrhnout regresní model?

Data: Počet obyvatel bytu starších 10 let x_1, počet místních telefonních hovorů x_2.

4.0	35.0,	5.0	92.0,	3.0	75.0,	1.0	5.0,	2.0	8.0,	4.0	120.0,
...	...	...	...	...	...	...	...	...	...	...	...
2.0	35.0,	3.0	62.0,	6.0	56.0,	4.0	57.0,	2.0	53.0,	3.0	50.0

Úloha S7.04 *Vztah mezi písemnou a ústní částí zkoušky*
Zkouška z matematiky se skládá z písemné x_1 a ústní části x_2, které se hodnotí nezávisle na sobě, písemná část x_1 0 až 20 body a ústní x_2 0 až 10 body. Na základě hodnocení 20 studentů vyšetřete, zda existuje korelace mezi oběma proměnnými na hladině významnosti $\alpha = 0.05$.

Data: Písemná část x_1 [0 až 20 bodů], ústní část x_2 [0 až 10 bodů].

6.0	4.0,	11.0	7.0,	8.0	6.0,	18.0	8.0,	6.0	3.0,	11.0	5.0,	6.0	6.0,
3.0	4.0,	14.0	9.0,	7.0	8.0,	17.0	10.0,	12.0	9.0,	8.0	6.0,	4.0	5.0,
15.0	7.0,	20.0	10.0,	13.0	8.0,	5.0	6.0,	10.0	7.0,	0	3.0		

Úloha S7.05 *Počet odpracovaných směn a počet výrobků*
V dílně pracuje 15 dělníků, u nichž byl zjištěn počet směn odpracovaných za měsíc x_1 a počet zhotovených výrobků x_2. Ověřte lineární regresní model.

Data: Počet směn odpracovaných za měsíc x_1, počet zhotovených výrobků x_2.

20.0	92.0,	21.0	93.0,	18.0	83.0,	17.0	80.0,
...	...	...	...	...	...	...	...
21.0	96.0,	15.0	64.0,	15.0	81.0		

Úloha S7.06 *Korelace kusů výrobku a jednicových nákladů*
Data obsahují údaje z osmi závodů o výrobě určitého výrobku v kusech x_1 a o jednicových nákladech x_2. Vyšetřete korelaci mezi oběma proměnnými a navrhněte regresní model.

Data: Počet kusů výrobku x_1 [ks], jednicové náklady x_2 [Kč/ks].

500	95,	800	88,	200	100,	1000	87,	400	96,
1500	84,	1200	84,	2000	81				

Úloha S7.07 *Vliv rychlosti jízdy autem Škoda 120 na spotřebu benzinu*
Vystihněte regresním modelem vztah spotřeby benzinu v přepočtu na 100 km x_2 a rychlosti jízdy automobilu Škoda 120 x_1.

Data: Rychlost jízdy x_1 [km/h], spotřeba benzínu x_2 [l/100 km].

42.0	8.11,	48.0	7.60,	55.0	7.30,	61.0	7.10,	67.0	6.80,
75.0	6.70,	80.0	6.50,	86.0	7.00,	93.0	7.80,	108.0	8.30,
115.0	8.60,	124.0	9.20,	130.0	9.60				

Úloha S7.08 *Měsíční příjem a vydání za potraviny v rodině*
Vybrané dvoučlenné domácnosti poskytly údaje o měsíčním příjmu x_1 v tisících Kč a vydání za potraviny x_2 v tisících Kč. Vyšetřete korelaci mezi oběma proměnnými. Ověřte, zda na hladině významnosti $\alpha = 0.05$ lze k vystižení regresního modelu užít např. logaritmickou funkci.

Data: Měsíční příjem x_1 [tisíce Kč], vydání za potraviny x_2 [tisíce Kč].

1.5	0.6,	2.1	0.9,	3.4	1.3,	3.5	1.2,	3.9	1.4,	4.2	1.6,	5.8	1.8,	6.4	1.6,	7.5	1.9,	9.0	2.1

Úloha S7.09 *Vliv zastavěné plochy domu na výši nájemného*
Vyšetřete na hladině významnosti $\alpha = 0.05$, zda existuje významná korelace tj. $\rho \$ 0.9$ zastavěné plochy domu x_1 ve stovkách čtverečních stop a nájemném x_2 v tisících dolarů.

Data: Plocha domu x_1 [stovky čtverečních stop], nájemné x_2 [tisíce US $].

15	1.9,	38	3.0,	23	1.4,	16	1.4,	16	1.5,	13	1.8,	20	2.4,	24	4.0,

Úloha S7.10 *Pracovní neschopnost pracovníků v závislosti na jejich průměrném věku a podílu zaměstnaných žen*. Od deseti vybraných závodů jsou k dispozici data o pracovní neschopnosti x_3 v procentech, průměrném věku pracovníků x_1 a podílu zaměstnaných žen na počtu pracovníků x_2 v procentech. Vyšetřete postupně korelaci dvojic všech tří proměnných: (a) x_3 na x_1, (b) x_3 na x_2, (c) x_1 na x_2. Za předpokladu, že procento pracovní neschopnosti závisí na průměrném věku pracovníků a dále i na podílu zaměstnaných žen, odhadněte regresní model.

Data: Průměrný věk pracovníků x_1 [roky], podíl žen na počtu pracovníků x_2 [%], pracovní neschopnost x_3 [%].

1.1	55.0	0.6,	2.5	54.6	1.0,	10.4	50.6	1.1,	4.5	51.1	2.0,
...	...	...	...	...	...	...	...	...	...	...	...
64.5	28.0	6.1,	70.8	24.6	6.3,	78.7	27.0	6.8,	90.2	17.6	7.5

Úloha S7.11 *Vliv výsledku testu zručnosti rukou, testu zručnosti prstů a počtu upevněných nýtů za jednotku času.* V podniku Loděnice se učni, kteří se ucházejí o přijetí na obor nýtovač, podrobují testu zručnosti rukou a prstů. Očekává se, že testy ukážou, jaké mají uchazeči předpoklady pro budoucí povolání. Po ukončení prvního roku učební doby se učni podrobují praktické zkoušce. Výsledky vstupního testu a praktické zkoušky, tj. počet upevněných nýtů za jednotku času pro 27 náhodně vybraných učňů jsou k dispozici. Vyšetřete, zda počet upevněných nýtů za jednotku času x_3 koreluje na hladině významnosti

$\alpha = 0.05$ s výsledky testu zručnosti rukou x_1 a zručnosti prstů x_2. Pokuste se i navrhnout regresní model.

Data: Výsledky testu zručnosti rukou x_1, testu zručnosti prstů x_2, počet upevněných nýtů za jednotku času x_3.

35.0	40.0	3.1,	33.0	44.0	4.0,	42.0	40.0	3.5,	34.0	38.0	3.0,
40.0	30.0	1.9,	36.0	32.0	2.0,	40.0	35.0	2.5,	38.0	36.0	3.0,
32.0	40.0	3.5,	40.0	35.0	2.5						

Úloha S7.12 *Vliv rychlosti auta na počet mil za hodinu na 1 galon benzinu*

Lze předpokládat, že spotřeba benzinu, vyjádřená v počtu ujetých mil na 1 galon benzinu x_2 [míle/galon], je závislá na rychlosti auta x_1 [míle/h]. Vyšetřete, zda lze pro vyjádření této závislosti použít lineární model.

Data: Rychlost auta x_1 [míle/h], počet mil na 1 galon x_2 [míle/galon].

15	14,	23	17,	30	20,	35	24,	42	26,	45	23,
50	18,	54	15,	60	11,	65	10,				

Úloha S7.13 *Vliv výkonu za směnu na procento vadných výrobků*

Jsou dány údaje o výkonu za směnu x_1 a procentu vadných výrobků x_2 u 20 náhodně vybraných pracovníků. Vyšetřete na hladině významnosti $\alpha = 0.05$, zda existuje korelace mezi oběma proměnnými a diskutujte její význam.

Data: Výkon za směnu x_1 [kusy], procento vadných výrobků x_2 [%].

18.0	5300.0,	3.00	2100.0,	20.0	6000.0,	15.0	5000.0,
...	...	...	...	...	...	...	...
12.0	4200.0,	7.00	3000.0				

Úloha S7.14 *Vztah mezi výsledky testu recitace u dětí a dosaženým věkem*

U náhodně vybraných dětí rozličného stáří byla bodována kvalita recitace x_2 v závislosti na dosaženém věku x_1. Vyšetřete, zda lze použít logaritmický model $x_2 = \beta_1 + \beta_2 \ln x_1$.

Data: Věk x_1, získané body x_2.

6.1	17.8,	7.2	47.4,	5.9	25.8,	6.3	24.3,	10.5	66.6,	11.0	91.4,

Úloha S7.15 *Vliv vzdálenosti od hasičské zbrojnice na škody způsobené požárem*

Pojišťovací společnost zkoumala závislost objemu škod požárem zničených objektů x_2 na předměstí vůči vzdálenosti k nejbližší hasičské zbrojnici x_1. Vyšetřete, zda existuje lineární vztah mezi proměnnými x_1 a x_2.

Data: Vzdálenost k hasičské zbrojnici x_1 [míle], objem požárem zničených objektů x_2 [10^3 \$].

3.4	26.2,	1.8	17.8,	4.6	31.3,	2.3	23.1,	3.1	27.5,
5.5	36.0,	0.7	14.1,	3.0	22.3,	2.6	19.6,	4.3	31.3,
2.1	24.0,	1.1	17.3,	6.1	43.2,	4.8	36.4,	3.8	26.1,

Úloha S7.16 *Maturitní výsledky, verbální a matematické skóre*
Maturitní známky se na amerických školách vyjadřují závěrečným SAT skóre verbálním x_1 a SAT skóre matematickým x_2 Vyšetřete, zda existuje lineární vytah mezi oběma skóre.
Data: Skóre verbální x_1, skóre matematické x_2.

421	476,	423	467,	422	465,	429	467,	424	470,	413	453,
437	470,	461	515,	460	512,	429	463,	430	465,		

Úloha S7.17 *Vliv velikosti domu na měsíční spotřebu elektřiny*
Elektrické závody se pokoušejí předpovědět měsíční spotřebu elektrické energie x_2 v kWh v závislosti na velikosti zastavěné plochy x_1, na které se obydlí nachází. Nalezněte regresní model.
Data: Velikost plochy x_1 [čtvereční stopy], měsíční spotřeba elektrické energie x_2 [kWh].

1290	1182,	1350	1172,	1470	1264,	1600	1493,
1710	1571,	1840	1711,	1980	1804,	2230	1840,
2400	1956,	2930	1954,				

Úloha S7.18 *Vliv velikosti motoru na ujeté míle na 1 galon benzinu*
Lze předpokládat, že spotřeba benzinu x_2 bude záviset na objemu válců benzinového motoru x_1. Vyšetřete, zda existuje významná korelace $\rho > 0.8$ mezi proměnnými x_1 a x_2, a pokuste se navrhnout vhodný regresní model.
Data: Typ auta, objem válců benzinového motoru x_1 [cm^3], ujeté míle na 1 galon benzinu x_2 [míle/galon].

WV Rabbit	970	42,	Mazda GLC	910	35,	Plymouth	1050	30,	
...	...	...	...	...	...	...	...	...	...
Datsun 200SX	1190	28,	Honda Civic	81	33,				

Úloha S7.19 *Vliv nahuštění pneumatik na ujeté míle na 1 galon benzinu*
Lze předpokládat, že spotřeba benzinu bude ovlivněna podhuštěnými či přehuštěnými pneumatikami. Vyšetřete, zda existuje typ závislosti mezi tlakem v pneumatice x_1 a ujetými mílemi na 1 galon benzinu x_2.
Data: Tlak v pneumatice x_1 [libra na palec^2], ujeté míle na 1 gallon benzinu x_2 [míle/galon].

30	29,	31	32,	32	36,	33	38,	34	37,	35	33,	36	26,

7.5 Kontrolní hodnoty (ADSTAT, NCSS2000)

Ve výsledcích značí *r* korelační koeficient, *D* koeficient determinace v procentech, regresní model $y = \beta_1(s_1) + \beta_2(s_2)\ .\ x$, počet odlehlých bodů *o*, počet extrémů *e*.

7.5.1 Analýza farmakologických a biochemických dat

B7.01 *r = 0.9433, D = 88.99 %, y = 4.44 (0.25) + 0.26 (0.03) x, 1o, 0e*
B7.02 *r = 0.9175, D = 84.18 %, y = 118.5 (8.4) + 9.03 (1.5) x, 1 e*
B7.03 *r = 0.7873, D = 61.99 %, y = 6.9 (5.6) + 0.0054 (0.0021) x, 0 o, 0 e*
B7.04 *r = 0.9267, D = 85.88 %, y = 0.76 (0.33) + 0.033 (0.008) x, 0 e*
B7.05 *(a) r = 0.9766, D = 95.38 %, x_4 = 0.13 (0.04) + 0.061 (0.003) x_1,* 2 o, 2 e, (b) *r = 0.9260, D = 85.74 %, x_4 = 1.67 (0.99) + 12.4 (1.1) x_2, 1 o, 2 e, (c) r = 0.4640, D = 21.53 %, x_4 = -11.8 (9.7) + 25.17 (10.0) x_3, 0 o, 3 e*
B7.06 *r = 0.9927, D = 98.54 %, y = -0.015 (0.010) + 0.029 (0.002)x, 1 o, 1 e*

B7.07 *r = 0.4653, D = 21.65 %, y = 178 (47) + 1.72 (1.23) x, 0 o, 1 e*
B7.08 *r = 0.6004, D = 36.05 %, y = -0.21 (0.27) + 1.24E-04 (6.73E-05) x, 1 o, 1 e*
B7.09 *r = 0.7633, D = 58.27 %, y = 10.6 (3.9) + 0.14 (0.0) x, 0 o, 0 e*
B7.10 *r = 0.5651, D = 31.93 %, y = 21.6 (13.1) + 0.30 (0.12) x, 0 o, 1 e*
B7.11 *r = 0.9323, D = 86.92 %, y = -2.9 (18.0) + 0.935 (0.128) x, 0 o, 1 e*
B7.12 *r = 0.6733, D = 45.33 %, y = 4.1 (0.6) + 1.49 (0.00) x, 0 o, 1 e*
B7.13 *r = 0.6109, D = 37.32 %, r_{12} = 0.2656, r_{13} = -0.2733, r_{23} = -0.6069, $R_{\bar{1}}$ = 0.3007, $r_{\bar{2}}$ 0.6156, R_3 = 0.6179, y = -0.55 (2.95) + 0.074 (0.045) x_1 + 0.193 (0.141) x_2 + 0.161 (0.556) x_3, 1 o, 1 e*
B7.14 *r = 0.6304, D = 39.74 %, r_{12} = 0.9522, R_1 = 0.9522, y = -7646 (16232) + 4.21 (8.40) x_1 + 0.565 (1.291) x_2, 1 o, 1 e*
B7.15 *r = 0.7364, D = 54.22 %, y = -1.45 (1.84) + 1.56 (0.28) x, 2 o, 3 e*
B7.16 *r = 0.8879, D = 78.84 %, y = -5.91E+05 (1.09E+05) + 301.3 (5.5) x, 1 o, 1 e*
B7.17 *r = 0.4330, D = 18.76 %, y = 921.1 (10.5) + 1.427 (0.408) x, 2 o, 4 e*
B7.18 *r = -0.6817, D = 46.46 %, y = 1667.0 (423.2) - 4.374E+04 (1.917E+04) x, 1 o, 1 e*
B7.19 *r = 0.8499, D = 72.23 %, y = -1.4 (17.0) + 1.87 (0.35) x, 1 o, 0 e*

7.5.2 Analýza chemických a fyzikálních dat

C7.01 *r = -0.6148, D = 37.80 %, y = 2.25 (0.61) - 0.0040 (0.0014) x, 0 o, 0 e*
C7.02 *r = 0.8814, D = 77.69 %, y = -0.331 (0.157) + 0.070 (0.013)x, 0 o, 0 e*
C7.03 *r = 0.9702, D = 94.13 %, r_{12} = -0.0458, r_{13} = -0.1606, r_{14} = -0.6670, r_{23} = 0.8683, r_{24} = -0.6501, r_{34} = -0.4352, R_1 = 0.9213, r_2 = 0.9608, R_3 = 0.8892, R_4 = 0.9580, 0 o, 0 e*
C7.04 *r = 0.9195, D = 84.55 %, y = 36.4 (5.0) + 0.059 (0.011) x, 1 o, 0 e*
C7.05 *r = 0.9982, D = 99.65 %, y = -2.10 (0.12) + 3.69E-03 (8.31E-05) x, 0 o, 0 e*
C7.06 *r = 0.9914, D = 98.29 %, y = -1.29 (0.16) + 0.142 (0.007) x, 1 o, 0 e*
C7.07 *r = -0.9233, D = 85.24 %, y = -257.3 (1.8) - 15.98 (2.35) x, 2 o, 0 e*
C7.08 *r = 0.9934, D = 98.68 %, y = -2.21 (1.70) + 1.31 (0.06) x, 1 o, 0 e*
C7.09 *r = 0.9836, D = 96.75 %, y = -0.41 (4.54) + 2.74 (0.25) x, 0 o, 0 e*
C7.10 *r = -0.8972, D = 80.49 %, y = 7.08 (0.75) - 0.166 (0.029) x, 0 o, 1 e*
C7.11 *r = -0.7735, D = 59.83 %, y = 1717 (151) - 150.3 (26.3) x, 1 o, 0 e*
C7.12 *r = 0.9304, D = 86.57 %, r_{12} = 0.2346, R_1 = 0.8933, R_2 = 0.4627, 1 o, 0 e*
C7.13 *r = 1.0000, D = 100.00 %, y = -5.32E-10 (2.50E-09) - 5.00E-03 (6.94E-11) x, 2 o, 0 e*

7.5.3 Analýza environmetálních, potravinářských a zemědělských dat

E7.01 *r = 0.1137, D = 1.29 %, y = 6.94 (0.88) + 0.101 (0.139) x, 1 o, 0 e*
E7.02 *r = -0.9264, D = 85.82 %, y = 8.00 (0.28) - 9.74 (0.00) x, 2 o, 0 e*
E7.03 *r = 0.9150, D = 83.72 %, y = -30.1 (16.0) + 33.30 (5.2) x, 2 o, 1 e*
E7.04 *r = 0.8319, D = 69.20 %, y = -0.539 (0.722) + 0.0188 (0.0047) x, 0 o, 1 e*
E7.05 *r = 0.9190, D = 84.46 %, y = 1.923 (0.230) + 0.292 (0.040) x, 1 o, 1 e*
E7.06 *r = 0.9131, D = 83.38 %, y = 6.514 (0.853) + 10.829 (1.709) x, 1 o, 1 e*
E7.07 *r = 0.8449, D = 71.38 %, y = -8.22 (4.64) + 30.12 (7.79) x, 0 o, 1 e*
E7.08 *r = 0.8587, D = 73.74 %, y = 89.62 (3.06) + 0.066 (0.006) x*
E7.09 *r = -0.9260, D = 85.74 %, y = -6.5 (3.5) - 0.438 (0.073) x*
E7.10 *r = 0.6957, D = 48.40 %, y = -13.86 (5.2) + 3.14 (0.90) x, 0 o, 2 e*
E7.11 *r = 0.7847, D = 61.58 %, y = -735 (787) + 3.30 (0.58) x, 1 o, 2 e*
E7.12 *r = 0.0918, D = 0.84 %, y = 29.1 (11.9) + 0.0088 (0.012) x*
E7.13 *r = -0.6691, D = 44.77 %, y = 173.6 (49.5) - 14.1 (7.0) x, 1 o, 0 e*
E7.14 *r = 0.2360, D = 5.57%, y = -0.390 (1.033, A) + 0.059 (0.036, A) x_1 + 1.09 10^{-4} (2.22 10^{-4}, A) x_2 , regresní model není statisticky významný, 4 o, 4 e,*
E7.15 *r = 0.9422, D = 88.77 %, y = 5.5 (12.5) + 0.982 (0.067) x, 4 o, 3 e*

7.5.4 Analýza hutnických a mineralogických dat

H7.01 *r = 0.9162, D = 83.94 %, y = 2045 (458) + 224 (30) x, 1 o, 0 e*
H7.02 *r = 0.8684, D = 75.42 %, y = 5.94 (1.12) + 0.00896 (0.00142)x, 1 o, 2 e*
H7.03 *r = 0.9121, D = 83.20 %, r_{12} = 0.9815, R_1 = 0.9043, R_2 = 0.9104, 0 o, 0 e*
H7.04 *r = 0.6871, D = 47.21 %, y = 1.21 (0.03) + 0.00376 (0.0008)x, 3 o, 3 e*

H7.05 *r = 0.5715, D = 32.66 %, r_{12} = 0.3080, R_{11} = -0.5350, R_{12} = -0.3558, 2 o, 0 e*

7.5.5 Analýza ekonomických a sociologických dat

S7.01 *r = 0.6907, D = 47.71 %, y =323.6 (147) + 132 (36) x*
S7.02 *r = 0.6264, D = 39.24 %, y = 20 (16) + 0.50 (0.25) x, 1 o, 1 e*
S7.03 *r = 0.5098, D = 25.99%, y = 20.07 (11.90, A) + 10.48 (3.03, Z), 2o, 0e,*
S7.04 *r = 0.8304, D = 68.95 %, y = 3.32 (0.58) + 0.333 (0.053 x, 2 o, 2 e*
S7.05 *r = 0.9272, D = 85.97 %, y = 5.0 (8.9) + 4.30 (0.48) x, 1 o,1 e*
S7.06 *r = -0.9429, D = 88.90 %, y = 99.45 (1.68) - 0.0106 (0.0015) x, 0 o, 1 e*
S7.07 *r = 0.6770, D = 45.84 %, y = 5.84 (0.66) + 0.023 (0.00) x, 1 o, 0 e*
S7.08 *r = 0.9405, D = 88.46 %, y = 0.584 (0.121) + 0.181 (0.023) x, 1 o, 1 e*
S7.09 *r = 0.5721, D = 32.73 %, y = 0.82 (0.84) + 0.0656 (0.0383) x, 1 o, 1 e*
S7.10 *(a) y na x_1: r = 0.8834, D = 78.04 %, y = 1.68 (0.39) + 0.065 (0.008) x_1, (b) y na x_2: r = -0.9626, D = 92.66 %, y = 11.07 (0.47) - 0.18 (0.01) x_2, (c) x_1 na x_2: r = -0.9321, D = 86.88 %, x_1 = 52.55 (1.63) - 0.37 (0.03) x_2*
S7.11 *r = 0.9747, D = 95.01 %, 0 o, 1 e, (a) y na x_1: r = -0.4504, D = 20.28 %, y = 6.16 (2.29) - 8.80 (0.06) x_1, (b) y na x_2: r = 0.9737, D = 94.81 %, y = -2.88 (0.48) + 0.16 (0.01) x_2, (c) x_1 na x_2: r = -0.5021, D = 25.21 %, x_1 = 59.6 (13.8) - 0.61 (0.37) x_2*
S7.12 *r = -0.3405, D = 11.60 %, y = 22.6 (5.0) - 0.115 (0.112) x, 0 o, 1 e*
S7.13 *r = 0.5990, D = 35.88 %, y = 2703 (485) - 109 (36) x, 2 o, 2 e*
S7.14 *r = 0.9298, D = 86.46 %, y = 1.62 (0.42) + 0.26 (0.01) ln x*
S7.15 *r = 0.9610, D = 92.35 %, y = 10.3 (1.4) + 4.9 (0.4) x, 0 o, 1 e*
S7.16 *r = 0.9293, D = 86.36 %, y = -41.8 (68.5) + 1.20 (0.16) x*
S7.17 *r = 0.9120, D = 83.17 %, y = 579 (167) + 0.54 (0.01) x, 1 o, 1 e*
S7.18 *r = -0.6472, D = 41.88 %, y = 40.7 (5.0) - 0.011 (0.004) x, 1 o, 2 e*
S7.19 *r = -0.1053, D = 1.10 %, y = 40.1 (30.0) - 0.214 (0.905) x, 1 o, 0 e*

7.6 Doporučená literatura

[1] Gubarev V. V.: *Algoritmy statističeskich izměrenij*. Energoatomizdat, Moskva 1986.
[2] Nimo-Smith I.: *Biometrika* **66**, 390 (1979).
[3] Kovalski Ch. J.: *Amer. Statist.* **27**, 103 (1973).
[4] Prescott P.: *Technometrics* **17**, 129 (1975).
[5] Mirskij G. J.: *Charakteristiki stochastičeskoj vzaimnosvjazi i ich izměrenije.* Energoizdat, Moskva 1982.
[6] Siotani M., Hyakawa T. a Fujikoshi Y.: *Modern Multivariate Statistical Analysis.* American Science Press, 1985.
[7] Kraemer H. Ch.: *J. Amer. Statist. Assoc.* **68**, 1004 (1973).
[8] Draper N. R., Smith H.: *Applied Regression Analysis*. Wiley New York, 1966.
[9] Mendenhall W., Sincich T.: *Statistics for the Engineering and Computer Sciences*. Dellen Publ. Comp., San Francisco 1988.
[10] Wonnacot T. H., Wonnacot R. J.: *Statistika pro obchod a hospodářství*. Victoria Publishing Praha 1993.

8

NELINEÁRNÍ REGRESNÍ MODELY

8.1 Tvorba nelineárního regresního modelu

Postup tvorby nelineárního regresního modelu se dá rozčlenit do těchto kroků:

1. Návrh regresního modelu. Obvykle se jako nelineární regresní model používá nějaká fyzikální nebo empirická závislost.

2. Odhadování parametrů. Na rozdíl od lineárních regresních modelů je třeba pro hledání minima kritéria regrese použít iterativních algoritmů. V naprosté většině případů se používá kritérium minima součtu čtverců odchylek (reziduí).

3. Posouzení kvality odhadů. Kvalita nalezených odhadů se standardně posuzuje podle jejich intervalů spolehlivosti nebo pouze jejich rozptylů $D(b_j)$. Příčinou vysokých rozptylů parametrů bývá také předčasné ukončení minimalizačního procesu před dosažením minima.

4. Grafické posouzení vhodnosti modelu. Zahrnuje řadu metod a charakteristik. Grafická analýza reziduí využívá grafu reziduí vs. predikce, ve kterém lze snadno odhalit:

a) odlehlé hodnoty,

b) trend v reziduích,

c) nedostatečné střídání znaménka u reziduí,

d) heteroskedasticitu.

K ověření normality rozdělení reziduí lze užít i rankitových grafů a vyčíslení koeficientu šikmosti $g_1(\hat{e})$ a špičatosti $g_2(\hat{e})$.

5. Základní statistické charakteristiky. O přiblížení navrženého modelu k experi-mentálním datům informuje hodnota sumy čtverců reziduí v minimu $U(\boldsymbol{b})$, ze které se vyčíslí *reziduální rozptyl* $\hat{\sigma}^2 = U(\boldsymbol{b})/(n - m)$. Jednoduchou charakteristikou, založenou na hodnotě $U(\boldsymbol{b})$, je *koeficient determinace D*, který je pro lineární regresní modely čtvercem vícenásobného korelačního koeficientu,

$$D = 1 - \frac{U(\boldsymbol{b})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \quad kde\ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i ,$$

Stonásobek koeficientu determinace se nazývá *regresní rabat*, 100 *D* [%]. V literatuře se často nesprávně užívá *Hamiltonův R-faktor*

$$R\&faktor \;=\; \sqrt{\frac{U(\boldsymbol{b})}{\sum_{i=1}^{n} y_i^2}} \;,$$

Pro $\bar{y} = 0$ platí, že R^2-*faktor* = 1 - *D*. Pro $\bar{y} \neq 0$ platí mezi *R-faktorem* a koeficientem determinace *D* vztah

$$R\&faktor \;=\; \sqrt{(1-D) - \frac{(1-D)\, n\, \bar{y}^2}{\sum_{i=1}^{n} y_i^2}} \;.$$

Hamiltonův R-faktor ukazuje na rozdíl mezi modelem $y = f(x, \boldsymbol{\beta})$ a modelem $y = 0$, což u modelů s absolutním členem nemá smysl a hodnoty *Hamiltonova R-faktoru* vycházejí v těchto případech *nesprávně nízké*. Je třeba upozornit, že *D* i *R-faktor* jsou funkcí počtu parametrů modelu, a to *D* je funkcí rostoucí s počtem parametrů, zatímco Hamiltonův R-*faktor* klesající. Ani *D*, ani *R-faktor* není proto vhodným rozlišovacím kritériem k porovnání modelů o různém počtu parametrů.

K rozlišení mezi modely je vhodnější užít *Akaikova informačního kritéria AIC,* pro které platí $AIC = -L(\boldsymbol{b}) + 2m$. Za optimální se považuje model, pro který dosahuje *AIC* minimální hodnoty. Při použití metody nejmenších čtverců a modelů nepatřících do téže třídy je

$$AIC \;=\; n \ln\left[\frac{U(\boldsymbol{b})}{n}\right] + 2m \;.$$

6. Regresní diagnostika. Obsahuje stejně jako u lineárních regresních modelů pomůcky a postupy analýzy regresního tripletu, tj. pro *kritiku dat*, *kritiku modelu* a *kritiku metody*. Analýzou vlivných bodů se identifikují body, které silně ovlivňují odhadované regresní parametry v modelu, což umožňuje určit vybočující pozorování nebo extrémy. Pro aditivní modely měření a užívanou metodu nejmenších čtverců jsou rezidua definována vztahem $\hat{e}_i = y_i - f(x_i, \mathbf{b})$. Popis je uveden v 6. kapitole.

A. Analýza klasických reziduí. Kritika dat se skládá z analýzy několika druhů grafických diagnostik a tabulek různých druhů reziduí. V řadě programů aplikované nelineární regrese je analýza reziduí hlavní diagnostickou pomůckou při rozlišení chemického modelu, a navíc těsnost dosaženého proložení experimentálními body je mírou vhodnosti navrženého modelu. Mezi nejčastěji užívané charakteristiky patří *směrodatná odchylka reziduí s(ê)*, která by se měla rovnat velikosti šumu závisle proměnné *y*, *koeficient šikmosti* $g_1(\hat{e})$ a *koeficient špičatosti* $g_2(\hat{e})$ reziduí.

K testování reziduí lze užít všech statistik, známých z lineárních regresních modelů. Potíže zde činí pouze určení rozdělení testačních statistik, které jsou závislé na nelinearitě modelu.

B. Analýza vlivných bodů. U lineárních regresních modelů (viz 6. kapitola) jsou k dispozici všechny charakteristiky k odhalení vlivných bodů pomocí reziduí $\hat{e}_i$ a dia-gonálních prvků P_{ii} projekční matice $\boldsymbol{P} = \boldsymbol{X}\,(\boldsymbol{X}^{\mathrm{T}}\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\mathrm{T}}$, zatímco u nelineárních modelů je rozdíl v matici $\boldsymbol{P}$. Matice $\boldsymbol{P} = \boldsymbol{J}\,(\boldsymbol{J}^{\mathrm{T}}\,\boldsymbol{J})^{-1}\,\boldsymbol{J}^{\mathrm{T}}$ totiž obsahuje $\boldsymbol{J}$ Jakobián čili derivaci modelové funkce podle jednotlivých parametrů v daných bodech.

U nelineárních regresních modelů je situace komplikována tím, že již nelze vyjádřit odhady parametrů a rezidua jako lineární kombinaci experimentálních dat. Pokud se užije linearizace nelineárního modelu, je možné užít přímo všech technik odhalení vlivných bodů v lineárních modelech. Vychází se z jednokrokové aproximace odhadu $\boldsymbol{b}_{(i)}$, pro kterou platí

$$\boldsymbol{b}_{(i)}^{1} \;=\; \boldsymbol{b} - \frac{(\boldsymbol{J}^{T}\,\boldsymbol{J})^{-1}\,\boldsymbol{J}_i\,\hat{e}_i}{1 - P_{ii}} \;,$$

kde P_{ii} jsou prvky projekční matice $\boldsymbol{P}$. Lze vyčíslit *charakteristiku* DFS_{ij}, která vyjadřuje vliv *i*-tého bodu na odhad *j*-tého parametru, vztahem

$$DFS_{ij} = \frac{b_j - b_{j(i)}^1}{\hat{s}_{(i)} \sqrt{V_{ii}}} ,$$

kde $s_{(i)}^2$ je odhad rozptylu vyčíslený při vynechání *i*-tého bodu, pro který platí

$$s_{(i)}^2 = \frac{U(\boldsymbol{b}) - \frac{\hat{e}_i^2}{1 - P_{ii}}}{n - m - 1} ,$$

Symbol V_{ii} značí prvky matice $\boldsymbol{V} = (\boldsymbol{J}^T \boldsymbol{J})^{-1}$. Při testování se považuje *i*-tý bod za vlivný, pokud je $DFS_{ij} > 2/\%n$.

Vlivné body lze také identifikovat na základě jednokrokové aproximace *Jackknife reziduí* $\hat{e}_{Ji}$, pro kterou platí vztah

$$\hat{e}_{Ji} = \frac{\hat{e}_i}{\hat{s}_{(i)} \sqrt{1 - P_{ii}}} ,$$

K vyjádření vlivu jednotlivých bodů na odhady parametrů lze použít i kvadratického rozvoje regresního modelu a vyčíslovat změny vektoru vychýlení $\boldsymbol{h}_{(i)}$ při vynechání
i-tého bodu nebo změny střední hodnoty *i*-tého rezidua při vynechání *i*-tého bodu. Mezi nelineární míry vlivu *i*-tého bodu na odhady parametrů patří *věrohodnostní vzdálenost*

$$LD_i = 2 [\ln L(\boldsymbol{b}) - \ln L(\boldsymbol{b}_{(i)})] .$$

Pro případ metody nejmenších čtverců bude věrohodnostní vzdálenost ve tvaru

$$LD_i = n \ln\left[\frac{U(\boldsymbol{b}_{(i)})}{U(\boldsymbol{b})}\right] .$$

Do obou vztahů lze dosadit buď odhady $\boldsymbol{b}_{(i)}$, určené regresí při vynechání *i*-tého bodu, nebo $\boldsymbol{b}^1_{(i)}$, určené z jednokrokové aproximace. Je-li $LD_i > \chi^2_{1-\alpha}(2)$, je daný bod silně vlivný. Obyčejně se volí $\alpha = 0.05$.

(a) Vlivné body ovlivňují nejenom odhady parametrů, ale také relativní vychýlení $\boldsymbol{h}_R$, které je značně citlivé na jejich výskyt.

(b) Charakteristiky založené na linearizaci nebo kvadratické aproximaci nelineárního modelu neindikují vždy správně přítomnost vlivných bodů. Hodí se především pro málo nelineární modely.

(c) Nejlepší indikaci vlivných bodů poskytuje věrohodnostní vzdálenost LD_i. Pouze tato charakteristika umožňuje indikaci celé skupiny vlivných bodů, kde může dojít k jejich vzájemnému "maskování".

(d) U praktických úloh postačuje aproximace LDS_i .

7. Mapa citlivostní funkce. Na rozdíl od lineárních regresních modelů je třeba u nelineárních modelů počítat s řadou komplikací, jako je neodhadnutelnost některých parametrů, existence minima funkce $U(\boldsymbol{\beta})$ jen pro některé regresní modely, výskyt lokálních minim a existence sedlových bodů, ovlivňujících kriteriální funkci $U(\boldsymbol{\beta})$ a špatnou podmíněnost parametrů v regresním modelu. Tyto problémy lze částečně indikovat na základě analýzy *normalizovaných citlivostních koeficientů*

$$C_{j(i)} = \beta_j \frac{\delta f(x_i, \boldsymbol{\beta})}{\delta \beta_j} \qquad \begin{matrix} j = 1, ..., m \\ i = 1, ..., n \end{matrix} .$$

Pro vizuální posouzení špatné podmíněnosti, vzniklé jako důsledek přibližné multikolinearity mezi parametry β_j, β_h , se konstruují *citlivostní grafy*. Obyčejně jde o závislosti $C_{j(i)}$ a $C_{h(i)}$ na x_i , $i = 1, ..., n$. Lze také vynášet závislost normalizovaných citlivostních koeficientů přímo na indexu i.

Pro vyjádření citlivosti regresních modelů na změnu parametru β_j je možné využít celkové *citlivostní funkce*

$$C_{cj} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\delta f(x_i, \boldsymbol{\beta})}{\delta \beta_j} \right]^2 .$$

Tato citlivostní funkce je nekonstantní pro takové parametry β_j, které jsou v modelu $f(x, \boldsymbol{\beta})$ nelineární.

Citlivostní grafy parametrů jsou pak závislosti C_{cj} na β_j v okolí bodů $\beta_j^{(0)}$ nebo b_j. Pokud jsou citlivostní grafy parametrů přibližně konstantní, indikuje to malou citlivost regresního modelu ke změnám j-tého parametru, nebo je model $f(x, \boldsymbol{\beta})$ vzhledem k parametru β_j *lineární*.

8. Predikční schopnost modelu. Predikční schopnost se může posoudit postupem "cross-validation": data se rozdělí na dvě podskupiny M_1 (s indexy $i = 1, ..., \text{int}(n/2)$) a M_2 (s indexy $i = \text{int}(n/2) + 1, ..., n$). Označí se odhady parametrů z bodů podskupiny M_1 jako $\boldsymbol{b}(M_1)$ a z bodů podskupiny M_2 jako $\boldsymbol{b}(M_2)$. Predikční schopnost modelu lze pak vyjádřit kritériem

$$K = \frac{U(\boldsymbol{b})}{\sum_{i \in M_1} [y_i - f(x_i, \boldsymbol{b}(M_2))]^2 + \sum_{i \in M_2} [y_i - f(x_i, \boldsymbol{b}(M_1))]^2} .$$

Predikční schopnost modelu je tím vyšší, čím víc se hodnota K blíží k jedné. Mezi další kritéria patří *střední kvadratická chyba predikce*

$$MEP = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, \boldsymbol{b}_{(i)}))^2 .$$

Místo odhadu $\boldsymbol{b}_{(i)}$ lze použít také jednokrokové aproximace $\boldsymbol{b}^1_{(i)}$. Čím je MEP nižší, tím je model věrohodnější a má lepší predikční schopnost.

9. Souhlas s požadavky fyzikálního smyslu. U navržených modelů jsou na odhady parametrů kladena omezení, vycházející z fyzikálního smyslu odpovídajících parametrů. Standardně se vyžaduje, aby odhady ležely v jisté předpokládané oblasti (např. koncentrace v oblasti kladných čísel, molární absorpční koeficienty ε v oboru čísel 10 až 10^6, konstanty stability $\log \beta_{pqr}$ v oboru čísel 0 až 50 atd.).

Program ADSTAT umožňuje numerickou a statistickou analýzu nelineárního regresního modelu $f(x, \boldsymbol{\beta})$ s využitím minimalizační hybridní strategie "double dog-leg". Vstupem je soubor experimentálních dat $\{x_i, y_i\}$, $i = 1, ..., n$, a nulté přiblížení odhadovaných parametrů $\boldsymbol{\beta}^{(0)}$. Uživatel zadává regresní model a může volit, zda se vybrané parametry zkonstantní.

Vzorová úloha 8.1 *Odhad tří parametrů rozšířeného Debyeova-Hückelova vztahu* Na vzorové **Úloze C8.08** ukážeme postup analýzy nelineárního regresního modelu se zadáním: Stanovte termodynamickou disociační konstantu pK^T_a (parametr β_1), efektivní průměr iontů *å* (parametr β_2) a vysolovací konstantu C (parametr β_3) závislosti smíšené disociační konstanty y na iontové síle x podle rozšířeného Debyeova-Hückelova vztahu[65] pro vybrané sulfonftaleiny. Mají-li oba ionty L^{Z-1} a HL^Z zhruba stejnou velikost *å* [10^{-10} m] a je-li celkový vysolovací koeficient $C = C^Z_{HL} - C^{Z-1}_L$, lze formulovat

Debyeův-Hückelův vztah tvarem

$$y \sim \beta_1 - \frac{(1 - 2Z)\, A \sqrt{x}}{(1 + B\, \beta_2 \sqrt{x})} + \beta_3\, x \quad ,$$

kde A = 0.5112 $mol^{-1/2}$. $l^{1/2}$. $K^{3/2}$, B = 0.3291 $mol^{-1/2}$. m^{-1}. $l^{1/2}$. $K^{1/2}$ jsou hodnoty konstant pro vodné roztoky a 25 EC. Předpokládejte aditivní model měření a normalitu chyb závisle proměnné y, zatímco nezávisle proměnná x je zatížena podstatně menší experimentální chybou.

Data: Bromkrezolová zeleň: Z = -1, *{x, y}*.

0.010	4.901	0.022	4.871	0.040	4.834	0.060	4.808	0.116	4.765	0.232	4.709
0.392	4.691	0.594	4.677	0.923	4.664	1.330	4.662	2.050	4.686	3.720	4.785

1. Návrh modelu: Označme parametr pK_a^T jako první parametr β_1 s jeho odhadem b_1, *å* jako druhý parametr β_2 s jeho odhadem b_2 a konečně C třetí parametr β_3 se odhadem b_3. Pro nulté přiblížení odhadovaných parametrů je voleno $b_1^{(0)}$ = 1.0, $b_2^{(0)}$ = 1.0, $b_3^{(0)}$ = 1.0.

2. Odhadování parametrů: Jsou nalezeny bodové a intervalové odhady b_j s polo-šířkami 95%ních intervalů spolehlivosti, vychýleními h_j a relativními vychýleními $h_{R,j}$.

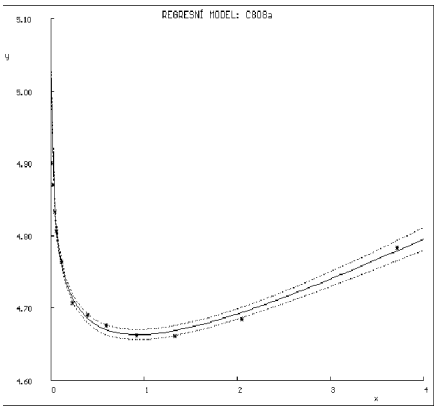
Bodové odhady parametrů:

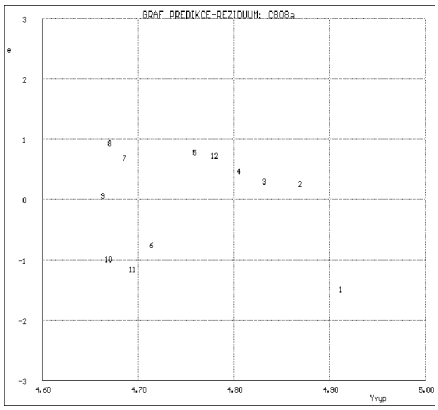
Parametr	**Bodový**	**Směrodatná**	**Absolutní**	**Relativní**
β_j	**odhad b_j**	**odchylka $s(b_j)$**	**vychýlení h_j**	**vychýlení $h_{R,j}$ [%]**
β_1	5.0336E+00	4.2468E-03	-1.9583E-05	-3.8905E-04
β_2	7.6010E+00	2.1432E-01	4.2364E-03	5.5735E-02
β_3	6.8337E-02	3.4380E-03	1.7995E-05	2.6332E-02

Intervalové odhady parametrů:

Parametr	**Bodový**	**Poloviční délka intervalu spolehlivosti spočtená z**	
β_j	**odhad b_j**	**délky poloos**	**maxim**
β_1	5.0336E+00	+-1.1915E-02	+-1.4456E-02
β_2	7.6010E+00	+-7.2957E-01	+-7.2957E-01
β_3	6.8337E-02	+-9.9546E-03	+-1.1703E-02

3. Graf regresní křivky: Danými experimentálními body a 95 % pásy spolehlivosti je zde zobrazen na obr. 8-1a. Graf analýzy klasických reziduí na obr. 8-1b ukazuje, že rezidua vytváří náhodný mrak, což potvrzuje vhodnost modelu.





Obr. 8-1 Vlevo: rozptylový graf, vpravo: analýza klasických reziduí, ***ADSTAT.***

4. Korelační matice parametrů: Ukazuje, že korelace mezi parametry je výrazná.

	b_1	b_2	b_3
$x[1, i]$	1.00000	-0.82415	0.52400
$x[2, i]$	-0.82415	1.00000	-0.85058
$x[3, i]$	0.52400	-0.85060	1.00000

5. Základní statistické charakteristiky: Jsou to hodnoty regresního rabatu 100 *D* [%], Akaikeho informačního kritéria *AIC* a střední kvadratické chyby predikce *MEP* k rozlišování mezi modely, obecné statistické momenty reziduí m_1 , m_2 , m_3 , a m_4 , odhady vnitřní křivosti Γ^T a vnější křivosti Γ^N (vysvětlení na str. 617 v cit.[100]). Bylo dosaženo výtečného proložení, regresní rabat 99.49% ukazuje, že vysoké procento bodů vyhovuje navrženému modelu Debyeovy-Hückelovy závislosti.

Regresní rabat *D* [%]	: 99.487
Akaikeho informační kriterium *AIC*	: -117.59
Střední kvadratická chyba predikce *MEP*	: 1.5516E-04
První statistický moment residuím m_1	: -1.9709E-07
Třetí statistický moment residuí m_3	: 1.1306E-07
Čtvrtý statistický moment residuí m_4	: 2.0362E-09
Parametr Γ^T	: 8.4501E-02
Parametr Γ^N	: 8.6108E+02

6. Regresní diagnostika: Obsahuje pomůcky pro kritiku dat, kritiku modelu a kritiku metody.

A. Analýza klasických reziduí: *Směrodatná odchylka reziduí* dosahuje hodnoty stejné velikosti, jako je odhad náhodných chyb (šumu) proměnné *y*, tj. $g(pK_{a,i}$. 0.01). Rozdělení reziduí je mírně asymetrické, sešikmené k nižším hodnotám, protože *odhad šikmosti* dosahuje záporné hodnoty. Rozdělení se blíží rovnoměrnému, protože *odhad špičatosti* je blízký hodnotě 1.80. S ohledem na malý počet dat nelze z výběrové šikmosti a špičatosti usuzovat na nenormalitu.

Bod	**Měřená hodnota**	**Predikovaná hodnota**	**Směrodatná odchylka**	**Vychýlení**	**Klasické reziduum**
i	y_i	$y_{vyp,\,i}$	$s(y_{vyp,\,i})$	h_i	e_i
1	4.9010	4.9115	3.7122E-03	-9.6193E-06	-1.0524E-02
2	4.8710	4.8691	3.3094E-03	-3.8063E-06	1.9192E-03
3	4.8340	4.8318	2.9072E-03	1.0885E-06	2.2400E-03
4	4.8080	4.8046	2.6222E-03	4.1321E-06	3.3797E-03
5	4.7650	4.7593	2.2938E-03	7.4433E-06	5.6887E-03
6	4.7090	4.7142	2.3949E-03	7.2323E-06	-5.2232E-03
7	4.6910	4.6860	2.7365E-03	4.1517E-06	5.0140E-03
8	4.6770	4.6703	2.9978E-03	3.8001E-07	6.7445E-03
9	4.6640	4.6635	3.1121E-03	-3.8612E-06	5.2314E-04
10	4.6620	4.6689	3.0511E-03	-6.3855E-06	-6.9445E-03
11	4.6860	4.6941	3.1750E-03	-6.2046E-06	-8.1348E-03
12	4.7850	4.7797	6.1717E-03	5.4489E-06	5.3175E-03
Reziduální součet čtverců *RSC*					: 4.0410E-04
Směrodatná odchylka reziduí *s(e)*					: 5.8030E-03
Odhad šikmosti g_1					: -0.579
Odhad špičatosti g_2					: 1.796
Hamiltonův *R-faktor* [%]					: 0.122

B. Tabulka vlivných bodů: Obsahuje základní charakteristiky k odhalení vlivných bodů: Jackknife rezidua $\hat{e}_{Ji}$,

Cookovy vzdálenosti D_i, diagonální prvky H_{ii} projekční matice, normalizovanou vzdálenost *FDA* a věrohodnostní vzdálenosti LDA_i.

Bod	Jackknife reziduum	Cookova vzdálenost	Diagonální prvky	Normalizovaná vzdálenost	Věrohodnostní vzdálenost
i	$e_{J,i}$	D_i	$H_{i,i}$	***FDA***	***LDA***
1	-2.2875E+00	5.2536E-01	3.0692E-01	1.2144E-01	2.6567E-01
2	3.1245E-01	1.1668E-02	2.4392E-01	3.8138E-03	2.8886E-03
3	3.5252E-01	1.0641E-02	1.8824E-01	4.6230E-03	7.9056E-03
4	5.2560E-01	1.8108E-02	1.5314E-01	9.3211E-03	8.6827E-03
5	8.9337E-01	3.6124E-02	1.1718E-01	2.1943E-02	1.4799E-02
6	-8.1923E-01	3.4006E-02	1.2774E-01	1.9626E-02	1.3011E-02
7	8.0345E-01	4.4839E-02	1.6679E-01	2.1813E-02	1.4450E-02
8	1.1447E+00	1.0565E-01	2.0015E-01	4.4082E-02	3.4816E-02
9	8.3151E-02	7.1250E-04	2.1571E-01	2.7009E-04	7.3859E-03
10	-1.1908E+00	1.1815E-01	2.0734E-01	4.5560E-02	3.7563E-02
11	-1.4634E+00	1.8341E-01	2.2452E-01	6.5850E-02	7.1201E-02
12	2.6179E+00	7.7430E+00	8.4834E-01	4.1027E-03	2.7646E-03

7. Mapa citlivostní funkce: Citlivostní funkce vyjadřuje změnu regresního modelu při změně parametru o ±5%. Hodnoty v tabulce ukazují, že parametr b_1 a b_3 jsou dobře podmíněny v modelu, jejich změna způsobí změnu účelové funkce 8 až 9 řádů. Druhý parametr b_2 je ve srovnání s předchozími parametry méně citlivý, hůře podmíněný v modelu, změna je podstatně menší.

Parametr	Relativní změna	Souhrnná citlivost	Relativní změna
j	C_{jR}**(-5%), [%]**	C_j**, [%]**	C_{jR} **(+5%), [%]**
1	8.0486E-09	1.0000	-7.2821E-09
2	1.5937E+01	1.0301E-03	-1.3244E+01
3	-3.3799E-07	1.7701	9.5080E-08

8. Predikční schopnost modelu: Pro $n = 12$ bude $M_1 = 1$ až 6, $M_2 = 7$ až 12, $RSC(M_1) = 2.8534E\text{-}05$, $RSC(M_2) = 5.3549E\text{-}05$, $U(\mathbf{b}) = 40.410E\text{-}05$, a proto $K = 4.923$. Jelikož se K neblíží k jedné, je predikční schopnost modelu slabší.

9. Souhlas s požadavky fyzikálního smyslu: První parametr představuje *termodyna-mickou disociační konstantu* $pK^T_a = 5.034$, která má fyzikální smysl a dále pak třetí parametr *vysolovací konstantu* $C = 0.068$, jež má rovněž fyzikální smysl. Druhý parametr představuje *efektivní průměr iontu å* $= 7.6 \times 10^{-10}$ m, což je v souhlase s hodnotami Kielandových tabulek.

8.2 Úlohy

Úlohy v počtu 57 jsou rozděleny do pěti kapitol: B8 (farmakologická a biochemická data), C8 (chemická a fyzikální data), E8 (environmentální, potravinářská a zemědělská data), H8 (hutní a mineralogická data) a S8 (ekonomická a sociologická data).

8.2.1 Analýza farmakologických a biochemických dat

Úloha B8.01 *Závislost hmotnosti očních čoček na stáří králíků*

Dudzinksi a Mykytowycz (1961) ukázali, že hmotnost vysušených očních čoček evropských králíků *Oryctolagus cuniculus* je užitečným indikátorem stáří králíka ve dnech. Nelineární regresní model je vyjádřen vztahem

$$y \text{ ' } \beta_1 \exp[\&\beta_2/(x \% \beta_3)]$$

a alternativní model transformovaný pro případ multiplikativních chyb má tvar

$$\ln y \text{ ' } \beta_1 \& \beta_2/(x \% \beta_3).$$

Vyšetřete regresní triplet a ověřte, který ze dvou navržených modelů lépe vyhovuje daným datům. K porovnání regresních modelů využijte střední kvadratickou chybu predikce *MEP* nebo Akaikeho informační kritérium *AIC*.

Doporučené nulté přiblížení:

Data: Stáří králíka *x* [dny], hmotnost vysušených očních čoček *y* [mg].

15.00	21.66	15.00	22.75	15.00	22.30	18.00	31.25	28.00	44.79
...	...	...	...	...	...	...	...	...	...
756.00	242.57	768.00	232.12	802.00	242.00	860.00	246.70		

Úloha B8.02 *Mozková činnost v závislosti na čase při intenzivní duševní činnosti*

Některé periody mozkové činnosti jsou ovlivňovány intenzitou duševní činnosti. Data CASY3 a CASY5 představují činnost jedince vystaveného od určitého okamžiku trvající duševní zátěži v časových jednotkách. Hodnoty jsou měřeny v čase, a to po 1 sekundě. Nalezněte především okamžik počátku zátěže. Ověřte model

$$y = a + b \exp(-c\,x)$$

a odhadněte maximální zátěž *a*. Pro odhad počátku zátěže použijeme předpoklad, že před zátěží je hodnota konstantní, $y = p$. Od začátku zátěže v $x = d$ začne klesat podle vztahu $y = a + b \exp(-c\,x)$ a platí, že $f_1(x) = y = p$ pro $x < d$ a $f_2(x) = y = a + b \exp(-c\,x)$ pro $x \ldots d$. Za předpokladu spojitosti je $f_1(d) = f_2(d)$ bude $p = y = a + b \exp(-c\,d)$.

Doporučené nulté přiblížení:

Užijte modul nelineární regrese systému ADSTAT 2.0 k odhadu *d* s modelem

(p[1] + p[2] * exp(x1 *p[3])) * (x1 > p[4]) + (p[1] + p[2] * exp(p[3] * p[4])) * (x1 < p[4]).

Podobně lze řešit i data CASY5. V datech existuje jedno evidentně vybočující měření (předposlední údaj).

Data: Duševní zátěž *y* měřena v čase po 1 sekundě (za *x* dosaďte krok 1 sekunda).

(a) Soubor CASY3:

395.870	392.606	392.143	393.150	390.346	389.847	395.871	392.858
...	...	...	...	...	...	...	...
333.354	332.850	331.846	331.116	329.084	328.080		

(b) Soubor CASY5:

595.450	595.790	593.740	593.720	592.860	582.480	567.020	573.060
...	...	...	...	...	...	...	...
458.530	458.040	455.150	456.050	453.220	455.970	452.720	490.390
457.310							

8.2.2 Analýza chemických a fyzikálních dat

Úloha C8.01 *Určení dvou parametrů Szyszkovského rovnice*

Závislost povrchového napětí $\Delta g = g^0 - g$ (závisle proměnná *y*) na koncentraci *x* organické látky ve vodném roztoku je popsána Szyszkovského rovnicí

$$y = \beta_1 \log (1 + \beta_2\, x),$$

kde g^0, *g* jsou povrchová napětí čisté vody a roztoku organické látky o molární koncentraci *x*. Parametr β_1 je stejný pro celou homologickou řadu a parametr β_2 zase pro členy se stejným počtem uhlíků. Určete parametry β_1 a β_2 pro vodný roztok při teplotě 18 EC a vyšetřete regresní triplet. Z intervalových odhadů odhadněte podmí-něnost parametrů v modelu. Adamcová (1989) uvádí pro kyselinu máselnou odhady $b_1 = 0.0298$ N. m^{-1} a $b_2 = 0.01964$ m^3. mol^{-1}.

Data: Koncentrace kyseliny máselné *x* [mol. m^{-3}], změna povrchového napětí *y* [N. m^{-1}].

$1.00 \cdot 10^{-3}$	$2.54 \cdot 10^{-7}$	$6.00 \cdot 10^{-3}$	$1.52 \cdot 10^{-6}$	$1.10 \cdot 10^{-2}$	$2.80 \cdot 10^{-6}$	$1.60 \cdot 10^{-2}$	$4.07 \cdot 10^{-6}$
...	...	...	...	...	...	...	...
$8.10 \cdot 10^{-2}$	$2.06 \cdot 10^{-5}$	$8.60 \cdot 10^{-2}$	$2.18 \cdot 10^{-5}$	$9.10 \cdot 10^{-2}$	$2.31 \cdot 10^{-5}$	$9.60 \cdot 10^{-2}$	$2.44 \cdot 10^{-5}$
$1.01 \cdot 10^{-1}$	$2.56 \cdot 10^{-5}$						

Úloha C8.02 *Parametry závislosti tenze par vody a dodekanu na teplotě*
Závislost tenze par vody y [kPa] a normálního dodekanu na teplotě x [EC] lze vystihnout rovnicí

$$\ln(y) = \beta_1 - \frac{\beta_2}{x + \beta_3} .$$

Odhadněte parametry β_1, β_2 a β_3 vody a normálního dodekanu. Jsou všechny tři parametry v modelu dobře podmíněny? Je nutné regresní model reparametrizovat? Normální dodekan je možné přehánět vodní parou. Určete, při jaké teplotě bude destilace probíhat? Adamcová (1989) uvádí odhady pro vodu b_1 =16.28861, b_2 = 3816.44, b_3 = 227.02 a pro dodekan b_1 = 14.09839, b_2 = 3774.56, b_3 = 181.83.
Data:
Voda: teplota x [EC], tenze par y [kPa].

20.00	2.31	25.00	3.14	30.00	4.22	35.00	5.60	40.00	7.36	45.00	9.57
...	...	...	...	...	...	...	...	...	...	...	...
80.00	47.37	85.00	57.81	90.00	70.11	95.00	84.53	100.00	101.32		

Normální dodekan: teplota x [EC], tenze par y [kPa].

20	0.010019	25	0.015746	30	0.024224	35	0.036533	40	0.054087	45	0.078702
...	...	...	...	...	...	...	...	...	...	...	...
80	0.727835	85	0.953567	90	1.236955	95	1.589551	100	2.024558		

Úloha C8.03 *Závislost molární tepelné kapacity plynné síry na teplotě*
Molární tepelná kapacita plynné síry y v intervalu teplot x [K] 717.7 až 2000 K je popsána rovnicí

$$y = \beta_1 + \beta_2 x + \frac{\beta_3}{x^2}.$$

Odhadněte parametry β_1 , β_2 a β_3. Jsou parametry dobře podmíněny v modelu? Jsou nalezené odhady parametrů dostatečně spolehlivé ?
Data: Teplota x [K], molární tepelná kapacita y [J. K^{-1}. mol^{-1}].

720	17.96733	770	18.03665	820	18.09886	870	18.15556	920	18.20790
...	...	...	...	...	...	...	...	...	...
1720	18.81533	1770	18.84769	1820	18.87980	1870	18.91168	1920	18.94336
1970	18.97486								

Úloha C8.04 *Parametry Antoineovy rovnice pro kyselinu sírovou a benzen*
Při rovnováze mezi fází plynnou a kondenzovanou (kapalina-pára) vystihuje závislost tlaku y na teplotě x [K] model Antoineovy rovnice

$$\log(y) = \beta_1 - \frac{\beta_2}{x + \beta_3} ,$$

kde β_1, β_2, β_3 jsou empirické parametry Antioneovy rovnice. Odhadněte tyto tři neznámé parametry β_1, β_2, β_3 pro kyselinu sírovou a benzen. Vyšetřete regresní triplet a diskutujte podmíněnost parametrů v modelu. Adamcová (1989) publikovala

odhady pro kyselinu sírovou $b_1 = 7.641563$, $b_2 = 3073.77$, $b_3 = 214.699$ a pro benzen $b_1 = 6.01907$, $b_2 = 204.682$, $b_3 = 220.078$. Jsou nalezené odhady spolehlivější?

Data: Teplota x [K], logaritmus tlaku, log y.

Kyselina sírová:

280	1.428148	285	1.490320	290	1.551260	295	1.611004	300	1.669587	305	1.727043
...	...	...	...	...	...	...	...	...	...	...	...
370	2.384550	375	2.429124	380	2.472948						

Benzen:

280	3.610082	285	3.633929	290	3.657310	295	3.680236	300	3.702721	305	3.724778
...	...	...	...	...	...	...	...	...	...	...	...
370	3.977506	375	3.994660	380	4.011528						

Úloha C8.05 *Závislost molární tepelné kapacity kyseliny dusičné na teplotě*

Závislost molární tepelné kapacity y [J. K^{-1}. mol^{-1}] na teplotě x [K] je dána vztahem

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \frac{\beta_4}{x^2},$$

Určete odhady parametrů β_1, β_2, β_3 a β_4 pro plynnou fázi kyseliny dusičné z přesných a z experimentálních, tzn. zašuměných, dat, když velikost šumu, čili náhodné chyby, je přibližně 0.001. Jaká je podmíněnost parametrů v modelu ? Posuďte také míru spolehlivosti nalezených odhadů parametrů. Adamcová (1989) publikovala pro přesná data odhady parametrů $b_1 = 91.826$, $b_2 = 0.00627$, $b_3 = 1.76110$, $b_4 = -9480500$. Jsou nalezené odhady spolehlivější?

Data: Teplota x [K], molární tepelná kapacita y [J. K^{-1}. mol^{-1}].

1. část: simulovaná data bez šumu:

330	8.755896	360	23.21345	390	34.61908	420	43.82147
...	...	...	...	...	...	...	...
1170	116.3426	1200	118.1247				

2. část: data s náhodnou chybou okolo 0.001:

330	8.755874	360	23.21373	390	34.61816	420	43.82199
...	...	...	...	...	...	...	...
1170	116.3417	1200	118.1249				

Úloha C8.06 *Parametry teplotní závislosti Ostwaldova absorpčního koeficientu*

Pro teplotní závislost Ostwaldova absorpčního koeficientu y na teplotě x [K] se pro systém radon-voda uvádí v literatuře vztah

$$y = \exp(\beta_0 + \frac{\beta_1}{x} + \beta_2 \ln x).$$

Stanovte odhady parametrů β_0 , β_1 a β_2 z dat uvedené teplotní závislosti. Z velikosti poslední vrstevnice sumy čtverců reziduí U odhadněte podmíněnost parametrů v modelu a odhadněte také jejich míry přesnosti. Jsou nalezené odhady spolehlivější?

Data: Teplota x [K], Ostwaldův absorpční koeficient y.

275	0.4960	280	0.4080	285	0.3411	290	0.2901	295	0.2506
...	...	...	...	...	...	...	...	...	...
375	0.1101								

Úloha C8.07 *Parametry teplotní závislosti rozpustnosti sádrovce*

Rozpustnost sádrovce $CaSO_4 \cdot 2H_2O$ v závislosti na teplotě x [K] je vyjádřena vztahem

$$y = \exp(\beta_1 + \frac{\beta_2}{x} + \beta_3 \ln x),$$

kde y je molární zlomek sádrovce ve vodném roztoku. Určete odhady parametrů β_1, β_2 a β_3 z dat uvedené teplotní závislosti a vyšetřete regresní triplet. Jsou odhady parametrů dostatečně věrohodné a zdůvodněte proč?

Data: Teplota x [K], molární zlomek y.

275	$2.380488 \cdot 10^{-4}$	280	$2.486899 \cdot 10^{-4}$	285	$2.578060 \cdot 10^{-4}$	290	$2.653324 \cdot 10^{-4}$
...	...	...	...	...	...	...	...
365	$2.208487 \cdot 10^{-4}$	370	$2.119828 \cdot 10^{-4}$	375	$2.029707 \cdot 10^{-4}$	380	$1.938900 \cdot 10^{-4}$

Úloha C8.08 *Odhad tří parametrů rozšířeného Debyeova-Hückelova vztahu*

Stanovte termodynamickou disociační konstantu pK^{T}_{a} (parametr β_1), efektivní průměr iontů $å$ (parametr β_2) a vysolovací konstantu C (parametr β_3) závislosti smíšené disociační konstanty y na iontové síle x podle rozšířeného Debyeova-Hückelova vztahu[65] pro vybrané sulfonftaleiny. Mají-li oba ionty L^{Z-1} a HL^{Z} zhruba stejnou velikost $å$ [10^{-10} m] a je-li celkový vysolovací koeficient $C = C^{Z}_{HL} - C^{Z-1}_{L}$, lze formulovat Debyeův-Hückelův vztah tvarem

$$y = \beta_1 - \frac{(1 - 2Z)\, A \sqrt{x}}{(1 + B\, \beta_2 \sqrt{x})} + \beta_3\, x \quad ,$$

kde A = 0.5112 $mol^{-1/2}$. $l^{1/2}$. $K^{3/2}$, B = 0.3291 $mol^{-1/2}$. m^{-1}. $l^{1/2}$. $K^{1/2}$ jsou hodnoty konstant pro vodné roztoky a 25EC. Předpokládejte aditivní model měření a normalitu chyb závisle proměnné y, zatímco nezávisle proměnná x je zatížena podstatně menší experimentální chybou. Jsou všechny tři parametry stejně podmíněny (tj. stejně citlivé) v modelu? Která kritéria spolehlivosti odhadnutých parametrů využijete?

Data:

(a) Bromkrezolová zeleň: Z = -1, *{x, y}*.

0.010	4.901	0.022	4.871	0.040	4.834	0.060	4.808	0.116	4.765	0.232	4.709
0.392	4.691	0.594	4.677	0.923	4.664	1.330	4.662	2.050	4.686	3.720	4.785

(b) Bromfenolová červeň: Z = -1, {x, y}.:

0.010	6.017	0.022	5.970	0.040	5.935	0.060	5.908	0.116	5.872	0.200	5.841
0.392	5.797	0.594	5.775	1.004	5.760	1.445	5.765	2.260	5.788	4.000	5.865

(c) Bromkrezolový purpur: Z = -1, {x, y}.

0.010	6.085	0.022	6.029	0.040	6.009	0.060	5.986	0.116	5.941	0.232	5.901
0.392	5.861	0.594	5.871	0.923	5.856	1.330	5.863	2.000	5.894	3.720	5.947

(d) Bromthymolová modř: Z = -1, *{x, y}*.

0.0025	7.413	0.011	7.075	0.024	7.029	0.045	6.979	0.067	6.954	0.127	6.924	0.251	6.878
0.423	6.844	0.635	6.838	1.004	6.822	1.445	6.826	2.260	6.843	4.000	6.915		

(e) Fenolová červeň: Z = -1, *{x, y}*.

0.039	7.801	0.061	7.758	0.128	7.677	0.261	7.575	0.353	7.538
0.635	7.465	0.780	7.455	1.420	7.450	2.250	7.465	3.610	7.520

(f) Thymolová modř: Z = -1, *{x, y}*:

0.022	9.065	0.060	9.008	0.116	8.965	0.392	8.908	0.594	8.892
0.923	8.875	1.330	8.863	2.050	8.885	3.750	8.985		

Úloha C8.09 *Disociační konstanty a molární absorpční koeficienty částic kyseliny HL*

Stanovte disociační konstantu pK_{a1} (parametr β_1) a dva molární absorpční koeficienty ε_{L^-} (parametr β_2) a ε_{HL} (parametr β_3) disociující kyseliny HL ⇄ H^+ + L^- regresní analýzou *A*-pH křivky[65,91] u vybraných indikátorů. V závislosti na pH (nezávisle proměnná *x*) vodný roztok indikátoru obsahuje postupně dvě různě protonové částice, L^-, a HL se smíšenou disociační konstantou, $K_{a1} = a_{H+}\,[L^-]/[HL]$. K měření pH byla užita skleněná elektroda G202B a SKE o směrnici článku 59.16 mV/pH při teplotě 25 °C v roztoku iontové síly 0.001 mol/l. Absorbance *y* byla změřena v kyvetě délky *d* [cm] na spektrofotometru SPEKOL 21 s chybou $s_{inst}(y) = 0.002$. Je předpokládán aditivní model měření a normalita chyb. Pokud obě částice absorbují záření dané vlnové délky, bude absorbance *y* v kyvetě délky *d* [cm] vyjádřena vztahem

$$y = d\,c\,\frac{\beta_2 + \beta_3\,10^{\beta_1 - x}}{1 + 10^{\beta_1 - x}} \quad ,$$

kde *c* je analytická koncentrace disociující kyseliny. Nulté přiblížení parametrů je uvedeno u každého indikátoru. Vyšetřete podmíněnost parametrů v modelu. Komentujte i spolehlivost odhadů parametrů z těsnosti proložení.

Data: $c = 1$ mol. dm^{-3}, $d = 1.000$ cm,

(a) Bromkrezolová zeleň: $\beta_1^{(0)} = 5.0$, $\beta_2^{(0)} = 0.1$, $\beta_3^{(0)} = 0.7$, $T = 298$ K, *{x, y}:*

7.370	0.0787	6.645	0.083	5.929	0.1051	5.542	0.1428	5.300	0.1820	5.150	0.2130
...	...	...	...	...	...	...	...	...	...	...	...
3.668	0.6450	3.428	0.6730	3.088	0.6933	2.633	0.7070				

(b) p-Nitroanilin: $\beta_1^{(0)} = 1.0$, $\beta_2^{(0)} = 0.1$, $\beta_3^{(0)} = 0.7$, $T = 293$ K, *{x, y}:*

1.67	0.233	1.47	0.276	1.26	0.335	1.06	0.400	0.87	0.462	0.67	0.552	0.47	0.570		

(c) Akridin: $\beta_1^{(0)} = 6.0$, $\beta_2^{(0)} = 0.02$, $\beta_3^{(0)} = 0.6$, $T = 293$ K, *{x, y}*:

6.30	0.125	6.10	0.170	5.89	0.235	5.68	0.299	5.47	0.367	5.27	0.429	5.08	0.474	4.85	0.523

(d) 8-Hydroxychinolin: $\beta_1^{(0)} = 10.0$, $\beta_2^{(0)} = 0.5$, $\beta_3^{(0)} = 0.05$, $T = 293$ K, *{x, y}:*

9.12	0.123	9.32	0.167	9.52	0.216	9.65	0.243	9.89	0.310	10.12	0.370	10.28	0.415	10.53	0.465

(e) Methylová oranž: $\beta_1^{(0)} = 3.0$, $\beta_2^{(0)} = 0.2$, $\beta_3^{(0)} = 0.8$, $T = 298$ K, *{x, y}*:

6.237	0.227	5.447	0.232	4.874	0.245	4.445	0.272	4.266	0.289	4.054	0.328	3.909	0.357	3.801	0.383
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2.890	0.680	2.803	0.703	2.711	0.727	2.594	0.750	2.363	0.786	2.062	0.817	1.730	0.835		

Úloha C8.10 *Disociační konstanty a molární absorpční koeficienty částic kyseliny H_2L*

Stanovte disociační konstanty pK_{a1}, pK_{a2} a tři molární absorpční koeficienty ε_L, ε_{HL} a ε_{H_2L} disociující kyseliny H_2L ⇄ H^+ + HL^- ⇄ H^+ + L^{-2} regresní analýzou *A*-pH křivky[65,91] u vybraných indikátorů. Byla užita skleněná elektroda G202B a SKE o směrnici článku 59.16 mV/pH při teplotě 25 °C. Je předpokládán aditivní model měření a normalita chyb. V závislosti na pH (nezávisle proměnná *x*) obsahuje v roztoku tři částice, L^{2-}, HL^-, H_2L, které jsou charakterizovány smíšenými disociačními konstantami, $K_{a1} = a_{H+}\,[L^{2-}]/[HL^-]$, $K_{a2} = a_{H+}\,[HL^-]/[H_2L]$. Pokud všechny rozličně protonované částice absorbují záření dané vlnové délky, bude absorbance *y* v kyvetě délky *d* [cm] vyjádřena vztahem

$$y = d\,c\,\frac{\varepsilon_L + \sum_{j=1}^{2} \varepsilon_{H_jL}\, 10^{\sum_{i=1}^{j} pK_{ai}\cdot j\,pH}}{1 + \sum_{j=1}^{2} 10^{\sum_{i=1}^{j} pK_{ai}\cdot j\,pH}} + d\,c\,\frac{\beta_3 + \beta_4\, 10^{\beta_1 \cdot x} + \beta_5\, 10^{\beta_1+\beta_2 \cdot 2x}}{1 + 10^{\beta_1 \cdot x} + 10^{\beta_1+\beta_2 \cdot 2x}},$$

kde c je analytická koncentrace indikátoru a pět stanovovaných parametrů je pK_{a1} (parametr β_1), pK_{a2} (parametr β_2), ε_L (parametr β_3), ε_{HL} (parametr β_4) a ε_{H2L} (parametr β_5). Nulté přiblížení parametrů $b_j^{(0)}$ je uvedeno u každé studované kyseliny. Vyšetřete podmíněnost parametrů v modelu. Které z kritérií spolehlivosti odhadů parametrů vypovídá nejlépe?

