Kapitola 2.: Diagnostické grafy a testy normality dat

Cíl kapitoly

Po prostudování této kapitoly budete

- znát způsob konstrukce krabicového diagramu, normálního pravděpodobnostního grafu, kvantil-kvantilového grafu, histogramu a dvourozměrného tečkového diagramu a budete umět tyto grafy vytvořit v systému STATISTICA
- schopni pomocí těchto diagnostických grafů orientačně posoudit povahu dat
- umět v systému STATISTICA provádět testy normality dat

Časová zátěž

Na prostudování této kapitoly a splnění úkolů s ní spojených budete potřebovat asi 20 hodin studia.

2.1. Motivace

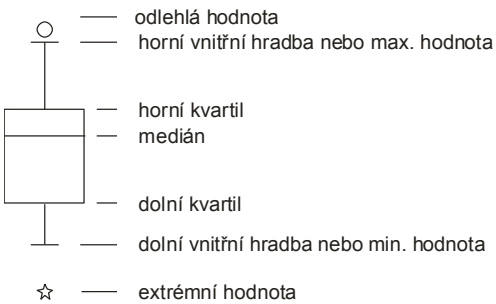
Diagnostické grafy slouží především k tomu, aby nám pomohly orientačně posoudit povahu dat a určit směr další statistické analýzy. Při zpracování dat se často předpokládá splnění určitých podmínek. V případě jednoho náhodného výběru je to především normalita (posuzujeme ji pomocí N-P plotu, Q-Q plotu, histogramu) a nepřítomnost vybočujících hodnot (odhalí je krabicový diagram). U dvou či více nezávislých náhodných výběrů sledujeme kromě normality též shodu středních hodnot nebo shodu rozptylů - homoskedasticitu (porovnáváme vzhled krabicových diagramů). V případě jednoho dvourozměrného náhodného výběru často posuzujeme dvourozměrnou normalitu dat (použijeme dvourozměrný tečkový diagram s proloženou $100(1-\alpha)\%$ elipsou konstantní hustoty pravděpodobnosti).

Vzhledem k důležitosti předpokladu normality se vedle grafického posouzení doporučuje též použití některého testu normality, např. Kolmogorovova – Smirnovova testu nebo Shapirova – Wilksova testu. K závěrům těchto testů však přistupujeme s určitou opatrností. Máme-li k dispozici rozsáhlejší datový soubor (orientačně $n > 30$) a test zamítne na obvyklé hladině významnosti 0,01 nebo 0,05 hypotézu o normalitě, i když vzhled diagnostických grafů svědčí jenom o lehkém porušení normality, nedopustíme se závažné chyby, pokud použijeme statistickou metodu založenou na normalitě dat.

2.2. Krabicový diagram

2.2.1. Popis diagramu

Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot. Způsob konstrukce je zřejmý z obrázku:



Odlehlá hodnota leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1{,}5q, x_{0,75} + 3q)$ či v intervalu $(x_{0,25} - 3q, x_{0,25} - 1{,}5q)$.
Extrémní hodnota leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q, \infty)$ či v intervalu $(-\infty, x_{0,25} - 3q)$.

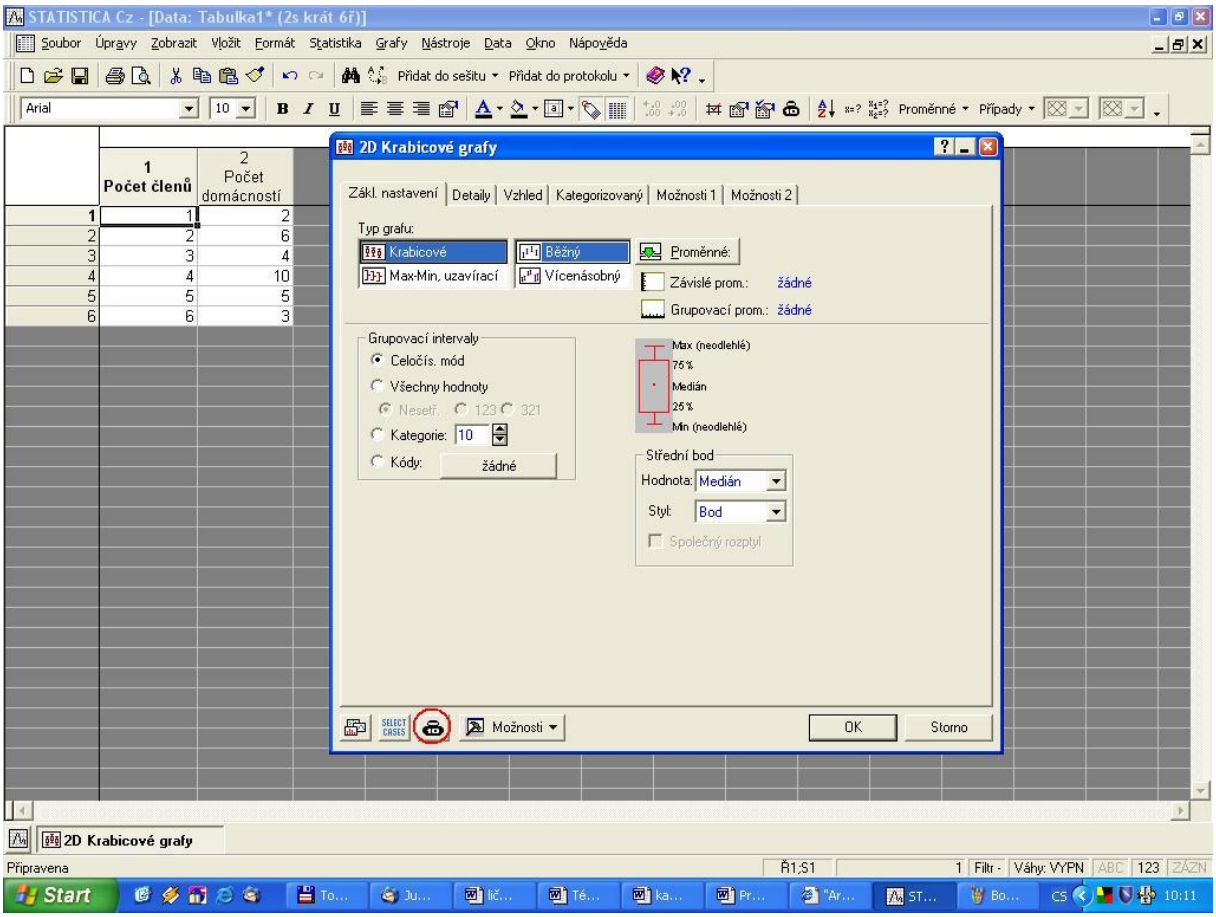
2.2.2. Příklad

U 30 domácností byl zjišťován počet členů.

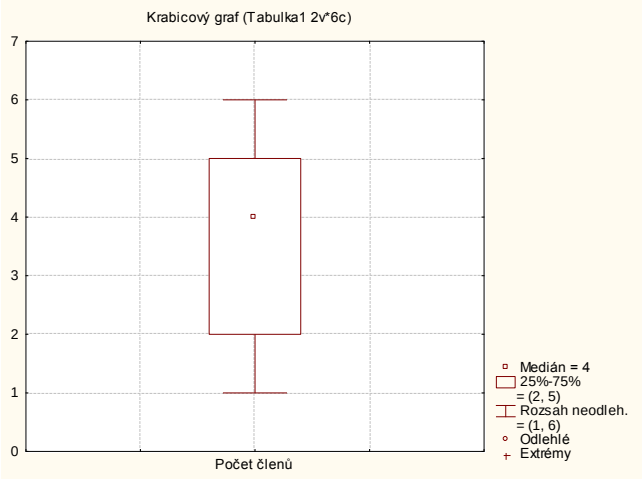
Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

Pro tyto údaje sestrojte krabicový diagram.

