Kapitola 3.: Diagnostické grafy a testy normality dat

3.1. Motivace

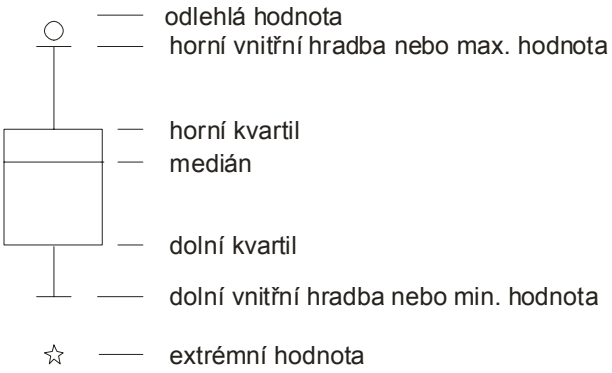
Diagnostické grafy slouží především k tomu, aby nám pomohly orientačně posoudit povahu dat a určit směr další statistické analýzy. Při zpracování dat se často předpokládá splnění určitých podmínek. V případě jednoho náhodného výběru je to především normalita (posuzujeme ji pomocí NP plotu, Q-Q plotu, 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-α)% 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 – Wilkova 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.

3.2. Krabicový diagram

3.2.1. 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 $(-\infty,\ x_{0,25} - 3q)$.

3.2.2. 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:

Připomeneme nejprve definici α-kvantilu. Je-li $\alpha \in (0;1)$, pak α-kvantil x_α je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl 1 – α všech dat. Pro výpočet α-kvantilu slouží algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo } c \Rightarrow x_\alpha = \dfrac{x_{(c)} + x_{(c+1)}}{2} \\ \text{necelé číslo} \Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } c \Rightarrow x_\alpha = x_{(c)} \end{cases}$$

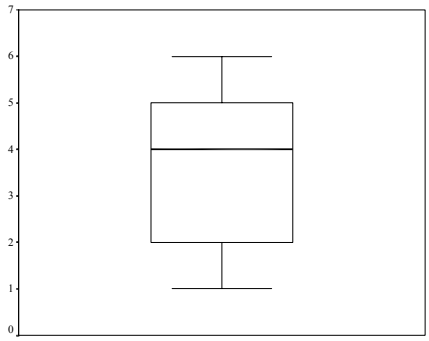
Pro speciálně zvolená α užíváme názvů: $x_{0,50}$ – medián, $x_{0,25}$ – dolní kvartil, $x_{0,75}$ – horní kvartil, $x_{0,1}$, ..., $x_{0,9}$ – decily, $x_{0,01}$, ..., $x_{0,99}$ – percentily. Jako charakteristika variability slouží kvartilová odchylka: $q = x_{0,75} - x_{0,25}$.

V našem případě rozsah souboru n = 30. Výpočty potřebných kvantilů uspořádáme do tabulky.

α	nα	c		x_α
0,25	7,5	8	$x_{(c)}=x_{(8)}$	2
0,50	15	15	$\dfrac{x_{(15)} + x_{(16)})}{2}$	4
0,75	22,5	23	$x_{(c)}=x_{(23)}$	5

Dolní kvartil je 2, tedy aspoň čtvrtina domácností má aspoň dva členy.
Medián je 4, tedy aspoň polovina domácností má aspoň 4 členy.
Horní kvartil je 5, tedy aspoň tři čtvrtiny domácností mají aspoň 5 členů.

Vypočteme kvartilovou odchylku: $q = x_{0,75} - x_{0,25} = 5 - 2 = 3$.
Dolní vnitřní hradba: $x_{0,25} - 1,5q = 2 - 1,5.3 = -2,5$
Horní vnitřní hradba: $x_{0,75} + 1,5q = 5 + 1,5.3 = 9,5$
Nakonec sestrojíme krabicový diagram:



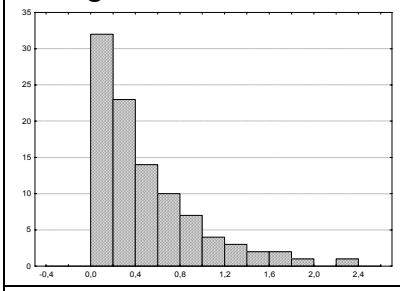
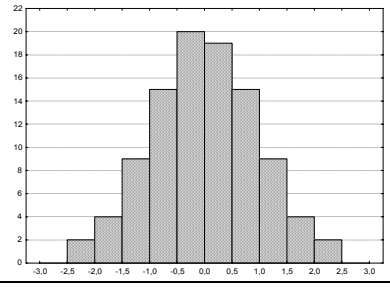
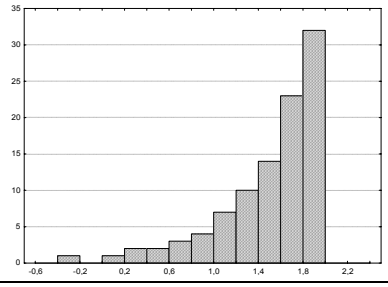
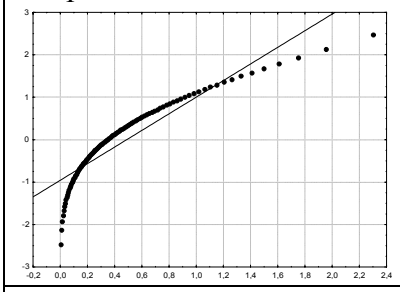
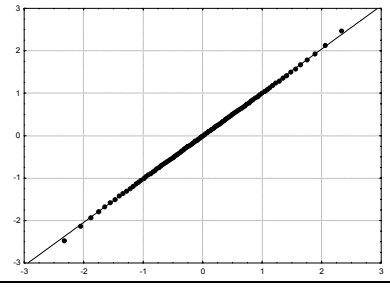
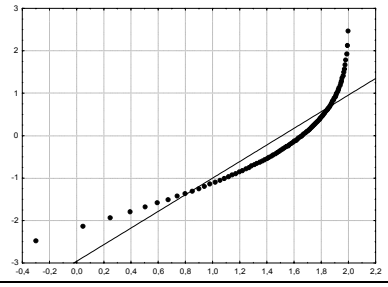
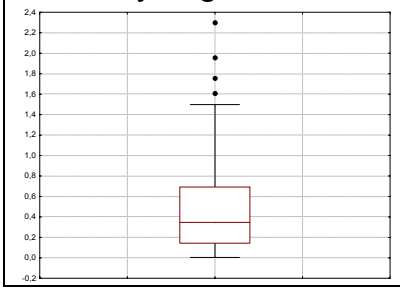
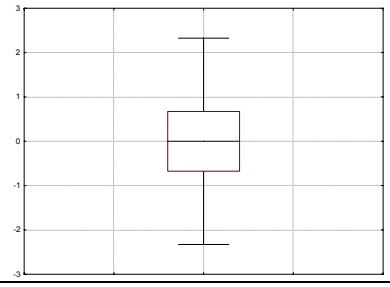
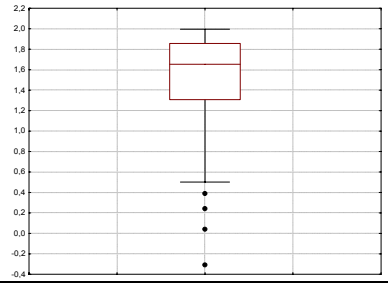
Vidíme, že datový soubor vykazuje určitou nesymetrii – medián je posunut směrem k hornímu kvartilu, soubor je tedy záporně zešikmen. V souboru se nevyskytují žádné odlehlé ani extrémní hodnoty.

3.3. Normální pravděpodobnostní graf (NP-plot)

3.3.1. Popis grafu

NP-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
		
NP plot	NP plot	NP plot
		
Krabicový diagram	Krabicový diagram	Krabicový diagram
		

3.3.2. 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:

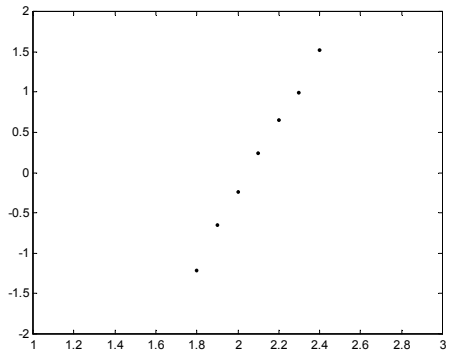
usp.hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

Vektor hodnot průměrného pořadí: j = (1,5 3 4,5 6,5 8 9 10),

vektor hodnot $\alpha_j = \frac{3j-1}{3n+1} = (0{,}1129;0{,}2581;0{,}4032;0{,}5968;0{,}7419;0{,}8387;0{,}9355)$,

vektor kvantilů $u_{\alpha_j} = (-1{,}2112;-0{,}6493;-0{,}245;0{,}245;0{,}6493;0{,}9892;1{,}5179)$.

