Základní pojmy matematické statistiky

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.

Definice náhodného výběru:

a) Nechť $X_1, ..., X_n$ jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení $L(\vartheta)$. Řekneme, že $X_1, ..., X_n$ je náhodný výběr rozsahu n z rozložení $L(\vartheta)$. (Číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ uspořádané do sloupcového vektoru odpovídají datovému souboru zavedenému v popisné statistice.)

b) Nechť $(X_1,Y_1), ..., (X_n,Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení $L_2(\vartheta)$. Řekneme, že $(X_1,Y_1), ..., (X_n,Y_n)$ je dvourozměrný náhodný výběr rozsahu n z dvourozměrného rozložení $L_2(\vartheta)$. (Číselné realizace $(x_1,y_1), ..., (x_n,y_n)$ náhodného výběru $(X_1,Y_1), ..., (X_n,Y_n)$ uspořádané do matice typu $n \times 2$ odpovídají dvourozměrnému datovému souboru zavedenému v popisné statistice.)

c) Analogicky lze definovat p-rozměrný náhodný výběr rozsahu n z p-rozměrného rozložení $L_p(\vartheta)$.

Důsledek

Je-li $X_1, ..., X_n$ náhodný výběr z rozložení s distribuční funkcí $\Phi(x)$, pak simultánní distribuční funkce náhodného vektoru $(X_1, ..., X_n)$ je $\Phi(x_1)\ldots \Phi(x_n)$.

Definice statistiky:

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

Definice důležitých statistik:

a) Nechť $X_1, ..., X_n$ je náhodný výběr, $n \geq 2$.
Označme

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

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

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

Pro libovolné, ale pevně dané reálné číslo x je statistikou též hodnota výběrové distribuční funkce $F_n(x) = \frac{1}{n}\mathrm{card}\{i; X_i \leq x\}$

b) Nechť je dáno $r \geq 2$ stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2, \ldots, n_r \geq 2$.

Celkový rozsah je $n = \sum_{j=1}^{r} n_j$.

Označme $M_1, \ldots, M_r$ výběrové průměry a $S_1^2, \ldots, S_r^2$ výběrové rozptyly jednotlivých výběrů.
Nechť $c_1, \ldots, c_r$ jsou reálné konstanty, aspoň jedna nenulová.

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

$S_*^2 = \dfrac{\sum_{j=1}^{r} (n_j - 1) S_j^2}{n - r}$ … vážený průměr výběrových rozptylů.

c) Nechť (X_1,Y_1), ..., (X_n,Y_n) je náhodný výběr z dvourozměrného rozložení o rozsahu n.

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 $F_n(x,y) = \frac{1}{n}\mathrm{card}\{i; X_i \le x \wedge Y_i \le y\}$.

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 $\frac{1}{n-1}$, nikoliv $\frac{1}{n}$, 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	$\sqrt{\frac{n-1}{n}}s$
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)

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 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2.

Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, ..., X_{10}$. Vypočtěte 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:

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

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

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

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.

$$x < 1{,}8 : F_{10}(x) = 0$$

$$1{,}8 \le x < 1{,}9 : F_{10}(x) = \frac{2}{10} = 0{,}2$$

$$1{,}9 \le x < 2 : F_{10}(x) = \frac{3}{10} = 0{,}3$$

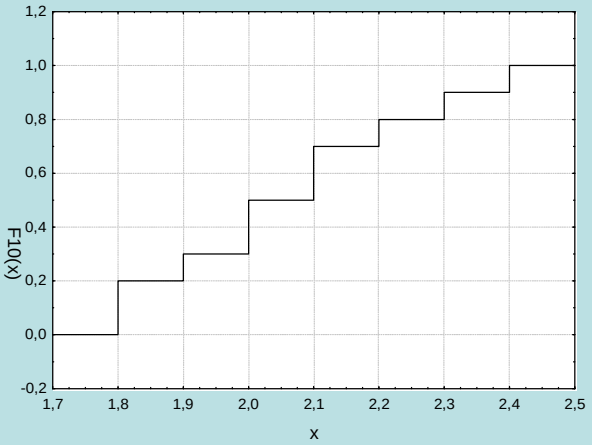
$$2 \le x < 2{,}1 : F_{10}(x) = \frac{5}{10} = 0{,}5$$

$$2{,}1 \le x < 2{,}2 : F_{10}(x) = \frac{7}{10} = 0{,}7$$

$$2{,}2 \le x < 2{,}3 : F_{10}(x) = \frac{8}{10} = 0{,}8$$

$$2{,}3 \le x < 2{,}4 : F_{10}(x) = \frac{9}{10} = 0{,}9$$

$$x \ge 2{,}4 : F_{10}(x) = 1$$



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).
Vypočtěte a interpretujte číselnou realizaci r_{12} výběrového koeficientu korelace R_{12}.