Řešení pomocí systému STATISTICA:
Data zapíšeme do datového okna programu STATISTICA. Po spuštění programu zadáme Soubor – Nový – Počet proměnných 2, Počet případů 6, OK. První proměnnou přejmenujeme na Počet členů, druhou na Počet domácností. (Přejmenování uskutečníme tak, že 2x klikneme myší na název proměnné a tím se otevře okno se specifikacemi proměnné.)
Vytvoření krabicového diagramu: Grafy – 2D Grafy – Krabicové grafy. Abychom systému STATISTICA sdělili, že pracujeme s údaji, pro které známe absolutní četnosti, klikneme myší na tlačítko s obrázkem závaží – na obrázku je v kroužku.



V okénku Váhy případů pro analýzu/graf zaškrtneme Status Zapnuto a zadáme Proměnná vah Počet domácností, OK. Na panelu 2D Krabicové grafy zadáme Proměnné – Závisle proměnné Počet členů, OK. Dostaneme krabicový diagram



Z obrázku lze vyčíst, že medián je 4 (aspoň polovina domácností má aspoň 4 členy), dolní kvartil 2 (aspoň čtvrtina domácností má aspoň 2 členy), horní kvartil 5 (aspoň tři čtvrtiny domácností mají aspoň 5 členů), minimum 1, maximum 6. Kvartilová odchylka je 5 – 2 = 3. Datový soubor vykazuje určitou nesymetrii – medián je posunut směrem k hornímu kvartilu, soubor je tedy záporně zešikmen. Odlehlé ani extrémní hodnoty se nevyskytují.

2.3. Normální pravděpodobnostní graf (N-P plot)

Před popisem tohoto grafu se musíme seznámit s pojmem pořadí čísla v posloupnosti čísel: Nechť $x_1, \ldots, x_n$ je posloupnost reálných čísel.
a) Jsou-li čísla navzájem různá, pak pořadím R_i čísla x_i rozumíme počet těch čísel $x_1, \ldots, x_n$, která jsou menší nebo rovna číslu x_i.
b) Vyskytují-li se mezi danými čísly skupinky stejných čísel, pak každé takové skupince přiřadíme průměrné pořadí.

2.3.1. Příklad

a) Jsou dána čísla 9, 4, 5, 7, 3, 1.
b) Jsou dána čísla 6, 7, 7, 9, 6, 10, 8, 6, 6, 9.
Stanovte pořadí těchto čísel.

Řešení

ad a)

usp. čísla	1	3	4	5	7	9
pořadí	1	2	3	4	5	6

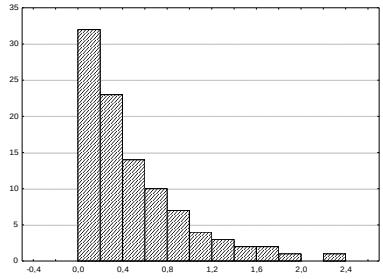
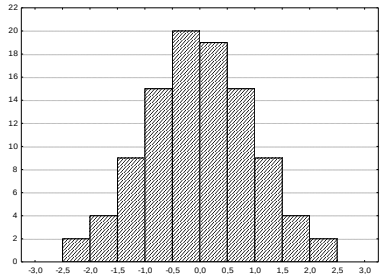
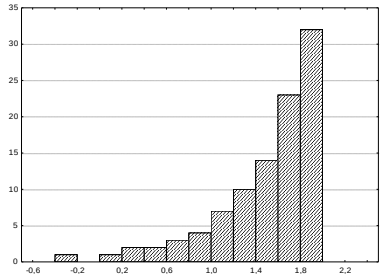
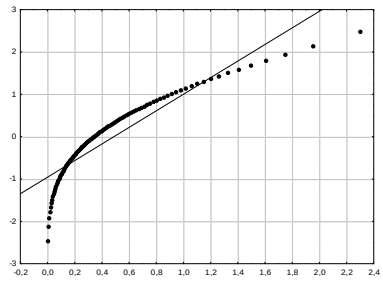
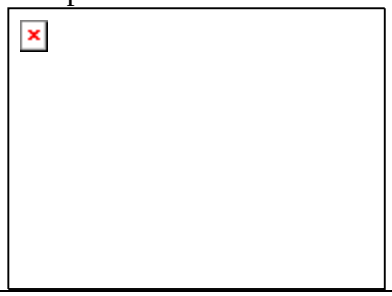
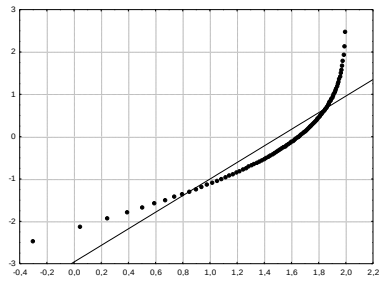
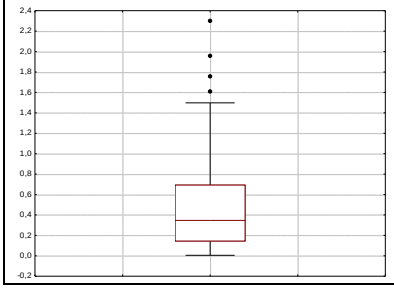
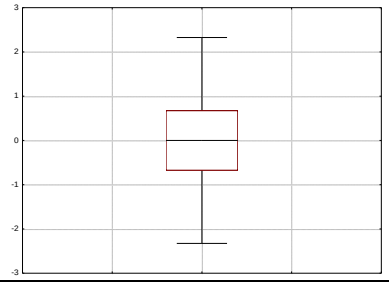
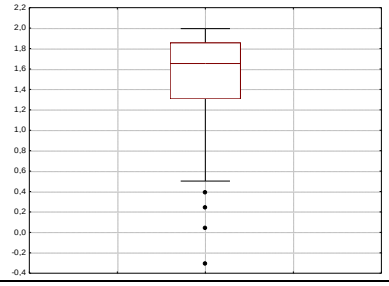
ad b)

usp. čísla	6	6	6	6	7	7	8	9	9	10
pořadí	1	2	3	4	5	6	7	8	9	10
prům. pořadí	2,25	2,25	2,25	2,25	5,5	5,5	7	8,5	8,5	10

2.3.2. Popis grafu

N-P plot umožňuje graficky posoudit, zda data pocházejí z normálního rozložení. Způsob konstrukce: na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \le ... \le x_{(n)}$ a na svislou osu kvantily u_{α_j}, kde $\alpha_j = \frac{3j-1}{3n+1}$, přičemž j je pořadí j-té uspořádané hodnoty (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince). Pocházejí-li data z normálního rozložení, pak všechny dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou ležet na přímce.

Pro data z rozložení s kladnou šikmostí se dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou řadit do konkávní křivky, zatímco pro data z rozložení se zápornou šikmostí se dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ budou řadit do konvexní křivky.

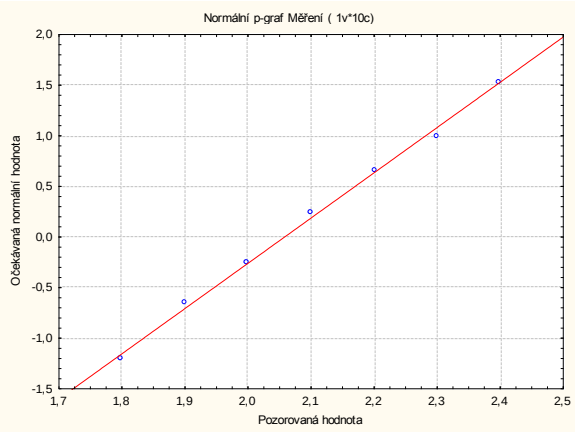
Rozložení s kladnou šikmostí	Normální rozložení	Rozložení se zápornou šikmostí
Histogram	Histogram	Histogram
		
N-P plot	N-P plot	N-P plot
		
Krabicový diagram	Krabicový diagram	Krabicový diagram
		

2.3.3. Příklad

Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Pomocí normálního pravděpodobnostního grafu posuďte, zda se tato data řídí normálním rozložením.