Normální pravděpodobnostní graf



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

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

3.4.1. 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 $(K_{\alpha_j}(X), x_{(j)})$ se metodou nejmenších čtverců proloží přímka.

Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

3.4.2. Příklad

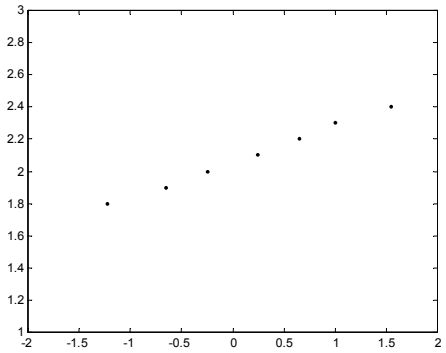
Pro data z příkladu 3.4.1. posuďte pomocí kvantil – kvantilového grafu, zda pocházejí z normálního rozložení.

Řešení:
Na základě tabulky vytvořené při řešení příkladu 3.4.1. stanovíme:
vektor hodnot průměrného pořadí: j = (1,5 3 4,5 6,5 8 9 10),

vektor hodnot $\alpha_j = \frac{j - 0{,}375}{n + 0{,}25} = (0{,}1098; 0{,}2561; 0{,}4024; 0{,}5976; 0{,}7439; 0{,}8415; 0{,}939)$,

vektor kvantilů $u_{\alpha_j} = (-1{,}2278; -0{,}6554; -0{,}247; 0{,}247; 0{,}6554; 1{,}0005; 1{,}566)$

Kvantil – kvantilový graf



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

3.5. Histogram

3.5.1. 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 zakreslí tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení. Kromě 8 typů rozložení uvedených u Q-Q plotu umožňuje STATISTICA použít ještě další 4 rozložení: Laplaceovo, logistické, geometrické, Poissonovo.

3.5.2. Příklad

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

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

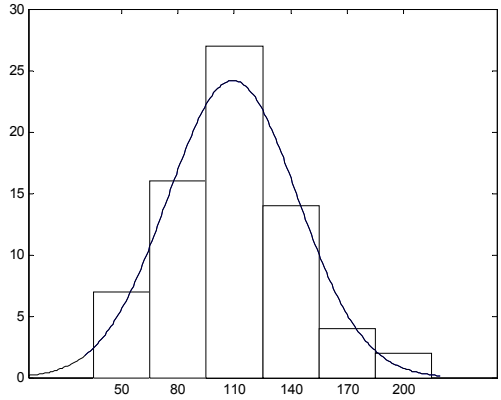
Nakreslete histogram s proloženou hustotou pravděpodobnosti normálního rozložení s parametry m a s^2, kde m je aritmetický průměr a s^2 rozptyl vypočtený z dat.

Řešení:
Pro výpočet průměru resp. rozptylu použijeme vzorec pro vážený aritmetický průměr resp. vážený rozptyl.

$$m = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]} = \frac{1}{70}(7 \cdot 50 + 16 \cdot 80 + 27 \cdot 110 + 14 \cdot 140 + 4 \cdot 170 + 2 \cdot 200) = 109{,}14$$

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{r} n_j (x_{[j]} - m)^2 = \frac{1}{69}\left[7 \cdot (50 - 109{,}14)^2 + \ldots + 2 \cdot (200 - 109{,}14)^2\right] = 1138{,}24$$

Hodnoty hustoty pravděpodobnosti normálního rozložení s parametry m a s^2 musíme vynásobit číslem 30.70 = 2100, kde 30 je délka třídicích intervalů a 70 rozsah datového souboru.
Histogram s proloženou hustotou pravděpodobnosti normálního rozložení



