

3. Základní pojmy matematické statistiky. Diagnosticke grafy.

3.1. Motivace: Matematická statistika je věda, která analyzuje a interpretuje data především za účelem získání předpovědi a zlepšení rozhodování v různých oborech lidské činnosti. Přitom se řídí principem statistické indukce, tj. na základě znalostí o náhodném výběru z určitého rozložení pravděpodobností se snaží učinit závěry o vlastnostech tohoto rozložení.

Ústředním pojmem matematické statistiky je tedy pojem náhodného výběru.

3.2. Definice:

Definice náhodného výběru:

- Nechť X_1, \dots, X_n jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení $L(\mathcal{Q})$. Řekneme, že X_1, \dots, X_n je **náhodný výběr rozsahu n z rozložení $L(\mathcal{Q})$** . (Číselné realizace x_1, \dots, x_n náhodného výběru X_1, \dots, X_n uspořádané do sloupcového vektoru odpovídají datovému souboru zavedenému v popisné statistice.)
- Nechť $(X_1, Y_1), \dots, (X_n, Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení $L_2(\mathcal{Q})$. Řekneme, že $(X_1, Y_1), \dots, (X_n, Y_n)$ je **dvourozměrný náhodný výběr rozsahu n z dvourozměrného rozložení $L_2(\mathcal{Q})$** . (Číselné realizace $(x_1, y_1), \dots, (x_n, y_n)$ náhodného výběru $(X_1, Y_1), \dots, (X_n, Y_n)$ uspořádané do matice typu $n \times 2$ odpovídají dvourozměrnému datovému souboru zavedenému v popisné statistice.)
- Analogicky lze definovat p -rozměrný **náhodný výběr rozsahu n z p -rozměrného rozložení $L_p(\mathcal{Q})$** .

Definice statistiky:

Libovolná funkce $T = T(X_1, \dots, X_n)$ náhodného výběru X_1, \dots, X_n (resp. $T = T(X_1, Y_1, \dots, X_n, Y_n)$ náhodného výběru $(X_1, Y_1), \dots, (X_n, Y_n)$) se nazývá (výběrová) **statistika**.

3.3. Důsledek:

Nechť X_1, \dots, X_n je náhodný výběr z rozložení s distribuční funkcí $\Phi(x)$. Pak simultánní distribuční funkce náhodného vektoru (X_1, \dots, X_n) je $\Phi(x_1) \dots \Phi(x_n)$.

3.4. Definice:

Definice důležitých statistik:

- Nechť X_1, \dots, X_n je náhodný výběr, $n \geq 2$.

$$M = \frac{1}{n} \sum_{i=1}^n X_i \quad \dots \quad \text{výběrový průměr,}$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M)^2 \quad \dots \quad \text{výběrový rozptyl,}$$

$S = \sqrt{S^2}$... výběrová směrodatná odchylka

Pro libovolné, ale pevně dané reálné číslo x je statistikou též hodnota výběrové

distribuční funkce $F_n(x) = \frac{1}{n} \text{card} \{X_i \leq x\}$

b) Necht' je dáno $r \geq 2$ stochasticky nezávislých náhodných výběrů o rozsazích

$n_1 \geq 2, \dots, n_r \geq 2$. Celkový rozsah je $n = \sum_{j=1}^r n_j$. Označme M_1, \dots, M_r výběrové

průměry a S_1^2, \dots, S_r^2 výběrové rozptyly jednotlivých výběrů. Necht' c_1, \dots, c_r jsou reálné konstanty, aspoň jedna nenulová.

$\sum_{j=1}^r c_j M_j$... lineární kombinace výběrových průměrů,

$S^2 = \frac{\sum_{j=1}^r c_j^2 S_j^2}{\sum_{j=1}^r c_j^2}$... vážený průměr výběrových rozptylů.

c) Necht' $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozložení .

Označme $M_1 = \frac{1}{n} \sum_{i=1}^n X_i$, $M_2 = \frac{1}{n} \sum_{i=1}^n Y_i$ výběrové průměry,

$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_1)^2$, $S_2^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - M_2)^2$ výběrové rozptyly.

