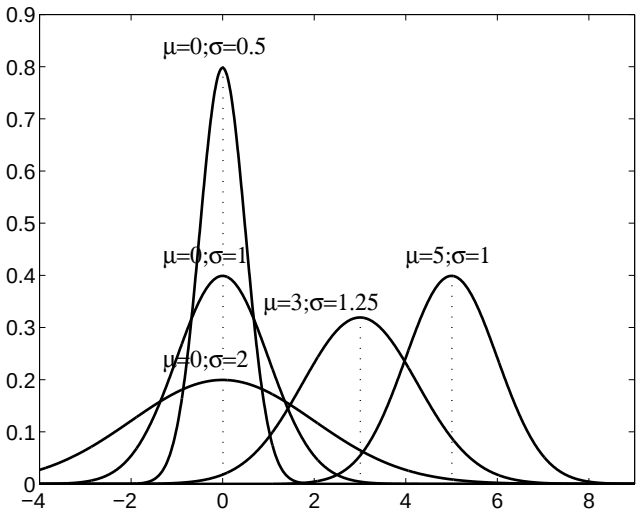
1 Normální rozložení a odvozená rozložení

Náhodná veličina s normálním rozložením $X \sim N(\mu, \sigma^2)$ má dominantní postavení v počtu pravděpodobnosti i v matematické statistice. Vyskytuje se v takových situacích, kdy se ke konstantní střední hodnotě μ přičítá velké množství nezávislých náhodných vlivů, které lehce kolísají kolem nuly. Takto vzniklá variabilita je charakterizována konstantou $\sigma \geq 0$. Normálně rozdělená náhodná veličina je tedy určena dvěma parametry μ a σ^2, kde μ je její střední hodnota a σ^2 je její rozptyl. Speciální případ, kde $\mu = 0$ a $\sigma^2 = 1$ nazýváme standardizované normální rozložení a značíme jej $U \sim N(0, 1)$. Příklady: procentové změny v cenách akcií na dobře fungujících trzích (Eugene Chama, 1960), devizové výplatní poměry měn,...

Ze standardizovaného normálního rozložení U lze různými transformacemi odvodit další rozložení, z nichž se seznámíme s Pearsonovým χ^2-rozložením, Studentovým $t-$rozložením a Fisher-Snedecorovým $F-$rozložením. Tato rozložení nacházejí velké uplatnění především v matematické statistice.

Definice 1.1
O spojité náhodné veličině X říkáme, že má normální rozložení s parametry μ a σ^2, když její hustota je dána vzorcem $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, $x \in \mathbf{R}$. Zkráceně píšeme $X \sim N(\mu, \sigma^2)$.



Distribuční funkci normální náhodné veličiny X vyjádříme
$\forall x \in \mathbf{R} : F(x) = \int\limits_{-\infty}^{x} f(t)\, dt = \int\limits_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\, dt.$

Definice 1.2
Náhodnou veličinu $U \sim N(0, 1)$ nazýváme standardizovaná normální náhodná veličina. Její hustota má tvar $f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$, $u \in \mathbf{R}$
a distribuční funkce má tvar $F(u) = \int\limits_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt.$

Následující věta uvede vlastnosti normálního rozložení.

Věta 1.3

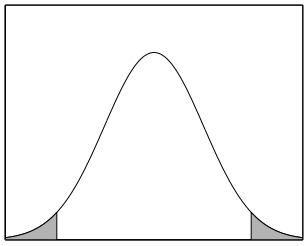
a) Jestliže $X \sim N(\mu, \sigma^2)$, pak $E(X) = \mu$, $D(X) = \sigma^2$.

b) Nechť $a, b \in \mathbf{R},\ b \neq 0$.
Jestliže $X \sim N(\mu, \sigma^2)$ a $Y = a + bX$, pak $Y \sim N(a + b\mu, b^2\sigma^2)$.
[Lineární transformace normální náhodné veličiny normalitu neporuší.]

c) Nechť $X_1, \dots X_n$ jsou stochasticky nezávislé náhodné veličiny,
$X_i \sim N(\mu_i, \sigma_i^2),\ i = 1, \dots n$. Pak $Y = \sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i,\ \sum_{i=1}^{n} \sigma_i^2)$
[Součet nezávislých normálních náhodných veličin je opět normální náhodná veličina.]

d) Nechť $X \sim N(\mu, \sigma^2)$. Pak $U = \frac{X-\mu}{\sigma} \sim N(0, 1)$
[Normální náhodnou veličinu X standardizujeme tak, že od ní odečteme její střední hodnotu a tento rozdíl pak dělíme její směrodatnou odchylkou.]

Poznámka 1.4
Distribuční funkce náhodné veličiny $U \sim N(0, 1)$ je tabelována ve statistických tabulkách pro $u \geq 0$. Jinak se užívá přepočtový vzorec $F(-u) = 1 - F(u)$. Kvantily náhodné veličiny $U \sim N(0, 1)$ se značí u_α a jsou tabelovány pro $\alpha \geq 0,5$. Jinak se užívá přepočtový vzorec $u_\alpha = -u_{1-\alpha}$.



Příklad 1.5
Výsledky u přijímací zkoušky na jistou VŠ jsou normálně rozloženy se střední hodnotou $\mu = 550$ bodů a směrodatnou odchylkou $\sigma = 100$ bodů. Jaká je pravděpodobnost, že náhodně vybraný uchazeč bude mít aspoň 600 bodů?

Řešení
Náhodná veličina X udává bodový výsledek náhodně vybraného uchazeče, $X \sim N(550, 100^2)$

$P(X \geq 600) = 1 - P(X < 600) = 1 - P(X \leq 600) + \overbrace{P(X = 600)}^{0} =$
$= 1 - P(\frac{X-550}{100} \leq \frac{600-550}{100}) = 1 - P(U \leq 0,5) = 1 - F(0,5) = 1 - 0,69146 \doteq 0,31.$

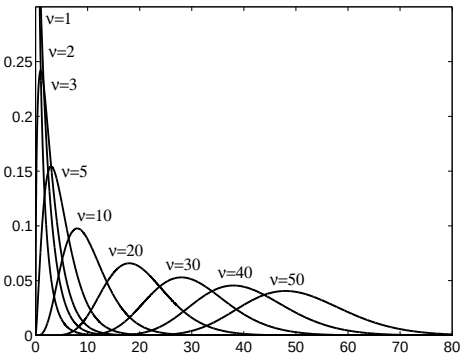
$F(0,5)$ je distribuční funkce standardizovaného normálního rozložení v bodě 0,5 - viz. tabulky.

Příklad 1.6
Nechť $X \sim N(-1, 4)$. Najděte kvantil $K_{0,025}(X)$.

Řešení
$U = \frac{X+1}{2} \sim N(0,1), \qquad K_{0,025}(X) = ?$

$0,025 = P(X \leq K_{0,025}(X)) = P(\frac{X+1}{2} \leq \frac{K_{0,025}(X)+1}{2}) = P(U \leq \frac{K_{0,025}(X)+1}{2})$.
Tedy $\frac{K_{0,025}(X)+1}{2} = u_{0,025}$
Proto $K_{0,025}(X) = 2u_{0,025} - 1 = 2 \cdot (-u_{1-0,025}) - 1 = -2 \cdot u_{0,975} - 1 = -2 \cdot 1,96 - 1 = -4,92$
Nyní budou následovat definice odvozených rozložení (Pearsonovo rozložení, Studentovo rozložení a Fisher-Snedecorovo rozložení) a související příklady. V definicích nepřehlédněte požadavky na nezávislost!

Definice 1.7
Nechť $U_1, \ldots, U_n$ jsou stochasticky nezávislé náhodné veličiny, $U_i \sim N(0,1),\ i = 1, \ldots, n$. Pak náhodná veličina $V = \sum_{i=1}^{n} U_i^2 \sim \chi^2(n)$.
Říkáme, že náhodná veličina V má Pearsonovo rozložení "chí kvadrát"a parametr n nazýváme stupně volnosti.
(Explicitní tvar hustoty lze nalézt např. v příloze A sbírky: Budíková,M.-Osecký,P.-Mikoláš,Š: *Teorie pravděpodobnosti a matematická statistika, Sbírka příkladů,* Masarykova Univerzita, Brno, (1998).)



Poznámka 1.8
$\alpha-$kvantil Pearsonova rozložení s n stupni volnosti značíme $\chi^2_\alpha(n)$. Tyto kvantily jsou tabelovány a pro $n > 30$ užíváme přibližný vztah $\chi^2_\alpha(n) \approx \frac{1}{2}(u_\alpha + \sqrt{2n-1})^2$

Příklad 1.9

a) Nechť $V \sim \chi^2(10)$. Najděte kvantil $\chi^2_{0,975}(10)$.

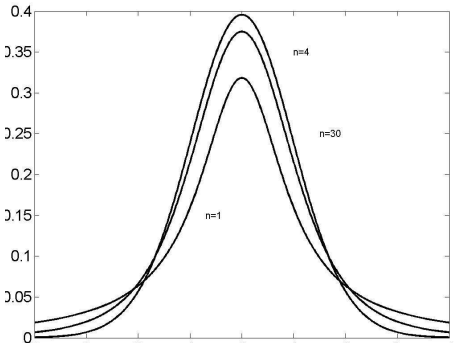
b) Nechť $V \sim \chi^2(3)$. Najděte kvantil $\chi^2_{0,05}(3)$.

Řešení

a) $\chi^2_{0,975}(10) = 20,483$.

b) $\chi^2_{0,05}(3) = 0,352$.

Definice 1.10
Nechť U, V jsou stochasticky nezávislé náhodné veličiny, $U \sim N(0,1),\ V \sim \chi^2(n)$.
Pak náhodná veličina $T = \frac{U}{\sqrt{\frac{V}{n}}} \sim t(n)$. Říkáme, že náhodná veličina T má Studentovo rozložení s n stupni volnosti.
(Explicitní tvar hustoty lze nalézt např. v příloze A sbírky: BUDÍKOVÁ,M.-OSECKÝ,P.-MIKOLÁŠ,Š: *Teorie pravděpodobnosti a matematická statistika, Sbírka příkladů,* Masarykova Univerzita, Brno, (1998).)



Poznámka 1.11
$\alpha-$kvantil Studentova rozložení s n stupni volnosti značíme $t_\alpha(n)$. Tyto kvantily jsou tabelovány. Pro $\alpha < 0,5$ se používá přepočtový vzorec $t_\alpha(n) = -t_{1-\alpha}(n)$ a pro distribuční funkci platí vztah $F(-x) = 1 - F(x)$.

Příklad 1.12

a) Nechť $T \sim t(8)$. Najděte kvantil $t_{0,9}(8)$.

b) Nechť $T \sim t(6)$. Najděte kvantil $t_{0,05}(6)$.

Řešení

a) $t_{0,9}(8) = 1,3968$.

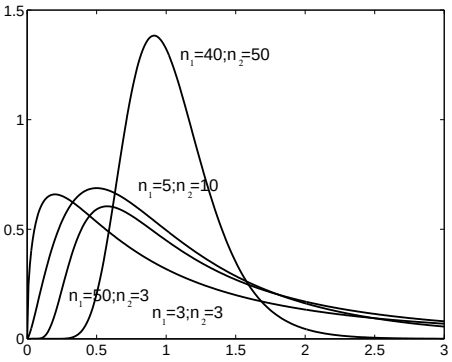
b) $t_{0,05}(6) = -t_{0,95}(6) = -1,9432$.

Příklad 1.13
Nechť $X \sim t(14)$. Určete konstantu c tak, aby platilo: $P(-c < X < c) = 0,9$.

Řešení
$0,9 = P(-c < X < c) = F(c) - F(-c) = F(c) - [1 - F(c)] = 2F(c) - 1$
Tedy $0,9 = 2F(c) - 1 \Rightarrow F(c) = \frac{1,9}{2} = 0,95 \Rightarrow c = t_{0,95}(14) = 1,7613$

Definice 1.14
Nechť V_1, V_2 jsou stochasticky nezávislé náhodné veličiny, $V_1 \sim \chi^2(n_1),\ V_2 \sim \chi^2(n_2)$.
Pak náhodná veličina $F = \frac{V_1/n_1}{V_2/n_2} \sim F(n_1, n_2)$. Říkáme, že náhodná veličina F má Fisher-Snedecorovo rozložení, kde n_1 je počet stupňů volnosti čitatele a n_2 je počet stupňů volnosti

jmenovatele.
(Explicitní tvar hustoty lze nalézt např. v příloze A sbírky: BUDÍKOVÁ,M.-OSECKÝ,P.-MIKOLÁŠ,Š: *Teorie pravděpodobnosti a matematická statistika, Sbírka příkladů,* Masarykova Univerzita, Brno, (1998).)



Poznámka 1.15
$\alpha-$kvantil Fisher-Snedecorova rozložení se stupni volnosti n_1, n_2 značíme $F_\alpha(n_1, n_2)$. Tyto kvantily jsou tabelovány. Pro $\alpha < 0,5$ se používá přepočtový vzorec $F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2,n_1)}$.

Příklad 1.16

a) Nechť $F \sim F(5, 7)$. Najděte kvantil $F_{0,975}(5, 7)$.

b) Nechť $F \sim F(8, 6)$. Najděte kvantil $F_{0,025}(8, 6)$.

Řešení

a) $F_{0,975}(5, 7) = 5,2852$.

b) $F_{0,025}(8, 6) = \frac{1}{F_{0,975}(6,8)} = \frac{1}{4,6517} = 0,215$.

Příklad 1.17
Nechť $X \sim F(5, 8)$. Určete konstantu c tak, aby platilo: $P(X < c) = 0,05$.

Řešení
$0,05 = P(X < c) = F(c)$
Tedy $c = F_{0,05}(5, 8) = \frac{1}{F_{0,95}(8,5)} = \frac{1}{4,8183} = 0,2075$.

Nyní se budeme věnovat náhodnému vektoru s $n-$rozměrným normálním rozložením, pro jednoduchost budeme uvažovat $n = 2$. Konvenci při zapisování náhodných vektorů ilustrujme následovně:

sloupcový vektor náhodných veličin značíme velkým tlustým písmenem, např. $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$

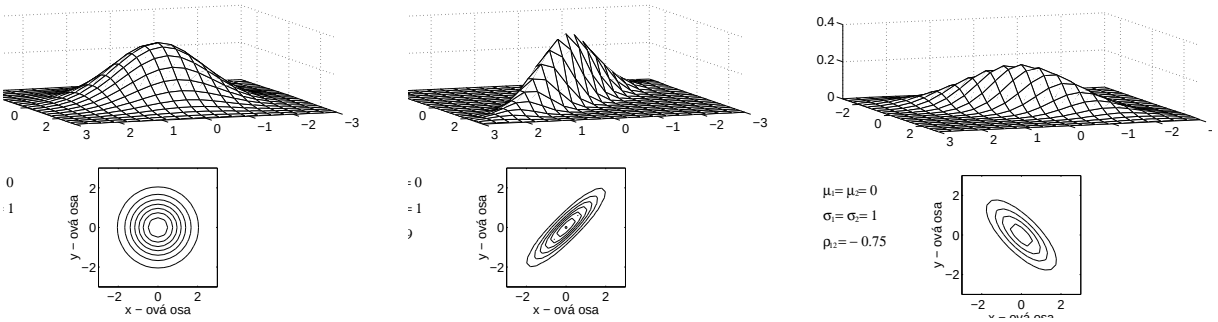
sloupcový vektor konstant značíme malým tlustým písmenem, např. $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$

Definice 1.18
O spojitém náhodném vektoru $\mathbf{X} = \binom{X_1}{X_2}$ říkáme, že má dvojrozměrné normální rozložení s parametry $\mu = \binom{\mu_1}{\mu_2}$ a $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, když jeho hustota je dána vzorcem

$$f(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\cdot\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}, \quad \mathbf{x} \in \mathbf{R}^2.$$

Zkráceně píšeme $\mathbf{X} = \binom{X_1}{X_2} \sim N_2(\mu, \mathbf{\Sigma})$.
Pro $\mu = \binom{0}{0}$ a $\mathbf{\Sigma} = \binom{1\ 0}{0\ 1}$ mluvíme o standardizovaném dvojrozměrném normálním rozložení.

Poznámka 1.19
Význam parametrů je následující:
$\mu_1 = E(X_1),\ \mu_2 = E(X_2),\ \sigma_1^2 = D(X_1),\ \sigma_2^2 = D(X_2),\ \rho = R(X_1, X_2)$



Věta 1.20
Nechť dvojrozměrný vektor $\mathbf{X}$ má dvojrozměrné normální rozložení
$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$.
Potom pro marginální rozložení skalární náhodné veličiny X_i platí: $X_i \sim N(\mu_i, \sigma_i^2),\ i = 1, 2$.
[Složky normálního náhodného vektoru normalitu "podědí".]

Věta 1.21
Nechť $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$ je normální náhodný vektor, nechť $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ je vektor reálných čísel, nechť $\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$ je matice reálných čísel. Potom transformovaný náhodný vektor $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X} \sim N_2(\mathbf{a} + \mathbf{B}\mu,\ \mathbf{B}\mathbf{\Sigma}\mathbf{B}')$.
[Lineární transformace zachovává normalitu.]

Příklad 1.22
Nechť kursy dvou akcií jsou náhodné veličiny $X_1 \sim N(600, 40^2)$, $X_2 \sim N(800, 30^2)$. Kore-

lace $R(X_1, X_2) = -0.4$. Jaká je pravděpodobnost, že index $X_1 + X_2$ nepoklesne pod 1300 bodů?

Řešení

$P(X_1 + X_2 \geq 1300) = ?$

Jelikož X_1, X_2 jsou korelované, nelze užít věty 1.3.c). Abychom mohli užít vět 1.20 a 1.21, musíme nejdříve určit rozložení náhodného vektoru $\mathbf{X} = \binom{X_1}{X_2}$.

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 600 \\ 800 \end{pmatrix}, \begin{pmatrix} 1600 & -480 \\ -480 & 900 \end{pmatrix} \right).$$

(Pro prvek σ_{12} matice Σ platí: $\sigma_{12} = \sigma_{21} = \rho\sigma_1\sigma_2 = -0,4 \cdot 40 \cdot 30 = -480$)

Užijeme-li ve větě 1.21 $\mathbf{a} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ a $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ potom transformovaný náhodný vektor $\mathbf{Y} = \mathbf{a} + \mathbf{BX} = \begin{pmatrix} X_1 + X_2 \\ 0 \end{pmatrix}$ zůstává normálně rozložený a dle věty 1.20 je normálně rozložená i každá jeho složka. Tedy náhodná veličina $Z = X_1 + X_2$ má normální rozložení a pro její parametry platí:

$E(Z) = E(X_1 + X_2) = E(X_1) + E(X_2) = 600 + 800 = 1400$

$D(Z) = D(X_1 + X_2) = D(X_1) + D(X_2) + 2C(X_1, X_2) = 1600 + 900 - 2 \cdot 480 = 1540$

$Z \sim N(1400, 1540)$

$$P(X_1 + X_2 \geq 1300) = P(Z \geq 1300) = 1 - P(Z \leq 1300) + \overbrace{P(Z = 1300)}^{0} =$$

$$= 1 - P\left(\frac{Z-1400}{\sqrt{1540}} \leq \frac{1300-1400}{\sqrt{1540}}\right) = 1 - P(U \leq -2,55) = P(U \leq 2,55) = 0,9946.$$

Index $X_1 + X_2$ nepoklesne pod 1300 bodů s pravděpodobností 0,9946.

2 Základní pojmy matematické statistiky

Počet pravděpodobnosti a matematická statistika jsou spolu úzce svázány předmětem svého zkoumání. Při popisování reality respektují působení stochasticky stabilních náhodných vlivů. Na rozdíl od počtu pravděpodobnosti se v matematické statistice setkáváme s větší neznalostí zkoumané reality v tom smyslu, že musíme pracovat ne s jednou pravděpodobností P, ale celou třídou předem přípustných pravděpodobností, aniž bychom věděli, která z nich odpovídá skutečnosti. Na základě pozorování a vyhodnocování statistických údajů se snažíme alespoň přibližně identifikovat tu pravděpodobnostní míru, která odpovídá pravdivé variantě zkoumané reality.

Uvažujme urnu s 10-ti koulemi o nichž víme jen to, že jsou buď černé, nebo bílé, ale nevíme, kolik je kterých a do urny se nesmíme podívat. Můžeme jen mnohokrát (s vracením) losovat jednu kouli a na základě výsledků losování odhadnout neznámý počet čerých koulí. Tento odhad bude věrohodný, pokud počet losování bude dostatečně velký.
Představme si, že jsme 100-krát losovali jednu kouli (s vracením) a jen v 9-ti případech jsme vylosovali černou kouli. Zdá se být velmi pravděpodobné, že černých koulí je méně, než bílých. Můžeme tvrdit i víc. Kandidáty na neznámý počet černých koulí jsou čísla:1,2,...,9. Nejvěrohodnějším kandidátem se v této situaci jeví číslo 1.

Matematická statistika se (mimo jiné) zabývá:

a) *teorií odhadu*, t.j. odhadováním některých parametrů (např. počtu černých koulí v urně).

b) *teorií testování hypotéz*, (např. testováním hypotézy, že v urně je c černých koulí).

Základem obou procedur jsou statistické údaje uspořádané do datových souborů. K tomu, aby zmiňované procedury dávali věrohodné závěry, je třeba, aby losování splňovalo jisté podmínky. Jde tedy o to, jak správně sbírat statistické údaje. S tím souvisí pojem náhodného výběru.

Definice 2.1

(i.) Nechť $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení $L(\vartheta)$, tedy $X_i \sim L(\vartheta),\ i = 1, \ldots, n$. Potom říkáme, že $X_1, \ldots, X_n$ je *náhodný výběr* rozsahu n z rozložení $L(\vartheta)$. Číselné realizace $x_1, \ldots, x_n$ uspořádané do sloupcového vektoru představují datový soubor.

(ii.) Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ jsou stochasticky nezávislé náhodné vektory, které mají všechny stejné rozložení $L_2(\vartheta)$. Potom říkáme, že $(X_1, Y_1), \ldots, (X_n, Y_n)$ je *náhodný výběr* rozsahu n z dvourozměrného rozložení $L_2(\vartheta)$. Číselné realizace $(x_1, y_1), \ldots, (x_n, y_n)$ uspořádané do matice $n \times 2$ představují dvourozměrný datový soubor.

(iii.) Analogicky lze definovat i náhodný výběr rozsahu n z $p-$rozměrného rozložení $L_p(\vartheta)$, $p \geq 3$.

(iv.) Libovolná funkce T náhodného výběru, (resp. několika náhodných výběrů) se nazývá *statistika*.

Důsledek 2.2
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení s distribuční funkcí $F(x)$. Pak pro simultánní distribuční funkce $F_*(\mathbf{x}) = F_*(x_1, \ldots, x_n)$ náhodného vektoru $(X_1, \ldots, X_n)$ platí: $F_*(\mathbf{x}) = F(x_1) \cdot F(x_2) \cdot \ldots \cdot F(x_n)$

Následující definice uvede důležité a často používané statistiky. Znalost těchto statistik, schopnost je používat a správně je interpretovat je v tomto kurzu zcela zásadní!!!

Definice 2.3

(i.) Nechť $X_1, \ldots, X_n$ je náhodný výběr, $n \geq 2$
- Statistika $M = \frac{1}{n} \sum_{i=1}^{n} X_i$ se nazývá *výběrový průměr*
- Statistika $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M)^2$ se nazývá *výběrový rozptyl*
- Statistika $S = \sqrt{S^2}$ se nazývá *výběrová směrodatná odchylka*
- Statistika $F_n(x) = \frac{1}{n} \cdot card\{i,\ X_i \leq x\},\ x \in \mathbf{R}$ se nazývá *hodnota výběrové distribuční funkce v bodě* x [pro libovolné, ale pevně zvolené reálné číslo x znamená $card\{i,\ X_i \leq x\}$ počet těch realizací veličin náhodného vektoru, které nepřekročí x.]

(ii.) Nechť $X_{11}, \ldots, X_{1n_1}; \ldots; X_{p1}, \ldots, X_{pn_p}$ je p stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2, \ldots, n_p \geq 2$,. Celkový rozsah je $n = \sum_{j=1}^{p} n_j$. Označme $M_1, \ldots, M_p$ výběrové průměry a $S_1^2, \ldots, S_p^2$ výběrové rozptyly jednotlivých výběrů. Nechť $c_1, \ldots, c_p$ jsou reálné konstanty, z nichž aspoň jedna je nenulová.
- Statistika $\sum_{j=1}^{p} c_j M_j$ se nazývá *lineární kombinace výběrových průměrů*
- Statistika $S_*^2 = \frac{\sum_{j=1}^{p} (n_j - 1) S_j^2}{n-p}$ se nazývá *vážený průměr výběrových rozptylů*

(iii.) Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozložení. Označme $M_1 = \frac{1}{n} \sum_{i=1}^{n} X_i,\ M_2 = \frac{1}{n} \sum_{i=1}^{n} Y_i$ výběrové průměry, $S_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_1)^2$, $S_2^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - M_2)^2$ výběrové rozptyly.
- Statistika $S_{12} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_1)(Y_i - M_2)$ se nazývá *výběrová kovariance*
- Statistika $R_{12} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(X_i - M_1)}{S_1} \frac{(Y_i - M_2)}{S_2} = \frac{S_{12}}{S_1 S_2}$ pro $S_1 \cdot S_2 > 0$
  se nazývá *výběrový koeficient korelace*. (Je-li některá z veličin S_1, S_2 rovna nule, výběrový koeficient korelace se nedefnuje.)

[$M, S^2, S, S_{12}, R_{12}$ jsou náhodné veličiny vzniklé transformací náhodného výběru. Až poté, co se náhodný výběr realizuje konkrétními čísly, získáme následně číselné realizace uvedených statistik. Tyto číselné realizace značíme malými písmeny $m, s^2, s, s_{12}, r_{12}$ a odpovídají číselným charakteristikám v popisné statistice s tím rozdílem, že v případě rozptylu,

směrodatné odchylky, kovariance a koeficientu korelace je multiplikativní konstanta před sumou $\frac{1}{n-1}$ a ne $\frac{1}{n}$, jak tomu bylo v popisné statistice.]