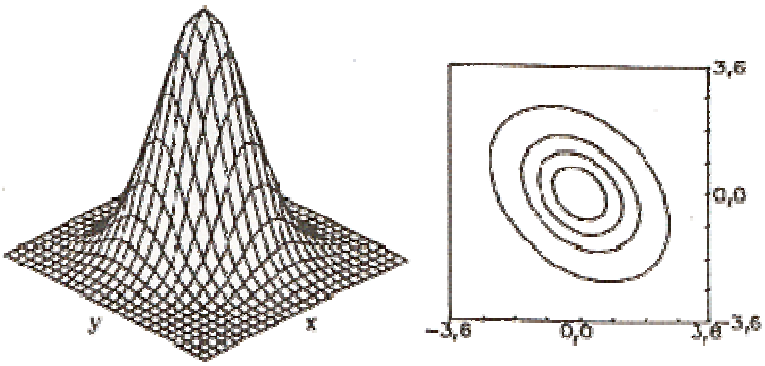
Vidíme, že tvar histogramu se poněkud odchyluje od tvaru hustoty pravděpodobnosti normálního rozložení. Malé hodnoty jsou četnější než velké – datový soubor je kladně zešikmen.

3.6. Dvourozměrný tečkový diagram

3.6.1. 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-α)% elipsu konstantní hustoty pravděpodobnosti. Bude-li více než 100α% 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.

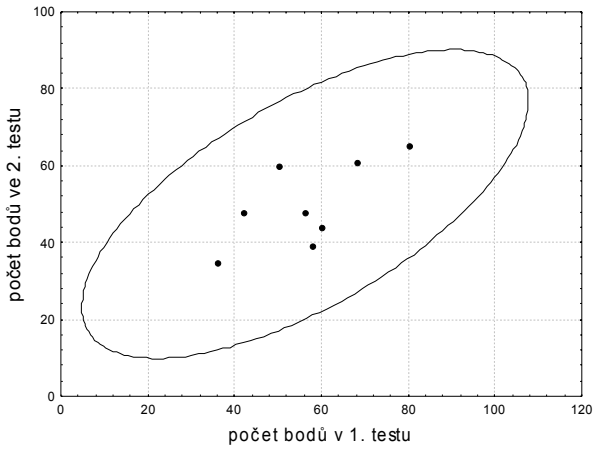
3.6.2. Příklad

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 posuďte, zda tato data lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení.

Řešení:



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.

3.7. Kolmogorovův – Smirnovův test normality dat

3.7.1. Popis testu

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, změní se rozložení testové statistiky D_n. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

3.7.2. Poznámka ke K-S testu ve STATISTICE

Test normality poskytuje hodnotu testové statistiky (ozn. d) a dvě p-hodnoty. První se vztahuje k případu, kdy μ a σ^2 známe předem, druhá (ozn. Liliefors 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.

3.7.3. Příklad

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

Řešení: Odhadem střední hodnoty je výběrový průměr m = 11, odhadem rozptylu je výběrový rozptyl $s^2 = 10$. Uspořádaný náhodný výběr je (8, 9, 10, 12, 16). Vypočteme hodnoty výběrové distribuční funkce:

$$x < 8 : F_5(x) = 0$$

$$8 \le x < 9 : F_5(x) = \frac{1}{5} = 0{,}2$$

$$9 \le x < 10 : F_5(x) = \frac{2}{5} = 0{,}4$$

$$10 \le x < 12 : F_5(x) = \frac{3}{5} = 0{,}6$$

$$12 \le x < 16 : F_5(x) = \frac{4}{5} = 0{,}8$$

$$x \ge 16 : F_5(x) = 1$$

Hodnoty teoretické distribuční funkce $\Phi_T(x)$ v bodech 8, 9, 10, 12, 16:

$$\Phi_T(8) = \Phi\left(\frac{8-11}{\sqrt{10}}\right) = \Phi(-0{,}95) = 1 - \Phi(0{,}95) = 1 - 0{,}82894 = 0{,}17106$$

$$\Phi_T(9) = \Phi\left(\frac{9-11}{\sqrt{10}}\right) = \Phi(-0{,}63) = 1 - \Phi(0{,}63) = 1 - 0{,}73565 = 0{,}26435$$

$$\Phi_T(10) = \Phi\left(\frac{10-11}{\sqrt{10}}\right) = \Phi(-0{,}32) = 1 - \Phi(0{,}32) = 1 - 0{,}62552 = 0{,}37448$$

$$\Phi_T(12) = \Phi\left(\frac{12-11}{\sqrt{10}}\right) = \Phi(0{,}32) = 0{,}62552$$

$$\Phi_T(16) = \Phi\left(\frac{16-11}{\sqrt{10}}\right) = \Phi(1{,}58) = 0{,}94295$$

(Φ je distribuční funkce rozložení N(0,1).)

