evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Počítačové cvičení

předmětu M6130 Výpočetní statistika

Cvičení 1.: Průzkumová analýza jednorozměrných dat

Vedení pojišťovny (zaměřené na pojištění automobilů) požádalo manažera oddělení marketingového výzkumu o provedení průzkumu, který by ukázal názory zákazníků na uvažovaný nový systém pojištění aut.

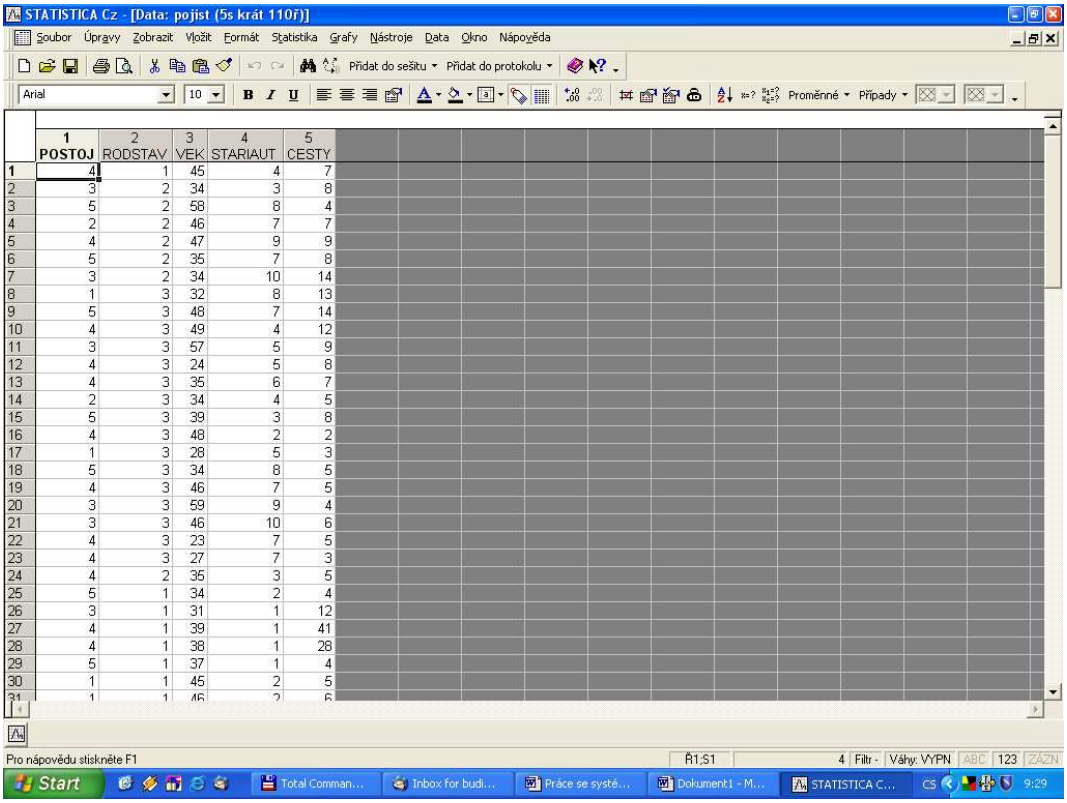
Náhodně bylo vybráno 110 současných zákazníků pojišťovny a ti byli telefonicky seznámeni s následujícím textem:

„Naše pojišťovna nabízí nový systém pojištění aut výhradně pro cesty nad 300 km. Za roční poplatek 12 tisíc Kč budete pojištěni pro případ libovolných potíží s autem při všech cestách nad 300 km. V případě nehody pojišťovna uhradí opravu, cestovní náklady a popř. i některé další výlohy, jako je ubytování a stravování v hotelu, telefon atd.

Stupnicí od 1 (jednoznačný nezájem) do 5 (jednoznačný zájem) laskavě vyjádřete svůj postoj k nabízenému novému typu pojištění. Dále uveďte svůj věk, počet cest nad 300 km v loňském roce, stáří vašeho auta a váš rodinný stav. Děkujeme."

Získané odpovědi byly zaznamenány do datového souboru pojist.sta a zakódovány takto:
POSTOJ ... postoj k novému typu pojištění (jednoznačný nezájem = 1, lehký nezájem = 2, neutrální postoj = 3, lehký zájem = 4, jednoznačný zájem = 5).
RODSTAV ... rodinný stav (svobodný = 1, rozvedený, ovdovělý = 2, ženatý = 3).
VEK ... věk v dokončených letech.
STARIAUT ... stáří auta v letech.
CESTY ... počet cest nad 300 km v předešlém roce.

Ukázka části datového souboru:



	1 POSTOJ	2 RODSTAV	3 VEK	4 STARIAUT	5 CESTY
1	4	1	45	4	7
2	3	2	34	3	8
3	5	2	58	8	4
4	2	2	46	7	7
5	4	2	47	9	9
6	5	2	35	7	8
7	3	2	34	10	14
8	1	3	32	8	13
9	5	3	48	7	14
10	4	3	49	4	12
11	3	3	57	5	9
12	4	3	24	5	8
13	4	3	35	6	7
14	2	3	34	4	5
15	5	3	39	3	8
16	4	3	48	2	2
17	1	3	28	5	3
18	5	3	34	8	5
19	4	3	46	7	5
20	3	3	59	9	4
21	3	3	46	10	6
22	4	3	23	7	5
23	4	3	27	7	3
24	4	2	35	3	5
25	5	1	34	2	4
26	3	1	31	1	12
27	4	1	39	1	41
28	4	1	38	1	28
29	5	1	37	1	4
30	1	1	45	2	5
31	1	1	46	2	6

Úkol 1.: Datový soubor pojist.sta načtěte do systému STATISTICA. Všem proměnným vytvořte návěští a popište význam jednotlivých variant proměnných POSTOJ a RODSTAV.

Návod: Soubor – Otevřít – pojist.sta – Otevřít.
Názvy a vlastnosti proměnných se upravují v okně, do něhož vstoupíme, když 2x klikneme myší na název proměnné. Návěští se píše do Dlouhého jména, význam variant do Text. hodnot.

Úkol 2. Zjistěte absolutní a relativní četnosti a absolutní a relativní kumulativní četnosti proměnných POSTOJ a RODSTAV.
Návod: Statistiky – Základní statistiky/Tabulky – Tabulky četností – OK – Proměnné POSTOJ, RODSTAV – OK – Výpočet.
Tabulky se uloží do pracovního sešitu, listovat v nich můžeme pomocí stromové struktury v levé části okna.

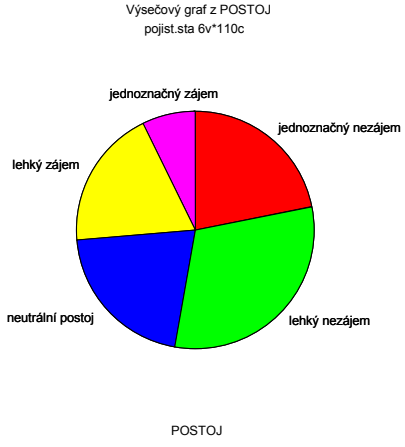
Tabulka četností pro POSTOJ

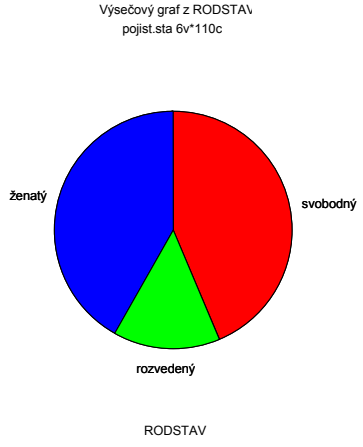
	Tabulka četností:POSTOJ: postoj k novému typu pojišt (pojist.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
jednoznačný nezájem	24	24	21,81818	21,8182
lehký nezájem	34	58	30,90909	52,7273
neutrální postoj	23	81	20,90909	73,6364
lehký zájem	21	102	19,09091	92,7273
jednoznačný zájem	8	110	7,27273	100,0000
ChD	0	110	0,00000	100,0000

Tabulka četností pro RODSTAV

	Tabulka četností:RODSTAV: rodinný stav zákazníka (pojist.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
svobodný	48	48	43,63636	43,6364
rozvedený	16	64	14,54545	58,1818
ženatý	46	110	41,81818	100,0000
ChD	0	110	0,00000	100,0000

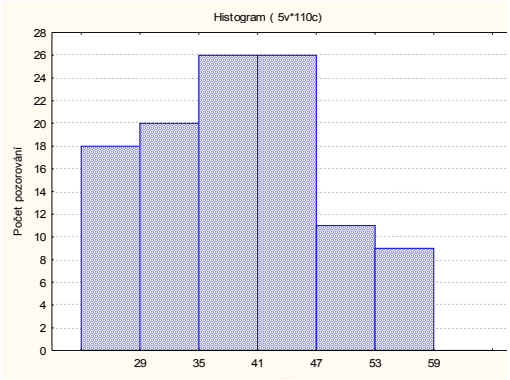
Úkol 3. Absolutní četnosti proměnných POSTOJ a RODSTAV znázorněte graficky pomocí výsečového diagramu.
Návod: V menu zvolíme Grafy – 2D Grafy – Výsečové grafy.
Vybereme proměnné POSTOJ, RODSTAV a dostaneme následující grafy:





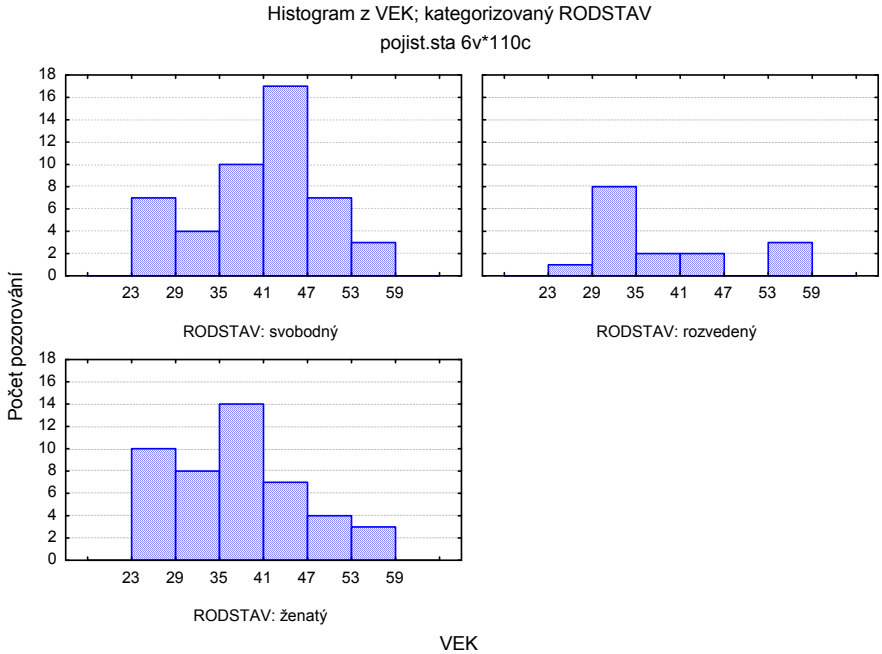
Z prvního diagramu je zřejmé, že nejméně zákazníků projevilo jednoznačný zájem o nový typ pojištění. Ostatní varianty jsou zastoupeny vcelku rovnoměrně.
Co se týká rodinného stavu zákazníků, vidíme, že v daném souboru jsou s přibližně stejnou četností zastoupeni ženatí a svobodní zákazníci. Rozvedených či ovdovělých je nejméně.

Úkol 4. Vytvořte histogram proměnné VEK se šesti třídicími intervaly <23,29>, (29,35>, (35,41>, (41,47>, (47,53>, (53,59>.
Návod: V menu vybereme Grafy – Histogramy – Proměnné VEK, OK, Detaily – zaškrtneme Hranice – Určit hranice – zaškrtneme Zadejte hraniční rozmezí, Minimum 23, Krok 6, Maximum 59 – OK – Vypneme normální proložení – OK. Dostaneme histogram v tomto tvaru:



Ze vzhledu histogramu lze soudit, že v souboru zákazníku jsou nejvíce zastoupeni lidé od 35 do 47 let. Soubor vykazuje kladné zešikmení, protože mladší věkové kategorie jsou zastoupeny s vyšší četností než starší věkové kategorie.

Úkol 5.: Vytvořte kategorizovaný histogram proměnné VEK podle proměnné RODSTAV.
Návod: Postupujeme stejně jako v předešlém případě a zvolíme Kategorizovaný – Kategorie X – Zapnuto – Změnit proměnnou RODSTAV – OK - OK.



Úkol 6.: Vypočtěte následující číselné charakteristiky:
POSTOJ (ordinální proměnná) – modus, medián, dolní a horní kvartil, kvartilová odchylka.
RODSTAV (nominální proměnná) – modus.
VEK, STARIAUT, CESTY (poměrové proměnné) – průměr, směrodatná odchylka, koeficient variace, šikmost, špičatost.
Návod: Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK, Proměnné – zadáme název příslušné proměnné, Detailní výsledky – vybereme příslušné charakteristiky.

	Popisné statistiky (pojist.sta)					
Proměnná	Medián	Modus	Četnost modu	Spodní kvartil	Horní kvartil	Kvartilové rozpětí
POSTOJ	2	2	34	2	4	2

Vidíme, že medián, modus a dolní kvartil jsou stejné – je to varianta 2 „lehký nezájem“. Horním kvartilem je varianta 4 „lehký zájem“.

	Popisné statistiky (pojist.sta)	
Proměnná	Modus	Četnost modu
RODSTAV	1	48

V našem datovém souboru je nejčetnější variantou rodinného stavu varianta 1 „svobodný“.

	Popisné statistiky (pojist.sta)				
Proměnná	Průměr	Sm.odch.	Koef.prom.	Šikmost	Špičatost
VEK	39,58182	8,823844	22,29267	0,191625	-0,59532
STARIAUT	4,16364	2,359938	56,67974	0,905405	0,35924
CESTY	7,16364	5,304537	74,04811	3,150711	15,99807

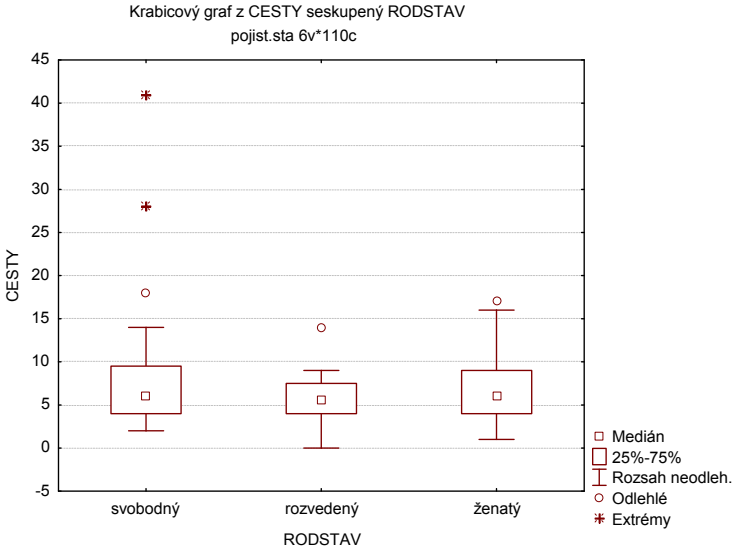
Průměrný věk zákazníka je 39 let a 7 měsíců se směrodatnou odchylkou 8 let a 10 měsíců. Rozložení věku vykazuje kladnou šikmost (podprůměrné hodnoty věku jsou četnější než nadprůměrné) a zápornou špičatost (rozložení věku je plošší než normální rozložení).
Průměrné stáří auta je 4 roky a 2 měsíce se směrodatnou odchylkou 2 roky a 4 měsíce. Rozložení stáří aut je kladně zešikmené a špičatější než normální rozložení.
Průměrný počet cest nad 300 km je 7,2 se směrodatnou odchylkou 5,3. Rozložení počtu cest na 300 km je značně kladně zešikmené a podstatně špičatější než normální rozložení.
Z porovnání variability uvedených tří proměnných pomocí koeficientů variace (koeficient variace je podíl směrodatné odchylky a průměru, často se udává v procentech) vyplývá, že nejvyšší variabilitu má proměnná CESTY, nejnižší VEK.

Úkol 7.: Zjistěte, jaký je průměrný počet cest nad 300 km pro svobodné, rozvedené , ženaté zákazníky pojišťovny. Výpočet doplňte krabicovým diagramem.
Návod: Statistiky – Základní statistiky/tabulky – Rozklad&jednofakt. ANOVA – OK – Proměnné – Závisle proměnné CESTY, Grupovací proměnná RODSTAV – OK – OK – Popisné statistiky – ponecháme jen N platných – Výpočet

Rozkladová tabulka popisných statistik (pojist.sta) N=110 (V seznamu záv. prom. nejsou ChD)		
RODSTAV	CESTY průměr	CESTY N
svobodný	7,895833	48
rozvedený	5,750000	16
ženatý	6,891304	46
Vš.skup.	7,163636	110

Vidíme, že nejvyšší průměrný počet cest nad 300 km mají svobodní zákazníci pojišťovny.

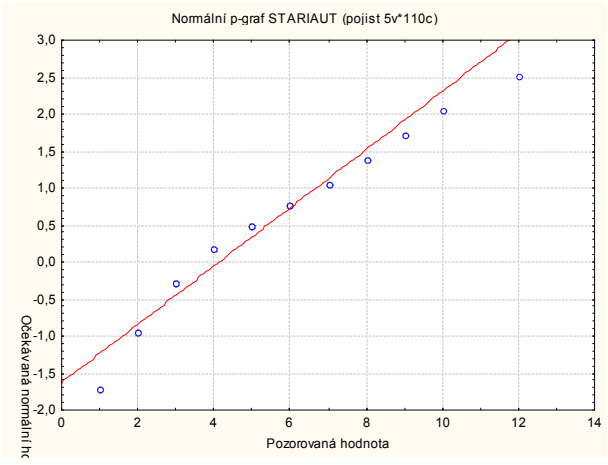
Vytvoření krabicového grafu: Grafy – 2D Grafy – Krabicové grafy – Proměnné – Závisle proměnné CESTY, Grupovací proměnná RODSTAV – OK – OK



Ve všech třech variantách rodinného stavu se vyskytují odlehlé hodnoty, u svobodných zákazníků pojišťovny jsou dokonce i extrémní hodnoty.

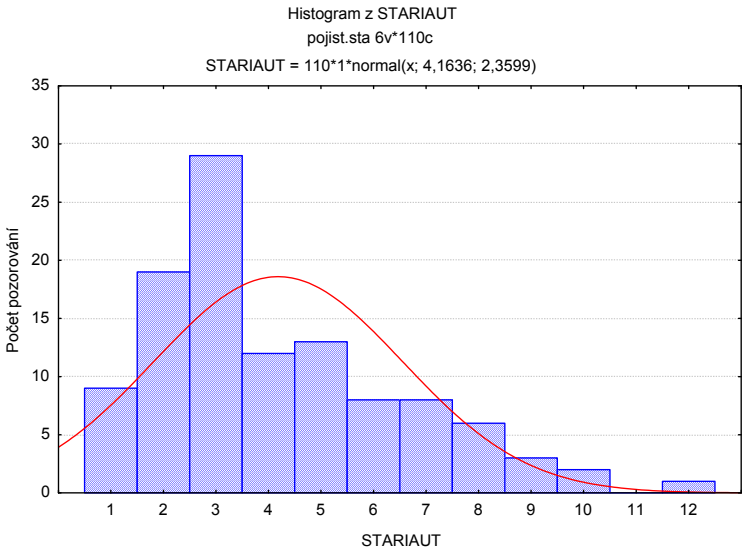
Úkol 8.: Pro proměnnou STARIAUT sestrojte N-P graf a s jeho pomocí posuďte normalitu této proměnné.

Návod: Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné STARIAUT – OK.



Tečky v NP grafu se značně odchylují od zakreslené přímky a řadí se do konkávního tvaru. Datový soubor vykazuje kladné zešikmení, nejedná se tedy o normální rozložení.

Úkol 9.: Pro proměnnou STARIAUT nakreslete histogram s proloženou hustotou normálního rozložení. Ponechejte implicitní počet třídicích intervalů.
Návod: Grafy – Histogramy – Proměnné STARIAUT – OK.

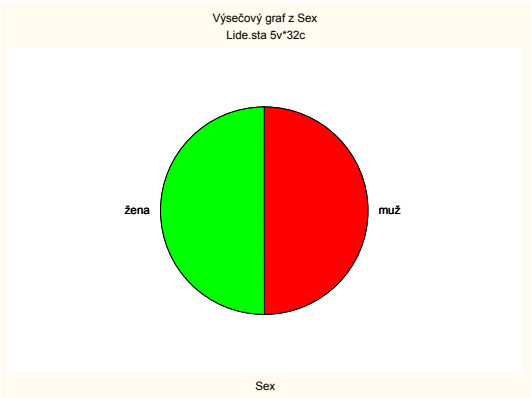


Tvar histogramu svědčí o kladně zešikmeném rozložení, jehož hustota neodpovídá hustotě normálního rozložení.

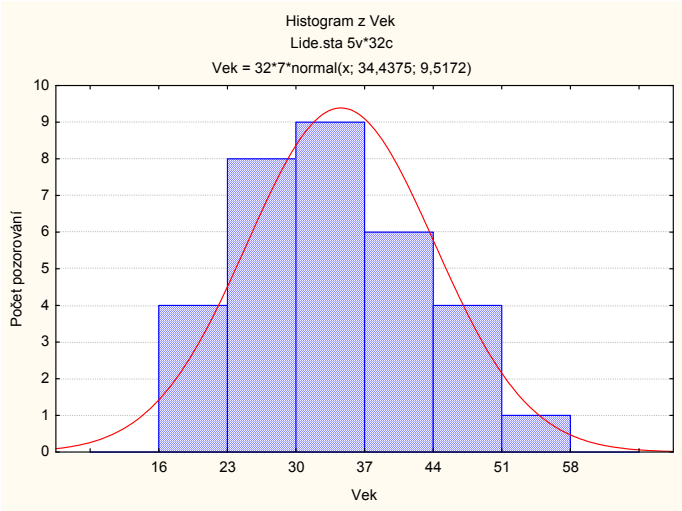
Příklad k samostatnému řešení:
Načtěte datový soubor lide.sta, s nímž jste pracovali v 1. cvičení.

1. Vytvořte tabulku absolutních a relativních četností proměnné SEX. Četnosti znázorněte pomocí výsečového diagramu.

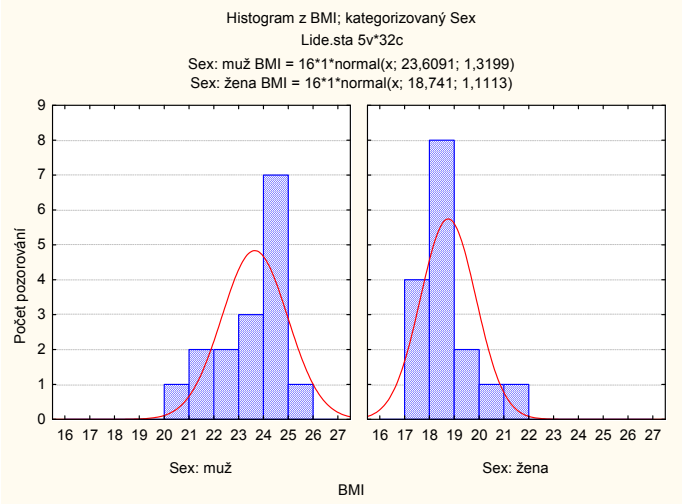
	Tabulka četností:Sex (Lide.sta)	
Kategorie	Četnost	Rel.četnost
muž	16	50
žena	16	50



2. Vytvořte histogram proměnné VEK se šesti třídicími intervaly (16,23>, (23,30>, (30,37>, (37,43>, (43,50>, (50,57> a zakreslenou Gaussovou křivkou.



3. Vytvořte kategorizované histogramy proměnné BMI pro muže a pro ženy.



4. Vypočtěte průměr, směrodatnou odchylku, koeficient variace, šikmost a špičatost proměnné BMI pro muže a pro ženy. Výsledky udávejte na dvě destinná místa.

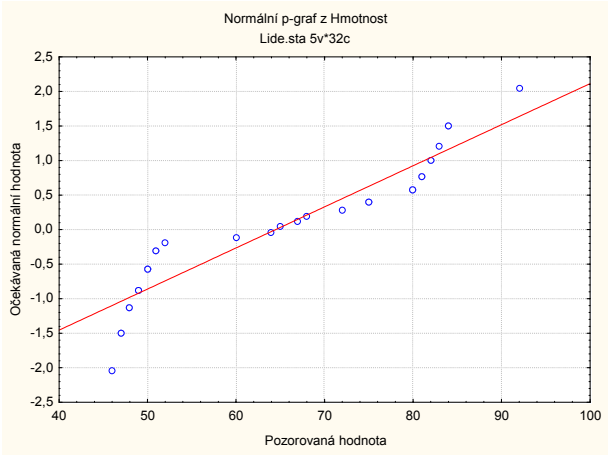
Pro muže:

	Popisné statistiky (Lide.sta) Zhrnout podmínku: Sex=1					
Proměnná	N platných	Průměr	Sm.odch.	Koef.prom.	Šikmost	Špičatost
BMI	16	23,61	1,32	5,59	-0,78	-0,25

Pro ženy

	Popisné statistiky (Lide.sta) Zhrnout podmínku: Sex=2					
Proměnná	N platných	Průměr	Sm.odch.	Koef.prom.	Šikmost	Špičatost
BMI	16	18,74	1,11	5,93	1,39	2,65

5. Sestrojte N-P plot pro proměnnou Hmotnost.



6. Vytvořte kategorizované krabicové diagramy pro proměnnou Vyska pro muže a pro ženy.

7. K extrémní hodnotě výšky umístěte jméno muže, kterému tato výška přísluší.
(Jan)

Cvičení 2.: Shluková analýza

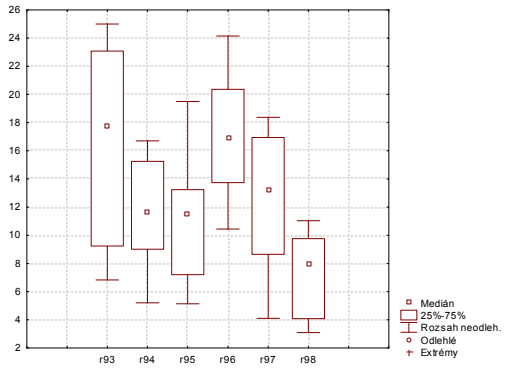
V souboru stanice.sta jsou uloženy údaje (v $\mu g/m^3$) o průměrných ročních koncentracích oxidu siřičitého v letech 1993 – 1998 na deseti brněnských měřicích stanicích: Dobrovského, Húskova, Krasová, Kroftova, Mendelova zemědělská a lesnická univerzita, Polní, Přízřenice, Skaunicové, Soběšice, Tuřany. Cílem je najít metodami shlukové analýzy skupiny stanic, které vykazují podobné rysy chování.

Datový soubor:

	1 Stanice	2 r93	3 r94	4 r95	5 r96	6 r97	7 r98
1	DOB	6,828	5,202	5,137	11,568	4,104	3,097
2	HUS	9,241	9,281	10,259	10,442	7,035	3,857
3	KRA	7,205	5,535	5,197	13,741	8,651	4,085
4	KRO	24,039	9,018	12,237	18,189	15,601	9,762
5	MZL	23,079	16,222	13,353	20,363	15,312	7,925
6	POL	25,005	14,568	10,723	15,76	11,068	4,916
7	PRI	15,874	15,251	13,241	19,435	16,943	8,081
8	SKA	14,297	9,49	7,209	14,434	10,961	8,063
9	SOB	19,728	13,772	12,943	20,948	17,564	11,039
10	TUR	22,524	16,708	19,502	24,144	18,377	11,024

Úkol 1.: Soubor stanice.sta upravte tak, aby případy 1 až 10 byly pojmenovány názvy stanic.
Návod: Data – Správce jmen případů – Délka jména příp. 5, Přenést jména případů z proměnné Stanice, OK.

Úkol 2.: Prozkoumejte proměnné r93 až r98 pomocí krabicových diagramů.
Návod: Grafy – 2D Grafy – Krabicové grafy –Typ grafu vícenásobný – Proměnné r93, ..., r98, OK, OK.

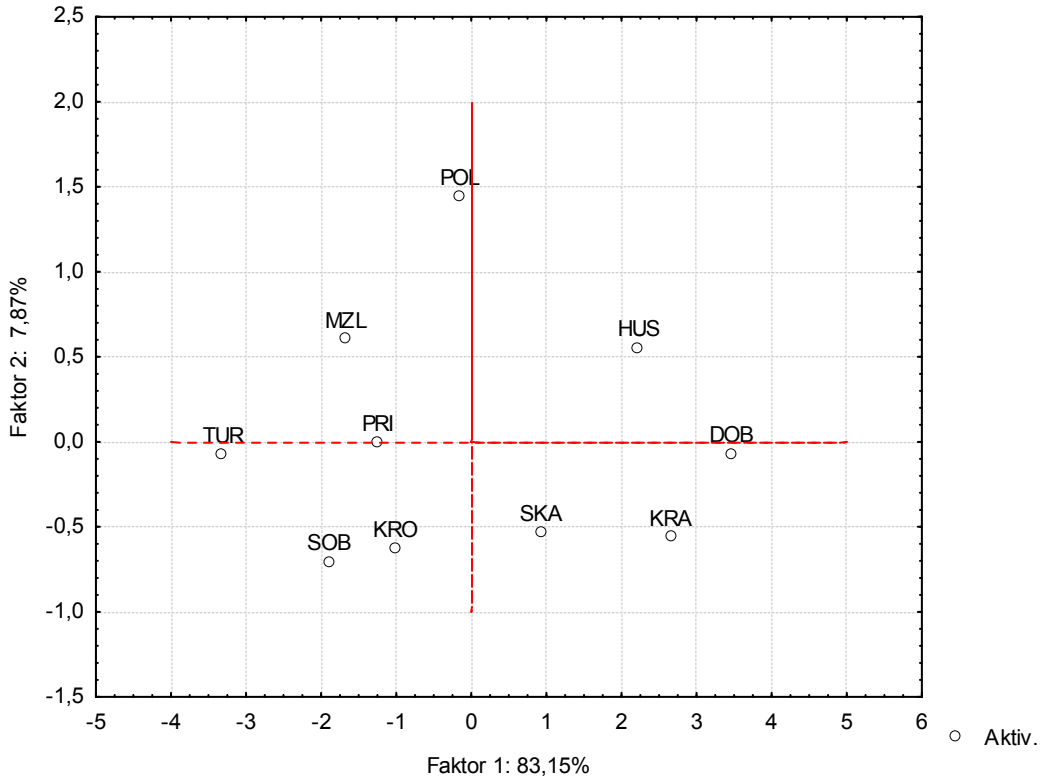


Interpretace: Z krabicových diagramů je vidět, že proměnné r93 až r98 vykazují velmi rozdílnou variabilitu. Nejvyšší variabilitu ve sledovaných deseti stanicích měly koncentrace oxidu siřičitého v roce 1993, naopak nejmenší v roce 1998.

Úkol 3.: Vzhledem k velmi rozdílné variabilitě proměnných r93 až r98 vytvořte standardizované proměnné a nadále pracujte s nimi.
Návod: Data – Standardizovat – Proměnné r93, ..., r98, OK.

	1 Stanice	2 r93	3 r94	4 r95	5 r96	6 r97	7 r98
DOB	DOB	-1,398	-1,4569	-1,3398	-1,2048	-1,7224	-1,3635
HUS	HUS	-1,0591	-0,514	-0,1653	-1,4591	-1,1255	-1,11
KRA	KRA	-1,3451	-1,3799	-1,326	-0,714	-0,7964	-1,0339
KRO	KRO	1,01924	-0,5748	0,28819	0,29058	0,61898	0,85957
MZL	MZL	0,88441	1,09043	0,54408	0,78159	0,56013	0,24685
POL	POL	1,15491	0,7081	-0,0589	-0,258	-0,3042	-0,7568
PRI	PRI	-0,1275	0,86598	0,5184	0,57199	0,89228	0,29889
SKA	SKA	-0,349	-0,4657	-0,8647	-0,5575	-0,326	0,29288
SOB	SOB	0,41376	0,5241	0,45007	0,91371	1,01875	1,2855
TUR	TUR	0,80646	1,20277	1,95397	1,63553	1,18432	1,2805

Úkol 4.: Z proměnných r93 až r98 vytvořte dvě hlavní komponenty a graficky znázorněte rozmístění stanic na ploše prvních dvou hlavních komponent.
Návod: Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné r93, ..., r98, OK, OK – Počet faktorů 2, zaškrtneme 2D graf fakt. souřadnic případů.