Řešení:

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

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{11}(85+103+\ldots+48) = 88{,}63$$

$$s_1^{\,2} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^{\,2} - n m_1^{\,2}\right) = \frac{1}{10}\left(5^2+4^2+\ldots+7^2-11\cdot 5{,}28^2\right) = 2{,}02$$

$$s_2^{\,2} = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^{\,2} - n m_2^{\,2}\right) = \frac{1}{10}\left(85^2+103^2+\ldots+48^2-11\cdot 88{,}63^2\right) = 970{,}85$$

$$s_{12} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n m_1 m_2\right) = \frac{1}{10}\left(5\cdot 85+4\cdot 103+\ldots+7\cdot 48-11\cdot 5{,}28\cdot 88{,}63\right) = -40{,}89$$

$$r_{12} = \frac{s_{12}}{s_1 \cdot s_2} = \frac{-40{,}82}{\sqrt{2{,}02}\cdot\sqrt{970{,}85}} = -0{,}92$$

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

Vlastnosti důležitých statistik

a) Případ jednoho náhodného výběru:

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení se střední hodnotou μ, rozptylem σ^2 a distribuční funkcí $\Phi(x)$. Nechť $n \geq 2$. Označme M_n výběrový průměr, S_n^2 výběrový rozptyl a pro libovolné, ale pevně dané $x \in R$ 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:

$E(M_n) = \mu$,

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

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

$$D(S_n^2) = \frac{\gamma_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$$, kde γ_4 je 4. centrální moment,

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

$$D(F_n(x)) = \frac{\Phi(x)[1-\Phi(x)]}{n}$$

b) Případ $r \geq 2$ stochasticky nezávislých náhodných výběrů:

Nechť $X_{11},\ldots,X_{1n_1}$, ..., $X_{r1},\ldots,X_{rn_r}$ je r stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2$, ..., $n_r \geq 2$ z rozložení se středními hodnotami μ_1, ..., μ_r a rozptylem σ^2. Celkový rozsah je $n = \sum_{j=1}^{r} n_j$. Nechť c_1, ..., c_r jsou reálné konstanty, aspoň jedna nenulová. Pak pro libovolné hodnoty parametrů μ_1, …, 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:

Nechť (X_1,Y_1), ..., (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:

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

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

Základní typy uspořádání pokusů

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 $X_1, ..., X_n$.

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 $X_{11},\ldots,X_{1n_1}$ a $X_{21},\ldots,X_{2n_2}$.

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 $(X_{11},X_{12}),\ldots,(X_{n1},X_{n2})$ z dvourozměrného rozložení. Přejdeme k rozdílovému náhodnému výběru $Z_i = X_{i1} - X_{i2}$, i = 1, …, 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 znáhodněn, 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 $X_{11},\ldots,X_{1n_1}$ až $X_{r1},\ldots,X_{rn_r}$.

Náhodně vylosujeme n_1, n_2, …, n_r selat téhož plemene, náhodně je rozdělíme na r souborů o n_1, n_2, …, 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 $(X_{11},\ldots,X_{1r}),\ldots,(X_{n1},\ldots,X_{nr})$ 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.

Diagnostické grafy

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.

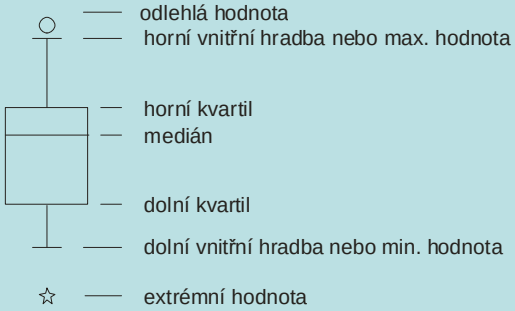
V případě jednoho náhodného výběru je to především normalita (posuzujeme ji pomocí NP plotu, Q-Q plotu, histogramu) a nepřítomnost vybočujících hodnot (odhalí je krabicový diagram).