$S_{12} = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_1)(Y_i - M_2)$... výběrová kovariance,

$R_{12} = \begin{cases} \frac{1}{n-1} \sum_{i=1}^n \frac{X_i - M_1}{S_1} \cdot \frac{Y_i - M_2}{S_2} = \frac{S_{12}}{S_1 S_2} \text{ pro } S_1 S_2 \neq 0 \\ 0 \text{ jinak} \end{cases}$... výběrový koeficient

korelace.

Pro libovolnou, ale pevně zvolenou dvojici reálných čísel x, y je statistikou též hodnota výběrové simultánní distribuční funkce

$F_n(x, y) = \frac{1}{n} \text{card} \{X_i \leq x \wedge Y_i \leq y\}$

Upozornění: Číselné realizace statistik $M, S^2, S, S_{12}, R_{12}$ odpovídají číselným charakteristikám $m, s^2, s, s_{12}, r_{12}$ zavedeným v popisné statistice, ale u rozptylu, směrodatné odchylky, kovariance a koeficientu korelace je multiplikativní konstanta $\frac{1}{n}$, nikoliv $\frac{1}{n-1}$, jak tomu bylo v popisné statistice. Jak uvidíme později, uvedené číselné realizace mohou být považovány za odhady číselných realizací náhodných veličin zavedených v počtu pravděpodobnosti.

Charakteristika vlastnosti	Počet pravděpodobnosti	Matematická statistika	Popisná statistika
poloha	$E(X) = \mu$	M	m
variabilita	$D(X) = \sigma^2$	S^2	$\frac{n-1}{n} s^2$
variabilita	$\sqrt{D(X)} = \sigma$	S	$\sqrt{\frac{n-1}{n} s^2}$
společná variabilita	$C(X_1, X_2) = \sigma_{12}$	S_{12}	$\frac{n-1}{n} s_{12}$
těsnost vztahu	$R(X_1, X_2) = \rho$	R_{12}	r_{12}
rozložení	$\Phi(x)$	$F_n(x)$	F(x)

3.5. Příklad:

(Výpočet realizací výběrového průměru, výběrového rozptylu a hodnot výběrové distribuční funkce):

Desetkrát nezávisle na sobě byla změřena jistá konstanta μ . Výsledky měření byly: 2,1, 1,8, 2,1, 2,4, 1,9, 2,1, 2,2, 1,8, 2,3, 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru X_1, \dots, X_{10} . Vypočítejte realizaci m výběrového průměru M , realizaci s^2 výběrového rozptylu S^2 , realizaci s výběrové směrodatné odchylky S a hodnoty výběrové distribuční funkce $F_{10}(x)$.

Řešení:

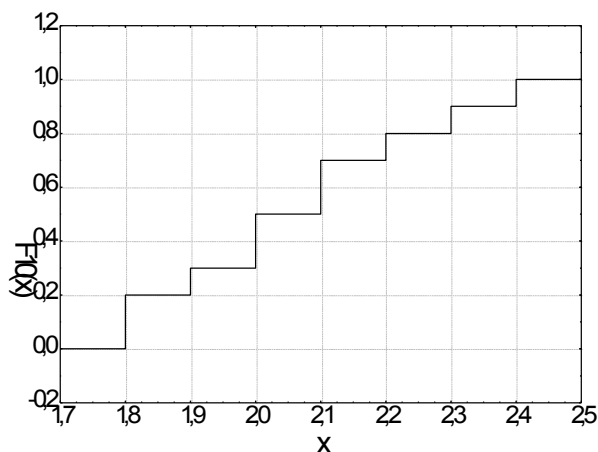
$$m = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10} (2,1 + 1,8 + 2,1 + 2,4 + 1,9 + 2,1 + 2,2 + 1,8 + 2,3 + 2,2) = 2,06$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{9} \left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) = \frac{1}{9} (2,1^2 + 1,8^2 + 2,1^2 + 2,4^2 + 1,9^2 + 2,1^2 + 2,2^2 + 1,8^2 + 2,3^2 + 2,2^2 - 10 \cdot 2,06^2) = 0,0404$$

$$s = \sqrt{s^2} = \sqrt{0,0404} = 0,2011$$

Pro usnadnění výpočtu hodnot výběrové distribuční funkce $F_{10}(x)$ uspořádáme měření podle velikosti: 1,8, 1,8, 1,9, 2, 2, 2,1, 2,1, 2,2, 2,3, 2,4.