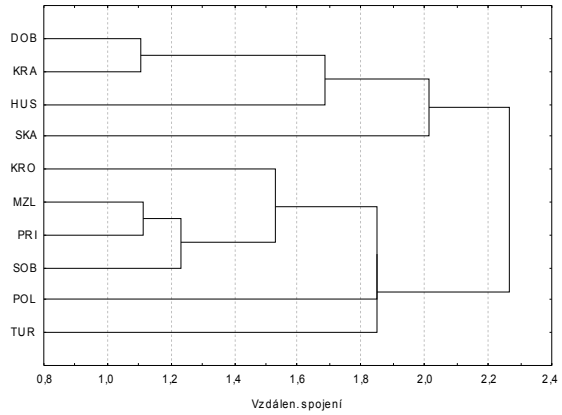


Interpretace: Z rozmístění stanic na ploše prvních dvou hlavních komponent lze usoudit, že stanice DOB, KRA, HUS, SKA mohou tvořit jeden shluk, stanice KRO, SOB, PRI, TUR, MZL druhý shluk a stanice POL se chová poněkud atypicky.

Úkol 5.: Pro standardizované proměnné r93 až r98 proveďte shlukovou analýzu s euklidovskou vzdáleností a třemi metodami: nejbližšího souseda, nejvzdálenějšího souseda a průměrné vazby. Výsledky znázorněte pomocí dendrogramu.

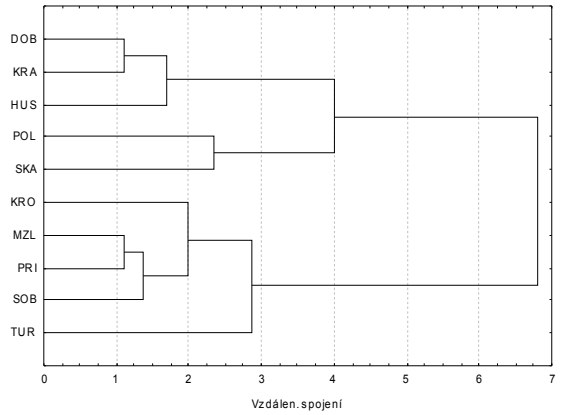
Návod: Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné r93 až r98 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky), pravidlo slučování ponecháme Jednoduché spojení, míru vzdálenosti ponecháme Euklidovské vzd. – OK – Horizontální graf hierarch. stromu. Pro další dvě metody na záložce Detaily vybereme pravidlo slučování Úplné spojení resp. Nevážený průměr skupin dvojic.

Dendrogram pro metodu nejbližšího souseda



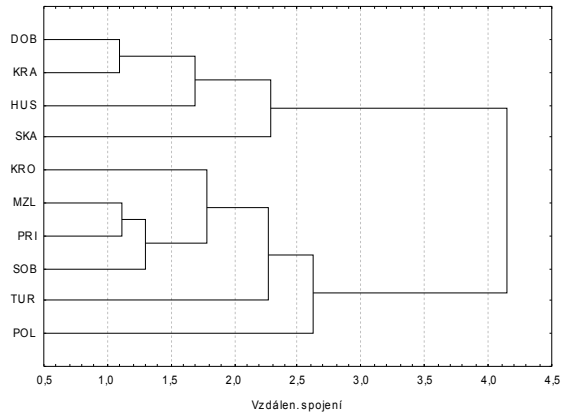
Interpretace: Stanice DOB, KRA, HUS a STA tvoří jeden shluk, stanice KRO, MZL, PRI, DOB, POL a TUR druhý shluk.

Dendrogram pro metodu nejvzdálenějšího souseda



Interpretace: Stanice DOB, KRA, HUS, POL a STA tvoří jeden shluk, stanice KRO, MZL, PRI, SOB a TUR druhý shluk.

Dendrogram pro metodu průměrné vazby



Interpretace: Stanice DOB, KRA, HUS a STA tvoří jeden shluk, stanice KRO, MZL, PRI, SOB, TUR a POL druhý shluk.

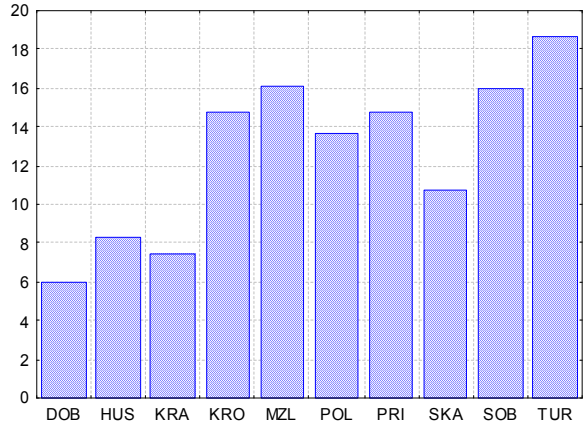
Shrneme-li výsledky všech tří metod, je zřejmé, že stanice DOB, KRA, HUS a STA zřejmě patří do jednoho shluku, zatímco stanice KRO, MZL, SOB a TUR patří do druhého shluku. Příslušnost stanice POL k jednomu či druhému shluku není jednoznačná.

Úkol 6.: Vypočtěte a pomocí sloupkových diagramů znázorněte průměrné roční koncentrace SO_2 a směrodatné odchylky za celé sledované období pro všech deset stanic.
Návod: Je nutné se vrátit k původním nestandardizovaným hodnotám, tj. znovu načíst soubor stanice.sta a pojmenovat případy názvy stanic – viz úkol 1. Pak je zapotřebí soubor transponovat – zaměnit řádky za sloupce: Data – Transponovat – Soubor. Vymažeme 1. řádek: Případy – Odstranit – Od případu 1 do případu 1, OK. Pomocí Popisných statistik vypočteme průměry a směrodatné odchylky proměnných DOB až TUR.

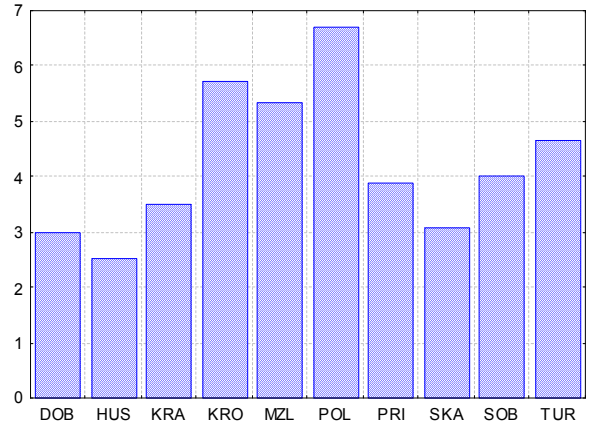
	Popisné statistiky (Tema4	
Proměnná	Průměr	Sm. odch.
DOB	5,98933	3,003043
HUS	8,35250	2,513866
KRA	7,40233	3,496625
KRO	14,80767	5,707322
MZL	16,04233	5,326765
POL	13,67333	6,719292
PRI	14,80417	3,873187
SKA	10,74233	3,083617
SOB	15,99900	3,993683
TUR	18,71317	4,645334

Vytvoření sloupkových diagramů pro průměry: ve workbooku klikneme pravým tlačítkem myši na sloupek Průměr: Grafy bloku dat – Vlastní graf bloku podle sloupce – Typ grafu – Sloupcové/pruhové grafy - OK. Podobně pro směrodatné odchylky.

Sloupkový diagram pro průměry



Slopupkový diagram pro sm. odchylky



Interpretace: Stanice v 1. shluku (DOB, HUS, KRA, SKA) vykazují za sledované období poměrně nízké průměrné koncentrace SO_2 (od 6 $\mu g/m^3$ po 11 $\mu g/m^3$) i malé směrodatné odchylky (od 2,5 $\mu g/m^3$ po 3,5 $\mu g/m^3$). Druhý shluk obsahuje stanice s vysokými koncentracemi (od 13 $\mu g/m^3$ po 19 $\mu g/m^3$) a velkými směrodatnými odchylkami (od 3,8 $\mu g/m^3$ po 6,8 $\mu g/m^3$).

Příklad k samostatnému řešení:
U 12 velmi slavných amerických hráčů košíkové byly v sezóně 1989 zjištěny hodnoty osmi proměnných.
Výška – výška hráče v cm
Hmotnost – hmotnost hráče v kg
FgPct – první antropometrická charakteristika
FtPct – druhá antropometrická charakteristika
Body – průměrný počet dosažených bodů
Doskoky - průměrný počet doskoků
Asistence – průměrný počet asistencí
Fauly – průměrný počet faulů
Data jsou uložena v souboru hraci_kosikove.sta.

	1 Jméno hráče	2 Vyska	3 Hmotnost	4 Fgpct	5 Ftpct	6 Body	7 Doskoky	8 Asistence	9 Fauly
1	Jabbar K.A.	218,6	105,0	55,9	72,1	24,6	11,2	3,6	3
2	Barry R.	200,8	93,6	44,9	90,0	23,2	6,7	4,9	3
3	Baylor E.	195,7	102,7	43,1	78,0	27,4	13,5	4,3	3,1
4	Bird L.	205,9	100,4	50,3	88,0	25,0	10,2	6,1	2,7
5	Chamberlain W.	216,0	125,5	54,0	51,1	30,1	22,9	4,4	2
6	Cousy B.	184,3	79,9	37,5	80,3	18,4	5,2	7,5	2,4
7	Erving J.	199,5	91,3	50,6	77,8	24,2	8,5	4,2	2,8
8	Johnson M.	205,9	98,1	53,0	83,4	19,5	7,4	11,2	2,4
9	Jordan M.	198,3	89,0	51,3	84,8	32,6	6,2	5,9	3,1
10	Robertson O.	195,7	95,8	48,5	83,8	25,7	7,5	9,5	2,8
11	Russell B.	207,1	100,4	44,0	56,1	15,1	22,6	4,3	2,7
12	West J.	189,4	82,2	47,4	81,4	27,0	5,8	6,7	2,6

Metodami shlukové analýzy najděte skupiny hráčů podobných vlastností.

(Příklad je převzat z knihy M. Meloun, J. Militký, M. Hill: Počítačová analýza vícerozměrných dat. Academia Praha 2005)

Cvičení 3: Základní pojmy matematické statistiky I

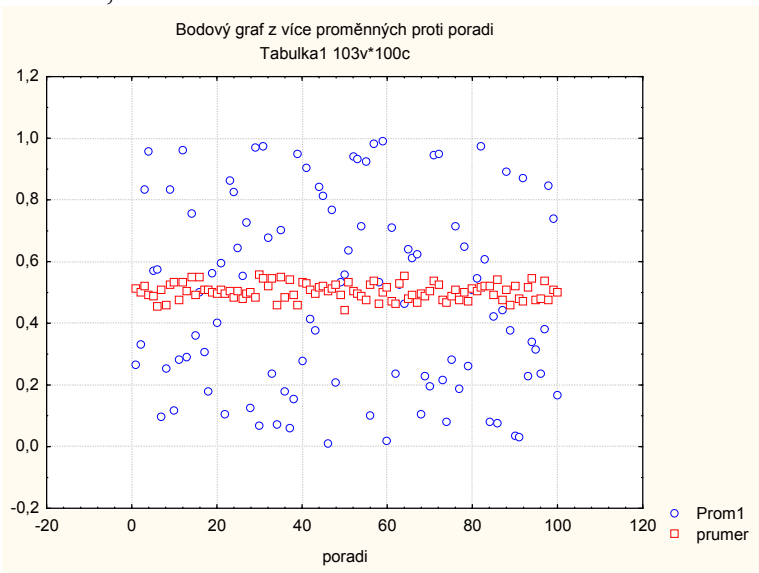
Úkol 1.: Průzkum chování výběrového průměru a výběrového rozptylu

1. Vytvořte nový datový soubor o 103 proměnných a 100 případech. Pomocí programu gener.svb, který si stáhnete z Učebních materiálů, se naplní prvních 100 proměnných 100 realizacemi náh. veličin $X_i \sim Rs(0,1)$, i=1, …, 100, do proměnné v101 se uloží pořadová čísla 1 až 100, do proměnné v102 (resp. v103) se uloží průměry (resp. rozptyly) proměnných v1 až v100.

```
Option Base 1
Sub Main
   Dim s As Spreadsheet
        Set s = ActiveSpreadsheet
        For i = 1 To 100
              s.Variable(i).FillRandomValues
              'do promennych v1 az v100 se ulozi nahodna cisla z intervalu(0,1)
        Next i
        s.VariableLongName(101) = "=v0"
        'do promenne v101 se ulozi poradova cisla 1 az 100
        s.VariableLongName(102) = "=mean(v1:v100)"
        'do promenne v102 se ulozi prumery promennych v1 az v100
        s.VariableLongName(103) = "=stdev(v1:v100)^2"
        'do do promenne v103 se ulozi rozptyly promennych v1 az v100
        s.Recalculate
End Sub
```

(Makro se spouští pomocí modré šipky na panelu nástrojů.)
Proměnnou v101 přejmenujte na PORADI, v102 na PRUMER a v103 na ROZPTYL. Vzniklý datový soubor uložte pod názvem uniform.sta.

2. Graficky znázorněte hodnoty některé z proměnných v1, …, v100 (např. v1) a hodnoty proměnné PRUMER.
Návod: Grafy – Bodové grafy – Typ grafu Vícenásobný – vypneme Lineární proložení – Proměnné X PORADI, Y v1, PRUMER, OK, OK. Vidíme, že hodnoty proměnné v1 se nacházejí mezi 0 a 1, zatímco hodnoty proměnné PRUMER se koncentrují v úzkém pásu kolem 0,5.

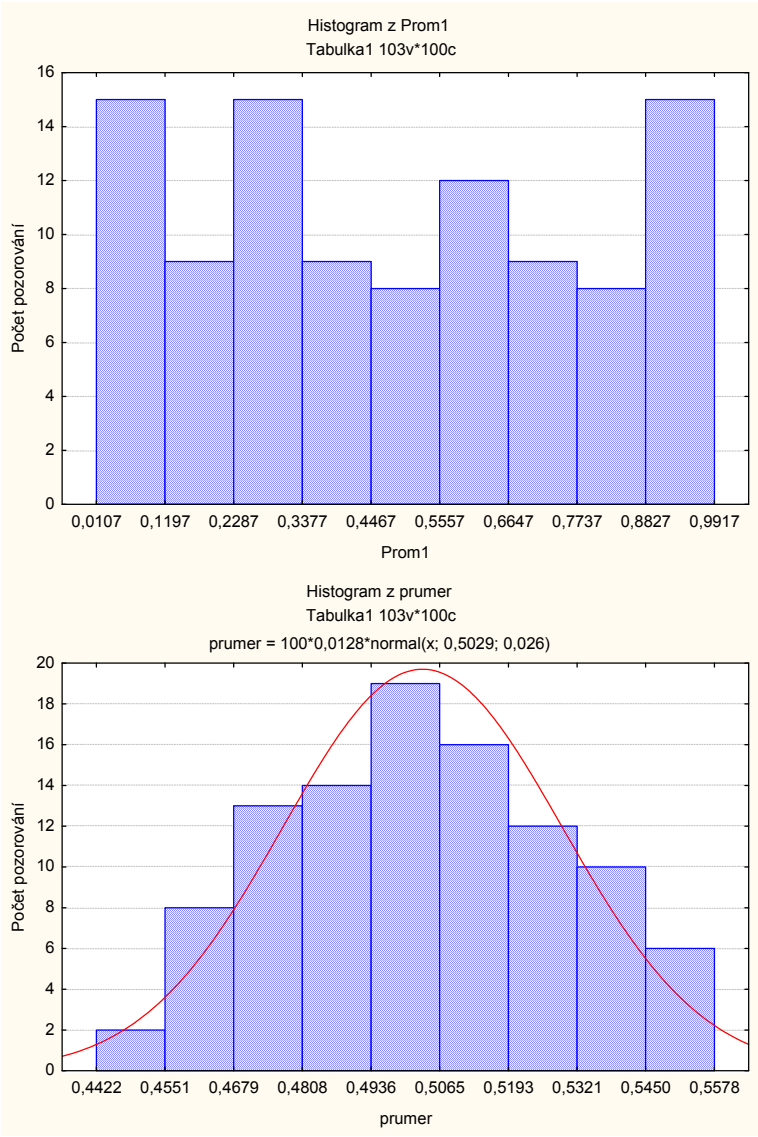


3. Vypočtěte průměr a rozptyl např. proměnné v1 a proměnné PRUMER. Průměr proměnné v1 by měl být blízký 0,5, rozptyl 1/12 = 0,083. Průměr proměnné PRUMER by se měl blížit 0,5, zatímco rozptyl by měl být 100 x menší než 1/12, tj. 0,00083. Dále vypočtěte průměr proměnné ROZPTYL. Měl by se blížit 1/12 = 0,083.

	Popisné statistiky (uniform)	
Proměnná	Průměr	Rozptyl
Prom1	0,536605	0,078676
PRUMER	0,503984	0,000783

	Popisné statistiky (uniform)
Proměnná	Průměr
ROZPTYL	0,083143

4. Nakreslete histogram pro proměnnou v1 a pro proměnnou PRUMER. První histogram se blíží úsečce, druhý Gaussově křivce.



5. Celý postup zopakujte pro exponenciální rozložení s parametrem λ=2. V programu gener.stb napište místo s.Variable(i).FillRandomValues
s.VariableLongName(i) = "=Vexpon(rnd(1);2) "

Připomeneme si, že průměr proměnné v1 i průměr proměnné PRUMER by se měl blížit 1/2, rozptyl proměnné v1 by měl být blízký 1/4 a rozptyl proměnné PRUMER by měl být 100 x menší, tj. 0,0025. Průměr proměnné ROZPTYL by se neměl příliš lišit od 1/4.

Úkol 2.: Ilustrace nestrannosti výběrové distribuční funkce

1. Vytvořte nový datový soubor o třech proměnných a 1000 případech.

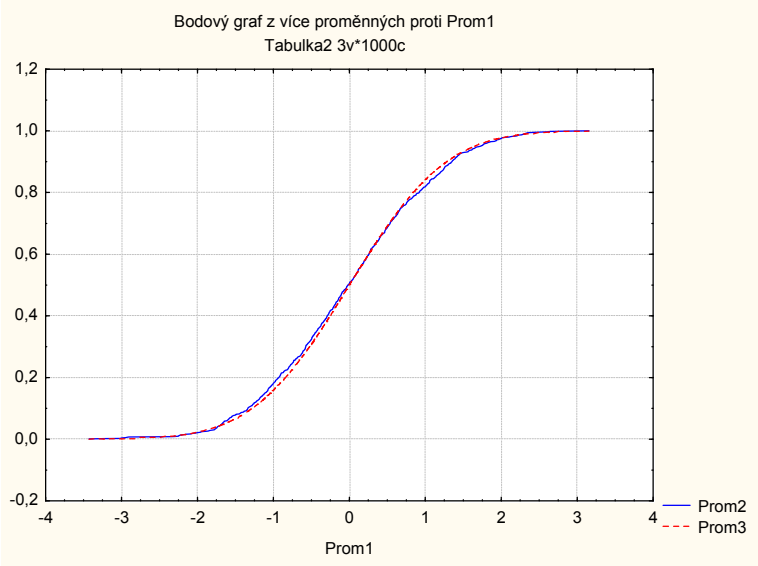
2. Do proměnné v1 uložte 1000 realizací náhodné veličiny s rozložením N(0,1) tak, že v Dlouhém jménu použijte příkaz =vnormal(rnd(1);0;1)

3. Hodnoty proměnné v1 setřiďte podle velikosti: Data - Setřídit.

4. Proměnnou v2 transformujte tak, že v Dlouhém jménu použijte příkaz =v0/1000.

5. Do proměnné v3 uložte hodnoty distribuční funkce rozložení N(0,1). Do Dlouhého jména napište příkaz =INormal(v1;0;1)

6. Nakreslete dvourozměrný tečkový diagram, kde na osu x vyneste v1 a na osu y v2 a v3.



Vidíme, že průběh výběrové distribuční funkce $F_{1000}(x)$ (modrá čára) je velmi podobný průběhu distribuční funkce $\Phi(x)$ (červená čára).

7. Postup zopakujte pro rozsah výběru n = 100. Uvidíte, průběh výběrové distribuční funkce $F_{100}(x)$ se od průběhu distribuční funkce $\Phi(x)$ liší výrazněji.

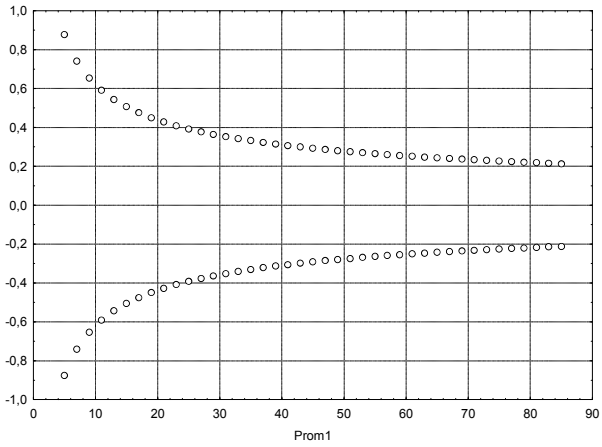
Úkol 3.: Sledování vlivu rozsahu výběru na šířku intervalu spolehlivosti (při α=0,05)

Pro hypotetické náhodné výběry rozsahu n (n = 5, 7, 9, …, 85) z rozložení N(0,1), jejichž výběrové průměry se vždy realizovaly hodnotou 0, vypočtěte dolní a horní meze 95% intervalů spolehlivosti pro μ a graficky znázorněte závislost těchto mezí na rozsahu n. Upozornění: Meze 100(1-α)% empirického intervalu spolehlivosti pro střední hodnotu při známém rozptylu se počítají podle vzorců: $d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$, $h = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$

Návod: Z Učebních materiálů stáhněte program intsp1.svb a otevřete ho v programovacím okně.

```
Option Base 1
Dim s As Spreadsheet
Sub Main
        alfa = 0.05
        'pevně zvolené riziko
        m = 0
        'pevně zvolený průměr
        sigma = 1
        'pevně zvolená směrodatná odchylka
        n = 3
        'počáteční rozsah výběru
        Set s = ActiveSpreadsheet
        For I = 1 To 41
                s.Cells(I, 2) = m - VNormal(1 - alfa / 2, 0, 1) / Sqrt(n + 2 * I)
                'dolní mez intervalu spolehlivosti
                s.Cells(I, 3) = m + VNormal(1 - alfa / 2, 0, 1) / Sqrt(n + 2 * I)
                'horní mez intervalu spolehlivosti
                s.Cells(I, 1) = n + 2 * I
                'zvětšení rozsahu výběru o 2
        Next I
End Sub
```

Vytvořte nový datový soubor o 3 proměnných a 41 případech. Po spuštění programu intsp1 se do proměnné v1 uloží rozsahy výběrů 5, 7, ..., 85, do v2 (resp. v3) dolní (resp. horní) meze 95% intervalů spolehlivosti pro μ. Vytvoření grafu: Grafy – Bodové grafy – Typ grafu Vícenásobný – vypneme Lineární proložení – Proměnné X v1, Y v2, v3 OK, OK.



Vidíme, že šířka intervalu spolehlivosti klesá se zvětšujícím se rozsahem náhodného výběru, zprvu rychle a pak stále pomaleji.

Úkol 4.: Sledování vlivu rizika na šířku intervalu spolehlivosti (při konstatním rozsahu výběru)

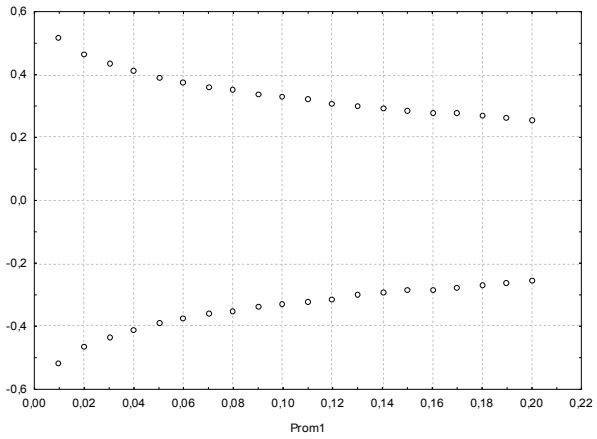
Pro hypotetický náhodný výběr rozsahu n=25 z rozložení N(0,1), jehož výběrový průměr se realizoval hodnotou 0, vypočtěte dolní a horní meze 100(1-α)% intervalů spolehlivosti (α=0,20, 0,19, …, 0,01) pro μ a graficky znázorněte závislost těchto mezí na riziku α.

Návod: Z Učebních materiálů stáhněte program intsp2.svb a otevřete ho v programovacím okně.

```
Option Base 1
```

```
Dim s As Spreadsheet
Sub Main
        alfa = 0.21
        'počáteční hodnota rizika
        m = 0
        'pevně zvolený průměr
        sigma = 1
        'pevně zvolená směrodatná odchylka
        n = 25
        'pevně zvolený rozsah výběru
        Set s = ActiveSpreadsheet
        For I = 1 To 20
                s.Cells(I, 2) = m - VNormal(1 - (alfa - I / 100) / 2, 0, 1) / Sqrt(n)
                'dolní mez intervalu spolehlivosti
                s.Cells(I, 3) = m + VNormal(1 - (alfa - I / 100) / 2, 0, 1) / Sqrt(n)
                'horní mez intervalu spolehlivosti
                s.Cells(I, 1) = alfa - I / 100
                'zmenšení rizika o 1/100
        Next I
End Sub
```

Vytvořte nový datový soubor o 3 proměnných a 20 případech. Po spuštění programu intsp2 se do proměnné v1 uloží rizika 0,20, 0,19, ..., 0,01, do v2 (resp. v3) dolní (resp. horní) meze $100(1-\alpha)\%$ intervalů spolehlivosti pro μ. Vytvoření grafu: stejným způsobem jako v předešlém případě.



Vidíme, že šířka intervalu spolehlivosti s rostoucím rizikem klesá.

Cvičení 4.: Základní pojmy matematické statistiky II, testy normality

Úkol: Měřením délky deseti válečků byly získány hodnoty (v mm): 5,38 5,36 5,35 5,40 5,41 5,34 5,29 5,43 5,42 5,32. Těchto deset hodnot považujeme za realizace náhodného výběru rozsahu 10 z normálního rozložení s neznámou střední hodnotou μ a známou směrodatnou odchylkou $\sigma = 0{,}04$.
Na hladině významnosti 0,1 testujte nulovou hypotézu, že střední hodnota délky válečků je 5,35 mm. Proti nulové hypotéze postavte
a) oboustrannou alternativu
b) levostrannou alternativu
c) pravostrannou alternativu.
Test proveďte pomocí
a) kritického oboru
b) intervalu spolehlivosti
c) p-hodnoty.
Systém STATISTICA použijte jako inteligentní kalkulačku.

Návod:
Formulace nulové hypotézy: H_0: $\mu = 5{,}35$, formulace alternativní hypotézy:
ad a) H_1: $\mu \neq 5{,}35$, ad b) H_1: $\mu < 5{,}35$, ad c) H_1: $\mu > 5{,}35$
Jedná se o jednovýběrový z-test.
Testová statistika $T_0 = \dfrac{M - c}{\dfrac{\sigma}{\sqrt{n}}}$ bude mít rozložení N(0, 1), pokud je nulová hypotéza pravdivá.

Provedení testu:

Ad a) Pomocí kritického oboru
Kritický obor pro oboustrannou alternativu:
$W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$.
Kritický obor pro levostrannou alternativu:
$W = \left\langle -\infty, -u_{1-\alpha}\right)$.
Kritický obor pro pravostrannou alternativu:
$W = \left\langle u_{1-\alpha}, \infty\right)$.
Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α.

Ad b) Pomocí intervalu spolehlivosti
Oboustranný interval spolehlivosti pro μ při známém σ:
$(d, h) = (m - \dfrac{\sigma}{\sqrt{n}}\, u_{1-\alpha/2},\ m + \dfrac{\sigma}{\sqrt{n}}\, u_{1-\alpha/2})$.
Pravostranný interval spolehlivosti pro μ při známém σ:
$(-\infty, h) = (-\infty,\ m + \dfrac{\sigma}{\sqrt{n}}\, u_{1-\alpha})$.
Levostranný interval spolehlivosti pro μ při známém σ:
$(d, \infty) = (m - \dfrac{\sigma}{\sqrt{n}}\, u_{1-\alpha},\ \infty)$.

Pokud číslo c (v našem případě 5,35) nepatří do 100(1-α)% intervalu spolehlivosti pro μ, H_0 zamítáme na hladině významnosti α.

Ad c) Pomocí p-hodnoty
Vzhledem k tomu, že testová statistika T_0 je spojitá náhodná veličina, můžeme použít úpravu $P(T_0 \geq t_0) = P(T_0 > t_0) = 1 - \Phi(t_0)$.
Vzorec pro výpočet p-hodnoty pro oboustrannou alternativu:
$p = 2 \min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\} = 2 \min\{\Phi(t_0), 1 - \Phi(t_0)\}$.
Vzorec pro výpočet p-hodnoty pro levostrannou alternativu:
$p = P(T_0 \leq t_0) = \Phi(t_0)$.
Vzorec pro výpočet p-hodnoty pro pravostrannou alternativu:
$p = P(T_0 \geq t_0) = 1 - \Phi(t_0)$.
Pokud $p \leq \alpha$, H_0 zamítáme na hladině významnosti α.

Zjištěné hodnoty zapíšeme do nového datového souboru o 10 případech a jedné proměnné, kterou nazveme X.
Pomocí Popisných statistik spočteme realizaci výběrového průměru: m = 5,37.
Pro pomocné výpočty otevřeme nový datový soubor o jednom případu a osmi proměnných, které nazveme t0, p1, p2, p3, kv1, kv2, d, h, d1, h2, p.
Do proměnné t_0 uložíme realizaci testové statistiky. Do jejího Dlouhého jména napíšeme vzorec pro výpočet testové statistiky:
= (5,37-5,35)/(0,04/sqrt(10)).
Zjistíme, že $t_0 = 1,5811$.

Nyní již můžeme provést test pomocí p-hodnoty.
Do Dlouhého jména proměnné p1 napíšeme vzorec pro výpočet p-hodnoty pro oboustrannou alternativu:
=2*min(INormal(t0;0;1);1-INormal(t0;0;1))
Vypočtená p-hodnota je 0,1138, což je větší než hladina významnosti 0,1 a nulovou hypotézu nelze na této hladině významnosti zamítnout ve prospěch oboustranné alternativy.
Do Dlouhého jména proměnné p2 napíšeme vzorec pro výpočet p-hodnoty pro levostrannou alternativu:
=INormal(t0;0;1)
I tato p-hodnota (0,9431) je větší než 0,1, což znamená, že nulovou hypotézu nelze na hladině významnosti 0,1 zamítnout ve prospěch levostranné alternativy.
Do Dlouhého jména proměnné p3 napíšeme vzorec pro výpočet p-hodnoty pro pravostrannou alternativu:
=1-INormal(t0;0;1)
Vyjde nám 0,0569, tedy na hladině významnosti 0,1 zamítáme nulovou hypotézu ve prospěch pravostranné alternativy. S rizikem omylu nejvýše 10 % jsme prokázali, že střední hodnota délky válečků je větší než 5,35 mm.

Dále provedeme test pomocí kritického oboru, nejprve pro oboustrannou alternativu.
Do proměnné kv1 uložíme kvantil $u_{1-\alpha/2} = u_{0,95}$:
= VNormal(0,95;0;1).
Vyjde nám 1,6449.
Kritický obor pro oboustrannou alternativu je tedy $W = (-\infty, -1,6449\rangle \cup \langle 1,6449, \infty)$.
Vidíme, že testová statistika nepatří do W, což znamená, že H_0 nezamítáme na hladině významnosti 0,1 ve prospěch oboustranné alternativy.