Příklad 2.4

10× nezávisle na sobě byla měřena jistá neznámá konstanta μ. Výsledky měření jsou: 2; 1,8; 2,1; 2,4; 1,9; 2,1; 2; 1,8; 2,3; 2,2. Tyto výsledky považujeme za realizace náhodného výběru $X_1, \ldots, X_{10}$. Vypočtěte m, s^2 a hodnoty výběrové distribuční funkce $F_{10}(x)$.

Řešení

$$\begin{aligned} m &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(2 + 1,8 + \ldots + 2,2) = 2,06 \\ s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2mx_i + m^2) = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} m^2] = \\ &= \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - 2mnm + nm^2] = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - nm^2] = \\ &= \frac{1}{9}(2^2 + 1,8^2 + \ldots + 2,2^2 - 10 \cdot 2,06^2) = 0,0404 \\ s &= \sqrt{s^2} = \sqrt{0,0404} = 0,2011 \end{aligned}$$

Pro usnadnění výpočtu $F_{10}(x)$ uspořádáme měření vzestupně:
1,8; 1,8; 1,9; 2; 2; 2,1; 2,1; 2,2; 2,3; 2,4;

pro $x < 1,8: \ F_{10}(x) = 0$
pro $1,8 \le x < 1,9: \ F_{10}(x) = 0,2$
pro $1,9 \le x < 2: \ F_{10}(x) = 0,3$
pro $2 \le x < 2,1: \ F_{10}(x) = 0,5$
pro $2,1 \le x < 2,2: \ F_{10}(x) = 0,7$
pro $2,2 \le x < 2,3: \ F_{10}(x) = 0,8$
pro $2,3 \le x < 2,4: \ F_{10}(x) = 0,9$
pro $x \ge 2,4: \ F_{10}(x) = 1$

Příklad 2.5

U 11 náhodně vybraných aut jisté značky bylo zjišťováno jejich stáří v letech (náhodná veličina X) a cena v korunách (náhodná veličina Y). Výsledky jsou v následující tabulce:

X	5	4	6	5	5	5	6	6	2	7	7
Y	85	103	70	82	89	98	66	95	169	70	48

Vypočtěte a interpretujte r_{12}.

Řešení

$$\begin{aligned} m_1 &= \frac{1}{11}(5 + 4 + \ldots + 7) = 5,28 \\ m_2 &= \frac{1}{11}(85 + \ldots + 48) = 88,63 \\ s_1^2 &= \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - nm_1^2] = \frac{1}{10}(5^2 + 4^2 + \ldots + 7^2 - 11 \cdot 5,28^2) = 2,02 \\ s_2^2 &= \frac{1}{n-1}[\sum_{i=1}^{n} y_i^2 - nm_2^2] = \frac{1}{10}(85^2 + 103^2 + \ldots + 48^2 - 11 \cdot 88,63^2) = 970,85 \\ s_{12} &= \frac{1}{n-1}[\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2)] = \frac{1}{n-1}[\sum_{i=1}^{n} x_i y_i - nm_1 m_2] = \\ &= \frac{1}{10}(5 \cdot 85 + 4 \cdot 103 + \ldots + 7 \cdot 48 - 11 \cdot 5,28 \cdot 88,63) = -40,89 \\ r_{12} &= \frac{s_{12}}{s_1 s_2} = \frac{-40,89}{\sqrt{2,02}\sqrt{970,85}} = -0,92 \end{aligned}$$

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

Poznámka 2.6
V řešení předchozích příkladů byl použit výpočetní tvar pro rozptyl a kovarianci:
$S^2 = \frac{1}{n-1}[\sum_{i=1}^{n} X_i^2 - nM^2]$ $\qquad$ $S_{12} = \frac{1}{n-1}[\sum_{i=1}^{n} X_i Y_i - nM_1M_2]$ $\square$

V následující větě budou uvedeny důležité vlastnosti často užívaných statistik. Vlastnosti uvedené v 1. odstavci budou odvozeny ve cvičení.

Věta 2.7

1.) Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení se střední hodnotou μ, s rozptylem σ^2 a distribuční funkcí $F(x)$. Potom platí:
$E(M) = \mu \qquad D(M) = \frac{\sigma^2}{n},\ n \geq 2 \qquad E(S^2) = \sigma^2$
pro pevné $x \in \mathbf{R}: \quad E[F_n(x)] = F(x), \qquad D[F_n(x)] = \frac{F(x)(1-F(x))}{n}$

2.) Nechť $X_{11}, \ldots, X_{1n_1}; \ldots; X_{p1}, \ldots, X_{pn_p}$ je p stochasticky nezávislých náhodných výběrů se středními hodnotami $\mu_1, \ldots, \mu_p$ a stejným rozptylem σ^2 pro všech p výběrů. Označme celkový rozsah $n = \sum_{j=1}^{p} n_j$. Dále nechť $c_1, \ldots, c_p$ jsou reálné konstanty, z nichž aspoň jedna je nenulová. Potom platí:
$E(\sum_{j=1}^{p} c_j M_j) = \sum_{j=1}^{p} c_j \mu_j \qquad E(S_*^2) = \sigma^2$

3.) Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Potom platí:
$E(S_{12}) = \sigma_{12}, \qquad$ avšak $\qquad E(R_{12}) \approx \rho$ (aproximace je vhodná pro $n \geq 30$)

Poznámka 2.8
Metody matematické statistiky velmi často slouží k vyhodnocování výsledků pokusů. Aby tyto výsledky byli správně vyhodnoceny, je důležité pokusy správně naplánovat. Uvedeme nejzákladnější typy uspořádání pokusu.

a) Jednoduché pozorování: náhodná veličina X je pozorována za týchž podmínek. Tuto situaci charakterizuje jeden náhodný výběr $X_1, \ldots, X_n$.

b) Dvojné pozorování: náhodná veličina X je pozorována za dvojích různých podmínek. Setkáváme se s dvěma odlišnými strategiemi uspořádání takovéhoto pokusu:

- Dvouvýběrové porovnávání: Tuto situaci charakterizují dva nezávislé náhodné výběry $X_{11}, \ldots, X_{1n_1};\ X_{21}, \ldots, X_{2n_2}$, které mohou být s různými rozsahy.
- Párové porovnávání: Tuto situaci charakterizuje jeden náhodný výběr $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ z dvourozměrného rozložení. V tomto případě přecházíme k rozdílovému náhodnému výběru $Z_1, \ldots, Z_n$; kde $Z_i = X_{i1} - X_{i2},\ i = 1, 2, \ldots, n$. Takto přejdeme k jednoduchému pozorování.

c) Mnohonásobné pozorování: náhodná veličina X je pozorována za $p \geq 3$ různých podmínek. Setkáváme se s dvěma odlišnými strategiemi uspořádání takovéhoto pokusu:

- Mnohovýběrové porovnávání: Tuto situaci charakterizuje p nezávislých náhodných výběrů $X_{11}, \ldots, X_{1n_1}; \ldots;\ X_{p1}, \ldots, X_{pn_p}$, které mohou být s různými rozsahy.
- Blokové porovnávání: Tuto situaci charakterizuje jeden náhodný výběr $(X_{11}, \ldots, X_{1p}), \ldots, (X_{n1}, \ldots, X_{np})$ z $p-$rozměrného rozložení.

3 Bodové a intervalové odhady parametrů a parametrických funkcí

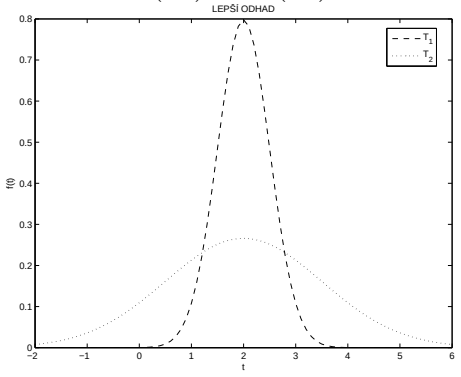
Uvažujme náhodný výběr $X_1, \ldots, X_n \sim L(\vartheta)$, přičemž parametr ϑ tohoto rozložení L je pro nás pozorovatele neznámá konstanta. Ovšem informace o této neznáme konstantě je ukryta a náhodnými vlivy "zamaskována"v našem náhodném výběru. Úkolem bodových i intervalových odhadů je tuto neznámou konstantu odhalit.
Bodový odhad spočívá v nahrazení neznáme hodnoty parametru ϑ takovou statistikou $T = T(X_1, \ldots, X_n)$, jejíž číselné realizace jsou "dostatečně blízko"neznámému parametru ϑ.
Jinou možností je sestrojit interval (D, H), který s velkou (a předem uživatelem stanovenou) pravděpodobností pokrývá neznámý parametr ϑ. Meze tohoto intervalu jsou statistikami náhodného výběru, tedy $D = D(X_1, \ldots, X_n),\ H = H(X_1, \ldots, X_n)$.

Definice 3.1
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\vartheta)$. Množina všech přípustných hodnot, jichž může parametr ϑ nabývat, se nazývá parametrický prostor a značí se Θ. Libovolná funkce $h(\vartheta)$ se nazývá parametrická funkce.

Definice 3.2
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\vartheta)$, nechť $h(\vartheta)$ je parametrická funkce a nechť $T, T_1, T_2, \ldots$ jsou statistiky.

(i.) Statistika T je *nestranným odhadem* parametrické funkce $h(\vartheta)$, jestliže
$\forall \vartheta \in \Theta :\ E(T) = h(\vartheta)$
[Při nestranném odhadu nedochází k systematickému nadhodnocování, nebo podhodnocování parametru, či parametrické funkce.]

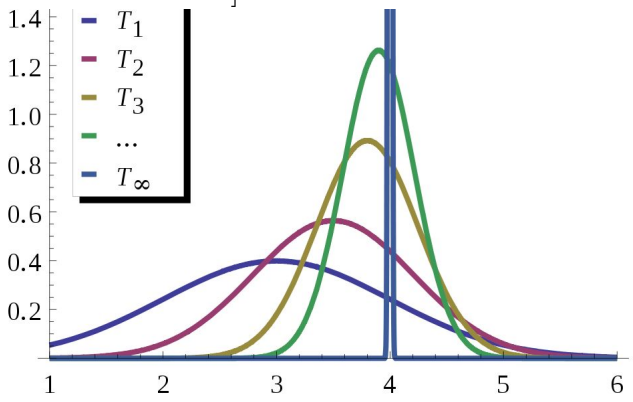
(ii.) Nechť T_1, T_2 jsou dva nestranné odhady téže parametrické funkce $h(\vartheta)$. Řekneme, že odhad T_1 je *lepší*, než odhad T_2, jestliže
$\forall \vartheta \in \Theta :\ D(T_1) < D(T_2)$



(iii.) Posloupnost $T_1, T_2, \ldots, T_n, \ldots$ se nazývá posloupnost *asymptoticky nestranných* odhadů parametrické funkce $h(\vartheta)$, jestliže
$\forall \vartheta \in \Theta :\ \lim_{n\to\infty} E(T_n) = h(\vartheta)$
[Když $E(T) \neq h(\vartheta)$, dochází ke zkreslení odhadu a mluvíme o vychýleném odhadu.

Pokud se toto zkreslení s rostoucím n zmenšuje, je statistika T asymptoticky nestranným odhadem parametru, či parametrické funkce.]

(iv.) Posloupnost $T_1, T_2, \ldots, T_n, \ldots$ se nazývá posloupnost *konzistentních* odhadů parametrické funkce $h(\vartheta)$, jestliže konverguje podle pravděpodobnosti k $h(\vartheta)$, t.j.
$\forall \vartheta \in \Theta,\ \forall \varepsilon > 0 : \lim_{n\to\infty} P(|T_n - h(\vartheta)| < \varepsilon) = 1$
[Statistika T je konzistentní, jestliže s rostoucím rozsahem výběru n roste pravděpodobnost, že absolutní odchylka odhadu T od parametrické funkce $h(\vartheta)$ klesne pod libovolně malé ε.]



[Výraz $\forall \vartheta \in \Theta$: můžete názorně číst "ať je pravda o parametru jakákoliv, platí..."Při prvním čtení definice si na místo parametrické funkce $h(\vartheta)$ dosaďte přímo jen parametr ϑ.]
Definice vyjmenovává žádoucí vlastnosti odhadů, přičemž požadavek na konzistenci odhadu je nejdůležitější.

Důsledek 3.3
Z nestrannosti odhadu T_n plyne i asymptotická nestrannost. Platí-li navíc $\lim_{n\to\infty} D(T_n) = 0$, pak z asymptotické nestrannosti plyne konzistence. Tedy:
$\lim_{n\to\infty} E(T_n) = h(\vartheta) \wedge \lim_{n\to\infty} D(T_n) = 0$, pak T_n je konzistentní odhad parametrické funkce $h(\vartheta)$.

Ve 3. kapitole jsem si zavedli statistiky $M, S^2, S_{12}, R_{12}, \ldots$. Nyní posoudíme, jestli tyto statistiky mají některé ze zmíněných žádoucích vlastností.

Věta 3.4
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení, které má střední hodnotu μ, rozptyl σ^2 a distribuční funkci $F(x)$. Nechť M_n je výběrový průměr, S_n^2 je výběrový rozptyl a $F_n(x)$ je hodnota výběrová distribuční funkce pro libovolné pevně dané x. Potom platí:

1. – M_n je nestranným odhadem parametru μ. [Tedy $\forall \mu \in \mathbf{R} : E(M_n) = \mu$]
   – S_n^2 je nestranným odhadem parametru σ^2. [Tedy $\forall \sigma \geq 0 : E(S_n^2) = \sigma^2$]
   – $F_n(x)$ je nestranným odhadem $F(x)$ pro libovolné pevně dané $x \in \mathbf{R}$.
   [Tedy $\forall x \in \mathbf{R} : E(F_n(x)) = F(x)$]

2. – $M_1, \ldots, M_n, \ldots$ je posloupnost konzistentních odhadů parametru μ.
   – $S_1^2, \ldots, S_n^2, \ldots$ je posloupnost konzistentních odhadů parametru σ^2.
   – $F_1(x), \ldots, F_n(x), \ldots$ je posloupnost konzistentních odhadů $F(x)$ pro libovolné pevně dané $x \in \mathbf{R}$.

Poznámka 3.5
Výběrová směrodatná odchylka S není !! nestranným odhadem parametru σ s jedinou výjimkou, kdy má S degenerované rozložení, tedy kdy je rovna konstantě.
[Kdyby S byla nestranným odhadem parametru σ, pak $E(S) = \sigma$. Potom $D(S) = E(S^2) - (E(S))^2 = \sigma^2 - \sigma^2 = 0$. Nulový rozptyl ukazuje na degenerované rozložení.]

Věta 3.6
Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12}. Potom výběrová kovariance S_{12} je nestranným odhadem parametru σ_{12}.
[Tedy $\forall \sigma_{12} \in \mathbf{R} : E(S_{12}) = \sigma_{12}$]

Dosud jsme se věnovali bodovým odhadům a jejich vlastnostem. Nyní přejdeme k intervalovým odhadům. Jejich předností proti bodovým odhadům je to, že pomocí pravděpodobnostní míry umíme číselně ohodnotit, "jak moc jim můžeme věřit".

Definice 3.7
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $h(\vartheta)$ je parametrická funkce, číslo $\alpha \in (0, 1)$, $D = D(X_1, \ldots, X_n)$, $H = H(X_1, \ldots, X_n)$ jsou statistiky.

(i.) Interval (D, H) se nazývá $100(1-\alpha)\%$ (oboustranný) interval spolehlivosti pro $h(\vartheta)$, jestliže
$\forall \vartheta \in \Theta : \; P(D < h(\vartheta) < H) \geq 1 - \alpha$
[Je skoro jisté, že náhodný interval obsahuje bod $h(\vartheta)$.]

(ii.) Interval (D, ∞) se nazývá $100(1-\alpha)\%$ (levostranný) interval spolehlivosti pro $h(\vartheta)$, jestliže
$\forall \vartheta \in \Theta : \; P(D < h(\vartheta)) \geq 1 - \alpha$

(iii.) Interval $(-\infty, H)$ se nazývá $100(1-\alpha)\%$ (pravostranný) interval spolehlivosti pro $h(\vartheta)$, jestliže
$\forall \vartheta \in \Theta : \; P(h(\vartheta) < H) \geq 1 - \alpha$

(iv.) Číslo α se nazývá riziko [volíme jej zpravidla blízké nule, nejčastěji 0,05; 0,01; 0,1], číslo $(1-\alpha)$ se nazývá spolehlivost. Statistiku D nazýváme dolní odhad, statistiku H nazýváme horní odhad.

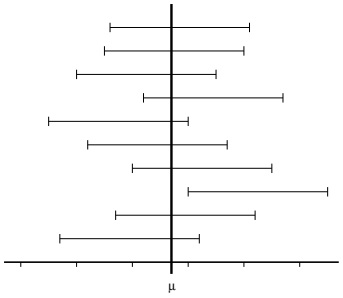
Poznámka 3.8
Volba oboustranného, levostranného, nebo pravostranného intervalu závisí na konkrétní situaci. Např. oboustranný interval spolehlivosti použije konstruktér, kterého zajímá dolní i horní hranice pro skutečnou délku μ nějaké součástky. Levostranný interval spolehlivosti použije výkupčí drahých kovů, který potřebuje znát dolní mez pro skutečný obsah zlata

μ v kupovaném slitku. Pravostranný interval spolehlivosti použije chemik, který potřebuje znát horní mez pro obsah nečistot μ v analyzovaném vzorku.

Poznámka 3.9 Postup při konstrukci intervalu spolehlivosti

1. Nalezneme statistiku V, která je nestranným bodovým odhadem parametrické funkce $h(\vartheta)$.
2. Nalezneme tzv. pivotovu statistiku W, která je monotónní funkcí bodového odhadu V, je monotónní funkcí $h(\vartheta)$, její rozložení je známé a na $h(\vartheta)$ nezávislé. Nalezneme její kvantily $w_{\alpha/2}$ a $w_{1-\alpha/2}$ tak, že platí:
   $\forall \vartheta \in \Theta: \ P(w_{\alpha/2} < W < w_{1-\alpha/2}) \geq 1 - \alpha$.
3. Nerovnost $w_{\alpha/2} < W < w_{1-\alpha/2}$ převedeme ekvivalentními úpravami na nerovnost $D < h(\vartheta) < H$.
4. Statistiky D a H nahradíme jejich číselnými realizacemi d a h a získáme tak $100(1-\alpha)\%$ empirický interval spolehlivosti pro $h(\vartheta)$, který pokrývá $h(\vartheta)$ s pravděpodobností alespoň $1 - \alpha$.

Poznámka 3.10
Jestliže $100\times$ nezávisle na sobě uskutečníme náhodný výber z rozložení se střední hodnotou μ a pokaždé sestrojíme 95% empirický interval spolehlivosti pro μ, pak přibližně v 95-ti případech bude ležet parametr μ v intervalech spolehlivosti a asi v 5-ti případech interval spolehlivosti μ nepokryje.



Postup při konstrukci intervalu spolehlivosti si ukážeme na následujícím příkladě.

Příklad 3.11
Nechť $X_1, \ldots, X_n$ je náhodný výběr z normálního rozložení $N(\mu, \sigma^2)$, kde $n \geq 2$ a numerická hodnota parametru σ je známá. Sestrojte $100(1-\alpha)\%$ interval spolehlivosti pro parametr μ.

Řešení

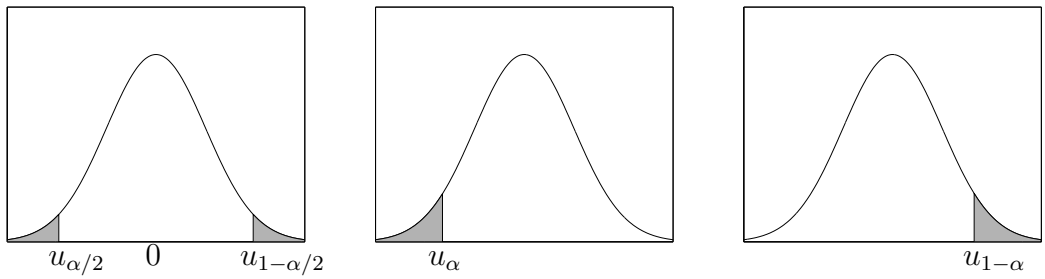
1. Nestranným bodovým odhadem V parametru μ je statistika
   $V = M = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$.
   $[E(M) = \mu, \ D(M) = \frac{\sigma^2}{n}$ a lineární kombinace nezávislého náhodného vektoru zachovává normalitu.]

2. Vhodnou pivotovou statistikou W pro naše zadání je $W = U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
   Kvantily $w_{\alpha/2}$, $w_{1-\alpha/2}$ jsou v tomto případě kvantily normálního standardizovaného rozložení, tedy $u_{\alpha/2} = -u_{1-\alpha/2}$ a $u_{1-\alpha/2}$. Proto
   $\forall \vartheta \in \Theta : 1 - \alpha \leq P(u_{\alpha/2} < U < u_{1-\alpha/2})$.

3. Uvedenou nerovnost budeme ekvivalentně upravovat tak, abychom odhadovaný parametr μ osamostatnili uprostřed nerovnosti.
   $1 - \alpha \leq P(u_{\alpha/2} < U < u_{1-\alpha/2}) = P(\underbrace{M - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2}}_{D} < \mu < \underbrace{M + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2}}_{H})$.

4. Pokud bychom měli reálná data, tak bychom pokračovali dosazením konkrétních čísel do odvozených odhadů.

V případě levo- nebo pravostranného intervalu spolehlivosti se riziko α nepůlí, ale zůstává soustředěné jenom na jednom okraji rozložení. Tak v předchozím příkladě by byl
levostranný interval spolehlivosti tvaru $(D, \infty) = (M - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha}, \infty)$
pravostranný interval spolehlivosti tvaru $(-\infty, H) = (-\infty, M + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha})$.



Příklad 3.12
Na základě náhodného výběru rozsahu 10 z rozložení $N(\mu; 0,04)$ byla vypočtena realizace výběrového průměru $m = 2,06$. Najděte 95% empirický interval spolehlivosti pro parametr μ, a to a) oboustranný, b) levostranný, c) pravostranný.

Řešení
$\sigma = 0,2;\ n = 10;\ m = 2,06;\ \alpha = 0,05;\ u_{1-\alpha/2} = u_{0,975} = 1,96;\ u_{1-\alpha} = u_{0,95} = 1,64$
ad a)
$d = m - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2} = 2,06 - \frac{0,2}{\sqrt{10}} \cdot 1,96 = 1,94$
$h = m + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2} = 2,06 + \frac{0,2}{\sqrt{10}} \cdot 1,96 = 2,18$
$P(1,94 < \mu < 2,18) \geq 0,95$
Tedy $\mu \in (1,94; 2,18)$ s pravděpodobností alespoň 0,95.

ad b)
$d = m - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha} = 2,06 - \frac{0,2}{\sqrt{10}} \cdot 1,64 = 1,96$
$P(1,96 < \mu) \geq 0,95$
Tedy $\mu > 1,96$ s pravděpodobností alespoň 0,95.

ad c)
$h = m + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha} = 2,06 + \frac{0,2}{\sqrt{10}} \cdot 1,64 = 2,16$
$P(\mu < 2,16) \geq 0,95$
Tedy $\mu < 2,16$ s pravděpodobností alespoň 0,95.

Poznámka 3.13
Nechť (d, h) je $100(1-\alpha)\%$ empirický interval spolehlivosti pro parametrickou funkci $h(\vartheta)$. Označme $\Delta = h - d$. Číslo Δ nazýváme šířka intervalu spolehlivosti.

a) Při konstantním riziku α klesá šířka Δ s rostoucím rozsahem výběru n.

b) Při konstantním rozsahu n klesá šířka Δ s rostoucím rizikem α.

Příklad 3.14
Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde σ^2 známe. Jaký musí být minimální rozsah výběru, aby šířka $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro parametr μ nepřekročila číslo δ?

Řešení
Požadujeme, aby šířka int. spol $\Delta \leq \delta$. Tedy
$\delta \geq \Delta = h - d = m + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2} - (m - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2}) = \frac{2\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2}$

$\sqrt{n} \geq \frac{2\sigma}{\delta} \cdot u_{1-\alpha/2}$

$n \geq \frac{4\sigma^2 u_{1-\alpha/2}^2}{\delta^2}$
Za rozsah výběru volíme nejmenší přirozené číslo, které vyhovuje poslední nerovnosti.

Příklad 3.15
V příkladu 3.12 a) se uživateli zdá 95% interval spolehlivosti pro μ (1,94; 2,18) příliš široký. Přál by si, aby šířka tohoto intervalu nepřesáhla číslo 0,16 a riziko α zvětšovat nechce. Co byste navrhli?

Řešení
Šířku intervalu ovlivníme změnou rozsahu výběru n.
$\delta = 0,16,\ n = ?,\ \sigma = 0,2,\ u_{0,975} = 1,96$
$n \geq \frac{4\sigma^2 u_{1-\alpha/2}^2}{\delta^2} = \frac{4 \cdot 0,04 \cdot 1,96^2}{0,16^2} = 24,01$

Pro $n = 25$ jsou požadavky uživatele splněny.

4 Úvod do testování hypotéz

Častým úkolem statistika je na základě dat ověřit předpoklady o parametrech, nebo typu rozložení, z něhož pochází náhodný výběr. Takovému tvrzení se říká nulová hypotéza a stanovujeme ji předem, bez přihlédnutí ke konkrétním datům. Obvykle vyjadřuje dosavadní představy o určité skutečnosti, nebo odráží dosavadní stav poznání v dané oblasti. Nesouhlas s nulovou hypotézou vyjadřuje alternativní hypotéza, která je formulována tak, aby mohla platit jenom jedna z těchto dvou hypotéz. Pravdivost alternativní hypotézy by znamenala objevení nějakých nových skutečností, nebo zásadnější změnu v dosavadních představách. Např. výzkumník by chtěl na základě dat prověřit tezi (nový objev), že pasivní kouření škodí zdraví. Jako nulovou hypotézu tedy položí tvrzení, že pasivní kouření neškodí zdraví a proti nulové hypotéze postaví alternativní, že pasivní kouření škodí zdraví. Testováním hypotéz se myslí rozhodovací postup, při kterém se na základě náhodného výběru provede rozhodnutí, zda nulovou hypotézu nezamítám, nebo zda ji zamítám ve prospěch alternativní hypotézy.

Definice 4.1
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\vartheta)$, kde parametr $\vartheta \in \Theta$ neznáme, nechť $h(\vartheta)$ je parametrická funkce a $c \in \mathbf{R}$ je konstanta.