$$\begin{aligned}
 x < 1,7 &: F_{10}(x) = 0 \\
 1,7 \leq x < 18 &: F_{10}(x) = 0,2 \\
 18 \leq x < 19 &: F_{10}(x) = 0,3 \\
 19 \leq x < 20 &: F_{10}(x) = 0,5 \\
 20 \leq x < 21 &: F_{10}(x) = 0,7 \\
 21 \leq x < 22 &: F_{10}(x) = 0,8 \\
 22 \leq x < 23 &: F_{10}(x) = 0,9 \\
 23 \leq x < 24 &: F_{10}(x) = 0,95 \\
 x \geq 24 &: F_{10}(x) = 1
 \end{aligned}$$



3.6. Příklad:

(Výpočet realizace výběrového koeficientu korelace):

U 11 náhodně vybraných aut jisté značky bylo zjišťováno jejich stáří (náhodná veličina X – v letech) a cena (náhodná veličina Y – v tisících Kč). Výsledky: (5, 85), (4, 103), (6, 70), (5, 82), (5, 89), (5, 98), (6, 66), (6, 95), (2, 169), (7, 70), (7, 48). Vypočítejte a interpretujte číselnou realizaci r_{12} výběrového koeficientu korelace R_{12} .

Řešení:

$$m_1 = \frac{1}{n} \sum_{i=1}^n x_i = \frac{5+4+6+5+5+5+6+6+2+7+7}{11} = 5,28$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n y_i = \frac{85+103+70+82+89+98+66+95+169+70+48}{11} = 93$$

$$s_1^2 = \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - n m_1^2 \right) = \frac{1}{11} (5^2 + 4^2 + 6^2 + 5^2 + 5^2 + 5^2 + 6^2 + 6^2 + 2^2 + 7^2 + 7^2 - 11 \cdot 5,28^2) = 12$$

$$s_2^2 = \frac{1}{n} \left(\sum_{i=1}^n y_i^2 - n m_2^2 \right) = \frac{1}{11} (85^2 + 103^2 + 70^2 + 82^2 + 89^2 + 98^2 + 66^2 + 95^2 + 169^2 + 70^2 + 48^2 - 11 \cdot 93^2) = 705$$

$$s_{12} = \frac{1}{n} \left(\sum_{i=1}^n x_i y_i - m_1 m_2 \right) = \frac{1}{11} (5 \cdot 85 + 4 \cdot 103 + 6 \cdot 70 + 5 \cdot 82 + 5 \cdot 89 + 5 \cdot 98 + 6 \cdot 66 + 6 \cdot 95 + 2 \cdot 169 + 7 \cdot 70 + 7 \cdot 48 - 5,28 \cdot 93) = 89$$

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{89}{\sqrt{12} \cdot \sqrt{705}} = -0,2$$

Mezi náhodnými veličinami X a Y existuje silná nepřímá lineární závislost. Čím starší auto, tím nižší cena.

3.7. Věta:

Vlastnosti důležitých statistik

a) **Případ jednoho náhodného výběru:** Necht' X_1, \dots, X_n je náhodný výběr z rozložení se střední hodnotou μ , rozptylem σ^2 a distribuční funkcí $\Phi(x)$. Necht' $n \geq 2$. Označme M_n výběrový průměr, S_n^2 výběrový rozptyl a pro libovolné, ale pevně dané x označme $F_n(x)$ hodnotu výběrové distribuční funkce. Pak pro libovolné hodnoty parametrů μ , σ^2 a libovolné, ale pevně dané reálné číslo x platí:

$$E(M_n) = \mu,$$

$$D(M_n) = \frac{\sigma^2}{n},$$

$$E(S_n^2) = \sigma^2,$$

$$D(S_n^2) = \frac{\gamma_4 - 3\sigma^4}{n}, \text{ kde } \gamma_4 \text{ je 4. centrální moment,}$$