Pro testování nulové hypotézy proti jednostranným alternativám musíme znát kvantil $u_{1-\alpha} = u_{0,9}$. Uložíme ho do proměnné kv2:
= VNormal(0,9;0;1).
Vyjde nám 1,2816.
Kritický obor pro levostrannou alternativu je tedy $W = \langle -\infty, -1{,}2816)$.
Vidíme, že testová statistika 1,5811 nepatří do W, což znamená, že H_0 nezamítáme na hladině významnosti 0,1 ve prospěch levostranné alternativy.
Kritický obor pro pravostrannou alternativu je tedy $W = \langle 1{,}2816, \infty)$
Vidíme, že testová statistika1,5811 patří do W, což znamená, že H_0 zamítáme na hladině významnosti 0,1 ve prospěch pravostranné alternativy.

Nakonec provedeme test pomocí intervalu spolehlivosti.
Pro oboustrannou alternativu:
Do Dlouhého jména proměnné d (resp. h) napíšeme vzorec pro dolní (resp. horní) mez oboustranného 90% intervalu spolehlivosti pro μ při známém σ:
=5,37-0,04*kv1/sqrt(10) (resp. =5,37+0,04*kv1/sqrt(10))
Zjistíme, že číslo c = 5,35 patří do intervalu (5,3492; 5,3908), tedy H_0 nezamítáme na hladině významnosti 0,1 ve prospěch oboustranné alternativy.
Pro levostrannou alternativu:
Do Dlouhého jména proměnné h2 napíšeme vzorec pro horní mez pravostranného 90% intervalu spolehlivosti pro μ při známém σ:
=5,37+0,04*kv2/sqrt(10)
Protože 5,35 patří do intervalu (-∞; 5,3862), H_0 nezamítáme na hladině významnosti 0,1 ve prospěch levostranné alternativy.
Pro pravostrannou alternativu:
Do Dlouhého jména proměnné d2 napíšeme vzorec pro dolní mez levostranného 90% intervalu spolehlivosti pro μ při známém σ:
=5,37-0,04*kv2/sqrt(10)
Protože 5,35 nepatří do intervalu (5,3538; ∞), H_0 zamítáme na hladině významnosti 0,1 ve prospěch pravostranné alternativy.

Kolmogorovův – Smirnovův test normality dat

Testujeme nulovou hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z normálního rozložení s parametry μ a σ^2. Distribuční funkci tohoto rozložení označme $\Phi_T(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika $D_n = \sup_{-\infty<x<\infty} |F_n(x) - \Phi_T(x)|$.

Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota.

V případě, že neznáme parametry μ a σ^2 normálního rozložení (což je nejčastější případ), změní se rozložení testové statistiky D_n. V takovém případě jde o Lilieforsovu modifikaci Kolmogorovova – Smirnovova testu. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

Poznámka ke K-S testu ve STATISTICE

Test normality poskytuje hodnotu testové statistiky (ozn. max D) a dvě p-hodnoty. (p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ podporují nulovou hypotézu, je-li pravdivá. P-hodnotu porovnáváme s námi zvolenou hladinou významnosti α. Jestliže p-hodnota $\leq \alpha$, pak H_0 zamítáme na hladině významnosti α, je-li p-hodnota $> \alpha$, pak H_0 nezamítáme na hladině významnosti α.) První p-hodnota se vztahuje k případu, kdy střední hodnotu μ a rozptyl σ^2 známe předem, druhá (ozn. Lilieforsovo p) se vztahuje k případu, kdy μ a σ^2 neznáme. Objeví-li se ve výstupu p = n.s. (tj. non significant), pak hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Shapirův – Wilkův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení $N(\mu, \sigma^2)$. Test je založen na zjištění, zda body v Q-Q grafu jsou významně odlišné od regresní přímky proložené těmito body.

(S-W test se používá především pro výběry menších rozsahů, n < 50, ale nyní již existuje modifikace pro velká n. V systému STATISTICA je implementováno rozšíření na n kolem 5000.)

Úkol 1. : U 45 studentek VŠE v Praze byla zjišťována výška a obor studia (1 – národní hospodářství, 2 – informatika). Hodnoty jsou uloženy v souboru vyska.sta. Pomocí Lilieforsovy modifikace K-S testu, pomocí S-W testu a pomocí testu dobré shody testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí N-P grafu posuďte vizuálně předpoklad normality.

Návod:

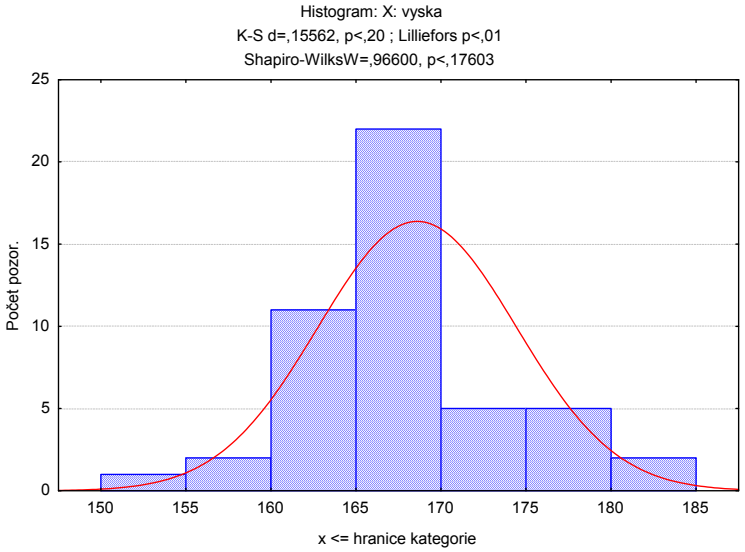
1. způsob provedení Lilieforsova a S-W testu: Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Proměnné X – OK – Normalita – zaškrtneme Lilieforsův test a S-W test – Testy normality.

	Testy normality (vyska.sta)				
Proměnná	N	max D	Lilliefors p	W	p
X: vyska	48	0,155621	p < ,01	0,965996	0,176031

Výstupní tabulka obsahuje počet pozorování, hodnotu testové statistiky Lilieforsovy modifikace K-S testu (max D = 0,155621), p-hodnotu (p < 0,01), testovou statistiku S-W testu (W = 0,965996) a odpovídající p-hodnotu (p = 0,176031). Vidíme, že Lilieforsův test zamítá hypotézu o normalitě na hladině významnosti 0,05, zatímco S-W test nikoli.

2. způsob provedení Lilieforsova a S-W testu: Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Normalita – zaškrtneme K-S test & Lilieforsův test a S-W test – Tabulky četností (nebo Histogram).

Kategorie	Tabulka četností:X: vyska (vyska.sta) K-S d=,15562, p<,20 ; Lilliefors p<,01 Shapiro-WilksW=,96600, p<,17603					
	Četnost	Kumulativní četnost	Rel.četn. (platných)	Kumul. % (platných)	Rel.četn. všech	Kumul. % všech
150,0000<x<=155,0000	1	1	2,08333	2,0833	2,08333	2,0833
155,0000<x<=160,0000	2	3	4,16667	6,2500	4,16667	6,2500
160,0000<x<=165,0000	11	14	22,91667	29,1667	22,91667	29,1667
165,0000<x<=170,0000	22	36	45,83333	75,0000	45,83333	75,0000
170,0000<x<=175,0000	5	41	10,41667	85,4167	10,41667	85,4167
175,0000<x<=180,0000	5	46	10,41667	95,8333	10,41667	95,8333
180,0000<x<=185,0000	2	48	4,16667	100,0000	4,16667	100,0000
ChD	0	48	0,00000		0,00000	100,0000



V tomto případě dostaneme v záhlaví tabulky či histogramu stejné informace jako pomocí předešlého způsobu.

Samostatný úkol: Testy normality a grafické ověření normality proveďte jak pro výšky studentek oboru národní hospodářství, tak pro výška studentek oboru informatiky.

Pro kontrolu:

Výsledky pro obor národní hospodářství:

Proměnná	Testy normality (vyska.sta) Zhrnout podmínku: z=1				
	N	max D	Lilliefors p	W	p
X: vyska	28	0,167473	p < ,05	0,970969	0,606793

Vidíme, že Lilieforsova varianta K-S testu zamítá hypotézu o normalitě na hladině významnosti 0,05 (p-hodnota je menší než 0,05), zatímco S-W test hypotézu o normalitě nezamítá (p-hodnota je větší než 0,05).

Výsledky pro obor informatika:

Proměnná	Testy normality (vyska.sta) Zhrnout podmínku: z=2				
	N	max D	Lilliefors p	W	p
X: vyska	20	0,172301	p < ,15	0,922747	0,111924

V tomto případě ani jeden z testů hypotézu o normalitě nezamítá na hladině významnosti 0,05.

Upozornění: V archivu závěrečných prací https://is.muni.cz/auth/th/77721/prif_m/ je uložena diplomová práce Dominika Grůzy „Ověřování normality“.
Zkoumání vlastností testů normality pomocí simulací je popsáno v bakalářské práci Marka Haičmana https://is.muni.cz/auth/th/150689/prif_b_a2/

Cvičení 5.: Parametrické úlohy o jednom náhodném výběru z normálního rozložení

Úkol 1.: Vlastnosti výběrového průměru z normálního rozložení

Předpokládejme, že velký ročník na vysoké škole má výsledky ze statistiky normálně rozloženy kolem střední hodnoty 72 bodů se směrodatnou odchylkou 9 bodů. Najděte pravděpodobnost, že průměr výsledků náhodného výběru 10 studentů bude větší než 80 bodů.

Návod:

$X_1, ..., X_{10}$ je náhodný výběr z N(72, 81). Počítáme P(M > 80), přičemž výběrový průměr M má normální rozložení se střední hodnotou E(M) = μ = 72 a rozptylem $D(M) = \frac{\sigma^2}{n} = \frac{81}{10} =$ 8,1 (viz skripta Základní statistické metody, věta 6.1.1.1., bod 2).

Tedy P(M > 80) = 1 - P(M ≤ 80) = 1 – Φ(80), kde Φ(80) je hodnota distribuční funkce rozložení N(72; 8,1) v bodě 80.

Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =1 – INormal(80;72;sqrt(8,1)). Zjistíme, že 1 - Φ(80) = 0,00247005. Funkce INormal(x;μ;σ) počítá hodnotu distribuční funkce rozložení $N(\mu,\sigma^2)$ v bodě x.

	1 Prom1
1	0,00247

Úkol 2.: Intervaly spolehlivosti pro parametry μ, σ^2 normálního rozložení

Z populace stejně starých selat téhož plemene bylo vylosováno šest selat a po dobu půl roku jim byla podávána táž výkrmná dieta. Byly zaznamenávány průměrné denní přírůstky hmotnosti v Dg. Z dřívějších pokusů je známo, že v populaci mívají takové přírůstky normální rozložení, avšak střední hodnota i rozptyl se měnívají. Přírůstky v Dg: 62, 54, 55, 60, 53, 58.

a) Najděte 95% empirický levostranný interval spolehlivosti pro neznámou střední hodnotu μ při neznámé směrodatné odchylce σ.
b) Najděte 95% empirický interval spolehlivosti pro směrodatnou odchylku σ.

Návod:

Vytvoříme nový datový soubor o jedné proměnné X a 6 případech. Do proměnné X napíšeme dané hodnoty.

Ad a) Statistiky – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. (ostatní volby zrušíme) – pro jednostranný interval změníme hodnotu na 90,00 - Výpočet. (Hodnotu změníme na 90, protože dolní mez levostranného 95% intervalu spolehlivosti pro μ je stejná jako dolní mez oboustranného 95% intervalu spolehlivosti pro μ.)

	Popisné statistiky (Tabulka1)	
Proměnná	Int. spolehl. -90,000%	Int. spolehl. 90,000
X	54,05683	59,94317

Vidíme, že μ > 54,06 Dg s pravděpodobností aspoň 0,95.

Ad b) Statistiky – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze sp. směr. odch., ponecháme implicitní hodnotu 95,00 – Výpočet.

Proměnná	Popisné statistiky (Tabulka1)	
	Spolehlivost Sm.Odch. -95,000%	Spolehlivost Sm.Odch. +95,000%
X	2,233234	8,774739

Dostáváme výsledek: 2,23 g < σ < 8,77 g s pravděpodobností aspoň 0,95.

Úkol 3.: Testování hypotézy o parametru μ normálního rozložení
Systematická chyba měřicího přístroje se eliminuje nastavením přístroje a měřením etalonu, jehož správná hodnota je μ = 10,00. Nezávislými měřeními za stejných podmínek byly získány hodnoty: 10,24 10,12 9,91 10,19 9,78 10,14 9,86 10,17 10,05, které považujeme za realizace náhodného výběru rozsahu 9 z rozložení $N(\mu, \sigma^2)$. Je možné při riziku 0,05 vysvětlit odchylky od hodnoty 10,00 působením náhodných vlivů?
Návod:
Na hladině významnosti 0,05 testujeme hypotézu H_0: μ = 10 proti oboustranné alternativě H_1: μ ≠ 10. Jde o úlohu na jednovýběrový t-test. Ten je ve STATISTICE implementován. Vytvoříme datový soubor o jedné proměnné a devíti případech, kam zapíšeme naměřené hodnoty. V Základních statistikách/tabulkách vybereme t-test, samostatný vzorek. Do Referenčních hodnot zapíšeme 10. Ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu. Pokud p-hodnota bude menší nebo rovna 0,05, zamítneme hypotézu H_0: μ = 10 ve prospěch oboustranné alternativní hypotézy H_1: μ ≠ 10 na hladině významnosti 0,05.
V opačném případě H_0 nezamítáme. V našem případě je

Proměnná	Test průměrů vůči referenční konstantě (hodnotě)							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Prom1	10,05111	0,162669	9	0,054223	10,00000	0,942611	8	0,373470

Protože p-hodnota 0,373470 > 0,05 nulovou hypotézu nezamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% lze tedy odchylky od hodnoty 10 vysvětlit působením náhodných vlivů.
Všimněme si ještě hodnoty testového kriteria: $t_0 = 0{,}942611$. Kritický obor

$$W = (-\infty, -t_{1-\alpha/2}(n-1)\rangle \cup \langle t_{1-\alpha/2}(n-1), \infty) = (-\infty, -t_{0,975}(8)\rangle \cup \langle t_{0,975}(8), \infty) =$$
$$= (-\infty, -2{,}306\rangle \cup \langle 2{,}306, \infty)$$

Protože $t_0 \notin W$, nezamítáme na hladině významnosti 0,05 hypotézu H_0 .

Úkol 4.: Interval spolehlivosti pro rozdíl parametrů μ_1 - μ_2 dvourozměrného rozložení
Bylo vylosováno 6 vrhů selat a z nich vždy dva sourozenci. Jeden z nich vždy dostal náhodně dietu č. 1 a druhý dietu č. 2. Přírůstky v Dg jsou následující: (62,52), (54,56), (55,49), (60,50), (53,51), (58,50). Za předpokladu, že rozdíly uvedených dvojic tvoří náhodný výběr z normálního rozložení se střední hodnotou μ_1 - μ_2, sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot.
Návod:
Vytvoříme datový soubor o třech proměnných a šesti případech. Do proměnných v1 a v2 zapíšeme naměřené přírůstky, do proměnné v3 uložíme rozdíly v1 - v2.
Ve STATISTICE je implementován výpočet oboustranného intervalu spolehlivosti pro μ, když σ^2 neznáme. Pomocí Popisných statistik zjistíme meze 95% intervalu spolehlivosti pro střední hodnotu proměnné v3 tak, že zaškrtneme Meze spoleh. prům.

	Popisné statistiky	
Proměnná	Int. spolehl. -95,000%	Int. spolehl. +95,000%
Prom3	0,626461	10,70687

Dostaneme výsledek: 0,63 Dg < μ < 10,71 Dg s pravděpodobností aspoň 0,95.

Úkol 5.: Testování hypotézy o rozdíl parametrů μ_1 - μ_2 dvourozměrného rozložení
Pro data z úkolu 4. testujte na hladině významnosti 0,05 hypotézu, že obě výkrmné diety mají stejný vliv.
Návod:
Označme $\mu = \mu_1 - \mu_2$. Na hladině významnosti 0,05 testujeme hypotézu H_0: $\mu = 0$ proti oboustranné alternativě H_1: $\mu \neq 0$. Jde o úlohu na párový t-test. Ten je ve STATISTICE implementován.Vytvoříme datový soubor o dvou proměnných a šesti případech. Do proměnných v1 a v2 zapíšeme naměřené přírůstky. V menu Základní statistiky/tabulky vybereme t-test, závislé vzorky. Zadáme názvy obou proměnných a ve výstupu se podíváme na hodnotu testového kritéria a na p-hodnotu.

	t-test pro závislé vzorky Označ. rozdíly jsou významné na hlad. p < ,05000							
Proměnná	Průměr	Sm.odch.	N	Rozdíl	Sm.odch. rozdílu	t	sv	p
Prom1	57,00000	3,577709						
Prom2	51,33333	2,503331	6	5,666667	4,802777	2,890087	5	0,034183

Protože p-hodnota 0,034183 < 0,05, zamítáme hypotézu H_0: $\mu = 0$ ve prospěch alternativní hypotézy H_1: $\mu \neq 0$ na hladině významnosti 0,05. Znamená to, že jsme s rizikem omylu nejvýše 5% prokázali rozdíl v účinnosti obou výkrmných diet.
Všimněme si ještě hodnoty testového kriteria: $t_0 = 2{,}890087$. Kritický obor

$$W = (-\infty, -t_{1-\alpha/2}(n-1)\rangle \cup \langle t_{1-\alpha/2}(n-1), \infty) = (-\infty, -t_{0,975}(5)\rangle \cup \langle t_{0,975}(5), \infty) =$$

$$= (-\infty, -2{,}5706\rangle \cup \langle 2{,}5706, \infty)$$

Protože $t_0 \in W$, zamítáme na hladině významnosti 0,05 hypotézu H_0.

Příklady k samostatnému řešení

Příklad 1.: Měřením délky deseti válečků byly získány hodnoty (v mm): 5,38 5,36 5,35 5,40 5,41 5,34 5,29 5,43 5,42 5,32. Těchto deset hodnot považujeme za realizace náhodného výběru rozsahu 10 z normálního rozložení $N(\mu, \sigma^2)$.

a) Sestrojte 99% interval spolehlivosti pro neznámou střední hodnotu μ
b) Sestrojte 99% interval spolehlivosti pro neznámou směrodatnou odchylku σ.
c) Na hladině významnosti 0,01 testujte hypotézu, že střední hodnota délky válečků je 5,3 mm proti oboustranné alternativě.

Výsledky:

ad a)

5,3228 mm < μ < 5,4172 mm s pravděpodobností aspoň 0,99

ad b)

0,0284 mm < σ < 0,1046 mm s pravděpodobností aspoň 0,99.

ad c) Testujeme H_0: $\mu = 5{,}3$ proti H_1: $\mu \neq 5{,}3$ na hladině významnosti 0,01. Nulovou hypotézu zamítáme na hladině významnosti 0,01 a přijímáme alternativní hypotézu.

Příklad 2.: Bylo náhodně vybráno 15 desetiletých chlapců a byla zjištěna jejich výška (v cm). Výsledky měření 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147 považujeme za realizace náhodného výběru rozsahu 15 z rozložení $N(\mu,\sigma^2)$. Podle názoru odborníků by střední hodnoty výšky desetiletých chlapců měla být 136,1 cm. Testujte tuto hypotézu na hladině významnosti 0,05.

Pomocí N-P plotu a S-W testu ověřte normalitu dat.

Výsledky:

S-W test poskytl p-hodnotu 0,7998, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05. Dále testujeme H_0: $\mu = 136{,}1$ proti H_1: $\mu \neq 136{,}1$ na hladině významnosti 0,05. Protože $p = 0{,}0947 > 0{,}05$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Příklad 3.: Pět mužů se rozhodlo, že budou hubnout. Zjistili svou hmotnost před zahájením diety a po ukončení diety.

Číslo osoby	1	2	3	4	5
Hmotnost před dietou	84	77,5	91,5	84,5	97,5
Hmotnost po dietě	78,5	73,5	88,5	80	97

Na hladině významnosti 0,05 testujte hypotézu, že dieta neměla vliv na hmotnost.

Výsledky:

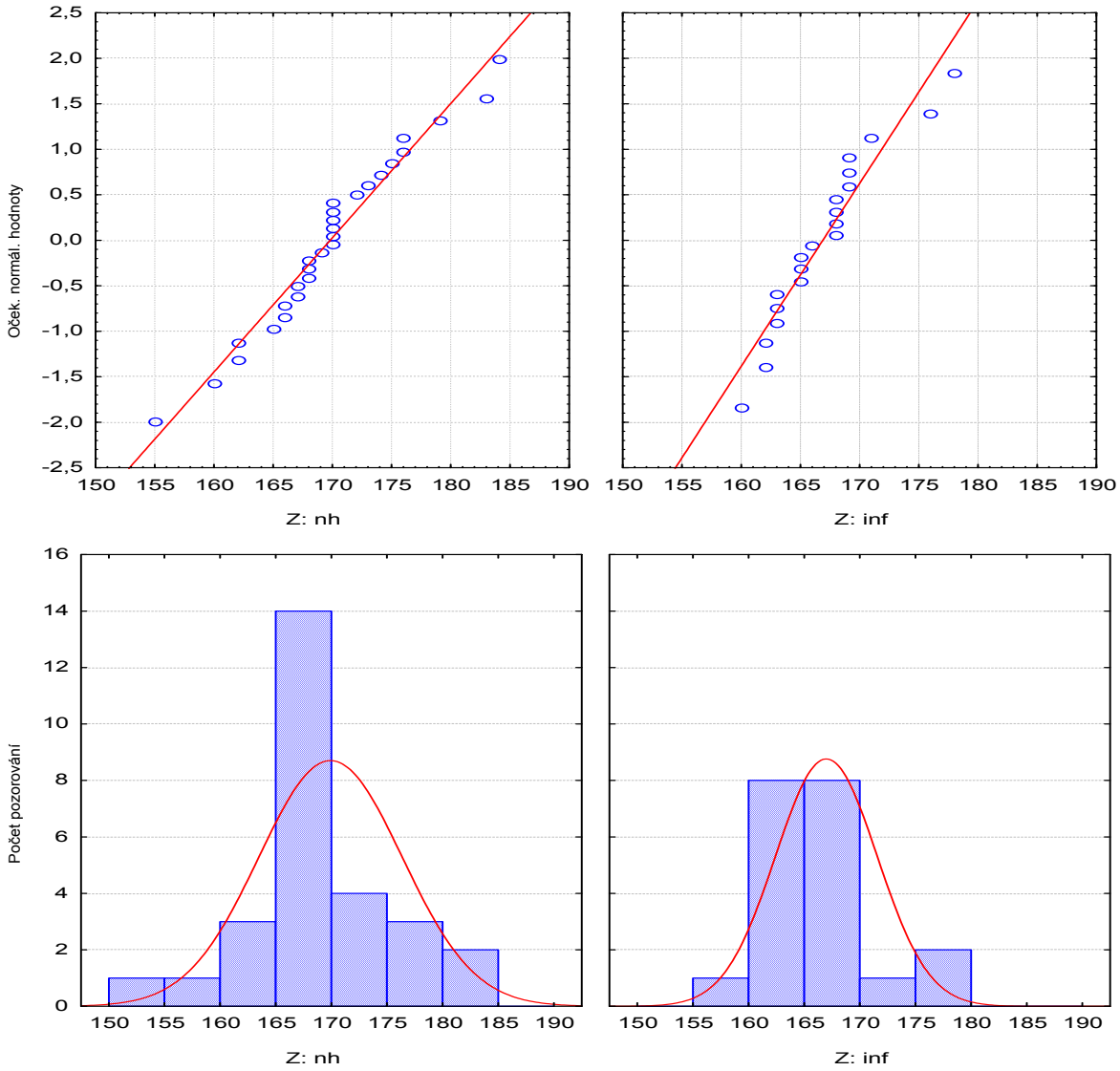
Testujeme H_0: $\mu_1 - \mu_2 = 0$ proti H_1: $\mu_1 - \mu_2 \neq 0$. Testová statistika nabývá hodnoty 4,1105, odpovídající p-hodnota je 0,0147, tedy nulovou hypotézu zamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že dieta má vliv na střední hodnotu hmotnosti.

Cvičení 6.: Parametrické úlohy
o dvou nezávislých výběrech z normálních rozložení

Úkol 1.: Do programu STATISTICA načtěte soubor studentky.sta, který obsahuje údaje o 48 náhodně vybraných studentkách VŠE v Praze:
1. sloupec – výška, 2. sloupec – známka z matematiky v 1. semestru, 3. sloupec – obor studia (1 – národní hospodářství, 2 – informatika).

Úkol 2.: Orientačně ověřte normalitu výšky ve skupině studentek oboru národní hospodářství a oboru informatika vykreslením N-P plotu a histogramu.

Návod:
Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné X – na záložce Kategorizovaný zaškrtneme Kategorie X Zapnuto – Změnit proměnnou – Z - OK – OK.
Podobně pro histogram.



Komentář: Grafy svědčí o mírném narušení normality, jedná se o mírné kladné zešikmení.

Nyní provedeme testy normality.
Statistiky – Základní statistiky/tabulky – Tabulky četností – OK – Select cases – Zapnout filtr – některé vybrané pomocí Z=1 – OK – Proměnná X – OK - Normalita - zaškrtneme Liliefors test, Shapiro-Wilk's test - Testy normality. Dostaneme tyto výsledky:

Pro studentky oboru nh

Proměnná	Testy normality (studentky.sta) Zhrnout podmínku: Z=1				
	N	max D	Lilliefors p	W	p
X: vyska	28	0,167473	p < ,05	0,970969	0,606793

Pro studentky oboru inf

Proměnná	Testy normality (studentky.sta) Zhrnout podmínku: Z=2				
	N	max D	Lilliefors p	W	p
X: vyska	20	0,172301	p < ,15	0,922747	0,111924

Komentář: Vypočtenou p-hodnotu porovnáváme se zvolenou hladinou významnosti testu (většinou volíme $\alpha = 0{,}05$). Je-li vypočtená p-hodnota $\leq \alpha$, pak hypotézu o normalitě zamítáme na hladině významnosti α. V našem případě dojde k zamítnutí hypotézy o normalitě výšky na hladině významnosti 0,05 pouze u Lilieforsova testu pro studentky oboru nh.

Úkol 3.: Sestrojte 95% empirický interval spolehlivosti pro střední hodnotu výšky
a) studentek oboru nh,
b) studentek oboru inf.

Návod:
Vzhledem k tomu, že data lze považovat za realizace náhodného výběru z normálního rozložení, můžeme použít postup pro konstrukci intervalu spolehlivosti pro střední hodnotu, když rozptyl neznáme. Výpočet je implementován ve STATISTICE. Meze 95% intervalu spolehlivosti pro střední hodnotu proměnné X zjistíme pomocí Popisných statistik, kde zaškrtneme Meze spoleh. prům.

Proměnná	Popisné statistiky (studentky.sta) Zhrnout podmínku: Z=1	
	Int. spolehl. -95,000%	Int. spolehl. 95,000
X	167,3328	172,3100

Proměnná	Popisné statistiky (studentky.sta) Zhrnout podmínku: Z=2	
	Int. spolehl. -95,000%	Int. spolehl. 95,000
X	164,7693	169,0307

Komentář: S pravděpodobností aspoň 95% lze očekávat, že střední hodnoty výška studentek oboru národní hospodářství leží v intervalu 167,3 cm až 172,3 cm, zatímco u studentek oboru informatika v intervalu 164,8 cm až 169 cm.

Úkol 4.: Sestrojte 95% interval spolehlivosti pro podíl rozptylů výšek studentek oboru nh a inf.

Návod:
K datovému souboru přidáme další dvě proměnné DM a HM pro výpočet dolní a horní meze intervalu spolehlivosti. Do Dlouhého jména těchto proměnných zapíšeme vzorce pro dolní a

horní mez intervalu spolehlivosti pro podíl rozptylů (viz skripta Základní statistické metody, Věta 7.1.2.1., bod 4 (a)). Výběrové rozptyly pro 1. a 2. výběr zjistíme pomocí Popisných statistik.
Interval spolehlivosti je

$$(d, h) = \left(\frac{s_1^2 / s_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{s_1^2 / s_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)} \right)$$

, přičemž první výběr tvoří studentky nh, druhý výběr studentky inf.

Proměnná	Popisné statistiky (Tema7) Include condition: z=1	
	N platných	Rozptyl
X	28	41,18915

Proměnná	Popisné statistiky (Tema7) Include condition: z=2	
	N platných	Rozptyl
X	20	20,72632

Do Dlouhého jména proměnné DM napíšeme:
=(41,18915/20,72622)/VF(0,975;27;19)
(Funkce VF(x;ný;omega) počítá x-kvantil Fisherova – Snedecorova rozložení F(ný, omega).)
Do Dlouhého jména proměnné HM napíšeme:
=(41,18915/20,72622)/VF(0,025;27;19)
Vyjde DM = 0,821186, HM = 4,513831.
S pravděpodobností aspoň 0,95 tedy platí: $0{,}821 < \sigma_1^2 / \sigma_2^2 < 4{,}514$.

Úkol 5.: Na hladině významnosti 0,05 testujte hypotézu, že rozptyly výšek studentek oboru nh a inf jsou shodné.

Návod:

Jedná se o F-test, kdy testujeme hypotézu $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti oboustranné alternativě

$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

1. způsob: lze využít výsledku 4. úkolu. 95% interval spolehlivosti pro podíl rozptylů obsahuje číslo 1, tedy hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.
2. způsob: F-test je implementován ve STATISTICE.
Statistiky – Základní statistiky/tabulky – t-test, nezávislé, podle skupin - OK, Proměnné – Závislé proměnné X, Grupovací proměnná Z – OK – Výpočet

Proměnná	t-testy; grupováno: Z: obor studia (Tema7) Skup. 1: nh: narodni hospodarstvi Skup. 2: inf: informatika										
	Průměr nh	Průměr inf	t	sv	p	Poč.plat nh	Poč.plat. inf	Sm.odch. nh	Sm.odch. inf	F-poměr rozptyly	p rozptyly
X	169,8214	166,9000	1,744008	46	0,087837	28	20	6,417878	4,552616	1,987288	0,124925

Komentář: Ve výstupní tabulce nás zajímá hodnota testové statistiky F-testu (v našem případě 1,987288) a odpovídající p-hodnota: 0,124925. Protože p-hodnota je větší než hladina významnosti $\alpha = 0{,}05$, nelze na hladině významnosti 0,05 zamítnout nulovou hypotézu.
S rizikem omylu nanejvýš 5% se tedy neprokázalo, že by rozptyly výšek studentek oborů nh a inf byly odlišné.

Úkol 6.: Sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot výšek studentek oboru nh a inf.

Návod:
K datovému souboru přidáme další dvě proměnné DM1 a HM1 pro výpočet dolní a horní meze intervalu spolehlivosti. Do Dlouhého jména těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro rozdíl středních hodnot (viz skripta Základní statistické metody, Věta 7.1.2.1., bod 2 (a)). Výběrové průměry a výběrové rozptyly pro první a druhý výběr zjistíme pomocí Popisných statistik.
Oboustranný interval spolehlivosti pro $\mu_1 - \mu_2$, když rozptyly σ_1^2, σ_2^2 neznáme, ale víme, že jsou shodné, je:

$(d, h) = (m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{1-\alpha/2}(n_1+n_2-2),\ m_1 - m_2 + s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{1-\alpha/2}(n_1+n_2-2))$, kde

$s_*^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ je vážený průměr výběrových rozptylů.

Do Dlouhého jména proměnné DM1 napíšeme
=169,8214-166,9-
sqrt((27*41,18915+19*20,72622)/46)*sqrt((1/28)+(1/20))*VStudent(0,975;46)
Do Dlouhého jména proměnné HM1 napíšeme
=169,8214-166,9+
sqrt((27*41,18915+19*20,72622)/46)*sqrt((1/28)+(1/20))*VStudent(0,975;46)
Vyjde DM1 = -0,450446, HM1 = 6,293246
S pravděpodobností aspoň 0,95 tedy -0,45 cm < $\mu_1 - \mu_2$ < 6,29 cm.

Úkol 7.: Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty výšek studentek oboru nh a inf jsou shodné. Výpočet doplňte krabicovými diagramy.