Data: c = 1 mol. dm^{-3}, d = 1.000 cm, T = 293 K,

(a) Pyrokatechinová violeť: $b_1^{(0)} = 8$, $b_1^{(0)} = 10$, $b_3^{(0)} = 0.1$, $b_4^{(0)} = 1.0$, $b_5^{(0)} = 0.1$, $\{x, y\}$:

4.133	0.008	5.242	0.019	6.465	0.101	6.783	0.169	7.009	0.246	7.194	0.326
...	....	...	...	...	...	...	...	...	...	...	...
10.112	0.606	10.327	0.535	10.584	0.472	10.882	0.425	11.324	0.389		

(b) SNAZOXS: $b_1^{(0)} = 7$, $b_2^{(0)} = 10$, $b_3^{(0)} = 0.1$, $b_4^{(0)} = 1$, $b_5^{(0)} = 1$, $\{x, y\}$:

11.223	0.651	10.515	0.651	8.858	0.649	8.390	0.644	8.045	0.638	7.924	0.633
...	...	...	...	...	...	...	...	...	...	...	...
1.838	0.294	1.708	0.290	1.617	0.287	1.553	0.286				

(c) Naftylazoxin 6S: $b_1^{(0)} = 7$, $b_2^{(0)} = 3$, $b_3^{(0)} = 0.1$, $b_4^{(0)} = 1$, $b_5^{(0)} = 1$, $\{x, y\}$:

10.380	0.630	9.750	0.637	8.630	0.632	8.080	0.618	7.760	0.595	7.520	0.574
...	...	...	...	...	...	...	...	...	...	...	...
1.840	0.294										

Úloha C8.11 *Odhad tří disociačních konstant sulfoazoxinů analýzou A-pH křivky*

Stanovte tři disociační konstanty, pK_{a1} (parametr β_1), pK_{a2} (parametr β_2), pK_{a3} (parametr β_3), a čtyři molární absorpční koeficienty, ε_L (parametr β_4), ε_{HL} (parametr β_5), ε_{H_2L} (parametr β_6) a ε_{H_3L} (parametr β_7), postupně disociující kyseliny H_3L regresní analýzou *A-pH* křivky[65,91] tří sulfoazoxinů 2-, 3- a 4-CAPAZOXS. V závislosti na *pH* (nezávisle proměnná x) obsahuje vodný roztok CAPAZOXSu postupně čtyři různě protonové částice, L^{3-}, HL^{2-}, H_2L^- a H_3L, které jsou charakterizovány smíšenými disociačními konstantami, $K_{a1} = a_{H^+}\,[L^{3-}]/[HL^{2-}]$, $K_{a2} = a_{H^+}\,[HL^{2-}]/[H_2L^-]$, $K_{a3} = a_{H^+}\,[H_2L^-]/[H_3L]$. Pokud všechny čtyři částice absorbují záření dané vlnové délky, bude absorbance y v kyvetě délky d [cm] vyjádřena vztahem

$$y = d\,c\,\frac{\varepsilon_L + \sum_{j=1}^{3} \varepsilon_{H_jL}\, 10^{\sum_{i=1}^{j} pK_{ai}\cdot j\,pH}}{1 + \sum_{j=1}^{3} 10^{\sum_{i=1}^{j} pK_{ai}\cdot j\,pH}},$$

$$y = d\,c\,\frac{\beta_4 + \beta_5\, 10^{\beta_1 \cdot x} + \beta_6\, 10^{\beta_1+\beta_2 \cdot 2x} + \beta_7\, 10^{\beta_1+\beta_2+\beta_3 \cdot 3x}}{1 + 10^{\beta_1 \cdot x} + 10^{\beta_1+\beta_2 \cdot 2x} + 10^{\beta_1+\beta_2+\beta_3 \cdot 3x}},$$

kde c je analytická koncentrace 3-CAPAZOXSu. Nulté přiblížení $b_j^{(0)}$ parametrů β_j je uvedeno u každého sulfoazoxinu.

Data: c = 1 mol dm^{-3}, d = 1.000 cm, T = 293 K,

(a) 2-CAPAZOXS: $b_1^{(0)} = 9$, $b_2^{(0)} = 5$, $b_3^{(0)} = 3.5$, $b_4^{(0)} = 0.05$, $b_5^{(0)} = 0.3$, $b_6^{(0)} = 0.4$, $b_7^{(0)} = 0.9$, $\{x, y\}$:

2.892	0.687	2.998	0.678	3.065	0.675	3.150	0.652
...	...	...	...	...	...	...	...

	10.101		0.104										
(b) 3-CAPAZOXS: $b_1^{(0)} = 9$, $b_2^{(0)} = 5$, $b_3^{(0)} = 3.5$, $b_4^{(0)} = 0.05$, $b_5^{(0)} = 0.3$, $b_6^{(0)} = 0.4$, $b_7^{(0)} = 0.9$, {*x*, *y*}:													
1.565	0.660	1.750	0.666	1.817	0.653	2.000	0.640	2.058	0.631	2.224	0.593	2.500	0.547
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9.855	0.140												
(c) 4-CAPAZOXS: $b_1^{(0)} = 9$, $b_2^{(0)} = 5$, $b_3^{(0)} = 3.5$, $b_4^{(0)} = 0.05$, $b_5^{(0)} = 0.3$, $b_6^{(0)} = 0.4$, $b_7^{(0)} = 0.9$, {*x*, *y*}:													
1.633	0.224	1.734	0.225	1.872	0.232	1.986	0.235	2.181	0.244	2.385	0.260	2.474	0.268
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7.609	0.408	7.777	0.414	7.910	0.422	8.145	0.430	8.640	0.440	8.968	0.442		

Úloha C8.12 *Odhad bodu ekvivalence regresí lineárních větví titrační křivky*

Nalezněte intervalový odhad bodu ekvivalence analýzou dvou přímkových úseků instrumentální (např. fotometrické) titrace. Rovnice přímkových úseků jsou

$$a + b\,V \quad \text{pro} \quad V < V_0$$

a

$$c + d\,V \quad \text{pro} \quad V \ldots V_0.$$

Dále platí podmínka spojitosti $a + b\,V_0 = c + d\,V_0$, kde V_0 je bod ekvivalence. Odtud se dosadí do $a = c + (d - b)V_0$ a model obou větví titrační křivky bude

$$I = c + (d - b)V_0 + b\,V \quad \text{pro} \quad V < V_0$$

a

$$I = c + d\,V \quad \text{pro} \quad V \ldots V_0.$$

Tento model je nelineární vzhledem k parametrům a bod ekvivalence v něm představuje rovněž neznámý parametr. K jeho odhadu užijeme nelineární regresi.

Data: Regresní model přepíšeme ve tvaru s dvěma větvemi titrační křivky, představujícími závisle proměnnou: $y = \beta_3 + (\beta_4 - \beta_2).\beta_5 + \beta_2\,x$ pro $x < \beta_5$ a $y - \beta_3 + \beta_4\,x$ pro $x > \beta_5$, kde neznámé parametry β_1 (parametr *a*), β_2 (parametr *b*), β_3 (parametr *c*), β_4 (parametr *d*) a β_5 (bod ekvivalence V_0) a přídavek titračního činidla *V* (nezávisle proměnná *x*). V ADSTATu bude regresní model zapsán ve tvaru:

(p3 + (p4 - p2) * p5 + p2 * x) * (x < p5) + ((p3 + p4 * x) * (x $ p5))

a počáteční odhady parametrů se zadají následovně: p3: 0, p4: 1, p2: 0.5, p5: 8.0. Při výpočtu se logické výrazy *pravdivé* vyhodnotí jako 1 a *nepravdivé* jako 0. Přídavek titračního činidla *x* [ml], *y* [dílky]:

1	4.67	2	5.02	3	5.46	4	6.04	5	6.36	6	6.82	7	7.64	8	7.97	9	9.10	10	9.91
11	10.98	12	12.00	13	12.77	14	13.93	15	15.08	16	16.1	17	16.98	18	17.9				

Úloha C8.13 *Odhad bodu ekvivalence regresí lineárních větví titrační křivky*

Nalezněte intervalový odhad bodu ekvivalence analýzou dvou přímkových úseků instrumentální, např. fotometrické titrace. Model je nelineární vzhledem k parametrům a bod ekvivalence je v něm jako neznámý parametr. K jeho odhadu proto užijete nelineární regresi. Využijte zadání úlohy **C8.12**.

Data: (a) Přídavek titračního činidla *x* [ml], *y* [dílky].

1	90.8	2	80.1	3	70.18	4	60.21	5	50.29	6	40.29	7	30.37	8	30.31	9	40.2		
10	47.7	11	54.9	12	61.9	13	62.9	14	76.2	15	83.9	16	91	17	98.5.				

(b) Přídavek titračního činidla *x* [ml], *y* [dílky].

1.012	44	1.985	87	3.822	166	5.528	244	7.117	316	8.599	379								
...	...	...	...	...	...	...	...	...	...	...	...								
16.257	698	16.735	717	17.199	734														

Úloha C8.14 *Odhad bodu ekvivalence regresí lineárních větví titrační křivky*

Nalezněte intervalový odhad bodu ekvivalence analýzou dvou přímkových úseků instrumentální, např. konduktometrické titrace. Model je nelineární vzhledem k para-metrům a bod ekvivalence je v něm jako neznámý parametr. K jeho odhadu proto užijete nelineární regresi. Využijte zadání úlohy **C8.12**.

Data: Přídavek titračního činidla x [ml], y [dílky].

0	6	1	5	2	6.5	3	9	4	12	5	15	6	17.5	7	21	8	24	9	30
10	38	11	46.5	12	55	13	63	14	71	15	79	16	87	17	95				

Úloha C8.15 *Hledání regresního modelu popisu závislosti výtěžku reakce na čase*

V chemické laboratoři se sleduje výtěžek chemické reakce na čase. Využitím modelů Úlohy **E8.06** a na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům, tj. spotřebě oxidu dusičného y [%] v závislosti na čase x [s]. Využijte kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*. Vyšetřete také podmíněnost parametrů v modelu. Do ADSTATu zadáme model ve tvaru

p[1]*[1+p[4]*exp(-p[3] * (x-p[2]))]^(-1/p[4]) pro p[4] = -1.

Data: Čas x [s], spotřeba oxidu dusičného y [%].

0.5	5.9	1.0	11.7	1.5	15.8	2.0	18.6	2.5	20.6	3.0	22.6	3.5	23.9
4.0	25.1	4.5	26.4	5.0	27.2	5.5	28.2	6.0	29.1	6.5	29.6	7.0	30.1

Úloha C8.16 *Hledání reakčního modelu katalytické dehydratace n-hexylalkoholu*

V chemické laboratoři se často sleduje rychlost chemické reakce y v závislosti na koncentraci reakčních komponent, např. na parciálním tlaku alkoholu x_1 a olefinu x_2. Byl navržen regresní model ve tvaru

$$y = \frac{\beta_1 \beta_3 x_1}{1 + \beta_1 x_1 + \beta_2 x_2},$$

který především ověřte. Vyšetřete regresní triplet a odhadněte neznámé parametry modelu. Komentujte kritéria spolehlivosti odhadů parametrů.

Data: Parciální tlak alkoholu x_1 a olefinu x_2, rychlost chemické reakce y [s^{-1}].

1.0	1.0	0.126	2.0	1.0	0.219	1.0	2.0	0.076	2.0	2.0	0.126	0.10	0.0	0.186

Úloha C8.17 *Odhady parametrů závislosti tlaku nasycené páry na teplotě*

Závislost mezi tlakem y a teplotou nasycené páry x je vyjádřena vztahem

$$y = \beta_1 \, 10^{\frac{\beta_2 x}{\beta_3 + x}},$$

kde β_1, β_2 a β_3 jsou neznámé parametry. Vyšetřete regresní triplet a stanovte 95 % intervalové odhady neznámých parametrů. Z šíře intervalových odhadů odhadněte podmíněnost parametrů v modelu a míru přesnosti odhadů parametrů.

Data: Teplota nasycené páry x [EC], tlak nasycené páry y [atm].

0.0	4.14	10.0	8.52	20.0	16.31	30.0	32.18	40.0	64.62
...	...	...	...	...	...	...	...	...	...
90.0	522.7	95.0	674.3	100.0	782.0	105.0	920.0		

Úloha C8.18 *Odhady parametrů absorbce oxidu dusičného v roztoku*

Oxid dusičný je absorbován v roztoku a chemickou reakcí s výchozí látkou vzniká reakční produkt. Vztah mezi koncentrací absorbovaného oxidu dusičného y a koncentrací výchozí látky x je $y = \beta_1 \exp[\beta_2 x] x^{\beta_3}$, kde β_1, β_2 a β_3 jsou neznámé parametry. Vyšetřete regresní triplet a stanovte 95 % intervalové odhady neznámých parametrů. Jsou všechny parametry stejně podmíněny v modelu? Jsou parametry v modelu dostatečně citlivé?

Data: Koncentrace výchozí látka x [g/l], koncentrace absorbovaného oxidu dusičného y [g/l].

0.09	15.10	0.25	43.30	0.32	57.30	0.44	71.60	0.50	79.70	0.69	103.3
...	...	...	...	...	...	...	...	...	...	...	...
3.39	178.7	3.63	172.3	3.77	167.5						

Úloha C8.19 *Odhady parametrů závislosti dynamické viskozity na hustotě a teplotě*
Hustota a dynamická viskozita anhydridu hydrazinu byly měřeny v rozsahu teplot od 288.16 K do 449.83 K. Vzorky čistoty 99.6 % byly připraveny z komerčního hydrazinu mícháním s oxidem barnatým, vakuovou destilací a dvojím přelitím přes molekulové síto. Obsah hydrazinu byl stanoven jodometricky. Hustota byla měřena v křemenném pyknometru. Vztah mezi dynamickou viskozitou y měřenou v centipoise, hustotou x_1 v g/cm^3 a teplotou x_2 v Kelvinech

$$y \sim x_1^{\beta_1} \exp\left(\beta_2 + \frac{\beta_3}{x_2} + \frac{\beta_4}{x_2^3}\right)$$

obsahuje čtyři neznámé parametry β_1, β_2, β_3 a β_4. Vyšetřete regresní triplet a stanovte 95 % intervalové odhady neznámých parametrů, vyšetřete jejich podmíněnost v modelu. Jsou všechny parametry v modelu stejně citlivé?
Data: Hustota x_1 [g/cm^3], teplota x_2 [K], absolutní viskozita y [centipoise].

1.0114	288.16	1.0275	0.9934	310.94	0.7268	0.9672	338.72	0.5363
...	...	...	...	...	...	...	...	...
0.8575	449.83	0.2344						

Úloha C8.20 *Obsah n-alkanů (a iso-alkanů) v parafinu na počtu uhlíkových atomů*
Ve vzorku parafinu byly chromatograficky stanoveny relativní obsahy *n*-alkanů y_1 a *iso*-alkanů y_2 v závislosti na počtu uhlíkových atomů v řetězci, a to v rozmezí 23 až 44 uhlíkových atomů. Pro stanovení byla použita methylsilikonová nepolární kapilární kolona 10 m/0.25 mm a FID detektor v teplotním režimu 240 - 340EC, při teplotě detektoru 300EC a teplotě dávkovače 280EC. Jako nosného plynu bylo užito dusíku o průtoku 0.7 ml/min s dělicím faktorem 1:10. Rozhodněte, který z následujících dvou regresních modelů lépe popisuje uvedenou závislost

1. model: $y_1 \sim \beta_1 \exp\{\beta_2 x^2\} + \beta_3 \exp\{\beta_4 x\}$,

2. model: $y_2 \sim \beta_1 \exp\{\beta_2 x^4\} + \beta_3 \exp\{\beta_4 x^2\}$,

Data: Počet atomů uhlíku v řetězci x snížený o 22, obsah *n*-alkanů y_1 (resp. *iso*-alkanů y_2) [%].

Počet atomů uhlíku - 22, x	Obsah *n*-alkanů y_1 [%]	Obsah *iso*-alkanů y_2 [%]	Počet atomů uhlíku - 22, x	Obsah *n*-alkanů y_1 [%]	Obsah *iso*-alkanů y_2 [%]
1	0.03	0.05	12	3.37	4.23
...	...	...	...	...	...
11	3.99	4.55	22	0.06	0.12

8.2.3 Analýza environmetálních, potravinářských a zemědělských dat

Úloha E8.01 *Úroda ovoce v závislosti na stáří ovocného stromu*
Při zkoumání závislosti úrody stromů y určité odrůdy na věku stromů x od jejich přesazení zjistil ovocnářský ústav o náhodně vybraných stromech následující údaje. Odhadněte parametry regresního modelu, popisujícího vztah mezi úrodou y a věkem stromů po přesazení x za předpokladu, že regresní model má tvar $y \sim \beta_1 \beta_2^x$. Vyšetřete regresní triplet a posuďte

věrohodnost nalezených odhadů dle kritérií těsnosti proložení.

Data: Věk stromů *x* [roky], úroda *y* [kg].

2	2	2	3	3	2	5	10	4	5	3	3
2	1	3	4	4	7	4	6	3	3	5	12

Úloha E8.02 *Model výnosu úrody cibule typu BIS na dvou lokalitách*

Pro obecný vztah mezi výnosem pěstované plodiny *y* a hustotou sadby zeleniny *x* byly navrženy následující empirické nelineární regresní modely

Model A (Bleasdale a Nelder, 1960): $y = (\beta_1 + \beta_2 x)^{-\frac{1}{\beta_3}}$,

Model B (Holliday, 1960): $y = \frac{1}{\beta_1 + \beta_2 x + \beta_3 x^2}$,

Model C (Farazdaghi a Harris, 1968): $y = \frac{1}{\beta_1 + \beta_2 x^{\beta_3}}$,

Model D (asymptotický, Mead, 1979): $y = \frac{1}{\beta_1 + \beta_2 x}$,

kde biologický výklad parametrů říká, že β_1 je *mírou genetického potenciálu plodiny*, nepotlačené vlivem konkurenčního plevele, zatímco β_2 je *mírou potenciálu konkurenčního plevele*. Na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům výnosu cibule **B**rown **I**mperial **S**panish (BIS) na dvou lokalitách jižní Austrálie. Využijte kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*. Komentujte i citlivost parametrů v modelu.

Data: Hustota sazenic na m^2 *x* [počet sazenic/m^2], výnos cibule *y* [g/sazenice].

Lokalita Mount Gambier:

20.64	176.58	26.91	159.07	26.91	122.41	28.02	128.32	32.44	125.77	34.28	126.81
...	...	...	...	...	...	...	...	...	...	...	...
108.75	65.19	115.38	57.10	150.77	52.68	152.24	47.01	155.19	44.28		

Lokalita Uraidla:

22.30	148.57	25.86	125.30	29.09	150.69	29.74	147.42	31.68	117.10	31.68	116.64
...	...	...	...	...	...	...	...	...	...	...	...
119.92	61.60	120.89	26.30	126.71	61.20	138.99	41.60	146.75	45.20	160.97	46.40

Úloha E8.03 *Model výnosu úrody cibule typu WIS ve dvou lokalitách*

Pro obecný vztah mezi výnosem pěstované plodiny *y* a hustotou sadby zeleniny *x* byly navrženy následující nelineární regresní modely, uvedené v úloze **E8.02**. Na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům výnosu cibule **W**hite **I**mperial **S**panish (WIS) na dvou lokalitách jižní Austrálie. Využijte kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*. Komentujte i citlivost parametrů v modelu.

Data: Hustota sazenic na m^2 *x* [počet sazenic/m^2], výnos cibule *y* [g/sazenice].

Lokalita Purnong Landing:

23.48	223.02	26.22	234.24	27.79	221.68	32.88	221.94	33.27	197.45
...	...	...	...	...	...	...	...	...	...
166.35	62.54	184.75	54.68						

Lokalita Virginia:

18.78	272.15	21.25	235.23	23.23	180.47	27.18	177.31	30.15	141.28
...	...	...	...	...	...	...	...	...	...

158.15	38.70	180.39	28.96

Úloha E8.04 *Růstový model časové závislosti narostlé trávy a cibule*
Pro obecný vztah, popisující nárůst, výnos nebo produkci v zemědělství, biologii, inženýrství a ekonomice byly navrženy nelineární regresní modely zvané také *růstové modely*:
Model A (Gompertz): $y = \beta_1 \exp\left[-\exp(\beta_2 - \beta_3 x)\right]$,
Model B (Logistic): $y = \dfrac{\beta_1}{1 + \exp\left(\beta_2 - \beta_3 x\right)}$,
Model C (Richards, 1959): $y = \dfrac{\beta_1}{[1 + \exp\left(\beta_2 - \beta_3 x\right)]^{1/\beta_4}}$,
Model D (Morgan-Mercer-Flodin, 1975): $y = \dfrac{\beta_2 \beta_3 + \beta_1 x^{\beta_4}}{\beta_3 + x^{\beta_4}}$,
Model E (Weibull, 1951): $y = \beta_1 - \beta_2 \exp\left[-\beta_3 x^{\beta_4}\right]$.
Na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených růstových modelů nejlépe odpovídá naměřeným datům: (a) závislost růstu trávy na pastvinách v čase, (b) závislost nárůstu cibulky a natě u cibule na čase. K rozlišení mezi modely využijte především kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*. Komentujte i citlivost jednotlivých parametrů v modelu.
Data:
(a) Čas x, růst trávy na pastvinách y.

9.0	8.93	14.0	10.80	21.0	18.59	28.0	22.33	42.0	39.35
57.0	56.11	63.0	61.73	70.0	64.62	79.0	67.08		

(b) Čas x, nárust cibulky a natě u cibule y.

1.0	16.01	2.0	33.01	3.0	65.08	4.0	97.02	5.0	191.55
...	...	...	...	...	...	...	...	...	...
11.0	724.93	12.0	699.56	13.0	689.96	14.0	637.56	15.0	717.41

Úloha E8.05 *Model časové závislosti velikosti okurek a obsahu vody ve fazolích*
Pro obecný vztah, popisující nárůst, výnos nebo produkci v zemědělství, biologii, inženýrství a ekonomice byly navrženy *růstové modely*, uvedené v úloze **E8.04**. Na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům: (a) závislost růstu plodů okurek y na čase x, (b) obsah vody ve fazolích y v závislosti na vzdálenosti od kořene x. K rozlišení mezi modely využijte především kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*. Komentujte i citlivost jednotlivých parametrů v modelu.
Data:
(a) Čas x, růst plodů okurek y.

0.0	1.23	1.0	1.52	2.0	2.95	3.0	4.34	4.0	5.26
5.0	5.84	6.0	6.21	8.0	6.50	10.0	6.83		

(b) Obsah vody ve fazolích x, vzdálenost od kořene y.

0.50	1.3	1.5	1.3	2.5	1.9	3.5	3.4	4.5	5.3	5.5	7.1
...	...	...	...	...	...	...	...	...	...	...	...
12.5	21.3	13.5	21.2	14.5	20.9						

Úloha E8.06 *Růstový model délky kapustňáka Ochechule bahenní v závislosti na stáří*
V zemědělství a biologii se k popisu rozličných empirických závislostí využívá asymptotického regresního modelu s

deterministickými proměnnými, známého jako *Mitscherlichův zákon*: $y = \beta_1 - \beta_2 \beta_3^x$ (*Model A*). Model je využíván např. i ve výzkumu živočichů, kde je znám pod názvem *růstová křivka podle von Bertalanffyho*, vystihující závislost délky jedince na jeho stáří. Často byl reparametrizován a v literatuře se často setkáváme s jeho modifikacemi:

Model B: $y = \beta_1 - \beta_2 \exp(-\beta_3 x)$,

Model C: $y = \beta_1 (1 - \exp\{-(x - \beta_2) \beta_3\})$,

Model D: $y = \beta_1 - \exp[-(\beta_2 - \beta_3 x)]$,

Model E: $y = \beta_1 - \exp(-\beta_2 \beta_3^x)$,

Model F: $y = \frac{1}{\beta_1} - \beta_2 \beta_3^x$,

Model G: $y = \exp\beta_1 - \beta_2 \beta_3^x$.

Na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům závislosti délky savce kapustňáka, zvaného *Ochechule bahenní (Dugong dugong)* na jeho stáří. K rozlišení mezi modely využijte především kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* nebo Akaikova informačního kritéria *AIC*. Komentujte i citlivost jednotlivých parametrů v modelu.

Data: Stáří kapustňáka *x* [roky], délka kapustňáka *Ochechule bahenní y* [stopa].

1.0	1.8	1.5	1.85	1.5	1.87	1.5	1.77	2.5	2.02	4.0	2.27	5.0	2.15
...	...	...	...	...	...	...	...	...	...	...	...	...	...
15.5	2.64	16.5	2.64	17.0	2.65	25.0	2.70	29.0	2.72	31.5	2.57		

Úloha E8.07 *Růstový model u počtu lístků stromu v závislosti na osvětlení*

V biologii se sleduje růst listů na stromech v závislosti na fotosyntéze, podmíněné intenzitou osvětlení. Využitím modelů úlohy **E8.06** a na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům, tj. počtu nových lístků vyrostlých na výhonku za den *y* v závislosti na intenzitě osvětlení *x* wattů na m^2 při 20EC. K rozlišení mezi modely využijte především kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*. Komentujte i citlivost jednotlivých parametrů v modelu.

Data: Intenzita osvětlení *x* [W/m^2], počet nových lístků *y*.

12.0	0.094	23.0	0.119	40.0	0.199	92.0	0.260	156.0	0.309	215.0	0.331

Úloha E8.08 *Modifikace Mitscherlichova zákona u výnosu obilí na intenzitě hnojení*

V zemědělství se sleduje výnos mlynářského obilí na intenzitě hnojení. Využitím modelů úlohy **E8.06** a na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům, tj. výnosu obilí *y* v [q/ha] na intenzitě hnojení *x*. K rozlišení mezi modely využijte především kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* a Akaikova informačního kritéria *AIC*. Komentujte i citlivost jednotlivých parametrů v modelu.

Data: Intenzita hnojení *x*, výnos obilí *y* [q/ha].

0	26.2	10.0	30.4	20.0	36.3	30.0	37.8	40.0	38.6

Úloha E8.09 *Růstový model závislosti obvodu kmene kaučukovníku na hnojení*

Při sběru surového kaučuku se také sleduje velikost obvodu kmene stromu kaučukovníku *y* na intenzitě hnojení *x*. Využitím modelů úlohy **E8.06** a na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům, tj. velikosti obvodu kaučukovníku na intenzitě hnojení. K rozlišení mezi modely využijte především kritérií těsnosti proložení, střední kvadratické chyby predikce *MEP* a Akaikova informačního

kritéria *AIC*. Komentujte i citlivost jednotlivých parametrů v modelu.

Data: Intenzita hnojení *x*, obvod stromu kaučukovníku *y*.

0	20.518	1.0	21.138	3.0	21.734	5.0	22.218	7.0	22.286

Úloha E8.10 *Parametry regresního modelu "větev-hyperbola"*

Griffiths a Miller (1973) navrhli regresní model pro závislosti, kdy dvě rozličné lineární větve přechází pozvolna (hyperbolicky) jedna v druhou. Analyzovaný nelineární regresní model má tvar

$$y = \beta_1 + \beta_2 (x - \beta_4) + \beta_3 \sqrt{(x - \beta_4)^2 + \beta_5}.$$

Vyšetřete regresní triplet a odhadněte parametry modelu pro soubor dat, popisující chování výšky stojaté povrchové vrstvy *y* v cm v závislosti na rychlosti řízeného toku vody nakloněným kanálem *x* [g/cm. s].

Data: Logaritmus rychlosti řízeného toku vody nakloněným kanálem (log *x*), logaritmus výšky nepohyblivé, stojaté povrchové vrstvy (log *y*).

-1.39	1.12	-1.39	1.12	-1.08	0.99	-1.08	1.03	-0.80	0.90
...	...	...	...	...	...	...	...	...	...
0.85	-0.30	0.85	-0.33	0.99	-0.46	0.99	-0.43		

Úloha E8.11 *Parametry regresního modelu "větev-hyperbola"*

Dvě rozličné lineární větve přechází pozvolna (hyperbolicky) jedna v druhou. Nelineární regresní model má tvar

$$y = \beta_1 + \beta_2 (x - \beta_4) + \beta_3 \sqrt{(x - \beta_4)^2 + \beta_5}.$$

Vyšetřete regresní triplet a odhadněte parametry modelu popisujícího závislost pro následující soubor dat.

Data: x, y:

1.0	290.426	2.0	295.632	3.0	299.183	4.0	302.900	5.0	307.454
...	...	...	...	...	...	...	...	...	...
21.0	345.731	22.0	346.648	23.0	345.517	24.0	346.544	25.0	346.899

Úloha E8.12 *Parametry regresního modelu "větev-hyperbola"*

Dvě rozličné lineární větve přechází pozvolna (hyperbolicky) jedna v druhou. Model má tvar

$$y = \beta_1 + \beta_2 (x - \beta_4) + \beta_3 \sqrt{(x - \beta_4)^2 + \beta_5}.$$

Vyšetřete regresní triplet a odhadněte parametry modelu pro následující soubor dat, popisující závislost.

Data: x, y:

1.0	113.978	2.0	115.288	3.0	116.709	4.0	117.624	5.0	118.776
...	...	...	...	...	...	...	...	...	...
26.0	137.724	27.0	137.128						

Úloha E8.13 *Parametry regresního modelu úrody obilí v závislosti na množství hnojiva*

Byla zkoumána závislost výnosu obilí *y* na množství použitého hnojiva *x*. Využitím modelů úlohy **E8.06** a na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům, tj. výnosu obilí *y* v [q/ha] na intenzitě hnojení *x*.

Data: Intenzita hnojení *x*, výnos obilí *y* [q/ha].

0.0	44.4	0.5	50.3	1.0	54.6	1.5	59.5	2.0	63.8
2.5	64.8	3.0	65.7	3.5	67.9	4.0	68.9	4.5	70.0

Úloha E8.14 *Parametry regresního modelu výnosu brambor na množství P_2O_5 v hnojivu*

Byla zkoumána závislost výnosu brambor *y* na množství oxidu fosforečného v použitém hnojivu *x*. Využitím modelů úlohy **E8.06** a na základě analýzy regresního tripletu a regresní diagnostiky rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům.

Data: Užitý P_2O_5 *x*, výnos brambor *y* [q/ha].

0.0	232.00	0.25	272.00	0.5	308.25	1.0	369.08	1.5	419.05
2.0	455.63	2.50	476.00	3.0	491.45	3.5	504.20	4.0	511.50

Úloha E8.15 *Růstový model závislosti obvodu kmene japonského jinanu na jeho stáří*

Obvod kmene japonského stromu jinan (*Nippon ginko*) *y* je sledováno v závislosti na stáří stromu *x*. Využitím modelů úlohy **E8.06** a na základě analýzy regresního tripletu rozhodněte, který z předložených modelů nejlépe odpovídá naměřeným datům, tj. velikosti kmene jinanu *y* v závislosti na stáří stromu *x*. Odhadněte stáří stromu, je-li jejich obvod je $y^* = 30, 60, 90$ cm. Jaký obvod se dá očekávat u stromu starého $x^* = 100, 200$ a 300 let?

Data: Stáří stromu *x* [roky], obvod stromu *y* [cm].

5	2	15	5	25	12	35	20	45	28	55	38	65	45
75	54	85	60	95	67	105	72	115	76	125	82		

Úloha E8.16 *Růstový model procenta matek s dítětem na stáří matky*

Gompertzův vztah byl doporučen pro popis procenta matek s alespoň jedním narozeným dítětem *y* v závislosti na stáří matky *x* v různých letech tohoto století. Na základě analýzy regresního tripletu rozhodněte, který z dalších možných modelů odpovídá křivkám čtveřice předložených dat.

Model A (Gompertz): $$y = \beta_1 \exp\left[-\exp\left\{\beta_2 - \beta_3 x\right\}\right],$$

Model B (Logistic): $$y = \frac{\beta_1}{1 + \exp\left\{\beta_2 - \beta_3 x\right\}},$$

Model C (Richards, 1959): $$y = \frac{\beta_1}{\left[1 + \exp\left(\beta_2 - \beta_3 x\right)\right]^{1/\beta_4}}.$$

Data: Stáří matky *x*, procento matek s alespoň 1 dítětem *y* [%].

(a) *Rok 1920:* 15 0, 20 7, 25 39, 30 67, 35 76, 40 78,
(b) *Rok 1930:* 15 0, 20 9, 25 48, 30 75, 35 83, 40 86, 45 86,
(c) *Rok 1940:* 15 0, 20 13, 25 59, 30 82, 35 87, 40 89, 45 89,
(d) *Rok 1945:* 15 0, 20 17, 25 60, 30 82, 35 88, 40 90.

Úloha E8.17 *Model biologické spotřeby kyslíku v závislosti na čase*

Experimentální data z knihovny NIST se týkají biologické spotřeby kyslíku *y* v závislosti na čase *x*. Závisle proměnnou je biologická spotřeba kyslíku *y* v mg/l a nezávisle proměnnou je inkubační čas *x* ve dnech. Autory[92] byl navržen nelineární regresní model

$$y = b_1 (1 - \exp[-b_2 x]).$$

Data: Nezávisle proměnnou je inkubační čas *x* [dny], závisle proměnnou je biologická spotřeba kyslíku *y* [mg/l]. Pro počáteční odhad parametrů b_1, b_2 jsou doporučeny hodnoty 1, 1 nebo 100, 0.75.

1	109	2	149	3	149	5	191	7	213	10	224

8.2.4 Analýza hutnických a mineralogických dat

Úloha H8.01 *Model časového průběhu poklesu teploty při kalení oceli*

Průběh poklesu teploty oceli byl měřen v pravidelných časových intervalech termočlánkem. Za předpokladu konstantního chování vzorku v měřeném teplotním intervalu je model ochlazování

$$y = \beta_1 + \beta_2 \exp(-\beta_3 x),$$

kde y je teplota, β_1 je počáteční teplota, β_2 je rozdíl výchozí a konečné teploty, β_3 je časová konstanta a x je doba od začátku chlazení.

Data: Čas x [s], teplota y [EC].

27	230.02	0	636.27	28	222.58	1	596.01	29	215.11	2	549.06	30	207.63
...	...	...	...	...	...	...	...	...	...	...	...	...	...
24	251.97	52	89.05	25	244.84	53	86.64	26	237.44	54	86.64		

Úloha H8.02 *Odhady parametrů závislosti vodivosti termistoru na teplotě*
Meyer a Roth odvodili závislost vodivosti y termistoru na teplotě x ve formě

$$y = \beta_1 \exp\left[\frac{\beta_2}{x + \beta_3}\right],$$

která říká, že vodivost termistoru y se zvyšuje s teplotou x [EC]. Regresní model by měl vykazovat aditivní chyby. Jelikož však rezidua vykazují evidentní heteroskedasticitu, jeví se zde pravděpodobnější případ multiplikativních chyb. Pak bude platit i alternativní model

$$\log y = -\beta_1 + \frac{\beta_2}{x + \beta_3},$$

který ověřte. Rozhodněte, který ze dvou navržených modelů platí, vyšetřete regresní triplet a odhadněte neznámé parametry vybraného modelu.

Data: Teplota x [EC], vodivost termistoru y [Ω^{-1}].

50.0	34.78	55.0	28.61	60.0	23.65	65.0	19.63
...	...	...	...	...	...	...	...
110.0	4.42	115.0	3.82	120.0	3.31	125.0	2.87

8.2.5 Analýza matematických modelů a fyzikálních dat

Úloha S8.01 *Hledání adekvátního empirického modelu pro předložená literární data*
Rozhodněte, který z předložených modelů nejlépe odpovídá předloženým datům:

Model A: $y = \beta_1 + \beta_2 \exp[-\beta_3 x],$

Model B: $y = \beta_1 + \beta_2 x^{\beta_3},$

Model C: $y = \beta_1 + \frac{\beta_2}{1 + \beta_3 x}.$

Data: Hodnoty nezávisle x a závisle proměnné y.

1	1.64	2	1.50	3	1.40	4	1.34	5	1.30	6	1.26
7	1.23	8	1.21	9	1.19	10	1.18	11	1.16		

Úloha S8.02 *Odhady parametrů zadaného regresního modelu*
Nalezněte nejlepší odhady parametrů β_1, β_2, β_3 a β_4 nelineárního regresního modelu

$$y = \frac{\beta_1 x_1 \ln(\beta_2/x_2)}{e^{\beta_3 x_3} + \beta_4}$$

pro daná data. Vyčíslete i intervalové odhady parametrů. Diskujte spolehlivost odhadů a působení vlivných bodů na tyto odhady.

Data: Hodnoty nezávisle proměnných x_1 , x_2 , x_3 a závisle proměnné y.

1.00	0.10	0.10	0.810280	10.00	0.10	0.10	8.102801	15.00	0.10	0.10	12.15400
...	...	...	...	...	...	...	...	...	...	...	...
15.00	1.00	0.10	4.567600	5.00	1.00	0.10	1.522500	75.00	1.00	0.10	22.83800

Úloha S8.03 *Odhady parametrů zadaného regresního modelu*

Nalezněte nejlepší odhady parametrů β_1, β_2, β_3 a β_4 nelineárního regresního modelu

$$y = (\beta_1 + \beta_2 x_1^2 + \cos(\beta_3 x_2))^{\sin(\beta_4 x_3)}$$

pro tato data. Vyčíslete i intervalové odhady parametrů. Diskutujte spolehlivost odhadů a působení vlivných bodů na tyto odhady.

Data: Hodnoty nezávisle proměnných x_1 , x_2 , x_3 a závisle proměnné y.

75.0	33.0	75.0	7.738501E-04	68.0	15.0	68.0	4.237200E-04
...	...	...	...	...	...	...	...
81.0	13.0	81.0	4.743500E-05	63.0	11.0	63.0	2.433600E+01

Úloha S8.04 *Sušicí koeficient v závislosti na tloušťce kůže a rychlosti sušicího vzduchu*

Sleduje se průběh sušení tak, že se v čase měří tloušťka kůže x_1 a rychlost proudícího sušicího vzduchu x_2 a sušicí koeficient y se počítá. Cílem je určit hodnotu sušicího koeficientu y v závislosti na tlouštce vysušované kůže x_1 a rychlosti proudícího sušicího vzduchu x_2 dle modelu $y = \beta_1 x_2^{\beta_2}/x_1$ s neznámými parametry β_1 a β_2. Vyšetřete regresní triplet a nalezněte nejlepší odhady parametrů β_1 a β_2 zadaného nelineárního regresního modelu pro následující soubor dat. Vyčíslete i intervalové odhady parametrů. Diskutujte spolehlivost odhadů a působení vlivných bodů na tyto odhady.

Data: Tloušťka kůže x_1 [mm], rychlost sušicího vzduchu x_2 [kg/(m^2. min)], sušicí koeficient y.

1.05	14.1	1.305	1.17	42.7	1.90	1.06	42.0	2.71
...	...	...	...	...	...	...	...	...
1.43	99.6	3.55						

Úloha S8.05 *Důležitost vlivných bodů u navrženého regresního modelu*

Při studiu kvantových vlastností částic byla změřena následující závislost imaginární složky hybnosti x a vzdálenosti y. Určete regresní model a vyšetřete regresní triplet. Nalezněte kladnou vzdálenost β_2 (teoreticky $\beta_2 = 2$) v modelu

$$y = \frac{\beta_1}{x + \beta_2}.$$

Data: Imaginární složka hybnosti x, vzdálenost y.

50	0.0504	40	0.0568	35	0.0699	30	0.0760	25	0.0884	20	0.1166
15	0.1471	12	0.1787	11	0.1920	7	0.2961	3	1.4951		

Úloha S8.06 *Stanovení časové konstanty zařízení při jeho oteplování*

Při vyhodnocování tepelně-setrvačných procesů výroby léčiv jsou stanovovány oteplovací konstanty. Podle nich jsou určeny časy předehřevu k zajištění konstantní podmínky sušicích a granulačních procesů. Teplota uvnitř fluidního granulátoru je měřena pomocí skupiny teplotních snímačů tak, že postupně jsou nastavovány teploty, které mají být regulačním systémem udržovány uvnitř zařízení. Měřicí aparaturou jsou monitorovány přechodové charakteristiky a jejich

průběh je vyhodnocen metodou nelineární regrese. Z takto naměřených přechodových charakteristik lze usuzovat, že měřenou závislost lze popsat statickou soustavou 2. řádu. Rozhodněte, který z navržených čtyř modelů pro rostoucí (oteplovací) přechodovou charakteristiku y v závislosti na čase x nejlépe vyhovuje naměřených datům:

1. model: $y = \beta_1 \exp[\beta_2 x] + \beta_3 \exp[\beta_4 x] + \beta_5$,

2. model: $y = \beta_1 \exp[\beta_2 x] \left\{ \left(\frac{\beta_2}{\beta_3} \sin[\beta_3 x] - \cos[\beta_3 x] \right) + 1 \right\} + \beta_4$,

3. model: $y = \beta_1 \exp[\beta_2 x] + \beta_3 \exp[\beta_4 x] + \beta_5 \exp[\beta_6 x] + \beta_7$,

4. model: $y = \beta_1 \exp[\beta_2 x] \sin[\beta_3 x + \beta_4] + \beta_5$.

Data: Data přechodové charakteristiky představují čas x [min] a teplotu y [EC].

x	0.00	0.33	0.67	1.00	1.33
y	18.20	18.70	20.50	22.90	25.50
...	...	...	...	...	...
x	16.66	16.99	17.33	17.66	
y	66.00	66.10	66.40	66.50	

Úloha S8.07 *Stanovení časové konstanty zařízení při jeho ochlazování*
Při vyhodnocování tepelně-setrvačných procesů výroby léčiv jsou stanovovány oteplovací konstanty. Podle nich jsou určeny časy předehřevu k zajištění konstantní podmínky sušicích a granulačních procesů. Rozhodněte, který z navržených čtyř modelů pro klesající (ochlazovací) přechodovou charakteristiku y v závislosti na čase x nejlépe vyhovuje naměřených datům:

1. model: $y = \beta_1 + \beta_2 \beta_3^x$,

2. model: $y = \beta_1 (1 - \exp[\beta_2 x]) + \beta_3 (1 - \exp[\beta_4 x]) + \beta_5$,

3. model: $y = \beta_1 (1 - \exp[\beta_2 x]) \left\{ \left(\frac{\beta_2}{\beta_3} \sin[\beta_3 x] - \cos[\beta_3 x] \right) + 1 \right\} + \beta_4$,

4. model: $y = \beta_1 (1 - \exp[\beta_2 x]) \sin(\beta_3 x + \beta_4) + \beta_5$.

Data: Data přechodové charakteristiky představují dobu x [min] a teplotu y [EC].

x	0.00	0.33	0.67	1.01	1.34
y	71.70	71.20	70.50	69.70	68.60
...	...	...	...	...	...
x	10.00	10.33	10.67	11.01	
y	30.20	29.70	29.40	29.10	

Úloha S8.08 *Stanovení časové konstanty a doby k ustálení při regulaci teploty*
Pro data naměřená při regulaci teploty substance najděte vhodný regresní model a určete časovou konstantu celého systému a dobu potřebnou k ustálení. Vyberte vhodnější z modelů:

Model 1. řádu: $y = b_1 + b_2\left\{1 - \exp\left[-\frac{x}{b_3}\right]\right\}$,

kde x je čas, y je hodnota teploty, b_1 je počáteční teplota, b_2 je součinitel, b_3 je časová konstanta, a

Model 2. řádu: $y = b_1 + b_2\left\{1 - \exp\left[-\frac{x}{b_3}\right]\right\}\left\{b_4 \sin(b_5 + b_6 x)\right\}$,

kde x je čas, y je hodnota teploty, b_1 je počáteční teplota, b_2 je součinitel, b_3 je časová konstanta, b_4 je amplituda zvlnění, b_5 je počáteční fáze, b_6 je vlastní kmitočet soustavy.

Data: Doba x [min], teplota y [EC].

0	8	130	56	300	64	430	60
...	...	...	...	...	...	...	...
120	60	290	60	420	68	550	60

Úloha S8.09 *Aproximace denních ranních teplot v roce křivkou*

Data představují nejnižší ranní teploty každého dne, naměřené v Praze v roce 1990. Osa x představuje pořadové číslo dne v roce, osa y je zmíněná ranní teplota. Pokuste se aproximovat uvedená data vhodnou křivkou. Denní teploty se během roku periodicky mění, což platí i pro denní minima teplot. Nejvhodnější křivkou se jeví harmonická funkce ve tvaru $y = b_1 + b_2 \sin(b_3 + b_4 x)$, kde y je denní minimum teploty, b_1 je konstanta (střední hodnota teploty v roce), b_2 je amplituda kolísání teploty, b_3 je počáteční fáze, b_4 je frekvence změn, x je pořadové číslo dne v roce.

Data: Pořadové číslo dne v roce x a ranní teplota y [EC].

1	-3.1	93	7.6	185	9.4	277	8.6
...	...	...	...	...	...	...	...
92	3.4	184	9.8	276	4.8		

Úloha S8.10 *Model supravodivosti a magnetismu NIST(1994)*

Experimentální data z knihovny NIST se týkají supravodivosti a magnetismu. Závisle proměnnou y je magnetismus a nezávisle proměnnou x je logaritmus doby v minutách. Autory[93] byl navržen nelineární regresní model $y = b_1 (b_2 + x)^{-1/b_3}$.

Data: Nezávisle proměnnou x představuje logaritmus doby [min], závisle proměnnou y magnetismus. Pro počáteční odhady parametrů b_1, b_2, b_3 jsou doporučeny -2000, 50, 0.8 nebo -1500, 45, 0.85.

7.447168	-34.8347	10.93126	-32.5832	11.60712	-32.1749	12.00674	-31.9403
...	...	...	...	...	...	...	...
10.90619	-32.5977	11.59442	-32.1836	11.99826	-31.9461		

Úloha S8.11 *Model kruhového rušení transmitance v závislosti na vlnové délce*

Data z knihovny NIST se týkají kruhového rušení transmitance v závislosti na vlnové délce. Závisle proměnnou y je transmitance a nezávisle proměnnou x je vlnová délka v nm. Autory[94] byl navržen nelineární regresní model

$$y = \frac{b_1}{b_2} \exp\left[-0.5\left\{\frac{(x - b_3)}{b_2}\right\}^2\right].$$

Data: Vlnová délka představuje nezávisle proměnnou x [nm] a transmitance závisle proměnnou y. Pro počáteční odhad parametrů b_1, b_2, b_3 jsou doporučeny hodnoty 1, 10, 500 nebo 1.5, 5, 450.

400.0	0.000158	442.5	0.040168	460.5	0.033720
...	...	...	...	...	...
441.0	0.023731	459.0	0.061676		

Úloha S8.12 *Model dvou Gaussovských píků na klesající základní linii*
Simulovaná data se týkají dvou Gaussovských píků superponovaných na klesající základní linii plus normálně rozdělený šum se střední hodnotou nula a rozptylem hodnoty 6.25. Autor[95] navrhl nelineární regresní model (NIST, 1996)

$$y = b_1 \exp(-b_2 x) + b_3 \exp\left[-\frac{(x - b_4)^2}{b_5^2}\right] + b_6 \exp\left[-\frac{(x - b_7)^2}{b_8^2}\right].$$

Data: Nezávisle proměnná x, závisle proměnná y. Pro počáteční odhad parametrů b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8 jsou doporučeny hodnoty 96.0, 0.009, 103.0, 106.0, 18.0, 72.0, 151.0, 18.0 nebo 98.0, 0.0105, 103.0, 106.0, 18.0, 72.0, 151.0, 18.0.

1	97.58776	46	54.38790	91	93.46992	136	77.60769	181	20.26812	226	6.118575
...	...	...	...	...	...	...	...	...	...	...	...
45	61.21785	90	95.86509	135	76.76062	180	23.69643	225	7.695971		

Úloha S8.13 *Model ultrazvukové kalibrace vzdálenosti kovů*
Experimentální data se týkají měření vzdálenosti mezi dvěma kovy ultrazvukem. Nezávisle proměnnou x je vzdálenost mezi kovy a závisle proměnnou y je signál ultrazvuku. Autor[96] navrhl nelineární regresní model

$$y = \exp\left[\frac{-1}{b_1/x + b_2}\right].$$

Data: Nezávisle proměnnou x je vzdálenost mezi kovy a závisle proměnnou y je signál ultrazvuku. Pro počáteční odhad parametrů b_1, b_2, b_3 jsou doporučeny hodnoty 0.1, 0.01, 0.02 nebo 0.15, 0.008, 0.010.

0.500	92.900	1.250	41.000	2.000	20.000	2.500	17.700
...	...	...	...	...	...	...	...
0.625	67.300	0.500	75.800	0.750	61.300		

Úloha S8.14 *Numerický model nelineární regrese k testování algoritmů*
Data byla generována na 6 správných cifer užitím nelineárního modelu

$$y = 0.0951 \exp(-x) + 0.8607 \exp(-3x) + 1.5576 \exp(-5x).$$

Takto generovanými daty byl prokládán nelineární regresní model tvaru[97]

$$y = b_1 \exp(-b_2 x) + b_3 \exp(-b_4 x) + b_5 \exp(-b_6 x).$$

Data: Nezávisle proměnná x, závisle proměnná y. Pro počáteční odhady parametrů b_1, b_2, b_3, b_4, b_5, b_6 jsou doporučeny hodnoty 1.2, 0.3, 5.6, 5.5, 6.5, 7.6 nebo 0.5, 0.7, 3.6, 4.2, 4.0, 6.3.

0.00	2.5134000	0.60	0.2720130
...	...	...	...
0.55	0.3197390	1.15	0.0623931

Úloha S8.15 *Numerický model nelineární regrese k testování algoritmů*
Data byla generována na 6 správných cifer a generovanými daty byl prokládám nelineární regresní model tvaru[98]

$$y = b_1 + b_2 \exp[-x\, b_4] + b_3 \exp[-x\, b_5].$$

Data: Nezávisle proměnná x, závisle proměnná y. Pro počáteční odhady parametrů b_1, b_2, b_3, b_4, b_5 jsou doporučeny hodnoty 50, 150, -100, 1, 2 nebo 0.5, 1.5, -1, 0.01, 0.02.

0	0.8440	90	0.7840	180	0.5380	270	0.4310
...	...	...	...	...	...	...	...
80	0.8180	170	0.5580	260	0.4380		

Úloha S8.16 *Model sigmoidální růstové křivky k testování algoritmů (Ratkowsky)*
Nelineární regresní model se týká prokládání sigmoidální růstové křivky daty, ve kterých nezávisle proměnnou tvoří čas a závisle proměnnou výtěžnost pastviny. Ratkowsky[99] navrhl následující model

$$y = \frac{b_1}{1 + \exp[b_2 - b_3 x]}.$$

Data: Nezávisle proměnná *x*, závisle proměnná *y*. Pro počáteční odhady parametrů b_1, b_2, b_3 jsou doporučeny hodnoty 100, 1.0, 0.1, nebo 75, 2.5, 0.07.

9	8.93	14	10.8	21	18.59	28	22.33	42	39.35
57	56.11	63	61.73	70	64.62	79	67.08		

8.3 Kontrolní hodnoty (ADSTAT, NCSS2000)

8.3.1 Analýza farmakologických a biochemických dat

B8.01 *(a) b_1 = 279.2 (4.2), b_2 = 127.2 (6.8), b_3 = 35.9 (4.4), RSC = 4382.0, D = 98.65%, AIC = 308.01, (b) b_1 = 5.64 (0.02), b_2 = 130.2 (5.9), b_3 = 37.5 (2.2), RSC = 0.2699, D = 99.055%, AIC = -409.4*

B8.02 *(a) b_1 = 244.5, b_2 = 169.1, b_3 = -0.015, b_4 = 8.45, RSC = 177.6, D = 99.21%, (b) b_1 = 428.6, b_2 = 206.2, b_3 = -0.046, b_4 = 4.68, RSC = 486.9, D = 99.54%.*

8.3.2 Analýza chemických a fyzikálních dat

C8.01 *b_1 = 9.2213E-03 N m^{-1}, b_2 = 6.3564E-02 m^3 mol^{-1}, RSC = 9.666E-15, s(e) = 2.25E-08*

C8.02 *(a) b_1 = 16.289, b_2 = 3816.4, b_3 = 227.02, RSC = 8.9168E-11,*

(b) b_1 = 14.098, b_2 = 3774.5, b_3 = 181.83,RSC = 9.1294E-13.

C8.03 *b_1 = 17.865, b_2 = 5.85E-04, b_3 = -1.653E+05, RSC = 1.6824E-10.*

C8.04 *b_1 = 7.641, b_2 = -3073.8, b_3 = 214.7, RSC = 9.2024E-13,*

benzen: b_1 = 6.019, b_2 = -1204.7, b_3 = 220.07, RSC = 1.6363E-12.

C8.05 *(a) b_1 = 91.826, b_2 = 6.270E-03, b_3 = 1.761E-05, b_4 = -9.7805E+06, RSC = 3.72E-09,*

(b) b_1 = 91.824, b_2 = 6.2737E-03, b_3 = 1.7608E-05, b_4 = -9.4804E+06, RSC = 8.178E-06.

C8.06 *b_1 = -241.78, b_2 = 12660.0, b_3 = 34.72, RSC = 1.638E-07.*

C8.07 *b_1 = 125.15, b_2 = -6161.7, b_3 = -19.778, RSC = 4.082E-12.*

C8.08 *(a) b_1 = 5.034 (0.004), b_2 = 7.60 (0.21), b_3 = 0.068 (0.003), RSC = 4.03E-04, D = 99.49%. (b) b_1 = 6.138 (0.003), b_2 = 7.85 (0.18), b_3 = 0.055 (0.003), RSC = 2.54E-04, D = 99.69%. (c) b_1 = 6.197 (0.006), b_2 = 8.80 (0.35), b_3 = 0.055 (0.004), RSC = 6.96E-04, D = 98.90%. (d) b_1 = 7.271 (0.040), b_2 = 5.45 (1.46), b_3 = 0.083 (0.040), RSC = 6.05E-02, D = 80.56%. (e) b_1 = 8.041 (0.009), b_2 = 3.60 (0.70), b_3 = 0.107 (0.031), RSC = 9.03E-04, D = 99.41%. (f) b_1 = 9.232 (0.011), b_2 = 8.04 (0.49), b_3 = 0.061 (0.006), RSC = 8.21E-04, D = 97.92%.*

C8.09 *(a) b_1 = 4.18 (0.04), b_2 = 0.088 (0.013), b_3 = 0.771 (0.014), RSC = 9.01E-03, D = 98.90%. (b) b_1 = 0.61 (0.06), b_2 = 0.098 (0.027), b_3 = 0.590 (0.044), RSC = 5.59E-04, D = 99.46%. (c) b_1 = 5.01 (0.03), b_2 = -9.0E-03 (0.01), b_3 = 0.449 (0.015), RSC = 1.97E-04, D = 99.87%. (d) b_1 = 10.65 (0.07), b_2 = -0.054 (0.019), b_3 = 0.270 (0.018), RSC = 1.30E-04, D = 99.88%. (e) b_1 = 3.10 (0.01), b_2 = 0.233 (0.005), b_3 = 0.875 (0.006), RSC = 2.55E-03, D = 99.72%.*

C8.10 *(a) b_1 = 9.84 (0.03), b_2 = 7.59 (0.02), b_3 = 0.020 (0.007), b_4 = 1.067 (0.011), b_5 = 0.366 (0.010), RSC = 6.42E-04, D = 99.97%. (b) b_1 = 6.98 (0.03), b_2 = 2.88 (0.02), b_3 = 0.278 (0.003), b_4 = 0.481 (0.002), b_5 = 0.649 (0.002), RSC = 3.72E-04, D = 99.95%. (c) b_1 = 7.31 (0.05), b_2 = 3.07 (0.06), b_3 = 0.286 (0.007), b_4 = 0.469 (0.003), b_5 = 0.637 (0.004), RSC = 2.66E-04, D = 99.89%.*

C8.11 *(a) b_1 = 9.11 (0.03), b_2 = 5.08 (5.90), b_3 = 3.56 (0.15), b_4 = 0.764 (0.010), b_5 = 0.363 (0.162), b_6 = 0.353 (0.003), b_7 = 0.078 (0.005), RSC = 2.51E-04, D = 99.98%. (b) b_1 = 7.34 (0.10), b_2 = 3.92 (0.77), b_3 = 2.72 (0.11), b_4 = 0.698 (0.006), b_5 = 0.298 (0.075), b_6 = 0.240 (0.005), b_7 = 0.138 (0.0037), RSC = 3.95E-04, D = 99.97%. (c) b_1 = 7.48 (0.04), b_2 = 4.14 0.56), b_3 = 2.73 (0.03), b_4 = 0.212*

(0.002), b_5 = 0.365 (0.003), b_6 = 0.355 (0.002), b_7 = 0.446 (0.002), RSC = 3.47E-05, D = 99.98%.

C8.12 *b_3 = -2.182E-03, b_4 = 9.980E-01, b_2 = 4.789E-01, b_5 = 7.875E00, RSC = 0.1814, D = 99.94%.*

C8.13 *(a) po vyloučení 1 bodu: b_3 = -2.716E01, b_4 = 7.402, b_2 = -1.003E01, b_5 = 7.32E00, RSC = 5.236, D = 99.93%. (b) b_3 = 6.736E01, b_4 = 3.880E01, b_2 = 4.406E01, b_5 = 1.287E01, RSC = 18.68, D = 99.99%.*

C8.14 *po vyloučení 1 bodu: b_3 = -4.301E01, b_4 = 8.133, b_2 = 2.786, b_5 = 8.27, RSC = 2.56, D = 99.98%.*

C8.15 *b_1 = 3.101E01, b_2 = -1.683E-02, b_3 = 4.390E-01, RSC = 2.552, D = 99.78%.*

C8.16 *b_1 = 3.131 (0.808), b_2 = 15.159 (0.631), b_3 = 0.780 (0.152), RSC = 4.36E-05, D = 99.66%.*

C8.17 *b_1 = 5.628 (2.28E+06), b_2 = 3.474 (3.50E+05), b_3 = 110.38 (58.13), RSC = 2.911E+03, D = 99.77%.*

C8.18 *b_1 = 5.267 (2.274), b_2 = 8.565 (2.043), b_3 = 295.0 (127.2), RSC = 1718.21, D = 99.86%, MEP = 335.61, AIC = 73.34.*

C8.19 *b_1 = 1.146 (4.17E-04), b_2 = -2.540 (9.47E-03), b_3 = 450.36 (3.03E-05), b_4 = 2.37E+07, RSC = 1.19E-04, D = 99.98%.*

C8.20 *y_1: 2. model b_1 = -9.25 (0.25), b_2 = -7.78E-03 (5.65E-04), b_3 = -9.225 (0.206), b_4 =-7.61E-03 (2.59E-04), RSC = 0.870, D = 99.37%, MEP = 0.0589, AIC = -63.06. y_2: 2. model b_1 = -8.256 (1.148), b_2 = -4.70E-04 (7.73E-05), b_3 = -8.928 (1.236), b_4 = -6.47E-03 (8.74E-04), RSC = 5.417, MEP = 0.378, D = 91.27%, AIC = -22.831.*

8.3.3 Analýza environmetálních, potravinářských a zemědělských dat

E8.01 *b_1 = 0.505 (0.137), b_2 = 1.852 (0.111), RSC = 8.398, D = 93.32%, MEP = 1.230, AIC = -0.282.*

E8.02 *(a) Model A: b_1 = 1.008 (0.016), b_2 = -1.42E-05 (3.16E-05), b_3 = -0.00155 (0.00300), RSC = 6386.8, D = 84.70%, MEP = 182.93, AIC = 212.98. (b) Model A: b_1 = 0.058 (0.124), b_2 = 5.3E-04 (5.31E-04), b_3 = 0.5307 (0.3827), RSC = 4909.2, D = 89.56%, MEP = 134.79, AIC = 205.97.*

E8.03 *(a) Model C: b_1 = 0.00275 (0.00050), b_2 = 3.29E-05 (2.51E-05), b_3 = 1.199 (0.156), RSC = 5370.3, D = 95.30%, MEP = 152.44, AIC = 209.74. (b) Model A: b_1 = -0.0001 (6.8E-05), b_2 = 1.09E-05 (1.0E-05), b_3 = 1.652 (0.213), RSC = 7686.5, D = 93.06%, MEP = 233.44, AIC = 224.8.*

E8.04 *(a) Model B: b_1 = 69.62 (2.13), b_2 = 4.255 (1.268), b_3 = 0.089 (0.019), RSC = 6.049, D = 99.87%, MEP = 1.509, AIC = 4.424. (b) Model B: b_1 = 702.8 (13.9), b_2 = 4.45 (0.35), b_3 = 0.689 (0.057), RSC = 8922.4, D = 99.17%, MEP = 961.25, AIC = 101.82.*

E8.05 *(a) Model D: b_1 = 6.986 (0.086), b_2 = 1.181 (0.062), b_3 = 12.959 (1.606), b_4 = 2.475 (0.120), RSC = 0.0237, D = 99.94%, MEP = 0.0157, AIC = -45.42. (b) Model B: b_1 = 21.509 (0.415), b_2 = 3.957 (0.262), b_3 = 0.622 (0.045), RSC = 6.210, D = 99.35%, MEP = 0.714, AIC = -7.228.*

E8.06 *Model A: b_1 = 2.667 (0.058), b_2 = 0.973 (0.065), b_3 = 0.873 (0.022), RSC = 0.1868, D = 90.48%, MEP = 0.00863, AIC = -128.29.*

E8.07 *Model B: b_1 = 0.335 (0.017), b_2 = 0.2955 (0.0196), b_3 = 0.0163 (0.0038), RSC = 6.06E-04, MEP = 3.28E-04, AIC = -49.206.*

E8.08 *Model F: b_1 = 0.024 (0.002), b_2 = 15.839 (3.496), b_3 = 0.956 (0.021), RSC = 2.9089, MEP = 16.265, AIC = 3.29.*

E8.09 *Model E: b_1 = 23.623 (0.197), b_2 = -1.132 (0.060), b_3 = 0.811 (0.044), RSC = 0.0123, D = 99.45%, MEP = 0.01083, AIC = -24.048.*

E8.10 *b_1 = 0.586 (0.036), b_2 = 0.735 (0.013), b_3 = -0.359 (0.025), b_4 = 0.063 (0.031), b_5 = 0.096 (0.053), RSC = 5.15E-03, D = 99.93%, MEP = 2.51E-04, AIC = -240.44.*

E8.11 *b_1 = 346.37 (1.148), b_2 = -2.239 (0.048), b_3 = -2.190 (0.086), b_4 = 13.547 (0.179), b_5 = 4.598 (2.184), RSC = 7.123, D = 99.92%, MEP = 0.435, AIC = -21.39.*

E8.12 *b_1 = 136.82 (0.96), b_2 = -0.696 (0.045), b_3 = -0.587 (0.062), b_4 = 18.751 (0.514), b_5 = 6.724 (5.594), RSC = 2.108, D = 99.87%, MEP = 0.140, AIC = -58.86.*

E8.13 *Model C: b_1 = 72.767 (1.157), b_2 = 1.833 (0.174), b_3 = 0.508 (0.052), MEP = 0.81275, AIC = -3.9199.*

E8.14 *Model A: b_1 = 539.86 (4.914), b_2 = -311.46 (4.560), b_3 = -0.623 (0.026), RSC = 82.184, D = 99.91%, MEP = 22.8457, AIC = 27.0638.*

E8.15 *Model F: b_1 = 0.0036 (0.0012), b_2 = 283.69 (93.52), b_3 = 0.997 (0.001), RSC = 51.778, D = 99.44%, MEP = 10.353, AIC = 23.97.*

E8.16 *(a) Model D: b_1 = 90.425 (0.16), b_2 = 6.149 (0.048), b_3 = 0.282 (0.002), RSC = 0.1184, D = 99.99%, MEP = 0.138, AIC = -17.55. (b) Model A: b_1 = 86.51 (0.26), b_2 = 6.288 (0.095), b_3 = 0.273 (0.004), RSC = 0.5988, D = 99.99%, MEP = 0.56554, AIC = -11.211. (c) Model A: $b_{\bar{1}}$ = 89.104 (0.196), $b_{\bar{2}}$ = 6.848 (0.80), b_3 = 0.310 (0.004), RSC = 0.4121, D = 99.99%, MEP = 0.29663, AIC = -13.826. (d) Model A: b_1 = 90.43 (0.16), b_2 = 6.15 (0.05), b_3 = 0.282 (0.002), RSC = 0.1184, D = 99.99%, MEP = 0.13827, AIC = -17.55.*

E8.17 *b_1 = 2.1381E+02(1.235E+01), b_2 = 5.4724E-01(1.0456E-01), RSC = 1.1680E+03, D = 88.05%, MEP = 422.42, AIC = 35.63.*

8.3.4 Analýza hutnických a mineralogických dat

H8.01 *b_1 = 8.240 (11.591), b_2 = -576.94 (9.682), b_3 = 0.037 (0.002), RSC = 7340.5, D = 99.35% a po odstranění prvních 9 bodů: b_1 = -55.106 (6.136), b_2 = 598.50, b_3 = -0.228 (6.45E-04), RSC = 2.529E+02, D = 99.94%*

H8.02 *(a) b_1 = 5.02E-05 (8.27E-07), b_2 = 4488.4 (4.8), b_3 = 283.7 (0.4), RSC = 0.05915,*

D = 99.99%, MEP = 0.00723, AIC = -83.60. (b) b_1 = 5.269 (0.018), b_2 = 2699.3 (15.5),
b_3 = 346.3 (1.2), RSC = 2.02E-04, D = 100.00%, MEP = 2.09E-05, AIC = -174.51.

8.3.5 Analýza matematických modelů a fyzikálních dat

S8.01 *Model C: b_1 = 0.9975 (0.0068), b_2 = 0.906 (0.009), b_3 = 0.408 (0.019), RSC = 9.23E-05,*
D = 99.96%, MEP = 2.06E-05, AIC = -122.57.

S8.02 *b_1 = 0.477 (0.291), b_2 = 4.000 (0.000), b_3 = 0.351 (0.178), b_4 = 1.138 (1.308), RSC = 4.30E-08, D = 100.00%, MEP = 1.35E-07, AIC = -287.07.*

S8.03 *b_1 = 11237.6 (739435.9), b_2 = 3.329 (771.6), b_3 = 1131.2 (19229.1), b_4 = 9.839 (0.032).* **S8.04** *b_1 = 0.275 (0.066), b_2 = 0.605 (0.054), RSC = 0.841, D = 94.43%, MEP = 0.0874,*
AIC = -31.59.

S8.05 *b_1 = 1.719 (0.085), b_2 = -1.849 (0.060), RSC = 0.003374, D = 99.81%, MEP = 0.4841,*
AIC = -84.41.

S8.06 *Model C: b_1 = 10.783 (0.026), b_2 = -3.762 (0.027), b_3 = -4.287 (0.027), b_4 = -88.0 (0.027), b_5 = -55.46 (0.026), b_6 = -0.222 (0.0009), b_7 = 67.163 (0.026), RSC = 6.878, AIC = -97.27, D = 99.94%, MEP = 0.135.*

S8.07 *Model D: b_1 = 48.787 (6.267), b_2 = -0.203 (0.0295), b_3 = 0.2945 (0.0110), b_4 = 0.857 (0.104), b_5 = 33.989 (1.684), RSC = 8.821, D = 99.89%, MEP = 0.4705, AIC = -35.88.*

S8.08 *Model A: b_1 = -3.782 (2.934), b_2 = 67.94 (2.90), b_3 = 63.42 (4.90), RSC = 864.02,*
AIC = 152.14, D = 93.78%, MEP = 25.757.

S8.09 *b_1 = 4.710 (0.355), b_2 = 6.956 (0.292), b_3 = -1.774 (0.153), b_4 = 0.0169 (0.00075),*
RSC = 4078.54, AIC = 888.96, D = 68.12%, MEP = 11.48.

S8.10 *b_1 = -1973.1 (74.16), b_2 = 44.12 (0.42), b_3 = 0.977 (0.007), RSC = 5.41E-04,*
AIC = -1928.1, D = 99.99%, MEP = 3.72E-06.

S8.11 *b_1 = 1.555 (0.015), b_2 = 4.090 (0.047), b_3 = 451.5 (0.047), RSC = 1.46E-03,*
AIC = -346.87, D = 99.70%, MEP = 6.52E-05.

S8.12 *b_1 = 99.02 (0.54), b_2 = 0.011 (0.000), b_3 = 101.88 (0.59), b_4 = 107.03 (0.15), b_5 = 23.58 (0.23), b_6 = 72.04 (0.62), b_7 = 153.27 (0.19), b_8 = 19.53 (0.26), RSC = 1247.53, AIC = 417.86, D = 99.65%, MEP = 5.327.*

S8.13 *b_1 = -1.666E-01(3.830E-02), b_2 = 5.165E-03(6.662E-04), b_3 = 1.215E-02(1.530E-03),*
RSC = 5.1305E+02, s(e) = 3.1717E+00.

S8.14 *b_1 = 0.096 (0.001), b_2 = 1.006 (0.003), b_3 = 0.864 (0.002), b_4 = 3.008 (0.004), b_5 = 1.553 (0.002), b_6 = 5.002 (0.001), RSC = 2.23E-11, AIC = -652.9, D = 100.00%, MEP = 2.99E-12.*

S8.15 *b_1 = 0.375 (0.002), b_2 = 1.936 (0.220), b_3 = -1.465 (0.222), b_4 = 0.013 (0.000),*
b_5 = 0.022 (0.001), RSC = 5.46E-05, AIC = -429.26, D = 99.99%, MEP = 8.10E-06.

S8.16 *b_1 = 72.46 (1.73), b_2 = 2.62 (0.09), b_3 = 0.067 (0.003), RSC = 8.0565, AIC = 5.003,*
D = 99.83%, MEP = 2.095.

8.4 Doporučená literatura

[1] Endrenyi L. (ed.): *Kinetic Data Analysis.* Plenum Press, New York 1983, str. 47.

[2] Magel R. C., Hertsgaard D.: *Commun. Statist.* **16**, 85 (1987).

[3] Criado J. M. a kol.: *J. Thermal. Anal.* **29**, 243 (1984).

[4] Anscombe F. J.: *J. Royal Stat. Soc.* **B29**, 1 (1967).

[5] Gallant A. R.: *Nonlinear Statistical Models.* J. Wiley, New York 1987.

[6] Bard Y.: *Nonlinear Parameter Estimation.* Academic Press, New York 1974.

[7] Bates D. M., Watts D. G.: *J. Roy. Stat. Soc.* **B42**, 1 (1980).

[8] Nash J. C.: *J. Inst. Math. Applics.* **19**, 231 (1977).

[9] Hiebert K.: *ACM Trans Math. Software* **7**, 1 (1981).

[10] Kuester J. L., Mize J. N.: *Optimization Techniques in FORTRAN.* McGraw Hill, New York 1973.

[11] Wolfe M. A: *Numerical Methods for Unconstrained Optimization.* Van Nostrand, New York 1978.

[12] Gill P. d., Murray W., Wright M. M.: *Practical Optimization.* Academic Press, London 1981.

[13] Schmidt R.: *Advances in Nonlinear Parameter Optimization.* Springer, Berlin 1982.

[14] Nakagawa T., Oyanagi Y.: *Program System SALS for Nonlinear Least Squares Fitting*, ISE-TR-13. University of Tsukuba, Japan 1980.

[15] Rosenbrock M. M., Storey C.: *Computational Techniques for Engineers.* Pergamon Press, Oxford 1966.

[16] Powell M. D. J.: *Computer J.* **7**, 155 (1964).

[17] Spendley W., Hext G. R., Himworth F. R.: *Technometrics* **4**, 441 (1962).

[18] Nelder J. A., Mead R.: *Computer J.* **7**, 308 (1965).

[19] Routh M. W., Schwartz P. A., Denton M. B.: *Anal. Chem.* **49**, 1422 (1977).

[20] Ryan P. B., Barr P. L., Tod M. D.: *Anal. Chem.* **49**, 1461 (1977).

[21] Marsili-Libelli S., Castelli M.: *Appl. Mathematics and Comput.* **23**, 341 (1987).

[22] Volkov I. A., Grabov P. I., Potapov A. B.: *Zavod. Labor.* **5**, 60 (1985).

[23] Lindstrom F. T.: *Amer. Statist.* **34**, 183 (1980).

[24] Spendley W., in Fletcher R. (ed.): *Optimization.* Academic Press London 1969.

[25] Price W. L.: *J. Opt. Theor. Appl.* **40**, 333 (1983).

[26] Bokachevsky I. O. a kol.: *Technometrics* **28**, 209 (1986).

[27] Henckroth M. V. a kol.: *AICHE Journal* **22**, 744 (1976).

[28] Pronzato L. a kol.: *Math. and Computation Simulation* **26**, 412 (1984).

[29] Sillén L. G., Ingri N.: *Acta Chem. Scand.* **16**, 173 (1962).

[30] Meloun M., Čermák J.: *Talanta* **31**, 947 (1984).

[31] Peckham G., *Computer J*. **13**, 418 (1970).

[32] Ralston M. L., Jennrich R. I.: *Technometrics* **20**, 7 (1978).

[33] Dennis J. E., Gay D. M., Welsch R. E.: *ACM Trans. Math. Software* **7**, 348 (1981).

[34] Ramsin H., Wedin P.: *BIT* **17**, 72 (1977).

[35] Jennrich R. I. Sampson P. F.: *Technometrics* **10**, 63 (1968).

[36] Schmidt R.: *Advances in Nonlinear Parameter Optimization*. Springer-Verlag, Berlin 1982.

[37] Dennis J. E., Welsch R. E.: *Commun. Statist.* **B7**, 345 (1978).

[38] Gill P. E., Murray W.: *SIAM J. Numer. Anal.* **15**, 977 (1978).