(i.) $H_0 : \ h(\vartheta) = c \quad$ proti $\quad H_1 : \ h(\vartheta) \neq c$

Tvrzení $H_0 : \ h(\vartheta) = c$ se nazývá jednoduchá nulová hypotéza,
tvrzení $H_1 : \ h(\vartheta) \neq c$ se nazývá složená oboustranná alternativní hypotéza.

(ii.) $H_0 : \ h(\vartheta) \geq c \quad$ proti $\quad H_1 : \ h(\vartheta) < c$

Tvrzení $H_0 : \ h(\vartheta) \geq c$ se nazývá složená pravostranná nulová hypotéza,
tvrzení $H_1 : \ h(\vartheta) < c$ se nazývá složená levostranná alternativní hypotéza.

(iii.) $H_0 : \ h(\vartheta) \leq c \quad$ proti $\quad H_1 : \ h(\vartheta) > c$

Tvrzení $H_0 : \ h(\vartheta) \leq c$ se nazývá složená levostranná nulová hypotéza,
tvrzení $H_1 : \ h(\vartheta) > c$ se nazývá složená pravostranná alternativní hypotéza.

Testováním H_0 proti H_1 rozumíme rozhodovací pravidlo, které na základě náhodného výběru $X_1, \ldots, X_n$ zamítne, či nezamítne platnost H_0.

Definice 4.2
Při testování H_0 proti H_1 se můžeme dopustit jedné ze dvou druhů chyb:
Chyba prvního druhu, jejíž pravděpodobnost značíme α znamená, že H_0 zamítneme, přestože platí.
Chyba druhého druhu, jejíž pravděpodobnost značíme β znamená, že H_0 nezamítneme, i když neplatí.

	H_0 nezamítáme	H_0 zamítáme
H_0 platí	správné rozhodnutí $P(H_0$ nezamítám$\|H_0$ platí$) = 1-\alpha$	chyba prvního druhu $P(H_0$ zamítám$\|H_0$ platí$) = \alpha$
H_0 neplatí	chyba druhého druhu $P(H_0$ nezamítám$\|H_0$ neplatí$) = \beta$	správné rozhodnutí $P(H_0$ zamítám$\|H_0$ neplatí$) = 1-\beta$

Pravděpodobnost chyby prvního druhu α nazýváme *hladina významnosti testu*.
Číslo $1-\beta$ označuje pravděpodobnost jevu, že nepravdivou hypotézu H_0 správně zamítneme. Toto číslo nazýváme *síla testu*. [Statistikovým přáním je, aby síla testu $1-\beta$ byla co největší a hladina α významnosti testu byla co nejmenší. Bohužel s klesajícím α roste β a síla testu slábne. Vžitý postup je takový, že se nejdříve zvolí nízké α a mezi různými testovacími postupy (pokud jich existuje víc) se volí takový, který minimalizuje β, tedy maximalizuje sílu testu.] □

Ilustrace chyb I. a II. druhu:

H_0: pacient je zdravý, H_1: pacient je nemocný.

Předpokládejme nejdříve, že pacient je ve skutečnosti zdravý, tedy H_0 platí.

Lékař to však nemůže vědět, proto pacienta vyšetří. Mohou nastat dvě možnosti:
•Nic nenajde a prohlásí pacienta za zdravého - pak je vše v pořádku a žádná chyba nenastala.
•Lékaři se po důkladném vyšetření něco nelíbí a prohlásí pacienta za nemocného - pak se dopouští chyby I. druhu (bude se léčit zdravý pacient). Toto riziko α nesprávného rozhodnutí volíme malé.

Předpokládejme nyní, že pacient je ve skutečnosti nemocný, tedy H_0 neplatí a platí H_1.

Lékař to však nemůže vědět, proto pacienta vyšetří. Opět mohou nastat dvě možnosti:
•Lékař po důkladném vyšetření něco závažného najde a prohlásí pacienta za nemocného - pak je vše v pořádku a žádná chyba nenastala.
•Lékař nic nenajde a prohlásí pacienta za zdravého - pak se dopouští chyby II. druhu (nebude se léčit nemocný pacient). Toto riziko β souvisí s nastavenou hodnotou α.
Pokud lékař nechce riskovat "léčbu zdravého pacienta"(tedy nastaví malé α), pak bude mít tendenci skoro každého pacienta ujišťovat, že je zdravý a tím omezí zbytečné případy léčení zdravých pacientů. Důsledkem ale je, že i nemocného pacienta dost možná označí za zdravého.

Poznámka 4.3
Testování H_0 proti H_1 na hladině významnosti α je možno provádět třemi způsoby: a) pomocí kritického oboru, b) pomocí intervalů spolehlivosti, c) pomocí $p-$hodnoty. Zásadní je porozumět testování pomocí kritického oboru. Ostatní postupy pak lze snadno odvodit.

Definice 4.4
Nechť $X_1, \ldots, X_n$ je náhodný výběr. Nechť číselná realizace statistiky $T_0 = T_0(X_1, \ldots, X_n)$ rozhoduje o tom, jestli H_0 zamítneme, nebo nezamítneme. Potom T_0 nazýváme *testovým*

kritériem (testovou statistikou).
Množinu všech hodnot, které může testové kritérium nabýt, rozložíme na dvě podmnožiny. Podmnožinu W, která obsahuje ty hodnoty testového kritéria, které vedou k zamítnutí testované hypotézy H_0 nazýváme *kritický obor*.
Podmnožinu V, která obsahuje ty hodnoty testového kritéria, které nevedou k zamítnutí testované hypotézy H_0 nazýváme *obor nezamítnutí nulové hypotézy*. Tyto dva obory jsou odděleny *kritickými hodnotami*, které lze pro danou hladinu α nalézt ve statistických tabulkách.
[Testové kritérium považujeme za ukazatel rozporu mezi testovanou hypotézou a reálnými daty, přičemž kritický obor představuje oblast, ve které je pro nás tento rozpor už nepřijatelný.]

Jestliže číselná realizace t_0 testového kritéria T_0 padne do kritického oboru W, pak H_0 zamítáme na hladině α ve prospěch alternativní hypotézy H_1 a znamená to skutečné vyvrácení H_0.
Jestliže číselná realizace t_0 testového kritéria T_0 padne do oboru nezamítnutí V, pak H_0 nezamítáme na hladině α, což neznamená, že hypotéza H_0 platí, ale jen to, že nemáme důvod ji zamítnout.
Pravděpodobnosti chyb prvního a druhého druhu zapíšeme následovně:
$P(T_0 \in W | H_0 \text{ platí}) = \alpha$
$P(T_0 \in V | H_1 \text{ platí}) = \beta$

Poznámka 4.5
Uveďme si, které hodnoty testového kritéria vzhledem k typu alternativní hypotézy svědčí v její prospěch a podle toho stanovme kritický obor. Ten zpravidla vychází (pro kritéria, které budeme používat):

$$\begin{array}{lllll} W_1 = & (t_{min},\ K_{\alpha/2}(T)\rangle & \cup & \langle K_{1-\alpha/2}(T), t_{max}) & \text{pro oboustr. alternativu } h(\vartheta) \neq c \\ W_2 = & & & \langle K_{1-\alpha}(T), t_{max}) & \text{pro pravostr. alternativu } h(\vartheta) > c \\ W_3 = & (t_{min},\ K_{\alpha}(T)\rangle & & & \text{pro levostr. alternativu } h(\vartheta) < c \quad , \end{array}$$

kde t_{min} je označením pro minimální hodnotu, kterou může testové kritérium T_0 nabýt, t_{max} je označením pro maximální hodnotu, kterou může testové kritérium T_0 nabýt a $K_\alpha(T)$ je α-kvantil testového kritéria T_0.

Doporučený postup testování H_0 proti H_1:

- Stanovíme nulovou hypotézu a alternativní hypotézu. Přitom je vhodné zvolit jako alternativní hypotézu ten předpoklad, jehož přijetí znamená "pokrok v poznání"a mělo by k němu dojít jen s malým rizikem omylu.
- Zvolíme hladinu významnosti α. Obvykle volíme $\alpha = 0,05$, méně často 0,1; nebo 0,01.
- Najdeme vhodné testové kritérium (pivotovu náh. veličinu) a na základě zjištěných dat vypočítáme jeho realizaci.

- Stanovíme kritický obor.
- Jestliže realizace testového kritéria náleží do kritického oboru, nulovou hypotézu zamítáme na hladině významnosti α. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti α.

Nyní přejdeme k testování hypotéz pomocí intervalů spolehlivosti. Volba vhodné pivotové statistiky při konstrukci intervalů spolehlivosti a volba testového kritéria spolu souvisí a tuto souvislost předvedeme na příkladech (ve cvičení).

Věta 4.6
Nechť (d, h) je empirický interval spolehlivosti pro parametrickou funkci $h(\vartheta)$. Pokryje-li tento interval konstantu c, pak H_0 nezamítáme na hladině α, v opačném případě H_0 zamítáme na hladině α.

1. Pro test proti oboustranné alternativě $H_1 : \ h(\vartheta) \neq c$ sestrojíme oboustranný interval spolehlivosti (d, h).
2. Pro test proti levostranné alternativě $H_1 : \ h(\vartheta) < c$ sestrojíme pravostranný interval spolehlivosti $(-\infty, h)$.
3. Pro test proti pravostranné alternativě $H_1 : \ h(\vartheta) > c$ sestrojíme levostranný interval spolehlivosti (d, ∞).

Volba levostranného, resp. pravostranného intervalu spolehlivosti v bodech 2., 3. je vázaná na tvar běžně používaných pivotových statistik.

Většina statistických balíků při testování hypotéz na vstupu nepožaduje zadání rizika α, ale místo toho na výstupu vypisuje tzv. $p-$hodnotu, jejíž podstata je obdobná hladině významnosti α. Obecněji podává více informací o výsledku testu.

Definice 4.7
$p-$hodnota udává nejnižší možnou hladinu významnosti, při které lze na základě realizace testové statistiky nulovou hypotézu zamítnout.

Věta 4.8
Testujeme-li nulovou hypotézu na hladině α, pak pro rozhodnutí o nulové hypotéze platí:
Je-li $p-$hodnota $p \leq \alpha$, pak H_0 zamítáme.
[Tedy realizace testové statistiky, pro kterou byla $p-$hodnota spočtena, náleží do kritického oboru odpovídajícího hladině α.]
Je-li $p-$hodnota $p > \alpha$, pak H_0 nezamítáme.
[Tedy realizace testové statistiky, pro kterou byla $p-$hodnota spočtena, nenáleží do kritického oboru, odpovídajícího hladině α.]
Podle typu alternativní hypotézy volíme způsob výpočtu $p-$hodnoty:

$p = 2\min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\}$	pro oboustr. alternativu $h(\vartheta) \neq c$
$p = P(T_0 \geq t_0)$	pro pravostr. alternativu $h(\vartheta) > c$
$p = P(T_0 \leq t_0)$	pro levostr. alternativu $h(\vartheta) < c$

[Nejdříve si rozmyslete, jak vypadá kritický obor pro levo-, resp. pravo- resp. oboustrannou alternativu, až poté Vám věta dá smysl.]

5 Parametrické úlohy o jednom náhodném výběru z normálního rozložení

Mnoho náhodných veličin, s kterými se v praxi setkáváme, se řídí normálním rozložením. Pomocí cenrální limitní věty můžeme za poměrně obecných podmínek aproximovat i jiná rozložení normálním rozložením. Proto je velmi důležité věnovat pozornost právě výběrům z normálního rozložení.
Normální rozložení je charakterizováno dvěma parametry, střední hodnotou μ a rozptylem σ^2. Budeme tedy řešit úlohy, týkající se těchto parametrů. Tyto úlohy spočívají v konstrukci odhadů a testování hypotéz.

Následující věta uvede rozložení běžně užívaných testových kritérií pro testy pro jeden náhodný výběr z normálního rozložení.

Věta 5.1
Nechť $X_1, \ldots, X_n$ je náhodný výběr z normálního rozložení $N(\mu, \sigma^2)$. Potom platí:

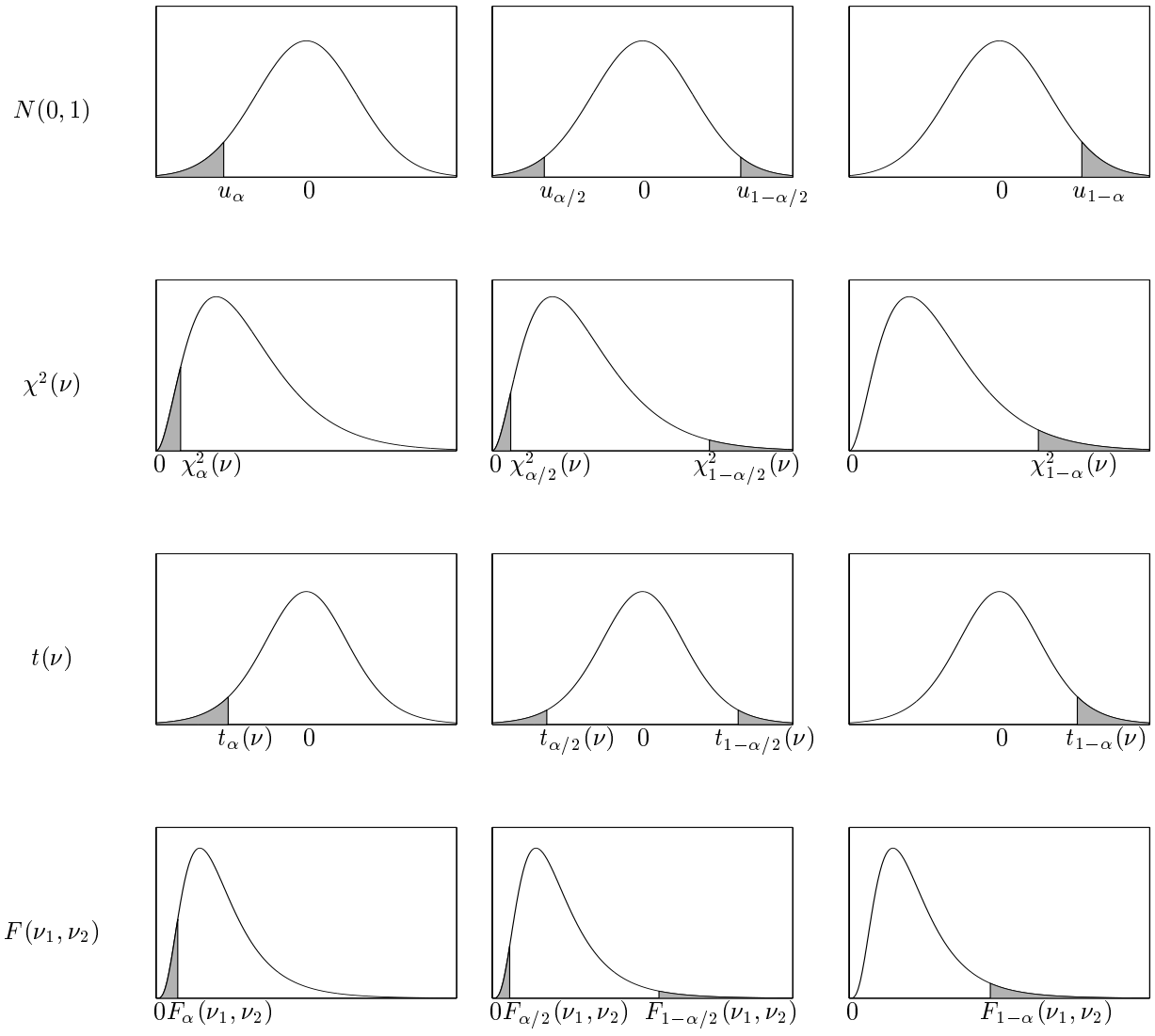
1. Výběrový průměr $M = \sum_{i=1}^{n} X_i$ a výběrový rozptyl $S^2 = \sum_{i=1}^{n} (X_i - M)^2$ jsou stochasticky nezávislé.

2. $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$, tedy $M \sim N(\mu,\ \frac{\sigma^2}{n})$
   [Pivotová statistika U slouží k řešení úloh o μ, když σ^2 známe.]

3. $K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
   [Pivotová statistika K slouží k řešení úloh o σ^2, když μ neznáme.]

4. $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$
   [Pivotová statistika T slouží k řešení úloh o μ, když σ^2 neznáme.]

5. $\frac{\sum_{i=1}^{n} (X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$
   [Tato pivotová statistika slouží k řešení úloh o σ^2, když μ známe.]



$$u_\alpha = -u_{1-\alpha} \qquad t_\alpha(\nu) = -t_{1-\alpha}(\nu) \qquad F_\alpha(\nu_1,\nu_2) = \frac{1}{F_{1-\alpha}(\nu_2,\nu_1)}$$

Příklad 5.2

Hmotnost balíčku krystalového cukru se řídí normálním rozložením $N(1002\,g,\ 64\,g^2)$. Kontrola náhodně vybírá 9 balíčků z jedné série a zjišťuje, zda jejich průměrná hmotnost je aspoň $999\,g$. Pokud ne, podnik musí zaplatit pokutu. Jaká je pravděpodobnost, že podnik bude muset pokutu zaplatit?

Řešení

$X_1, \ldots, X_9 \sim N(1002,\ 64)$, $M \sim N(1002,\ \frac{64}{9})$, $P(M \leq 999) = ?$

$P(M \leq 999) = P(\frac{M-1002}{\sqrt{\frac{64}{9}}} \leq \frac{999-1002}{\sqrt{\frac{64}{9}}}) = P(U \leq \frac{-9}{8}) = 1 - \Phi(\frac{9}{8}) = 1 - \Phi(1,125) = 1 - 0,87076 = 0,12924.$

Pravděpodobnost, že podnik bude platit pokutu je přibližně 12,9%.

Častým úkolem statistika je odvodit intervaly spolehlivosti pro neznámé parametry. V případě normálního rozložení se jedná o parametry μ a σ^2. Mohou tedy nastat čtyři situace: hledáme interval spolehlivosti 1. pro μ, když σ^2 známe; 2. pro σ^2, když μ neznáme; 3. pro μ, když σ^2 neznáme a 4. pro σ^2, když μ známe. Při konstrukci intervalů spolehlivosti je potřeba vědět, která pivotová statistika je vhodná pro zvolenou z uvedených čtyř možností. Potom je snadné pomocí postupu v 3.9 odvodit samotný dolní, resp. horní odhad. Toto odvození pro první možnost - pro μ, když σ^2 známe - je v příkladě 3.11. Jak by pomocí zmíněného postupu dopadli odhady pro zbývající možnosti, uvede následující věta.

Věta 5.3
Nechť $X_1, \ldots, X_n$ je náhodný výběr z normálního rozložení $N(\mu, \sigma^2)$. Uvažujme $100(1-\alpha)$ procentní empirický interval spolehlivosti.

1. Interval spolehlivosti pro μ, když σ^2 známe odvozujeme z pivotové statistiky $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Potom meze jsou pro:

$$\begin{array}{lcl} \text{oboustranný int. spol. } (d,h) & = & \left(m - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2} \quad , \quad m + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2}\right) \\ \text{levostranný int. spol. } (d,\infty) & = & \left(m - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha} \quad , \quad \infty\right) \\ \text{pravostranný int. spol. } (-\infty,h) & = & \left(-\infty \quad , \quad m + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha}\right) \end{array}$$

2. Interval spolehlivosti pro σ^2, když μ neznáme odvozujeme z pivotové statistiky $K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. Potom meze jsou pro:

$$\begin{array}{lcl} \text{oboustranný int. spol. } (d,h) & = & \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \quad , \quad \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right) \\ \text{levostranný int. spol. } (d,\infty) & = & \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)} \quad , \quad \infty\right) \\ \text{pravostranný int. spol. } (-\infty,h) & = & \left(-\infty \quad , \quad \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)}\right) \end{array}$$

3. Interval spolehlivosti pro μ, když σ^2 neznáme odvozujeme z pivotové statistiky $T = \frac{M-\mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$. Potom meze jsou pro:

$$\begin{array}{lcl} \text{oboustranný int. spol. } (d,h) & = & \left(m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) \ , \ m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1)\right) \\ \text{levostranný int. spol. } (d,\infty) & = & \left(m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha}(n-1) \quad , \quad \infty\right) \\ \text{pravostranný int. spol. } (-\infty,h) & = & \left(-\infty \quad , \quad m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha}(n-1)\right) \end{array}$$

4. Interval spolehlivosti pro σ^2, když μ známe odvozujeme z pivotové statistiky $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$. Potom meze jsou pro:

$$\text{oboustranný int. spol. } (d,h) = \left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)} \quad , \quad \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)} \right)$$

$$\text{levostranný int. spol. } (d,\infty) = \left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha}(n)} \quad , \quad \infty \right)$$

$$\text{pravostranný int. spol. } (-\infty,h) = \left(-\infty \quad , \quad \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha}(n)} \right)$$

Příklad 5.4
10krát nezávisle na sobě byla změřena určitá konstanta μ. Výsledky měření byly:
$2 \quad 1,8 \quad 2,1 \quad 2,4 \quad 1,9 \quad 2,1 \quad 2 \quad 1,8 \quad 2,3 \quad 2,2$
Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, \ldots, X_n$ z rozloženi $N(\mu, \sigma^2)$, kde parametry μ, σ^2 neznáme. Najděte 95% interval spolehlivosti pro parametr μ, a to a) oboustranný, b) levostranný, c) pravostranný.

Řešení
Jedná se o interval spolehlivosti pro μ, když σ^2 neznáme. K odvození mezí využijeme statistiku $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$, jejíž $\alpha-$kvantily nalezneme v tabulkách.
$n = 10 \qquad \alpha = 0,05 \qquad t_{1-\alpha/2}(n-1) = t_{0,975}(9) = 2,2622$
$t_{1-\alpha}(n-1) = t_{0,95}(9) = 1,8331$
$m = 2,06 \qquad s^2 = 0,0404 \qquad s = 0,2011$

---

ad a)
$d = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 2,06 - \frac{0,2011}{\sqrt{10}} \cdot 2,2622 = 1,92$
$h = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 2,06 + \frac{0,2011}{\sqrt{10}} \cdot 2,2622 = 2,20$

$1,92 < \mu < 2,2$ s pravděpodobností aspoň 0,95

ad b)
$d = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha}(n-1) = 2,06 - \frac{0,2011}{\sqrt{10}} \cdot 1,8331 = 1,94$

$1,94 < \mu$ s pravděpodobností aspoň 0,95

ad c)
$h = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 2,06 + \frac{0,2011}{\sqrt{10}} \cdot 1,8331 = 2,18$

$\mu < 2,18$ s pravděpodobností aspoň 0,95

Dosud jsme se věnovali intervalům spolehlivosti pro parametry normálního rozložení, nyní se budeme věnovat testování hypotéz o parametrech μ a σ^2. Budeme se věnovat testování pomocí kritického oboru, další způsoby testování lze lehce odvodit.

Definice 5.5
Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde σ^2 známe. Nechť $n \geq 2$ a c je konstanta. Test $H_0: \ \mu = c$ proti $H_1: \ \mu \neq c$ (resp. $H_1: \ \mu < c$ resp. $H_1: \ \mu > c$) se nazývá *z-test*.

Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde σ^2 neznáme. Nechť $n \geq 2$ a c je konstanta. Test $H_0: \ \mu = c$ proti $H_1: \ \mu \neq c$ (resp. $H_1: \ \mu < c$ resp. $H_1: \ \mu > c$) se nazývá jednovýběrový *t-test*.

Nechť $X_1, \ldots, X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde μ neznáme. Nechť $n \geq 2$ a c je konstanta. Test $H_0: \ \sigma^2 = c$ proti $H_1: \ \sigma^2 \neq c$ (resp. $H_1: \ \sigma^2 < c$ resp. $H_1: \ \sigma^2 > c$) se nazývá *test o rozptylu*.

Poznámka 5.6
Volba vhodného testovacího kritéria pro zvolený test je obdobná volbě vhodné pivotové náhodné veličiny v 5.3, tedy pro z-test volím testovací kritérium T_0 odvozené ze statistiky U, pro t-test ze statistiky T a pro test o rozptylu ze statistiky K.
Pozor na dvojsmyslnost písmena T. Obecně značí T_0 jakékoliv testovací kritérium, v případě t-testu značí T statistiku se Studentovým rozložením. Tedy můžeme psát $T_0 = U$, $T_0 = T$, $T_0 = K$ za předpokladu, že H_0 platí.

Věta 5.7
Nechť $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, $c \in \mathbf{R}$, $n \geq 2$

1. V případě z-testu se na hladině α nulová hypotéza H_0 zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium $T_0 = \frac{M-c}{\frac{\sigma}{\sqrt{n}}}$ realizuje v oboru W, kde

pro oboustrannou alternativu	$H_1: \ \mu \neq c$	je $W = (-\infty, \ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2}, \ \infty)$
pro levostrannou alternativu	$H_1: \ \mu < c$	je $W = (-\infty, \ -u_{1-\alpha}\rangle$
pro pravostrannou alternativu	$H_1: \ \mu > c$	je $W = \langle u_{1-\alpha}, \ \infty)$

2. V případě t-testu se na hladině α nulová hypotéza H_0 zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium $T_0 = \frac{M-c}{\frac{S}{\sqrt{n}}}$ realizuje v oboru W, kde

pro oboustrannou alt.	$H_1: \ \mu \neq c$	je $W = (-\infty, \ -t_{1-\alpha/2}(n-1)\rangle \bigcup \langle t_{1-\alpha/2}(n-1), \ \infty)$
pro levostrannou alt.	$H_1: \ \mu < c$	je $W = (-\infty, \ -t_{1-\alpha}(n-1)\rangle$
pro pravostrannou alt.	$H_1: \ \mu > c$	je $W = \langle t_{1-\alpha}(n-1), \ \infty)$

3. V případě testu o rozptylu se na hladině α nulová hypotéza H_0 zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium $T_0 = \frac{(n-1)S^2}{c}$ realizuje v oboru W, kde

pro oboustrannou alt.	$H_1: \ \sigma^2 \neq c$	je $W = (0, \ \chi^2_{\alpha/2}(n-1)\rangle \bigcup \langle \chi^2_{1-\alpha/2}(n-1), \ \infty)$
pro levostrannou alt.	$H_1: \ \sigma^2 < c$	je $W = (0, \ \chi^2_{\alpha}(n-1)\rangle$
pro pravostrannou alt.	$H_1: \ \sigma^2 > c$	je $W = \langle \chi^2_{1-\alpha}(n-1), \ \infty)$

Příklad 5.8
Podle údajů na obalu čokolády by její čistá hmotnost měla být 125 g. Výrobce dostal několik stížností od kupujících, ve kterých tvrdili, že hmotnost čokolád je nižší, než deklarovaných 125 g. Z tohoto důvodu oddělení kontroly náhodně vybralo 50 čokolád a zjistilo,

že jejich průměrná hmotnost je 122 g a směrodatná odchylka 8,6 g. Za předpokladu, že hmotnost čokolád se řídí normálním rozložením, můžeme na hladině významnosti $\alpha = 0,01$ považovat stížnosti kupujících za oprávněné?