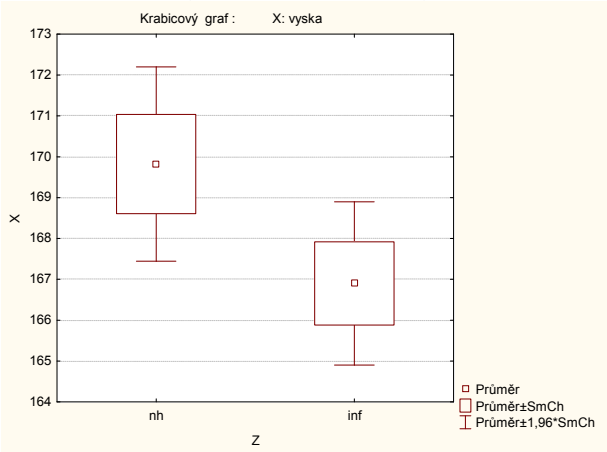
Návod:
Jedná se o dvouvýběrový t-test, kdy testujeme hypotézu $H_0 : \mu_1 - \mu_2 = 0$ proti oboustranné alternativě $H_1 : \mu_1 - \mu_2 \neq 0$

1. způsob: lze využít výsledku 6. úkolu. 95% interval spolehlivosti pro rozdíl středních hodnot obsahuje číslo 0, tedy hypotézu o shodě středních hodnot nezamítáme na hladině významnosti 0,05.
2. způsob: dvouvýběrový t-test je implementován ve STATISTICE.
Statistiky – Základní statistiky/tabulky – t-test, nezávislé, podle skupin - OK, Proměnné – Závislé proměnné X, Grupovací proměnná Z – OK – Výpočet

	t-testy; grupováno: Z: obor studia (Tema7) Skup. 1: nh: narodni hospodarstvi Skup. 2: inf: informatika										
Proměnná	Průměr nh	Průměr inf	t	sv	p	Poč.plat nh	Poč.plat. inf	Sm.odch. nh	Sm.odch. inf	F-poměr rozptyly	p rozptyly
X	169,8214	166,9000	1,744008	46	0,087837	28	20	6,417878	4,552616	1,987288	0,124925

Komentář: Ve výstupní tabulce najdeme hodnotu testového kritéria ($t_0 = 1{,}744006$) a odpovídající p-hodnotu. Protože p-hodnota = 0,087837 je větší než hladina významnosti 0,05, nulovou hypotézu nezamítáme na hladině významnosti 0,05. S rizikem omylu nanejvýš 5% se tedy neprokázal rozdíl mezi středními hodnotami výšek studentek oborů nh a inf.

Konstrukce krabicových diagramů: V tabulce t-test, nezávislé, podle skupin zvolíme Krabicový diagram. Dostaneme graf:



Komentář: Ze vzhledu krabicových diagramů je vidět, že rozložení výšek v obou skupinách je vcelku symetrické kolem průměru, odlehlé ani extrémní hodnoty se nevyskytují, variabilita vyjádřená směrodatnou odchylkou se liší jen nepatrně a průměrná výška ve skupině studentek oboru inf je o něco menší než ve skupině studentek oboru nh.

Poznámka: Protože F-test neprokázal odlišnost rozptylů, mohli jsme ve STATISTICE použít variantu dvouvýběrového t-testu se shodnými rozptyly. Pokud by však F-test zamítl na dané hladině významnosti hypotézu o shodě rozptylů, museli bychom zvolit variantu dvouvýběrového t-testu se separovanými odhady rozptylů.

Úkol k samostatnému řešení: Hejtman Jihomoravského kraje chtěl porovnat situaci svého kraje s ostatními moravskými kraji vzhledem ke znečištění ovzduší oxidem siřičitým, oxidy dusíku a oxidem uhelnatým. Požádal proto Stranu zelených, aby na základě údajů ze Statistické ročenky ČSÚ za léta 2000 až 2006 její experti provedli příslušnou analýzu. Roční měrné emise jsou uvedeny v tunách na km^2. Data jsou uložena v souboru znecisteni.sta. Vaším úkolem bude provést srovnání středních hodnot znečištění oxidem siřičitým v Jihomoravském kraji a Olomouckém kraji. Na hladině významnosti 0,05 ověřte normalitu dat, homogenitu rozptylů a proveďte test shody středních hodnot. Výpočty doplňte krabicovými grafy a rovněž vypočtěte Cohenův koeficient věcného účinku.

Výsledek:

Průměrné znečištění oxidem siřičitým v Jihomoravském kraji v letech 2000 – 2006 je 0,51, v Olomouckém 1,23. Testová statistika pro test shody rozptylů se realizuje hodnotou 1,94117, odpovídající p-hodnota je 0,4397, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě rozptylů.

(Upozornění: v případě zamítnutí hypotézy o shodě rozptylů je zapotřebí v tabulce t-testu pro nezávislé vzorky dle skupin na záložce Možnosti zaškrtnout volbu Test se samostatnými odhady rozptylu.)

Testová statistika pro test shody středních hodnot se realizuje hodnotou -12,247, počet stupňů volnosti je 12, odpovídající p-hodnota je velmi blízká 0, tedy hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05. Znamená to, že s rizikem omylu nejvýše 5% se prokázal rozdíl ve středních hodnotách znečištění oxidem siřičitým v Jihomoravském a Olomouckém kraji.

Cohenův koeficient nabyl hodnoty 6,55, vliv kraje na velikost znečištění oxidem siřičitým je tedy velký. (Výpočet Cohenova koeficientu je možno provést pomocí programu Cohen.svb.)

Cvičení 7.: Parametrické úlohy o jednom výběru a dvou nezávislých výběrech z alternativního rozložení

Úkol 1.: Vlastnosti výběrového průměru z alternativního rozložení
Mezi americkými voliči 60% osob volí republikány a 40% demokraty. Jaká je pravděpodobnost, že v náhodném výběru 100 amerických voličů budou voliči republikánů v menšině? Výpočet proveďte jak přesně, tak pomocí aproximace normálním rozložením.

Návod:
$X_1, ..., X_{100}$ je náhodný výběr z A(0,6), $X_i = 1$, když i-tá osoba volí republikány, $X_i = 0$ jinak, i = 1, ..., 100. Zavedeme statistiku $Y_{100} = X_1 + ... + X_{100}$, $Y_{100} \sim Bi(100; 0,6)$ (viz skripta Teorie pravděpodobnosti a matematická statistika, sbírka příkladů, příklad 8.10.), $E(Y_{100}) = n\vartheta = 100.0,6 = 60$, $D(Y_{100}) = n\vartheta(1-\vartheta) = 100.0,6.0,4 = 24$ Označme $\Phi_{100}(y)$ distribuční funkci náhodné veličiny Y_{100}, $\Phi_{100}(y) = \sum_{t=0}^{y} \binom{100}{t} 0,6^t 0,4^{100-t}$.
Přesný výpočet: $P(Y_{100} < 50) = P(Y_{100} \leq 49) = \Phi_{100}(49) = 0,016761686$.
Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =IBinom(49;0,6;100). Funkce IBinom(x;p;n) počítá hodnotu distribuční funkce rozložení Bi(n,p) v bodě x.
Přibližný výpočet: užijeme důsledek Moivreovy - Laplaceovy integrální věty (viz skripta Základní statistické metody, věta 6.3.1.1.). Nejdříve ověříme splnění podmínky dobré aproximace $n\vartheta(1-\vartheta) = 100.0,6.0,4 = 24 > 9$. Podmínka je splněna.
$P(Y_{100} < 50) = P(Y_{100} \leq 49) \approx \Phi(49)$, kde $\Phi(49)$ je hodnota distribuční funkce rozložení N(60; 24) v bodě 49.
Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =INormal(49;60;sqrt(24)).
Zjistíme, že $\Phi(49) = 0,012372$.

Přesný výpočet

	1 Prom1
1	0,016762

Aproximativní výpočet

	1 Prom1
1	0,012372

Úkol 2.: Asymptotický interval spolehlivosti pro parametr ϑ alternativního rozložení
Může politická strana, pro niž se v předvolebním průzkumu vyslovilo 60 z 1000 dotázaných osob, očekávat se spolehlivostí aspoň 0,95, že by v této době ve volbách překročila 5% hranici pro vstup do parlamentu?

Návod:
Zavedeme náhodné veličiny $X_1, ..., X_{1000}$, přičemž $X_i = 1$, když i-tá osoba se vysloví pro danou politickou stranu a $X_i = 0$ jinak, i = 1, ..., 1000. Tyto náhodné veličiny tvoří náhodný výběr z rozložení $A(\vartheta)$. V tomto případě n = 1000, m = 60/1000 = 0,06, $\alpha = 0,05$, $u_{1-\alpha} = u_{0,95} = 1,645$.
Ověření podmínky $n\vartheta(1-\vartheta) > 9$: parametr ϑ neznáme, musíme ho nahradit výběrovým průměrem. Pak $1000.0,06.0,94 = 56,4 > 9$.
95% levostranný interval spolehlivosti pro ϑ je $\left(m - \sqrt{\frac{m(1-m)}{n}} u_{1-\alpha} \; ; \; \infty\right) = \left(0,06 - \sqrt{\frac{0,06(1-0,06)}{1000}} u_{0,95} \; ; \; \infty\right)$ (viz skripta Základní statistické metody, důsledek 6.3.2.2.)

Postup ve STATISTICE:
1. možnost: Vytvoříme datový soubor o jedné proměnné a o jednom případu. Do Dlouhého jména této proměnné napíšeme =0,06-sqrt(0,06*0,94/1000)*VNormal(0,95;0;1). Vyjde 0,047647.
2. možnost: Statistiky – Analýza síly testu – Odhad intervalu – Jeden podíl, Z, Chí-kvadrát test – OK – Pozorovaný podíl p: 0,06, Velik. Vzorku (N): 1000, Spolehlivost: 0,9 – Vypočítat. Dostaneme 0,0476.
S pravděpodobností přibližně 0,95 tedy $\vartheta > 0{,}047647$. Protože tento interval zahrnuje i hodnoty nižší než 0,05, nelze vyloučit, že strana získá méně než 5% hlasů.

Úkol 3: Testování hypotézy o parametru ϑ alternativního rozložení
Určitá cestovní kancelář organizuje zahraniční zájezdy podle individuálních přání zákazníků. Z několika minulých let ví, že 30% všech takto organizovaných zájezdů má za cíl zemi X. Po zhoršení politických podmínek v této zemi se cestovní kancelář obává, že se zájem o tuto zemi mezi zákazníky sníží. Ze 150 náhodně vybraných zákazníků v tomto roce má 38 za cíl právě zemi X. Potvrzují nejnovější data pokles zájmu o tuto zemi? Volte hladinu významnosti 0,05.

Návod:
Máme náhodný výběr $X_1, ..., X_{150}$ z rozložení A(0,3). Testujeme H_0: $\vartheta = 0{,}3$ proti levostranné alternativě H_1: $\vartheta < 0{,}3$. V tomto případě je testovým kritériem statistika

$$T_0 = \frac{M - c}{\sqrt{\frac{c(1-c)}{n}}}$$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1)

(viz skripta Základní statistické metody, věta 6.3.3.1.). Musíme ověřit splnění podmínky $n\vartheta(1-\vartheta) > 9$: $150.0{,}3.0{,}7 = 31{,}5 > 9$. Vypočteme realizaci testového kritéria:

$$\frac{m - c}{\sqrt{\frac{c(1-c)}{n}}} = \frac{\frac{38}{150} - 0{,}3}{\sqrt{\frac{0{,}3(1-0{,}3)}{150}}} = -1{,}24722$$. Kritický obor: $W = \langle -\infty, -u_{1-\alpha} \rangle = \langle -\infty, -1{,}645 \rangle$.

Protože testové kritérium nepatří do kritického oboru, H_0 nezamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% tedy naše data neprokázala pokles zájmu zákazníků cestovní kanceláře o zemi X.

Postup ve STATISTICE:
Asymptotický způsob: Vytvoříme datový soubor o dvou proměnných (nazveme je t0 a kvantil) a jednom případu. Vypočteme realizaci testového kritéria tak, že do Dlouhého jména proměnné t0 napíšeme
=(38/150-0,3)/sqrt(0,3*0,7/150)
Do Dlouhého jména proměnné kvantil napíšeme
=VNormal(0,95;0;1)
Tím získáme kvantil $u_{0,95}$.

	1 t0	2 kvantil
1	-1,24721913	1,644854

Jelikož realizace testového kritéria $t_0 = -1{,}24721913$ nepatří do kritického oboru $W = \langle -\infty, -1{,}644854 \rangle$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Přibližný způsob: Do nového datového souboru o jedné proměnné X a 150 případech uložíme 38 jedniček (indikují zájem o danou zemi) a 112 nul (indikují nezájem o danou zemi). Statistika – Základní statistiky a tabulky – t-test, samost. vzorek – OK – Proměnné X – OK, Test všech průměrů vůči 0,3 – Výpočet.

	Test průměrů vůči referenční konstantě (hodnotě) (Tabulka4)							
Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
X	0,253333	0,436377	150	0,035630	0,300000	-1,30976	149	0,192294

Hodnota testové statistiky je při tomto přibližném způsobu -1,30976. Odpovídající p-hodnota je 0,1923, ovšem to je p-hodnota pro oboustranný test. Tuto p-hodnotu tedy musíme dělit dvěma a dostaneme 0,0961. Na asymptotické hladině významnosti 0,05 nelze zamítnout hypotézu, že zájem o danou zemi se nezměnil.

Úkol 4.: Asymptotický interval spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$

Při výstupní kontrole bylo náhodně vybráno 150 výrobků vyrobených na ranní směně a rovněž 150 výrobků vyrobených na odpolední směně. U ranní směny bylo zjištěno 16 zmetků a u odpolední 12 zmetků. Sestrojte 95% asymptotického interval spolehlivosti pro rozdíl pravděpodobností vyrobení zmetku v obou směnách.

Návod: Zavedeme náhodnou veličinu X_{1i}, která bude nabývat hodnoty 1, když i-tý výrobek z ranní směny je zmetek, 0 jinak, i = 1, …, 150. Náhodné veličiny $X_{1,1}$, …, $X_{1,150}$ tvoří náhodný výběr z rozložení $A(\vartheta_1)$. Dále zavedeme náhodnou veličinu X_{2i}, která bude nabývat hodnoty 1, když i-tý výrobek z odpolední směny je zmetek, 0 jinak, i = 1, …, 150.. Náhodné veličiny $X_{2,1}$, …, $X_{2,150}$ tvoří náhodný výběr z rozložení $A(\vartheta_2)$.
$n_1 = 150$, $n_2 = 150$, $m_1 = 16/150 = 0{,}1067$, $m_2 = 12/150 = 0{,}08$.
Ověření podmínek $n_1 \vartheta_1 (1-\vartheta_1) > 9$ a $n_2 \vartheta_2 (1-\vartheta_2) > 9$: Parametry ϑ_1 a ϑ_2 neznáme, nahradíme je odhady m_1 a m_2: $16.(1-16/150) = 14{,}29 > 9$, $12.(1-12/150) = 11{,}04 > 9$.
Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$ jsou:

$$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\, u_{1-\alpha/2} =$$

$$= \frac{16}{150} - \frac{12}{150} - \sqrt{\frac{\frac{16}{150}\left(1-\frac{16}{150}\right)}{150} + \frac{\frac{12}{150}\left(1-\frac{12}{150}\right)}{150}}\, 1{,}96 = -0{,}039$$

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\, u_{1-\alpha/2} =$$

$$= \frac{16}{150} - \frac{12}{150} + \sqrt{\frac{\frac{16}{150}\left(1-\frac{16}{150}\right)}{150} + \frac{\frac{12}{150}\left(1-\frac{12}{150}\right)}{150}}\, 1{,}96 = 0{,}092$$

Zjistili jsme tedy, že s pravděpodobností přibližně 0,95: $-0{,}039 < \vartheta_1 - \vartheta_2 < 0{,}092$.

Postup ve STATISTICE:
Otevřeme nový datový soubor se dvěma proměnnými d a h a o jednom případu. Do Dlouhého jména proměnné d napíšeme:
=16/150-12/150-sqrt((16/150)*(134/150)/150+(12/150)*(138/150)/150)*VNormal(0,975;0;1)
Do Dlouhého jména proměnné h napíšeme:

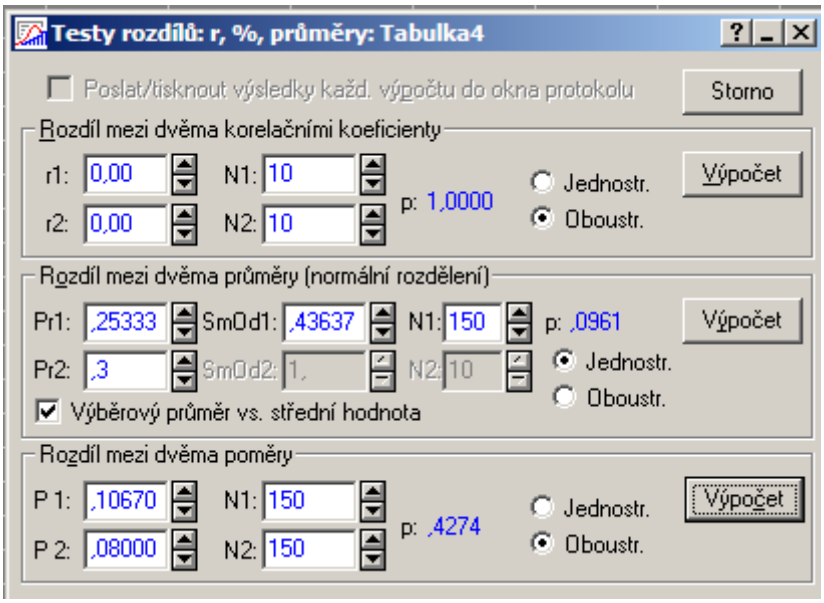
=16/150-
12/150+sqrt((16/150)*(134/150)/150+(12/150)*(138/150)/150)*VNormal(0,975;0;1)
Dostaneme tabulku

	1 d	2 h
1	-0,0391	0,092433

S pravděpodobností přibližně 0,95 se rozdíl pravděpodobností vyrobení zmetku na ranní a odpolední směně nachází v intervalu (-0,039; 0,092).

Úkol 5.: Testování hypotézy o parametrické funkci $\vartheta_1 - \vartheta_2$
Pro údaje z úkolu 4 testujte na asymptotické hladině významnosti 0,05 hypotézu, že pravděpodobnost vyrobení zmetků v obou směnách je táž.

Postup ve STATISTICE:
Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,1067, do políčka N1 napíšeme 150, do políčka P 2 napíšeme 0,08, do políčka N2 napíšeme 150 –Výpočet. Dostaneme p-hodnotu 0,4274, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.



Úkol k samostatnému řešení: Přírůstky cen akcií na burze (v %) u 10 náhodně vybraných společností dosáhly těchto hodnot: 10, 16, 5, 10, 12, 8, 4, 6, 5, 4. Sestrojte 95% asymptotický empirický interval spolehlivosti pro pravděpodobnost, že přírůstek ceny akcie překročí 8,5%.
Výsledek: $0,096 < \vartheta < 0,704$ s pravděpodobností aspoň 0,95.
Znamená to, že pravděpodobnost, že přírůstek ceny akcie překročí 8,5%, je aspoň 9,6% a nanejvýš 70,4% (při spolehlivosti 95%.)

Úkol k samostatnému řešení: Z 28 studentek oboru národní hospodářství mělo z matematiky trojku 17 studentek, zatímco z 20 studentek oboru informatika mělo z matematiky trojku jen 6

studentek. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že pravděpodobnost získání trojky z matematiky je obě skupiny studentek stejná.

Výsledek: Testová statistika se realizuje hodnotou $t_0 = 2{,}100009$, kritický obor je $W = (-\infty;-1{,}96\rangle \cup \langle 1{,}96;\infty)$. Protože $t_0 \in W$, zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

Použijeme-li v systému STATISTICA aplikaci Testy rozdílů, dostaneme p-hodnotu 0,0358, tedy na asymptotické hladině významnosti 0,05 zamítáme nulovou hypotézu.

Cvičení 8: Parametrické úlohy o více nezávislých náhodných výběrech

Úkol 1.: V jisté továrně se měřil čas, který potřeboval každý ze tří dělníků k uskutečnění téhož pracovního úkonu. Čas v minutách:

1. dělník: 3,6 3,8 3,7 3,5
2. dělník: 4,3 3,9 4,2 3,9 4,4 4,7
3. dělník: 4,2 4,5 4,0 4,1 4,5 4,4.

Na hladině významnosti 0,05 testujte hypotézu, že výkony těchto tří dělníků jsou stejné. Zamítnete-li nulovou hypotézu, určete, výkony kterých dělníků se liší na dané hladině významnosti 0,05.

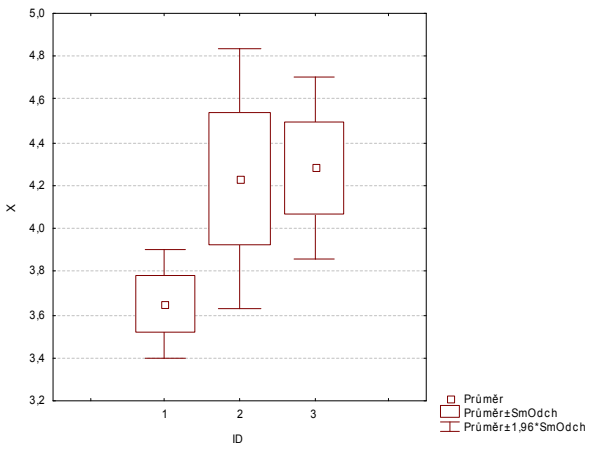
Návod:
Úloha vede na analýzu rozptylu jednoduchého třídění. Postupujeme podle skript Základní statistické metody, odstavec 8.1.
Načteme datový soubor cas_delniku.sta. Proměnná X obsahuje zjištěné časy, proměnná ID nabývá hodnoty 1 pro 1. dělníka, hodnoty 2 pro 2. dělníka a hodnoty 3 pro 3. dělníka.
Statistiky – Základní statistiky/tabulky – Rozklad & jednofakt. ANOVA – Proměnné - Závislé X, Grupovací ID, OK, Kódy pro grupovací proměnné – Vše, OK, Výpočet: Tabulka statistik (zobrazí se průměry, směrodatné odchylky a rozsahy všech tří výběrů).

Rozkladová tabulka popisných statistik (cas_delniku.sta)
N=16 (V seznamu záv. prom. nejsou ChD)

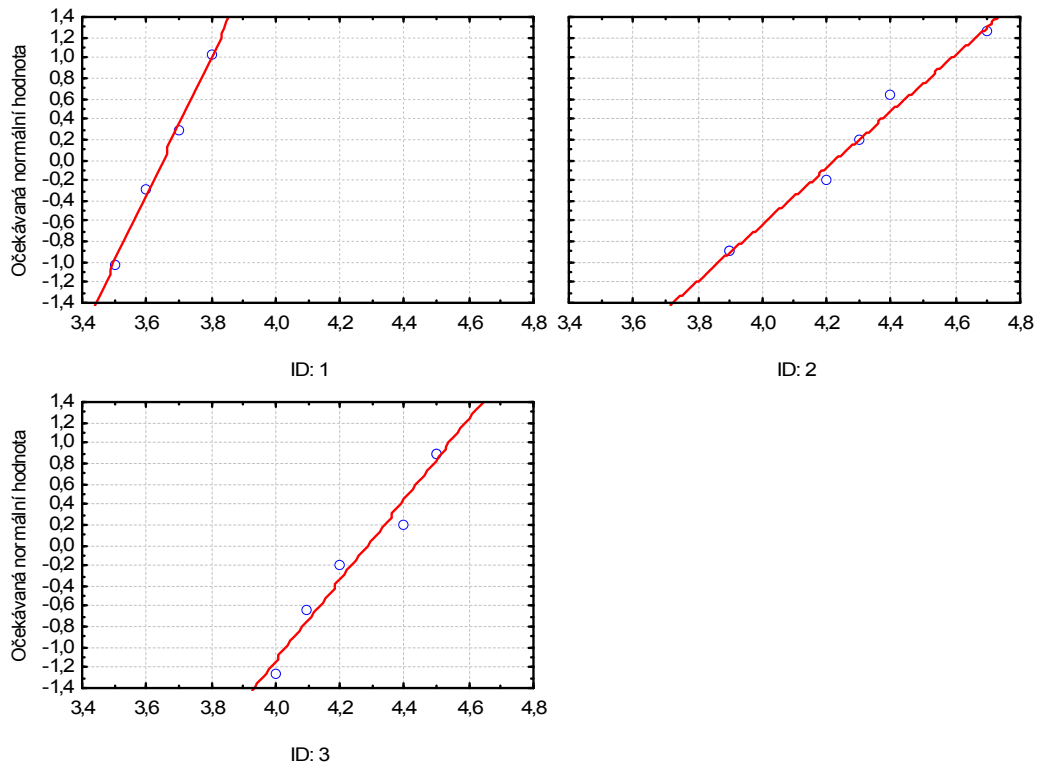
ID	X průměr	X N	X Sm.odch.
1	3,650000	4	0,129099
2	4,233333	6	0,307679
3	4,283333	6	0,213698
Vš.skup.	4,106250	16	0,353023

Komentář: Na uskutečnění daného pracovního úkonu potřebuje nejkratší čas 1. dělník. Podává také nejvyrovnanější výkony – směrodatná odchylka proměnné X je u něj nejmenší. Naopak nejpomalejší je 3. dělník.

Nyní vytvoříme krabicové diagramy: Návrat do Statistiky podle skupin – Kategoriz. krabicový graf (současné zobrazení krabicových diagramů pro všechny tři výběry)



Pomocí N-P plot orientačně posoudíme normalitu všech tří výběrů:
Návrat do Statistiky podle skupin – ANOVA & testy – Kategoriz. norm. pravd. grafy



Komentář: Ve všech třech případech se tečky jen málo odchylují od přímky, lze soudit, že data pocházejí z normálního rozložení.

Provedení testu o shodě rozptylů:
Návrat do Statistiky podle skupin – Leveneovy testy

	Leveneův test homogenity rozpylů (cas_delniku.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
X	0,042708	2	0,021354	0,183333	13	0,014103	1,514205	0,256356

Komentář: Testová statistika Levenova testu nabývá hodnoty 1,5142, stupně volnosti čitatele = 2, jmenovatele = 13, odpovídající p-hodnota = 0,256, tedy na hladině významnosti 0,05 se nezamítá hypotézu o shodě rozptylů.

Provedení testu o shodě středních hodnot:
Návrat do Statistiky podle skupin – Analýza rozptylu.

	Analýza rozptylu (cas_delniku.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
X	1,117708	2	0,558854	0,751667	13	0,057821	9,665327	0,002680

Komentář: Skupinový součet čtverců $S_A = 1{,}1177$, počet stupňů volnosti $f_A = 2$, reziduální součet čtverců $S_E = 0{,}7517$, počet stupňů volnosti $f_E = 13$, testová statistika $F_A = \frac{S_A / f_A}{S_E / f_E}$ nabývá hodnoty 9,6653, počet stupňů volnosti čitatele = 2, jmenovatele = 13, odpovídající p-hodnota = 0,00268, tedy na hladině významnosti 0,05 se zamítá hypotéza o shodě středních hodnot .

Provedení metody mnohonásobného porovnávání (Scheffého test – viz skripta Základní statistické metody, věta 8.2.2.1.):
Návrat do do Statistiky podle skupin – Post- hoc – Schefféův test.

	Scheffeho test; proměn.:X (cas_delniku.sta) Označ. rozdíly jsou významné na hlad. p < ,05000		
ID	{1} M=3,6500	{2} M=4,2333	{3} M=4,2833
1 {1}		0,008391	0,004705
2 {2}	0,008391		0,937504
3 {3}	0,004705	0,937504	

Komentář: Tabulka obsahuje p-hodnoty pro testování hypotéz o shodě středních hodnot všech dvojic výběrů. Výsledek Scheffého metody ukazuje, že na hladině významnosti 0,05 se liší výkony dělníků (1,2), (1,3) a neliší se (2,3).

Úkol 2.: V cestovní kanceláři zkoumali u 609 náhodně vybraných klientů, o jaké ubytování měli zájem (varianty apartmán, bungalov, hotel, stan) a zjišťovali též pohlaví klienta.

Typ ubytování	apartmán	bungalov	hotel	stan
Počet žen	12	27	208	33
Počet mužů	100	68	36	152

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozdíly v typech ubytování mezi muži a ženami jsou způsobeny pouze náhodnými vlivy.

Návod:
Postupujeme podle skript Základní statistické metody, Věta 8.5.1.1. Testujeme hypotézu H_0: $\vartheta_1 = \ldots = \vartheta_4$ proti alternativní hypotéze H_1: aspoň jedna dvojice parametrů je různá.
Načteme datový soubor klienti_CK.sta. Proměnná POHLAVI obsahuje hodnotu 0 pro ženu, 1 pro muže. Proměnná TYP UBYTOVANI má hodnotu 1 pro apartmán, hodnotu 2 pro bungalov, hodnotu 3 pro hotel a hodnotu 4 pro stan.
Nejprve zjistíme podíly mužů v jednotlivých typech ubytování.
Statistiky – Základní statistiky/tabulky - Rozklad & jednofakt. ANOVA - OK - Proměnné - Závislé POHLAVI, Grupovaci TYP UBYTOVANI, OK, Kódy pro grupovací proměnné – Vše, OK – Popisné statistiky - Výpočet: Tabulka statistik – ponecháme zaškrtnuto N - OK.

typ ubytovani	pohlavi průměr	pohlavi N
apartmán	0,892857	112
bungalov	0,602941	68
hotel	0,147541	244
stan	0,821622	185
Vš.skup.	0,540230	609

Komentář: Vidíme, že z těch klientů, kteří se ubytovali v apartmánu, bylo 89,3% mužů, mezi obyvateli bungalovů bylo 60,3% mužů, z ubytovaných v hotelu bylo mužů pouze 14,7% a z těch, kteří bydleli pod stanem, bylo 82,1% mužů.

Ověříme splnění podmínek dobré aproximace: $n_j m_* > 5$ pro všechna j = 1, ..., r. Vážený průměr m_* se nachází v posledním řádku výstupní Rozkladové tabulky popisných statistik. Jeho hodnotu okopírujeme do políček pro průměry relativní četnosti ubytovaných v jednotlivých typech ubytování, poslední řádek odstraníme a k tabulce přidáme jednu novou proměnnou, do jejíhož Dlouhého jména napíšeme =v2*v3.

typ ubytovani	pohlavi průměr	pohlavi N	NProm =v2*v3
apartmán	0,540230	112	60,505747
bungalov	0,540230	68	36,735632
hotel	0,540230	244	131,816092
stan	0,540230	185	99,942529

Komentář: Vidíme, že podmínky dobré aproximace jsou splněny.

Dále provedeme testování hypotézy o shodě parametrů čtyř alternativních rozložení. Statistiky – Základní statistiky/tabulky – Kontingenční tabulky – OK - Specif. tabulky – List 1 POHLAVI, List 2 TYP UBYTOVANI, OK– Možnosti - Statistiky dvourozm tabulek - zaškrtneme Pearson & M-L Chi –square – Detailní výsledky – Detailní 2-rozm. tabulky

Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	267,6070	df=3	p=0,0000
M-V chí-kvadr.	294,9782	df=3	p=0,0000

Komentář: Testová statistika Q (viz skripta Základní statistické metody, vzorec 8.15.) se realizuje hodnotou 267,6070, počet stupňů volnosti je 3, odpovídající p-hodnota = 0,0000, tedy na asymptotické hladině významnosti 0,05 hypotézu H_0 zamítáme. S rizikem omylu nejvýše 0,05 jsme tedy prokázali, že rozdíly v podílech klientů a klientek ubytovaných v různých typech ubytovacích zařízení nelze vysvětlit pouze náhodnými vlivy.

Nakonec provedeme metodu mnohonásobného porovnávání, abychom zjistili, které dvojice typů ubytování se liší na asymptotické hladině významnosti 0,05.
Návrat do Statistiky podle skupin – Post- hoc – Schefféův test.

typ ubytovani	{1} M=,89286	{2} M=,60294	{3} M=,14754	{4} M=,82162
apartmán {1}		0,000016	0,000000	0,471207
bungalov {2}	0,000016		0,000000	0,000797
hotel {3}	0,000000	0,000000		0,000000
stan {4}	0,471207	0,000797	0,000000	

Komentář: Tabulka obsahuje p-hodnoty pro testování hypotéz o shodě středních hodnot všech dvojic výběrů. Výsledek Scheffého metody ukazuje, že z hlediska podílu mužů se na hladině významnosti 0,05 neliší pouze ubytování v apartmánu a ve stanu.