[39] Chambers J. M.: *Biometrika* **60**, 1 (1973).

[40] Nash J. C.: *Compact Numerical Methods for Computers*, Adam Hilger LTD., Bristol 1979.

[41] Wharton M., Olson D. K.: *A generalized nonlinear least-squares fitting*, Program Rept. ORNL ITM-6545, Oak Ridge Natl. Lab., 1978.

[42] Moré J. J.: in *Lecture Notes in Mathematics* **630**, EE: D. Watson Springer-Verlag, Berlin 1978, str. 105.

[43] Linquist S. G.: *Proc. Conf. COMPSTAT 80*, Physica Verlag, Wien 1980.

[44] Meyer R. R., Roth D. M.: *J. Inst. Math. Applics* **9**, 218 (1973).

[45] Dennis J. E., Mei H. H. W.: *J. Opt. Theor. Appl.* **28**, 453 (1979).

[46] Militký J., Čáp J.: *Proc. Conf. CEF 87*, Taormina, Sicilia, May 1987.

[47] Beck J. V., Arnold K. J.: *Parameter Estimation in Engineering and Science*. J Wiley, New York 1977.

[48] Gallant A. R.: *J. Amer. Statist. Assoc.* **72**, 523 (1977)

[49] Demidenko E. Z.T.: *Linějnaja i nelinějnaja regresija*. Finansy i Statistika. Moskva 1981.

[50] Stanley G. M., Mah R. S. H.: *Chem Eng. Sci.* **36**, 259 (1981).

[51] Gorskij V. G.: *Zavod. Labor.* No. **1**, 50 (1987).

[52] Brown K. M., Dennis J. E.: *Numer. Math* **18**, 289 (1972) a Miller A. J. in McNeil D., ed.: *Interactive Statistics.* North Holland, Amsterdam 1979, str. 39.

[53] Ratkowsky D. A.: *Nonlinear Regression Modelling.* Marcel Dekker Inc., New York 1983.

[54] Lukšan L.: *SPONA - Soubor programů pro optimalizaci a nelineární aproximaci*. Výzkumná zpráva č. V-4, Centrální výpočetní středisko ČSAV, Praha 1976.

[55] James F., Ross M.: *Comp. Phys. Commun.*, **10**, 343 (1976).

[56] Bates D. M., Wats D. G.: *J. Roy. Stat. Soc.* **B42**, 1 (1986).

[57] Donaldson J. R., Schanabel R. B.: *Technometrics* **29**, 67 (1987).

[58] Cook R. D. a kol.: *Biometrika* **73**, 615 (1986).

[59] Box M. J.: *J. Roy. Stat. Soc.* **B32**, 171 (1971).

[60] Morton R.: *Biometrika* **74**, 679 (1987).

[61] Clarke G. P. Y.: *J. Amer. Statist. Assoc.* **82**, 221 (1987).

[62] Schwartz L.: *Anal. Chim. Acta* **122**, 291 (1980).

[63] Lyoness R. M.: *Commun. Statist.* **16**, 997 (1987).

[64] Himmelbau D. M.: *Process Analysis by Statistical Methods*. Wiley New York 1970.

[65] Meloun M., Havel J., Högfeldt E.: *Computation of Solution Equilibria*, Ellis Horwood, Chichester 1988.

[66] Hamilton W. C.: *Statistics in Physical Science*. Ronald Press, New York, 1964.

[67] Militký J. a kol.: *Proc. 2nd Int. Statist. Conference*, Tampere University Press, 1987.

[68] White H., Dorniwotz I.: *Ecometrica* **52**, 143 (1984).

[69] Hestens M. R., Stiefel E.: *J. Res. of the NBS*, **49**, 409 (1952).

[70] Dixon L. C. W.: *J. Inst. Math. Appl.* **15**, 9 (1975).

[71] Fletcher R.: *Comput J.* **13**, 317 (1970).

[72] Gill P. E., Murray W.: *J. Inst. Maths. Appl.* **9**, 91 (1972).

[73] Oren S. S., Luenberger D. G.: *Management Sci.* **20**, 845 (1974).

[74] Oren S. S.: *Management Sci.* **20**, 863 (1974).

[75] Huang H. Y., Chambliss J. P.: *J. Optimization Theory Appl.* **13**, 620 (1974).

[76] Bass R.: *Math. of Comp.* **26**, 129 (1972).

[77] Jacobson D. H., Oksma W.: *J. Math. Anal. Appl.* **38**, 535 (1972).

[78] Davison E. J., Wong P.: *Automatica* **11**, 297 (1975).

[79] Ritter K.: *Computing* **14**, 79 (1975).

[80] Davidon W. C.: *Math. Programming* **9**, 1 (1975).

[81] Swann W. H.: *Central Instrument Laboratory Research Note* **64**, 3 (1964).

[82] Powell M. J. D.: *Comput J.* **7**, 155 (1964).

[83] Zangwill W. I.: *Comput J.* **10**, 293 (1967).

[84] Brodlie K. W.: *J. Inst. Maths. Appl.* **15**, 385 (1975).

[85] Nazareth L.: *Res. Report LBL 2692*, University of California 1973.

[86] Mifflin R.: *Math. Programming* **9**, 100 (1975).

[87] Dennemeyer R. F., Mookini E. H.: *J. Optimization Theory Appl.* **16**, 67 (1975).

[88] Gill P. E., Murray: *Math. Programming* **7**, 311 (1974).

[89] Bosarge W. E., Falb P. L.: *J. Optimization Theory Appl.* **4**, 156 (1969).

[90] Fletcher R.: *Res. Report R-6799*, AERE Harwell, 1971.

[91] Meloun M., Javůrek M.: *Talanta* **32**, 973 (1985).

[92] Box G. P., W. G. Hunter, J. S. Hunter: *Statistics for Experimenters*. J. Wiley, New York, 1978. str. 483-487.

[93] Bennett L., Swartzendruber L., Brown H.: *Superconductivity Magnetization Modeling,* NIST 1994.

[94] Eckerle K.: *Circular Interference Transmittance Study,* NIST (1975).

[95] Rust B.: NIST 1996.

[96] Chwirut D.: *Ultrasonic Reference Block Study,* NIST (1975).

[97] Lanczos C.: *Applied Analysis.* Englewood Cliffs, NJ., Prentice Hall, 1956, str. 272-280.

[98] More J. J., Garbow B. S., Hillstrom K. E.: *Testing unconstrained optimization software. ACM Transactions on Mathematical Software.* 7(1), (1981), s. 17-41. v knize Osborne, M. R.: *Some aspects of nonlinear least squares calculations*. *In Numerical Methods for Nonlinear Optimization,* Lootsma (ed). Academic Press, New York 1972, s. 171-189.

[99] Ratkowsky D.A.: *Nonlinear Regression Modeling*. Marcel Dekker, New York 1983, str. 61-88.

[100] Meloun M., Militký J.: *Statistické zpracování experimentálních dat*. EAST PUBLISHING, Praha 1998.

9

INTERPOLACE A APROXIMACE

Interpolace a aproximace funkcí nebo experimentálních dat zahrnuje řadu technik. Účelem je provést náhradu funkce *f(x),* zadané hodnotami $\{x_i , y_i\}$, $i = 1, ..., n,$ vhodnou aproximující funkcí *g(x)*. Za aproximující funkci *g(x)* se často volí lineární kombinace *m*-tice elementárních funkcí $g_j(x)$

$$g(x) \quad ' \quad \sum_{j=1}^{m} c_j \, g_j(x) \quad .$$

Příkladem elementárních funkcí $g_j(x)$ jsou polynomy, racionální funkce, podíly polynomů, trigonometrické funkce, exponenciální funkce atd. Aproximující funkce souvisí se zadáním dané úlohy a ovlivňuje stupeň aproximace. Ten se obyčejně vyjadřuje jako vzdálenost mezi *aproximující funkcí g(x)* a *aproximovanou funkcí f(x)*, resp. diskrétními hodnotami y_i.

Zvláštním případem aproximace je *interpolace*: při interpolaci závislostí se sestrojuje funkce *g(x)* tak, aby procházela zadanými body $\{x_i, y_i\}$, $i = 1, ..., n,$ a splňovala přitom podmínky týkající se jejího tvaru.

Při *interpolaci funkcí* musí být v definovaných bodech ξ_i, $i = 1, ..., n$, nazvaných *uzlové body interpolace*, funkce *f(x)* a *g(x)* spojité ve funkčních hodnotách a hodnotách zvolených derivací

$$f^{(j)}(\xi_i) \quad ' \quad g^{(j)}(\xi_i), \qquad i \ ' \ 1, \ ..., \ n, \qquad j \ ' \ 0, \ ..., \ r_i \quad .$$

Zde $f^{(j)}$ označuje *j*-tou derivaci a r_i je maximální derivace v *i*-tém uzlu, ve které jsou obě, aproximovaná a aproximující funkce, totožné.

Interpolace se v technické praxi využívá pro

(a) *zespojitění tabelárních údajů*, například teplotní závislosti fyzikálně chemických konstant, jako jsou rozpustnost, hustota, iontový součin, relativní permitivita, součin rozpustnosti atd.;

(b) *náhradu složitých funkcí f(x)* nebo funkcí, které nelze přímo vyčíslit. Příkladem jsou Besselovy funkce, funkce Gamma, nekonečné řady atd.;

(c) *numerickou derivaci a integraci*;

(d) *kreslení grafu* závislosti zadané tabulkou.

Předpokladem interpolace závislosti jsou deterministické hodnoty souřadnice *x*

a jim odpovídající hodnoty na ose *y*. V praxi je však častější případ, kdy x_i jsou volené (nastavované) hodnoty (např. čas, teplota) a y_i jsou jim odpovídající experimentálně změřené hodnoty (např. koncentrace, absorbance, napětí, proud atd.). Experimentální hodnoty *y* jsou pak zatíženy náhodnými chybami.

Při aproximaci závislosti se předpokládá aditivní působení chyb typu

$$y_i = g(x_i) + \varepsilon_i \quad .$$

Pokud je druh funkce *g(x)* předem znám, přechází úloha aproximace na úlohu *(ne)lineární regrese*. Pokud se volí *g(x)* ve tvaru lineární kombinace elementárních funkcí, jde o úlohu *lineární regrese*. Rovněž úloha *aproximace funkce* se převádí na úlohu regrese, kde se však součty nahrazují integrály.

Aproximace se v technické praxi využívá k

(a) *vyhlazování závislostí*, tj. k eliminaci náhodných chyb ε_i. Pokud se data y_i pouze nahrazují hodnotami $g(x_i)$ a $x_{i+1} - x_i = \Delta$, $i = 1, ..., n - 1$, jde o úlohu *číslicové filtrace*. Vyhlazení se užívají k určení vyhlazených hodnot *g(x)* nebo ke kreslení grafů;

(b) *náhradě rozsáhlých souborů dat* funkcemi, obsahujícími méně parametrů, k účelům uchování informací o datech, např. v paměti počítače;

(c) *numerické derivaci a integraci* experimentálních dat, zatížených náhodnými chybami;

(d) *tvorbě speciálních empirických modelů* regresního typu, jako je spline-regrese.

V řadě technických úloh je interpolace a aproximace dílčí části postupu zpracování dat. V této kapitole jsou uvedeny vybrané, nejvíce užívané techniky aproximace a interpolace funkcí, resp. závislostí. Vedle klasických postupů jsou uvedeny i postupy, které využívají *po částech definovaných funkcí* (piecewise-funkcí).

9.1 Klasické interpolační postupy

Úlohou interpolace je nalezení funkce *g(x)* tak, aby pro *n* hodnot $x_1 < x_2 < ... < x_n$ (v případě závislostí), nebo pro *n* uzlů $\xi_1 < \xi_2 < ... < \xi_n$ (v případě funkcí) platila úvodní rovnice. Protože platí $x_i = \xi_i$, budeme označovat uzly také symbolem x_i. Mezi nejznámější postupy patří ***polynomická interpolace***, která hledá polynom *g*(x), splňující úvodní podmínku $f^{(j)}(\xi_i) = g^{(j)}(\xi_i)$. Hledaný polynom je stupně nejvýše

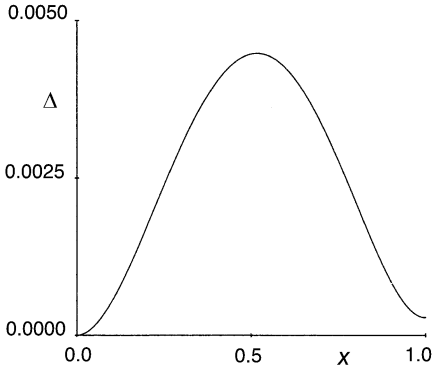
$$m = \sum_{i=1}^{n} r_i - n - 1 .$$

Pokud je požadavkem shoda pouze ve funkčních hodnotách, jsou $r_i = 0$, $i = 1, ..., n$, a *n*-tice bodů je interpolována jednoznačně polynomem (n - 1)ního stupně. Z úvodní podmínky se sestaví *m* lineárních rovnic, ze kterých se vypočtou odpovídající koeficienty c_j .

Vzorová úloha 9.1 *Náhrada funkce exp(x)*
Nalezněte interpolační polynom, který aproximuje funkci exp(x) v intervalu {0, 1} tak, že v krajních bodech $x_1 = 0$ a $x_2 = 1$ souhlasí s touto funkcí ve funkčních hodnotách a hodnotách prvních derivací.
Řešení: Určíme stupeň interpolačního polynomu $m = 2 + 2 - 1 = 3$. Koeficienty c_1, c_2 , c_3 a c_4 interpolačního polynomu $g(x) = c_1 + c_2x + c_3x^2 + c_4x^3$ určíme po dosazení do rovnice podmínky. Na základě zadání lze úvodní rovnici vyjádřit ve tvaru

$\exp(0) = 1 = c_1$,
$\exp'(0) = 1 = c_2$,
$\exp(1) = 2.718 = c_1 + c_2(1) + c_3(1)^2 + c_4(1)^3$
a $\exp'(1) = 2.718 = c_2 + 2\,c_3(1) + 3\,c_4(1)^2$,

kde $\exp'(x) = \mathrm{d}\exp(x)/\mathrm{d}x$ je první derivace funkce exp(x). Z prvních dvou rovnic jsou $c_1 = 1$ a $c_2 = 1$. Zbylé koeficienty se vyčíslí ze dvou lineárních rovnic o dvou neznámých $0.718 = c_3 + c_4$, $1.718 = 2c_3 + 3c_4$. Řešením je $c_3 = 0.436$ a $c_4 = 0.282$. Interpolační polynom má pak tvar $g(x) = 1 + x + 0.436\,x^2 + 0.282\,x^3$.



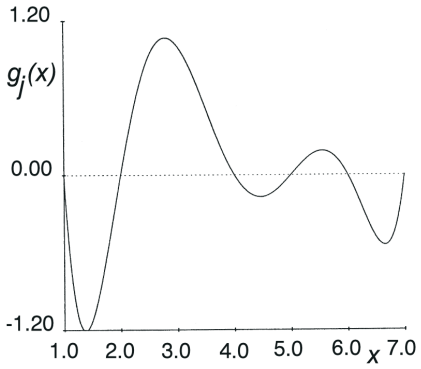
Obr. 9.1 Graf chyby Δ aproximace funkce exp(x) polynomem g(x).

Na obr. 9.1 je pro tento polynom znázorněn graf chyby aproximace $\Delta = \exp(x) - g(x)$.
Závěr: Úloha interpolace zde vede na úlohu hledání řešení soustavy lineárních rovnic.

9.1.1 Lagrangeova a Newtonova interpolační formule

Formule se užívají pro případ $r_i = 0$, kdy se konstruuje polynom stupně nejvýše $m = n - 1$, interpolující n uzlových bodů, a kdy platí $y_i = f(x_i) = g(x_i)$. Interpolační polynom splňující tyto podmínky lze vyjádřit jako lineární kombinaci všech y hodnot $L_m(x) = \sum_{i=1}^{n} y_i\, g_j(x)$,

kde $g_j(x)$ jsou polynomy stupně $(n - 1)$ takové, že $g_j(x_i) = 0$ pro všechna $j \neq i$, $g_j(x_j) = 1$. Tyto podmínky zajišťují, že $L_m(x)$ je interpolační polynom $(n - 1)$. stupně. Funkce $g_j(x)$ je znázorněna na obr. 9.2.



Obr. 9.2 Jednoduchý Lagrangeův polynom.

Polynomy g_j(x), splňující tyto podmínky, lze vyjádřit ve tvaru

$$g_j(x) = w_j (x - x_1)(x - x_2) \dots (x - x_{j-1})(x - x_{j+1}) \dots (x - x_n) =$$

$$= w_j \prod_{\substack{i=1 \\ i \neq j}}^{n} (x - x_i).$$

kde normalizační koeficient w_j je roven $w_j = \dfrac{1}{\prod_{\substack{i=1 \\ i \neq j}} (x_j - x_i)}$. Lagrangeův interpolační polynom má potom tvar

$$L_m(x) = \sum_{j=1}^{n} y_j w_j \prod_{\substack{i=1 \\ i \neq j}}^{n} (x - x_i) = \sum_{j=1}^{n} y_j \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{x - x_i}{x_i - x_j} .$$

Tato formulace Lagrangeova interpolačního polynomu se hodí pouze pro malá n a jednoduché ruční výpočty. Pro výpočty s využitím počítače se doporučuje tzv. *barycentrická reprezentace* ve tvaru

$$L_m(x) = \frac{\sum_{j=1}^{n} y_j \dfrac{w_j}{x - x_j}}{\sum_{j=1}^{n} \dfrac{w_j}{x - x_j}} \quad \text{pro} \quad x \neq x_j ,$$

$$L_m(x) = y_j \quad \text{pro } x = x_j .$$

Tato reprezentace interpolačního polynomu je numericky stabilní a navíc lze pro různá dělení uzlů x_i , $i = 1, ..., n$, určit normalizační koeficienty w_i analyticky[2].

V případě, že byl Lagrangeův interpolační polynom použit pro interpolaci n-tice funkčních hodnot $f(x_i)$ známé funkce $f(x)$, platí pro chybu interpolace v libovolném bodě

x vztah

$$f(x) - L_{n-1}(x) = \frac{f^{(n)}(\alpha)}{n!} \prod_{j=1}^{n} (x - x_j) \quad ,$$

kde $x_1 < \alpha < x_n$. Nevýhodou tradičního vyjádření interpolačního polynomu v Lagrangeově tvaru je požadavek na opětovné přepočítání všech členů při přidání dalšího bodu x_{n+1}, y_{n+1}. Z tohoto hlediska je při postupném přidávání uzlů výhodnější *Newtonova interpolační formule*, pro kterou platí

$$P_m(x) = \sum_{j=1}^{n} a_j \prod_{k=1}^{j-1} (x - x_k) \quad .$$

Přidání bodu x_{n+1}, y_{n+1} pak vede k interpolačnímu polynomu

$$P_{m+1}(x) = P_m(x) + a_{n+1} \prod_{k=1}^{n} (x - x_k) \quad .$$

Z této rovnice vychází při dosazení za $x = x_j$ pro koeficient a_{n+1} vztah

$$a_{n+1} = \sum_{i=1}^{n+1} w_i y_i = \sum_{i=1}^{n+1} \frac{y_i}{\prod_{\substack{j=1 \\ j \neq i}}^{n+1} (x_i - x_j)} \quad .$$

Pro ostatní koeficienty a_k, $k = 1, 2, ..., n$ polynomu P_{m+1} platí

$$a_k = \sum_{i=1}^{k} w_i y_i \prod_{j=k+1}^{n+1} (x_i - x_j) \quad .$$

Koeficienty a_k se nazývají postupné diference funkce *f(x)* a nultá postupná diference je $a_1 = y_1$. Ostatní postupné diference lze definovat rekurentně. Pro první postupnou diferenci platí

$$a_2 = \frac{y_1}{x_1 - x_2} + \frac{y_2}{x_2 - x_1} = \frac{y_2 - y_1}{x_2 - x_1} = [x_2, x_1]f \quad .$$

Pro druhou postupnou diferenci je

$$a_3 = \frac{[x_3, x_2]f - [x_2, x_1]f}{x_3 - x_1} = \frac{\frac{y_3 - y_2}{x_3 - x_2} - \frac{y_2 - y_1}{x_2 - x_1}}{x_3 - x_1}$$

a pro i-tou postupnou diferenci a_{i+1} platí

$$a_{i+1} = \frac{[x_{i+1}, \ ..., \ x_2]f - [x_i, \ ... \ ,x_1]f}{x_{i+1} - x_1} \quad ,$$

$$\begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_{n\&1} \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{j=2}^{n} (x_1 \,\&\, x_j) & 0 & . & . & . & . & 0 \\ \sum_{j=3}^{n} (x_1 \,\&\, x_j) & \sum_{j=3}^{n} (x_2 \,\&\, x_j) & . & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ (x_1 \,\&\, x_n) & (x_2 \,\&\, x_n) & . & . & . & . & 0 \\ 1 & 1 & . & . & . & . & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ . \\ . \\ . \\ d_{n\&1} \\ d_n \end{bmatrix}.$$

Při sestavování postupných diferencí se s výhodou sestavuje tabulka, jejíž diagonálu tvoří koeficienty a_j. Tak jsou definovány vztahy mezi koeficienty a_k Newtonovy formule a koeficienty $w_k y_k = d_k$ Lagrangeovy formule. Při znalosti koeficientů a_k můžeme stanovit koeficienty d_k řešením soustavy n lineárních rovnic[3].

Postup efektivního výpočtu d_i , $i = 1, ..., n$, z této soustavy rovnic je popsán v práci[3]. Na jeho základě byl odvozen algoritmus pro efektivní výpočet normalizačních koeficientů w_j , $j = 1, ..., n$, který lze vyjádřit posloupností vztahů:

$$\begin{aligned} &a_1^{(1)} = 1, \; a_k^{(1)} = 0, && k = 2, ..., n, \\ &a_k^{(i)} = a_k^{(i-1)}/(x_k - x_i), && k = 1, 2, ..., i - 1, \\ &a_i^{(k+1)} = a_i^{(k)} - a_k^{(i)}, && i = 2, 3, ..., n, \\ &w_i = a_i^{(n)}. \end{aligned}$$

Užitím tohoto schématu je počet operací potřebných pro vyčíslení polynomu v Lagrangeově tvaru shodný s počtem operací pro vyčíslení koeficientů v Newtonově tvaru[3].

Vytváření postupných diferencí

Data	První diference	Druhé diference	... (n-1)ní diference
$x_1\, y_1$			
	$[x_2, x_1]f$		
$x_2\, y_2$		$[x_3, x_2, x_1]f$	
	$[x_3, x_2]f$		
$x_3\, y_3$		$[x_4, x_3, x_2]f$	
	$[x_4, x_3]f$		
$x_4\, y_4$			
.	.		$[x_n, x_{n-1}, ..., x_1]f$
$x_n\, y_n$			

Vzorová úloha 9.2 *Náhrada funkce exp(x)*

Nalezněte interpolační polynom, který aproximuje funkci exp(x) a prochází uzly o hodnotách $x_1 = 0$, $x_2 = 0.5$ a $x_3 = 1$.

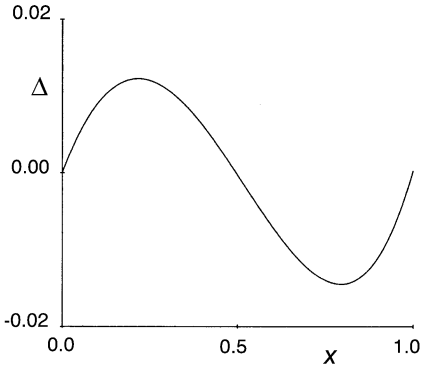
Řešení: Pro určení interpolačního polynomu druhého stupně využijeme Newtonovy formule.

		Postupné diference	
x_i	y_i	První diference	Druhá diference
0	1		
		1.298	
0.5	1.649		0.84
		2.138	
1	2.718		

Postupné diference jsou v tabulce. Hledaný interpolační polynom má potom tvar

$$P(x) = 1 + 1.298(x - 0) + 0.84(x - 0)(x - 0.5) =$$
$$= 1 + 0.878\,x + 0.84\,x^2.$$

Průběh chyby aproximace Δ funkce *f(x)* tímto interpolačním polynomem je znázorněn na obr. 9.3.



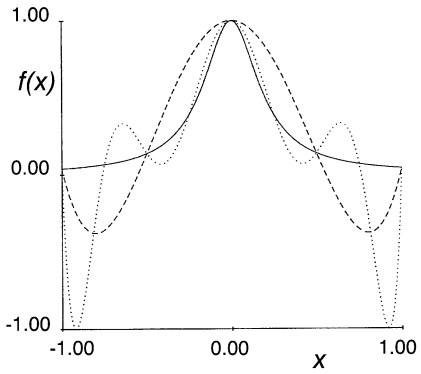
Obr. 9.3 Chyba Δ aproximace funkce exp(x) polynomem *P(x)*.

Závěr: Použití Newtonovy interpolační formule je s využitím tabulky proměnných diferencí jednoduché. Přidání dalšího bodu znamená pouze přidání dalšího členu do interpolační formule.

Vzorová úloha 9.3 *Aproximace racionální funkce*

Aproximujte funkci $f(x) = 1/(1 + 25\,x^2)$ interpolačními polynomy *L(x)* na intervalu {-1, 1} při volbě $n = 10$ (polynom stupně $m = 9$) a $n = 16$ (polynom stupně $m = 15$). Dělení uzlových bodů volte ekvidistantní.

Řešení: Programem ADSTAT byly určeny oba interpolační polynomy, které jsou spolu se skutečným průběhem *f(x)* znázorněny na obr. 9.4.



Obr. 9.4 Interpolace funkce *f(x)* (plnou čarou) polynomy 9. (čárkovaně) a 15. (tečkovaně) stupně.

Závěr: Použití interpolačních polynomů vyšších stupňů nemusí ještě znamenat zpřesnění aproximace funkce *f(x).*

9.1.2 Hermitovská interpolace

Při této interpolaci se požaduje, aby interpolační polynom H_m se svou první derivací souhlasil ve všech uzlových bodech s danou funkcí a její první derivací. To znamená, že $r\ i = 1$, $i = 1, ..., n$ a interpolační polynom je stupně $(2\ n - 1)$. Označíme-li hodnoty derivací v uzlových bodech x_i jako y'_i, můžeme psát

$$H_m(x) = \sum_{i=1}^{n} y_i\ h_i(x) + \sum_{i=1}^{n} y_i'\ \bar{h}_i(x) \quad ,$$

kde $h_i(x) = [1 - 2(x - x_i)\ g_i'(x_i)]\ g_i^2(x)$ a $\bar{h}_i(x) = (x - x_i)\ g_i^2(x)$. V těchto vztazích jsou $g_i(x)$ elementární Lagrangeovy polynomy. Pro Hermitovskou interpolaci je vhodné použít barycentrického tvaru

$$H_m(x) = \frac{\sum_{i=1}^{n} \left[\frac{w_i}{x - x_i} \left(\frac{w_i}{x - x_i} - v_i \right) y_i + w_i \frac{w_i}{x - x_i}\ y_i' \right]}{\sum_{i=1}^{n} \frac{w_i}{x - x_i} \left(\frac{w_i}{x - x_i} - v_i \right)} \quad \text{pro} \quad x \neq x_j$$

$$H_m(x) = y_j \quad \text{pro} \quad x = x_j \quad .$$

Pro koeficienty v_i platí[2] $v_i = 2\, w_i \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{x_i - x_j}$, $i = 1, 2, ..., n$. V práci[3] je uveden efektivní algoritmus simultánního postupného určování koeficientů w_i a v_i.

Vzorová úloha 9.4 *Hermitovská interpolace funkce exp(x)*
Řešte vzorovou úlohu 9.1 využitím Hermitovské interpolační formule.

Řešení: Plyne, že $g_1(x) = \frac{x - 1}{0 - 1} = 1 - x, \quad g_1'(x) = -1,$

$$g_2(x) = \frac{x}{1 - 0} = x, \quad g_2'(x) = 1$$

a dále platí $h_1(x) = [1 - 2x]\,(x - 1)^2, \quad h_2(x) = [1 + 2(x - 1)]\,x^2$ a

$$\bar{h}_1(x) = x\,(x - 1)^2, \quad \bar{h}_2(x) = (x - 1)\,x^2 .$$

Po přímém dosazení vychází

$$H_3(x) = [1 + 2x]\,(1 - x)^2 + 2.718\,[1 - 2(x - 1)]\,x^2 +$$
$$+ x(1 - x)^2 + 2.718\,(x - 1)\,x^2 = 1 + x + 0.436\,x^2 + 0.282\,x^3.$$

Závěr: Polynom $H_3(x)$ je pochopitelně totožný s polynomem $g(x)$, nalezeným ve vzorové úloze 9.1.

9.1.3 Racionální interpolace

Při této aproximaci je interpolující funkce $R_{m,l}(x)$ definována jako podíl polynomu stupně m (v čitateli) a polynomu stupně l (ve jmenovateli)

$$R_{m,l}(x) = \frac{P_m(x)}{P_l(x)} .$$

Tato aproximace nahrazuje klasickou polynomickou interpolaci stupně $(m + 1)$. S výhodou se používá *racionální aproximace typu Padé*

$$R(x) = b_1 + \cfrac{x - x_1}{b_2 + \cfrac{x - x_3}{b_4 + \cfrac{x - x_5}{b_6 + \ldots}}} .$$

Místo tohoto zápisu se používá i zkrácená forma

$$R(x) = b_1 + \frac{x - x_1}{b_2} \; \frac{x - x_2}{+ b_3} \; \ldots \; \frac{x - x_{n-1}}{+ b_n} .$$

Pro určení koeficientů $b_1, \ldots, b_n$ tak, aby $R(x)$ interpolovala zadanou funkci v n uzlech, se používá rekurentních formulí

$$
\begin{aligned}
R_1(x) &= b_1 + \frac{x - x_1}{R_2(x)} \\
R_2(x) &= b_2 + \frac{x - x_2}{R_3(x)} \\
&\vdots \\
R_i(x) &= b_i + \frac{x - x_i}{R_{i+1}(x)} \\
&\vdots \\
R_n(x) &= b_n .
\end{aligned}
$$

Za předpokladu, že $R_{i+1}(x) \neq 0$, dostáváme, že

$$b_i = R_i(x_i), \quad \text{pro } i = 1, \ldots, n \ .$$

Pro interpolaci musí ještě platit omezení, že

$$R_1(x_i) = y_i, \quad i = 1, \ldots, n \ .$$

Z této rovnice přímo plyne, že $b_1 = y_1$. Z předešlých rovnic lze také určit, že platí

$$R_{j+1}(x_i) = \frac{x_i - x_j}{R_j(x_i) - b_j}, \ i = j + 1, \ j + 2, \ldots, n \ .$$

Využitím této rovnice lze pro $j = 1, 2, \ldots, n - 1$, počítat $R_{j+1}(x_i)$ pro $i = j + 1, j + 2, \ldots, n$, a určovat $b_{j+1} = R_{j+1}(x_{i+1})$. Tímto postupem se jednoduše určí koeficienty Padé interpolace. Pokud vyjde $R_{j+1}(x_j)$ pro $j = 1, \ldots, n - 2$ rovno nule, nelze tento postup použít.

Vzorová úloha 9.5 *Racionální interpolace funkce exp(x)*

Nalezněte racionální interpolační polynom, který aproximuje funkci $\exp(x)$ a prochází uzly o hodnotách $x_1 = 0$, $x_2 = 0.5$ a $x_3 = 1$.

Řešení: Podle uvedeného postupu je

$$
\begin{aligned}
R_1(x_1) &= b_1 = 1, \\
R_1(x_2) &= 1.649, \\
R_1(x_3) &= 2.718,
\end{aligned}
$$

$$b_2 = R_2(x_2) = \frac{0.5 - 0}{1.649 - 1} = 0.7704,$$

pak

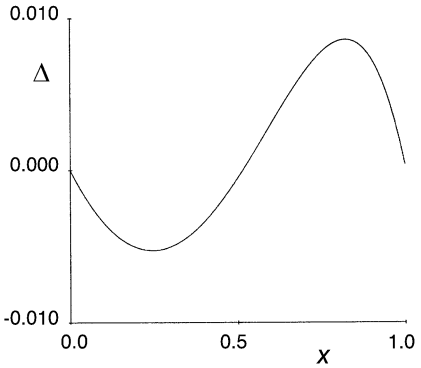
$$R_2(x_3) = \frac{1 - 0}{2.718 - 1} = 0.58207$$

a konečně $$b_3 = R_3(x_3) = \frac{1 - 0.5}{0.582 - 0.77} = -2.66 \ .$$

Po úpravě vyjde

$$R(x) = \frac{b_1\ b_2\ b_3 + b_3\ (x - x_1) + b_1\ (x - x_2)}{b_2\ b_3 + (x - x_2)} =$$

$$= \frac{1.659\,x + 2.5478}{2.5478 - x}.$$



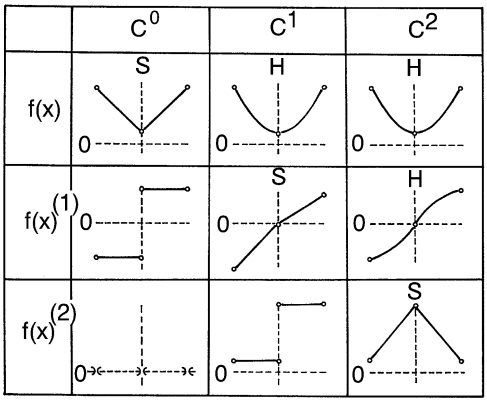
Obr. 9.5 Graf chyby aproximace funkce exp(*x*) racionální lomenou funkcí.

Závěr: Na obr. 9.5 je znázorněn průběh chyby aproximace $\Delta = R(x) - \exp(x)$, vystihující kvalitu provedené aproximace.

9.2 Spline interpolace

Užívání polynomiálních interpolačních formulí má řadu nevýhod. Jsou totiž složeny z elementárních funkcí definovaných na celé reálné ose, což vede u interpolačních formulí vyšších řádů ke vzniku řady lokálních minim, maxim a inflexních bodů, které neodpovídají průběhu funkce *f(x)* či tabelované závislosti $\{x_i, y_i\}$, $i = 1, ..., n$. Při interpolaci fyzikálních závislostí se stává, že chování v jistém intervalu se výrazně liší od jejich chování v intervalech sousedních. Jde o závislost tzv. ***neasociativní povahy***. Z těchto úvah plyne, že pro účely interpolace, ale i aproximace, bude výhodnější volit lokálně definované funkce, které budou v místech vzájemného styku, tj. v uzlech, spojité ve funkčních hodnotách a hodnotách zadaných derivací.

Vhodné interpolační funkce tohoto typu jsou složeny z polynomických úseků a platí pro ně, že jsou ze třídy $C^m[a, b]$. Obecně jsou funkce třídy $C^m[a, b]$ na intervalu $[a, b]$ spojité v prvních m derivacích a funkčních hodnotách. Na obr. 9.6 jsou schematicky znázorněny funkce třídy C^0, C^1, C^2 a odpovídající první a druhé derivace. Z obr. 9.6 plyne, že hladké jsou všechny funkce od třídy C^1. Pro funkce třídy C^m platí, že m-tá derivace je lineární lomená závislost, $(m + 1)$ derivace je po částech konstantní a $(m + 2)$ derivace je po částech nulová, tj. není definovaná v uzlových bodech ξ_i.



Obr. 9.6 Příklady funkcí C^0, C^1, C^2 a jejich derivací: H značí hladká, S značí spojitá křivka.

Využitím uvedených vlastností funkcí ze třídy $C^m[a, b]$ můžeme definovat obecně polynomický spline $S_m(x)$ s uzly $a = \xi_1 < \xi_2 < \xi_3 < ... < \xi_n = b$. Tento spline je na každém úseku $[\xi_j, \xi_{j+1}], j = 1, ..., n - 1$, reprezentován polynomem maximálně m-tého stupně. Pokud je v nějakém bodě x_i některá derivace $S_m^{(l)}(\xi_i)$ nespojitá, jde o ***defektní spline***. Vlastnosti spline $S_m(\xi_i)$ závisí na

(a) řádu polynomu m, přičemž se obvykle volí kubický spline $m = 3$;
(b) počtu a polohách uzlů $\xi_1 < \xi_2 < ... < \xi_n$;
(c) defektech v uzlových bodech.

Z defektních spline se omezíme na klasické spline, které mají *minimální defekt* roven $k = 1$, tj. patří do třídy $C^{m-1}[a, b]$. Pro účely Hermitovské interpolace je výhodné použít spline defektu $k = 2$, který patří do třídy $C^{m-2}[a, b]$. Defekt $k = 1$ umožňuje zadat podmínky interpolace a defekt $k = 2$ ještě navíc podmínky týkající se hodnot prvních derivací.

Klasické spline polynomy $S_m(x)$ ze třídy $C^{m-1}[a, b]$ je možno definovat několika způsoby. Nejjednodušší je pro každý interval $I_j \in [\xi_{i+1}, \xi_i], j = 1, 2, ..., n - 1$, definovat lokální polynom $P_j(x) = c_0 + \sum_{k=1}^{m} c_k (x - \xi_i)^k$, pro $x \in I_j$. Tento zápis je však redundantní, protože $P_j(x)$ obsahuje v každém $(m + 1)$ intervalu parametrů c_k a celkově pak $n (m + 1)$ parametrů.

Nejefektivnější je vyjádření spline $S_m(x)$ jako lineární kombinace bázových B-spline s minimální podporou. *Bázový B-spline* $B_{m,j}$ je definován v $(m + 1)$ uzlových bodech $\xi_{j-m} < \xi_{j-m+1} < ... < \xi_j$ jako normalizovaná m-tá poměrná diference useknutého polynomu $g(\xi) = (\xi - x)_+^{m-1}$. Využitím formálního zápisu postupné diference můžeme psát $B_{m,j} = (\xi_j - \xi_{j-m}) [\xi_{j-m}, ..., \xi_j] g$. V praxi se pro výpočet normalizovaných B-spline používá rekurentní formule

$$B_{m,j}(x) = \frac{x - \xi_{j-m}}{\xi_{j-1} - \xi_{j-m}} B_{m-1,j-1}(x) + \frac{\xi_j - x}{\xi_j - \xi_{j-m+1}} B_{m-1,j}(x) \quad .$$

Začíná se od $B_{1,j}(x)$, pro které platí

$$B_{1,j}(x) = \begin{cases} 1 & \text{pro } \xi_{j\&1} \# x \# \xi_j \\ 0 & \text{jinde} \end{cases} .$$

Bázové *B*-spline *m*-tého řádu mají zajímavé vlastnosti:

(a) jsou kladné pouze v intervalu $\xi_{j-m} < x < \xi_j$ a všude jinde nabývají nulových hodnot,

$$\begin{aligned} B_{m,j}(x) &> 0 \quad \text{pro } \xi_{j\&m} < x < \xi_j \\ B_{m,j}(x) &= 0 \quad \text{jinde} \end{aligned} ;$$

(b) jsou normalizované, tj. $\sum_{(j)} B_{m,j}(x) = 1$ pro všechna $\xi_1 < x < \xi_n$;

(c) na intervalu (ξ_{j-m}, ξ_j) je $B_{m,j}(x)$ spline polynom stupně $(m - 2)$ s uzlovými body $(\xi_{j-m}, \xi_{j-m+1}, \ldots \xi_j)$. To znamená, že v každém intervalu je mezi dvojicí uzlových bodů vyjádřen spoj polynomem stupně maximálně $(m - 1)$ a patří do třídy funkcí $C^{m-2}[\xi_{j-m}, \xi_j]$. Tato vlastnost platí, pokud jsou všechny uzlové body ξ_j navzájem různé.

Vzorová úloha 9.6 *Lineární B-spline*

Odvoďte tvar *B*-spline $B_{2,j}$ a zakreslete všechny spline patřící do intervalu $\langle \xi_{j-2}, \xi_j \rangle$.

Řešení: Z definice je patrné, že $B_{2,j}$ je definováno na intervalu $\xi_{j-2} < \xi_{j-1} < \xi_j$ vztahem

$$B_{2,j} = \frac{(\xi_j - x)_+}{\xi_j - \xi_{j\&1}} \cdot \frac{(x - \xi_{j\&1})_+}{\xi_j - \xi_{j\&1}} + \frac{(\xi_{j\&1} - x)_+}{\xi_{j\&1} - \xi_{j\&2}} \cdot \frac{(\xi_{j\&2} - x)_+}{\xi_{j\&1} - \xi_{j\&2}} =$$

$$= \frac{(\xi_{j\&2} - x)_+}{\xi_{j\&1} - \xi_{j\&2}} \cdot \left[\frac{1}{\xi_{j\&1} - \xi_{j\&2}} - \frac{1}{\xi_j - \xi_{j\&1}}\right] (\xi_{j\&1} - x)_+ - \frac{(\xi_j - x)_+}{\xi_j - \xi_{j\&1}} .$$

Při určení $B_{2,j}(x)$ bylo použito rekurentního vztahu pro postupné diference

$$\left[\xi_{j\&2}, \xi_{j\&1}, \xi_j\right] g = \frac{[\xi_{j\&1}, \xi_j]\, g - [\xi_{j\&2}, \xi_{j\&1}]\, g}{\xi_j - \xi_{j\&2}} .$$

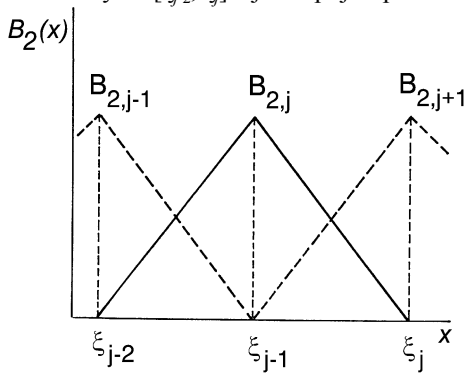
Přepišme si $B_{2,j}(x)$ jako funkce definované v intervalu $I_{j-1} = (\xi_{j-2}, \xi_{j-1})$ a $I_j = (\xi_{j-1}, \xi_j)$. Pro interval I_{j-1} platí

$$B_{2,j}(x) = 1 - \frac{x}{\xi_{j\&1} - \xi_{j\&2}} + \frac{\xi_{j\&1}}{\xi_{j\&1} - \xi_{j\&2}}, \quad \text{pro } \xi_{j\&2} \# x \# \xi_{j\&1} .$$

Jde o rovnici přímky nabývající v místě ξ_{j-2} hodnoty $B_{2,j}(\xi_{j-2}) = 0$ a v místě ξ_{j-1} hodnoty $B_{2,j}(\xi_{j-1}) = 1$. Pro interval I_j platí

$$B_{2,j}(x) = \frac{\xi_j}{\xi_j - \xi_{j\&1}} - \frac{x}{\xi_j - \xi_{j\&1}}, \quad \text{pro } \xi_{j\&1} \# x \# \xi_j .$$

Jde také o rovnici přímky nabývající v místě ξ_{j-1} hodnoty $B_{2,j}(\xi_{j-1}) = 1$ a v místě ξ_j hodnoty $B_{2,j}(\xi_j) = 0$. Ve smyslu definice je tedy $B_{m,j}$ pro $m = 2$ spline polynom $S_1(x)$ určený polynomy prvního stupně, které jsou ze třídy C^0 $[\xi_{j-2}, \xi_j]$ a jsou spojité pouze ve funkčních hodnotách.



Obr. 9.7 Elementární lineární B-spline.

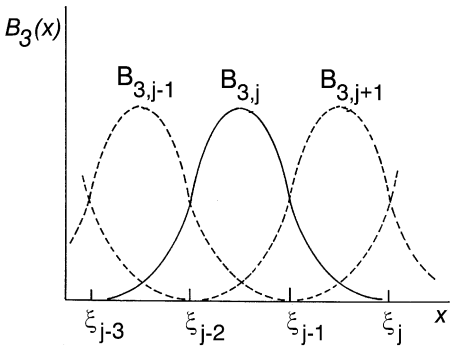
Na obr. 9.7 je znázorněn *B*-spline $B_{2,j}(x)$ spolu se sousedním *B*-spline, které jsou nenulové na intervalu ξ_{j-2}, ξ_j.

Závěr: Konstrukce *B*-spline je pro nízké hodnoty *m* možná přímo z definice. Při použití počítače je však výhodnější rekurentní formule.

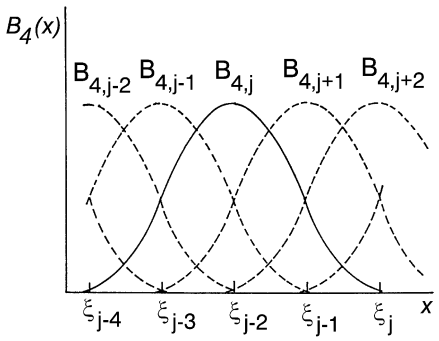
Obdobně lze odvodit vztahy pro kvadratická $B_{3,j}$ a kubická $B_{4,j}$ spline, které jsou pro praktické účely nejpoužívanější:

(1) Pro *kvadratické spline* $B_{3,j}$ platí, že jsou definovány na čtyřech uzlových bodech ξ_{j-3}, ξ_{j-2}, ξ_{j-1}, ξ_j jako parabolické úseky spojité v těchto uzlových bodech ve funkčních hodnotách a hodnotách první derivace. Na každém intervalu $+\xi_{j-1}$, ξ_j, jsou nenulová pouze tři *B*-spline $B_{3,j}$, $B_{3,j+1}$ a $B_{3,j+2}$.

(2) Pro *kubická spline* $B_{4,j}$ platí, že jsou definována na pěti uzlových bodech ξ_{j-4}, ξ_{j-3}, ξ_{j-2}, ξ_{j-1}, ξ_j, a to jako kubické úseky spojité v těchto uzlových bodech ve funkčních hodnotách a hodnotách prvních dvou derivací.



Obr. 9.8 Elementární kvadratické spline $B_3(x)$.



Obr. 9.9 Elementární kubické spline $B_4(x)$.

Pomocí normalizovaných *B*-spline lze vyjádřit na intervalu [*a, b*] s uzlovými body $a = \xi_1 < ... < \xi_n = b$ spline $S_m(x)$ v jednoduchém tvaru

$$S_m(x) = \sum_{j=1}^{n+m-1} c_j \, B_{m,j}(x) \quad .$$

Vzhledem k tomu, že v intervalu $\langle \xi_j, \xi_{j+1} \rangle$ je právě *m* bázových *B*-spline nenulových, je nutné pro úplnou definici všech bázových spline ještě definovat na každé straně intervalu [*a, b*] celkem *m* přídavných uzlových bodů $\xi_{-m+1} \le \xi_{-m+2} \le ... \le \xi_0$ a $\xi_{n+1} \le \xi_{n+2} \le ... \le \xi_{n+m}$, cit[7]. *B*-spline reprezentace je výhodná jak pro úlohy interpolace, tak i pro úlohy spline regrese.

9.2.1 Lokální Hermitovská interpolace

Pro účely rekonstrukce závislostí z daných tabelárních hodnot $\{x_i, y_i\}$, $i = 1, ..., n$, se používá kubických spline $S_3^2(x)$ defektu $k = 2$. Jde o lokální kubické polynomy, které jsou v uzlových bodech spojité v hodnotách funkce a první derivace. Patří tedy do třídy $C^1[a, b]$. Protože jsou zde uzlové body interpolace totožné se zadanými souřadnicemi na ose *x* rekonstruované závislosti, tj. $x_i = \xi_i$, $i = 1, ..., n$, budeme používat označení x_i i pro uzly interpolace. Spline $S_3^2(x) = P(x)$ lze vyjádřit pro $x_i \le x \le x_{i+1}$ ve tvaru $P(x) = c_1 + (x - x_i)\, c_2 + (x - x_i)^2\, c_3 + (x - x_i)^3\, c_4$. Konstanty c_1 až c_4 se určí z podmínek interpolace $P(x_i) = y_i$, $P(x_{i+1}) = y_{i+1}$, a podmínek spojitosti v první derivaci $P^{(1)}(x_i) = d_i$, $P^{(2)}(x_{i+1}) = d_{i+1}$. Při znalosti hodnot prvních derivací d_i, $i = 1, ..., n$, lze pak konstanty c_1 až c_4 určit z jednoduchých vztahů

$$c_1 = y_i, \quad c_2 = d_i \ , \quad c_3 = \frac{3\,\Delta_i - d_{i+1} - 2\,d_i}{h_i} \ , \quad c_4 = \frac{2\,\Delta_i - d_{i+1} + d_i}{h_i^2} \ ,$$

kde je použito označení $h_i = x_{i+1} - x_i$, $\Delta_i = \frac{y_{i+1} - y_i}{h_i}$. Pokud jsou k dispozici hodnoty funkce a první derivace v uzlových bodech x_i , $i = 1, ..., n$, lze pouhým dosazením určit koeficienty polynomů $P(x)$ pro všechny intervaly $[x_{i+1}, x_i]$, $i = 1, ..., n - 1$, postupně, což je výhodné při zpracování rozsáhlejších úloh nebo interaktivním kreslení závislostí. Tento typ Hermitovské interpolace má v porovnání s klasickou Hermitovskou interpolací výhodu, že jde o lokální kubické polynomy, které jsou dosti flexibilní, ale nemají tendenci tvořit nadbytečné extrémy. Jejich tvar lze "řídit" úpravou derivací, kdy však již nejde o typickou Hermitovskou interpolaci funkce.

Kubická C^1-interpolace je výhodná při rekonstrukci závislostí, protože se zde první derivace d_i stanovují lokálně na základě několika uzlových bodů v okolí bodu $\{x_i, y_i\}$. Není pak problémem modifikovat d_i tak, aby interpolující funkce $s_3^2(x)$ zachovávala lokální chování dat. Využitím směrnice Δ_i lze definovat *lokálně monotónní data* na intervalu $[x_{i+2}, x_i]$, pro která platí, že $\Delta_i \, \Delta_{i+1} > 0$. Je-li $\Delta_i \le \Delta_{i+1}$, jde o *lokálně konvexní* a pro $\Delta_i > \Delta_{i+1}$ *lokálně konkávní data*.

Požadavek lokální monotónnosti dat omezuje hodnoty d_i a d_{i+1} . Fritsch a Carlson[8] vyšli z nutné podmínky lokální monotónnosti

$$\mathrm{sign}(d_i) = \mathrm{sign}(d_{i+1}) = \mathrm{sign}(\Delta_i),$$

kde sign(x) je znaménková funkce, pro kterou je

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{pro } x > 0 \\ 0 & \text{pro } x = 0 \\ -1 & \text{pro } x < 0 \end{cases}.$$

Dále nalezli vztahy, které musí platit k zajištění lokální monotónnosti. Ke spojení případů monotónního růstu a poklesu je vhodné zavést parametry $\alpha_i = \frac{d_i}{\Delta_i}$, $\beta_i = \frac{d_{i+1}}{\Delta_i}$. Pokud platí nerovnost $\alpha_i + \beta_i - 2 \le 0$, je interpolace lokálně monotónní. Jestliže však neplatí, musí být k zajištění lokální monotónnosti splněny následující podmínky:

(a) $2\alpha_i + \beta_i - 3 \le 0$,

(b) $\alpha_i + 3\beta_i - 3 \le 0$,

(c) $\alpha_i + \frac{(2\alpha_i + \beta_i - 3)^2}{3\alpha_i + 3\beta_i - 6} \ge 0$.

Těmito podmínkami je definována přípustná oblast kombinací d_i, d_{i+1}, zajišťující lokální monotónnost. Místo podmínky (c) se pro jednoduchost uvažuje oblast vymezená v rovině α, β čtvrtkružnicí s poloměrem 3 ležící v prvním kvadrantu. Vlastní algoritmus úprav derivací pracuje ve dvou krocích:

1. Provede se úprava derivací. Vyjde-li $\Delta_i = 0$, položí se také $d_i = d_{i+1} = 0$.

2. Pro každý interval, kde neleží α_i, β_i v přípustné oblasti splňující podmínky (a), (b), (c), se naleznou přípustné derivace $d_i^{(}$, $d_{i+1}^{(}$, odpovídající hraničním hodnotám $\alpha_i^{(}$, $\beta_i^{(}$ na hranici přípustné oblasti. Využitím náhrady podmínky (c) čtvrtkružnicí lze psát

$$\alpha_i^{(} = \frac{3\alpha_i}{\sqrt{\alpha_i^2 + \beta_i^2}}, \qquad \beta_i^{(} = \frac{3\beta_i}{\sqrt{\alpha_i^2 + \beta_i^2}}.$$

Tento postup je využit u všech našich programů pro C^1-kubickou interpolaci jako standardní metoda. Popsaný algoritmus úprav prvních derivací vychází z hodnot všech směrnic d_i, cit.[8], což je při interaktivní interpolaci často omezením.

Hyman[9] navrhl použití lokální techniky, která provádí úpravu daných derivací d_i podle vztahů $d_i^{(} = \min[\max(0, d_i); 3\min(|\Delta_{i-1}|, |\Delta_i|)]$ pro případ $\sigma > 0$ a $d_i^{(} = \max[\min(0, d_i); -3\min(|\Delta_{i-1}|, |\Delta_i|)]$ pro případ $\sigma \ne 0$. Parametr σ má význam $\sigma = \mathrm{sign}(\Delta_{i-1})$, pokud je $|\Delta_{i-1}| > |\Delta_i|$ a $\sigma = \mathrm{sign}(\Delta_i)$, pokud je $|\Delta_{i-1}| \le |\Delta_i|$. Hymanův algoritmus lze tedy opět jednoduše zabudovat do všech programů pro C^1-kubickou a popř. i C^2-kubickou interpolaci.

Mezi nejjednodušší postupy určování derivací d_i patří tříbodová parabolická *Besselova interpolační formule*

$$d_i = \frac{h_{i\&1}\,\Delta_i + h_i\,\Delta_{i\&1}}{x_{i+1} - x_i}, \quad i = 2, \ldots, n-1 .$$

K zajištění lokální monotónnosti je výhodné použít přímo alternativní formule

$$d_i = \frac{3\,\Delta_i\,\Delta_{i\&1}}{\Delta_i^2 + \Delta_{i\&1}^2 + \Delta_i\,\Delta_{i\&1}} \cdot \frac{h_{i\&1}\,\Delta_i + h_i\,\Delta_{i\&1}}{x_{i+1} - x_{i\&1}}, \quad i = 2, \ldots, n-1 .$$

K určení derivací d_1, d_n je možno volit různé techniky. Mezi základní patří

(a) volba $d_1 = d_n = 0$,

(b) konstrukce přidaných uzlů, kde postačuje při využití parabolické interpolace z prvních a posledních tří bodů definovat x_0 a x_{n+1},

(c) použití interpolačního kvadratického polynomu pro první a poslední interval,

(d) zadání d_1 a d_n dle požadavku uživatele.

Volba d_1, d_n ovlivní chování interpolující funkce jen lokálně, takže není pro praktické účely rozhodující. V programech pro C^1-kubickou interpolaci volíme standardně postup založený na konstrukci přídavných bodů. Ze tříbodových formulí patří mezi nejlepší taková formule, která vychází z interpolace pomocí oskulační kružnice procházející třemi sousedními body. Zde se derivace určují ze vztahů

$$d_i = \frac{q_{i\&1}(h_i^2 + q_i^2) + q_i(h_{i\&1}^2 + q_{i\&1}^2)}{h_{i\&1}(h_i^2 + q_i^2) + h_i(h_{i\&1}^2 + q_{i\&1}^2)} ,$$

kde $q_i = y_{i+1} - y_i$. Pro určení d_1 a d_2 se používá vztahů

(a) pro $\Delta_1 > 0$ a $\Delta_1 > d_2$ nebo pro $\Delta_1 < 0$ a $\Delta_1 < d_2$ se volí $d_1 = 2\Delta_1 - d_2$,

(b) v opačném případě se volí $d_1 = \Delta_1 + \frac{|\Delta_1|\,(\Delta_1 - d_2)}{|\Delta_1| + |\Delta_1 - d_2|}$.

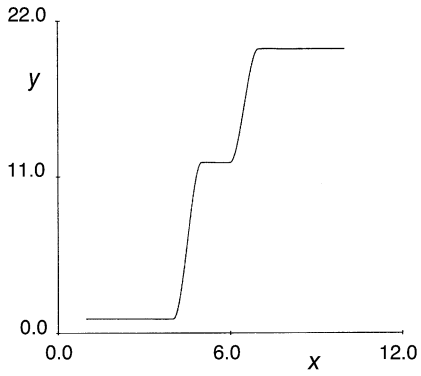
Obdobně se postupuje při volbě d_n.

Vzorová úloha 9.7 *Lokální kubická interpolace stupňovité závislosti*

Nalezněte lokální C^1-kubickou interpolaci pro zadaná data s využitím derivací počítaných z rovnice pro veličinu d_i. Určete i hodnoty derivace a integrálu ve všech uzlových bodech.

Data: $n = 10$

x_i	1	2	3	4	5	6	7	8	9	10
y_i	1	1	1	1	12	12	20	20	20	20



Obr. 9.10 C^1-interpolace využívající derivací .

Řešení: Byl určen průběh interpolační funkce pro případ bez omezení derivací. Výsledek je znázorněn na obr. 9.10. Při znalosti koeficientů c_1 až c_4 pro všechny lokální kubické polynomy je snadné určit analyticky jak první derivaci, tak i integrál v libovolném bodě intervalu $x_1 \# x \# x_n$. V tabulce 9.3 jsou uvedeny hodnoty první derivace a integrálu, odpovídající C^1-interpolaci znázorněné pro uzlové body na obr. 9.10.

Tabulka 9.3 Hodnoty derivací a integrálu v uzlových bodech.

x_i	1	2	3	4	5	6	7	8	9	10
Derivace	0	0	0	0.089	0.089	0.121	0.121	0	0	0
Integrál	0	1	2	2.99	9.49	21.5	37.5	57.5	77.5	97.5

Závěr: Z obr. 9.10 je patrné, že použití tříbodové formule vede pro tento případ k závislosti zachovávající lokální monotónnost dat.

Jednu z nejpoužívanějších pětibodových formulí pro lokální určování derivací navrhl Akima[10]. Derivace d_i se zde určují jako vážený "průměr" směrnic Δ_{i-1} a Δ_i podle vztahu

$$d_i = \frac{\Delta_{i\&1} * \Delta_{i\%1} \& \Delta_i * \% \Delta_i * \Delta_{i\&1} \& \Delta_{i\&2} *}{* \Delta_{i\%1} \& \Delta_i * \% * \Delta_{i\&1} \& \Delta_{i\&2} *} .$$

Pokud je $\Delta_{i+1} = \Delta_i$ a zároveň $\Delta_{i-1} = \Delta_{i-2}$, používá se jednodušší vztah $d_i = 0.5 * \Delta_i \% \Delta_{i\%1} *$. Klasický postup Akimovy interpolace využívá dvou dvojic přidaných uzlových bodů x_{-1}, x_0 *a* x_{n+1}, x_{n+2} na obou koncích intervalu interpolace. Souřadnice těchto bodů se pro dolní konec intervalu interpolace určují z rovnosti $x_{-1} - x_1 = x_0 - x_2 = x_1 - x_3$. Příslušné hodnoty y_i se pak určují z interpolační paraboly procházející prvními třemi uzlovými body $\{x_1 , y_1\}$, $\{x_2 , y_2\}$ a $\{x_3 , y_3\}$. Stejným způsobem jsou definovány přídavné uzlové body x_{n+1} , x_{n+2}. V programu AKIMA je použito jednodušších vztahů

$$x_{\&1} = x_1 \& 0.2\,(x_2 \& x_1), \qquad x_0 = x_1 \& 0.1\,(x_2 \& x_1) ,$$

$$x_{n\%1} \;'\; x_n \,\%\, 0.1\,(x_n \,\&\, x_{n\&1}), \qquad x_{n\%2} \;'\; x_n \,\%\, 0.2\,(x_n \,\&\, x_{n\&1}) \quad .$$

Uvedenou techniku testoval Akima spolu s několika dalšími interpolačními postupy a "průměrnými" křivkami, získanými subjektivním proložením "od oka", na několika testovacích příkladech. Proložení "od oka" se nejvíce blížila právě jeho metoda.

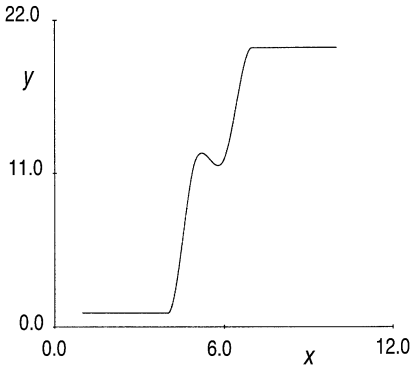
Vzorová úloha 9.8 *Akimova interpolace schodovité závislosti*

Pro data ze vzorové úlohy 9.7 nalezněte C^1-interpolační formuli využitím Akimova vztahu pro derivace a vypočítejte derivace a integrály ve všech uzlových bodech.

Řešení: Byl vypočten průběh interpolační funkce pro případ bez omezení derivací (obr. 9.11). V tabulce jsou uvedeny hodnoty derivace a integrálu této závislosti v uzlových bodech. Je patrné, že v tomto případě již C^1-interpolace neodpovídá lokálnímu chování dat. Úpravou derivací dle Fritche a Carlsona[8] vyjdou všechna $d_i = 0$ a průběh interpolované závislosti je pak shodný s obr. 9.10.

Hodnoty derivací a integrálu v uzlových bodech.

x_i	1	2	3	4	5	6	7	8	9	10
Derivace	0	0	0	0	4.63	4.63	0	0	0	0
Integrál	0	1	2	3	9.11	21.11	37.5	57.5	77.5	97.5



Obr. 9.11 C^1-Akimova interpolace.

Závěr: Je patrné, že Akimova interpolace nezajišťuje souhlas s lokálním chováním dat. Tento problém lze snadno odstranit použitím technik pro úpravu derivací.

9.2.2 Kubické spline

Kubické spline, nazývané také *křivítkové funkce*, patří k nejznámějším představitelům spline polynomů. Jejich základní výhodou je, že jsou spojité v prvních dvou derivacích, což umožňuje sestrojení hladké křivky v první derivaci (obr. 9.6). Používají se hojně ve všech oblastech počítačové grafiky, v systémech CAD/CAM i pro aproximaci funkcí, kde mají řadu vhodných vlastností.

Platí, že při aproximaci funkce $f(x)$ kubickým splinem $S_3(x)$ je zajištěna minimální norma $\|f^{(2)}(x) - S_3^{(2)}(x)\|^2$. Dále $S_3(x)$ splňují i podmínku

$$\int_a^b [S_3(x)]^2 \, dx \to \min,$$

což znamená, že jistým způsobem minimalizují celkovou křivost interpolující funkce. Fyzikálně si lze spline představit jako ideální elastický nosník, se zanedbatelnou hmotností, který je zatížen nebo podepřen v uzlových bodech $\{x_i, y_i\}$, $i = 1, ..., n$. Tento nosník zaujme tvar odpovídající minimu potenciální energie. Interpolační kubický spline lze sestrojit při volbě $m = 3$ přímo z definice. Numericky nejvýhodnější je však použití B-spline $B_4(x)$, znázorněných na obr. 9.9. Pro ilustraci je výhodné vyjít přímo z definice $S_3(x)$ jako polynomické funkce třídy $C^2[a, b]$. Kubický spline je pak definován podmínkami

(a) podmínkou interpolace, tj. $S_3(x_i) = y_i, \quad i = 1, ..., n$,

(b) podmínkami spojitosti ve funkčních hodnotách a hodnotách první i druhé derivace. Funkce $S(x)$, $S^{(1)}(x)$ a $S^{(2)}(x)$ jsou spojité v celém intervalu $[a, b]$;

(c) podmínkou po částech konstantní třetí derivace $S^{(3)}(x)$ všude kromě uzlových bodů x_i, $i = 1, ..., n$;

(d) podmínkou nulové čtvrté derivace $S^{(4)}(x) = 0$ všude kromě uzlových bodů x_i, $i = 1, ..., n$.

Z těchto podmínek plyne, že $S_3(x)$ je v každém intervalu $I_j \in [x_{i+1}, x_i]$ definována kubickým polynomem. Pro jednoznačné určení všech čtveřic koeficientů c_1 až c_4 ve všech $(n - 1)$ intervalech I_j je potom nutné sestavit $4(n - 1)$ nezávislých podmínek. Z podmínek spojitosti plyne, že

$$P_j(x_i) = P_{j-1}(x_i), \quad P_j^{(1)}(x_i) = P_{j-1}^{(1)}(x_i), \quad P_j^{(2)}(x_i) = P_{j-1}^{(2)}(x_i)$$

pro všechna $i = 2, ..., n - 1$, což vede na $3(n - 2)$ vazebných rovnic. Z podmínky interpolace vychází dalších n rovnic. Celkově vede použití podmínek definice $S_3(x)$ k sestavení $(4n - 6)$ lineárních rovnic. Dvě rovnice pro jednoznačné určení $S_3(x)$ však ještě scházejí. Tyto vazebné rovnice se využívají pro definici *okrajových podmínek* určujících chování spline v místě x_1 a x_n. Často bývají voleny tzv. přirozené okrajové podmínky $S_3^{(2)}(x_1) = S_3^{(2)}(x_n) = 0$. Odpovídající podmínky pro první derivace

$$d_1 = S_3^{(1)}(x_1), \quad d_n = S_3^{(1)}(x_n)$$

mají tvar $d_1 = 0.5 \left[\frac{3 (y_2 - y_1)}{h_1} - d_2 \right]$, $d_n = 0.5 \left[\frac{3 (y_n - y_{n-1})}{h_{n-1}} - d_{n-1} \right]$. Název "přirozené" zde vystihuje fyzikální smysl prostého podepření, kdy vně poslední podpory x_1, resp. x_n zaujímá nosník přímkový tvar.

Obecně lze definovat *okrajové podmínky typu I*, kdy se určují d_1, d_n a *okrajové podmínky typu II*, kdy se určují druhé derivace $S_3^{(2)}(x_1)$ a $S_3^{(2)}(x_n)$. Přehled různých typů okrajových podmínek, které ovlivňují chování spline $S_3(x)$ pouze lokálně, je uveden

v práci[12].

Označme druhé derivace kubického spline $M_i = S_3^{(2)}(x_i)$ a první derivace $d_i = S_3^{(1)}(x_i)$. Pro určení kubického spline postačuje nalezení derivací d_2 , ..., d_{n-1} . Vyjádříme-li koeficienty c_1 až c_4 pomocí derivací d_i , d_{i+1} , M_i , M_{i+1} , můžeme z podmínky spojitosti druhých derivací v místě x_i dospět po úpravách ke vztahu

$$\alpha_i\, d_{i\&1} \,\%\, \beta_i\, d_i \,\%\, \gamma_i\, d_{i\%1} \;'\; \delta_i, \quad i \;'\; 2, \ldots, n \,\&\, 1 \;,$$

kde

$$\alpha_i \;'\; \frac{1}{h_{i\&1}}, \quad \beta_i \;'\; \frac{2}{h_{i\&1} \,\%\, h_i}, \quad \gamma_i \;'\; \frac{1}{h_i}$$

a

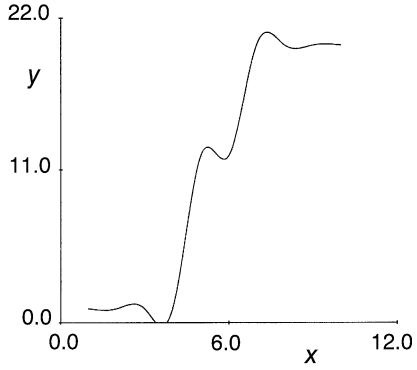
$$\delta_i \;'\; 3\left[\frac{y_i \,\&\, y_{i\&1}}{h_{i\&1}^2} \,\%\, \frac{y_{i\%1} \,\&\, y_i}{h_i^2}\right] .$$

Zápis představuje tridiagonální soustavu (n - 2) lineárních rovnic pro neznámé d_1 , ..., d_n. Při použití okrajových podmínek pak dostáváme soustavu n lineárních rovnic s n neznámými, kdy je matice koeficientů tridiagonální. Pro její řešení lze využít např. kompaktních algoritmů vycházejících z Gaussovy eliminace[1].

Vzorová úloha 9.9 *Spline interpolace schodovité závislosti*

Pro data uvedená ve vzorové úloze 9.7 nalezněte kubický spline s okrajovými podmínkami a vypočítejte první derivace v uzlových bodech.

Řešení: Využitím programu SPLINE byl určen průběh kubického interpolačního spline $S_3(x)$, který je zakreslen na obr. 9.12. První derivace v uzlových bodech jsou v tabulce.



Obr. 9.12 C^2-kubický spline interpolace.

Hodnoty první derivace v uzlových bodech

x_i	1	2	3	4	5	6	7	8	9	10
Derivace	0	0.56	-1.97	7.34	5.61	3.20	5.57	-1.5	0.429	-0.21

Závěr: Kubický spline tvoří všude tam, kde dochází k náhlým změnám křivosti interpolované závislosti, falešné extrémy.

Základním problémem při použití kubických spline pro rekonstrukci závislostí je

jejich tendence k překmitávání při náhlých změnách křivosti interpolované závislosti. Pro odstranění této nevýhody je vhodné použít tzv. *spline pod napětím* (exponenciální spline), která umožňují lokální řízení tvaru interpolující funkce pomocí parametru napětí. Spline pod napětím jsou řešením diferenciální rovnice[14]

$$S_T^{(4)}(x) - \lambda^2 S_T^{(2)}(x) = 0$$

pro $x = x_i$, $i = 1, ..., n$. Symbol λ označuje *parametr napětí.* Ten může být obecně v každém uzlovém bodě jiný a roven λ_i. Řešením uvedené diferenciální rovnice je spline pod napětím s těmito vlastnostmi:

(a) v každém intervalu $I_j \in [x_{i+1}, x_i]$ je

$$S_T(x) = a + bx + \exp(\lambda x) + \exp(-\lambda x) \quad .$$

(b) $S_T(x)$ je ze třídy funkcí $C^2[a, b]$,
(c) platí podmínka interpolace $S_T(x_i) = y_i$, $i = 1, ..., n$,
(d) $S_T^{(2)}(x_1) = S_T^{(2)}(x_n) = 0$, tj. platí přirozené okrajové podmínky.

Pro hraniční hodnoty parametru napětí λ platí, že

(a) pro $\lambda \to 0$ je $S_T(x)$ klasický kubický spline;
(b) pro $\lambda \to \infty$ je $S_T(x)$ lineární spline, tj. lomená čára spojující uzlové body.

Pro praktické účely se využívá toho, že rozdíl $S_T^{(2)}(x) - \lambda^2 S_T(x)$ je v každém intervalu I_j lineární funkcí x. Po dvojí integraci lze pak $S_T(x)$ vyjádřit ve tvaru

$$S_T(x) = y_{i+1}\, t + y_i\, (1 - t) + \frac{M_{i+1}}{\lambda_i^2}\left[\frac{\sinh(\mu_i\, t)}{\sinh(\mu_i)} - t\right] +$$

$$+ \frac{M_i}{\lambda_i^2}\left[\frac{\sinh(\mu_i\, (1 - t))}{\sinh(\mu_i)} - (1 - t)\right].$$

kde $t = (x - x_i)/h_i$ a $\mu_i = \lambda_i h_i$. Symboly $M_i = S_T^{(2)}(x_i)$ označují druhé derivace. Z této rovnice je patrné, že

(a) $S_T(x)$ je vzhledem k M_i, $i = 2, ..., n - 1$, lineární;
(b) při znalosti M_i, $i = 2, ..., n - 1$, je $S_T(x)$ jednoznačně určeno pro zvolená λ_i, $i = 1, ..., n$.