Řešení

$X_1, \ldots, X_{50} \sim N(\mu, \sigma^2)$. Testujeme $H_0 : \mu = 125$ proti levostranné alternativní hypotéze $H_1 : \mu < 125$. Parametr σ^2 neznáme, tedy úloha vede na jednovýběrový t-test.

Testovací kritérium je $T_0 = \frac{M-c}{\frac{S}{\sqrt{n}}}$.

Jeho číselná realizace je $t_0 = \frac{122-125}{\frac{8,6}{\sqrt{50}}} = -2,4667$.

Kritický obor je $W = (-\infty,\ -t_{1-\alpha}(n-1)\rangle = (-\infty,\ -t_{0,99}(49)\rangle = (-\infty;\ -2,4049\rangle$

Protože $t_0 \in W,\ H_0$ zamítáme na hladině významnosti 0,01.

Stížnosti kupujících lze považovat za oprávněné s rizikem omylu nejvýše 1%.

Máme- li jeden náhodný výběr z dvourozměrného normálního rozložení, můžeme jej převést na výběr z jednorozměrného normálního rozložení a poté můžeme pro intervaly spolehlivosti i pro testování hypotéz použít dosud odvozené postupy.

Poznámka 5.9 o jednom výběru z dvourozměrného normálního rozložení

Nechť $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n} \sim N_2\left(\binom{\mu_1}{\mu_2},\ \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right),\ n \geq 2.$

Pomocí lineární transformace převedeme náhodný vektor $\binom{X}{Y}$ na skalární náhodnou veličinu $Z = (X - Y) \sim N((\mu_1 - \mu_2),\ (\sigma_1^2 - 2\sigma_{12} + \sigma_2^2))$

Označíme $\mu = \mu_1 - \mu_2 \qquad \sigma^2 = \sigma_1^2 - 2\sigma_{12} + \sigma_2^2$

Nyní náš náhodný výběr $(X_1 - Y_1), \ldots, (X_n - Y_n) = Z_1, \ldots, Z_n$ je z normálního rozložení $N(\mu, \sigma^2)$ a říkáme mu *rozdílový náhodný výběr.*

Věta 5.10

Nechť $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n} \sim N_2\left(\binom{\mu_1}{\mu_2},\ \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right),\ n \geq 2$ a rozptylová matice Σ není známa. Meze $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro parametrickou funkci $\mu = \mu_1 - \mu_2$ jsou:

$d = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1)$

$h = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1)$

Příklad 5.11

Dvěma různými laboratorními metodami se zjišťoval obsah chemické látky v roztoku (údaje jsou v procentech). Bylo vybráno 5 vzorků:

číslo vzorku	1	2	3	4	5
1. metoda	2,3	1,9	2,1	2,4	2,6
2. metoda	2,4	2,0	2,0	2,3	2,5

Za předpokladu, že data jsou z dourozměrného normálního rozložení, sestrojte 90% empirický interval spolehlivosti pro rozdíl středních hodnot výsledků obou metod.

Řešení

Přejdeme k rozdílovému náhodnému výběru, kde:

$z_1 = -0,1 \quad z_2 = -0,1 \quad z_3 = 0,1 \quad z_4 = 0,1 \quad z_5 = 0,1$

$m = 0,02 \qquad s^2 = 0,012 \qquad s = 0,109545 \qquad n = 5 \qquad t_{1-\alpha/2}(n-1) = t_{0,95}(4) = 2,1318$

$d = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 0,02 - \frac{0,109545}{\sqrt{5}} \cdot 2,1318 = -0,0844$

$h = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 0,02 + \frac{0,109545}{\sqrt{5}} \cdot 2,1318 = 0,1244$

S pravděpodobností alespoň 0,95 platí $-0,0844 < \mu < 0,1244$

Definice 5.12
Nechť $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n} \sim N_2\left(\binom{\mu_1}{\mu_2}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right),\ n \geq 2.$
Test $H_0 : \mu_1 - \mu_2 = 0$ proti $H_1 : \mu_1 - \mu_2 \neq 0$ se nazývá *párový t-test*. Přechodem k rozdílovému náhodnému výběru převedeme párový t-test na jednovýběrový t-test.

Příklad 5.13
V následující tabulce jsou údaje o výnosnosti dosažené 12-ti náhodně vybranými firmami při investování do mezinárodního podnikání (X) a do domácího podnikání (Y):

číslo firmy	1	2	3	4	5	6	7	8	9	10	11	12
X	10	12	14	12	12	17	9	15	9	11	7	15
Y	11	14	15	11	13	16	10	13	11	17	9	19

Výnosnost je vyjádřena v procentech a představuje podíl na zisku vložených investic za rok. Za předpokladu, že data pocházejí z dvourozměrného normálního rozložení na hladině významnosti 0,1 testujte hypotézu, že neexistuje rozdíl mezi investováním do domácího a do mezinárodního podnikání. Test proveďte pomocí a) intervalu spolehlivosti, b) pomocí kritického oboru.

Řešení
Nejdříve přejdeme k rozdílovému náhodnému výběru $Z_i = X_i - Y_i,\ i = 1, \ldots 12$. Realizace výběrových charakteristik pak jsou $m = -1,33, \quad s^2 = 4,\overline{78}$
Testujeme hypotézu $H_0 : \mu = 0$ proti $H_1 : \mu \neq 0$,
budeme potřebovat kvantil $t_{0.95}(11) = 1,7959$

ad a)

$d = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = -1,\overline{3} - \frac{\sqrt{4,\overline{78}}}{\sqrt{12}} \cdot 1,7959 = -2,4677$

$h = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = -1,\overline{3} + \frac{\sqrt{4,\overline{78}}}{\sqrt{12}} \cdot 1,7959 = -0,1989$

•Protože $0 \notin (-2,4677\ ,\ -0,1989)$, H_0 zamítáme na hladině významnosti 0,1.

ad b)
Testovací kritérium je $T_0 = \frac{M-c}{\frac{S}{\sqrt{n}}}$.
Jeho číselná realizace je $t_0 = \frac{-1,\overline{3}-0}{\frac{\sqrt{4,\overline{78}}}{\sqrt{12}}} = -2,11085.$
Kritický obor je $W = (-\infty,\ -t_{1-\alpha/2}(n-1)\rangle \bigcup \langle t_{1-\alpha/2}(n-1),\ \infty) =$
$= (-\infty,\ -1,7959\rangle \bigcup \langle 1,7959,\ \infty)$
•Protože $t_0 \in W$, H_0 zamítáme na hladině významnosti 0,1.

6 Parametrické úlohy o dvou nezávislých náhodných výběrech z normálního rozložení.

Vycházíme ze situace, kdy máme dva nezávislé náhodné výběry; první je z rozložení $N(\mu_1, \sigma_1^2)$ a druhý je z rozložení $N(\mu_2, \sigma_2^2)$. Našim úkolem bude konstruovat intervaly spolehlivosti pro parametrickou funkci $\mu_1 - \mu_2$, či $\frac{\sigma_1^2}{\sigma_2^2}$ a testovat o těchto parametrických funkcích hypotézy.

Nyní si uvedeme větu o rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů těchto dvou výběrů.

Věta 6.1
Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z normálního rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr z normálního rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2,\ n_2 \geq 2$. Označme M_1, M_2 výběrové průměry, $S_1^2,\ S_2^2$ výběrové rozptyly a $S_*^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ vážený průměr výběrových rozptylů. Potom platí:

1. Statistiky $(M_1 - M_2)$ a S_*^2 jsou stochasticky nezávislé.

2. $U = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$, tedy $M_1 - M_2 \sim N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
   [Pivotová statistika U slouží k řešení úloh o $\mu_1 - \mu_2$, když $\sigma_1^2,\ \sigma_2^2$ známe.]

3. Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak
   $K = \frac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$
   [Pivotová statistika K slouží k řešení úloh o společném neznámém rozptylu σ^2.]

4. Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak
   $T = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{S_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$
   [Pivotová statistika T slouží k řešení úloh o $\mu_1 - \mu_2$, když $\sigma_1^2,\ \sigma_2^2$ neznáme, ale víme, že jsou shodné.]

5. $F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1,\ n_2 - 1)$
   [Pivotová statistika slouží k řešení úloh o σ_1^2/σ_2^2.]

Pomocí uvedených pivotových statistik lze zkonstruovat intervaly spolehlivosti např. pro následující parametrické funkce: $\mu_1 - \mu_2$ a σ_1^2/σ_2^2. Při konstrukci intervalu spolehlivosti pro $\mu_1 - \mu_2$ musíme rozlišit, jestli jsou rozptyly známé, nebo neznámé. Jsou-li neznámé, musíme zjistit, jestli jsou shodné, či nikoliv. Shodu rozptylů ověřujeme pomocí F-testu, který bude později uveden.
Následující věta uvede již odvozené dolní a horní meze pro zmíněné parametrické funkce.

Věta 6.2
Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z normálního rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21}, \ldots, X_{2n_2}$ je

na něm nezávislý náhodný výběr z normálního rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2,\ n_2 \geq 2$. Uvažujme $100(1-\alpha)$ procentní empirický interval spolehlivosti.

1. Interval spolehlivosti pro $\mu_1 - \mu_2$, když $\sigma_1^2,\ \sigma_2^2$ známe odvozujeme z pivotové statistiky
$U = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$. Potom meze jsou pro:

$$\begin{aligned}
\text{oboustr. } (d,h) &= \left(m_1 - m_2 - \sqrt{\tfrac{\sigma_1^2}{n_1}+\tfrac{\sigma_2^2}{n_2}} \cdot u_{1-\alpha/2} \quad , \quad m_1 - m_2 + \sqrt{\tfrac{\sigma_1^2}{n_1}+\tfrac{\sigma_2^2}{n_2}} \cdot u_{1-\alpha/2}\right) \\
\text{levostr. } (d,\infty) &= \left(m_1 - m_2 - \sqrt{\tfrac{\sigma_1^2}{n_1}+\tfrac{\sigma_2^2}{n_2}} \cdot u_{1-\alpha} \quad , \quad \infty\right) \\
\text{pravostr. } (-\infty,h) &= \left(-\infty \quad , \quad m_1 - m_2 + \sqrt{\tfrac{\sigma_1^2}{n_1}+\tfrac{\sigma_2^2}{n_2}} \cdot u_{1-\alpha}\right)
\end{aligned}$$

2. Interval spolehlivosti pro společný neznámý rozptyl σ^2 odvozujeme z pivotové statistiky
$K = \frac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$. Potom meze jsou pro:

$$\begin{aligned}
\text{oboustr. } (d,h) &= \left(\tfrac{(n_1+n_2-2)s_*^2}{\chi^2_{1-\alpha/2}(n_1+n_2-2)} \quad , \quad \tfrac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha/2}(n_1+n_2-2)}\right) \\
\text{levostr. } (d,\infty) &= \left(\tfrac{(n_1+n_2-2)s_*^2}{\chi^2_{1-\alpha}(n_1+n_2-2)} \quad , \quad \infty\right) \\
\text{pravostr. } (-\infty,h) &= \left(-\infty \quad , \quad \tfrac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha}(n_1+n_2-2)}\right)
\end{aligned}$$

3. Interval spolehlivosti pro $\mu_1 - \mu_2$, když $\sigma_1^2,\ \sigma_2^2$ neznáme, ale víme, že jsou shodné odvozujeme z pivotové statistiky
$T = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{S_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$. Potom meze jsou pro:

$$\begin{aligned}
\text{oboustr. } (d,h) &= \Big(m_1 - m_2 - s_*\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}} \cdot t_{1-\alpha/2}(n_1+n_2-2) \quad , \\
&\qquad\qquad m_1 - m_2 + s_*\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}} \cdot t_{1-\alpha/2}(n_1+n_2-2)\Big) \\
\text{levostr. } (d,\infty) &= \Big(m_1 - m_2 - s_*\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}} \cdot t_{1-\alpha}(n_1+n_2-2) \quad , \quad \infty\Big) \\
\text{pravostr. } (-\infty,h) &= \Big(-\infty \quad , \quad m_1 - m_2 + s_*\sqrt{\tfrac{1}{n_1}+\tfrac{1}{n_2}} \cdot t_{1-\alpha}(n_1+n_2-2)\Big)
\end{aligned}$$

4. Interval spolehlivosti pro podíl rozptylů σ_1^2/σ_2^2 odvozujeme z pivotové statistiky
$F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$. Potom meze jsou pro:

$$\begin{aligned}
\text{oboustr. } (d,h) &= \left(\tfrac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1,n_2-1)} \quad , \quad \tfrac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1,n_2-1)}\right) \\
\text{levostr. } (d,\infty) &= \left(\tfrac{s_1^2/s_2^2}{F_{1-\alpha}(n_1-1,n_2-1)} \quad , \quad \infty\right) \\
\text{pravostr. } (-\infty,h) &= \left(-\infty \quad , \quad \tfrac{s_1^2/s_2^2}{F_{\alpha}(n_1-1,n_2-1)}\right)
\end{aligned}$$

Poznámka 6.3
Není-li v bodě 3. předchozí věty splněn požadavek shody rozptylů, lze sestrojit alespoň

přibližný $100(1-\alpha)\%$ interval spolehlivosti pro $\mu_1 - \mu_2$. V tomto případě má statistika T přibližně rozložení $t(\nu)$, kde pro počet stupňů volnosti ν platí:

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \qquad \text{tzv. Welchova aproximace}$$

Není-li ν celé číslo, použijeme lineární interpolaci.

Příklad 6.4
Ve dvou nádržích se zkoumal obsah chlóru (v g/l). Z první nádrže bylo odebráno 25 vzorků, z druhé nádrže 10 vzorků. Byly vypočteny realizace výběrových průměrů a rozptylů: $m_1 = 34,48, \quad m_2 = 35,59, \quad s_1^2 = 1,7482, \quad s_2^2 = 1,7121$ Hodnoty zjištěné z odebraných vzorků považujeme za realizace dvou nezávislých náhodných výběrů z rozložení $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$. Sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Řešení
Jedná se o interval spolehlivosti pro $\mu_1 - \mu_2$, když $\sigma_1^2,\ \sigma_2^2$ neznáme, ale víme, že jsou shodné. Interval spolehlivosti odvozujeme z pivotové statistiky
$T = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{S_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$.
Budeme potřebovat kvantil $t_{1-\alpha/2}(n_1+n_2-2) = t_{0,975}(33) = 2,035$
a realizaci váženého průměru výběrových rozptylů $s_*^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{24.1,7482+9.1,7121}{33} = 1,7384$.

Potom meze jsou :
$d = m_1 - m_2 - s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \cdot t_{1-\alpha/2}(n_1+n_2-2) =$
$= 34,48 - 35,59 - \sqrt{1,7384} \cdot \sqrt{\frac{1}{25}+\frac{1}{10}} \cdot 2,035 = -2,114$

$h = m_1 - m_2 + s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \cdot t_{1-\alpha/2}(n_1+n_2-2) =$
$= 34,48 - 35,59 + \sqrt{1,7384} \cdot \sqrt{\frac{1}{25}+\frac{1}{10}} \cdot 2,035 = -0,106$

Zjistili jsme, že $\mu_1 - \mu_2 \in (-2,114\ g/l\ ,\ -0,106\ g/l)$ s pravděpodobností aspoň 0,95.

Příklad 6.5
V příkladě 6.4 nyní předpokládejme, že dané dva náhodné výběry pocházejí z rozložení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$. Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů.

Řešení
Jedná se o interval spolehlivosti pro podíl rozptylů σ_1^2/σ_2^2. Interval spolehlivosti odvozujeme z pivotové statistiky
$F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$.
Budeme potřebovat kvantil $F_{1-\alpha/2}(n_1-1, n_2-1) = F_{0,972}(24,9) = 3,6142$ a
$F_{\alpha/2}(n_1-1, n_2-1) = F_{0,025}(24,9) = \frac{1}{F_{0,975}(9,24)} = \frac{1}{2,7027}$

.

Potom meze jsou :
$d = \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1,n_2-1)} = \ldots = 0,28$
$h = \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1,n_2-1)} = \ldots = 2,76$

Tedy $P\left(\sigma_1^2/\sigma_2^2 \in (0,28\ ;\ 2,76)\right) \geq 0,95$

Definice 6.6
Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z normálního rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr z normálního rozložení $N(\mu_2, \sigma_2^2)$. Nechť $n_1 \geq 2,\ n_2 \geq 2$ a c je konstanta.

(i.) Předpokládejme, že σ_1^2, σ_2^2 známe.
Test $H_0 :\ \mu_1 - \mu_2 = c$ proti $H_1 :\ \mu_1 - \mu_2 \neq c$ (resp. $H_1 :\ \mu_1 - \mu_2 < c$ resp. $H_1 :\ \mu_1 - \mu_2 > c$) se nazývá *dvouvýběrový z-test.*

(ii.) Předpokládejme, že σ_1^2, σ_2^2 neznáme, ale víme, že $\sigma_1^2 = \sigma_2^2$.
Test $H_0 :\ \mu_1 - \mu_2 = c$ proti $H_1 :\ \mu_1 - \mu_2 \neq c$ (resp. $H_1 :\ \mu_1 - \mu_2 < c$ resp. $H_1 :\ \mu_1 - \mu_2 > c$) se nazývá *dvouvýběrový t-test.*

(iii.) Test $H_0 :\ \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti $H_1 :\ \frac{\sigma_1^2}{\sigma_2^2} \neq 1$ (resp. $H_1 :\ \frac{\sigma_1^2}{\sigma_2^2} < 1$ resp. $H_1 :\ \frac{\sigma_1^2}{\sigma_2^2} > 1$) se nazývá *F-test.*

Poznámka 6.7
Volba vhodného testovacího kritéria pro zvolený test je stejná, jako volba vhodné pivotové náhodné veličiny v 6.2, tedy pro dvouvýběrový z-test volím jako testovací kritérium T_0 statistiku U, pro dvouvýběrový t-test volím statistiku T a pro F-test volím statistiku F.

Věta 6.8
Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z normálního rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr z normálního rozložení $N(\mu_2, \sigma_2^2)$. Nechť $n_1 \geq 2,\ n_2 \geq 2$ a c je konstanta.

1. V případě dvouvýběrového z-testu se na hladině α nulová hypotéza H_0 zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium
$T_0 = \frac{M_1 - M_2 - c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ realizuje v oboru W, kde
pro oboustr. $H_1 :\ \mu_1 - \mu_2 \neq c$ je $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$
pro levostr. $H_1 :\ \mu_1 - \mu_2 < c$ je $W = (-\infty,\ -u_{1-\alpha}\rangle$
pro pravostr. $H_1 :\ \mu_1 - \mu_2 > c$ je $W = \langle u_{1-\alpha},\ \infty)$

2. V případě dvouvýběrového t-testu se na hladině α nulová hypotéza H_0 zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium

$T_0 = \frac{M_1 - M_2 - c}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ realizuje v oboru W, kde

pro ob. $H_1 : \ \mu_1 - \mu_2 \neq c$ je $W = (-\infty, \ -t_{1-\alpha/2}(n_1 + n_2 - 2)\rangle \bigcup \langle t_{1-\alpha/2}(n_1 + n_2 - 2), \ \infty)$

pro lev. $H_1 : \ \mu_1 - \mu_2 < c$ je $W = (-\infty, \ -t_{1-\alpha}(n_1 + n_2 - 2)\rangle$

pro pra. $H_1 : \ \mu_1 - \mu_2 > c$ je $W = \langle t_{1-\alpha}(n_1 + n_2 - 2), \ \infty)$

3. V případě F-testu se na hladině α nulová hypotéza H_0 zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium
$T_0 = \frac{S_1^2/S_2^2}{1}$ realizuje v oboru W, kde

pro ob. $H_1 : \ \sigma_1^2/\sigma_2^2 \neq 1$ je $W = (0, \ F_{\alpha/2}(n_1 - 1, n_2 - 1)\rangle \bigcup \langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \ \infty)$

pro lev. $H_1 : \ \sigma_1^2/\sigma_2^2 < 1$ je $W = (0, \ F_{\alpha}(n_1 - 1, n_2 - 1)\rangle$

pro prav. $H_1 : \ \sigma_1^2/\sigma_2^2 > 1$ je $W = \langle F_{1-\alpha}(n_1 - 1, n_2 - 1), \ \infty)$

Příklad 6.9

V restauraci "U bílého koníčka"měřili ve 20 případech čas obsluhy zákazníka. Výsledky jsou v minutách: 6, 8, 11, 4, 7, 6, 10, 6, 9, 8, 5, 12, 13, 10, 9, 8, 7, 11, 10, 5. V restauraci "Zlatý lev"bylo dané pozorování uskutečněno v 15 případech s těmito výsledky: 9, 11, 10, 7, 6, 4, 8, 13, 5, 15, 8, 5, 6, 8 ,7. Za předpokladu, že uvedené hodnoty pocházejí ze dvou normálních rozložení, na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty doby obsluhy jsou v obou restauracích stejné.

Řešení

Řešení: Na hladině významnosti 0,05 testujeme nulovou hypotézu $H_0 : \mu_1 - \mu_2 = 0$ proti oboustranné alternativě $H_0 : \mu_1 - \mu_2 \neq 0$. Jedná se o dvouvýběrový t-test. Předpokladem tohoto testu je však shoda rozptylů, kterou musíme nejdříve ověřit. K tomu užijeme F-test.
$m_1 = 8,25; \quad m_2 = 8,13; \quad s_1^2 = 6,307; \quad s_2^2 = 9,41;$
$s_* = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{19.6,307 + 14.9,41}{19+14} = 7,623$

Na hladině významnosti 0,05 tedy nejdříve testujeme hypotézu
$\bullet H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$.
Testové kritérium je
$T_0 = \frac{S_1^2/S_2^2}{1}$, jeho realizace je $t_0 = \frac{6,307}{9,41} = 0,6702$.
Kritický obor je
$W = \langle 0, \ F_{\alpha/2}(n_1 - 1, n_2 - 1)\rangle \bigcup \langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \ \infty) =$
$\langle 0, \ F_{0,025}(19, 14)\rangle \bigcup \langle F_{0,975}(19, 14), \ \infty) =$
$\langle 0, \ \frac{1}{F_{0,975}(14,19)}\rangle \bigcup \langle 2,8607, \ \infty) = \langle 0, \ 0,3778\rangle \bigcup \langle 2,8607, \ \infty)$

$t_0 \notin W$, tedy H_0 o zhodě rozptylů nezamítám na hladině významnosti 0,05 a mohu pokračovat t-testem.

$\bullet H_0 : \mu_1 - \mu_2 = 0$ proti $H_1 : \mu_1 - \mu_2 \neq 0$
Testové kritérium je
$T_0 = \frac{M_1 - M_2 - c}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, jeho realizace je $t_0 = \frac{8,25 - 8,13}{\sqrt{7,623}\sqrt{\frac{1}{20} + \frac{1}{15}}} = 0,124$
Kritický obor je
$W = (-\infty, \ -t_{1-\alpha/2}(n_1 + n_2 - 2)\rangle \bigcup \langle t_{1-\alpha/2}(n_1 + n_2 - 2), \ \infty) =$

$= (-\infty,\ -t_{0,975}(33)\rangle \bigcup \langle t_{0,972}(33),\ \infty) = (-\infty,\ -2,035\rangle \bigcup \langle 2,035\ ,\ \infty)$

Protože $t_0 \notin W,\ H_0$ nezamítáme na hladině významnosti 0,05. [Tedy není důvod pochybovat o tom, že v obou restauracích obsluhují stejně rychle.]

7 Parametrické úlohy o jednom náhodném výběru a dvou nezávislých náhodných výběrech z alternativního rozložení

Věta 7.1
Nechť $X_1, \ldots, X_n$ je náhodný výběr z alternativního rozložení $A(\vartheta)$ a nechť je splněna podmínka $n\vartheta(1-\vartheta) > 9$. Nechť $M = \frac{1}{n}\sum_{i=1}^{n} X_i$ je výběrový průměr. Pak statistika $U = \frac{M-\vartheta}{\sqrt{\frac{M(1-M)}{n}}} \approx N(0,1)$. Čteme: statistika U má asymptoticky normální standardizované rozložení.
Potom meze $100(1-\alpha)\%$ asymptotického empirického intervalu spolehlivosti pro parametr ϑ alternativního rozložení jsou:
$d = m - \sqrt{\frac{m(1-m)}{n}} u_{1-\alpha/2}$
$h = m + \sqrt{\frac{m(1-m)}{n}} u_{1-\alpha/2}$

Poznámka 7.2
Je důležité uvědomit si interpretaci průměru M. Náhodná veličina X_i nabývá pouze dvě hodnoty: jedničky a nuly, kde jednička znamená, že nastal úspěch. Potom $\sum_{i=1}^{n} X_i$ znamená počet úspěchů v n nezávislých pokusech a $\frac{1}{n}\sum_{i=1}^{n} X_i$ je tedy relativní četnost úspěchů. Relativní četnost úspěchů je statistika, která odhaduje parametr pravděpodobnosti úspěchu ϑ.

Příklad 7.3
Oddělení marketingu podniku chce odhadnout, jaký podíl na trhu s výrobky, které podnik vyrábí, má konkurence. Náhodným výběrem 100 spotřebitelů jsme zjistili, že 34 z nich používá výrobky konkurence, zbytek výrobky podniku. Určete 95%-ní interval spolehlivosti pro podíl konkurenčních výrobků na trhu.