Řešení:

Po zapsání dat do proměnné nazvané Měření zvolíme Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné Měření, OK.



Protože dvojice $\left(x_{(j)}, u_{\alpha_j}\right)$ téměř leží na přímce, lze usoudit, že data pocházejí z normálního rozložení.

2.4. Kvantil-kvantilový graf (Q-Q plot)

2.4.1. Popis grafu

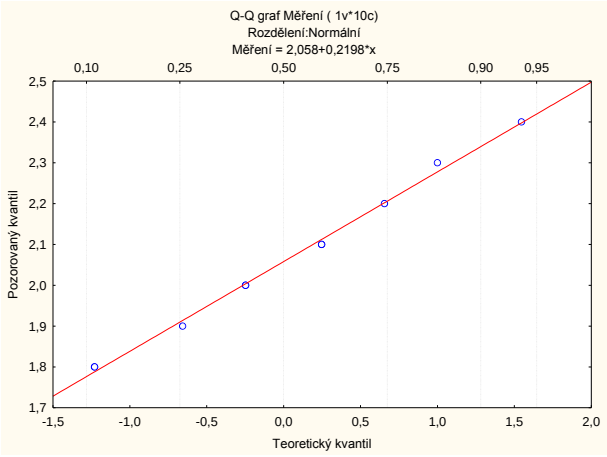
Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení (např. systém STATISTICA nabízí 8 typů rozložení: beta, exponenciální, Gumbelovo, gamma, log-normální, normální, Rayleighovo a Weibulovo). Pro nás je nejdůležitější právě normální rozložení.

Způsob konstrukce: na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \le ... \le x_{(n)}$ a na vodorovnou osu kvantily $K_{\alpha_j}(X)$ vybraného rozložení, kde $\alpha_j = \frac{j - r_{adj}}{n + n_{adj}}$, přičemž r_{adj} a n_{adj} jsou korigující faktory $\le 0{,}5$, implicitně $r_{adj} = 0{,}375$ a $n_{adj} = 0{,}25$. (Jsou-li některé hodnoty $x_{(1)} \le ... \le x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.) Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadnou z dat nebo je může zadat uživatel. Body $\left(K_{\alpha_j}(X), x_{(j)}\right)$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

2.4.2. Příklad

Pro data z příkladu 3.3.3. posuďte pomocí kvantil – kvantilového grafu, zda pocházejí z normálního rozložení.

Řešení:
Zvolíme Grafy – 2D Grafy – Grafy typu Q-Q – ponecháme implicitní nastavení na normální rozložení (pokud bychom chtěli změnit nastavení na jiný typ rozložení, zvolili bychom ho na záložce Detaily) – Proměnné Měření, OK.



Vzhled grafu nasvědčuje tomu, že data pocházejí z normálního rozložení.

2.5. Histogram

2.5.1. Popis grafu

Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. (Ve STATISTICE je pojem histogramu širší, skrývá se za ním i sloupkový diagram.)

Způsob konstrukce ve STATISTICE: na vodorovnou osu se vynášejí třídicí intervaly (implicitně 10, jejich počet lze změnit, stejně tak i meze třídicích intervalů) či varianty znaku a na svislou osu absolutní nebo relativní četnosti třídicích intervalů či variant. Do histogramu se může zakreslit tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení. Kromě osmi typů rozložení uvedených u Q-Q plotu umožňuje STATISTICA použít ještě další čtyři rozložení: Laplaceovo, logistické, geometrické, Poissonovo.

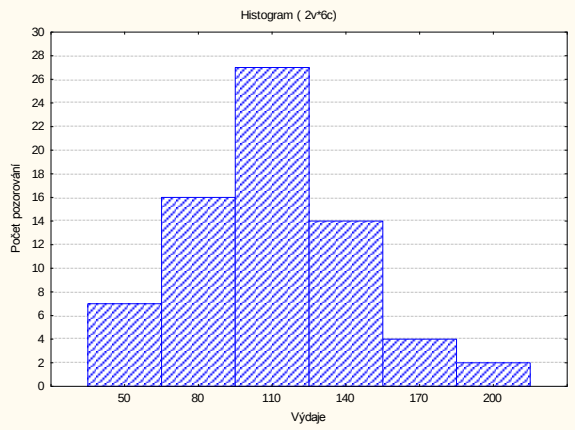
2.5.2. Příklad

U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	$\langle 35,65\rangle$	$\langle 65,95\rangle$	$\langle 95,125\rangle$	$\langle 125,155\rangle$	$\langle 155,185\rangle$	$\langle 185,215\rangle$
Počet dom.	7	16	27	14	4	2

Nakreslete histogram

Řešení pomocí systému STATISTICA:
Vytvoříme nový datový soubor s dvěma proměnnými Výdaje a Počet domácností. Do proměnné Výdaje zapíšeme středy třídicích intervalů, do proměnné Počet domácností odpovídající absolutní četnosti třídicích intervalů. V menu zvolíme Grafy – Histogramy – pomocí tlačítka s obrázkem závaží zadáme proměnnou vah Počet domácností – OK, Proměnná Výdaje – zapneme volbu Všechny hodnoty – OK. Dostaneme histogram:



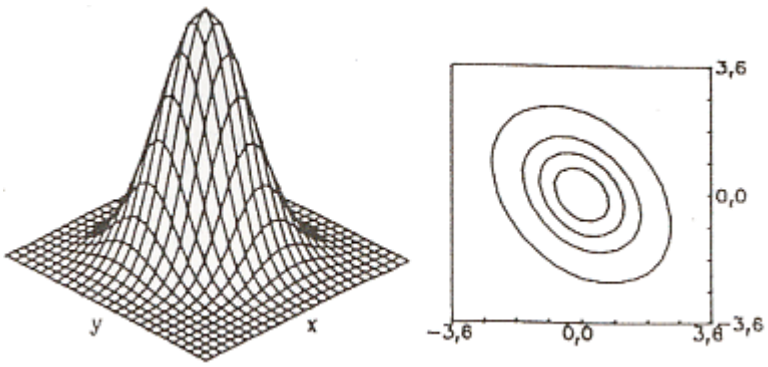
Vidíme, že tvar histogramu není symetrický. Malé hodnoty jsou četnější než velké – datový soubor je kladně zešikmen.

2.6. Dvourozměrný tečkový diagram

2.6.1. Popis diagramu

Máme dvourozměrný datový soubor $(x_1, y_1), \ldots , (x_n, y_n)$, který je realizací dvourozměrného náhodného výběru $(X_1, Y_1), \ldots , (X_n, Y_n)$ z dvourozměrného rozložení. Na vodorovnou osu vyneseme hodnoty x_j , na svislou hodnoty y_k a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dvojice (x_j, y_k). Jedná-li se o náhodný výběr z dvourozměrného normálního rozložení, měly by tečky zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy – viz následující obrázek.

Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0{,}75$:



Do dvourozměrného tečkového diagramu můžeme ještě zakreslit $100(1-\alpha)\%$ elipsu konstantní hustoty pravděpodobnosti. Bude-li více než $100\alpha\%$ teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi veličinami X a Y existuje určitý stupeň přímé resp. nepřímé lineární závislosti.

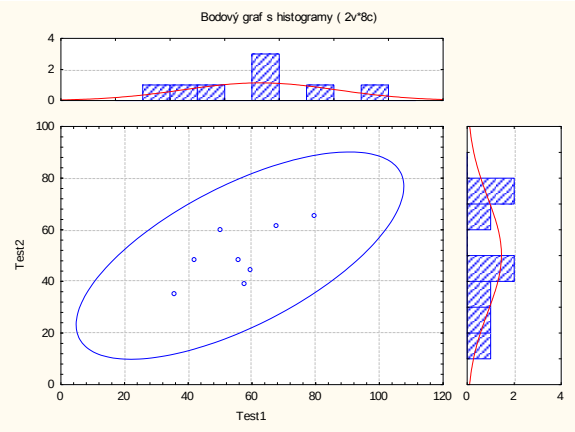
2.6.2. Příklad

Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Pomocí dvourozměrného tečkového diagramu se zakreslenou 95% elipsou konstantní hustoty pravděpodobnosti a histogramy pro počty bodů v 1. a 2. testu posuďte, zda tato data lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení.