Uveďme, že funkce $\sinh(x)$ a $\cosh(x)$ lze vyjádřit využitím funkce $\exp(x)$ ve tvaru

$$\sinh(x) = 0.5\, [\exp(x) - \exp(-x)],$$
$$\cosh(x) = 0.5\, [\exp(x) + \exp(-x)].$$

Analogicky jako u klasických kubických spline lze z podmínky spojitosti prvních derivací $S_T^{(1)}(x)$ v uzlových bodech nalézt tridiagonální soustavu lineárních rovnic

$$\alpha_i^+ M_{i-1} + (\beta_i^+ + \beta_{i-1}^+)\, M_i + \alpha_{i+1}^+ M_{i+1} = \delta_i^+ \quad , \quad \text{kde} \quad \alpha_{i+1}^+ = \frac{\sinh(\mu_i) - \mu_i}{\mu_i^2 \sinh(\mu_i)}\, h_i \quad ,$$

$$\beta_i^{\%} \leftarrow \frac{\mu_i \cosh(\mu_i) - \sinh(\mu_i)}{\mu_i^2 \sinh(\mu_i)} h_i \ , \quad \delta_i^{\%} \leftarrow \frac{y_{i\%1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}} \ .$$

Řešení této soustavy rovnic lze provést stejně jednoduše jako u klasických polynomických spline. Samostatným problémem souvisejícím s použitím spline pod napětím je volba parametrů napětí. Rentrop[15] využívá ve své proceduře testu, zda druhé postupné diference $[x_{i-1}, x_i, x_{i+1}]\ y = \Delta^2 y_i$ souhlasí co do znaménka s druhou derivací M_i. Pokud vyjde, že $(M_i \ \Delta^2 y_i)\ (M_{i\%1} \ \Delta^2 y_{i\%1}) > 0$, dosazuje se λ_i . 0. Není-li tato podmínka splněna, rozlišují se dva případy:

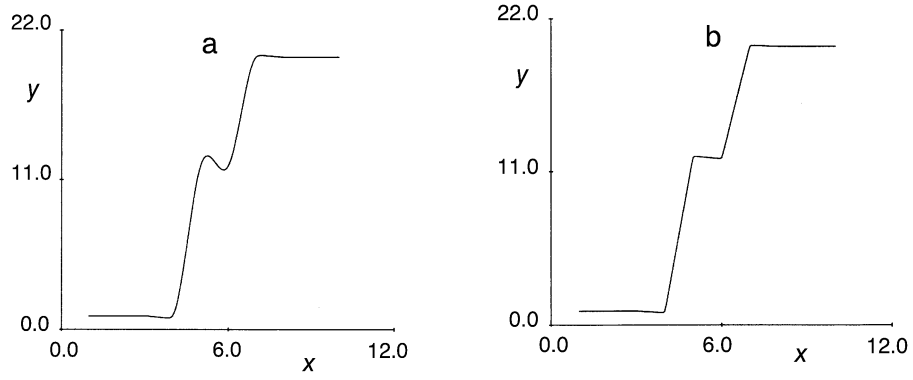
a) je-li $y_i = y_{i+1} = 0$, volí se $\lambda_i = 15/h_i$;

b) je-li $y_i = y_{i+1}$, volí se $\lambda_i \leftarrow \frac{1}{h_i}\left[4 \% (0.1 \% *y_{i\%1} - y_i* \max(*y_i*; *y_{i\%1}*))^{-1}\right]$.

Vzorová úloha 9.10 *Interpolace pomocí spline pod napětím*

Pro data uvedená ve vzorové úloze 9.7 nalezněte spline pod napětím $S_T(x)$ při volbě $\lambda = 50$ a dále při optimální volbě podle Rentropa.

Řešení: Byly určeny průběhy $S_T(x)$ pro optimální λ_i podle Rentropa (obr. 9.13a) a dále pro $\lambda_i = 50$, $i = 1, ..., n$, (obr. 9.13b). V tabulce jsou uvedeny hodnoty první derivace v uzlových bodech.



Obr. 9.13 Interpolace pomocí spline pod napětím při (a) volbě λ_i dle Rentropa, (b) volbě $\lambda_i = 50$, $i = 1, ..., n$.

Hodnoty první derivace v uzlových bodech

x_i	1	2	3	4	5	6	7	8	9	10
Derivace (λ dle Rentropa)	0	0.004	-0.16	-2.7	6.6	3.82	1.91	-0.008	0.003	$-2 \cdot 10^{-4}$
Derivace ($\lambda = 50$)	0	0.004	-0.047	5.54	5.51	3.99	4.03	-0.034	$2 \cdot 10^{-4}$	$-4 \cdot 10^{-6}$

Při porovnání obr. 9.13 s 9.12 (spline pod napětím pro $\lambda = 0$) je patrné, že Rentropův postup vede v výraznému zvýšení hladkosti. Není však zajištěna lokální monotónnost interpolující funkce. Při veliké hodnotě napětí ($\lambda = 50$) může dojít až ke stavu, kdy se interpolující funkce jeví jako lineární lomená závislost a poloměry křivosti jsou příliš malé. *Závěr:* Vhodnou volbou napětí λ_i lze tvar interpolujícího spline pod napětím "řídit" v širokých mezích.

9.3 Aproximace funkcí

Při aproximaci funkce *f(x)* vhodnou aproximující funkcí *g(x)* je třeba řešit dvě základní úlohy: (a) výběr typu funkce *g(x)*; (b) výběr kritéria pro vyjádření blízkosti funkcí *f(x)* a *g(x)*. S ohledem na jednoduchost zpracování se často volí *g(x)* ve tvaru $g_j(x) = x^{j-1}$, tj. polynomická aproximace. Blízkost funkcí *f(x)* a *g(x)* se vyjadřuje pomocí normy ve zvoleném L_P-prostoru, pro kterou platí

$$S_P \;=\; 2w(x)\;[f(x)\;\&\;g(x)]2_P \;=\; \left[\int_a^b w(x)\;*f(x)\;\&\;g(x)*^P\;\mathrm{d}x\right]^{1/P} .$$

V této rovnici je *w(x)* vhodná váhová funkce a interval $a \# x \# b$ určuje oblast, ve které se hledá aproximace funkce *f(x)* funkcí *g(x)*. Koeficienty c_j se pak hledají tak, aby bylo S_P minimální. Při volbě $p = 1$ jde o L_1-*aproximaci* a minimalizuje se integrál absolutních odchylek mezi *f(x)* a *g(x);* při volbě $p = 2$ se minimalizuje integrál čtverců odchylek a jde o L_2-*aproximaci*, odpovídající kritériu metody nejmenších čtverců pro diskrétní data. Konečně při volbě $p \to \infty$ jde o *minimaxní (Čebyševovu) aproximaci*, minimalizující kritérium

$$S_4 \;=\; \max[w(x)\,*f(x)\;\&\;g(x)*], \quad x \in [a,\, b] .$$

Minimalizace kritéria S_P vede obecně na úlohu nelineární optimalizace. Zde se omezíme na L_2-normu. Jde o spojitou analogii úlohy lineární regrese, která je podrobně popsána v kap. 6.

Uvažujme pro jednoduchost *w(x)* = 1. Pro odhad koeficientů c_1 až c_m se podobně, jako v diskrétním, případě vychází z analytické minimalizace S_2

$$\frac{\delta S_2}{\delta c_j} \;=\; 0, \quad j \;=\; 1, \ldots, m .$$

Po dosazení obdržíme soustavu normálních rovnic ve tvaru

$$\begin{bmatrix} \int_a^b f\, g_1\, \mathrm{d}x \\ \int_a^b f\, g_2\, \mathrm{d}x \\ \vdots \\ \int_a^b f\, g_j\, \mathrm{d}x \\ \vdots \\ \int_a^b f\, g_m\, \mathrm{d}x \end{bmatrix} = \begin{bmatrix} \int_a^b g_1^2\, \mathrm{d}x & \int_a^b g_1\, g_2\, \mathrm{d}x & \dots & \int_a^b g_1\, g_m\, \mathrm{d}x \\ \int_a^b g_1\, g_2\, \mathrm{d}x & \int_a^b g_2^2\, \mathrm{d}x & \dots & \int_a^b g_2\, g_m\, \mathrm{d}x \\ \vdots & \vdots & & \vdots \\ \int_a^b g_1\, g_j\, \mathrm{d}x & \int_a^b g_2\, g_j\, \mathrm{d}x & \dots & \int_a^b g_j\, g_m\, \mathrm{d}x \\ \vdots & \vdots & & \vdots \\ \int_a^b g_1\, g_m\, \mathrm{d}x & \int_a^b g_2\, g_m\, \mathrm{d}x & \dots & \int_a^b g_m^2\, \mathrm{d}x \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_j \\ \vdots \\ c_m \end{bmatrix} .$$

Je použito označení $f = f(x)$ a $g_j = g_j(x)$. Pokud máme analytické vyjádření funkce $f(x)$ a zvolíme vhodně $g_j(x)$, můžeme určit jednotlivé integrály analyticky a řešit pak soustavu m lineárních rovnic o m neznámých stejně jako v diskrétním případě (viz odd. 9.6). Při polynomické aproximaci je výhodné provést transformaci nezávisle proměnné tak, aby byl integrační obor v rozmezí [-1, 1]. Toho lze docílit volbou

$$x' = \frac{2x - a - b}{b - a} .$$

Úlohou je pak určení koeficientů polynomu $g(x') = \sum_{j=1}^{m} b_j\, x'^{(j-1)}$ tak, aby byla ve smyslu normy L_2 nejlépe aproximována funkce $f\left(\frac{a+b}{2} + \frac{b-a}{2}\, x'\right)$ v intervalu [-1, 1].

Při sestavování matice koeficientů uvedené maticové rovnice lze využít známých integrálů

$$\int_{-1}^{1} x^{2j}\, \mathrm{d}x = \frac{2}{2j+1} , \quad \int_{-1}^{1} x^{2j+1}\, \mathrm{d}x = 0 .$$

Pro *(m*-1*) sudé* přechází tato maticová rovnice do tvaru

$$\begin{bmatrix} 1 & 0 & \frac{1}{3} & 0 & \frac{1}{5} & \dots & \frac{1}{m-1} \\ 0 & \frac{1}{3} & 0 & \frac{2}{5} & 0 & \dots & 0 \\ \frac{1}{3} & 0 & \frac{1}{5} & 0 & \frac{1}{7} & \dots & \frac{1}{m-3} \\ \vdots & & & & & & \\ \frac{1}{m-1} & 0 & \frac{1}{m-3} & 0 & \frac{1}{m-5} & \dots & \frac{1}{2m-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} I_0 \\ I_1 \\ I_2 \\ \vdots \\ I_{m-1} \end{bmatrix},$$

$$\text{kde} \quad I_j = 0.5 \int_{-1}^{1} x^j f\left(\frac{a+b}{2} + \frac{b-a}{2} x\right) dx \,.$$

V této rovnici jsou oproti předešlé rovnici všechny koeficienty děleny dvěma tak, aby došlo ke zjednodušení matice koeficientů. Pro zvolený stupeň polynomické aproximace m lze určit koeficienty b_1 až b_m jako lineární kombinace známých integrálů. Obdobně lze postupovat i pro (m - 1) liché. V tabulce jsou pro m = 2, 3, 4, 5 uvedeny výrazy pro koeficienty b_i.

Koeficienty aproximačních polynomů $g(x^*)$ různých stupňů (m - 1)

m	b_1	b_2	b_3
2	I_0	$3I_1$	0
3	$\frac{3}{4}(3I_0 - 5I_2)$	$3I_1$	$\frac{15}{4}(3I_2 - I_0)$
4	$\frac{3}{4}(3I_0 - 5I_2)$	$\frac{15}{4}(5I_1 - 7I_3)$	$\frac{15}{4}(3I_2 - I_0)$
5	$\frac{15}{64}(15I_0 - 70I_2 + 63I_4)$	$\frac{15}{4}(5I_1 - 7I_3)$	$\frac{105}{32}(42I_2 - 45I_4 - 5I_0)$

m	b_4	b_5
2	0	0
3	0	0
4	$\frac{35}{4}(5I_3 - 3I_1)$	0
5	$\frac{35}{4}(5I_3 - 3I_1)$	$\frac{315}{64}(3I_0 - 30I_2 + 35I_4)$

Postačuje-li tedy aproximace polynomem maximálně čtvrtého stupně, lze určit jeho

koeficienty přímo z tabulky. Pro vyjádření kvality aproximace se počítá střední kvadratická odchylka[17]

$$SE = \sqrt{\frac{1}{b-a}\int_a^b (f(x)-g(x))^2\,\mathrm{d}x} = \sqrt{\frac{1}{b-a}\int_a^b f^2(x)\,\mathrm{d}x - \sum_{i=1}^{m} b_i I_{i-1}}.$$

Vzorová úloha 9.11 *Aproximace funkce exp(x)*

Stanovte kvadratický aproximační polynom, který ve smyslu L_2-normy nejlépe aproximuje funkci exp(x) v intervalu (0, 2).

Řešení: Pro určení aproximační paraboly využijeme tabulky. Pro integrály I_0 až I_2 při $x^* = x - 1$ platí

$$I_0 = 0.5\int_{-1}^{1}\exp(1+x^*)\,\mathrm{d}x = \frac{\exp(2)-1}{2} = 3.1945\ ,$$

$$I_1 = 0.5\int_{-1}^{1} x^*\exp(1+x^*)\,\mathrm{d}x = [\mathrm{e}^x\,(x+1)]_0^2 - [\mathrm{e}^x]_0^2 = 1\ ,$$

$$I_2 = 0.5\int_{-1}^{1} x^{*2}\exp(1+x^*)\,\mathrm{d}x = \frac{\exp(2)-5}{2} = 1.1945\ .$$

Použitím druhého řádku tabulky, pro $m = 3$, pak přímo dostaneme $b_1 = 2.70825$, $b_2 = 3$ a $b_3 = 1.45875$. Aproximační polynom má tvar $g(x^*) = 2.70825 + 3\,x^* + 1.45875\,x^{*2}$. Po zpětné transformaci na proměnnou x pak vyjde $g(x) = 1.167 + 0.08248\,x + 1.45875\,x^2$. Pro výpočet střední kvadratické odchylky je třeba ještě určit integrál

$$\int_0^2 \exp(2x)\,\mathrm{d}x = 0.5\,(\exp(4)-1)$$

a pak dosadit do rovnice

$$\mathrm{SE} = \sqrt{0.5\,(\exp(4)-1) - 2.70825\cdot 3.1945 - 3 - 1.45875\cdot 1.1945} = 0.0745\ .$$

Závěr: Aproximace funkce *f(x)* je při použití tabulky velmi jednoduchá. Vyžaduje pouze analytické či numerické určení integrálů.

V řadě praktických případů je základním problémem určování integrálů I_j, které

lze vyčíslit pouze numericky a nikoliv analyticky. Pokud pracujeme s funkcí zadanou pouze tabulkou hodnot $\{x_i , f(x_i)\}$, $i = 1, ..., n$, je úloha její aproximace shodná s úlohou aproximace závislosti.

9.4 Aproximace tabelárních závislostí

Úloha aproximace závislostí, zadaných tabulkou $\{x_i , y_i\}$, $i = 1, ..., n$, se od úlohy aproximace funkcí liší pouze v tom, že místo integrálu se v rovnici k vyjádření kritéria S_P užívá sumy. Pro známé $g(x)$ a $p = 2$ jde o úlohu nelineární nebo lineární regrese, která je podrobně popsána v kap. 8 a kap. 6.

9.4.1 Polynomická aproximace

V kap. 6 je pojednáno o odhadu parametrů v polynomických modelech, hledání vhodného stupně polynomu a využití ortogonálních polynomů na dané posloupnosti hodnot $x_1 , x_2 , ..., x_n$. Jde vždy o metodu nejmenších čtverců odchylek, tj. L_2-aproximaci. V řadě úloh aproximace metoda nejmenších čtverců odchylek nevyhovuje a požaduje se minimalizace maximální odchylky. Toho lze při polynomické aproximaci docílit využitím *ortogonálních Čebyševových polynomů.* Čebyševovy polynomy $T_m(x)$ lze generovat podle rekurentní formule

$$T_{m+1}(x) = 2 x T_m(x) - T_{m-1}(x) ,$$

Platí, že $T_0(x) = 1$ a $T_1(x) = 1$. Čebyševovy polynomy mají tyto základní vlastnosti[18]:

(a) koeficient u maximální mocniny x^m je roven 2^{m-1} - 1 pro $m \geq 1$ nebo 1 pro $m = 0$,

(b) Čebyševovy polynomy jsou symetrické kolem počátku, tj. platí

$$T_m(-x) = (-1)^m T_m(x) ,$$

(c) Čebyševův polynom $T_m(x)$ má v intervalu [-1, 1] právě m nulových bodů $T_m(x) = 0$ v místech x_j^*, které se nazývají čebyševovské uzlové body.

(d) Čebyševův polynom $T_m(x)$ má v intervalu [-1, 1] právě $(m + 1)$ extrémů x_j^+, pro které platí

$$x_j^+ = \frac{\cos(j \pi)}{n}; \quad T_m(x_j^+) = (-1)^j \quad \text{pro } j = 0, 1, ..., m ,$$

(e) při zavedení váhové funkce $w(x) = 1/\sqrt{1 - x^2}$ jsou Čebyševovy polynomy vzájemně ortogonální na celém intervalu [-1, 1].

(f) pokud se definuje $(m + 1)$ bodů x_j^*, které jsou nulovými body Čebyševova polynomu $T_{m+1}(x)$, platí, že

$$\sum_{j=1}^{m+1} T_i(x_j^*) T_k(x_j^*) = \begin{cases} 0 & \text{pro } i \neq k \\ 0.5 (m + 1) & \text{pro } i = k \neq 0 \\ m + 1 & \text{pro } i = k = 0 \end{cases} .$$

Tato rovnice platí pro všechna *i, k* = 0, ..., m, a ukazuje, že na Čebyševových uzlech jsou Čebyševovy polynomy vzájemně ortogonální.

(g) ze všech polynomů *m*-tého stupně, které mají koeficient u mocniny x^m roven 1, má normalizovaný Čebyševův polynom $T_m(x)/2^{m-1}$ minimální hodnotu normy S_4 v intervalu [-1, 1]. Pomocí Čebyševových polynomů lze aproximující funkci *g(x)* vyjádřit ve tvaru

$$g(x) \doteq \sum_{j=0}^{m} c_j \, T_j(x) \qquad \text{pro } -1 \le x \le 1 \ .$$

Pro zajištění ortogonality funkcí $T_j(x)$ se nejdříve určí Čebyševovy uzly dvoufázovým postupem:

(a) pro zvolené *n* se určí uzlové body x^*_{j+1}, $j = 0, ..., n - 1$, v intervalu {-1, 1},

(b) využitím vztahu $Z_j^* = 0.5\,(a + b) + 0.5\,(b - a)\,x_j^*$ se určí Čebyševovy uzly Z_j^* v intervalu [*a, b*].

Pro hodnoty Z_j^* se následně určí buď funkční hodnoty $f(Z_j^*)$, nebo hodnoty závislosti y_j. Vzhledem k ortogonalitě jednotlivých $T(x_j)$ pro Čebyševovy uzly lze určit koeficienty c_j snadno ze vztahů

$$c_0 \doteq \sum_{j=1}^{n} \frac{f(Z_j^*)}{n} \quad \text{a} \quad c_j \doteq 2 \sum_{i=1}^{n} \frac{T_j(x_i^*)\, f(Z_i^*)}{n} \ .$$

Obě rovnice odpovídají použití klasické metody nejmenších čtverců.

Vzorová úloha 9.12 *Čebyševova aproximace funkce exp(x)*

Stanovte Čebyševovu aproximaci funkce exp(*x*) v intervalu ⟨0, 2⟩ polynomem druhého stupně pro $n = 5$ bodů.

Řešení: V tabulce jsou uvedeny Čebyševovy uzly x_j^*, Z_j^* a hodnoty funkce $\exp(Z_j^*)$.

Zadání hodnot pro aproximace exp(*x*)

j	x_j^*	Z_j^*	$\exp(Z_j^*)$
1	-0.9511	0.0489	1.0502
2	-0.5878	0.4122	1.5102
3	0	1	2.7183
4	0.5878	1.5878	4.8929
5	0.9511	1.9511	7.0361

Plyne, že $T_0 = 1$, $T_1 = x$ a $T_2 = 2x^2 - 1$. Rovnice má konkrétní tvar

$$g(x) \doteq c_0 + c_1 x + c_2 (2x^2 - 1) \ .$$

Z rovnic pak vyčíslíme $c_0 \doteq \sum_{j=1}^{5} \frac{\exp(Z_j^*)}{5} \doteq 3.4415$,

$$c_1 = 2 \sum_{j=1}^{5} x_j \frac{\exp(Z_j)}{5} = 3.0725\,, \quad c_2 = 2 \sum_{j=1}^{5} \frac{(2x_j^2 - 1)\exp(Z_j)}{5} = 0.7380\,.$$

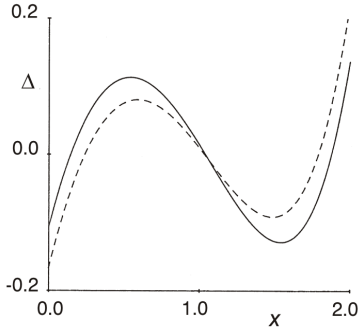
Po úpravách vyjde pro interval [-1, 1] aproximační polynom

$$g(x) = 2.7035 + 3.0725\,x + 1.476\,x^2.$$

Převedením do původního intervalu ($0 \le x \le 2$) dostáváme

$$g(x) = 1.107 + 0.1205\,x + 1.476\,x^2.$$

Střední kvadratická odchylka je rovna $SE = 0.0843$. Na obr. 9.14 je znázorněn průběh chyby aproximace $\Delta = \exp(x) - g(x)$, z $g(x)$ vyjádřené ve vzorové úloze 9.11.



Obr. 9.14 Chyba Δ aproximace exp(*x*) pomocí Čebyševovy (plná čára) a L_2-aproximace (čárkovaně).

Z hlediska celkového přiblížení je lepší polynom určený z integrálního kritéria L_2-aproximace. Čebyševovská aproximace však vede k minimální maximální odchylce.
Závěr: Pokud lze předem volit souřadnice aproximované funkce či závislosti na ose *x,* je snadné určit aproximační polynom optimální v minimaxním smyslu, tj. minimalizující maximální absolutní odchylku.

Vzorová úloha 9.13 *Hledání nejlepšího poměru polynomů*

Nalezení nejlepšího modelu poměru dvou polynomů, mezi stovkami všech možných transformačních modelů, se provede na základě kritéria dosažení co nejtěsnějšího proložení. Nalezený model je podroben detailní regresní analýze. Na nezávisle proměnnou *x* a závisle proměnnou *y* lze užít také rozličné transformace, čímž se paleta testovaných modelů rozšíří až na několik set. Obecný model poměru polynomů zapíšeme vztahem

$$g(y) = \frac{a_0 + a_1 f(x) + a_2 f^2(x) + a_3 f^3(x) + a_4 f^4(x) + a_5 f^5(x)}{1 + b_1(x) + b_2 f^2(x) + b_3 f^3(x) + b_4 f^4(x) + b_5 f^5(x)} + \varepsilon\,,$$

kde $g(y)$ a $f(x)$ představuje mocninné transformace *y* a *x* nebo logaritmy, odmocniny, atd. Neznámé parametry $a_0, a_1, ..., a_5$ a $b_1, b_2, ..., b_5$ jsou odhadovány z dat, ε značí náhodnou chybu. Řada parametrů však může být nulových a model se pak zjednoduší.
Pravidla výstavby modelu: 1. První zásadou je nalézt model co nejjednodušší, s co nejmenším počtem neznámých parametrů. 2. Druhou zásadou je požadavek, aby testovaný model měl v čitateli vždy polynom nižšího stupně než má polynom ve jmenovateli.
Data: jsou uvedena v tabulce predikce; jde o popis dvou chromatografických píků.

Řešení:

Průběh minimalizačního procesu

It.	Suma chyb					
	λ	λ	a_0	a_1	a_2	a_3
0	70.81162	0.00004	11.78766	-0.2798519	5.809555E-03	-4.172031E-05
1	70.73225	0.16	11.78725	-0.2798166	5.810308E-03	-4.172031E-05
..	...	...	...	...	...	...
9	70.50236	0.1048576	11.79268	-0.2795659	5.810308E-03	-4.172031E-05
10	70.49983	0.4194304	11.79228	-0.279557	5.810308E-03	-4.172031E-05

Bylo dosaženo maximálního počtu povolených iterací.

Odhady parametrů modelu

Parametr	Odhad parametru	Asymptotická směr. odchylka	Dolní mez 95 % i. s.	Horní mez 95 % i. s.
a_0	11.79254	1.012715	9.750206	13.83488
a_1	-0.2795364	9.166186E-02	-0.4643901	-9.468261E-02
a_2	5.810308E-03	3.516727E-03	-1.281847E-03	1.290246E-02
a_3	-4.172031E-05	2.863182E-05	-9.946187E-05	1.602125E-05
b_1	-0.0783304	2.020337E-03	-8.240479E-02	-0.074256
b_2	2.391857E-03	4.689093E-05	2.297292E-03	2.486421E-03
b_3	-2.86472E-05	4.332001E-06	-3.738351E-05	-1.991088E-05
b_4	1.171313E-07	1.025966E-06	-1.951927E-06	2.186189E-06

Závisle proměnná: y

Nezávisle proměnná: x

Model: $y = (a_0 + a_1x + a_2x^2 + a_3x^3) / (1 + b_1x + b_2x^2 + b_3x^3 + b_4x^4)$

R^2 0.990159

Počet iterací: 10

Model numericky: ((11.79254-(.2795364)*(*x*)+(5.810308E-03)*(*x*)^2-(4.172031E-05)*(*x*)^3)/(1-(.0783304)*(*x*)+(2.391857E-03)*(*x*)^2-(2.86472E-05)*(*x*)^3+(1.171313E-07)*(*x*)^4))

Asymptotická korelační matice parametrů

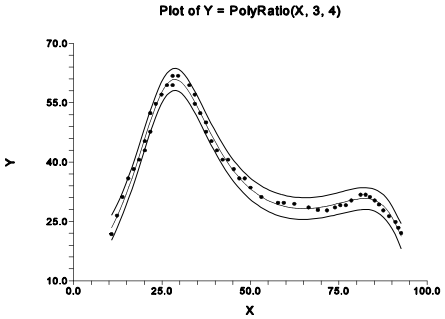
	a_0	a_1	a_2	a_3	b_1	b_2	b_3	b_4
a_0	1.000000	-0.668342	0.251252	-0.126446	0.045937	0.033491	-0.444426	0.769239
a_1	-0.668342	1.000000	-0.826039	0.725516	0.586512	-0.664699	-0.127771	-0.364811
a_2	0.251252	-0.826039	1.000000	-0.986996	-0.937223	0.967391	0.655601	0.217490
a_3	-0.126446	0.725516	-0.986996	1.000000	0.978499	-0.991658	-0.765633	-0.176658
b_1	0.045937	0.586512	-0.937223	0.978499	1.000000	-0.989097	-0.867018	-0.095811
b_2	0.033491	-0.664699	0.967391	-0.991658	-0.989097	1.000000	0.801153	0.123138
b_3	-0.444426	-0.127771	0.655601	-0.765633	-0.867018	0.801153	1.000000	-0.139159
b_4	0.769239	-0.364811	0.217490	-0.176658	-0.095811	0.123138	-0.139159	1.000000

Je-li absolutní hodnota korelace vyšší než 0.95, je přesnost parametru podezřelá.

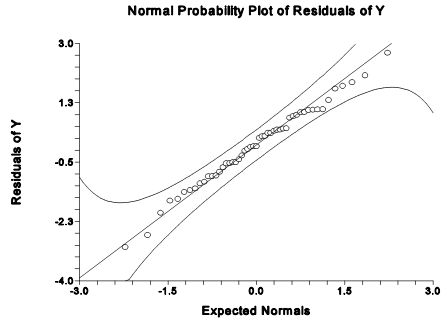
Predikované hodnoty a analýza klasických reziduí

Řádek			Predikce	Dolní mez	Horní mez	
	x	y	y_P	**95.0%** i. s.	**95.0%** i. s.	**Reziduum**
1	10.69182	21.76471	23.39933	20.2282	26.57045	-1.634617
2	12.26415	26.47059	26.25687	23.30506	29.20868	0.2137172
3	13.83648	31.17647	29.50871	26.70471	32.31271	1.667763
4	15.4088	35.88235	33.16126	30.38701	35.93551	2.721093

5	16.98113	38.23529	37.18533	34.38747	39.98318	1.04997
6	18.55346	40.58823	41.49773	38.6993	44.29615	-0.9094926
7	20.12579	42.94118	45.94416	43.17859	48.70974	-3.002987
8	20.12579	45.29412	45.94416	43.17859	48.70974	-0.6500469
9	21.69811	47.64706	50.29161	47.55496	53.02825	-2.644546
10	23.27044	54.70588	54.24173	51.50516	56.97831	0.4641487
11	24.84277	57.05882	57.47352	54.71762	60.22942	-0.4146998
12	26.41509	59.41177	59.71032	56.93428	62.48635	-0.2985483
13	27.98742	61.76471	60.78881	58.00365	63.57397	0.9758998
14	29.55975	61.76471	60.69903	57.92286	63.4752	1.065677
15	27.98742	59.41177	60.78881	58.00365	63.57397	-1.37704
16	21.69811	52.35294	50.29161	47.55496	53.02825	2.061334
17	32.7044 5	9.41177 5	7.6542	54.93171	60.37669	1.757568
18	34.27673	57.05882	55.19875	52.49294	57.90457	1.860069
19	34.27673	54.70588	55.19875	52.49294	57.90457	-0.4928702
20	35.84906	52.35294	52.45885	49.75431	55.16339	-0.1059121
21	37.42138	50.000	49.63521	46.9237	52.34673	0.364788
22	37.42138	47.64706	49.63521	46.9237	52.34673	-1.988152
23	38.99371	45.29412	46.8722	44.15504	49.58937	-1.578087
24	40.56604	42.94118	44.26291	41.54681	46.97902	-1.321735
25	42.13836	40.58823	41.86002	39.15155	44.56849	-1.271782
26	43.71069	40.58823	39.68738	36.98952	42.38524	0.9008505
27	45.28302	38.23529	37.74997	35.06122	40.43872	0.4853272
28	46.85535	35.88235	36.04123	33.35659	38.72587	-0.158876
29	48.42767	35.88235	34.54837	31.86112	37.23561	1.33399
30	50.000	33.52941	33.25569	30.5592	35.95218	0.273719
31	53.14465	31.17647	31.2057	8.47747	33.93393	-2.923354E-02
32	57.86164	29.70588	29.2488	26.47584	32.02176	0.4570828
33	59.43396	29.70588	28.84581	26.06691	31.62471	0.8600686
34	62.45283	29.41177	28.36928	25.59633	31.14223	1.042489
35	66.47799	28.52941	28.26572	25.5267	31.00475	0.2636867
36	73.92453	28.52941	29.31248	26.57718	32.04779	-0.7830742
37	81.16982	31.76471	30.7306	27.96259	33.49862	1.034105
38	78.59119	30.29412	30.34191	27.55837	33.12545	-4.778932E-02
39	77.01887	29.11765	30.00762	27.23381	32.78143	-0.8899736
40	75.44654	29.11765	29.65103	26.89622	32.40585	-0.5333864
41	69.09434	27.94118	28.47773	25.75909	31.19637	-0.5365493
42	71.71069	27.79412	28.87128	26.15429	31.58828	-1.077167
43	82.57861	31.76471	30.78054	28.0286	33.53248	0.9841667
44	83.78616	31.17647	30.67746	27.92962	33.4253	0.4990151
45	85.19497	30.29412	30.32524	27.55765	33.09282	-3.111753E-02
46	86.40252	29.26471	29.76715	26.96264	32.57165	-0.5024396
47	87.61006	27.79412	28.91596	26.07486	31.75706	-1.121842
48	89.22012	26.32353	27.23518	24.39041	30.07995	-0.9116499
49	91.03145	24.85294	24.48924	21.65272	27.32576	0.3637006
50	91.83648	23.38235	22.96273	20.02536	25.90011	0.4196216
51	92.64151	22.05882	21.2528	18.05387	24.45173	0.8060208



Obr. 9.15a Těsnost proložení experimentálních bodů modelem.



Obr. 9.15b *Q-Q* graf reziduí.

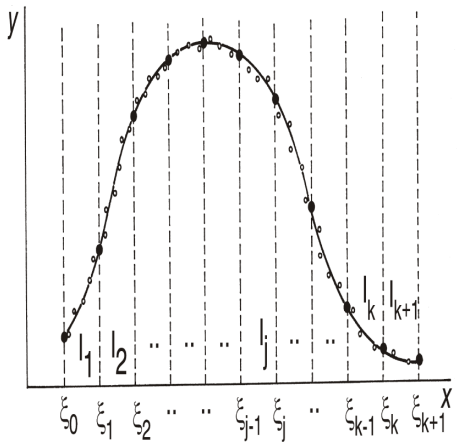
Závěr: Jelikož jsou pásy intervalu spolehlivosti predikce poměrně úzké a rovnoměrné, lze považovat nalezené odhady parametrů a regresní model za konečné.

9.4.2 Úseková regrese

Použití klasických polynomů jako aproximačních modelů je nevhodné při aproximaci fyzikálních závislostí, které nejsou asociativní povahy. Pro komplikovanější průběhy s několika extrémy mají navíc tendenci oscilovat, či výrazně zkreslovat aproximovanou závislost. V těchto případech je výhodnější použít po částech definovaných funkcí. Kromě zadaných n bodů $\{x_i, y_i\}$, i = 1, ..., n, kde se předpokládá, že y_i jsou náhodné veličiny (měřené hodnoty), se ještě určují uzlové body ξ_j, $j = 1, ..., k$ (resp. ještě ξ_0, ξ_{k+1}). Uzlové body tvoří hranice intervalů, kde jsou definovány jednotlivé funkce. V každém intervalu I_j, ohraničeném uzlovými body ξ_{j-1}, ξ_j, lze aproximující funkci $g(x)$ vyjádřit modelem $g_j(x)$, takže platí:

$$
\begin{aligned}
g(x) &= g_1(x) && \text{pro } x \,\varepsilon\, I_1 \\
&\vdots && \vdots \\
g(x) &= g_j(x) && \text{pro } x \,\varepsilon\, I_j \\
&\vdots && \vdots \\
g(x) &= g_{k+1}(x) && \text{pro } x \,\varepsilon\, I_{k+1}.
\end{aligned}
$$

Funkce $g_j(x)$ jsou lokálně definovány pouze na intervalech I_j . Kvalita aproximace závisí na počtu a polohách jednotlivých uzlových bodů ξ_j, typu funkcí $g_j(x)$ a na tom, ze které třídy C^m má být aproximující funkce $g(x)$. Úlohu lze převést na úlohu nelineární regrese, kde se hledá počet uzlových bodů, jejich polohy a koeficienty všech lokálně definovaných funkcí $g_j(x)$ metodou nejmenších čtverců nebo obecněji maximální věrohodnosti. Takto definovaná úloha je značně rozsáhlá, a proto se používá řada zjednodušení.



Obr. 9.16 Zadání úsekové regrese.

Obyčejně si uživatel volí počet uzlových bodů a často i jejich polohy předem na základě průběhu aproximované závislosti. Pokud jsou navíc $g_j(x)$, $j = 1, ..., k + 1$, lineární vzhledem k parametrům (polynomy), jde vlastně o úlohu lineární regrese s omezeními, definovanými podmínkami spojitosti ve funkčních hodnotách a hodnotách derivací s ohledem na třídu C^m

$$g_j^{(l)}(\xi_j) \quad ' \quad g_{j\%1}^{(l)}(\xi_j) \ , \qquad j \ ' \ 1, \ ..., \ k \ , \quad l \ ' \ 0, \ ..., \ m \ .$$

Jak bylo ukázáno v odd. 9.2, splňují podmínku spojitosti ve funkčních hodnotách a hodnotách m derivací spline polynomy $(m + 1)$ stupně $S_{m+1}(x)$, které jsou definovány jako polynomy maximálního stupně $(m + 1)$. Za aproximující funkci $g(x)$ lze použít vhodnou definici spline $S_{m+1}(x)$ a hledat jeho koeficienty metodou nejmenších čtverců.

Pro ilustraci vyjdeme z předpokladu, že funkce $g(x)$ je požadována ze třídy C^0. Jako $g(x)$ použijeme vyjádření lineárního spline ve tvaru useknutého polynomu.

$$g(x) \quad ' \quad \beta_1 \ \% \ \beta_2 x \ \% \ \sum_{j \ 1}^{k} \beta_{j\%2} (x \ \& \ \xi_j)_{\%} \ ,$$

Pokud platí aditivní model měření $y_i \ ' \ g(x_i) \ \% \ \varepsilon_i$, $i \ ' \ 1, \ ..., \ n$, a chyby ε_i jsou nezávislé a stejně rozdělené náhodné veličiny s konstantním rozptylem, lze získat odhady b_j parametrů β_j, $j = 1, ..., k + 1$ minimalizací kritéria metody nejmenších čtverců $U \ ' \ \sum_{i \ 1}^{n} [y_i \ \& \ g(x_i)]^2$. Při znalosti počtu a poloh uzlových bodů ξ_j jde o úlohu lineární regrese. Derivací kritéria U podle jednotlivých parametrů lze dospět k soustavě normálních rovnic $\boldsymbol{M}\,\boldsymbol{b} \ ' \ \boldsymbol{Z}$. Soustava představuje $(k + 2)$ lineárních rovnic vzhledem k hledaným $b_1, ..., b_{k+2}$. Struktura matice $\boldsymbol{M}$ a vektoru $\boldsymbol{Z}$ je však ovlivněna speciálním typem funkce $g(x)$. První řádek matice $\boldsymbol{M}$ má tvar

$$[n \quad \sum x_i \quad \sum (x_i \,\&\, \xi_1)_{\%} \quad ..... \quad \sum (x_i \,\&\, \xi_k)_{\%}] \ ,$$

a druhý řádek je $[\sum x_i \quad \sum x_i^2 \quad \sum x_i (x_i \,\&\, \xi_1)_{\%} \quad ..... \quad \sum x_i (x_i \,\&\, \xi_k)_{\%}]$.

V dalších řádcích má obecně první prvek tvar

$$M_{j\%2,1} = \sum (x_i \,\&\, \xi_j)_{\%} \ , j = 1, ..., k \ ,$$

druhý prvek má tvar $M_{j\%2,2} = \sum x_i (x_i \,\&\, \xi_j)_{\%}$, $j = 1, ..., k$ . .

Další prvky $M_{l+2,j+2}$ jsou

$$M_{l\%2,j\%2} = \sum (x_i \,\&\, \xi_j)_{\%} (x_i \,\&\, \xi_l)_{\%} \ , \quad j = 1, ..., k \ , \quad l = 1, ..., k \ .$$

Vektor **Z** má složky $\boldsymbol{Z} = [\sum y_i \quad \sum y_i x_i \quad \sum y_i (x_i \,\&\, \xi_1)_{\%} \ ..... \ \sum y_i (x_i \,\&\, \xi_k)_{\%}]^{\mathrm{T}}$.

Jednotlivé složky matice ***M*** a vektoru ***Z*** jsou ovlivněny také tím, že se pracuje s useknutými polynomy. Pro zlepšení numerické stability se doporučuje transformace souřadnic x do intervalu [1, 2]. Pokud se požaduje aproximace ze třídy C^1, lze volit kvadratický spline a pro aproximaci ze třídy C^2 kubický spline atd. Pro všechny modely tohoto typu lze sestavit matici ***M*** i vektor ***Z*** a nalézt odhady parametrů $\beta_1, ..., \beta_{m+k+1}$ aproximačního spline $S_m(x)$. Z numerického hlediska však není použití reprezentace spline ve tvaru useknutých polynomů příliš vhodné, protože pro větší počet uzlových bodů je matice ***M*** špatně podmíněná. Výhodnější je použití *B*-spline reprezentace.

Demonstrujme si použití této reprezentace spline na příkladu, kdy má být $g(x)$ ze třídy C^0. Při použití lineárních *B*-spline je třeba definovat ještě přídavné body ξ_0 a ξ_{k+1} (viz obr. 9.16). Aproximující model má pak tvar

$$g(x) = \sum_{j=1}^{k\%2} b_j \, B_{2,j}(x) \ ,$$

kde $B_{2,j}(x)$ jsou konkrétně definovány ve vzorové úloze 9.6 a zakresleny na obr. 9.7. Zaveďme zkrácené označení $N_{j,i} = B_{2,j}(x_i)$. Po dosazení $g(x)$ do kritéria U a analytické minimalizaci dospějeme opět k soustavě rovnic $\boldsymbol{M}\,\boldsymbol{b} = \boldsymbol{Z}$. V tomto případě má vektor Z složky

$$Z_j = \sum_{i=1}^{n} y_i \, N_{j,i} \ , \qquad j = 1, ..., k \ \% \ 2 \ .$$

Matice ***M*** má vzhledem k lokální definovanosti lineárních *B*-spline tridiagonální strukturu. Pro její první řádek platí $M_{1,1} = \sum N^2_{1,i}$, $M_{1,2} = \sum N_{1,i} N_{2,i}$ *a* $M_{1,j} = 0, j = 3, ..., k + 2$. V j-tém řádku jsou nenulové pouze diagonální a první poddiagonální resp. naddiagonální prvky, pro které platí $M_{j,j} = \sum N^2_{j,i}$, $M_{j,j+1} = \sum N_{j,i} N_{j+1,i}$, $M_{j,j-1} = \sum N_{j,i} N_{j-1,i}$. Konečně v posledním ($k$+2). řádku jsou nenulové pouze poslední dva prvky.

$$M_{k\%2,k\%1} \;'\; \mathrm{j}\; N_{k\%2,i}\; N_{k\%1,i} \quad \mathrm{a} \quad M_{k\%2,k\%2} \;'\; \mathrm{j}\; N^{2}_{k\%2,i}\;.$$

Tridiagonální soustava rovnic se dá řešit kompaktními algoritmy. Pro případ C^1-aproximace vede použití kvadratických *B*-spline (viz obr. 9.8) k matici ***M*** s pětidiagonální strukturou. Případ C^2-aproximace vede při použití kubických *B*-spline (viz obr. 9.9) k matici ***M*** se sedmidiagonální strukturou. Také při volbě C^m-aproximace pro větší *m* je výhodné použití *B*-spline reprezentace. Přehled dalších možností použití spline regrese a způsob statistické analýzy těchto speciálních lineárních modelů je popsán v Eubankově práci[20].

Samostatným problémem je *volba uzlových bodů* ξ_j. V programu ADSTAT je možné vybrat mezi 4 alternativami:

(a) konstantním dělením uzlových bodů,

(b) umístěním uzlových bodů tak, aby v každém intervalu I_j byl stejný počet experimentálních bodů,

(c) volbou poloh uzlových bodů uživatelem,

(d) hledáním uzlového bodu programem, a to regresní optimalizací.

Volba uzlových bodů: Při volbě uzlových bodů uživatelem lze v případě kubické spline regrese, kdy je aproximační funkce ze třídy C_2 , použít následující rámcová pravidla[21]:

I. Nejvhodnější je volit co nejméně uzlových bodů s tím, že v každém intervalu I_j by mělo být nejméně 4 až 5 bodů.

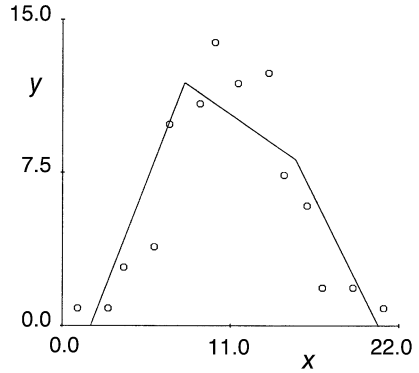
II. V intervalu I_j by měl být maximálně jeden extrém (minimum nebo maximum) a jeden inflexní bod.

III. Pokud je v I_j extrém, měl by ležet přibližně uprostřed.

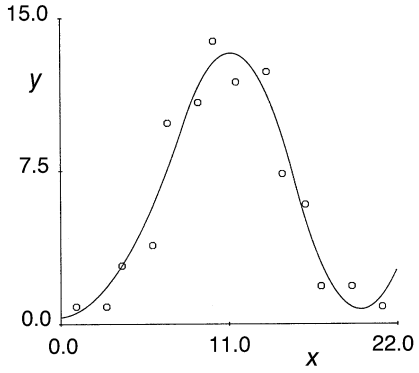
IV. Pokud je v intervalu I_j inflexní bod, měl by ležet v blízkosti uzlového bodu.

Vzorová úloha 9.14 *Aproximace píku*

Aproximujte pík, zadaný diskrétními hodnotami, využitím lineární, kvadratické a kubické spline regrese pro případ, že v každém intervalu I_j má ležet pět bodů. Stanovte také plochu pod tímto píkem.



Obr. 9.17a Lineární spline aproximace.

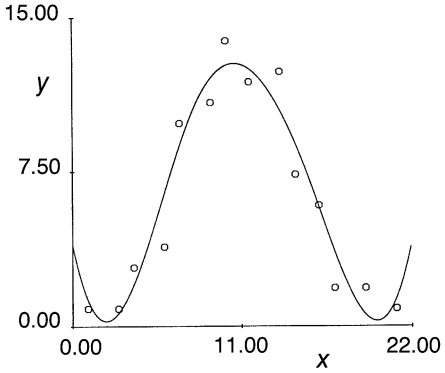


Obr. 9.17b Kvadratická spline aproximace.

Data: n = 14

x_i	1	3	4	6	7	9	10	11.5	13.5	14.5	16	17	19	21
y_i	1	1	3	4	10	11	14	12	12.5	7.5	6	2	2	1

Řešení: Výsledek C^0-regrese s modelem ve tvaru lineárního spline je znázorněn na obr. 9.17a. Při znalosti koeficientů lineárního spline lze snadno analyticky určit integrál I od $x = 1$ do $x = 21$, $I = 130.269$. Výsledek C^1-regrese s modelem ve tvaru kvadratického spline je znázorněn na obr. 9.17b. Stejným způsobem jako u lineárních spline byla určena plocha pod píkem $I = 126.068$. Na obr. 9.18 je znázorněn výsledek C^2-regrese s modelem ve tvaru kubického spline. Plocha pod píkem je v tomto případě rovna $I = 111.43$.



Obr. 9.18 Kubická spline aproximace.

Závěr: I když není zvolené dělení ani počet uzlových bodů optimální, ukazuje tento příklad rozdíly mezi spline modely různých stupňů. Pro případ, kdy se požaduje určení derivace, je vhodné volit kvadratický či kubický model.

Jistou nevýhodou modelů ve tvaru spline polynomů je při *hledání vhodného počtu a poloh uzlových bodů* opakované řešení metody nejmenších čtverců. Tyto modely jsou vhodné při interaktivní práci s obrazovkou počítače, kde uživatel snadno generuje různé způsoby rozmístění uzlových bodů tak, aby byl s výsledkem spokojen. V některých případech se však požaduje jednoprůchodová metoda, kdy se jednotlivé uzlové body zařazují postupně podle zvoleného statistického kritéria: v těchto případech se volí jak třída funkcí C^m, tak i kritérium regrese. Demonstrujme si takový postup na případu, kdy má být aproximující funkce ze třídy C^2 a jsou splněny podmínky pro užití metody nejmenších čtverců. Pro úsekovou regresi se pak používá polynomů stupně $(m + 3)$, tj. pátého.

Pro zvolené ξ_1 lze odpovídající polynom $p_1(x)$ vyjádřit ve tvaru

$$p_1(x) = \sum_{k=1}^{6} \beta_k (x - \xi_1)^{k-1} .$$

Pro odhad parametrů β_1 až β_6 lze použít metody nejmenších čtverců, což vede k řešení soustavy rovnic $\boldsymbol{\beta}_1 = \boldsymbol{H}_1^{-1} \boldsymbol{Z}_1$, kde vektor $\boldsymbol{Z}_1$ má složky

$$Z_j = \sum_{x_i \in I_1} y_i (x_i - \xi_1)^{j-1} , \qquad j = 1, \ldots 6 ,$$

a matice $\boldsymbol{H}_1$ má prvky

$$H_{lj} = \sum_{x_i \in I_1} (x_i - \xi_1)^{l-1} (x_i - \xi_1)^{j-1} , \qquad l = 1, ..., 6 , \quad j = 1, ..., 6 .$$

Při konstrukci ostatních polynomů je třeba vzít v úvahu omezení plynoucí z požadavku spojitosti ve funkčních hodnotách a hodnotách prvních dvou derivací v uzlových bodech. Polynom $p_k(x)$ pro interval $(\xi_{k-1} \# x \# \xi_k) \cap I_k$ je nutné vyjádřit ve tvaru

$$p_k(x) = \sum_{l=0}^{2} \frac{(x - \xi_{k-1})^l}{l!} \left. \frac{d^l p_{k-1}(x)}{dx^l} \right|_{x = \xi_{k-1}} + \sum_{r=4}^{6} \beta_r (x - \xi_{j-1})^{r-1} .$$

V této rovnici jsou pouze tři neznámé parametry β_4, β_5, β_6. Parametry lze určit metodou nejmenších čtverců. Označme první sčítance této rovnice jako $K(x)$. Formálně lze pak odhad parametrů $\boldsymbol{b}_k$ vyjádřit ve tvaru $\boldsymbol{b}_k = \boldsymbol{H}_k^{-1} \boldsymbol{Z}_h$, kde $\boldsymbol{b}_k$ má nyní pouze tři složky. Vektor $\boldsymbol{Z}_k$ má prvky

$$Z_j = \sum_{x_i \in I_k} [y_i (x_i - \xi_{k-1})^j - K(x_i))] , \quad j = 3, 4, 5 ,$$

a matice $\boldsymbol{H}_k$ má prvky $H_{l,j} = \sum_{x_i \in I_k} (x_i - \xi_{k-1})^{l-1} (x_i - \xi_{k-1})^{j-1}$ pro $j, l = 4, 5, 6$. Pro výpočet

koeficientů polynomu $p_k(x)$ postačuje znalost pouze předchozích uzlů ξ_1, ..., ξ_{k-1} a volba nového uzlu ξ_k.

K určení vhodného počtu a polohy ξ_k existuje řada postupů vycházejících z různých kritérií. Základní statistické kritérium sleduje, aby nevznikl nenáhodný trend v reziduích v intervalu I_j. V případě, že v reziduích e_i není trend, mělo by platit

$$\sum_{i=p}^{q-1} \hat{e}_i \, \hat{e}_{i+1} \# \sum_{i=p}^{q} \frac{\hat{e}_i^2}{\sqrt{q - p}} ,$$

kde p je index nejmenšího a q index největšího bodu v intervalu I_k. Tedy $x_p = \min(x_i)$ pro $x_i \in I_k$. Podobně lze definovat i x_q. Z dalších kritérií, popsaných v kap. 6, u výběru vhodného modelu se často používá *MEP* nebo *AIC* statistika.

Úseková regrese programem NCSS2000

1. Model úsekové regrese: Úseková regrese v často užívaném software NCSS2000 je konstruována kombinací přímek a kvadratických parabol, např. úsekový polynomický regresní model *lineární-lineární* se týká modelu o dvou lineárních rovnicích, když každá platí v jiném úseku proměnné *x*. Modelů je celá řada: model, kde větve jsou (a) *lineární-lineární,* (b) *lineární-kvadratická,* (c) *kvadratická-lineární,* (d) *kvadratická-kvadratická,* (e) *lineární-lineární-lineární.*

2. Uzlové body (body zvratu): bod zvratu nemusí být uživateli znám, často bývá cílem výpočtu. Přechod jedné větve křivky do druhé v bodu zvratu může být (a) *ostrý*, (b) *vnitřní*, hladký přechod uvnitř průsečíku křivek, (c) *vnější,* hladký, ale vně průsečíku křivek.

3. *Proměnné:* proměnné x a y mohou být předem transformovány mocninnou transformací, např. závisle proměnná y je předem transformována do tvaru $1/y^2$, $1/y$, $1/y^{0.5}$, $\ln y$, $y^{0.5}$, y^2 a nezávisle proměnná x do tvaru $1/x^2$, $1/x$, $1/x^{0.5}$, $\ln x$, $x^{0.5}$, x^2.

4. *Tabulka regresních modelů větví prokládané křivky:*

1. Model: lineární-lineární větve:

Regresní model: $y = A + B\,x + C\,(x - D)\,\text{sign}\,(x - D)$

Rovnice: $y = a_1 + b_1\,x,\ x < \xi$

$y = a_2 + b_2\,x,\ x > \xi$

Odhadované parametry:

$A = (a_1 + a_2)/2$, $a_1 = A + D\,C$, $a_2 = A - D\,C$,

$B = (b_1 + b_2)/2$, $b_1 = B - C$, $b_2 = B + C$,

$C = (b_2 - b_1)/2$, $\xi = \mathrm{D}$

$D = \xi$

2. Model: lineární-kvadratické větve:

Regresní model: $y = A + B\,x + C\,x^2 + (x - D)\,\text{sign}\,(x - D)[C(x + D) + E]$

Rovnice: $y = a_1 + b_1\,x,\ x \# \xi$

$y = a_2 + b_2\,x + c_2\,x^2,\ x > \xi$

Odhadované parametry:

$A = (a_1 + a_2)/2$, $a_1 = A + D\,C\,2 + D\,E$, $a_2 = A - D\,C\,2 - D\,E$,

$B = (b_1 + b_2)/2$, $b_1 = B - E$, $b_2 = B + E$,

$C = c_2/2$, $\xi = D$, $c_2 = 2C$

$D = \xi$

$E = (b_2 - b_1)/2$

3. Model: kvadratické-lineární větve:

Regresní model: $y = A + B\,x + C\,x^2 + (x - D)\,\text{sign}\,(x - D)[E - C(x + D)]$

Rovnice: $y = a_1 + b_1\,x + c_1\,x^2,\ x \# \xi$

$y = a_2 + b_2\,x,\ x > \xi$

Odhadované parametry:

$A = (a_1 + a_2)/2$, $a_1 = A - D\,C\,2 + D\,E$, $a_2 = A + D\,C\,2 - D\,E$,

$B = (b_1 + b_2)/2$, $b_1 = B - E$, $b_2 = B + E$,

$C = c_1/2$, $\xi = \mathrm{D}$, $c_1 = 2C$

$D = \xi$

$E = (b_2 - b_1)/2$

4. Model: kvadratické-kvadratické větve:

Regresní model: $y = A + B\,x + C\,x^2 + (x - D)\,\text{sign}\,(x - D)[E\,(x + D) + F]$

Rovnice: $y = a_1 + b_1\,x + c_1\,x^2,\ x \# \xi$

$y = a_2 + b_2\,x + c_2\,x^2,\ x > \xi$

Odhadované parametry:

$A = (a_1 + a_2)/2$, $a_1 = A - E\,D\,2 + D\,F$, $a_2 = A + E\,D\,2 - D\,F$,

$B = (b_1 + b_2)/2$, $b_1 = B - F$, $b_2 = B + F$,

$C = (c_1 + c_2)/2$, $\xi = D$

$D = \xi$, $c_1 = C - E$, $c_2 = C + E$

$E = (c_2 - \mathrm{c}_1)/2$

$F = (b_2 - b_1)/2$

5. Model: lineární-lineární-lineární větve:

Regresní model: $y = A + B\,x + C\,(x - D)\,\text{sign}\,(x - D) + E\,(x - F)\,\text{sign}\,(x - F)$

Rovnice: $y = a_1 + b_1\,x,\ x < \xi_1$

$y = a_2 + b_2\,x,\ \xi_1 < x \# \xi_2$

$y = a_3 + b_3\,x,\ x > \xi_2$

Odhadované parametry:

$A = (a_1 + a_3)/2$, $a_1 = A + DC + EF$, $a_2 = A - DC - EF$, $a_3 = A - DC + EF$,

$B = (b_1 + b_3)/2$, $b_1 = B - C - E$, $b_2 = B + C - E$, $b_3 = B + C + E$

$C = (b_2 - b_1)/2$, $\xi_1 = D$, $\xi_2 = F$

$D = \xi_1$

$E = (b_3 - b_2)/2$
$F = \xi_2$

Vzorová úloha 9.15 *Aplikace postupu úsekové polynomické regrese*

Na datech **úlohy S9.08** je třeba aproximovat body neasociativní závislosti a nalézt oba uzlové body zvratu u tří větví prokládané křivky typu lineární-lineární-lineární větve.
Data: použijeme data úlohy S9.08.
Řešení: Je uveden minimalizační proces postupného zjemňování odhadů neznámých parametrů a postupné snižování hodnoty minimalizované sumy čtverců reziduí U až k dosažení minima U_{min}.

Průběh minimalizačního procesu sumy čtverců reziduí

Iterace	**Suma čtverců Reziduí**	**A**	**B**	**C**	**D**
0	91978.17	13.73333	1.396706	0.0000	32.7044
1	128.7423	-22.38974	1.439365	-1.41563	32.7044
2	98.84669	-18.73867	1.417875	-1.451389	31.90939
3	97.88834	-18.97554	1.433407	-1.477436	31.87549
4	97.88829	-18.97609	1.43342	-1.477446	31.87606

Dosaženo konvergenčního kritéria.

Po dosažení minima sumy čtverců reziduí je tištěna tabulka nejlepších odhadů stanovovaných parametrů regresního modelu všech tří větví prokládané křivky a hodnoty bodů zvratu první a druhé větve a druhé a třetí větve křivky.

Odhady parametrů modelu

Parameter	**Odhad parametru**	**Asympt. směr. odch.**	**Dolní mez 95% i. s.**	**Horní mez 95% i. s.**
A	-18.97609	2.371069	-23.79468	-14.1575
B	1.43342	4.792372E-02	1.336028	1.530813
C	-1.477446	5.087956E-02	-1.580845	-1.374046
D	31.87606	0.5111715	30.83723	32.91488
E	1.428675	4.940346E-02	1.328275	1.529075
F	55.26949	0.4470693	54.36093	56.17804

Závisle proměnná: y
Nezávisle proměnná: x
Model: y = Linear-Linear-Linear (x)
Koeficient determinace R^2: 0.975292

Rovnice modelu: -18.97609+1.43342*x - 1.477446*(x - 31.87606)*SIGN(x - (31.87606) +
+1.428675*(x - 55.26949)*SIGN(x - 55.26949)

Odhadované parametry:
$y = a_1 + b_1 x$, když $x < \xi_1$,
$y = a_2 + b_2 x$, když $\xi_1 < x \# \xi_2$,
$y = a_3 + b_3 x$, když $x > \xi_2$,
kde $a_1 = 12.89089$, $a_2 = 107.0812$, $a_3 = -50.84307$, $b_1 = 1.482191$, $b_2 = -1.4727$, $b_3 = 1.384649$.

Hodnoty souřadnice x bodů zvratu: $\xi_1 = 31.87606$ $\xi_2 = 55.26949$

Dolní a horní mez 95% intervalu spolehlivosti odhadovaných parametrů byla vyčíslena dle vzorce pro velké

výběry, platící pro více než 25 bodů. **Rovnice modelu** s odhady neznámých parametrů umožňuje predikovat závisle proměnnou *y* pro libovolné hodnoty nezávisle proměnné *x*. **Odhadované parametry** přináší rovnice přímek všech tří větví prokládané křivky a ukazuje na souřadnici *x* obou bodů zvratu $\xi_1 = 31.87606$ a $\xi_2 = 55.26949$.

Asymptotická korelační matice odhadů parametrů

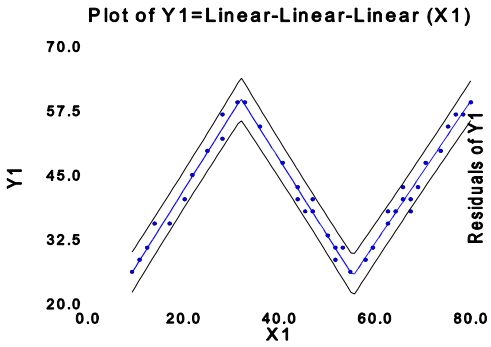
	A	B	C	D	E	F
A	1.000000	-0.859286	0.199576	0.127212	-0.628010	-0.671159
B	-0.859286	1.000000	-0.501307	-0.411324	0.453763	0.463477
C	0.199576	-0.501307	1.000000	-0.110813	-0.543587	0.394458
D	0.127212	-0.411324	-0.110813	1.000000	0.513127	-0.257576
E	-0.628010	0.453763	-0.543587	0.513127	1.000000	0.043351
F	-0.671159	0.463477	0.394458	-0.257576	0.043351	1.000000

Jsou-li korelace vysoké, např. absolutní hodnota korelačního koeficientu je vyšší než 0.95, přesnost parametrů je podezřelá.

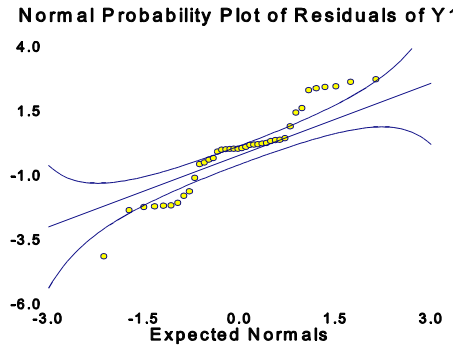
Predikované hodnoty a analýza reziduí

Řádek	***x***	***y***	**Predikovaná hodnota**	**Dolní mez 95 % i. s.**	**Horní mez 95 % i. s.**	**Reziduum**
1	9.119497	26.47059	26.40772	22.50505	30.3104	6.286404E-02
2	10.69182	28.82353	28.73821	24.91603	32.5604	8.531865E-02
3	12.26415	31.17647	31.0687	27.31541	34.82199	0.1077657
4	13.83648	35.88235	33.39919	29.70256	37.09583	2.483162
5	16.98113	35.88235	38.06017	34.43795	41.6824	-2.177817
6	20.12579	40.58823	42.72115	39.11903	46.32327	-2.132915
7	21.69811	45.29412	45.05164	41.43882	48.66445	0.2424791
8	24.84277	50.0000 4	9.71262	46.03753	53.38771	0.2873817
9	27.98742	57.05882	54.3736 5	0.58421	58.16299	2.685227
10	27.98742	52.35294	54.3736 5	0.58421	58.16299	-2.020656
11	31.13208	59.41177	59.03458	55.08336	62.98579	0.3771898
12	32.7044 5	9.41177 5	8.91739 5	4.82736	63.00743	0.4943722
13	35.84906	54.70588	54.28626	50.40296	58.16956	0.4196204
14	40.56604	47.64706	47.33957	43.67289	51.00624	0.3074946
15	43.71069	42.94118	42.70843	39.11507	46.30179	0.2327485
16	43.71069	40.58823	42.70843	39.11507	46.30179	-2.120195
17	45.28302	38.23529	40.39286	36.81327	43.97246	-2.157569
18	46.85535	40.58823	38.0773	34.4959	41.65869	2.510936
19	46.85535	38.23529	38.0773	34.4959	41.65869	0.1579967
20	50.0000	33.52941	33.44617	29.81478	37.07755	0.0832449
21	51.57233	31.17647	31.1306	27.45165	34.80955	0.0458671
22	51.57233	28.82353	31.1306	27.45165	34.80955	-2.307069
23	53.14465	31.17647	28.81503	25.07418	32.55589	2.361433
24	54.71698	26.47059	26.49947	22.68307	30.31587	-2.888087E-02
25	57.86164	28.82353	29.27502	25.44972	33.10032	-0.4514848
26	59.43396	31.17647	31.45214	27.69921	35.20506	-0.2756702
27	62.57862	35.88235	35.80638	32.16622	39.44654	7.597418E-02
28	62.57862	38.23529	35.80638	32.16622	39.44654	2.428914
29	64.15094	38.23529	37.9835	34.38272	41.58427	0.2517979
30	65.72327	40.58823	40.16063	36.58738	43.73388	0.4276057

31	65.72327	42.94118	40.16063	36.58738	43.73388	2.780549
32	67.2956	40.58823	42.33775	38.77988	45.89562	-1.749516
33	67.2956	38.23529	42.33775	38.77988	45.89562	-4.102455
34	68.86793	42.94118	44.51487	40.96008	48.06966	-1.573693
35	70.44025	47.64706	46.69199	43.12796	50.25602	0.9550684
36	73.58491	50.0000	51.04623	47.42723	54.66524	-1.046234
37	75.15723	54.70588	53.22335	49.55917	56.88755	1.482527
38	76.72956	57.05882	55.40048	51.67984	59.12111	1.658346
39	78.30189	57.05882	57.5776	53.78976	61.36543	-0.5187755
40	79.87421	59.41177	59.75472	55.88948	63.61995	-0.3429533



Obr. 9.19a Proložení tří lineárních větví křivky úsekovou regresí a hledání 2 bodů zvratu.



Obr. 9.19b *Q-Q* graf při analýze reziduí.

Závěr: Proložení tří lineárních větví nalezeným regresním modelem zadanými body je dostatečně těsné. Současně byly nalezeny i oba body zvratu ξ_1 a ξ_2.

Vzorová úloha 9.16 *Určení bodu ekvivalence u dvou větví titrační křivky*

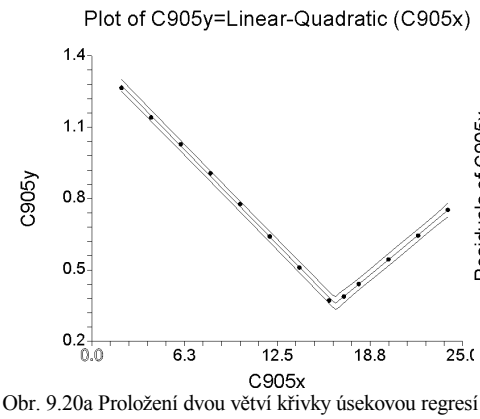
Na datech **úlohy C9.05** je třeba určit bod ekvivalence, čili uzlový bod zvratu dvou větví titrační křivky v instrumentální analýze. Protože však okolí bodu ekvivalence může být i nelineárního (zakřiveného) charakteru, je třeba vyšetřit, zda lze experimentálními body titrační závislosti aproximovat model s větvemi lineární-lineární, lineární-kvadratickou, kvadratickou-lineární a kvadratickou-kvadratickou. Titrační křivka se týká konduktometrické titrace 0.1 M kyseliny chlorovodíkové titrantem 0.1M hydroxidem sodným.

Data: použijeme data **úlohy C9.05,** kde x je objem přidávaného titračního činidla 0.1M NaOH a $y = (100 - a)/a$ a a je odečtená hodnota délky na odporovém můstku [mm].

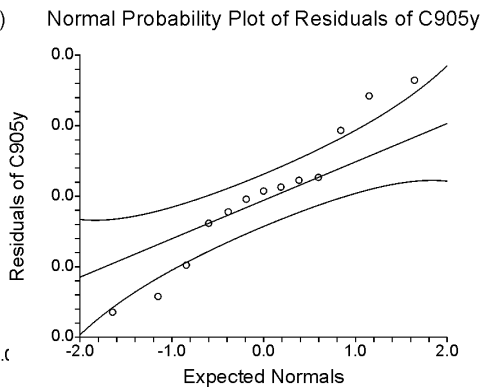
Řešení: Byly testovány čtyři regresní modely a výsledky přináší tabulka. Protože R^2 není dostatečné rozlišovací kritérium mezi testovanými modely, byla dána přednost grafické analýze klasických reziduí kvantil-kvantilovým *Q-Q* grafem. Tento graf ověřuje normalitu reziduí, které je dosaženo jedině v případě správného regresního modelu.

Typ větví modelu regrese	Bod zvratu [ml]	Dolní mez 95% i. s. [ml]	Horní mez 95% i. s. [ml]	100 R^2	Model je

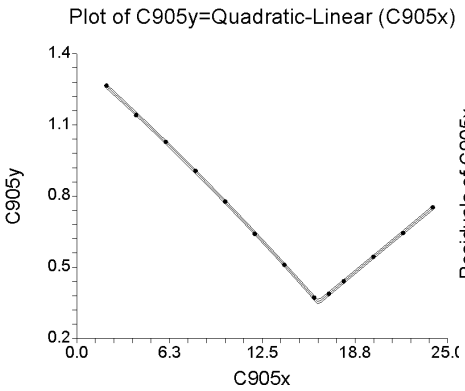
Lineární-lineární	16.414	16.239	16.588	99.935	Zamítnut
Lineární-kvadratická	16.402	16.152	16.652	99.935	Zamítnut
Kvadratická-lineární	16.285	16.214	16.355	99.993	Přijat
Kvadratická-kvadratická	16.273	16.174	16.371	99.993	Přijat



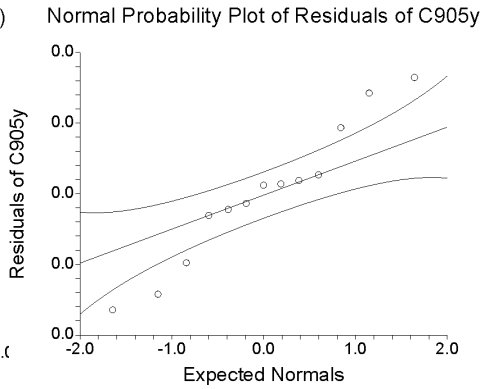
Obr. 9.20a Proložení dvou větví křivky úsekovou regresí typu lineární - lineární a hledání bodu zvratu.



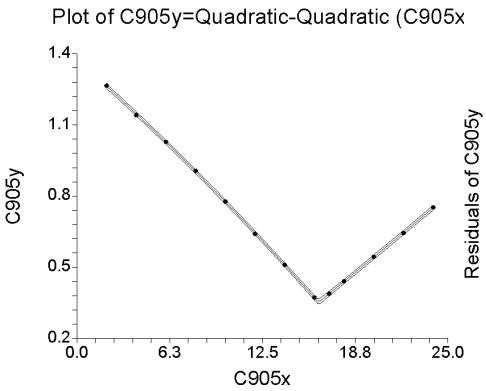
Obr. 9.20b *Q-Q* graf při analýze reziduí.



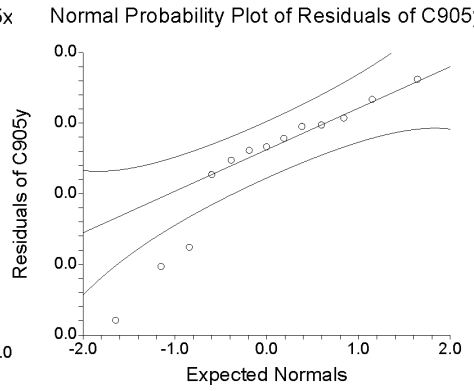
Obr. 9.21a Proložení dvou větví křivky úsekovou regresí typu lineární - kvadratická a hledání bodu zvratu.



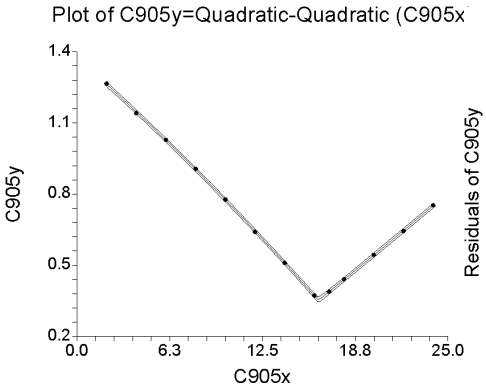
Obr. 9.21b *Q-Q* graf při analýze reziduí.



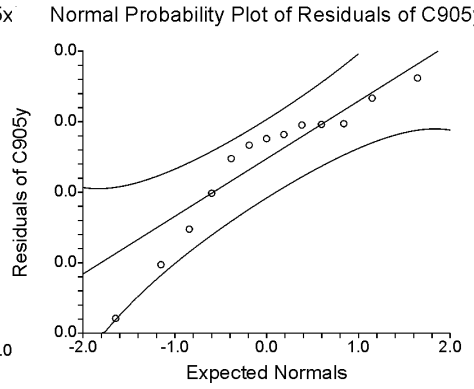
Obr. 9.22a Proložení dvou větví křivky úsekovou regresí typu kvadratická - lineární a hledání bodu zvratu.



Obr. 9.22b *Q-Q* graf při analýze reziduí.



Obr. 9.23a Proložení dvou větví křivky úsekovou regresí typu kvadratická - kvadratická a hledání bodu zvratu.



Obr. 9.23b *Q-Q* graf při analýze reziduí.

Závěr: Nejlepší regresní model je model s větvemi kvadratická-lineární s bodem ekvivalence (16.29 ± 0.07) ml.

9.5 Numerické vyhlazování

Účelem numerického vyhlazování je odstranění náhodných šumů g_i. Požaduje se, aby vyhlazující funkce *g(x)* měla pouze obecné vlastnosti, jako je spojitost ve zvoleném počtu derivací. Nejde v pravém slova smyslu o nalezení konkrétního funkčního tvaru aproximující funkce *g(x),* ale o nalezení *rekonstruované bezšumové závislosti* především k zobrazení, numerické derivaci a integraci. Pro numerické vyhlazení lze použít především *spline vyhlazování*, pro které je charakteristické, že uzlové body ξ_i jsou totožné s souřadnicemi x zadaných experimentálních dat $\{x_i , y_i\}$ $i = 1, ..., n$. S vyhlazujícími spline úzce souvisí tzv. *neparametrická regrese*, kdy funkce *g(x)* je vhodně vážená lineární kombinace veličin y_i , $i = 1, ..., n$, s vahami závislými na vzdálenostech $x - x_i$.

V řadě případů postačuje pouze nalezení posloupnosti vyhlazených hodnot $g(x_i)$

z původních hodnot y_i. Pro ekvidistantní dělení experimentálních bodů na ose x, kdy $h_i = x_{i+1} - x_i$ = konst., $i = 1, ..., n - 1$, jde o úlohu *číslicové filtrace*. Ta je vhodná pro předzpracování signálů z různých měřicích přístrojů. V technické praxi je výhodné využití těchto postupů všude tam, kde nezávisí na typu vyhlazovací funkce $g(x)$, a kde je třeba získat závislost s odstraněnou šumovou složkou (chybami ε_i). Často jde o úlohy derivace nebo integrace dat zatížených náhodnými chybami.