Příklady k samostatnému řešení

Příklad 1.: Studenti byli vyučováni předmětu za využití pěti pedagogických metod: tradiční způsob, programová výuka, audiotechnika, audiovizuální technika a vizuální technika.

Z každé skupiny byl vybrán náhodný vzorek studentů a všichni byli podrobeni témuž písemnému testu. Výsledky testu:

metoda	počet bodů							
tradiční	76,2	48,3	85,1	63,7	91,6	87,2		
programová	85,2	74,3	76,5	80,3	67,4	67,9	72,1	60,4
audio	67,3	60,1	55,4	72,3	40			
audiovizuální	75,8	81,6	90,3	78	67,8	57,6		
vizuální	50,5	70,2	88,8	67,1	77,7	73,9		

Na hladině významnosti 0,05 testujte hypotézu, že znalosti všech studentů jsou stejné a nezávisí na použité pedagogické metodě. V případě zamítnutí hypotézy zjistěte, které výběry se liší na hladině významnosti 0,05.

Řešení:
Načteme datový soubor pet_metod.sta. Proměnná BODY obsahuje dosažené počty bodů a proměnná METODA označení příslušné pedagogické metody.

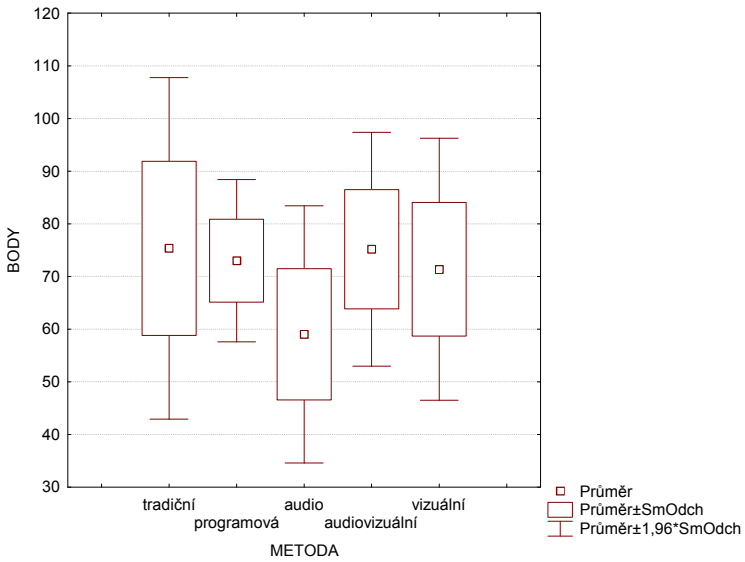
Nejprve vypočteme průměry, směrodatné odchylky a rozsahy všech tří výběrů:

Rozkladová tabulka popisných statistik (pet_metod.sta)
N=31 (V seznamu záv. prom. nejsou ChD)

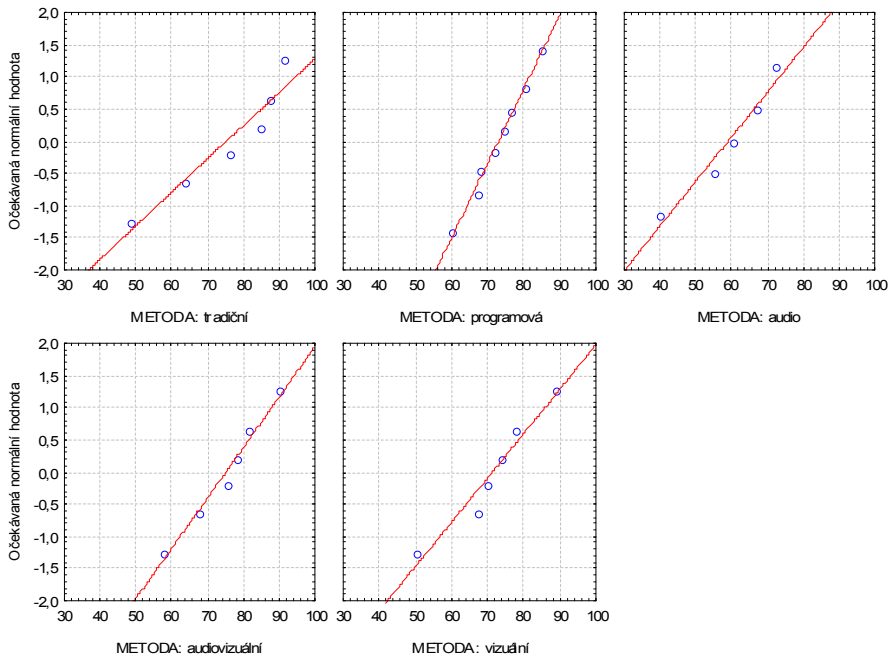
METODA	BODY průměr	BODY N	BODY Sm.odch.
tradiční	75,35000	6	16,53901
programová	73,01250	8	7,86501
audio	59,02000	5	12,45941
audiovizuální	75,18333	6	11,32862
vizuální	71,36667	6	12,69199
Vš.skup.	71,30968	31	12,69534

Komentář: Nejlepších výsledků dosahují studenti vyučovaní tradiční metodou, podávají však nejméně vyrovnané výkony (počty bodů v této skupině mají největší směrodatnou odchylku). Naopak nejhoršího výsledku dosáhli studenti vyučovaní audio metodou. Nejvyrovnanější výkony pozorujeme u studentů vyučovaných programovou metodou.

Vytvoříme krabicové diagramy:



Pomocí N-P grafů vizuálně posoudíme normalitu všech pěti výběrů:



Komentář: Ze vzhledu N-P grafů je patrné, že předpoklad normality je ve všech pěti případech oprávněný.

Provedeme Levenův test (testování homogenity rozptylů všech pěti výběrů)

	Leveneův test homogenity rozpylů (pet_metod.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
BODY	162,4883	4	40,62208	1289,544	26	49,59783	0,819029	0,524791

Komentář: Testová statistika F se realizuje hodnotou 0,819, počet stupňů volnosti čitatele = 4, jmenovatele = 26, odpovídající p-hodnota = 0,5248, na hladině významnosti 0,05 tedy nezamítáme hypotézu o shodě rozptylů.

Budeme testovat hypotézu o shodě středních hodnot všech pěti výběrů:

	Analýza rozptylu (pet_metod.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
BODY	966,3737	4	241,5934	3868,773	26	148,7990	1,623623	0,198252

Komentář: Testová statistika F se realizuje hodnotou 1,6236, počet stupňů volnosti čitatele = 4, jmenovatele = 26, odpovídající p-hodnota = 0,1983, na hladině významnosti 0,05 tedy nezamítáme hypotézu o shodě středních hodnot. Znamená to, že s rizikem omylu nejvýše 5% se neprokázal rozdíl v účinnosti jednotlivých pedagogických metod..

Příklad 2.: Pan Novák může cestovat z místa bydliště do místa pracoviště třemi různými způsoby: tramvají (způsob A), autobusem (způsob B) a metrem s následným přestupem na tramvaj (způsob C). Máme k dispozici jeho naměřené časy cestování do práce v době ranní špičky (včetně čekání na příslušný spoj) v minutách:

způsob A: 32, 39, 42, 37, 34, 38:
způsob B: 30, 34, 28, 26, 32,
způsob C: 40, 37, 31, 39, 38, 33, 34

Pro všechny tři způsoby dopravy vypočtěte průměrné časy cestování. Na hladině významnosti 0,05 testujte hypotézu, že doba cestování do práce nezávisí na způsobu dopravy. V případě zamítnutí nulové hypotézy zjistěte, které způsoby dopravy do práce se od sebe liší na hladině významnosti 0,05.

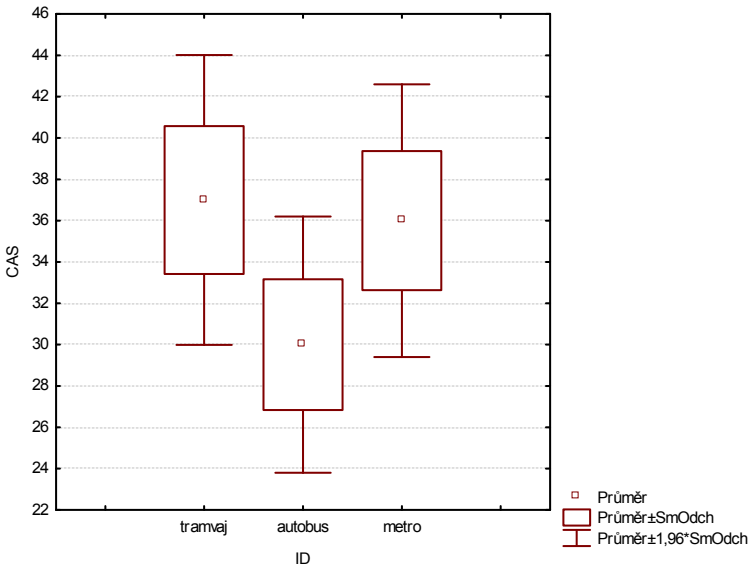
Řešení:
Načteme datový soubor doby_cestovani.sta. Proměnná CAS obsahuje zjištěné doby cestování a proměnná ID označení příslušného způsoby dopravy.

Nejprve vypočteme průměry, směrodatné odchylky a rozsahy všech tří výběrů:

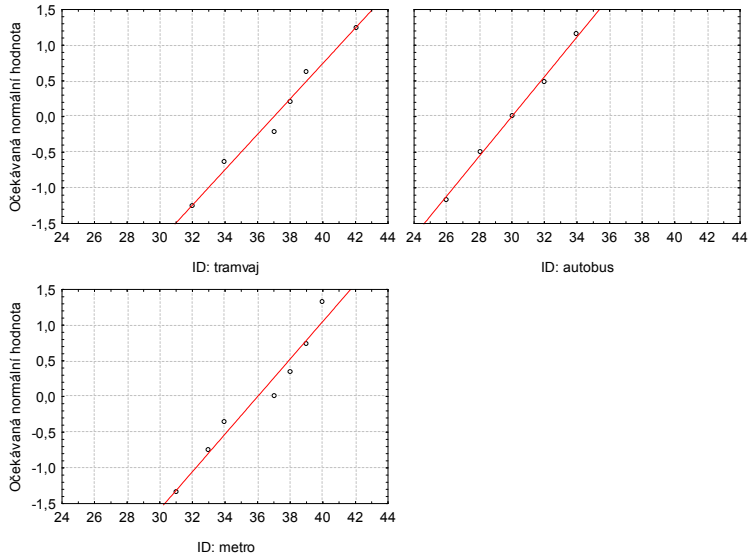
Rozkladová tabulka popisných statistik (doby_cestovani.sta) N=18 (V seznamu záv. prom. nejsou ChD)			
ID	CAS průměr	CAS N	CAS Sm.odch.
tramvaj	37,00000	6	3,577709
autobus	30,00000	5	3,162278
metro	36,00000	7	3,366502
Vš.skup.	34,66667	18	4,379095

Komentář: Nejkratší průměrnou dobu do zaměstnání pan Novák cestuje, když použije autobus, naopak nejdéle cestuje tramvají. Variabilita dob jednotlivých způsobů cestování je vcelku vyrovnaná.

Vytvoříme krabicové diagramy:



Pomocí N-P grafů vizuálně posoudíme normalitu všech tří výběrů:



Komentář: Ze vzhledu N-P grafů je patrné, že předpoklad normality je ve všech třech případech oprávněný.

Provedeme Levenův test (testování homogenity rozptylů všech tří výběrů)

Proměnná	Leveneův test homogenity rozpylů (doby_cestovani.sta) Označ. efekty jsou význ. na hlad. p < ,05000 SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
CAS	0,609524	2	0,304762	43,39048	15	2,892698	0,105356	0,900665

Komentář: Testová statistika F se realizuje hodnotou 0,1054, počet stupňů volnosti čitatele = 2, jmenovatele = 15, odpovídající p-hodnota = 0,9007, na hladině významnosti 0,05 tedy nezamítáme hypotézu o shodě rozptylů.

Budeme testovat hypotézu o shodě středních hodnot všech tří výběrů:

	Analýza rozptylu (doby_cestovani.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
CAS	154,0000	2	77,00000	172,0000	15	11,46667	6,715116	0,008267

Komentář: Testová statistika F se realizuje hodnotou 6,7151, počet stupňů volnosti čitatele = 2, jmenovatele = 15, odpovídající p-hodnota = 0,0083, na hladině významnosti 0,05 tedy zamítáme hypotézu o shodě středních hodnot. Znamená to, že s rizikem omylu nejvýše 5% se prokázal rozdíl v dobách cestování pana Nováka do zaměstnání autobusem, tramvají a metrem.

Scheffého metodou mnohonásobného porovnávání zjistíme, které dvojice způsobů cestování do zaměstnání se liší na hladině významnosti 0,05:

	Scheffeho test; proměn.:CAS (doby_cestovani.sta) Označ. rozdíly jsou významné na hlad. p < ,05000		
ID	{1} M=37,000	{2} M=30,000	{3} M=36,000
tramvaj {1}		0,013410	0,869732
autobus {2}	0,013410		0,028046
metro {3}	0,869732	0,028046	

Komentář: Z tabulky vyplývá, že s rizikem omylu nejvýše 5% se neliší pouze cestování tramvají a metrem.

Příklad 3.: U 856 žáků ZŠ bylo zjišťováno celkové IQ (proměnná IQ_CELK). Na asymptotické hladině významnosti 0,05 testujte hypotézu, že pravděpodobnost výskytu dítěte s nadprůměrným IQ_CELK (tj. nad 100 bodů) je stejná ve skupinách matek se základním, středoškolským a vysokoškolským vzděláním (proměnná VZDEL_M).

Řešení:

Máme tři nezávislé náhodné výběry, j-tý pochází z rozložení $A(\vartheta_j)$, j = 1, 2, 3. Testujeme hypotézu H_0: $\vartheta_1 = \vartheta_2 = \vartheta_3$.

$n_1 = 361$, $n_2 = 386$, $n_3 = 109$, $n = 856$

$m_1 = 111/361 = 30{,}75\%$, $m_2 = 227/386 = 58{,}81\%$, $m_3 = 85/109 = 77{,}98\%$, $m_* = (111+227+85)/856 = 423/856 = 49{,}42\%$.

Podmínky dobré aproximace:

$$361 \cdot \frac{423}{856} = 178{,}39\,,\ 386 \cdot \frac{423}{856} = 190{,}75\,,\ 109 \cdot \frac{423}{856} = 53{,}86$$

Testová statistika

$$Q = \frac{1}{M_*(1-M_*)} \sum_{j=1}^{r} n_j M_j^{\,2} - n\frac{M_*}{1-M_*} = 99{,}53$$

Kritický obor: $W = \langle \chi^2_{0,95}(2), \infty) = \langle 5{,}991, \infty)$.

Protože testové kritérium se realizuje v kritickém oboru, H_0 zamítáme na asymptotické hladině významnosti 0,05.
Metoda mnohonásobného porovnávání prokázala, že na asymptotické hladině významnosti 0,05 se liší všechny tři skupiny.

Cvičení 9: Neparametrické úlohy o mediánech

Úkol 1.: **Párový znaménkový test a párový Wilcoxonův test**
Při zjišťování kvality jedné složky půdy se používají dvě metody označené A a B. Výsledky:

Vzorek	1	2	3	4	5	6	7	8	9	10	11	12
A	0,275	0,312	0,284	0,3	0,365	0,298	0,312	0,315	0,242	0,321	0,335	0,307
B	0,28	0,312	0,288	0,298	0,361	0,307	0,319	0,315	0,242	0,323	0,341	0,315

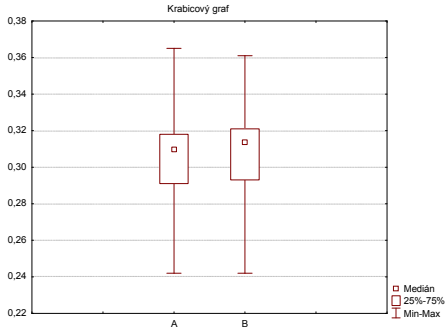
Na hladině významnosti 0,05 testujte pomocí párového znaménkového testu a poté pomocí párového Wilcoxonova testu hypotézu, že metody A a B dávají stejné výsledky.

Návod:
Načteme datový soubor kvalita_pudy.sta. Proměnná A obsahuje výsledky metody A, proměnná B výsledky metody B.
Nejprve budeme testovat nulovou hypotézu pomocí párového znaménkového testu (viz skripta Základní statistické metody, věta 9.3.1.3. a 9.3.1.1.).
Statistiky –Neparametrická statistika – Porovnání dvou závislých vzorků (proměnné) – OK – 1. seznam proměnných A, 2. seznam proměnných B – OK – Znaménkový test.

	Znaménkový test (kvalita_pudy.sta) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet různých	procent v < V	Z	Úroveň p
A & B	9	77,77778	1,333333	0,182422

Komentář: Vidíme, že nenulových hodnot n = 9. Z nich záporných je $77,\overline{7}\,\%$, tj.7. Kladných je tedy 9 – 7 = 2, což je hodnota testové statistiky S_Z^+. Asymptotická testová statistika U_0 (zde označená jako Z) se realizuje hodnotou $1,\overline{3}$. Odpovídající asymptotická p-hodnota je 0,18422, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu, že obě metody dávají stejné výsledky.
Upozornění: V tomto případě není splněna podmínka pro využití asymptotické normality statistiky S_Z^+, tj. n > 20. Je tedy vhodnější najít v tabulkách kritické hodnoty pro znaménkový test (viz skripta Základní statistické metody, tabulka na straně 156). Pro n = 9 a $\alpha = 0,05$ jsou kritické hodnoty $k_1 = 1$, $k_2 = 8$. Protože kritický obor $W = \langle 0,1\rangle \cup \langle 8,9\rangle$ neobsahuje hodnotu 2, nezamítáme H_0 na hladině významnosti 0,05. Dostáváme stejný výsledek při použití asymptotického testu

Nyní graficky znázorníme výsledky obou metod: Návrat do Porovnání 2 proměnných - Krabicový graf všech proměnných – OK – A, B – OK.



Komentář: Z krabicových diagramů je vidět, že obě metody se poněkud liší v úrovni, ale neliší se ve variabilitě.

Dále provedeme Wilcoxonův párový test (viz skripta Základní statistické metody, věta 9.3.2.1): Návrat do Porovnání 2 proměnných – Wilcoxonův párový test.

	Wilcoxonův párový test (kvalita_pudy) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	p-hodn.
A & B	9	5,000000	2,073221	0,038153

Komentář: Výstupní tabulka poskytne hodnotu testové statistiky S_W^+ (zde označena T), hodnotu asymptotické testové statistiky U_0 a p-hodnotu pro U_0. V tomto případě je p-hodnota 0,038153, tedy nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. Ze srovnání asymptotických p-hodnot pro znaménkový test a pro Wilcoxonův test plyne, že Wilcoxonův test je silnější.
Upozornění: V tomto případě není splněna podmínka pro využití asymptotické normality statistiky S_W^+, tj. $n \geq 30$. Je tedy vhodnější najít v tabulkách kritickou hodnotu pro Wilcoxonův párový test (viz skripta Základní statistické metody, tabulka na straně 157). Pro $n = 9$ a $\alpha = 0{,}05$ je kritická hodnota rovna 5. Protože kritický obor $W = \langle 0,5 \rangle$ obsahuje hodnotu 5, zamítáme H_0 na hladině významnosti 0,05. To souhlasí s výsledkem asymptotického testu.

Úkol 2.: **Znaménkový test a jednovýběrový Wilcoxonův test**
Vyráběné ocelové tyče mají kolísavou délku s předpokládanou hodnotou mediánu 10 m. Náhodný výběr 10 tyčí poskytl tyto výsledky:
9,83 10,10 9,72 9,91 10,04 9,95 9,82 9,73 9,81 9,90
Na hladině významnosti 0,05 testujte hypotézu, že předpoklad o mediánu délky tyčí je oprávněný.

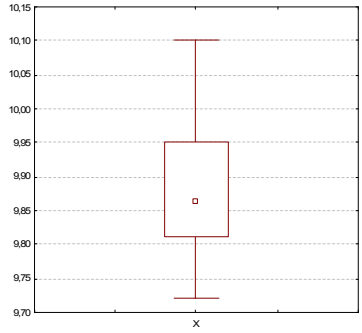
Návod:
Načteme datový soubor ocelove_tyce.sta. Proměnná X obsahuje naměřené hodnoty a proměnná Y obsahuje konstantu 10. Provedení znaménkového a Wilcoxonova testu je nyní stejné jako v předešlém případě.

	Znaménkový test (ocelove_tyce.sta) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet různých	procent v < V	Z	Úroveň p
X & Y	10	80,00000	1,581139	0,113846

Dvojice proměnných	Wilcoxonův párový test (ocelove_tyce) Označené testy jsou významné na hladině p <,05000 Počet platných	T	Z	Úroveň p
X & Y	10	5,500000	2,242448	0,024933

Komentář: Znaménkový test poskytl asymptotickou p-hodnotu 0,113846, tedy nulová hypotéza se nezamítá na hladině významnosti 0,05. Wilcoxonův test dává asymptotickou p-hodnotu 0,024933, tedy nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. Podobně jak v úkolu 1 by bylo vhodnější najít kritické hodnoty v tabulkách. V případě znaménkového testu jsou kritické hodnoty pro n = 10 a $\alpha = 0,05$ rovny 1 a 9, testová statistika $S_Z^+ = 2$. Protože S_Z^+ nepatří do kritického oboru $W = \langle 0,1 \rangle \cup \langle 9,10 \rangle$, nelze nulovou hypotézu zamítnout na hladině významnosti 0,05, což je v souladu s výsledkem asymptotického testu. V případě Wilcoxonova testu je kritická hodnota pro n = 10 $\alpha = 0,05$ rovna 8. Protože kritický obor $W = \langle 0,8 \rangle$ obsahuje hodnotu 5,5, zamítáme H_0 na hladině významnosti 0,05. I zde tedy existuje soulad mezi výsledkem přesného a asymptotického testu.

Podobně jako v úkolu 1. znázorníme výsledky měření pomocí krabicového diagramu:

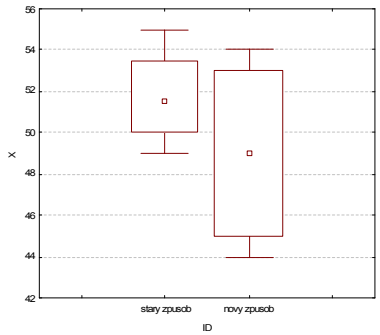


Úkol 3.: **Dvouvýběrový Wilcoxonův test, dvouvýběrový Kolmogorovův - Smirnovův test**
Bylo vybráno 10 polí stejné kvality. Na čtyřech z nich se zkoušel nový způsob hnojení, zbylých šest bylo ošetřeno starým způsobem. Pole byla oseta pšenicí a sledoval se její hektarový výnos. Je třeba testovat na hladině významnosti 0,05, zda nový způsob hnojení má týž vliv na průměrné hektarové výnosy pšenice jako starý způsob hnojení.
hektarové výnosy při novém způsobu: 51 52 49 55
hektarové výnosy při starém způsobu: 45 54 48 44 53 50

Návod:
Načteme datový soubor hnojeni_poli.sta. Proměnná X udává výnosy pšenice při obou způsobech hnojení a proměnná ID nabývá hodnoty 1 pro starý způsob hnojení, hodnoty 2 pro nový způsob hnojení.
Dvouvýběrový Wilcoxonův test (viz skripta Základní statistické metody, věta 9.4.1.1): Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků (skupiny) – OK - Seznam závislých proměnných X, Grupovací proměnná ID - OK - Mann – Whitneyův U test.

Proměnná	Mann-Whitneyův U test (hnojeni_poli) Dle proměn. ID Označené testy jsou významné na hladině p <,05000 Sčt poř. st. zp.	Sčt poř. n. zp.	U	Z	p-hodn.	Z uprav.	p-hodn.	N platn. st. zp.	N platn. n. zp.	2*1str. přesné p
X	27,00000	28,00000	7,000000	0,959403	0,337356	0,959403	0,337356	4	6	0,352381

Komentář: Ve výstupní tabulce jsou součty pořadí T_1, T_2, hodnota testové statistiky $\min(U_1, U_2)$ označená U, hodnota asymptotické testové statistiky U_0 (označená Z), asymptotická p-hodnota pro U_0 a přesná p-hodnota (ozn. 2*1 str. přesné p – ta se používá pro rozsahy výběrů pod 30). V našem případě přesná p-hodnota = 0,352381, tedy H_0 nezamítáme na hladině významnosti 0,05.
Výpočet je vhodné doplnit krabicovým diagramem.



Komentář: Je zřejmé, že výnosy při novém způsobu hnojení jsou vesměs nižší než při starém způsobu a také vykazují mnohem větší variabilitu.

Dvouvýběrový Kolmogorovův – Smirnovův test (viz skripta Základní statistické metody, věta 9.4.3.1.) poskytne tabulku:

Proměnná	Kolmogorov-Smirnovův test (hnojeni_poli.sta) Dle proměn. ID Označené testy jsou významné na hladině p <,05000 Max záp rozdíl	Max klad rozdíl	p-hodn.	Průměr stary zpusob	Průměr novy zpusob	Sm.odch. stary zpusob	Sm.odch. novy zpusob	N platn. stary zpusob
X	-0,083333	0,500000	p > .10	51,75000	49,00000	2,500000	4,098780	4

Komentář: Dostaneme maximální záporný a maximální kladný rozdíl mezi hodnotami obou výběrových distribučních funkcí, dolní omezení pro p-hodnotu (p > 0,1), průměry, směrodatné odchylky a rozsahy obou výběrů.
Protože p > 0,05, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Úkol 4.: **Kruskalův – Wallisův test a mediánový test**
Voda po holení jisté značky se prodává ve čtyřech různých lahvičkách stejného obsahu. Údaje o počtu prodaných lahviček za týden v různých obchodech:
1. typ: 50 35 43 30 62 52 43 57 33 70 64 58 53 65 39
2. typ: 31 37 59 67 44 49 54 62 34 42 40
3. typ: 27 19 32 20 18 23
4. typ: 35 39 37 38 28 33.
Posuďte na 5% hladině významnosti, zda typ lahvičky ovlivňuje úroveň prodeje.

Návod:

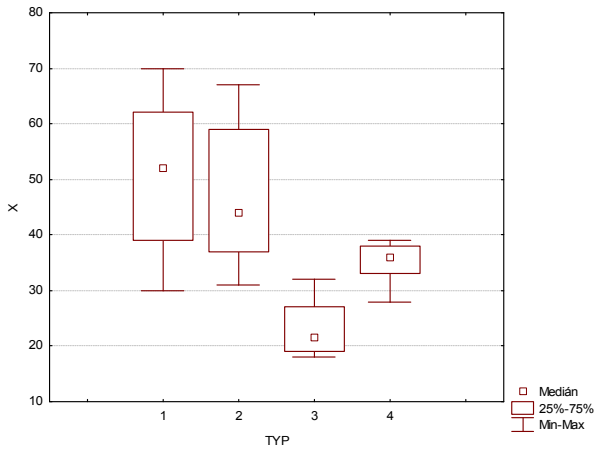
Načteme datový soubor voda_po_holeni.sta. Proměnná X udává počet prodaných lahviček a proměnná TYP udává typ lahvičky. Úloha vede na Kruskalův – Wallisův test nebo mediánový test (viz skripta Základní statistické metody, věta 9.5.1.1. resp. věta 9.5.3.1.): Statistiky – Neparametrická statistika – Porovnání více nezávislých vzorků (skupiny) – OK – Proměnné – Seznam závislých proměnných X, Grupovací proměnná TYP – OK – Shrnutí: Kruskal-Wallis ANOVA & Mediánový test. Ve dvou výstupních tabulkách se objeví výsledky K-W testu a mediánového testu.

Závislá: X	Kruskal-Wallisova ANOVA založ. na poř.; X (voda_po_holeni.sta Nezávislá (grupovací) proměnná :TYP Kruskal-Wallisův test: H (3, N= 38) =18,80199 p =,0003		
	Kód	Počet platných	Součet pořadí
1	1	15	379,0000
2	2	11	257,0000
3	3	6	24,0000
4	4	6	81,0000

Závislá: X	Mediánový test, celk. medián = 39,5000; X (voda_po_holeni.sta Nezávislá (grupovací) proměnná :TYP Chi-Kvadr. = 17,53939 sv = 3 p = ,0005				
	1	2	3	4	Celkem
<= Medián: pozorov.	4,00000	3,00000	6,00000	6,00000	19,00000
očekáv.	7,50000	5,50000	3,00000	3,00000	
poz.-oč.	-3,50000	-2,50000	3,00000	3,00000	
> Medián: pozorov.	11,00000	8,00000	0,00000	0,00000	19,00000
očekáv.	7,50000	5,50000	3,00000	3,00000	
poz.-oč.	3,50000	2,50000	-3,00000	-3,00000	
Celkem: oček.	15,00000	11,00000	6,00000	6,00000	38,00000

Komentář: Oba testy zamítají hypotézu o shodě mediánů v daných čtyřech skupinách. Testová statistika K-W testu je 18,802, počet stupňů volnosti 3, odpovídající asymptotická p-hodnota 0,0003. Testová statistika mediánového testu je 17,539, počet stupňů volnosti 3, odpovídající asymptotická p-hodnota 0,0005. K-W test je poněkud silnější (p-hodnota = 0,0003, zatímco p-hodnota pro mediánový test je 0,0005).

Grafické znázornění výsledků: návrat do Shrnutí: Kruskal-Wallis ANOVA & Mediánový test – Krabicový graf – Vyberte proměnnou: X – OK – Typ krabicového grafu: Medián/Kvartily/Rozpětí – OK.



Komentář: Je vidět, že úroveň prodeje pro 1. typ je nevyšší, zatímco pro 3. typ nejnižší. Pro 1. a 2. typ je variabilita prodeje značná, pro 3. a 4. typ naopak malá.

Vzhledem k tomu, že jsme zamítli nulovou hypotézu o shodě mediánů na asymptotické hladině významnosti 0,05, provedeme metoda mnohonásobného porovnávání:Vícenás. porovnání průměrného pořadí pro vš. sk.

	Vícenásobné porovnání p hodnot (oboustr.);X (voda_po_holeni.sta Nezávislá (grupovací) proměnná :TYP Kruskal-Wallisův test: H (3, N= 38) =18,80199 p =,0003			
Závislá: X	1 R:25,267	2 R:23,364	3 R:4,0000	4 R:13,500
1		1,000000	0,000447	0,170297
2	1,000000		0,003579	0,481908
3	0,000447	0,003579		0,832208
4	0,170297	0,481908	0,832208	

Komentář: Tabulka obsahuje p-hodnoty pro test hypotézy, že l-tý a k-tý výběr pocházejí z téhož rozložení. Vidíme, že na hladině významnosti 0,05 zamítáme nulovou hypotézu pro 1. a 3. typ lahvičky a pro 2. a 3. typ lahvičky.

Příklady k samostatnému řešení

Příklad 1.: U osmi osob byl změřen systolický krevní tlak před pokusem a po něm.

č. osoby	1	2	3	4	5	6	7	8
tlak před	130	185	162	136	147	181	128	139
tlak po	139	190	175	135	155	175	158	149

Na hladině významnosti 0,05 testujte hypotézu, že pokus neovlivní systolický krevní tlak.