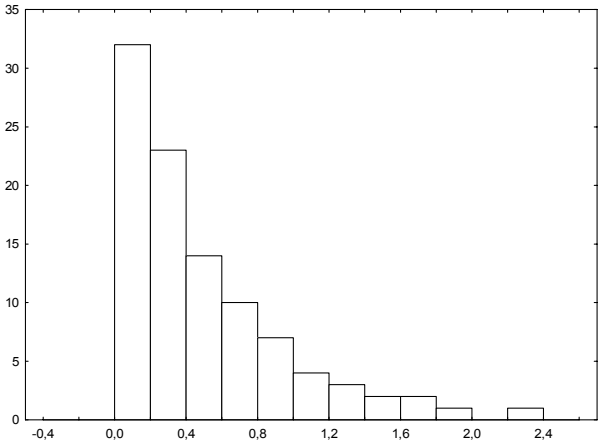
Rozdíly mezi výběrovou distribuční funkcí $F_5(x)$ a teoretickou distribuční funkcí $\Phi_T(x)$:
$d_1 = 0{,}2 - 0{,}17106 = 0{,}02894$; $d_2 = 0{,}4 - 0{,}26435 = 0{,}13565$; $d_3 = 0{,}6 - 0{,}37448 = 0{,}22552$;
$d_4 = 0{,}8 - 0{,}62552 = 0{,}17448$; $d_5 = 1 - 0{,}94295 = 0{,}05705$.
Testová statistika: $D_5 = 0{,}22552$, modifikovaná kritická hodnota pro n = 5, $\alpha = 0{,}05$ je 0,343.
Protože $0{,}22552 < 0{,}343$, hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

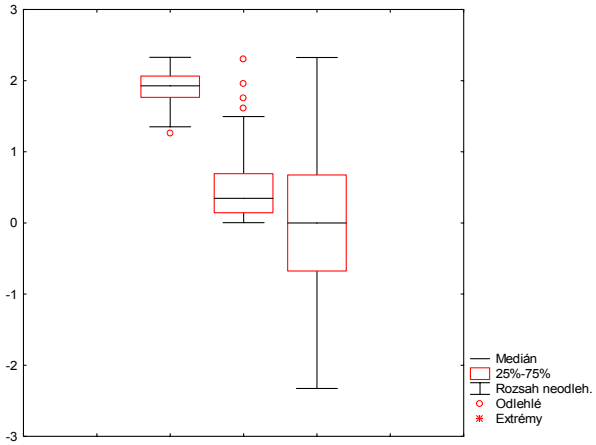
3.8. Shapirův – Wilkův test normality dat

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 kvantil-kvantilovém grafu jsou významně odlišné od regresní přímky proložené těmito body. S-W test se používá především pro výběry menších rozsahů, $n < 50$.

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

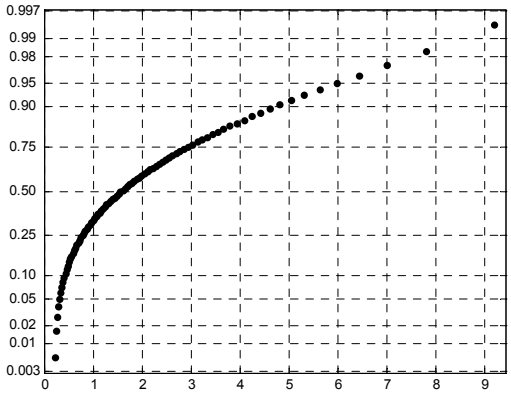




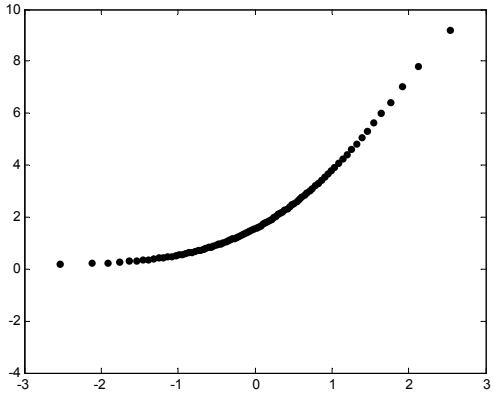
Výsledek: Jedná se o druhý krabicový diagram.

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.

Příklady

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. (S) Pro počet bodů z 1. příkladu sestrojte normální pravděpodobnostní graf.

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

4. (S) 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. (S) 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. (S) 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	20,0	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.