9.5.1 Spline vyhlazování

Při konstrukci vyhlazujících spline se vychází z požadavku, aby se funkce $g(x)$ přibližovala co nejvíce k experimentálním datům, a to ve smyslu zvolené L_p-normy (kap. 6). Jsou-li chyby ε_i nezávislé a stejně rozdělené náhodné veličiny s konstantním rozptylem, je vhodné volit L_2-normu, vedoucí ke kritériu metody nejmenších čtverců

$$U(g) = \sum_{i=1}^{n} w_i \, [y_i - g(x_i)]^2 \ ,$$

kde w_i jsou váhy jednotlivých bodů, závislé na jejich "přesnosti" vyjádřené např. přes rozptyly. Dalším požadavkem je, aby vyhlazující funkce byla dostatečně hladká a spojitá ve zvoleném počtu derivací. Omezme se na nejčastější případ, kdy se požaduje, aby $g(x)$ byla dvakrát diferencovatelná, tzn. ze třídy $C^2[a, b]$, kde $a = x_1$ a $b = x_n$. Jak bylo uvedeno v odd. 9.2, je možno kritérium hladkosti vyjádřit integrálem

$$I(g) = \int_a^b [g^{(2)}(x)]^2 \, \mathrm{d}x \ ,$$

kde $g^{(2)}(x)$ je druhá derivace vyhlazující funkce. Integrál $I(g)$ souvisí s normou druhé derivace a označuje se jako *míra hladkosti v křivosti funkce g(x)*. Účelem je nalézt takovou funkci $g(x)$, která by měla dostatečně malou hodnotu $U(g)$, tj. byla v blízkosti experimentálních dat a přitom měla malou hodnotu $I(g)$, tj. byla dostatečně hladká a její průběh by neměl být nadměrně zvlněný. Pro stanovení optimální vyhlazující funkce $g(x)$ můžeme sestavit dvě základní minimalizační úlohy:

I. Jde o minimalizaci modifikovaného součtu čtverců odchylek

$$K_1 = U(g) + \alpha \, I(g) \ ,$$

kde $0 \leq \alpha \leq 4$ je *parametr vyhlazení*, který "řídí" poměr mezi hladkostí $g(x)$ a jejím přiblížením k experimentálním bodům.

II. Při splnění podmínky $U(g) = S$ se hledá vyhlazující funkce $g(x)$ minimalizující integrál $I(g)$. Pro dané S existuje takový *parametr vyhlazení* $\alpha = \alpha(S)$, že řešení úlohy I je zároveň i řešením úlohy II. Důvodem zavedení úlohy II je fakt, že parametr S má význam reziduálního součtu čtverců a souvisí přímo s odhadem rozptylu náhodných chyb ε.

V řadě prací bylo odvozeno, že funkce $g(x)$ ze třídy $C^2[a, b]$, která pro dané α minimalizuje K_1 z výše uvedené rovnice, má následující vlastnosti[22]:

1. Funkce $g(x)$ je na každém intervalu $I_j \in [x_{i+1}, x_i]$ polynomem třetího stupně.
2. Ve všech místech x_i čili uzlových bodech je funkce $g(x)$ spojitá ve funkčních

hodnotách a hodnotách prvních dvou derivací, což se zapíše rovnicí

$$g^{(k)}(x_i^{-}) = g^{(k)}(x_i^{+}), \quad i = 2, \ldots, n-1, \quad k = 0, 1, 2,$$

kde $g(x_i^{+})$ je limita v bodě x_i zprava a $g(x_i^{-})$ je limita v bodě x_i zleva.

3. Ve třetí derivaci je vyhlazující funkce nespojitá a platí pro ni, že

$$g^{(3)}(x_i^{+}) - g^{(3)}(x_i^{-}) = \frac{w_i}{\alpha}\,[y_i - g(x_i)] .$$

4. V rozmezí $(-\infty, a]$ a $[b, \infty)$ jsou druhé derivace $g^{(2)}(x) = 0$. To znamená, že funkce $g(x)$ je mimo interval (a, b) lineární.

Všechny funkce vyhovující těmto podmínkám jsou kubické spline $S_3(x)$ s uzly x_i, které jsou při znalosti parametru vyhlazení α jednoznačně určeny podmínkou 3, nahrazující klasickou podmínku interpolace u interpolačních spline. Pokud se v rovnici $I(g)$ nahradí druhá derivace m-tou derivací, vychází jako optimální $g(x)$ ve třídě $C^{2m-2}[a, b]$ polynomický spline $S_{2m-1}(x)$ stupně $(2m - 1)$. Podmínka pak platí pro $(2k - 1)$. derivaci.

Místo kritéria metody nejmenších čtverců $U(g)$ lze použít i jiná kritéria. Pokud se místo čtverců použije pomaleji rostoucí funkce (viz kap. 6), rezultují *robustní vyhlazující spline*. Podobně jako při spline interpolaci, lze i zde řídit hladkost vyhlazující funkce zavedením *parametrů napětí* ρ_i. Pro vyhlazující spline pod napětím lze místo rovnice z podmínky 3 psát

$$[g^{(3)}(x_i^{+}) - \rho_i\, g^{(1)}(x_i^{+})] - [g^{(3)}(x_i^{-}) - \rho_i\, g^{(1)}(x_i^{-})] =$$

$$= \frac{w_i}{\alpha}\,[y_i - g(x_i)].$$

Detailní postup konstrukce vyhlazujících spline pod napětím je uveden v práci[23].

Ukažme postup konstrukce kubického vyhlazujícího spline, jež minimalizuje rovnici K_1 při známé hodnotě *parametru vyhlazení* α. Nejjednodušší je hledat vyhlazené funkční hodnoty $g(x_i) = g_i$ a druhé derivace $g^{(2)}(x_i) = M_i$. Při znalosti těchto hodnot je možné určit koeficienty kubických polynomů, které procházejí body $\{x_i, g_i\}$, $i = 1, \ldots, n$. V odd. 9.2 bylo ukázáno, že pro druhé derivace M_i u klasických interpolačních spline platí soustava lineárních rovnic

$$M_1 = 0, \quad M_n = 0,$$

$$\frac{h_{j-1}}{6}\,M_{j-1} + \frac{h_j + h_{j-1}}{3}\,M_j + \frac{h_j}{6}\,M_{j+1} =$$

$$= \frac{g_{j+1}}{h_j} + g_j\left(-\frac{1}{h_j} - \frac{1}{h_{j-1}}\right) + \frac{g_{j-1}}{h_{j-1}}, \qquad j = 2, \ldots, n-1.$$

V maticovém zápisu má tato soustava tvar $\boldsymbol{A}\,\boldsymbol{M} = \boldsymbol{D}\,\boldsymbol{g}$, kde matice $\boldsymbol{A}$ i matice $\boldsymbol{D}$ jsou

tridiagonální a vektor $\boldsymbol{g} = (g_1, ..., g_n)$ je vektor vyhlazených hodnot. Matice $\boldsymbol{A}$ má prvky

$$A_{1,1} = A_{n,n} = 1, \qquad A_{i,i\&1} = \frac{h_{i\&1}}{6}, \qquad A_{i,i} = \frac{h_i + h_{i\&1}}{3}, \qquad A_{i,i+1} = \frac{h_i}{6} .$$

Ostatní prvky této matice jsou nulové. Matice $\boldsymbol{D}$ má první a poslední řádek složen ze samých nul. Dále je

$$D_{i,i\&1} = \frac{1}{h_{i\&1}}, \qquad D_{i,i} = \left(\frac{-1}{h_i} - \frac{1}{h_{i\&1}} \right), \qquad D_{i,i+1} = \frac{1}{h_i} .$$

Ostatní prvky této matice jsou opět nulové. S využitím faktu, že třetí derivace kubického spline v intervalu I_j je rovna

$$g^{(3)}(x) = \frac{M_{i+1} - M_i}{h_i} ,$$

je možno podmínku 3. vyjádřit ve tvaru $\boldsymbol{g} = \boldsymbol{y} - \boldsymbol{v}\,\boldsymbol{D}^{\mathrm{T}}\,\boldsymbol{M}$, kde $\boldsymbol{v} = (\alpha/w_1, ..., \alpha/w_i, ..., \alpha/w_n)^{\mathrm{T}}$ je vektor parametrů vyhlazení a $\boldsymbol{y} = (y_1, ..., y_n)^{\mathrm{T}}$ je vektor původních nevyhlazených hodnot. Po dosazení za $\boldsymbol{g}$ můžeme nalézt vektor druhých derivací $\boldsymbol{M}$ ze soustavy lineárních rovnic

$$(\boldsymbol{A} + \boldsymbol{D}\,\boldsymbol{v}\,\boldsymbol{D}^{\mathrm{T}})\,\boldsymbol{M} = \boldsymbol{D}\,\boldsymbol{y} .$$

Soustava je pětidiagonální a má jednoznačné řešení. K určení $\boldsymbol{M}$ lze použít kompaktních algoritmů podobně jako u spline interpolace. Při znalosti hodnot druhých derivací **M** lze určit vyhlazené hodnoty **g** přímo dosazením dle vztahu

$$g_i = y_i - v_i\,L_i , \quad i = 1, ..., n ,$$

kde $$L_1 = \frac{M_2 - M_1}{h_1} , \quad L_n = \frac{M_n - M_{n\&1}}{h_{n\&1}}$$

a $$L_i = \frac{1}{h_i}\,(M_{i+1} - M_i) + \frac{1}{h_{i\&1}}\,(M_i - M_{i\&1}) , \quad i = 2, ..., n - 1 .$$

Z této rovnice plyne, že vyhlazující spline je lineární odhad, protože platí

$$\boldsymbol{g} = \boldsymbol{H}(\alpha)\,\boldsymbol{y} .$$

Matice $\boldsymbol{H}(\alpha)$ vyjde dle vztahu

$$\boldsymbol{H}(\alpha) = [\boldsymbol{E} + \boldsymbol{v}\,\boldsymbol{D}^{\mathrm{T}}\,\boldsymbol{A}^{-1}\,\boldsymbol{D}]^{-1} .$$

Reinsch[25] určil matici $\boldsymbol{E} - \boldsymbol{H}(\alpha)$ v kompaktním tvaru

$$\boldsymbol{E} \;\&\; \boldsymbol{H}(\alpha) \;\;'\;\; \boldsymbol{Q}\,(\boldsymbol{Q}^{\mathrm{T}}\,\boldsymbol{Q} \;\%\; p\;\boldsymbol{T})^{\&1}\;\boldsymbol{Q}^{\mathrm{T}},$$

kde $\boldsymbol{Q}$ je $n\times(n-2)$ dimenzionální tridiagonální matice, která vznikne vynecháním prvního a posledního sloupce matice $\boldsymbol{D}^{\mathrm{T}}$. Matice $\boldsymbol{T}$ je $(n-2)\times(n-2)$ dimenzionální tridiagonální matice, která vznikne vynecháním prvního a posledního sloupce i řádku matice $\boldsymbol{A}$ a vynásobením ostatních prvků matice $\boldsymbol{A}$ dvěma. Parametr $p = 1/\alpha$ platí pro případ stejných vah $w_i = 1$, $i = 1, ..., n$. Matice $\boldsymbol{H}(\alpha)$ má řadu vlastností shodných s projekční maticí $\boldsymbol{H}$, definovanou v kap. 6. Pro její diagonální prvky platí dle Eubanka[26] $0 \# H_{ii}(\alpha) \# 1$ a mimodiagonální prvky jsou

$$\&1 \# H_{ij}(\alpha) \# 1 \quad \text{pro } j \ldots i \;.$$

Navíc platí, že $\sum_{i\,1}^{n} H_{ij}(\alpha) \;'\; 1$. Podobně jako u klasické projekční matice, je i zde $H_{ii}(\alpha) = 1$, pokud $H_{ij} \ldots 0$ pro všechna $i \ldots j$. Chování matice $\boldsymbol{H}(\alpha)$ souvisí úzce s projekční maticí $\boldsymbol{H}^{*}$ pro regresní přímku (viz kap. 6). Projekční matice $\boldsymbol{H}^{*}$ závisí pouze na hodnotách x_i , $i = 1, ..., n$. Platí, že

A. Pro $\alpha \to 0$: $H_{ii}(\alpha) \to 1$ a $H_{ij}(\alpha) \;'\; 0$. Dále plyne, že v tomto případě jsou $y_i = g(x_i)$. Vyhlazující funkce $g(x)$ je totožná s klasickým interpolačním kubickým spline $S_3(x)$, který je nejhladší.

B. Pro $\alpha \to \infty$: $H_{ii}(\alpha) \to H_{ii}^{(}$ a $H_{ij}(\alpha) \;'\; H_{ij}^{(}$. Vyhlazující funkce $g(x)$ je totožná s regresní přímkou aproximující experimentální body ve smyslu nejmenších čtverců odchylek. Pro případ, že se použije m-tá derivace a $\alpha \to \infty$, je výsledkem regresní polynom stupně $(m - 1)$. Vyhlazující funkce se proto označují jako *zobecněné polynomické regresní modely*.

Späthův algoritmus. Späth[24] použil při konstrukci algoritmu pro vyhlazující kubický spline postup, který vychází z rovnice pro $\boldsymbol{g}$. Vyhlazující spline vyjádřil ve tvaru lokálních kubických polynomů a k řešení pětidiagonální soustavy lineárních rovnic využil kompaktní varianty Choleského metody. V programu SPÄTH jsou použity *lokální parametry vyhlazení* $\beta_i = w_i/\alpha_i$, takže platí

(a) pro $\beta_i \to 0$, $i = 1, ..., n$, je potlačena podmínka hladkosti a rezultuje funkce $g(x)$ jako regresní přímka;

(b) pro $\beta_i \to \infty$ prochází vyhlazující spline bodem $\{x_i , y_i\}$. Pokud je $\beta_i \to \infty$, $i = 1, ..., n$, rezultuje funkce $g(x)$ jako kubický interpolační spline $S_3(x)$.

Volbou β_i lze proto řídit lokální přiblížení vyhlazující funkce k experimentálním bodům.

Vzorová úloha 9.17 *Vyhlazování píku algoritmem SPÄTH*

Využitím Späthova algoritmu určete vyhlazující spline pro instrumentální data píku

za předpokladu, že:

a) váhy $\beta_i = 1$, $i = 1, ..., n$.

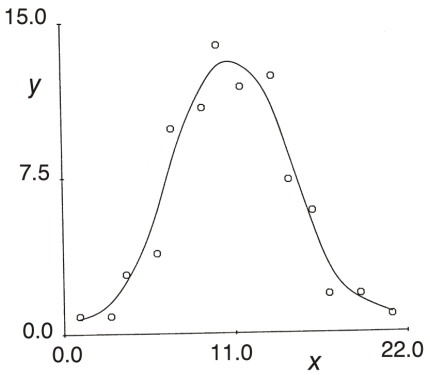
b) váhy $\beta_7 = \beta_9 = 100$ a ostatní $\beta_i = 1$, tj. případ, kdy má vyhlazující funkce procházet body č. 7 a č. 9.

Data: jsou uvedena v následující tabulce $\{x, y\}$.

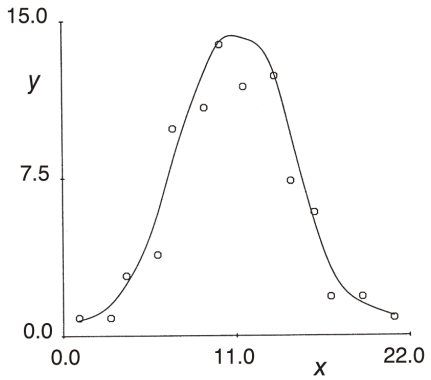
Řešení: Vyhlazující spline pro případ (a) je spolu s experimentálními body znázorněn na obr. 9.24 a pro případ (b) na obr. 9.25. V tabulce jsou uvedeny hodnoty vyhlazené funkce, první a druhé derivace a integrálu v jednotlivých bodech pro případ (a).

Vyhlazení derivace a integrál v zadaných bodech, kde $I(x_i) = \int_a^{x_i} g(x)\,dx$

x_i	y_i	$g(x_i)$	$g^{(1)}(x_i)$	$g^{(2)}(x_i)$	$I(x_i)$
1	1	0.729	0.210	0	0
3	1	1.511	0.753	0.543	2.058
4	3	2.496	1.176	0.304	4.026
6	4	5.809	2.314	0.834	11.951
7	10	8.282	2.375	-0.710	18.991
9	11	11.845	1.304	-0.361	39.475
10	14	12.857	0.608	-1.081	51.884
11.5	12	12.875	-0.407	-0.323	71.374
13.5	12.5	10.877	-1.36	-1.129	95.61
14.5	7.5	8.656	-2.379	0.091	105.42
16	6	5.226	-2.17	0.187	115.79
17	2	3.289	-1.564	1.025	120.0
19	2	1.61	-0.417	0.122	124.51
21	1	0.939	-0.295	0	127.024



Obr. 9.24 Kubický vyhlazovací spline ($\beta_i = 1$, $i = 1, ..., n$).



Obr. 9.25 Kubický vyhlazovací spline ($\beta_7 = \beta_9 = 100$; $\beta_i = 1$ jinde).

Závěr: Volbou parametrů β_i lze měnit jak globální, tak i lokální vyhlazení dle předběžných znalostí o vyhlazované závislosti.

Reinschův algoritmus: Reinsch[25] řešil minimalizaci *I(g)* za podmínky *U(g) = S,* a to využitím metody Lagrangeových multiplikátorů, což vede k minimalizaci funkcionálu

$$K_2 \;'\; I(g) \;\%\; p\;(U(g) \;\%\; Z^2 \;\&\; S) \quad ,$$

kde p je Lagrangeův multiplikátor a Z je pomocná proměnná. Minimalizace funkcionálu K_2 vede ke kubickému vyhlazujícímu spline, což je pro známé p úloha hledání řešení soustavy lineárních rovnic s pětidiagonální maticí koeficientů. Optimální p pro zadané S se hledá iterativním řešením nelineární rovnice Newtonovou metodou. Přesto, že je tento algoritmus komplikovanější, je v praxi rozšířenější. V programu Reinsche[25] se vedle hodnoty S zadávají i váhy jednotlivých bodů w_i . Platí, že

a) čím je S větší, tím více se vyhlazující funkce $g(x)$ blíží k regresní přímce,

b) čím jsou váhy w_i větší, tím více se vyhlazující funkce $g(x)$ blíží k experimentálním bodům.

Vzorová úloha 9.18 *Vyhlazování píku algoritmem REINSCH*

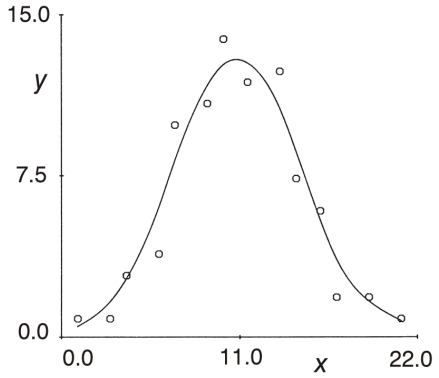
Využitím Reinschova algoritmu určete vyhlazující spline pro data uvedená ve vzorové úloze 9.17. Vypočítejte i hodnoty prvních dvou derivací a integrálu ve všech bodech x_i

Data: jsou uvedena v následující tabulce $\{x, y\}$.

Řešení: Na základě předběžných experimentů bylo zvoleno $S = 18.72$. Výsledný vyhlazující spline je spolu s experimentálními body znázorněn na obr. 9.26. V tabulce jsou uvedeny hodnoty vyhlazené funkce, prvních dvou derivací a integrálu.

Vyhlazení, derivace a integrál v zadaných bodech, kde $I(x_i) \;'\; \int_a^{x_i} g(x)\;\mathrm{d}x$

x_i	y_i	$g(x_i)$	$g^{(1)}(x_i)$	$g^{(2)}(x_i)$	$I(x_i)$
1	1	0.505	0.43	0.0	0
3	1	1.653	0.862	0.432	2.014
4	3	2.72	1.26	0.363	4.168
6	4	6.088	2.093	0.47	12.648
7	10	8.226	2.144	-0.368	19.775
9	11	11.697	1.285	-0.492	39.985
10	14	12.675	0.61	-0.859	52.228
11.5	12	12.743	-0.439	-0.54	71.488
13.5	12.5	10.634	-1.745	-0.765	95.30
14.5	7.5	8.623	-2.159	-0.062	104.96
16	6	6.435	-2.159	0.256	115.479
17	2	3.526	-1.527	0.716	119.969
19	2	1.669	-0.598	0.213	124.954
21	1	0.756	-0.385	0.0	127.308



Obr. 9.26 Kubický vyhlazující spline $S = 18.72$.

Závěr: Při znalosti reziduálního rozptylu, odpovídajícího rozptylu chyb σ^2, lze volit $S = \sigma^2(n - 1)$. Jinak lze použitím Reinschova algoritmu dojít téměř ke stejným výsledkům jako Späthovým algoritmem.

Volba parametru vyhlazení α. Samostatným problémem je *volba parametru vyhlazení* α a parametru S s ohledem na to, aby ve zvoleném statistickém smyslu vyhlazující funkce *g(x)* co nejlépe aproximovala experimentální data. Jsou-li vhodně vybrány váhy w_i , jež odpovídají reciprokým hodnotám rozptylů v jednotlivých bodech, je možno volit parametr S v intervalu

$$(n - 1) - \sqrt{2(n - 1)} \le S \le (n - 1) + \sqrt{2(n - 1)} .$$

Dobré výsledky poskytuje volba $S = n + 1$. K určení optimálního parametru α je nejčastěji používána *střední kvadratická chyba predikce MEP(α),* která je definována vztahem

$$MEP(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - g(x_i))^2}{(1 - H_{ii}(x))^2} .$$

Místo kritéria *MEP(α)* lze užít *zobecněnou střední kvadratickou chybu predikce CEP(α),* kde se nahrazuje $H_{ii}(\alpha)$ střední hodnotou

$$T(\alpha) = \frac{1}{n} \sum_{i=1}^{n} H_{ii}(\alpha) = \frac{1}{n} Tr(\boldsymbol{H}(\alpha)) ,$$

kde *Tr(.)* je stopa matice. Rovnice pro *MEP(α)* pak přechází na tvar

$$CEP(\alpha) = \frac{1}{n} \frac{\sum_{i=1}^{n} (y_i - g(x_i))^2}{(1 - T(\alpha))^2} ,$$

Optimální parametr vyhlazení α je pak takový, pro který nabývá *CEP(α)* své minimální hodnoty. S využitím této rovnice lze kritérium *CEP(α)* vyjádřit pouze jako funkci hodnot $\{x_i, y_i\}$, $i = 1, ..., n$, ve tvaru

$$CEP(\alpha) = \frac{1}{n} \frac{2(\boldsymbol{E} - \boldsymbol{H}(\alpha))\, \boldsymbol{y}2^2}{\left[1 - \frac{1}{n} Tr(\boldsymbol{H}(\alpha))\right]^2} .$$

Postačuje nalézt pouze matice $\boldsymbol{H}(\alpha)$, protože čitatel v této rovnici je reziduální součet čtverců odchylek $e_i = y_i - g(x_i)$, který lze pro daný parametr vyhlazení α snadno vyčíslit. Efektivní postup vyčíslení stopy matice $\boldsymbol{E} - \boldsymbol{H}(\alpha)$ je popsán v Hutchinsonových a de Hoogových pracích[27,28]. Při ekvidistantním dělení bodů na ose x je možno $T(\alpha)$ vyjádřit ve tvaru

$$T(\alpha) = \frac{1}{n} \sum_{i=1}^{n} (1 + \alpha \lambda_i)^{-1} .$$

Konstanty λ_i lze s minimální ztrátou přesnosti vypočítat ze vztahů

$$\lambda_1 = \lambda_2 = 0,$$

$$\lambda_i = \left(\frac{\pi}{n}\right)^4 \frac{(i - 1.5)^4}{h^3}, \qquad i = 3, 4, ..., n.$$

Při znalosti $T(\alpha)$ lze vypočítat i $CEP(\alpha)$ a vhodnou numerickou metodou hledat jeho minimum. Zobecnění tohoto postupu pro libovolné dělení bodů na ose x navrhl Silverman[29]. Veličina $T(\alpha)$ se zde určuje podle vztahu

$$T(\alpha) = \frac{2}{n} + \frac{1}{n} \sum_{i=3}^{n} \left[1 + \frac{\pi^4}{n} \alpha (i - 1.5)^4 c_0\right]^{-1} ,$$

kde konstanta c_0 se počítá z přibližného vzorce

$$c_0 = \left[\int_a^b \hat{f}^{1/4}(t)\, dt\right]^{-4} \cdot \left[\frac{b - a}{32} \sum_{i=1}^{32} \hat{f}^{1/4}(x_j^{(})\right]^{-4} .$$

Zde $f(t)$ je odhad hustoty pravděpodobnosti, určený z hodnot x_i, $i = 1, ..., n$. Postup lze formulovat ve dvou krocích:

(1) Určí se hodnoty $x_j^{(} = a + \frac{(b - a)\,(j - 0.5)}{32}$, $j = 1, ..., 32$. Mezní hodnoty $[a, b]$ se počítají podle vztahů

$$a = x_1 - \frac{x_n - x_1}{n}, \qquad b = x_n + \frac{x_n - x_1}{n} .$$

(2) Vypočtou se odhady hustoty pravděpodobnosti

$$\hat{f}(x_j^{(}) \quad ' \quad \frac{1}{n\ \Delta\ \sqrt{2\pi}}\ \sum_{i'1}^{n} \exp\left[\& \ 0.5 \left(\frac{x_j^{(} \ \& \ x_i}{\Delta} \right)^2 \right] \ .$$

Parametr Δ určuje hladkost odhadu hustoty pravděpodobnosti. Pro praktické případy postačuje volba $\Delta = 1.06\ n^{-1/5}\ s$, kde s je směrodatná odchylka počítaná z hodnot x_i, $i = 1$, ..., n. Uvedený postup je sice přibližný, ale pro praktické učely postačuje. Parametr c_0 nezávisí na α a lze jej určit pouze jednou. V Silvermanově práci[29] je ukázáno, že při volbě $T(\alpha)$ vychází $CEP(\alpha)$ větší než při použití přesného vztahu, rozdíl je však výraznější pouze pro malé α. V jiné Silvermanově práci[31] je uvedeno, jak rozšířit tento přístup i na případ nekonstantních vah w_i. Využitím aproximace prvků $H_{ii}(\alpha)$ ve tvaru[31]

$$H_{ii}(\alpha) \quad . \quad \alpha^{\&1/4}\ n^{\&3/4}\ 2^{\&3/2}\ \hat{f}^{\&3/4}(x_i)$$

je možné konstruovat i *přibližné pásy spolehlivosti predikce*. Pro *95% pásy spolehlivosti* platí

$$L_{1,2}(x_i) \quad ' \quad g(x_i) \pm 1.96\ \hat{\sigma}\ \sqrt{H_{ii}(\alpha)} \quad .$$

Tuto rovnici lze použít pro konstrukci pásů spolehlivosti pro libovolné x. Vlastně to znamená vyčíslit pouze $\hat{f}(x)$. Zbývá ještě nalézt *odhad rozptylu* $\hat{\sigma}^2$. Byl navržen vztah

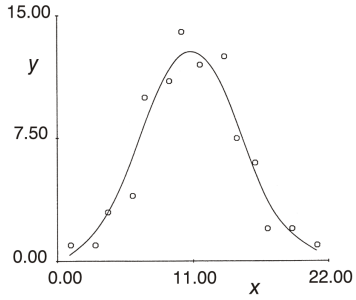
$$\hat{\sigma}^2 \quad ' \quad \frac{\sum_{i'1}^{n} (y_i \ \& \ g(x_i))^2}{n\ (1 \ \& \ T(\alpha))} \quad ,$$

kde $n\ T(\alpha)$ jsou analogicky jako u lineární regrese stupně volnosti, odpovídající vyhlazujícímu spline. Pro určení $\hat{\sigma}^2$ se dosazuje α, jež minimalizuje kritérium $CEP(\alpha)$. Veličina $T(\alpha)$ se pak vyčíslí.

Vzorová úloha 9.19 *Optimální vyhlazení píku*

Nalezněte optimální parametr vyhlazení pro data ze vzorové úlohy 9.17 při užití Silvermanova postupu.

Řešení: Bylo určeno $c_0 = 1.5 \cdot 10^{-5}$. Při použití Späthova algoritmu bylo voleno $w_i = 1$, takže $\beta_i = 1/\alpha$. Pro určení optimálního parametru vyhlazení α byla volena metoda půlení intervalu pro logaritmické dělení ln α. Vyšlo $\alpha = 3.3446$, tj. $\beta_i = 0.2988$, $i = 1$, ..., n. Průběh optimálního vyhlazujícího spline je spolu s experimentálními body zobrazen na obr. 9.27.



Obr. 9.27 Optimální vyhlazující spline.

Závěr: Aproximativní Silvermanův postup poskytuje při své jednoduchosti v praxi použitelné výsledky a je vhodný pro automatizovaný výběr vhodného parametru vyhlazení s využitím počítače.

9.5.2 Neparametrická regrese

Vyhlazující spline je lineární kombinací všech měření. Existuje taková váhová funkce $G_\alpha(x, x_i)$, pro kterou je

$$g(x) \quad ' \quad \frac{1}{n} \sum_{i}^{n} y_i \; G_\alpha(x, x_i) \quad .$$

Váhová funkce závisí na konkrétních hodnotách x_i a parametru vyhlazení α. Předpokládejme, že lokální hustota souřadnic na ose x je $f(x)$, takže počet bodů v intervalu dx je $f(x)$ dx. Za předpokladu, že α není ani příliš veliké, ani příliš malé, a x je dostatečně vzdálené od konců intervalu $[a, b]$, platí podle Silvermana[31], že pro dostatečně veliká n je

$$G_\alpha(x, x_i) \quad . \quad \frac{1}{f(x_i)} \; \frac{1}{\delta(x_i)} \; K\left(\frac{x \; \& \; x_i}{h(x_i)} \right) \quad .$$

Symbolem $K(Z)$ je označena tzv. *jádrová funkce,* která má pro tento případ tvar

$$K(Z) \quad ' \quad \frac{1}{2} \; \exp\left(\frac{\&*Z*}{\sqrt{2}} \right) \; \sin\left(\frac{*Z*}{\sqrt{2}} \; \% \; 14 \; \pi \right) \quad .$$

Parametr $\delta(x_i)$ určuje *lokální vyhlazení* a platí pro něj vztah

$$\delta(x_i) \quad ' \quad \alpha^{1/4} \; n^{\&1/4} \; f^{\&1/4}(x_i) \quad .$$

Z tohoto zápisu vyhlazujícího spline plynou následující důležité závěry[31]:

(a) Vliv bodu $\{x_i , y_i\}$ se projevuje pouze na lokálním chování vyhlazující funkce $g(x)$ pro dostatečně blízká x k hodnotě x_i.

(b) Z poslední rovnice pro $\delta(x_i)$ plyne, že parametr $\delta(x)$ je úměrný čtvrté odmocnině α. Velké změny α se proto příliš neprojeví na velikosti lokálního vyhlazení $\delta(x_i)$.

V dalším výkladu předpokládejme, že souřadnice experimentálních bodů na ose x jsou lineárně transformovány, takže $x_1 = 0$ a $x_n = 1$, a dále platí $x_{i+1} > x_i$, $i = 1, ..., n - 1$.

1. Pro *ekvidistantně rozdělená data* se vyhlazující neparametrický regresní model vyjadřuje ve tvaru

$$p(x) = \frac{1}{n\delta} \sum_{i=1}^{n} y_i \, K\left(\frac{x - x_i}{\delta} \right) .$$

2. Pro *neekvidistantně rozdělená data* se používá modifikovaný vyhlazující neparametrický regresní model

$$p(x) = \sum_{i=1}^{n} y_i \left(\frac{x_i - x_{i-1}}{\delta} \right) K\left(\frac{x - x_i}{\delta} \right) ,$$

kde jádrová funkce *K(Z)* musí mít tyto vlastnosti:

(a) je nezáporná $K(Z) \geq 0$,

(b) je symetrická kolem nuly $K(Z) = K(-Z)$,

(c) má vlastnosti hustoty pravděpobnosti, tj.

$$\int_{-\infty}^{\infty} K(Z)\, dZ = 1 \quad a \quad \int_{-\infty}^{\infty} K^2(Z)\, dZ < \infty .$$

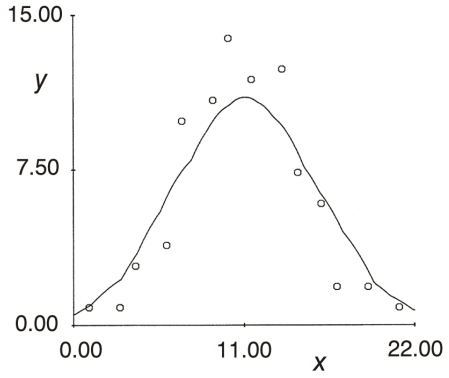
Optimální *K(Z)* s ohledem na minimalizaci střední kvadratické chyby predikce je ve tvaru $K(Z) = 0.75\ (1 - Z^2)_{+}$, kde $(1 - Z^2)_{+}$ je nenulové jen pro $|Z| < 1$. Přehled dalších druhů jádrových funkcí je uveden v práci Bendettiové[32]. K určení *optimálního parametru vyhlazení* δ, označeného jako *šířka pásu*, lze použít jak kritéria *MEP*, tak i kritéria *CEP* nebo řady dalších[33]. Pro modifikovaný vyhlazující neparametrický regresní model *p(x)* má kritérium *střední kvadratické chyby predikce* tvar

$$MEP(\delta) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i \left(1 + \frac{K(0)}{\delta\, n} \right) - \frac{1}{\delta\, n} K\left(\frac{1}{\delta\, n} \right) y_{i+1} - p(x_i) \right]^2 .$$

Vzorová úloha 9.20 *Neparametrická regrese píku*

Nalezněte neparametrický regresní model *p(x)* pro data ze vzorové úlohy 9.17 při využití rovnice modifikovaného vyhlazujícího neparametrického regresního modelu.

Řešení: Na obr. 9.28 je znázorněna modifikovaná neparametrická regrese pro δ = 5.9, které bylo určeno na základě vizuálního porovnání výsledků pro několik hodnot δ.



Obr. 9.28 Neparametrická regrese.

Závěr: Také jednoduchý model neparametrické regrese umožňuje numerické vyhlazení, jež vyhovuje praktickým potřebám.

9.5.3 Číslicová filtrace

Číslicová filtrace umožňuje průběžnou eliminaci šumové složky ve zpracovávaných signálech. V technické praxi se taková úloha vyskytuje při digitalizaci údajů ze zapisovačů u spektrofotometrů, chromatografických přístrojů, polarografů atd. Vychází se z dat y_i, měřených po ekvidistantních, a to obyčejně časových nebo délkových intervalech $s = x_{i+1} - x_i$. Uvažuje se zde aditivní model měření

$$y_i = Z_i^* + \varepsilon_i ,$$

kde Z_i^* jsou skutečné deterministické hodnoty a ε_i jsou náhodné chyby. Použitím číslicové filtrace se získá sekvence filtrovaných hodnot Z_i, které "rekonstruuje" neznámé veličiny Z_i^*:

1. *Lineární číslicový filtr* je možno obecně vyjádřit ve tvaru

$$Z_i = \sum_{j=-k}^{i} c_j \, y_{i-j} + \sum_{j=1}^{i} d_j \, Z_{i-j} ,$$

kde konstanty c_j a d_j určují typ filtru.

2. Pro *nerekurzivní filtry* platí, že všechna $d_j = 0$.

3. Pokud je alespoň jedno $d_j \neq 0$, jde o *filtr rekurzivní*.

Klasické digitální filtry, které jsou náhradou analogových filtrů, jsou *fyzikálně realizovatelné*. Tyto filtry používají pro určení filtrovaných hodnot pouze hodnot y_{i-j} pro $j > 0$, které byly získány až do daného časového okamžiku x_i. Pro tyto filtry je vždy $c_j = 0$ pro všechna $j < 0$. Pokud se při výpočtu filtrované hodnoty používají i "budoucí" údaje y_{i+k}, $k = 1, 2, ...$, jde o *fyzikálně nerealizovatelné filtry* označované jako "smoothers".

Nerekurzivní filtry.

(a) Z nerekurzívních filtrů pro účely předzpracování experimentálních dat

doporučuje Marmet[34] opakované použití *jednoduchého Marmetova filtru*

$$Z_i \;=\; \frac{1}{4}\,(y_{i\&1} \,\%\, 2\,y_i \,\%\, y_{i\%1}) \quad ,$$

(b) Dobré vyhlazovací vlastnosti má *Hippeho filtr*[35]

$$Z_i \;=\; \frac{\&\,(y_{i\&2} \,\%\, y_{i\%2}) \,\%\, 4\,(y_{i\&1} \,\%\, y_{i\%1}) \,\%\, 6\,y_i}{12} \quad ,$$

který byl využit pro předzpracování chromatografických měření.

Rekurzivní filtry. Rekurzívní filtry se obyčejně používají k vyhlazování časových řad a vstupů do číslicových regulátorů.

(a) Nejjednodušší je *exponenciální filtr*

$$Z_i \;=\; K\,y_i \,\%\, (1 \,\&\, K)\,Z_{i\&1} \quad ,$$

kde K je stupeň zesílení filtru $0 < K < 1$.

(b) Mezi rekurzívní patří také *dvoustupňový Holtův filtr*, definovaný vztahy

$$Z_i \;=\; K\,q_i \,\%\, (1 \,\&\, K)\,Z_{i\&1} \quad ,$$

$$q_i \;=\; K\,y_i \,\%\, (1 \,\&\, K)\,q_{i\&1} \quad ,$$

kde K je *konstanta zesílení*. Společnou nevýhodou rekurzívních filtrů je nutnost volby parametru zesílení a dalších konstant.

Robustní nelineární filtry. Pro případ, kdy lze v datech očekávat i hrubé nenáhodné chyby (outliers), jsou vhodné *robustní nelineární filtry*. Jsou to varianty robustních vyhlazovacích metod[36].

(a) Mezi nejjednodušší patří nelineární filtry *L*-typu[37], založené na pohyblivých mediánech. Medián *S(v, i)* lichého stupně je definován vztahem

$$S(v,\, i) \;=\; \mathrm{med}(y_{i\&u},\, \ldots,\, y_i,\, \ldots,\, y_{i\%u}) \quad ,$$

kde $u = (v - 1)/2$ a symbol med(.) označuje střed podle velikosti setříděných hodnot y. Užívá se mediánu třetího stupně ($v = 3$) a pátého stupně ($v = 5$). Mediány lichého stupně lze kombinovat s pohyblivými aritmetickými průměry.

(b) *Jednoduchý filtr 53H* je dán výrazem

$$Z_i \;=\; \frac{S(5,\, i\&2)}{4} \,\%\, \frac{S(5,\, i\&1)}{2} \,\%\, \frac{S(5,\, i)}{4} \quad .$$

K zajištění dokonalejšího vyhlazení se mediány používají opakovaně.

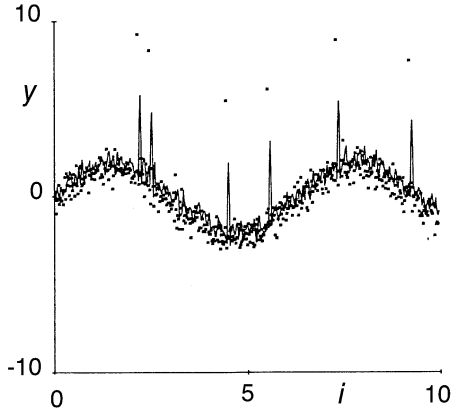
(c) Z této skupiny je nejjednodušší *filtr 3T*, pro který je

$$Z_i \;=\; \mathrm{med}[S(3,\, i\&2),\; S(3,\, i\&1)\;\; S(3,\, i)] \quad .$$

Další varianty pohyblivých mediánů obsahuje Vellemannovy práce[38]. Na základě simulační studie bylo zjištěno, že mezi nejvhodnější patří *filtr 53H*, který je dostatečně robustní

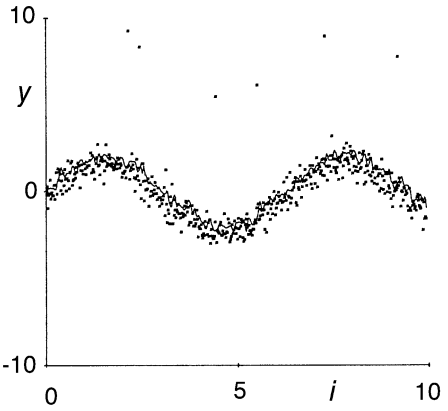
a přitom neposkytuje "nadměrně" vyhlazené úseky.

Vzorová úloha 9.21 *Porovnání vlastností lineárních a nelineárních filtrů*
Na sinusoidálních datech, zatížených jak náhodnými normálně rozdělenými chybami, tak i hrubými chybami, ověřte vlastnosti Hippeho filtru, dále filtru 53H a filtru 3T.
Data: n = 50. Pro stoupající hodnoty i byly generovány hodnoty závisle proměnné y_i dle vzorce $y_i = 2 \sin(i) + 0.5\, N(0, 1) + \delta\, R_i$, kde $N(0, 1)$ jsou náhodné veličiny s normálním rozdělením, nulovou střední hodnotou a jednotkovým rozptylem. R_i je náhodná veličina nabývající hodnot 0 a 1 v závislosti na hodnotách generátoru pseudonáhodných čísel a $\delta = 7.5$.

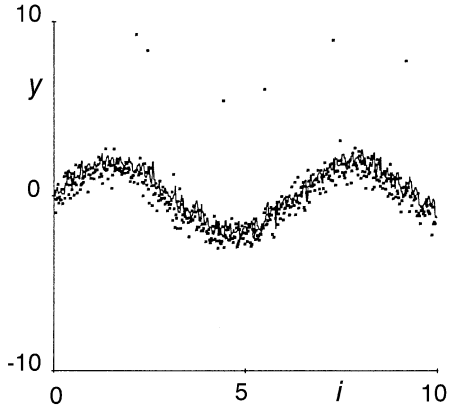


Obr. 9.29 Vyhlazení pomocí Hippeho filtru.

Řešení: Výsledky vyhlazení jsou uvedeny na obrázcích 9.29, 9.30 a 9.31. Pro ilustraci jsou hodnoty Z_i spojeny lineárními úseky. Obr. 9.29 ukazuje vyhlazení pomocí Hippeho filtru. Je patrné, že hrubé chyby způsobují značné překmitávání a ani vyhlazení pro gaussovské chyby není příliš dokonalé. Obr. 9.30 ukazuje vyhlazení pomocí nelineárního filtru 53H. Je patrná necitlivost na přítomnost hrubých chyb. Obr. 9.31 ukazuje vyhlazení pomocí nelineárního filtru 3T. Také zde neovlivňují hrubé chyby proces filtrace. Vznikají však lokální lineární úseky.



Obr. 9.30 Výsledek vyhlazení filtrem 53H.



Obr. 9.31 Výsledek vyhlazení filtrem 3T.

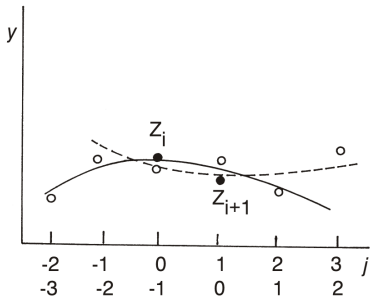
Závěr: Lineární filtry jsou obecně nerobustní. Z nelineárních se jako nejvhodnější jeví filtr 53H, který lze jednoduše zařadit do programů pro předzpracování analytických signálů, pokud lze očekávat výskyt hrubých chyb.

Lineární regresní filtry Sawitzkého-Golaye. Pozornost je věnována také lineárním regresním filtrům. Ty jsou často užívány, např. pod názvem *vyhlazení Sawitzkého-Golaye*[45]. Tyto filtry jsou fyzikálně nerealizovatelné a lze je vyjádřit ve tvaru

$$Z_i = \sum_{j=-N}^{N} c_j \, y_{i-j} \quad ,$$

kde hodnota N označuje *řád filtru* a $(2N+1)$ je *délka filtru*, která určuje počet naměřených dat y_{i-j}, jež byla užita k rekonstrukci hodnoty Z_i. *Filtr stupně d* odpovídá polynomickému regresnímu modelu stupně *d.* Pokud je $d \geq 2N$, platí, že existuje pouze jeden filtr stupně *d,* pro který platí, že $c_0 = 1$ a ostatní $c_k = 0$. To znamená, že $Z_i = y_i$ a nedochází potom k filtraci. V těchto případech prochází polynomický model všemi hodnotami y_{i-j}. Pro $d < 2N$ existuje nekonečně mnoho filtrů řádu N a stupně *d,* a to v závislosti na konkrétních hodnotách $(2N - d)$ nastavitelných parametrů c_k .

Za *lineární regresní filtr* pro kritérium nejmenších čtverců odchylek se označuje takový filtr, kterému odpovídá nejmenší součet čtverců koeficientů c_j , $j = -N, ..., 0, ..., N$. Výsledek filtrace pomocí tohoto filtru odpovídá postupu, kdy je sekvence $2N+1$ bodů $\{j, y_{i-j}\}$, $j = -N, ..., 0, ..., N,$ proložena polynomem stupně d ve smyslu metody nejmenších čtverců a za Z_i se bere hodnota tohoto regresního polynomu v místě $j = 0$. Tento postup se označuje v literatuře jako *pohyblivé nejmenší čtverce* (moving least squares). Schematicky je pro $d = 2$ (tj. regresní paraboly) a $N = 2$ (tj. délka filtru rovna 5) znázorněna na obr. 9.32.



Obr. 9.32 Princip činnosti filtru stupně 2 a délky 5.

Pro realizaci číslicových filtrů lze přímo použít metodu nejmenších čtverců a odhadnout koeficienty regresního polynomu a nalézt predikci (vyhlazenou hodnotu) Z_i. Postup je však jednodušší, protože lze snadno určit koeficienty c_j vzhledem ke speciální volbě souřadnic x.

Pokud je Z_i polynom stupně d a chyby g_i jsou stejně rozdělené nezávislé náhodné veličiny s nulovou střední hodnotou a konstantním rozptylem σ^2, platí v souladu s teorií lineární regrese, že

$$E(Z_i) = Z_i \quad a \quad D(Z_i) = \sigma^2 \sum_{j=-N}^{N} c_j^2 \ .$$

Výsledky Z_i lineárních regresních filtrů jsou nevychýlené odhady s minimálním rozptylem. Hodnota Z_i odpovídá absolutnímu členu b_0 regresního polynomu

$$Z_i = b_0 + \sum_{k=1}^{d} b_k j^k \ ,$$

protože v místě i je $j = 0$. Ostatní koeficienty b_k , $k = 1, 2, ..., d$, pak odpovídají hodnotám první, druhé až d-té derivace dělené faktoriálem 1!, 2! až d!. S ohledem na speciální sekvenci souřadnic nezávisle proměnné $j = -N, ..., 0, ..., N$ platí, že

$$\sum_{j=-N}^{N} j^q = 0 \quad \text{pro} \quad q = 2u + 1, \quad \text{kde} \quad u = 0, 1, 2, ...$$

Důsledkem je, že odhad b_0 pro polynom stupně $2d$ je totožný s odhadem a_0 pro polynom stupně $(2d + 1)$. Regresní filtr sudého stupně je zároveň regresním filtrem větším o jeden lichý stupeň. Thrall[39] ukázal, že pro koeficienty lineárních regresních filtrů c_j , $j = -N, ..., 0, ..., N$, stupně $2d$ platí $c_j = \sum_{q=0}^{d} \alpha_q j^{2q}$. Vektor $\boldsymbol{\alpha} = (\alpha_0, ..., \alpha_d)^T$ je dán řešením soustavy rovnic $\boldsymbol{H}\,\boldsymbol{\alpha} = \boldsymbol{e}$, kde $\boldsymbol{e} = (1, 0, ..., 0)^T$ a $\boldsymbol{H}$ je symetrická matice rozměru $(d + 1)\times(d + 1)$ s prvky

$$\boldsymbol{H}_{ik} = \sum_{j=-N}^{N} j^{2i+2k}, \quad i, j = 0, \ldots, d.$$

Mezi koeficienty regresních filtrů platí řada vztahů, plynoucích přímo z jejich definice. Tyto vztahy umožňují snadnou kontrolu jejich správnosti. Pro koeficienty c_j filtru řádu N a stupně d platí, že

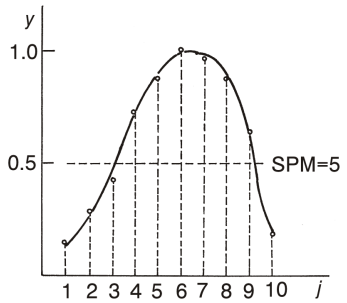
$$\sum_{j=-N}^{N} j^{q} c_j = \begin{cases} 1 & \text{pro } q = 0 \\ 0 & \text{pro } q = 1, \ldots, d \end{cases}.$$

V práci Bromby a Zieglera[40] jsou podrobně rozebrány vlastnosti regresních filtrů. Je ukázáno, že filtry jsou optimální pro signály, které se dají nahradit na délce filtru $(2N + 1)$ Taylorovým rozvojem do stupně d. Stupeň vyhlazení regresním filtrem bude růst s délkou filtru $(2N + 1)$ a klesat s růstem stupně filtru d. Při dostatečné délce filtru a nízkém stupni d lze očekávat i odstranění hrubších chyb.

V technické praxi se často filtruje signál, kde Z^* je ve tvaru píku, tj. jako součet gaussovských nebo lorenzovských křivek. Pro vyhlazení se používá kvadratických nebo kubických filtrů, kdy je $d = 2$. Pro optimální vyhlazení je třeba, aby délka filtru $F = (2N + 1)$ byla menší než *šířka píku v polovině maxima SPM*. Proctor a Sherwood[41] doporučují volit $F \doteq 0.7\ SPM$, kde SPM je udáno v počtech bodů užitých na tuto vzdálenost. Frank[44] však doporučuje volit délku filtru F při filtraci píku podle vztahu

$$F = A \frac{PM}{\Delta},$$

kde $\Delta = x_{i+1} - x_i$ je skutečná vzdálenost mezi filtrovanými hodnotami a PM je *šířka píku v polovině maxima v jednotkách x*. Pro gaussovské píky se doporučuje $A \doteq 1$ a pro lorentzovské $A \doteq 0.7$.



Obr. 9.33 Určení délky filtru při filtraci píku.

Detailní analýza výběru vhodné délky filtru pro různé stupně regresních filtrů je uvedena v práci Bromby a Zieglera[40]. Při konstrukci regresních filtrů postačuje pro zadaná N a d určit koeficienty c_j, $j = -N, \ldots, 0, \ldots, N$. Problémem je, že matice $\boldsymbol{H}$ je pro větší d špatně podmíněná.

Plyne, že vektor $\boldsymbol{c}$ koeficientů regresního filtru lze snadno určit na základě koeficientů $\boldsymbol{\alpha}$ speciálního regresního polynomického modelu[39]

$$\delta_i = \alpha_0 + \sum_{k=1}^{d} f_i^k \alpha_k + \varepsilon_i ,$$

kde $\delta_i = 1$ pro $i = 0$ a $\delta_i = 0$ pro $i \neq 0$. Funkce $f_i = i^2$. Soustava rovnic je pak soustavou normálních rovnic, ze které je $\boldsymbol{\alpha} = \boldsymbol{H}^{-1} \boldsymbol{e} = (\boldsymbol{F}^{\mathrm{T}} \boldsymbol{F})^{-1} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{\delta}$, kde matice $\boldsymbol{F}$ o rozměru $(2N + 1) \times (d + 1)$ má prvky $F_{ij} = i^{2j}$ pro $i = -N, ..., N$ a $j = 0, 1, ..., d$. Místo proměnných i^{2j} je výhodnější použít ortogonálních polynomů pro dané dělení $-N, ..., 0, ..., N$. Thrall[39] nahradil pro velká N tyto polynomy Legendrovými polynomy a odvodil vztahy pro koeficienty regresního filtru stupně $2d$. Pro kubické a kvadratické regresní filtry (d = 2, resp. 3) lze počítat c_j v závislosti na velikosti N podle vztahu[44]

$$c_j = \frac{(3N^2 + 3N - 1) - 5j^2}{(2N - 1)(2N + 1)(2N + 3)/3}$$

nebo přibližně podle Thrallova vztahu[39]

$$c_j \doteq \frac{1}{2N + 1} \left[-\frac{15}{4} \left(\frac{j}{N} \right) + \frac{9}{4} \right] .$$

Pro d = 4, resp. 5, tj. regresní filtr čtvrtého a pátého stupně, je možno použít vztah[44]

$$c_j = \frac{(15N^4 + 30N^3 - 35N^2 - 50N + 12) - 35(2N^2 + 2N - 3)j^2 + 63j^4}{4(2N - 3)(2N - 1)(2N + 1)(2N + 3)(2N + 5)/15} .$$

Regresní filtry lze použít také pro získávání vyhlazených hodnot derivací. Koeficienty kubického filtru pro první derivaci mají tvar

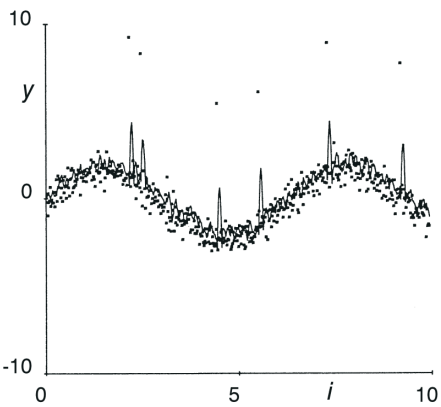
$$c_j^1 = \frac{5[5(3N^4 + 6N^3 - 3N + 1)j - 7(3N^2 + 3N - 1)j^3]}{(2N + 3)(2N + 1)(2N - 1)(N + 2)(N + 1)N(N - 1)} .$$

Analytické vztahy pro případ d = 6, 7 a první až páté derivace jsou uvedeny v práci Maddena[42].

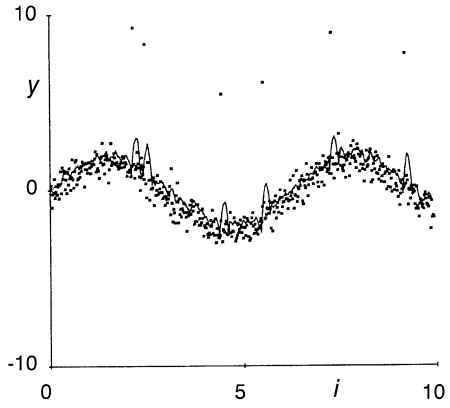
Vzorová úloha 9.22 *Vliv délky regresního filtru na vyhlazující vlastnosti*

Pro data generovaná ve vzorové úloze 9.21 sestrojte číslicový kvadratický filtr (d = 2) o délce $F = 2N + 1 = 7$ a také o délce $F = 13$.

Řešení: Výsledek vyhlazení pro $F = 7$ je znázorněn plnou čarou na obr. 9.34, jež vznikla spojením vyhlazených hodnot lineárními úseky.



Obr. 9.34 Vyhlazení kvadratickým regresním filtrem délky $N = 7$.



Obr. 9.35 Vyhlazení kvadratickým regresním filtrem délky $N = 13$.

Vyhlazení pro $F = 13$ je znázorněno na obr. 9.35.
Závěr: S růstem délky filtru dochází k omezení vlivu hrubých chyb. Na druhé straně však při nadměrném růstu N roste nebezpečné "převyhlazení", vedoucí až k odstranění i nenáhodné lokální změny tvaru.

Regresní filtry neumožňují filtraci prvních N a posledních N bodů. To je ovšem při vyhlazování menšího počtu dat nevýhodné. Obyčejně však postačuje počítat hodnoty prvních a posledních N bodů na základě regresních polynomů pro prvních a posledních ($2N + 1$) bodů. Dosazují se obecně $j = 0$. Pro případ kvadratického regresního filtru byl odvozen[41] pro výpočet $Z(j)$ v intervalu $-N \# j \# N$ vztah

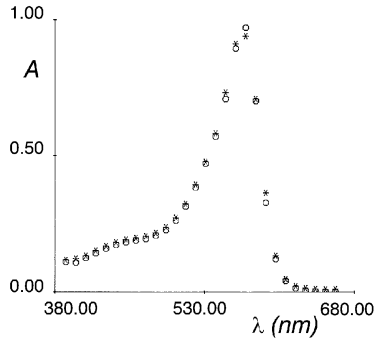
$$Z(k) = \sum_{j=-N}^{N} y_j \left\{ \frac{5(3j^2 - N(N-1))k^2 - (2N+1)2N-3)jk}{N(4N^2-1)(2N-3)(N-1)/3} - \right.$$

$$\left. - \frac{N(N-1)[3N(N-1)-1-5j^2]}{N(4N^2-1)(2N-3)(N-1)/3} \right\}.$$

Pro prvních N a posledních N bodů se počítají hodnoty $Z(k)$ pro různá k, tj. od $-N$ do 0 a od 0 do N. Ostatní body se vyhlazují dle centrální formule $Z(0) = Z_i$. Derivací této rovnice dle k se získá závislost pro určení prvních derivací. Postup využívající této rovnice se označuje *vyhlazení pomocí klouzavých parabol*.

Vzorová úloha 9.23 *Filtrace absorpčního spektra fenolové červeně*

Bylo proměřeno spektrum fenolové červeně při pH = 8.1 v rozsahu vlnových délek 390 až 660 nm s intervalem 10nm. Proveďte vyhlazení pomocí klouzavých parabol.
Data: hodnoty absorbance začínají od 390 nm: 0.1114, 0.1091, 0.1259, 0.1438, 0.1606, 0.1742, 0.1837, 0.1906, 0.1959, 0.2083, 0.2284, 0.2632, 0.3149, 0.3845, 0.4723, 0.5717, 0.7103, 0.8960, 0.9735, 0.7030, 0.3291, 0.1224, 0.0424, 0.0159, 0.0069, 0.0037, 0.0025, 0.0020.



Obr. 9.36 Vyhlazení bodů spektra (kolečka) pomocí klouzavých parabol $N = 2$ (hvězdičky).

Závěr: Kvalita vyhlazení závisí na volbě délky *(2N+1)*.

Při on-line digitální filtraci s využitím jednoduchých programovatelných prostředků je vhodné použít filtrů rekurzivních. Obecný postup konstrukce rekurzivních regresních filtrů z regresních filtrů nerekurzivních je popsán v práci Bromby a Zieglera[43]. Příkladem *rekurzivní verze regresního filtru* pro $d = 2$ tzv. *kvadratického filtru* o délce filtru 15, tj. $N = 7$, je vztah

$$Z_i \;'\; Z_{i\&3} \;\&\; 3\,(Z_{i\&2} \;\&\; Z_{i\&1}) \;\&\; \frac{78}{1105}\,(y_{i\%7} \;\&\; y_{i\&10}) \;\%$$

$$\%\; \frac{221}{1105}\,(y_{i\%6} \;\&\; y_{i\&9}) \;\&\; \frac{153}{1105}\,(y_{i\%5} \;\&\; y_{i\&8})$$

Regresní filtry lze konstruovat poměrně jednoduše, a to s ohledem na použitý výpočetní prostředek. V literatuře se většinou vychází z původní práce Savitzkého a Golaye[45], která však obsahuje chyby v numericky vyčíslených koeficientech c_j. S problematikou regresních filtrů úzce souvisí techniky pohyblivých nejmenších čtverců nebo pohyblivé regrese, kdy se regresní modely určují lokálně pouze z informací o sousedních bodech vyhlazovaného bodu $\{x_i , y_i\}$. Tyto techniky se používají při hledání trendů v rozptylových grafech nebo grafech reziduí[46].

Vzorová úloha 9.24 *Výpočet hustoty kyseliny fosforečné*

Pro laboratorní výpočet je třeba znát hustotu 68%ní kyseliny fosforečné. V tabulkách jsou však uvedeny hustoty s intervalem 5 hm. %. Určete požadovaný údaj využitím spline interpolace.

Data: V rozmezí 60 až 80 hm. % byly z tabulek odečteny následující hodnoty

c [hm. %]	60	65	70	75	80
h 10^{-3} [kg m^{-3}]	1.426	1.475	1.526	1.579	1.633

Řešení: Programem Spline byla určena hustota 68%ní H_3PO_4 rovna 1505 kg. m^{-3}.

Závěr: Program SPLINE lze použít nejenom pro znázornění grafu interpolující funkce, výpočet derivací, resp. integrálu, ale také pro interpolaci v tabulkách.

Vzorová úloha 9.25 *Určení chybějící hodnoty v infračerveném spektru*

Při měření infračerveného spektra methylsulfonylchloridu došlo k výpadku registračního zařízení tiskárny a nebyla zaznamenána hodnota pro vlnočet $\hat{\nu}$= 1165 cm^{-1}. Určete chybějící hodnotu.

Data: n = 10

$\hat{\nu}$ [cm^{-1}]	1160	1161	1162	1163	1164	1166	1167	1168	1169	1170
A	0.0466	0.0539	0.0631	0.0744	0.0883	0.1254	.1482	0.17112	0.1907	0.2023

Řešení: Protože lze očekávat, že naměřené hodnoty absorbance nebudou zcela přesné, byl použit program pro spline regresi, využívající parabolických spline. Stupeň blízkosti modelu k experimentálním bodům se řídí výběrem uzlových bodů. Byla zvolena strategie automatického vkládání uzlů tak, aby mezi nimi byly stejné vzdálenosti (tzv. konstantní uzlové intervaly).

Vliv počtu uzlů kvadratické spline regrese na hodnotu absorbance při vlnočtu $\hat{\nu}$ = 1165 cm^{-1}

Počet uzlů	2	3	4	5	6	7	8
A pro $\hat{\nu}$=1165	0.10462	0.10555	0.10507	0.105288	0.105201	0.105271	0.105264

Závěr: Z tabulky je patrné, že s růstem počtu uzlů se stabilizuje hodnota absorbance. Výhodou spline regrese je její flexibilita, což umožňuje použití i pro komplikované nelineární závislosti.

9.6 Postup při interpolaci a aproximaci

V první fázi je třeba rozhodnout o tom, zda jde o úlohu *interpolace* či *aproximace*. Pro interpolaci platí, že hodnoty y jsou nenáhodné veličiny a aproximující funkce prochází všemi zadanými body. V případě aproximace jsou hodnoty y na ose x zatíženy náhodnými chybami nebo nepřesné a účelem je nalezení aproximující funkce, která je potlačuje:

Postup interpolace a aproximace

1. Interpolace funkcí: podle požadavků na shodu zadané a aproximující funkce lze volit buď klasickou polynomickou, nebo Hermitovskou interpolaci. Polohy uzlů interpolace je vhodné volit podle formule. Pro libovolné dělení uzlů interpolace je výhodnější použití spline interpolace defektu k = 1 (požadavky na shodu ve funkčních hodnotách) nebo defektu k = 2 (požadavky také na shodu v prvních derivacích).

2. Interpolace závislostí: pro tento případ se doporučuje použití spline interpolace vhodného stupně (defektu k = 1) tak, aby byly splněny podmínky spojitosti ve funkčních hodnotách a hodnotách derivací. Volí se třída funkcí C^m a dle toho se vybírá vhodná metoda. Pokud není požadována znalost derivací aproximující funkce, je vhodné použít lokální kubické interpolační postupy C^1 s automatickou úpravou lokální monotónnosti. Tyto metody umožňují poměrně kvalitní rekonstrukci závislosti. Pokud jsou požadovány znalosti první derivace rekonstruované závislosti, je vhodné použití kubických spline pod napětím s lokálním řízením tvaru aproximující funkce.

3. Aproximace funkcí: rozhodující je volba normy S_p. Pro případ p = 2 (metoda nejmenších čtverců) nebo p 6 4 (Čebyševova minimaxní aproximace) a polynomické aproximační modely je postup odhadu parametrů poměrně jednoduchý.

4. Aproximace závislostí: k řešení této úlohy existuje řada možností. Je možné volit mezi následujícími základními postupy:

(a) aproximace polynomy pro zvolenou normu S_p,
(b) spline regrese s volbou stupně spline a uzlových bodů,
(c) úseková regrese,
(d) spline vyhlazování s volbou parametrů vyhlazení,
(e) neparametrická regrese s volbou parametrů vyhlazení,
(f) číslicová filtrace pro ekvidistantní dělení souřadnic x s volbou stupně a délky filtru.

Výběr vhodného postupu zde závisí na cíli zpracování dat. Odstranění náhodných šumů umožňují prakticky všechny postupy a liší se zejména definicí aproximující funkce a její složitostí. Pro účely tvorby empirických modelů vyhovuje obyčejně nejlépe spline regrese. Spline vyhlazování je účelné pro případy, u kterých se data budou následně numericky derivovat nebo integrovat.

Předběžné zpracování naměřených signálů je účelné provádět využitím robustního vyhlazování nebo regresních filtrů. Pro stejný účel je možné použít i neparametrických regresních modelů.

9.7 Úlohy

Metody interpolace, aproximace a vyhlazování se v praxi často používají jako součást předzpracování rozsáhlých datových souborů. Řešte samostatně následující úlohy s využitím interaktivní práce se statistickým softwarem, např. programovým systémem ADSTAT. Vysvětlete jednotlivé diagnostiky a učiňte své závěry o výběru dat. Úlohy jsou rozděleny do dvou kapitol: C9 (chemická a fyzikální data) a S9 (ekonomická a sociologická data). Jelikož řešení úloh této kapitoly je většinou grafického charakteru, nejsou u úloh uváděny *kontrolní hodnoty*.

9.7.1 Analýza chemických a fyzikálních dat

Úloha C9.01 *Kalibrace čidla vlhkosti*
Při kalibraci čidla AVK byly pro předem nastavené vlhkosti, charakterizované rosným bodem T_R [EC], měřeny hodnoty elektrického odporu R [Ω]. Z předběžných experimentů plyne, že lze tuto kalibrační závislost vyjádřit pomocí lomené po částech lineární funkce. Nalezněte využitím C^0-regrese pomocí lineárních spline kalibrační model.
Data: Hodnoty elektrického odporu R [Ω], rosný bod T_R [EC]:

T_R[EC]	-32	-21	-9	-3.5	2	6.5	10	14.5
R[Ω]	87	97	107	112	122	132	142	152

Úloha C9.02 *Kinetika barvení modifikovaných PES vláken*
Byla sledována závislost koncentrace kationtového barviva C [mg. g^{-1}] na modifikovaném PES vlákně v různých časech t [min]. Pro vyjádření kinetiky barvení je třeba určit rychlost dC/dt v časech t = 20 a 90 min. Použijte C^2-regrese pomocí kubických spline a spline

vyhlazování s konstantními vahami.

Data: Koncentrace kationtového barviva C [mg. g^{-1}], čas t [min]:

C [mg. g^{-1}]	0	9.24	12.38	14.54	16.77	17.95
t [min]	0	4	10	30	60	100

Úloha C9.03 *Interpolace závislosti rozpustnosti na teplotě*

Ve fyzikálně chemických tabulkách jsou uvedeny hodnoty rozpustnosti chloru y ve vodě v závislosti na teplotě x. Interpolací najděte rozpustnost při 14EC.

Data: Závislost rozpustnosti chloru y na teplotě x:

Teplota x [EC]	6	10	13	15
Rozpustnost y [g/l]	1.0800	0.9972	0.9050	0.8495

Úloha C9.04 *Interpolace závislosti denní ranní teploty v roce 1990*

Data představují nejnižší ranní teplotu dne, naměřenou v Praze v roce 1990. Osa x představuje pořadové číslo dne v roce, osa y je ranní teplota. Pokuste se aproximovat uvedená data vhodnou křivkou.

Data: x je pořadové číslo dne v roce 1990, y je ranní teplota EC,

1	-3.1	2	-3.0	3	-3.2	4	-7.0	5	-7.9	6	-12.0	7	-12.5	8	-15.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
361	0.8	362	-1.1	363	-0.2	364	4.2	365	1.5						

Úloha C9.05 *Vyhodnocení bodu ekvivalencí dvou lineárních větví titrační křivky*

Titrační křivka se týká konduktometrické titrace titrandu 0.1 M HCl titrantem 0.1M NaOH. Je třeba určit bod ekvivalence čili uzlový bod zvratu dvou větví titrační křivky v instrumentální analýze. Protože však okolí bodu ekvivalence může být i nelinárního (zakřiveného) charakteru, je třeba vyšetřit, zda lze experimentálními body titrační závislosti aproximovat model s větvemi lineární-lineární, lineární-kvadratickou, kvadratickou-lineární nebo kvadratickou-kvadratickou.

Data:

(a) **Data C905a**, kde x je objem přidávaného titračního činidla 0.1M NaOH (v datech označené *C905ax*) a $y = (100 - a)/a$ a a je odečtená hodnota na odporovém můstku [mm] (v datech označené *C905ay*):

C905ax	2	4	6	8	10	12	14	16	17	18	20	22	24
C905ay	1.265	1.141	1.028	0.906	0.777	0.641	0.51	0.372	0.388	0.441	0.544	0.644	0.752

(b) **Data C905b**, kde x je objem přidávaného titračního činidla (v datech značené *C905bx*) a y je hodnota měřeného signálu (v datech značené *C905by*):

C905bx	0	1	...	...	17	18
C905by	0	2	...	...	16.02	16.19

(c) **Data C905c,** kde x je objem přidávaného titračního činidla (v datech značené *C905cx*) a y je hodnota měřeného signálu (v datech značené *C905cy*):

C905cx	1	2	...	...	16	17
C905cy	90.8	80.12	...	...	91.0	98.5

(d) **Data C905d,** kde x je objem přidávaného titračního činidla (v datech značené *C905dx*)

a *y* je hodnota měřeného signálu (v datech značené *C905dy*):

C905dx	1.012	1.985	...	...	15.766	16.257
C905dy	44	87	...	...	679.0	698.0

(e) **Data C905e,** kde *x* je objem přidávaného titračního činidla (v datech značené *C905ex*) a *y* je hodnota měřeného signálu (v datech značené *C905ey*):

C905ex	0	1	...	...	17	18
C905ey	0.29	0.58	...	...	11.86	11.93

Úloha C9.06 *Vyhodnocení bodu ekvivalencí dvou lineárních větví titrační křivky*
Titrační křivka se týká monitorování signálu libovolné instrumentální (fotometrické, ampérometrické, konduktometrické, coulometrické atd.) titrace titrandu titračním činidlem. Je třeba určit bod ekvivalence čili uzlový bod zvratu dvou větví titrační křivky v instrumentální analýze. Protože však okolí bodu ekvivalence může být i nelinárního (zakřiveného) charakteru, je třeba vyšetřit, zda lze experimentálními body titrační závislosti proložit model s větvemi lineární-lineární, lineární-kvadratickou, kvadratickou-lineární nebo kvadratickou-kvadratickou.
Data:
(a) **Data C906a**, kde *x* je objem přidávaného titračního činidla (v datech označené *C906ax*) a *y* hodnota měřeného signálu (v datech označené *C906ay*):

C906ax	0	1	...	...	17	18
C906ay	38	35.18	...	...	19.4	20.55

(b) **Data C906b**, kde *x* je objem přidávaného titračního činidla (v datech značené *C906bx*) a *y* je hodnota měřeného signálu (v datech značené *C906by*):

C906bx	0.545	0.643	...	...	1.652	1.714
C906by	47	56	...	...	107	109

(c) **Data C906c,** kde *x* je objem přidávaného titračního činidla (v datech značené *C906cx*) a *y* je hodnota měřeného signálu (v datech značené *C906cy*):

C906cx	0	1	...	...	16	17
C906cy	6	5	...	...	87	95

(d) **Data C906d,** kde *x* je objem přidávaného titračního činidla (v datech značené *C906dx*) a *y* je hodnota měřeného signálu (v datech značené *C906dy*):

C906dx	0	1	...	...	17	18
C906dy	23.5	22	...	...	15	17

(e) **Data C906e,** kde *x* je objem přidávaného titračního činidla (v datech značené *C906ex*) a *y* je hodnota měřeného signálu (v datech značené *C906ey*):

C906ex	0	1	...	...	11	12
C906ey	12.78	10.65	...	...	5.53	6.76

9.7.2 Analýza ekonomických a ostatních dat

Úloha S9.01 *Vyrovnání časové řady živě narozených dětí a zemřelých lidí*
Pokuste se vyrovnat následující časové intervalové řady následujících intervalových ukazatelů: počet y_1 živě narozených v letech 1970 - 1982, počet y_2 zemřelých v letech 1970 - 1982 proti času x. Sestrojte také sloupcové grafy, spojnicové grafy.
Data: Základní demografické údaje ze statistické ročenky v letech 1970 až 1982:

Rok x	Narození y_1	Zemřelí y_2
1970	228531	165567
...	...	...
1982	233284	179983

Úloha S9.02 *Vyrovnání časové řady výroby oceli v roce 1982*
Měsíční výroba oceli závisí na počtu dní v měsíci. Časová řada takových údajů nemůže proto sloužit jako podklad pro sledování vývoje výroby oceli v Československu v jednotlivých měsících roku 1982. Abychom mohli porovnávat výrobu oceli v jednotlivých měsících x, musíme údaje časové řady upravit tak, aby byly z hlediska délky intervalu x [měsíce] srovnatelné, tzn. přepočet údajů na měsíc o 30 dnech. Pokuste se vyrovnat časovou řadu původních dat y a přepočtených dat měsíční výroby oceli z.
Data: Výroba oceli v Československu v roce 1982:

Měsíc x	Výroba oceli y [10^3 t]	Přepočtená výroba oceli z [10^3 t]
1	1285	1243.55
...	...	...
12	1328	1285.16

Úloha S9.03 *Vyrovnání časové řady výroby televizorů v letech 1966 - 1982*
Pokuste se odhalit vývojovou tendenci v časové řadě počtu vyrobených televizorů y v letech 1966 - 1982 v Československu. Užijte postupnou lineární interpolaci 1., 2. a 3. řádu.
Data: Výroba televizorů y v Československu v letech 1966 až 1982

Rok x	Počet televizorů y [tisíce ks]
1966	227.9
...	...
1982	391.4

Úloha S9.04 *Vyrovnání časové řady výroby piva klouzavými průměry*
U časové řady výroby piva v Československu. za jednotlivá čtvrtletí roku 1977 až 1982 byla délka periody jednoznačně dána obdobím jednoho roku. Pomocí klouzavých průměrů nalezněte informaci o vývojové tendenci sledovaného ukazatele.
Data: Čtvrtletní výroba piva v ČSSR v letech 1977 až 1982

Rok	Čtvrtletí	Výroba piva [mil. l]
1977	1	482.3
	2	629.2
	3	630.9
	4	501.8
...	...	...
1982	1	508.8

2	689.2
3	732.1
4	561.1

Úloha S9.05 *Interpolace signálu analogově-číselného převodníku*
Při testování dynamických vlastností velmi rychlých analogově-číselných (AČ) převodníků se používá testovacích signálů ve tvaru sinusoidy, trojúhelníkového nebo lichoběžníkového průběhu nebo zdroje bílého šumu, generovaného např. Zenerovými diodami v závěrném směru. Na výstupu AČ jsou náhodným výběrem odebírány digitální hodnoty, jejichž četnost je dána kvalitou převodní charakteristiky AČ převodníku. Jedna z metod, používaná pro vyhodnocení převodní charakteristiky a dynamických vlastností, spočívá ve zpětné rekonstrukci vstupního signálu z digitálního výstupu AČ převodníku pomocí ideálního ČA převodníku. Rekonstrukce je prováděna pomocí počítače, který simuluje činnost ideálního ČA převodníku. Porovnáním vlastností vstupního signálu a rekonstruované křivky se dají velmi přesně určit dynamické vlastnosti daného převodníku. Vyřešte následující dva úkoly:

(a) Vstupním signálem je sinusoida, tj. body rekonstruované křivky, zkreslené vlastnostmi vlastního AČ převodníku a měřicím a vyhodnocovacím mechanismem. Aproximujte rekonstruovanou křivku pomocí metod aproximace a numerického vyhlazování.