Řešení
Nechť X_i je náhodná veličina, která nabývá hodnotu 1, když $i-$tá osoba používá výrobek konkurence, a hodnotu 0 jinak.; $i = 1, 2, \ldots, 100$.
Potom $X_i \sim A(\vartheta)$ a $X_1, \ldots, X_n$ je náhodný výběr z alternativního rozložení. Našim úkolem je určit interval spolehlivosti pro parametr ϑ tohoto rozložení.
$n = 100 \quad m = \frac{34}{100} \quad u_{1-\alpha/2} = u_{0,975} = 1,96$
Podmínka aproximace normálním rozložením je $n\vartheta(1-\vartheta) > 9$. Jelikož parametr ϑ není znám, nahradíme ho odhadem m.
$100.0, 34.0, 66 = 22, 44 > 9$. Tedy pro odhad m je podmínka splněna. Potom:

$$d = 0,34 - \sqrt{\tfrac{0,34.0,66}{100}} \cdot 1,96 = 0,2472 \qquad h = 0,34 + \sqrt{\tfrac{0,34.0,66}{100}} \cdot 1,96 = 0,4328$$

Tedy $0,2472 < \vartheta < 0,4328$ s pravděpodobností přibližně 0,95.

[ϑ je pravděpodobnost, že náhodně vybraná osoba bude užívat výrobek konkurence, tato pravděpodobnost je z intervalu (0,2472; 0,4328). Tomuto intervalu můžeme věřit na přibližných 95%.]

Věta 7.4
Nechť $X_1, \ldots, X_n$ je náhodný výběr z $A(\vartheta)$, $c \in (0,1)$, M je výběrový průměr a nechť je splněna podmínka
$n\vartheta(1-\vartheta) > 9$.
Na asymptotické hladině významnosti α se nulová hypotéza $H_0 : \vartheta = c$ zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium
$T_0 = \frac{M-c}{\sqrt{\frac{c(1-c)}{n}}}$ realizuje v oboru W, kde

pro oboustrannou alternativu	$H_1 : \vartheta \neq c$	je $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$
pro levostrannou alternativu	$H_1 : \vartheta < c$	je $W = (-\infty,\ -u_{1-\alpha}\rangle$
pro pravostrannou alternativu	$H_1 : \vartheta > c$	je $W = \langle u_{1-\alpha},\ \infty)$

[Platí-li H_0, pak $T_0 \approx N(0,1)$.]

Poznámka 7.5
Testovací kritérium je odvozeno z Moivre-Laplaceovy věty. Při platné H_0 je $T_0 = U = \frac{M-\vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}} \approx N(0,1)$

Poznámka 7.6
Statistiky, z níchž jsme odvozovali intervalové odhady a testovací kritérium v předchozí větě, nejsou stejné. Testování hypotézy pomocí buď testovacího kritéria, nebo intervalu spolehlivosti, nemusí nutně vést ke stejnému závěru.

Příklad 7.7
Deklarovaná pravděpodobnost vyrobení zmetku při výrobě určité součástky je $\vartheta = 0,01$. Bylo náhodně vybráno 1000 výrobků a zjistilo se, že mezi nimi je 16 zmetků. Na asymptotické hladině významnosti 0,05 testujte hypotézu $H_0 : \vartheta = 0,01$ proti $H_1 : \vartheta \neq 0,01$.

Řešení
Podmínka aproximace normálním rozložením $n\vartheta(1-\vartheta) > 9$ je při neznámém ϑ nahrazena podmínkami $nm(1-m) > 9$ a $nc(1-c) > 9$
$1000 \cdot \frac{16}{1000} \cdot \frac{984}{1000} = 15,744 > 9$ a $1000 \cdot 0,01 \cdot 0,99 = 9,9 > 9$, tedy aproximace normálním rozložením je možná.
Realizace testového kritéria je $t_0 = \frac{\frac{16}{1000} - 0,01}{\sqrt{\frac{0,01 \cdot 0,99}{1000}}} = 1,907$
Kritický obor je $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty) = (-\infty;\ -1,96\rangle \bigcup \langle 1,96;\ \infty)$.
Protože $1,907 \notin W$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.
[Tedy není na základě naměřených hodnot důvod pochybovat o tom, že pravděpodobnost vyrobení zmetku je 0,01.]

Věta 7.8
Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z alternativního rozložení $A(\vartheta_1)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr z $A(\vartheta_2)$. Nechť je splněna podmínka $n_i\vartheta_i(1-\vartheta_i) > 9,\ i = 1,2$.

Nechť $M_1,\ M_2$ jsou výběrové průměry. Pak statistika

$$U = \frac{(M_1-M_2)-(\vartheta_1-\vartheta_2)}{\sqrt{\frac{M_1(1-M_1)}{n_1}+\frac{M_2(1-M_2)}{n_2}}} \approx\ N(0,1).$$

Potom meze $100(1-\alpha)\%-$ního asymptotického empirického intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$ jsou:

$$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1}+\frac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha/2}$$

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1}+\frac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha/2}$$

Příklad 7.9

Management supermarketu vyhlásil týden slev a sledoval, zda toto vyhlášení má vliv na podíl větších nákupů (nad 500 Kč). Na základě náhodného výběru 200 zákazníků v týdnu bez slev bylo zjištěno 97 velkých nákupů, zatímco v týdnu se slevou z 300 náhodně vybraných zákazníků učinilo velký nákup 162 zákazníků. Sestrojte 95% asymptotický interval spolehlivosti pro rozdíl pravděpodobností uskutečnění většího nákupu v týdnu bez slevy a v týdnu se slevou.

Řešení

Zavedeme náhodnou veličinu $X_{1,i}$, která bude nabývat hodnoty 1, když v týdnu bez slevy $i-$tý náhodně vybraný zákazník uskuteční větší nákup a hodnoty 0 jinak, $i = 1, \ldots, 200$. Náhodné veličiny $X_{1,1}, \ldots, X_{1,200}$ tvoří náhodný výběr z rozložení $A(\vartheta_1)$. Dále zavedeme náhodnou veličinu X_{2i}, která bude nabývat hodnoty 1, když v týdnu se slevou $i-$tý náhodně vybraný zákazník uskuteční větší nákup a hodnoty 0 jinak, $i = 1, \ldots, 300$. Náhodné veličiny $X_{2,1}, \ldots, X_{2,300}$ tvoří náhodný výběr z rozložení $A(\vartheta_2)$ a je na předchozím výběru nezávislý. $n_1 = 200, \quad n_2 = 300, \quad m_1 = \frac{97}{200}, \quad m_2 = \frac{162}{300}$. V podmínkách aproximace normálním rozložením $n_i\vartheta_i(1-\vartheta_i) > 9,\ i = 1, 2$ neznáme parametry $\vartheta_1,\ \vartheta_2$, ale můžeme je nahradit odhady m_1 a m_2. Tedy podmínky jsou splněny:
$200 \cdot \frac{97}{200} \cdot \frac{103}{200} = 49,955 > 9,\ \ 300 \cdot \frac{162}{300} \cdot \frac{138}{300} = 74,52 > 9.$
Meze $100(1-\alpha)\%$ asymptotického empirického intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$ tedy jsou:

$$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1}+\frac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha/2} =$$

$$= \frac{97}{200} - \frac{162}{300} - \sqrt{\frac{\frac{97}{200}\left(1-\frac{97}{200}\right)}{200}+\frac{\frac{162}{300}\left(1-\frac{162}{300}\right)}{300}} \cdot 1,96 = -0,1443$$

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1}+\frac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha/2} =$$

$$= \frac{97}{200} - \frac{162}{300} + \sqrt{\frac{\frac{97}{200}\left(1-\frac{97}{200}\right)}{200}+\frac{\frac{162}{300}\left(1-\frac{162}{300}\right)}{300}} \cdot 1,96 = 0,0343$$

Zjistili jsme tedy, že s pravděpodobností přibližně 0,95 je parametrická funkce
$\vartheta_1 - \vartheta_2 \in (-0,1443\ ,\ 0,0343)$.

Věta 7.10

Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z alternativního rozložení $A(\vartheta_1)$ a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr z $A(\vartheta_2)$, $M_1,\ M_2$ jsou výběrové průměry.
Nechť je splněna podmínka $n_i\vartheta_i(1-\vartheta_i) > 9;\ i = 1, 2$.

Na asymptotické hladině významnosti α se nulová hypotéza $H_0 : \vartheta_1 - \vartheta_2 = c$ zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium
$T_0 = \frac{(M_1-M_2)-c}{\sqrt{\frac{M_1(1-M_1)}{n_1}+\frac{M_2(1-M_2)}{n_2}}}$ realizuje v oboru W, kde

pro oboustrannou alternativu	$H_1 : \vartheta_1 - \vartheta_2 \neq c$	je $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$
pro levostrannou alternativu	$H_1 : \vartheta_1 - \vartheta_2 < c$	je $W = (-\infty,\ -u_{1-\alpha}\rangle$
pro pravostrannou alternativu	$H_1 : \vartheta_1 - \vartheta_2 > c$	je $W = \langle u_{1-\alpha},\ \infty)$

[Platí-li H_0, pak $T_0 \approx N(0,1)$.]

Poznámka 7.11
Je-li speciálně $H_0 : \vartheta_1 - \vartheta_2 = 0$, tedy $c = 0$, pak vhodnějším testovacím kritériem je

$$T_0 = \frac{M_1-M_2}{\sqrt{M_*(1-M_*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \qquad \text{kde } M_* = \frac{n_1M_1+n_2M_2}{n_1+n_2}.$$

Platí-li H_0, pak $T_0 \approx N(0,1)$.

Příklad 7.12
Pro údaje z příkladu 7.9 testujte na asymptotické hladině významnosti 0,05 hypotézu, že týden se slevami nezvýší pravděpodobnost uskutečnění většího nákupu.

Řešení
Testujeme hypotézu $H_0 : \vartheta_1 - \vartheta_2 = 0$ proti levostranné alternativě $H_1 : \vartheta_1 - \vartheta_2 < 0$ na asymptotické hladině významnosti 0,05.
$n_1 = 200,\ n_2 = 300,\ m_1 = \frac{97}{200},\ m_2 = \frac{162}{300},\ m_* = \frac{97+162}{500} = 0,518.$
Podmínky aproximace normálním rozložením byly ověřeny v příkladu 7.9.
ad a) Testování pomocí intervalu spolehlivosti:
Pro levostrannou alternativu použijeme pravostranný interval spolehlivosti:
$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}} \cdot u_{1-\alpha} =$
$= \frac{97}{200} - \frac{162}{300} + \sqrt{\frac{\frac{97}{200}\left(1-\frac{97}{200}\right)}{200} + \frac{\frac{162}{300}\left(1-\frac{162}{300}\right)}{300}} \cdot 1,645 = 0,02$
Protože číslo $c = 0$ je obsaženo v intervalu$(-\infty\ ;\ 0,02)$, H_0 nezamítáme na asymptotické hladině významnosti 0,05, tedy slevy nemají vliv na podíl větších nákupů.

ad b) Testování pomocí kritického oboru:
Testové kritérium je:
$T_0 = \frac{M_1-M_2}{\sqrt{M_*(1-M_*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \qquad \text{kde } M_* = \frac{n_1M_1+n_2M_2}{n_1+n_2}$
$m_* = \frac{200\cdot\frac{97}{200}+300\cdot\frac{162}{300}}{200+300} = 0,518$
$t_0 = \frac{\frac{97}{200}-\frac{162}{300}}{\sqrt{0,518(1-0,518)\left(\frac{1}{200}+\frac{1}{300}\right)}} = -1,2058$
Kritický obor je:
$W = (-\infty\ ,\ -u_{1-\alpha}\rangle = (-\infty\ ,\ -u_{0,95}\rangle = (-\infty\ ,\ -1,645\rangle.$

Protože $t_0 \notin W,\ H_0$ nezamítáme na asymptotické hladině významnosti 0,05.

8 Analýza rozptylu jednoduchého třídění

Problematiku uvedeme dvěma příklady. Zajímá nás, jestli je produktivita manuálního pracovníka (vyjádřená počtem vyrobených kusů) ovlivněna denní dobou (ráno, dopoledne, odpoledne, večer, v noci).
Nebo jestli je doba kojení (vyjádřená počtem týdnů) ovlivněna vzděláním matky (vzdělání základní, středoškolské, vysokoškolské). Obecně analýza rozptylu řeší problém, jestli má náhodná veličina nominálního typu (faktor A) vliv na náhodnou veličinu X intervalového, či poměrového typu. V prvním příkladě se náhodná veličina A - "denní doba"- realizuje pěti hodnotami, říkáme, že faktor A má 5 úrovní. Obdobně v druhém příkladě má faktor "vzdělání matky"3 úrovně.
Abychom zjistili, jestli má faktor A vliv na náhodnou veličinu X, pořídíme pro každou úroveň i faktoru A příslušných n_i nezávislých pozorování náhodné veličiny X. Tedy nabývá-li faktor A právě r úrovní, pak ke každé úrovni přiřadíme jeden náhodný výběr a dále požadujeme, aby těchto r výběrů bylo navzájem stochasticky nezávislých:

faktor A	Náhodný výběr
úroveň 1	$X_{11}, \ldots, X_{1n_1} \sim N(\mu_1, \sigma^2)$
úroveň 2	$X_{21}, \ldots, X_{2n_2} \sim N(\mu_2, \sigma^2)$
$\vdots$	$\vdots$
úroveň r	$X_{r1}, \ldots, X_{rn_r} \sim N(\mu_r, \sigma^2)$

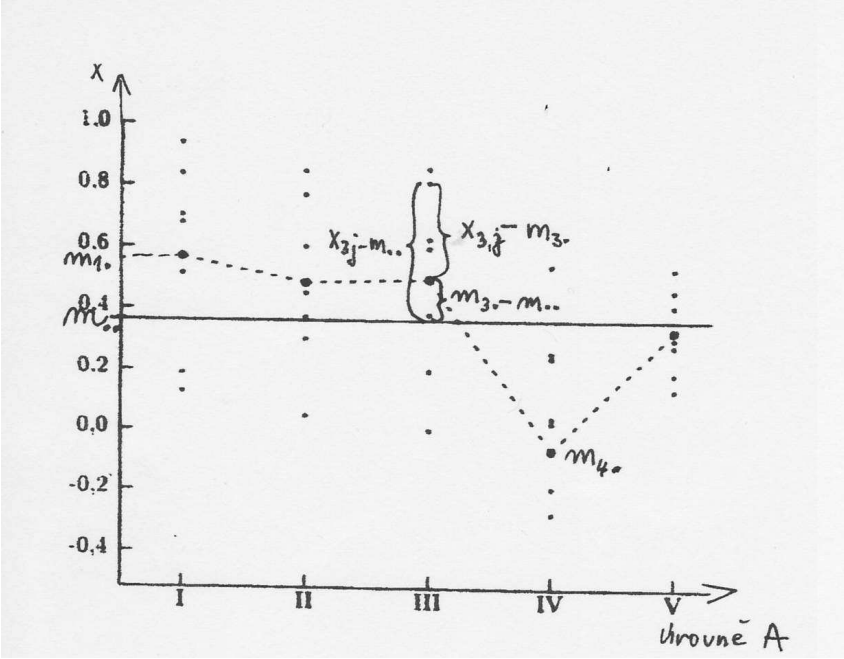
Pokud faktor A nemá vliv na náhodnou veličinu X, pak by střední hodnoty $\mu_1,\ \mu_2, \ldots, \mu_r$ měli být stejné. Tedy testujeme na hladině α hypotézu:

$H_0 :\ \mu_1 = \mu_2 = \ldots = \mu_r$ proti alternativní hypotéze
H_1 : Aspoň jedna dvojice středních hodnot se liší.

Všimněte si, že nabývá-li faktor A právě dvou hodnot, jedná se vlastně o dvouvýběrový $t-$test na hladině α. (Je věk ovlivněn pohlavím?)
S tím souvisí otázka, jestli by hypotézu H_0 nešlo testovat pomocí $\binom{r}{2}$ separátních dvouvýběrových t-testů, každý na hladině α? Pokud by alespoň jedna dvojice zamítla rovnost středních hodnot, pak bychom nulovou hypotézu o rovnosti všech středních hodnot zamítli a současně bychom věděli, které dvojice se liší. Tento postup ovšem nesplňuje podmínku, že pravděpodobnost chyby prvního druhu má být nejvýše α. (Chyba by byla podstatně větší.)
Metoda ANOVA (Analysis of variance), kterou si uvedeme, danou podmínku splňuje. Tato metoda slouží k testování H_0 o shodě středních hodnot. Neslouží tedy k porovnávání rozptylů, jak by se mohlo zdát z názvu. Rozklad (analýza) rozptylů je pouze prostředkem k rozhodnutí o H_0 o shodě středních hodnot. Zamítneme-li na hladině α nulovou hypotézu, pak nás dále zajímá, které dvojice středních hodnot se od sebe liší. K tomu slouží metody mnohonásobného porovnávání např. Scheffého, nebo Tukeyova metoda. Než přejdeme k samotnému testování, uvedeme si označení obvyklé pro analýzu rozptylu.

Označení 8.1

$n = \sum_{i=1}^{r} n_i$	celkový rozsah všech r výběrů
$M_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$	výběrový průměr v $i-$tém výběru
$M_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij}$	celkový průměr všech r výběrů
$S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{\cdot\cdot})^2$	celkový součet čtverců
	(statistika S_T má $f_T = n - 1$ stupňů volnosti)
$S_A = \sum_{i=1}^{r} n_i \cdot (M_{i\cdot} - M_{\cdot\cdot})^2$	meziskupinový součet čtverců
	(statistika S_A má $f_A = r - 1$ stupňů volnosti)
$S_E = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - M_{i\cdot})^2$	vnitroskupinový, neboli reziduální součet čtverců
	(statistika S_E má $f_E = n - r$ stupňů volnosti)



Poznámka 8.2

•Statistika S_T charakterizuje celkovou variabilitu všech pozorování X_{ij} kolem společného průměru $M_{\cdot\cdot}$, tedy až na koeficient $\frac{1}{n-1}$ představuje rozptyl náhodné veličiny X.

•Statistika S_A charakterizuje variabilitu mezi jednotlivými r výběry, tedy charakterizuje vliv faktoru A.

•Statistika S_E charakterizuje variabilitu uvnitř jednotlivých výběrů způsobenou náhodnými vlivy, tedy nevysvětlenou faktorem A.

Ke každé sumě čtverců určujeme tzv. *stupně volnosti*, dané počtem veličin, které jsou

v rámci sumy nezávislé. Uvážíme-li statistiku S_T, pak n pozorováním odpovídá n sčítanců sumy. Ovšem tyto sčítance nejsou zcela libovolné - musí vyhovovat vedlejší podmínce $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-M_{..})=0$. Proto S_T má $f_T=n-1$ stupňů volnosti. Obdobně statistika S_A má r pozorování, které musí vyhovovat vedlejší podmínce $\sum_{i=1}^{r} n_i(M_{i.}-M_{..})=0$. Proto S_A má $f_A=r-1$ stupňů volnosti.
Pro počet stupňů volnosti statistiky S_E platí vztah $f_E = f_T - f_A$. (Plyne ze vztahu $S_T = S_A + S_E$, který bude uveden v následující větě.) Tedy $f_E = n - r$.

Věta 8.3
Vzhledem k označení 9.1 platí:

1.) $S_T = S_A + S_E$.

2.) $S_*^2 = \frac{S_E}{n-r}$, kde S_*^2 je vážený průměr výběrových rozptylů.

3.) $\frac{S_E}{\sigma^2} \sim \chi^2(n-r)$. $[E(\frac{S_E}{n-r}) = \sigma^2]$

4.) Veličiny $\frac{S_E}{\sigma^2}$ a $\frac{S_A}{\sigma^2}$ jsou stochasticky nezávislé.

Platí-li $H_0:\ \mu_1 = \mu_2 = \ldots = \mu_r$ pak

5.) $\frac{S_A}{\sigma^2} \sim \chi^2(r-1)$. $[E(\frac{S_A}{r-1}) = \sigma^2$ platí-li $H_0.]$

8.4 Testování hypotézy o shodě středních hodnot

Testujeme na hladině α hypotézu:

$H_0:\ \mu_1 = \mu_2 = \ldots = \mu_r$ proti alternativní hypotéze
H_1 : Aspoň jedna dvojice středních hodnot se liší.

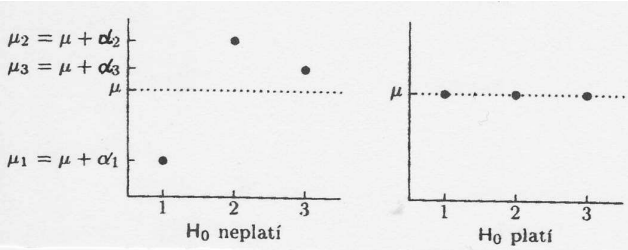
Pro náhodnou veličinu $X_{ij},\ i = 1,\ldots,r;\ j = 1,\ldots,n_i$ platí: $X_{ij} \sim N(\mu_i, \sigma^2)$.
Proto X_{ij} můžeme zapsat také jako:

$$\boxed{\begin{aligned} X_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij}, \end{aligned}}$$

kde ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0,\sigma^2)$
μ je část střední hodnoty X společná všem r vyšetřovaným náhodným výběrům
α_i je efekt faktoru A na úrovni i.
(Parametry μ a α_i jsou neznámé a požadujeme platnost reparametrizační rovnice $\sum_{i=1}^{r} n_i\alpha_i = 0$)

Nulová hypotéza platí, když na faktoru A nezáleží. Můžeme ji tedy přepsat do tvaru
$H_0:\ \alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$
Při platné H_0 potom pro X_{ij} platí:

$$\boxed{X_{ij} = \mu + \varepsilon_{ij}}$$



Statistika $M_{i\cdot}$ je bodovým odhadem střední hodnoty μ_i
Statistika $M_{\cdot\cdot}$ je bodovým odhadem střední hodnoty μ
Statistika $M_{i\cdot} - M_{\cdot\cdot}$ je bodovým odhadem efektu $\alpha_i = \mu_i - \mu$

Samotné rozhodování o nulové hypotéze je založeno na porovnání průměrných čtverců S_A/f_A a S_E/f_E, jejichž střední hodnoty jsou při platné nulové hypotéze stejné a testovací statistika $F_A = \frac{S_A/f_A}{S_E/f_E}$ se řídí Fisher-Snedocerovým rozložením $F(f_A, f_E)$.

Proti testované nulové hypotéze svědčí zejména případ, kdy se statistiky $M_{i\cdot}$ hodně liší od $M_{\cdot\cdot}$. Proto na platnost, či neplatnost nulové hypotézy ukazuje statistika S_A (variabilita mezi výběry); statistika S_E slouží k odhadu rozptylu σ^2 a současně dává měřítko pro hodnocení velikosti variability mezi výběry. Proto nulovou hypotézu o shodě středních hodnot (tedy o nevýznamnosti faktoru A) zamítáme na hladině α, když:

$$F_A = \frac{S_A/f_A}{S_E/f_E} \geq F_{1-\alpha}(r-1, n-r)$$

Je zvykem zapisovat výsledky výpočtů do tabulky analýzy rozptylu jednoduchého třídění.

Zdroj variability	součet čtverců	stupně volnosti	průměrný součet čtverců	testová statistika
faktor	S_A	$f_A = r - 1$	S_A/f_A	$F_A = \frac{S_A/f_A}{S_E/f_E}$
rezidua	S_E	$f_E = n - r$	S_E/f_E	
celkový	S_T	$f_T = n - 1$		

Jestliže jsme zamítli nulovou hypotézu na hladině α, znamená to, že se aspoň jedna dvojice středních hodnot liší. Takovéto dvojice můžeme identifikovat pomocí metod *mnohonásobného porovnávání.*

8.5 Tukeyova metoda

Tato metoda je vhodná pro vyvážené třídění, kdy rozsahy všech výběrů jsou stejné, tedy když $p := n_1 = n_2 = \ldots = n_r$.
Testujeme na hladině α hypotézu:

$H_0 : \ \mu_k = \mu_l$ proti alternativní hypotéze
$H_1 : \ \mu_k \neq \mu_l$
Hypotézu o rovnosti $\mu_k = \mu_l$ zamítáme na hladině α, když

$$|M_{k\cdot} - M_{l\cdot}| \geq q_{1-\alpha}(r, n-r)\frac{S_*}{\sqrt{p}},$$

kde hodnoty $q_{1-\alpha}(r, n-r)$ jsou tabelované a nazýváme je *kvantily studentizovaného rozpětí*. Tímto postupem vyznačíme všechny dvojice k, l, pro něž se střední hodnoty $\mu_k,\ \mu_l$ na hladině α liší.

8.6 Scheffého metoda

Tuto metodu užijeme, jsou-li rozsahy r výběrů různé.
Testujeme na hladině α hypotézu:

$H_0:\ \mu_k = \mu_l$ proti alternativní hypotéze
$H_1:\ \mu_k \neq \mu_l$
Hypotézu o rovnosti $\mu_k = \mu_l$ zamítáme na hladině α, když

$$|M_{k\cdot} - M_{l\cdot}| \geq S_* \sqrt{(r-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right) F_{1-\alpha}(r-1, n-r)}$$

Může nastat situace, že hypotézu $H_0:\ \mu_1 = \mu_2 = \ldots = \mu_r$ jsme zamítli a přesto metody mnohonásobného porovnávání nenašly významný rozdíl u žádné dvojice středních hodnot. Pak je významně rozdílná některá složitější kombinace středních hodnot, tzv. *kontrast*.

Připomeňme si nyní předpoklady analýzy rozptylu a dále si uvedeme testy těchto předpokladů.

8.7 Předpoklady analýzy rozptylu

Vzhledem k zavedenému značení požadujeme, aby náhodné výběry splňovali následující vlastnosti:

1.) Normalita: $X_{i1}, \ldots, X_{in_i} \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, r.$

2.) Nezávislost: Jednotlivé náhodné výběry jsou navzájem nezávislé.

3.) Homoskedasticita: Rozptyly jednotlivých výběrů jsou shodné,
tedy $\sigma^2 := \sigma_1^2 = \ldots = \sigma_r^2$

Normalita je buď známa ze zkušeností, nebo užijeme již zmíněné testy normality. (Celkově analýza rozptylu není příliš citlivá na porušení normality.) Nezávislost výběrů musí vyplývat z organizace pokusu. Zbývá ověřit homoskedasticitu, tedy testovat hypotézu, že rozptyly jsou pro všech r porovnávaných výběrů stejné. K tomu slouží následující testy:

8.8 Levenův test

Levenův test je vlastně analýzou rozptylu jednoduchého třídění formálně provedenou na veličinách $|X_{ij} - M_{i\cdot}|$.
Na hladině α testujeme hypotézu:

$H_0:\ \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2 := \sigma^2$ proti alternativní hypotéze
H_1 : Aspoň jedna dvojice rozptylů se liší.