Řešení pomocí systému STATISTICA:
Vytvoříme nový datový soubor se dvěma proměnnými Test1 a Test2 a osmi případy. Nyní nakreslíme dvourozměrný tečkový diagram: Grafy – 2D Grafy - Bodové grafy s histogramy. V typu proložení pro bodový graf vypneme lineární proložení. Proměnné – X – Test1, Y – Test2 – OK. Dostaneme dvourozměrný tečkový diagram pro vektorovou proměnnou (Test1, Test2) a histogramy pro Test1 a Test2. Nyní do diagramu zakreslíme 95% elipsu konstantní hustoty pravděpodobnosti: 2x klikneme na pozadí grafu a otevře se okno s názvem Vš. možnosti. Vybereme Graf: Elipsa, zvolíme Přidat novou elipsu. Po vykreslení elipsy změníme měřítko: na vodorovné ose bude minimum 0, maximum 120, na svislé ose bude minimum 0, maximum 100. (Stačí 2x kliknout na číselný popis osy a na záložce Měřítka vybrat manuální mód.)



Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti, tzn., že u studentů, kteří měli vysoký resp. nízký počet bodů v 1. testu, lze očekávat vysoký resp. nízký počet bodů ve 2. testu.

2.7. Testy normality dat

K ověřování normality dat slouží celá řada testů, které jsou podrobně popsány ve statistické literatuře. Zde se omezíme na dva testy, které jsou implementovány v systému STATISTICA, a to Kolmogorovův – Smirnovův test a Shapirův – Wilksův test. V systému STATISTICA lze hypotézu o normalitě testovat také pomocí testu dobré shody, kterým se budeme zabývat v 11. kapitole.

2.7.1. Kolmogorovův – Smirnovův test a jeho Lilieforsova varianta

Testujeme 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. Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem $\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}$.

V případě, že neznáme parametry μ a σ^2 normálního rozložení, musíme je odhadnout z dat (střední hodnotu odhadneme pomocí m a rozptyl pomocí s^2). Tím se změní rozložení testové statistiky D_n. Příslušné modifikované kvantily byly určeny pomocí simulačních studií. V této situaci používáme Lilieforsovu variantu Kolmogorovova – Smirnovova testu.

2.7.2. Shapirův – Wilksův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z normálního rozložení $N(\mu, \sigma^2)$.
Testová statistika má tvar:

$$W = \frac{\sum_{i=1}^{m} a_i^{(n)} \left[X_{(n-i+1)} - X_{(i)}\right]^2}{\sum_{i=1}^{m} (X_i - M)^2},$$

kde $m = n/2$ pro n sudé a $m = (n-1)/2$ pro n liché. Koeficienty $a_i^{(n)}$ jsou tabelovány.

Na testovou statistiku W lze pohlížet jako na korelační koeficient mezi uspořádanými pozorováními a jim odpovídajícími kvantily standardizovaného normálního rozložení.
V případě, že data vykazují perfektní shodu s normálním rozložením, bude mít W hodnotu 1. Hypotézu o normalitě tedy zamítneme na hladině významnosti α, když se na této hladině neprokáže korelace mezi daty a jim odpovídajícími kvantily rozložení N(0,1).
Lze také říci, že S – W 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 v systému STATISTICA je implementováno jeho rozšíření i na výběry velkých rozsahů, kolem 2000.)

2.7.3. Příklad

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí K- S testu a S – W testu zjistěte na hladině významnosti 0,05, zda tato data pocházejí z normálního rozložení.

Řešení pomocí systému STATISTICA:
Vytvoříme nový datový soubor o jedné proměnné nazvané X a pěti případech. Do proměnné X zapíšeme uvedené hodnoty. V menu vybereme Statistika – Základní statistiky/tabulky – Tabulky četností – OK, Proměnné X – OK. Na záložce zvolíme Normalita a zaškrtneme Lilieforsův test a Shapiro – Wilksův W test – Testy normality.

	Testy normality (Tabulka1)				
Proměnná	N	max D	Lilliefors p	W	p
X	5	0,22408	p > .20	0,91240	0,48215

Vidíme, že testová statistika K-S testu je d = 0,22409, odpovídající Lilieforsova p-hodnota je větší než 0,2, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.
Testová statistika S-W testu je W = 0,9124, odpovídající p-hodnota je 0,48215, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

2.8. Vzorový příklad

Zadání příkladu:

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.

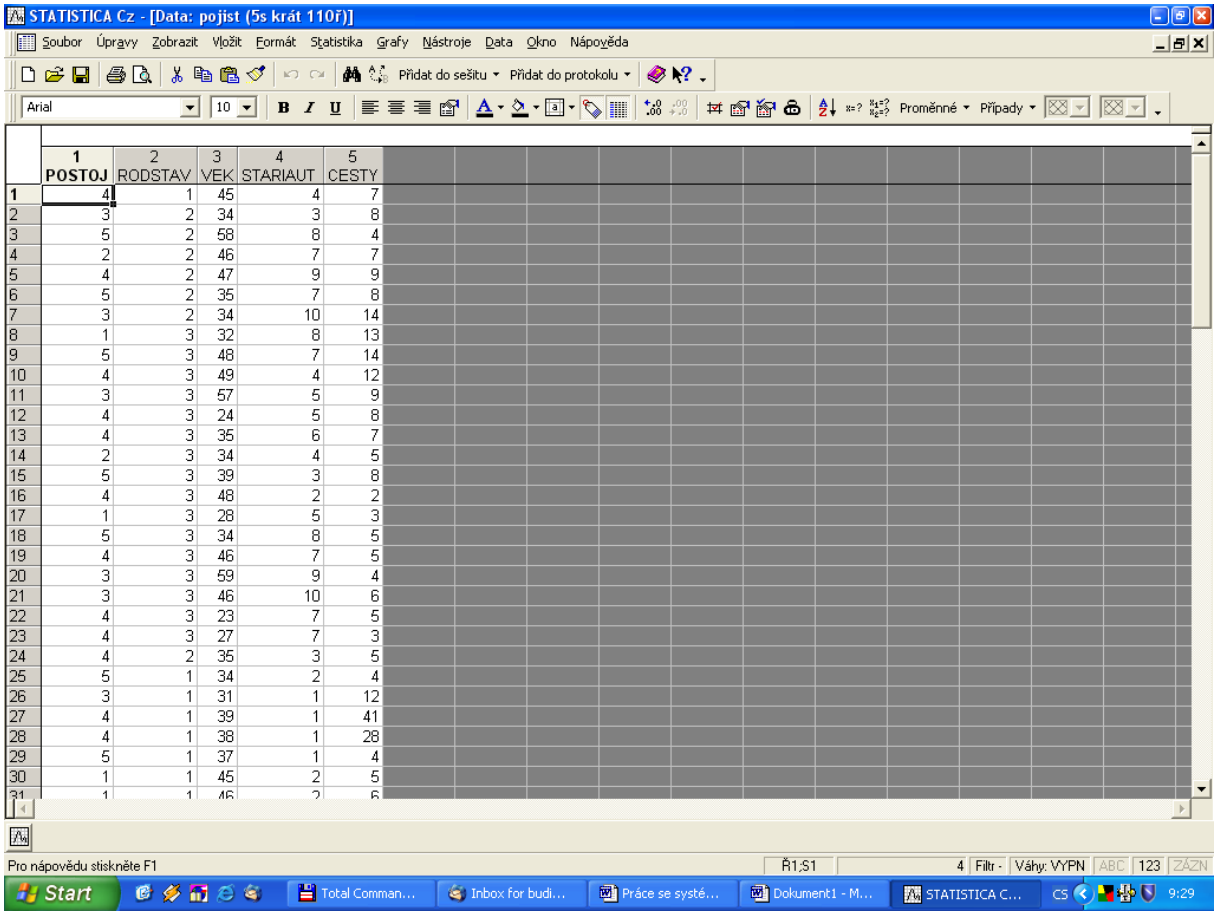
Náhodně bylo vybráno 110 současných zákazníků pojišťovny a ti byli telefonicky seznámeni s následujícím textem:

„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 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:

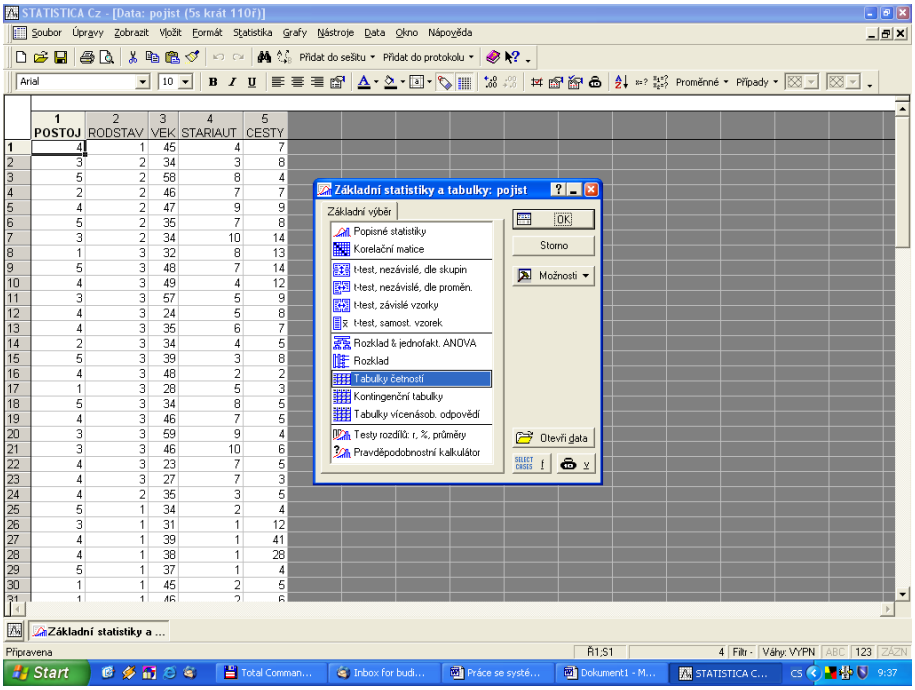


	1 POSTOJ	2 RODSTAV	3 VEK	4 STARIAUT	5 CESTY
1	4	1	45	4	7
2	3	2	34	3	8
3	5	2	58	8	4
4	2	2	46	7	7
5	4	2	47	9	9
6	5	2	35	7	8
7	3	2	34	10	14
8	1	3	32	8	13
9	5	3	48	7	14
10	4	3	49	4	12
11	3	3	57	5	9
12	4	3	24	5	8
13	4	3	35	6	7
14	2	3	34	4	5
15	5	3	39	3	8
16	4	3	48	2	2
17	1	3	28	5	3
18	5	3	34	8	5
19	4	3	46	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

Úkol 1. Zjistěte absolutní a relativní četnosti a absolutní a relativní kumulativní četnosti proměnných POSTOJ a RODSTAV.

Návod:

V menu zvolíme položku Statistika – Základní statistiky/tabulky – Tabulky četností – OK.



Pro analýzu vybereme proměnné POSTOJ, RODSTAV – OK. Zvolíme Výpočet: Tabulky četností. Získáme tabulku četností pro POSTOJ

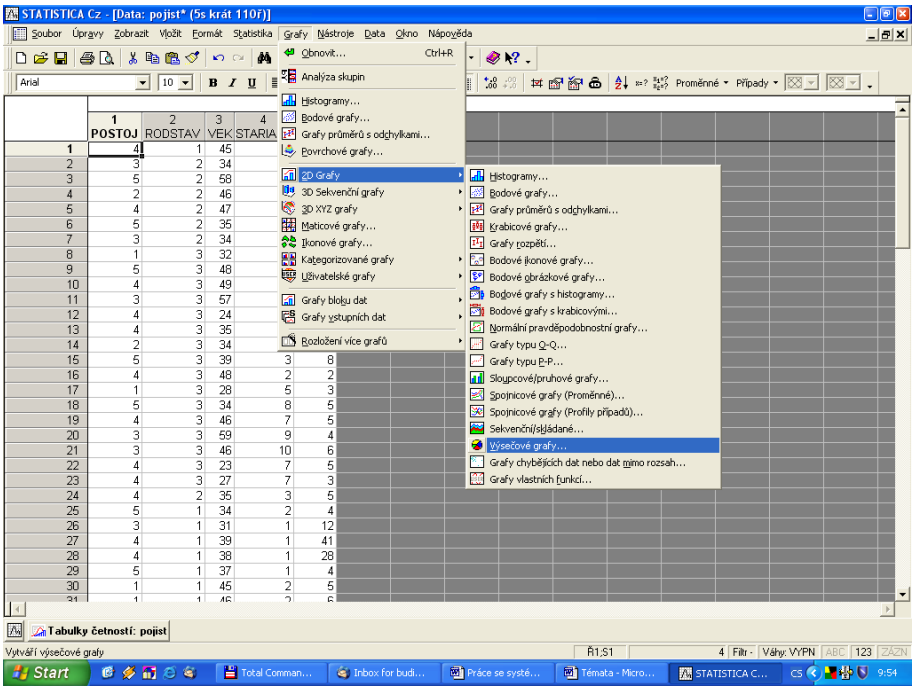
	Tabulka četností:POSTOJ: Postoj k novému typ			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
ednoznačný nezájem	8	8	7,27273	7,2727
ehký nezájem	21	29	19,09091	26,3636
eutrální postoj	23	52	20,90909	47,2727
ehký zájem	34	86	30,90909	78,1818
ednoznačný zájem	24	110	21,81818	100,000(
ChD	0	110	0,0000(	100,000(

a pro RODSTAV

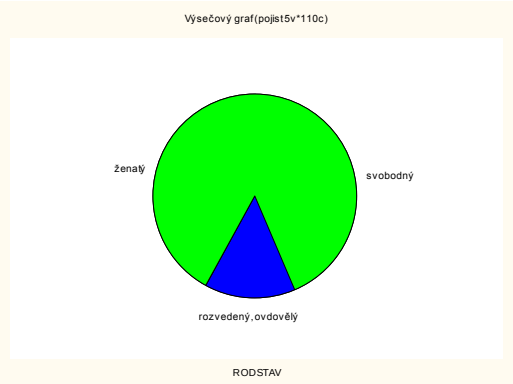
	Tabulka četností:RODSTAV: Rodinný stav (pojis			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
svobodný	48	48	43,6363(	43,636
rozvedený, ovdovělý	16	64	14,5454	58,1818
ženatý	46	110	41,81818	100,000(
ChD	0	110	0,0000(	100,000(

Úkol 2. Absolutní četnosti proměnných POSTOJ a RODSTAV znázorněte graficky pomocí výsečového diagramu.

Návod:
V menu zvolíme Grafy – 2D grafy – Výsečové grafy.