(b) Proveďte obdobné vyhodnocení pro případ vstupního signálu ve tvaru bílého šumu.

Data: Proměnné: *x* čas, *y* (sin) je rekonstruovaná sinusoida po ideálním AČ a ČA převodu, *y* (sum) je rekostruovaný bílý šum po ideálním AČ a ČA převodu, *y* (spline) je rekonstrouvaná sinusoida po reálném AČ a ideálním ČA převodu,

x	*y* (sin)	*y* (sum)	*y* (spline)	*x*	*y* (sin)	*y* (sum)	*y* (spline)
0	0.006000	4.929755	4.935755,	1	-0.945630	1.043386	0.097753,
...	...	...	...	...	...	...	...
92	-5.223980	5.602445	0.378462,	93	-4.689820	4.850651	0.160833,

Úloha S9.06 *Nalezení regresního modelu vývozu chmele z Československa*
Nalezněte regresní model pro odhad vývozu chmele z Československa, když jsou k dispozici data časové řady o vývozu chmele a o sklizni chmele od roku 1962 do roku 1978. Porovnejte obě závislosti a komentujte. Je třeba užít klasickou regresi nebo vyrovnání časové řady?

Data: Sklizeň a vývoz chmele v Československu v letech 1962 až 1978

Rok	Vývoz chmele [tis. tun]	Sklizeň chmele [tis. tun]
1962	3.7	5.9
...	...	...
1978	7.7	12.2

Úloha S9.07 *Vyrovnání časové řady demografických dat v Československu*
Vyrovnejte časové řady demografických dat obyvatel Československu: (a) počtu obyvatel, (b) počtu sňatků, (c) počtu rozvodů, (d) počtu živě narozených dětí a (e) počtu zemřelých.

Data: Demografická data v Československu:

Rok	Obyvatel	Sňatků	Rozvodů	Narození	Zemřelých
1945	14 151 970	106 143	10 033	280 444	252 449

...	...	...	...	...	...
1983	15 414 360	120 547	36 254	230 660	186 907

9.8 Doporučená literatura

[1] Vitásek E.: *Numerické metody.* SNTL Praha 1987.
[2] Henrici P.: *Essentials of Numerical Analysis.* Wiley, New York 1982.
[3] Werner W.: *Math. Comput.* **43**, 205 (1984).
[4] Rice J. R.: *Numerical Methods, Software and Analysis.* McGraw Hill,Auckland 1983.
[5] Graves-Morris P. R.: *Padé Approximations and Its Application.* Springer Verlag, Berlin 1980.
[6] Hayes J. G., ed.: *Numerical Approximation and Data.* AlthonePress, London 1970, Kap. 2.
[7] De Boor C.: *SIAM J. Numer. Anal.*, **14**, 441 (1977).
[8] Fritsch F. N. a Carlson R. F.: *SIAM J. Numer. Anal.*, **17**, 238 (1980).
[9] Hyman J. M.: *Accurate Monotoniaty Preserving Cubic Internation*, Rept.LA-8796-MS Los Alamos Natl. Lab., 1982.
[10] Akima H.: *J. ACM*, **17**, 589 (1970).
[11] Brodlie K. W., ed.: *Mathematical Methods in Computer Graphics and Design.* Academic Press, London 1980.
[12] De Boor C.: *A Practical Guide to Splines*. Springer Verlag, New York1978.
[13] Sard A., a Weitraub S.: *A Book of Splines.* Wiley, New York 1971.
[14] Schweikert D. G.: *J. Math. Phys.*, **45**, 312 (1966).
[15] Rentrop P.: *Numer. Math.* **35**, 81 (1980).
[16] Reed M. B.: *Trans. Civil. Engn. Australia*, **CE25**, 57 (1983).
[17] Angot A.: *Užitá matematika pro elektrotechnické inženýry.* SNTL, Praha1972.
[18] Dahlquist G., Bjorck A.: *Numerical Methods.* Englewood Cliffs, New Jersey 1974.
[19] Watson G. A.: *Approximation Theory and Numerical Methods.* Wiley, Chichester 1980.
[20] Enbank R. L.: *Commun. Statist.* **13**, 433 (1984).
[21] Wold S.: *Technometrics* **16**, 1 (1974).
[22] Schvenberg I. J.: *Proc. Natl. Acad. Sci.* USA, **52**, 947 (1964).
[23] Pruess S.: *Computing* **19**, 365 (1978).
[24] Späth H.: *Spline Algorithmen zur Konstruktion glatter Kurven und Flächen*: R. Oldenburg Verlag, Munchen 1973.
[25] Reinsch C. M.: *Numer. Math.* **10**, 177 (1967).
[26] Eubank P. L.: *Statistics and Probab. Letters*, **2**, 9 (1984).
[27] De Hoog F. R., Hutchinson M. F.: *Numer. Math.* **50**, 312 (1987).
[28] Hutchinson M. F., de Hoog F. R.: *Numer. Math.*, **47**, 99 (1985).
[29] Silverman B. W.: *J. Amer. Statist. Assoc.*, **79**, 584 (1984).
[30] Utreras D. F.: *Numer. Math.* **34**, 15 (1980).
[31] Silverman B. W.: *J. R. Stat. Soc.*, **B47**, 1 (1985).
[32] Bendetti J. K.: *J. R. Stat. Soc.*, **B39**, 248 (1977).
[33] Rice J.: *The Annals Statist.*, **12**, 1215 (1984).

[34] Marmet P.: *Rev. Sci. Inst.*, **50**, 79 (1979).
[35] Hippe Z a kol.: *Anal. Chim. Acta* **122**, 279 (198).
[36] Martin R. D., in: *Smoothing Techniques for Curve Estimation.* Heidelberg 1979.
[37] Mallows C. L., in: *Time Series.* North Holland, Amsterdam 1980.
[38] Vellemann P. F.: *J. Am. Statist. Assoc.*, **75**, 609 (1986).
[39] Thrall AD.: *SIAM J. Appl. Math.* **40**, 169 (1981).
[40] Bromba M. V. A., Ziegler H.: *Anal. Chem.*, **53**, 1583 (1981).
[41] Proctor A., Sherwood M. A.: *Anal. Chem.*, **52**, 2315 (1980).
[42] Madden H. H.: *Anal. Chem.*, **50**, 1383 (1978).
[43] Bromba M. V. A. a Ziegler H.: *Anal. Chem.*, **51**, 1762 (1979).
[44] Frank L.: *Czech. J. Physics*, **B38**, 241 (1988).
[45] Savitzki A., Golay M. J. E.: *Anal. Chem.*, **36**, 1627 (1964).
[46] Cleveland W. S.: *J. Am. Statist. Assoc.*, **74**, 829 (1979).

10

KONTROLA A ŘÍZENÍ JAKOSTI

10.1 Podstata úloh řízení jakosti

Pojem jakost (kvalita) se používá jak v celé hierarchii řízení podniků, tak v obchodní sféře a sféře spotřebitelů. Přitom se ukazuje, že již vlastní definice tohoto pojmu je často zmatená a neumožňuje jakost stanovit. Dobře je tento fakt demonstrován na výsledcích ankety mezi špičkovými britskými manažery. Na základě vyhodnocení této ankety bylo zjištěno, že:

(a) Většina řídicích pracovníků cítí potřebu řízení jakosti;

(b) Nikdo z řídicích pracovníků nemá pocit, že by měl za jakost osobně zodpovídat;

(c) Většina řídicích pracovníků nedovede přesně vyjádřit, co to jakost vlastně je.

Závěrem uvedené ankety bylo konstatováno, že manažery zajímají tři problémy: *peníze, jak je získat a jak je neztratit.*

V této kapitole je učiněn pokus o naznačení základních přístupů k hodnocení jakosti s aplikací na průmysl[1-6]. Je uveden základní matematicko-statistický aparát pro konstrukci regulačních diagramů, indexů způsobilosti procesu a souvisejících úloh. Konstrukce a použití regulačních diagramů je nejlépe známou statistickou technikou řízení jakosti. Přes velké rozšíření regulačních diagramů se často setkáváme s jejich nesprávným použitím, způsobeným nerespektováním některých zásad, anebo zanedbáním základních předpokladů o datech, jako je normalita, nezávislost, stabilita. Důsledkem je nesprávná interpretace diagramů, nedůvěra ke statistickým metodám, upouštění od statistické regulace, někdy i zkreslování výsledků a nedovolená manipulace s daty.

Definice jakosti: V literatuře lze nalézt celou řadu více či méně obecných definic jakosti. Populární Juranova učebnice jakosti[7] uvádí definici: *"Jakost je vyjádřením vhodnosti k užívání"*. Podrobnější definici lze nalézt v normě ANSI/ASCQ z r. 1978: *"Jakost je souhrn rysů a charakteristik produktu nebo služby, který zajišťuje jeho schopnost vyhovět daným požadavkům"*. Velmi výstižná a obecná je definice z našich ČSN: *"Jakost výrobku je souhrnem vlastností podmiňujících způsobilost uspokojit potřeby odpovídající jeho účelu použití"*. Tato definice obsahuje v praxi často opomíjený fakt, že jakost je vždy spjatá s *účelem použití.* Nelze tedy říci, že se vyrábí jakostní výrobky, aniž je známo k čemu budou použity. To je např. v případě posuzování jakosti polotovarů často omezující.

Užitná hodnota: Existují jisté znaky jakosti, vyjádřené tzv. *užitnými vlastnostmi,* které mohou být buď jednoduše měřitelné (délkové rozměry, pevnost, tažnost, vlhkost,

koncentrace), nebo přímo neměřitelné, většinou subjektivní (vůně, chuť, tvar, barevnost, komfort při použití, vzhled), které jsou zejména pro výrobky spotřebního charakteru rozhodující. Předpokládejme, že lze obecně specifikovat K-tici užitných vlastností $R_1, ..., R_K$. Na základě přímých a nepřímých měření lze pak stanovit *ukazatele jakosti* (průměr, rozptyl, kvantily, podíl prvků mimo meze atd.) $x_1, ..., x_K$. Tyto ukazatele charakterizují užitné vlastnosti. Z hodnot x_i pro i-tou užitnou vlastnost lze pomocí vhodné funkční transformace definovat dílčí úroveň jakosti

$$u_i = f(x_i, K_D, K_H) ,$$

kde K_D je předepsaná hodnota užitné vlastnosti pro právě nevyhovující ($u_i = 0$) a K_H pro právě vyhovující ($u_i = 1$) výrobek. Celková úroveň jakosti, označovaná jako *užitná hodnota výrobku*, je pak vhodný vážený obecný průměr dílčích úrovní

$$u = ave(u_i, w_i) ,$$

kde w_i jsou váhy definující význam dané užitné vlastnosti a související s účelem použití výrobku[6]. S ohledem na své vlastnosti (pro nulové u_i vychází také $u = 0$) se obyčejně používá *vážený geometrický průměr*. Popsaný postup vycházející z obecné definice jakosti vyžaduje:

(a) nalezení pokud možno úplné množiny významných užitných vlastností,
(b) stanovení jejich velikosti (měření),
(c) nalezení vhodných vah.

Při konstrukci užitné hodnoty se projeví *hledisko hodnocení*.

(1) *Výrobce* bude zřejmě preferovat dodržení technologických parametrů výroby a snažit se omezovat variabilitu produktů.

(2) *Zpracovatel* bude hodnotit zpracovatelské vlastnosti vstupujícího "meziproduktu" a jejich vliv na jakost vyráběného produktu.

(3) *Spotřebitele* budou zřejmě zajímat užitné vlastnosti, které nemusí přímo souviset s jakostí, vyjádřenou z hlediska výrobce a zpracovatele (organoleptické vlastnosti, vzhled, životnost atd.).

Protože je užitná hodnota u stanovena na základě experimentálních údajů, jde o náhodnou veličinu, pro kterou lze určit střední hodnotu $E(u)$, rozptyl $D(u)$ a interval spolehlivosti střední hodnoty. Na základě těchto údajů lze pak porovnávat rozdíly mezi užitnými hodnotami výrobků s ohledem na přesnosti měření jednotlivých charakteristik. Komplexní charakteristika jakosti, *užitná hodnota*, se při řízení jakosti přímo ve výrobě uplatňuje velmi obtížně. Hodí se spíše pro komparaci finálních výrobků.

Ztrátová funkce: Moderní postupy "inženýrství kvality" využívají Taguchiho definice jakosti[5]: *"Jakost produktu je úměrná ztrátě způsobené společnosti odchylkou od předepsaných (cílových) hodnot"*. Přitom ztráta, způsobená společnosti, zahrnuje různé opravy, čištění, přerušení výroby, likvidace, odpady, nespokojenost zákazníka, ztráty trhu, náklady na reklamaci, arbitráž atd. a vyjadřuje se v peněžních jednotkách. Pro jeden parametr jakosti x lze uvedenou ztrátu definovat jednoduše pomocí *kvadratické ztrátové funkce*

$$L(x) = K(x - T)^2 ,$$

kde T je předepsaná cílová hodnota a K je parametr určený z odchylky x od T právě o definovanou, předepsanou toleranci, stanovenou výrobcem nebo akceptovanou

spotřebitelem. Ztrátovou funkci *L(x)* je třeba pro některé parametry jakosti poněkud modifikovat:

a) pro případy kdy "nižší hodnota je lepší" (nestejnoměrnost, vady, obsah škodlivin) se používá jednoduchá *ztrátová funkce konvexně rostoucí*

$$L(x) \quad ' \quad K x^2 \ ,$$

b) pro případy, kdy "vyšší hodnota je lepší" (pevnost, stálost, zralost) se používá *konvexně klesající ztrátová funkce*, která má v nejjednodušším případě tvar

$$L(x) \quad ' \quad \frac{K}{x^2} \ ,$$

c) pro případy, kdy "nominální hodnota je nejlepší" (rozměr, plošná hmotnost, tažnost atd.), se často místo funkce $K(x - T)^2$ používá asymetrická varianta ztrátové funkce

$$\begin{aligned} L(x) \quad ' \quad & K_1 (x \ \& \ T)^2 \qquad \text{pro} \quad x \ \# \ T, \\ L(x) \quad ' \quad & K_2 (x \ \& \ T)^2 \qquad \text{pro} \quad x \ > \ T. \end{aligned}$$

Tyto tvary ztrátových funkcí odrážejí obecně to, jakým způsobem ovlivňují jednotlivé parametry jakosti ztrátu způsobenou společnosti. Uvedené definice ztrátové funkce vycházejí z předpokladu, že je parametr jakosti x stanoven přesně, čili jde o deterministickou proměnnou. Zejména ve zpracovatelském průmyslu, kde je variabilita složení a struktury vždy značná, je třeba uvažovat, že na základě výběru (z různých míst, v různých časech) lze získat pouze odhad střední hodnoty x_s (aritmetický průměr) a odpovídající rozptyl s^2 (výběrový rozptyl). Ztrátová funkce má pak tvar

$$L(x) \quad ' \quad K\left[s^2 \ \% \ (x_s \ \& \ T)^2\right] .$$

Při vlastním řízení jakosti pomocí ztrátové funkce lze na základě této rovnice určit, zda je výhodnější snižovat variabilitu s^2 nebo se blížit předepsané hodnotě T. Je také zřejmé, že i při dodržení předepsané hodnoty parametru jakosti může být ztráta způsobená variabilitou výroby nebo výrobku poměrně značná.

Techniky řízení jakosti: S ohledem na historii řízení jakosti lze specifikovat tři základní koncepce:

(a) *Přejímací plány*, které určují pravidla, podle nichž se na základě analýzy části výrobků usuzuje, zda je celá dodávka přijatelné kvality či nikoliv,

(b) *Statistické řízení procesů,* kdy se monitorují znaky jakosti s využitím regulačních diagramů (on-line regulace),

(c) *Inženýrství jakosti,* kdy se provádí off-line projektování hodnot procesních proměnných s využitím koncepce ztrátové funkce.

V současné době se využívá kombinace všech tří koncepcí řízení jakosti přesto, že zde existuje logický nesoulad. Přejímací plány a regulační diagramy vycházejí z předpokladu *skokové funkce jakosti,* tj. pokud jsou parametry jakosti v zadaných mezích, je *jakost přijatelná,* a pokud jsou mimo tyto meze, je *jakost nepřijatelná.* Inženýrství jakosti využívá spojité funkce jakosti (např. ztrátové funkce výše uvedené rovnice), kdy se každá odchylka od ideálního stavu ($x_s = T$, $s^2 = 0$) projeví ztrátou vyjádřenou finančně.

Teorie *přejímacích plánů* je v současné době zpracována jak pro přejímky srovnáváním

(diskrétní znaky jakosti), tak i měřením (spojité znaky jakosti). Obecně je statistická přejímka soubor postupů výběrové kontroly jakosti, prováděné na dodávkách surovin, polotovarů a výrobků. Postup statistické přejímky je vlastně testem statistické hypotézy o parametru jakosti. Pravděpodobnost zamítnutí vyhovující dodávky α se označuje jako *riziko dodavatele* (chyba prvního druhu), pravděpodobnost přijetí nevyhovující dodávky β se označuje jako *riziko odběratele* (chyba druhého druhu). Při statistické přejímce srovnáváním i měřením je třeba určit tzv. *rozhodné číslo* (odpovídá kritické hodnotě u testů) *c* a rozsah výběru *N*. Dvojice (*c, N*) se označuje jako *přejímací plán*. Demonstrujme si určení plánu např. u přejímky měřením. Jakostní znak *x* nechť je spojitá náhodná veličina s normálním rozdělením. Výrobek se považuje za vadný, pokud překročí veličina *x* horní toleranční mez *USL*. Označme P_1 jako přípustný podíl výrobků, pro které bude $x > USL$ a P_2 jako nepřípustný podíl takto definovaných vadných výrobků (obyčejně se volí $P_1 = 0.001$ a $P_2 = 0.01$).

Riziko dodavatele pak je $\alpha = P(x_S + c\,s \geq USL*P_1)$

a *riziko odběratele* je $\beta = P(x_S + c\,s < USL*P_2)$,

kde x_S je výběrová střední hodnota a s^2 je výběrový rozptyl. Za předpokladu, že veličina $x_S + c\,s$ má přibližně normální rozdělení, lze nalézt parametry přejímacího plánu

$$c = \frac{u_{1\&P_1}\,u_{1\&\beta} + u_{1\&P_2}\,u_{1\&\alpha}}{u_{1\&\alpha} + u_{1\&\beta}} ,$$

$$N = \frac{u_{1\&\alpha}{}^2\,(2 + c^2)}{2\,(u_{1\&P_1} - c)^2} ,$$

kde u_r je $100r$ %ní kvantil normovaného normálního rozdělení. Při volbě $\alpha = \beta = 0.05$ a $P_1 = 0.001$, $P_2 = 0.01$ vyjde $c = 2.71$ a $N = 174$. Při provedení přejímacího plánu se tedy vybere 174 výrobků, stanoví se hodnoty jakostních znaků x_i, $i = 1, ..., 171$, a určí se odhady x_S, s^2. Pokud je splněna nerovnost $x_S + 2.71\,s \leq USL$, zásilka se přijímá. V opačném případě se zamítá. Kritickým je zde především předpoklad normality a nezávislosti prvků výběru. Obdobným způsobem se konstruují také přejímací plány pro přejímku srovnáváním.

Pro přímé řízení jakosti výroby se používají *toleranční meze*, *LSL* (dolní) a *USL* (horní), definující interval, ve kterém leží se zvolenou pravděpodobností předepsané procento výsledků, hodnot parametru jakosti. Na základě tolerančních mezí se definuje tzv. *parametr způsobilosti procesu* (process capability index)

$$C_p = \frac{USL - LSL}{6\sigma} ,$$

kde σ je směrodatná odchylka parametru jakosti. Hodnota $C_p > 1$ svědčí o tom, že jakost je přijatelná čili v podstatě celá výroba je v tolerančních mezích. Je zajímavé, že i ve vyspělých státech, jako jsou USA, vychází často C_p menší než 1. Na druhé straně v Japonsku již v r. 1980 docílili průměrného $C_p = 1.33$, a u "high tech" produktů dokonce

C_p = 2. Je zřejmé, že pro případ normálního rozdělení leží v mezích ± 3σ přibližně 99.73 % hodnot parametru jakosti. Hodnota C_p = 1 pak ukazuje že 99.73 % výrobků je v tolerančních mezích. Hodnota 6 ve jmenovateli rovnice pro C_p je obecně závislá na velikosti výběru, z něhož se odhaduje σ, a na rozdělení znaku jakosti.

Omezení problému s nenormalitou rozdělení znaku jakosti lze docílit vhodnou volbou procenta výrobků ležících v tolerančních mezích. Pokud zvolíme tento parametr 99 %, můžeme použít ve jmenovateli hodnotu 5.15, platnou pro řadu rozdělení, a to se šikmostí od 0 do 3.111 a špičatostí od 1 do 5.997.

Pro případ, že lze předpokládat normální rozdělení parametru jakosti, je také možno vypočítat *LSL* a *USL* (pokud nejsou zadány) relativně jednoduše (popsáno v ČSN 01 0230) a C_p lze pak snadno určit. V obecném případě je třeba použít komplikovanější postupy[10]. Je patrné, že přejímací plány a parametry způsobilosti jsou off-line charakteristiky umožňující sledovat jakost mimo vlastní proces výroby. Pro přímé ovlivňování výroby, monitorování kvality, se používají regulační diagramy. Jejich základní myšlenka vychází z intervalů spolehlivosti.

Účelem je ovlivňovat proces výroby tak, aby hodnoty parametrů jakosti ležely uvnitř těchto intervalů: *regulační diagram* je graf, v němž je obyčejně vyznačena jistá předepsaná (průměrná) hodnota a regulační meze (příp. jiné kontrolní meze). Do tohoto grafu se postupně vyznačují hodnoty parametru jakosti určované přímo v procesu výroby. V případě, kdy se projeví nenáhodné trendy, následuje seřízení procesu nebo zásah do výroby.

Pro případ spojitého znaku jakosti se obyčejně používají *regulační diagramy Shewhartovy.* Při jejich konstrukci se vychází z měřených dat (výběrů) a počítá se vhodná *statistika* S_i (což může být např. průměr x_{Si} , směrodatná odchylka s_i , rozpětí R_i atd.). Statistiky S_i se vynášejí do grafu, kde jsou znázorněny regulační meze, odpovídající obyčejně intervalům spolehlivosti $E(S_i) \pm 3\ \%[D(S_i)]$. Zde $E(x)$ značí střední hodnotu a $D(x)$ rozptyl parametru x. Na základě znázorněných bodů se pak usuzuje, zda je proces *statisticky stabilní*. Pokud mají statistiky S_i přibližně normální rozdělení, je pravděpodobnost překročení těchto 3σ-mezí rovna pouze 0.0027. V ostatních případech je třeba stanovit regulační meze ze znalosti rozdělení statistiky S_i . Pokud je do regulačních diagramů vynášeno více hodnot současně (neprovádí se přímo on-line kontrola), je pravděpodobnost P toho, že jedna nebo více hodnot padne mimo regulační meze, závislá na počtu vynášených bodů N. Tato pravděpodobnost se dá modelovat pomocí binomického rozdělení. Platí, že

$$P = 1 - (1 - p)^N ,$$

kde p je pravděpodobnost, že hodnota jednoho znaku padne mimo toleranční meze, tj. p = 0.0027. Pro malá N pak přibližně platí, že

$$P = 1 - (1 - p)^N \doteq pN .$$

Je patrné, že při vynášení více bodů současně se odpovídající pravděpodobnost P zvyšuje. Rovnice pro P se dá použít pro určení p, pro které bude dodržena pravděpodobnost P = 0.0027 při zadaném N, např. pro N = 20 je p = 0.00014 a odpovídající $100(1-p/2)$ %ní kvantil normálního rozdělení je 3.81. Pro tuto situaci je tedy vhodné konstruovat regulační meze $LCL = x_s - 3.81s$ a $UCL = x_s + 3.81s$.

Regulační diagramy jsou velmi populární pro svoji jednoduchost a lze je konstruovat

pro téměř všechny typy parametrů jakosti (spojité, diskrétní, ordinální, nominální). Při praktickém sestrojování činí potíže zejména nerobustnost odhadů x_s a s^2, nenormalita rozdělení znaku jakosti a závislost mezi jednotlivými měřeními.

Kromě diagramů Shewhartova typu jsou používány také diagramy kumulativních součtů (CUSUM), pohyblivých charakteristik polohy a různé kombinace[1]. Regulační diagramy se dají použít také pro případ simultánního sledování více znaků jakosti, Hotellingovy karty[1]. Přejímací plány a regulační diagramy vycházejí z představy, že proces, který sledujeme, je v ustáleném stavu, tj. jeho řiditelné parametry (teplota, koncentrace, rychlost atd.) jsou na optimální úrovni.

Inženýrství jakosti umožňuje off-line nastavení podmínek výroby tak, aby bylo dosaženo jakostní výroby. Využívá se principů plánování experimentů s několika modifikacemi:

(a) Parametry výroby se dělí na *ovladatelné* (ty se optimalizují) a *šumy* (omezuje se jejich variabilita).

(b) Měřítkem kvality experimentu je *poměr* signálu a šumu *S/N*, který se maximalizuje.

(c) Používá se plánů ve tvaru ortogonálních polí.

Poměr signálu a šumu *S/N* souvisí s definicí ztrátové funkce *L(x)*. Pro případ, kdy "nominální je nejlepší", lze použít vztahu

$$S/N \text{ ' } 10 \log(x_s^2/s^2) \quad .$$

Inženýrství jakosti využívá obecně technik plánovaných experimentů pro modelování chování procesů výroby s ohledem na nalezení optima. Při klasickém plánování experimentů se vychází z náhodné veličiny y (vysvětlující proměnná), která je výstupem procesu. Ta je funkcí vektoru vstupních parametrů $\boldsymbol{x}$. Účelem je stanovit vhodnou regresní funkci $f(\boldsymbol{x}, \boldsymbol{\beta})$ a odhadnout její parametry $\boldsymbol{\beta}$ tak, aby bylo dosaženo jistých kritérií optimality. Tato kritéria úzce souvisejí s přesností parametrů a rozptylem σ_y^2 veličiny y. Plánované experimenty se s výhodou používají také pro hledání optimálních podmínek minimalizujících tzv. ztrátovou funkci.

Taguchi[27] používá jako vstupní tzv. *parametry plánu* a hledá optimální podmínky, aby byly splněny tyto požadavky:

a) Minimální rozptyly některých znaků jakosti, jejichž střední hodnota je rovna cílové (požadované) hodnotě.

b) Minimalizace citlivosti výstupu na externí a interní fluktuace neřízených faktorů.

Pokud má náhodná veličina y rozdělení se střední hodnotou μ_y a rozptylem σ_y^2, lze poměr signálu a šumu *S/N* určit ze vztahu

$$S/N \text{ ' } 10 \log(\mu_y^2/\sigma_y^2) \text{ ' } \&10 \log(v^2) \ ,$$

kde $v = \sigma_y / \mu_y$ je variační koeficient. Podle vlivu na poměr *S/N* lze jednotlivé vstupní parametry rozdělit do tří kategorií:

1) *Řídicí faktory,* ovlivňující variabilitu procesu vyjádřenou *S/N.*

2) *Nastavované parametry,* které mají zanedbatelný vliv na *S/N,* ale významně ovlivňují střední hodnotu procesu.

3) *Šumové faktory,* které neovlivňují ani *S/N,* ani střední hodnotu procesu.

Faktory prvních dvou skupin patří do *parametrů plánu.*

Klasický postup návrhu experimentů vychází z minimalizace očekávané ztrátové

funkce

$$M(\boldsymbol{x}) \quad ' \quad E[\,(y(\boldsymbol{x}) \;\&\; T)^2\,] \quad ' \quad \sigma_y^2(\boldsymbol{x}) \;\%\; (\mu_y(\boldsymbol{x}) \;\&\; T)^2 \;,$$

kde jak střední hodnota $\mu_y(\boldsymbol{x})$, tak i rozptyl $\sigma_y^2(\boldsymbol{x})$ jsou funkcí parametrů plánů a T je požadovaná cílová hodnota. Využitím technik plánovaných experimentů lze nalézt podmínky $\boldsymbol{x}_0$, pro které je proces nejblíže požadovanému stavu. Tedy $M(\boldsymbol{x}_0)$ je minimem. Speciálně *Taguchiho přístup* předpokládá, že systém má jisté specifické vlastnosti vedoucí ke zjednodušení problému. Tyto vlastnosti plynou ze vztahu mezi σ_y a μ_y.

Předpokládá se, že závislost mezi σ_y a μ_y je takového druhu, že lze nalézt funkci $f[\mu_y(x)]$, pro kterou je $\sigma_y^2(\boldsymbol{x}) \;/\; \{f[\mu_y(\boldsymbol{x})]\}^2$ měřítkem rozptylu $P(\boldsymbol{x}_1)$. Funkce $P(\boldsymbol{x}_1)$ závisí pouze na podmnožině $\boldsymbol{x}_1$ všech parametrů plánu $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$. Pak je $P(\boldsymbol{x}_1)$ nezávislá na μ_y, protože pro daná x_1 je μ_y závislé pouze na parametrech $\boldsymbol{x}_2$. Funkce $P(\boldsymbol{x}_1)$ nezávisí také na $\boldsymbol{x}_2$. Pak $\boldsymbol{x}_1$ jsou řídicí parametry ovlivňující rozptyl a $\boldsymbol{x}_2$ jsou parametry nastavované, jejichž změny neovlivní rozptyl. Na základě těchto předpokladů lze vyjádřit ztrátovou funkci ve tvaru

$$M(\boldsymbol{x}) \quad ' \quad (f[\,\mu(\boldsymbol{x})\,])^2 \;\; P(\boldsymbol{x}_1) \;\%\; (\mu(\boldsymbol{x}) \;\&\; T)^2 \;.$$

Pro zvolené μ je pak $M(\boldsymbol{x})$ minimální, pokud se nalezne $\boldsymbol{x}_{10}$, pro které je $P(\boldsymbol{x}_{10})$ minimální. Po nalezení $\boldsymbol{x}_{10}$ lze určit absolutní minimum $M(\boldsymbol{x})$ změnami parametrů $\boldsymbol{x}_2$. Pak je možno vyjádřit $\mu_y(\boldsymbol{x}_0) = \mu_0$, pro které nabývá $M(\boldsymbol{x})$ minima ve tvaru

$$\mu_0 \quad ' \quad T \;\&\; f^2(\mu_0) \;\; P(\boldsymbol{x}_1) \;.$$

Speciální případ funkce $f(\mu)$ je obecná mocnina μ^{α}. Pak platí, že

$$P(\boldsymbol{x}_1) \quad ' \quad \sigma_y^2 \,/\, \mu_y^{2\alpha} \quad \text{a} \quad \mu_0 \quad ' \quad T \,/\, (1 \;\%\; \alpha v_0^2) \;,$$

kde v_0 je variační koeficient v minimu. Uvažujme dva speciální případy:

a) $\alpha = 0$, pak σ_y není funkcí μ_y, $P(\boldsymbol{x}_1) = \sigma^2(\boldsymbol{x}_1)$ a $\mu_0 = T$. V tomto případě se $M(\boldsymbol{x})$ minimalizuje nejdříve nalezením $\boldsymbol{x}_{10}$ minimalizujícího $\sigma(\boldsymbol{x}_1)$ a pak se nalezne $\boldsymbol{x}_{20}$ tak, aby $\mu_0 = T$.

b) $\alpha = 1$, pak $P(\boldsymbol{x}_1) = v^2$ a minimalizace $P(\boldsymbol{x}_1)$ je shodná s maximalizací poměru signálu a šumu *S/N*. Při hledání $\boldsymbol{x}_{20}$ se hledá μ_0, pro které je $\mu_0 \;' \; T(1 \;\%\; v_0^2)$. Je známo, že pokud je směrodatná odchylka lineární funkcí střední hodnoty, vede logaritmická transformace y ke zkonstantnění rozptylu. Minimalizace směrodatné odchylky v *logaritmické transformaci* je tedy ekvivalentní maximalizaci *S/N* v původních proměnných. Je patrné, že Taguchiho postup je možné převést na standardní úlohu plánovaných experimentů ve vhodné (logaritmické) transformaci. Podrobnosti lze nalézt např. v cit.[27].

Inženýrství jakosti je logicky prvním stadiem celého procesu řízení jakosti, protože umožňuje stanovení podmínek, které je třeba při výrobě dodržovat. Následně se používají regulační diagramy pro monitorování jakosti. Tento přístup k řízení jakosti úzce souvisí s restrukturalizací informačních a řídicích informací (total quality management) a začíná se uplatňovat při zabezpečení jakosti ve vyspělých státech[9].

10.2 Regulační diagramy

Regulační diagramy patří k základním nástrojům pro regulaci jakosti při výrobních procesech. Dají se však použít zcela obecně všude tam, kde jsou postupně v čase získávány informace o jakosti. Umožňují pro procesy, které jsou *statisticky regulovatelné* a měřený znak jakosti má stejné v čase neměnné rozdělení, modifikovat výrobní procesy tak, aby procento zmetků, kdy znak jakosti leží mimo předepsané meze, bylo dostatečně malé. Hodí se velmi dobře zejména pro monitorování procesů pomocí počítače. V této kapitole jsou uvedeny základní typy regulačních diagramů včetně způsobů jejich konstrukce a statistických vlastností.

V květnu 1924 navrhl W. A. Shewhart z Bell Telephone Laboratory první regulační diagram pro posouzení, zda variabilita sledovaného procesního parametru je způsobena náhodným kolísáním, nebo speciálními příčinami (seřízení strojů, změna surovin atd.). V současné době představují v praxi nejrozšířenější typ, i když nejsou zdaleka univerzální. Typický regulační diagram Shewhartova typu je diagram, kde je znázorněna centrální linie, značící standardní, očekávanou cílovou hodnotu charakteristiky znaku jakosti a obě regulační meze, tj. dolní *LCL* a horní *UCL*. Tyto meze určují interval, ve kterém se s velkou pravděpodobností pohybují charakteristiky znaku jakosti, pokud je proces v *požadovaném stavu.* Pokud padnou hodnoty charakteristiky znaku jakosti mimo regulační meze, je *proces "mimo" požadovaný stav* a je třeba provést korekce (seřízení strojů, atd.).

Jako charakteristiky znaků jakosti se používají zejména *průměr* x_S, *směrodatná odchylka s*, *variační rozpětí R*, *podíl nestandardních výrobků P*, *počet defektních výrobků C* a *počet defektů na výrobek u*. Speciální typy regulačních diagramů jako jsou *kumulativní součty* CUSUM, *pohyblivé průměry* MA atd. vycházejí z *metod analýzy časových řad* a konstrukčně se poněkud liší od regulačních diagramů Shewhartova typu. Konečně je možné sestrojovat také vícerozměrné Hotellingovy regulační diagramy pro více znaků jakosti současně[1].

10.2.1 Regulační diagramy pro dílčí výběry

V řadě případů je měření znaku jakosti jednoduché a rychlé, takže lze provést opakovaná měření, téměř ve stejném čase, a získat v různých časech několik výběrů $V_1,...,V_M$. Uvažujme pro jednoduchost, že pro každý výběr V_j stejné velikosti N jsou určeny výběrové průměry x_{sj} a rozptyly s_j^2. Předpokládejme, že rozdělení výběrových průměrů x_s je normální $N(d, \sigma^2/N)$ a výběrové průměry jsou vzájemně nezávislé. Odhadem střední hodnoty d je pak generální průměr d^*, definovaný vztahem

$$d^{(} \quad ' \quad \frac{1}{M} \sum_{j}^{M} \sum_{1} x_{sj} ,$$

a odhadem směrodatné odchylky σ je např. průměrná směrodatná odchylka

$$\sigma^{(} \quad ' \quad \frac{1}{M C_4} \sum_{j}^{M} \sum_{1} s_j ,$$

kde C_4 je konstanta zajišťující nevychýlenost. S využitím známého faktu, že

$$\frac{(N \& 1)\, s^2}{\sigma^2} \quad . \quad \chi^2 (N \& 1) \ ,$$

kde symbol $\chi^2(.)$ označuje rozdělení χ^2-kvadrát, lze určit, že střední hodnota výběrové směrodatné odchylky je

$$E(s) \quad ' \quad C_4 \, \sigma \ ,$$

kde $$C_4 \quad ' \quad \sqrt{\frac{2}{N \& 1}} \ \frac{\Gamma(N/2)}{\Gamma(N \& 1)/2)} \quad . \quad 1 \ \& \ \frac{1}{4N \& 1} \ ,$$

kde $\Gamma(.)$ je gamma funkce. Odpovídající rozptyl je pak roven $D(s) \ ' \ \sigma^2 (1 \ \& \ C_4^2)$.

Pro odhad směrodatné odchylky σ se používá vztah

$$\sigma^{\%} \quad ' \quad \sqrt{\frac{\sum_{j \, 1}^{M} s_j^2}{M}} \ .$$

Odhad σ^+ však již není nevychýlený.

10.2.2 Regulační diagramy typu "x s pruhem"

Tyto regulační diagramy využívají k posuzování stavu sledovaného procesu aritmetické průměry x_{Si} (resp. obecněji *parametr polohy*). Vyžadují ke své konstrukci buď znalost parametrů d, σ^2 normálního rozdělení, ze kterého data pocházejí, resp. pouze znalost vhodných odhadů d^* a σ^{*2}. Regulační meze se konstruují tak, aby bylo zajištěno, že pravděpodobnost jejich překročení pro proces v požadovaném stavu je dostatečně malá. Z vlastností normálního rozdělení plyne, s jakou pravděpodobností se vyskytuje veličina x_s v mezích

$$d \quad \pm \quad \frac{K \sigma}{\sqrt{N}} \ .$$

Pro K = 1.96 to je pravděpodobnost 0.95 a pro K = 3.09 je to 0.999. V praxi se pro konstrukci regulačních mezí volí běžně hodnota K = 3, které odpovídá pravděpodobnost 0.9973. Obyčejně však nejsou parametry d a σ známy a nahrazují se svými odhady d^* a σ^*.

Regulační diagram "*x* s pruhem" má centrální linii d^* a regulační meze

$$L\hat{C}L \quad ' \quad d(\ \& \ \frac{3 \, \sigma(}{\sqrt{N}} \ ,$$

$$U\hat{C}L \;=\; d^{*} + \frac{3\sigma^{*}}{\sqrt{N}} \,.$$

Takto definované regulační meze jsou odhady regulačních mezí *LCL* a *UCL*. Je zřejmé, že veličina d^* má normální rozdělení a veličina σ^{*2} (jako průměr *M* nezávislých proměnných) má také přibližně normální rozdělení. Pak $L\hat{C}L$ a $U\hat{C}L$ jako lineární kombinace d^* a σ^* mají přibližně normální rozdělení. Lze ukázat, že pro $U\hat{C}L$ platí

$$E(U\hat{C}L) \;=\; d + \frac{3\sigma}{\sqrt{N}} \;=\; UCL$$

a

$$D(U\hat{C}L) \;=\; \frac{\sigma^2}{MN}\left[1 + \frac{9(1 - C_4^{\,2})}{C_4^{\,2}}\right] \;=\; \frac{K_1\sigma^2}{MN} \,.$$

Obdobně lze snadno určit střední hodnotu a rozptyl pro $L\hat{C}L$. Je zřejmé, že pravděpodobnost *p*, s jakou x_S překročí regulační meze $L\hat{C}L$ a $U\hat{C}L$, je závislá na *M* a *N*. Podle uvedeného výkladu totiž platí, že

$$p \;=\; 2\left[1 - F_N\left(\frac{3}{\sqrt{1 + \frac{1}{M}K_1}}\right)\right] \,.$$

Pomocí tohoto vztahu lze pak snadno určit zkreslení, ke kterému dochází vlivem použití odhadů d^* a σ^*. Tak např. pro případ, že $M = 30$ a $N = 5$, vyjde $C_4 = 0.94$ a $p = 0.00378$. Je tedy zřejmé, že pro tento případ je pravděpodobnost *P*, s jakou se x_S vyskytuje v regulačních mezích, rovna $P = 1 - p = 0.9962$. Pro menší *M* a *N* může být pokles výraznější, což negativně ovlivní *použití regulačních diagramů*, když totiž vzroste četnost nesprávného rozhodnutí o zásahu do výroby.

Pravděpodobnost *p* je jednou z charakteristik kvality regulačních diagramů. Souvisí přímo s *rizikem I. druhu,* tj. pravděpodobností, že extrém vyvolaný náhodným kolísáním bude interpretován jako důsledek speciálních příčin. Je zřejmé, že čím je *p* vyšší, tím je riziko I. druhu menší.

Riziko II. druhu je pravděpodobnost jevu, že x_S bude v regulačních mezích přesto, že došlo ke změně úrovně procesu, a to v důsledku speciálních příčin. Toto riziko je přímo úměrné velikosti *p* a nepřímo úměrné velikosti změny úrovně procesu.

Pro účely návrhu a posouzení regulačních diagramů je vhodné sledovat počet hodnot v regulačním diagramu *L,* který je třeba k tomu, aby bylo indikováno překročení regulačních mezí. Pokud jsou výběrové průměry nezávislé, má veličina *L* geometrické rozdělení s pravděpodobnostní funkcí

$$P(L = i) = p(1 - p)^{i-1}, \quad i = 1, 2, 3, \ldots,$$

kde *p* je pravděpodobnost toho, že jeden výběrový průměr x_S překročí regulační meze.

Střední hodnota *E(L)* se označuje jako *ARL* a je rovna

$$E(L) \;=\; ARL \;=\; 1/p$$

a pro rozptyl pak platí $D(L) = D(ARL) = \frac{1-p}{p^2}$. Pro regulační diagramy "*x* s pruhem" je v případě exaktních regulačních mezí *LCL* a *UCL* veličina *ARL* = 370.37 a *D(L)* = 136803.84. Pro případ přibližných regulačních mezí, počítaných z odhadů střední hodnoty a rozptylu, jsou však

$$e_j \;=\; x_{Sj} - \hat{UCL}$$

a

$$e_k \;=\; x_{Sk} - \hat{UCL}$$

vždy pozitivně korelované a pro korelační koeficient platí

$$\rho(e_j, e_k) \;=\; [1 + M(1 + K_1)^{-1}]^{-1} \quad .$$

Nejjednodušším typem změny stavu procesu vlivem speciálních příčin je posun střední hodnoty *d* na velikost d_p. Pro tento případ je možno snadno určit, že pravděpodobnost, s jakou bude ležet x_S v mezích *LCL* a *UCL,* je rovna

$$p \;=\; F_N(A - 3) + 1 - F_N(A + 3) \;,$$

kde $A = ((d - d_p)\%N)/\sigma$ je standardizovaný posun střední hodnoty. S využitím *p* lze snadno určit *ARL*. Tak např. pro *A* = 1 vyjde *ARL* = 43.89 a pro *A* = 2 je *ARL* = 6.3. Také pro složitější změny stavu procesu lze buď analyticky, nebo na základě simulací určit *ARL*. Při konstrukci regulačních diagramů "*x* s pruhem" se vychází z předpokladů:

a) Rozdělení dat je alespoň přibližně normální.
b) Velikosti výběrů jsou stejné.
c) Měření jsou nezávislá.
d) V datech nejsou vybočující měření (hrubé chyby).

Obecně je třeba, jak ve fázi konstrukce, tak i ve fázi použití regulačních diagramů, testovat předpoklady o datech, podobně jako při statistické analýze jednorozměrných výběrů[16].

Nenormalita. V řadě případů má sledovaný parametr sešikmené rozdělení (např. pro cirkularitu, pevnost, koncentrace ve stopové analýze atd.). Pro menší a střední šikmosti se nenormalita výrazně neprojeví, pokud je počet prvků v jednotlivých výběrech $N \geq 5$. Pro větší šikmosti je možno použít následujících technik:

1. Nalézt vhodnou normalizační transformaci (např. ve třídě Boxových-Coxových mocninných transformací[16]) a realizovat regulační diagramy v transformovaných proměnných. Jednoduché empirické pravidlo doporučuje použití logaritmické transformace dat, pokud jsou prvky ve výběrech řádově rozdílné.

2. Nalézt vhodnou teoretickou hustotu pravděpodobnosti, resp. distribuční funkci F_T a určit meze *LCL* a *UCL* tak, aby platilo

$$F_T(LCL) \;=\; 1 - F_T(UCL) \;=\; p/2 \;,$$

kde standardně *p* = 0.0027. Pro hledání F_T je možné použít jak teoretických úvah, tak i celé

řady exploratorních metod (např. *Q-Q* grafy, atd.)[16].

3. Použít heuristické techniky, vycházející z pravděpodobnosti P_x , že výběrové průměry x_{Sj} leží pod generálním průměrem d^*. Tato pravděpodobnost se dá odhadnout z počtu výběrových průměrů x_s , ležících pod d^*

$$\hat{P}_x = \frac{1}{M} \sum_j^M I_{[x_{Sj}]}$$

kde indikátorová funkce $\begin{cases} I_{[x_{Sj}]} = 1 & \text{pro } x_{Sj} < d^* \\ I_{[x_{Sj}]} = 0 & \text{pro } x_{Sj} \geq d^* \end{cases}$.

Pro rozdělení sešikmené k vyšším hodnotám je pak

$$UCL = d^* + 3\sigma^* \sqrt{\frac{2P_x}{N}}$$

a

$$LCL = d^* - 3\sigma^* \sqrt{\frac{2(1-P_x)}{N}} \ .$$

Meze *UCL* a *LCL* byly sestaveny na základě náhrady původní hustoty pravděpodobnosti dvěma segmenty v místě *d*. Každý segment je použit pro sestavení symetrické hustoty pravděpodobnosti[18].

Nestejné velikosti výběrů. Pokud se počet prvků výběru *N* mění, je třeba pouze upravovat meze, definované rovnicí pro $L\hat{C}L$ a $U\hat{C}L$. Regulační meze jsou pak tvořeny po částech konstantními úseky.

Autokorelace. Vlivem autokorelace dochází ke zkreslení regulačních diagramů "*x* s pruhem". Např. v případě pozitivní autokorelace roste počet případů, překračujících regulační meze, i když je proces v požadovaném stavu (falešný poplach). Pro případ autokorelace prvního řádu s autokorelačním koeficientem ρ platí pro střední hodnotu výběrového rozptylu vztah

$$E(s^2) = \frac{\sigma^2}{(1-2\rho/N)} \ .$$

Při použití s^2 tedy v případě pozitivní autokorelace vyjde rozptyl střední hodnoty podhodnocený, totiž nesprávně menší. Někteří autoři doporučují pro omezení vlivu autokorelace zvětšit regulační meze o faktor $1/\sqrt{1-\rho^2}$. Obecně však nejsou vhodné, pokud jsou měření silně korelována. V některých případech lze zdroje autokorelace indikovat a odstranit.

Vybočující měření. Přítomnost vybočujících měření obecně zkresluje odhady d^* a σ^* a vede k rozšiřování regulačních mezí. K velmi dobrým robustním výsledkům vede náhrada

výběrových rozptylů interkvartilovým rozmezím

$$IQR \quad ' \quad x_{(b)} \ \& \ x_{(a)} \ ,$$

kde a = int[N / 4] + 1 a $b = N - a + 1$. Zde symbol int[x] označuje celočíselnou část čísla x a $x_{(i)}$ je i-tá pořádková statistika. Obecně se může zrobustňovat celá řada parametrů. Místo aritmetických průměrů x_{Sj} lze použít robustní odhady polohy G_j, místo výběrových rozptylů lze použít robustní odhady rozptýlení s_{Rj} a místo aritmetických průměrů d^* a σ^* lze použít robustní charakteristiky polohy $T(G)$, $T(s_R)$. Pro regulační meze lze pak použít vztahy

$$UCL \quad ' \quad T(G) \ \% \ 3 \ T(s_R) \ D(G) / E(\ T(s_R)) \ ,$$

$$LCL \quad ' \quad T(G) \ \& \ 3 \ T(s_R) \ D(G) / E(T(s_R)) \ ,$$

kde $E(.)$ a $D(.)$ označují střední hodnotu a rozptyl. Na základě detailnějšího testování bylo zjištěno, že dobré vlastnosti mají *mediánové regulační diagramy,* pro které jsou G_j mediány, s_{Rj} jsou interkvartilová rozmezí a $T(.)$ je aritmetický průměr. Místo mediánu lze také použít uřezaný průměr a stupeň uřezání je 0.25, cit.[16]. V případě, že data pocházejí z normálního rozdělení, jsou robustní regulační diagramy méně efektivní. To znamená, že regulační meze jsou široké. Tak např. pro mediánové diagramy a $N = 5$ je směrodatná odchylka o faktor 1.2 větší, tj. meze jsou o 20 % širší.

Klasické regulační diagramy "x s pruhem" jsou také málo citlivé na malé systematické změny střední hodnoty d (trend). Pro tyto účely se konstruují ještě výstražné meze ve vzdálenostech $\pm 2\sigma^*/\%N$ a $\pm \sigma^*/\%N$ od generálního průměru d^*. Pro indikaci trendu se pak používá celá řada heuristických pravidel, např.:

1. Jedna hodnota x_S leží mimo regulační meze (3σ).
2. Dva ze tří po sobě následujících bodů leží mimo výstražné meze (2σ) na stejné straně od d^*.
3. Čtyři z pěti po sobě následujících hodnot x_{Sj} leží mimo výstražné meze (1σ) na stejné straně od d^*.
4. Osm po sobě následujících bodů leží na stejné straně od průměru.

Existuje ještě celá řada dalších heuristických pravidel, která mohou být pro speciální případy užitečná. Obyčejně se do regulačních diagramů Shewhartova typu vynášejí x_{Si} po konstantních časových intervalech. S ohledem na rychlost odhalení systematických změn je často výhodné použít také nekonstantní časové intervaly. Při konstrukci regulačních diagramů je třeba důkladně ověřit, zda použité výběry $V_1,...,V_M$ netvoří nenáhodná seskupení či trendy. Podrobnosti o konstrukci a ověřování regulačních "diagramů x s pruhem" lze nalézt v ISO 8258.

10.2.3 Regulační diagramy pro posouzení variability

Tyto regulační diagramy umožňují posouzení úrovně variability procesu. Vycházejí opět z předpokladu normality nezávislosti výběru a konstantnosti rozptylů.

Regulační diagram "s" má centrální linii $s_p = \sigma^* C_4$ a regulační meze

$$LCL = s_p \sqrt{\frac{\chi^2_{0.001}(N-1)}{(N-1)}},$$

$$UCL = s_p \sqrt{\frac{\chi^2_{0.999}(N-1)}{(N-1)}},$$

kde $\chi_\nu^2(N-1)$ je 100ν %ní kvantil χ^2-rozdělení s $N-1$ stupni volnosti. Volba $\nu = 0.001$ a $\nu = 0.999$ zajišťuje, že v regulačních mezích bude ležet (pokud je proces v požadovaném stavu) 99.8 % všech výběrových směrodatných odchylek.

V předpočítačové éře se často pro vyjádření variability používalo místo směrodatné odchylky s variační rozpětí R dat

$$R = x_{(N)} - x_{(1)},$$

kde $x_{(N)}$ je nejvyšší a $x_{(1)}$ nejmenší prvek výběru. Při konstrukci *regulačního diagramu* "R" lze pak použít průměrné rozpětí R_p z výběrů $V_1,\ldots,V_M$. Pro regulační meze platí

$$LCL = D_3 R_p,$$

$$UCL = D_4 R_p.$$

Hodnoty D_3, D_4 souvisejí pouze s rozsahem výběru a jsou tabelovány např v cit.[1].
Pro $N = 3$ je $D_3 = 0$, $D_4 = 2.575$, pro $N = 5$ je $D_3 = 0$, $D_4 = 2.115$ a pro $N = 10$ je $D_3 = 0.233$, $D_4 = 1.773$.

Je zřejmé, že snahou výrobců je snižování variability výroby. Z tohoto pohledu ztrácejí dolní meze pro regulační diagramy "s" a regulační diagramy "R" smysl a často se definuje pouze horní regulační mez. Při korektní konstrukci meze *UCL* pro regulační diagram "s" je nutné použít takový 100ν% kvantil χ^2-rozdělení, který odpovídá požadované pravděpodobnosti s jakou nemají výběrové směrodatné odchylky překročit regulační mez. Tedy pro případ, že minimálně 99.8 % výběrových směrodatných odchylek má ležet pod *UCL* se volí $\nu = 0.998$. Je zřejmé, že takto určené *UCL* bude *nižší* než *UCL*.

10.2.4 Regulační diagramy kumulativních součtů, CUSUM

Je zřejmé, že Shewhartovy diagramy jsou vlastně ekvivalentní opakovaným testům významnosti pro konstantní velikosti výběru. Jejich základní výhodou je jednoduchost a rychlá indikace velkých změn stavu procesu. Na střední a malé změny (resp. trendy) reagují pomalu. Základním omezením je předpoklad, že jednotlivé dílčí výběry jsou nezávislé. Pokud je účelem z co nejmenšího počtu výběrů zachytit nenáhodný trend (indikovaný např. posunem střední hodnoty), je vhodné použít regulačních diagramů typu CUSUM.

Page v roce 1954 navrhl konstrukci regulačních diagramů, založenou na součtu všech hodnot, takže i-tá hodnota v grafu obsahuje všechny "historické" informace, obsažené v hodnotě $(i-1)$ a změnu způsobenou přechodem ze stavu $(i-1)$ do stavu (i). Tyto

regulační diagramy slouží tedy obyčejně k rychlé detekci malých a středních změn průměrné úrovně znaku jakosti x. Ten má střední hodnotu $E(x) = d$ a rozptyl $D(x) = \sigma^2$. Nechť d_A označuje úroveň *přijatelné jakosti* (acceptable quality level, *AQL*) a d_R úroveň *nepřijatelné jakosti* (rejectable quality level, *RQL*). Veličina *AQL* se také často označuje jako *požadovaná, cílová hodnota*. Veličina $D = |d_A - d_R|$ pak určuje velikost změny průměrné úrovně znaku jakosti, kterou je nutno detekovat. Při klasickém postupu navrženém Pagem se v konstantních časových intervalech $i = 1, 2, \ldots$ získávají výběry V_i a počítají se vhodné charakteristiky T_i (obyčejně aritmetické průměry x_{Si}) a kumulativní součty těchto statistik se vynášejí proti indexu i.

Výpočet kumulativních součtů závisí na tom, zda je účelem detekovat systematickou změnu v jednom směru (pozitivní nebo negativní odchylky) nebo v obou směrech. Běžně se při výpočtu kumulativních součtů odečítá jistá *referenční hodnota* K a kumulativní součet po j-tém časovém intervalu (z výběru $V_1,\ldots,V_j$) je roven

$$S_j = \sum_{i=1}^{j} (x_{S,i} - K) \quad .$$

Referenční hodnota K se volí s ohledem na to, jaké systematické odchylky se mají detekovat. Obyčejně se volí $K = (d_A + d_R)/2$. Regulační diagram CUSUM je pak graf S_j v závislosti na j. Pro detekci kladné odchylky $(d_{R1} > d_A)$ se volí referenční hodnota $K_1 = (d_A + d_{R1})/2$ a podobně pro detekci záporné odchylky $(d_{R2} < d_A)$ se volí $K_2 = (d_A + d_{R2})/2$. Zde d_{R1} a d_{R2} jsou úrovně nepřijatelné jakosti ve směru kladných a záporných odchylek. Kladná odchylka od průměrné úrovně je pak detekována, pokud

$$S_r - \min_{0,j,r} (S_j) \geq h^- ,$$

a záporná odchylka je detekována, pokud

$$\max_{0\#j\#r} (S_j) - S_r \geq h^+ .$$

Veličiny h^+ a h^- jsou tzv. *rozhodné meze.* Při výpočtu S_r a S_j se používá referenční hodnota K_1 a při výpočtu charakteristik se používá referenční hodnota K_2. Tato verze diagramu CUSUM je vhodná zejména tam, kde je účelem pouze indikovat změnu úrovně. Vlastní postup je velmi blízký Waldově sekvenčnímu testu podílu věrohodností. Barnardova metoda konstrukce regulačních diagramů CUSUM spočívá v náhradě referenční hodnoty přímo veličinou d_A, takže se vynesou kumulativní součty

$$S_j = \sum_{i=1}^{j} (x_{Si} - d_A)$$

proti indexu j. Pokud zůstává proces na úrovni d_A, kolísají hodnoty S_j kolem nuly. Posun střední hodnoty se projeví trendem v diagramu. Pro jeho indikaci se často konstruuje tzv. *V-maska*, tj. výseč ve tvaru "<" umístěná u posledního vyneseného bodu (S_r, r). Pokud nějaká S_j $(j < r)$ leží mimo tuto výseč, došlo k posunu střední hodnoty u bodu (S_r, r). Pro použití V-masky je nezbytné zvolit vhodné měřítko na ose y kumulativních součtů. Obyčejně se předpokládá, že jednotková vzdálenost na této ose je rovna 2σ. V případě neznalosti směrodatné odchylky se využívá jejího odhadu $\hat{\sigma}_e$ z rozptylu výběrových

průměrů

$$\hat{\sigma}_e = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(x_{Si}-d_0)^2}\ .$$

Pokud nejsou výrazné odchylky mezi výběrovými průměry a v datech je pouze náhodné kolísání σ, volí se $\hat{\sigma}_e = \sigma_0/\sqrt{N}$. Konstrukce *V-masky* vyžaduje znalost vzdálenosti jejího vrcholu od posledního vyneseného bodu *w* a úhlu θ, který svírají ramena této masky s její horizontální osou. Parametry V-masky úzce souvisí s posuzováním odchylek od standardního stavu s využitím referenčních hodnot K_1, K_2 a rozhodných mezí h^- a h^+. Pro případ symetrických úrovní nepřijatelné jakosti lze psát

$$d_{R1} = d_A + \delta\sigma\ ,$$

$$d_{R2} = d_A - \delta\sigma\ ,$$

kde δ je standardizovaný posun střední hodnoty, který má být detekován. Je zřejmé, že velikost celkové změny procesu, která se má detekovat při simultánním sledování je

$$D = K_1 + K_2 = (d_{R1} - d_{R2})/2 = \delta\sigma\ .$$

V případě neznalosti σ se používá jeho odhad, definovaný rovnicí pro $\hat{\sigma}_e$. Obyčejně se ještě zvolí $*h^-* = *h^+* = h$, tj. jedna rozhodná mez. Pak pro vzdálenost vrcholu V-masky od posledního bodu platí

$$w = \frac{h}{2\sigma\,\mathrm{tg}\,\Theta}$$

a pro úhel θ je možno psát

$$\mathrm{tg}\ \theta = \frac{D}{2\sigma} = \frac{\delta}{2}\ .$$

Z těchto rovnic je zřejmé, že existuje úzká souvislost mezi referenčními hodnotami K_1 a K_2 rozhodnou mezí *h* a parametry (*w*, θ), charakterizujícími V-masku. Snadno lze určit, že rozhodná mez

$$h = w\ 2\sigma\ \mathrm{tg}\ \theta$$

je vlastně průsečíkem a referenční hodnota $K = \delta\sigma$ je směrnice přímky v souřadném systému, jehož počátek je posunut do bodu (S_r , *r*). Druhá přímka má úsek -*h* a směrnici -*K* a indikuje negativní odchylky. Existuje ještě celá řada různých variant V-masky (uříznutá, semiparabolická atp.), které vedou ke snížení průměrného počtu dat do indikace odchylky od referenčního stavu, tj. *ARL*.

Standardně začínají regulační diagramy CUSUM od nuly, tj. $S_0 = 0$. Modifikace FIR (fast initial response), která je citlivá na počáteční změny stavu procesu ve zvoleném směru, začíná s nenulovou hodnotou $S_0 = h/2$. Zrobustnění lze docílit např. tím, že se hodnoty x_{Sj}, převyšující specifikované meze, do výpočtu S_j nezahrnují. Obyčejně se x_{Sj} považuje za silně vybočující, pokud

$$\frac{*x_{Sj} \;\&\; d(*}{\sigma(} \;>\; 4\ .$$

Velmi dobré se jeví pravidlo *dvou extrémů,* kdy se x_{Sj} do výpočtu nezahrnuje, ale pokud vyjde totéž i pro x_{Sj+1} , je indikována výrazná změna stavu procesu[20].

Odpovídající výpočet průměrného počtu hodnot *E(L)* do indikace změny stavu procesu, tj. *ARL,* je dosti komplikovaný. Pokud je proces ve stavu se střední hodnotou d_A*(AQL)* a měření mají normální rozdělení, může být rozdělení *L* považováno za geometrické. To však již neplatí pro změnu stavu, kdy má střední hodnota úroveň d_R*(RQL).* Obecně je pro případ, kdy $S_0 = 0$,

$$ARL \;'\; \frac{N(0)}{1 \;\&\; P(0)}\ ,$$

kde $N(0)$ a $P(0)$ jsou Fredholmovy integrální rovnice druhého druhu, které je nutné pro zadané rozdělení veličiny *x* řešit numericky[19].

Moderní regulační diagramy typu **CUSUM** využívají kumulativních součtů standardizovaných odchylek od generálního průměru d^*. Předpokládejme, že známe odhady d^* a σ^*. Pak můžeme pro výběrový průměr x_{Sj} sestrojit normalizovanou náhodnou veličinu

$$z_j \;'\; \frac{(x_{Sj} \;\&\; d(\sqrt{N}}{\sigma(}$$

a určit dvojici kumulativních součtů

$$S_{H,j} \;'\; \max\left[0;\ (Z_j \;\&\; K) \;\%\; S_{H,j\&1}\right],$$

$$S_{L,j} \;'\; \max\left[0;\ (\&Z_j \;\&\; K) \;\%\; S_{L,j\&1}\right].$$

Hodnota *K* určuje polovinu průměrného posunu v *Z*-transformaci, která se má detekovat. Obyčejně se volí $K = 0.5$, což odpovídá posunu o jedno σ. Začíná se od $S_{H,0} = S_{L,0} = 0$. Suma S_H slouží k detekci pozitivního posunu a suma S_L k detekci negativního posunu střední hodnoty. Obě sumy jsou kladné. Pokud $S_{H,j}$, resp. $S_{L,j}$ překročí rozhodnou mez h_j, došlo k indikaci nenáhodného posunu střední hodnoty. Běžně se volí $h_j = 4$ nebo 5, i když se dá dokázat, že pro malé *j* by mělo být h_j menší. Přesnější funkce $h_j = f(j)$ lze nalézt v literatuře[2].

Zajímavou možností je kombinace *CUSUM a regulačního diagramu "x s pruhem".* Zde se počítají hodnoty Z_j, $S_{H,j}$, $S_{L,j}$. Indikací posunu střední hodnoty je:

a) $Z_j > 3$, resp. $Z_j < -3$,

b) $S_{H,j}$, resp. $S_{L,j} > h_j$.

Obecně má použití regulačních diagramů typu CUSUM následující přednosti:

1. Vyšší efektivnost oproti Shewhartovým diagramům pro případ standar-dizovaného posunu střední hodnoty $d_A - d_R = \delta\,\sigma$, δ 0 (0.5÷2).

2. Snadnou vizuální detekci posunu velikosti střední hodnoty z průměrné směrnice vynášených bodů.

3. Snadné určení místa, kde došlo k posunu střední hodnoty.

4. Vhodné pro případy, kdy náklady na získání experimentálních údajů jsou vysoké.

Na druhé straně lze nalézt také některé nevýhody. Mezi základní patří, že je třeba pro každý případ konstruovat *speciální V-masku.* Také regulační diagramy CUSUM vycházejí z celé řady předpokladů, jejichž nesplnění vede ke ztrátě efektivnosti. Rozhodující je především dobrý odhad σ. Pokud je tento odhad podhodnocený, snižuje se *ARL,* a pokud je nadhodnocený, *ARL* roste. Výrazný vliv má také nenormalita rozdělení veličiny *x*. Na *RQL* se efekt nenormality výrazně neprojeví. Kladná šikmost však výrazně sníží *ARL* na *AQL.* Negativní šikmost zvyšuje *ARL.* Výrazně zkreslující vliv má také autokorelace mezi původními daty. V případě kladné autokorelace se snižuje *ARL* a v případě záporné autokorelace se zvyšuje. Tedy ani diagramy CUSUM nejsou robustní vůči narušení předpokladů normality a nezávislosti. Podobně jako u Shewhartových diagramů lze i zde využít různých technik pro zlepšení, resp. eliminaci, těchto narušení, příp. modifikovat vlastní proces sestavení regulačních diagramů. Regulační diagramy CUSUM se dají použít také pro posouzení velikosti rozptýlení znaku jakosti *x*. Problémy činí zejména volba vhodné standardizace (modifikace rovnice pro z_j), která by zajistila alespoň přibližnou normalitu. Pro případ konstrukce diagramu CUSUM pro směrodatnou odchylku *s* se volí $Z_i = (s/\sigma')^{0.625}$, což zajišťuje přibližnou normalitu *Z* pro velikosti výběrů $N = 3 \div 20$.

10.2.5 Regulační diagramy na bázi lokálního vyhlazení

Existuje celá řada regulačních diagramů, využívajících různých typů vyhlazování. Mezi jednoduché patří *exponenciálně vážený pohyblivý průměr* (EWMA), označovaný také jako pohyblivý geometrický průměr

$$W_j = r x_{Sj} + (1 - r) W_{j-1},$$

kde pro $j = 0$ je $W_0 = d^*$ a r $(0 < r \leq 1)$ je parametr definující váhu, obyčejně se doporučuje $r = 0.25$. Rozptyl tohoto parametru je roven[4]

$$S_{Wj}^2 = \frac{\sigma^2}{N} \frac{r}{2 - r} (1 - (1 - r)^{2j}).$$

Pro $r > 0.2$ a $j \geq 5$ lze použít aproximaci

$$S_{Wj}^2 = \frac{\sigma^2}{N} \frac{r}{2 - r}.$$

Regulační meze pro pohyblivý geometrický průměr pak jsou

$$LCL = d' - 3\sqrt{\frac{r\sigma'^2}{N(2 - r)}},$$

$$UCL = d' + 3\sqrt{\frac{r\sigma'^2}{N(2 - r)}}.$$

Regulační diagram pohyblivého geometrického průměru obsahuje pouze regulační meze *LCL* a *UCL.* Vynášejí se do ní hodnoty W_j. Tento regulační diagram je výhodný pro případy, kdy je třeba detekovat s vysokou přesností malé změny stavu procesu. Je opět

nerobustní vůči odchylkám od normality a nezávislosti. Podobně jako u diagramů CUSUM, lze provést snadno kombinaci se Shewhartovými regulačními mezemi a provést zrobustnění s využitím identifikace odlehlých hodnot. Za jistých podmínek lze použít EWMA pro případy, kdy jsou data autokorelována. Tak např. pro proces popsaný schématem

$$x_t = x_{t-1} + g_t - \theta g_{t-1}$$

je predikce stavu $x_{t+1}(t)$ v čase t (tzv. jednokroková predikce) rovná vztahu

$$W_t = \hat{x}_{t+1}(t) .$$

Odpovídající chyby predikce $e_t = x_t - W_{t-1}$ jsou pak nezávislé náhodné veličiny s $E(e_t) = 0$ a $D(e_t) = \sigma_P^2$. Regulační diagram pro tento případ má regulační meze

$$U\hat{C}L = 3\hat{\sigma}_P ,$$

$$L\hat{C}L = -3\hat{\sigma}_P .$$

Pro odhad směrodatné odchylky chyby jednokrokové predikce σ_P lze použít průměrné absolutní odchylky $\Delta(t)$, počítané ze vztahu

$$\Delta(t) = \alpha |e_t| + (1 - \alpha) \Delta(t - 1) ,$$

kde α je váhový parametr. Protože pro normální rozdělení je směrodatná odchylka o 25 % vyšší než $\Delta(t)$, odhaduje se $\hat{\sigma}_P$ ze vztahu

$$\sigma_P(t) \doteq 1.25\ \Delta(t) .$$

Další možností je přímý výpočet vyhlazeného odhadu rozptylu chyby jednokrokové predikce

$$\hat{\sigma}_P^2(t) = \alpha\, e_t^2 + (1 - \alpha)\, \hat{\sigma}_P^2(t - 1) .$$

Doporučuje se použít rozmezí $0.03 \le \alpha \le 0.1$ s tím, že nižší hodnoty jsou vhodnější. Standardně se doporučuje $\alpha \doteq 0.05$. Další možností je použití přímo hodnot W_t s tím, že se postupně upravují regulační meze s ohledem na to, že jde o jednokrokovou predikci

$$U\hat{C}L_{t+1} = W_t + 3\hat{\sigma}_P ,$$

$$L\hat{C}L_{t+1} = W_t - 3\hat{\sigma}_P .$$

Centrální linie tohoto diagramu pro časovou periodu $(t + 1)$ je W_t a vynáší se do něho přímo x_{t+1} (příp. x_{St+1}). Pokud lze očekávat, že data tvoří spíše časové řady, je použití těchto modifikací regulačních diagramů EWMA výhodné pro posouzení dynamiky dějů[21]. Uvedené vztahy lze použít buď pro průměrné hodnoty x_{St} v čase t, resp. přímo pro jednotlivá pozorování x_t v čase t. Regulační diagramy EWMA se dají také snadno modifikovat pro neparametrické charakteristiky polohy, resp. variability, kdy je zajištěna necitlivost na typ rozdělení původních dat. Jednoduché je použití standardizovaného pořadí,

kdy se jednotlivé hodnoty x_t nahrazují odpovídajícími pořadími R_t. Lze použít také Wilcoxonova pořadí se znaménky. Pro hodnoty $x_{t-g+1},...,x_{t-1}$, x_t je pořadí definováno vztahem

$$R_t^{(} \;=\; 1 \,+\, \sum_{(i)} I_{[x_t>x_i]} \;,$$

kde indikátorová funkce $I_{[x_t>x_i]} = 1$ pro $x_t > x_i$ a v ostatních případech je nulová. Standardizované pořadí je pak

$$R_t \;=\; \frac{2}{g}\,(R_t^{(} - \frac{g+1}{2}) \;.$$

Lze ukázat, že $E(R_t) = 0$ a $D(R_t) = \frac{g^2-1}{3g^2}$ pro všechna t.

Využitím rovnice pro W_j je pak možno definovat exponenciálně vážený průměr pořadí

$$T_t \;=\; (1-r)\,T_{t-1} + r\,R_t \;,$$

kde pro $t = 0$ je $T_0 = 0$ a $0 < r < 1$ je parametr vyhlazení. Doporučená hodnota je $r = 0.3$. Pro případ větších g (obyčejně $g = 30$) a r nepříliš blízkého k jedné je rozdělení T_t přibližně normální se střední hodnotou $E(T_t) = 0$ a rozptylem

$$D(T_t) \;=\; \frac{r\,(g^2-1)}{3\,g^2\,(2-r)} \;.$$

Regulační meze pro tento typ regulačních diagramů jsou

$$LCL \;=\; -3\,\sqrt{\frac{r(g^2-1)}{3\,g^2\,(2-r)}} \;,$$

$$UCL \;=\; 3\,\sqrt{\frac{r\,(g^2-1)}{3\,g^2\,(2-r)}}$$

a vynáší se hodnoty T_t.