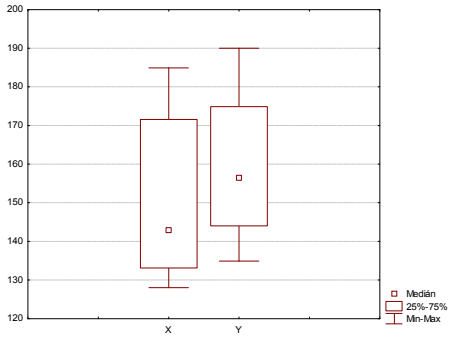
Řešení:

Stejně jako v úkolu 1 provedeme párový znaménkový a párový Wilcoxonův test. Načteme soubor tlak.sta. Proměnná X obsahuje hodnoty tlaku před pokusem, proměnná Y po pokusu. Výstupní tabulka:

	Znaménkový test (tlak.sta) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet různých	procent v < V	Z	Úroveň p
X & Y	8	75,00000	1,060660	0,288844

Komentář: Jelikož p-hodnota = 0,288844 > 0,05, nelze nulovou hypotézu zamítnout na hladině významnosti 0,05. Vzhledem k malému rozsahu výběru je však vhodnější najít v tabulkách kritické hodnoty pro znaménkový test (viz skripta Základní statistické metody, tabulka na straně 156). Pro n = 8 a $\alpha = 0,05$ jsou kritické hodnoty $k_1 = 0$, $k_2 = 8$. Hodnotu testové statistiky S_Z^+ získáme jako 75% z 8, což je 6. Protože kritický obor neobsahuje hodnotu 6, nezamítáme H_0 na hladině významnosti 0,05. Dostáváme stejný výsledek při použití asymptotického testu

Grafické znázornění výsledků pomocí krabicového diagramu:



Komentář: Úroveň tlaku před pokusem byla poněkud nižší než po pokusu, variabilita je jen nepatrně odlišná.

Výstupní tabulka Wilcoxonova testu:

	Wilcoxonův párový test (tlak.sta) Označené testy jsou významné na hladině p <,05000			
Dvojice proměnných	Počet platných	T	Z	Úroveň p
X & Y	8	4,000000	1,960392	0,049951

Vidíme, že asymptotická p-hodnota = 0,049951, nulová hypotéza se tedy zamítá na asymptotické hladině významnosti 0,05. Rozsah souboru je pouze 8, není splněna podmínka dobré aproximace standardizovaným normálním rozložením (n > 30). Proto zjistíme ve skriptech Základní statistické metody v tabulce na str. 157 kritickou hodnotu pro n = 8 a α = 0,05. Kritická hodnota je rovna 3, hodnota testové statistiky (ve výstupní tabulce označena T) je 4 > 3, tedy nulovou hypotézu nezamítáme na hladině významnosti 0,05, což je v souladu s výsledkem znaménkového testu.

Příklad 2.: Majitel obchodu chtěl zjistit, zda velikost nákupů (v dolarech) placených kreditními kartami Master/EuroCard a Visa jsou přibližně stejné. Náhodně vybral 7 nákupů placených Master/EuroCard: 42 77 46 73 78 33 37 a 9 placených Visou: 39 10 119 68 76 126 53 79 102. Lze na hladině významnosti 0,05 tvrdit, že nákupů placených těmito dvěma typy karet se shodují?

Řešení:
Stejně jako úkolu 3 použijeme dvouvýběrový Wilcoxonův test a Kolmogorovův - Smirnovův test.
Načteme datový soubor kreditni_karty.sta. Proměnná X obsahuje hodnoty nákupů, proměnná ID má hodnotu 1 pro kartu Master/EuroCard a hodnotu 2 pro kartu Visa.
Výstupní tabulka pro dvouvýběrový Wilcoxonův test:

	Mann-Whitneyův U test (kreditni_karty.sta) Dle proměn. ID Označené testy jsou významné na hladině p <,05000									
Proměnná	Sčt poř. M/E Card	Sčt poř. Visa	U	Z	Úroveň p	Z upravené	Úroveň p	N platn. M/E Card	N platn. Visa	2*1str. přesné p
X	48,00000	88,00000	20,00000	-1,21729	0,223495	-1,21729	0,223495	7	9	0,252273

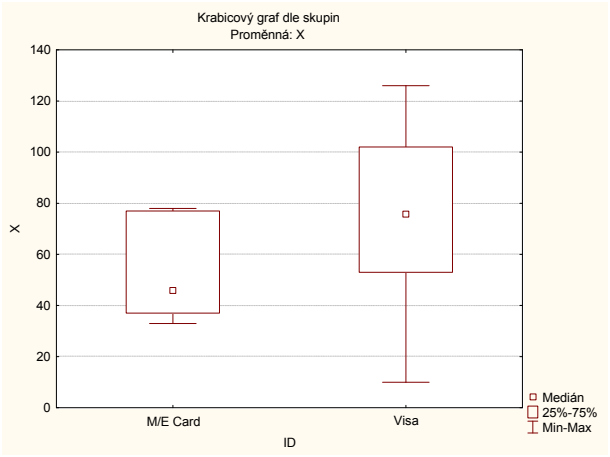
Komentář: Zajímá nás především přesná p-hodnota (ozn. 2*1 sided exact p – ta se používá pro rozsahy výběrů pod 30). V našem případě přesná p-hodnota = 0,252273, tedy H_0 nezamítáme na hladině významnosti 0,05.

Výstupní tabulka pro Kolmogorovův – Smirnovův test:

	Kolmogorov-Smirnovův test (kreditni_karty.sta) Dle proměn. ID Označené testy jsou významné na hladině p <,05000								
Proměnná	Max záp rozdíl	Max klad rozdíl	Úroveň p	Průměr M/E Card	Průměr Visa	Sm.odch. M/E Card	Sm.odch. Visa	N platn. M/E Card	N platn. Visa
X	-0,444444	0,111111	p > .10	55,14286	74,66667	19,97856	37,64306	7	9

Komentář: Dostaneme maximální záporný a maximální kladný rozdíl mezi hodnotami obou výběrových distribučních funkcí, dolní omezení pro p-hodnotu (p > 0,1), průměry, směrodatné odchylky a rozsahy obou výběrů. Protože p > 0,05, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Výpočet je vhodné doplnit krabicovým diagramem typu Medián/Kvartily/Rozpětí.



Komentář: Vidíme, že platby za nákupy kartou Master/EuroCard mají nižší úroveň, ale přibližně stejnou variabilitu jako platby kartou Visa.

Příklad 3.: Z produkce tří podniků vyrábějících televizory bylo vylosováno 10, 8 a 12 kusů. Byly získány následující výsledky zjišťování citlivosti těchto televizorů v mikrovoltech:
1. podnik: 420 560 600 490 550 570 340 480 510 460
2. podnik: 400 420 580 470 470 500 520 530
3. podnik: 450 700 630 590 420 590 610 540 740 690 540 670
Testujte na hladině významnosti 0,05 hypotézu o shodě úrovně citlivosti televizorů v jednotlivých podnicích.

Řešení:
Stejně jako v úkolu 4 provedeme Kruskalův-Wallisův test a mediánový test. Načteme datový soubor televizory.sta. Proměnná X obsahuje hodnoty citlivosti televizorů, proměnná ID udává číslo podniku.
Ve dvou výstupních tabulkách máme výsledky mediánového testu a K-W testu.

Závislá: X	Kruskal-Wallisova ANOVA založ. na poř.; X (televizory.sta) Nezávislá (grupovací) proměnná : ID Kruskal-Wallisův test: H (2, N= 30) =8,204318 p =,0165		
	Kód	Počet platných	Součet pořadí
1. podnik	1	10	127,0000
2. podnik	2	9	101,5000
3. podnik	3	11	236,5000

Závislá: X	Mediánový test, celk. medián = 535,000; X (televizory.sta) Nezávislá (grupovací) proměnná : ID Chi-Kvadr. = 7,632323 sv = 2 p = ,0220			
	1. podnik	2. podnik	3. podnik	Celkem
<= Medián: pozorov.	6,00000	7,00000	2,00000	15,00000
očekáv.	5,00000	4,50000	5,50000	
poz.-oč.	1,00000	2,50000	-3,50000	
> Medián: pozorov.	4,00000	2,00000	9,00000	15,00000
očekáv.	5,00000	4,50000	5,50000	
poz.-oč.	-1,00000	-2,50000	3,50000	
Celkem: oček.	10,00000	9,00000	11,00000	30,00000

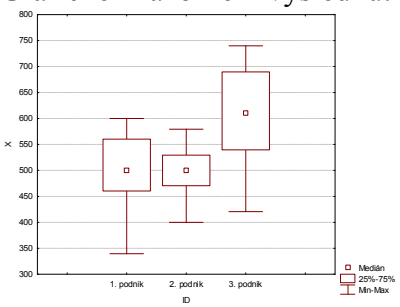
Komentář: Protože zjištěné p-hodnoty jsou menší než zvolená hladina významnosti 0,05, oba testy zamítají hypotézu o shodě mediánů v daných třech skupinách.

Výsledky metody mnohonásobného porovnávání:

Závislá: X	Vícenásobné porovnání p hodnot (oboustr.); X (televizory.sta Nezávislá (grupovací) proměnná : ID Kruskal-Wallisův test: H (2, N= 30) =8,204318 p =,0165		
	1. podnik R:12,700	2. podnik R:11,278	3. podnik R:21,500
1. podnik		1,000000	0,066447
2. podnik	1,000000		0,029347
3. podnik	0,066447	0,029347	

Komentář: Na hladině významnosti 0,05 se liší citlivost televizorů vyráběných ve 2. a 3. podniku.

Grafické znázornění výsledků:



Komentář: Je vidět, že citlivost televizorů ze 3. podniku je nevyšší, zatímco ze 2.podniku nejnižší. Citlivost výrobků 3. podniku však vykazuje největší variabilitu.

Cvičení 10: Porovnání empirického a teoretického rozložení

Úkol 1.: Ze souboru rodin s pěti dětmi bylo náhodně vybráno 84 rodin a byl zjišťován počet chlapců:

Počet chlapců	0	1	2	3	4	5
Počet rodin	3	10	22	31	14	4

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozložení počtu chlapců se řídí binomickým rozložením Bi(5; 0,5).

Řešení:
Pravděpodobnost, že náhodná veličina s rozložením Bi(5; 0,5) bude nabývat hodnot p_0, ..., p_5 je $p_j = \binom{5}{j}\frac{1}{32}, j = 0,1,\ldots,5$.
Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j
0	3	0,03125	84.0,03125=2,625
1	10	0,15625	84.0,15625=13,125
2	22	0,3125	84.0,3125=26,25
3	31	0,3125	84.0,3125=26,25
4	14	0,15625	84.0,15625=13,125
5	4	0,03125	84.0,03125=2,625

Podmínky dobré aproximace nejsou splněny, sloučíme tedy první dvě varianty a poslední dvě varianty.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
0 a 1	13	0,1875	84.0,1875=15,75	0,480159
2	22	0,3125	84.0,3125=26,25	0,688095
3	31	0,3125	84.0,3125=26,25	0,859524
4 a 5	18	0,1875	84.0,1875=15,75	0,321429

Vypočteme realizaci testové statistiky: K = 0,48059 + 0,688095 + 0,859524 + 0,321429 = 2,3492, počet tříd r = 4, počet odhadovaných parametrů p = 0, r – p - 1 = 3, kritický obor $W = \langle \chi^2_{1-\alpha}(r-p-1), \infty) = \langle \chi^2_{0,95}(3), \infty) = \langle 7,8147; \infty)$. Protože $K \notin W$, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:
Vytvoříme datový soubor se dvěma proměnnými a čtyřmi případy. Proměnná nj obsahuje zjištěné četnosti (po sloučení variant), proměnná npj pak teoretické četnosti.
Statistiky – Neparametrická statistika – Pozorované vs. očekávané χ2 – OK – Proměnné – Pozorované četnosti nj, očekávané četnosti npj – OK – Výpočet.

Případ	Pozorované vs. očekávané četnosti (Tabulka1 Chi-Kvadr. = 2,349206 sv = 3 p = ,503161 pozorov. nj	očekáv. npj	P - O	(P-O)^2 /O
C: 1	13,00000	15,75000	-2,75000	0,480159
C: 2	22,00000	26,25000	-4,25000	0,688095
C: 3	31,00000	26,25000	4,75000	0,859524
C: 4	18,00000	15,75000	2,25000	0,321429
Sčt	84,00000	84,00000	0,00000	2,349206

V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (2,349206), počet stupňů volnosti = 3 a p-hodnota (0,503161). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Úkol 2.: U 48 studentek VŠE v Praze byla zjišťována výška (v cm):

165	170	170	179	170	168	174	162	167	165	170	173	183	176	165	168
171	178	168	168	169	163	172	184	176	175	176	169	168	170	166	160
167	162	162	166	170	168	155	162	169	166	160	169	165	163	168	163

Pomocí testu dobré shody testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí histogramu posuďte vizuálně předpoklad normality.

Výpočet pomocí systému STATISTICA:

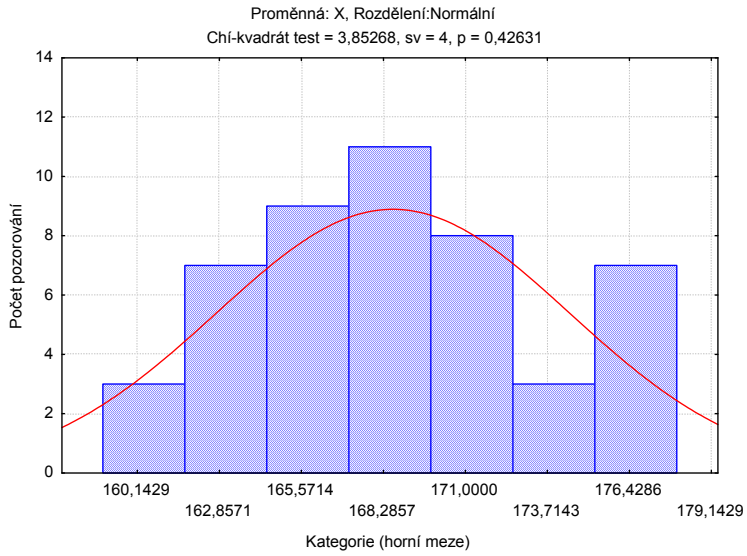
Načteme datový soubor vyska.sta. Statistiky - Prokládání rozdělení – ponecháme implicitní nastavení na normální rozložení – OK – Proměnná X – OK – na záložce Parametry změníme Počet kategorií na 7 (podle Sturgesova pravidla) – Výpočet.

Horní hranice	Proměnná: X, Rozdělení:Normální (vyska.sta) Chí-kvadrát = 1,09280, sv = 1 (uprav.) , p = 0,29585 Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 157,14286	1	1	2,08333	2,0833	1,19706	1,19706	2,49387	2,4939
162,28571	6	7	12,50000	14,5833	5,51484	6,71189	11,48924	13,9831
167,42857	12	19	25,00000	39,5833	13,46220	20,17409	28,04624	42,0293
172,57143	19	38	39,58333	79,1667	15,89146	36,06555	33,10721	75,1366
177,71429	6	44	12,50000	91,6667	9,07700	45,14255	18,91042	94,0470
182,85714	2	46	4,16667	95,8333	2,50365	47,64620	5,21594	99,2629
< Nekonečno	2	48	4,16667	100,0000	0,35380	48,00000	0,73708	100,0000

Při tomto roztřídění dat do 7 intervalů nejsou splněny podmínky dobré aproximace, ve třech intervalech jsou teoretické četnosti pod 5. Změníme tedy dolní mez na 159 a horní na 178.

Horní hranice	Proměnná: X, Rozdělení:Normální (vyska.sta) Chí-kvadrát = 3,85268, sv = 4, p = 0,42631 Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 161,71429	3	3	6,25000	6,2500	5,722996	5,72300	11,92291	11,9229
164,42857	7	10	14,58333	20,8333	5,675946	11,39894	11,82489	23,7478
167,14286	9	19	18,75000	39,5833	7,862633	19,26157	16,38048	40,1283
169,85714	11	30	22,91667	62,5000	8,812455	28,07403	18,35928	58,4876
172,57143	8	38	16,66667	79,1667	7,991516	36,06555	16,64899	75,1366
175,28571	3	41	6,25000	85,4167	5,863558	41,92910	12,21575	87,3523
< Nekonečno	7	48	14,58333	100,0000	6,070896	48,00000	12,64770	100,0000

V tomto případě jsou podmínky dobré aproximace splněny. Testová statistika se realizuje hodnotou 3,85268, p-hodnota je 0,42631, tedy na asymptotické hladině významnosti 0,05 hypotézu o normalitě nezamítáme. Podívejme se ještě na histogram s proloženou Gaussovou křivkou: Na záložce Základní výsledky zvolíme Graf pozorovaného a očekávaného rozdělení.



Samostatný úkol: Tentýž úkol proveďte zvlášť pro studentky oboru informatika a národní hospodářství.

Úkol 3.: Jsou známy počty občanů města Brna podle měsíce narození (stav k 31.12.2001).

měsíc narození	počet osob
leden	32309
únor	30126
březen	35010
duben	34761
květen	34955
červen	32883
červenec	33255
srpen	31604
září	31173
říjen	30536
listopad	28571
prosinec	29467
celkem	384650

Na asymptotické hladině významnosti 0,05 ověřte hypotézu, že pravděpodobnost narození je pro všechny měsíce stejná. (Pravděpodobnost narození pro libovolný měsíc získáte tak, že počet dnů v tomto měsíci podělíte počtem dnů v roce.) Počty narozených lidí v jednotlivých měsících roku rovněž znázorněte graficky.

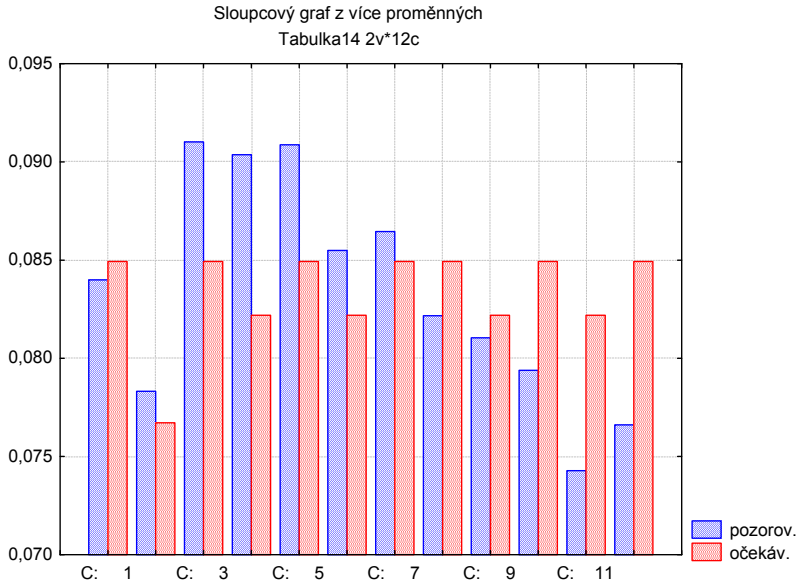
Výpočet pomocí systému STATISTICA:
Načteme datový soubor obyvatele_brna.sta. Tento soubor má tři proměnné (X, X1 a Y) a 12 případů. Proměnná X obsahuje absolutní četnosti z předchozí tabulky. Proměnné X1

obsahuje relativní četnosti, tj. v jejím Dlouhém jméně je napsáno = X/384650. Proměnná Y obsahuje očekávané relativní četnosti, tj. její hodnoty jsou vždy počet dní v měsíci/365. Statistiky – Neparametrická statistika – Pozorované versus očekávané χ^2 – OK - Pozorované četnosti X1, Očekávané četnosti Y - OK – Výpočet. Dostaneme tabulku:

Případ	Pozorované vs. očekávané četnosti (obyvatele_Brna.sta Chi-Kvadr. = ,0039156 sv = 11 p = 1,000000			
	pozorov. X1	očekáv. Y	P - O	(P-O)^2 /O
C: 1	0,083996	0,084932	-0,000936	0,000010
C: 2	0,078321	0,076712	0,001608	0,000034
C: 3	0,091018	0,084932	0,006086	0,000436
C: 4	0,090370	0,082192	0,008179	0,000814
C: 5	0,090875	0,084932	0,005943	0,000416
C: 6	0,085488	0,082192	0,003296	0,000132
C: 7	0,086455	0,084932	0,001524	0,000027
C: 8	0,082163	0,084932	-0,002769	0,000090
C: 9	0,081043	0,082192	-0,001149	0,000016
C: 10	0,079386	0,084932	-0,005545	0,000362
C: 11	0,074278	0,082192	-0,007914	0,000762
C: 12	0,076607	0,084932	-0,008324	0,000816
Sčt	1,000000	1,000000	0,000000	0,003916

Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}(r-1-p)$. V našem případě je r = 12, p = 0. $\chi^2_{0,95}(11) = 19,675$. Protože K = 0,0039282 < 19,675,nezamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

Výpočet doplníme sloupkovým diagramem pozorovaných relativních četností a očekávaných relativních četností.



Komentář: Největší rozdíly mezi pozorovanými a očekávanými relativními četnostmi jsou v prosinci, dubnu a listopadu, naopak nejmenší v lednu a září.

Úkol 4.: Firma, která vlastní několik supermarketů, se zajímá, zda zákazníci dávají přednost některému dnu v týdnu pro nákup. Náhodně bylo vybráno 300 zákazníků, kteří měli říci, který den v týdnu nejčastěji nakupují v supermarketu.
Výsledky:

Den	pondělí	úterý	středa	čtvrtek	pátek	sobota	neděle
Počet	10	20	40	40	80	60	50

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že žádný den v týdnu nemá při nakupování v supermarketu přednost před jinými dny.

Návod:
Načteme datový soubor nakupy.sta. Proměnná X obsahuje pozorované absolutní četnosti a Y vypočítané teoretické četnosti (v našem případě 300/7).
Statistiky – Neparametrické statistiky – Pozorované vs. očekávané χ^2 – Proměnné Pozorované X, Očekávané Y, OK – Výpočet. Dostaneme tabulku:

	Pozorované vs. očekávané četnosti (nakupy.sta) Chi-Kvadr. = 78,00000 sv = 6 p = ,000000			
Případ	pozorov. X	očekáv. Y	P - O	(P-O)^2 /O
C: 1	10,0000	42,8571	-32,8571	25,19048
C: 2	20,0000	42,8571	-22,8571	12,19048
C: 3	40,0000	42,8571	-2,8571	0,19048
C: 4	40,0000	42,8571	-2,8571	0,19048
C: 5	80,0000	42,8571	37,1429	32,19048
C: 6	60,0000	42,8571	17,1429	6,85714
C: 7	50,0000	42,8571	7,1429	1,19048
Sčt	300,0000	300,0000	0,0000	78,00000

Komentář: Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Square = 78) a odpovídající p-hodnotu, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota velmi malá, takřka nulová, takže nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že zákazníci nakupují během týdne nerovnoměrně.

Příklad k samostatnému řešení: D rybníka bylo umístěno 5 pastí, přičemž každá past svítila jiným světlem (bílým, žlutým, modrým, zeleným, červeným). Do těchto pastí se chytilo 56, 72, 41, 53 a 38 jedinců. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že barva světla v pasti nemá vliv na počet chycených jedinců.
Výsledek: Testová statistika nabývá hodnoty 14,1154, kritický obor je $W = \langle 9{,}488; \infty)$, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme. S rizikem omylu nejvýše 0,05 jsme prokázali, že barva světla v pasti má vliv na počet chycených jedinců.

Cvičení 11: Hodnocení kontingenčních tabulek

Úkol 1.: Testování hypotézy o nezávislosti, měření síly závislosti
V roce 1950 zkoumali Yule a Kendall barvu očí a vlasů u 6800 mužů.

Barva očí	Barva vlasů			
	světlá	kaštanová	černá	rezavá
modrá	1768	807	180	47
šedá nebo zelená	946	1387	746	53
hnědá	115	438	288	16

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti barvy očí a barvy vlasů. Vypočtěte Cramérův koeficient. Simultánní četnosti znázorněte graficky.

Návod:
Testujeme hypotézu H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti
H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny. Testová statistika má tvar:

$$K = \sum_{j=1}^{r}\sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$$. Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$,

kde r, s jsou počty variant jednotlivých proměnných.
Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$.
V našem případě zjistíme, že K = 1088,15, r = 3 , s = 4, $\chi^2_{1-\alpha}((r-1)(s-1) = \chi^2_{0,95}(6) =$ 12,592 a protože hodnota testové statistiky K = 1088,15 ≥ 12,592, zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde m = min{r,s}. Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi X a Y, čím blíže je 0, tím je tato závislost volnější.
Význam hodnot Cramérova koeficientu:
mezi 0 až 0,1 … zanedbatelná závislost,
mezi 0,1 až 0,3 … slabá závislost,
mezi 0,3 až 0,7 … střední závislost,
mezi 0,7 až 1 … silná závislost.

Otevřeme datový soubor oci_vlasy.sta. Před provedením testu je zapotřebí ověřit podmínky dobré aproximace: Statistiky – Základní statistiky/tabulky – Kontingenční tabulky - Specif. tabulky – List 1 OCI, List 2 VLASY, OK, Váhy - CETNOST, Stav zapnuto, OK – na záložce Možnosti zaškrtneme Očekávané četnosti – Výpočet.

Souhrnná tab.: Očekávané četnosti (oci_vlasy.sta) Četnost označených buněk > 10 Pearsonův chí-kv. : 1088,15, sv=6, p=0,00000					
OCI	VLASY světlá	VLASY kaštanová	VLASY černá	VLASY rezavá	Řádk. součty
modrá	1167,259	1085,976	500,902	47,8622	2802,000
šedá nebo zelená	1304,731	1213,875	559,895	53,4990	3132,000
hnědá	357,010	332,149	153,202	14,6388	857,000
Vš.skup.	2829,000	2632,000	1214,000	116,0000	6791,000

Podmínky dobré aproximace jsou splněny. Všechny teoretické četnosti jsou větší než 5. Nyní budeme testovat hypotézu o nezávislosti proměnných OCI, VLASY.

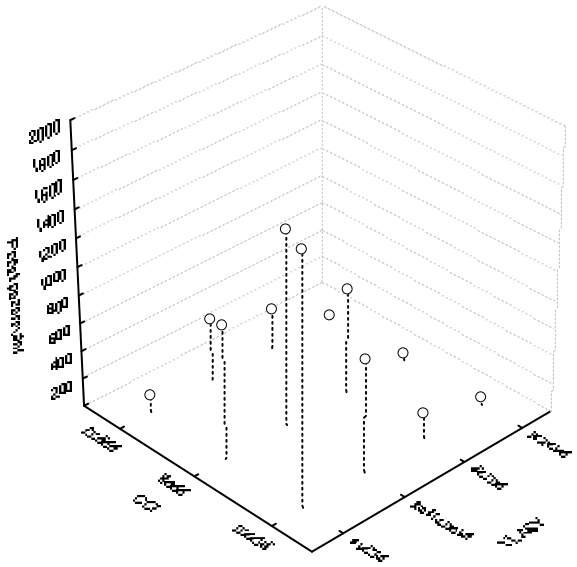
Návrat do Výsledky; kontingenční tabulky – na záložce Detaily zaškrtneme Pearsonův & M-L Chi - kvadrát, Phi & Cramerovo V – Detailní výsledky – Detailní 2 rozm. tabulky.

Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	1088,149	df=6	p=0,0000
M-V chí-kvadr.	1155,669	df=6	p=0,0000
Fí	,4002923		
Kontingenční koeficient	,3716246		
Cramér. V	,2830494		

Ve výstupní tabulce najdeme mj. hodnotu testové statistiky (Pearsonův chí-kv = 1088,149) s počtem stupňů volnosti (sv = 6) a odpovídající p-hodnotou (p = 0,0000), dále Cramérův koeficient (V = 0,283). Protože p-hodnota je mnohem menší než 0,05, nulovou hypotézu o nezávislosti barvy očí a barvy vlasů zamítáme na asymptotické hladině významnosti 0,05. Cramérův koeficient svědčí o slabé závislosti barvy očí a vlasů.

Pro grafické znázornění četností se vrátíme do Výsledky; kontingenční tabulky – Detailní výsledky – 3D histogramy. Po vytvoření grafu 2 krát poklepeme levým tlačítkem myši na pozadí grafu:
Rozvržení grafu – Typ Šipky – OK. Graf lze natáčet pomocí volby Zorný bod.

Dvourozměrné rozdělení:
OCI x VLASY



Úkol 2.: Fisherův faktoriálový test
100 náhodně vybraných osob bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

preferovaný nápoj	*pohlaví*	
	muž	žena
A	20	30
B	30	20

Na hladině významnosti 0,05 testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Návod: Vytvoříme nový datový soubor o třech proměnných NAPOJ, POHLAVI, CETNOST a čtyřech případech. Do proměnné NAPOJ napíšeme dvakrát pod sebe 1 (nápoj A) a dvakrát pod sebe 2 (nápoj B). Do proměnné POHLAVI napíšeme jedničku (1 – muž) a dvojku (2 – žena) a znovu jedničku a dvojku. D proměnné CETNOST napíšeme uvedené četnosti. Statistiky – Základní statistiky/tabulky – Kontingenční tabulky - Specif. tabulky – List 1 NAPOJ, List 2 POHLAVI, OK, Váhy - CETNOST, Stav zapnuto, OK – na záložce Možnosti zaškrtneme Fisher exakt, Yates, McNemar (2x2) – Detailní výsledky – Detailní 2-rozm. tabulky.

Statist.	Statist. : POHLAVI(2) x NAPOJ(2) (kap11_2)		
	Chí-kvadr.	sv	p
Pearsonův chí-kv.	4,000000	df=1	p=,04550
M-V chí-kvadr.	4,027103	df=1	p=,04478
Yatesův chí-kv.	3,240000	df=1	p=,07186
Fisherův přesný, 1-str.			p=,03567
2-stranný			p=,07134
McNemarův chí-kv. (A/D)	,0250000	df=1	p=,87437
(B/C)	,0166667	df=1	p=,89728

Ve výstupní tabulce je mimo jiné uvedena p-hodnota pro oboustranný a jednostranný test. V našem případě se jedná o oboustranný test (nevíme, zda muži více preferují nápoj A či nápoj B než ženy), zajímáme se tedy o Fisherův přesný, 2-str. Ta je 0,07134. Protože p-hodnota je větší než 0,05, nezamítáme na hladině významnosti 0,05 hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Úkol 3.: Podíl šancí

Pro údaje z úkolu 2 vypočtěte podíl šancí a sestrojte 95% asymptotický interval spolehlivosti pro podíl šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti 0,05 hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Návod: Nejprve zopakujme teorii:

Ve čtyřpolních tabulkách používáme charakteristiku $OR = \frac{ad}{bc}$, která se nazývá podíl šancí (odds ratio). Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		$n_{j.}$
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
$n_{.k}$	a+c	b+d	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za 1. okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$. Podíl šancí je $OR = \frac{ad}{bc}$. Považujeme ho za odhad skutečného podílu šancí oρ. Pomocí 100(1-α)% asymptotického intervalu spolehlivosti pro logaritmus skutečného podílu šancí ln oρ lze na asymptotické hladině významnosti α testovat hypotézu o nezávislosti nominálních veličin X a Y. Asymptotický 100(1-α)% interval spolehlivosti pro přirozený logaritmus skutečného podílu šancí má meze:

$\ln OR \pm \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}$. Jestliže interval spolehlivosti nezahrne 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

V našem případě podíl šancí vypočteme ručně. $OR = \frac{ac}{bd} = \frac{20 \cdot 20}{30 \cdot 30} = \frac{4}{9} = 0,\overline{4}$. Dolní a horní mez intervalu spolehlivosti pro OR zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a dvou případech. Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:

=log(4/9)-sqrt(1/20+1/30+1/30+1/20)*VNormal(0,975;0;1)

a analogicky do Do Dlouhého jména proměnné HM napíšeme vzorec pro horní mez:
=log(4/9)+sqrt(1/20+1/30+1/30+1/20)*VNormal(0,975;0;1)

	1 DM	2 HM
1	-1,61108	-0,01078

Výsledek: -1,61108 < ln oρ < -0,01078 s pravděpodobností přibližně 0,95. Protože tento interval spolehlivosti neobsahuje 0**,** na asymptotické hladině významnosti 0,05 zamítáme hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.
Tento výsledek je v rozporu s výsledkem, ke kterému dospěl Fisherův přesný test. Je to způsobeno tím, že test pomocí asymptotického intervalu spolehlivosti je pouze přibližný.