$$E(F_n(x)) = \Phi(x),$$

$$D(F_n(x)) = \frac{\sigma^2}{n} \phi^2(x),$$

b) **Případ $r \geq 2$ stochasticky nezávislých náhodných výběrů:** Necht' $X_{11}, \dots, X_{1n_1}, \dots, X_{r1}, \dots, X_{rn_r}$ je r stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2, \dots, n_r \geq 2$ z rozložení se středními hodnotami μ_1, \dots, μ_r a rozptylem σ^2 .

Celkový rozsah je $n = \sum_{j=1}^r n_j$. Necht' c_1, \dots, c_r jsou reálné konstanty, aspoň jedna nenulová. Pak pro libovolné hodnoty parametrů μ_1, \dots, μ_r a σ^2 platí:

$$E\left(\sum_{j=1}^r c_j M_j\right) = \sum_{j=1}^r c_j \mu_j,$$

$$E(S_*^2) = \sigma^2.$$

c) **Případ jednoho náhodného výběru z dvourozměrného rozložení:** Necht' $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ . Pak pro libovolné hodnoty parametrů σ_{12} a ρ platí:

$$E(S_{12}) = \sigma_{12},$$

$$E(R_{12}) \approx \rho \text{ (shoda je vyhovující pro } n \geq 30).$$

3.8. Poznámka:

Metody matematické statistiky často slouží k vyhodnocování výsledků pokusů. Aby mohl být pokus správně vyhodnocen, musí být dobře naplánován. Uvedeme zde nejjednodušší typy uspořádání pokusů

Předpokládejme například, že sledujeme hmotnostní přírůstky selat téhož plemene při různých výkrmných dietách.

a) **Jednoduché pozorování:** Náhodná veličina X je pozorována za týchž podmínek. Situace je charakterizována jedním náhodným výběrem X_1, \dots, X_n .
Náhodně vylosujeme n selat téhož plemene, podrobíme je jediné výkrmné dietě a zjistíme u každého selete hmotnostní přírůstek. Tím dostaneme realizaci jednoho náhodného výběru.

b) **Dvojné pozorování:** Náhodná veličina X je pozorována za dvojích různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

Dvouvýběrové porovnávání: situace je charakterizována dvěma nezávislými náhodnými výběry X_{11}, \dots, X_{1n_1} a X_{21}, \dots, X_{2n_2} .

Náhodně vylosujeme n_1 a n_2 selat téhož plemene, náhodně je rozdělíme na dva soubory o n_1 a n_2 jedincích, první podrobíme výkrmné dietě č. 1 a druhý výkrmné dietě číslo 2. Tak dostaneme realizace dvou nezávislých náhodných výběrů.

Párové porovnávání: situace je charakterizována jedním náhodným výběrem $(X_{11}, X_{12}), \dots, (X_{n1}, X_{n2})$ z dvourozměrného rozložení. Přejdeme k rozdílovému náhodnému výběru $Z_i = X_{i1} - X_{i2}$, $i = 1, \dots, n$ a tím dostaneme jednoduché pozorování.

Náhodně vylosujeme n vrhů stejně starých selat téhož plemene, z každého odebereme dva sourozence a náhodně jim přiřadíme první a druhou výkrmnou dietu. Tak dostaneme realizaci jednoho dvourozměrného náhodného výběru, kde první složka odpovídá první dietě a druhá složka druhé dietě.

(Párové porovnávání je efektivnější, protože skutečný rozdíl v účinnosti obou diet je překrýván pouze náhodnými vlivy při samotném krmení a trvání, kdežto vliv různých dědičných vloh, který byl losováním znárodněn, je u sourozeneckého páru selat částečně vyloučen.)

c) **Mnohonásobné pozorování:** Náhodná veličina X je pozorována za $r \geq 3$ různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

Mnohovýběrové porovnávání: situace je charakterizována r nezávislými náhodnými výběry X_{11}, \dots, X_{1n_1} až X_{r1}, \dots, X_{rn_r} .