10.2.6 Regulační diagramy pro jednotlivé hodnoty

V některých případech není možné provést ve stejném čase N-tici měření pro sestavení výběru V. V řadě případů je také kolísání charakteristik sledovaného procesu příliš rychlé ve srovnání s měřením, takže "průměrování" postrádá smysl. Pak se konstruují regulační diagramy pro jednotlivá měření. Ty mají celou řadu nevýhod:

(a) Jsou citlivé na nenormalitu rozdělení znaku x (rozdělení průměru x_s se více blíží normálnímu než rozdělení původních hodnot).
(b) Jsou málo citlivé na posun střední hodnoty.
(c) Jsou negativně ovlivněny trendy v datech.

(d) Jsou citlivé na velikosti výběru ze kterých se odhadují parametry rozdělení.

Vyjděme z předpokladu, že znak jakosti x má normální rozdělení $N(d, \sigma^2)$. Parametry d a σ^2 se odhadují z výběru velikosti N. Vzhledem k tomu, že jde pouze o jeden výběr, je třeba aby $N \geq 50$ a před vlastní analýzou bylo provedeno ověření normality, resp. identifikace vybočujících měření.

Výhodné je použití Q-Q grafů, kdy lze simultánně ověřit jak normalitu, tak i přítomnost vybočujících měření. Principem je hodnocení linearity v grafu $x_{(i)}$ vs. $F_N^{-1}(P_i)$. Symbolem $x_{(i)}$ jsou označeny pořádkové statistiky, čili setříděné prvky výběru, a $P = i/(N + 1)$ je pořadová pravděpodobnost. Parametr d se odhaduje jako aritmetický průměr x_s a odhadem parametru σ^2 je výběrový rozptyl s^2. Vzhledem k tomu, že odpovídající směrodatná odchylka s je vychýleným odhadem, používá se místo ní nevychýlený odhad

$$s' = s / C_4 .$$

Regulační diagram "x" má centrální linii x_s a regulační meze

$$LCL = x_S - 3s' ,$$

$$UCL = x_S + 3s' .$$

Do tohoto grafu se vynášejí přímo naměřené hodnoty znaku x. Místo výběrové směrodatné odchylky se v praxi s oblibou používá *průměrného pohyblivého rozpětí MR*, definovaného vztahem

$$MR = \frac{1}{N-1} \sum_{i=2}^{N} |x_i - x_{i-1}| .$$

Pomocí MR lze definovat odhad směrodatné odchylky pro případ normálního rozdělení

$$\tilde{\sigma} = \frac{\sqrt{\pi}\, MR}{2} = 0.8865\, MR .$$

Regulační meze jsou pak definovány vztahy

$$\hat{UCL} = x_S + 3\tilde{\sigma} = x_S + 2.6595\, MR ,$$

$$\hat{LCL} = x_S - 3\tilde{\sigma} = x_S - 2.6595\, MR .$$

Doporučená velikost výběru je $N = 300$. Podobně jako u regulačních diagramů "x s pruhem", lze i zde v případě nenormality použít normalizační transformace a v případě vybočujících měření robustních odhadů. Pro jednotlivé hodnoty lze také snadno aplikovat diagramy CUSUM. Stačí pouze počítat veličinu Z_j ze vztahu

$$Z_j = \frac{x_j - x_S}{s'} ,$$

kde x_j je j-tá naměřená hodnota. Hawkins[3] navrhuje pro regulaci variability na základě

jednotlivých pozorování CUSUM regulační diagram. Pro tento diagram se volí Z_j ve tvaru

$$Z_j = \frac{\sqrt{*x_j - x_{j-1}*}\,\sigma^{-1} - 0.82218}{0.34914},$$

Tato rovnice je motivována faktem, že $E\left(\frac{\sqrt{*x_j - x_{j-1}*}}{\sigma}\right) = 0.82218$ a odpovídající rozptyl je roven $(0.34914)^2$, pokud má znak jakosti x rozdělení $N(0, \sigma^2)$. Velmi snadné je také použití různých typů EWMA regulačních karet pro jednotlivé hodnoty.

10.2.7 Regulační diagramy pro distrétní znaky

V řadě případů lze sledovaný znak jakosti rozdělit pouze do dvou kategorií: *vyhovující* a *nevyhovující* (zmetek). V různých časech lze získat z výběru velikosti N celkový počet x (nevyhovujících). Podíl nevyhovujících výrobků je pak zřejmě

$$P_N' = \frac{x}{N}.$$

Tento podíl je odhadem pravděpodobnosti výskytu nevyhovujících výrobků P_N. Je však třeba použít dostatečně vysoké N (obyčejně $N > 500$). Lze ukázat, že pro P_N nepříliš vzdálené od 0.5 a střední N mají veličiny

$$Z_1 = \frac{x - NP_N}{\sqrt{NP_N(1 - P_N)}}$$

a

$$Z_2 = \frac{P_N' - P_N}{\sqrt{P_N(1 - P_N)/N}}$$

přibližně normované normální rozdělení N(0, 1).

Regulační diagramy "np" pro počet nevyhovujících jednotek mají centrální linii NP_N^* a regulační meze

$$LCL = NP_N' - 3\sqrt{NP_N'(1 - P_N')},$$

$$UCL = NP_N' + 3\sqrt{NP_N'(1 - P_N')}.$$

Vynáší se do nich počet nevyhovujících výrobků určený z výběrů velikosti N.

Regulační diagramy "*p*" pro podíl nevyhovujících výrobků mají centrální linii P_N^* a regulační meze

$$LCL \quad ' \quad P_N(\; \& \; 3\sqrt{P_N((1 \; \& \; P_N()/N} \; ,$$

$$UCL \quad ' \quad P_N(\; \% \; 3\sqrt{P_N((1 \; \& \; P_N()/N} \; .$$

Problém při použití těchto diagramů spočívá v tom, že rovnice pro Z_1 a Z_2 platí velmi přibližně, zejména pro malá N a P_N vzdálená od 0.5. Pro malé výběry je možné použít korekční faktor $(2N)^{-1}$ a pak

$$LCL \quad ' \quad P_N(\; \& \; 3\sqrt{P_N((1 \; \& \; P_N()/N \; \& \; 1/(2N)} \; ,$$

$$UCL \quad ' \quad P_N(\; \% \; 3\sqrt{P_N((1 \; \& \; P_N()/N \; \% \; 1/(2N)} \; .$$

Přiblížení k normalitě lze docílit například použitím arkussinové transformace

$$A(P_N) \quad ' \quad \arcsin\sqrt{\frac{P_N \; \% \; 3/8}{N \; \% \; 3/4}} \; ,$$

pro kterou platí, že střední hodnota

$$E(A(P_N)) \quad ' \quad \arcsin\sqrt{P_N}$$

a rozptyl $$D(A(P_N)) \quad ' \quad 1/4N \; .$$

Lze tedy sestrojit regulační diagram *p,* do kterého se vynášejí $A(P_N)$ se střední linií arcsin [$\%P^*_N$] a regulačními mezemi

$$LCL \quad ' \quad \arcsin(\sqrt{P_N(}) \; \& \; 1.5\sqrt{N} \; ,$$

$$UCL \quad ' \quad \arcsin(\sqrt{P_N(}) \; \% \; 1.5\sqrt{N} \; .$$

Také pro počet, resp. podíl, nevyhovujících jednotek lze použít diagramy kumulativních součtů CUSUM. V některých případech je regulovanou veličinou počet vad výrobku *C*. Při konstrukci regulačního diagramu se vychází z předpokladu, že parametr *C* má Poissonovo rozdělení. Velikost *C* se v podstatě odhaduje jako průměrný počet vad C_P určený z *N* výrobků.

Regulační diagramy "*c*" mají centrální linii C_P a regulační meze jsou

$$LCL \quad ' \quad C_P \; \& \; 3\sqrt{C_P} \; ,$$

$$UCL \quad ' \quad C_P \; \% \; 3\sqrt{C_P} \; .$$

Tyto limity vycházejí opět z aproximace Poissonova rozdělení rozdělením normálním. Zlepšení aproximace lze docílit vhodnou transformací. Jednoduchá je transformace

$$B(C) = \sqrt{C} \pm \sqrt{C + 1},$$

kdy toleranční meze jsou $B(C_P) \pm 3$. Také pro počet vad lze využít diagramy CUSUM, kdy se dosazuje za Z_j veličina

$$Z = \frac{C - C_P}{\sqrt{C_P}}.$$

Další typy regulačních diagramů pro diskrétní znaky lze nalézt např. v cit.[1]. Obecně je důležité ověřit, zda diskrétní data pocházejí z binomického nebo Poissonova rozdělení a podle toho volit další zpracování:

A. Pro Poissonovo rozdělení platí, že $E(x) = \lambda = D(x)$, tj. střední hodnota je totožná s rozptylem. Parametr λ se odhaduje jako aritmetický průměr $\hat{\lambda} = \bar{x}_S$. Při konstrukci regulačních diagramů Shewhartova typu se pak používá buď normalizační aproximace, nebo vztahu mezi Poissonovým a χ^2-rozdělením.

B. Pro binomické rozdělení platí, že

$$E(x) = n\,p \quad \text{a} \quad D(x) = n\,p\,(1 - p).$$

To znamená, že rozptyl se nerovná střední hodnotě. Parametr p se odhaduje s pomocí výběrového aritmetického průměru $\hat{p} = \bar{x}_S / n$. Při konstrukci regulačních diagramů se používá aproximace Poissonovým nebo normálním rozdělením. Jednoduše lze ověřit shodu rozdělení diskrétních dat s Poissonovým a binomickým rozdělením využitím grafu poměru frekvencí[16]. Orientačně lze použít poměru rozptylů $V = s^2 / D(x)$, kde $D(x)$ se dosazuje podle toho, které rozdělení se ověřuje. Pokud leží V mimo interval $0.8 \leq V \leq 1.25$, znamená to, že dané rozdělení neaproximuje dobře experimentální data.

10.2.8 Regulační diagramy pro více proměnných

V řadě reálných situací se kvalita procesu vyjadřuje vektorem znaků jakosti měřených simultánně (ve stejném čase), nebo se pro vyjádření (nepřímé) jednoho znaku používá více různých metod. Pro případ, že jsou jednotlivé znaky jakosti vzájemně nezávislé, je pro ně možné sestrojit regulační diagramy individuálně, aniž dojde ke zkreslení. Čím více jsou znaky jakosti vzájemně korelovány, tím více jsou jednoduché regulační diagramy "zkresleny". Předpokládejme, že výsledkem sledování procesu jsou náhodné vektory $\boldsymbol{x}_1$, $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$. Každý vektor obsahuje q složek, reprezentujících q-tici znaků jakosti. Složky těchto vektorů mohou být buď jednotlivé hodnoty x_{ij}, $j = 1, \ldots, q$, nebo výběrové průměry x_{Sij}. Při konstrukci vícerozměrných regulačních diagramů se vychází z předpokladu, že vektory $\boldsymbol{x}_j$ jsou nezávislé a mají vícerozměrné normální rozdělení $N(\boldsymbol{\mu}, \boldsymbol{C})$ s vektorem středních hodnot $\boldsymbol{\mu}$ a kovarianční maticí $\boldsymbol{C}$. Pokud je proces ve standardním stavu, je $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ a $\boldsymbol{C} = \boldsymbol{C}_0$. Parametry $\boldsymbol{\mu}_0$ a $\boldsymbol{C}_0$ jsou buď zadány, nebo se odhadují z dat, kdy je proces ve standardním stavu. Přirozeným zobecněním Shewhartových regulačních diagramů typu "x s pruhem" jsou *Hotellingovy regulační diagramy,* které využívají *T^2-statistiky*, která je ekvivalentní Mahalanobisově vzdálenosti

$$T_i^2 \;=\; (\boldsymbol{x}_i - \boldsymbol{\mu}_0)^T\, \boldsymbol{C}_0^{-1}\, (\boldsymbol{x}_i - \boldsymbol{\mu}_0)\ .$$

Rozdělení náhodné veličiny T_i^2 souvisí s náhodnou veličinou mající F-rozdělení známým vztahem

$$T_\alpha^2(N,\ q) \;=\; \frac{N\,q}{N-q+1}\ F_\alpha(q,\ N-q+1)\ ,$$

kde N je počet vektorů, použitý k určení odhadu kovarianční matice $\boldsymbol{C}_0$, resp. vektoru středních hodnot $\boldsymbol{\mu}_0$. Jednodušší je aproximace rozdělení T_i^2 pomocí χ^2-rozdělení s q stupni volnosti. Výsledný regulační diagram je nesymetrický. Začíná na úrovni $T = 0$ a má pouze *UCL* definovanou vztahem

$$UCL \;=\; \chi^2_{0.9975}(q)\ .$$

Do tohoto grafu se vynášejí hodnoty T_i^2. Pokud se pracuje s výběrovými vektory x_{Sij}, lze definovat zobecněnou Hottelingovu statistiku[22]. Je možné také použít vícerozměrné analogie regulačních diagramů CUSUM. Jedna konstrukce diagramů CUSUM využívá veličiny

$$L_i \;=\; \sqrt{(\boldsymbol{s}_{i-1} + \boldsymbol{x}_i)^T\, \boldsymbol{C}_0^{-1}\, (\boldsymbol{s}_{i-1} + \boldsymbol{x}_i)}\ .$$

Vektory $\boldsymbol{s}_i$ se generují podle schématu

$$\boldsymbol{s}_i \;=\; \boldsymbol{0}\ ,\quad \text{pokud}\quad C_i \le k_1\ ,$$

nebo $$\boldsymbol{s}_i \;=\; (\boldsymbol{s}_{i-1} + \boldsymbol{x}_i)\ (1 - k_1 / L_i)\ .$$

Zde $\boldsymbol{0}$ je nulový vektor a $k_1 > 0$. Do této varianty diagramu CUSUM se vynášejí hodnoty

$$Y_i \;=\; \sqrt{\boldsymbol{s}_i^{\,T}\, \boldsymbol{C}_0^{-1}\, \boldsymbol{s}_i}\ .$$

Podrobnosti o volbě k_1 a indikaci změny stavu procesu jsou popsány v práci[24]. Další možností je použití vektorů *kumulativních součtů*

$$\boldsymbol{d}_i \;=\; \sum_{j=i-l_i+1}^{i} \boldsymbol{x}_j$$

a statistiky $MC_i \;=\; \max\left[0,\ \sqrt{\boldsymbol{d}_i^{\,T}\, \boldsymbol{C}_0^{-1}\, \boldsymbol{d}_i} - k_2\, l_i\right]$. Zde $k_2 > 0$ a dále

$l_i = l_{i-1} + 1$, pokud $MC_{i-1} > 0$, resp. $l_i = 1$, pokud $MC_{i-1} \le 0$, pro $i = 1$, 2, 3,... Do diagramu se vynášejí hodnoty MC_i. Podrobnosti o interpretaci tohoto diagramu lze nalézt v cit.[25]. Mezi účinné vícerozměrné regulační diagramy patří *vícerozměrná varianta exponenciálně váženého pohyblivého průměru (MEWMA)*

$$\boldsymbol{z}_i \;=\; \boldsymbol{R}\, \boldsymbol{x}_i + (\boldsymbol{E} - \boldsymbol{R})\, \boldsymbol{z}_{i-1}\ ,$$

kde $z_0 = 0$, $\boldsymbol{R}$ je diagonální matice s prvky r_i na hlavní diagonále a $\boldsymbol{E}$ je jednotková matice. Pokud není důvod pro různé "vážení" rozličných složek sledovaného vektoru znaků jakosti, volí se prvky $r_i = r$, $i = 1, ..., q$. Tato rovnice pak přechází na tvar

$$\boldsymbol{z}_i = r\,\boldsymbol{x}_i + (1 - r)\,\boldsymbol{z}_{i-1} .$$

Odpovídající kovarianční matice má tvar

$$\boldsymbol{C}_{Zi} = \frac{r[1 - (1 - r)^{2i}]}{2 - r}\,\boldsymbol{C}_0 .$$

Z těchto veličin se sestavuje statistika Hotellingova typu $T_i^2 = \boldsymbol{z}_i^{\boldsymbol{T}}\,\boldsymbol{C}_{\boldsymbol{Zi}}^{\boldsymbol{-1}}\,\boldsymbol{z}_i$, která se vynáší do příslušného diagramu. Pokud vyjde $T_i^2 > h_4$, je proces mimo standardní stav. Vhodné volby parametrů, určené z analýzy *ARL,* jsou $r = 0.1$ a $h_4 = 8.79$.

Je také možné sestavovat vícerozměrné analogie regulačních diagramů "s" nebo "s^2". Vychází se z předpokladu, že je k dispozici požadovaná kovarianční matice $\boldsymbol{C}_0$. Účelem je, aby se kovarianční matice určená z dat $\boldsymbol{C}$ od $\boldsymbol{C}_0$ výrazně nelišila. Je známo, že skalární míra vícerozměrného rozptylu je zobecněný rozptyl det($\boldsymbol{C}$), kde det(.) označuje determinant.

Vícerozměrná analogie "s" diagramů pak využívá zobecněné směrodatné odchylky $s_g = \sqrt{\det(\boldsymbol{C})}$. Je možné odvodit, že $E(s_g) = s_{g0}\,b_1$ a $D(s_g) = s_{g0}^2\,(b_1 - b_3^2)$, kde $s_g = \sqrt{\det(\boldsymbol{C_0})}$. Parametry b_1 a b_3 jsou dány vztahy

$$b_1 = \frac{\sum_{i=1}^{g}(N - q)}{(N - 1)^q} \quad \text{a} \quad b_3 = \left(\frac{2}{N - 1}\right)^{q/2} \cdot \frac{\Gamma(N/2)}{\Gamma((N - q)/2)} .$$

Za předpokladu přibližné normality zobecněné směrodatné odchylky s_g lze konstruovat regulační diagramy se střední linií $CL = s_{g0}\,b_3$ a regulačními mezemi

$$UCL = s_{g0}\,(b_3 + 3\sqrt{b_1 - b_3^2}),$$

$$LCL = s_{g0}\,(b_3 - 3\sqrt{b_1 - b_3^2}).$$

Do tohoto diagramu se vynášejí přímo hodnoty s_g. Vícerozměrná analogie regulačních diagramů s^2 vychází ze statistiky

$$W^* = -q(N - 1) - (N - 1)\ln s_g + (N - 1)\ln s_{g0} + (N - 1)\mathrm{tr}(\boldsymbol{C_0^{-1}C}),$$

kde tr(.) značí stopu matice. Veličina W^* odpovídá testovací statistice testu shody kovariančních matic věrohodnostním poměrem. Pro konstrukci regulačních diagramů se využívá toho, že W^* má asymptoticky χ^2-rozdělení s $q\,(q + 1)/2$ stupni volnosti. Speciálním případem jsou regresní regulační diagramy, resp. regulační diagramy založené na sledování vstupní a výstupní jakosti. Uvažujme pro jednoduchost, že v *první fázi* sledovaného procesu je jakostním znakem veličina x (může to být jakostní znak suroviny) a ve *druhé fázi* sledovaného procesu je jakostním znakem veličina y (může to být jakostní znak produktu). Jednotlivá měření (x_i , y_i) jsou realizována na stejné položce produkce (např. při kusové výrobě), takže je zřejmé, že $y = f(x, \boldsymbol{\beta})$, kde $f(x, \boldsymbol{\beta})$ je obecně neznámá modelová funkce.

Při sestavování regulačního diagramu se pak postupuje v těchto krocích[26]:

a) Metodami regresní analýzy se z počáteční N-tice bodů (y_i , x_i), $i = 1, ..., N$ (kdy je proces ve standardním stavu) určí regresní model a jeho parametry $f(x, \mathbf{b})$,

b) stanoví se rezidua $e_i = y_i - f(x_i, \boldsymbol{b})$ a použije se analogie Shewhartova diagramu pro jednotlivá pozorování. Střední úroveň v tomto diagramu je

$$CL = \sum_{i=1}^{N} e_i/N.$$

Při použití lineární regrese a metody nejmenších čtverců je vždy $CL = 0$. Pro regulační meze pak platí

$$UCL = CL + 2.66\ MR,$$
$$LCL = CL - 2.66\ MR,$$

kde MR je odhad směrodatné odchylky reziduí počítaný z pohyblivého rozpětí

$$MR = \frac{1}{N-1} \sum_{i=1}^{N-1} |e_{i+1} - e_i| .$$

Do takto konstruovaného grafu se vynášejí rezidua $e_j = y_i - f(x_j, \boldsymbol{b})$, určená na základě dalších dvojic (y_j , x_j) získávaných postupně v průběhu sledování daného procesu. Tuto variantu regresních regulačních diagramů lze bez potíží zobecnit na případ vektoru vstupních znaků jakosti $\boldsymbol{x}$. Funkce $f(x, \boldsymbol{b})$ může být lineární i nelineární a při její konstrukci se využívá všech technik budování regresních modelů[16].

V některých případech je výhodnější použít přímo regresní model $f(x, \boldsymbol{b})$ a jako regulační meze konfidenční pásy. Jejich konstrukce je popsána např. v cit.[16] (intervaly spolehlivosti predikce). Při praktickém použití je vhodné kombinovat regresní regulační diagramy s regulačními diagramy typu "x s pruhem" pro obě proměnné. Také u vícerozměrných regulačních diagramů jsou kritickými předpoklady normalita a nezávislost vektorů $\boldsymbol{x}_i$. Kromě uvedených regulačních diagramů lze využít hlavních komponent, resp. Andrewsovy techniky znázornění vícerozměrných dat[23].

10.2.9 Používání regulačních diagramů

Regulační diagram má obecně sloužit jako diagnostický nástroj k posouzení, zda se sledovaný proces (představovaný nějakou měřenou veličinou nebo veličinami, které jej charakterizují) chová tak, jak očekáváme, zvláště pak, nedošlo-li k nečekané změně procesu. Došlo-li k takové změně, je třeba ji interpretovat - vysvětlit a případně přistoupit k nějakému zásahu. Příkladem měřených veličin jsou spojité veličiny jako pevnost, koncentrace, rozměr, elektrický odpor nebo diskrétní veličiny jako podíl zmetků na 1000 výrobků, počet povrchových vad na laku nebo počet uzlíků na 1 m^2 tkaniny. Kromě samotné hodnoty je nutno v případě spojité veličiny sledovat také její variabilitu (míru kolísání či rozptylu), která je pro posouzení procesu stejně důležitá. Proto Shewhartův regulační diagram musí vždy obsahovat informace jak o sledované hodnotě samotné, tak o její variabilitě.

10.2.10 Konstrukce regulačních diagramů

Postup konstrukce regulačního diagramu

1. Zvolíme takovou část procesu, která odpovídá naší představě, předpisu nebo zkušenosti a připravíme příslušná procesní data.

2. Na základě těchto dat stanovíme jejich statistický model. Obyčejně však máme pouze střední hodnotu (aritmetický průměr) a směrodatnou odchylku a ověříme platnost statistických předpokladů Shewhartova diagramu.

3. Z těchto dvou parametrů se zkonstruuje vlastní regulační diagram, který má podobu základní linie *ZL* (angl. central line, *CL*) a horní a spodní regulační meze *LCL* a *UCL* (angl. lower control level a upper control level).

4. Do tohoto regulačního diagramu se pak vynášejí další data z procesu a sleduje se výskyt 'zvláštních případů', signalizujících nečekanou změnu chování procesu, z nichž základní je překročení regulační meze.

5. Výskyt zvláštních případů se eviduje a hledá se tzv. přiřaditelná příčina (pokud se ji podaří identifikovat) a opatření, které bylo přijato.

Základními předpoklady pro Shewhartův regulační diagramu měřením jsou:

a) Normalita rozdělení dat, symetrie.
b) Konstantní střední hodnota procesu.
c) Konstantní rozptyl (směrodatná odchylka) dat.
d) Nezávislost, nekorelovanost dat.
e) Nepřítomnost vybočujících hodnot.
f) Vhodně zvolené podskupiny.

Tyto předpoklady je nutno testovat před konstrukcí regulačního diagramu postupy pro analýzu jednorozměrného výběru. Pokud se nepodaří ověřit předpoklady pro použití diagramu v bodě b), je nutno zdroje porušení předpokladů ověřit. V případě, že je zdroj náhodný a není předpoklad, že by se měl opakovat, je možné "problematická data" (např. vybočující hodnoty) ze souboru vyloučit a diagram konstruovat bez nich. Pokud je ale porušení předpokladů systematické, je inherentní vlastností procesu, nebo se jej nepodaří uspokojivě vysvětlit, není možné příslušná data vylučovat. Pak je třeba uvažovat o jiném typu regulačních diagramů. Zvláštní případy jsou takové situace, které jsou při optimálním průběhu procesu velmi nepravděpodobné. Historicky prvním případem je překročení regulačních mezí. Jsou-li regulační meze pro normálně rozdělená data stanoveny jako ±3s, je pravděpodobnost jejich překročení 0.27 %. To znamená, že k překročení dojde v průměru jednou z 1/0.0027 = 370 případů, tedy přibližně jednou za rok, máme-li jedno měření denně.

Vzorová úloha 10.1 *Aplikace regulačního diagramu pro průměry a směrodatné odchylky*. Při konstrukci diagramu *"x* s pruhem" se vychází z průměrů a směrodatných odchylek tzv. logických podskupin. Aby bylo možné sledovat jak úroveň procesu, tak i průběh jeho variability, je nutné používat dva diagramy. První je založen na průměrech (diagram *x* s pruhem), druhý na směrodatných odchylkách (diagram *s*).

Typickou strukturu dat uvádí následující tabulka. Ze vzorku představujícího jeden bod regulačního diagramu se vypočítá aritmetický průměr a směrodatná odchylka.

Data: Data pro konstrukci diagramu "*x* s pruhem" a *s* jsou obvykle v tabulce tvaru

x_1	x_2	x_3	...	x_n	Průměr	Směr. odch.
x_{11}	x_{21}	x_{31}	...	x_{n1}	$\bar{x}_1$	s_1
x_{12}	x_{22}	x_{32}	...	x_{n2}	$\bar{x}_2$	s_2
x_{13}	x_{23}	x_{33}	...	x_{n3}	$\bar{x}_3$	s_3
...	...	...	...	...	...	...
x_{1m}	x_{2m}	x_{3m}	...	...	$\bar{x}_m$	...
					$\bar{\bar{x}}$	$\bar{s}$

Ukázka dat pro konstrukci regulačního diagramu "*x* s pruhem" a *s:*

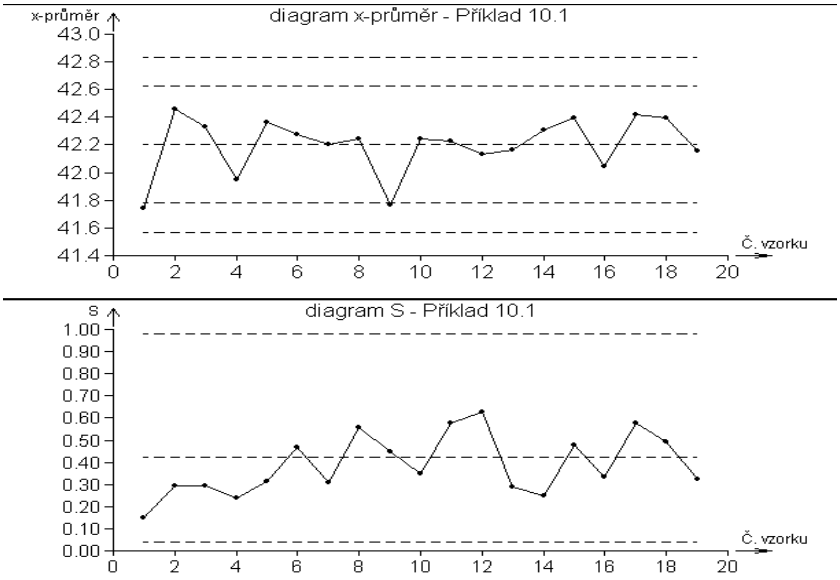
A1	*A2*	*A3*	*A4*	Průměr	Sm. odch.
41.56	41.82	41.90	41.68	41.740	0.151
42.05	42.44	42.69	42.66	42.46	0.295
42.31	42.42	42.65	41.95	42.333	0.292
41.87	41.68	42.25	41.99	41.948	0.239
42.47	42.66	41.92	42.41	42.365	0.315
42.95	41.95	42.25	41.96	42.278	0.469
42.43	42.51	41.92	41.95	42.203	0.311
41.62	42.89	42.49	41.99	42.248	0.557
41.36	41.68	41.63	42.41	41.77	0.449
41.96	42.66	42.41	41.96	42.248	0.347
42.97	41.95	42.37	41.63	42.23	0.579
41.36	41.99	42.86	42.31	42.13	0.627
41.96	42.41	42.41	41.87	42.163	0.288
42.44	41.94	42.37	42.47	42.305	0.247
41.76	42.64	42.86	42.31	42.393	0.478
41.72	42.09	42.49	41.87	42.043	0.335
42.59	43	41.63	42.47	42.423	0.575
42.47	41.75	42.41	42.95	42.395	0.493
42.12	41.72	42.37	42.43	42.16	0.323
				42.202	**0.388**

Řešení: Vyčíslí se základní linie a regulační meze diagramu "*x* s pruhem". Nestranný odhad směrodatné odchylky *s* se získá z průměru směrodatných odchylek. Hodnoty *LCL* a *UCL* zde představují 0.135 % a 99.865 % kvantily. Interval *(LCL, UCL)* tak vymezuje 99.73 % očekávaných naměřených dat. Pravděpodobnost překročení regulačních mezí je tak malá (0.27 %), že považujeme překročení za indikaci poruchy procesu. Vyčíslí se rovněž základní linie a regulační meze diagramu *s*. Při výpočtu regulačních mezí pro směrodatnou odchylku jsme využili kvantilů, které odpovídají pravidlu 3*s*. Symbol $c^2_a(n)$ označuje *a*-kvantil χ^2-rozdělení s *n* stupni volnosti. Hodnoty těchto kvantilů pro $n = 1$ až 20 jsou uvedeny v příloze. Jsou také běžně dostupné na kalkulačkách, v tabulkových procesorech (např. funkce CHIINV v Excelu), statistickém softwaru (např. QC-Expert, nebo funkce chisq v S-Plus) a podobně. Tabulka uvádí hodnoty faktoru c_4 pro velikosti podskupiny

n = 2 až 19.

Hodnoty faktoru c_4 pro n = 2 až 19

n	$c4(n)$	n	$c4(n)$
2	0.79788	11	0.97535
3	0.88623	12	0.97756
4	0.92132	13	0.97941
5	0.93999	14	0.98097
6	0.95153	15	0.98232
7	0.95937	16	0.98348
8	0.96503	17	0.98451
9	0.96931	18	0.98541
10	0.97266	19	0.98621



Obr. 10.1 Regulační diagram "x s pruhem" (horní část) a s (dolní část).

Závěr: Výsledný diagram je uveden na obr. 10.1.

Vzorová úloha 10.2 *Aplikace diagramu R*

Diagram R pro rozpětí (angl. range) lze použít jako alternativu diagramu s. Rozpětí podskupiny je rozdíl největší a nejmenší hodnoty v podskupině, $R_i = x_{\max,i} - x_{\min,i}$. S omezenou přesností lze R použít pro výpočet odhadu směrodatné odchylky. Pro data v následující tabulce lze pak konstruovat diagramy "x s pruhem" a R s podobnými vlastnostmi jako mají diagramy "x s pruhem" a s.

Data: Data pro konstrukci diagramu "x s pruhem" a R

x_1	x_2	x_3	...	x_n	Průměr	Rozpětí
x_{11}	x_{21}	x_{31}	...	x_{n1}	$\bar{x}_1$	R_1
x_{12}	x_{22}	x_{32}	...	x_{n2}	$\bar{x}_2$	R_2
x_{13}	x_{23}	x_{33}	...	x_{n3}	$\bar{x}_3$	R_3
...	...	...	...	...	...	...
x_{1m}	x_{2m}	x_{3m}	...	...	$\bar{x}_m$	...
					$\bar{\bar{x}}$	$\bar{R}$

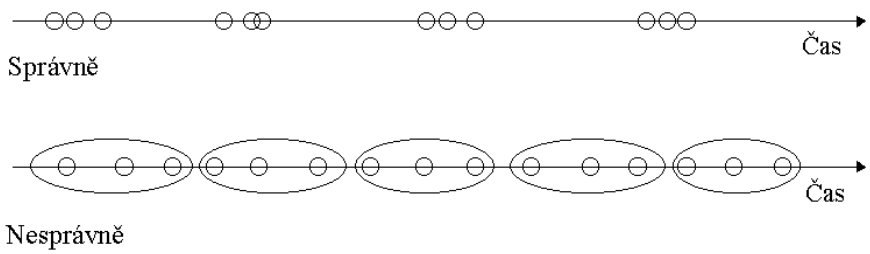
Pro diagram "x s pruhem" bude základní linie a regulační meze vyčísleny dle vztahů $UCL = \bar{\bar{x}} + A_2\bar{R}$, $CL = \bar{\bar{x}}$, $LCL = \bar{\bar{x}} - A_2\bar{R}$ a dále d_2 , d_3 , D_3 , D_4 a A_2 jsou tabelované koeficienty, když přibližné hodnoty uvádí tabulka:

Koeficienty pro diagramy x s pruhem a R:

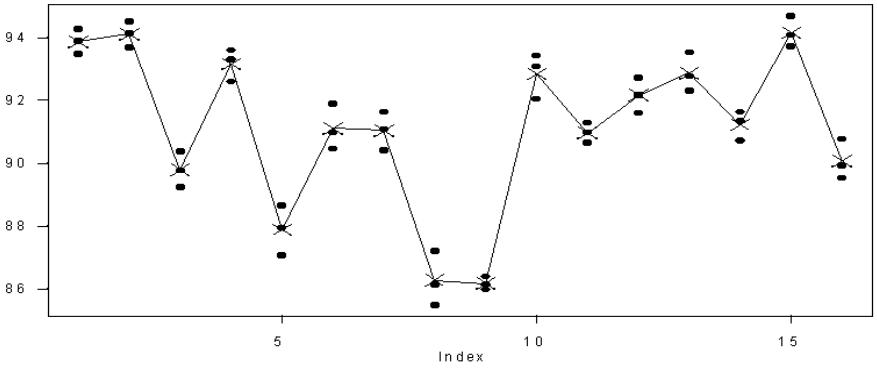
n	d_2	d_3	D_3	D_4	A_2
2	1.128	0.853	0	3.269	1.881
3	1.693	0.888	0	2.574	1.023
4	2.059	0.880	0	2.282	0.729
5	2.326	0.864	0	2.114	0.577
6	2.534	0.848	0	2.004	0.483
7	2.704	0.833	0.076	1.924	0.419
8	2.847	0.820	0.136	1.864	0.373
9	2.970	0.808	0.184	1.816	0.337
10	3.078	0.797	0.223	1.777	0.308
11	3.173	0.787	0.256	1.744	0.285
12	3.258	0.778	0.284	1.716	0.266
13	3.336	0.770	0.308	1.692	0.249
14	3.407	0.763	0.328	1.672	0.235
15	3.472	0.756	0.347	1.653	0.223
16	3.532	0.750	0.363	1.637	0.212
17	3.588	0.744	0.378	1.622	0.203
18	3.64	0.739	0.391	1.609	0.194
19	3.689	0.734	0.403	1.597	0.187
20	3.735	0.729	0.414	1.586	0.180

Řešení: Diagramy používající rozpětí místo směrodatné odchylky jsou méně efektivní (zvláště pro větší podskupiny), neboť využívají informace pouze o dvou hodnotách z celé podskupiny. Pro $n = 10$ je efektivita (přesnost) odhadu R jen 85 % ve srovnání se směrodatnou odchylkou s. V případě $n = 2$ je efektivita R stejná jako s. Diagram R má tedy opodstatnění v případě diagramu pro individuální hodnoty. Rozpětí má výhodu jednoduchého výpočtu ve srovnání se směrodatnou odchylkou, což může mít význam, není-li možno použít počítače. *Racionální podskupina* je pojem, který má zásadní význam pro správnou funkci diagramu "x s pruhem". Nevhodná volba podskupiny může mít za následek vážné problémy až selhání Shewhartova diagramu. Podskupina je obvykle opakované měření procesní veličiny odpovídající jednomu časovému okamžiku. Časový rozsah jednotlivých hodnot podskupiny se nerozlišuje, proto je vhodné, aby časový rozsah měření

v rámci podskupiny byl malý ve srovnání s časovým intervalem mezi podskupinami (obr. 10.2).

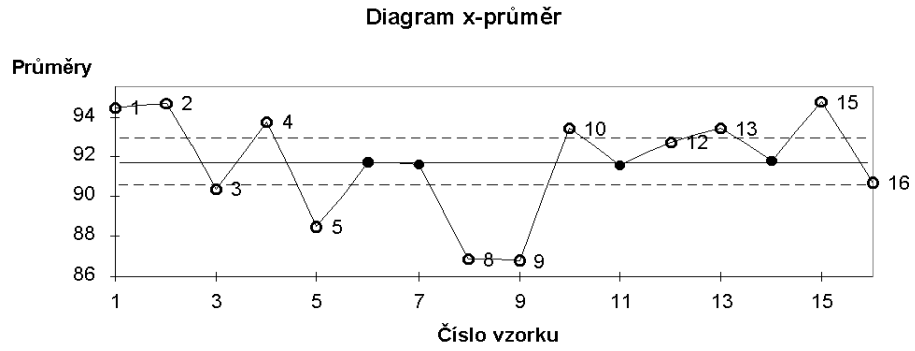


Obr. 10.2 Správná a nesprávná volba racionálních podskupin.

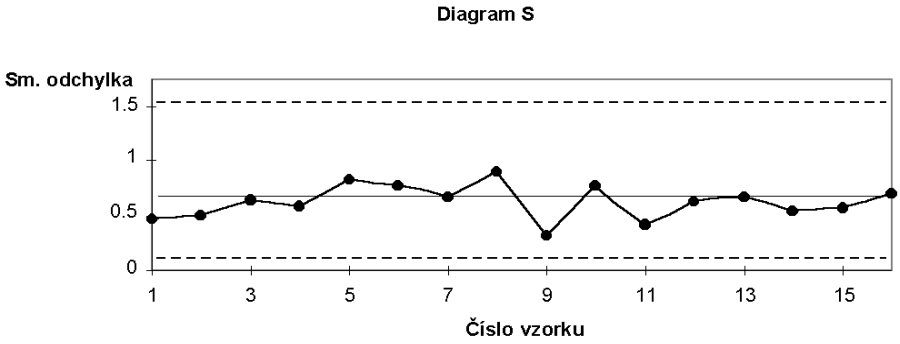


Obr. 10.3 Malá variabilita uvnitř podskupin, graf individuálních hodnot.

Zároveň však musíme dbát na to, aby hodnoty v podskupině odrážely dostatečně variabilitu měřené veličiny. Nedodržení této podmínky má za následek zúžení regulačních mezí a nepoužitelnost regulačního diagramu. Důsledek takové situace je znázorněn na obrázcích 10.3 až 10.5. Pro návrh podskupiny je důležité pochopení funkce diagramů "x s pruhem" a s. První indikuje především změny mezi jednotlivými podskupinami, kdežto druhý především změny a nestabilitu uvnitř podskupiny. Významná výhoda použití průměru podskupiny místo jednotlivých hodnot, na níž se často zapomíná, je výrazně lepší normalita dat, zvláště pro větší podskupiny ($n > 5$) díky platnosti centrální limitní věty. V některých případech je opakování měření těžko proveditelné (např. u destruktivních zkoušek), příliš nákladné, nebo časově náročné (např. chemické analýzy, zkoušky životnosti). Volba podskupin může být obtížná a násilná. Pak lze doporučit spíše použití diagramu pro individuální hodnoty, popřípadě regulačních diagramů CUSUM nebo EWMA.



Obr. 10.4 Diagram *"x* s pruhem" při příliš malé variabilitě uvnitř podskupiny pro data z obr. 10.3.



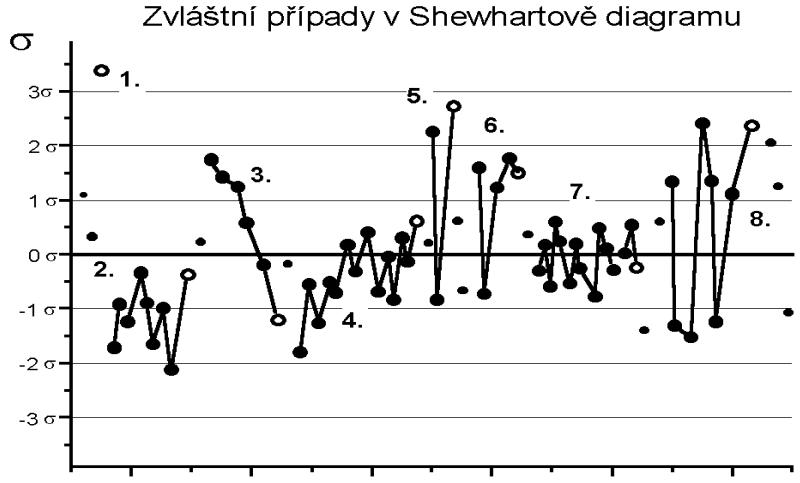
Obr. 10.5 Diagram *s* pro data z obr. 10.3.

10.2.11 Pravidla pro určování zvláštních případů

Zvláštní případy rozšiřují diagnostické možnosti Shewhartova diagramu a umožňují detekci poruch a změn, které se neprojeví překročením regulačních mezí, nebo by se projevily se zpožděním. Jedná se o osm nejpoužívanějších a ustálených situací. Pravděpodobnost jejich výskytu pro ideální data je srovnatelná s pravděpodobností překročení mezí. Pravděpodobnost výskytu následujících případů v normálně rozdělených nezávislých datech byla zjištěna výpočtem nebo simulacemi a je přibližně 0.25 %. V těchto osmi případech je nutno uvažovat o hledání přiřaditelné příčiny, případně regulačním zásahu. V případě jiného než normálního rozdělení může být tato pravděpodobnost o něco větší. Každý z těchto případů ukazuje na poruchu určitého druhu a lze jej použít jako užitečné vodítko při hledání přiřaditelné příčiny. Povahu možné poruchy uvádíme u jednotlivých případů. Grafické znázornění je na obr. 10.6.

Pravidla k odhalení zvláštních případů v diagramu "x s pruhem" a "x-individual" (podle ISO 8258) s komentářem

Pravidlo 1.	*Jedna hodnota je mimo regulační meze.* Lokální porucha procesu, chybné měření, výpadek. Chybně stanovené regulační meze, malá variabilita uvnitř podskupiny při konstrukci diagramu. Opakuje-li se na téže straně, může jít o posunutí střední hodnoty nebo o asymetrické rozdělení dat. Opakuje-li se na obou stranách, může jít o zvýšení nestability nebo rozptylu dat.
Pravidlo 2.	*9 hodnot je na téže straně od centrální linie.* Pravděpodobné posunutí střední hodnoty, snížení variability mezi podskupinami, asymetrie dat, příliš široké nebo neodpovídající regulační meze.
Pravidlo 3.	*6 hodnot monotónně roste či klesá.* Autokorelovaný proces, závislá měření. Lineární trend, způsobený opotřebením nebo výpadkem. Příliš široké regulační meze. Odstraněním přiřaditelné příčiny lze někdy zvýšit C_p.
Pravidlo 4.	*14 alternujících hodnot.* Přeregulovaný nebo nestabilní proces. Autokorelovaná měření se záporným *r*. Odstraněním přiřaditelné příčiny lze někdy zvýšit C_p. Podvádění operátorem, vymyšlená čísla.
Pravidlo 5.	*2 ze 3 hodnot je mimo interval ±2s.* Varování před možným překročením regulačních mezí.
Pravidlo 6.	*4 z 5 hodnot mimo interval ±s na téže straně centrální linie.* Pravděpodobné posunutí střední hodnoty. Varování před možným překročením regulačních mezí.
Pravidlo 7.	*15 hodnot je uvnitř intervalu ±s.* Snížení variability mezi podskupinami. Při opakování uvažovat o nových regulačních mezích. Nesprávná volba regulačních mezí. Podvádění operátorem, vymyšlená čísla.
Pravidlo 8.	*8 hodnot je mimo interval ±s na obou stranách centrální linie.* Zvýšení variability mezi podskupinami. Varování před překročením regulačních mezí. Porucha procesu.



Obr. 10.6 Znázornění pravidel pro určování zvláštních případů.

Vzorová úloha 10.3 *Aplikace regulačního diagramu pro jednotlivé hodnoty*

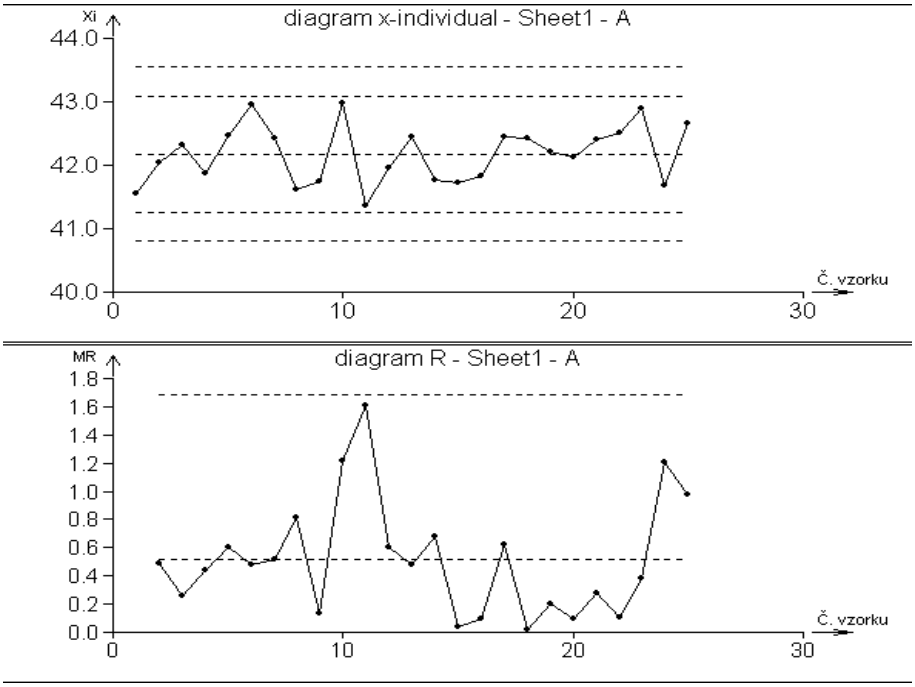
x. V případech, kdy z nějakého důvodu není účelné stanovování podskupin, lze použít Shewhartův diagram pro jednotlivé hodnoty x_i , zvaný také "x-individual". Místo průměrů podskupin se pracuje přímo s naměřenými hodnotami x_i .

Data: Příklad dat pro konstrukci diagramu "x-individual" a *MR*

i	x_i	MR_i	i	x_i	MR_i
1	41.56	-	14	41.76	0.68
2	42.05	0.49	15	41.72	0.04
3	42.31	0.26	16	41.82	0.1
4	41.87	0.44	17	42.44	0.62
5	42.47	0.6	18	42.42	0.02
6	42.95	0.48	19	42.22	0.2
7	42.43	0.52	20	42.12	0.1
8	41.62	0.81	21	42.4	0.28
9	41.75	0.13	22	42.51	0.11
10	42.97	1.22	23	42.89	0.38
11	41.36	1.61	24	41.68	1.21
12	41.96	0.6	25	42.66	0.98
13	42.44	0.48		$\bar{x}$ = **42.175**	$\overline{MR}$ = **0.515**

Řešení: Jako příslušný diagram pro variabilitu se používá *R*-diagram popsaný v předchozím odstavci. Místo rozpětí podskupiny se však použijí rozpětí mezi po sobě následujícími hodnotami. Tato hodnota se nazývá klouzavé rozpětí a označuje se *MR* (angl. moving range) a vyčíslí se dle vztahu $MR_i = *x_i - x_{i-1}*$. První hodnota MR_1 se nedefinuje. Statistické vlastnosti klouzavého rozpětí jsou stejné jako u rozpětí podskupiny pro $n = 2$. Koeficient d_2 má hodnotu 1.128. Pro základní linii a regulační meze diagramu X se použijí vztahy $UCL = \bar{x} \,\%\, 3\,\overline{MR}/d_2$, $LCL = \bar{x} \,\&\, 3\,\overline{MR}/d_2$, $CL = \bar{x}$ a pro diagram *MR* bude $UCL = D_4\overline{MR}$, $CL = \overline{R}$, $LCL = 0$. Koeficient D_4 je zde 3.269.

Závěr: Potíž s diagramem *MR* je v tom, že tyto vztahy platí pro nezávislé hodnoty *R*. Klouzavá rozpětí však nezávislá nejsou, pro výpočet MR_i a MR_{i-1} se použila společná hodnota x_{i-1} . Pro normálně rozdělené hodnoty x je autokorelační koeficient r_{MR} zhruba roven 0.22. Z tohoto důvodu někteří autoři nedoporučují diagram MR konstruovat.



Obr. 10.7 Regulační diagram *"x*-individual" (horní část) a *MR* (dolní část).

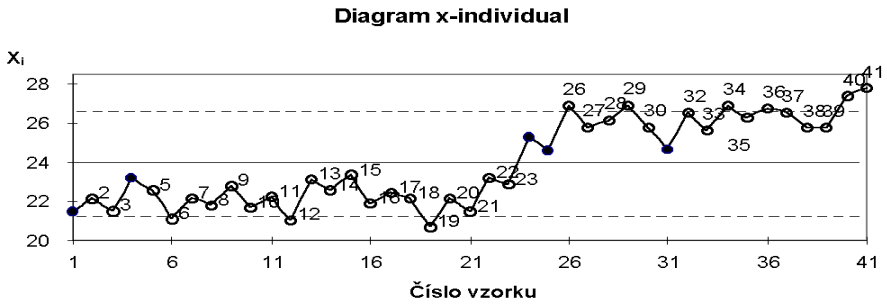
10.2.12 Porušení předpokladů o datech

Regulační diagram pro individuální hodnoty je na různá porušení předpokladů o datech obvykle citlivější než ostatní Shewhartovy diagramy. Především to platí o normalitě dat. V praxi se často setkáváme s procesy a procesními veličinami, které nejsou nezávislé, nemají konstantní střední hodnotu nebo rozptyl, nejsou normálně rozdělené atd. Ze zkušeností víme, že statisticky "dobré" chování většinou vykazují strojírenské procesy, kde jsou měřenými veličinami rozměry nebo hmotnosti. V případě měření dalších fyzikálních veličin, jako pevnosti, viskozity, se setkáme s asymetrickým rozdělením. Při sledování spojitých procesů v chemii, farmacii, potravinářstní, metalurgii vykazují data často silnou závislost. Těžko ovlivnitelná jakost suroviny (např. horniny) může mít za následek kolísání nebo nekonstantnost střední hodnoty. Při sledování emisí a stopových koncentrací nečistot se setkáme s logaritmicko-normálním, asymetrickým rozdělením atd. Ve všech těchto případech se jedná o vlastnost procesu, která se buď nedá ovlivnit, nebo se s ní v technologii počítá. Konstrukce Shewhartových diagramů může však v těchto případech selhat. Velmi hrubě jsme se pokusili shrnout zkušenosti s několika stovkami typických reálných datových souborů z různých odvětví a technologií. Uvádí je následující tabulka, ve které je kroužkem označeno převažující splnění předpokladu a křížkem pak jeho porušení.

Typická porušení předpokladů v různých technologiích

Odvětví / technologie / veličina	Normalita	Nezávislost	Konstantnost střední hodnoty	Homogenita, vybočující body
Mechanické strojírenství, automobilový průmysl (rozměry)	o	o	o	o
Mechanické zkušebny (pevnost, pružnost, ...)	x	o	o	x
Chemie, metalurgie, hutnictví (koncentrace, obsahy)	o	x	x	x
Chemie, metalurgie, hutnictví (ostatní fyzikální parametry)	o	x	x	x
Životní prostředí, hygiena (nízké koncentrace)	x	x	x	x
Elektrické veličiny, součástky	o	o	o	x
Energetika	x	x	x	x
Plasty, polymery, textil, fyzikálně-mechanické veličiny	x	o	x	o
Biochemie, farmacie, potravinářství	x	x	o	o
Vnitropodnikové ekonomické a finanční ukazatele	x	x	x	o
Sociologie, lidské zdroje	x	x	x	x

Tato tabulka naznačuje, že ve většině případů je na místě určitý pesimismus vzhledem k jednoduchému mechanickému použití klasických regulačních diagramů. Je nutné mít k dispozici a používat takové typy a techniky konstrukce regulačního diagramu, které odpovídají reálným datům a povaze procesu či technologie. V opačném případě může vést použití regulačních diagramů ke chybné interpretaci a nedůvěře k technikám statistického řízení procesu.

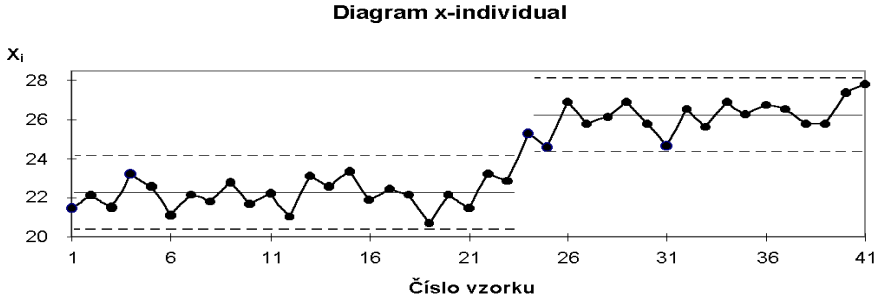


Obr. 10.8 Změna střední hodnoty ve Shewhartově regulačním diagramu typu *x*-individual.

Obr. 10.8, *nekonstantní střední hodnota,* ukazuje případ, kdy se změnila střední hodnota. Ke skokové změně dochází např. při přechodu na jinou surovinu nebo technologii. K posouzení významnosti rozdílu středních hodnot lze použít *t*-test.

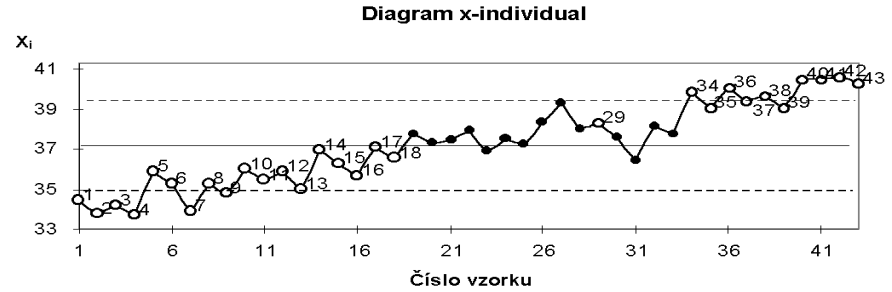
Obr. 10.9, *trvalá změna střední hodnoty*, ukazuje, kdy je tato změna trvalá a je třeba

konstruovat regulační diagram s novou základní linií.



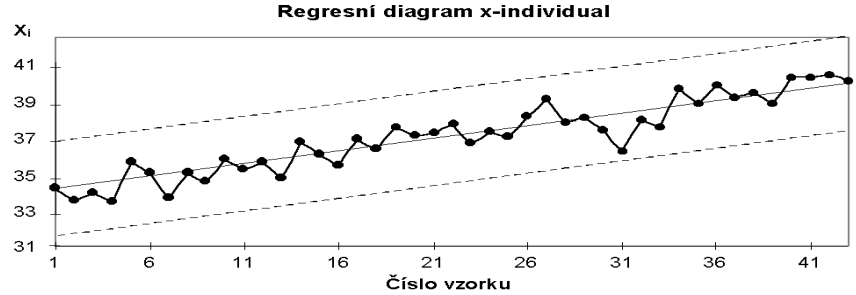
Obr. 10.9 Data z obr. 10.8 o dvou úrovních základní linie.

Obr. 10.10, *data vykazují stálý statisticky významný trend.* Trend je možné prokázat testem významnosti směrnice.



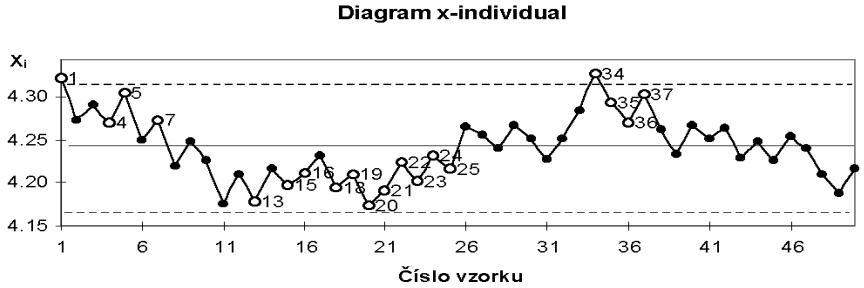
Obr. 10.10 Lineární trend v datech.

Obr. 10.11, *základní linie představuje regresní přímku* v případě lineárního trendu v datech. Alternativou je konstrukce regulačního diagramu pro rezidua (odchylky od regresní přímky).



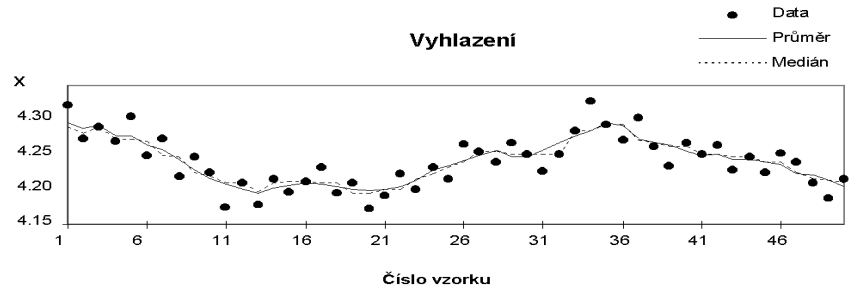
Obr. 10.11 Regresní kontrolní diagram pro data z obr. 10.10 .

Obr. 10.12, *kolísání dat a autokorelace v datech* se dá prokázat např. znaménkovým testem nebo testem významnosti autokorelačního koeficientu.

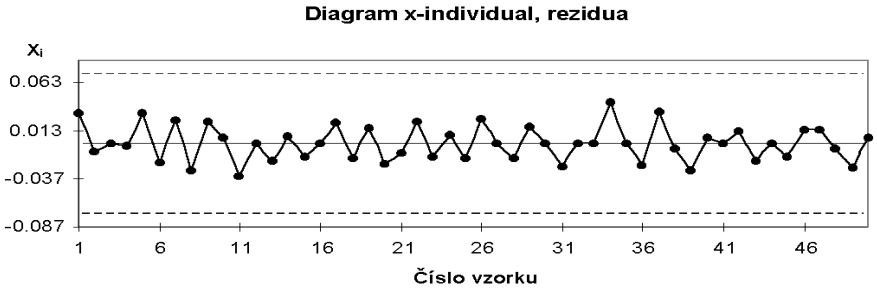


Obr. 10.12 Kolísání v datech, pokus o diagram "*x*-individual".

Obrázky 10.13 a 10.14, *vyhlazená data,* lze zpracovat dynamickým regulačním diagramem EWMA, případně vyhladit a pro konstrukci diagramu použít rezidua.

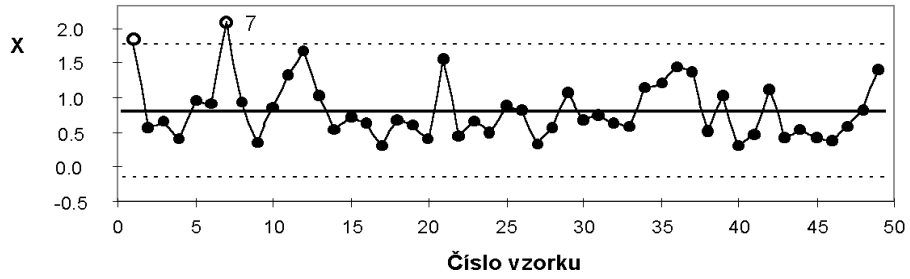


Obr. 10.13 Vyhlazená data z obr. 10.12.



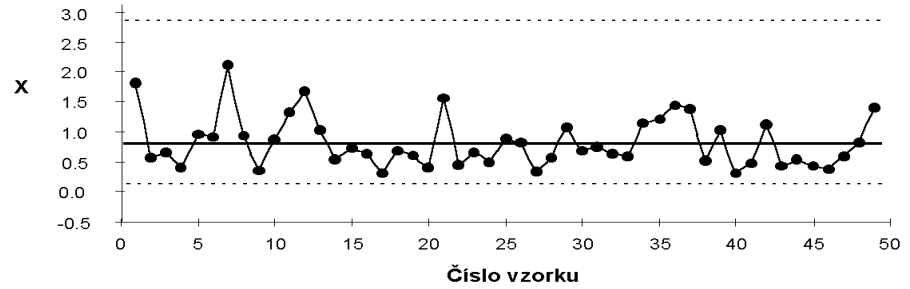
Obr. 10.14 Rezidua z vyhlazených dat na obr. 10.13 použitá pro diagram "*x*-individual".

Obr. 10.15, *asymetrické rozdělení dat* (koncentrací). Jako kritérium lze použít test normality nebo test významnosti mocninné transformace. Spodní regulační mez vyšla záporná, což je v daném případě nesmyslné.



Obr. 10.15 Klasický Shewhartův regulační diagram pro asymetrická data.

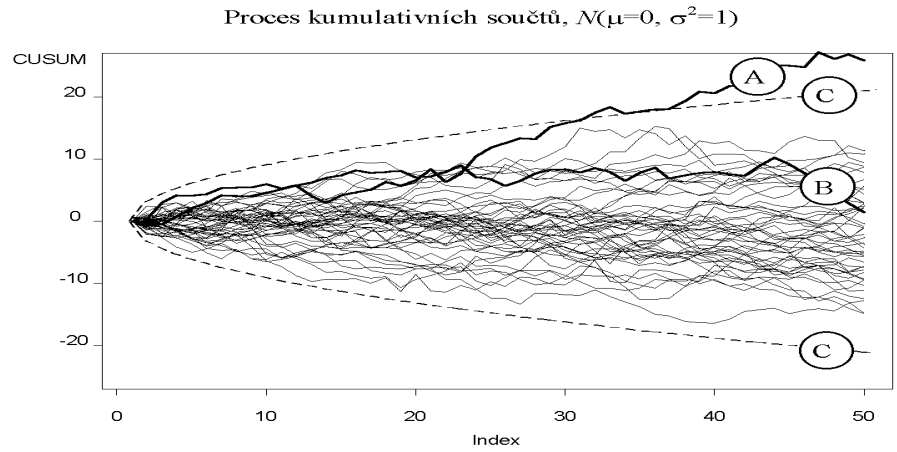
Obr. 10.16, *asymetrické regulační meze:* Regulační meze jsou konstruované pomocí retransformovaných kvantilů po nelineární Boxově-Coxově transformaci dat. Asymetrické regulační meze zde odpovídají povaze dat, spodní mez je kladná.



Obr. 10.16 Shewhartův regulační diagram s retransformovanými mezemi.

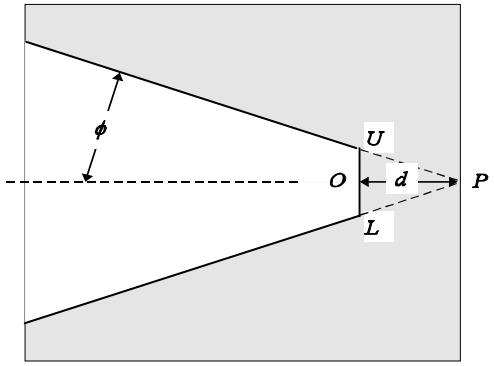
10.2.13 Pomůcky diagramů kumulativních součtů CUSUM

Diagramy, založené na kumulativních součtech (anglicky CUmulative SUMs) jsou rychlou detekcí relativně malého posunutí střední hodnoty procesu. Ve srovnání s Shewhartovými diagramy je tato detekce až o řád rychlejší. Ke konstrukci těchto diagramů se používá postupných součtů odchylek měřené veličiny od předepsané nebo očekávané konstantní cílové hodnoty *K*. Takový proces se ve statistice nazývá náhodné kráčení. Na obr. 10.17a je ukázka 50 diagramů, tvořených vygenerovanými posloupnostmi S_i. Má-li X_j konstantní střední hodnotu *K* a směrodatnou odchylku *s*, pak výsledná křivka na obr. 10.17a bude s velkou pravděpodobností ležet uvnitř parabolické oblasti C. Použijeme-li pravidla 3*s*, bude totiž oblast, v níž leží 99.73 % bodů s_i , ležet uvnitř paraboly $\pm 3s_X\%N$. Dojde-li k odchylce střední hodnoty *x* od hodnoty *K*, (křivka A na obr. 10.17a), proces tyto meze velmi rychle překročí. Reakce směrodatné odchylky s_i na odchylku od hodnoty *K* je mnohem citlivější než u Shewhartových diagramů, proto jsou diagramy CUSUM velmi oblíbené a rozšířené.



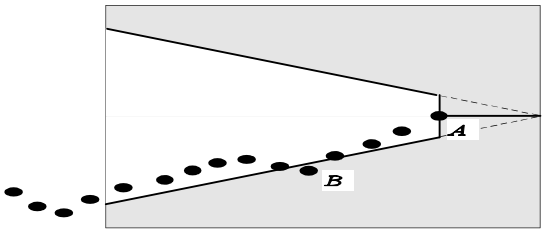
Obr. 10.17a Povaha procesu kumulativních součtů, *n*=50, A-vychýlená data, *m*=+0.6; B-nevychýlená data, *m*=0; C-parabolická mez.

Při praktickém použití se využívalo, a někde ještě využívá, speciální masky, která je zjednodušením paraboly, takzvané *V-masky*. V-maska, která je znázorněna na 10.17b, se vystřihla z papíru, a přikládala na graf kumulativních součtů.



Obr. 10.17b V-maska.

Geometrické parametry V-masky *d* a *f* lze vypočítat podle vztahů upravených do tvaru $d = (2/\delta^2)$ ln $(1-\beta)/\alpha$ φ tg^{-1} (Δx/2k), kde α, β jsou rizika I. resp. II. druhu, ΔX je odchylka od hodnoty *K,* kterou je třeba detekovat a *d* je její standardizovaná hodnota, $\delta = \Delta X/\sigma_X$, k je faktor měřítka v grafu, poměr velikosti jednotky na svislé a vodorovné ose. Hodnota δ (tedy citlivost na odchylku v jednotkách σ_X) se volí obyčejně mezi 0.5 a 1.5. Pro menší hodnotu δ je diagram příliš citlivý a je nebezpečí, že bude hlásit falešné poplachy, pro větší hodnotu δ je diagram málo citlivý a ztrácí se jeho přednost před Shewhartovým diagramem. Hodnota *ARL*, tedy průměrný počet dat mezi dvěma následnými překročeními regulačních mezí je pro ideální data u diagramu CUSUM 500 oproti 370 u srovnatelného Shewhartova diagramu.



Obr. 10.18 Použití V-masky, A-poslední vynesený bod; B-bod mimo výseč.

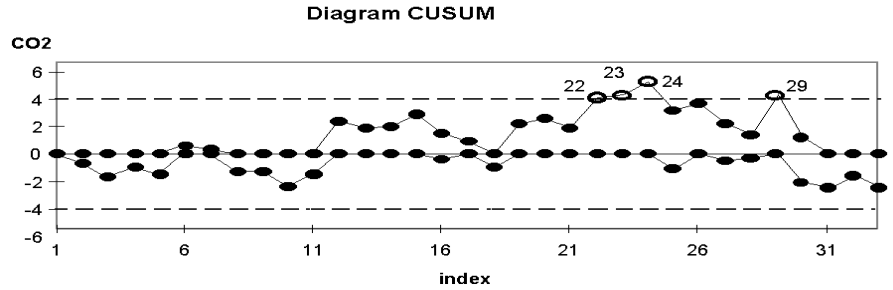
Při použití pomůcky V-masky se postupuje podle obr. 10.18. Maska se přiloží vodorovně bodem na poslední vynesený bod diagramu (A). Ocitne-li se některý z předchozích vynesených bodů mimo výseč (B), je bod A (nikoli B!) označen a proces se považuje za odchýlený od *K*. Výhodou tohoto postupu je jednoduchost výpočtu vynášených hodnot, což mělo význam na pracovištích bez výpočetní techniky. Nevýhodou je to, že vidíme vždy jen jediný bod (v našem případě bod A). Když chceme testovat jiný bod, musíme posunout masku. Nemáme tak přehled o delším úseku procesu, jak jsme na to byli zvyklí u Shewhartových diagramů. Tuto nevýhodu nemá např. rozšířený, modernější Lucasův postup konstrukce diagramu CUSUM.

Vzorová úloha 10.4 *Lucasova modifikace regulačního diagramu CUSUM*

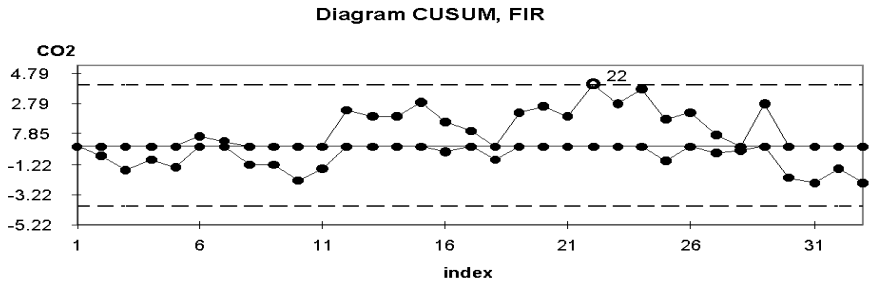
Regulační diagram CUSUM v efektivní modifikaci podle Lucase z roku 1976 (není nutné použití posuvné V-masky) se doporučuje především tam, kde je možné využít počítač a kde není k dispozici papírová V-maska, popř. příslušný nomogram. Tento diagram je rovněž založen na metodě kumulativních součtů odchylek od cílové střední hodnoty. Převažují-li významné odchylky na jednu stranu od centrální linie nad odchylkami na druhou stranu, indikuje diagram tuto skutečnost velmi rychle. O citlivosti diagramu rozhoduje parametr *K*, který zadává uživatel. Hodnota parametru *K* udává, na jak veliké posunutí v jednotkách *s* má diagram reagovat. Pokud je splněn předpoklad normality a nezávislosti původních dat, je tato metoda velmi účinná. Za porušení pravidel se považuje překročení mezí $\pm h$, které se zde nazývají rozhodné meze. Pokud ihned po překročení rozhodné meze provádíme zásah, můžeme použít techniku FIR. Tato technika umožní velmi rychlé ověření, zda byl zásah úspěšný, tím, že posune následující bod diagramu pod rozhodnou mez na té straně, kde došlo k překročení. Pokud pak po zásahu dojde k opětnému překročení meze, zásah nevedl k nápravě. Použití techniky FIR je patrné z obrázků 10.19 a 10.20. Do diagramu se vynášejí hodnoty $S_H(i)$ a $S_L(i)$, kde $S_H(0) = S_L(0) = 0$. Za K se nejčastěji volí 0.5. Pro konstrukci rozhodných mezí se používá $h_j = 4$, nebo $h_j = 5$. Následující tabulka ukazuje hodnoty *ARL* tohoto diagramu ve srovnání se Shewhartovým diagramem (bez uplatnění pravidel pro zvláštní případy).

Porovnání ARL u dvou diagramů, CUSUM a Shewhartova regulačního

	ARL		
Posunutí, *s*	*h* = 4	*h* = 5	Shewhart
0	168	465	370
0.25	74.2	139	281
0.5	26.6	38.0	155
0.75	13.3	17.0	81.2
1	8.38	10.4	43.9
1.5	4.75	5.75	15
2	3.34	4.01	6.3
2.5	2.62	3.11	3.2
3	2.19	2.57	2



Obr. 10.19 Diagram CUSUM bez FIR.

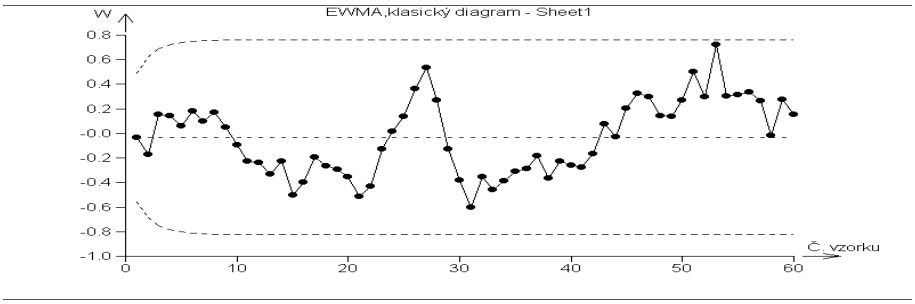


Obr. 10.20 Diagram CUSUM ze stejných dat jako v obr. 10.19, ale s technikou FIR.