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 $100(1-\alpha)\%$ elipsou konstantní hustoty pravděpodobnosti).

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 $(-\infty, x_{0,25} - 3q)$.
Pro speciálně zvolená α užíváme názvů: $x_{0,50}$ – medián, $x_{0,25}$ – dolní kvartil, $x_{0,75}$ – horní kvartil, $x_{0,1}, \ldots, x_{0,9}$ – decily, $x_{0,01}, \ldots, x_{0,99}$ – percentily. Jako charakteristika variability slouží kvartilová odchylka: $q = x_{0,75} - x_{0,25}$.

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:

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

Algoritmus:

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

Data:

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

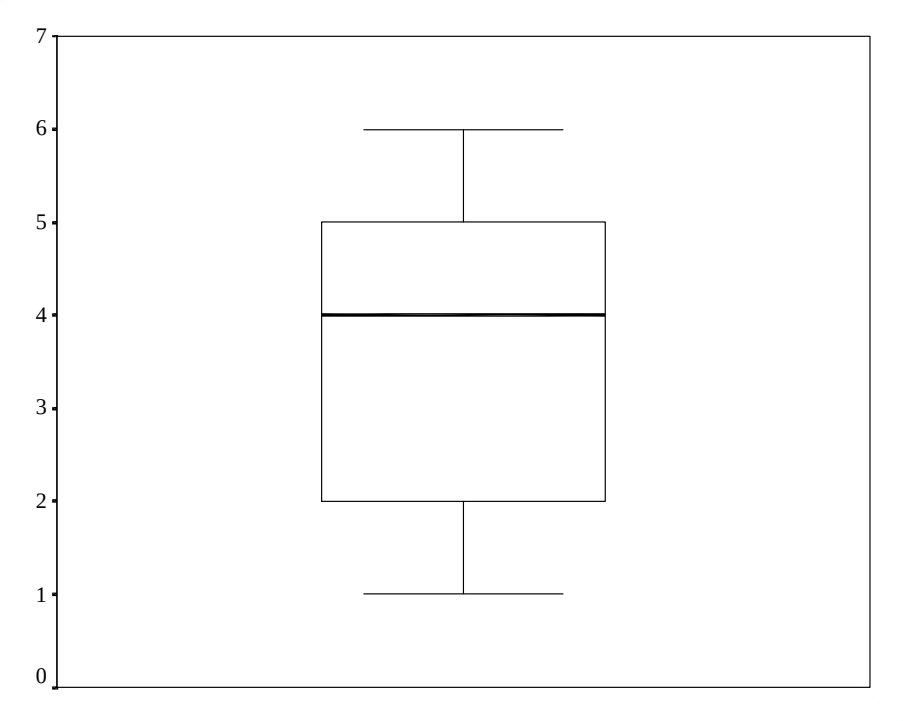
V našem případě rozsah souboru n = 30. Výpočty potřebných kvantilů uspořádáme do tabulky.

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 nejvýše dva členy.
Medián je 4, tedy aspoň polovina domácností má nejvýše 4 členy.
Horní kvartil je 5, tedy aspoň tři čtvrtiny domácností mají nejvýše 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: $x_{0,75} + 1,5q = 5 + 1,5.3 = 9,5$

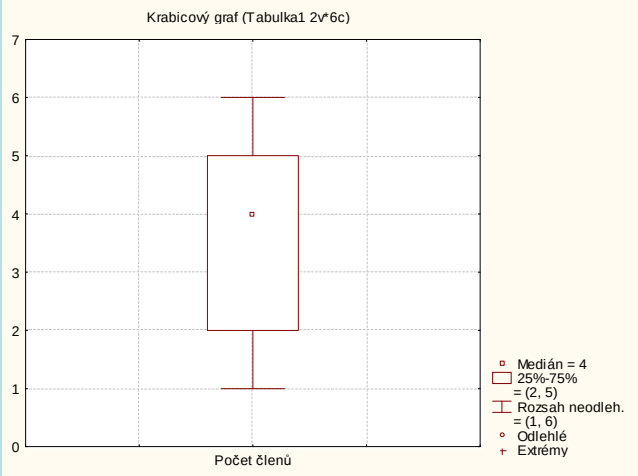
Nakonec sestrojíme krabicový diagram:



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.

Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor se dvěma proměnnými Počet členů a Počet domácností a šesti případy. Vytvoření krabicového diagramu: Grafy – 2D Grafy – Krabicové grafy. Aktivujeme váhy - 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 nejvýš 4 členy), dolní kvartil 2 (aspoň čtvrtina domácností má nejvýš 2 členy), horní kvartil 5 (aspoň tři čtvrtiny domácností mají nejvýš 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.

Normální pravděpodobnostní graf (NP-plot)

NP-plot umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.

Způsob konstrukce: na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq ... \leq x_{(n)}$ 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.

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 NP plotu posuďte, zda se tato data řídí normálním rozložením.

Řešení:

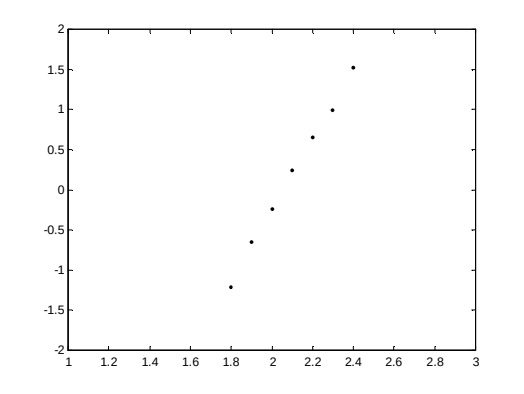
uspořádané hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

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

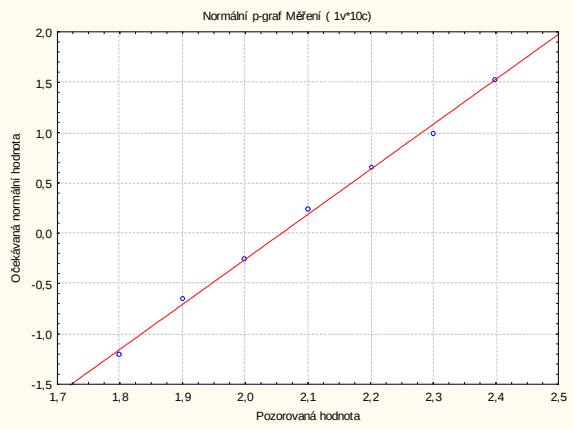


Závěr:

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

Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor s jednou proměnnou X a deseti případy.
Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné X, OK.



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

Kvantil-kvantilový graf (Q-Q plot)

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

Způsob konstrukce: na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \leq ... \leq x_{(n)}$ 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 $\leq 0{,}5$, implicitně $r_{adj} = 0{,}375$ a $n_{adj} = 0{,}25$. (Jsou-li některé hodnoty $x_{(1)} \leq ... \leq x_{(n)}$ 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.

Příklad

Pro údaje o měření konstanty posuďte pomocí kvantil – kvantilového grafu, zda pocházejí z normálního rozložení.

Řešení:

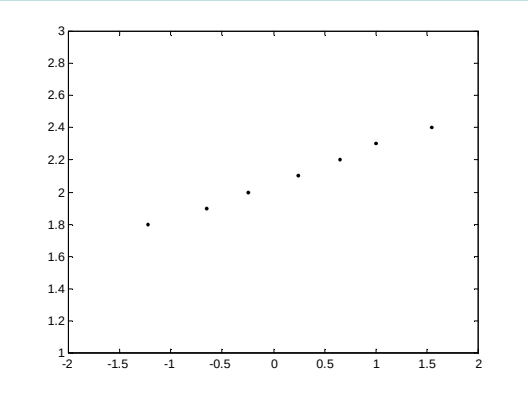
Na základě tabulky vytvořené při řešení předešlého příkladu stanovíme:

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

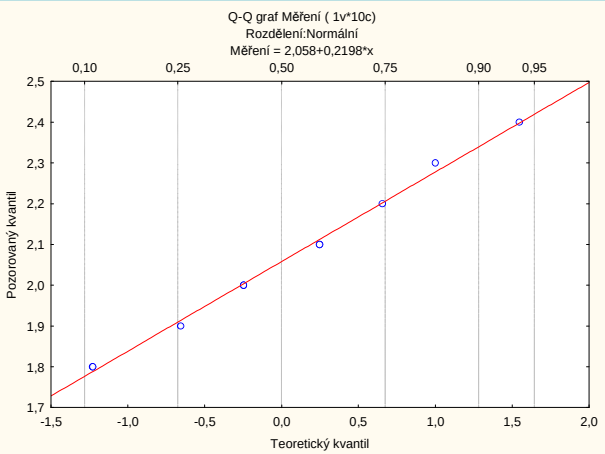
Kvantil – kvantilový graf



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

Řešení pomocí systému STATISTICA:

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.