Náhodně vylosujeme n_1, n_2, \dots, n_r selat téhož plemene, náhodně je rozdělíme na r souborů o n_1, n_2, \dots, n_r jedincích, první podrobíme výkrmné dietě č. 1, druhý výkrmné dietě číslo 2 atd. až r -tý podrobíme výkrmné dietě číslo r . Tak dostaneme realizace r nezávislých náhodných výběrů.

Blokové porovnávání: situace je charakterizována jedním náhodným výběrem $(X_{11}, \dots, X_{1r}), \dots, (X_{n1}, \dots, X_{nr})$ z r -rozměrného rozložení.

Náhodně vylosujeme n vrhů stejně starých selat téhož plemene, z každého odebereme r sourozenců a náhodně jim přiřadíme první až r -tou výkrmnou dietu. Tak dostaneme realizaci jednoho r -rozměrného náhodného výběru, kde první

složka odpovídá první dietě , druhá složka druhé dietě atd. až r-tá složka odpovídá r-té dietě.

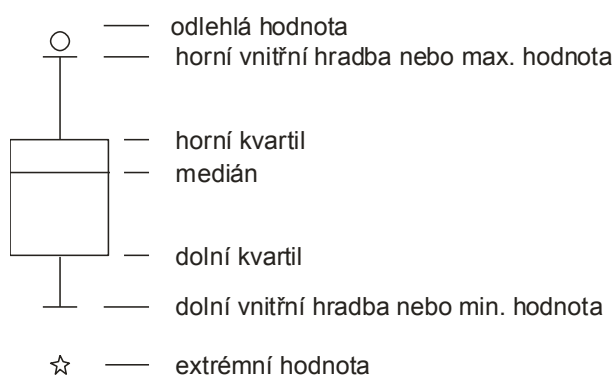
3.9. Motivace:

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 či histogramu) a nepřítomnost vybočujících hodnot (odhalí je krabicový diagram).

3.10. Krabicový diagram

Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot.

Způsob konstrukce



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

Příklad na krabicový diagram

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 $1 - \alpha$ všech dat. Pro výpočet α -kvantilu slouží algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo} \Rightarrow x_{(c)} & \text{pro } \alpha = \frac{X_{(c)} + X_{(c+1)}}{2} \\ \text{ne celé číslo} \Rightarrow \text{zaokrouhláme na nejblíže číslo} \Rightarrow x \end{cases}$$

Pro speciálně zvolená α užíváme názvů: $x_{0,50}$ – medián, $x_{0,25}$ – dolní kvartil, $x_{0,75}$ – horní kvartil, $x_{0,1}, \dots, x_{0,9}$ – decily, $x_{0,01}, \dots, x_{0,99}$ – 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	$n\alpha$	c		x_α
0,25	7,5	8	$x_{(c)} = x_{(8)}$	2
0,50	15	15	$\frac{x_{(15)} + x_{(16)}}{2}$	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.

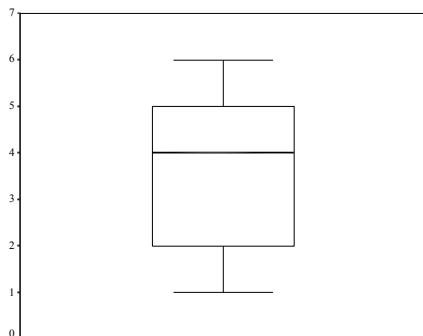
Horní kvartil je 5, tedy aspoň tři čtvrtiny domácností mají aspoň 5 členů.

Vypočteme kvartilovou odchylku: $q = x_{0,75} - x_{0,25} = 5 - 2 = 3$.