Příklady k samostatnému řešení

Příklad 1.: Zajímá nás, zda má lokalita v ČR vliv na objem exportu do sousedních zemí. Sledujeme lokality: Ostrava, Brno, Plzeň, Praha a země: Slovensko, Rakousko, Německo, Polsko, USA). Máme k dispozici tato data:

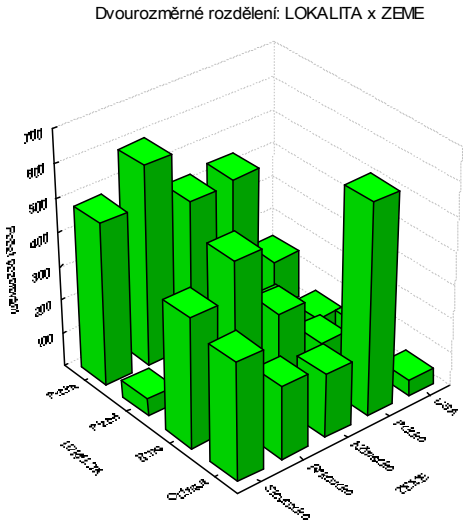
	Kam:				
Odkud:	Slovensko	Rakousko	Německo	Polsko	USA
Ostrava	350	216	189	626	46
Brno	387	489	274	126	115
Plzeň	52	83	264	132	51
Praha	484	594	737	447	141

Řešení:
Načteme datový soubor export.sta. Proměnná EXPORT obsahuje objem exportu pro zvolenou kombinaci LOKALITA, ZEMĚ.
Testová statistika K ze vzorce (11.1) nabývá hodnoty 821,59, odpovídající p-hodnota je velmi blízká nule, tedy na asymptotické hladině významnosti 0,05 považujeme za prokázanou závislost objemu exportu na lokalitě v České republice. Podmínky dobré aproximace jsou splněny, jak vidíme z následující tabulky:

Souhrnná tab.: Očekávané četnosti (export.sta) Četnost označených buněk > 10 Pearsonův chí-kv. : 821,587, sv=12, p=0,00000						
LOKALITA	ZEME Slovensko	ZEME Rakousko	ZEME Německo	ZEME Polsko	ZEME USA	Řádk. součty
Ostrava	330,106	358,371	301,840	345,146	91,5375	1427,000
Brno	321,778	349,330	294,226	336,438	89,2282	1391,000
Plzeň	134,633	146,161	123,105	140,767	37,3335	582,000
Praha	486,484	528,138	444,829	508,649	134,9008	2103,000
Vš.skup.	1273,000	1382,000	1164,000	1331,000	353,0000	5503,000

Cramérův koeficient nabývá hodnoty 0,223, tedy mezi sledovanými proměnnými existuje slabá závislost.

Zjištěná data ještě znázorníme graficky:



Příklad 2.: 200 respondentů, z nichž bylo 73 žen, hodnotilo úroveň jistého časopisu. 34 žen ji hodnotilo kladně, stejně jako 47 mužů. Ostatní respondenti se o úrovni časopisu vyjádřili záporně. Vypočtěte a interpretujte podíl šancí časopisu na kladné hodnocení a na asymptotické hladině významnosti 0,05 testujte pomocí asymptotického intervalu spolehlivosti pro podíl šancí hypotézu, že hodnocení úrovně časopisu nezávisí na pohlaví respondenta. Proveďte též Fisherův přesný test a vypočtěte Cramérův koeficient.

Řešení:

Sestavíme čtyřpolní kontingenční tabulku simultánních absolutních četností:

hodnocení časopisu	pohlaví respondenta		$n_{j.}$
	muž	žena	
kladné	47	34	81
záporné	80	39	119
$n_{.k}$	127	73	200

Kladné hodnocení časopisu pozorujeme u 37% mužů a u 46,6 % žen.

Vypočítáme podíl šancí časopisu na kladné hodnocení.

$\mathrm{OR} = \frac{ad}{bc} = \frac{47 \cdot 39}{34 \cdot 80} = 0{,}673897,$ což znamená, že u mužů je 0,674 x menší šance na kladné hodnocení časopisu než u žen.

Dále provedeme výpočty pro stanovení intervalu spolehlivosti.

$\ln \mathrm{OR} = -0{,}39468,\ \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{47}+\frac{1}{34}+\frac{1}{80}+\frac{1}{39}} = 0{,}298,\ u_{0{,}975} = 1{,}96$

$\ln d = -0{,}39468 - 0{,}298 \cdot 1{,}96 = -0{,}97876,\ \ln h = -0{,}39468 + 0{,}298 \cdot 1{,}96 = 1{,}89476$

Protože interval (-0,97876; 1,89476) obsahuje číslo 0, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti hodnocení úrovně časopisu na pohlaví respondenta.

Další výsledky máme v tabulce:

Statist.	Statist. : hodnoceni(2) x pohlavi(2) (Tabulka13		
	Chí-kvadr.	sv	p
Pearsonův chí-kv.	1,760835	df=1	p=,18452
M-V chí-kvadr.	1,752654	df=1	p=,18555
Yatesův chí-kv.	1,386184	df=1	p=,23905
Fisherův přesný, 1-str.			p=,11967
2-stranný			p=,23131
McNemarův chí-kv. (A/D)	17,76316	df=1	p=,00003
(B/C)	,5697674	df=1	p=,45035
Fí pro tabulky 2 x 2	,0938306		
Tetrachorická korelace	,1507792		
Kontingenční koeficient	,0934202		

Fisherův přesný test poskytl p-hodnotu 0,23131, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti hodnocení úrovně časopisu na pohlaví respondenta. Cramérův koeficient je 0,0938, což svědčí o zanedbatelné závislosti mezi sledovanými veličinami.

Příklad 3.: Na hladině významnosti 0,05 testujte hypotézu o nezávislosti pedagogické hodnosti a pohlaví a vypočtěte Cramérův koeficient vyjadřující intenzitu závislosti pedagogické hodnosti na pohlaví, jsou-li k dispozici následující údaje:

pohlaví	pedagogická hodnost		
	odb. asistent	docent	profesor
muž	32	15	8
žena	34	8	3

Výsledek: Podmínky dobré aproximace jsou splněny, pouze jediná teoretická četnost klesne po 5. Testová statistika K nabývá hodnoty 3,5, p = 0,1739, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti pedagogické hodnosti a pohlaví. Cramérův koeficient: V = 0,187.

Příklad 4.: 18 mužů onemocnělo určitou chorobou. Někteří z nich se léčili, jiní ne. Někteří se uzdravili, jiní zemřeli. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

přežití	léčení	
	ano	ne
ano	5	3
ne	6	4

Vypočtěte a interpretujte podíl šancí. Pomocí intervalu spolehlivosti pro podíl šancí testujte na asymptotické hladině významnosti 0,05 hypotézu, že přežití nezávisí na léčení proti tvrzení, že léčení zvyšuje šance na přežití.

Výsledek: $OR = 1,\bar{1}$, nulovou hypotézu nezamítáme asymptotické hladině významnosti 0,05, protože levostranný 95% asymptotický interval spolehlivosti pro logaritmus podílu šancí je $(-1,49785; \infty)$.

Cvičení 12: Jednoduchá korelační analýza

Úkol 1.: **Testování nezávislosti ordinálních veličin**

12 různých softwarových firem nabízí speciální programové vybavení pro vedení účetnictví. Jednotlivé programy byly posouzeny odbornou komisí složenou z počítačových odborníků a komisí složenou z profesionálních účetních. Úkolem bylo doporučit vhodný program na základě stanovení pořadí jednotlivých programů. Výsledky posouzení:

Produkt firmy číslo	1	2	3	4	5	6	7	8	9	10	11	12
Pořadí dle odborníků	6	7	1	8	4	2,5	9	12	10	2,5	5	11
Pořadí dle účetních	4	5	2	10	6	1	7	11	8	3	12	9

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou komisí jsou nezávislá.

Návod:

Testujeme vlastně nulovou hypotézu, že koeficient pořadové korelace je roven nule proti oboustranné alternativě.

Načteme datový soubor vedeni_ucetnictvi.sta o dvou proměnných X (hodnocení 1. komise), Y (hodnocení 2. komise) a 12 případech.

Statistiky – Neparametrické statistiky – Korelace – OK – vybereme Vytvořit detailní report - Proměnné X, Y – OK – Spearmanův koef. R. Dostaneme tabulku

	Spearmanovy korelace (vedeni_ucetnictvi.sta) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000			
Dvojice proměnných	Počet plat.	Spearman R	t(N-2)	p-hodn.
X & Y	12	0,714537	3,229806	0,009024

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,7145, testová statistika se realizuje hodnotou 3,2298, odpovídající p-hodnota je 0,009024, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení dvou komisí ve prospěch oboustranné alternativy.

Upozornění: Systém STATISTICA používá při testování hypotézy o pořadové nezávislosti veličin X, Y asymptotickou variantu testu bez ohledu na rozsah náhodného výběru. Pokud rozsah výběru nepřesáhne 20, měli bychom systém STATISTCA použít jen k výpočtu r_S a testování bychom měli provést pomocí tabelované kritické hodnoty. V našem případě pro $n = 12$ a $\alpha = 0,05$ je kritická hodnota 0,5804. Vidíme, že nulovou hypotézu zamítáme na hladině významnosti 0,05, protože $0,7145 \geq 0,5804$.

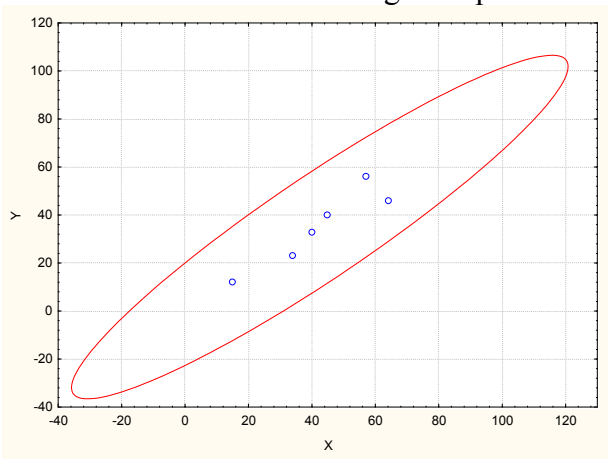
Úkol 2.: **Testování nezávislosti intervalových a poměrových veličin – oboustranná alternativa**

Zjišťovalo se, kolik mg kyseliny mléčné je ve 100 ml krve matek prvorodiček (veličina X) a u jejich novorozenců (veličina Y) těsně po porodu. Byly získány tyto výsledky:

Číslo matky	1	2	3	4	5	6
x_i	40	64	34	15	57	45
y_i	33	46	23	12	56	40

Nakreslete dvourozměrný tečkový diagram, vypočtěte výběrový korelační koeficient, sestrojte 95% interval spolehlivosti pro korelační koeficient a na hladině významnosti 0,05 testujte hypotézu o nezávislosti výsledků obou měření.

Návod: Načteme datový soubor kyselna_mlecna.sta o dvou proměnných X a Y a šesti případech. Obvyklým způsobem zobrazíme dvourozměrný tečkový diagram, s jehož pomocí posoudíme dvourozměrnou normalitu dat. Tedy:
Grafy – Bodové grafy – vypneme lineární proložení - Proměnné X, Y – OK – Detaily - Elipsa normální – OK. Ve vzniklém grafu upravíme měřítka na vodorovné a svislé ose:



Testování hypotézy o nezávislosti: Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 2 seznamy proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Prom. X & prom. Y	Korelace (kyselina_mlecna.sta) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD) Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	42,50000	17,39828									
Y	35,00000	15,89969	0,934832	0,873912	5,265339	0,006232	6	-1,30823	0,854311	6,696994	1,022943

Ve výstupní tabulce je mj. hodnotu výběrového korelačního koeficientu R_{12} (r=0,9348), tzn. že mezi X a Y existuje silná přímá lineární závislost), hodnota testové statistiky (t = 5,2653) a p-hodnotu pro test hypotézy o nezávislosti (p=0,006232), H_0 tedy zamítáme na hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že mezi oběma koncentracemi existuje závislost.
Pro testování pomocí intervalu spolehlivosti zopakujme nejprve teorii:
Nechť dvourozměrný náhodný výběr rozsahu n pochází z dvourozměrného normálního rozložení s koeficientem korelace ρ. Meze 100(1-α)% asymptotického intervalu spolehlivosti pro ρ jsou:

$$d = \operatorname{tgh}\left(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right),\ h = \operatorname{tgh}\left(Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right) \text{ přičemž } \operatorname{tgh} x = \frac{e^x - e^{-x}}{e^x + e^{-x}},\ Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}.$$

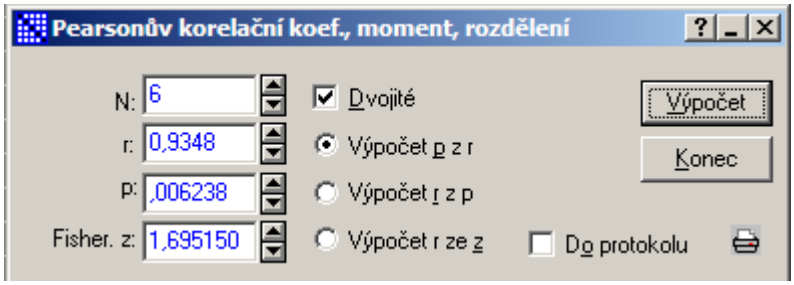
Výpočet mezí intervalu spolehlivosti: vytvoříme nový datový soubor s proměnnými DM a HM. Do Dlouhého jména proměnné DM zapíšeme příkaz
= TanH(0,5*log((1+0,9348)/(1-0,9348))-VNormal(0,975;0;1)/sqrt(6-3))
a do Dlouhého jména proměnné HM zapíšeme příkaz
= TanH(0,5*log((1+0,9348)/(1-0,9348))+VNormal(0,975;0;1)/sqrt(6-3))

	1 DM	2 HM
1	0,510617	0,993014

95% interval spolehlivosti pro ρ má tedy meze 0,5106 a 0,9930, nepokrývá hodnotu 0 a tudíž hypotézu o nezávislosti veličin X, Y zamítáme na hladině významnosti 0,05.

Využití modulu Analýza síly testu
Statistiky – Analýza síly testu – Odhad intervalu - Jedna korelace, t-test – OK – Pozorované R: 0,9348, N: 6, Spolehlivost: 0,95 – Výpočetní algoritmus: zaškrtneme Fisherovo Z (původní) – Vypočítat.
Zjistíme, že Dolní mez = 0,5106, Horní mez = 0,993.

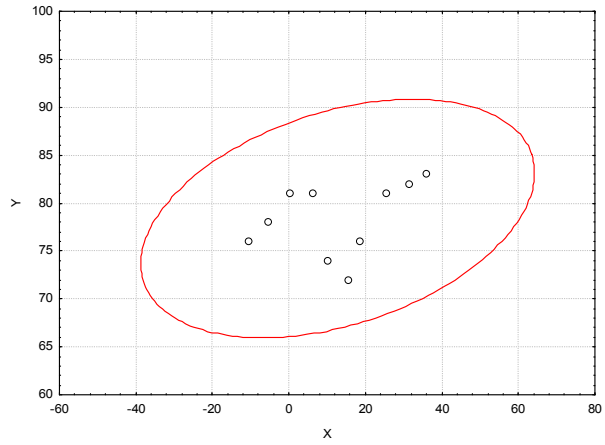
Poznámka: Pokud známe výběrový koeficient korelace a rozsah výběru, můžeme test nezávislosti veličin X, Y provést pomocí Pravděpodobnostního kalkulátoru**.**
Statistiky – Pravděpodobnostní kalkulátor – Korelace – zadáme n a r, zaškrtneme Výpočet p z r – Výpočet.



Úkol 3.: **Testování nezávislosti intervalových a poměrových veličin – jednostranná alternativa**
Při průzkumu příčin dopravních nehod bylo provedeno měření diastolického tlaku 10 skupin řidičů autobusů při různých teplotách vnějšího ovzduší. Data znázorněte graficky, posuďte jejich dvourozměrnou normalitu, vypočtěte realizaci výběrového koeficientu korelace a na hladině významnosti 0,05 testujte hypotézu, že teplota ovzduší neovlivňuje krevní tlak řidičů proti alternativě, že že mezi teplotou a tlakem existuje kladná korelace.

Teplota ovzduší (ve ° C):	-10,5	-5,4	0,2	6,4	10,2	15,6	18,5	25, 5	28,9	31,5	35,8
průměrný tlak (v mm Hg):	76	78	81	81	74	72	76	81	82	83	84

Návod:
Načteme datový soubor ridici_autobusu.sta. Proměnná X obsahuje teploty, proměnná Y tlaky. Vytvoříme dvourozměrný tečkový diagram s 95% elipsou konstantní hustoty pravděpodobnosti:



Komentář: Vzhled diagramu svědčí o dvourozměrné normalitě dat.

Číselná realizace výběrového koeficientu korelace: $r_{12} = 0{,}3823$ svědčí o existenci poměrně slabé přímé lineární závislosti mezi vnější teplotou a diastolickým krevním tlakem řidičů autobusů – s rostoucí teplotou poněkud roste krevní tlak.
Na hladině významnosti 0,05 testujeme hypotézu $H_0 : \rho = 0$ proti pravostranné alternativě $H_1 : \rho > 0$. Pomocí Pravděpodobnostního kalkultátoru zjistíme p-hodnotu pro tuto jednostrannou alternativu: p = 0,1378. Na hladině významnosti 0,05 tedy nezamítáme hypotézu, že vztah mezi teplotou a tlakem je pouze náhodný.

Úkol 4.: **Porovnání dvou korelačních koeficientů**
V psychologickém výzkumu bylo vyšetřeno 426 hochů a 430 dívek. Ve skupině hochů činil výběrový koeficient korelace mezi verbální a performační složkou IQ 0,6033, ve skupině dívek činil 0,5833. Za předpokladu dvourozměrné normality dat testujte na hladině významnosti 0,05 hypotézu, že korelační koeficienty se neliší.
Návod: Nejprve zopakujeme teorii:
Nechť jsou dány dva nezávislé náhodné výběry o rozsazích n a n^* z dvourozměrných normálních rozložení s korelačními koeficienty ρ a ρ^*. Testujeme H_0: $\rho = \rho^*$ proti H_1: $\rho \neq \rho^*$. Označme R_{12} výběrový korelační koeficient 1. výběru a R_{12}^* výběrový korelační koeficient 2. výběru. Položme $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ a $Z^* = \frac{1}{2}\ln\frac{1+R_{12}^*}{1-R_{12}^*}$. Platí-li H_0, pak testová statistika $U = \frac{Z - Z^*}{\sqrt{\frac{1}{n-3} + \frac{1}{n^*-3}}}$ má asymptoticky rozložení N(0,1). Kritický obor pro test H_0 proti oboustranné alternativě tedy je $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$. H_0 zamítáme na asymptotické hladině významnosti α, když $U \in W$.

Výpočet pomocí systému STATISTICA:
Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,6033, do políčka N1 napíšeme 426, do políčka r2 napíšeme 0,5833, do políčka N2 napíšeme 430 - Výpočet. Dostaneme p-hodnotu 0,6528, tedy nezamítáme nulovou hypotézu o shodě dvou koeficientů korelace na asymptotické hladině významnosti 0,05.

Úkoly k samostatnému řešení

Příklad 1.: Bylo sledováno 10 žáků. Na základě psychologického vyšetření byli tito žáci seřazeni podle nervové lability (čím byl žák labilnější, tím dostal vyšší pořadí R_i). Kromě toho sledování žáci dostali pořadí Q_i na základě svých výsledků v matematice (nejlepší žák v matematice dostal pořadí 1). Výsledky jsou uvedeny v tabulce:

Pořadí R_i	1	2	3	4	5	6	7	8	9	10
Pořadí Q_i	9	3	8	5	4	2	10	1	7	6

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že nervová labilita a výsledky v matematice jsou nezávislé.
Výsledek: $r_S = -0{,}127$, H_0 nezamítáme na hladině významnosti 0,05.

Příklad 2.: V náhodném výběru 10 dvoučlenných domácností byl zjišťován měsíční příjem (veličina X, v tisících Kč) a vydání za potraviny (veličina Y, v tisících Kč).

x_i	15	21	34	35	39	42	58	64	75	90
y_i	3	4,5	6,5	6	7	8	9	8	9,5	10,5

Vypočtěte výběrový koeficient korelace. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti veličin X, Y. Sestrojte 95% asymptotický interval spolehlivosti pro ρ
Výsledek: $r_{12} = 0{,}9405$, H_0 zamítáme na hladině významnosti 0,05, s pravděpodobností aspoň 0,95 platí: $0{,}7623 < \rho < 0{,}9862$

Příklad 3.: U určitého výrobku hodnotil expert dvě vlastnosti na desetibodové stupnici tak, že nula je nejhorší a desítka nejlepší hodnocení. Máte k dispozici výsledky hodnocení 11 náhodně vybraných výrobků:

1. vlastnost	3,1	2,8	4,4	5,8	5,1	4,3	4,7	2,9	5,3	5,4	5,9
2. vlastnost	7,2	6,5	6,9	8,4	76	4,4	3,8	7,1	4,3	4,7	8,9

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou vlastností jsou pořadově nezávislá.
Výsledek: $r_S = 0{,}282$, H_0 nezamítáme na hladině významnosti 0,05.

Příklad 4.: V následující tabulce jsou uvedeny číselné realizace a absolutní četnosti náhodného výběru $(X_1, Y_1), (X_1, Y_2), ..., (X_{62}, Y_{62})$ z dvourozměrného rozložení:

x	y						
	1	3	5	7	9	11	13
15	0	0	0	0	1	2	1
25	0	0	0	5	4	2	0
35	0	0	5	8	2	0	0
45	0	5	6	4	0	0	0
55	3	5	3	0	0	0	0
65	4	2	0	0	0	0	0

Podle vzhledu dvourozměrného tečkového diagramu orientačně posuďte dvourozměrnou normalitu dat. Vypočtěte výběrový koeficient korelace a interpretujte ho. Na hladině významnosti 0,05 testujte hypotézu o nezávislosti veličin X a Y.

Výsledek:

Protože tečky v dvourozměrném tečkovém diagramu vytvářejí elipsovitý obrazec, lze připustit dvourozměrnou normalitu. Výběrový koeficient korelace nabývá hodnoty –0,8699, což znamená, že mezi veličinami X a Y existuje dosti silná nepřímá lineární závislost. Testová statistika se realizuje hodnotou -13,6613, odpovídající p-hodnota je velmi blízká 0, nulovou hypotézu zamítáme na hladině významnosti 0,05.

Cvičení 13: Jednoduchá regresní analýza

Úkol 1.: U šesti obchodníků byla zjišťována poptávka po určitém druhu zboží loni (veličina X - v kusech) a letos (veličina Y - v kusech).

číslo. obchodníka	1	2	3	4	5	6
poptávka loni (X)	20	60	70	100	150	260
poptávka letos (Y)	50	60	60	120	230	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Sestavte regresní matici, vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry a zjistěte relativní chyby odhadů regresních parametrů.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.

g) Vypočtěte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

i) Spočtěte střední absolutní procentuální chybu predikce (MAPE)

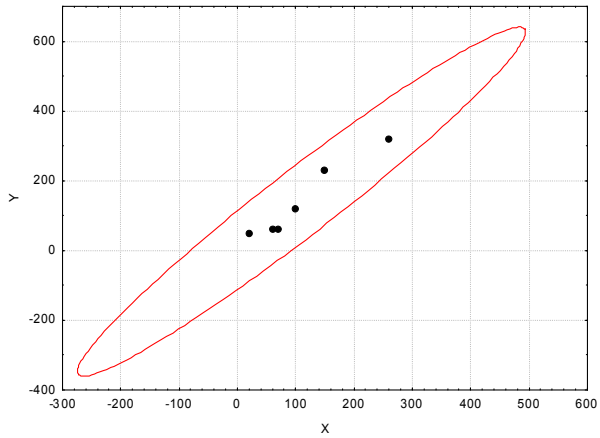
j) Proveďte analýzu reziduí.

Návod:
Načteme nový datový soubor obchodnici.sta se dvěma proměnnými X a Y a 6 případy:

	1 X	2 Y
1	20	50
2	60	60
3	70	60
4	100	120
5	150	230
6	260	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení. Vypočtěte výběrový koeficient korelace mezi X a Y, interpretujte jeho hodnotu a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny.

Zobrazíme dvourozměrný tečkový diagram s proloženou elipsou 95% konstantní hustoty pravděpodobnosti, s jehož pomocí posoudíme dvourozměrnou normalitu dat: Grafy – Bodové grafy – vypneme Typ proložení – Proměnné X, Y - OK . Na záložce Detaily vybereme Elipsa Normální – OK. Ve vzniklém dvourozměrném tečkovém diagramu změníme rozsah zobrazených hodnot na vodorovné a svislé ose, abychom viděli celou elipsu



Ze vzhledu diagramu je patrné, že předpoklad dvourozměrné normality je oprávněný a že mezi loňskou a letošní poptávkou existuje vcelku silná přímá lineární závislost.

Testování hypotézy o nezávislosti: Statistika – Základní statistiky /Tabulky - Korelační matice – OK – 2 seznamy proměnných X, Y, OK. Na záložce Možnosti zaškrtneme Zobrazit detailní tabulku výsledků – Souhrn.

Prom. X & prom. Y	Korelace (Tabulka1) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD) Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	110,0000	85,3229									
Y	140,0000	111,1755	0,971977	0,944739	8,269474	0,001167	6	0,686813	1,266484	5,566343	0,745955

Ve výstupní tabulce najdeme hodnotu výběrového korelačního koeficientu R_{12} ($r = 0{,}971977$, tzn. že mezi X a Y existuje velmi silná přímá lineární závislost), realizaci testové statistiky $t = 8{,}269474$ a p-hodnotu pro test hypotézy o nezávislosti ($p = 0{,}001167$, H_0 tedy zamítáme na hladině významnosti 0,05).

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X - OK – OK – Výpočet: Výsledky regrese.

N=6	Výsledky regrese se závislou proměnnou : Y (Tabulka1) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219 Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

Ve výstupní tabulce najdeme koeficient b_0 ve sloupci B na řádku označeném Abs. člen, koeficient b_1 ve sloupci B na řádku označeném X. Rovnice regresní přímky:
y = 0,686813 + 1,266484 x.
Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

Efekt	Analýza rozptylu (Tabulka1)				
	Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	58384,89	1	58384,89	68,38420	0,001167
Rezid.	3415,11	4	853,78		
Celk.	61800,00				

Odhad rozptylu najdeme na řádku Rezid., ve sloupci Průměr čtverců, tedy $s^2 = 853,78$.
Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2. V našem případě $ID^2 = 0,9447$, tedy variabilita letošní poptávky je z 94,5% vysvětlena regresní přímkou.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry) a hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry). Do Dlouhého jména proměnné dm resp. hm napíšeme: =v3-v4*VStudent(0,975;4) resp. =v3+v4*VStudent(0,975;4)

N=6	Výsledky regrese se závislou proměnnou : Y (Tabulka1) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219							
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p	dm =v3-v4*V	hm =v3+v4*
Abs.člen			0,686813	20,64236	0,033272	0,975052	-56,6256	57,99918
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167	0,841266	1,691701

Vidíme, že $-56,63 < \beta_0 < 58$ s pravděpodobností aspoň 0,95 a $0,841 < \beta_1 < 1,692$ s pravděpodobností aspoň 0,95.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

Testovou statistiku F-testu a odpovídající p-hodnotu najdeme v záhlaví výstupní tabulky regrese. Zde F = 68,384, p-hodnota < 0,00117, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku. (Výsledky F-testu jsou rovněž uvedeny v tabulce ANOVA.)

f) Na hladině významnosti 0,05 proveďte dílčí t-testy a vypočtěte relativní chyby odhadů regresních parametrů.

Výsledky dílčích t-testů jsou uvedeny ve výstupní tabulce regrese. Testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je 0,033272, p-hodnota je 0,975052. Hypotézu o nevýznamnosti úseku regresní přímky tedy nezamítáme na hladině významnosti 0,05. Testová statistika pro test hypotézy H_0: $\beta_1 = 0$ je 8,269474, p-hodnota je 0,001167. Hypotézu o nevýznamnosti směrnice regresní přímky tedy zamítáme na hladině významnosti 0,05.
K upravené výstupní tabulce s mezemi intervalů spolehlivosti přidáme proměnnou chyba. Do jejího Dlouhého jména napíšeme
=100*abs(0,5*(hm-dm)/v3)

N=6	Výsledky regrese se závislou proměnnou : Prom2 (Tabulka1) R= ,97197702 R2= ,94473932 Upravené R2= ,93092415 F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219								
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p	dm =v3-v4*V	hm =v3+v4*	chyba =100*abs
Abs.člen			0,686813	20,64236	0,033272	0,975052	-56,6256	57,99918	8344,681
Prom1	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167	0,841266	1,691701	33,57463

Výsledek pro parametr β_0: Protože p = 0,975 < 0,05, hypotézu o nevýznamnosti regresního parametru β_0 (tj. posunutí regresní přímky) nezamítáme na hladině významnosti 0,05.
Výsledek pro parametr β_1: Protože p = 0,0012 < 0,05, hypotézu o nevýznamnosti regresního parametru β_1 (tj. směrnice regresní přímky) zamítáme na hladině významnosti 0,05.

g) Vypočtěte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

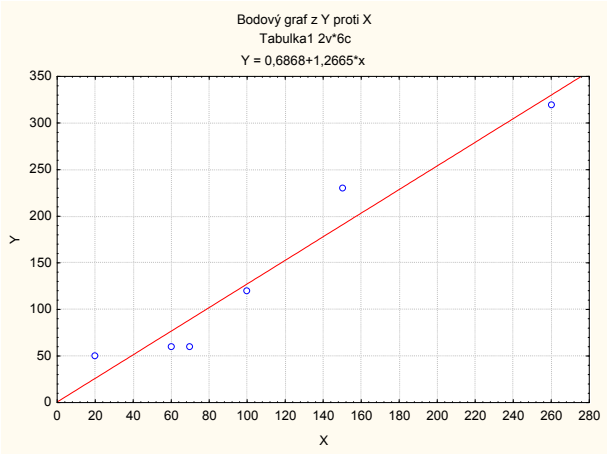
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi Předpovědi závisle proměnné X: 110 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď.

Proměnná	Předpovězené hodnoty (Tabulka1 proměnné: Y		
	B-váž.	Hodnota	B-váž. * Hodnot
X	1,266484	110,0000	139,3132
Abs. člen			0,6868
Předpověď			140,0000
-95,0%LS			106,8803
+95,0%LS			173,1197

Při loňské poptávce 110 kusů je predikovaná hodnota letošní poptávky 140 kusů.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Bodové grafy zvolíme Typ proložení: Lineární, OK.



i) Vypočtěte střední absolutní procentuální chybu predikce (MAPE)

Ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua / předpoklady / předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat vše – OK. Ve vzniklé tabulce odstraníme proměnné 5 – 10, přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme
=100*abs(v4/v2)
Pak spočteme průměr této proměnné a zjistíme, že MAPE = 25,17%.

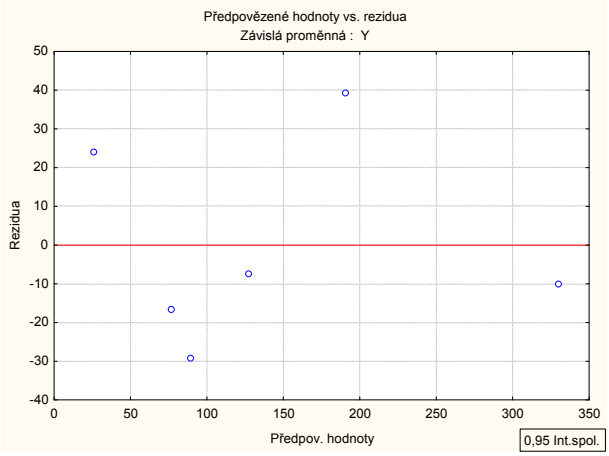
j) Proveďte analýzu reziduí.

Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbin-Watson.d	Sériové korelace
Odhad	2,022847	-0,113505

Hodnota této statistiky je blízká 2, svědčí o tom, že rezidua jsou nekorelovaná.

Posouzení homoskedasticity reziduí
Reziduální analýza – Bodové grafy – Předpovědi vs. rezidua



Rezidua jsou kolem 0 rozmístěna náhodně.

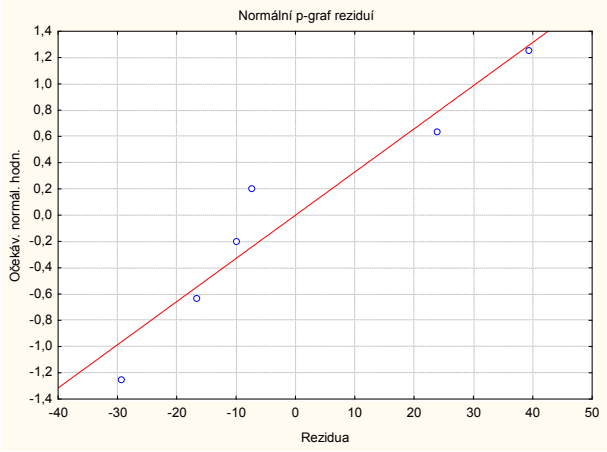
Testování nulovosti střední hodnoty reziduí:

Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000003	26,13469	6	10,66944	0,00	-0,000000	5	1,000000

Na hladině významnosti 0,05 nezamítáme hypotézu, že střední hodnota reziduí je 0.