Vzorová úloha 10.5 *Aplikace diagramů exponenciálně vážených klouzavých průměrů, EWMA*

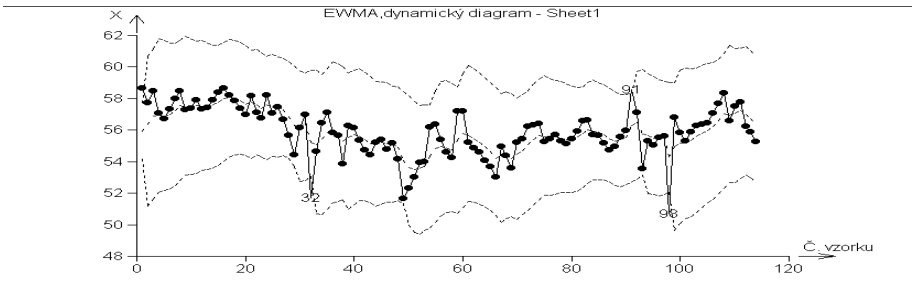
Diagram EWMA je zkratkou anglického Exponentially Weighted Moving Average, exponenciálně vážené klouzavé průměry, zvané někdy také exponenciální zapomínání. Jeho použití je podobné jako u Shewhartových diagramů. Každý bod diagramu W_j je váženým průměrem nově naměřené hodnoty x_j, případně průměru podskupiny velikosti N a posledního zaznamenaného bodu diagramu W_{j-1} . Základním volitelným parametrem diagramu EWMA je právě váha r, která může nabývat hodnot mezi 0 a 1. Hodnota $r = 1$ odpovídá Shewhartově diagramu, čím nižší je r, tím pomaleji reagují vynášené hodnoty W_j

na lokální změny sledovaného procesu. Vhodnou volbou parametru *r* lze diagram nastavit tak, aby nereagoval na lokální odchylky od předepsané hodnoty tak rychle jako diagram Shewhartův. Používá se tedy s výhodou tam, kde k takovým odchylkám dochází, aniž by se jednalo o poruchu. Na druhé straně má diagram tendenci zvýraznit systematickou dlouhodobější odchylku tím, že se vrací zpět k předepsané hodnotě pomaleji než vlastní měřená veličina. Tato vlastnost je tím výraznější, čím je *r* menší.
Řešení: Různí autoři se rozcházejí v doporučovaných hodnotách pro *r*. Obvykle se hodnota *r* volí mezi 0.15 a 0.4, nejčastěji $r = 0.25$. Podobně jako u spline lze k určení *r* pro daná data použít metody minimální střední kvadratické chyby predikce *MEP*. K určení regulačních mezí se využívá odhad rozptylu W_j . Vzhledem k přibližně normálnímu rozdělení W_j lze pro konstrukci regulačních mezí použít pravidlo 3 sigma. Takto definované meze pak odpovídají mezím v Shewhartově diagramu. Mají-li data normální rozdělení s konstantním rozptylem a střední hodnotou, je pravděpodobnost překročení mezí asi 0.25 %.



Obr. 10.21 Klasický diagram EWMA s nekonstantními regulačními mezemi.

Závěr: Výhodná je modifikace diagramu EWMA nazývaná *dynamický diagram EWMA* s jednokrokovou predikcí střední hodnoty a rozptylu. Tato modifikace je určená pro autokorelovaná data a umožňuje konstrukci regulačního diagramu pro procesy s nekonstantní střední hodnotou a nekonstantním rozptylem. Překročení regulačních mezí způsobí pouze náhlá změna střední hodnoty nebo zvýšení rozptylu. Pomalé změny procesu jsou tolerovány. Váhový parametr α se volí 0.05.

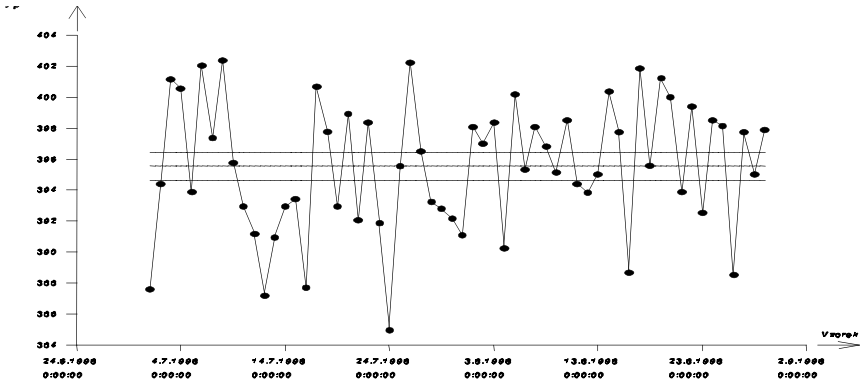


Obr. 10.22 Dynamický regulační diagram EWMA dovoluje pomalé změny střední hodnoty a rozptylu.

Vzorová úloha 10.6 *Kontrola tavby v metalurgickém provozu regulačním*

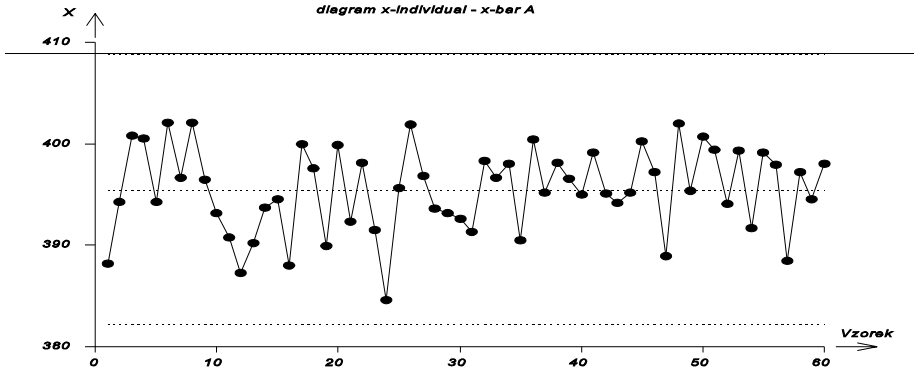
diagramem

Na úloze **C10.1** ukážeme kontrolu tavby v metalurgickém provoze regulačním diagramem, kdy z každé tavby byly odebrány 3 vzorky. Z těchto trojic se má sestrojit Shewhartův regulační diagram "*x* s pruhem". Vysvětlete jeho nepoužitelnost a navrhněte řešení.
Řešení: Vypočítáme *CL*=395.518; *LCL* = 394.603; *UCL* = 396.434 a sestrojíme regulační diagram, do něhož zakreslíme příslušné průměry, obr. 10.23.



Obr. 10.23 Diagram *"x* s pruhem".

Řešení: Téměř všechny body jsou mimo regulační meze. Regulační meze se konstruují na základě variability uvnitř podskupiny, která musí být srovnatelná s variabilitou mezi skupinami. Vyčíslíme-li tyto variability pomocí směrodatné odchylky, vyjde pro podskupinu hodnota 0.305, použitá ke konstrukci regulačních mezí, a mezi podskupinami hodnota 4.202.



Obr. 10.24 Diagram *"x*-individual".

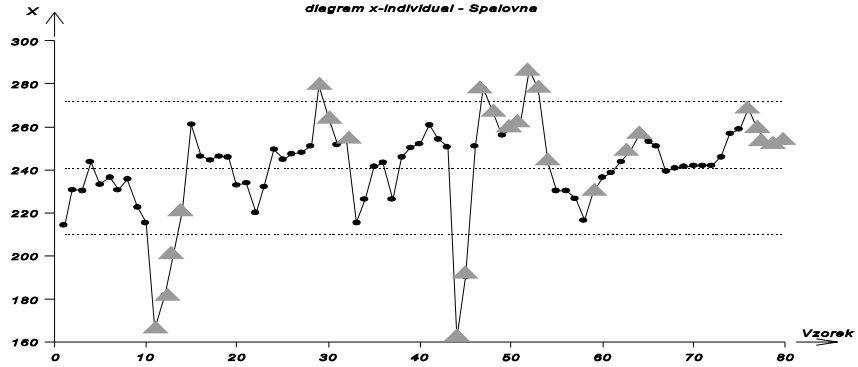
Závěr: Není třeba *F*-testu shody rozptylů, abychom viděli řádový rozdíl, který způsobil nereálnost regulačních mezí. Možnou příčinou tohoto stavu byl chybný odběr vzorku pouze z jednoho místa, což může snížit reprezentativnost podskupiny. Pokud byla tato možnost vyloučena, bylo by zřejmě výhodnější použít diagram pro individuální hodnoty, což by navíc ušetřilo zbytečné odběry, které z hlediska regulačního diagramu přinášely jen málo

informace. Diagram *"x*-individual" je na obr. 10.24.

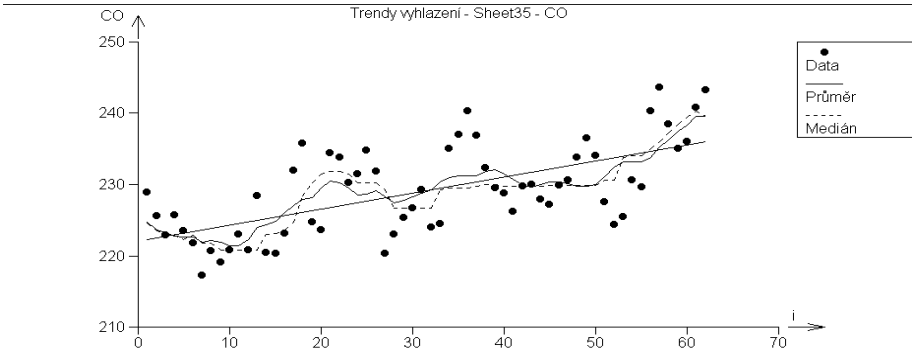
Vzorová úloha 10.7 *Chemická analýza složení plynných splodin*

Na úloze **C10.2** ukážeme chemickou analýzu složení plynných zplodin, která má být monitorována pomocí regulačního diagramu s cílem detekovat odchylky od statistické stability.

Řešení: Protože se jedná o individuální hodnoty bez opakování, bylo by formálně možné použít diagramu "*x*-individual". Jako centrální linii lze použít průměr dat, *CL* = 240.9, a regulační meze lze vypočítat z klouzavého rozpětí, *UCL* = 271.9; *LCL* = 209.8. Po konstrukci mezí a doplnění dat získáme následující diagram, obr. 10.25.



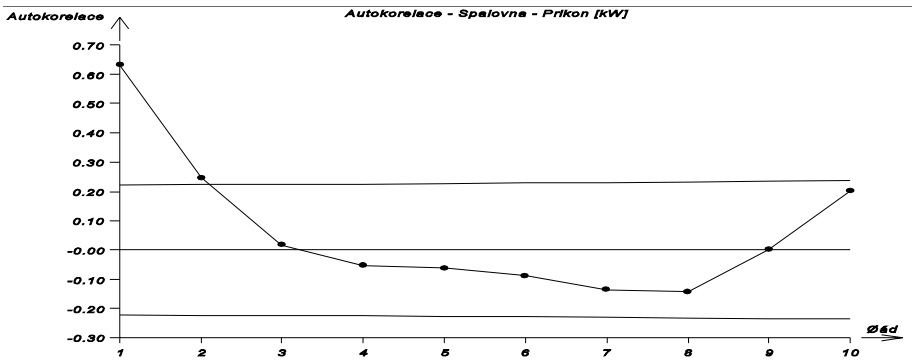
Obr. 10.25 Diagram *"x*-individual".



Obr. 10.26 Data s vyznačeným trendem a vyhlazením.

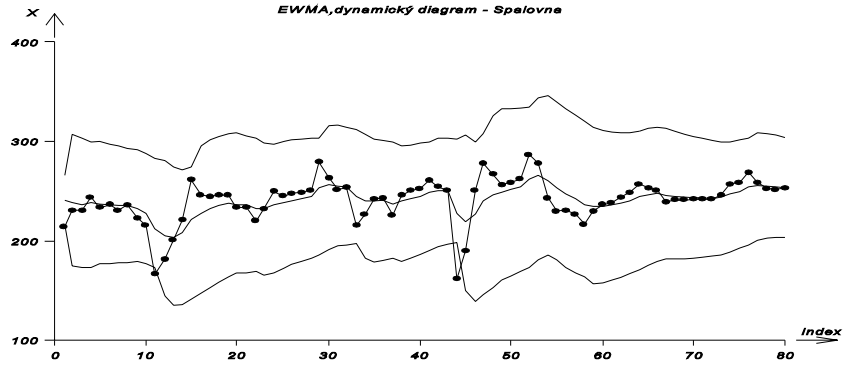
Několikeré překročení regulačních mezí je snad ještě možné vysvětlit poruchami nebo jiným hrubým porušením stability, avšak vysvětlení pro dalších 15(!) dat, která porušují podmínky dané zvláštními případy uvnitř regulačních mezí, je nutno hledat v porušení některého z předpokladů o datech pro konstrukci regulačních diagramů. V grafu na obr. 10.26 je zřejmý statisticky významný lineární trend znázorněný přímkou a kolísání hodnot, které je typické pro autokorelovaná data. Tento předpoklad potvrzuje i graf autokorelace na obr. 10.27, kde autokorelační koeficient prvního a druhého řádu je statisticky významný. Lze tedy konstatovat, že statistickým modelem našich dat s největší

pravděpodobností není nezávislé normální rozdělení s konstantní střední hodnotou, jak to vyžaduje konstrukce Shewhartova diagramu "*x*-individual".



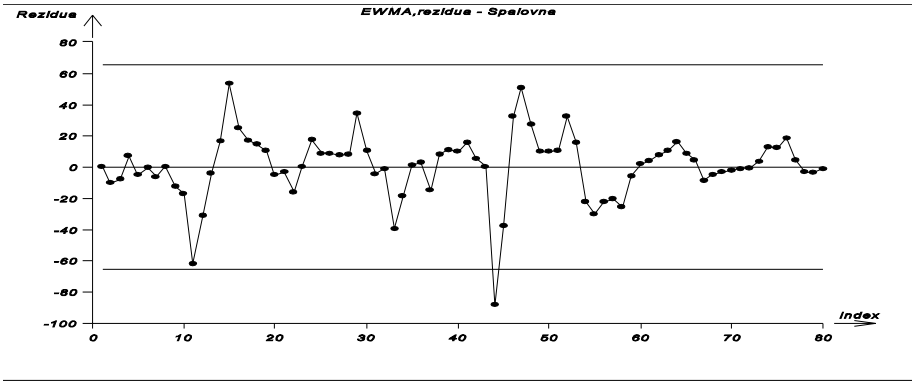
Obr. 10.27 Graf autokorelace s mezemi významnosti korelačních koeficientů.

Regulačním diagramem, který připouští autokorelaci i nekonstantnost střední hodnoty, může být například některý model časové řady jako autoregrese *AR*, klouzavý průměr *MA*. V praxi často je používaný dynamický diagram EWMA s jednokrokovou predikcí obr. 10.28, případně rezidua tohoto diagramu, obr. 10.29.



Obr. 10.28 Dynamický diagram EWMA s jednokrokovou predikcí.

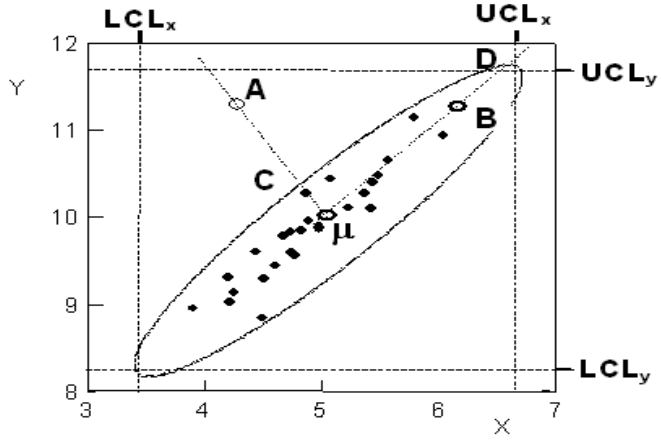
Závěr: Oba poslední diagramy ukazují dva hlavní problémy: bod 11 a bod 44, na něž je třeba se zaměřit při interpretaci. Ostatních 45 problémů detekovaných diagramem "*x*-individual" jsou zřejmě důsledkem závislosti a trendu v datech. Tento závěr však neznamená, že závislostí nebo trendy v datech se není třeba zabývat. Pokud by se podařilo autokorelaci technologicky vysvětlit a odstranit, dosáhli bychom snížení variability, zvýšení indexu způsobilosti, a tím zlepšení jakosti o faktor $\sqrt{(1 - r^2)}$, tedy v tomto případě zhruba o 25 %, protože autokorelační koeficient 1. řádu je r_1=0.632.



Obr. 10.29 Diagram EWMA, rezidua.

Vzorová úloha 10.8 *Aplikace Hotellingova regulačního diagramu*

Tabulka představuje dvourozměrná data *X* a *Y* (plné body), která jsou poměrně silně korelována. Hodnoty *UCL* a *LCL* představují regulační meze Shewhartových diagramů, které by se konstruovaly zvlášť pro *X* a pro *Y*.



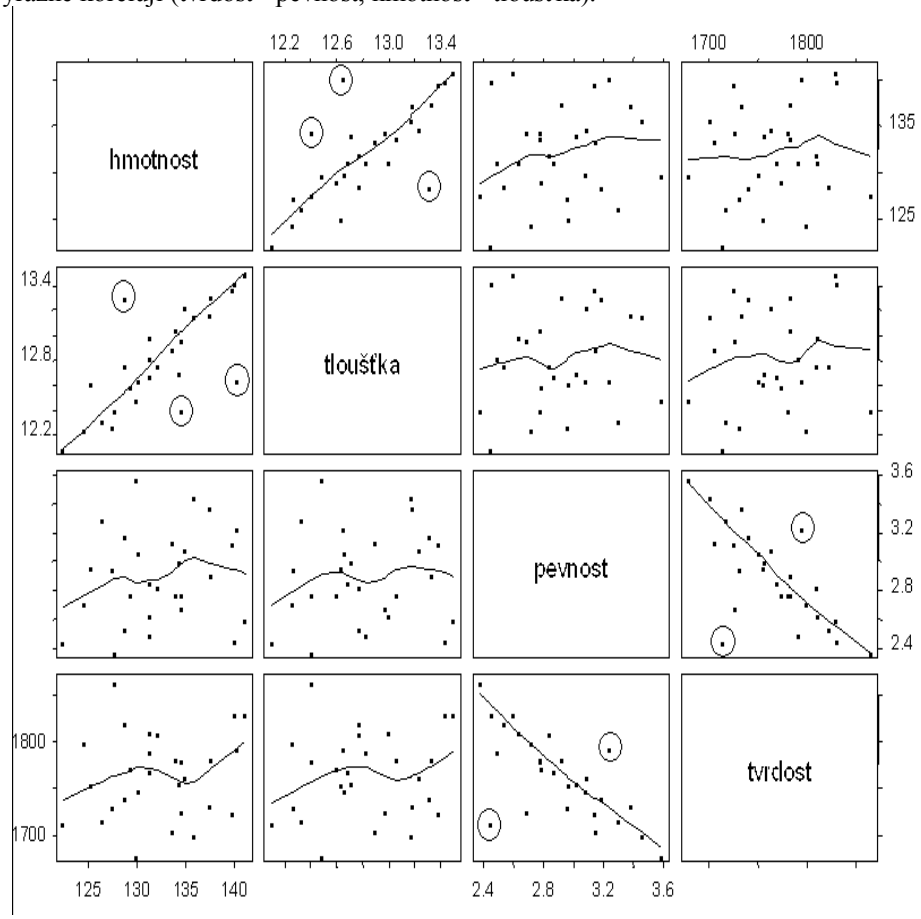
Obr. 10.30 Podstata Hotellingovy statistiky pro korelovaná data.

Předpokládáme-li, že data představují výběr z dvourozměrného normálního rozdělení, pak lze určit eliptickou oblast odpovídající 99.73 % intervalu spolehlivosti dat, tedy oblast, jejíž překročení má pravděpodobnost 0.27 %, což odpovídá kontrolním mezím regulačního diagramu. Bod **B** se nachází uvnitř této elipsy i uvnitř mezí *LCL* a *UCL*. Bod **A** je však daleko mimo přípustnou oblast dat, je tedy krajně nepravděpodobný a v Hotellingově diagramu vyvolá výrazné překročení horní meze. Kdybychom však bod **A** zaznamenali v obou Shewhartových diagramech, nacházel by se uvnitř mezí a tato výrazná porucha by byla ignorována. Monitorujte statistickou stabilitu fyzikálních parametrů tablet ve farmaceutickém provozu.

Data:

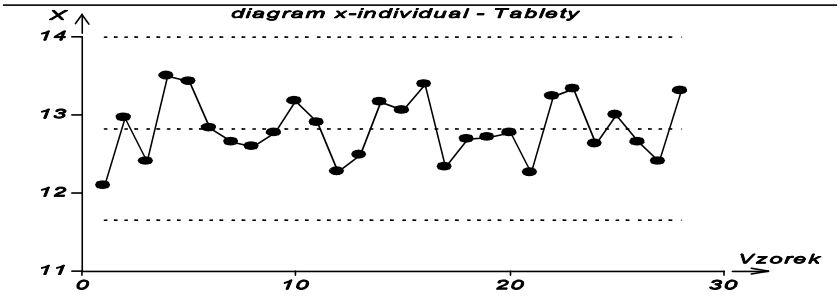
Hmotnost	Tloušťka	Pevnost	Tvrdost
122.4	12.10	2.45	1715
...	...	...	...
128.7	13.31	3.19	1741

Řešení: Tabulka představuje data z lisovny tablet, kde se měřila hmotnost, tloušťka, pevnost a tvrdost. Na obrázcích 10.32 až 10.35 jsou klasické Shewhartovy diagramy pro jednotlivé veličiny. Tyto diagramy nesignalizují žádnou podstatnou odchylku. Na obr. 10.36 je Hotellingův diagram pro všechny čtyři veličiny, který odhaluje výrazné překročení regulační meze na začátku a na konci směny. Obr. 10.31 ukazuje, že měřené veličiny spolu výrazně korelují (tvrdost - pevnost, hmotnost - tloušťka).

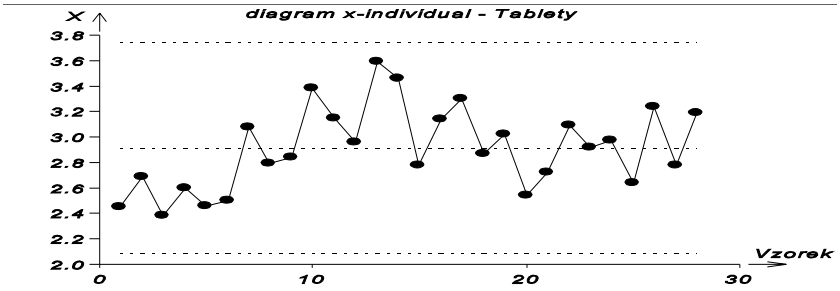


Obr. 10.31 Korelační struktura dat z lisovny (párové korelace).

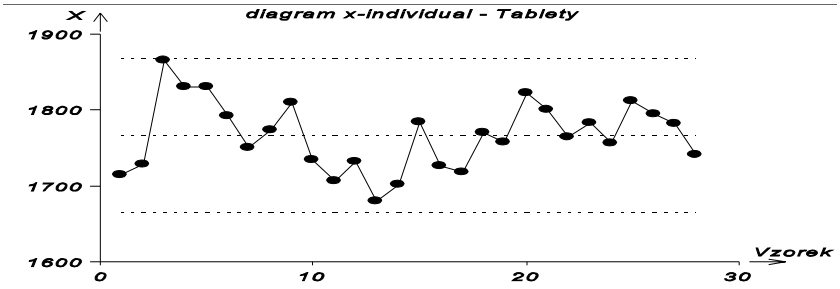
Na následujících čtyřech Shewhartových diagramech, obrázky 10.32 až 10.35, se neprojevily žádné problémy. V párových rozptylových grafech jsou však patrné body, které se vymykají převažujícímu trendu, na obr. 10.31 jsou označeny kroužkem. Tyto body jsou odhaleny pouze Hotellingovým diagramem - obr. 10.36.



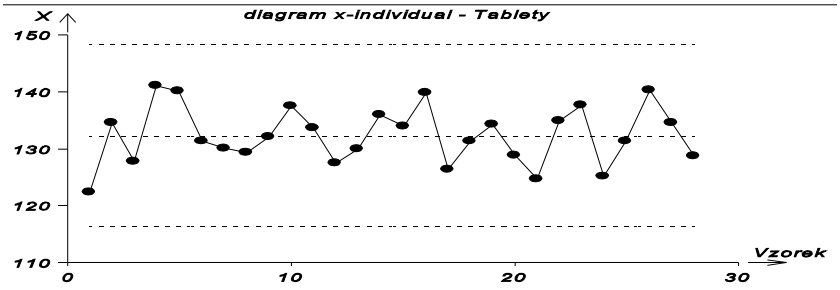
Obr. 10.32 Regulační diagram *"x*-individual" pro tloušťku.



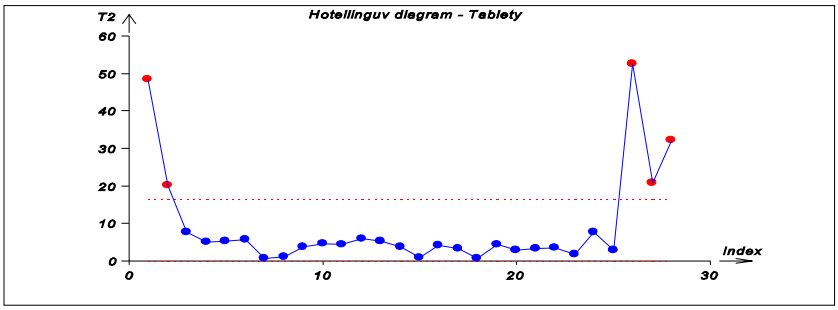
Obr. 10.33 Regulační diagram *"x*-individual" pro pevnost.



Obr. 10.34 Regulační diagram *"x*-individual" pro tvrdost.



Obr. 10.35 Regulační diagram *"x*-individual" pro hmotnost.



Obr. 10.36 Hotellingův diagram pro všechny parametry.

Závěr: Hotellingův diagram jednoznačně odhalil dva fatální body na začátku a tři fatální body na konci směny, které odpovídají zakroužkovaným bodům v obr. 10.31. Shewhartovy diagramy zcela selhaly a žádné fatální body nenalezly.

10.3 Indexy způsobilosti procesu

Je přirozenou snahou řady výrobců charakterizovat nějakým způsobem kvalitu výrobního procesu s ohledem na rezervy při zvyšování jakosti výrobků. Obecně zahrnuje analýza způsobilosti procesů celou řadu fází:

1. Definici výrobního procesu jako systému (stanovení všech vazeb vývojového diagramu).
2. Specifikaci *vstupních proměnných, procesních proměnných* a charakteristik produktu a způsobu jejich měření.
3. Výběr znaků jakosti, *které* se budou sledovat, *kde* se budou sledovat a *kdy*.
4. Posouzení stavu výrobního procesu na základě indexů způsobilosti pro vybrané znaky jakosti.

Pro tento účel byla navržena v posledních 20 letech celá řada různých (snadno určitelných, ale hůře interpretovatelných) indexů. Nejznámější z nich je tzv. *index způsobilosti procesu* (proces capability index, *PCI*) označovaný symbolem C_P . V této kapitole je ukázán statistický charakter tohoto a souvisejících indexů způsobilosti.

Index C_P: Předpokládejme, že známe měřitelný spojitý znak x, který charakterizuje kvalitu výrobního procesu (např. nestejnoměrnost příze, rozměry strojní součásti, elektrický odpor elektronické součástky atp.). Na základě požadované funkčnosti výrobku nebo polotovaru můžeme obyčejně určit dolní *LSL* a horní *USL* meze specifikace jakosti. Pokud padne hodnota x do těchto mezí, je výrobek považován za jakostní, a pokud padne mimo tyto meze, jde o zmetek (nevyhovující výrobek, *NC*). Mírou toho, nakolik je výrobní proces schopen zajistit splnění požadavků $LSL < x < USL$, je index způsobilosti C_P. Je zřejmé, že čím bude C_P vyšší, tím bude proces výroby kvalitnější. Pro dobré výrobní procesy by mělo platit, že $C_P > 1$, tj. statistická variabilita procesu výroby je dostatečně malá ve srovnání s tolerančním intervalem $[USL, LSL]$. Pro případ, že znak x má normální rozdělení $N(d, \sigma^2)$, kde

$$d = (LSL + USL)/2,$$

je zřejmě očekávaný podíl zmetků, tj. hodnot mimo tolerance, roven $P_C = 2F_N(-m/\sigma)$, kde $F_N(.)$ označuje distribuční funkci normovaného normálního rozdělení a $m = (USL - LSL)/2$. Po dosazení vyjde

$$P_C = 2F_N(-3C_P).$$

Pro případ, že index způsobilosti $C_P = 1$, je z tabulek možno nalézt $F_N(-3) = 0.00135$. Pak $P_C = 0.0027$ a tedy maximálně 0.27 % výrobků budou tvořit zmetky. Při znalosti P_C lze za předpokladu normality definovat C_P z rovnice ve tvaru[12]

$$C_P = 0.33\, F_N^{-1}((1 + P_C)/2,$$

kde $F_N^{-1}(.)$ je kvantilová funkce normovaného normálního rozdělení. Na základě výběrového průměru x_S a rozptylu s^2 lze určit přirozený odhad P_C ve tvaru

$$P_C' = 1 - F_N\left(\frac{USL - x_S}{s}\right) + F_N\left(\frac{LSL - x_S}{s}\right).$$

Tento odhad je však pro reálné výběry vychýlený. Poměrně komplikovaný nevychýlený odhad je publikován v práci[11]. Pokud se směrodatná odchylka σ nahrazuje klasickou výběrovou směrodatnou odchylkou s, je odhad C^*_P náhodnou veličinou. Rozptyl tohoto odhadu je přibližně roven

$$\mathrm{var}(C_P^*) \doteq C_P^2/(2(N - 3)),$$

kde N je velikost výběru, ze kterého byla počítána směrodatná odchylka s. Odpovídající 100(1-ν) %ní interval spolehlivosti pro populační hodnotu C_P má tvar

$$\frac{\chi^2_{\nu/2}\,(N - 1)\,C_P'}{\sqrt{N - 1}} \le C_P \le \frac{\chi^2_{1-\nu/2}\,(N - 1)\,C_P'}{\sqrt{N - 1}},$$

kde $\chi_\nu^2(N - 1)$ je 100ν %ní kvantil χ^2-kvadrát rozdělení s $N - 1$ stupni volnosti. Pro malé výběry je C^*_P vychýlený odhad parametru C_P. Nevychýlený odhad C^N_P má pro $N > 15$ tvar

$$C_P^N = C_P'(1 - 0.75/(N - 1))$$

a jeho rozptyl je roven $\mathrm{var}(C_P^N) = \dfrac{8(N - 1) + 9}{16(N - 1)(N - 3)} C_P^2$.

Další indexy způsobilosti procesů: Dosavadní úvahy byly založeny na platnosti vztahu $d = (LSL + USL)/2$. Obecně však střední hodnota d_T znaku jakosti x není rovna d určenému z tohoto vztahu. Modifikovaný index C_{PK} má pak tvar[13]

$$C_{PK} = \min[(USL - d_T)/3\sigma,\ (d_T - LSL)/3\sigma],$$

resp. po úpravě $C_{PK} = C_P(1 - *d_T - d*/m)$. Je patrné, že C_{PK} je vždy menší nebo rovno C_P. Očekávaný podíl zmetků P_C je pak roven

$$P_C = F_N\left(\frac{LSL - d_T}{\sigma}\right) + F_N\left[1 - \left(\frac{USL - d_T}{\sigma}\right)\right].$$

Odhad C^*_{PK} lze určit snadno při znalosti výběrového průměru x_S a rozptylu s^2

$$C_{PK}^* = (m - *x_S - d*)/3s.$$

Tento odhad je opět pro malé rozsahy výběru N vychýlený. Přibližný 100(1-ν) %ní interval spolehlivosti parametru C^*_{PK} je roven

$$\frac{(1 - u_{1-v/2})\, C_{PK}^{(}}{\sqrt{2N - 2}} \leq C_{PK} \leq \frac{(1 + u_{1-v/2})\, C_{PK}^{(}}{\sqrt{2N - 2}} .$$

Z uvedeného výkladu je patrné, že pokud lze přijmout předpoklad normality rozdělení znaku x, lze stanovit intervaly spolehlivosti pro C_P, resp. C_{PK}, podle uvedených vztahů. Komplikace činí především nutnost odhadu směrodatné odchylky a průměru z výběrových hodnot. V souvislosti se ztrátovou funkcí používanou při Taguchiho přístupu k hodnocení jakosti byl navržen index[14]

$$C_{Pm} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (d_T - T)^2}} ,$$

kde T je předepsaná (cílová hodnota) parametru x. Obyčejně platí, že $T = d$, tj. jde o střed tolerančního intervalu.

Ztrátová funkce $L(x)$ je standardně uvažována ve tvaru

$$L(x) = K (x - T)^2 .$$

Výraz $\sigma^2 + (d_T - T)^2$ je pak *mírou průměrné ztráty,* způsobené nedodržením podmínek kvalitní výroby ($\sigma^2 = 0$, $T = d_T$). Je zřejmé, že $C_{Pm} \neq C_P$. Lze odvodit, že

$$C_{Pm} = C_P \sqrt{1 + Z^2} , \quad \text{kde } Z = (d_T - T)/\sigma.$$

Při použití C_{Pm} je třeba brát v úvahu to, že je vhodný pro vyjádření způsobilosti procesů jen, pokud je cílová hodnota T rovna středu tolerančního intervalu d. Nevhodnost použití C_{Pm} pro obecné situace je patrná z tohoto jednoduchého příkladu. Mějme tři procesy s následujícími charakteristikami

A $d_T = 50.00$ $\sigma = 5$,
B $d_T = 57.50$ $\sigma = 2.5$,
C $d_T = 61.25$ $\sigma = 1.25$.

Nechť $T = 50$, $USL = 65$ a $LSL = 35$. Pro tyto hodnoty jsou C_P, C_{PK} a C_{Pm} rovny

Proces	C_P	C_{PK}	C_{Pm}
A	1	1	1
B	2	1	0.64
C	4	1	0.44

Je patrné, že každý z uvedených indexů ukazuje zcela něco jiného. Základní rozdíl mezi C_{Pm} a C_{PK} spočívá v relativní významnosti mezí specifikace *(LSL, USL)* a cílové hodnoty T.

Funkcí indexu C_{PK} je indikace, do jaké míry je proces uvnitř mezí specifikace. Pokud $LSL < d_T < USL$, je pro $\sigma \to 0$ hodnota $C_{PK} \to \infty$. Vysoké hodnoty C_{PK} však neukazují na rozdíly mezi T a d_T.

Index C_{Pm} ukazuje na míru, nakolik je proces blízký cílové hodnotě T. Pro případ, že $\sigma \to 0$ a $d_T \to T$ je pak $C_{Pm} \to \infty$. Zde meze specifikace slouží jen jako škála pro ztrátovou funkci. Lze ukázat, že pokud T není rovno d, vede posun průměrné hodnoty znaku x směrem k cílové hodnotě T (zvyšující C_{PK} a C_{Pm}) ke zvýšení podílu zmetků, počtu výrobků mimo meze specifikace.

Je patrné, že C_{PK} vzniklo modifikací čitatele a C_{Pm} modifikací jmenovatele vztahu pro C_P. Kombinací těchto dvou modifikací lze dospět k *indexu* C_{PmK} definovanému výrazem[15]

$$C_{PmK} = \frac{m - |x_S - d|}{3\sqrt{\sigma^2 + (d_T - T)^2}} .$$

Je zřejmé, že $C_{PmK} = C_{PK} C_{Pm}/CP$ a pro případ, kdy $d_T = T = d$, jsou tyto indexy stejné. *Přirozený odhad indexu* C_{PmK} má tvar

$$\hat{C}_{PmK} = \frac{m - |x_S - d|}{3\sqrt{\frac{\sum_{i=1}^{N} (x_i - T)^2}{N}}} .$$

Všechny dosavadní úvahy jsou založeny na předpokladu normality. Pokud je rozdělení znaku x silně nenormální, jsou odhady indexů způsobilosti vychýlené a mají nesprávný rozptyl. Pokud lze pro rozdělení znaku x specifikovat velikost šikmosti g_1 a špičatosti g_2 , lze použít tabulek (viz cit.[11]) pro nahrazení čísla 6 ve jmenovali vztahu pro C_P takovým číslem, pro které bude $P_C = 0.0027$. Pokud se místo $P_C = 0.0027$ použije vyšší hodnota $P_C = 0.01$ (tj. připouští se max. 1 % zmetků), je ve jmenovateli vztahu pro C_P nutno použít číslo 5.15. Zajímavé je, že tato hodnota vyhovuje také pro celou řadu dalších rozdělení ($0 < g_1 < 9$ a $1 < g_2 < 6$). Z uvedeného textu vidíme, že při použití indexů způsobilosti procesů je nezbytné uvažovat také jejich meze a jejich statistickou povahu. Jen tak lze omezit jejich nevhodné aplikace. Pro případ více znaků, charakterizujících jakost procesu lze použít zobecněných parametrů, založených např. na Hotellingově T^2 statistice[11].

10.4 Software pro řízení jakosti

Jak je patrné, není vlastní použití regulačních diagramů a indexů způsobilosti procesu nikterak náročné. Velká většina metod byla konstruována tak, že nevyžaduje žádné složitější výpočty a tedy ani počítač. S tím, jak se i do provozů dostává výpočetní technika, sběr dat se provádí automaticky a řada strojů se jak monitoruje, tak i řídí počítačem, rostou i požadavky na vhodný software. Bohužel velká většina programů nevyužívá nikterak speciálně možností výpočetní techniky, takže pouze nahrazuje rutinní práci praktiků. Tento stav je často považován za vyhovující. Potíže však nastávají, jakmile nejsou splněny předpoklady, za kterých byly tyto metody odvozeny (normalita, nepřítomnost hrubých chyb, dostatečná velikost výběru, statistická stabilita procesu atd.). Lze říci, že řada speciálních programů pro řízení jakosti je orientována pouze na konstrukci regulačních diagramů a různých indexů, aniž ověřuje kvalitu dat. Z tohoto hlediska jsou výhodnější některé statistické programy nebo jejich kombinace, které umožňují komplexní statistickou analýzu dat a navíc konstrukci diagramů resp. dalších pomůcek pro řízení jakosti. Přehled současného softwarového vybavení pro oblast jakosti je uveden např. v cit.[17].

Pro komplexní řešení úloh souvisejících s jakostí a jejím řízením je nutná celá řada statistických metod. Mezi základní patří:

(1) Analýza jednorozměrných výběrů.

(2) Porovnání shody výběrového rozdělení se zvoleným teoretickým.

(3) Analýza rozptylu.

(4) Regresní analýza.

(5) Konstrukce pravděpodobnostních modelů.

Ze všech statistických programových systémů, které umožňují řešení uvedených úloh, jsou dále diskutovány pouze ty, které se jeví jako pro praxi vhodné a jsou na našem trhu běžně k dispozici.

Jak bylo ukázáno, souvisí problém řízení jakosti velmi úzce s tvorbou intervalů spolehlivosti vybraných charakteristik jakostních znaků. Tyto intervaly vycházejí ze znalosti dobrého odhadu směrodatné odchylky a vyžadují splnění základních předpokladů o datech, jako jsou nezávislost, nepřítomnost vybočujících měření, normalita. Vhodný statistický software by měl tedy umožnit ověření těchto předpokladů nesložitými "robustními" technikami, které jsou pokud možno distribučně nezávislé.

Je zřejmé, že vlivem kolísání vlastností surovin, podmínek zpracování, stavu okolí atd., mají průmyslové procesy náhodný charakter. Variabilita procesů je jedním ze základních ukazatelů jeho stavu a její odhad se používá jak při konstrukci regulačních diagramů, tak i při výpočtech indexů způsobilosti. Základní příčiny variability průmyslových procesů lze rozdělit do těchto skupin:

(a) *Náhodné šumy,* které se vyskytují i za podmínek, že je proces v optimálním (standardním) stavu. Jsou způsobeny nekontrolovanými příčinami a lze je obecně snížit pouze změnou procesu (strojního zařízení).

(b) *Externí zdroje variability,* způsobené změnami podmínek okolí.

(c) *Procesní zdroje variability,* způsobené nekonstantností procesních parametrů.

(d) *Přiřaditelné zdroje variability,* způsobené např. změnou jakosti suroviny, špatným seřízením strojů, opotřebením pracovních orgánů, stárnutím materiálů a různými skokovými změnami externích, resp. procesních proměnných ovlivňujících stav procesu.

Techniky řízení jakosti se snaží postihnout právě *přiřaditelné zdroje variability,* které lze provedením "regulačního" zásahu eliminovat a převést zpět do stavu, kdy proces ovlivňuje jen náhodné šumy. Standardně se při posuzování variability procesů vychází z (nezávislých) výběrů $V_1, \ldots, V_i, \ldots, V_M$, z nichž má každý velikost N a je charakte-rizován výběrovým průměrem x_{Si} a rozptylem s_j^2. Symbol x_{ij} označuje j-tý prvek v i-tém výběru V_i. Je účelné sestavit typické modely ukazující zdroje variability:

(a) *Nejjednodušší model* je model náhodného šumu, pro který platí

$$x_{ij} = d + \sigma_C \, g_{ij}, \; i = 1, \ldots, M, j = 1, \ldots, N,$$

kde d je celkový průměr, σ_C^2 je rozptyl (uvnitř výběrů) a g_{ij} jsou nezávislé náhodné veličiny s normovaným normálním rozdělením $N(0, 1)$. Odhadem d je d^* a odhadem σ_C je σ^*. Tento model je tedy standardně používán při konstrukci regulačních diagramů.

(b) Poněkud *komplexnější model* uvažuje, že kromě šumů se projeví i další zdroje variability, ovlivňující každý výběr jako celek (procesní, resp. externí zdroje variability). Pro tento model platí

$$x_{ij} = d + \sigma_B \, \omega_i + \sigma_C \, g_{ij}, \; i = 1, \ldots, M, j = 1, \ldots, N,$$

kde σ_B^2 je rozptyl mezi výběry a ω_i jsou nezávislé náhodné veličiny s normovaným normálním rozdělením $N(0, 1)$. Pro tento model platí, že rozptyl měření je roven

$$D(x_{ij}) = \sigma_B^2 + \sigma_C^2$$

a rozptyl výběrových průměrů je roven

$$D(x_{Si}) = \sigma_B^2 + \sigma_C^2/N.$$

Je patrné, že zvětšováním počtu prvků ve výběru N se sníží pouze část variability (uvnitř výběrů), ale neovlivní rozptyl mezi výběry. Odhad s_B^2 rozptylu mezi skupinami

σ_B^2 se provádí obyčejně podle vztahu pro klasický výběrový rozptyl pro jednotlivé průměry

$$s_B^2 = \frac{1}{M-1} \sum_{i=1}^{M} (x_{Si} - d)^2 \quad .$$

Tento odhad je však vychýlený, protože obsahuje příspěvek z variability uvnitř výběrů. Platí, že

$$E(s_B) \approx \sqrt{\sigma_B^2 + \sigma_C^2 / N} \quad .$$

Často je výhodnější použít odhad $\hat{\sigma}_B^2$, kde se odečte příspěvek rozptylu uvnitř skupin, tedy

$$\sigma_B^2 = s_B^2 - \sigma^2/N \quad .$$

Pokud vyjde rozdíl na pravé straně této rovnice roven nule, dosadí se definitoricky $\hat{\sigma}_B^2 = 0$. Při konstrukci klasických Shewhartových regulačních diagramů se obyčejně předpokládá, že $\hat{\sigma}_B^2 = 0$. To lze snadno ověřit na základě podílu

$$V = s_B^2 N / \sigma^{2*},$$

který má v případě platnosti nulové hypotézy H_0: $\sigma_B^2 = 0$ *F*-rozdělení s (*M* - 1) a *M(N - 1)* stupni volnosti. Pomocí *F*-testu lze tedy ověřit platnost hypotézy H_0. Pro případ konstrukce diagramů CUSUM je třeba odhadnout směrodatnou odchylku skupinových průměrů σ_e. Pokud lze použít $\sigma_B = 0$, je $\hat{\sigma}_e = \sigma/\sqrt{N}$, ale pokud je $\sigma_B > 0$, je nutné uvažovat, že $\hat{\sigma}_e = s_B$. Posouzení velikosti σ_B^2 je tedy pro řadu situací zcela *nezbytné*. Dalším zdrojem variability může být chyba měření, vyjádřená rozptylem σ_M^2, která se obyčejně slučuje na rozptyl σ_C^2. Celkový rozptyl je pak jejich součtem a odpovídající směrodatná odchylka je zvětšena o faktor $f = \sqrt{1 + \sigma_M^2/\sigma_C^2}$. Pokud bude platit, že $\sigma_M \approx \sigma_C/3$, neprojeví se chyby měření výrazně na úrovni rozptylu, vyjde maximální hodnota $f = \sqrt{(1 + 1/9)} = 1.054$.

(c) Posledním *frekventovaným modelem* je případ autokorelace prvního řádu. To vlastně znamená porušení předpokladu nezávislosti jednotlivých výběrů. To je typický model situace, kdy data tvoří časovou řadu, nebo když se některé procesní parametry resp. externí parametry mění zvolna. Model autokorelace prvního řádu má tvar

$$x_{ij} = d + W_i + \sigma_C \varepsilon_{ij}, \quad \text{kde} \quad W_i = \rho W_{i-1} + \sigma_B \omega_i \quad .$$

Platí, že $W_0 = 0$. Autokorelačním koeficientem je korelační koeficient mezi dvojicemi x_{Si} a x_{Si+1}, $i = 1, ..., M - 1$. Z této rovnice lze snadno určit, že

$$D(W) = \sigma_B^2/(1 - \rho^2) \quad .$$

Využitím výsledků pro předchozí modely lze pak vyjádřit rozptyl měření

$$D(x_{ij}) = \frac{\sigma_B^2}{1 - \rho^2} + \sigma_C^2$$

a rozptyl výběrových průměrů $D(x_{Si}) = \frac{\sigma_B^2}{1 - \rho^2} + \sigma_C^2 / N$. Je patrné, že v případě výrazné autokorelace dojde ke zvýšení rozptylu výběrových průměrů, což by mělo za následek zvětšení regulačních mezí a ztrátu citlivosti regulačních diagramů vůči změnám stavu procesu. Autokorelační koeficient ρ se obyčejně odhaduje pomocí vztahu

$$\hat{\rho} = \sum_{i=1}^{M-1} \frac{(x_{Si} - d^{(})(x_{Si+1} - d^{(})}{s_B^2 (M - 1)} .$$

Orientačně platí, že pokud je $-2/\sqrt{M} \leq \hat{\rho} < 2/\sqrt{M}$, lze považovat ρ za nevýznamné. Přesnější testy významnosti jsou uvedeny v cit.[16]. Pro posouzení autokorelace lze použít také modifikovaný von Neumanův poměr

$$V^2 = \frac{\sum_{i=1}^{M-1} (x_{Si} - x_{Si+1})^2}{2 \sum_{i=1}^{M} (x_{Si} - d^{(})^2} .$$

Lze ukázat, že pro velká M má veličina V^2 přibližně normální rozdělení s parametry

$$E(V^2) = 1 , \quad D(V^2) = \frac{M - 2}{M^2 - 1} . \frac{1}{M + 2} .$$

Orientačně platí, že pokud leží V^2 mimo interval $1 \pm 2/\sqrt{(M + 2)}$, nejsou výběry nezávislé Pokud vyjde V^2 menší než spodní mez tohoto intervalu, projevuje se v datech trend nebo pomalé cyklické změny. Pokud vyjde V^2 větší než horní mez tohoto intervalu, projevují se v datech rychlé cyklické změny.

Z uvedeného výkladu je patrné, že při určování směrodatné odchylky výběrových průměrů je třeba postupovat opatrně a nejdříve testovat autokorelaci, resp. významnost rozptylu σ_B^2.

10.5 Úlohy

Konstrukce a použití regulačních diagramů je nejznámější statistickou technikou řízení jakosti. Přes velké rozšíření regulačních diagramů se často setkáváme s jejich nesprávným použitím, způsobeným nerespektováním zásad, anebo zanedbáním základních předpokladů o datech, jako je normalita, nezávislost, stabilita. Důsledkem je pak nesprávná interpretace diagramů, nedůvěra ke statistickým metodám, upouštění od statistické regulace, někdy izkreslování výsledků a nedovolená manipulace s daty. V následujících úlohách ukážeme situace, které při používání regulačních diagramů vedou k problémům. V řešení úloh naznačíme postupy formou kontrolních mezivýsledků, které mohou tyto potíže zmírnit nebo vyřešit. Většina úloh je z oblastí, v nichž se problémy nejčastěji vyskytují, což jsou především odvětví chemie, metalurgie, polymerů a plastů, farmacie a klinické medicíny, potravinářství, mechanické zkušebny, monitoring životního prostředí, geologie.

Úloha C10.1 *Kontrola tavby v metalurgickém provoze regulačním diagramem*
Při kontrole tavby v metalurgickém provoze regulačním diagramem byly z každé tavby odebrány 3 vzorky, na nichž byla měřena pevnost. Z těchto trojic se má sestrojit Shewhartův regulační diagram "x s pruhem". Vysvětlete jeho (ne-)použitelnost a navrhněte druh jiného regulačního diagramu, resp. transformaci dat. Ověřte předpoklady o normalitě.
Data: pevnost:

Datum	Vzorek1	Vzorek2	Vzorek3	Datum	Vzorek1	Vzorek2	Vzorek3
1.7.1998	388.2	387.3	387.4	1.8.1998	398.3	397.9	398.1
...	...	...	...	...	...	...	...
28.7.1998	393.6	392.8	393.3	28.8.1998	394.5	394.9	395.7
31.7.1998	391.3	391.1	390.9				

Úloha C10.2 *Chemická analýza složení plynných zplodin*
Chemická analýza složení plynných zplodin má být monitorována pomocí regulačního diagramu s cílem detekovat odchylky od statistické stability. Komentujte použitelnost regulačních diagramů.
Data: Obsah CO [mg]:

214.3	245	256.2	246.2
...	...	...	...
249.5	266.9	242	

Úloha C10.3 *Sledování rozměrů polyuretanového výlisku automobilní součástky*
Použijte sloupce *A, B, C, D* pro konstrukci Hotellingova diagramu. Pomocí tohoto diagramu posuzujte pak i data A_1 , B_1 , C_1 , D_1 z provozu. Jednotlivé sloupce představují šířku, délku, výšku a rozteč otvorů výlisku z měkčeného polyuretanu.
Data: šířka, délka, výška a rozteč otvorů výlisku z měkčeného polyuretanu [mm]:

Čas	A	B	C	D	A_1	B_1	C_1	D_1
1.3.97 8:00	13.14	26.62	18.66	9.83	15.54	28.65	18.66	11.06
1.3.97 9:00	14.38	26.54	19.03	9.76	15.39	27.89	18.26	10.67
...	...	...	...	...	...	...	...	...
2.3.97 10:00	13.88	27.24	17.77	9.05	15.87	26.92	18.39	9.64

Úloha C10.4 *Sledování obsahu CaO (%) v surovině pro výrobu cementu*
Data představují obsah CaO (%) v meziproduktu při výrobě cementu. Na základě dat rozhodněte o vhodnosti Shewhartova diagramu "*x*-individual" pro tento proces, případně navrhněte jinou techniku.

Data: Obsah CaO [%] v surovině pro výrobu cementu:

42.53	44.08	42.55
42.49	44.4	42.42
...	...	...
43.34	42.96	44.29

Úloha C10.5 *Sledování počtu zákrutů na metr délky textilie*
V textilním podniku jsou vyráběna textilní vlákna a jako indikátor kvality je měřen počet zákrutů na jeden metr délky. Úkolem je vytvořit regulační diagramy a zjistit kvalitu těchto vláken. Vykazují data symetrické, normální rozdělení? Sestrojte regulační diagram typu "*x* s pruhem" a *s*. Jak se určí základní linie a regulační meze v diagramu *s*? Diagram *R* pro rozpětí lze použít jako alternativu diagramu *s*. Rozpětí podskupiny je rozdíl největší a nejmenší hodnoty v podskupině. Jak vypočtete základní linii a regulační meze diagramu *R*? Hodnoty D_3 a D_4 jsou tabelovány; pro $N = 5$ vycházejí $D_3 = 0$, $D_4 = 2.114$. Specifikační meze jsou stanoveny takto: *LSL*=600; *USL*=900; *T*=750. K čemu zde slouží indexy způsobilosti procesu C_p a C_{pk}? Určují do jaké míry se daří dodržovat předepsané specifikační meze? Co popisuje toto jediné bezrozměrné číslo? K jakým závěrům lze dospět na základě regulačních diagramů a indexu způsobilosti procesu?

Data: 42 vzorků po pěti měřeních (počet zákrutů na jeden metr délky):

Vzorek	x_1	x_2	x_3	x_4	x_5
1	824	708	834	800	794
...	...	...	...	...	...
42	770	780	842	760	756

Úloha C10.6 *Sledování vlhkosti sklářského kmene*

Technologický proces výroby skla je založen na tavení sklářského kmene za vysokých teplot. Sklářský kmen představuje směs vstupních surovin o vhodné zrnitosti, které jsou míseny ve vhodném molárním či hmotnostním poměru s ohledem na vlastnosti připravovaného skla. Vlhkost sklářského kmene je parametrem, který ovlivňuje technologický proces tavení po stránce energetické i kvalitu utavené skloviny a její vlastnosti. Po 18 dní bylo v pravidelných intervalech odebíráno 6 vzorků sklářského kmene před vstupem do tavícího agregátu a vlhkost stanovována gravimetricky jako hmotnostní úbytek po vysušení kmene při 110 ^{0}C do konstantní hmotnosti. Získané úbytky vyjádřené v % hmotn. přináší datový výběr: 1. Konstruujte regulační diagram pro *"x*-individual" a hodnoty rozpětí (x_i - *MR*). 2. Pro šestice dat sestrojte regulační diagram "*x* s pruhem"-rozpětí ($\bar{x}$ - *R*). Na základě výpočtu indexů způsobilosti diskutujte stabilitu procesu přípravy sklářského kmene z hlediska obsahu vody.

Data: Vlhkost (v % hmont.) šesti vzorků sklářského kmene gravimetrickou metodou:

i	x_1	x_2	x_3	x_4	x_5	x_6
1	3.1	3.2	3.1	2.8	3.2	3.5

...	...	...	...	...	...	...
18	3.6	3.7	3.4	3.5	2.8	2.4

Úloha C10.7 *Velikost částic frakce hrubě mletého žulového štěrku*

Navrhněte regulační diagram pro individuální hodnoty pro velikost částic frakce hrubě mletého stavebního materiálu (hmotnost v gramech). Je vhodný klasický Shewhartův diagram? Jaké je rozdělení hmotností? Je v datech trend nebo jiná závislost?

Data: Hmotnost částic [g] frakce hrubě mletého stavebního materiálu:

0.4	0.32	0.22	0.2	0.39	1.02	0.43	0.23
...	...	...	...	...	...	...	...
0.6	0.3	0.21	0.6	0.49	0.26	0.58	

Úloha C10.8 *Kontrola hmotnosti tablet léčiva*

Navrhněte Shewhartův regulační diagram pro individuální hodnoty podle ISO 8258 pro hmotnosti tablet uvedené v mg. Vyjadřují vypočtené regulační meze očekávanou variabilitu dat? Mají data normální rozdělení? Co je příčinou příliš širokých regulačních mezí?

Data: Hmotnost tablet [mg]:

211	212	209	208
...	...	...	...
209	207	214	206

Úloha C10.9 *Teplota trysky pro produkci vláken PES*

Ověřte možnost použití regulačního diagramu CUSUM ke sledování hodnot teploty trysky (stupně Celsia, EC) při produkci vláken PES. Cílová hodnota je 173. Během produkce je třeba rychle diagnostikovat odchylku od cílové hodnoty. Srovnejte efektivitu diagramů Shewhartových a CUSUM. Směrodatnou odchylku odhadněte z dat.

Data: Sledování hodnot teploty trysky (EC) pro produkci vláken PES:

173.7	172.9	172.9	171.2
...	...	...	...
173.6	171.5	172.5	

Úloha C10.10 *Obsah CaO ve slinku*

Při sledování obsahu CaO ve slinku v cementárně byly každou hodinu odebrány z kontinuálního procesu 3 vzorky (Vzorek 1, 2 a 3) a stanoven obsah CaO. Z těchto trojic analýz zkonstruujte Shewhartův regulační diagram "x s pruhem" a diagram s, s vypočítanými mezemi o velikosti podskupiny 3 podle ISO 8258. Vysvětlete selhání regulačního diagramu a navrhněte řešení.

Data: Obsah CaO ve slinku [%]:

Vzorek1	Vzorek2	Vzorek3	Vzorek1	Vzorek2	Vzorek3
42.53	42.89	43.77	43.96	44.12	44.74
...	...	...	...	...	...
43.72	43.43	43.02	44.29	44.4	44.63

Úloha C10.11 *Sledování hygienického stavu mlékárenského stroje bioluminiscenční technikou*. Bioluminiscenční technikou se sleduje adenosintrifosfát ATP tak, že se měří počet emitovaných světelných jednotek *RLU*, které jsou testačním kritériem sledované hygienické normy: čím více jednotek *RLU*, tím více ATP a tím horší hygiena. Po 92 dny bylo vždy před zahájením směny mlékárenské plničky provedeno jedno měření *RLU*. V regulačním diagramu CUSUM nebo "x-individual" sledujte dny, kdy byla hygienická norma porušena. U obou diagramů je třeba nejprve prověřit předpoklad normality. Při jeho nesplnění je vhodné použít symetrizační transformaci dat (např. Boxovu-Coxovu) a tím rovněž redukovat náhodný šum v datech. U diagramu CUSUM je vhodné použít V-masku. Vykazují data nějaký trend?

Data: i je pořadové číslo dne, *RLU* je počet relativních jednotek světla: 1 23, 2 46, ... , 92 67.

Úloha C10.12 *Obsah Si v surovině*

Ve sklářské surovině byl stanovován obsah křemíku c v procentech v celkem 75 po sobě následujících vzorcích ke zjištění, zda z hlediska technologie naměřené obsahy vyhovují normě. Ověřte normalitu a přítomnost vybočujících měření. Navrhněte druh regulačního diagramu ke sledování statistické stability tohoto parametru. Jak a proč selhává Shewhartův diagram pro jednotlivé hodnoty "x-individual"? Ověřte možnost použití diagramu EWMA s predikcí (dynamický diagram).

Data: Obsah křemíku c [%]: 34.51, 34.52, ..., 33.95.

Úloha C10.13 *Hmotnost při plnění*

Při plnění dvou druhů pudinkového prášku do sáčku je průběžně kontrolována čistá hmotnost m v gramech. Odstraňte vybočující měření a vypočítejte kontrolní meze a sestrojte Shewhartův regulační diagram pro oba druhy prášku. Vypočítejte index způsobilosti při jednostranné specifikační mezi $LSL = 40$ g. Jsou obě linky nastaveny na stejnou střední hodnotu? Mají obě linky stejný rozptyl?

Data: Čistá hmotnost pudinkového prášku v sáčku *m* [g]:

Čokoláda				Jahoda			
40.69	40.51	40.92	41.73	40.14	40.9	40.64	40.06
...	...	...	...	...	...	...	...
41.31	40.09	41.56	40.33	41.06	40.42	40.24	40.55

Úloha C10.14 *Sledování finačních nákladů při výrobě litiny o různém obsahu C-Si*

Analýzou řady proměnných se ukázalo, že 4 proměnné významně monitorují finační náklady výroby litiny v hutním závodě: x_1 počet odlitků za hodinu, s jehož počtem náklady rostou; x_2 hodinové náklady na kvalitu litiny u řízeného výrobního procesu, x_3 hodinové náklady na kvalitu litiny u neřízeného výrobního procesu, x_4 indikovatelné odchýlení neřízené výroby, vystižené násobkem směrodatné odchylky, je v negativní korelaci s náklady - je dražší indikovat malé změny v kvalitě. Uvedené čtyři proměnné představují 90 % celkové proměnlivosti finančních nákladů. V regulačním diagramu CUSUM sledujte změny finačních nákladů. Užijte také Hottelingova regulačního diagramu více proměnných.

Data: x_1 počet odlitků za hodinu, x_2 hodinové náklady na kvalitu litiny u řízeného výrobního procesu, x_3 hodinové náklady na kvalitu litiny u neřízeného výrobního procesu, x_4 indikovatelné odchýlení neřízené výroby,

i	$x_{1,i}$	$x_{2,i}$	$x_{3,i}$	$x_{4,i}$
1	0.04	80	1250	0.25
2	0.02	150	1250	0.25
...	...	...	...	...
65	0.04	150	650	1.25

Úloha C10.15 *Sledování faktorů, ovlivňujících obsah sušiny v másle*

Byl stanoven obsah sušiny v máslové smetaně z náhodně vybraných odběrů 400 litrového zásobníku. Při obsahu vyšším než dolní hranice 45 % byla norma splněna a máselná hmota se transponovala do plnicího a balicího stroje. Jelikož jediné měření se ukázalo jako nedostatečné, byl vyvinut způsob, který bere v úvahu vliv proměn-livosti α_i při r sloučeninách, na jejichž základě se obsah sušiny x_{ijk} stanovuje, dále vliv proměnlivosti β_j z m odběrů vzorku a vliv proměnlivosti $\beta_{(ij)}$ z n měření a konečně γ_{ijk} je náhodná chyba dle modelu

$$x_{ijk} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk}, \quad i = 1,..., r, \quad j = 1,..., m, \quad k = 1,..., n,$$

kde x_{ijk} představuje k-té měření sušiny v j-tém vzorku pro i-tou sloučeninu v máslové smetaně. Zvolte vhodný typ regulačního diagramu, ověřte předpoklad normality rozdělení. Je nutná transformace dat? Existuje v datech nějaký trend?

Data: obsah sušiny v máslové smetaně, vyčíslený z 20 vybraných sloučenin pro 3 paralení měření u 5 náhodně vybraných vzorků:

i	j	x_1	x_2	x_3
1	1	47.66	47.33	47.76
1	2	47.43	47.95	47.9
1	3	48.03	48.32	47.92
1	4	47.68	47.89	47.51
1	5	48.64	48.27	48.56
2	1	46.54	47.04	46.84
2	2	46.81	47.22	47.2
2	3	46.83	46.22	46.32
2	4	46.19	46.35	46.43
2	5	46.98	47.12	47.21
...	...	...	...	...
20	1	48.84	48.56	48.61
20	2	48.05	48.32	48.51
20	3	47.86	48.32	48.28
20	4	47.65	47.89	47.96
20	5	48.21	48.29	48.02

Úloha C10.16 *Sledování tvrdosti materiálu ze dvou výrobních linek*

Byla sledována tvrdost materiálu, pocházejícího ze dvou výrobních linek. Oba materiály byly určeny ke spojení svárem a jejich tvrdost by měla být přibližně stejná. Zvolte vhodný typ regulačního diagramu, ověřte předpoklad normality rozdělení. Je nutná transformace dat? Existuje v datech nějaký trend? Je možné užít jeden společný regulační diagram pro obě výrobní linky?

Data: i index, x_1 tvrdost materiálu z první výrobní linky, x_2 tvrdost materiálu z první výrobní linky:

i	$x_{1,i}$	$x_{2,i}$
1	50.9	44.3
2	44.8	25.7
...	...	...
59	50.1	34.4

Úloha C10.17 *Sledování intenzity barvy chemického produktu ve výrobě*

Ve výrobě byla sledována intenzita barvy každé šarže chemického produktu, měřená hodnotou absorbance při dané vlnové délce. Zvolte vhodný typ regulačního diagramu a nalezněte číslo šarže výrobního produktu, která je mimo toleranční meze. Jsou splněny předpoklady kladené na výběr? Existuje v datech nějaký trend?

Data: i index šarže, hodnota absorbance chemického produktu A_i: 1 0.670, 2 0.630, ..., 102 0.602.

Úloha C10.18 *Sledování odporu elektrického izolátoru*

Byl sledován odpor elektrického izolátoru, který však může být ovlivněn i povětrnostními podmínkami, teplotou atd. Je třeba posoudit, jsou-li tyto vlivy statisticky významné a zda-li je třeba je brát vůbec do úvahy. Zvolte vhodný typ regulačního diagramu, ověřte předpoklady kladené na výběr. Odstraňte vybočující hodnoty a regulační diagram vyčíslete znovu. Mají odlehlé hodnoty významný vliv na průběh diagramu?

Data: i index izolátoru, měřený odpor R_i izolátoru [Mohm]: 1 5.045, 2 4.635, ..., 204 5.000.

Úloha C10.19 *Sledování průměru vnějšího kroužku kuličkového ložiska*

Byl sledován průměr vnějšího kroužku při výrobě kuličkového ložiska o hodnotě 26.60 mm. Celkově bylo proměřeno 12 skupin po 6 kroužcích, přičemž byly sledovány pouze odchylky od jmenovitého rozměru 26.60 mm. Tyto odchylky byly zaznamenány v setinách mm. Na základě vhodného regulačního diagramu zjistěte, zda je výrobní zařízení schopné dodržet předepsanou toleranci a zda je seřízeno na střed tolerančního pole. Vyšetřete také, zda přesnost výrobního zařízení vyhovuje předepsané toleranci. Mohli bychom k řízení uvedeného výrobního zařízení použít statistické regulace?

Data: odchylky v mm od hodnoty 26.60 mm byly vynásobeny 100:

i	x_1	x_2	x_3	x_4	x_5	x_6
1.00	-4.45	6.06	11.26	3.24	16.82	3.34
...	...	...	...	...	...	...
12.00	-11,26	10.07	-3.05	8.33	-5.46	12.66

10.6 Kontrolní hodnoty (ADSTAT, NCSS2000)

(Výsledky jsou uvedeny pouze u vybraných úloh).

Úloha C10.1 Vektor středních hodnot: μ_0 :1.452, 1.584, 7.723, 1.290; kovarianční matice:

(A) 0.37-0.05 0.14 -0.05, (B) -0.05 0.14 -0.05 0, (C) 0.14-0.05 0.28 -0.07, (D) -0.05 0 -0.07 0.17

LCL: 0, *UCL*: 16.424, data překračující *UCL*: 1, 2, 3, 4, 7, 27.

Úloha C10.2 Asymetrické rozdělení. Trend v datech. Pozitivní autokorelace.: μ_0 :224.62; Centrální linie: 224.62; *LCL*: 207.92; *UCL*: 241.33, Diagram *R*: Centrální linie: 6.28; *LCL*: 0; *UCL*: 20.52 data překračující *UCL*: 43, 50, 51, 60; Dalších 19 dat porušuje některé ze zvláštních pravidel. Vhodné použití dynamického EWMA diagramu.

Úloha C10.3 Vektor středních hodnot: μ_0 :1.452, 1.584, 7.723, 1.290; kovarianční matice:

(A) 0.37-0.05 0.14 -0.05, (B) -0.05 0.14 -0.05 0, (C) 0.14-0.05 0.28 -0.07, (D) -0.05 0 -0.07 0.17,

LCL: 0, *UCL*: 16.424, data překračující *UCL*: 1, 2, 3, 4, 7, 27.

Úloha C10.4 Diagram *x*: Centrální linie: 18.184; *LCL*: 16.457; *UCL*: 19.911, Diagram *s*: Centrální linie: 0.997; *LCL*: 0.0315; *UCL*: 2.621, Indexy způsobilosti: *Cp:* 0.738; spodní mez: 0.653; horní mez: 0.823, *Cpk*:0.719;spodní mez: 0.616; horní mez: 0.821 (dle Heavlina), *Cpm* :0.737; spodní mez: 0.652; horní mez: 0.821

Úloha C10.5 c_4 0.93999 -1s 745.1277, CL 765.99 +1s 786.8532, UCL 828.5787 -2s 724.265, LCL 703.4023 +2s 807.7159,

χ^2-kvantil(0.00135) 17.79993, χ^2-kvantil(0.99865) 0.105763, CL 43.85, LCL 92.50359, UCL 7.130452, D4 2.114,D3 0, CL 109.24, UCL 230.9293, LCL 0

Úloha C10.6 Regulační diagram s asymetrickými mezemi, *LCL* = 0, *UCL* = 1.77, na základě transformace dat; Shewhartův diagram nevhodný; sešikmené rozdělení, kladná šikmost; není lineární trend, závislost: významná pozitivní autokorelace 2. řádu.

Úloha C10.8 *LCL* = 20.107, *UCL* = 21.909, *CL* = 21.008; všechna data jsou uvnitř varovných mezí ±2σ; významná negativní korelace, ρ_1= -0.581, data mají normální rozdělení.

Úloha C10.9 Diagram CUSUM detekuje výraznou negativní odchylku od *t* = 171 pro data *i* = 32 až *i* = 47. Odhad sigma = 1.02. Shewhartův diagram nedekuje žádný problém. Pro rychlou detekci odchylky je výhodnější diagram CUSUM. Pro Shewhartův diagram je *UCL* = 169.65, *UCL* = 175.93.

Úloha C10.10 Shewhartův diagram, *LCL* = 42.530, *UCL* = 43.664, *ZL* = 43.097, příliš úzké přirozené regulační meze, podskupiny jsou korelované, možným řešením je diagram pro individuální hodnoty.

Úloha C10.12 dokonalá shoda s normálním rozdělením, žádné vybočující měření, přesto velký počet dat překračuje přirozené regulační meze *LCL* = 33.687, *UCL* = 34.492, *CL* = 34.089 z důvodu významné pozitivní autokorelace dat ρ_1 = 0.693. EWMA-dynamický diagram (w = 0.4, α = 0.1) je vhodný.

Úloha C10.13 Vybočující body: 2. v druhé lince; Shewhartovy diagramy LSL_1 = 39.99, USL_1 = 42.31, CL_1 = 41.15, LSL_2 = 39.96, USL_2 = 41.14, CL_1 = 40.55; C_{pk1} = 0.71, C_{pk2} = 0.81; *t*-test: 8.729 t_{krit} = 1.98 proto rozdílné průměry; poměr rozptylů: 2.565, F_{krit} = 1.58 proto rozdílné rozptyly.

10.7 Doporučená literatura

[1] Ryan T. P.: *Statistical Methods for Quality Improvement.* J. Wiley, New York 1989.

[2] Lucas J. M.: J. Quality Technology **14** , 51 (1982).

[3] Hawkins D. M.: J. Quality Technology **13** , 228 (1981).

[4] Alwan L. C., Roberts H. V.: J. of Business and Economic Statistics **6** , 87 (1988).

[5] Taguchi G.: *Quality Evaluation for Quality Assurance.* American-Supplier Institute, Romulus MI, 1984.

[6] Militký J.: *Metody hodnocení jakosti plošných textilií.* Sborník přednášek z mezinárodní konference Jakost 93, Osava, duben 1993.

[7] Juran J. M.: *Quality Control Handbook.* Mc Graw Hill, New York 1977 (3. vyd.).

[8] El Mogahzy Y.: "Using Off-line Quality Engineering in Textile Processing". Tex. Res. J. **62**, 622 (1992).

[9] Sullivan L. P.: *The Power of Taguchi Methods.* Quality Progres, s. 76-79 (1987).

[10] Jílek J.: *Statistické toleranční meze.* SNTL, Praha 1988.

[11] Kotz S., Johnson N. L.: *Process Capability Indices,* Champan and Hall, 1993.

[12] Lam C. T., Litting S. J.: Tech. Rept. 92, An Arbor University, 1992.

[13] Kane V. E.: J. Qual. Technol. **18**, 41 (1986).

[14] Chan L. K. a kol.: J. Qual. Technol. **20**, 160 (1988).

[15] Pearn W. L. a kol.: J. Qual. Technol. **24**, 216 (1992).

[16] Meloun M., Militký J.: *Statistické zpracování experimentálních dat.* Plus, Praha 1994.

[17] Anonym: Quality Progress, March 1994, s. 61.

[18] Choobinech F., Ballard J. L.: IEEE Trans. on Reliability, **R36**, 473 (1987).

[19] Gan F. F.: J. Quality Technology **25**, 205 (1993).

[20] Lucas J. M., Crosier R. B.: Commun. Statist. **A11**, 2669 (1982).

[21] Montgomery D. C., Mastrangelo Ch. H.: J. Quality Technol., **23**, 179 (1991).

[22] Wetheril G. B., Brown D. W.: *Statistical Process Control, Theory and Practice.* Chapman and Hall, London 1991.

[23] Jackson J. E.: Commun. Statist. **A14**, 2657 (1985).

[24] Crosier R. B.: Technometrics **30**, 291 (1988).

[25] Pignatello J. J., Runger G. C.: J. Quality Technol. **22**, 173 (1990).

[26] Wade M.R., Woodall W.H.: J. Quality Technol. **25**, 161 (1993).

[27] Box G. E. P.: *Studies in Quality Improvement.* Rept. No. 26, University of Wisconsin, July 1987.

[28] Hayes G. D., Scallan A. J., Wong J. H. F.: Food Control **8,** 173 - 176 (1997).

Rejstřík:

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

Z