Dolní vnitřní hradba: $x_{0,25} - 1,5q = 2 - 1,5 \cdot 3 = -2,5$

Horní vnitřní hradba: $x_{0,75} + 1,5q = 5 + 1,5 \cdot 3 = 9,5$

Nakonec sestrojíme krabicový diagram:



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

3.11. Pravděpodobnostně – pravděpodobnostní graf (P – P plot)

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

Vypočtou se standardizované hodnoty $Z_j = \frac{x_{(j)} - \bar{x}}{s}$, $j = 1, \dots, n$. Na vodorovnou osu se vynesou hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = j/n$. (Jsou-li některé hodnoty $x_{(1)} \leq \dots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.) Pokud se body $(\Phi(z_{(j)}), F(z_{(j)}))$ řadí kolem hlavní diagonály čtverce $[0,1] \times [0,1]$, lze usuzovat na dobrou shodu empirického a teoretického rozložení.

Pro posouzení normality dat se používá normální pravděpodobnostní graf (N – P plot): na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \dots \leq x_{(n)}$ a na svislou osu kvantily U_α , kde $\alpha = \frac{j}{n+1}$ (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince).

Pocházejí-li data z normálního rozložení, pak dvojice $(x_{(j)}, U_\alpha)$ budou ležet na přímce.

Pocházejí-li data z rozložení s kladnou šikmostí, pak dvojice $(x_{(j)}, U_\alpha)$ se budou řadit do konkávní křivky.

Pocházejí-li data z rozložení se zápornou šikmostí, pak dvojice $(x_{(j)}, U_\alpha)$ se budou řadit do konvexní křivky.

Příklad na N – P plot:

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:

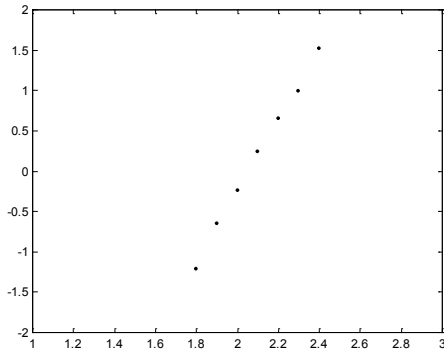
usp. hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

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

vektor hodnot $\alpha = \frac{j}{n+1} = (0,125 \ 0,25 \ 0,403 \ 0,596 \ 0,741 \ 0,838 \ 0,935)$,

vektor kvantilů $U_\alpha = (0,11 \ 0,24 \ 0,49 \ 0,49 \ 0,64 \ 0,64 \ 0,89 \ 0,95 \ 1)$.

Normální pravděpodobnostní graf



Protože dvojice (X_i, U_{α}) téměř leží na přímce, lze usoudit, že data pocházejí z normálního rozložení.

3.12. Histogram

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 třídící intervaly (implicitně 10, jejich počet lze změnit, stejně tak i meze třídících intervalů) či varianty znaku a na svislou osu absolutní nebo relativní četnosti třídících intervalů či variant. Do histogramu se zakreslí tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení.

Příklad na histogram

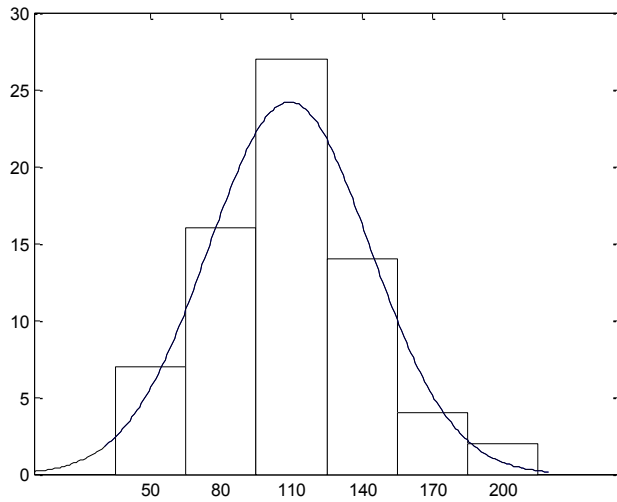
U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	356\$	699\$	952\$	1235\$	1235\$	1821\$
Počet dom.	7	16	27	14	4	2

Nakreslete histogram.

Řešení:

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 sešikmen.

Vlastnosti rozložení četností datového souboru se projeví ve vzhledu histogramu, N–P plotu a krabicového diagramu, jak vidíme na následujícím obrázku:

