Vysoké učení technické v Brně

Fakulta strojního inženýrství

STATISTICKÁ ANALÝZA

Doc. RNDr. Zdeněk Karpíšek, CSc.

Přehledový učební text pro doktorské studium

BRNO 2008

Přednášející:

Doc. RNDr. Zdeněk Karpíšek, CSc.
Centrum pro jakost a spolehlivost ve výrobě
Odbor statistiky a optimalizace
Ústav matematiky
FSI VUT v Brně
E-mail: karpisek@fme.vutbr.cz

OBSAH

PŘEDMLUVA

Učební text obsahuje přehled metod nejčastěji používaných metod matematické statistiky a je pouze základní pomůckou pro studium. Pro individuální přípravu ke zkoušce jsou do každé kapitoly zařazeny neřešené příklady k procvičení a kontrolní otázky. K prohloubení znalostí se doporučuje literatura citovaná v textu a uvedená v závěrečné části. Tabulková část má sloužit k řešení úloh na odhady parametrů a testování statistických hypotéz. Dodatky 1 a 2 doplňují učební text o základní informace z popisné statistiky a teorie pravděpodobnosti.

Děkuji všem, kteří mně pomohli připomínkami a radami k přípravě tohoto vydání učebního textu. Rád přijmu všechny podněty a doporučení k jeho obsahu i zpracování.

Brno, říjen 2008 Zdeněk Karpíšek

ELEMENTY MATEMATICKÉ STATISTIKY

1 NÁHODNÝ VÝBĚR A JEHO CHARAKTERISTIKY

Matematická (inferenční, indukční) statistika poskytuje metody pro popis veličin náhodného charakteru pomocí jejich pozorovaných hodnot. Jedná se vlastně o určení vlastností rozdělení pravděpodobnosti náhodné veličiny nebo náhodného vektoru na základě jejich pozorovaných hodnot a v podstatě jde o řešení dvou základních úloh matematické statistiky:

- ***odhady parametrů a rozdělení***,
- ***testování statistických hypotéz o parametrech a rozděleních***.

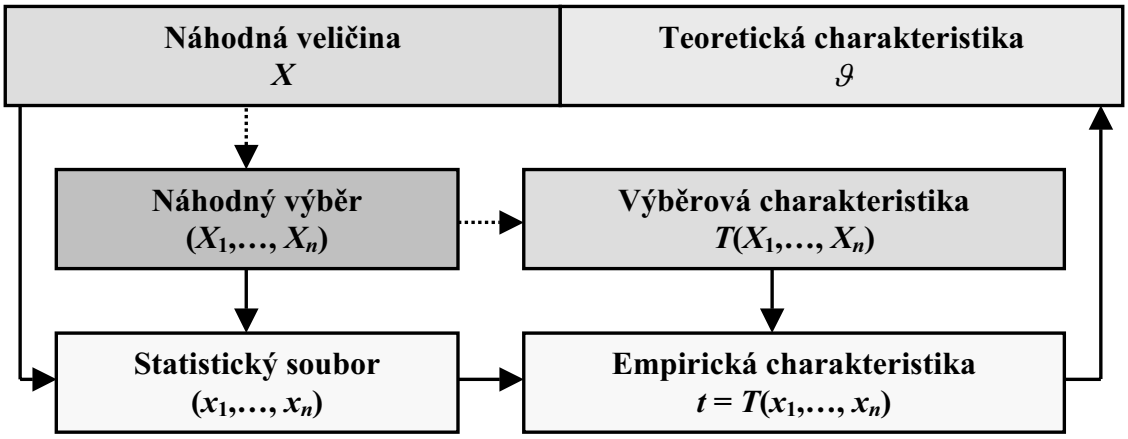
Tyto úlohy se dle potřeby kombinují, když např. odhadujeme nebo testujeme číselné charakteristiky rozdělení, vyšetřujeme závislosti náhodných veličin apod. Metody matematické statistiky jsou založeny na následujících pojmech.

Opakujeme-li n-krát nezávisle pokus, jehož výsledkem je hodnota náhodné veličiny X s distribuční funkcí $F(x, \vartheta)$, kde ϑ je reálný parametr (případně vektor parametrů anebo jejich funkce) daného rozdělení pravděpodobnosti, pozorujeme vlastně náhodný vektor $\mathbf{X} = (X_1,...,X_n)$ a předpokládáme, že jeho složky jsou nezávislé náhodné veličiny X_i se stejnou distribuční funkcí jakou má pozorovaná náhodná veličina X. Náhodný vektor $\mathbf{X} = (X_1,...,X_n)$ se nazývá ***náhodný výběr*** (z náhodné veličiny X nebo z jejího rozdělení pravděpodobnosti) a číslo n je ***rozsah*** náhodného výběru. Analogicky definujeme náhodný výběr z náhodného vektoru. Náhodný výběr má simultánní distribuční funkci $F(\mathbf{x};\vartheta) = F(x_1,...,x_n;\vartheta) = \prod_{i=1}^{n} F(x_i;\vartheta)$.

Číselný vektor $\mathbf{x} = (x_1,...,x_n)$, který získáme při realizaci náhodného výběru, kde x_i je pozorovaná hodnota složky X_i, $i = 1,...,n$, je ***statistický soubor*** s ***rozsahem*** n. Statistický soubor $\mathbf{x} = (x_1,...,x_n)$ je jinak řečeno pozorovaná hodnota náhodného výběru $\mathbf{X} = (X_1,...,X_n)$, což znamená, že při opakovaných realizacích náhodného výběru obdržíme obecně (a náhodně) různé statistické soubory. Množina všech hodnot náhodného výběru, tj. množina všech statistických souborů, tvoří tzv. ***výběrový prostor.***

Funkce náhodného výběru $T(X_1,...,X_n)$ je ***výběrová charakteristika*** nebo ***statistika***. Její hodnota na statistickém souboru $t = T(x_1,...,x_n)$ je ***empirická charakteristika*** nebo ***pozorovaná hodnota statistiky*** T. Výběrovou charakteristiku (statistiku) T (a tím také

empirickou charakteristiku t) volíme tak, nabývala na výběrovém prostoru s velkou pravděpodobností hodnot blízkých neznámé nebo předpokládané teoretické charakteristice, např. parametru ϑ pozorované náhodné veličiny X. Z toho vyplývá základní princip statistické indukce v matematické statistice, který je schematicky vyjádřen na obr. 1.1.



Obr. 1.1

Používáme zejména tyto výběrové charakteristiky:

1) ***výběrový průměr*** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

2) ***výběrový rozptyl*** $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$,

3) ***výběrová směrodatná odchylka*** $S = \sqrt{S^2}$,

4) ***výběrový koeficient korelace*** $R = \dfrac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{S(X)\,S(Y)}$ pro náhodný výběr z náhodného vektoru (X, Y), kde $S(X)$ a $S(Y)$ jsou výběrové směrodatné odchylky náhodných veličin X a Y.

Základní vlastnosti výběrového průměru $\bar{X}$ a výběrového rozptylu S^2 jsou:

a) Jestliže pozorovaná náhodná veličina X má střední hodnotu $E(X)$, pak

$$E\left(\bar{X}\right) = E(X).$$

b) Jestliže pozorovaná náhodná veličina X má rozptyl $D(X)$, pak

$$D\left(\bar{X}\right) = \frac{D(X)}{n},\ \sigma\left(\bar{X}\right) = \frac{\sigma(X)}{\sqrt{n}},\ E\left(S^2\right) = \frac{n-1}{n}D(X).$$

Hodnoty výběrových charakteristik jsou empirické charakteristiky, které získáme po zpracování statistického souboru. Např. aritmetický průměr $\bar{x}$ je pozorovaná hodnota

výběrového průměru $\overline{X}$ apod. Tyto hodnoty jsou však náhodné, jinak řečeno, empirické charakteristiky se při opakovaných realizacích náhodného výběru náhodně mění. Avšak z předcházejícího plyne, že např. pro $n \to \infty$ rozptyl výběrového průměru $D\left(\overline{X}\right) \to 0$, takže pro dostatečně velké *n* je takřka jistě aritmetický průměr $\overline{x}$ blízký neznámé střední hodnotě $E\left(X\right)$. Přitom ale $\sigma\left(\overline{X}\right) \to 0$ pouze s rychlostí $n^{1/2}$, což znamená, že např. pro dosažení dvojnásobné přesnosti aproximace neznámé střední hodnoty $E\left(X\right)$ aritmetickým průměrem $\overline{x}$ musíme zvýšit rozsah náhodného výběru čtyřikrát atd. Ve statistické literatuře se hovoří o tzv. ***statistické kletbě***.

Protože $\frac{n-1}{n} < 1$, je $E\left(S^2\right) < D\left(X\right)$, takže empirické hodnoty s^2 se vzhledem ke skutečnému (a obvykle neznámému) rozptylu $D\left(X\right)$ častěji vychylují doleva (do menších hodnot) od $D\left(X\right)$. Proto se mnohdy definuje výběrový rozptyl $\hat{S}^2$ ve tvaru

$$\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

a pro tento výběrový rozptyl je $E\left(\hat{S}^2\right) = D\left(X\right)$. Odpovídající rozptyl statistického souboru pak je

$$\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 .$$

Statistika $\hat{S}^2$ má však větší rozptyl než statistika S^2, ale pro velká *n* (řádově 100 a více) je rozdíl mezi těmito statistikami zanedbatelný. Analogicky definujeme výběrovou směrodatnou odchylku $\hat{S}$ a směrodatnou odchylku statistického souboru $\hat{s}$. Různé definice uvedených charakteristik je nutno respektovat při zpracování statistického souboru na PC pomocí statistických programů a také ve vzorcích jak pro odhady parametrů, tak i pro testování statistických hypotéz.

Nejčastěji řešené úlohy při aplikacích metod matematické statistiky se týkají pozorovaných náhodných veličin s normálním rozdělením pravděpodobnosti. Jestliže pozorovaná náhodná veličina X má normální rozdělení $\mathrm{N}(\mu; \sigma^2)$, pak statistika:

a) $\overline{X}$ má normální rozdělení $\mathrm{N}(\mu; \frac{\sigma^2}{n})$,

b) $\frac{\overline{X} - \mu}{\sigma}\sqrt{n}$ má normální rozdělení $\mathrm{N}(0;1)$,

c) $\frac{\overline{X}-\mu}{S}\sqrt{n-1}$ má tzv. ***Studentovo rozdělení*** S(n – 1) s $n-1$ stupni volnosti, nazývané též ***t-rozdělení***,

d) $\frac{nS^2}{\sigma^2}$ má tzv. ***Pearsonovo rozdělení*** $\chi^2(n-1)$ s $n-1$ stupni volnosti, nazývané též ***chí-kvadrát rozdělení***.

Jestliže pozorovaná náhodná veličina X má normální rozdělení $\mathrm{N}\left(\mu(X);\sigma^2(X)\right)$ a pozorovaná náhodná veličina Y má normální rozdělení $\mathrm{N}\left(\mu(Y);\sigma^2(Y)\right)$, X a Y jsou nezávislé a také náhodné výběry $\left(X_1,...,X_{n_1}\right)$, $\left(Y_1,...,Y_{n_2}\right)$ jsou nezávislé, pak statistika:

a) $\frac{\overline{X}-\overline{Y}-\left(\mu(X)-\mu(Y)\right)}{\sqrt{\frac{\sigma^2(X)}{n_1}+\frac{\sigma^2(Y)}{n_2}}}$ má normální rozdělení $\mathrm{N}(0;1)$,

b) $\frac{\overline{X}-\overline{Y}-\left(\mu(X)-\mu(Y)\right)}{\sqrt{n_1S^2(X)+n_2S^2(Y)}}\sqrt{\frac{n_1n_2\left(n_1+n_2-2\right)}{n_1+n_2}}$ má pro stejné rozptyly $\sigma^2(X)=\sigma^2(Y)$ Studentovo rozdělení $\mathrm{S}\left(n_1+n_2-2\right)$,

c) $\frac{\frac{n_1S^2(X)}{n_1-1}}{\frac{n_2S^2(Y)}{n_2-1}}$ má pro stejné rozptyly $\sigma^2(X)=\sigma^2(Y)$ tzv. ***Fisherovo-Snedecorovo rozdělení*** F(n_1-1, n_2-1) s n_1-1 a n_2-1 stupni volnosti.

Jestliže $X_1, X_2, ...$ je posloupnost nezávislých náhodných veličin s libovolným stejným rozdělením pravděpodobnosti (např. i asymetrickým nebo diskrétním), které má střední hodnotu μ_0 a směrodatnou odchylku σ_0, pak posloupnost náhodných veličin

$$\frac{\frac{1}{n}\sum_{i=1}^{n}X_i-\mu_0}{\sigma_0}\sqrt{n}$$

konverguje (v distribuci) k náhodné veličině U s normovaným normálním rozdělením $\mathrm{N}(0;1)$. Odtud plyne, že při dostatečně velkém rozsahu náhodného výběru n můžeme rozdělení pravděpodobnosti výběrového aritmetického průměru $\overline{X}$ pro libovolnou pozorovanou náhodnou veličinu X se střední hodnotou μ_0 a rozptylem σ_0^2 aproximovat

normálním rozdělením $\mathrm{N}(\mu_0;\frac{\sigma_0^2}{n})$. To také znamená, že při dostatečně velkém rozsahu n můžeme stanovit intervalový odhad např. střední hodnoty μ_0 pozorované náhodné veličiny X s jiným než normálním (dokonce i neznámým) rozdělením pravděpodobnosti. Tento interval zkonstruujeme ze získaného statistického souboru a jeho spolehlivost (tj. pravděpodobnost zachycení μ_0) pak vyjádříme pomocí normálního rozdělení pravděpodobnosti.

Výše uvedená tzv. statistická rozdělení pravděpodobnosti jsou tabelována (viz Statistické tabulky na konci tohoto učebního textu) a je také možno určit jejich hodnoty pomocí Excelu, profesionálních statistických softwarů a statistických apletů na Internetu. Detailní informace o výše uvedených a dalších používaných statistikách, jejich rozděleních pravděpodobnosti a asymptotických vlastnostech lze nalézt např. v [2], [3], [8], [15], [17], [30].

Kontrolní otázky

1. Jaké dvě základní úlohy se řeší v matematické statistice? Uveďte konkrétní příklady.
2. Definujte náhodný výběr a jeho realizaci.
3. Definujte výběrovou charakteristiku a empirickou charakteristiku.
4. Popište princip statistické indukce.
5. Popište základní vlastnosti výběrového průměru a výběrového rozptylu.
6. Jaká základní tzv. statistická rozdělení pravděpodobnosti používáme?
7. Jaké rozdělení pravděpodobnosti má výběrový průměr, jestliže pozorovaná náhodná veličina má normální rozdělení?
8. Jakým rozdělením pravděpodobnosti můžeme pro dostatečně velký rozsah náhodného výběru aproximovat rozdělení výběrového průměru?

2 ODHADY PARAMETRŮ

Bodové a intervalové odhady

Předpokládáme, že pozorovaná náhodná veličina X (případně náhodný vektor) má distribuční funkci $F(x,\vartheta)$ známého tvaru, kde ϑ je ***parametr*** (reálné číslo nebo reálný vektor) rozdělení pravděpodobnosti X. Skutečnou hodnotu parametru ϑ obvykle neznáme a odhadujeme ji pomocí získaného statistického souboru. Parametrem ϑ může také být číselná charakteristika náhodné veličiny (náhodného vektoru), např. střední hodnota $E(X)$, rozptyl $D(X)$, koeficient korelace $\rho(X,Y)$ apod., případně tzv. ***parametrická funkce***, tj. funkce parametrů rozdělení. Množina všech uvažovaných hodnot parametru ϑ se nazývá ***parametrický prostor***. Podle způsobu provedení rozdělujeme odhady na ***odhady bodové*** a ***intervalové***.

Odhadem T parametru ϑ je statistika $T(X_1,...,X_n)$, která na celém parametrickém prostoru nabývá hodnot blízkých parametru ϑ. Používáme zejména tyto odhady:

1. Odhad T parametru ϑ je ***nestranný*** (***nevychýlený***), jestliže jeho střední hodnota $E(T) = \vartheta$. Pokud je $E(T) \neq \vartheta$, jde o ***stranný*** (***vychýlený***) odhad.
2. Je-li rozptyl nestranného odhadu T nejmenší z rozptylů všech nestranných odhadů téhož parametru ϑ, je T ***nejlepší nestranný odhad***.
3. Odhad T je ***konzistentní***, jestliže $\lim_{n\to\infty} P\left(|T-\vartheta| < \varepsilon\right) = 1$ pro libovolné reálné číslo $\varepsilon > 0$.

Platí:

a) $\overline{X}$ je nestranný konzistentní odhad střední hodnoty $E(X)$,

b) $\frac{n}{n-1}S^2$ je nestranný konzistentní odhad rozptylu $D(X)$,

c) odhady a) a b) jsou pro normální rozdělení X také nejlepší.

Další typy odhadů (např. ***maximálně věrohodné odhady***) jsou popsány v [2], [3], [8], [15], [17], [30].

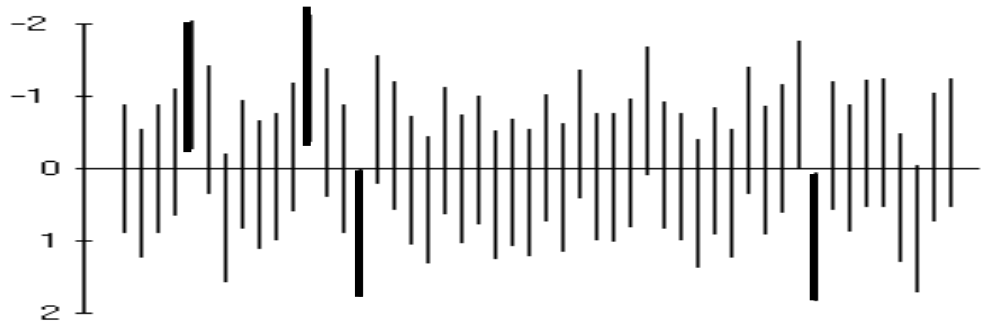
Bodový odhad parametru ϑ je pozorovaná hodnota $t = T\left(x_1,...,x_n\right)$ odhadu T na statistickém souboru $\left(x_1,...,x_n\right)$. Bodové odhady základních číselných charakteristik jsou

$$E\left(X\right) = \overline{x},\ D\left(X\right) = \frac{n}{n-1}s^2,\ \sigma\left(X\right) = \sqrt{\frac{n}{n-1}}s,\ \rho\left(X,Y\right) = r,$$

kde $\overline{x}, s^2, s, r$ jsou empirické charakteristiky získané ze statistického souboru $\left(x_1,...,x_n\right)$, resp. $\left(\left(x_1,y_1\right),\ldots,\left(x_n,y_n\right)\right)$, a znaménko = vyjadřuje pouze odhad a nikoli rovnost hodnot.

Interval spolehlivosti (***konfidenční interval***) pro parametr ϑ ***se spolehlivostí*** $1-\alpha$, kde $\alpha \in (0;1)$, je dvojice takových statistik $(T_1; T_2)$, že $P(T_1 \le \vartheta \le T_2) = 1-\alpha$ pro každou hodnotu parametru ϑ. ***Intervalový odhad*** parametru ϑ ***se spolehlivostí*** $1-\alpha$ je interval $\langle t_1; t_2 \rangle$ a píšeme $\vartheta \in \langle t_1; t_2 \rangle$, kde t_1, t_2 jsou hodnoty statistik T_1, T_2 na daném statistickém souboru $(x_1, ..., x_n)$, resp. $((x_1, y_1), \ldots, (x_n, y_n))$.

Spolehlivost $1-\alpha$ volíme blízkou jedné, podle konvence obvykle 0,95 nebo 0,99, a uvádíme ji také v %. Spolehlivost $1-\alpha$ znamená, že při mnoha opakovaných výběrech s konstantním rozsahem n z daného základního souboru zhruba $(1-\alpha)100$ % všech intervalových odhadů obsahuje skutečnou hodnotu parametru ϑ a naopak $\alpha 100$ % jich tuto hodnotu neobsahuje. Situaci ilustruje počítačově simulovaný příklad na obr. 2.1, kde $\vartheta = 0$ a tučně jsou vyznačeny případy odpovídající riziku chybného odhadu α, tj. intervalové odhady, které nezachytily hodnotu parametru ϑ.



4 intervalové odhady z 50 provedených intervalových odhadů se spolehlivostí 0,95 neobsahují odhadovanou hodnotu 0, tj. pozorovaná spolehlivost odhadů je 0,92

Obr. 2.1

Snížení rizika α, tedy zvýšení spolehlivosti $1-\alpha$, vede při zachování rozsahu výběru n ke zvětšení velikosti intervalového odhadu. Pro $\alpha = 0$, tedy pro 100 % spolehlivost, je intervalovým odhadem celý parametrický prostor a to nemá v aplikacích rozumný význam. Zmenšit velikost intervalového odhadu je možno:

a) snížením spolehlivosti, což není vhodné, protože se tím vlastně nepřesnost odhadu zvětší,
b) zvýšením rozsahu výběru n, ovšem s ohledem na "kletbu statistiky", neboť velikost intervalového odhadu se zmenší víceméně úměrně $n^{1/2}$,
c) volbou jiného a současně "užšího" intervalu spolehlivosti pro daný parametr, pokud takovou statistiku T známe.

Na druhé straně je zřejmé, že bodový odhad má spolehlivost nulovou anebo blízkou nule (pro

diskrétní rozdělení pravděpodobnosti pozorované náhodné veličiny X). Intervalové odhady proto poskytují významně dokonalejší pohled na vlastnosti pozorované náhodné veličiny než odhady bodové a navíc bodový odhad obsahují.

Intervalové odhady dělíme na ***dvoustranné*** (***oboustranné***) a ***jednostranné*** podle toho, zda je ohraničujeme oboustranně anebo jednostranně. Často volíme statistiky T_1, T_2 ve tvaru $T_1 = T - \delta_1$ a $T_2 = T + \delta_2$, kde $\delta_1 \geq 0$ a $\delta_2 \geq 0$ jsou vhodná reálná čísla (závisející na spolehlivosti $1 - \alpha$ a rozsahu náhodného výběru n) a T je nějaký odhad parametru ϑ. Poznamenejme, že z předem dané délky Δ dvoustranného odhadu intervalového odhadu a spolehlivosti $1 - \alpha$ je možno určit potřebný rozsah výběru.

Odhady parametrů normálního rozdělení

Předpokládáme, že pozorovaná náhodná veličina X, resp. náhodný vektor (X, Y), má normální rozdělení pravděpodobnosti s parametry μ, σ^2, resp. ρ.

Bodové odhady jsou

$$\mu = \overline{x},\ \sigma^2 = \frac{n}{n-1}s^2,\ \sigma = \sqrt{\frac{n}{n-1}}s,\ \rho = r\ .$$

Intervalový odhad střední hodnoty μ při neznámém rozptylu σ^2 je

$$\left\langle \overline{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n-1}}; \overline{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n-1}} \right\rangle,$$

kde $t_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$- kvantil Studentova rozdělení S(k) s $k = n - 1$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T2**.

Intervalový odhad rozptylu σ^2 je

$$\left\langle \frac{ns^2}{\chi^2_{1-\alpha/2}}; \frac{ns^2}{\chi^2_{\alpha/2}} \right\rangle,$$

kde χ^2_P je P - kvantil Pearsonova rozdělení $\chi^2(k)$ s $k = n - 1$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T3**. Z uvedeného intervalového odhadu získáme po odmocnění jeho mezí ***intervalový odhad směrodatné odchylky*** σ.

Příklad 2.1

Měřením délky 10 válečků byl získán statistický soubor s empirickými charakteristikami $\overline{x} = 5,37$ mm, $s^2 = 0{,}0019$ mm^2 a $s = 0{,}044$ mm (viz příklad 2.1 z učebního textu MME 2). Určete bodové odhady střední hodnoty, rozptylu a směrodatné odchylky. Za předpokladu, že

naměřená délka X má normální rozdělení pravděpodobnosti, určete intervalové odhady těchto číselných charakteristik se spolehlivostí 0,95.

Ř e š e n í:

Bodové odhady jsou:

střední délka válečku $\mu = 5{,}37$ mm,

rozptyl délky válečku $\sigma^2 = \frac{10}{9} 0{,}0019 = 0{,}00211$ mm^2,

směrodatná odchylka délky válečku $\sigma = \sqrt{0{,}00211} \approx 0{,}046$ mm.

Intervalový odhad střední délky válečku μ se spolehlivostí 0,95 je, neboť $t_{0{,}975} = 2{,}262$ pro 9 stupňů volnosti z tabulky **T2**,

$$\mu \in \langle 5{,}37 - 2{,}262 \frac{\sqrt{0{,}0019}}{\sqrt{10-1}};\ 5{,}37 + 2{,}262 \frac{\sqrt{0{,}0019}}{\sqrt{10-1}} \rangle \approx \langle 5{,}337;\ 5{,}403 \rangle \text{ mm}.$$

Intervalový odhad rozptylu délky válečku σ^2 se spolehlivostí 0,95 je, neboť $\chi^2_{0{,}025} = 2{,}700$ a $\chi^2_{0{,}975} = 19{,}023$ pro 9 stupňů volnosti z tabulky **T3**,

$$\sigma^2 \in \langle \frac{10.0{,}0019}{19{,}023};\ \frac{10.0{,}0019}{2{,}700} \rangle \approx \langle 0{,}00100;\ 0{,}00704 \rangle \text{ mm}^2,$$

takže intervalový odhad směrodatné odchylky délky válečku σ je

$$\sigma \in \langle \sqrt{0{,}00100};\ \sqrt{0{,}00704} \rangle \approx \langle 0{,}0316;\ 0{,}0839 \rangle \text{ mm}.$$

Intervalový odhad koeficientu korelace ρ pro $n \geq 10$ a $r \neq \pm 1$ je

$$\langle tgh\ z_1; tgh\ z_2 \rangle,$$

kde

$$z_1 = w - \frac{u_{1-\alpha/2}}{\sqrt{n-3}},\quad z_2 = w + \frac{u_{1-\alpha/2}}{\sqrt{n-3}},\quad w = \frac{1}{2}\left(\ln \frac{1+r}{1-r} + \frac{r}{n-1} \right),\quad tgh\ z = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{e^{2z} - 1}{e^{2z} + 1},$$

a $u_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$- kvantil normovaného normálního rozdělení N(0;1), jehož hodnoty lze získat z tabulky **T1** s hodnotami distribuční funkce $\Phi(u)$. Pro $1 - \alpha = 0{,}95$ je $u_{0{,}975} = 1{,}960$ a pro $1 - \alpha = 0{,}99$ je $u_{0{,}995} = 2{,}576$. Uvedený odhad je pouze přibližný, avšak jeho přesnost je v praktických úlohách zcela postačující (přesný odhad není znám).

Příklad 2.2

Sledováním nákladů X a ceny stejného výrobku Y u 10 výrobců byl získán dvourozměrný statistický soubor s koeficientem korelace $r = 0{,}82482$ (viz příklad 2.3 z učebního textu

MME 2). Určete bodový odhad a intervalový odhad se spolehlivosti 0,99 koeficientu korelace ρ základního souboru.

Ř e š e n í:

Bodový odhad koeficientu korelace nákladů a ceny je $\rho = 0{,}82482$. Po dosazení je

$$w = \frac{1}{2}\left(\ln \frac{1+0{,}82482}{1-0{,}82482} + \frac{0{,}82482}{10-1} \right) \approx 1{,}21753 .$$

Z tabulky **T1** je $u_{0{,}995} = 2{,}576$, takže

$$z_1 = 1{,}21753 - \frac{2{,}576}{\sqrt{10-3}} \approx 0{,}24397 , \; z_2 = 1{,}21753 + \frac{2{,}576}{\sqrt{10-3}} \approx 2{,}19110$$

a intervalový odhad koeficientu korelace nákladů a ceny ρ se spolehlivostí 0,99 je

$$\rho \in \langle tgh\ 0{,}24397; tgh\ 2{,}19110 \rangle \approx \langle 0{,}239242; 0{,}975313 \rangle .$$

Odhady parametru binomického rozdělení

Předpokládáme, že pozorovaná náhodná veličina X má alternativní rozdělení pravděpodobnosti s parametrem p, tedy binomické rozdělení Bi(1; p). Při odhadu parametru p jde vlastně o odhad velikosti podílu prvků základního souboru majících sledovanou vlastnost. Přitom X_i nabývá hodnotu $x_i = 1$, resp. 0, jestliže i-tý náhodně vybraný prvek má, resp. nemá, sledovanou vlastnost, $i = 1,\ldots, n$. Nechť x je počet prvků se sledovanou vlastností z n náhodně vybraných prvků, tedy $x = \sum_{i=1}^{n} x_i$.

Bodový odhad je $p = \frac{x}{n}$.

Intervalový odhad p je pro $n > 30$

$$\left\langle \frac{x}{n} - u_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}};\ \frac{x}{n} + u_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} \right\rangle ,$$

kde $u_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$- kvantil normovaného normálního rozdělení, jehož hodnoty lze získat z tabulky **T1**. Uvedený odhad je pouze přibližný, avšak jeho přesnost je pro velká n v praktických úlohách obvykle postačující.

Příklad 2.3

Při průzkumu zájmu o nový výrobek odpovědělo ze 400 dotázaných zákazníků supermarketu

STAMET kladně na otázku, zda si nový výrobek koupí, 80 zákazníků. Určete bodový a intervalový odhad podílu zákazníků p ze základního souboru všech zákazníků supermarketu STAMET.

Ř e š e n í:

Protože $x = 80$ a $n = 400$, je bodový odhad $p = \frac{80}{400} = 0,2$, tedy 20 % všech zákazníků supermarketu STAMET si chce koupit nový výrobek.

Z tabulky **T1** pro spolehlivost 0,95 je $u_{0,975} = 1,960$, takže intervalový odhad podílu zákazníků p se spolehlivostí 0,95 je

$$p \in \left\langle \frac{80}{400} - 1,960\sqrt{\frac{\frac{80}{400}\left(1-\frac{80}{400}\right)}{400}}; \frac{80}{400} + 1,960\sqrt{\frac{\frac{80}{400}\left(1-\frac{80}{400}\right)}{400}} \right\rangle = \ldots = < 0,1608;\ 0,2392 >.$$

Pro spolehlivost 0,99 obdržíme analogickým způsobem intervalový odhad

$$p \in < 0,1485;\ 0,2515 >.$$

Se spolehlivostí 0,95, resp. 0,99, si nový výrobek koupí přibližně 16 až 24 %, resp. 15 až 25 %, všech zákazníků supermarketu STAMET. Pokud má STAMET celkem 10 000 zákazníků, lze víceméně očekávat, že prodá cca 2 000 nových výrobků. Z intervalového odhadu můžeme pak se spolehlivostí 0,95 usuzovat, že STAMET prodá přibližně 10 000·0,16 = 1 600 až 10 000·0,24 = 2 400 nových výrobků.

Příklady k procvičení

Příklad 2.4

Určete bodový a intervalový odhad se spolehlivostí 0,99 parametrů μ a σ^2 normálního rozdělení, jestliže realizací náhodného výběru byl získán statistický soubor o rozsahu $n = 18$ s aritmetickým průměrem $\bar{x} = 50,1$ a s rozptylem $s^2 = 17,64$.

V ý s l e d e k: $\mu = 50,1$; $\sigma^2 = 18,678$; $\mu \in <47,09;\ 53,10>$; $\sigma^2 \in <8,894;\ 55,705>$

Příklad 2.5

Statistický soubor o rozsahu $n = 12$ má aritmetický průměr $\bar{x} = 77,55$ a rozptyl $s^2 = 1045,65$. Určete bodový a intervalový odhad μ a σ základního souboru se spolehlivostí 0,99.

V ý s l e d e k: $\mu = 77,55$; $\sigma = 33,78$; $\mu \in <47,267;\ 107,833>$; $\sigma \in <21,638;\ 69,47>$

Příklad 2.6

U sta náhodně vybraných pracovníků stejné kategorie byla zjištěna hodinová tarifní mzda (Kč) a vypočteny empirické charakteristiky $\bar{x}$ = 98,64 Kč a s^2 = 1,1979 Kč. Určete bodové a intervalové odhady střední hodinové tarifní mzdy μ a směrodatné odchylky σ se spolehlivostí 99% za předpokladu, že základní soubor má normální rozdělení.

V ý s l e d e k: μ = 98,64 Kč; σ = 1,10 Kč;

$\mu \in$ <98,35; 98,93> Kč; $\sigma \in$ <0,93; 1,34> Kč

Příklad 2.7

Z patnácti nezávislých pozorování byl vypočten bodový odhad střední hodnoty 424,7 ms^{-1} a směrodatné odchylky 8,7 ms^{-1} maximální rychlosti letadla. Určete intervalový odhad střední hodnoty a směrodatné odchylky maximální rychlosti se spolehlivostí 95% za předpokladu normálního rozdělení maximální rychlosti.

V ý s l e d e k: $\mu \in$ <419,88; 429,52> ms^{-1} ; $\sigma \in$ <6,37; 13,72> ms^{-1}

Příklad 2.8

Bylo provedeno 5 nezávislých a stejně přesných měření ke stanovení objemu nádoby: 4,781; 4,792; 4,795; 4,779; 4,769 (v litrech). Stanovte intervalový odhad střední hodnoty objemu nádoby se spolehlivostí 0,99 za předpokladu normálního rozdělení.

V ý s l e d e k: <4,761; 4,805> l

Příklad 2.9

Při kontrole záručních listů určitého druhu masové konzervy ve skladu hypermarketu bylo náhodně vybráno 320 konzerv a zjištěno, že 59 jich má prošlou záruční lhůtu. Stanovte bodový a intervalový odhad se spolehlivostí 95% procenta konzerv s prošlou záruční lhůtou ve skladech hypermarketu firmy. Totéž určete pro roční sklad hypermarketu s počtem 20 000 konzerv.

V ý s l e d e k: p = 0,184 = 18,4 %; $p \in$ <0,142; 0,226> = <14,2; 22,6> %; N = 3680;

$N \in$ <2840; 4520>

Příklad 2.10

Při náhodném výběru pneumatik vyráběných velkou evropskou nadnárodní společností 10% pneumatik nevyhovělo nové normě. Pro rozsah výběru (a) n = 100, (b) n = 400, (c) n = 1600 určete 95%-ní interval spolehlivosti pro podíl p pneumatik vyráběných touto společností, které nevyhovují nové normě.

V ý s l e d e k: (a) <0,041; 0,159>; (b) <0,071; 0,129>; (c) <0,085; 0,115>

Kontrolní otázky

1. Definujte pojem odhadu parametru a jeho druhy.
2. Definujte bodový odhad a uveďte bodové odhady základních číselných charakteristik.
3. Popište interval spolehlivosti a intervalový odhad parametrů.
4. Jaký význam má spolehlivost intervalového odhadu?
5. Jaké druhy intervalových odhadů používáme?
6. Jaký vliv má změna spolehlivosti na velikost intervalového odhadu při zachování rozsahu náhodného výběru?
7. Jaký obecný vliv má změna rozsahu náhodného výběru na velikost intervalového odhadu při zachování jeho spolehlivosti?
8. Jakou spolehlivost má bodový odhad?

3 TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

Statistická hypotéza a její test

Při sledování náhodných veličin a náhodných vektorů jsme často nuceni ověřit určité předpoklady či domněnky o jejich vlastnostech pomocí jejich pozorovaných hodnot. Jedná se např. o rozhodnutí, zda nová technologie, seřízení stroje, reklama, změna financování, řízení firmy apod. vedly ke změně ve sledovaných parametrech výrobku, obratu, zisku apod., anebo zda jakost dodávky výrobků či surovin má dohodnutou úroveň.

Statistická hypotéza H je tvrzení o vlastnostech rozdělení pravděpodobnosti pozorované náhodné veličiny X s distribuční funkcí $F(x,\vartheta)$ nebo náhodného vektoru (X, Y) se simultánní distribuční funkcí $F(x,y,\vartheta)$ apod. Postup, jímž ověřujeme danou hypotézu, se nazývá ***test statistické hypotézy***. Proti testované hypotéze H, nazývané také ***nulová hypotéza***, stavíme tzv. ***alternativní hypotézu*** $\overline{H}$, kterou volíme dle požadavků úlohy. Jestliže H je hypotéza, že parametr ϑ má hodnotu ϑ_0, píšeme $H:\vartheta=\vartheta_0$. Případ $\overline{H}:\vartheta\neq\vartheta_0$ je ***dvoustranná*** alternativní hypotéza a $\overline{H}:\vartheta>\vartheta_0$, resp. $\overline{H}:\vartheta<\vartheta_0$, je ***jednostranná*** alternativní hypotéza. Hypotéza může být ***jednoduchá***, jestliže uvažujeme jedinou hypotetickou hodnotu $\vartheta=\vartheta_0$ anebo naopak ***složená***, např. $\vartheta\neq\vartheta_0$. Dále rozdělujeme hypotézy na ***parametrické***, kdy jde tvrzení o parametrech pozorované náhodné veličiny X, a na ***neparametrické***, kdy jde o tvrzení o kvalitativních vlastnostech této náhodné veličiny.

Testovaná hypotéza H se někdy v literatuře, resp. aplikacích na PC, označuje symbolem H_0, resp. *H0*, a alternativní hypotéza $\overline{H}$ symbolem H_1, H_A, resp. *HA*.

Pro testování hypotézy $H:\vartheta=\vartheta_0$ proti nějaké zvolené alternativní hypotéze $\overline{H}$ se konstruuje vhodná statistika $T(X_1,...,X_n)$, tzv. ***testové kritérium***. Obor hodnot testového kritéria $T(X_1,...,X_n)$ se za předpokladu, že platí hypotéza $H:\vartheta=\vartheta_0$, rozdělí na dvě disjunktní podmnožiny: ***kritický obor*** W_α a jeho doplněk $\overline{W}_\alpha$ (viz obr. 8.2). Kritický obor W_α se vzhledem k alternativní hypotéze $\overline{H}$ stanoví tak, aby pravděpodobnost toho, že testové kritérium $T(X_1,...,X_n)$ nabude hodnotu z kritického oboru W_α, byla α (přesněji pro diskrétní náhodnou veličinu T nejvýše α). Číslo $\alpha\in(0;1)$ je ***hladina významnosti testu*** a volíme ji blízkou nule, obvykle 0,05 anebo 0,01. Hladina významnosti se někdy uvádí také v % (např. v některých softwarových aplikacích pro PC), tedy obvykle 5 % anebo 1 %.

Rozhodnutí o hypotéze H pomocí pozorovaných hodnot náhodné veličiny X je pak založeno na následující konvenci. Jestliže tzv. ***pozorovaná hodnota testového kritéria*** $t = T(x_1,...,x_n)$ na získaném statistickém souboru $(x_1,...,x_n)$ padne do kritického oboru, tedy $t \in W_\alpha$, ***zamítáme*** hypotézu H a současně ***nezamítáme*** hypotézu $\overline{H}$ na hladině významnosti α. Jestliže naopak nepadne t do kritického oboru, tedy $t \in \overline{W}_\alpha$, ***nezamítáme*** hypotézu H a současně ***zamítáme*** hypotézu $\overline{H}$ na hladině významnosti α. Nezamítnutí hypotézy H, resp. $\overline{H}$, neznamená ještě prokázání její platnosti, neboť jsme na základě realizace náhodného výběru získali pouze informace, které nestačí na její zamítnutí. Je-li to možné, je vhodné před ***přijetím*** dané hypotézy zvětšit rozsah statistického souboru a znovu hypotézu H testovat.

Při testování hypotézy H mohou nastat čtyři možnosti znázorněné na obr. 3.1. Jestliže zamítáme neplatnou hypotézu anebo nezamítáme platnou hypotézu, je vše v pořádku, avšak při rozhodnutí o hypotéze H na základě testu se můžeme dopustit jedné ze dvou chyb:

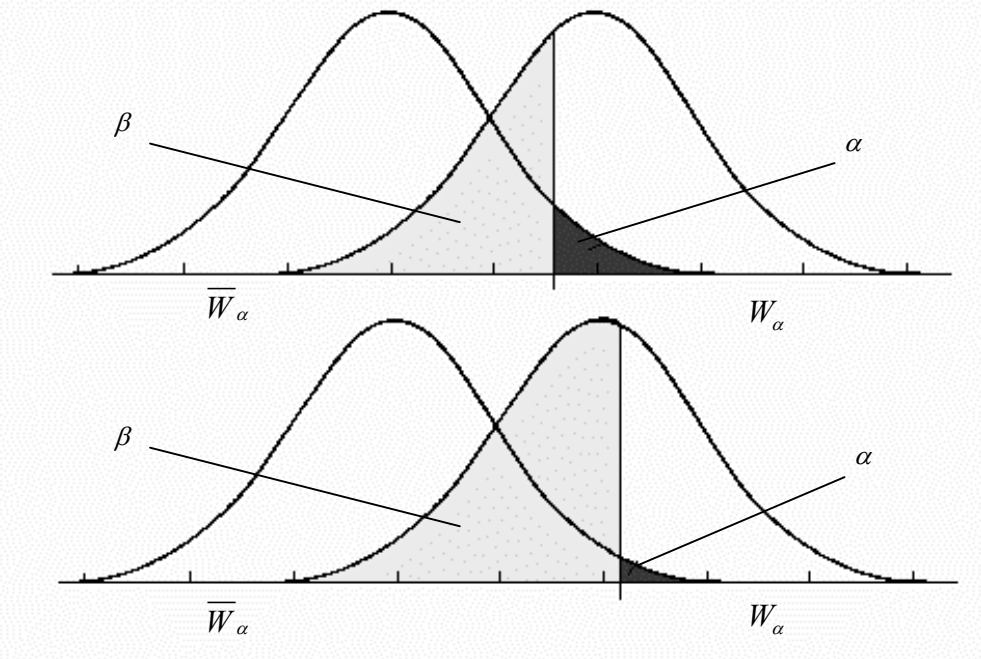
1. ***Chyba prvního druhu*** nastane, jestliže hypotéza H platí, avšak $t \in W_\alpha$, takže hypotézu H zamítáme. Pravděpodobnost této chyby je hladina významnosti $\alpha = P(T \in W_\alpha / H)$.
2. ***Chyba druhého druhu*** nastane, jestliže hypotéza H neplatí, avšak $t \notin W_\alpha$ (tj. $t \in \overline{W}_\alpha$), takže hypotézu H nezamítáme. Pravděpodobnost této chyby je $\beta = P(T \notin W_\alpha / \overline{H})$ a pravděpodobnost $1 - \beta = P(T \in W_\alpha / \overline{H})$ je tzv. ***síla testu***.

H	PLATÍ	NEPLATÍ
ZAMÍTÁME	CHYBA 1. DRUHU	-------
NEZAMÍTÁME	-------	CHYBA 2. DRUHU

Obr. 3.1

Hladina významnosti, tj. pravděpodobnost chyby prvního druhu α má ten praktický význam, že při mnoha opakovaných realizacích náhodného výběru (např. řádově v tisících) a současné platnosti testované hypotézy H se v přibližně $100\alpha\,\%$ testech této hypotézy zmýlíme, tedy zamítneme platnou hypotézu. Podobně když hypotéza H neplatí, tak se v přibližně $100\beta\,\%$ testech zmýlíme a nezamítneme ji. Avšak snížením hladiny významnosti α se při nezměněném rozsahu statistického souboru n zvýší β a naopak, takže pro zvolenou hladinu významnosti α zajišťujeme snížení β zvýšením rozsahu n. Riziko chyb prvního i druhého druhu nelze v reálných úlohách eliminovat, pouze je můžeme snížit. Vztah mezi α a

β je ilustrován na obr. 3.2, kde pro jednoduchost je i alternativní hypotéza $\overline{H}$ jednoduchá. Na tomto obrázku křivky vlevo odpovídají hustotě (pravděpodobnostní funkci) testového kritéria T při platnosti hypotézy H a křivky vpravo odpovídají hustotě (pravděpodobnostní funkci) testového kritéria T při platnosti hypotézy $\overline{H}$.



Obr. 3.2

Vzhledem k tomu, že testové kritérium T je náhodná veličina, bývá obor $\overline{W}_\alpha$ ve tvaru intervalu, např. $\langle t_1; t_2 \rangle$, kde t_1, t_2 jsou kvantily statistiky T stejně jako u intervalových odhadů. Při testování statistických hypotéz se jim také říká ***kritické hodnoty***. Poznamenejme, že intervalové odhady lze přímo použít k testování statistických hypotéz. Např. při testu hypotézy $H : \vartheta = \vartheta_0$ proti alternativě $\overline{H} : \vartheta \neq \vartheta_0$ na hladině spolehlivosti α, můžeme místo testového kritéria vzít oboustranný intervalový odhad parametru ϑ se spolehlivostí $1 - \alpha$.. Jestliže tento intervalový odhad obsahuje hodnotu ϑ_0, hypotézu H nezamítáme na hladině významnosti α a naopak. Více o statistických hypotézách a jejich testech lze nalézt např. v [2], [3], [8], [15], [17], [30].

Při testování statistických hypotéz na PC pomocí statistického software se místo kritického oboru $\overline{W}_\alpha$ obvykle používá následující tzv. ***P-hodnota***. Jestliže např. testujeme hypotézu $H : \mu = \mu_0$ proti dvoustranné alternativní hypotéze $\overline{H} : \mu \neq \mu_0$, pak pro pozorovanou hodnotu t testového kritéria T je P-hodnotou je číslo $1 - P(-t \leq T \leq t)$. Výše

uvedené konvenci rozhodnutí o daných hypotézách pomocí kritického oboru, resp. oboru nezamítnutí, odpovídá následující adekvátní postup. Jestliže $P < \alpha$, pak ***zamítáme*** hypotézu H a současně ***nezamítáme*** hypotézu $\overline{H}$ na hladině významnosti α. Jestliže naopak $P \geq \alpha$, pak ***nezamítáme*** hypotézu H a současně ***zamítáme*** hypotézu $\overline{H}$ na hladině významnosti α.

Testy hypotéz o parametrech normálního rozdělení

Předpokládáme, že náhodné veličiny X a Y, resp. náhodný vektor (X, Y), mají normální rozdělení pravděpodobnosti. Předpoklad o normálním rozdělení pravděpodobnosti lze testovat pomocí testů popsaných v dalším odstavci této kapitoly. Dále uvádíme pouze testová kritéria pro dvoustranné alternativní hypotézy, např. $\overline{H} : \mu \neq \mu_0$ apod. Testy hypotéz H pro jednostranné alternativní hypotézy $\overline{H} : \mu > \mu_0$ a $\overline{H} : \mu < \mu_0$ se provádějí pomocí stejných testových kritérií a odlišují se pouze jednostrannými kritickými obory, resp. obory nezamítnutí, a odpovídajícími kritickými hodnotami - viz např. [2], [3], [8], [15], [17], [30].

Test hypotézy $H : \mu = \mu_0$ ***při neznámém rozptylu*** σ^2. Pozorovaná hodnota testového kritéria je

$$t = \frac{\overline{x} - \mu_0}{s} \sqrt{n-1}$$

a $\overline{W}_\alpha = \left\langle -t_{1-\alpha/2}; t_{1-\alpha/2} \right\rangle$, kde $t_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$-kvantil Studentova rozdělení S(k) s $k = n - 1$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T2**. Jedná se o tzv. ***t - test*** nebo ***Studentův test pro jeden výběr***.

Příklad 3.1

Měřením délky 10 válečků byly získány empirické charakteristiky $\overline{x} = 5{,}37$ mm a $s^2 = 0{,}0019$ mm^2 (viz příklad 2.1). Na hladině významnosti 0,05 testujeme hypotézu, že střední naměřená délka válečku je 5,40 mm, tedy $H : \mu = 5{,}40$.

Ř e š e n í:

Pozorovaná hodnota testového kritéria je

$$t = \frac{5{,}37 - 5{,}40}{\sqrt{0{,}0019}} \sqrt{10-1} = -2{,}0647.$$

Pro $10 - 1 = 9$ stupňů volnosti je $t_{0{,}975} = 2{,}262$ z tabulky **T2**, takže $\overline{W}_{0{,}05} = \langle -2{,}262;\ 2{,}262 \rangle$. Protože $t \in \overline{W}_{0{,}05}$, hypotézu nezamítáme. Pro testování této hypotézy bylo možno použít také intervalový odhad se spolehlivostí 0,95 z příkladu 2.1. Protože tento odhad obsahuje

hypotetickou hodnotu 5,40, nezamítáme danou hypotézu na hladině významnosti 1 – 0,95 = = 0,05.

Test hypotézy $H : \sigma^2 = \sigma_0^2$. Pozorovaná hodnota testového kritéria je

$$t = \frac{ns^2}{\sigma_0^2}$$

a $\overline{W}_\alpha = \left\langle \chi^2_{\alpha/2}; \chi^2_{1-\alpha/2} \right\rangle$, kde χ^2_P je *P*-kvantil Pearsonova rozdělení $\chi^2(k)$ s $k = n - 1$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T3**. Jedná se o tzv. ***Pearsonův test***.

Příklad 3.2

Na hladině významnosti 0,05 testujte hypotézu, že rozptyl naměřené délky válečku z příkladu 2.1 je 0,0025 mm^2, tedy $H : \sigma^2 = 0{,}0025$.

Ř e š e n í:

Pozorovaná hodnota testového kritéria je

$$t = \frac{10 \cdot 0{,}0019}{0{,}0025} = 7{,}6.$$

Pro 10 – 1 = 9 stupňů volnosti je $\chi^2_{0{,}025}$ = 2,700 a $\chi^2_{0{,}975}$ = 19,023 z tabulky **T3**, takže $\overline{W}_{0{,}05}$ = <2,700; 19,023>. Protože $t \in \overline{W}_{0{,}05}$, hypotézu nezamítáme.

Test hypotézy $H : \rho = \rho_0$. Pozorovaná hodnota testového kritéria pro $n \geq 10$, $|r| \neq 1$ a $|\rho_0| \neq 1$ je

$$t = \left(\ln \frac{1+r}{1-r} - \ln \frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{n-1} \right) \frac{\sqrt{n-3}}{2}$$

a $\overline{W}_\alpha = \left\langle -u_{1-\alpha/2}; u_{1-\alpha/2} \right\rangle$, kde $u_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$-kvantil normálního rozdělení N(0; 1), jehož hodnoty lze získat z tabulky **T1**.

Příklad 3.3

Sledováním nákladů X a ceny Y stejného výrobku u deseti výrobců byl získán dvourozměrný statistický soubor s koeficientem korelace r = 0,82482 (viz řešený příklad 2.2). Na hladině významnosti 0,01 testujte hypotézu, že veličiny X a Y jsou nekorelované (tj. vzhledem k normálnímu rozdělení nezávislé), tedy $H : \rho = 0$.

Ř e š e n í:

Pozorovaná hodnota testového kritéria je

$$t = \left(\ln \frac{1+0,82482}{1-0,82482} - \ln \frac{1+0}{1-0} - \frac{0}{10-1} \right) \frac{\sqrt{10-3}}{2} \approx 3,1001.$$

Pro danou hladinu významnosti je $u_{0,995} = 2,576$ z tabulky **T1**, takže $\overline{W}_{0,01} = <-2,576; 2,576>$. Protože $t \notin \overline{W}_{0,01}$, hypotézu zamítáme a považujeme X, Y za závislé.

Test hypotézy $H : \mu(X - Y) = 0$ ***pro dvojice*** (X, Y) za předpokladu, rozdíl $X - Y$ má normální rozdělení pravděpodobnosti. Označme pro pozorované dvojice (x_i, y_i), kde $i = 1,\ldots, n$, jejich rozdíly $d_i = x_i - y_i$ a odpovídající empirické charakteristiky $\overline{d}$ a $s^2(d)$. Pozorovaná hodnota testového kritéria je

$$t = \frac{\overline{d}}{s(d)} \sqrt{n-1}$$

a $\overline{W}_\alpha = \langle -t_{1-\alpha/2}; t_{1-\alpha/2} \rangle$, kde $t_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$-kvantil Studentova rozdělení S(k) s $k = n - 1$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T2**. Uvedený test se také nazývá ***t - test*** (***Studentův test***) ***pro párové hodnoty***.

Příklad 3.4

Měřením teploty dvěma přístroji byly během osmi dnů získány dvojice (x_i, y_i) = (51,8; 49,5), (54,9; 53,3), (52,2; 50,6), (53,3; 52,0), (51,6; 46,8), (54,1; 50,5), (54,2; 52,1), (53,3; 53,0) (^{o}C). Na hladině významnosti 1% testujte hypotézu, že střední hodnota rozdílu pozorovaných dvojic teplot rozdíl středních hodnot je nevýznamný, tedy $H : \mu(X) = \mu(Y)$.

Ř e š e n í:

Pro $d_i = x_i - y_i$, $i = 1,..., 8$, dostaneme $\overline{d} = 2,2\ ^{o}$C a $s(d) = 1,3172\ ^{o}$C. Pozorovaná hodnota testového kritéria je

$$t = \frac{2,2}{1,3172} \sqrt{8-1} \approx 4,4190.$$

Pro 8 – 1 = 7 stupňů volnosti je $t_{0,995} = 3,499$ z tabulky **T2**, takže $\overline{W}_{0,01} = <-3,499; 3,499>$. Protože $t \notin \overline{W}_{0,01}$, hypotézu zamítáme na hladině významnosti 1 % a považujeme rozdíl naměřených hodnot za statisticky významný.

U dalších testů předpokládáme, že pozorováním dvou nezávislých náhodných veličin X a Y s normálními rozděleními s parametry $\mu(X)$, $\sigma^2(X)$ a $\mu(Y)$, $\sigma^2(Y)$ byly získány realizace nezávislých náhodných výběrů s rozsahy n_1 a n_2.

Test hypotézy $H: \mu(X) - \mu(Y) = \mu_0$ ***při neznámých rozptylech*** $\sigma^2(X) = \sigma^2(Y)$.

Pozorovaná hodnota testového kritéria je

$$t = \frac{\overline{x} - \overline{y} - \mu_0}{\sqrt{n_1\ s^2(x) + n_2\ s^2(y)}}\ \sqrt{\frac{n_1\, n_2\,(n_1 + n_2 - 2)}{n_1 + n_2}}$$

a $\overline{W}_\alpha = \left\langle -t_{1-\alpha/2};\, t_{1-\alpha/2} \right\rangle$, kde $t_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$-kvantil Studentova rozdělení S(k) s $k =$ $= n_1 + n_2 - 2$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T2**. Jedná se o tzv. ***t - test*** nebo ***Studentův test pro dva výběry při stejných rozptylech***.

Příklad 3.5

Zkouškami pevnosti drátů vyrobených dvěma různými technologiemi byly získány dva statistické soubory s charakteristikami $n_1 = 33$, $\overline{x} = 5{,}4637$ kN, $s^2(x) = 0{,}3302$ kN^2, $n_2 = 28$, $\overline{y} = 6{,}1179$ kN, $s^2(y) = 0{,}4522$ kN^2. Na hladině významnosti 0,05 testujte hypotézu, že rozdílné technologie nemají vliv na střední pevnost drátu (za předpokladu stejných rozptylů $\sigma^2(X)$ a $\sigma^2(Y)$, tedy $H: \mu(X) - \mu(Y) = 0$.

Ř e š e n í:

Pozorovaná hodnota testového kritéria je

$$t = \frac{5{,}4637 - 6{,}1179 - 0}{\sqrt{33 \cdot 0{,}3302 + 28 \cdot 0{,}4522}} \sqrt{\frac{33 \cdot 28\,(33 + 28 - 2)}{33 + 28}} \approx 4{,}030.$$

Pro 33 + 28 – 2 = 59 stupňů volnosti je $t_{0{,}975} = 2{,}001$ interpolací z tabulky **T**2, takže $\overline{W}_{0{,}05} =$ $=$ <–2,001; 2,001>. Protože $t \notin \overline{W}_{0{,}05}$, hypotézu zamítáme. Rozdílné technologie mají vliv na střední pevnost drátu.

Test hypotézy $H: \mu(X) - \mu(Y) = \mu_0$ ***při neznámých rozptylech*** $\sigma^2(X) \neq \sigma^2(Y)$.

Pozorovaná hodnota testového kritéria je

$$t = \frac{\overline{x} - \overline{y} - \mu_0}{\sqrt{\frac{s^2(x)}{n_1 - 1} + \frac{s^2(y)}{n_2 - 1}}}$$

a $\overline{W}_\alpha = \left\langle -\overline{t}_{1-\alpha/2};\, \overline{t}_{1-\alpha/2} \right\rangle$, kde

$$\overline{t}_{1-\alpha/2} = \frac{\frac{s^2(x)}{n_1 - 1} t(x) + \frac{s^2(y)}{n_2 - 1} t(y)}{\frac{s^2(x)}{n_1 - 1} + \frac{s^2(y)}{n_2 - 1}}$$

a $t(x)$, resp. $t(y)$, je $\left(1-\frac{\alpha}{2}\right)$-kvantil Studentova rozdělení S(k) s $k = n_1 - 1$, resp. $n_2 - 1$, stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T2**. Jedná se o tzv. ***t - test*** nebo ***Studentův test pro dva výběry při různých rozptylech***.

Příklad 3.6

Při vyšetřování životnosti výrobků v různých systémech extrémních provozních podmínek byly získány dva statistické soubory s charakteristikami $n_1 = 21$, $\overline{x} = 3{,}581$, $s^2(x) = 0{,}114$, $n_2 = 23$, $\overline{y} = 3{,}974$, $s^2(y) = 0{,}041$ (životnost výrobků je v hodinách). Za předpokladu různých rozptylů $\sigma^2(X)$ a $\sigma^2(Y)$ testujte na hladině významnosti 0,05, že druhý systém extrémních provozních podmínek zvyšuje oproti prvnímu systému extrémních provozních podmínek střední životnost výrobku o 0,5 hod., tedy hypotézu $H : \mu(X) - \mu(Y) = -0{,}5$.

Ř e š e n í:

Pozorovaná hodnota testového kritéria je

$$t = \frac{3{,}581 - 3{,}974 - (-0{,}5)}{\sqrt{\frac{0{,}114}{21-1} + \frac{0{,}041}{23-1}}} \approx 1{,}2303.$$

Z tabulky **T2** pro $1 - \alpha/2 = 0{,}975$ je $t(x) = 2{,}086$ pro $21 - 1 = 20$ stupňů volnosti a $t(y) = 2{,}074$ pro $23 - 1 = 22$ stupňů volnosti, takže

$$\overline{t}_{0{,}975} = \frac{\frac{0{,}114}{21-1}2{,}086 + \frac{0{,}041}{23-1}2{,}074}{\frac{0{,}114}{21-1} + \frac{0{,}041}{23-1}} \approx 2{,}083.$$

a $\overline{W}_{0{,}05} = \langle -2{,}083; 2{,}083 \rangle$. Protože $t \in \overline{W}_{0{,}05}$, hypotézu o zvýšení střední životnosti o 0,5 hod. nezamítáme.

Test hypotézy $H : \sigma^2(X) = \sigma^2(Y)$. Pozorovaná hodnota testového kritéria je

$$t = \frac{\max\left(\frac{n_1 s^2(x)}{n_1 - 1}; \frac{n_2 s^2(y)}{n_2 - 1}\right)}{\min\left(\frac{n_1 s^2(x)}{n_1 - 1}; \frac{n_2 s^2(y)}{n_2 - 1}\right)},$$

kde klademe $\overline{W}_\alpha = \langle 1; F_{1-\alpha/2} \rangle$ a $F_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$-kvantil Fisherova - Snedecorova rozdělení F(k_1, k_2) se stupni volnosti $k_1 = n_1 - 1$ a $k_2 = n_2 - 1$ pro $\frac{n_1 s^2(x)}{n_1 - 1} \geq \frac{n_2 s^2(y)}{n_2 - 1}$ anebo $k_1 = n_2 - 1$

a $k_2 = n_1 - 1$ pro $\frac{n_1 s^2(x)}{n_1 - 1} \le \frac{n_2 s^2(y)}{n_2 - 1}$. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T4**.

Jedná se o tzv. ***F - test*** nebo ***Fisherův test***. Pomocí něho lze testovat předpoklady o rozptylech v obou předcházejících testech.

Příklad 3.7

Na hladině významnosti 0,05 ověřte předpoklad o různých rozptylech v řešeném příkladu 3.6, tedy že $\sigma^2(X) \neq \sigma^2(Y)$, kde $s^2(x) = 0{,}114$, $n_1 = 21$, $s^2(y) = 0{,}041$, $n_2 = 23$.

Ř e š e n í:

Testujeme naopak hypotézu $H: \sigma^2(X) = \sigma^2(Y)$. Pozorovaná hodnota testového kritéria je

$$t = \frac{\max\left(\frac{21.0{,}114}{21-1}; \frac{23.0{,}041}{23-1}\right)}{\min\left(\frac{21.0{,}114}{21-1}; \frac{23.0{,}041}{23-1}\right)} \approx \frac{\max\left(0{,}11970; 0{,}04286\right)}{\min\left(0{,}11970; 0{,}04286\right)} = \frac{0{,}11970}{0{,}04286} \approx 2{,}7928.$$

Z tabulky **T4** je pro $k_1 = 21 - 1 = 20$ a $k_2 = 23 - 1 = 22$ stupňů volnosti $F_{0{,}975} = 2{,}389$, takže $\overline{W}_{0{,}05} = \langle 1;\ 2{,}389 \rangle$. Protože $t \notin \overline{W}_{0{,}05}$, hypotézu zamítáme a předpoklad o různých rozptylech v příkladu 3.6 považujeme za správný.

Testy hypotéz o parametru binomického rozdělení

Předpokládáme, že pozorovaná náhodná veličina X má alternativní rozdělení pravděpodobnosti s parametrem p, tedy binomické rozdělení Bi(1; p). Při testování hypotézy $H: p = p_0$ jde vlastně o test hypotézy, že podíl prvků p_0 základního souboru má sledovanou vlastnost na základě zjištění, že x prvků z n náhodně vybraných prvků ze základního souboru má sledovanou vlastnost. Dále uvádíme pouze testová kritéria pro dvoustranné alternativní hypotézy, neboť testy hypotéz pro jednostranné alternativní hypotézy se odlišují pouze tím, že mají jednostranné kritické obory a odpovídající kritické hodnoty. Testy o parametru binomického rozdělení se používají často v jakosti (test podílu neshodných výrobků nebo zmetků v celkové produkci) a při průzkumu zájmu o výrobek, služby apod.

Test hypotézy $H: p = p_0$. Pozorovaná hodnota testového kritéria pro $n > 30$ je

$$t = \frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

a $\overline{W}_\alpha = \left\langle -u_{1-\alpha/2};\ u_{1-\alpha/2} \right\rangle$, kde $u_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$-kvantil normálního rozdělení N(0; 1), jehož

hodnoty lze získat z tabulky **T1**. Uvedený test je pouze přibližný, avšak jeho přesnost je pro velká *n* v praktických úlohách obvykle postačující.

Příklad 3.8

Podle expertního předpokladu bude mít zájem o nový výrobek 20 % zákazníků. Ze 400 dotázaných zákazníků projevilo zájem 62 zákazníků. Na hladině významnosti 0,05 testujme hypotézu o reálnosti předpokladu, tedy $H : p = 0{,}2$.

Ř e š e n í:

Rozsah obou výběru je dostatečně velký a pro $x = 62$ a $n = 400$ je pozorovaná hodnota testového kritéria

$$t = \frac{\frac{62}{400} - 0{,}2}{\sqrt{\frac{0{,}2(1-0{,}2)}{400}}} = \frac{-0{,}045}{0{,}02} = -2{,}25 .$$

Z tabulky **T1** je $u_{0{,}975} = 1{,}960$. Protože $t = -2{,}25 \notin \overline{W}_{0{,}05} = \langle -1{,}960;\ 1{,}960 \rangle$, hypotézu o předpokladu 20 % zájmu zamítáme na hladině významnosti 0,05. Skutečný zájem bude pravděpodobně menší. Na hladině významnosti 0,01 však hypotézu nezamítáme, neboť $u_{0{,}995} = 2{,}576$.

U dalšího testu předpokládáme, že pozorováním dvou nezávislých náhodných veličin *X*, *Y* s alternativními rozděleními s parametry p_1, p_2 byly získány realizace vzájemně nezávislých náhodných výběrů s rozsahy n_1, n_2 a počty *x*, *y* prvků se sledovanou vlastností.

Test hypotézy $H : p_1 = p_2$. Pozorovaná hodnota testového kritéria za předpokladu $n_1 > 50$ a $n_2 > 50$ je

$$t = \frac{\frac{x}{n_1} - \frac{y}{n_2}}{\sqrt{\overline{f}(1-\overline{f})}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

pro $\overline{f} = \frac{x+y}{n_1+n_2}$ a $\overline{W}_\alpha = \left\langle -u_{1-\alpha/2};\ u_{1-\alpha/2} \right\rangle$, kde $u_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$-kvantil normálního rozdělení N(0; 1), jehož hodnoty lze získat z tabulky **T1**. Uvedený test je pouze přibližný, avšak jeho přesnost je pro velké rozsahy n_1 a n_2 v praktických úlohách obvykle postačující.

Příklad 3.9

Obchodní inspekce provedla 250 kontrolních nákupů potravinářského zboží a 200 kontrolních nákupů průmyslového zboží. Zjistila přitom nedostatky u 108 nákupů potravinářského zboží a u 73 nákupů průmyslového zboží. Na hladině významnosti 0,05 testujme, zda kvalita nákupů

je stejná u obou druhů zboží, tedy hypotézu $H: p_1 = p_2$, kde p_1, p_2 jsou teoretické podíly (pravděpodobnosti) nákupů s nedostatky u daných druhů zboží.

Ř e š e n í:

Rozsahy obou výběrů jsou dostatečně velké a pro $x = 108$, $n_1 = 250$, $y = 73$, $n_2 = 200$ je

$$\overline{f} = \frac{108 + 73}{250 + 200} = 0{,}40222,$$

takže pozorovaná hodnota testového kritéria je

$$t = \frac{\frac{108}{250} - \frac{73}{200}}{\sqrt{0{,}40222(1 - 0{,}40222)}} \sqrt{\frac{250 \cdot 200}{250 + 200}} \approx \frac{0{,}067 \cdot 10{,}5409}{0{,}49035} \approx 1{,}4403 .$$

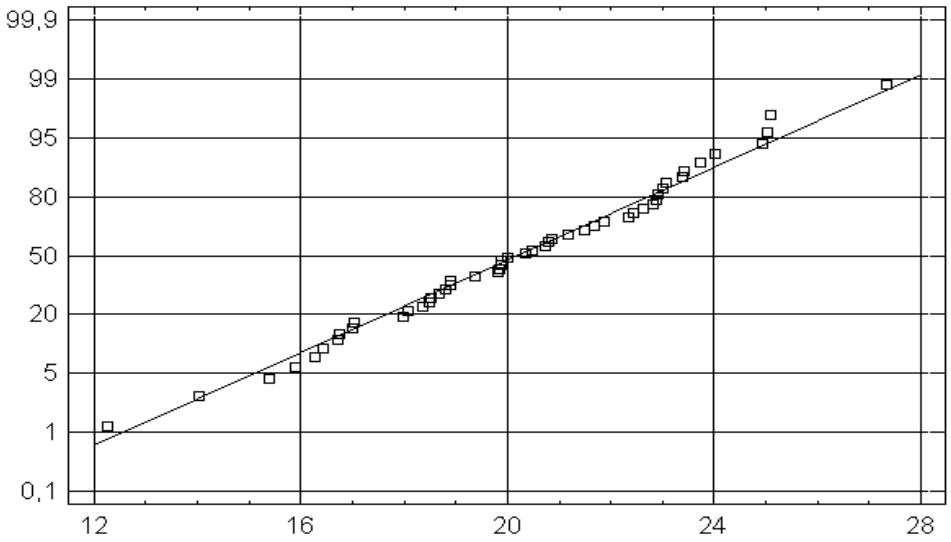
Z tabulky **T1** je $u_{0{,}975} = 1{,}960$. Protože $t = 1{,}4403 \in \overline{W}_{0{,}05} = \langle -1{,}960; 1{,}960 \rangle$, hypotézu o rovnosti podílů nákupů s nedostatky nezamítáme na hladině významnosti 0,05 a považujeme prodej obou druhů zboží za stejně nekvalitní.

Testy hypotéz o rozdělení

Vzhledem k tomu, že testy o parametrech rozdělení (a také intervalové odhady parametrů) závisejí na tvaru pozorovaných rozdělení, je zapotřebí testovat, zda pozorovaná náhodná veličina (náhodný vektor) má předpokládané rozdělení pravděpodobnosti. Nejčastěji se užívají následující ***testy hypotéz o rozdělení*** (***testy dobré shody***).

Grafická metoda **je** orientační test pomocí tzv. ***pravděpodobnostního papíru***, který obsahuje síť dvou navzájem kolmých soustav rovnoběžných přímek. Měřítko ve svislém směru (souřadná osa y) je zvoleno vzhledem k měřítku ve vodorovném směru (souřadná osa x) tak, aby grafem uvažované distribuční funkce $F(x, \vartheta)$ byla pro libovolné (v našem případě obvykle neznámé) hodnoty ϑ přímka. Na osu y se vynáší hodnoty distribuční funkce, někdy i v % a někdy jsou na této ose vyznačeny také hodnoty odpovídající střední hodnotě a celočíselným násobkům směrodatné odchylky základního souboru. Na pravděpodobnostním papíru znázorňujeme graf tzv. ***empirické distribuční funkce*** statistického souboru $(x_1, ..., x_n)$ následujícím způsobem. Uspořádáme původní statistický soubor podle velikosti, takže získáme uspořádaný soubor $(x_{(1)}, ..., x_{(n)})$, kde $x_{(i)} \le x_{(i+1)}$ pro $i = 1, ..., n$. Do souřadného systému pak vyneseme body $\left[x_{(i)}; (i - 0{,}5)/n\right]$, resp. $\left[x_{(i)}; i/(n+1)\right]$, pro $i = 1, ..., n$. Je-li statistický soubor realizací náhodného výběru ze základního souboru s rozdělením pravděpodobnosti pro daný pravděpodobnostní papír, leží výše uvedené body přibližně na přímce a naopak. V současné době se obvykle nepoužívá pravděpodobnostní papír, ale

metoda se realizuje na PC. Na obr. 3.3 je ukázka grafického výstupu z PC pro normální rozdělení pravděpodobnosti. Z grafu usuzujeme, že pozorovaná náhodná veličina má normální rozdělení pravděpodobnosti.



Obr. 3.3

Test chí-kvadrát (***Pearsonův test***) ***o rozdělení***, tj. hypotézy *H*, že pozorovaná náhodná veličina *X* má distribuční funkci *F*(*x*), proti alternativní hypotéze $\overline{H}$, že *X* nemá distribuční funkci *F*(*x*). Roztřídíme získaný statistický soubor $(x_1,...,x_n)$ do *m* tříd s četnostmi f_j a vypočteme teoretické absolutní četnosti $\hat{f}_j$, $j = 1,...,m$, resp. jejich odhady, pro hypotetické rozdělení. Statistický soubor roztřídíme tak, aby ve všech třídách byly dostatečně velké teoretické absolutní četnosti - obvykle požadujeme, aby $\hat{f}_j > 5$. Toho lze při dostatečně velkém rozsahu *n* dosáhnout vhodnou volbou tříd nebo sloučením již získaných sousedních tříd. Pozorovaná hodnota testového kritéria je

$$t = \sum_{j=1}^{m} \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j} = \left(\sum_{j=1}^{m} \frac{f_j^2}{\hat{f}_j} \right) - n$$

a $\overline{W}_\alpha = \left\langle 0 ; \chi^2_{1-\alpha} \right\rangle$, kde $\chi^2_{1-\alpha}$ je (1 − α)-kvantil Pearsonova rozdělení $\chi^2(k)$ s $k = m - q - 1$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T3**. Číslo *q* je počet parametrů hypotetického rozdělení náhodné veličiny *X*, které jsme nuceni odhadnout z roztříděného statistického souboru pro určení hodnot distribuční funkce *F*(*x*). Uvedený test je asymptotický (tj. vhodný pro dostatečně velké rozsahy výběru *n*, řádově aspoň desítky) a zjednodušenou, ale obvykle používanou variantou přesného testu chí-kvadrát, který se realizuje pomocí statistického softwaru na PC. Více o tomto a dalších testech dobré shody v [2], [3], [8], [15], [17], [30].

Příklad 3.10

Bylo provedeno 120 hodů se šestistěnnou hrací kostkou se stěnami očíslovanými od 1 do 6. Získané výsledky jsou v následující tabulce:

x_j^*	1	2	3	4	5	6
f_j	11	18	15	21	24	31

Na hladině významnosti 0,05 testujte hypotézu, že kostka není falešná, tj. pravděpodobnosti padnutí každého ze všech 6 čísel jsou stejné.

Ř e š e n í:

Testujeme hypotézu *H*, že pozorovaná náhodná veličina *X* má tzv. klasické (uniformní) rozdělení pravděpodobnosti s pravděpodobnostní funkcí $p(x) = \frac{1}{6}$ pro $x = 1,\ldots,6$. V našem případě je $x_j^* = x$ a $\hat{f}_j = np\left(x_j^*\right) = 120 \cdot \frac{1}{6} = 20$ pro $j = 1,\ldots,6$. Další potřebné výpočty jsou v tabulce:

j	f_j	$\hat{f}_j$	$\frac{(f_j - \hat{f}_j)^2}{\hat{f}_j}$
1	11	20	4,05
2	18	20	0,20
3	15	20	1,25
4	21	20	0,05
5	24	20	0,80
6	31	20	6,05
$\sum$	120	120	12,40

Podmínka $\hat{f}_j > 5$ je pro všechna *j* splněna a hodnota testového kritéria je $t = 12,40$. Neodhadujeme žádný parametr rozdělení pravděpodobnosti, takže $q = 0$ a počet stupňů volnosti je $k = 6 - 0 - 1 = 5$. Z tabulky **T3** je pro hladinu významnosti 0,05 a daný počet stupňů volnosti kvantil $\chi^2_{0,95} = 11,070$. Protože $t = 12,40 \notin \overline{W}_{0,05} = \langle 0;11,070 \rangle$, zamítáme na hladině významnosti 0,05 hypotézu, že kostka není falešná. Na hladině významnosti 0,01 ale tuto hypotézu nezamítáme, neboť $\chi^2_{0,99} = 15,086$. Oba zdánlivě protichůdné závěry můžeme také získat z *P*–hodnoty 0,02969946, kterou vypočteme např. pomocí statistické funkce CHIDIST v Excelu.

Neparametrické testy hypotéz

Neparametrické testy statistických hypotéz se používají v případech, kdy neznáme rozdělení pozorované náhodné veličiny X, resp. náhodného vektoru (X, Y), anebo pro známé rozdělení nemáme potřebná testová kritéria. Omezením neparametrických metod je obvykle požadavek, že pozorované náhodné veličiny mají spojitá rozdělení, avšak v některých případech stačí znát pouze pořadí uspořádaných hodnot daného statistického souboru, tj. hodnoty odpovídajícího ordinálního statistického znaku. Slabší předpoklady o rozdělení (na rozdíl od parametrických testů - viz např. výše uvedené testy parametrů normálního a binomického rozdělení) mají za následek, že neparametrické metody nejsou tak silné, jako jejich parametrické protějšky. Základním principem neparametrických testů je nahrazení původních pozorovaných hodnot jejich pořadími co do velikosti a proto se také v literatuře hovoří o ***pořadových testech***.

Jestliže pozorovaný statistický soubor $(x_1, ..., x_n)$ sestává pouze z navzájem různých reálných čísel, pak ***pořadím*** R_i prvku x_i, $i = 1, ..., n$, rozumíme počet prvků z daného statistického souboru, jejichž hodnota je menší nebo rovna x_i. Nahrazením prvku x_i jeho pořadím R_i tak získáme soubor pořadí $(R_1, ..., R_n)$. Např. statistickému souboru

$$(x_1, ..., x_7) = (5; 8; -2; -3; 0; 2; 1)$$

odpovídá uspořádaný statistický soubor

$$(x_{(1)}, ..., x_{(7)}) = (-3; -2; 0; 1; 2; 5; 8),$$

takže soubor pořadí je

$$(R_1, ..., R_7) = (6; 7; 2; 1; 3; 5; 4).$$

Jestliže nejsou všechna čísla x_i navzájem různá, pak všem stejným číslům x_i přiřadíme aritmetický průměr takových pořadí, jakoby následovala těsně za sebou. Např. ve statistickém souboru

$$(x_1, ..., x_7) = (5; 8; -2; -3; 0; 2; 0)$$

je číslo 0 dvakrát, takže soubor pořadí je

$$(R_1, ..., R_7) = (6; 7; 2; 1; 3,5; 5; 3,5).$$

I v případě shodných prvků je součet všech pořadí $\sum_{i=1}^{n} R_i = \frac{n(n+1)}{2}$.

Při neparametrických testech pracujeme s testovými kritérii (statistikami), které nabývají diskrétních hodnot. Jde proto o testy s hladinou významnosti nejvýše rovnu α. Je

proto na rozdíl od běžné definice kvantilu vhodné definovat jejich kritické hodnoty pro nezamítnutí anebo zamítnutí hypotéz tak, že *P*-kvantilem (kritickou hodnotou) daného diskrétního rozdělení je takové maximální číslo t_P, pro které je pravděpodobnost náhodného jevu $T \le t_P$ menší nebo rovna číslu *P*. V našem případě jde o dále používané binomické, Wilcoxonovo a Mannovo-Whitneyovo rozdělení (tabulka **T5** a **T6**). Poznamenejme ještě, že níže použitá asymptotická testová kritéria mají normované normální rozdělení, které je spojité, takže naše definice *P*-kvantilu dává tytéž hodnoty jako definice běžně používaná.

Znaménkový test $H : x_{0,5} = c$. Testujeme hypotézu, že medián $x_{0.5}$ spojité náhodné veličiny *X* je roven *c*. Jde o neparametrickou verzi odpovídající Studentovu testu střední hodnoty normálního rozdělení, které je symetrické a proto má střední hodnotu rovnu mediánu. Označme *y* počet kladných rozdílů $x_i - c$. Případy $x_i = c$ vynecháváme. Jestliže hypotéza *H* platí, pak má náhodná veličina *Y* nabývající hodnot *y* binomické rozdělení Bi(*n*;0,5). Číslo *y* je přímo pozorovaná hodnota testového kritéria Y a obory nezamítnutí hypotézy *H* jsou:

a) $\overline{W}_\alpha = \langle k_{\alpha/2} + 1, n - (k_{\alpha/2} + 1) \rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} \ne c$,

b) $\overline{W}_\alpha = \langle k_\alpha + 1, n \rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} < c$,

c) $\overline{W}_\alpha = \langle 0, n - (k_\alpha + 1) \rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} > c$,

kde k_P je *P*-kvantil uvedeného binomického rozdělení, tj. je maximální číslo splňující nerovnost $2^{-n} \sum_{k=0}^{k_P} \binom{n}{k} \le P$. Hodnoty k_P jsou pro $\alpha = 0,05$ a $\alpha = 0,01$ tabelovány a je možno je také vypočítat pomocí statistické funkce BINDIST v Excelu anebo „ručně“. Pro $n \ge 20$ můžeme použít asymptotickou verzi testu s testovým kritériem

$$u = \frac{2y - n}{\sqrt{n}}$$

a obory nezamítnutí hypotézy *H* jsou

a) $\overline{W}_\alpha = \langle -u_{1-\alpha/2}, u_{1-\alpha/2} \rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} \ne c$,

b) $\overline{W}_\alpha = \langle -u_{1-\alpha}, \infty)$ pro alternativní hypotézu $\bar{H} : x_{0,5} < c$,

c) $\overline{W}_\alpha = (-\infty, u_{1-\alpha} \rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} > c$,

kde u_P je *P*-kvantil normovaného normálního rozdělení N(0;1) – viz tabulku **T1**.

Znaménkový test se často používá pro tzv. ***párové hodnoty*** (X_1, X_2), kdy testujeme hypotézu, že medián rozdílu $X = X_1 - X_2$ je roven hodnotě *c* (nejčastěji pro $c = 0$). Existuje

také obecnější varianta znaménkového testu (tzv. ***kvantilový test***), když testujeme hypotézu $H : x_q = c$, kde x_q je q-kvantil pozorované náhodné veličiny X.

Příklad 3.11

Při přípravě nové písemné práce pro zkoušku ze statistiky chceme ověřit správnost předpokladu, že medián získaných bodů je roven 60. Vyskytly se námitky, že písemná práce je těžká a počty získaných bodů jsou převážně nižší než 60. K ověření bylo náhodně vybráno 25 výsledků z minulé zkoušky a v nich byla zjištěna tato bodová hodnocení: 62; 61; 27; 84; 50; 90; 49; 32; 48; 43; 55; 54; 53; 34; 68; 80; 39; 56; 52; 91; 45; 47; 78; 46; 74. Pro test hypotézy zvolme hladinu významnosti 0,05.

Ř e š e n í:

Znaménkovým testem testujeme nulovou hypotézu $H : x_{0,5} = 60$ proti alternativní hypotéze $\overline{H} : x_{0,5} < 60$. Přípravný výpočet je v tabulce:

i	x_i	$x_i - 60$	Znaménko	i	x_i	$x_i - 60$	Znaménko
1	62	2	+	14	34	-26	-
2	61	1	+	15	68	8	+
3	27	-33	-	16	80	20	+
4	84	24	+	17	39	-21	-
5	50	-10	-	18	56	-4	-
6	90	30	+	19	52	-8	-
7	49	-11	-	20	91	31	+
8	32	-28	-	21	45	-15	-
9	48	-12	-	22	47	-13	-
10	43	-17	-	23	78	18	+
11	55	-5	-	24	46	-14	-
12	54	-6	-	25	74	14	+
13	53	-7	-				

Z tabulky získáme počet kladných znamének $y = 9$. Postupným součtem zjistíme, že maximální číslo $k_{0.05}$ splňující nerovnost $2^{-25}\sum_{k=0}^{k_{0,05}}\binom{25}{k} \le 0,05$, je $k_{0,05} = 7$. Např. pomocí funkce BINOMDIST v Excelu snadno ověříme, že pro horní mez sumace 7 je levá strana nerovnosti rovna 0,021642625 a pro 8 je 0,053876072. Kvantil $k_{0,05} = 7$ můžeme také najít v tabulce **T7**. Protože $y = 9 \in \overline{W}_{0,05} = \langle 8; 25 \rangle$, nezamítáme na hladině významnosti 0,05 hypotézu $H : x_{0,5} = 60$ proti alternativní hypotéze $\overline{H} : x_{0,5} < 60$ a zamítáme námitku, že

statisticky významně převažují písemné práce s hodnocením menším než 60 bodů. Protože rozsah souboru je 25, můžeme použít také asymptotický test. Dostaneme tentýž závěr, neboť $u = \frac{2 \cdot 9 - 25}{\sqrt{25}} = -1,4 \in \overline{W}_{0,05} = \langle -1,645; \infty)$, kde kvantil $u_{0,95} = 1,645$ získáme z tabulky **T1**.

K přesnějšímu závěru pomocí obou testových kritérií bychom dospěli zvýšením rozsahu výběru, neboť tak bychom zvětšili sílu testu, tj. snížili pravděpodobnost chyby druhého druhu (nezamítnutí neplatné nulové hypotézy).

Wilcoxonův jednovýběrový test $H : x_{0,5} = c$. Testujeme hypotézu, že medián $x_{0,5}$ spojité náhodné veličiny *X*, která má symetrické rozdělení vzhledem k mediánu, je roven *c*. Jde opět o neparametrickou verzi odpovídající Studentovu testu střední hodnoty normálního rozdělení. Předpokládáme, že je $x_i \neq c$ pro všechna $i = 1, ..., n$. Případy $x_i = c$ vynecháváme. Vytvořme rozdíly $x_i - c$ a jejich absolutní hodnoty $|x_i - c|$. Nechť R_i^+ značí pořadí hodnot $|x_i - c|$, kde respektujeme případné shody pořadí. Označme dále součty pořadí $S^+ = \sum_{x_i - c > 0} R_i^+$ a $S^- = \sum_{x_i - c < 0} R_i^+$. Platí, že $S^+ + S^- = n(n+1)/2$. Hypotézu $H : x_{0,5} = c$ nezamítáme, jestliže:

a) $S^+ \in \overline{W}_\alpha = \left\langle w_{\alpha/2} + 1, \frac{n(n+1)}{2} - (w_{\alpha/2} + 1) \right\rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} \neq c$,

b) $S^+ \in \overline{W}_\alpha = \left\langle w_\alpha + 1, \frac{n(n+1)}{2} \right\rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} < c$,

c) $S^+ \in \overline{W}_\alpha = \left\langle 0, \frac{n(n+1)}{2} - (w_\alpha + 1) \right\rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} > c$,

kde w_P je *P*-kvantil Wilcoxonova rozdělení, které je tabelováno – viz tabulku **T5**. Pro velká *n* můžeme také použít asymptotickou verzi testu s testovým kritériem

$$u = \frac{S^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

a obory nezamítnutí hypotézy *H* jsou

a) $\overline{W}_\alpha = \langle -u_{1-\alpha/2}, u_{1-\alpha/2} \rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} \neq c$,

b) $\overline{W}_\alpha = \langle -u_{1-\alpha}, \infty)$ pro alternativní hypotézu $\bar{H} : x_{0,5} < c$,

c) $\overline{W}_\alpha = (-\infty, u_{1-\alpha} \rangle$ pro alternativní hypotézu $\bar{H} : x_{0,5} > c$,

kde u_P je *P*-kvantil normovaného normálního rozdělení N(0;1) – viz tabulku **T1**.

Wilcoxonův jednovýběrový test a také znaménkový test se často používá pro tzv. ***párové hodnoty*** (X_1, X_2), kdy testujeme hypotézu, že medián rozdílu $X = X_1 - X_2$ je roven hodnotě *c* (nejčastěji pro $c = 0$).

Příklad 3.12

Na hladině významnosti 0,05 testujte pomocí Wilcoxonova jednovýběrového testu hypotézu $H : x_{0,5} = 60$ proti alternativní hypotézu $\overline{H} : x_{0,5} < 60$ pro data z příkladu 3.11.

Ř e š e n í:

Přípravný výpočet je v tabulce:

i	x_i	$x_i - 60$	$\lvert x_i - 60 \rvert$	R_i^+	R_i^+ pro $x_i - 60 > 0$	R_i^+ pro $x_i - 60 < 0$
1	62	2	2	2	2	
2	61	1	1	1	1	
3	27	-33	33	25		25
4	84	24	24	20	20	
5	50	-10	10	9		9
6	90	30	30	23	23	
7	49	-11	11	10		10
8	32	-28	28	22		22
9	48	-12	12	11		11
10	43	-17	17	16		16
11	55	-5	5	4		4
12	54	-6	6	5		5
13	53	-7	7	6		6
14	34	-26	26	21		21
15	68	8	8	7,5	7,5	
16	80	20	20	18	18	
17	39	-21	21	19		19
18	56	-4	4	3		3
19	52	-8	8	7,5		7,5
20	91	31	31	24	24	
21	45	-15	15	15		15
22	47	-13	13	12		12
23	78	18	18	17	17	
24	46	-14	14	13,5		13,5
25	74	14	14	13,5	13,5	
Σ	---	---	---	325	126	199

Z tabulky je $S^+ = 126$ a $S^- = 199$. Protože $S^+ = 126 \in \overline{W}_{0,05} = \langle 101; 325 \rangle$, kde pro $n = 25$ je

kvantil $w_{0,05} = 100$ z tabulky **T5**, nezamítáme hypotézu $H : x_{0,5} = 60$. Naopak zamítáme hypotézu, že převažují písemné práce s nižším bodovým hodnocením. Součet S^- jsme nemuseli počítat, ale při „ručním“ výpočtu a malém počtu hodnot R_i^+ pro $x_i - 60 < 0$ je někdy vhodné využít toho, že $S^+ + S^- = n(n+1)/2$. Protože rozsah souboru je dostatečně velký, můžeme také aplikovat asymptotický test dané hypotézy. Dostaneme tentýž závěr, neboť $u = \dfrac{126 - \dfrac{25 \cdot 26}{4}}{\sqrt{\dfrac{25 \cdot 26 \cdot 51}{24}}} \doteq -0,98210 \in \overline{W}_{0,05} = \langle -1,645; \infty)$, přičemž kvantil $u_{0,95} = 1,645$ je z tabulky **T1**.

Wilcoxonův dvouvýběrový test a ***Mannův-Whitneyův test***. Předpokládáme, že jsme pozorováním náhodné veličiny X se spojitým rozdělením s distribuční funkcí F získali statistický soubor $(x_1, ..., x_m)$ a pozorováním náhodné veličiny Y se spojitým rozdělením s distribuční funkcí G statistický soubor $(y_1, ..., y_n)$. Testujeme hypotézu $H : F = G$, tj. X a Y mají stejné rozdělení pravděpodobnosti, proti alternativní hypotéze $\overline{H} : F \neq G$, tj. X a Y nemají stejné rozdělení pravděpodobnosti. Sloučíme oba statistické soubory do jednoho statistického souboru o rozsahu $m + n$, uspořádáme tento soubor vzestupně podle velikosti a určíme pořadí všech $m + n$ hodnot. Označme T_1 součet všech pořadí odpovídajících statistickému souboru $(x_1, ..., x_m)$ a T_2 součet všech pořadí odpovídajících statistickému souboru $(y_1, ..., y_n)$. Zřejmě je $T_1 + T_2 = (m+n)(m+n+1)/2$. Statistika T_1 je testovým kritériem ***Wilcoxonova dvouvýběrového testu*** a její kritické hodnoty jsou tabelovány, ale v současné době se pro testování převážně používá ekvivalentní varianta nazývaná ***Mannův-Whitneyův test***. Pro tento test vypočteme hodnotu statistiky

$$U_1 = mn + \frac{m(m+1)}{2} - T_1$$

a hypotézu $H : F = G$ nezamítáme, jestliže $U_1 \in \overline{W}_\alpha = \langle v_{\alpha/2} + 1, mn - (v_{\alpha/2} + 1) \rangle$, kde $v_{\alpha/2}$ je $(\alpha/2)$-kvantil Mannovy-Whitneyovy statistiky – viz tabulku **T6**. Hodnotu statistiky U_1 můžeme také určit bez sloučení původních statistických souborů a výpočtu součtu pořadí T_1 přímo ze vztahu

$$U_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} h_{ij} \,,$$

kde klademe $h_{ij} = 1$ pro $x_i \le y_j$ a $h_{ij} = 0$ pro $x_i > y_j$. Jestliže $m > 10$ a $n > 10$, můžeme také použít asymptotickou verzi testu s testovým kritériem

$$u = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}.$$

Oborem nezamítnutí hypotézy *H* je pak $\overline{W}_\alpha = \langle -u_{1-\alpha/2}, u_{1-\alpha/2} \rangle$, kde $u_{1-\alpha/2}$ je $\left(1 - \frac{\alpha}{2}\right)$-kvantil normovaného normálního rozdělení N(0;1) – viz tabulku **T1**. Poznamenejme, že v Mannovu-Whitneyovu testu můžeme také použít místo U_1 druhou statistiku $U_2 = mn + \frac{n(n+1)}{2} - T_2$.

Příklad 3.13

Byly vybrány dvě skupiny $m = 13$ a $n = 12$ firem, které vyrábějí tytéž výrobky. Firmy v první skupině nevyužívají statistické metody řízení jakosti, naopak firmy ve druhé skupině tyto metody využívají. U obou skupin byl zjištěn zisk v Kč získaný prodejem jednoho výrobku. Na hladině významnosti 0,05 posuďte, zda aplikace metod řízení jakosti má statisticky významný vliv na zisk u daného výrobku. Získané hodnoty jsou tabulce, kde x_i je zisk *i*-té firmy z první skupiny a y_j je zisk *j*-té firmy ze druhé skupiny:

i	x_i	*j*	y_j
1	66,7	1	67,7
2	57,7	2	67,2
3	58,8	3	69,3
4	66,1	4	65,8
5	57,1	5	61,6
6	62,2	6	67,3
7	64,6	7	65,3
8	58,4	8	68,8
9	59,6	9	64,1
10	60,5	10	61,3
11	61,8	11	67,1
12	59,2	12	63,3
13	66,9		

Ř e š e n í:

Pomocí Mannova-Whitneyova testu testujeme hypotézu, že náhodná veličina X (zisk firmy z první skupiny) s neznámou distribuční funkcí F má stejné rozdělení jako náhodná veličina Y (zisk firmy ze druhé skupiny) s neznámou distribuční funkcí G, tedy $H : F = G$ proti alternativní hypotéze $\overline{H} : F \neq G$. Sloučíme oba soubory do jednoho souboru s rozsahem $m + n = 13 + 12 = 25$ a uspořádáme jej vzestupně podle velikosti. Další výpočty jsou v následující tabulce, kde podtržená čísla odpovídají druhému souboru, tj. Y:

k	Sloučený soubor	Uspořádaný sloučený soubor	Hodnoty prvního souboru	Hodnoty druhého souboru	Pořadí pro první soubor	Pořadí pro druhý soubor
1	66,7	57,1	57,1		1	
2	57,7	57,7	57,7		2	
3	58,8	58,4	58,4		3	
4	66,1	58,8	58,8		4	
5	57,1	59,2	59,2		5	
6	62,2	59,6	59,6		6	
7	64,6	60,5	60,5		7	
8	58,4	61,3		61,3		8
9	59,6	61,6		61,6		9
10	60,5	61,8	61,8		10	
11	61,8	62,2	62,2		11	
12	59,2	63,3		63,3		12
13	66,9	64,1		64,1		13
14	67,7	64,6	64,6		14	
15	67,2	65,3		65,3		15
16	69,3	65,8		65,8		16
17	65,8	66,1	66,1		17	
18	61,6	66,7	66,7		18	
19	67,3	66,9	66,9		19	
20	65,3	67,1		67,1		20
21	68,8	67,2		67,2		21
22	64,1	67,3		67,3		22
23	61,3	67,7		67,7		23
24	67,1	68,8		68,8		24
25	63,3	69,3		69,3		25
Σ	---	---	---	---	117	208

Z tabulky vidíme, že $T_1 = 117$. Odtud $U_1 = 13 \cdot 12 + \frac{13 \cdot 14}{2} - 117 = 130$ a z tabulky **T6** je $v_{0,025} = 41$. Protože $U_1 = 130 \notin \overline{W}_{0,05} = \langle 41 + 1; 156 - (41 + 1) \rangle = \langle 42; 114 \rangle$, zamítáme na

hladině významnosti 0,05 hypotézu $H : F = G$. Aplikace statistických metod řízení jakosti má patrně vliv na výši zisku a po jejich nasazení můžeme očekávat jeho vyšší úroveň, samozřejmě pokud se nevyskytují ve firmách ze druhé skupiny další faktory, které zisk pozitivně ovlivňují. Vzhledem k dostatečně velkým rozsahům obou souborů můžeme také použít asymptotický test. Pak je $u = \dfrac{130 - \dfrac{13 \cdot 12}{2}}{\sqrt{\dfrac{13 \cdot 12 \cdot 26}{12}}} \doteq 2,828$ a z tabulky **T1** $u_{0,975} = 1,960$.

Hypotézu $H : F = G$ opět zamítáme, protože $u = 2,828 \notin \overline{W}_{0,05} = \langle -1,960; 1,960 \rangle$.

Wilcoxonův dvouvýběrový test a také Mannův-Whitneyův test vychází z porovnání mediánů dvou nezávislých pozorovaných náhodných veličin a oba testy jsou neparametrickou obdobou Studentova dvouvýběrového testu rovnosti středních hodnot těchto veličin, kdy ale předpokládáme, že obě mají normální rozdělení. V aplikacích se úspěšně používá řada dalších neparametrických testů – viz např. [2], [3], [4], [6], [10], [22], [28].

Příklady k procvičení

Příklad 3.14

Statistický soubor o rozsahu n = 10 má aritmetický průměr $\bar{x}$ = 32 a rozptyl s^2 = 15. Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota pozorované náhodné veličiny s normálním rozdělením je μ = 30.

V ý s l e d e k: t = 1,549; $t_{0,975}$ = 2,262; hypotézu nezamítáme

Příklad 3.15

Realizací náhodného výběru z normálního rozdělení byl po roztřídění získán statistický soubor:

x_j^*	-2	-1	0	1	2	3
f_j	1	4	7	3	3	2

Na hladině významnosti 0,05 testujte hypotézu, že μ = 0,1.

V ý s l e d e k: $\bar{x}$ = 0,45; s = 1,3592; t = 1,1224; $t_{0,975}$ = 2,093; hypotézu nezamítáme

Příklad 3.16

Požadovaná střední hodnota vlhkosti v pražené kávě je 4,2 % a směrodatná odchylka 0,4 %. Ve 20 vzorcích byly analýzou zjištěny tyto skutečné hodnoty vlhkosti v %: 4,5; 4,3; 4,1; 4,9; 4,6; 3,2; 4,4; 5,1; 4,8; 4,0; 3,7; 4,4; 3,9; 4,1; 4,2; 4,1; 4,7; 4,3; 4,2; 4,4. Na hladině

významnosti 5% testujte hypotézy, že základní soubor s normálním rozdělením, z něhož vzorky pocházejí, má (a) požadovanou střední hodnotu vlhkosti a (b) variabilitu.

V ý s l e d e k: (a) $t = 1{,}033$; $t_{0,975} = 2{,}093$; hypotézu nezamítáme

(b) $t = 22{,}25$; $\chi^2_{0,025} = 8{,}907$; $\chi^2_{0,975} = 32{,}852$; hypotézu nezamítáme

Příklad 3.17

Pomocí statistického souboru o rozsahu 10 a s rozptylem $s^2 = 2{,}0$ testujte na hladině významnosti 0,01 hypotézu, že základní soubor s normálním rozdělením má rozptyl $\sigma^2 = 0{,}2$.

V ý s l e d e k: $t = 100$; $\chi^2_{0,005} = 1{,}735$; $\chi^2_{0,995} = 23{,}589$; hypotézu zamítáme

Příklad 3.18

Pro posouzení přesnosti dvou měřících metod bylo provedeno 8 měření a byly určeny rozdíly dvojic (odchylky) odpovídajících si výsledků. Odtud pak byla určena průměrná odchylka $\overline{d} = 0{,}244$ a směrodatná odchylka $s(d) = 0{,}192$. Zjistěte na hladině významnosti 0,05, zda obě metody můžeme považovat za stejně přesné, jestliže rozdíly mají normální rozdělení.

V ý s l e d e k: $t = 3{,}362$; $t_{0,975} = 2{,}365$; hypotézu zamítáme

Příklad 3.19

Na dvou váhách bylo provedeno vážení 10 vzorků s výsledky (x_i, y_i) = (25; 28), (30; 31), (28; 26), (50; 52), (20; 24), (40; 36), (32; 33), (36; 35), (42; 45), (38; 40) (g). Na hladině významnosti 0,01 zjistěte, zda rozdílné výsledky jsou statisticky nevýznamné za předpokladu, že rozdíly získaných dvojic hodnot mají normální rozdělení.

V ý s l e d e k: $t = -1{,}13$; $t_{0,995} = 3{,}250$; hypotézu nezamítáme, tedy rozdíly jsou statisticky nevýznamné

Příklad 3.20

Před seřízením a po seřízení váhy na balícím automatu byly získány statistické soubory s charakteristikami $n_1 = 12$, $\overline{x} = 31{,}2$ g, $s^2(x) = 0{,}770$ g^2 a $n_2 = 18$, $\overline{y} = 29{,}2$ g, $s^2(y) =$ $= 0{,}378$ g^2. Za předpokladu stejných rozptylů a normálního rozdělení testujte na hladině významnosti 0,05 hypotézu, že se střední hodnota nastavení váhy seřízením nezměnila.

V ý s l e d e k: $t = 7{,}1$; $t_{0,975} = 2{,}048$; hypotézu zamítáme

Příklad 3.21

Studijní průměry 20 studijních skupin daného ročníku jsou:

x_j^*	1,70	1,86	2,01	2,23	2,27	2,411
f_j	2	3	5	7	2	1

Celkový studijní průměr v minulém ročníku byl $\overline{y}$ = 2,201 a rozptyl $s^2(y)$ = 0,012 pro 20 studijních skupin. Testujte hypotézu, že se střední hodnoty studijních výsledků mezi oběma ročníky neliší, předpokládáme-li normální rozdělení studijních průměrů se stejnými rozptyly.

V ý s l e d e k: $\overline{x}$ = 2,0795; $s^2(x)$ = 0,0399; t = - 2,325;

$t_{0,975}$ = 2,023 (lineární interpolací); hypotézu zamítáme;

$t_{0,995}$ = 2,712 (lineární interpolací); hypotézu nezamítáme

Příklad 3.22

Bylo provedeno po 18 zkouškách pevnosti v tahu na vzorcích dvou druhů lan s výsledky: $\overline{x}$ = = 3389,3 N, $s^2(x)$ = 1144,4 N^2, $\overline{y}$ = 3339,2 N, $s^2(y)$ = 3453,5 N^2. Za předpokladu různých rozptylů pevností v tahu s normálním rozdělením testujte na hladině významnosti 0,05 hypotézu, že střední pevnosti v tahu obou druhů lan jsou stejné.

V ý s l e d e k: t = 3,046; $\overline{t}_{0,975}$ = 2,110; hypotézu zamítáme

Příklad 3.23

Dva statistické soubory s rozsahy n_1 = 20 a n_2 = 10 a charakteristikami $\overline{x}$ = 10,24; $\overline{y}$ = 11,09; $s^2(x)$ = 4,231 a $s^2(y)$ = 18,457 byly získány nezávislými náhodnými výběry z nezávislých normálních rozdělení s různými rozptyly. Testujte na hladině významnosti 1% hypotézu, že uvedená rozdělení mají stejné střední hodnoty.

V ý s l e d e k: t = -0,5637; $\overline{t}_{0,975}$ = 3,212; hypotézu nezamítáme

Příklad 3.24

Při určování tuku v mléce byly použity dvě různé metody. První metoda při provedení 12 analýz dala rozptyl naměřených hodnot $s^2(x)$ = 0,0224 a druhá metoda dala rozptyl při provedení 8 analýz $s^2(y)$ = 0,0263. Testujte na hladině významnosti 0,01 hypotézu, že obě metody jsou vzhledem k rozptylu stejně přesné, jestliže mají naměřené hodnoty normální rozdělení.

V ý s l e d e k: t = 1,23; $F_{0,975}$ = 3,759; hypotézu nezamítáme

Příklad 3.25

Testujte předpoklad o stejných rozptylech základních souborů z příkladu 3.20 na hladině významnosti 0,05.

V ý s l e d e k: t = 2,1; $F_{0,975}$ = 2,87 (lineární interpolací); hypotézu nezamítáme

Příklad 3.26

Představenstvo velké akciové společnosti zvažuje prodej akcií svým zaměstnancům a odhaduje, že asi 20 % z nich si je zakoupí. Při průzkumu u náhodně vybraných 400 zaměstnanců projevilo zájem o akcie 66 zaměstnanců. Testujte na hladině významnosti 0,05, zda předpoklad představenstva je reálný.

V ý s l e d e k: $t = -1{,}75$; $u_{0,975} = 1{,}960$; hypotézu nezamítáme, předpoklad je reálný

Příklad 3.27

Z 200 výrobků vyrobených novou technologií bylo 31 zmetků. Ověřte, že na hladině významnosti 1 % nová technologie změnila zmetkovitost oproti dřívějším dlouhodobě zjištěným 10 % zmetkovitosti.

V ý s l e d e k: $t = 2{,}593$; $u_{0,995} = 2{,}576$; hypotézu zamítáme, nová technologie změnila zmetkovitost

Příklad 3.28

Ve dvou závodech vyrábějí tentýž výrobek. Podíl vadných výrobků v obou závodech by měl být stejný, protože používají týchž technologií výroby. V prvním závodě bylo 10 vadných výrobků mezi 200 kontrolovanými a ve druhém závodě bylo 23 vadných výrobků mezi 250 kontrolovanými. Na hladině významnosti 0,01 ověřte, zda mezi oběma závody je statisticky významný rozdíl v jakosti výroby těchto výrobků.

V ý s l e d e k: $t = -1{,}699$; $u_{0,995} = 2{,}576$; hypotézu nezamítáme, mezi závody není statisticky významný rozdíl v jakosti výroby

Příklad 3.29

Mezi 58 zemědělci z jisté lokality bylo zjištěno 23 nemocných a mezi 43 dělníky z téže lokality 28 nemocných. Testujte na hladině významnosti 5 % hypotézu, že u dělníků je stejná nemocnost jako u zemědělců.

V ý s l e d e k: $t = -2{,}534$; $u_{0,975} = 1{,}960$; hypotézu zamítáme, výskyt onemocnění je u dělníků spíše větší než u zemědělců

Příklad 3.30

Deset osob mělo nezávisle na sobě bez předchozího nácviku odhadnout, kdy od daného signálu uplyne jedna minuta. Byly získány výsledky v sekundách: 53, 48, 45, 55, 63, 51, 66, 56, 50, 58. Testujte na hladině významnosti 0,05 znaménkovým testem hypotézu, že polovina lidské populace délku jedné minuty podhodnotí a polovina ji nadhodnotí, proti hypotéze, že je to jinak.

V ý s l e d e k: $y = 2 \in \overline{W}_{0,05} = \langle 2;8 \rangle$; hypotézu nezamítáme

Příklad 3.31

Pomocí náhodného výběru 16 firem ověřte domněnku, že burzovní experti systematicky podhodnocují odhady cen akcií na burze. Odhady expertů a skutečně dosažené ceny jsou v tabulce:

Firma	1	2	3	4	5	6	7	8
Odhad x_{1i}	123	764	905	3200	1356	724	254	2255
Cena x_{2i}	113	680	901	3310	1280	733	330	2358
Firma	9	10	11	12	13	14	15	16
Odhad x_{1i}	55	173	894	2784	142	423	674	3556
Cena x_{2i}	57	185	866	2890	153	431	688	3560

Zvolte hladinu významnosti α = 0,05. (Návod: Použijte párový Wilcoxonův a znaménkový test hypotézy, že medián rozdílu $X = X_1 - X_2$ je roven 0 proti alternativě, že je menší než 0.)

V ý s l e d e k: $S^+ = 47 \in \overline{W}_{0,05} = \langle 36;136 \rangle$, resp. $u = -1,2669 \in \overline{W}_{0,05} = \langle -1,645;\infty)$;

Wilcoxonovým testem nezamítáme nulovou hypotézu (tj. domněnku o podceňování cen zamítáme)

$y = 5 \in \overline{W}_{0,05} = \langle 5;16 \rangle$, resp. $u = -1,5 \in \overline{W}_{0,05} = \langle -1,645;\infty)$; znaménkovým testem nezamítáme nulovou hypotézu (tj. domněnku o podceňování cen zamítáme)

Příklad 3.32

Výrobce určitého výrobku se má rozhodnout mezi dvěma dodavateli polotovarů vyrábějících je různými technologickými postupy. Rozhodující je procentní obsah účinné látky. Pro ověření, zda procentní obsah této látky je při použití obou technologií stejný, bylo náhodně vybráno 5 kusů vyrobených první technologií a 9 kusů vyrobených druhou technologií:

$x_i =$ 1,52 1,57 1,71 1,34 1,68

$y_j =$ 1,75 1,67 1,56 1,66 1,72 1,79 1,64 1,55 1,65

Testujte na 5% hladině významnosti hypotézu, že obě technologie poskytují stejné procento účinné látky.

V ý s l e d e k: $U_1 = 31 \in \overline{W}_{0,05} = \langle 8;37 \rangle$; hypotézu nezamítáme

$u \doteq 1,13333 \in \overline{W}_{0,05} = \langle -1,960;1,960 \rangle$; hypotézu nezamítáme (rozsahy výběrů jsou ale dosti malé!)

Kontrolní otázky

1. Definujte statistickou hypotézu a popište její druhy.
2. Co je testové kritérium a kritický obor?
3. Jakou konvenci používáme při testování statistické hypotézy?
4. Popište chybu 1. druhu při testování statistické hypotézy. Jaký je její praktický význam?
5. Popište chybu 2. druhu při testování statistické hypotézy. Jaký je její praktický význam?
6. Jaký je vztah mezi pravděpodobnostmi chyb 1. a 2. druhu a rozsahem náhodného výběru?
7. Jak souvisejí intervalové odhady s testy parametrických hypotéz?
8. Jakým způsobem používáme tzv. *P*-hodnotu při testování parametrické hypotézy na PC?
9. Popište grafickou metodu testu hypotézy o rozdělení pravděpodobnosti pozorované náhodné veličiny.
10. Proč používáme neparametrické testy a co omezuje jejich použití?
11. Popište princip transformace původního souboru na soubor pořadí a to i s ohledem na shodu pořadí.

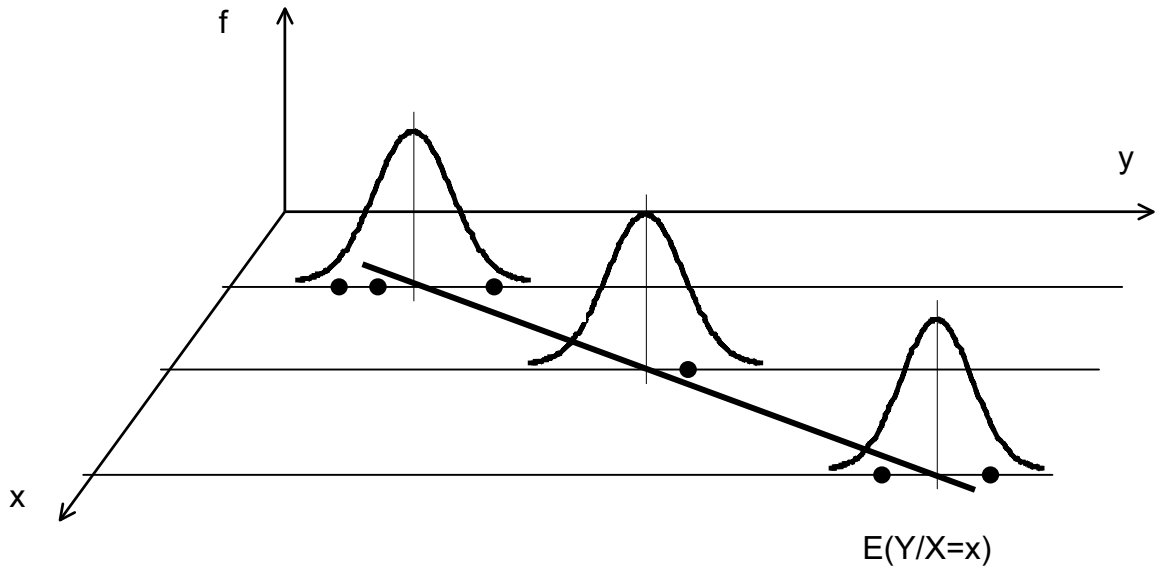
4 REGRESNÍ ANALÝZA

Regresní funkce

Důležitou statistickou úlohou v ekonomických aplikacích je hledání a zkoumání závislostí proměnných, jejichž hodnoty získáme při realizaci experimentů. Jde o stanovení závislosti pozorované náhodné veličiny Y na reálném vektoru nezávisle proměnných $\mathbf{X} = (X_1, ..., X_k)$, který může ale nemusí být náhodný (jeho případná náhodnost není v našem případě podstatná). Náhodnou veličinou Y může být např. výsledná cena výrobku a složky $X_1, ..., X_k$ vektoru $\mathbf{X}$ tvoří: ceny materiálu a energie, mzdy, daně a zisk. K popisu, stanovení a vyšetřování závislosti Y na $\mathbf{X}$ užíváme ***regresní analýzu***, přičemž tuto závislost vyjadřuje ***regresní funkce***

$$y = \varphi(\mathbf{x}, \boldsymbol{\beta}) = E(Y \mid \mathbf{X} = \mathbf{x}),$$

kde $\mathbf{x} = (x_1, ..., x_k)$ je vektor nezávisle proměnných (pozorovaná hodnota vektoru $\mathbf{X}$), y je závisle proměnná (pozorovaná hodnota náhodné veličiny Y) a $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)$ je vektor reálných parametrů, tzv. ***regresních koeficientů*** β_j, $j = 1, ..., m$. $E(Y \mid \mathbf{X} = \mathbf{x})$ je podmíněná střední hodnota náhodné veličiny Y, tj. její střední hodnota pro $\mathbf{x} = (x_1, ..., x_k)$.



Obr. 4.1

Při vyšetřování závislosti Y na $\mathbf{X}$ získáme realizací n experimentů $(k + 1)$-rozměrný statistický soubor $((\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)) = ((x_{11}, ..., x_{k1}, y_1), ..., (x_{1n}, ..., x_{kn}, y_n))$ s rozsahem n, kde y_i je pozorovaná hodnota náhodné veličiny Y_i (Y_i odpovídá i-tému pozorování Y) a

$\mathbf{x}_i = (x_{1i}, ..., x_{ki})$ je pozorovaná hodnota vektoru nezávisle proměnných **X**, $i = 1, ..., n$. Na obr. 4.1 je znázorněn případ pro k = 1, tedy pro **x** = x_1 = x (jde o tzv. ***regresní přímku***), a s opakovanými pozorováními. Opakování pozorování pro danou hodnotu nezávisle proměnné **x** však není v regresní analýze nezbytné. Pro určení odhadů neznámých regresních koeficientů β_j minimalizujeme tzv. ***reziduální součet čtverců***

$$S^* = \sum_{i=1}^{n} \left[y_i - \varphi(\mathbf{x}_i, \boldsymbol{\beta}) \right]^2$$

a hovoříme o tzv. ***metodě nejmenších čtverců***.

Před výpočtem regresních koeficientů volíme obvykle takový tvar regresní funkce, který co nejvíce odpovídá vyšetřované nebo uvažované závislosti. Bývá zvykem volit regresní funkci s co nejmenším počtem regresních koeficientů a jednoduchým předpisem, avšak dostatečně flexibilní a s požadovanými vlastnostmi: monotonie, předepsané hodnoty, asymptoty aj. Vychází se přitom povětšinou ze zkušenosti, avšak v současné době se při realizaci regresní analýzy na PC dají často úspěšně použít vhodné databáze regresních funkcí.

Regresní funkce rozdělujeme na ***lineární*** a ***nelineární regresní funkce***, a to vzhledem k regresním koeficientům, nikoli k vektoru nezávisle proměnných **x**. Některé nelineární regresní funkce můžeme vhodnou linearizací převést na lineární (např. mocninnou nebo exponenciální funkci logaritmujeme). Jde sice o běžně používaný postup, kdy ale řešíme jiný regresní model nežli původně uvažovaný. Blíže o linearizaci nelineární regresní funkce je pojednáno např. v [2], [3], [17], [19], [21], [29].

Lineární regresní model

Lineární regresní funkce má tvar

$$y = \sum_{j=1}^{m} \beta_j f_j(\mathbf{x}),$$

kde $f_j(\mathbf{x})$ jsou známé funkce neobsahující regresní koeficienty $\beta_1, ..., \beta_m$.

Uvažujeme tzv. ***lineární regresní model*** založený na předpokladech:

1. Funkce $f_j(\mathbf{x})$ nabývají hodnot $f_{ji} = f_j(\mathbf{x}_i)$ pro $j = 1, ..., m$ a $i = 1, ..., n$.
2. Matice $\mathbf{F} = \begin{pmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mn} \end{pmatrix}$ typu (*m*, *n*) s prvky f_{ji} má hodnost $m < n$.

3. Náhodná veličina Y_i má střední hodnotu $E(Y_i) = \sum_{j=1}^{m} \beta_j f_{ji}$ a konstantní rozptyl $D(Y_i) = \sigma^2 > 0$ pro $i = 1,...,n$.

4. Náhodné veličiny Y_i jsou nekorelované a mají normální rozdělení pravděpodobnosti pro $i = 1,...,n$.

Předpoklady 1 a 2 zaručují jednoznačnou existenci minima uvedeného reziduálního součtu čtverců, tj. určení bodových odhadů regresních koeficientů. Předpoklady 3 a 4 pak slouží k intervalovým odhadům a testováním hypotéz. V literatuře se místo popsaného lineárního regresního modelu také uvádí ekvivalentní model ve tvaru

$$Y_i = \sum_{j=1}^{m} \beta_j f_j(\mathbf{x}_i) + E_i, \quad i = 1,...,n,$$

kde E_i jsou nekorelované náhodné veličiny (vyjadřující např. náhodné chyby měření) s normálním rozdělením pravděpodobnosti N(0, σ^2).

Odhady regresních koeficientů, rozptylu a funkčních hodnot, a také testy statistických hypotéz o regresních koeficientech provádíme pomocí následujících vztahů. Označíme-li matice

$$\mathbf{H} = \mathbf{F}\mathbf{F}^{\mathrm{T}} = \begin{pmatrix} \sum_{i=1}^{n} f_{1i} f_{1i} & \cdots & \sum_{i=1}^{n} f_{1i} f_{mi} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} f_{mi} f_{1i} & \cdots & \sum_{i=1}^{n} f_{mi} f_{mi} \end{pmatrix}, \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{g} = \mathbf{F}\mathbf{y} = \begin{pmatrix} \sum_{i=1}^{n} f_{1i} y_i \\ \vdots \\ \sum_{i=1}^{n} f_{mi} y_i \end{pmatrix},$$

kde horní index $^{\mathrm{T}}$ označuje transpozici matice, pak platí:

1. ***Bodový odhad regresního koeficientu*** β_j je číslo b_j, $j = 1,...,m$, kde matice **b** je řešení soustavy lineárních algebraických rovnic (tzv. ***soustavy normálních rovnic***)

$$\mathbf{H}\mathbf{b} = \mathbf{g}.$$

2. ***Bodový odhad lineární regresní funkce*** je funkce

$$\hat{y} = \sum_{j=1}^{m} b_j f_j(\mathbf{x}),$$

jejíž konkrétní hodnota pro dané **x** je bodový odhad jak střední hodnoty, tak i individuální (predikované) hodnoty náhodné veličiny *Y*.

3. ***Bodový odhad rozptylu*** σ^2 náhodné veličiny Y je

$$s^2 = \frac{S^*_{\min}}{n - m},$$

kde $S^*_{\min} = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} b_j f_{ji} \right)^2 = \sum_{i=1}^{n} y_i^2 - \sum_{j=1}^{m} b_j g_j$ a g_j je prvek matice **g**.

4. ***Intervalový odhad regresního koeficientu*** β_j se spolehlivostí $1 - \alpha$, $j = 1,...,m$, je

$$\left\langle b_j - t_{1-\alpha/2} s\sqrt{h^{jj}}; b_j + t_{1-\alpha/2} s\sqrt{h^{jj}} \right\rangle,$$

kde h^{jj} je j-tý diagonální prvek matice $\mathbf{H}^{-1}$ a $t_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$-kvantil Studentova rozdělení s $n - m$ stupni volnosti - viz tabulku **T2**.

5. ***Intervalový odhad střední funkční hodnoty*** y regresní funkce (***konfidenční interval*** pro $E(Y \mid \mathbf{X} = \mathbf{x})$) se spolehlivostí $1 - \alpha$ je

$$\left\langle \sum_{j=1}^{m} b_j f_j(\mathbf{x}) - t_{1-\alpha/2} s\sqrt{h^*};\ \sum_{j=1}^{m} b_j f_j(\mathbf{x}) + t_{1-\alpha/2} s\sqrt{h^*} \right\rangle,$$

kde $h^* = \mathbf{f}(\mathbf{x})^{\mathrm{T}} \mathbf{H}^{-1} \mathbf{f}(\mathbf{x})$, přičemž $\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}$, a $t_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$-kvantil Studentova rozdělení s $n - m$ stupni volnosti - viz tabulku **T2**. ***Intervalový odhad individuální funkční hodnoty*** y regresní funkce (***predikční interval*** pro $Y \mid \mathbf{X} = \mathbf{x}$) se spolehlivostí $1 - \alpha$ obdržíme analogicky, avšak místo h^* vezmeme $1 + h^*$.

6. ***Test hypotézy*** $H : \beta_j = \beta_{j0}$ proti alternativní hypotéze $\bar{H} : \beta_j \neq \beta_{j0}$ na hladině významnosti α, kde j je jeden pevně zvolený index, $j = 1,...,m$, provádíme pomocí pozorované hodnoty testového kritéria

$$t = \frac{b_j - \beta_{j0}}{s\sqrt{h^{jj}}},$$

$\overline{W}_\alpha = \left\langle -t_{1-\alpha/2}; t_{1-\alpha/2} \right\rangle$ a $t_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$-kvantil Studentova rozdělení s $n - m$ stupni volnosti - viz tabulku **T2**. Tento test je možno také provést pomocí výše uvedeného intervalového odhadu koeficientu β_j se spolehlivostí $1 - \alpha$.

Z intervalových odhadů střední funkční hodnoty, resp. individuální funkční hodnoty, se konstruuje ***pás spolehlivosti pro střední hodnotu*** (***konfidenční pás***), resp. ***pás spolehlivosti pro individuální hodnotu*** (***predikční pás***) – viz např. užší, resp. širší, pás kolem regresní přímky na obr. 4.2. Poznamenejme ještě, že test hypotézy $H : \beta_j = \beta_{j0}$ se týká pouze

jednoho (byť libovolného) regresního koeficientu. Současný test více regresních koeficientů je nutno provést pomocí tzv. ***sdružené hypotézy*** - viz např. [2], [3], [17], [19], [21], [29].

Orientační mírou vhodnosti vypočtené regresní funkce pro získaná data je ***koeficient vícenásobné korelace***

$$r = \sqrt{1 - \frac{S^*_{\min}}{\sum y_i^2 - n\left(\overline{y}\right)^2}},$$

nazývaný také ***index*** (***koeficient***) ***determinace*** r^2 ($\overline{y}$ je aritmetický průměr hodnot y_i), který nabývá hodnot z intervalu $\langle 0;1 \rangle$. Číslo $r^2 100\,\%$ vyjadřuje procentuální podíl z rozptylu hodnot y_i "vysvětlený" vypočtenou regresní funkcí. Hodnoty r (a tím také r^2) blízké 1 naznačují vhodnost zvoleného tvaru regresní funkce. Pro bližší posouzení vhodnosti vypočtené regresní funkce se provádí její grafický rozbor vzhledem k pozorovaným bodům $[\mathbf{x}_1, y_1], ..., [\mathbf{x}_n, y_n]$. Pro rigorózní závěr je však nutné provést tzv. ***regresní diagnostiku*** a testovat další statistické hypotézy - viz např. [2], [3], [17], [19], [21], [29].

Nejvíce užívanou lineární regresní funkcí pro pozorovaný dvourozměrný statistický soubor $(x_1, y_1), ..., (x_n, y_n)$ je funkce

$$y = \beta_1 + \beta_2 x,$$

jejímž grafem je ***regresní přímka***. Pro tuto funkci je $k = 1$, $\mathbf{x} = x_1 = x$ (píšeme x místo x_1), $m = 2$, $f_1(x) = 1$, $f_2(x) = x$, takže

$$\mathbf{F} = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Při „ručním“ výpočtu můžeme pro regresní funkci $y = \beta_1 + \beta_2 x$ použít ***explicitní vztahy***, kde det **H** značí determinant matice **H**:

a) $\mathbf{H} = \begin{pmatrix} \sum_{i=1}^n 1\,; & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i; & \sum_{i=1}^n x_i^2 \end{pmatrix}$, $\mathbf{g} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$, $\sum_{i=1}^n 1 = n$,

b) $\det \mathbf{H} = n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2$, $b_2 = \dfrac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\det \mathbf{H}}$, $b_1 = \overline{y} - b_2 \overline{x}$, $\overline{x}$ a $\overline{y}$ jsou aritmetické průměry,

c) $S^*_{\min} = \sum_{i=1}^{n} \left(y_i - b_1 - b_2 x_i \right)^2 = \sum_{i=1}^{n} y_i^2 - b_1 \sum_{i=1}^{n} y_i - b_2 \sum_{i=1}^{n} x_i y_i$, $s^2 = \frac{S^*_{\min}}{n-2}$,

d) $h^{11} = \frac{\sum_{i=1}^{n} x_i^2}{\det \mathbf{H}}$, $h^{22} = \frac{n}{\det \mathbf{H}}$,

e) $h^* = \frac{1}{n} + \frac{\left(x - \overline{x} \right)^2}{\sum_{i=1}^{n} x_i^2 - n \left(\overline{x} \right)^2} = \frac{1}{n} + \frac{n \left(x - \overline{x} \right)^2}{\det \mathbf{H}}$,

f) $r = \left| \, r(x, y) \, \right|$, kde $r\left(x, y \right) = \frac{\sum_{i=1}^{n} x_i y_i - \overline{xy}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n \left(\overline{x} \right)^2 \right) \left(\sum_{i=1}^{n} y_i^2 - n \left(\overline{y} \right)^2 \right)}}$ je koeficient korelace statistického souboru $\left(\left(x_1, y_1 \right), ..., \left(x_n, y_n \right) \right)$.

V ekonomických úlohách se také často potkáváme s lineárními regresními funkcemi:

a) ***regresní rovina*** $y = \beta_1 + \beta_2 x_1 + \beta_3 x_2$, kde $k = 2$, $\mathbf{x} = \left(x_1, x_2 \right)$, $m = 3$, $f_1 \left(x_1, x_2 \right) = 1$, $f_2 \left(x_1, x_2 \right) = x_1$, $f_3 \left(x_1, x_2 \right) = x_2$,

b) ***regresní parabola*** $y = \beta_1 + \beta_2 x + \beta_3 x^2$, kde $k = 1$, $\mathbf{x} = x_1 = x$, $m = 3$, $f_1 \left(x \right) = 1$, $f_2 \left(x \right) = x$, $f_3 \left(x \right) = x^2$.

Jejich „ruční“ výpočet je však namáhavý a je lépe aplikovat profesionální statistický software (Minitab, Statistica, Statgraphics, QC Expert, SPSS, SAS aj.) anebo použít potřebné funkce a maticové operace v Excelu.

Příklad 4.1

U osmi náhodně vybraných firem poskytujících konzultace v oblasti jakosti výroby byly v roce 1993 zjištěny počty zaměstnanců x a roční obraty y (mil. Kč):

x_i	3	5	5	8	9	11	12	15
y_i	0,8	1,2	1,5	1,9	1,8	2,4	2,5	3,1

Vyjádřete závislost ročního obratu firmy na počtu zaměstnanců ve tvaru $y = \beta_1 + \beta_2 x$, vypočtěte intervalový odhad β_2 se spolehlivostí 0,95, testujte na hladině významnosti 0,05 hypotézu $H: \beta_1 = 0{,}2$, určete bodový a intervalový odhad $y(10)$ se spolehlivostí 0,95. Pomocí grafu a koeficientu korelace r posuďte vhodnost regresní funkce. Předpokládejte, že roční obrat má podmíněné normální rozdělení s konstantním rozptylem vzhledem k počtu zaměstnanců.

Ř e š e n í:

V tabulce jsou pomocné výpočty:

i	x_i	y_i	x_i^2	x_iy_i	y_i^2
1	3	0,8	9	2,4	0,64
2	5	1,2	25	6,0	1,44
3	5	1,5	25	7,5	2,25
4	8	1,9	64	15,2	3,61
5	9	1,8	81	16,2	3,24
6	11	2,4	121	26,4	5,76
7	12	2,5	144	30,0	6,25
8	15	3,1	225	46,5	9,61
Σ	68	15,2	694	150,2	32,80

Vlastní výpočty provedeme v následujících krocích.

1) Jde o regresní přímku, takže s využitím výše uvedených vzorců obdržíme pro $n = 8$ z tabulky matici $\mathbf{H} = \begin{pmatrix} 8 & 68 \\ 68 & 694 \end{pmatrix}$, jejíž determinant je $\det \mathbf{H} = 8 \cdot 694 - 68^2 = 928$, takže bodový odhad β_2 je

$$b_2 = \frac{8 \cdot 150,2 - 68 \cdot 15,2}{928} = 0,1810344 \approx 0,181.$$

Dále je $\overline{x} = 68/8 = 8,5$, $\overline{y} = 15,2/8 = 1,9$, takže bodový odhad β_1 je

$$b_1 = 1,9 - 0,1810344 \cdot 8,5 = 0,3612068 \approx 0,361.$$

Potom bodový odhad regresní funkce je $y = 0,361 + 0,181x$.

2) Minimální hodnota reziduálního součtu čtverců je

$$S^*_{\min} = 32,80 - 0,3612068.15,2 - 0,1810344 \cdot 150,2 \approx 0,1182758$$

a bodový odhad rozptylu σ^2, resp. směrodatné odchylky σ, je

$$s^2 = 0,1182758/(8 - 2) = 0,0197126, \text{ resp. } s = \sqrt{0,0197126} \approx 0,1404017.$$

3) Diagonální prvky matice $\mathbf{H}^{-1}$ jsou

$$h^{11} = 694/928 \approx 0,7478448, \quad h^{22} = 8/928 \approx 0,00862069.$$

Z tabulky **T2** je pro $8 - 2 = 6$ stupňů volnosti $t_{0,975} = 2,447$. Intervalový odhad regresního koeficientu β_2 je

$$\beta_2 \in < 0,1810344 - 2,447 \cdot 0,1404017 \sqrt{0,00862069};$$

$$0,1810344 + 2,447 \cdot 0,1404017 \sqrt{0,00862069} > = < 0,1491353;\ 0,2129334 > \approx$$

$$\approx < 0,149;\ 0,213 >.$$

Bodový odhad přírůstku ročního obratu odpovídajícího zvýšení počtu zaměstnanců firmy o jednoho je tedy 181 000 Kč a intervalový odhad tohoto přírůstku se spolehlivostí 0,95 je 149 000 Kč až 213 000 Kč.

4) Pozorovaná hodnota testového kritéria pro $H: \beta_1 = 0{,}2$ je

$$t = \frac{0{,}3612068 - 0{,}2}{0{,}1404017\sqrt{0{,}7478448}} \approx 1{,}3277.$$

Pro alternativní hypotézu $\bar{H}: \beta_1 \neq 0{,}2$ je $\bar{W}_{0{,}05} = < -2{,}447;\ 2{,}447 >$. Vzhledem k tomu, že $t \in \bar{W}_{0{,}05}$, hypotézu $\beta_1 = 0{,}2$ na hladině významnosti 0,05 nezamítáme. Na dané hladině významnosti vlastně nezamítáme hypotézu, že firma bez zaměstnanců (pracují jen majitelé), neboť $y(0) = \beta_1$, bude mít roční obrat okolo 200 000 Kč.

5) Bodový odhad střední i individuální hodnoty ročního obratu firmy pro 10 zaměstnanců je

$$y(10) = 0{,}3612068 + 0{,}1810344 \cdot 10 = 2{,}1715508 \approx 2{,}172.$$

U dané firmy lze tedy očekávat roční obrat okolo 2 172 000 Kč. Protože

$$h^* = \frac{1}{8} + \frac{8(10 - 8{,}5)^2}{928} = 0{,}1443965,$$

je intervalový odhad se spolehlivostí 0,95 střední hodnoty ročního obratu firmy s 10 zaměstnanci

$$y(10) \in < 2{,}1715508 - 2{,}447 \cdot 0{,}1404017\sqrt{0{,}1443965}\ ;$$

$$2{,}1715508 + 2{,}447 \cdot 0{,}1404017\sqrt{0{,}1443965} > = < 2{,}0409985;\ 2{,}3021031 > \approx$$

$$\approx < 2{,}040;\ 2{,}302 >.$$

Se spolehlivostí 0,95 lze očekávat, že střední hodnota ročního obratu takové firmy bude od 2 040 000 Kč do 2 302 000 Kč. Jestliže použijeme ve výpočtu $1 + h^*$ místo h^*, dostaneme intervalový odhad se spolehlivostí 0,95 individuální hodnoty ročního obratu firmy s 10 zaměstnanci

$$y(10) \in < 2{,}1715508 - 2{,}447 \cdot 0{,}1404017\sqrt{1{,}1443965}\ ;$$

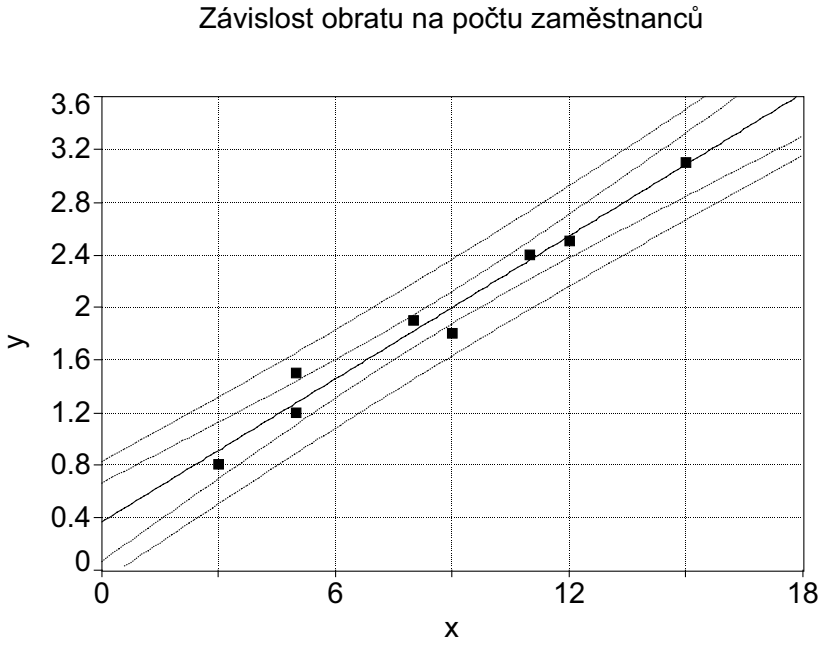
$$2{,}1715508 + 2{,}447 \cdot 0{,}1404017\sqrt{1{,}1443965} > = < 1{,}8040193;\ 2{,}5390823 > \approx$$

$$\approx < 1{,}804;\ 2{,}539 >.$$

Se spolehlivostí 0,95 lze očekávat, že individuální hodnota ročního obratu takové firmy bude od 1 804 000 Kč do 2 539 000 Kč.

6) Koeficient korelace je $r = 0{,}984798$, takže index determinace je $r^2 \approx 0{,}969827$.

Z grafu na obr. 4.2 a velikosti koeficientu korelace vidíme, že zvolený tvar regresní funkce vcelku dobře vystihuje danou závislost. Podle často používané konvence lze říci, získaná regresní funkce vyjadřuje celkem $r^2 100\,\% \approx 96,98\,\%$ změn (variability) pozorovaného obratu firmy.



Obr. 4.2

Příklady k procvičení

Příklad 4.2

Při sledování průměrných cen y (Kč) v roce 2005 a průměrných cen x (Kč) v roce 2004 u 6 vybraných druhů zboží byly získány následující hodnoty:

x_i	3,4	4,3	5,4	6,7	8,7	10,6
y_i	4,5	5,8	6,8	8,1	10,5	12,7

Určete regresní funkci $y = \beta_1 + \beta_2 x$, bodový odhad $y(5,4)$, intervalové odhady β_1, β_2 a $y(5,4)$ se spolehlivostí 0,95, a koeficient korelace.

V ý s l e d e k: $y \approx 0,7744 + 1,1190\,x$; $\beta_1 \in < 0,3095;\ 1,2394 >$; $\beta_2 \in < 1,0524;\ 1,1856 >$;

$y(5,4) \approx 6,8171$; $y(5,4) \in < 6,6350;\ 6,9992 >$, resp. $< 6,3710;\ 7,2632 >$;

$r \approx 0,999082$

Příklad 4.3

Poptávka po určitém výrobku y^* (v tis. ks) při jeho různých cenách x^* (Kč) zjištěná statistickým průzkumem uvedena v tabulce:

x_i^*	100	110	140	160	200
y_i^*	120	89	56	41	22

Vyjádřete závislost poptávky na ceně mocninnou regresní funkcí $y^* = \gamma x^{*\,\delta}$, určete bodové a intervalové odhady (se spolehlivostí 0,95) regresních koeficientů a poptávky pro cenu výrobku 120 Kč. (Návod: logaritmujte mocninnou funkci.)

V ý s l e d e k: $\ln y^* \approx 15{,}64395 - 2{,}36035 \ln x^*$; $y^* = 6{,}224 \cdot 10^6 x^{*\,-2{,}36}$;

$\ln \gamma = \beta_1 \in < 13{,}95342;\ 17{,}33448 >$; $\delta = \beta_2 \in < -2{,}37817;\ -2{,}34253 >$;

$y^*(120) \approx 77$; $y^*(120) \in < 69{,}8;\ 84{,}9 >$, resp. $< 62{,}0;\ 95{,}6 >$

Příklad 4.4

U 6 výrobků jedné firmy byly zjištěny náklady y (Kč) a ceny x (Kč):

x_i	40	64	34	15	57	45
y_i	33	46	23	12	56	40

Určete regresní funkci $y = \beta_1 + \beta_2 x$, bodový odhad rozptylu σ^2, intervalový odhad koeficientu β_2 se spolehlivostí 0,95 a testujte hypotézu $\beta_2 = 0$ na hladině významnosti 0,05.

V ý s l e d e k: $y = -1{,}3082 + 0{,}8543\, x$; $\sigma^2 \approx 39{,}8439$; $\beta_2 \in \langle 0{,}404;\ 1{,}305 \rangle$, takže hypotézu zamítáme

Příklad 4.5

Pro posouzení závislosti letošní poptávky y na loňské poptávce x na jistý druh zboží byly u 6 obchodníků zjištěny údaje (ks):

x_i	20	60	70	100	150	260
y_i	50	60	60	120	230	320

Určete bodové a intervalové odhady (se spolehlivostí 95 %) koeficientů regresní přímky a hodnoty letošní poptávky pro loňskou poptávku 110 kusů. Na hladině významnosti 5 % testujte hypotézu, že $\beta_1 = 0$ a určete koeficient korelace.

V ý s l e d e k: $y \approx 0{,}687 + 1{,}266x$; $\beta_1 \in < -57{,}194;\ 58{,}568 >$; $\beta_2 \in < 0{,}836;\ 1{,}696 >$;

$y(110) \approx 140$; $y(110) \in < 106{,}55;\ 173{,}45 >$, resp. $< 51{,}50;\ 228{,}50 >$;

hypotézu nezamítáme; $r \approx 0{,}97198$

Příklad 4.6

Pozorováním množství y prodaných akcií v závislosti na odchylce ceny x (kč) jedné akcie firmy STAMET od emisní hodnoty byla získána data:

x_i	-60	-32	-15	1	15	30	55
y_i	781	824	840	855	868	882	897

Vypočtěte regresní funkci $y = \beta_1 + \beta_2 x$, bodový odhad rozptylu σ^2, intervalový odhad koeficientu β_2 se spolehlivostí 0,95 a bodový i intervalový odhad hodnoty y pro $x = -30$ a $x = 15$.

V ý s l e d e k: $y = 850{,}4 + 0{,}991x$; $\sigma^2 \approx 46{,}67$; $\beta_2 \in \langle 0{,}773;\ 1{,}208\rangle$; $y(-30) \approx 820{,}7$;

$y(-30) \in \langle 810{,}6;\ 830{,}8\rangle$; $y(15) \approx 865{,}3$; $y(15) \in \langle 861{,}8;\ 868{,}8\rangle$

Příklad 4.7

Velikost čistého zisku y^* (tis. Kč) firmy STATEX v prvních 6 letech x^* její činnosti je v následující tabulce:

x_i^*	1	2	3	4	5	6
y_i^*	112	149	238	354	580	867

Aproximujte data exponenciální regresní funkcí $y^* = \gamma \exp(\delta x^*)$ a určete bodové i intervalové odhady (se spolehlivostí 95 %) regresních koeficientů a předpovědi zisku v 7. roce činnosti firmy, a koeficient korelace. (Návod: logaritmujte exponenciální funkci.)

V ý s l e d e k: $\ln y^* \approx 4{,}22798 + 0{,}42020\, x^*$; $y^* = 68{,}57875 \exp(0{,}42020\, x^*)$;

$\gamma = \exp(\beta_1) \in < 59{,}40715;\ 79{,}16632 >$; $\delta = \beta_2 \in < 0{,}38333;\ 0{,}45706 >$;

$y^*(7) \approx 1299{,}04$; $y^*(7) \in < 1125{,}30;\ 1499{,}59 >$, resp. $< 1052{,}24;\ 1603{,}71 >$;

$r \approx 0{,}99801$ pro linearizovanou regresní funkci

Příklad 4.8

Měřením byly získány hodnoty:

x_i	0,75	1,50	2,25	3,00	3,75	4,50	5,10	6,10	6,70	7,50
y_i	0,017	0,046	0,075	0,110	0,142	0,167	0,188	0,224	0,262	0,282

Určete regresní funkci $y = \beta_1 + \beta_2 x$, vypočtěte bodový odhad rozptylu σ^2, testujte hypotézu

$\beta_1 = 0$ na hladině významnosti 5 % a vypočtěte intervalový odhad koeficientu β_2 se spolehlivostí 95 %.

V ý s l e d e k: $y = -0{,}012009 + 0{,}039686x$; $\sigma^2 \approx 2{,}07 \cdot 10^{-5}$; hypotézu $\beta_1 = 0$ zamítáme;

$\beta_2 \in \langle 0{,}038066;\ 0{,}041064 \rangle$

Příklad 4.9

Na souřadnicové vrtačce byla za teploty 20 °C nastavena vzdálenost 1 m od počátku souřadné soustavy a měřena diference y (m) mezi skutečnou a nastavenou vzdáleností v závislosti na přírůstku teploty x (°C):

x_i	10	20	30	40	50	60
y_i	0,00018	0,00035	0,00048	0,00065	0,00084	0,00097

Pomocí regresní funkce $y = \beta_1 + \beta_2 x$ vypočtěte bodový a intervalový odhad chyby počátečního nastavení β_1, koeficientu tepelné roztažnosti β_2 a skutečné vzdálenosti $d = y + 1$ od počátku souřadné soustavy pro teplotu 35 °C se spolehlivostí 95 %.

V ý s l e d e k: $\beta_1 \approx 1{,}93333 \cdot 10^{-5}$ m; $\beta_1 \in \langle -2{,}31756 \cdot 10^{-5};\ 6{,}18423 \cdot 10^{-5} \rangle$ m;

$\beta_2 \approx 1{,}59714 \cdot 10^{-5}$ m·°C^{-1}; $\beta_2 \in \langle 1{,}48799 \cdot 10^{-5};\ 1{,}70630 \cdot 10^{-5} \rangle$ m·°C^{-1};

$d(35) \approx 1{,}000578333$ m; $d(35) \in \langle 1{,}000559669 \cdot 10^{-5};\ 1{,}000596998 \cdot 10^{-5} \rangle$ m,

resp. $d(35) \in \langle 1{,}000528953 \cdot 10^{-5};\ 1{,}000627714 \cdot 10^{-5} \rangle$ m;

Příklad 4.10

Určete odhad regresní funkce $y = \beta_1 + \beta_2 x_1 + \beta_3 x_2$ a vypočtěte intervalové odhady koeficientu β_2, β_3 se spolehlivostí 0,95, jestliže pro každou dvojici (x_1, x_2) je Y náhodná veličina s normálním rozdělením a veličiny Y jsou pro různé dvojice (x_1, x_2) nezávislé. Experimentem byla získána data uvedená v tabulce:

x_{1i}	1,0	3,0	3,0	5,0	7,0	7,0	9,0	11,0	11,0	13,0
x_{2i}	0,2	0,7	0,1	0,3	0,2	0,6	0,2	0,2	0,7	0,5
y_i	2,0	2,8	5,3	5,9	7,4	5,6	8,7	11,2	10,4	13,2

V ý s l e d e k: $y = 1{,}82 + 0{,}918x_1 - 2{,}7x_2$; $\sigma^2 = 0{,}5419$; $\beta_2 \in \langle 0{,}768; 1{,}068 \rangle$;

$\beta_3 \in \langle -5{,}291; -0{,}119 \rangle$

Příklad 4.11

U osobního automobilu byla měřena spotřeba paliva y (v litrech na 100 km) v závislosti na

jeho rychlosti x (km/hod.) za konstantních podmínek. Byly získány hodnoty:

x_i	40	50	60	70	80	90	100
y_i	6,4	6,1	6,3	6,8	7,1	8,4	10,3

Určete regresní funkci $y = \beta_1 + \beta_2 x + \beta_3 x^2$, bodový odhad rozptylu σ^2 a na hladině významnosti 0,05 testujte hypotézu, že závislost je lineární (tj. $\beta_3 = 0$).

V ý s l e d e k: $y = 11,693 - 2,073 \cdot 10^{-1} x + 1,917 \cdot 10^{-2} x^2$; $\sigma^2 = 5,202 \cdot 10^{-2}$; hypotézu $\beta_3 = 0$ zamítáme, neboť $0 \notin \langle 1,590 \cdot 10^{-3} ; 2,242 \cdot 10^{-3} \rangle$ pro spolehlivost $1 - \alpha - 0,95$

Kontrolní otázky

1. Co se rozumí regresní analýzou a jaký je statistický princip regresní analýzy?
2. Definujte regresní funkci a lineární regresní funkci?
3. Na jakých předpokladech je založen lineární regresní model?
4. Jaké odhady a testy statistických hypotéz používáme v regresní analýze?
5. Jaký je rozdíl mezi odhady střední a individuální funkční hodnoty regresní funkce?
6. Jak posuzujeme vhodnost vypočtené regresní funkce?
7. Uveďte konkrétní příklady lineární a nelineární regresní funkce.
8. Uveďte konkrétní aplikaci regresní analýzy ve svém oboru.

5 ANALÝZA ROZPTYLU

Motivace a základní pojmy

V ekonomických, finančních a výrobních aplikacích statistických metod se často setkáváme s problémy posouzení vlivu nějakých faktorů na pozorovanou náhodnou veličinu. Jde například o ověření vlivu výše vzdělání na velikost příjmu jedince, druhu obchodu na cenu daného zboží, typu reklamy a věkové kategorie zákazníka na objem jím nakupovaného zboží, dne v týdnu a směny na kvalitu výroby, banky a času na kurz měnové jednotky apod. Uvažované faktory mají obvykle charakter kategoriálního znaku, který nabývá známých a rozlišitelných hodnot. V dále popsaných základních metodách jde sice převážně o posouzení vlivu faktorů na střední hodnotu pozorované náhodné veličiny, ale vlastní analýza vychází z rozptylu pozorovaných hodnot této veličiny, takže hovoříme o ***analýze rozptylu***, jejíž zkratka je ***ANOVA*** (z anglického „analysis of variance“). Analýzu rozptylu rozlišujeme podle počtu ovlivňujících faktorů (třídících znaků). V případě jednoho znaku A hovoříme o ***analýze rozptylu jednoduchého třídění***, v případě dvou znaků A a B jde o ***analýzu rozptylu dvojného třídění***. Analýzu rozptylu dvojného třídění se dvěma třídícími znaky A, B dále rozdělujeme na analýzu ***bez interakce*** těchto znaků, když nepředpokládáme jejich společné působení, a na analýzu ***s interakcí*** těchto znaků, když uvažujeme jejich společné působení, tj. jakoby třetího znaku označeného AB. V případě většího počtu třídících znaků pak jde o modely s dalšími možnými interakcemi. Poznamenejme ještě, že „ruční“ zpracování analýzy rozptylu je únosné nejvýše pro analýzu rozptylu s jedním nebo dvěma třídícími znaky. Metody analýzy rozptylu jsou velmi rozpracované a implementované do profesionálního statistického softwaru a částečně i do Excelu.

Analýza rozptylu jednoduchého třídění (ANOVA 1)

Předpokládáme, že pozorováním náhodné veličiny X byl získán statistický soubor $(x_1,...,x_n)$ s rozsahem n a dále, že znak A nabývá I různých kvalitativních hodnot $A_1,...,A_I$, kde $I \geq 2$. Přitom hodnotě A_i daného znaku odpovídá skupina $(x_{i1},...,x_{in_i})$ s rozsahem n_i, $i = 1,...,I$, prvků původního statistického souboru tak, že je původní soubor statistický soubor $(x_1,...,x_n)$ rozdělen do I disjunktních skupin (podsouborů). Zřejmě je $\sum_{i=1}^{I} n_i = n$. Pro zpracování analýzy rozptylu používáme tyto číselné charakteristiky:

a) aritmetický průměr *i*-té skupiny $\overline{x}_{i\bullet} = \frac{x_{i\bullet}}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, kde $x_{i\bullet} = \sum_{j=1}^{n_i} x_{ij}$ je součet prvků *i*-té skupiny, $i = 1,...,I$,

b) celkový průměr $\overline{x}_{\bullet\bullet} = \frac{x_{\bullet\bullet}}{n} = \frac{1}{n}\sum_{i=1}^{I} n_i \overline{x}_{i\bullet}$, kde $x_{\bullet\bullet} = \sum_{i=1}^{I} x_{i\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} x_{ij}$ je součet všech prvků původního souboru.

Analýza rozptylu jednoduchého třídění vychází z modelu ve tvaru

$$X_{ij} = \mu + \alpha_i + E_{ij},$$

kde E_{ij} jsou nezávislé náhodné veličiny s normálním rozdělením $N(0,\sigma^2)$ a μ, α_i, σ^2 jsou neznámé parametry. Hypotéze, že znak *A* nemá vliv na pozorovanou náhodnou veličinu *X*, odpovídá sdružená hypotéza $H : \alpha_1 = \cdots = \alpha_I = 0$ s alternativní hypotézou $\overline{H}$, že aspoň jedno α_i je různé od ostatních α_k, tj. že znak *A* má vliv na náhodnou veličinu *X*.

Pro testování použijeme rozklad součtu čtverců

$$S_t = S_A + S_e,$$

kde

a) ***celkový součet čtverců*** $S_t = \sum_{i=1}^{I}\sum_{j=1}^{n_I}\left(x_{ij} - \overline{x}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_I} x_{ij}^{\,2} - \frac{\left(x_{\bullet\bullet}\right)^2}{n}$,

b) ***součet čtverců mezi skupinami*** $S_A = \sum_{i=1}^{I} n_i \left(\overline{x}_{i\bullet} - \overline{x}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{I} \frac{\left(x_{i\bullet}\right)^2}{n_i} - \frac{\left(x_{\bullet\bullet}\right)^2}{n}$,

c) ***reziduální součet čtverců*** $S_e = \sum_{i=1}^{I}\sum_{j=1}^{n_I}\left(x_{ij} - \overline{x}_{i\bullet}\right)^2 = S_t - S_A$.

Hypotézu *H* testujeme pomocí testovacího kritéria

$$F = \frac{\frac{S_A}{I-1}}{\frac{S_e}{n-I}}$$

s oborem nezamítnutí $\overline{W}_\alpha = \langle 0; F_{1-\alpha} \rangle$, kde $F_{1-\alpha}$ je $(1-\alpha)$-kvantil Fisherova-Snedecorova rozdělení s $k_1 = I - 1$ a $k_2 = n - I$ stupni volnosti – viz tabulku **T4**. Pro $I = 2$ můžeme použít Studentův dvouvýběrový test, avšak nikoli pro $I > 2$ všechny dvouvýběrové testy, protože vzniká problém s nastavením hladiny významnosti a závislostí testových kritérií.

Testování zapisujeme obvykle do následující ***tabulky analýzy rozptylu***:

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testové kritérium
Znak A	S_A	$I-1$	$S_A/(I-1)$	$\dfrac{S_A/(I-1)}{S_e/(n-I)}$
Reziduální	S_e	$n-I$	$S_e/(n-I)$	---
Celkový	S_t	$n-1$	---	---

Při výpočtu na PC bývá tabulka zprava doplněna o sloupec obsahující *P*-hodnotu, která umožňuje test bez použití kvantilu $F_{1-\alpha}$.

Jestliže přijmeme alternativní hypotézu, že daný třídící znak má vliv na třídění, pak obvykle testujeme tzv. ***kontrasty***, tj. hledáme dvojice A_i a A_k, které vliv třídícího znaku způsobují. Použijeme k tomu postupně hypotézy $H: \alpha_i = \alpha_k$ s alternativami $\bar{H}: \alpha_i \neq \alpha_k$ pro $i = 1,...,I$, $k = 1,...,I$, $i < k$. Tyto hypotézy můžeme testovat Studentovým dvouvýběrovým testem anebo pomocí adekvátního testového kritéria

$$F = \frac{\dfrac{\left(\bar{x}_{i\bullet} - \bar{x}_{k\bullet}\right)^2}{I-1}}{\dfrac{S_e}{n-I}} \frac{n_i n_k}{n_i + n_k}$$

se stejným oborem nezamítnutí $\overline{W}_\alpha = \langle 0; F_{1-\alpha} \rangle$ jako má původní sdružená hypotéza o vlivu znaku *A*.

Pro úplnost analýzy rozptylu je zapotřebí rozhodnout, zda všechny rozptyly σ_i^2 náhodných veličin odpovídajících jednotlivým skupinám jsou stejné. Jde o test sdružené hypotézy $H: \sigma_1^2 = \cdots = \sigma_I^2$ s alternativou, že aspoň dva rozptyly jsou různé. Nejčastěji se k tomu používá ***Bartlettův test*** s kritériem

$$B = \frac{1}{C}\left[(n-I)\ln s^2 - \sum_{i=1}^{I}(n_i - 1)\ln s_i^2\right],$$

kde

$$C = 1 + \frac{1}{3(I-1)}\left(\sum_{i=1}^{I}\frac{1}{n_i - 1} - \frac{1}{n-I}\right),$$

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_{i\bullet}\right)^2 = \frac{1}{n_i - 1}\left(\sum_{j=1}^{n_i} x_{ij}^2 - \frac{\left(x_{i\bullet}\right)^2}{n_i}\right),$$

$$s^2 = \frac{1}{n-I}\sum_{i=1}^{I}(n_i - 1)s_i^2 = \frac{S_e}{n-I}\ .$$

Obor nezamítnutí hypotézy H je $\overline{W}_\alpha = \left\langle 0, \chi^2_{1-\alpha} \right\rangle$, kde $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil Pearsonova rozdělení chí-kvadrát s $k = I - 1$ stupni volnosti – viz tabulku **T3**. Jde o přibližný, ale plně dostačující test.

Poznamenejme, že zamítnutí sdružené hypotézy o skupinových rozptylech má také za následek odhalení vlivu znaku *A* na pozorovanou náhodnou veličinu *X*. Další testy používané při analýze rozptylu jednoduchého třídění (včetně neparametrického Kruskalova-Wallisova testu) a metodách analýzy rozptylu s více třídícími znaky bez i s interakcemi lze nalézt např. v [2], [3], [4], [6], [10], [22], [28].

Příklad 5.1

Sledováním měsíčních platů (v tisících Kč) tří pracovníků vykonávajícím stejnou práci během půl roku byly získány následující údaje, kde A_i odpovídá *i*-tému pracovníku, $i = 1, 2, 3$, a x_{ij} jsou jeho měsíční platy:

$A_1 \ \ldots \ x_{1j} = 22; 20; 19; 20; 21; 19,$

$A_2 \ \ldots \ x_{2j} = 20; 22; 21; 22; 24; 23,$

$A_3 \ \ldots \ x_{3j} = 29; 28; 26; 26; 27; 25.$

Pomocí ANOVA 1 testujeme na hladině významnosti 0,05 hypotézu, že střední měsíční platy všech tří pracovníků jsou stejné.

Ř e š e n í:

Pomocné výpočty jsou v tabulce:

j	x_{1j}	x_{2j}	x_{3j}	x_{1j}^2	x_{2j}^2	x_{3j}^2
1	22	20	29	484	400	841
2	20	22	28	400	484	784
3	19	21	26	361	441	676
4	20	22	26	400	484	676
5	21	24	27	441	576	729
6	19	23	25	361	529	625
Σ	121	132	161	2447	2914	4331
ΣΣ	414			9692		

Ze zadání $I = 3$, $n = 18$, $n_1 = n_2 = n_3 = 6$ a z tabulky pomocných výpočtů dostaneme

$$S_t = \sum_{i=1}^{I} \sum_{j=1}^{n_I} x_{ij}^{\ 2} - \frac{(x_{..})^2}{n} = 9692 - \frac{414^2}{18} = 170,$$

$$S_A = \sum_{i=1}^{I} \frac{(x_{i\bullet})^2}{n_i} - \frac{(x_{\bullet\bullet})^2}{n} = \frac{121^2}{6} + \frac{132^2}{6} + \frac{161^2}{6} - \frac{414^2}{18} \doteq 142,33333,$$

$$S_e = S_t - S_A \doteq 170 - 142,33333 = 27,66667,$$

Počty stupňů volnosti jsou $I-1=2$ a $n-I=15$, takže

$$S_A/(I-1) = 142,33333/2 = 71,16667,$$

$$S_e/(n-I) \doteq 27,66667/15 \doteq 1,84444,$$

$$F = \frac{\frac{S_A}{I-1}}{\frac{S_e}{n-I}} \doteq \frac{71,16667}{1,84444} \doteq 38,58434.$$

Tabulka analýzy rozptylu pak je:

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testové kritérium
Znak *A*	142,33333	2	71,66667	38,58434
Reziduální	27,66667	15	1,84444	---
Celkový	170,00000	17	---	---

Pro $k_1 = I-1=2$ a $k_2 = n-I=15$ stupňů volnosti je $F_{0,975}=4,765$ z tabulky **T4**. Tabulka **T4** neobsahuje kvantily $F_{0,95}$, ale je $F_{0,95} < F_{0,975}$, což k našemu testu stačí, ale např. z Excelu pomocí funkce FINV dostaneme $F_{0,95}=3,682$. Protože $F = 38,58434 \notin \overline{W}_{0,.05} = \langle 0;3,862 \rangle$, zamítáme sdruženou hypotézu *H* o stejných středních měsíčních platech na hladině významnosti 0,05. Testujeme proto dále kontrasty, tj. rozdíly středních měsíčních platů dvojic pracovníků.

Z tabulky pomocných výpočtů je

$$\overline{x}_{1\bullet} = \frac{x_{1\bullet}}{n_1} = \frac{121}{6} \doteq 20,16667,\ \overline{x}_{2\bullet} = \frac{x_{2\bullet}}{n_2} = \frac{132}{6} = 22,00000,\ \overline{x}_{3\bullet} = \frac{x_{3\bullet}}{n_3} = \frac{161}{6} \doteq 26,83333.$$

Při testech kontrastů obdržíme:

a) 1. pracovník $\leftrightarrow$ 2. pracovník:

$$F = \frac{\frac{(\overline{x}_{1\bullet} - \overline{x}_{2\bullet})^2}{I-1}}{\frac{S_e}{n-I}} \frac{n_1 n_2}{n_1+n_2} \doteq \frac{\frac{(20,16667-22,00000)^2}{2}}{1,84444} \frac{6\cdot 6}{6+6} \doteq 2,73343 \in \overline{W}_{0,.05} = \langle 0;3,862 \rangle,$$

takže hypotézu o rovnosti středních měsíčních platů 1. a 2. pracovníka nezamítáme,

b) 1. pracovník ↔ 3. pracovník:

$$F = \frac{\frac{\left(\overline{x}_{1.} - \overline{x}_{3.}\right)^2}{I-1}}{\frac{S_e}{n-I}} \frac{n_1 n_3}{n_1 + n_3} \doteq \frac{\frac{\left(20{,}16667 - 26{,}83333\right)^2}{2}}{1{,}84444} \frac{6 \cdot 6}{6+6} \doteq 36{,}14458 \notin \overline{W}_{0{,}05} = \left\langle 0; 3{,}862 \right\rangle,$$

takže hypotézu o rovnosti středních měsíčních platů 1. a 3. pracovníka zamítáme,

c) 2.pracovník ↔ 3. pracovník:

$$F = \frac{\frac{\left(\overline{x}_{2.} - \overline{x}_{3.}\right)^2}{I-1}}{\frac{S_e}{n-I}} \frac{n_2 n_3}{n_2 + n_3} \doteq \frac{\frac{\left(22{,}00000 - 26{,}83333\right)^2}{2}}{1{,}84444} \frac{6 \cdot 6}{6+6} \doteq 38{,}99849 \notin \overline{W}_{0{,}05} = \left\langle 0; 3{,}862 \right\rangle,$$

takže hypotézu o rovnosti středních měsíčních platů 2. a 3. pracovníka zamítáme.

Pro Bartlettův test rovnosti skupinových rozptylů je:

$$C = 1 + \frac{1}{3\left(I-1\right)} \left(\sum_{i=1}^{I} \frac{1}{n_i - 1} - \frac{1}{n-I} \right) = 1 + \frac{1}{3 \cdot 2} \left(\frac{1}{5} + \frac{1}{5} + \frac{1}{5} - \frac{1}{15} \right) \doteq 1{,}08889,$$

$$s_1^2 = \frac{1}{n_1 - 1} \left(\sum_{j=1}^{n_1} x_{1j}^2 - \frac{\left(x_{1.}\right)^2}{n_1} \right) = \frac{1}{5} \left(2447 - \frac{121^2}{6} \right) \doteq 1{,}36667,$$

$$s_2^2 = \frac{1}{n_2 - 1} \left(\sum_{j=1}^{n_2} x_{2j}^2 - \frac{\left(x_{2.}\right)^2}{n_2} \right) = \frac{1}{5} \left(2914 - \frac{132^2}{6} \right) = 2{,}00000,$$

$$s_3^2 = \frac{1}{n_3 - 1} \left(\sum_{j=1}^{n_3} x_{3j}^2 - \frac{\left(x_{3.}\right)^2}{n_3} \right) = \frac{1}{5} \left(4331 - \frac{161^2}{6} \right) \doteq 2{,}16667,$$

$$s^2 = \frac{S_e}{n-I} \doteq 1{,}84444,$$

takže

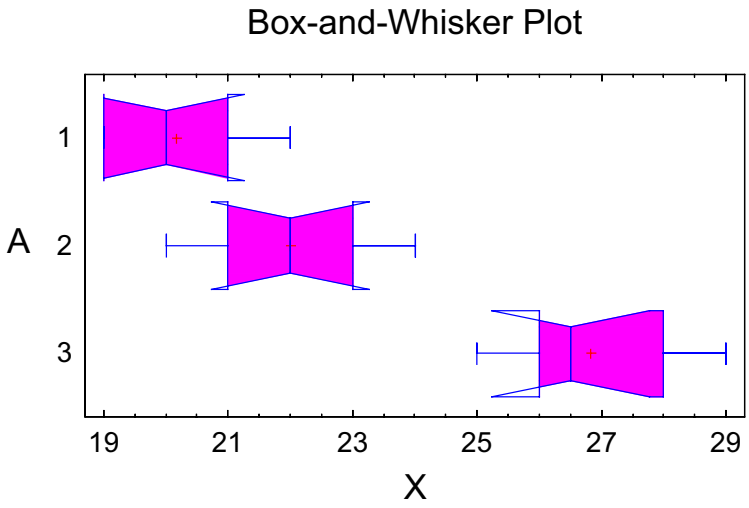
$$B = \frac{1}{C} \left[\left(n-I\right) \ln s^2 - \sum_{i=1}^{I} \left(n_i - 1\right) \ln s_i^2 \right] \doteq$$

$$\doteq \frac{1}{1{,}08889} \left[15 \ln 1{,}84444 - \left(5 \ln 1{,}36667 + 5 \ln 2 + 5 \ln 2{,}16667 \right) \right] \doteq 0{,}26548.$$

Protože $B = 0{,}26548 \in \overline{W}_{0{,}95} = \left\langle 0; 5{,}991 \right\rangle$, kde $\chi^2_{0{,}95} = 5{,}991$ pro $k = I - 1 = 2$ stupňů volnosti z tabulky **T3**, nezamítáme sdruženou hypotézu o rovnosti skupinových rozptylů.

Na obr. 5.2 jsou pro ilustraci krabicové grafy skupin (měsíčních platů jednotlivých pracovníků), které naznačují nenormální rozdělení pravděpodobnosti (kladná asymetrie pro 1. a 3. pracovníka), takže by bylo pro analýzu rozptylu adekvátnější použít neparametrický

Kruskalův-Wallisův test [2], [3]. Vzhledem k tomu, že rozsahy skupin jsou poměrně malé, to ale není zcela nezbytné.



Obr. 5.2

Test sdružené hypotézy H o rovnosti středních hodnot v analýze rozptylu s jedním nebo dvěma třídícími znaky (faktory) můžeme také realizovat snadno v Excelu, kde zvolíme ***Nástroje/Analýza dat/Analýza: jeden faktor***. Ukázka kompletního výstupu této analýzy pro náš příklad 5.1 je v následující tabulce:

Anova: jeden faktor						
Faktor						
Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*		
A1	6	121	20,16667	1,366667		
A2	6	132	22	2		
A3	6	161	26,83333	2,166667		
ANOVA						
Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Mezi výběry	142,3333	2	71,16667	38,58434	1,22E-06	3,682317
Všechny výběry	27,66667	15	1,844444			
Celkem	170	17				

Příklady k procvičení

Příklad 5.2

U čtyř odrůd brambor označených A_1, A_2, A_3, A_4 se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky (v kg) jsou v následující tabulce:

Odrůda	Hmotnost
A_1	0,9 0,8 0,6 0,9
A_2	1,3 1,0 1,3
A_3	1,3 1,5 1,6 1,1 1,5
A_4	1,1 1,2 1,0

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

V ý s l e d e k: $k_1 = 3$, $k_2 = 11$, $F = 9,97 \notin \overline{W}_{0,95} = \langle 0; 3,59 \rangle$; hypotézu o nezávislosti na odrůdě zamítáme; statisticky významně se liší pouze odrůdy A_1 a A_3

Příklad 5.3

Ve firmě PRASTAT se měřil čas, který potřeboval každý ze tří dělníků D1, D2 a D3 k uskutečnění téhož pracovního úkonu. Dosažené časy v minutách:

D1	3,6	3,8	3,7	3,5		
D2	4,3	3,9	4,2	3,9	4,4	4,7
D3	4,2	4,5	4,0	4,1	4,5	4,4

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

V ý s l e d e k: $k_1 = 2$, $k_2 = 13$, $F = 9,665 \notin \overline{W}_{0,05} = \langle 0; 3,806 \rangle$; hypotézu o stejných výkonech zamítáme; liší se výkony dvojic dělníků (D1,D2), (D1,D3) a neliší se (D2,D3).

Příklad 5.4

Pracovníci vybrané firmy byly školeni z metod řízení jakosti za využití pěti výukových metod: tradiční způsob, programová výuka, audiotechnika, audiovizuální technika a vizuální technika. Z každé skupiny byl vybrán náhodný vzorek pracovníků a všichni byli podrobeni témuž písemnému testu. Na hladině významnosti 0,05 testujte hypotézu, že znalosti všech pracovníků jsou stejné a nezávisí na použité výukové metodě. V případě zamítnutí hypotézy zjistěte, které metody se liší na hladině významnosti 0,05. Dosažené body dle metod jsou v následující tabulce:

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

V ý s l e d e k: $k_1 = 4$, $k_2 = 26$, $F = 1,624 \in \overline{W}_{0,05} = \langle 0; 2,743 \rangle$; hypotézu nezamítáme, znalosti nezávisí na použité výukové metodě

Příklad 5.4

Student soukromé vysoké školy Akademie Sting v Brně může cestovat ze svého brněnského bydliště do školy třemi různými způsoby: trolejbusem (A), autobusem (B) a osobním autem (C). Máme k dispozici jeho naměřené časy cestování do školy v době ranní špičky (včetně čekání na příslušný spoj) v minutách:

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

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

V ý s l e d e k: $k_1 = 2$, $k_2 = 15$, $F = 6,715 \notin \overline{W}_{0,05} = \langle 0; 3,682 \rangle$; hypotézu zamítáme, způsob dopravy má vliv na dobu cestování; neliší se způsoby (A,C) a liší se způsoby (A,B) a (B,C)

Kontrolní otázky

1. Popište motivaci analýzy rozptylu a uveďte příklady na ANOVA 1 a ANOVA 2 bez i s interakcí.
2. Popište model ANOVA 1 pro sdruženou hypotézu středních hodnot.
3. Co rozumíme rozkladem celkového součtu čtverců?
4. Kdy a jakým způsobem testujeme kontrasty a rovnost skupinových rozptylů?

6 KATEGORIÁLNÍ ANALÝZA

Motivace

Při statistickém vyhodnocování průzkumu např. zájmu o výrobky, služby, zboží a úspěšnosti reklamy jde často o posouzení a postižení závislosti a vzájemného ovlivňování sledovaných vícerozměrných kategoriálních (kvalitativních) znaků jak nominálního, tak i ordinálního typu. Vycházíme přitom převážně pouze z absolutních četností nastoupení náhodných jevů, které odpovídají uvažovaným kategoriálním znakům. Byla proto vypracována řada efektivních metod tzv. ***kategoriální analýzy*** pro aplikace v různých oblastech: sociologie, marketing, psychologie, medicína, pedagogika apod. Tyto metody jsou povětšinou implementovány do profesionálního statistického softwaru, neboť při statistických šetřeních dostáváme v současné době velmi rozsáhlé databázové soubory, pro něž není „ruční“ zpracování únosné. V této kapitole je pouze nepatrný nástin těchto metod a více můžeme nalézt v [2], [3], [22], [28].

Pearsonův test nezávislosti a homogenity

Mějme náhodný vektor (X,Y) s konečným diskrétním sdruženým rozdělením pravděpodobnosti, přičemž náhodná veličina X nabývá hodnot $i = 1,...,r$ a náhodná veličina Y hodnot $j = 1,...,c$, kde $r \geq 2$ a $c \geq 2$. Předpokládejme, že se uskutečnil náhodný výběr o rozsahu $n \geq 4$ z (X,Y) a n_{ij} je počet případů, kdy se ve výběru vyskytla dvojice (i,j). Matice absolutních četností n_{ij} má pak multinomické rozdělení pravděpodobnosti s parametrem n a s pravděpodobnostmi p_{ij}. Pozorované hodnoty n_{ij} zapisujeme do tzv. ***kontingenční tabulky***:

X	Y			
	1	…	c	Σ
1	n_{11}	⋯	n_{1c}	$n_{1\bullet}$
…	…	…	…	…
r	n_{r1}	⋯	n_{rc}	$n_{r\bullet}$
Σ	$n_{\bullet 1}$	⋯	$n_{\bullet c}$	n

kde $n_{i\bullet} = \sum_{j=1}^{c} n_{ij}$, $n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$ jsou marginální četnosti a platí $n = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$.

Test nezávislosti X a Y je ekvivalentní testu sdružené hypotézy $H: p_{ij} = p_{i\bullet}p_{\bullet j}$ pro všechny dvojice (i, j), kde $p_{i\bullet} = \sum_{j=1}^{c} p_{ij}$ a $p_{\bullet j} = \sum_{i=1}^{r} p_{ij}$ jsou tzv. ***marginální pravděpodobnosti*** složek X a Y náhodného vektoru (X, Y). Hypotézu H testujeme pomocí Pearsonova testového kritéria

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} = n\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - n\,.$$

Hypotézu H nezamítáme na hladině významnosti α, jestliže $\chi^2 \in \overline{W}_\alpha = \langle 0; \chi^2_{1-\alpha} \rangle$, kde $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil Pearsonova (chí-kvadrát) rozdělení s $k = (r-1)(c-1)$ stupni volnosti – viz tabulku **T3**. Test je asymptotický a obvykle požadujeme, aby pro všechny dvojice (i, j) bylo $\frac{n_{i\bullet}n_{\bullet j}}{n} > 5$.

Uvedený test lze také použít k tzv. ***testu homogenity***, kdy testujeme hypotézu, že pozorované četnosti ve všech řádcích kontingenční tabulky mají multinomická rozdělení pravděpodobnosti s parametry $n_{i\bullet}$ a se stejnými pravděpodobnostmi $q_j = p_{1j} = \cdots = p_{rj}$, $j = 1, ..., c$. Místo řádků můžeme se stejným výsledkem testu zapsat pozorované četnosti do sloupců kontingenční tabulky.

Jestliže $r = 2$ a $c = 2$, jde o tzv. ***čtyřpolní tabulku*** pro ***alternativní*** (***dichotomické***) statistické znaky X a Y (např. pro odpovědi respondentů „ano“ anebo „ne“). Pro dostatečně velké četnosti n_{ij} můžeme opět použít Pearsonův test nezávislosti X a Y s výše uvedeným testovým kritériem anebo ve tvaru

$$\chi^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}\,.$$

Počet stupňů volnosti je $k = 1$. Tento test lze také aplikovat při testování hypotézy o rovnosti parametrů dvou binomických rozdělení místo testu uvedeného v kapitole 3.

Příklad 6.1

Průzkumem bylo zjištěno hodnocení televizního seriálu u jednotlivých skupin televizních diváků. Hodnocení mělo škálu: výborný, velmi dobrý, dobrý, špatný. Byly zvoleny skupiny diváků podle jejich nejvyššího dosaženého vzdělání: ZŠ, SŠ, VŠ. Pomocí Pearsonova testu

posuďte na hladině významnosti 0,05 závislost hodnocení televizního seriálu a nejvyššího dosaženého vzdělání televizních diváků, jestliže byly získány četnosti:

Vzdělání	Hodnocení televizního seriálu			
	Výborný	Velmi dobrý	Dobrý	Špatný
ZŠ	9	15	4	4
SŠ	6	11	14	5
VŠ	5	7	10	13

Ř e š e n í:

Pomocné výpočty byly „ručně“ provedeny v Excelu a jsou uvedeny v následující tabulce, kde vnitřní buňky (ve 2. až 4. řádku a 2. až 5. sloupci) obsahují tyto hodnoty:

$$n_{ij}$$

$$\frac{n_{i\bullet}n_{\bullet j}}{n}$$

$$n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}$$

$$\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2 \Big/ \frac{n_{i\bullet}n_{\bullet j}}{n}$$

Y / X	Výborný	Velmi dobrý	Dobrý	Špatný	Σ
ZŠ	9 6,21359223 2,78640777 1,24952973	15 10,2524272 4,74757282 2,19844991	4 8,69902913 -4,6990291 2,53831484	4 6,83495146 -2,8349515 1,17586055	32 32 0 7,162155
SŠ	6 6,99029126 -0,9902913 0,14029126	11 11,5339806 -0,5339806 0,02472132	14 9,78640777 4,21359223 1,81418554	5 7,68932039 -2,6893204 0,94058301	36 36 0 2,9197811
VŠ	5 6,7961165 -1,7961165 0,47468793	7 11,2135922 -4,2135922 1,5832892	10 9,51456311 0,48543689 0,02476719	13 7,47572816 5,52427184 4,08222166	35 35 0 6,164966
Σ	20	33	28	22	103
χ^2					16,246902

Z tabulky vidíme, že hodnota testového kritéria je

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} \doteq 16,247\,.$$

Počet stupňů volnosti $k = (3-1)(4-1) = 6$ a kritická hodnota pro hladinu významnosti 0,05, tj. 0,95-kvantil chí-kvadrát rozdělení se 6 stupni volnosti, je z tabulky **T3** $\chi^2_{0,95} = 12,592$, takže na této hladině významnosti hypotézu o nezávislosti zamítáme. Pro významnost 0,01 je však kritická hodnota 16,812, takže na této hladině významnosti hypotézu o nezávislosti nezamítáme.

Příklady k procvičení

Příklad 6.2

Celkem 180 náhodně vybraných matek bylo dotázáno, zda jejich kojenec dostává dudlík. Zjišťoval se též nejvyšší stupeň dosaženého vzdělání matky. Zjištěné četnosti jsou v tabulce:

Vzdělání matky	Počet matek	Počet dětí s dudlíkem
Základní	39	27
Středoškolské	47	34
Vysokoškolské	18	15

Na hladině významnosti 0,05 testujte hypotézu, že podíly dětí s dudlíkem nezávisí na vzdělání matky.

V ý s l e d e k: $k = 2$, $\chi^2 = 0,19 \in \overline{W}_{0,95} = \langle 0; 5,992 \rangle$; hypotézu o nezávislosti na vzdělání nezamítáme

Příklad 6.3

Na soukromou vysokou školu bylo přijato 142 studentů. Ti byli náhodně rozděleni do skupin A, B, C, D. V každé skupině předmět M vyučován jinou metodou. Na konci semestru roku psali všichni studenti stejnou písemnou práci a byly zaznamenány počty studentů z jednotlivých skupin, kteří vyřešili všechny zadané úkoly:

Skupina	A	B	C	D
Počet studentů	35	36	37	34
Počet úspěšných studentů	9	12	27	32

Na hladině významnosti 0,05 testujte hypotézu, že rozdíly mezi skupinami jsou způsobeny pouze náhodnými vlivy.

V ý s l e d e k: $k = 3$, $\chi^2 = 12,66 \notin \overline{W}_{0,95} = \langle 0; 7,815 \rangle$; hypotézu zamítáme, metoda výuky ovlivnila výsledky (šlo také současně o test homogenity rozdělení pravděpodobnosti odpovídajících řádkům tabulky)

Příklad 6.3

Průzkumem byl zjišťován zájem mezi potenciálními zákazníky o nový typ mobilu s kamerou. Výsledky průzkumu u 140 respondentů jsou po roztřídění podle zájmu a toho, zda dotázaný je či není majitel mobilu, v tabulce:

Majitel mobilu	Zájem	
	Ano	Ne
Je	49	25
Není	30	36

Testujte, zda zájem o nový typ mobilu závisí na tom, zda zákazník již mobil vlastní.

V ý s l e d e k: $\chi^2 = 6,12 \notin \overline{W}_{0,95} = \langle 0; 3,841 \rangle$; na hladině významnosti 0,05 hypotézu o závislosti zájmu na vlastnictví mobilu zamítáme,

$\chi^2 = 6,12 \notin \overline{W}_{0,99} = \langle 0; 6,654 \rangle$; na hladině významnosti 0,01 hypotézu o závislosti zájmu na vlastnictví mobilu nezamítáme,

(pro snížení pravděpodobnosti chyby druhého druhu by bylo vhodné zvýšit rozsah výběru a testovat hypotézu znovu)

Kontrolní otázky

1. Popište motivaci a princip kategoriální analýzy na konkrétním příkladu ze svého okolí.
2. Co je kontingenční tabulka a jak se konstruuje?
3. Jaká omezení má Pearsonův test nezávislosti?
4. Co rozumíme testem homogenity?
5. Kdy se používá čtyřpolní tabulka?

LITERATURA

Učebnice a monografie

1. Aczel, A. D. *Complete Business Statistics*. Chicago : IRWIN, 1989.
2. Anděl, J. *Matematická statistika*. 1. vyd. Praha : SNTL/ALFA, 1978.
3. Anděl, J. *Statistické metody.* 1. vyd. Praha : MATFYZPRESS, 1993.
4. Bowerman, B. L. - O´Connell, R. T. *Applied Statistics - Improving Business Processes*. Chicago : IRWIN, 1997.
5. Cyhelský, L. - Kahounová, J. - Hindls, R. *Elementární statistická analýza.* 1. vyd. Praha : Management Press, 1996.
6. Dowdy, S. - Wearden, S. *Statistics for Research.* New York : John Wiley & Sons, Inc., 1983.
7. Hahn, G. J. - Shapiro, S. S. *Statistical Models in Engineering*. New York : John Wiley & Sons, Inc., 1994.
8. Hátle, J. - Likeš, J. *Základy počtu pravděpodobnosti a matematické statistiky*. 1. vyd. Praha : SNTL/ALFA, 1974.
9. Hebák, P. - Hustopecký, J. *Vícerozměrné statistické metody.* 1. vyd. Praha : SNTL/ALFA, 1987.
10. Hebák, P. - Hustopecký, J. *Průvodce moderními statistickými metodami*. 1. vyd. Praha : SNTL, 1990.
11. Chatterjee, S. - Price, B. *Regression Analysis by Example*. New York : John Wiley & Sons, Inc., 1991.
12. Kupka, K. *Statistické řízení jakosti*. 1. vyd. Pardubice : TriloByte, 1997.
13. Lamoš, F. - Potocký, R. *Pravdepodobnosť a matematická štatistika*. 1. vyd. Bratislava : ALFA, 1989.
14. Likeš, J. - Machek, J. *Počet pravděpodobnosti*. 1. vyd. Praha : SNTL, 1981.
15. Likeš, J. - Machek, J. *Matematická statistika*. 1. vyd. Praha : SNTL, 1983.
16. Meloun, M. - Militký, J. *Statistické zpracování experimentálních dat*. 1. vyd. Praha : PLUS, 1994.
17. Montgomery, D. C. - Renger, G. *Probability and Statistics*. New York : John Wiley & Sons, Inc., 1996.
18. Potocký, R. et. al. *Zbierka úloh z pravdepodobnosti a matematickej štatistiky.* 1. vyd. Bratislava : ALFA/SNTL, 1986.
19. Rao, C. R. *Lineární metody statistické indukce a jejich aplikace.* Praha : Academia, 1978.
20. Rényi, A. *Teorie pravděpodobnosti.* 1. vyd. Praha : Academia, 1972.
21. Ryan, T. P.: *Modern Regression Methods*. New York : John Wiley & Sons, Inc., 1997.
22. Seger, J. - Hindls, R. *Statistické metody v tržním hospodářství*. 1. vyd. Praha : Victoria Publishing, 1995.
23. Swoboda, H. *Moderní statistika*. 1. vyd. Praha : Svoboda, 1977.

24. Štěpán, J. *Teorie pravděpodobnosti*. 1. vyd. Praha : Academia, 1987.

25. Šťastný, Z. *Matematické a statistické výpočty v Excelu.* 1. vyd. Brno : Computer Press, 1999.

26. Sprinthall, R. C. *Basic Statistical Analysis*. 5th ed. Boston : Allyn and Bacon, 1997.

27. Triola, M. F. *Elementary Statistics*. Redwood City : B/C Publishing Comp., 1989.

28. Wonnacot, T. H. - Wonnacot, R. J. *Statistika pro obchod a hospodářství*. 1. vyd. Praha : Victoria Publishing, 1993.

29. Zvára, K. *Regresní analýza*. 1. vyd. Praha : Academia, 1989.

30. Zvára, K. - Štěpán, J. *Pravděpodobnost a matematická statistika*. 1. vyd. Praha : MATFYZPRESS, 1997.

31. Plesník, J. – Dupačová, J. – Vlach, M. *Lineárne programovanie*. Bratislava : Alfa, 1990.

32. Taha, H. A. *Operations Research: An Introduction*. 8th r. e. New York : Macmillan, 2006.

Učební texty

33. Budíková, M. - Mikoláš, Š. - Osecký, P. *Teorie pravděpodobnosti a matematická statistika - Sbírka příkladů.* 1. vyd. Brno : MU, 1996.

34. Jarošová, E. *Statistika B - Řešené příklady*. 1. vyd. Praha : VŠE, 1994.

35. Karpíšek, Z. *Pravděpodobnostní metody*. 6. vyd. Brno : FP VUT u vydavatele Ing. Zdeněk Novotný, CSc., 2003.

36. Karpíšek, Z. - Drdla, M. *Statistické metody*. 7. vyd. Brno : FP VUT u vydavatele Ing. Zdeněk Novotný, CSc., 2003.

37. Karpíšek, Z. - Drdla, M. *Applied Statistics*. 1. vyd. Brno : FP VUT v PC - DIR, 1999.

38. Karpíšek, Z. - Drdla, M. *Aplikovaná statistika*. 2. vyd. Brno : BIBS, 2003.

39. Karpíšek, Z. – Popela, P. – Bednář, J. *Statistika a pravděpodobnost. Učební pomůcka - studijní opora pro kombinované studium.* FSI VUT v CERM Brno, Brno 2002.

40. Koutková, H. - Moll, I. *Úvod do pravděpodobnosti a matematické statistiky.* 1. vyd. Brno : ES VUT, 1990.

41. Kropáč, J. *Úvod do počtu pravděpodobnost a matematické statistiky.* 1. vyd. Brno : VA, 2000.

42. Likeš, J. - Cyhelský, L. - Hindls, R. *Úvod do statistiky a pravděpodobnosti - Statistika A*. 1. vyd. Praha : VŠE, 1995.

43. Michálek, J. *Matematická statistika pro informatiky.* 1. vyd. Praha : SPN, 1987.

44. Reif, J. *Metody matematické statistiky*. 1. vyd. Plzeň : Západočeská univerzita, 2000.

45. Seberová, H. *Statistika I, II*. 1. vyd. Vyškov : VVŠ PV, 1995.

46. Šikulová, M. - Karpíšek, Z. *Matematika IV - Pravděpodobnost a matematická statistika.* 6. vyd. Brno : ES VUT, 1996.

47. Zapletal, J. *Základy počtu pravděpodobnosti a matematické statistiky*. 1. vyd. Brno : ES VUT, 1995.

48. Klapka, J. – Dvořák, J. – Popela, P. *Metody operačního výzkumu.* Brno: PC-DIR, 1996.

WWW odkazy

49. http://badame.vse.cz/
50. http://davidmlane.com/hyperstat/
51. http://home.zcu.cz/~friesl/Vyuka/Odkazy.html
52. http://math.uc.edu/~brycw/classes/147/blue/tools.htm#texts
53. http://www.graphpad.com/welcome.htm
54. http://www.math.csusb.edu/faculty/stanton/m262/index.html
55. http://www.md-stat.com/
56. http://www.psychstat.smsu.edu/sbk00.htm
57. http://www.ruf.rice.edu/~lane/rvls.html
58. http://www.stat.sc.edu/rsrch/gasp/
59. http://www.statsoft.com/textbook/stathome.html
60. http://www.statsoft.cz/
61. http://www.trilobyte.cz/
62. http://www.fme.vutbr.cz/opory/
63. http://home.eunet.cz/berka/o/
64. http://www.mujweb.cz/www/januska/
65. http://www.fm.vslib.cz/~ksi/cz/mater/oa/linprog

STATISTICKÉ TABULKY

T1 Hodnoty distribuční funkce $\Phi(u)$ normovaného normálního rozdělení N(0;1)

u	0	1	2	3	4	5	6	7	8	9
0,0	0,50000	50399	50798	51197	51596	51994	52392	52791	53188	53586
0,1	53983	54380	54776	55172	55567	55962	56356	56750	57143	57535
0,2	57926	58317	58707	59096	59484	59871	60257	60642	61026	61409
0,3	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0,4	65542	65910	66276	66640	67003	67365	67724	68082	68439	68793
0,5	69146	69498	69847	70195	70540	70884	71226	71566	71904	72241
0,6	72575	72907	73237	73565	73892	74216	74537	74857	75175	75490
0,7	75804	76115	76424	76731	77035	77337	77637	77935	78231	78524
0,8	78815	79103	79389	79673	79955	80234	80511	80785	81057	81327
0,9	81594	81859	82121	82382	82639	82894	83147	83398	83646	83891
1,0	84135	84375	84614	84850	85083	85314	85543	85769	85993	86214
1,1	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1,2	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1,3	90320	90490	90658	90824	90988	91149	91309	91466	91621	91774
1,4	91924	92073	92220	92364	92507	92647	92786	92922	93056	93189
1,5	93319	93448	93574	93699	93822	93943	94062	94179	94295	94408
1,6	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1,7	95543	95637	95728	95819	95907	95994	96080	96164	96246	96327
1,8	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1,9	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2,0	97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2,1	98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2,2	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2,3	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2,4	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2,5	99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2,6	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2,7	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2,8	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2,9	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3,0	99865	99869	99874	99878	99882	99886	99889	99893	99896	99900
3,1	99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3,2	99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3,3	99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3,4	99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3,5	99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3,6	99984	99985	99985	99986	99986	99987	99987	99988	99988	99989
3,7	99989	99990	99990	99990	99991	99991	99992	99992	99992	99992
3,8	99993	99993	99993	99994	99994	99994	99994	99995	99995	99995
3,9	99995	99995	99996	99996	99996	99996	99996	99996	99997	99997

4,00	99997	4,10	99998	4,20	99999	4,30	99999	4,40	99999	4,50	99999

Poznámka: $\Phi(-u) = 1 - \Phi(u)$; $u_{0,95} \approx 1{,}645$; $u_{0,975} \approx 1{,}960$; $u_{0,99} \approx 2{,}326$; $u_{0,995} \approx 2{,}576$.

T2 Kvantily t_P Studentova rozdělení S(k)

k \ P	0,95	0,975	0,99	0,995	0,999	0,9995
1	6,314	12,706	31,821	63,656	318,289	636,578
2	2,920	4,303	6,965	9,925	22,328	31,600
3	2,353	3,182	4,541	5,841	10,214	12,924
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	4,032	5,894	6,869
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,408
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,947	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	4,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,610	3,922
19	1,729	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,552	3,850
21	1,721	2,080	2,518	2,831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,505	3,792
23	1,714	2,069	2,500	2,807	3,485	3,768
24	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,689
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,660
30	1,697	2,042	2,457	2,750	3,385	3,646
35	1,690	2,030	2,438	2,724	3,340	3,591
40	1,684	2,021	2,423	2,704	3,307	3,551
45	1,679	2,014	2,412	2,690	3,281	3,520
50	1,676	2,009	2,403	2,678	3,261	3,496
60	1,671	2,000	2,390	2,660	3,232	3,460
70	1,667	1,994	2,381	2,648	3,211	3,435
80	1,664	1,990	2,374	2,639	3,195	3,416
90	1,662	1,987	2,368	2,632	3,183	3,402
100	1,660	1,984	2,364	2,626	3,174	3,390
120	1,658	1,980	2,358	2,617	3,160	3,373
140	1,656	1,977	2,353	2,611	3,149	3,361
160	1,654	1,975	2,350	2,607	3,142	3,352
180	1,653	1,973	2,347	2,603	3,136	3,345
200	1,653	1,972	2,345	2,601	3,131	3,340
300	1,650	1,968	2,339	2,592	3,118	3,323
500	1,648	1,965	2,334	2,586	3,107	3,310
1000	1,646	1,962	2,330	2,581	3,098	3,300
∞	1,645	1,960	2,326	2,576	3,090	3,290

Poznámka: Pro $0 \le P \le 0{,}5$ použijeme vztah $t_P = -t_{1-P}$.

T3 Kvantily χ^2_P Pearsonova rozdělení $\chi^2(k)$

k \ P	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
1	0,000	0,000	0,001	0,004	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	44,985	48,232	52,191	55,002
32	15,134	16,362	18,291	20,072	46,194	49,480	53,486	56,328
33	15,815	17,073	19,047	20,867	47,400	50,725	54,775	57,648
34	16,501	17,789	19,806	21,664	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	49,802	53,203	57,342	60,275
36	17,887	19,233	21,336	23,269	50,998	54,437	58,619	61,581
37	18,586	19,960	22,106	24,075	52,192	55,668	59,893	62,883
38	19,289	20,691	22,878	24,884	53,384	56,895	61,162	64,181
39	19,996	21,426	23,654	25,695	54,572	58,120	62,428	65,475
40	20,707	22,164	24,433	26,509	55,758	59,342	63,691	66,766
41	21,421	22,906	25,215	27,326	56,942	60,561	64,950	68,053
42	22,138	23,650	25,999	28,144	58,124	61,777	66,206	69,336
43	22,860	24,398	26,785	28,965	59,304	62,990	67,459	70,616
44	23,584	25,148	27,575	29,787	60,481	64,201	68,710	71,892
45	24,311	25,901	28,366	30,612	61,656	65,410	69,957	73,166

T3 Kvantily χ^2_P Pearsonova rozdělení $\chi^2(k)$

(pokračování)

k \ P	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
46	25,041	26,657	29,160	31,439	62,830	66,616	71,201	74,437
47	25,775	27,416	29,956	32,268	64,001	67,821	72,443	75,704
48	26,511	28,177	30,754	33,098	65,171	69,023	73,683	76,969
49	27,249	28,941	31,555	33,930	66,339	70,222	74,919	78,231
50	27,991	29,707	32,357	34,764	67,505	71,420	76,154	79,490
51	28,735	30,475	33,162	35,600	68,669	72,616	77,386	80,746
52	29,481	31,246	33,968	36,437	69,832	73,810	78,616	82,001
53	30,230	32,019	34,776	37,276	70,993	75,002	79,843	83,253
54	30,981	32,793	35,586	38,116	72,153	76,192	81,069	84,502
55	31,735	33,571	36,398	38,958	73,311	77,380	82,292	85,749
56	32,491	34,350	37,212	39,801	74,468	78,567	83,514	86,994
57	33,248	35,131	38,027	40,646	75,624	79,752	84,733	88,237
58	34,008	35,914	38,844	41,492	76,778	80,936	85,950	89,477
59	34,770	36,698	39,662	42,339	77,930	82,117	87,166	90,715
60	35,534	37,485	40,482	43,188	79,082	83,298	88,379	91,952
61	36,300	38,273	41,303	44,038	80,232	84,476	89,591	93,186
62	37,068	39,063	42,126	44,889	81,381	85,654	90,802	94,419
63	37,838	39,855	42,950	45,741	82,529	86,830	92,010	95,649
64	38,610	40,649	43,776	46,595	83,675	88,004	93,217	96,878
65	39,383	41,444	44,603	47,450	84,821	89,177	94,422	98,105
66	40,158	42,240	45,431	48,305	85,965	90,349	95,626	99,330
67	40,935	43,038	46,261	49,162	87,108	91,519	96,828	100,554
68	41,714	43,838	47,092	50,020	88,250	92,688	98,028	101,776
69	42,493	44,639	47,924	50,879	89,391	93,856	99,227	102,996
70	43,275	45,442	48,758	51,739	90,531	95,023	100,425	104,215
71	44,058	46,246	49,592	52,600	91,670	96,189	101,621	105,432
72	44,843	47,051	50,428	53,462	92,808	97,353	102,816	106,647
73	45,629	47,858	51,265	54,325	93,945	98,516	104,010	107,862
74	46,417	48,666	52,103	55,189	95,081	99,678	105,202	109,074
75	47,206	49,475	52,942	56,054	96,217	100,839	106,393	110,285
80	51,172	53,540	57,153	60,391	101,879	106,629	112,329	116,321
85	55,170	57,634	61,389	64,749	107,522	112,393	118,236	122,324
90	59,196	61,754	65,647	69,126	113,145	118,136	124,116	128,299
95	63,250	65,898	69,925	73,520	118,752	123,858	129,973	134,247
100	67,328	70,065	74,222	77,929	124,342	129,561	135,807	140,170
110	75,550	78,458	82,867	86,792	135,480	140,916	147,414	151,948
120	83,852	86,923	91,573	95,705	146,567	152,211	158,950	163,648
130	92,223	95,451	100,331	104,662	157,610	163,453	170,423	175,278
150	109,142	112,668	117,985	122,692	179,581	185,800	193,207	198,360
200	152,241	156,432	162,728	168,279	233,994	241,058	249,445	255,264
500	422,303	429,387	439,936	449,147	553,127	563,851	576,493	585,206
1000	888,563	898,912	914,257	927,594	1074,68	1089,53	1106,97	1118,95

T4 Kvantily F_P Fisherova – Snedecorova rozdělení F(k_1,k_2) pro P = 0,975

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10
1	647,793	799,482	864,151	899,599	921,835	937,114	948,203	956,643	963,279	968,634
2	38,506	39,000	39,166	39,248	39,298	39,331	39,356	39,373	39,387	39,398
3	17,443	16,044	15,439	15,101	14,885	14,735	14,624	14,540	14,473	14,419
4	12,218	10,649	9,979	9,604	9,364	9,197	9,074	8,980	8,905	8,844
5	10,007	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681	6,619
6	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,523	5,461
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823	4,761
8	7,571	6,059	5,416	5,053	4,817	4,652	4,529	4,433	4,357	4,295
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026	3,964
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779	3,717
11	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,588	3,526
12	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,436	3,374
13	6,414	4,965	4,347	3,996	3,767	3,604	3,483	3,388	3,312	3,250
14	6,298	4,857	4,242	3,892	3,663	3,501	3,380	3,285	3,209	3,147
15	6,200	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,123	3,060
16	6,115	4,687	4,077	3,729	3,502	3,341	3,219	3,125	3,049	2,986
17	6,042	4,619	4,011	3,665	3,438	3,277	3,156	3,061	2,985	2,922
18	5,978	4,560	3,954	3,608	3,382	3,221	3,100	3,005	2,929	2,866
19	5,922	4,508	3,903	3,559	3,333	3,172	3,051	2,956	2,880	2,817
20	5,871	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,837	2,774
21	5,827	4,420	3,819	3,475	3,250	3,090	2,969	2,874	2,798	2,735
22	5,786	4,383	3,783	3,440	3,215	3,055	2,934	2,839	2,763	2,700
23	5,750	4,349	3,750	3,408	3,183	3,023	2,902	2,808	2,731	2,668
24	5,717	4,319	3,721	3,379	3,155	2,995	2,874	2,779	2,703	2,640
25	5,686	4,291	3,694	3,353	3,129	2,969	2,848	2,753	2,677	2,613
26	5,659	4,265	3,670	3,329	3,105	2,945	2,824	2,729	2,653	2,590
27	5,633	4,242	3,647	3,307	3,083	2,923	2,802	2,707	2,631	2,568
28	5,610	4,221	3,626	3,286	3,063	2,903	2,782	2,687	2,611	2,547
29	5,588	4,201	3,607	3,267	3,044	2,884	2,763	2,669	2,592	2,529
30	5,568	4,182	3,589	3,250	3,026	2,867	2,746	2,651	2,575	2,511
35	5,485	4,106	3,517	3,179	2,956	2,796	2,676	2,581	2,504	2,440
40	5,424	4,051	3,463	3,126	2,904	2,744	2,624	2,529	2,452	2,388
45	5,377	4,009	3,422	3,086	2,864	2,705	2,584	2,489	2,412	2,348
50	5,340	3,975	3,390	3,054	2,833	2,674	2,553	2,458	2,381	2,317
55	5,310	3,948	3,364	3,029	2,807	2,648	2,528	2,433	2,355	2,291
60	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,334	2,270
70	5,247	3,890	3,309	2,975	2,754	2,595	2,474	2,379	2,302	2,237
80	5,218	3,864	3,284	2,950	2,730	2,571	2,450	2,355	2,277	2,213
90	5,196	3,844	3,265	2,932	2,711	2,552	2,432	2,336	2,259	2,194
100	5,179	3,828	3,250	2,917	2,696	2,537	2,417	2,321	2,244	2,179
120	5,152	3,805	3,227	2,894	2,674	2,515	2,395	2,299	2,222	2,157
150	5,126	3,781	3,204	2,872	2,652	2,494	2,373	2,278	2,200	2,135
250	5,085	3,744	3,169	2,837	2,618	2,459	2,338	2,243	2,165	2,100
500	5,054	3,716	3,142	2,811	2,592	2,434	2,313	2,217	2,139	2,074
∞	5,024	3,689	3,116	2,786	2,566	2,408	2,288	2,192	2,114	2,048

T4 Kvantily F_P Fisherova – Snedecorova rozdělení F(k_1,k_2) pro P = 0,975 (pokračování)

k_2 \ k_1	12	15	20	24	30	40	60	100	250	∞
1	976,725	984,874	993,081	997,272	1001,40	1005,60	1009,79	1013,16	1016,22	1018,26
2	39,415	39,431	39,448	39,457	39,465	39,473	39,481	39,488	39,494	39,498
3	14,337	14,253	14,167	14,124	14,081	14,036	13,992	13,956	13,924	13,902
4	8,751	8,657	8,560	8,511	8,461	8,411	8,360	8,319	8,282	8,257
5	6,525	6,428	6,329	6,278	6,227	6,175	6,123	6,080	6,041	6,015
6	5,366	5,269	5,168	5,117	5,065	5,012	4,959	4,915	4,876	4,849
7	4,666	4,568	4,467	4,415	4,362	4,309	4,254	4,210	4,170	4,142
8	4,200	4,101	3,999	3,947	3,894	3,840	3,784	3,739	3,698	3,670
9	3,868	3,769	3,667	3,614	3,560	3,505	3,449	3,403	3,361	3,333
10	3,621	3,522	3,419	3,365	3,311	3,255	3,198	3,152	3,109	3,080
11	3,430	3,330	3,226	3,173	3,118	3,061	3,004	2,956	2,912	2,883
12	3,277	3,177	3,073	3,019	2,963	2,906	2,848	2,800	2,755	2,725
13	3,153	3,053	2,948	2,893	2,837	2,780	2,720	2,671	2,626	2,595
14	3,050	2,949	2,844	2,789	2,732	2,674	2,614	2,565	2,519	2,487
15	2,963	2,862	2,756	2,701	2,644	2,585	2,524	2,474	2,427	2,395
16	2,889	2,788	2,681	2,625	2,568	2,509	2,447	2,396	2,349	2,316
17	2,825	2,723	2,616	2,560	2,502	2,442	2,380	2,329	2,280	2,247
18	2,769	2,667	2,559	2,503	2,445	2,384	2,321	2,269	2,220	2,187
19	2,720	2,617	2,509	2,452	2,394	2,333	2,270	2,217	2,167	2,133
20	2,676	2,573	2,464	2,408	2,349	2,287	2,223	2,170	2,120	2,085
21	2,637	2,534	2,425	2,368	2,308	2,246	2,182	2,128	2,077	2,042
22	2,602	2,498	2,389	2,332	2,272	2,210	2,145	2,090	2,039	2,003
23	2,570	2,466	2,357	2,299	2,239	2,176	2,111	2,056	2,004	1,968
24	2,541	2,437	2,327	2,269	2,209	2,146	2,080	2,024	1,972	1,935
25	2,515	2,411	2,300	2,242	2,182	2,118	2,052	1,996	1,942	1,906
26	2,491	2,387	2,276	2,217	2,157	2,093	2,026	1,969	1,915	1,878
27	2,469	2,364	2,253	2,195	2,133	2,069	2,002	1,945	1,891	1,853
28	2,448	2,344	2,232	2,174	2,112	2,048	1,980	1,922	1,867	1,829
29	2,430	2,325	2,213	2,154	2,092	2,028	1,959	1,901	1,846	1,807
30	2,412	2,307	2,195	2,136	2,074	2,009	1,940	1,882	1,826	1,787
35	2,341	2,235	2,122	2,062	1,999	1,932	1,861	1,801	1,743	1,702
40	2,288	2,182	2,068	2,007	1,943	1,875	1,803	1,741	1,680	1,637
45	2,248	2,141	2,026	1,965	1,900	1,831	1,757	1,694	1,631	1,586
50	2,216	2,109	1,993	1,931	1,866	1,796	1,721	1,656	1,592	1,545
55	2,190	2,083	1,967	1,904	1,838	1,768	1,692	1,625	1,559	1,511
60	2,169	2,061	1,944	1,882	1,815	1,744	1,667	1,599	1,532	1,482
70	2,136	2,028	1,910	1,847	1,779	1,707	1,628	1,558	1,488	1,436
80	2,111	2,003	1,884	1,820	1,752	1,679	1,599	1,527	1,455	1,400
90	2,092	1,983	1,864	1,800	1,731	1,657	1,576	1,503	1,428	1,371
100	2,077	1,968	1,849	1,784	1,715	1,640	1,558	1,483	1,407	1,347
120	2,055	1,945	1,825	1,760	1,690	1,614	1,530	1,454	1,374	1,310
150	2,032	1,922	1,801	1,736	1,665	1,588	1,502	1,423	1,340	1,271
250	1,997	1,886	1,764	1,697	1,625	1,546	1,457	1,374	1,282	1,201
500	1,971	1,859	1,736	1,669	1,596	1,515	1,423	1,336	1,235	1,137
∞	1,945	1,833	1,708	1,640	1,566	1,484	1,388	1,296	1,183	1,000

T4 Kvantily F_P Fisherova – Snedecorova rozdělení F(k_1,k_2) pro P = 0,995

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10
1	16212,5	19997,4	21614,1	22500,8	23055,8	23439,5	23715,2	23923,8	24091,5	24221,8
2	198,503	199,012	199,158	199,245	199,303	199,332	199,361	199,376	199,390	199,390
3	55,552	49,800	47,468	46,195	45,391	44,838	44,434	44,125	43,881	43,685
4	31,332	26,284	24,260	23,154	22,456	21,975	21,622	21,352	21,138	20,967
5	22,785	18,314	16,530	15,556	14,939	14,513	14,200	13,961	13,772	13,618
6	18,635	14,544	12,917	12,028	11,464	11,073	10,786	10,566	10,391	10,250
7	16,235	12,404	10,883	10,050	9,522	9,155	8,885	8,678	8,514	8,380
8	14,688	11,043	9,597	8,805	8,302	7,952	7,694	7,496	7,339	7,211
9	13,614	10,107	8,717	7,956	7,471	7,134	6,885	6,693	6,541	6,417
10	12,827	9,427	8,081	7,343	6,872	6,545	6,303	6,116	5,968	5,847
11	12,226	8,912	7,600	6,881	6,422	6,102	5,865	5,682	5,537	5,418
12	11,754	8,510	7,226	6,521	6,071	5,757	5,524	5,345	5,202	5,085
13	11,374	8,186	6,926	6,233	5,791	5,482	5,253	5,076	4,935	4,820
14	11,060	7,922	6,680	5,998	5,562	5,257	5,031	4,857	4,717	4,603
15	10,798	7,701	6,476	5,803	5,372	5,071	4,847	4,674	4,536	4,424
16	10,576	7,514	6,303	5,638	5,212	4,913	4,692	4,521	4,384	4,272
17	10,384	7,354	6,156	5,497	5,075	4,779	4,559	4,389	4,254	4,142
18	10,218	7,215	6,028	5,375	4,956	4,663	4,445	4,276	4,141	4,030
19	10,073	7,093	5,916	5,268	4,853	4,561	4,345	4,177	4,043	3,933
20	9,944	6,987	5,818	5,174	4,762	4,472	4,257	4,090	3,956	3,847
21	9,829	6,891	5,730	5,091	4,681	4,393	4,179	4,013	3,880	3,771
22	9,727	6,806	5,652	5,017	4,609	4,322	4,109	3,944	3,812	3,703
23	9,635	6,730	5,582	4,950	4,544	4,259	4,047	3,882	3,750	3,642
24	9,551	6,661	5,519	4,890	4,486	4,202	3,991	3,826	3,695	3,587
25	9,475	6,598	5,462	4,835	4,433	4,150	3,939	3,776	3,645	3,537
26	9,406	6,541	5,409	4,785	4,384	4,103	3,893	3,730	3,599	3,492
27	9,342	6,489	5,361	4,740	4,340	4,059	3,850	3,687	3,557	3,450
28	9,284	6,440	5,317	4,698	4,300	4,020	3,811	3,649	3,519	3,412
29	9,230	6,396	5,276	4,659	4,262	3,983	3,775	3,613	3,483	3,376
30	9,180	6,355	5,239	4,623	4,228	3,949	3,742	3,580	3,451	3,344
35	8,976	6,188	5,086	4,479	4,088	3,812	3,607	3,447	3,318	3,212
40	8,828	6,066	4,976	4,374	3,986	3,713	3,509	3,350	3,222	3,117
45	8,715	5,974	4,892	4,294	3,909	3,638	3,435	3,276	3,149	3,044
50	8,626	5,902	4,826	4,232	3,849	3,579	3,376	3,219	3,092	2,988
55	8,554	5,843	4,773	4,181	3,800	3,531	3,330	3,173	3,046	2,942
60	8,495	5,795	4,729	4,140	3,760	3,492	3,291	3,134	3,008	2,904
70	8,403	5,720	4,661	4,076	3,698	3,431	3,232	3,076	2,950	2,846
80	8,335	5,665	4,611	4,028	3,652	3,387	3,188	3,032	2,907	2,803
90	8,282	5,623	4,573	3,992	3,617	3,352	3,154	2,999	2,873	2,770
100	8,241	5,589	4,542	3,963	3,589	3,325	3,127	2,972	2,847	2,744
120	8,179	5,539	4,497	3,921	3,548	3,285	3,087	2,933	2,808	2,705
150	8,118	5,490	4,453	3,878	3,508	3,245	3,048	2,894	2,770	2,667
250	8,021	5,412	4,382	3,812	3,444	3,183	2,987	2,833	2,709	2,607
500	7,950	5,355	4,330	3,763	3,396	3,137	2,941	2,789	2,665	2,562
∞	7,879	5,298	4,279	3,715	3,350	3,091	2,897	2,744	2,621	2,519

T4 Kvantily F_P Fisherova – Snedecorova rozdělení F(k_1,k_2) pro P = 0,995 (pokračování)

k_2 \ k_1	12	15	20	24	30	40	60	100	250	∞
1	24426,7	24631,6	24836,5	24937,1	25041,4	25145,7	25253,7	25339,4	25413,9	25466,1
2	199,419	199,434	199,449	199,449	199,478	199,478	199,478	199,478	199,507	199,507
3	43,387	43,085	42,779	42,623	42,466	42,310	42,150	42,022	41,906	41,829
4	20,705	20,438	20,167	20,030	19,892	19,751	19,611	19,497	19,394	19,325
5	13,385	13,146	12,903	12,780	12,656	12,530	12,402	12,300	12,206	12,144
6	10,034	9,814	9,589	9,474	9,358	9,241	9,122	9,026	8,938	8,879
7	8,176	7,968	7,754	7,645	7,534	7,422	7,309	7,217	7,132	7,076
8	7,015	6,814	6,608	6,503	6,396	6,288	6,177	6,087	6,006	5,951
9	6,227	6,032	5,832	5,729	5,625	5,519	5,410	5,322	5,242	5,188
10	5,661	5,471	5,274	5,173	5,071	4,966	4,859	4,772	4,692	4,639
11	5,236	5,049	4,855	4,756	4,654	4,551	4,445	4,359	4,279	4,226
12	4,906	4,721	4,530	4,431	4,331	4,228	4,123	4,037	3,958	3,904
13	4,643	4,460	4,270	4,173	4,073	3,970	3,866	3,780	3,700	3,647
14	4,428	4,247	4,059	3,961	3,862	3,760	3,655	3,569	3,490	3,436
15	4,250	4,070	3,883	3,786	3,687	3,585	3,480	3,394	3,314	3,260
16	4,099	3,920	3,734	3,638	3,539	3,437	3,332	3,246	3,166	3,111
17	3,971	3,793	3,607	3,511	3,412	3,311	3,206	3,119	3,039	2,984
18	3,860	3,683	3,498	3,402	3,303	3,201	3,096	3,009	2,929	2,873
19	3,763	3,587	3,402	3,306	3,208	3,106	3,000	2,913	2,832	2,776
20	3,678	3,502	3,318	3,222	3,123	3,022	2,916	2,828	2,747	2,690
21	3,602	3,427	3,243	3,147	3,049	2,947	2,841	2,753	2,671	2,614
22	3,535	3,360	3,176	3,081	2,982	2,880	2,774	2,685	2,602	2,546
23	3,474	3,300	3,116	3,021	2,922	2,820	2,713	2,624	2,541	2,484
24	3,420	3,246	3,062	2,967	2,868	2,765	2,658	2,569	2,486	2,428
25	3,370	3,196	3,013	2,918	2,819	2,716	2,609	2,519	2,435	2,377
26	3,325	3,151	2,968	2,873	2,774	2,671	2,563	2,473	2,389	2,330
27	3,284	3,110	2,927	2,832	2,733	2,630	2,522	2,431	2,346	2,287
28	3,246	3,073	2,890	2,794	2,695	2,592	2,483	2,392	2,307	2,247
29	3,211	3,038	2,855	2,759	2,660	2,557	2,448	2,357	2,270	2,210
30	3,179	3,006	2,823	2,727	2,628	2,524	2,415	2,323	2,237	2,176
35	3,048	2,876	2,693	2,597	2,497	2,392	2,282	2,188	2,099	2,036
40	2,953	2,781	2,598	2,502	2,401	2,296	2,184	2,088	1,997	1,932
45	2,881	2,709	2,527	2,430	2,329	2,222	2,109	2,012	1,918	1,851
50	2,825	2,653	2,470	2,373	2,272	2,164	2,050	1,951	1,855	1,786
55	2,779	2,608	2,425	2,327	2,226	2,118	2,002	1,902	1,804	1,733
60	2,742	2,570	2,387	2,290	2,187	2,079	1,962	1,861	1,761	1,689
70	2,684	2,513	2,329	2,231	2,128	2,019	1,900	1,797	1,694	1,618
80	2,641	2,470	2,286	2,188	2,084	1,974	1,854	1,748	1,643	1,563
90	2,608	2,437	2,253	2,155	2,051	1,939	1,818	1,711	1,602	1,520
100	2,583	2,411	2,227	2,128	2,024	1,912	1,790	1,681	1,570	1,485
120	2,544	2,373	2,188	2,089	1,984	1,871	1,747	1,636	1,521	1,431
150	2,506	2,335	2,150	2,050	1,944	1,830	1,704	1,590	1,471	1,374
250	2,446	2,275	2,089	1,989	1,882	1,765	1,636	1,516	1,387	1,274
500	2,402	2,230	2,044	1,943	1,835	1,717	1,584	1,460	1,319	1,184
∞	2,358	2,187	2,000	1,898	1,789	1,669	1,533	1,402	1,245	1,000

T5 Kvantily w_P Wilcoxonova rozdělení (n = 5,…, 30)

n \ P	0,005	0,01	0,025	0,05
5	---	---	---	0
6	---	---	0	2
7	---	0	2	3
8	0	1	3	5
9	1	3	5	8
10	3	5	8	10
11	5	7	10	13
12	7	9	13	17
13	9	12	17	21
14	12	15	21	25
15	15	19	25	30
16	19	23	29	35
17	23	27	34	41
18	27	32	40	47
19	32	37	46	53
20	37	43	52	60
21	42	49	58	67
22	48	55	65	75
23	54	62	73	83
24	61	69	81	91
25	68	76	89	100
26	75	84	98	110
27	83	92	107	119
28	91	101	116	130
29	100	110	126	140
30	109	120	137	151

T6 Kvantily v_P Mannova – Whitneyova rozdělení pro $P = 0{,}025$ ($m = 2,\ldots, 20$; $n = 9,\ldots, 20$)

m \ n	9	10	11	12	13	14
2	0	0	0	1	1	1
3	2	3	3	4	4	5
4	4	5	6	7	8	9
5	7	8	9	11	12	13
6	10	11	13	14	16	17
7	12	14	16	18	20	22
8	15	17	19	22	24	26
9	17	20	23	26	28	31
10	20	23	26	29	33	36
11	23	26	30	33	37	40
12	26	29	33	37	41	45
13	28	33	37	41	45	50
14	31	36	40	45	50	55
15	34	39	44	49	54	59
16	37	42	47	53	59	64
17	39	45	51	57	63	67
18	42	48	55	61	67	74
19	45	52	58	65	72	78
20	48	55	62	69	76	83

m \ n	15	16	17	18	19	20
2	1	1	2	2	2	2
3	5	6	6	7	7	8
4	10	11	11	12	13	13
5	14	15	17	18	19	20
6	19	21	22	24	25	27
7	24	26	28	30	32	34
8	29	31	34	36	38	41
9	34	37	39	42	45	48
10	39	42	45	48	52	55
11	44	47	51	55	58	62
12	49	53	57	61	65	69
13	54	59	63	67	72	76
14	59	64	67	74	78	83
15	64	70	75	80	85	90
16	70	75	81	86	92	98
17	75	81	87	93	99	105
18	80	86	93	99	106	112
19	85	92	99	106	113	119
20	90	98	105	112	119	127

T7 Kvantily k_P binomického rozdělení Bi(n;0,5)

n \ P	0,005	0,01	0,025	0,05
6	-	-	0	0
7	-	0	0	0
8	0	0	0	1
9	0	0	1	1
10	0	0	1	1
11	0	1	1	2
12	1	1	2	2
13	1	1	2	3
14	1	2	2	3
15	2	2	2	3
16	2	2	3	4
17	2	3	4	4
18	3	3	4	5
19	3	4	4	5
20	3	4	5	5
21	4	4	5	6
22	4	5	5	6
23	4	5	6	7
24	5	5	6	7
25	5	6	7	7
26	6	6	7	8
27	6	7	7	8
28	6	7	8	9
29	7	7	8	9
30	7	8	9	10

DODATEK 1 – ZÁKLADY POPISNÉ STATISTIKY

1.1 Základní pojmy

Při statistickém zkoumání se zabýváme jevy a procesy, které mají hromadný charakter a vyskytují se u rozsáhlého souboru individuálních objektů (výrobky, osoby apod.), nazývaného ***základní soubor*** nebo také ***populace***. Zkoumané objekty jsou tzv. ***statistické jednotky*** a sledujeme u nich vytypované vlastnosti - ***statistické znaky*** (veličiny, parametry atd.), které nabývají pozorovatelných ***hodnot*** (***úrovní***).

Podle druhu hodnot dělíme statistické znaky na ***kvantitativní***, které nabývají číselných hodnot (hmotnost, délka, pevnost, cena, doba, životnost, ...) a ***kvalitativní***, které nemají číselný charakter a lze je vyjádřit slovně (barva, jakostní třída, podmínky provozu, tvar, ...). Sledujeme-li jen jeden znak, hovoříme o ***jednorozměrném*** znaku, naopak o ***vícerozměrném*** znaku.

Kvantitativní znaky dělíme na ***diskrétní***, jestliže nabývají pouze oddělených číselných hodnot (počet zmetků, počet vad, kusová produkce apod.) a ***spojité***, které nabývají všech hodnot z nějakého intervalu reálných čísel (rozměr výrobku, doba do poruchy, cenový index apod.).

Kvalitativní znaky dělíme na ***ordinální***, jejichž slovní hodnoty má smysl uspořádat (jakostní třídy, klasifikace apod.) a ***nominální***, jejichž slovní hodnoty postrádají význam pořadí (barva, tvar, dodavatelé apod.).

Podstatou statistických metod je, že informace o základním souboru nezjišťujeme u všech jeho jednotek, ale jen u některých, které získáme tzv. ***výběrem***. Vedou nás k tomu různá omezení, např. dosažitelnost všech jednotek, velký rozsah základního souboru, způsob získávání informací (zkoušky životnosti, ověření opotřebení atd.), náklady na statistické sledování a další. Počet vybraných jednotek je ***rozsah*** výběru. Dle rozsahu dělíme výběry na ***malé*** (obvykle do 30 až 50) a ***velké*** (řádově stovky, tisíce i více). Toto dělení je relativní a závisí na okolnostech statistického sledování. Výběr by měl být ***reprezentativní*** (poskytovat informace bez omezení) a ***homogenní*** (bez vlivu dalších různých faktorů). To však často nelze v plné míře verifikovatelně zajistit a proto obvykle vybíráme statistické jednotky do výběru ***náhodně***, ovšem s rizikem, že výběr může poskytnout více či méně zkreslené informace o základním souboru. Podle způsobu provedení rozlišujeme výběry:

- ***bez opakování*** (každá jednotka může být vybrána nejvýše jednou),
- ***s opakováním*** (každá jednotka může být vybrána vícekrát),

– ***záměrný*** (vybíráme typické jednotky),
– ***oblastní*** (základní soubor rozdělíme na podmnožiny a z nich provedeme části výběru),
– ***systematický*** nebo ***mechanický*** (vybíráme vždy několikátou jednotku co do pořadí při realizaci výběru).

Hodnoty znaku, pozorované či zjištěné na statistických jednotkách z výběru o rozsahu *n*, tvoří ***statistický soubor s rozsahem*** *n*. Pro jednorozměrný znak *X* získáme ***jednorozměrný statistický soubor*** $(x_1,...,x_n)$, kde x_i je ***pozorovaná hodnota*** znaku *X* u *i*–té statistické jednotky, $i = 1,...,n$. Analogicky pro dvourozměrný znak (*X*, *Y*) obdržíme ***dvourozměrný statistický soubor*** $((x_1, y_1),...,(x_n, y_n))$ apod.

1.2 Jednorozměrný statistický soubor s kvantitativním znakem

Získaný statistický soubor $(x_1,...,x_n)$ s rozsahem *n* se také nazývá ***neroztříděný statistický soubor***. Dle potřeby jej můžeme uspořádat podle rostoucích hodnot x_i a obdržíme ***uspořádaný statistický soubor*** $(x_{(1)},...,x_{(n)})$, kde $x_{(i)} \le x_{(i+1)}$ pro všechny indexy *i*. Interval $\langle x_{(1)}; x_{(n)} \rangle$ je ***variační obor*** a jeho délka $x_{(n)} - x_{(1)}$ je ***rozpětí*** statistického souboru.

Při velkém rozsahu statistického souboru nebo z důvodu dalšího zpracování (některá grafická vyjádření anebo užití matematicko - statistických metod) původní soubor ***roztřídíme***. ***Roztříděný statistický soubor*** získáme pokrytím variačního oboru systémem disjunktních intervalů (obvykle zleva otevřených a zprava uzavřených), tzv. ***tříd*** o počtu *m*, které mají obvykle stejnou ***délku*** *h*. Každá třída je reprezentována uspořádanou dvojicí (x_j^*, f_j), kde x_j^* je ***střed*** *j*-té třídy, $x_j^* < x_{j+1}^*$, a f_j je ***absolutní četnost*** *j*-té třídy, $j = 1,...,m$. Absolutní četnost f_j je počet prvků x_i původního neroztříděného statistického souboru, které leží v *j*-té třídě. Číslo $\frac{f_j}{n}$ je ***relativní četnost*** a uvádí se též v %. Platí $\sum_{j=1}^{m} f_j = n$.

Počet tříd *m* volíme obvykle přibližně $1 + 3,3 \log n$ (pro statistický soubor symetrického charakteru) anebo $\sqrt{n}$ až $2\sqrt{n}$ (pro statistický soubor asymetrického charakteru). ***Délka*** třídy je $h \approx \frac{x_{(n)} - x_{(1)}}{m}$ a stanovujeme ji tak, aby odpovídala přesnosti získání hodnot x_i a aby střed třídy x_j^* byl zaokrouhlené číslo. U diskrétního znaku volíme obvykle za středy tříd přímo hodnoty, kterých tento znak může nabývat. Pokud třídění

provádíme na PC, měli bychom zkontrolovat, zda nastavení parametrů *m*, resp. *h* použitého statistického software odpovídá našim požadavkům.

Číslo $F_j = \sum_{k=1}^{j} f_k$ je ***kumulativní absolutní četnost***, číslo $\frac{F_j}{n}$ je ***kumulativní relativní četnost***, $j = 1,...,m$, a uvádí se též v %. Platí, že $F_{j+1} = F_j + f_{j+1}$ pro $j = 1,...,m-1$, kde $F_1 = f_1$, takže $F_m = n$.

Roztříděný statistický soubor zapisujeme do tzv. ***četnostní tabulky*** pro různé typy četností, např. pro absolutní četnosti:

x_j^*	x_1^*	. . .	x_m^*
f_j	f_1	. . .	f_m

Významné vlastnosti statistického souboru vyjadřují v koncentrované formě jeho následující ***číselné*** (***empirické***) ***charakteristiky***. Jde zejména o ***charakteristiky polohy, proměnlivosti a souměrnosti***.

Základní ***charakteristiky polohy*** statistického souboru jsou:

1. ***Aritmetický průměr***

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{pro neroztříděný soubor,}$$

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{m} f_j x_j^* \qquad \text{pro roztříděný soubor.}$$

Vlastnosti aritmetického průměru jsou:

a) $y = ax + b \Rightarrow \overline{y} = a\overline{x} + b$ pro reálné konstanty *a*, *b*,

b) $\overline{x + y} = \overline{x} + \overline{y}$,

c) $x_{(1)} \le \overline{x} \le x_{(n)}$,

d) $\overline{x}$ má tentýž rozměr jako znak X.

Někdy se užívá též ***vážený aritmetický průměr***

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

kde $w_i \ge 0$ jsou ***váhy*** (vhodně stanovená reálná čísla, z nichž aspoň jedno je nenulové) hodnot x_i, které vyjadřují jejich význam, např. přesnost.

2. ***Medián*** pro neroztříděný statistický soubor

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{pro lichá } n\,, \\ \dfrac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right] & \text{pro sudá } n\,. \end{cases}$$

Vlastnosti mediánu:

a) $y = ax + b \Rightarrow \tilde{y} = a\tilde{x} + b$ pro reálné konstanty *a*, *b*,

b) $x_{(1)} \le \tilde{x} \le x_{(n)}$,

c) $\tilde{x}$ má tentýž rozměr jako znak X.

Medián rozděluje statistický soubor na "dolní polovinu" a "horní polovinu" hodnot x_i (viz obr. 1.1). Jde o ***robustní*** charakteristiku, která je oproti aritmetickému průměru málo citlivá na extrémně odchýlené hodnoty. Pro roztříděný soubor se k výpočtu mediánu užívá vhodná aproximace.

3. ***Modus*** $\hat{x}$ je číslo, v jehož okolí je nejvíce hodnot x_i, resp. je to střed x_j^* třídy s největší absolutní četností f_j. Modus má tytéž vlastnosti jako aritmetický průměr i medián a dle potřeby se počítá vhodnou aproximací (např. pro roztříděný soubor).

Základní ***charakteristiky proměnlivosti*** (***variability***) statistického souboru jsou:

1. ***Rozptyl*** (***disperze***, ***variance***)

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2 \qquad \text{pro neroztříděný soubor,}$$

$$s^2 = \frac{1}{n}\sum_{j=1}^{m} f_j \left(x_j^* - \overline{x}\right)^2 = \left(\frac{1}{n}\sum_{j=1}^{m} f_j x_j^{*2}\right) - \overline{x}^2 \quad \text{pro roztříděný soubor.}$$

Dle potřeby a také pro zdůraznění znaku *X* někdy píšeme $s^2(x)$ apod. Vlastnosti rozptylu jsou:

a) $s^2 \ge 0$,

b) $y = ax + b \Rightarrow s^2(y) = a^2 s^2(x)$ pro reálné konstanty *a*, *b*,

c) $s^2 = 0 \Leftrightarrow x_1 = \cdots = x_n$, resp. $x_1^* = \cdots = x_m^*$,

d) s^2 má rozměr rovný kvadrátu rozměru znaku X.

Větší proměnlivosti znaku *X* odpovídá větší rozptyl a naopak. Při výpočtech se také užívá jiný

vzorec pro rozptyl, když výraz $\frac{1}{n}$ zaměníme výrazem $\frac{1}{n-1}$. Takto vypočtený rozptyl je roven číslu $\frac{n}{n-1}s^2 > s^2$ (pro $s^2 \neq 0$). Zdůvodnění výrazu $\frac{1}{n-1}$ plyne z požadavků uvedených v kapitole 6 a 7.

2. ***Směrodatná odchylka*** $s = \sqrt{s^2}$.

Dle potřeby také píšeme $s(x)$. Vlastnosti směrodatné odchylky jsou:

a) $s \geq 0$,

b) $y = ax + b \Rightarrow s(y) = |a|\, s(x)$ pro reálné konstanty a, b,

c) $s = 0 \Leftrightarrow x_1 = \cdots = x_n$, resp. $x_1^* = \cdots = x_m^*$

d) s má tentýž rozměr jako znak X.

Větší proměnlivosti znaku X odpovídá větší směrodatná odchylka a naopak.

3. ***Variační koeficient*** $v = \frac{s}{\overline{x}}$.

Dle potřeby také píšeme $v(x)$. Vlastnosti variačního koeficientu jsou:

a) $v(ax) = \frac{a}{|a|}v(x)$ pro reálnou konstantu $a \neq 0$,

b) v je bezrozměrné číslo.

Jde o relativní míru variability znaku X a uvádí se též v %. Má smysl pouze pro znak X, který nabývá pouze kladných anebo záporných hodnot. Není proto např. vhodný pro znak X vyjadřující odchylky od nějaké nominální hodnoty.

4. ***Rozpětí*** $x_{(n)} - x_{(1)}$. Rozpětí má stejné vlastnosti jako směrodatná odchylka.

Základní ***charakteristikou souměrnosti*** statistického souboru je ***koeficient šikmosti*** (***koeficient asymetrie***)

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{s^3} \qquad \text{pro neroztříděný soubor,}$$

$$A = \frac{\frac{1}{n}\sum_{j=1}^{m} f_j\left(x_j^* - \overline{x}\right)^3}{s^3} \qquad \text{pro roztříděný soubor.}$$

Dle potřeby také píšeme $A(x)$. Vlastnosti koeficientu šikmosti jsou:

a) $A > 0 \Leftrightarrow$ většina hodnot x_i je menší než (leží pod) $\overline{x}$,

b) $A = 0 \Leftrightarrow$ hodnoty x_i jsou rozloženy souměrně vzhledem k $\overline{x}$,

c) $A < 0 \Leftrightarrow$ většina hodnot x_i je větší než (leží nad) $\overline{x}$,

d) $y = ax + b \Rightarrow A(y) = \dfrac{a}{|a|} A(x)$ pro reálné konstanty a, b, $a \neq 0$,

e) A je bezrozměrné číslo.

Existuje řada dalších číselných charakteristik statistického souboru. Např. pro poměrové znaky (cenové a objemové indexy, úrokové míry apod.) se místo aritmetického průměru užívá ***geometrický průměr***

$$\overline{x}_g = \sqrt[n]{x_1 \ldots x_n}$$

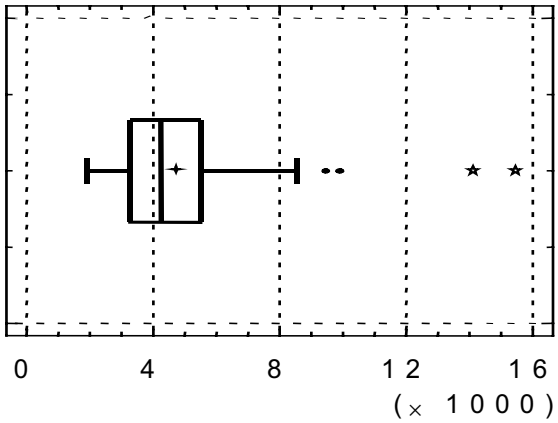
a ve speciálních případech (např. pro znaky vyjadřující rychlost nějakého děje) počítáme ***harmonický průměr***

$$\overline{x}_h = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}.$$

Dle potřeby se také někdy počítá ***koeficient špičatosti*** (***koeficient excesu***)

$$\frac{\dfrac{1}{n} \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^4}{s^4} - 3,$$

který vyjadřuje specifickým způsobem míru koncentrace hodnot statistického souboru.

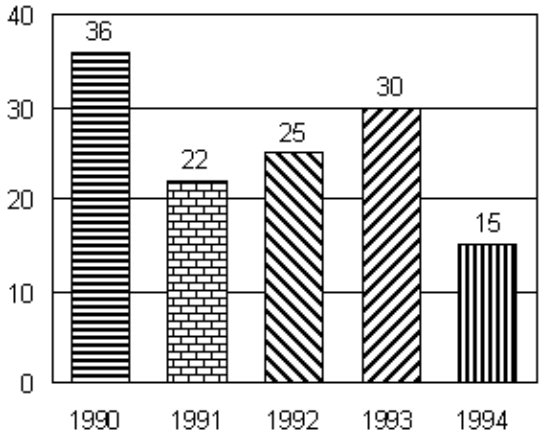


Obr. 1.1

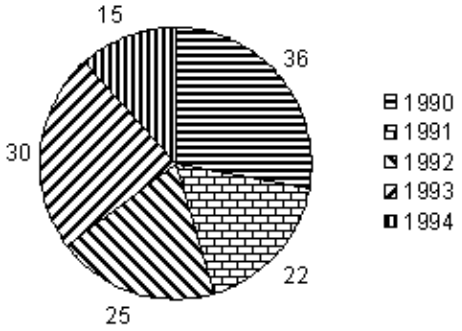
Mnoho rychlých a cenných informací poskytují o statistických souborech jejich ***grafická vyjádření***. Pro jednorozměrný neroztříděný resp. uspořádaný statistický soubor se zejména užívá ***krabicový graf*** - obr. 1.1, kde tučně vyznačený obdélník obsahuje střední část uspořádaného souboru (cca polovinu všech jeho hodnot) tak, že nalevo a napravo od

obdélníku leží vždy cca čtvrtina hodnot uspořádaného souboru. Levá (pravá) svislá strana obdélníku odpovídá tzv. ***dolnímu*** (***hornímu***) ***kvartilu*** statistického souboru a svislá čára uvnitř je v místě ***mediánu***. Výška obdélníku je úměrná rozsahu souboru a úsečky ("vousy") vlevo a vpravo zakončené krátkými svislými čarami vyjadřují přijatelné obory pro zbývající dolní a horní čtvrtinu souboru. Hodnoty mimo tyto úsečky jsou považovány za podezřelé, případně extrémně odchýlené. Existují další modifikace tohoto grafu a jiná vyjádření.

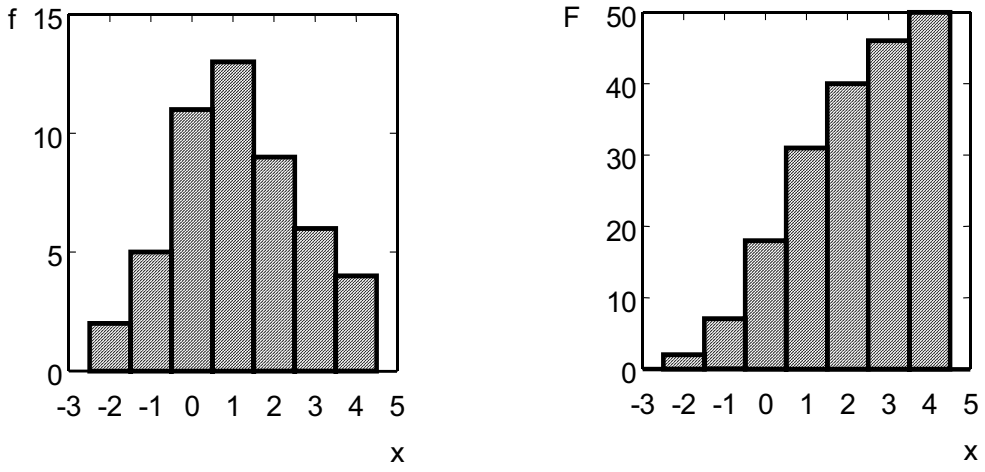
Pro jednorozměrný roztříděný statistický soubor s diskrétním znakem X se užívají obvykle následující grafy. ***Sloupcový graf*** na obr. 1.2 je podobný histogramu z obr. 1.4, avšak vyznačené obdélníky na sebe nenavazují a někdy se kreslí ve vodorovné poloze. ***Koláčový*** (***výsečový***) ***graf*** na obr. 1.3 je kruh rozdělený na výseče, jejichž úhel odpovídá četnostem tříd, případně jsou některé zvolené výseče vysunuty z kruhu. V uvedených grafech se různými barvami nebo šrafováním zvýrazňují potřebné informace a mnohdy se dále geometricky a výtvarně prezentačně modifikují.



Obr. 1.2

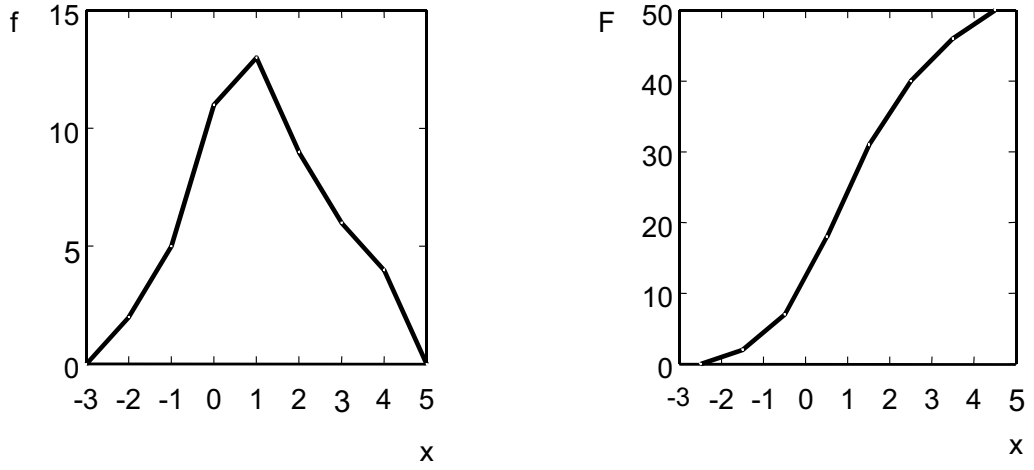


Obr. 1.3



Obr. 1.4

Pro jednorozměrný roztříděný statistický soubor se v případě spojitého znaku X užívají nejčastěji následující dva typy grafů. ***Histogram*** na obr. 1.4 je soustava obdélníků v kartézské souřadné soustavě, jejichž základny jsou třídy a výšky jsou četnosti tříd (absolutní, relativní, kumulativní atd.). ***Polygon*** na obr. 1.5 je lomená čára v kartézské souřadné soustavě spojující body, jejichž x-ová souřadnice je střed třídy, příp. horní hranice třídy pro kumulativní četnosti, a y-ová souřadnice je četnost třídy.



Obr. 1.5

Řešený příklad 1.1

Měřením délky X (mm) 10 válečků byly získány hodnoty: 5,38; 5,36; 5,35; 5,40; 5,41; 5,34; 5,29; 5,43; 5,42; 5,32. Určete rozsah, variační obor, variační rozpětí, aritmetický průměr, rozptyl, směrodatnou odchylku, variační koeficient a medián statistického souboru.

Ř e š e n í:

Rozsah daného souboru je $n = 10$, takže nemá smysl jej třídit. Protože $x_{(1)} = 5{,}29$ mm a $x_{(10)} = 5{,}43$ mm, je variační obor <5,29; 5,43> mm a variační rozpětí je 5,43 – 5,29 = 0,14 mm. Dále je:

$\bar{x}$ = (5,38 +…+ 5,32)/10 = 53,70/10 = 5,37 mm … průměrná délka,

$s^2 = (5{,}38^2 + \ldots + 5{,}32^2)/10 - 5{,}37^2 = 288{,}388/10 - 28{,}8369 = 0{,}0019$ mm^2,

$s = \sqrt{0{,}0019} \approx 0{,}0435889894 \approx 0{,}044$ mm,

$v = \sqrt{0{,}0019}/5{,}37 \approx 0{,}0435889894/5{,}37 \approx 0{,}00811713 \approx 0{,}8117$ %,

$\tilde{x}$ = (5,36 + 5,38)/2 = 5,37 mm …medián délky.

Pro grafické vyjádření tohoto statistického souboru by byl vhodný krabicový graf.

Řešený příklad 1.2

Při kontrole byl zjišťován objem nápoje X v 50 lahvích a byly naměřeny následující odchylky (ml) od hodnoty na etiketě:

1,2;	2,1;	1,7;	0,9;	0,3;	2,0;	-1,3;	-0,1;	3,2;	2,8;
0,8;	4,4;	2,9;	1,2;	0,0;	-2,3;	1,2;	0,9;	2,3;	-0,2;
0,1;	1,9;	-1,9;	-0,2;	-1,3;	1,5;	0,5;	2,0;	-1,3;	3,7;
0,9;	1,0;	0,4;	1,9;	1,4;	-1,3;	1,6;	1,4;	3,1;	-0,1;
1,8;	0,0;	4,1;	1,3;	3,0;	0,4;	3,8;	-0,8;	3,1;	0,9.

Roztřiďte daný statistický soubor, graficky jej znázorněte a vypočtěte $\overline{x}$, s^2, s, $\hat{x}$, A.

Ř e š e n í:

Rozsah souboru $n = 50$; $x_{(1)} = -2{,}3$ ml a $x_{(50)} = 4{,}4$ ml, takže variační obor je $<-2{,}3; 4{,}4>$ ml a rozpětí je $4{,}4 - (-2{,}3) = 6{,}7$ ml. Volíme počet tříd $m = 7$ (tj. asi $\sqrt{50}$) a délku třídy $h = 1$ (tj. asi 6,7/7). Volba tříd a jejich středů, roztřídění do tříd a výpočet absolutních a kumulativních četností je v následující tabulce, kde např. // značí 2 hodnoty a ### značí 5 hodnot ležících v dané třídě:

j	třída	x_j^*	zařazení do tříd	f_j	F_j
1	-2,5; -1,5	-2	//	2	2
2	-1,5; -0,5	-1	###	5	7
3	-0,5; 0,5	0	### ### /	11	18
4	0,5; 1,5	1	### ### ///	13	31
5	1,5; 2,5	2	### ////	9	40
6	2,5; 3,5	3	### /	6	46
7	3,5; 4,5	4	////	4	50

Histogramy a polygony tohoto statistického souboru jsou na obr. 1.4 a 1.5. Další výpočty jsou pro přehlednost znázorněny v následující tabulce, ze které dostaneme:

$\overline{x} = 56/50 = 1{,}12$ ml; $s^2 = 180/50 - 1{,}12^2 = 2{,}3456$ ml^2; $s = \sqrt{2{,}3456} \approx 1{,}532$ ml;

střed třídy s největší četností $\hat{x} = 1$ ml; dalším výpočtem obdržíme $A \approx 0{,}098502$.

j	x_j^*	f_j	$f_j x_j^*$	$f_j x_j^{*2}$
1	-2	2	-4	8
2	-1	5	-5	5
3	0	11	0	0
4	1	13	13	13
5	2	9	18	36
6	3	6	18	54
7	4	4	16	64
$\sum$	—	50	56	180

1.3 Dvourozměrný statistický soubor s kvantitavními znaky

Získaný statistický soubor $((x_1, y_1),...,(x_n, y_n))$ s rozsahem n je ***neroztříděný statistický soubor***. Vynecháním první, resp. druhé, hodnoty v každé dvojici obdržíme jednorozměrné statistické soubory $(x_1,...,x_n)$ a $(y_1,...,y_n)$. Zpracováním těchto souborů získáme jejich číselné charakteristiky $\overline{x}$, $\overline{y}$, $s^2(x)$, $s^2(y)$ atd.

Roztříděný dvourozměrný statistický soubor získáme roztříděním jednorozměrných statistických souborů $(x_1,...,x_n)$ a $(y_1,...,y_n)$, přičemž oba roztříděné soubory mohou mít různé počty tříd i jejich délky. Dostaneme tak dvourozměrné třídy se ***středy*** (x_j^*, y_k^*) a ***absolutními četnostmi*** f_{jk}, $j = 1,...,m_1$ a $k = 1,...,m_2$. Dle potřeby se dále určují ***relativní četnosti*** $\frac{f_{jk}}{n}$, ***kumulativní četnosti*** F_{jk} atd.

Roztříděný dvourozměrný statistický soubor zapisujeme do ***četnostní tabulky*** pro různé typy četností. Následující tabulka je pro absolutní četnosti f_{jk}, kde čísla f_{xj} a f_{yk} jsou ***marginální*** (***okrajové***) ***četnosti*** a platí

$$f_{xj} = \sum_{k=1}^{m_2} f_{jk} \ , \quad f_{yk} = \sum_{j=1}^{m_1} f_{jk} \ , \quad \sum_{j=1}^{m_1} f_{xj} = \sum_{k=1}^{m_2} f_{yk} = \sum_{j=1}^{m_1}\sum_{k=1}^{m_2} f_{jk} = n \, .$$

y_k^* / x_j^*	y_1^*	...	$y_{m_2}^*$	f_{xj}
x_1^*	f_{11}	...	$f_{1\,m_2}$	f_{x1}
...	...	...	...	...
$x_{m_1}^*$	$f_{m_1 1}$	...	$f_{m_1\,m_2}$	$f_{x\,m_1}$
f_{yk}	f_{y1}	...	$f_{y\,m_2}$	n

Pro roztříděné jednorozměrné statistické soubory $\left(x_j^*, f_{xj}\right)$, $j = 1,...,m_1$, a $\left(y_k^*, f_{yk}\right)$, $k = 1,...,m_2$, obdržíme jejich číselné charakteristiky $\overline{x}$, $\overline{y}$, $s^2(x)$, $s^2(y)$ atd.

Mírou závislosti znaků X a Y je ***koeficient korelace*** (***korelační koeficient***)

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{s(x)s(y)} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}}{s(x)s(y)} \qquad \text{pro neroztříděný soubor,}$$

$$r = \frac{\frac{1}{n}\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} f_{jk}\left(x_j^* - \overline{x}\right)\left(y_k^* - \overline{y}\right)}{s(x)s(y)} = \frac{\frac{1}{n}\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} f_{jk} x_j^* y_k^* - \overline{xy}}{s(x)s(y)} \qquad \text{pro roztříděný soubor,}$$

přičemž čitatelé ve všech zlomcích vyjadřují tzv. ***kovarianci***, kterou značíme *cov*. Někdy pro zdůraznění znaků X, Y píšeme $r(x, y)$, resp. $cov(x, y)$. Vlastnosti koeficientu korelace:

a) $u = ax + b,\ v = cy + d \Rightarrow r(u, v) = \frac{ac}{|ac|} r(x, y)$ pro reálné konstanty a, b, c, d, $a \neq 0,\ c \neq 0$,

b) $r(y, x) = r(x, y)$,

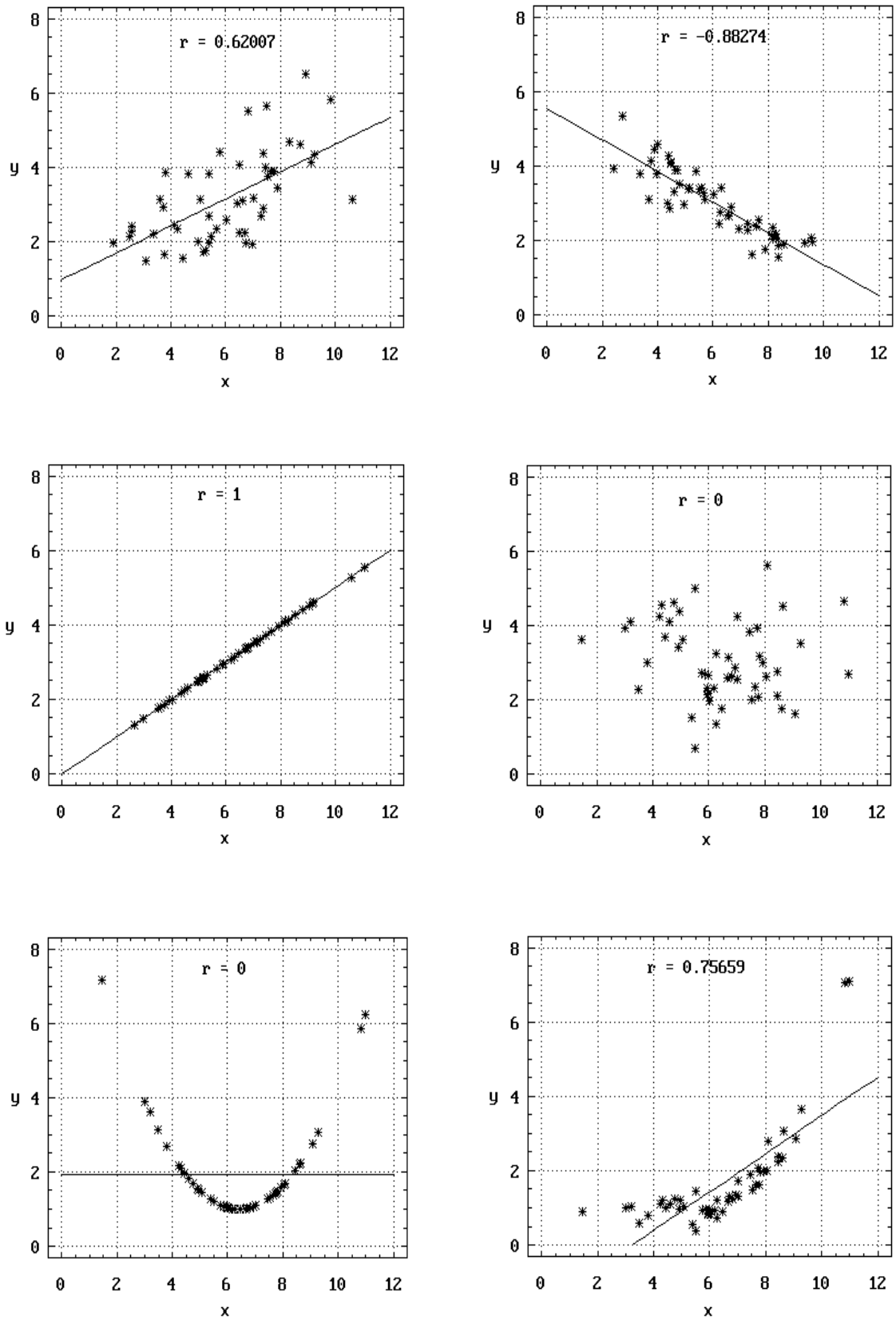
c) $-1 \leq r \leq 1$,

d) $r = \pm 1 \Leftrightarrow y = ax + b,\ a \neq 0$,

e) r je bezrozměrné číslo.

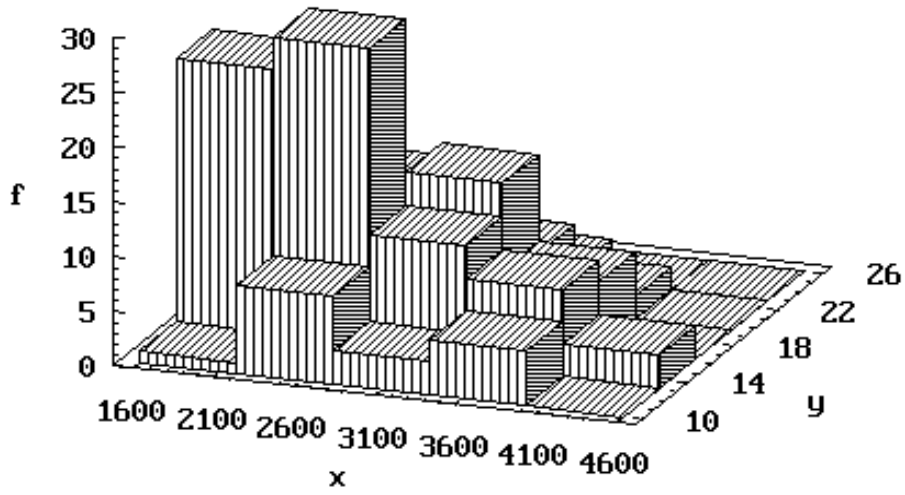
Koeficient korelace r je pouze mírou lineární závislosti mezi znaky X a Y. Čím je jeho hodnota bližší 1 anebo -1, tím je závislost bližžší lineární závislosti a body $\left(x_i, y_i\right)$ bližší přímce. Jeho kladná (záporná) hodnota odpovídá celkově rostoucí (klesající) závislosti mezi X

a *Y*. Hodnota blízká 0 vyjadřuje, že závislost není lineární a znaky *X*, *Y* mohou být nezávislé.



Obr. 1.6

Pro grafické vyjadření dvourozměrného neroztříděného statistického souboru se užívá ***rozptylový graf*** na obr.1.6, kde jsou rovněž uvedeny pro ilustraci hodnoty koeficientu korelace, a pro dvourozměrný roztříděný statistický soubor třírozměrný ***histogram*** na obr. 1.7, případně třírozměrný ***sloupcový graf*** pro diskrétní znaky *X*, *Y*.



Obr. 1.7

Řešený příklad 1.3

Statistickým šetřením nákladů *X* (Kč) a cen *Y* (Kč) pro stejný výrobek u 10 výrobců byl získán dvourozměrný statistický soubor:

(30,18; 50,26), (30,19; 50,23), (30,21; 50,27), (30,22; 50,25), (30,25; 50,22),

(30,26; 50,32), (30,26; 50,33), (30,28; 50,29), (30,30; 50,37), (30,33; 50,42).

Vypočtěte $\overline{x}$, $\overline{y}$, $s^2(x)$, $s^2(y)$, $s(x)$, $s(y)$, c, r.

Ř e š e n í:

Vzhledem k malému rozsahu $n = 10$ soubor netřídíme. Použitím výše uvedených vztahů dostaneme:

$\overline{x}$ = (30,18 + ... + 30,33)/10 = 30,248 Kč … průměrné náklady,

$\overline{y}$ = (50,26 + ... + 50,42)/10 = 50,296 Kč … průměrná cena,

$s^2(x)$ = (0,18^2 + ... + 30,33^2)/10 - 30,248^2 = 0,002096 Kč^2,

$s^2(y)$ = (50,26^2 + ... + 50,42^2)/10 - 50,296^2 = 0,003684 Kč^2,

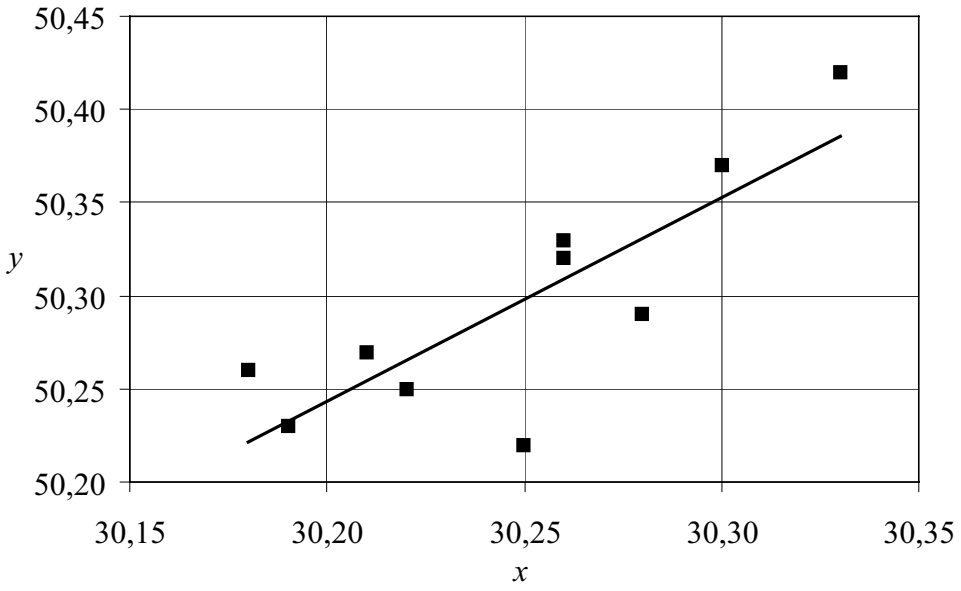
$s(x) = \sqrt{0,002096} \approx 0,0457821$ Kč ≈ 0,0458 Kč,

$s(y) = \sqrt{0,003684} \approx 0,0606960$ Kč ≈ 0,0607 Kč,

cov = (30,18.50,26 + ... + 30,33.50,42)/10 - 30,248.50,296 = 0,002292 Kč^2,

r = 0,002292/(0,0457821·0,0606960) = 0,82481996263 ≈ 0,8248.

Vzhledem k velikosti koeficientu korelace r lze předpokládat, že mezi oběma znaky X a Y (náklady a cenou) je závislost víceméně blízká lineární. Jeho kladná hodnota odpovídá tomu, že s rostoucími náklady roste cena výrobku. Rozptylový graf daného statistického souboru je na obr. 1.8.



Obr. 1.8

1.4 Statistické soubory s kvalitativními znaky

Jednorozměrný statistický soubor s kvalitativním znakem $(x_1,...,x_n)$ s rozsahem n vyjadřujeme pomocí ***četnostní tabulky***, kde x_j^* jsou možné slovní hodnoty znaku X a f_j jsou četnosti těchto hodnot v původním souboru, $j = 1,...,m$. Číselné charakteristiky se až na výjimky (variabilitu) nepoužívají - viz např. [40]. Ke grafickému vyjádření souboru slouží ***sloupcový graf***, ***koláčový graf*** apod. ***Dvourozměrný statistický soubor s kvalitativními znaky*** $((x_1, y_1),...,(x_n, y_n))$ s rozsahem n vyjadřujeme pomocí ***četnostní tabulky*** podobně jako pro kvantitativní znaky, kde (x_j^*, y_k^*) jsou dvojice možných slovních hodnot dvourozměrného kvalitativního znaku (X, Y) a f_{jk} jsou četnosti těchto hodnot v původním souboru pro $j = 1,...,m_1$ a $k = 1,...,m_2$. Z číselných charakteristik se užívají především různé míry závislosti znaků X a Y - viz např. [2], [3], [8], [15], [17], [30]. Ke grafickému vyjádření souboru slouží třírozměrný ***sloupcový graf*** podobný třírozměrnému sloupcovému grafu pro dvourozměrný diskrétní kvantitativní znak.

DODATEK 2 – ELEMENTY TEORIE PRAVDĚPODOBNOSTI

2.1 Náhodné jevy

Náhodný jev je výsledek pokusu (realizace určitého systému podmínek), který může, ale nemusí nastat. Míru možnosti jeho nastoupení vyjadřuje v číselné formě jeho pravděpodobnost. U náhodných jevů požadujeme hromadnost a stabilitu, tj. dostatečnou opakovatelnost a neměnnost pokusu. Nezbytným předpokladem je také rozpoznatelnost náhodných jevů.

Jednotlivým možným (uvažovaným) výsledkům pokusu odpovídají elementární jevy, které vyjadřujeme pomocí jednoprvkových množin $\{\omega\}$. Všechny možné výsledky pokusu tvoří množinu Ω nazývanou základní prostor, přičemž $\omega \in \Omega$. Při pokusu nastane právě takový náhodný jev A, který obsahuje pozorovaný elementární jev $\{\omega\}$. Náhodné jevy A, B, A_1, A_i,… proto vyjadřujeme jako podmnožiny Ω. Jistý jev nastane při každém pokusu a je ekvivalentní základnímu prostoru Ω. Nemožný jev nenastane při žádném pokusu a vyjadřuje jej prázdná množina $\varnothing$.

Vztahy mezi náhodnými jevy vyjadřujeme pomocí množinových inkluzí:

a) $A \subseteq B$ znamená, že nastoupení náhodného jevu A má za následek nastoupení náhodného jevu B.

b) $A = B$ vyjadřuje rovnost náhodných jevů A a B.

Operace s náhodnými jevy vyjadřujeme pomocí množinových operací:

a) Sjednocení $A \cup B$ nastane, jestliže nastane aspoň jeden z náhodných jevů A a B, tedy A nebo B. Analogicky definujeme $\bigcup_{i=1}^{n} A_i$ a $\bigcup_{i=1}^{\infty} A_i$, které nastanou, jestliže nastane aspoň jeden jev A_i.

b) Průnik $A \cap B$ nastane, jestliže nastanou oba náhodné jevy A a B. Analogicky definujeme $\bigcap_{i=1}^{n} A_i$ a $\bigcap_{i=1}^{\infty} A_i$, které nastanou, jestliže nastanou všechny jevy A_i.

c) Rozdíl $A - B$ nastane, jestliže nastane náhodný jev A a nenastane náhodný jev B.

d) Opačný náhodný jev $\overline{A} = \Omega - A$ k náhodnému jevu A nastane, jestliže nenastane jev A.

e) Náhodné jevy A a B jsou disjunktní, jestliže $A \cap B = \varnothing$.

Vlastnosti operací s náhodnými jevy jsou samozřejmě totožné s vlastnostmi operací s množinami. Abychom mohli definovat pravděpodobnost náhodného jevu, zabýváme se jenom takovými náhodnými jevy na Ω, které tvoří následující strukturu.

Jevové pole Σ je množina náhodných jevů (systém podmnožin základního prostoru Ω) s vlastnostmi:

1. $\emptyset \in \Sigma, \ \Omega \in \Sigma$.
2. Pro každý náhodný jev $A \in \Sigma$ také $\overline{A} \in \Sigma$.
3. Pro každou posloupnost náhodných jevů $A_i \in \Sigma$, i = 1, 2, … také $\bigcap_{i=1}^{\infty} A_i \in \Sigma$.

Příklad 2.1

Náhodný pokus spočívá v jednom hodu hrací kostkou ve tvaru krychle se stěnami očíslovanými od 1 do 6. Náhodný jev A nastoupí, jestliže padne sudé číslo a náhodný jev B nastoupí, jestliže padne číslo větší než 4. Určete $\Omega, \overline{A}, \overline{B}$, $A \cup B$, $A \cap B$, $A - B$, $B - A$, Σ.

Řešení:

Základní prostor je Ω = {1, 2, 3, 4, 5, 6} je konečný a elementární náhodné jevy jsou {1}, {2}, {3}, {4}, {5}, {6}. Dále je A = {2, 4, 6} a B ={5, 6}, takže

$\overline{A}$ = {1, 3, 5} ... padne liché číslo,

$\overline{B}$ = {1, 2, 3, 4} ... padne číslo menší než 5,

$A \cup B$ = {2, 4, 6} $\cup$ {5, 6} = {2, 4, 5, 6} ... nepadne číslo 1 a 3,

$A \cap B$ = {2, 4, 6} $\cap$ {5, 6} = {6} ... padne číslo 6,

$A - B$ = {2, 4, 6} – {5, 6} ={2, 4} ... padne číslo 2 nebo 4,

$B - A$ = {5, 6} – {2, 4, 6} = {5} ... padne číslo 5.

Protože nejsou stanovena žádná omezení na náhodné jevy, můžeme uvažovat maximální jevové pole (tj. množinu všech podmnožin základního prostoru Ω)

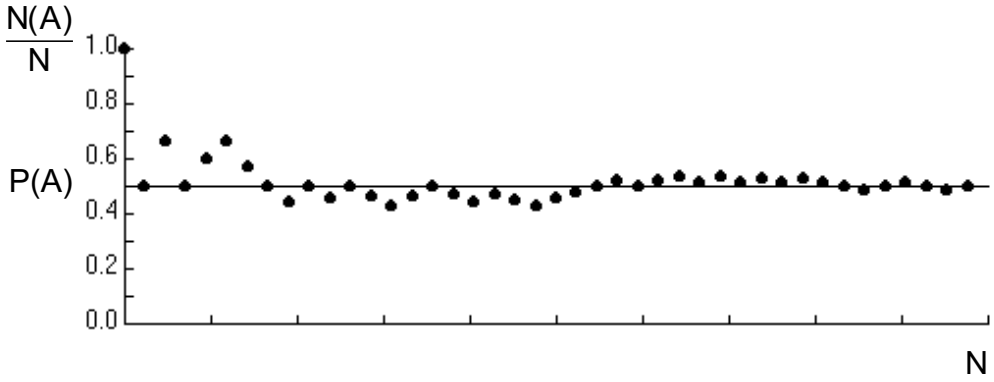
Σ = {$\emptyset$, {1}, {2}, {3}, {4}, {5}, {6}, {1, 2}, {1, 3},…,{5, 6},…, {2, 3, 4, 5, 6}, Ω },

které obsahuje 2^6 = 64 náhodných jevů.

2.2 Pravděpodobnost a její vlastnosti

Jestliže při opakovaných sériích náhodných pokusů, které sestávají vždy z N pokusů, sledujeme chování relativní četnosti nastoupení náhodného jevu A, tj. posloupností

čísel $\frac{N(A)}{N}$, kde N(A) je počet nastoupení jevu A v dané sérii N pokusů, pak vidíme, že posloupnosti relativních četností mají ve skoro všech sériích snahu konvergovat pro dostatečně velký počet pokusů N k jisté pevné hodnotě P(A) – viz na obr. 2.1.



Obr. 2.1. Příklad posloupnosti $\frac{N(A)}{N}$

Teoretická hodnota P(A) vyjadřuje míru možnosti nastoupení náhodného jevu A v jednotlivém pokusu a hovoříme o tzv. „statistické definici pravděpodobnosti“ náhodného jevu A. Z jakékoliv realizované série N pokusů však můžeme pravděpodobnost P(A) náhodného jevu A pomocí zjištěné relativní četnosti $\frac{N(A)}{N}$ pouze více či méně přesně odhadnout. Naopak pravděpodobnost P(A) znamená, že při mnoha pokusech (řádově tisíce a více) nastoupí náhodný jev A zhruba ve 100P(A) % pokusů. Na vlastnostech relativní četnosti

$$0 \le \frac{N(A)}{N} \le 1,\ A \cap B = \varnothing \ \Rightarrow\ \frac{N(A \cup B)}{N} = \frac{N(A)}{N} + \frac{N(B)}{N},$$

je založena následující obecná (axiomatická) definice pravděpodobnosti náhodného jevu.

Pravděpodobnost P(A) náhodného jevu $A \in \Sigma$ je reálná funkce definovaná na Σ s vlastnostmi:

1. $P(A) \ge 0$ pro všechny náhodné jevy $A \in \Sigma$.
2. $P(\Omega) = 1$.
3. Pro každou posloupnost disjunktních náhodných jevů $A_i \in \Sigma$, i = 1, 2,… , je

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Uspořádaná trojice (Ω, Σ, P) se nazývá pravděpodobnostní prostor.

Platí:

a) $P(\overline{A}) = 1 - P(A)$; $P(\emptyset) = 0$; $0 \le P(A) \le 1$.

b) $A \subseteq B \Rightarrow P(A) \le P(B)$, $P(B - A) = P(B) - P(A)$.

c) $$P(A_1 \cup ... \cup A_n) = 1 - P(\overline{A}_1 \cap ... \cap \overline{A}_n) =$$

$$= \sum_{i=1}^{n} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{n} P(A_i \cap A_j) + ... + (-1)^{n-1} P(A_1 \cap ... \cap A_n),$$

speciálně pro n = 2 je $P(A \cup B) = 1 - P(\overline{A} \cap \overline{B}) = P(A) + P(B) - P(A \cap B)$.

Pro konečný nebo spočetný základní prostor Ω (tj. elementární jevy $\{\omega\}$ lze uspořádat do posloupnosti) je

$$P(A) = \sum_{\omega \in A} P(\{\omega\}).$$

Speciálně pro základní prostor Ω s n stejně pravděpodobnými elementárními jevy je

$$P(A) = \frac{m}{n},$$

kde m je počet elementárních jevů $\{\omega\}$, z nichž sestává náhodný jev A. Říkáme, že „m je počet příznivých výsledků pokusu“ a „n je počet výsledků pokusu“ a že jde o tzv. klasickou definici pravděpodobnosti.

P ř í k l a d 2.2

Vypočtěte pravděpodobnosti P(A), P(B), $P(\overline{A})$, $P(\overline{B})$, $P(A \cup B)$, $P(A \cap B)$, P(A − B), P(B − A) náhodných jevů z příkladu 1, jestliže kostka je z homogenního materiálu.

Ř e š e n í:

Elementární náhodné jevy mají vzhledem k pravidelnosti a homogennosti hrací kostky stejnou pravděpodobnost $P(\{\omega\}) = \frac{1}{6}$ a n = 6. Přímým výpočtem z „klasické definice pravděpodobnosti“ obdržíme

$$P(A) = \frac{3}{6} = \frac{1}{2}, \qquad P(B) = \frac{2}{6} = \frac{1}{3},$$

$$P(\overline{A}) = \frac{3}{6} = \frac{1}{2}, \qquad P(\overline{B}) = \frac{4}{6} = \frac{2}{3},$$

$$P(A \cup B) = \frac{4}{6} = \frac{2}{3}, \qquad P(A \cap B) = \frac{1}{6},$$

$$P(A - B) = \frac{2}{6} = \frac{1}{3}, \qquad P(A - B) = \frac{1}{6}.$$

Z vlastností pravděpodobnosti lze např. určit

$$P(\overline{A}) = 1 - P(A) = 1 - \frac{1}{2} = \frac{1}{2},$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3}.$$

Příklad 2.3

V dodávce 100 hřídelí nemá požadovaný průměr 10 kusů, požadovanou délku nemá 20 kusů a současně nemá požadovaný průměr i délku 5 kusů. Určete pravděpodobnost toho, že náhodně vybraný hřídel má požadovaný průměr i délku.

Řešení:

Jestliže A, popř. B, značí, že náhodně vybraný hřídel nemá požadovaný průměr, popř. délku, potom pravděpodobnost toho, že náhodně vybraný hřídel má požadovaný průměr i délku je

$$P(\overline{A} \cap \overline{B}) = 1 - P\left(\overline{\overline{A} \cap \overline{B}}\right) = 1 - P(A \cup B) = 1 - \left[P(A) + P(B) - P(A \cap B)\right] =$$
$$= 1 - (0{,}10 + 0{,}20 - 0{,}05) = 0{,}75.$$

2.3 Podmíněná pravděpodobnost a nezávislé jevy

Pravděpodobnost náhodného jevu $A \in \Sigma$ za podmínky (předpokladu), že nastane náhodný jev $B \in \Sigma$, $P(B) \neq 0$, je podmíněná pravděpodobnost

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

Platí:

a) $P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2/A_1) \ldots P(A_n/A_1 \cap \ldots \cap A_{n-1})$,
speciálně je $P(A \cap B) = P(A)P(B/A) = P(B)P(A/B)$,

b) Pro náhodný jev $A \subseteq \bigcup_{i=1}^{n} B_i$, kde B_i, jsou disjunktní náhodné jevy,
$i = 1, \ldots, n$, je tzv. úplná pravděpodobnost

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A/B_i)$$

a pro $P(A) \neq 0$ platí Bayesův vzorec

$$P(B_j / A) = \frac{P(B_j)P(A/B_j)}{\sum_{i=1}^{n} P(B_i)P(A/B_i)}, \qquad j = 1, \ldots, n.$$

Příklad 2.4

Ze skupiny 100 výrobků, která obsahuje 10 zmetků, vybereme náhodně bez vracení 3 výrobky. Pravděpodobnost toho, že první výrobek není zmetek - náhodný jev A_1, druhý výrobek není zmetek – náhodný jev A_2 a třetí výrobek je zmetek – náhodný jev $\overline{A}_3$, je

$$P(A_1 \cap A_2 \cap \overline{A}_3) = P(A_1)P(A_2/A_1)P(\overline{A}_3/A_1 \cap A_2) = \\ = (90/100)(89/99)(10/98) \approx 0{,}08256 .$$

Příklad 2.5

Do obchodu s potravinami dodávají rohlíky 3 pekárny v počtech 500, 1000 a 1500 kusů denně. Zmetkovitost jejich dodávek je 5%, 4% a 3%. Jejich dodávky jsou v obchodě smíchány do celkové zásoby. Určete pravděpodobnost, že

a) náhodně vybraný rohlík z celkové zásoby je zmetek,

b) náhodně vybraný rohlík z celkové zásoby, který je zmetek, byl dodán druhou pekárnou.

Řešení:

Označme náhodné jevy

A ... vybraný rohlík je zmetek,

B_i ... rohlík byl dodán i-tou pekárnou, i = 1, 2, 3.

Pravděpodobnosti jsou

$$P(B_1) = \frac{500}{500+1000+1500} = \frac{1}{6}, \; P(A/B_1) = 0{,}05,$$

$$P(B_2) = \frac{1000}{500+1000+1500} = \frac{2}{6}, \; P(A/B_2) = 0{,}04,$$

$$P(B_3) = \frac{1500}{500+1000+1500} = \frac{3}{6}, \; P(A/B_3) = 0{,}03 .$$

a) Podle vzorce pro úplnou pravděpodobnost je

$$P(A) = 0{,}05\frac{1}{6} + 0{,}04\frac{2}{6} + 0{,}03\frac{3}{6} = \frac{0{,}22}{6} = 0{,}03\overline{6} \approx 0{,}03667,$$

takže zmetkovitost z hlediska zákazníka je přibližně 3,667%.

b) Z Bayesova vzorce je pro j = 2

$$P(B_2 / A) = \frac{0,04\frac{2}{6}}{\frac{0,22}{6}} = \frac{0,08}{0,22} = 0,\overline{36} \approx 0,36364 .$$

Analogicky lze získat $P(B_1 /A) \approx 0,22727$ a $P(B_3 /A) \approx 0,40909$, takže největší podíl na zmetkovitosti celkové zásoby má 3. pekárna. Přitom má absolutně nejmenší zmetkovitost ze všech tří dodavatelů, avšak dodává největší počet rohlíků.

Náhodné jevy $A, B \in \Sigma$ jsou <u>nezávislé</u>, jestliže P(A/B) = P(A) anebo P(B) = 0. Náhodné jevy $A_1,\ldots, A_n \in \Sigma$ jsou <u>vzájemně nezávislé</u>, jestliže jsou nezávislé všechny náhodné jevy ve dvojicích

A_i, A_j pro $i \neq j$,

$A_i, A_j \cap A_k$ pro $i \neq j, i \neq k$,

$A_i, A_j \cap A_k \cap A_m$ pro $i \neq j, i \neq k$ a $i \neq m$,

atd.

Platí:

a) A, B jsou nezávislé, právě když $P(A \cap B) = P(A)P(B)$.

b) Jestliže $A_1, \ldots, A_n$ jsou vzájemně nezávislé, pak

$P(A_1 \cap \ldots \cap A_n) = P(A_1) \ldots P(A_n)$,

$P(A_1 \cup \ldots \cup A_n) = 1 - [1 - P(A_1)] \ldots [1 - P(A_n)]$,

$B_1, \ldots, B_n$ jsou vzájemně nezávislé pro libovolné varianty $B_i = A_i, \overline{A}_i, \Omega$.

<u>P ř í k l a d 2.6</u>

Jaká je pravděpodobnost, že v prvním hodu pravidelnou homogenní šestistěnnou kostkou padne sudé číslo (náhodný jev A) a ve druhém hodu touto kostkou padne liché číslo (náhodný jev B)?

Ř e š e n í:

Náhodné jevy A a B jsou nezávislé a jejich pravděpodobnosti jsou P(A) = P(B) = 1/2, takže $P(A \cap B) = (1/2).(1/2) = 1/4$.

P ř í k l a d 2.7

Výrobek prochází třemi nezávislými operacemi, při kterých jsou pravděpodobnosti výroby zmetku $P(A_1) = 0{,}05$, $P(A_2) = 0{,}08$ a $P(A_3) = 0{,}03$. Určete pravděpodobnost výroby zmetku po všech třech operacích.

Ř e š e n í:

Vzhledem k nezávislosti operací jsou vzájemně nezávislé náhodné jevy A_1, A_2, A_3 a výrobek je zmetek, jestliže nastane aspoň jeden z těchto jevů, takže

$P(A_1 \cup A_2 \cup A_3) = 1 - [1 - P(A_1)][1 - P(A_2)][1 - P(A_3)] = 1 - 0{,}95.0{,}92.0{,}97 = 0{,}15222.$

2.4 Náhodná veličina a její funkční charakteristiky

Náhodná veličina (náhodná proměnná) X je reálná proměnná, která nabývá náhodně reálných číselných hodnot x – blíže v [1], [2], [3]. Její distribuční funkce je

$$F(x) = P(X < x) = P[X \in (-\infty; x)],\ x \in (-\infty;+\infty).$$

Distribuční funkce má vlastnosti:

a) $0 \le F(x) \le 1$ pro všechna $x \in (-\infty;+\infty)$,

b) $F(x)$ je neklesající a zleva spojitá na $(-\infty;+\infty)$.

c) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$,

d) $P(a \le X < b) = F(b) - F(a)$ pro libovolná reálná čísla $a < b$,

e) $P(X = c) = \lim_{x \to c+} F(x) - F(c)$ pro libovolné reálné číslo c .

Někdy se distribuční funkce definuje vztahem $F(x) = P(X \le x)$. Tato distribuční funkce je zprava spojitá, $P(a < X \le b) = F(b) - F(a)$ a $P(X = c) = F(c) - \lim_{x \to c-} F(x)$. Potkáme se s ní zejména ve statistických softwarových produktech.

Náhodná veličina X je diskrétní a říkáme, že má diskrétní rozdělení pravděpodobnosti, jestliže nabývá nejvýše spočetně mnoha hodnot $x = x_1, x_2, \ldots$. Její pravděpodobnostní funkce je posloupnost

$$p(x) = P(X = x) > 0 \text{ pro } x = x_1, x_2, \ldots.$$

Platí:

a) $\sum_x p(x) = 1$,

b) $F(x) = \sum_{t<x} p(t)$ pro všechna $x \in (-\infty;+\infty)$,

c) $P(X \in M) = \sum_{x \in M} p(x)$ pro libovolnou množinu reálných čísel M.

Distribuční funkce diskrétní náhodné veličiny má “schodovitý tvar” – viz obr. 2.2.

P ř í k l a d 2.8

Pravděpodobnost poruchy každé ze tří nezávisle pracujících výrobních linek je $0 < p < 1$. Diskrétní náhodná veličina X, která vyjadřuje počet výrobních linek v poruše, nabývá hodnot x = 0, 1, 2, 3 a hodnoty její pravděpodobnostní funkce jsou

$$p(0) = (1 - p)^3,$$

$$p(1) = 3p(1 - p)^2,$$

$$p(2) = 3p^2(1 - p),$$

$$p(3) = p^3.$$

Její distribuční funkce je

$$F(x) = 0 \text{ pro } x \in (-\infty, 0\rangle,$$

$$F(x) = p(0) = (1 - p)^3 \quad \text{pro } x \in (0, 1\rangle,$$

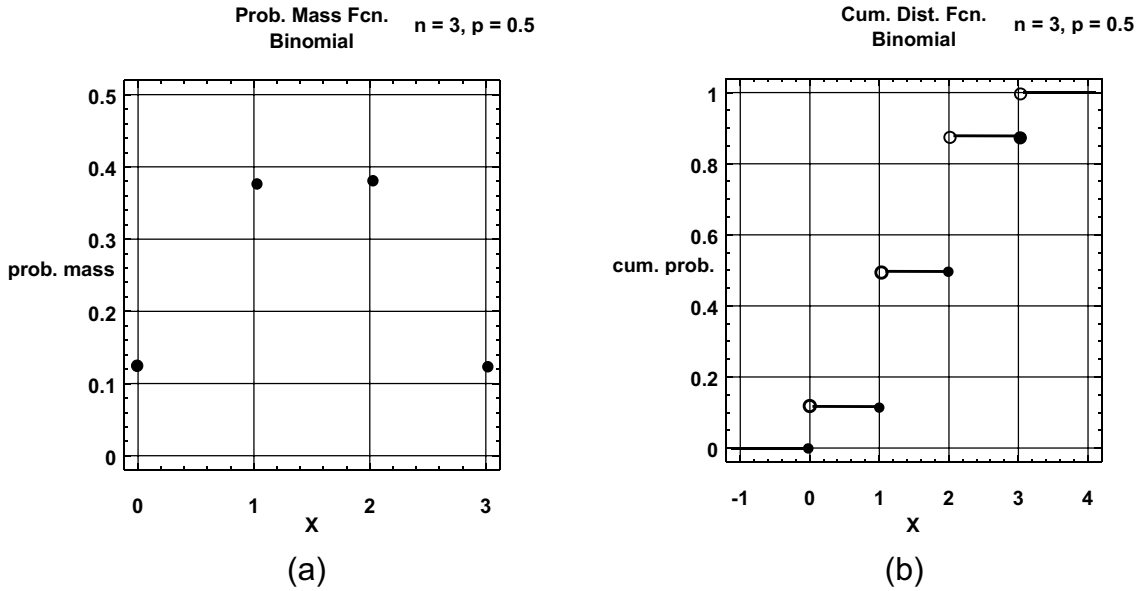
$$F(x) = p(0) + p(1) = (1 + 2p)(1 - p)^2 \quad \text{pro } x \in (1, 2\rangle,$$

$$F(x) = p(0) + p(1) + p(2) = (1 + p + p^2)(1 - p) = 1 - p^3 \quad \text{pro } x \in (2, 3\rangle,$$

$$F(x) = p(0) + p(1) + p(2) + p(3) = 1 \quad \text{pro } x \in (3; \infty).$$

Na obr. 2.2 jsou grafy p(x) a F(x) pro p = 0,5. Pravděpodobnost toho, že alespoň jedna linka má poruchu je

$$P(X \geq 1) = P(1 \leq X < +\infty) = F(+\infty) - F(1) = 1 - (1 - p)^3.$$



(a) (b)

Obr. 2.2 Grafy pravděpodobnostní funkce (a) a distribuční funkce (b) diskrétního rozdělení pravděpodobnosti

Náhodná veličina X je spojitá a říkáme, že má spojité rozdělení pravděpodobnosti, jestliže má spojitou distribuční funkci (tedy X nabývá všech hodnot z nějakého intervalu apod.). Její hustota pravděpodobnosti, je taková nezáporná funkce f(x), že

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad \text{pro všechna } x \in (-\infty;+\infty).$$

Platí:

a) $\int_{-\infty}^{+\infty} f(x)dx = 1,$

b) $f(x) = F'(x)$, pokud derivace existuje,

c) $F(x)$ je spojitá funkce pro všechna $x \in (-\infty;+\infty)$,

d) $P(a \le X \le b) = P(a < X < b) = P(a < X \le b) = P(a \le X < b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$

pro libovolná reálná čísla $a \le b$,

e) $P(X = c) = 0$ pro libovolné reálné číslo c.

P ř í k l a d 2.9

Náhodná veličina X má hustotu pravděpodobnosti $f(x) = cx$ pro $x \in \langle 0; 2\rangle$ a 0 pro $x \notin \langle 0; 2\rangle$. Z vlastností spojité náhodné veličiny získáme následující výsledky. Je

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{2} cxdx + \int_{2}^{+\infty} 0dx = \ldots = 2c = 1,$$

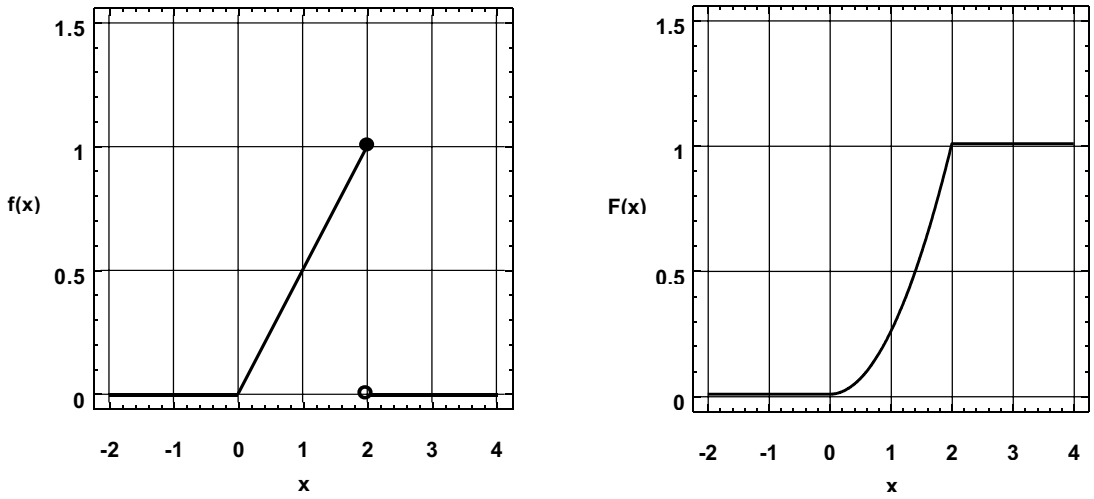
takže $c = 1/2$ a $f(x) = \frac{x}{2}$ pro $x \in \langle 0; 2\rangle$. Distribuční funkce náhodné veličiny X je

$$F(x) = \int_{-\infty}^{x} 0dt = 0 \quad \text{pro } x \in (-\infty; 0\rangle,$$

$$F(x) = \int_{-\infty}^{0} 0dt + \int_{0}^{x} \frac{t}{2}dt = \ldots = \frac{x^2}{4} \quad \text{pro } x \in \langle 0; 2\rangle,$$

$$F(x) = \int_{-\infty}^{0} 0dt + \int_{0}^{2} \frac{t}{2}dt + \int_{2}^{x} 0dt = \ldots = 1 \quad \text{pro } x \in \langle 2;+\infty).$$

Na obr. 2.3 jsou grafy f(x) a F(x). Pravděpodobnost toho, že náhodná veličina nabude hodnotu $x \in \langle 1; 3\rangle$ je $P(1 \le X \le 3) = F(3) - F(1) = 1 - (1^2/4) = 0{,}75$.



(a) (b)

Obr. 2.3 Grafy hustoty pravděpodobnosti (a) a distribuční funkce (b) spojitého rozdělení pravděpodobnosti

2.5 Číselné charakteristiky náhodné veličiny

Číselné charakteristiky náhodné veličiny X jsou reálná čísla, která koncentrovaně vyjadřují její důležité vlastnosti.

Polohu rozdělení pravděpodobnosti charakterizuje střední hodnota náhodné veličiny X

$$E(X) = \sum_x xp(x) \quad \text{pro diskrétní náhodnou veličinu X,}$$

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \quad \text{pro spojitou náhodnou veličinu X,}$$

pokud sumace, příp. integrál, konverguje absolutně.

Střední hodnota má vlastnosti:

a) $E(aX + b) = aE(X) + b$ pro libovolná reálná čísla a, b,

b) $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$ pro náhodné veličiny $X_1, \ldots, X_n$.

Míru kolísání hodnot náhodné veličiny X kolem její střední hodnoty E(X) vyjadřuje její rozptyl (disperze, variance) $D(X) = E\left[(X - E(X))^2\right]$.

Rozptyl má vlastnosti:

a) $D(X) = \sum_x (x - E(X))^2 p(x) = \sum_x x^2 p(x) - (E(X))^2$ pro diskrétní náhodnou

veličinu X,

$D(X) = \int_{-\infty}^{+\infty}(x - E(X))^2 f(x)dx = \int_{-\infty}^{+\infty}x^2 f(x)dx - (E(X))^2$ pro spojitou náhodnou

veličinu X, pokud sumace , příp. integrál, konvergují,

b) $D(X) \geq 0$,

c) $D(aX + b) = a^2 D(X)$ pro libovolná reálná čísla a, b,

e) $D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i)$ pro nezávislé náhodné veličiny $X_1, \ldots, X_n$.

Směrodatná odchylka náhodné veličiny X je $\sigma(X) = \sqrt{D(X)}$.

Směrodatná odchylka má vlastnosti:

a) $\sigma(X) \geq 0$;

b) $\sigma(aX + b) = |a|\, \sigma(X)$ pro libovolná reálná čísla a, b.

Střední hodnota, popř. rozptyl, náhodné veličiny X je speciální případ tzv. obecného, popř. centrálního momentu. Blíže o momentových charakteristikách (variačním koeficientu, koeficientech šikmosti a špičatosti) v [1], [2], [3].

P-kvantil nebo také 100P%-kvantil náhodné veličiny X je pro $0 < P < 1$ její hodnota $x_P = \inf\{x; F(x) \geq P\}$. Pro spojitou náhodnou veličinu X s rostoucí distribuční funkcí je $F(x_P) = P$. Medián náhodné veličiny X je její kvantil $x_{0,5}$ a charakterizuje její polohu. Další kvantilové charakteristiky jsou v [2], [3].

Modus $\hat{x}$ náhodné veličiny X je její hodnota, v níž nabývá pravděpodobnostní funkce nebo hustota pravděpodobnosti maximum, příp. suprémum.

P ř í k l a d 2.10

Náhodná veličina X z příkladu 2.9 má střední hodnotu

$$E(X) = \int_{-\infty}^{0} x \cdot 0dx + \int_{0}^{2} x \frac{x}{2} dx + \int_{2}^{+\infty} x \cdot 0dx = \ldots = \frac{4}{3} \approx 1{,}33333,$$

rozptyl

$$D(X) = \int_{-\infty}^{0} x^2 \cdot 0dx + \int_{0}^{2} x^2 \frac{x}{2} dx + \int_{2}^{+\infty} x^2 \cdot 0dx - \left(\frac{4}{3}\right)^2 = 2 - \frac{16}{9} = \frac{2}{9} \approx 0{,}22222,$$

a směrodatnou odchylku

$$\sigma(X) = \sqrt{\frac{2}{9}} \approx 0{,}47140 \ .$$

P-kvantil x_P je kořen rovnice $\frac{x^2}{4} = P$ z intervalu $\langle 0; 2\rangle$, tedy $x_P = 2\sqrt{P}$. Odtud medián náhodné veličiny X je $x_{0,5} = 2\sqrt{0,5} \approx 1,41421$. Z grafu f(x) na obr. 3 vidíme, že modus náhodné veličiny X je $\hat{x} = 2$.

2.6 Některá významná rozdělení pravděpodobnosti

Diskrétní rozdělení pravděpodobnosti

a) Binomické rozdělení Bi(n, p), kde n je přirozené číslo, p je reálné číslo, $0 < p < 1$:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n;$$

$$E(X) = np; \quad D(X) = np(1 - p); \quad (n + 1)p - 1 \le \hat{x} \le (n + 1)p.$$

Toto rozdělení má počet nastoupení sledovaného náhodného jevu v posloupnosti n vzájemně nezávislých pokusů (např. počet zmetků x mezi n výrobky, když p je pravděpodobnost výroby zmetku). Jedná se také o popis tzv. výběru s vracením, kdy např. postupně vybíráme z dodávky n výrobků a každý vybraný výrobek vracíme zpět do dodávky.

Příklad 2.11

V sérii 50 výrobků je 5 zmetků. Ze série jsou náhodně vybrány 3 výrobky. Počet zmetků mezi vybranými výrobky je náhodná veličina X. Určete typ jejího rozdělení pravděpodobnosti, její pravděpodobnostní funkci p(x), střední hodnotu E(X), rozptyl D(X), směrodatnou odchylku $\sigma(X)$, medián $x_{0,5}$, modus $\hat{x}$ a $P(1 < X \le 3)$. Předpokládejte, že každý vybraný výrobek se vrátí nazpět do série, takže jde o náhodný výběr s vracením.

Řešení:

Náhodná veličina X má rozdělení Bi(n,p), kde n = 3 a p = 5/50 = 0,1. X nabývá hodnot x = 0, 1, 2, 3. Pravděpodobnostní funkce je

$$p(x) = \binom{3}{x} 0,1^x.0,9^{3-x} \text{ pro } x = 0, 1, 2, 3.$$

Střední hodnota je E(X) = np = 3.0,1 = 0,3,

rozptyl je D(X) = np(1 - p) = 3.0,1.0,9 = 0,27,

směrodatná odchylka je $\sigma(X) = \sqrt{D(X)} = \sqrt{0,27} \approx 0,51962$,

medián $x_{0,5} = 0$, neboť $p(0) = 0{,}729$,

modus $\hat{x} = 0$, neboť $(n + 1)p - 1 = -0{,}6$ a $(n + 1)p = 0{,}4$,

$P(1 < X \leq 3) = p(2) + p(3) = 0{,}027 + 0{,}001 = 0{,}028$.

b) Hypergeometrické rozdělení H(N,M,n), kde N, M a n jsou přirozená čísla, $1 \leq n \leq N$, $1 \leq M \leq N$:

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad x = \max\{0, M - N + n\}, \ldots, \min\{M, N\};$$

$$E(X) = n\frac{M}{N}; \quad D(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}; \quad a - 1 \leq \hat{x} \leq a, \text{ kde } a = \frac{(M+1)(n+1)}{N+2}.$$

Toto rozdělení popisuje tzv. náhodný výběr bez vracení, kdy např. N je celkový počet výrobků, M počet zmetků a vybereme náhodně (bez vracení) n výrobků, mezi nimiž je x zmetků.

Příklad 2.12

V sérii 50 výrobků je 5 zmetků. Ze série jsou náhodně vybrány 3 výrobky. Počet zmetků mezi vybranými výrobky je náhodná veličina X. Určete typ jejího rozdělení pravděpodobnosti, její pravděpodobnostní funkci p(x), střední hodnotu E(X), rozptyl D(X), směrodatnou odchylku $\sigma(X)$, medián $x_{0,5}$, modus $\hat{x}$ a $P(1 < X \leq 3)$. Předpokládejte (na rozdíl od příkladu 2.11), že vybraný výrobek se nevrací nazpět, takže jde o náhodný výběr bez vracení.

Řešení:

Náhodná veličina X má rozdělení H(N,M,n), kde N = 50, M = 5 a n = 3. X nabývá hodnot x = 0, 1, 2, 3. Pravděpodobnostní funkce je

$$p(x) = \frac{\binom{5}{x}\binom{45}{3-x}}{\binom{50}{3}} \quad \text{pro } x = 0, 1, 2, 3.$$

Střední hodnota je $E(X) = n\frac{M}{N} = 3.0{,}1 = 0{,}3$,

rozptyl je D(X) = $D(X) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$ = 3.0,1.0,9. (47/49) ≈ 0,25898,

směrodatná odchylka je $\sigma(X) = \sqrt{D(X)} \approx \sqrt{0{,}25898} \approx$ 0,50890,

medián $x_{0,5}$ = 0, neboť max p(x) = p(0) ≈ 0,724,

modus $\hat{x}$ = 0, neboť $a = \frac{(M+1)(n+1)}{N+2} \approx 0{,}46154$, a – 1 ≈ -0,53846,

P(1 < X ≤ 3) = p(2) + p(3) ≈ 0,023 + 0,0005 = 0,0235.

c) Poissonovo rozdělení Po(λ), kde λ je reálné číslo, λ > 0:

$$p(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x = 0, 1, \ldots ; \quad E(X) = \lambda; \quad D(X) = \lambda; \quad \lambda - 1 \le \hat{x} \le \lambda.$$

Toto rozdělení se obvykle užívá pro vyjádření pravděpodobnosti počtu nastoupení sledovaného jevu v určitém časovém intervalu (počet poruch, nehod, katastrof, zmetků apod.) s malou pravděpodobností výskytu.

Příklad 2.13

Během 1 minuty navštíví prodejnu průměrně 3 zákazníci. Najděte vhodný typ rozdělení pravděpodobnosti náhodné veličiny X vyjadřující počet zákazníků, kteří navštíví prodejnu za 1 minutu, střední počet zákazníků, rozptyl jejich počtu a nejpravděpodobnější počet zákazníků za 1 minutu. Určete dále pravděpodobnost, že během 1 minuty přijde a) právě 1 zákazník, b) aspoň 1 zákazník.

Řešení:

Nahradíme-li střední počet zákazníků, kteří navštíví prodejnu během 1 min, jejich průměrným počtem, můžeme vyjádřit náhodnou veličinu X pomocí Poissonova rozdělení pravděpodobnosti Po(λ) s pravděpodobnostní funkcí

$$p(x) = \frac{3^x}{x!}e^{-3}, \quad x = 0, 1, \ldots .$$

Střední hodnota E(X) = λ = 3, rozptyl D(X) = λ = 3, pro modus je λ - 1 ≤ $\hat{x}$ ≤ λ, takže $\hat{x}$ = 2 a 3,

P(X = 1) = p(1) = $\frac{3^1}{1!}e^{-3} \approx$ 0, 14936,

P(X ≥ 1) = p(1) + p(2) + … = 1 – p(0) = 1 – $\frac{3^0}{0!}e^{-3} \approx 1 -$ 0, 04979 = 0,95021.

Spojitá rozdělení pravděpodobnosti

a) Rovnoměrné rozdělení R(a, b), kde $a < b$ jsou reálná čísla:

$$f(x) = \frac{1}{b-a} \text{ pro } x \in \langle a;b \rangle,$$
$$= 0 \quad \text{pro } x \notin \langle a;b \rangle,$$

$$F(x) = 0 \text{ pro } x \in (-\infty;a),$$
$$= \frac{x-a}{b-a} \text{ pro } x \in \langle a;b \rangle,$$
$$= 1 \text{ pro } x \in (b;+\infty),$$

$$E(X) = x_{0,5} = \frac{a+b}{2}; \qquad D(X) = \frac{(b-a)^2}{12}.$$

Toto rozdělení slouží především k simulaci reálných procesů nebo numerickým výpočtům tzv. metodou Monte Carlo na počítači a pro výpočty pomocí tzv. geometrické pravděpodobnosti.

P ř í k l a d 2.14

K přerušení optického kabelu v délce 500 m může dojít v libovolné vzdálenosti od jeho počátku, přičemž pravděpodobnost náhodného jevu, že dojde k přerušení v nějakém úseku je přímo úměrná délce úseku a nezávisí na jeho poloze. Určete rozdělení pravděpodobnosti náhodné veličiny X vyjadřující vzdálenost místa přerušení kabelu od jeho počátku, její hustotu pravděpodobnosti a základní číselné charakteristiky a pravděpodobnost, že k přerušení kabelu dojde v úseku od 300 m do 400 m od počátku.

Ř e š e n í:

Náhodná veličina X má rozdělení R(a, b), kde $a = 0$ a $b = 500$ s hustotou pravděpodobnosti $f(x) = \frac{1}{500}$ pro $x \in \langle 0;\ 500 \rangle$ a $f(x) = 0$ pro $x \notin \langle 0;\ 500 \rangle$.

Střední vzdálenost a medián $E(X) = x_{0,5} = \frac{0+500}{2} = 250$ m,

rozptyl $D(X) = \frac{(500-0)^2}{12} \approx 20833{,}3 \text{ m}^2$,

směrodatná odchylka $\sigma(X) = \sqrt{D(X)} \approx \sqrt{20833{,}3} \approx 144{,}34$ m,

pravděpodobnost $P(300 \le X \le 400) = F(400) - F(300) = \frac{400}{500} - \frac{300}{500} = 0{,}2.$

b) Normální rozdělení $N(\mu, \sigma^2)$, kde μ, σ^2 jsou reálná čísla, $\sigma^2 > 0$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right], \quad x \in (-\infty, +\infty);$$

$$E(X) = x_{0,5} = \hat{x} = \mu, \quad D(X) = \sigma^2.$$

Toto nejrozšířenější rozdělení (nazývané také Gaussovo rozdělení) se užívá k vyjádření náhodných veličin, které lze interpretovat jako aditivní výsledek mnoha nezávislých vlivů (např. chyba měření, odchylka rozměru výrobku apod.). Někdy se také hovoří o zákonu chyb.

Transformací

$$U = \frac{X - \mu}{\sigma}$$

dostaneme normované (základní) normální rozdělení N(0;1), jehož distribuční funkce $\Phi(x)$ je tabelována (viz tabulku T1) anebo její hodnoty určíme výpočtem na PC, např. pomocí software Excel. Platí

$$\Phi(-x) = 1 - \Phi(x).$$

Pro náhodnou veličinu X s normálním rozdělením $N(\mu, \sigma^2)$ je

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma} \right),$$

a např. $P(\mu - 3\sigma \le X \le \mu + 3\sigma) \approx 0{,}9973$ (tzv. pravidlo tří sigma).

P ř í k l a d 2.15

Jaká je pravděpodobnost, že náhodná veličina X, která má rozdělení N(20;16), nabude hodnotu a) menší než 16, b) větší než 20, c) v mezích od 12 do 28, d) menší než 12 nebo větší než 28 ?

Ř e š e n í:

Ze vztahu $F(x) = \Phi\left(\frac{x - \mu}{\sigma} \right)$ a tabulky T1 dostaneme

a) $P(X < 16) = F(16) - F(-\infty) = F(16) - 0 = \Phi((16 - 20) / 4) = \Phi(-1) = 1 - \Phi(1) \approx$

$\approx 1 - 0{,}84135 = 0{,}15865$;

b) $P(X > 20) = 1 - P(X \le 20) = 1 - F(20) = 1 - \Phi((20 - 20) / 4) = 1 - \Phi(0) =$

$= 1 - 0{,}5 = 0{,}5$;

c) $P(12 \leq X \leq 28) = F(28) - F(12) = \Phi((28 - 20) / 4) - \Phi((12 - 20) / 4) =$

$= \Phi(2) - \Phi(-2) = \Phi(2) - (1 - \Phi(2)) = 2\Phi(2) - 1 \approx 2.0{,}97725 - 1 =$

$= 0{,}9545$;

d) $P((X < 12) \vee (X > 28)) = 1 - P(12 \leq X \leq 28) \approx 1 - 0{,}9545 = 0{,}0455$.

Informace o dalších v praxi často užívaných rozděleních pravděpodobnosti a náhodných vektorech lze najít např. v [1], [2], [3].