Označme $Z_{ij} = |X_{ij} - M_{i\cdot}|$. Potom v souladu se značením ANOVA je

$M_{Zi} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij}$ průměr odchylek v $i-$tém výběru

$M_Z = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Z_{ij}$ celkový průměr odchylek

$S_{ZA} = \sum_{i=1}^{r} n_i \cdot (M_{Zi} - M_Z)^2$ meziskupinový součet čtverců

$S_{ZE} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Z_{ij} - M_{Zi})^2$ vnitroskupinový součet čtverců

Platí-li hypotéza o shodě rozptylů, pak statistika

$$F_{ZA} = \frac{S_{ZA}/(r-1)}{S_{ZE}/(n-r)} \sim F(r-1, n-r)$$

Nulovou hypotézu o shodě rozptylů zamítáme na hladině α, když $F_{ZA} \geq F_{1-\alpha}(r-1, n-r)$.

8.9 Bartlettův test

Bartlettův test shody rozptylů lze použít tehdy, když rozsahy všech výběrů jsou alespoň 7. Jeho nevýhodou je značná citlivost vůči porušení předpokladů normality pro jednotlivé výběry.

Platí-li hypotéza o shodě rozptylů, pak statistika

$$B = \frac{1}{C} \left((n-r) \ln S_*^2 - \sum_{i=1}^{r} (n_i - 1) \ln S_i^2 \right) \approx \chi^2(r-1), \text{ kde}$$

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

$S_i^2 = \sum_{j=1}^{n_i} \frac{1}{n_i - 1} (X_{ij} - M_{i\cdot})^2$

$S_*^2 = \frac{1}{n-r} \sum_{i=1}^{r} (n_i - 1) S_i^2 = \frac{S_E}{n-r}$

Nulovou hypotézu o shodě rozptylů zamítáme na hladině α, když $B \geq \chi^2_{1-\alpha}(r-1)$.

8.10 Shrnutí závěrem

Postup při analýze rozptylu jednoduchého třídění:

1.) Ověříme předpoklady ANOVY; pro ověření homoskedasticity užijeme *Levenův test*, nebo *Bartlettův test*.

2.) Pomocí tabulky analýzy rozptylu rozhodneme o nulové hypotéze o shodě středních hodnot.

3.) Byla-li hypotéza o shodě středních hodnot zamítnuta, užijeme metody mnohonásobného porovnávání, abychom identifikovali ty dvojice, které způsobili zamítnutí hypotézy o shodě středních hodnot. K tomu můžeme užít *Tukeyovu metodu*, nebo *Scheffeho metodu*.

Příklad 8.11
U čtyř odrůd brambor (označených římskými číslicemi) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky v kg jsou v následující tabulce:

odrůda	hmotnost
I.	0,9 0,8 0,6 0,9
II.	1,3 1,0 1,3
III.	1,3 1,5 1,6 1,1 1,5
IV.	1,1 1,2 1,0

Na hladině významnosti 5% testujte hypotézu, že střední hodnota hmotnosi trsu nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

Řešení
Data považujeme za realizace čtyř nezávislých náhodných výběrů ze čtyř normálních rozložení se stejným rozptylem. Tedy $X_{i1}, \ldots, X_{in_i} \sim N(\mu_i, \sigma^2);\ i = 1, 2, 3, 4.$
Testujeme hypotézu, že všechny čtyři střední hodnoty jsou stejné, tedy:
$H_0 :\ \mu_1 = \mu_2 = \mu_3 = \mu_4$ proti alternativní hypotéze
H_1 : Aspoň jedna dvojice středních hodnot se liší.

Určíme realizace potřebných statistik:
$m_{1\cdot} = 0,8; \quad m_{2\cdot} = 1,2; \quad m_{3\cdot} = 1,4; \quad m_{4\cdot} = 1,1; \quad m_{\cdot\cdot} = 1,14$
$S_E = 0,3; \quad S_A = 0,816; \quad S_T = 1,116$ a dále $r = 4; \quad n = 15$

Zdroj variability	souč. čtverců	stupně volnosti	prům. souč. čtverců	testová statistika
faktor	$S_A = 0,816$	$f_A = r - 1 = 3$	$S_A/3 = 0,272$	$F_A = \frac{S_A/f_A}{S_E/f_E} = 9,97$
rezidua	$S_E = 0,3$	$f_E = n - r = 11$	$S_E/11 = 0,02727$	
celkový	$S_T = 1,116$	$f_T = n - 1 = 14$		

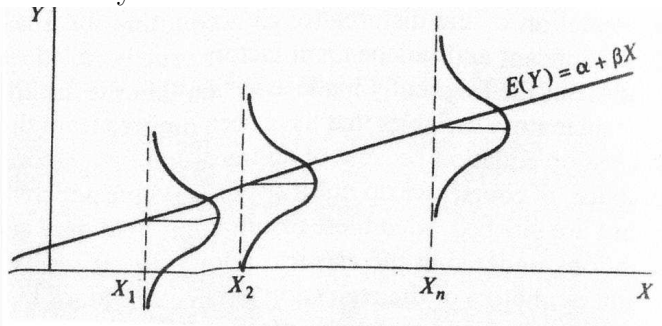
Kritický obor je $W = \langle F_{0,95}(3, 11); \infty) = \langle 3,59; \infty)$. Realizace testové statistiky $9,97 \in W$, proto H_0 o shodě středních hodnot zamítáme na hladině α.
Nyní pomocí Scheffého metody zjistíme, které dvojice odrůd se liší na hladině $\alpha = 0,05$.

Srovnávané odrůdy	Rozdíly $\lvert M_{k\cdot} - M_{l\cdot}\rvert$	Pravá strana nerovnosti
I., II.	0,4	0,41
I., III.	0,67∗	0,36
I., IV.	0,3	0,41
II., III.	0,2	0,40
II., IV.	0,1	0,44
III., IV.	0,3	0,40

Na hladině významnosti 0,05 se liší odrůdy I. a III. (Hvězdičkou v tabulce je vyznačena dvojice, kde rozdíl $\lvert M_{k\cdot} - M_{l\cdot}\rvert$ je významný.

9 Jednoduchá lineární regrese

V této kapitole se budeme věnovat studiu závislostí mezi proměnnými. Příkladem takových závislostí může být vztah mezi nabídkou a poptávkou; vztah mezi příjmy a výdaji; cenou, jako funkcí více proměnných; produktivitou, jako funkcí více proměnných a pod. Ve fyzice a v matematice často pracujeme s deterministickou závislostí popsanou funkcí $y = f(x)$, kde ke každé hodnotě definičního oboru nezávislé proměnné x existuje právě jedna hodnota y závisle proměnná. V ekonomické praxi je takováto závislost velmi vzácná. Běžně se setkáváme se závislostí, která je stochastická, tedy kde ke každé hodnotě nezávislé proměnné x existuje celé rozložení pravděpodobnosti hodnot závislé proměnné Y. To znamená, že hodnotu Y nelze předvídat přesně - je ovlivněna náhodnými vlivy. Regresní analýza se zabývá těmito stochastickými závislostmi. Jejím úkolem je a) jednak určení typu funkce, která danou závislost popisuje, b) a dále pro zvolenou funkci odhadnout její parametry.



ad a)
Při určení typu funkce vycházíme jednak z logického rozboru situace (např. ze známé ekonomické teorie), nebo se hledanou závislost snažíme odhadnout z dvourozměrného tečkového diagramu. (Toto lze pouze tehdy, když vysvětlovaná (závislá) proměnná je funkcí jen jedné vysvětlující (nezávislé) proměnné.)
Uveďme si často užívané typy regresních funkcí:

- regresní přímka: $E(Y|x) = \beta_0 + \beta_1 x$
- regresní parabola: $E(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$
- regresní polynom p-tého stupně: $E(Y|x) = \beta_0 + \beta_1 x + \ldots + +\beta_p x^p$
- regresní hyperbola: $E(Y|x) = \beta_0 + \beta_1 \frac{1}{x}$
- regresní logaritmická funkce: $E(Y|x) = \beta_0 + \beta_1 \ln x$

Všechny uvedené typy jsou příklady jednoduché lineární regresní funkce. (O lineární regresní funkci mluvíme tehdy, když tato funkce je lineární vzhledem k parametrům $\beta_0,\ \beta_1,\ \beta_2, \ldots$. Jednoduchá je tehdy, když závislá proměnná je vysvětlovaná jednou nezávislou proměnnou. Pokud vysvětlujících proměnných je více, muvíme o vícenásobné regresi.)
ad b)
Neznámé parametry $\beta_0,\ \beta_1,\ \beta_2, \ldots$ odhadujeme na základě znalosti n dvojic hodnot $(x_1, y_1), \ldots, (x_n, y_n)$. U lineárních modelů se k nalezení odhadů parametrů nejčastěji užívá metoda nejmenších čtverců.

9.1 Specifikace klasického modelu jednoduché lineární regrese

Model sestává z regresní rovnice a základních předpokladů. Uveďme si nejdříve rovnici:

•$Y = \beta_0 + \beta_1 f_1(x) + \ldots + \beta_p f_p(x) + \varepsilon$, kde:

Y je závisle proměnná náhodná veličina, je pozorovatelná
x je nezávisle proměnná nenáhodná veličina, je pozorovatelná
ε je náhodná odchylka, která zahrnuje působení náhodných vlivů, je nepozorovatelná
$\beta_0 + \beta_1 f_1(x) + \ldots + \beta_p f_p(x)$ je **teoretická** regresní funkce s neznámými parametry $\beta_0, \beta_1, \ldots, \beta_p$

Pro n pozorování regresní rovnici přepíšeme:
$y_1 = \beta_0 + \beta_1 f_1(x_1) + \ldots + \beta_p f_p(x_1) + \varepsilon_1$
$\vdots$
$y_i = \beta_0 + \beta_1 f_1(x_i) + \ldots + \beta_p f_p(x_i) + \varepsilon_i$
$\vdots$
$y_n = \beta_0 + \beta_1 f_1(x_n) + \ldots + \beta_p f_p(x_n) + \varepsilon_n$
$i = 1, \ldots, n$ jsou indexy objektů, na nichž byla náhodná veličina pozorována, nebo v případě časových řad představuje i časový okamžik, kdy bylo pozorování učiněno.

•Předpoklady kladené na náhodné odchylky $\varepsilon_i,\ i = 1, \ldots, n$

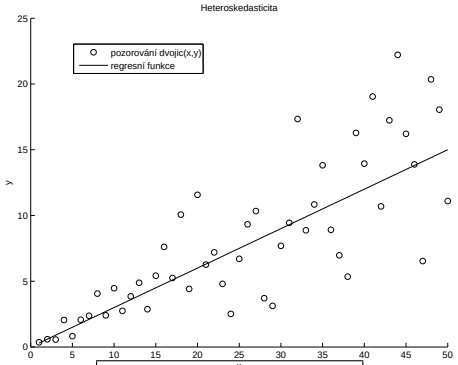
a) $E(\varepsilon_i) = 0$ [odchylky nejsou systematické]

b) $D(\varepsilon_i) = \sigma^2 > 0$ [všechna pozorování jsou se stejnou přesností]

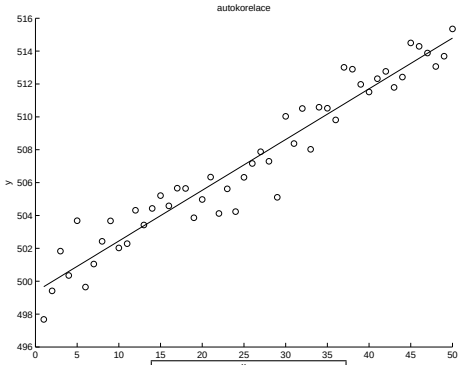
c) $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$ [mezi odchylkami není žádný lineární vztah]

d) $\varepsilon_i \sim N(0, \sigma^2)$ [odchylky jsou normálně rozložené]

Příklady porušení některých předpokladů ukazují následující dva obrázky. V prvním případě je porušen předpoklad b), mluvíme pak o heteroskedasticitě náhodných odchylek; v druhém případě je porušen předpoklad c), mluvíme pak o autokorelaci náhodných odchylek.

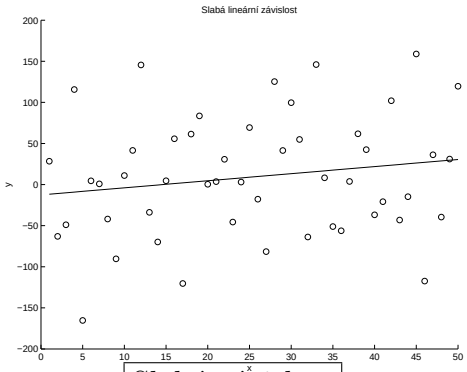


Heteroskedasticita

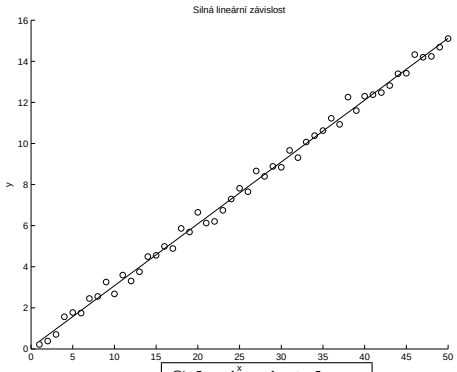


Autokorelace

Další obrázky ukazují příklady slabé a silné lineární závislosti při splněných předpokladech:



Slabá závislost



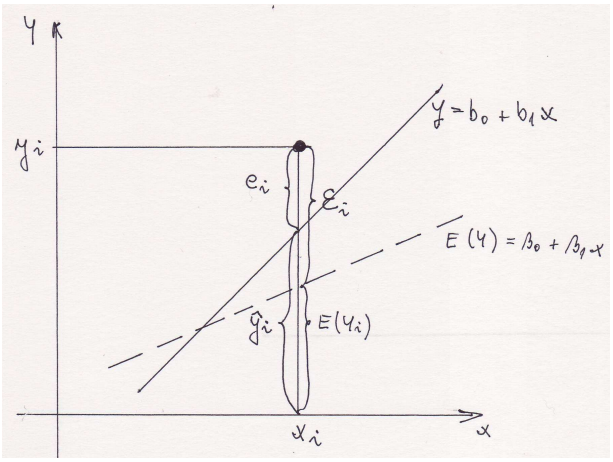
Silná závislost

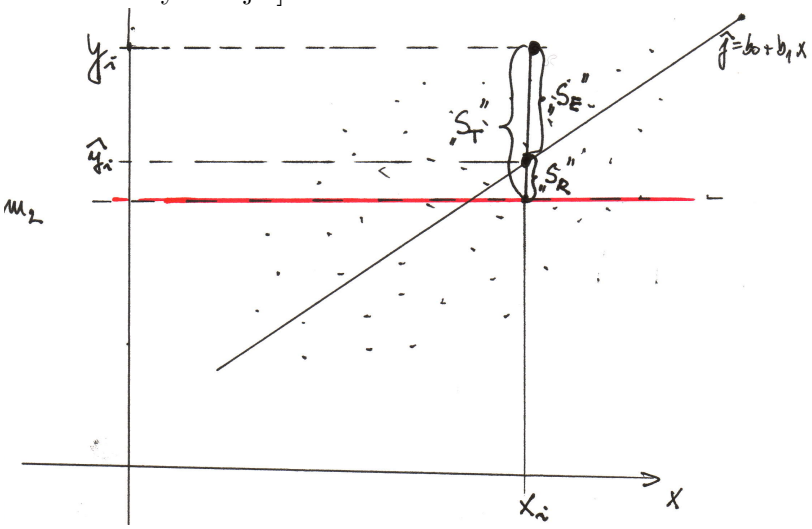
Poté, co je model specifikován je potřeba odhadnout jeho neznámé parametry $\beta_0, \beta_1 \ldots, \beta_p$

9.2 Odhady regresních parametrů a související označení

$b_0, b_1, \ldots, b_p$	odhady regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$
$b_0 + b_1 f_1(x) + \ldots + b_p f_p(x)$	empirická (**odhadnutá**) regresní funkce
$\hat{y}_i = b_0 + b_1 f_1(x_i) + \ldots + b_p f_p(x_i)$	regresní odhad i-té hodnoty náhodné veličiny Y
$e_i = y_i - \hat{y}_i$	i-té reziduum
$S_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$	reziduální součet čtverců
$s^2 = \frac{S_E}{n-p-1}$	odhad rozptylu σ^2
$S_R = \sum_{i=1}^{n} (\hat{y}_i - m_2)^2; \quad m_2 = \frac{1}{n} \sum_{i=1}^{n} y_i$	regresní součet čtverců
$S_T = \sum_{i=1}^{n} (y_i - m_2)^2;$	celkový součet čtverců [platí $S_T = S_R + S_E$]
$ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}$	index determinace [$ID^2 \in (0, 1)$]

[Index determinace udává, jaká část variability náhodné veličiny Y je vysvětlena regresním modelem; čím blíže je ID^2 k 1, tím lépe model data vystihuje.]





9.3 Metoda nejmenších čtverců

Metodou nejmenších čtverců hledáme odhady $b_0, b_1, \ldots, b_p$ regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$ tak, aby součet druhých mocnin reziduí byl co nejmenší. Tedy

$$S(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 f_1(x_i) + \ldots + \beta_p f_p(x_i))\right]^2 \rightarrow \ min$$

Naším úkolem je tedy najít minimum funkce $S(\beta_0, \beta_1, \ldots, \beta_p)$, která závisí jen na neznámých parametrech regresního modelu. (Víme, že v bodech extrému funkce více proměnných jsou parciální derivace - podle všech parametrů - rovny nule). Postup je tedy následující:

1. Určíme derivace $S(\beta_0, \beta_1, \ldots, \beta_p)$ podle všech regresních parametrů.
2. Položíme tyto derivace rovny nule. Tím dostáváme soustavu n rovnic o n neznámých. Tuto soustavu nazýváme *systém normálních rovnic*.
3. Řešením systému normálních rovnic získáme hledané odhady $b_0, b_1, \ldots, b_p$ regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$.

Systém normálních rovnic má pak tvar:

$$\begin{array}{lllllllllll}
\beta_0 \sum 1 & + & \beta_1 \sum f_1 & + & \beta_2 \sum f_2 & + & \ldots & + & \beta_p \sum f_p & = & \sum y_i \\
\beta_0 \sum f_1 & + & \beta_1 \sum f_1^2 & + & \beta_2 \sum f_1 f_2 & + & \ldots & + & \beta_p \sum f_1 f_p & = & \sum y_i f_1 \\
\vdots & & & & & & & & & & \vdots \\
\beta_0 \sum f_p & + & \beta_1 \sum f_p f_1 & + & \beta_2 \sum f_p f_2 & + & \ldots & + & \beta_p \sum f_p^2 & = & \sum y_i f_p
\end{array}$$

kde symbol $\sum$ znamená $\sum_{i=1}^{n}$ a symbol $\sum f_j$ znamená $\sum_{i=1}^{n} f_j(x_i)$.

Takové $\beta_0, \beta_1, \ldots, \beta_p$, které řeší systém normálních rovnic, označíme $b_0, b_1, \ldots, b_p$.

Příklad 9.4

Pro regresní přímku určete odhady b_0, b_1 koeficientů β_0, β_1.

Řešení

Odhady $b_0,\ b_1$ regresních koeficientů přímky získáme ze soustavy rovnic:

$$\begin{array}{lllll}
b_0 \sum_{i=1}^{n} 1 & + & b_1 \sum_{i=1}^{n} x_i & = & \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_i & + & b_1 \sum_{i=1}^{n} x_i^2 & = & \sum_{i=1}^{n} y_i x_i
\end{array} .$$

Řešením soustavy je

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad b_1 = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Odhad regresní přímky je tedy $\hat{y} = b_0 + b_1 x$.

(Uvědomte si, že $b_0,\ b_1$ jsou náhodné veličiny; jsou závislé na realizacích $(x_i,\ y_i)$, zatímco parametry $\beta_0,\ \beta_1$ jsou konstanty.)

9.5 Maticový zápis klasického modelu lineární regrese a jeho řešení

•Model: $y_i = \beta_0 + \beta_1 f_1(x_i) + \ldots + \beta_p f_p(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$

můžeme zapsat v maticovém tvaru: $\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$, tedy

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & f_1(x_1) & \ldots & f_p(x_1) \\ 1 & f_1(x_2) & \ldots & f_p(x_2 1) \\ \vdots & & & \vdots \\ 1 & f_1(x_n) & \ldots & f_p(x_n) \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Užíváme názvy:

$\mathbf{y}$	vektor pozorování vysvětlované veličiny Y
$\mathbf{X}$	regresní matice [předpoklad o hodnosti: $h(X) = p + 1 < n$]
$\boldsymbol{\beta}$	vektor regresních parametrů
$\boldsymbol{\varepsilon}$	vektor náhodných odchylek

•Předpoklady modelu pak lze zapsat: $\boxed{\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})}$

Jak jsme již uvedli v 10.3, odhady $b_0, b_1, \ldots, b_p$ regresních parametrů $\beta_0, \beta_1, \ldots, \beta_p$ získáme řešením systému normálních rovnic. Tento systém i odhady parametrů lze vyjádřit v maticovém zápisu následovně:

$\mathbf{X'X}\boldsymbol{\beta} = \mathbf{X'y}$	systém normálních rovnic
$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$	odhad vektoru $\boldsymbol{\beta}$ získaný metodou nejmenších. čtverců
$\hat{\mathbf{y}} = \mathbf{Xb}$	vektor regresních odhadů
$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$	vektor reziduí

9.6 Vlastnosti odhadu $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$

1. odhad $\mathbf{b}$ je lineární; je lineární funkcí náhodného vektoru $\mathbf{y}$
2. odhad $\mathbf{b}$ je nestranný; platí $E(\mathbf{b}) = \boldsymbol{\beta}$
3. odhad $\mathbf{b}$ má varianční matici $var\ \mathbf{b} = \sigma^2(\mathbf{X'X})^{-1}$
4. odhad $\mathbf{b} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X'X})^{-1})$; normalita plyne z $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ a vlastnosti 1.
5. odhad $\mathbf{b}$ je nejlepší lineární nestranný odhad vektoru $\boldsymbol{\beta}$. (Odhad $\mathbf{b}$ je BLUE - Best Linear Unbiased Estimator)

Poznámka 9.7

Vlastnost 5. je známá jako *Gaussova-Markovova věta.*

To, že odhad $\mathbf{b}$ je nejlepší znamená, že má "nejmenší"možnou varianční matici, přesněji:

rozdíl varianční matice libovolného jiného nestranného odhadu vektoru $\boldsymbol{\beta}$ a varianční matice odhadu $\mathbf{b}$ je matice pozitivně semidefinitní.

Známe-li rozložení nějaké vhodné pivotové náhodné veličiny, můžeme provádět běžné statistické procedury. Nás by zajímali intervaly spolehlivosti a testy pro jednotlivé složky vektoru $\boldsymbol{\beta}$. Rozložení odhadu $\mathbf{b}$ známe. Ovšem parametr σ, který vystupuje ve varianční matici, je neznámý. Proto potřebujeme získat alespoň jeho odhad a dále odhady rozptylů pro jednotlivé složky vektoru $\mathbf{b}$.

$$\text{Varianční matice } var\ \mathbf{b} = \begin{pmatrix} var(b_0) & cov(b_0,b_1) & \dots & cov(b_0,b_p) \\ cov(b_1,b_0) & var(b_1) & \dots & cov(b_1,b_p) \\ \vdots & & \ddots & \vdots \\ cov(b_p,b_0) & cov(b_p,b_1) & \dots & var(b_p) \end{pmatrix} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Je tedy zřejmé, že rozptyly $D(b_j),\ j = 0,1,\dots,p$ jsou diagonální prvky matice $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Připoměňme si z 9.2, že $s^2 = \frac{S_E}{n-p-1}$ je bodovým odhadem parametru σ^2. Potom je $s^2(\mathbf{X}'\mathbf{X})^{-1}$ odhadem varianční matice $var\ \mathbf{b}$ a diagonální prvky matice $s^2(\mathbf{X}'\mathbf{X})^{-1}$ jsou odhady rozptylů $D(b_j)$. Obvykle se užívá následující označení :

v_{jj} $\quad j-$tý diagonální prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$

$s_{b_j} = s \cdot \sqrt{v_{jj}}$ $\quad$ směrodatná chyba odhadu b_j

9.8 Intervaly spolehlivosti pro regresní parametry

Statistika $T_j = \frac{b_j - \beta_j}{s_{b_j}}$ má rozložení $t(n-p-1)$ pro $j = 0,1,\dots,p$.

Proto $100(1-\alpha)\%$-ní interval spolehlivosti pro β_j má meze: $b_j \pm s_{b_j} t_{1-\alpha/2}(n-p-1)$

9.9 Testování významnosti regresních parametrů (dílčí $t-$testy)

Na hladině α pro $j = 0,1,\dots,p$ testujeme:

$H_0:\ \beta_j = 0$ proti $H_1:\ \beta_j \neq 0$.

[Nulová hypotéza tvrdí, že vektor $\mathbf{y}$ vůbec nezávisí na $j-$tém sloupci regresní matice $\mathbf{X}$. Zamítnutí hypotézy H_0 znamená, že odhadnutý parametr je statisticky významný a má smysl ho v modelu uvažovat.]

Testová statistika $T_j = \frac{b_j}{s_{b_j}}$ má rozložení $t(n-p-1)$, pokud H_0 platí.

Kritický obor: $W = (-\infty;\ -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1);\ \infty)$

9.10 Testování významnosti modelu jako celku (celkový F-test)

Na hladině α testujeme:

$H_0:\ (\beta_1,\beta_2,\dots,\beta_p) = (0,0,\dots,0)$ proti $H_1:\ (\beta_1,\beta_2,\dots,\beta_p) \neq (0,0,\dots,0)$.

[Nulová hypotéza tvrdí, že postačující je model s konstantou, tedy $Y = \beta_0 + \varepsilon$. Pokud H_0 nezamítneme, znamená to, že regresní koeficienty $\beta_1,\beta_2,\dots,\beta_p$ můžeme z modelu vypustit a tedy že byl použit nevhodný model.]

Testová statistika: $F = \frac{S_R/p}{S_E/(n-p-1)} \sim F(p,n-p-1)$, pokud H_0 platí.