Posouzení normality reziduí:
Na záložce Pravděpodobnostní grafy zvolíme Normální pravděpodobnostní graf reziduí:

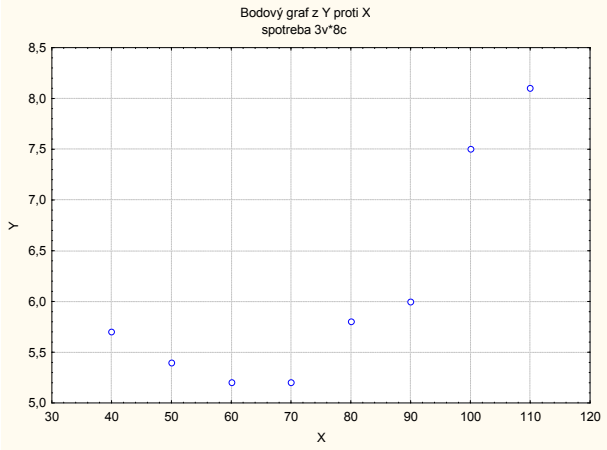


Rezidua se řadí kolem ideální přímky, lze tedy soudit, že se řídí normálním rozložení.

Úkol 2.: (Příklad je převzat z knihy Jiří Anděl: Matematická statistika, SNTL/Alfa, Praha, 1978, str. 111)
U automobilu Škoda 120 byla změřena spotřeba benzínu (v l/100 km) v závislosti na rychlosti (v km/h).

rychlost	40	50	60	70	80	90	100	110
spotřeba	5,7	5,4	5,2	5,2	5,8	6,0	7,5	8,1

a) Data znázorněte graficky dvourozměrným tečkovým diagramem a najděte vhodnou regresní funkci.
Načteme datový soubor spotreba_benzinu.sta se dvěma proměnnými X a Y a 8 případy.
Grafy – Bodové grafy – vypneme Typ proložení – Proměnné X, Y - OK



Z dvourozměrného tečkového diagramu je patrno, že vhodnou regresní funkcí bude parabola: $m(x;\beta_0,\beta_1,\beta_2)=\beta_0+\beta_1 x+\beta_2 x^2$.
K datovému souboru tedy přidáme novou proměnnou Xkv a do jejího Dlouhého jména napíšeme = X^2

	1 X	2 Y	3 Xkv
1	40	5,7	1600
2	50	5,4	2500
3	60	5,2	3600
4	70	5,2	4900
5	80	5,8	6400
6	90	6	8100
7	100	7,5	10000
8	110	8,1	12100

b) Vypočtěte odhady regresních parametrů a napište rovnici regresní paraboly.
Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnné X, Xkv - OK – OK – Výpočet: Výsledky regrese.

N=8	Výsledky regrese se závislou proměnnou : Y (spotreba) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p
Abs.člen			9,751786	0,945689	10,31183	0,000148
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905

Rovnice regresní paraboly: y = 9,751786 – 0,150536 x + 0,001244xkv

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

Vrátíme se do Výsledky – vícenásobná regrese – Detailní výsledky – ANOVA.

Efekt	Analýza rozptylu (spotreba)				
	Součet čtverců	sv	Průměr čtverců	F	Úroveň p
Regres.	8,064881	2	4,032440	76,40988	0,000179
Rezid.	0,263869	5	0,052774		
Celk.	8,328750				

Odhad rozptylu najdeme na řádku Rezid., ve sloupci Průměr čtverců, tedy $s^2 = 0{,}05277$.
Index determinace je uveden v záhlaví původní výstupní tabulky pod označením R2. V našem případě $ID^2 = 0{,}9683$, tedy variabilita spotřeby benzínu je z 96,8% vysvětlena regresní parabolou.

d) Určete 95 % intervaly spolehlivosti pro regresní parametry.

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti pro regresní parametry) a hm (pro horní meze 95% intervalů spolehlivosti pro regresní parametry). Do Dlouhého jména proměnné dm resp. hm napíšeme: =v3-v4*VStudent(0,975;5) resp. =v3+v4*VStudent(0,975;5)

N=8	Výsledky regrese se závislou proměnnou : Y (spotreba) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973							
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p	dm =v3-v4*	hm =v3+v4
Abs.člen			9,751786	0,945689	10,31183	0,000148	7,320815	12,18276
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483	-0,21948	-0,08159
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905	0,000788	0,0017

Vidíme, že
$7{,}320815 < \beta_0 < 12{,}18276$ s pravděpodobností aspoň 0,95,
$-0{,}21948 < \beta_1 < -0{,}08159$ s pravděpodobností aspoň 0,95,
$0{,}000788 < \beta_2 < 0{,}0017$ s pravděpodobností aspoň 0,95

e) Na hladině významnosti 0,05 proveďte celkový F-test.
Testovou statistiku F-testu a odpovídající p-hodnotu najdeme v záhlaví výstupní tabulky regrese. Zde F = 76,41, p-hodnota < 0,00018, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku. (Výsledky F-testu jsou rovněž uvedeny v tabulce ANOVA.)

f) Na hladině významnosti 0,05 proveďte dílčí t-testy a vypočtěte relativní chyby odhadů regresních parametrů.
Výsledky dílčích t-testů jsou uvedeny ve výstupní tabulce regrese.
Testová statistika pro test hypotézy H_0: $\beta_0 = 0$ je 10,31183, p-hodnota je 0,000148. Hypotézu o nevýznamnosti parametru β_0 tedy zamítáme na hladině významnosti 0,05.
Testová statistika pro test hypotézy H_0: $\beta_1 = 0$ je -5,61264, p-hodnota je 0,002483. Hypotézu o nevýznamnosti parametru β_1 tedy zamítáme na hladině významnosti 0,05.
Testová statistika pro test hypotézy H_0: $\beta_2 = 0$ je 7,01912, p-hodnota je 0,000905. Hypotézu o nevýznamnosti parametru β_2 tedy zamítáme na hladině významnosti 0,05.
K upravené výstupní tabulce s mezemi intervalů spolehlivosti přidáme proměnnou chyba. Do jejího Dlouhého jména napíšeme
=100*abs(0,5*(hm-dm)/v3)

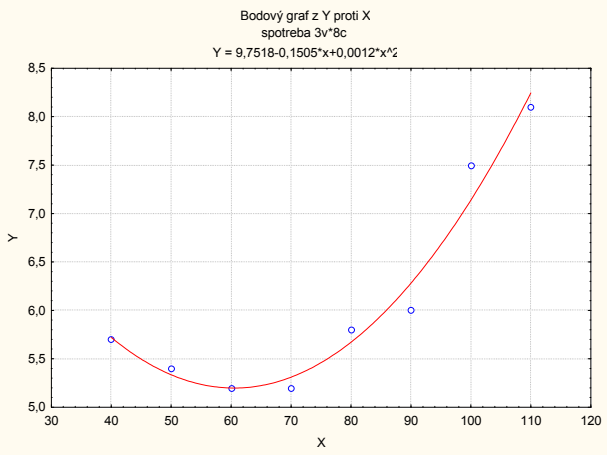
N=8	Výsledky regrese se závislou proměnnou : Y (spotreba) R= ,98403165 R2= ,96831829 Upravené R2= ,95564561 F(2,5)=76,410 p<,00018 Směrod. chyba odhadu : ,22973								
	Beta	Sm.chyba beta	B	Sm.chyba B	t(5)	Úroveň p	dm =v3-v4*	hm =v3+v4	chyba =100*a
Abs.člen			9,751786	0,945689	10,31183	0,000148	7,320815	12,18276	24,92847
X	-3,38045	0,602292	-0,150536	0,026821	-5,61264	0,002483	-0,21948	-0,08159	45,79987
Xkv	4,22756	0,602292	0,001244	0,000177	7,01912	0,000905	0,000788	0,0017	36,62259

Vidíme, že chyby odhadů jsou velké, v řádu desítek procent.

g) Určete regresní odhad spotřeby benzínu při rychlosti 80 km/h.
Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi - Předpovědi závisle proměnné X: 80, Xkv 6400 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď: 5,6708

h) Znázorněte data s proloženou regresní funkcí.

Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Bodové grafy zvolíme na záložce Detaily Typ proložení: Polynomiální, OK. Stupeň polynomu je implicitně nastaven na 2, lze změnit na záložce Možnosti 2.



i) Vypočtěte střední absolutní procentuální chybu predikce (MAPE)

Ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua/předpoklady/předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat X, Y – OK. Ve vzniklé tabulce odstraníme proměnné 5 – 10, přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme
=100*abs(v4/v2)
Pak spočteme průměr této proměnné a zjistíme, že MAPE = 2,15%.

Cvičení 14: Úvod do analýzy časových řad

Úkol 1.: Časová řada vyjadřuje počet obyvatelstva ČSSR (v tisících) v letech 1965 až 1974 vždy ke dni 31.12.

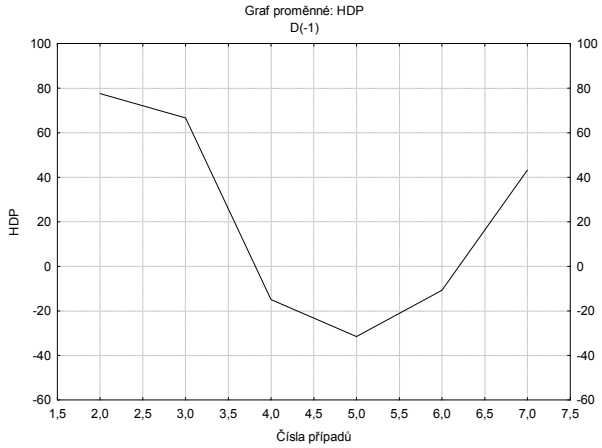
Rok	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
počet	14194	14271	14333	14387	14443	14345	14419	14576	14631	14738

Charakterizujte tuto časovou řadu chronologickým průměrem.

Návod: Načteme datový soubor obyvatele_CSSR.sta o 11 proměnných a jednom případu. Do Dlouhého jména poslední proměnné napíšeme
=(v1/2+sum(v2:v9)+v10/2)/9
Dostaneme výsledek 14430,11.

Úkol 2.: Pro časovou řadu HDP ČR v letech 1994 až 2000 (v miliardách Kč) vypočtěte základní charakteristiky dynamiky a graficky znázorněte relativní přírůstky a koeficienty růstu.

Návod: Načteme datový soubor HDP. sta**.**
Výpočet 1. diferencí: $\Delta y_i = y_i - y_{i-1}$ pro i = 2,...,n
Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné Y – OK – OK (transformace, autokorelace, kříž. korelace, grafy) – Oddělit-sloučit - OK (transformovat vybrané řady) – vykreslí se graf.



Vrátíme se do Transformace proměnných – Uložit proměnné. Otevře se nové datové okno, kde v proměnné HDP_1 jsou uloženy 1. diference.

	HDP	HDP_1
1	1303,600	
2	1381,100	77,500
3	1447,700	66,600
4	1432,800	-14,900
5	1401,300	-31,500
6	1390,600	-10,700
7	1433,800	43,200

Výpočet relativních přírůstků: $\delta_i = \frac{\Delta y_i}{y_{i-1}}$ pro i = 2,...,n

Vrátíme se do Transformace proměnných – označíme proměnnou, kterou chceme transformovat (HDP) – vybereme Posun – OK, (Transformovat vybrané řady) – vykreslí se graf.
Vrátíme se do Transformace proměnných – Uložit proměnné. Tato transformovaná veličina se uloží do tabulky pod názvem HDP_1 (proměnná s 1. diferencemi se přejmenuje na HDP_2).
Přidáme novou proměnnou RP a do jejího Dlouhého jména napíšeme vzorec =HDP_2/HDP_1.

Výpočet koeficientů růstu: $k_i = \frac{y_i}{y_{i-1}}$ pro i = 2,...,n

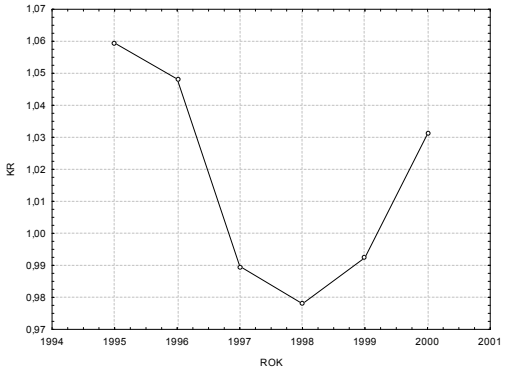
Do tabulky přidáme proměnnou KR a do jejího Dlouhého jména napíšeme vzorec =HDP/HDP_1.
Získáme tabulku

	HDP	HDP_2	HDP_1	RP	KR
1	1303,600				
2	1381,100	77,500	1303,600	0,059451	1,059451
3	1447,700	66,600	1381,100	0,048222	1,048222
4	1432,800	-14,900	1447,700	-0,010292	0,989708
5	1401,300	-31,500	1432,800	-0,02198	0,978015
6	1390,600	-10,700	1401,300	-0,00764	0,992364
7	1433,800	43,200	1390,600	0,031066	1,031066
8			1433,800		

Pomocí Grafy - 2D Grafy – Spojnicové grafy (Proměnné) vykreslíme průběh relativních přírůstků a koeficientů růstu.

Graf relativních přírůstků

Graf koeficientů růstu



Průměrný absolutní přírůstek a průměrný koeficient růstu vypočteme na kalkulačce pomocí vzorců $\overline{\Delta} = \frac{1433,8 - 1303,6}{6} = 21,7$ a $\overline{k} = \sqrt[6]{\frac{1433,8}{1303,6}} = 1,016$.

Úkol 3.: Je dána časová řada potratů (v tisících) v ČR v letech 1986 až 1996: 99,5 126,7 129,3 126,5 126,1 120,1 109,3 85,4 67,4 61,6 60.

Předpokládejte, že tato časová řada má kvadratický trend. Odhadněte parametry trendové funkce.

Vypočtěte index determinace ID^2.

Proveďte celkový F-test. (Popis celkového F- testu: Na hladině významnosti α testujeme H_0: $(\beta_1,\ldots,\beta_p)' = (0,\ldots,0)'$ proti H_1: $(\beta_1,\ldots,\beta_p)' \neq (0,\ldots,0)'$, přičemž p je počet odhadovaných regresních parametrů (bez parametru β_0)
(Nulová hypotéza říká, že dostačující je model konstanty.)
Testová statistika $F = \frac{S_R/p}{S_E/(n-p-1)}$ má rozložení F(p, n-p-1), pokud H_0 platí. Přitom

$S_E = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2$ je reziduální součet čtverců a $S_R = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2$ je regresní součet čtverců,

kde $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$.
Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.
$F \in W \Rightarrow H_0$ zamítáme na hladině významnosti α.

Proveďte dílčí t-testy. (Popis dílčích t-testů: Na hladině významnosti α pro j = 0,1, ..., p testujeme hypotézu
H_0: $\beta_j = 0$ proti H_1: $\beta_j \neq 0$.
Testová statistika: $T_j = \frac{b_j}{s_{b_j}}$ j má rozložení t(n-p-1), pokud H_0 platí. Přitom s_{b_j} je směrodatná chyba odhadu b_j.
Kritický obor: $W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty)$.
$T_j \in W \Rightarrow H_0$ zamítáme na hladině významnosti α.)

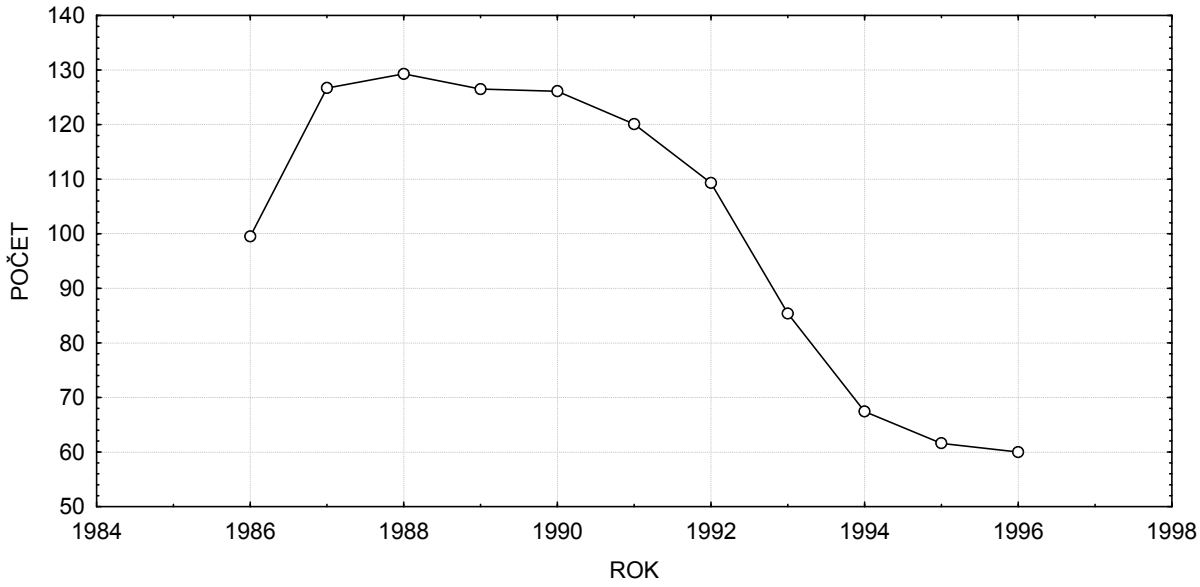
Ověřte normalitu reziduí.

Sestrojte 95% intervaly spolehlivosti pro parametry trendové funkce. (Vzorec pro meze 100(1- α)% intervalu spolehlivosti pro β_j: $b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$)

Stanovte střední absolutní procentuální chybu predikce (MAPE). MAPE se počítá podle vzorce $MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$.

Graficky znázorněte průběh časové řady s odhadnutým trendem, 95% pásem spolehlivosti a 95% predikčním pásem.

Návod:
Načteme datový soubor potraty.sta. Pro lepší orientaci znázorníme časovou řadu graficky.
Grafy – Bodové grafy – Proměnné X ROK, Y POCET – OK – vypneme Lineární proložení– OK.
Formát – Všechny možnosti – Graf: Obecné – zaškrtneme Spojnice – OK. Vznikne spojnicový diagram.



Trendová funkce $\hat{f}(t) = \beta_0 + \beta_1 t + \beta_2 t^2$

Odhady parametrů:

Statistiky – Vícenásobná regrese – Proměnné Závislé, Nezávislé t, tkv - OK

N=11	Výsledky regrese se závislou proměnnou : POCET (potraty.sta) R= ,94015284 R2= ,88388736 Upravené R2= ,85485920 F(2,8)=30,449 p<,00018 Směrod. chyba odhadu : 10,629					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(8)	Úroveň p
Abs.člen			103,2418	11,67235	8,84499	0,000021
t	1,30140	0,531476	10,9470	4,47060	2,44866	0,040020
tkv	-2,16020	0,531476	-1,4748	0,36285	-4,06453	0,003611

Odhadnutá trendová funkce má tedy tvar:

$\hat{f}(t) = 103{,}2418 + 10{,}947t - 1{,}4748t^2$, kde t = 1, …, 11.

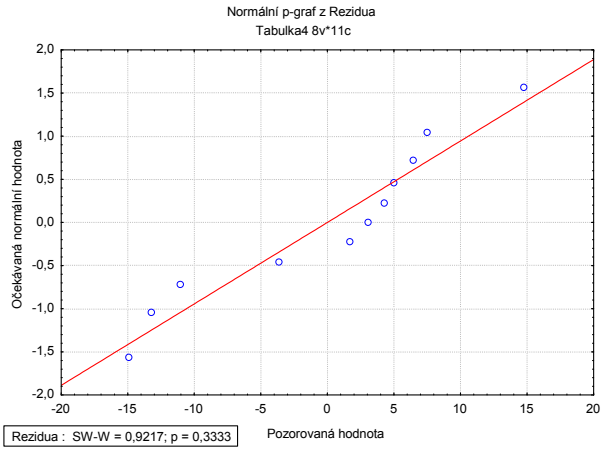
Index determinace je 0,884, což znamená, že kvadratická trendová funkce vysvětluje variabilitu dané časové řady z 88,4%.

Testová statistika celkového F-testu je 30,449, p-hodnota je blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku.

Všechny tři dílčí t-testy mají p-hodnoty menší než 0,05, tedy na hladině významnosti 0,05 zamítáme hypotézy o nulovosti parametrů β_0, β_1, β_2.

Ověření normality reziduí:

Na záložce Rezidua/předpoklady/předpovědi zvolíme Reziduální analýza – Uložit – Uložit rezidua & předpovědi. Sestrojíme N-P plot reziduí a současně provedeme S-W test:



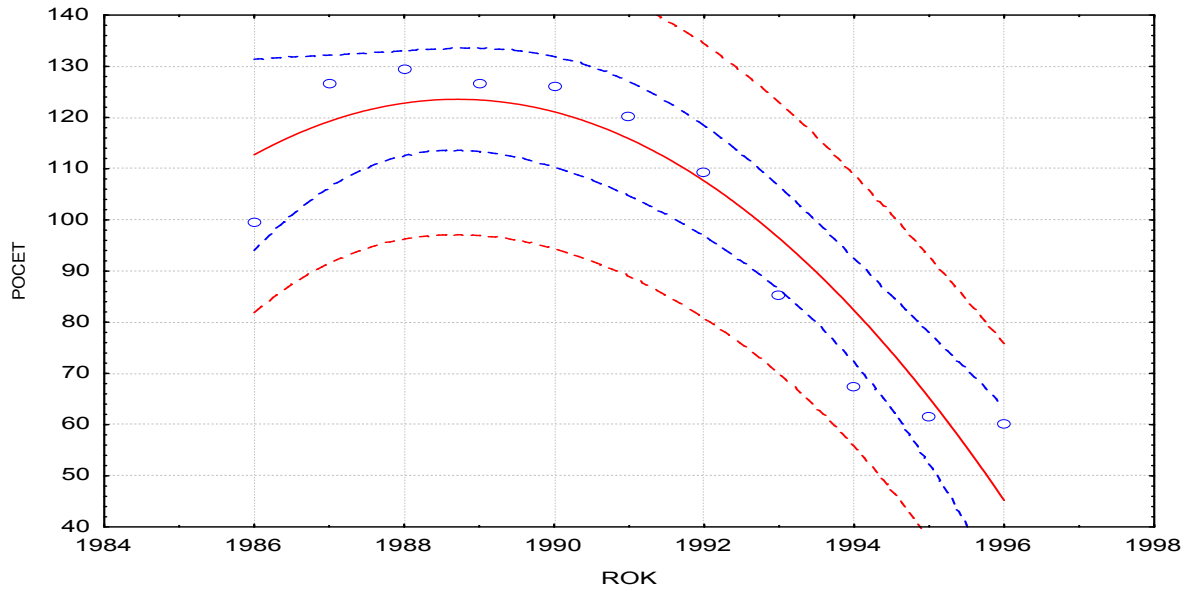
S-W test poskytuje p-hodnotu 0,333, tedy na hladině významnosti 0,05 nezamítáme hypotézu o normalitě reziduí.
Sestrojení 95% intervalů spolehlivosti pro parametry trendu:
Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti) a hm (pro horní meze 95% intervalů spolehlivosti). Do Dlouhého jména proměnné dm resp. hm napíšeme:
=v3-v4*VStudent(0,975;8) resp. =v3+v4*VStudent(0,975;8)

N=11	Výsledky regrese se závislou proměnnou : POCET (potraty.sta) R= ,94015284 R2= ,88388736 Upravené R2= ,85485920 F(2,8)=30,449 p<,00018 Směrod. chyba odhadu : 10,629							
	Beta	Sm.chyba beta	B	Sm.chyba B	t(8)	Úroveň p	dm =v3-v4*	hm =v3+v4
Abs.člen			103,2418	11,67235	8,84499	0,000021	76,32533	130,1583
t	1,30140	0,531476	10,9470	4,47060	2,44866	0,040020	0,637767	21,25622
tkv	-2,16020	0,531476	-1,4748	0,36285	-4,06453	0,003611	-2,31156	-0,63809

Vidíme, že $76,32 < \beta_0 < 130,16$ s pravděpodobností aspoň 0,95, $0,64 < \beta_1 < 21,26$ a $-2,31 < \beta_2 < -0,64$ s pravděpodobností aspoň 0,95.

Výpočet MAPE:
Ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua / předpoklady / předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat vše – OK. Ve vzniklé tabulce odstraníme proměnné 7 – 12, přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme =100*abs(v6/v2). Pak spočteme průměr této proměnné a zjistíme, že MAPE = 9,21%.

Graf časové řady s proloženým kvadratickým trendem získáme takto:
Grafy – Bodové grafy – Proměnné X ROK, Y POCET – OK – Detaily Proložení Polynomiální. Ve vytvořeném grafu 2x klikneme na pozadí, vybereme Graf: Regresní pásy – Přidat nový pár pásů – Typ Spolehlivostní – OK. Totéž provedeme ještě jednou a nyní zaškrtneme Typ Predikční.

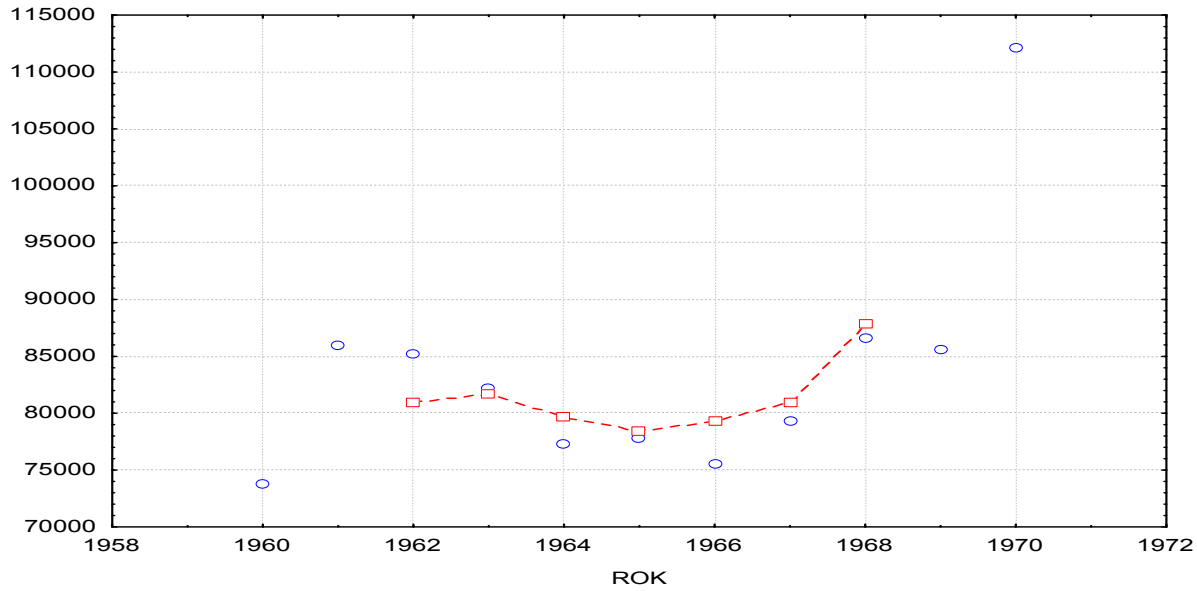


Úkol 4.: Máme k dispozici údaje o počtu bytů předaných do užívání v Československu v letech 1960 až 1970: 73 766 86 032 85 221 82 189 77 301 77 818 75 576 79 297 86 571 85 656 112 135. Odhadněte trend této časové řady pomocí klouzavých průměrů s vyhlazovacím okénkem šířky 5 a graficky znázorněte.

Návod:
Načteme datový soubor byty.sta o dvou proměnných ROK a POCET a jedenácti případech. Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné POCET – OK– OK (transformace, autokorelace, kříž. korelace, grafy) – Vyhlazování – zaškrtneme N-bod. klouzavý průměr, N = 5 – OK (Transformovat vybrané řady) – vykreslí se graf, vrátíme se do Transformace proměnných – Uložit proměnné. Otevře se nový spreadsheet, kde v proměnné POCET_1 jsou uloženy klouzavé průměry pro N = 5. Proměnnou POČET_1 okopírujeme do původního datového souboru do nové proměnné KP5 (pozor – roky 1960, 1961, 1969 a 1970 nemají přiřazený odhad).

	1 ROK	2 POCET	3 KP5
1	1960	73766	
2	1961	86032	
3	1962	85221	80901,8
4	1963	82189	81712,2
5	1964	77301	79621,0
6	1965	77818	78436,2
7	1966	75576	79312,6
8	1967	79297	80983,6
9	1968	86571	87847,0
10	1969	85656	
11	1970	112135	

Pomocí Grafy – Bodové grafy – Vícenásobný graf vytvoříme graf časové řady počtu bytů s odhadnutým trendem.



Příklady k samostatnému řešení:

Příklad 1.: V průběhu jednoho roku byla čtyřikrát provedena inventarizace skladových zásob. Určete průměrný stav zásob ve sledovaném roce, jestliže hodnoty jsou uvedeny v tabulce. Pro jednoduchost počítejte, že každý měsíc má 30 dní.

inventarizace	1	2	3	4
datum	2.1.	2.3.	12.9.	30.12.
zásoby [tis. Kč]	752	652	925	426

Výsledky: použijeme vážený chronologický průměr a zjistíme, že průměrný stav zásob je 739,9 [tis. Kč].

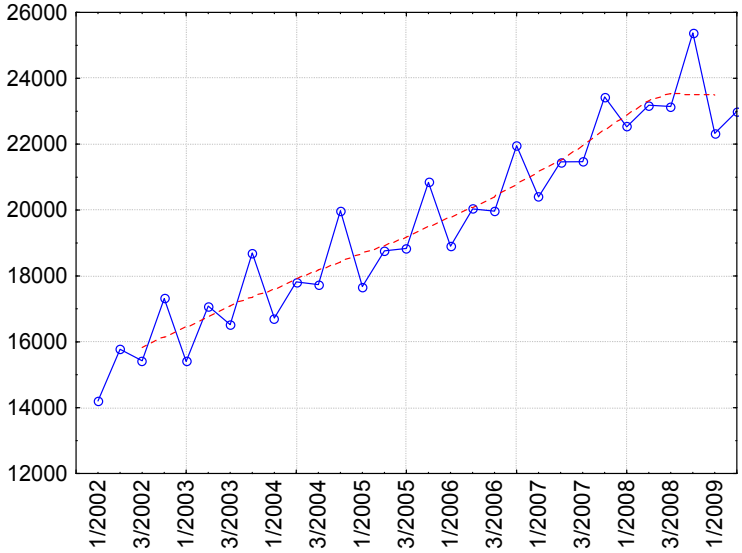
Příklad 2.: Máme k dispozici čtvrtletní časovou řadu průměrných měsíčních mezd v České republice v době od 1/2001 do 3/2009 (datový soubor ctvrtletni_ mzda.sta):

čas	mzda	čas	mzda	čas	mzda
1/2002	14204	4/2004	19980	3/2007	21470
2/2002	15772	1/2005	17678	4/2007	23435
3/2002	15422	2/2005	18763	1/2008	22531
4/2002	17315	3/2005	18833	2/2008	23182
1/2003	15407	4/2005	20841	3/2008	23144
2/2003	17084	1/2006	18903	4/2008	25381
3/2003	16522	2/2006	20036	1/2009	22328
4/2003	18697	3/2006	19968	2/2009	22992
1/2004	16722	4/2006	21952	3/2009	23350
2/2004	17817	1/2007	20399		
3/2004	17738	2/2007	21462		

a) Odhadněte trend této časové řady pomocí klouzavých průměrů s vyhlazovacím okénkem šířky 4.

b) Graficky znázorněte průběh časové řady s odhadnutým trendem.

Výsledky:



Vidíme, že díky vhodné volbě šířky vyhlazovacího okénka se podařilo odhadnout trend dané časové řady.