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 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.

Příklad

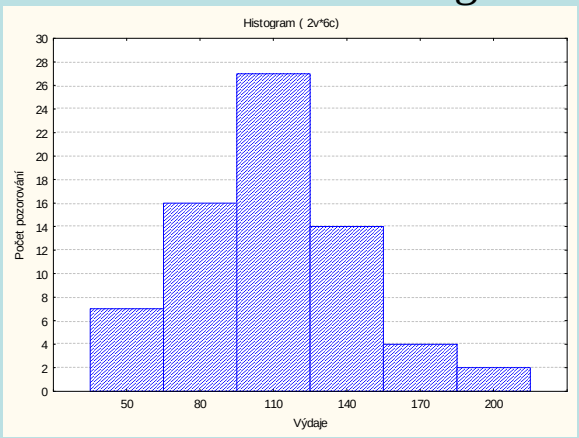
U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

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 domácností	7	16	27	14	4	2

Nakreslete histogram.

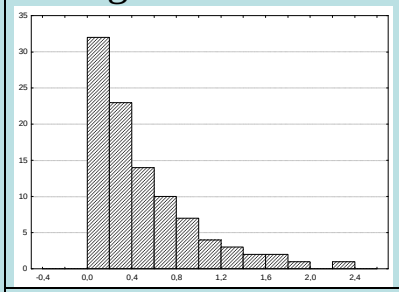
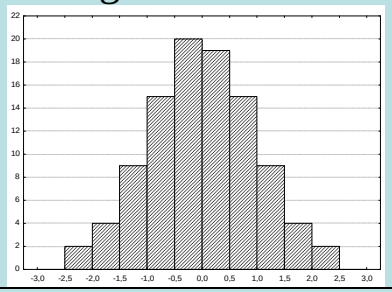
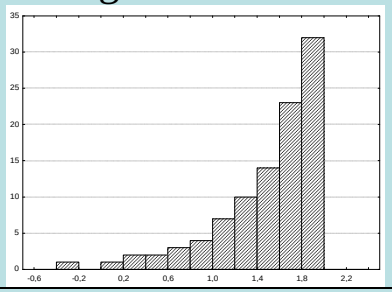
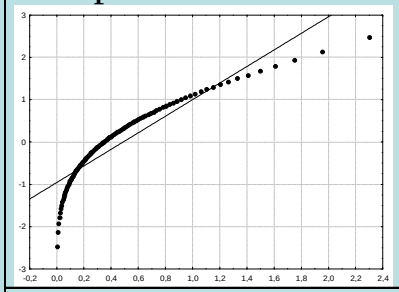
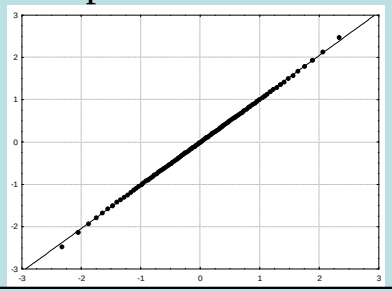
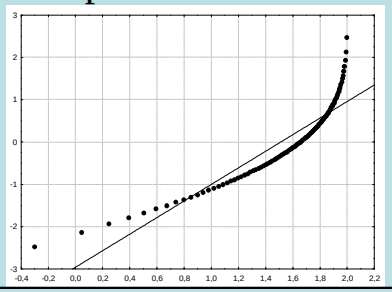
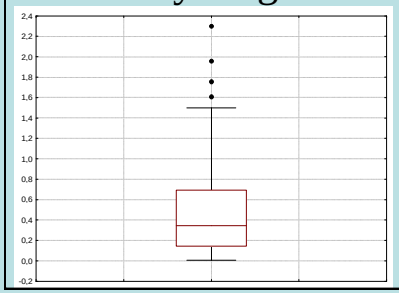
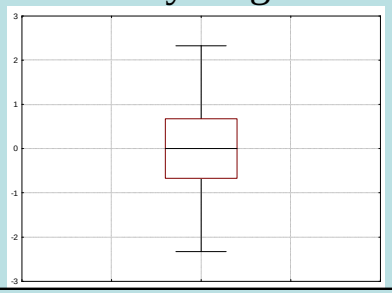
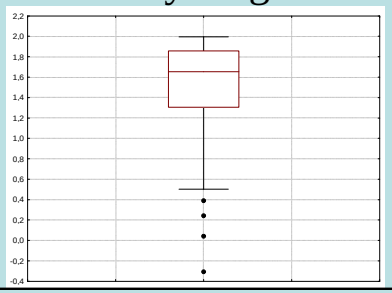
Řešení pomocí systému STATISTICA:

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.