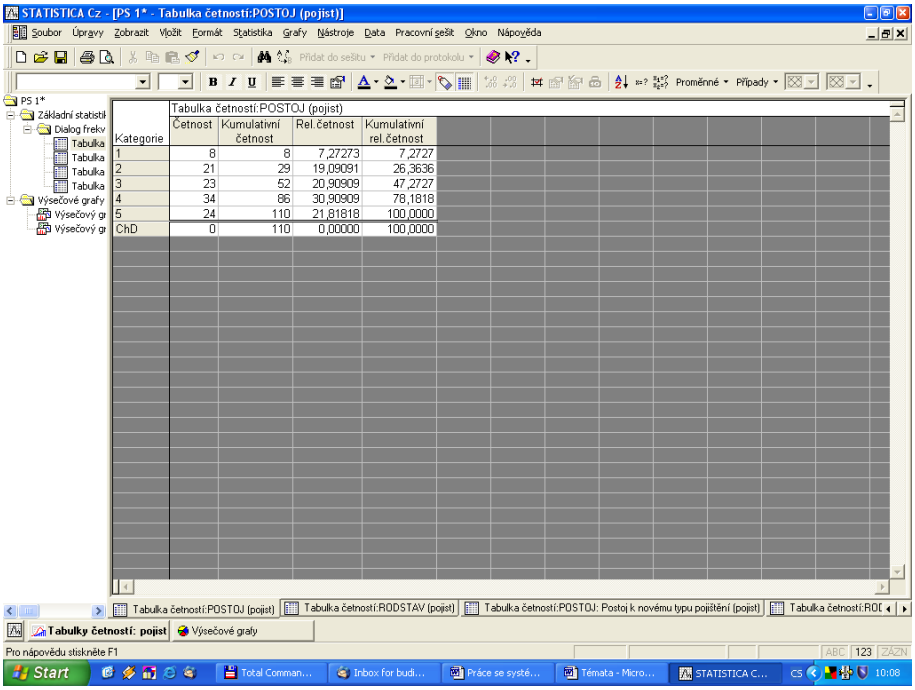
Vybereme proměnné POSTOJ, RODSTAV a dostaneme následující grafy:



Z prvního diagramu je zřejmé, že nejméně zákazníků projevilo jednoznačný nezájem 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.

Všechny tabulky a grafy se ukládají do pracovního sešitu. Listovat v nich lze pomocí stromové struktury v levém okně.



Úkol 3. Vypočtěte následující číselné charakteristiky:
a) POSTOJ (ordinální proměnná) – modus, medián, dolní a horní kvartil, kvartilová odchylka.
b) RODSTAV (nominální proměnná) – modus.
c) VEK, STARIAUT, CESTY (poměrové proměnné) – průměr, směrodatná odchylka, šikmost, špičatost.

Návod:
ad a) Statistika – Základní statistiky/tabulky – Popisné statistiky – Proměnné POSTOJ – OK. Na záložce Detaily vybereme Medián, Modus, Dolní & horní kvartily, Kvartilové rozpětí – Souhrn. Dostaneme tabulku

	Popisné statistiky (pojist)					
Proměnná	Medián	Modus	Četnost modu	Spodní kvartil	Horní kvartil	Kvartilové rozpětí
POSTOJ	4,00000(	4,00000(	34	2,00000(	4,00000(	2,00000(

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

ad b) V tabulce Popisné statistiky změníme proměnnou na RODSTAV – OK. Na záložce Detaily vybereme Modus – Souhrn. Dostaneme tabulku

	Popisné statistiky (	
Proměnná	Modus	Četnost modu
RODSTAV	1,00000(	48

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

ad c) V tabulce Popisné statistiky změníme proměnné na VEK, STARIAUT, CESTY – OK. Na záložce Detaily vybereme Průměr, Směrodat. odchylka, Šikmost, Špičatost – Souhrn. Dostaneme tabulku

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

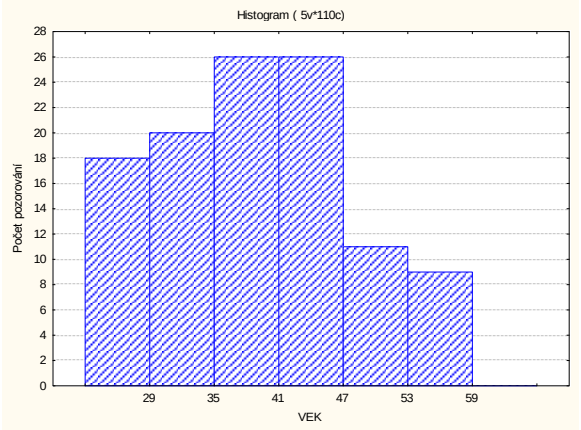
Průměrný věk zákazníků je 39,6 roku, směrodatná odchylka věku činí 8,8 roku. Rozložení věku vykazuje kladnou šikmost (podprůměrné hodnoty věku jsou četnější než nadprůměrné) a zápornou špičatost (rozložení věku je plošší než normální rozložení).
Průměrné stáří auta je 4,2 roku se směrodatnou odchylkou 2,4 roku. Rozložení stáří aut je kladně zešikmené a špičatější než normální rozložení.
Průměrný počet cest v předešlém roce činil 7,2 se směrodatnou odchylkou 5,3. Rozložení počtu cest je značně kladně zešikmené a podstatně špičatější než normální rozložení.
Poznámka: Pokud bychom chtěli porovnat variabilitu uvedených tří proměnných, mohli bychom vypočítat koeficienty variace (koeficient variace je podíl směrodatné odchylky a průměru). Do tabulky s vypočítanými číselnými charakteristikami přidáme další proměnnou nazvanou CV: Proměnné – Přidat – Kolik 1 – Za Špičatost – Jméno CV – do okénka Dlouhé jméno napíšeme =v2/v1 – OK. Dostaneme tabulku

Proměnná	Popisné statistiky (pojist)				
	Průměr	Sm. odch.	Šikmost	Špičatost	CV
VEK	39,58182	8,823844	0,191625	-0,59532	0,222927
STARIAUT	4,16364	2,359938	0,905405	0,35924	0,566797
CESTY	7,16364	5,304537	3,150711	15,99807	0,740481

Vidíme, že nejvyšší variabilitu má proměnná CESTY, nejnižší VEK.

Úkol 4. Vytvořte histogram proměnné VEK se šesti třídicími intervaly $\langle 23,29\rangle, (29,35\rangle, (35,41\rangle, (41,47\rangle, (47,53\rangle, (53,59\rangle$.

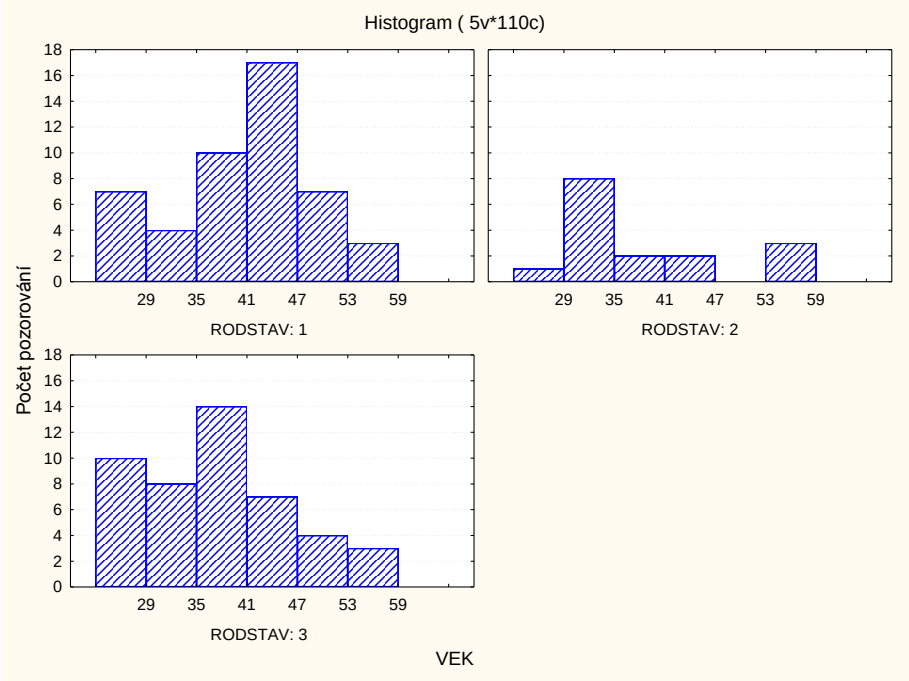
Návod:
V menu vybereme Grafy – Histogramy – Proměnné VEK, OK. Odškrtneme Typ proložení: Normální. V záložce Detaily vybereme Hranice – Určit hranice – zadáme horní meze intervalů, tj. 29 35 41 47 53 59, OK, OK.
Dostaneme histogram v tomto tvaru:



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.

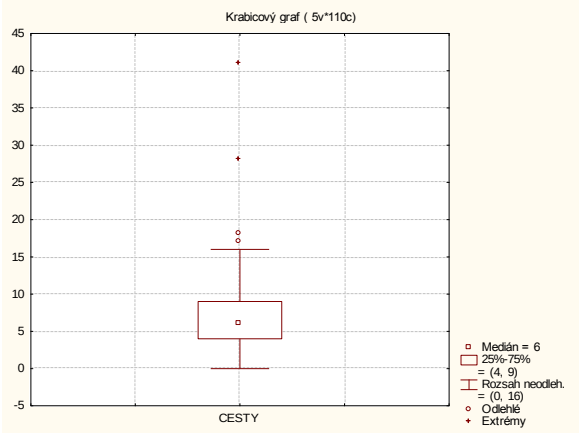
Úkol 5. Vytvořte kategorizovaný histogram proměnné VEK podle proměnné RODSTAV.

Návod
Postupujeme stejně jako v předešlém případě, jenom na záložce Kategorizovaný zvolíme Kategorie X – Zapnuto, Změnit proměnnou – RODSTAV, OK, OK Dostaneme tři histogramy:



Úkol 6. Sestrojte krabicový diagram proměnné CESTY. S jeho pomocí zjistěte, zda proměnná CESTY obsahuje odlehlé či extrémní hodnoty.

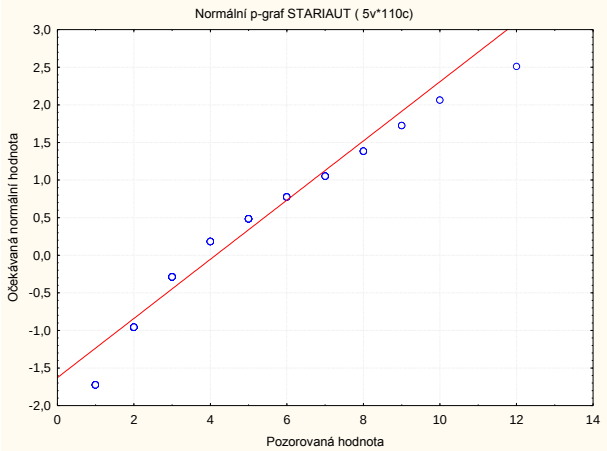
Návod:
V menu Grafy zvolíme 2D Grafy – Krabicové grafy – Proměnné – Závisle proměnné – CESTY – OK, OK.



Medián je posunut k dolnímu kvartilu, což svědčí o kladně zešikmeném rozložení. Vyskytují se odlehlé i extrémní hodnoty, jedná se tedy o špičaté rozložení.

Úkol 7. 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.

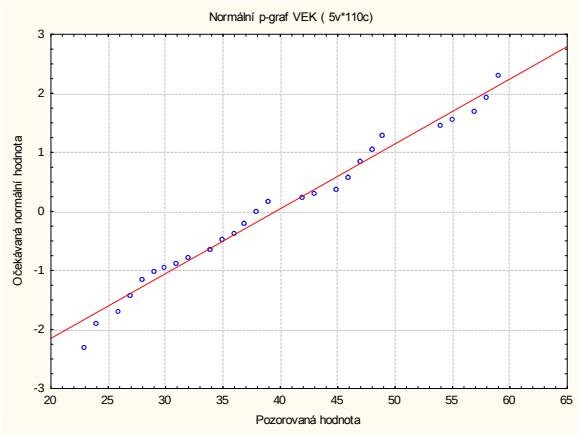


Tečky v NP grafu se značně odchylují od zakreslené přímky a řadí se do konkávního tvaru. Datový soubor vykazuje kladné zešikmení, nejedná se tedy o normální rozložení.

Úkol 8. Rozhodněte pomocí K-S testu a S-W testu na hladině významnosti 0,05, zda lze údaje o věku zákazníků považovat za realizace náhodného výběru z normálního rozložení.

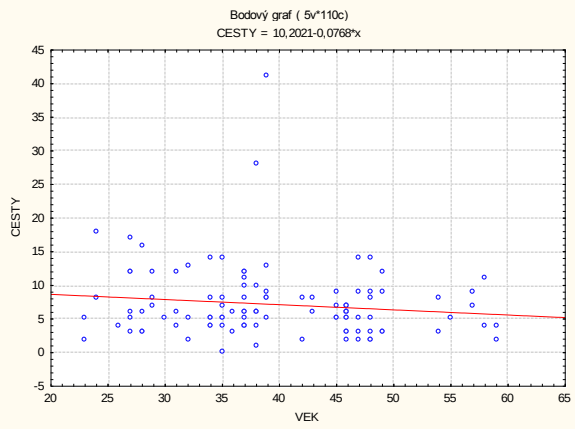
Návod:
Statistika – Základní statistiky/tabulky – Tabulky četností – OK, Proměnné X – OK. Na záložce zvolíme Normalita a zaškrtneme Lilieforsův test a Shapiro – Wilksův W test – Testy normality

Ve výstupu se objeví tabulka, v níž je uvedena hodnota testové statistiky pro K-S test (d = 0,11222) a S-W test (W = 0,96695) a odpovídající p-hodnoty. U K-S testu uvažujeme Lilieforsovo p, které je počítáno na základě parametrů odhadnutých z dat. V našem případě $p < 0,01$ a pro S-W test $p = 0,00783$, tedy oba testy zamítají na hladině významnosti 0,05 hypotézu o normalitě. Výpočet je vhodné doplnit NP grafem:



Úkol 9. Pomocí dvourozměrného tečkového diagramu posuďte, zda mezi věkem zákazníka a počtem cest nad 300 km v předešlém roce existuje nějaká lineární závislost.

Návod:
Grafy – Bodové grafy – Proměnné X – VEK, Y – CESTY – OK. OK. Dostaneme tento graf:



Vidíme, že s rostoucím věkem zákazníka poněkud klesá počet cest, mezi proměnnými VEK a CESTY tedy dosti slabá nepřímá lineární závislost.

Shrnutí

Při určení směru statistické analýzy dat používáme *diagnostické grafy*, které umožní posoudit

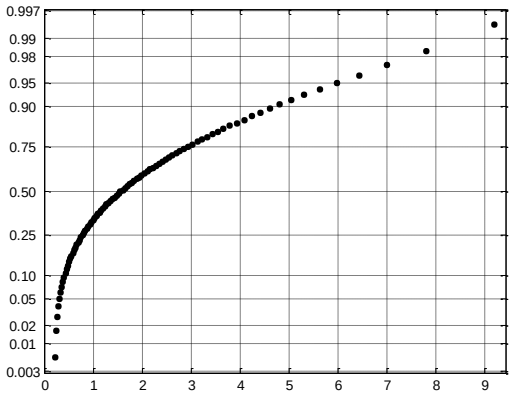
- normalitu dat či tvar rozložení (*N-P plot, Q- Q plot, histogram*)
- existenci odlehlých či extrémních hodnot (*krabicový graf*)
- dvourozměrnou normalitu dat (*dvourozměrný tečkový diagram*)

Kromě grafického znázornění dat používáme *testy normality dat*, např. *Kolmogorovův – Smirnovův test* (ve většině reálných situací jeho variantu poskytující *Lilieforsovu p-hodnotu*) nebo *Shapirův – Wilksův* test. Musíme si být ovšem vědomi toho, že pro výběry větších rozsahů (orientačně n > 30) i malé odchylky od normality mohou být statisticky významné, i když věcně nikoliv. V takovém případě se nedopustíme závažné chyby, pokud použijeme metodu založenou na předpokladu normality dat.

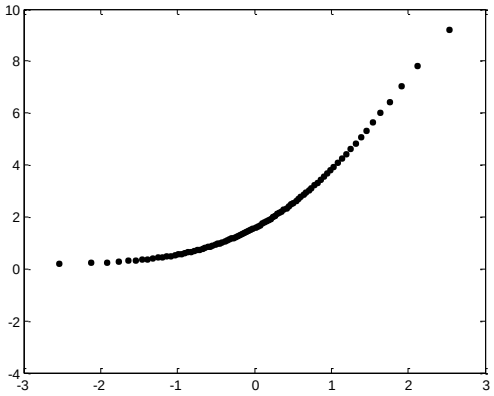
Kontrolní otázky

1. K čemu slouží diagnostické grafy?
2. Popište způsob konstrukce krabicového diagramu.
3. Jak budete interpretovat situaci, kdy v krabicovém diagramu je medián posunut směrem k dolnímu kvartilu?
4. V dvourozměrném tečkovém diagramu jsou tečky zhruba rovnoměrně rozptýleny uvnitř kruhového obrazce. Co lze říci o vztahu veličin X a Y?
5. Jak se liší provedení K-S testu normality dat v případě, kdy známe parametry normálního rozložení od případu, kdy je neznáme?
6. Jak souvisí S-W test normality dat s kvantil-kvantilovým grafem?
7. Pro datový soubor o rozsahu n = 50 byl vytvořen normální pravděpodobnostní graf a kvantil-kvantilový graf. Pomocí těchto grafů posuďte, zda se data mohou řídit normálním rozložením.

NP plot



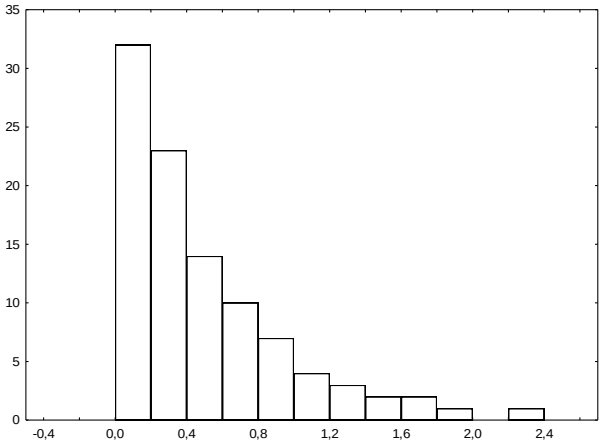
Q-Q plot

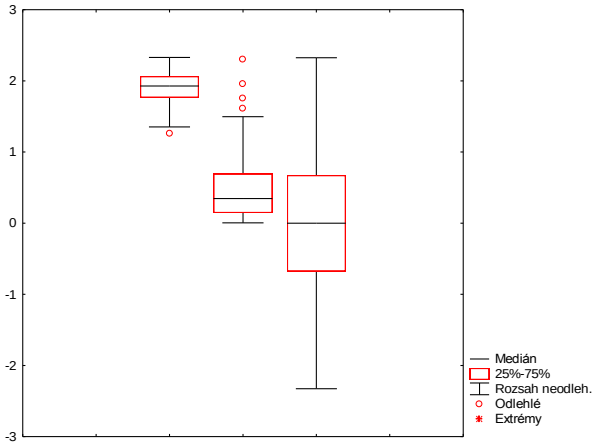


Výsledek: Data nepocházejí z normálního rozložení, vzhled obou diagramů svědčí o značném kladném zešikmení.

Autokorekční test

1. Z 99 hodnot byl sestrojen histogram. Určete, který ze tří uvedených krabicových diagramů byl sestrojen ze stejných hodnot.





a) První krabicový diagram.
b) Druhý krabicový diagram.
c) Třetí krabicový diagram.

2. Určete, která tvrzení jsou pravdivá:
a) Odlehlá hodnota v datovém souboru leží za vnějšími hradbami.
b) Extrémní hodnota v datovém souboru leží mezi vnitřními a vnějšími hradbami.
c) Extrémní hodnota je více vzdálena od mediánu než odlehlá hodnota.

3. Určete, která tvrzení jsou pravdivá:
a) Pocházejí-li data z normálního rozložení, budou se tečky v normálním pravděpodobnostním grafu řadit do přímky.
b) Pocházejí-li data z rozložení s kladnou šikmostí, budou se tečky v normálním pravděpodobnostním grafu řadit do konvexní křivky.

c) Pocházejí-li data z rozložení se zápornou šikmostí, budou se tečky v normálním pravděpodobnostním grafu řadit do konkávní křivky.

4. Určete, která tvrzení jsou pravdivá:
a) Pokud se v dvourozměrném tečkovém diagramu seskupují tečky do elipsovitého útvaru, jehož hlavní osa je přímka s kladnou směrnicí, lze usoudit, že mezi veličinami X a Y existuje určitý stupeň přímé lineární závislosti.
b) Pokud se v dvourozměrném tečkovém diagramu seskupují tečky do kruhovitého útvaru, lze usoudit, že mezi veličinami X a Y existuje určitý stupeň nelineární závislosti.
c) Pokud v dvourozměrném tečkovém diagramu leží všechny tečky na přímce se zápornou směrnicí, lze usoudit, že mezi veličinami X a Y existuje úplná nepřímá lineární závislost.

Správné odpovědi: 1b) 2c) 3a) 4a), c)

Příklady

1. Během semestru se studenti podrobili písemnému testu z matematiky, v němž bylo možno získat 0 až 10 bodů. Výsledky jsou uvedeny v tabulce:

Počet bodů	0	1	2	3	4	5	6	7	8	9	10
Počet studentů	1	4	6	7	11	15	19	17	12	6	3

Pro počet bodů sestrojte krabicový diagram. Je počet bodů symetricky rozložen kolem mediánu? Vyskytují se v datech odlehlé nebo extrémní hodnoty?
Výsledek: $x_{0,25} = 1$, $x_{0,50} = 6$, $x_{0,75} = 7$, medián je posunut k hornímu kvartilu, data vykazují zápornou šikmost. Odlehlé ani extrémní hodnoty se nevyskytují.

2. Pro počet bodů z 1. příkladu sestrojte normální pravděpodobnostní graf.

3. Pro počet bodů z 1. příkladu sestrojte kvantil-kvantilový graf pro normální rozložení.

4. Pro počet bodů z 1. příkladu testujte pomocí K-S testu na hladině významnosti 0,05 hypotézu, že se řídí normálním rozložením. Zjistěte hodnotu testové statistiky a odpovídající p-hodnotu.
Výsledek:
Testová statistika = 0,12895, Liliefors $p < 0,01$, hypotézu o normalitě zamítáme na hladině významnosti 0,05.

5. Pro počet bodů z 1. příkladu testujte pomocí S-W testu na hladině významnosti 0,05 hypotézu, že se řídí normálním rozložením. Zjistěte hodnotu testové statistiky a odpovídající p-hodnotu.
Výsledek:
Testová statistika = 0,96906, $p < 0,01784$, hypotézu o normalitě zamítáme na hladině významnosti 0,05.

6. Na 10 automobilech stejného typu se testovaly dva druhy benzínu lišící se oktanovým číslem. U každého automobilu se při průměrné rychlosti 90 km/h měřil dojezd (tj. dráha, kterou ujede na dané množství benzínu) při použití každého z obou druhů benzínu. Výsledky:

číslo auta	1	2	3	4	5	6	7	8	9	10
benzín A	17,5	20	18,9	17,9	16,4	18,9	17,2	17,5	18,5	18,2
benzín B	17,8	20,8	19,5	18,3	16,6	19,5	17,5	17,9	19,1	18,6

Pro uvedená data sestrojte dvourozměrný tečkový diagram se zakreslenou 95% elipsou konstantní hustoty pravděpodobnosti. Mohou data pocházet z dvourozměrného normálního rozložení?
Výsledek: ano.