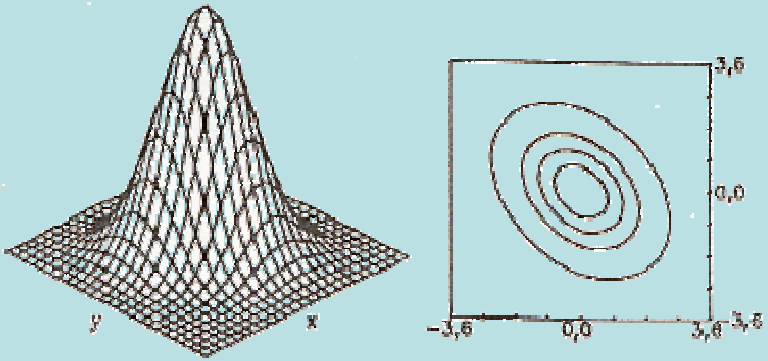
Vlastnosti rozložení četností datového souboru se projeví ve vzhledu diagnostických grafů:

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 

Dvourozměrný tečkový diagram

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.

Příklad

V dílně pracuje 15 dělníků. Byl u nich zjištěn počet směn odpracovaných za měsíc (náhodná veličina X) a počet zhotovených výrobků (náhodná veličina Y):

X 20 21 18 17 20 18 19 21 20 14 16 19 21 15 15
Y 92 93 83 80 91 85 82 98 90 60 73 86 96 64 81.

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 pomocí systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 15 případy.

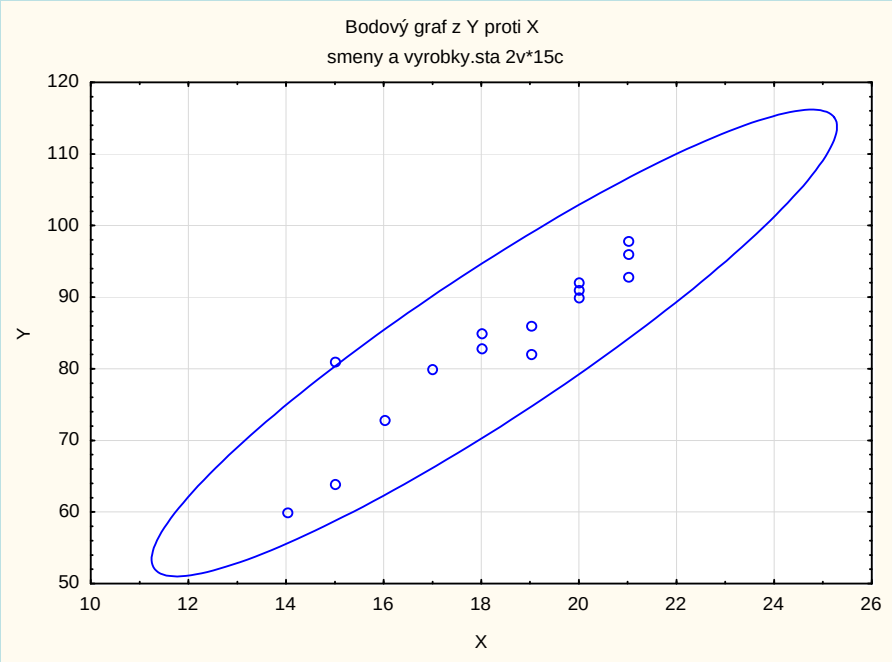
Nakreslíme dvourozměrný tečkový diagram: Grafy – 2D Grafy - Bodové grafy. Vypneme lineární proložení. Zadáme Proměnné – X – X, Y – Y – OK.

Dostaneme dvourozměrný tečkový diagram.

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 10, maximum 26, na svislé ose bude minimum 40, maximum 120. (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 počtem směn a počtem výrobků bude existovat určitý stupeň přímé lineární závislosti, tzn., že u dělníků, kteří odpracovali vysoký resp. nízký počet směn, lze očekávat vysoký resp. nízký počet zhotovených výrobků.