Kritický obor: $W = \langle F_{1-\alpha}(p,n-p-1);\infty)$

Výsledky $F-$testu se obvykle zapisují ve formě tabulky rozptylu:

zdroj variability	součet čtverců	stupně volnosti	průměrný součet čtverců	testová statistika
model	S_R	p	S_R/p	$\frac{S_R/p}{S_E/(n-p-1)}$
rezidua	S_E	$n-p-1$	$S_E/(n-p-1)$	
celkový	S_T	$n-1$		

Příklad 9.11
U šesti obchodníků byla zjišťována poptávka po určitém druhu zboží loni (veličina X v kusech) a letos (veličina Y v kusech).

číslo. obchodníka	1	2	3	4	5	6
poptávka loni (X)	20	60	70	100	150	260
poptávka letos (Y)	50	60	60	120	230	320

a) Orientačně ověřte předpoklad, že data pocházejí z dvourozměrného normálního rozložení.

b) Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Sestavte regresní matici, vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

c) Najděte odhad rozptylu, vypočtěte index determinace a interpretujte ho.

d) Najděte 95% intervaly spolehlivosti pro regresní parametry.

e) Na hladině významnosti 0,05 proveďte celkový F-test.

f) Na hladině významnosti 0,05 proveďte dílčí t-testy.

g) Vypočtěte regresní odhad letošní poptávky při loňské poptávce 110 kusů.

h) Nakreslete dvourozměrný tečkový diagram s proloženou regresní přímkou.

Řešení
viz. Studijní materiály

10 Úvod do korelační analýzy

Při zpracování dat se velmi často setkáváme s úkolem zjistit, zda dvě náhodné veličiny jsou nezávislé. V případě, že nezávislost vyloučíme, pak nás následně zajímá intenzita závislosti těchto dvou náhodných veličin. Zkoumáme-li závislost náhodných veličin intervalového, či poměrového typu, hovoříme o korelační analýze. [Regresní analýza a korelační analýza řeší příbuzné úlohy. V případě regrese jde o závislost náhodné veličiny na jedné, nebo několika veličinách; v případě korelace jde o sílu vzájemné závislosti dvou "rovnocenných" veličin.] V této přednášce se budeme věnovat pouze lineární závislosti a uvedené testy budou předpokládat normalitu náhodných výběrů.

Poznámka 10.1
V definici 9.12 v prvním semestru byl zaveden koeficient korelace a dále byly uvedeny i jeho vlastnosti.

$$R(X,Y) = E\left(\frac{X-E(X)}{\sqrt{D(X)}} \cdot \frac{Y-E(Y)}{\sqrt{D(Y)}}\right) \text{ pro } \sqrt{D(X)}\sqrt{D(Y)} > 0$$

Připomeneme jeho vlastnosti:

1. $R(X,Y) = \frac{C(X,Y)}{\sqrt{D(X)\cdot\sqrt{D(Y)}}}$ pro $\sqrt{D(X)}\sqrt{D(Y)} > 0$
2. $R(X,X) = 1$ pro $D(X) \neq 0$
3. $R(X,Y) = R(Y,X)$
4. $-1 \leq R(X,Y) \leq 1$
5. $R(X,Y) = 1$, pak exist. konstanty $a, b \in \mathbf{R}, b > 0$ takové, že $P(Y = a + bX) = 1$,
   $R(X,Y) = -1$, pak exist. konstanty $a, b \in \mathbf{R}, b < 0$ takové, že $P(Y = a + bX) = 1$,
6. $R(a + bX, c + dY) = \text{sgn}(bd)R(X,Y)$
7. Jsou-li náhodné veličiny X, Y stochasticky nezávislé, pak $R(X,Y) = 0$.
   (Opačná implikace obecně neplatí!)

Je tedy zřejmé, že pokud je mezi náhodnými veličinami lineární závislost, pak koeficient korelace je vhodný ukazatel intenzity tohoto lineárního vztahu. Čím je hodnota $|R(X,Y)|$ blíže k 1, tím těsnější je lineární závislost mezi náhodnými veličinami X, Y. Kladné hodnoty korelačního koeficientu odpovídají přímé lineární závislosti (k velkým hodnotám jedné z veličin očekáváme spíše velké hodnoty druhé náhodné veličiny); záporné hodnoty korelačního koeficientu odpovídají nepřímé lineární závislosti (k velkým hodnotám jedné z veličin očekáváme spíše malé hodnoty druhé náhodné veličiny). Jsou-li náhodné veličiny nezávislé, pak koeficint korelace je roven 0. (Nulový může být i při některých nelineárních závislostech.) Je obvyklé značit koeficient korelace $R(X,Y)$ řeckým znakem ρ. (Stejně, jako obvykle μ používáme pro $E(X)$ a σ^2 pro $D(X)$).
Určit koeficient korelace většinou nejde přímo, jelikož obvykle není známé simultánní rozložení náhodného vektoru (X,Y). Proto jej musíme odhadnout z náhodného výběru.

Definice 10.2
Uvažujme náhodný výběr $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ z rozložení, kterým se řídí náhodný vektor $\binom{X}{Y}$.

Nechť M_1, M_2 jsou výběrové průměry,
$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)^2$; $S_2^2\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M_2)^2$ jsou výběrové rozptyly a
$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ je výběrová kovariance.
Potom

$$R_{12} = \frac{S_{12}}{S_1 \cdot S_2} \text{ pro } S_1 \cdot S_2 > 0$$

se nazývá výběrový koeficient korelace. Je-li některá z veličin $S_1,\ S_2$ rovna nule, výběrový koeficient korelace se nedefinuje.

Poznámka 10.3
Výběrovou kovarianci lze upravit na tvar $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n} X_iY_i - \frac{n}{n-1}M_1M_2$

Výběrový koeficient korelace lze upravit na tvar $R_{12} = \frac{\sum_{i=1}^{n} X_iY_i - nM_1M_2}{\sqrt{(\sum_{i=1}^{n} X_i^2 - nM_1^2)(\sum_{i=1}^{n} Y_i^2 - nM_2^2)}}$

Poznámka 10.4
Výběrový koeficient korelace R_{12} není nestranným odhadem koeficientu korelace $R(X, Y)$, je odhadem vychýleným. Pro $n > 30$ je toto vychýlení zanedbatelné.
Vlastnosti výběrového koeficientu korelace R_{12} jsou analogické vlastnostem koeficientu korelace $R(X, Y)$. Je ale obtížné správně interpretovat hodnotu výběrového koeficientu korelace, pokud jsou v datovém souboru **odlehlé hodnoty**. Proto kdykoliv je to možné, je potřeba sledovat dvojrozměrný tečkový diagram. Ten také může naznačit i **jinou, než lineární závislost** - v tomto případě korelační koeficient není vhodným ukazatelem závislosti. Nakonec je potřeba zdůraznit, že korelační koeficient měří sílu vzájemné závislosti, která ale nutně **nemusí být příčinná**. Jeho hodnota může naznačovat, že obě proměnné jsou simultánně ovlivněny nějakou třetí proměnnou.

V dalším textu budeme předpokládat, že náhodné výběry $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ pochází z dvourozměrného normálního rozložení.

Věta 10.5
Nechť náhodný vektor X, Y má dvourozměrné normální rozložení. Pak náhodné veličiny X a Y jsou stochasticky nezávislé právě tehdy, když koeficient korelace $\varrho = R(X, Y) = 0$.
[Obecně z nekorelovanosti nezávislost neplyne. Ovšem v případě dvourozměrného normálního rozložení je nekorelovanost a nezávislost ekvivalentní.]

Věta 10.6
Nechť $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ je náhodný výběr z dvourozměrného normálního rozložení a nechť $\varrho = 0$. Potom statistika

$$T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1 - R_{12}^2}}$$

má Studentovo rozložení $t(n-2)$.
Statistiku T z předchozí věty lze užít k testování nezávislosti náhodných veličin X, Y, pocházejících z dvourozměrného normálního rozložení. Z věty 10.5 víme, že jsou-li X, Y nekorelované, pak jsou také nezávislé.

Věta 10.7
Pro výběr z dvourozměrného normálního rozložení se na hladině α nulová hypotéza $H_0: \ \varrho = 0$ zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium $T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$ realizuje v oboru W, kde

pro oboustr. alt.	$H_1:$	$\varrho \neq 0$	je $W = (-\infty, \ -t_{1-\alpha/2}(n-2)\rangle \bigcup \langle t_{1-\alpha/2}(n-2), \ \infty)$
pro levostr. alt.	$H_1:$	$\varrho < 0$	je $W = (-\infty, \ -t_{1-\alpha}(n-2)\rangle$
pro pravostr. alt.	$H_1:$	$\varrho > 0$	je $W = \langle t_{1-\alpha}(n-2), \ \infty)$

Příklad 10.8
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

1	2	3	4	5	6	7	8
80	50	36	58	42	60	56	68
65	60	35	39	48	44	48	61

Na hladině významnosti 0,05 testujte hypotézu, že výsledky obou testů nejsou kladně korelované.

Řešení
Viz. studijní materiály

Dosud jsme testovali nezávislost prostřednictvím hypotézy $H_0: \ \varrho = 0$. Nyní nás bude zajímat, jak velká (těsná) je případná lineární závislost; k tomu užijeme testy o výběrovém korelačním koeficientu. Tyto testy používají Fisherovu z-transformaci výběrového korelačního koeficientu R_{12}.

Věta 10.9
Nechť $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ je náhodný výběr z dvourozměrného normálního rozložení s koeficientem korelace $R(X,Y) = \varrho$. Statistika

$$Z = \frac{1}{2} \ln \frac{1 + R_{12}}{1 - R_{12}}$$

se nazývá *Fisherova z-transformace* a má střední hodnotu a rozptyl:
$E(Z) = \frac{1}{2} \ln \frac{1+\varrho}{1-\varrho} + \frac{\varrho}{2(n-1)}$
$D(Z) = \frac{1}{n-3}$.
Potom standardizovaná statistika $U = \frac{Z - E(Z)}{\sqrt{D(Z)}} \approx N(0,1)$.
(S rostoucím n hodnota výrazu $\frac{\varrho}{2(n-1)}$ klesá. Např pro $\rho = 1$ a $n = 31$ má uvedený výraz hodnotu 1/60=0,01667. Asymptotická statistika v 10.12 tento výraz zanedbává.)

Věta 10.10
Nechť $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}, \ n \geq 10$ je náhodný výběr z dvourozměrného normálního rozložení s

koeficientem korelace ϱ. Nechť R_{12} je výběrový koeficient korelace, nechť $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ je jeho Fisherova z-transformace a nechť $c \in (-1,1)$ je daná konstanta.
Na asymptotické hladině α se nulová hypotéza
$H_0: \ \varrho = c$ zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium

$U = \frac{Z - \frac{1}{2}\ln\frac{1+c}{1-c} - \frac{c}{2(n-1)}}{\sqrt{\frac{1}{n-3}}}$ realizuje v oboru W, kde

pro oboustr. alt.	$H_1:$	$\varrho \neq c$	je $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$
pro levostr. alt.	$H_1:$	$\varrho < c$	je $W = (-\infty,\ -u_{1-\alpha}\rangle$
pro pravostr. alt.	$H_1:$	$\varrho > c$	je $W = \langle u_{1-\alpha},\ \infty)$

Příklad 10.11
U 600 vzorků rudy byl stanoven obsah železa dvěma analytickými metodami s výběrovým koeficientem korelace 0,85. V literatuře se uvádí, že koeficient korelace těchto dvou metod má být 0,9. Na asymptotické hladině významnosti 0,05 testujte hypotézu $H_0 : \varrho = 0,9$ proti $H_1 : \varrho \neq 0,9$.

Řešení
Viz. studijní materiály
Pomocí asymptotické statistiky U lze odvodit meze asymptotického intervalu spolehlivosti pro parametr ϱ. Nejdříve se odvodí meze intervalu spolehlivosti pro konstantu $\frac{1}{2}\ln\frac{1+\varrho}{1-\varrho}$. Potom se takto odvozené meze transformují na meze intervalu spolehlivosti pro parametr ϱ, a to pomocí hyperbolického tangens.

Věta 10.12
Nechť platí předpoklady věty 10.9. Potom pro meze $100(1-\alpha)\%-$ního asymptotického intervalu spolehlivosti
•pro výraz $\frac{1}{2}\ln\frac{1+\varrho}{1-\varrho}$ platí:
$\frac{1}{2}\ln\frac{1+\varrho}{1-\varrho} \in \left(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\ ,\ Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$ s pravděpodobností přibližně $1-\alpha$.
•pro parametr ϱ platí:
$\varrho \in \left(\mathrm{tgh}(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}})\ ,\ \mathrm{tgh}(Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}})\right)$ s pravděpodobností přibližně $1-\alpha$.

Poznámka 10.13
$\mathrm{tgh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ pro $x \in \mathbf{R}$. (Na kalkulačkách s funkcemi je obvykle tlačítko *hyp*, které lze kombinovat s dalšími goniometrickými funkcemi.)

Příklad 10.14
Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi počtem dní absence za rok (veličina Y) a věkem pracovníka (veličina X). Proto náhodně vybral údaje o 10 pracovnících.

pracovník	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Za předpokladu, že uvedené údaje tvoří číselné realizace náhodného výběru rozsahu 10 z dvourozměrného normálního rozložení, vypočtěte výběrový korelační koeficient a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny. Sestrojte 95% asymptotický interval spolehlivosti pro skutečný korelační koeficient ϱ.

Řešení
Viz. studijní materiály

Poznámka 10.15
Máme-li dva výběrové koeficienty korelace R_{12}, R_{12}^*, odpovídající dvěma nezávislým dvourozměrným normálním výběrům, může nás zajímat, jestli jsou koeficienty ϱ, ϱ^* stejné. Tomuto úkolu se věnuje následující věta.

Věta 10.16
Nechť jsou dány dva nezávislé náhodné výběry o rozsazích n a n^* z dvourozměrných normálních rozložení s koeficienty korelace ϱ, ϱ^*. Označme R_{12}, R_{12}^* výběrové koeficienty korelace, Z, Z^* jejich Fisherovy z-transformace.
Na asymptotické hladině α se nulová hypotéza
$H_0: \ \varrho = \varrho^*$ zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium

$U = \frac{Z - Z^*}{\sqrt{\frac{1}{n-3} + \frac{1}{n^*-3}}}$ realizuje v oboru W, kde

pro oboustr. alt. $H_1: \ \varrho \neq \varrho^*$ je $W = (-\infty, \ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2}, \ \infty)$
pro levostr. alt. $H_1: \ \varrho < \varrho^*$ je $W = (-\infty, \ -u_{1-\alpha}\rangle$
pro pravostr. alt. $H_1: \ \varrho > \varrho^*$ je $W = \langle u_{1-\alpha}, \ \infty)$

Příklad 10.17
Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých jedinců činil výběrový korelační koeficient mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že korelační koeficienty v obou skupinách se neliší.

Řešení
Viz studijní materiály

11 Testování nezávislosti nominálních a ordinálních náhodných veličin

Nominální náhodné veličiny se realizují čísly, která představují pouze číselné kódy pro sledovanou vlastnost. Např. "rodinný stav"(svobodná, vdaná, rozvedená, ovdovělá) je nominální náhodná veličina, která nabývá čtyř variant. "Způsob platby"(převodem, poukázkou, hotově) je také nominální náhodná veličina. Tedy jediná možná obsahová interpretace nominálních veličin je u relace rovnosti. (Je realizace rovna zvolenému kódu, nebo není?) V první části této kapitoly se budeme věnovat testům nezávislosti nominálních náhodných veličin; pokud testy nezávislost zamítnou, bude nás zajímat síla (těsnost) případné závislosti.

Ordinální náhodné veličiny se realizují čísly u nichž kromě relace rovnosti je smysluplná i relace uspořádání. Např. *školní klasifikace* (známky 1, 2, 3, 4, 5) představuje větší, nebo menší znalosti zkoušených. Přitom ale nelze říct, že jedničkář je "o dva lepší"než trojkař a trojkař je "o dva lepší"než pětkař. K testování nezávislosti ordinálních náhodných veličin slouží *Spearmanův koeficint pořadové korelace*, který zároveň slouží k měření síly závislosti. Tímto se budeme zabývat v druhé části této kapitoly.

Testování nezávislosti nominálních náhodných veličin

Definice 11.1
Nechť X, Y jsou nominální náhodné veličiny, kde X nabývá varianty $x_{[1]}, \dots, x_{[r]}$ a Y nabývá varianty $y_{[1]}, \dots, y_{[s]}$. Uvažujme náhodný výběr $\binom{X_1}{Y_1}, \dots, \binom{X_n}{Y_n}$ z rozložení, kterým se řídí náhodný vektor $\binom{X}{Y}$. Označme n_{jk} simultánní absolutní četnost dvojice variant $(x_{[j]}, y_{[k]})$. Tabulka obsahující tyto simultánní absolutní četnosti se nazývá *kontingenční tabulka*.

	$y_{[k]}$	$y_{[1]}$	$\dots$	$y_{[s]}$	$n_{j\cdot}$
$x_{[j]}$	n_{jk}				
$x_{[1]}$		n_{11}	$\dots$	n_{1s}	$n_{1\cdot}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_{[r]}$		n_{r1}	$\dots$	n_{rs}	$n_{r\cdot}$
$n_{\cdot k}$		$n_{\cdot 1}$	$\dots$	$n_{\cdot s}$	n

Četnosti $n_{j\cdot} = \sum_{k=1}^{s} n_{jk}$, $n_{\cdot k} = \sum_{j=1}^{r} n_{jk}$ se nazývají marginální četnosti.

Věta 11.2
Uvažujme nulovou hypotézu:
$H_0:$ X, Y jsou stochasticky nezávislé náhodné veličiny proti alternativní hypotéze
$H_1:$ X, Y nejsou stochasticky nezávislé náhodné veličiny.
Platí-li H_0, pak testovací kritérium

$$K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j\cdot} n_{\cdot k}}{n}\right)^2}{\frac{n_{j\cdot} n_{\cdot k}}{n}} \approx \chi^2((r-1)(s-1)).$$

[Říkáme, že K má asymptoticky χ^2 rozložení s $(r-1)(s-1)$ stupni volnosti.]
Hypotézu o nezávislosti veličin X, Y zamítáme na asymptotické hladině významnosti α, když realizace $K > \chi^2_{1-\alpha}((r-1)(s-1))$. Tedy kritický obor $W = \langle \chi^2_{1-\alpha}((r-1)(s-1)), \infty)$.

Poznámka 11.3
Rozložení statistiky K lze aproximovat Pearsonovým rozložením $\chi^2((r-1)(s-1))$, pokud pro výraz $\frac{n_{j\cdot}.n_{\cdot k}}{n}$ platí:
alespoň v 80% případů je $\frac{n_{j\cdot}.n_{\cdot k}}{n} \geq 5$ a ve zbylých nejvýše 20% případů je $\frac{n_{j\cdot}.n_{\cdot k}}{n} \geq 2$.
Nejsou-li podmínky dobré aproximace splněny, doporučuje se slučování některých variant.

Poznámka 11.4
Označme $p_{jk} = P(X = x_{[j]} \wedge Y = y_{[k]}) \quad p_{j\cdot} = \sum_{k=1}^{s} p_{jk} \quad p_{\cdot k} = \sum_{j=1}^{r} p_{jk}$
Veličiny X a Y jsou nezávislé právě tehdy, když platí multiplikativní vztah $p_{jk} = p_{j\cdot} \cdot p_{\cdot k}$.
Testovací kritérium pro nulovou hypotézu o nezávislosti veličin X a Y vychází z myšlenky, že rozdíl absolutních četností n_{jk} a očekávaných teoretických četností za předpokladu nezávislosti $n \cdot p_{jk} = n \cdot p_{j\cdot} \cdot p_{\cdot k}$ by měl být "velmi malý".
Jelikož marginální pravděpodobnostní funkce $p_{j\cdot}$, $p_{\cdot k}$ obvykle nejsou známé, odhadujeme je prostřednictvím marginálních absolutních četností: $\hat{p}_{j\cdot} = \frac{n_{j\cdot}}{n}$ a $\hat{p}_{\cdot k} = \frac{n_{\cdot k}}{n}$. Proto očekávané teoretické četnosti $n \cdot p_{j\cdot} \cdot p_{\cdot k}$. lze odhadnout četnostmi $n \cdot \frac{n_{j\cdot}}{n} \cdot \frac{n_{\cdot k}}{n} = \frac{n_{j\cdot}.n_{\cdot k}}{n}$.
Ve prospěch alternativy tedy svědčí velký rozdíl mezi hodnotami n_{jk} a odhady očekávaných teoretických četností $\frac{n_{j\cdot}.n_{\cdot k}}{n}$. Proto je kritický obor soustředěn na pravém konci Pearsonova rozložení.
Při určení počtu stupňů volnosti si musíme uvědomit, že dvojitá suma $\sum_{j=1}^{r}\sum_{k=1}^{s}$ má sice $r \cdot s$ sčítanců, ale obě marginální pravděpodobnostní funkce jsou vázány vztahem $\sum_{j=1}^{r} p_{j\cdot} = 1$ a $\sum_{k=1}^{s} p_{\cdot k} = 1$. Tedy v první sumě máme celkem $r-1$ nezávislých sčítanců a v druhé sumě máme celkem $s-1$ nezávislých sčítanců. Dvojitá suma má tedy $(r-1)\cdot(s-1)$ nezávislých sčítanců. □

Definice 11.5
Intenzitu závislosti nominálních náhodných veličin X, Y měří *Cramérův koeficient*
$V = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r, s\}$.
Cramérův koeficient je monotónní funkcí statistiky K a nabývá hodnot z intervalu $\langle 0, 1\rangle$.
Pro $V \in \langle 0; 0,1\rangle$ mluvíme o *zanedbatelné závislosti.*
Pro $V \in (0,1; 0,3\rangle$ mluvíme o *slabé závislosti.*
Pro $V \in (0,3; 0,7\rangle$ mluvíme o *střední závislosti.*
Pro $V \in (0,7; 1\rangle$ mluvíme o *silné závislosti.*

Příklad 11.6
V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází a typ školy, na kterou se hlásí. Výsledky jsou zaznamenány v kontingenční tabulce:

	Sociální skupina	I	II	III	IV	$n_{j\cdot}$
Typ školy	n_{jk}					
univerzitní		50	30	10	50	140
technický		30	50	20	10	110
ekonomický		10	20	30	50	110
$n_{\cdot k}$		90	100	60	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérův koeficient.

Řešení

Nejprve vypočteme všech 12 odhadů teoretických četností:

$$\frac{n_{1\cdot}n_{\cdot 1}}{n} = \frac{140\cdot 90}{360} = 35 \quad \frac{n_{1\cdot}n_{\cdot 2}}{n} = \frac{140\cdot 100}{360} = 38,9 \quad \frac{n_{1\cdot}n_{\cdot 3}}{n} = \frac{140\cdot 60}{360} = 23,3 \quad \frac{n_{1\cdot}n_{\cdot 4}}{n} = \frac{140\cdot 110}{360} = 42,8$$
$$\frac{n_{2\cdot}n_{\cdot 1}}{n} = \frac{110\cdot 90}{360} = 27,5 \quad \frac{n_{2\cdot}n_{\cdot 2}}{n} = \frac{110\cdot 100}{360} = 30,6 \quad \frac{n_{2\cdot}n_{\cdot 3}}{n} = \frac{110\cdot 60}{360} = 18,3 \quad \frac{n_{2\cdot}n_{\cdot 4}}{n} = \frac{110\cdot 110}{360} = 33,6$$
$$\frac{n_{3\cdot}n_{\cdot 1}}{n} = \frac{110\cdot 90}{360} = 27,5 \quad \frac{n_{3\cdot}n_{\cdot 2}}{n} = \frac{110\cdot 100}{360} = 30,6 \quad \frac{n_{3\cdot}n_{\cdot 3}}{n} = \frac{110\cdot 60}{360} = 18,3 \quad \frac{n_{3\cdot}n_{\cdot 4}}{n} = \frac{110\cdot 110}{360} = 33,6$$

Podmínky aproximace Pearsonovým rozložením jsou tedy splněny - všechny odhady teoretických četností převyšují číslo 5.

•Nyní určíme realizaci testovacího kritéria:

$$K = \sum_{j=1}^{3}\sum_{k=1}^{4} \frac{\left(n_{jk} - \frac{n_{j\cdot}n_{\cdot k}}{n}\right)^2}{\frac{n_{j\cdot}n_{\cdot k}}{n}} = \frac{(50-35)^2}{35} + \frac{(30-38,9)^2}{38,9} + \ldots + \frac{(50-33,6)^2}{33,6} = 76,84$$

•Kritický obor:

$W = \langle \chi^2_{1-\alpha}((3-1)(4-1)); \infty) = \langle \chi^2_{0.95}(6); \infty) = \langle 12,6; \infty)$

Protože $K \in W$, hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05.

•Dále určíme Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r, s\}$,

tedy $V = \sqrt{\frac{76,84}{360(3-1)}} = 0,3267$.

Hodnota Cramérova koeficientu svědčí o tom, že mezi veličinami X a Y existuje středně silná závislost.

Definice 11.7

Je-li kontingenční tabulka typu 2×2, tedy $r = s = 2$, nazýváme ji *čtyřpolní tabulka*. V tomto případě se obvykle používá označení absolutních četností: $n_{11} = a; n_{12} = b; n_{21} = c; n_{22} = d$.

	$y_{[k]}$	$y_{[1]}$	$y_{[2]}$	$n_{j\cdot}$
$x_{[j]}$	n_{jk}			
$x_{[1]}$		a	b	$a+b$
$x_{[2]}$		c	d	$c+d$
$n_{\cdot k}$		$a+c$	$b+d$	n

○

Pro čtyřpolní tabulky máme k testování hypotézy o nezávislosti nominálních veličin k dispozici tři testy.

1.) Asymptotický χ^2 test ve čtyřpolních tabulkách.
Jeho nevýhodou je, že shoda s limitním rozdělením nastane pouze za splnění dále uvedených předpokladů. Pokud předpoklady nejsou splněny, ve čtyřpolní tabulce již nemůžeme spojovat řádky, nebo sloupce a tento test pak nelze použít.

2.) Asymptotický test, vycházející ze statistily OR, která se nazývá podíl šancí (angl. odds ratio).
Tato statistika neslouží pouze pro test nezávislosti, ale také popisuje sílu případné závislosti; je "obdobou korelačního koeficientu". I tento test je asymptotický a lze ho použít pouze při dostatečně velkých četnostech.

3.) Fisherův přesný faktoriálový test.
Tento test lze použít i v případě, kdy nejsou splněny předpoklady předchozích testů. Jeho nevýhodou však je, že skutečná hladina tohoto testu je menší, než tolerovaná pravděpodobnost α; je to důsledek diskrétního charakteru této testovací procedury.

Věta 11.8
Ve čtyřpolní tabulce lze pro test nezávislosti upravit testovací kritérium K z věty 11.2. na tvar

$$K = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Při platné nulové hypotéze je $K \approx \chi^2(1)$.

Poznámka 11.9
Rozložení statistiky K lze aproximovat Pearsonovým rozložením $\chi^2(1)$, pokud platí podmínky:
$a+b>5;\ c+d>\frac{a+c}{3}$.

Příklad 11.10
U 125 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

	dojem	dobrý	špatný	$n_{j.}$
přijetí	n_{jk}			
ano		17	11	28
ne		39	58	97
$n_{.k}$		56	69	125

Řešení
Nejdříve ověříme splnění podmínek aproximace Pearsonovým rozložením:
$a+b=28>5;\ 97=c+d>\frac{(a+c)}{3}=56/3=18,66$. Podmínky jsou tedy splněny.
•Nyní určíme realizaci testového kritéria:
$K=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}=\frac{125(17\cdot58-11\cdot39)^2}{28\cdot97\cdot56\cdot69}=3,6953$
•Kritický obor:
$W=\langle\chi^2_{0,95}(1),\ \infty)=\langle 3,841;\ \infty)$

Protože testová statistika se nerealizuje v kritickém oboru, nulovou hypotézu o nezávislosti "přijetí"a "dojmu"nezamítáme na asymptotické hladině významnosti 0,05.

Poznámka 11.11
Jiný přístup k hodnocení čtyřpolních tabulek vychází z následující představy. Určitý pokus provádíme za dvojích okolností a může skončit úspěchem, nebo neúspěchem. Lze jej tedy popsat čtyřpolní tabulkou, kde náhodná veličina X nabývá dvou hodnot: úspěch - neúspěch; a náhodná veličina Y nabývá dvou hodnot: nastala okolnost I - nastala okolnost II.

	Okolnost	I	II	$n_{j.}$
Výsledek pokusu	n_{jk}			
úspěch		a	b	$a+b$
neúspěch		c	d	$c+d$
$n_{.k}$		$a+c$	$b+d$	n

Poměr úspěchů k neúspěchům, neboli "šance"za okolnosti I je tedy $\frac{a}{c}$ a za okolnosti II je $\frac{b}{d}$. Nemá-li okolnost vliv na výsledek pokusu, pak podíl šancí za zmíněných dvou okolností $\frac{\frac{a}{c}}{\frac{b}{d}}$ by měl být "blízko"jedné.

Definice 11.12
Uvažujme čtyřpolní tabulku. Statistiku $OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$ nazýváme podíl šancí.
Konstantu $o\rho = \frac{p_{11}p_{22}}{p_{12}p_{21}}$ nazýváme teoretický podíl šancí.

Poznámka 11.13
Jsou-li veličiny X, Y nezávislé, pak $p_{jk} = p_{j.}p_{.k}$. Tedy v případě nezávislosti vychází teoretický podíl šancí $o\rho = 1$. Závislost veličin bude tím větší, čím více se bude $o\rho$ vzdalovat od jedné. Uvědomme si ale, že $o\rho \in \langle 0, \infty)$, tedy hodnoty $o\rho$ jsou kolem bodu 1 rozmístěny nesymetricky. Proto se užívají logaritmické podíly šancí $\ln o\rho$ a $\ln OR$.

Věta 11.14
Uvažujme čtyřpolní tabulku pro dvě nominální náhodné veličiny X, Y.
Statistika $U = \frac{\ln OR - \ln o\rho}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \approx N(0,1)$.

Na asymptotické hladině α se nulová hypotéza
$H_0 :\ \ln o\rho = 0$ [je ekvivalentní s tím, že X, Y jsou stoch. nezávislé] zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium

$U = \frac{\ln OR}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$ realizuje v oboru W, kde

pro oboustr. alt. $H_1 : \ln o\rho \neq 0$ je $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$
pro levostr. alt. $H_1 : \ln o\rho < 0$ je $W = (-\infty,\ -u_{1-\alpha}\rangle$
pro pravostr. alt. $H_1 : \ln o\rho > 0$ je $W = \langle u_{1-\alpha},\ \infty)$ □

Je potřeba si uvědomit, že $\ln o\rho > 0$ právě tehdy, když "šance: úspěch k neúspěchu"je větší za okolnosti I a $\ln o\rho < 0$ právě tehdy, když "šance: úspěch k neúspěchu"je větší za okolnosti II.

Věta 11.15
Uvažujme čtyřpolní tabulku pro dvě nominální náhodné veličiny X, Y.
Potom meze $100(1-\alpha)\%$ asymptotického empirického intervalu spolehlivosti pro teoretický podíl šancí $o\rho$ jsou:
$d = e^{\ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}}$
$h = e^{\ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}}$ □

Nulovou hypotézu o nezávislosti X, Y [ekvivalentní s tím, že $o\rho = 1$] zamítáme, když v asymptotickém intervalu spolehlivosti pro teoretický podíl šancí $o\rho$ hodnota 1 neleží.

Příklad 11.16
Pro údaje z příkladu 11.10. vypočtěte a interpretujte podíl šancí, sestrojte asymptotický interval spolehlivosti pro teoretický podíl šancí a s jeho pomocí testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

Řešení
•Podíl šancí $OR = \frac{ad}{bc} = \frac{17\cdot 58}{11\cdot 39} = 2,298$.
Realizace statistiky OR nám říká, že poměr "přijetí"ku "nepřijetí"je $2,3\times$ větší u uchazeče, který zapůsobil na komisi dobrým dojmem, než u uchazeče, který zapůsobil špatným dojmem.
•Nyní určíme meze $100(1-\alpha)\%$ asymptotického empirického intervalu spolehlivosti pro teoretický podíl šancí:
$d = e^{\ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}} = e^{\ln 2,298 - \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}}\cdot 1,96} = e^{-0,028} = 0,972$
$h = e^{\ln OR + \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}u_{1-\alpha/2}} = e^{\ln 2,298 + \sqrt{\frac{1}{17}+\frac{1}{11}+\frac{1}{39}+\frac{1}{58}}\cdot 1,96} = e^{1,692} = 5,433$

Tedy $o\rho \in (0,972\ ;\ 5,433)$ na asymptotické hladině 5%.
Jelikož interval $(0,972\ ;\ 5,433)$ obsahuje číslo 1, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

Poznámka 11.17
Popis Fisherova faktoriálového testu přesahuje rámec tohoto kurzu. Poznamenejme ale, že test nezávislosti nominálních veličin X, Y formuluje opět prostřednictvím podílu šancí a proti nulové hypotéze $\ln o\rho = 0$ staví nejen oboustrannou, ale i jednostranné alternativy. Vychází-li $p-$hodnota pro zvolenou alternativu $\leq \alpha$, pak nulovou hypotézu o nezávislosti X, Y zamítáme na hladině významnosti α.

Testování nezávislosti ordinálních náhodných veličin

Definice 11.18
Nechť X, Y jsou dvě ordinální náhodné veličiny. Uvažujme náhodný výběr $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ ze spojitého rozložení, kterým se řídí náhodný vektor $\binom{X}{Y}$. Označme R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny Y_i; $i = 1, 2, \ldots, n$. Ukazatelem intenzity pořadové závislosti veličin X, Y je statistika

$$r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2,$$

která se nazývá *Spearmanův koeficient pořadové korelace.*

Poznámka 11.19

Spearmanův koeficient r_S nabývá hodnot z intervalu $\langle -1, 1\rangle$. Čím je jeho hodnota bližší 1 (resp. -1), tím je silnější přímá (resp. nepřímá) pořadová závislost veličin X, Y. Čím je jeho hodnota bližší 0, tím je pořadová závislost veličin X, Y slabší.

Statistika r_S je vlastně obyčejný výběrový koeficient korelace počítaný z pořadí R_i, Q_i, místo z původních hodnot náhodných veličin X_i, Y_i. Úpravou definičního vztahu z 11.2 speciálně pro R_i a Q_i dostaneme vztah z 11.18.

Příkladem „dokonalé“ přímé pořadové závislosti je např vztah mezi R_i, Q_i popsaný tabulkou:

R_i	1	2	3	4	5
Q_i	1	2	3	4	5

Příkladem „dokonalé“ nepřímé pořadové závislosti je např vztah mezi R_i, Q_i popsaný tabulkou:

R_i	1	2	3	4	5
Q_i	5	4	3	2	1

Poznámka 11.20

Spearmanův koeficient pořadové korelace lze použít i tehdy, když jsou data intervalového, či poměrového typu. Např. v 10. kapitole všechny testy nezávislosti předpokládali normalitu. Při porušení normality lze užít testy odvozené od Spearmanova koeficientu. Také jsou situace, kdy v náhodném výběru $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ nelze hodnoty uvedených náhodných veličin přesně stanovit, je k dispozici jen jejich pořadí. Jsou-li pořadí $X-$ových a $Y-$ových veličin hodně podobná, svědčí to o jisté závislosti mezi X_i a Y_i.

Věta 11.21

Nechť X, Y jsou dvě ordinální náhodné veličiny. Uvažujme náhodný výběr $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ z rozložení, kterým se řídí náhodný vektor $\binom{X}{Y}$.

Testujeme hypotézu $H_0:$ X, Y jsou pořadově nezávislé náhodné veličiny.

Na hladině α se nulová hypotéza H_0 zamítá ve prospěch alternativní hypotézy H_1, když se testovací kritérium Spearmanův koeficient r_S realizuje v oboru W, kde

pro oboustr. alt.	$H_1 : X, Y$ jsou pořadově závislé	je $W = (-1,\ -r_{S,1-\alpha/2}(n)\rangle \bigcup \langle r_{S,1-\alpha/2}(n),\ 1)$
pro levostr. alt.	H_1 : mezi X, Y ex. nepřímá p. závisl.	je $W = (-1,\ -r_{S,1-\alpha}(n)\rangle$
pro pravostr. alt.	H_1 : mezi X, Y ex. přímá p. závisl.	je $W = \langle r_{S,1-\alpha}(n),\ 1)$

a $r_{S,1-\alpha}(n)$ je tabelovaná kritická hodnota pro daná $\alpha = 0,05$ a $n = 5, 6, \ldots, 30$. □

Pro větší výběry jsou k dispozici asymptotické testy, kde testovací kritéria mají při nezávislých pořadích asymptoticky normální, či asymptoticky Studentovo rozložení.

Věta 11.22

Nechť platí předpoklady a formulace nulové hypotézy věty 11.21.

Nechť $n > 30$ a nechť platí H_0. Pak testovací kritérium

$$U_0 = r_S\sqrt{n-1} \approx N(0,1)$$

a kritický obor je $W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$. Hypotézu o nezávislosti X, Y zamítáme ve prospěch oboustranné alternativy, když realizace $U_0 \in W$.

Poznámka 11.23
Testové kritérium, které používá sw., má tvar $T_0 = r_S\sqrt{\frac{n-2}{1-r_S^2}}$ a při platné nulové hypotéze má asymptoticky Studentovo rozložení s $(n-2)$ stupni volnosti $t(n-2)$.
Toto asymptotické Studentovo rozložení lze použít pro $n > 20$, tuto podmínku ovšem software neověřuje. Proto pro menší výběry je vhodné použít tabulky pro kritické hodnoty Spearmanova korelačního koeficientu.

Příklad 11.24
Dva lékaři hodnotili stav sedmi pacientů po stejném chirurgickém zákroku. Postupovali tak, že nejvyšší pořadí dostal nejtěžší případ.

i číslo pacienta	1	2	3	4	5	6	7
R_i hodnocení 1. lékaře	4	1	6	5	3	2	7
Q_i hodnocení 2. lékaře	4	2	5	6	1	3	7

Vypočtěte Spearmanův koeficient r_S a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.

Řešení
H_0 : Hodnocení obou lékařů jsou pořadové nezávislá
H_1 : Hodnocení obou lékařů jsou pořadové závislá
•Testovací kritérium je:
$r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2 = 1 - \frac{6}{7(7^2-1)}\{(4-4)^2 + (1-2)^2 + \ldots + (7-7)^2\} = 0.857$
•Kritická hodnota je $r_{S,1-\alpha}(n) = r_{S,0,95}(7) = 0,745$
•Kritický obor je $W = (-\infty,\ -0,745\rangle \bigcup \langle 0,745,\ \infty)$
Jelikož testovací kritérium $r_{S,0,95}(7) = 0,857 \in W$, nulovou hypotézu o nezávislosti hodnocení obou lékařů zamítáme na hladině významnosti 0,05.

12 Neparametrické testy o mediánech

Běžně užívané t-testy (jednoduchý, párový, dvojvýběrový), či Analýza rozptylu jednoduchého třídění (tedy zobecněný t-test), formulují nulovou hypotézu pomocí parametrů normálního, či dvourozměrného normálního rozložení. Proto těmto a podobným testům říkáme parametrické testy. Někdy je ale nelze použít s ohledem na porušení předpokladů. Náhodný výběr nemusí být normální, není splněna homogenita rozptylů (u dvouvýběrového t-testu, či u ANOVY), nebo data mají pouze ordinální charakter. V tomto případě mohou pomoct neparametrické testy.

Již jejich název naznačuje, že budou formulovat hypotézy bez použití parametrů nějakého rozložení. Pokud jsme se původně zajímali např. o parametr μ normálního rozložení, ale nesplněné předpoklady zabránily použití parametrického testu, můžeme svůj zájem přesunout k mediánu rozložení, z něhož náhodný výběr pochází a testovat hypotézy o mediánu. Takovéto neparametrické testy mají však jednu velkou nevýhodu - síla (tedy schopnost zamítnout nepravdivou nulovou hypotézu) těchto testů je menší, než síla parametrických testů. Proto při možnosti volby volíme raději testy parametrické. Co je příčinou menší síly těchto testů? Je to způsobeno tím, že neparametrické testy „zapomenou“ původní hodnoty náhodných výběrů a nahradí je pouze jejich pořadími, či dokonce jen znaménky "+"a "-", nebo znaménky i pořadími zároveň. Právě ztráta části informace obsažené v původním výběru vede k slabší síle neparametrických testů v srovnání s testy parametrickými (ty se snaží využívat veškerou informaci z náhodných výběrů).

V tabulce 12.1. je přehled nejčastěji užívaných neparametrických testů včetně jejich předpokladů. V následujícím textu se budeme podrobně věnovat jen dvěma z nich, ostatní testy jsou podrobněji v učebnici *Průvodce základními statistickými metodami*, 16. kapitola.

Jeden výběr a párový výběr (Analogie parametrického t-testu)	Dva výběry (Analogie parametrického dvouvýběr. t-testu)	$r \geq 3$ výběrů (Analogie ANOVy jednoduch. třídění)
$H_0: \ x_{0,5} = c$ **Znaménkový test** •Výběr pochází ze spojitého rozložení	$H_0: \ x_{0,5} - y_{0,5} = 0$ **Dvouvýběrový Wilcoxonův test** •Výběry pochází ze spojitých na sobě nezávislých rozložení •hustoty $f_1(x)$ a $f_2(y)$ se liší jenom posunutím, tedy rozptyly mají stejné.	H_0 : mediány všech r rozložení jsou stejné **Kruskal-Walisův test** •Výběry pochází ze spojitých na sobě nezávislých rozložení •hustoty se liší jenom posunutím, tedy rozptyly mají stejné.
Wilcoxonův test •Výběr pochází ze spojitého rozložení •Rozložení musí být symetrické dle mediánu	**Kolmogorov-Smirnovův test** •Výběry pochází ze spojitých na sobě nezávislých rozložení •hustoty $f_1(x)$ a $f_2(y)$ mohou mít odlišný tvar	**Mediánový test** •Výběry pochází ze spojitých na sobě nezávislých rozložení

Tabulka 12.1

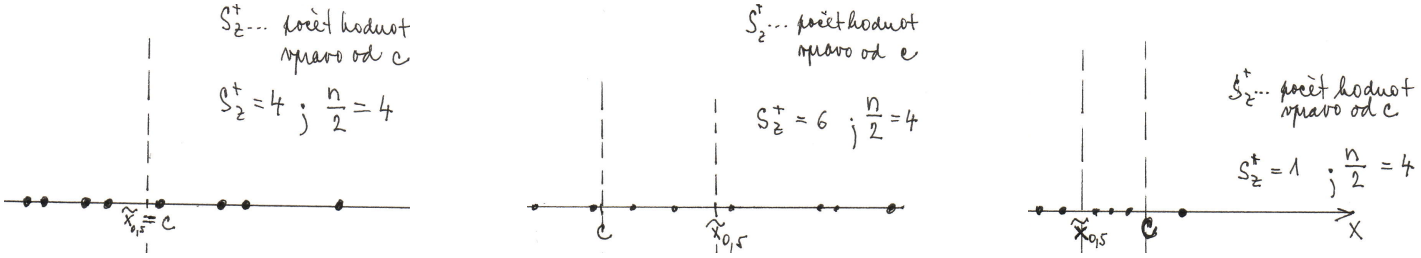
Znaménkový test 12.1

Znaménkový test používáme pro testování hypotéz o mediánu $x_{0,5}$ proti jednostranné i oboustranným alternativám. Tedy

$H_0: \ x_{0,5} = c$ proti $H_1: \ x_{0,5} \neq c$; resp. $H_1: \ x_{0,5} \geq c$; resp. $H_1: \ x_{0,5} \leq c$

Princip testu: Sledujeme, kolik realizací náhodné veličiny X v náhodném výběru rozsahu n je „vpravo“ od konstanty c a „vlevo“ od ní. Je-li c rovno skutečnému mediánu, pak realizací vpravo od c by mělo být „zhruba“ stejně jako realizací vlevo od c. Pokud tomu tak není, pak se zřejmě skutečný medián $x_{0,5}$ od c liší, jak je vidět na následujících obrázcích.

Do statistiky S_Z^+ „vložíme“ počet realizací větších, než c. Pokud je c skutečným mediánem, tak tento počet S_Z^+ by měl být blízký $\frac{n}{2}$. Pokud je hodnota S_Z^+ o hodně větší, či o hodně menší, než $\frac{n}{2}$, pak můžeme těžko věřit tomu, že c je mediánem a nulovou hypotézu zamítneme. Tedy statistika S_Z^+ slouží jako testovací kritérium. *Upozornění:* Je-li nějaká hodnota v náhodném výběru přímo rovna hodnotě c, pak ji z výběru odstraníme a teprve snížený počet pozorování označíme n.



Statistika $S_Z^+ \in \{0, 1, \ldots, n\}$. Pokud realizaci náhodné veličiny X_i vpravo od c budeme považovat za úspěch, pak S_Z^+ udává počet úspěchů a má binomické rozložení. Při platné nulové hypotéze je pravděpodobnost úspěchu $\vartheta = 0,5$ a tedy $S_Z^+ \sim Bi(n; 0,5)$. Jak již bylo řečeno, pozorované hodnoty S_Z^+ blízké k $\frac{n}{2}$ jsou v souladu s nulovou hypotézou.

Známe-li rozložení, můžeme testovat a díky dostupnosti sw. lze přímo počítat p-hodnotu. (Binomické rozložení není tabelované, tedy bez sw. by bylo potřeba používat tabulky pro kritické hodnoty S_Z^+ a pro větší rozsahy odvodit asymptotická testovací kritéria.)

$H_0: \ x_{0,5} = c \Leftrightarrow S_Z^+ = \frac{n}{2}$ •p-hodnota = $P($ statistika $S_Z^+ \geq$ pozorovaný počet úspěchů $S_Z^+)$
$H_1: \ x_{0,5} > c \Leftrightarrow S_Z^+ > \frac{n}{2}$ pomocí sw.: *1-IBINOM(počet pozorovaných úspěchů-1;0,5;n)*

$H_0: \ x_{0,5} = c \Leftrightarrow S_Z^+ = \frac{n}{2}$ •p-hodnota = $P($ statistika $S_Z^+ \leq$ pozorovaný počet úspěchů $S_Z^+)$
$H_1: \ x_{0,5} < c \Leftrightarrow S_Z^+ < \frac{n}{2}$ pomocí sw.: *IBINOM(počet pozorovaných úspěchů ;0,5;n)*

Při výpočtu p-hodnoty pro oboustrannou alternativu musíme rozlišit, zda realizace S_Z^+ je větší, či menší, než $\frac{n}{2}$.

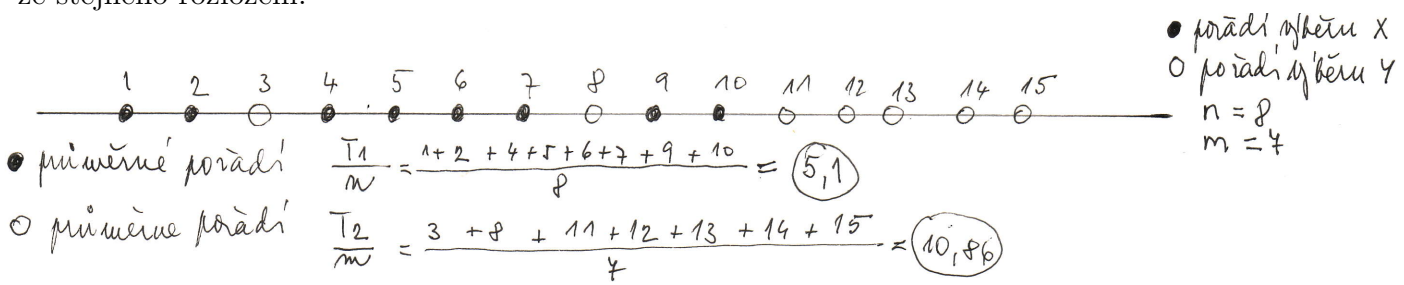
Je-li pozorovaná hodnota $S_Z^+ > \frac{n}{2}$
•p-hodnota = $2P($ statistika $S_Z^+ \geq$ pozorovaný počet úspěchů $S_Z^+)$
$H_0: \ x_{0,5} = c \Leftrightarrow S_Z^+ = \frac{n}{2}$ pomocí sw.: *2*(1-IBINOM(počet pozorovaných úspěchů-1;0,5;n))*
$H_1: \ x_{0,5} \neq c \Leftrightarrow S_Z^+ \neq \frac{n}{2}$ Je-li pozorovaná hodnota $S_Z^+ < \frac{n}{2}$
•p-hodnota = $2P($ statistika $S_Z^+ \leq$ pozorovaný počet úspěchů $S_Z^+)$
pomocí sw.: *2*IBINOM(počet pozorovaných úspěchů;0,5;n)*

Poznámka 12.2
Velmi častá je párová varianta znaménkového testu. Máme-li např. testovat, že se příjmy manželů X a manželek Y neliší, pak nejdříve původně dvojrozměrný výběr převedeme na jednorozměrný výběr rozdílů v příjmech $Z = X - Y$. Pokud se opravdu příjmy partnerů neliší, pak medián proměnné Z je roven nule, tedy $H_0 : \ z_{0,5} = 0$. Takto koncipovaný párový znaménkový test je implementován v sw. Statistica, který počítá pouze asymptotickou p-hodnotu. Přesnější je tedy použít dlouhého jména, jak uvedeno výše. (Podrobnosti o asymptotické variantě testu jsou v učebnici *Průvodce základními statistickými metodami*, 16.2.)

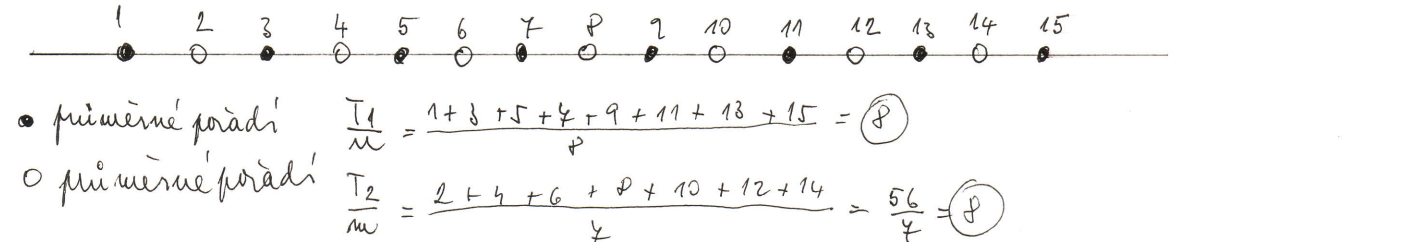
Dvouvýběrový Wilcoxonův test 12.3

V tabulce 12.1 jsou u dvouvýběrového Wilcoxonova testu následující předpoklady: výběr rozsahu n proměnné X je nezávislý na výběru rozsahu m proměnné Y; oba výběry pochází ze spojitých rozložení a hustoty $f_1(x)$ a $f_2(y)$ se liší jenom posunutím. Za těchto předpokladů je nulová hypotéza $H_0 : \ x_{0,5} - y_{0,5} = 0$ ekvivalentní s hypotézou, že oba výběry pochází ze stejného rozložení.

Obr. 1.



Obr. 2.



Princip testu: Oba výběry smícháme dohromady, ale zapamatujeme si původ jednotlivých objektů. Tento smíchaný výběr seřadíme. Od této chvíle "zapomeneme"původní hodnoty veličin X a Y a zapamatujeme si pouze jejich pořadí. Pochází-li oba výběry ze stejného rozložení (- tedy platí H_0), pak jsou původně x-ová a původně y-ová pořadí dokonale promíchána, jak je vidět na Obr.2. Naopak na prvním obrázku Obr.1. jsou takové dva výběry, kde zřejmě hustota proměnné Y je „vpravo“ od hustota proměnné X.
Náhodná veličina T_1 je rovna součtu x-ových pořadí a náhodná veličina T_2 je rovna součtu y-ových pořadí. Jsou-li průměrná pořadí $\frac{T_1}{n}$ a $\frac{T_2}{m}$ velmi podobná, nemáme důvod pochybovat o H_0. Pokud se ale výrazně liší, pak zřejmě uvažované výběry pochází z různých rozložení. Zbývá rozhodnout, jak velká musí odlišnost být, aby bylo nutné nulovou hypotézu zamítnout.

Princip, jak se počítá p-hodnota osvětlíme na následujícím příkladě. Představme si, že

máme dva náhodné výběry, oba rozsahu 3. Všechna možná pořadí jak mohl smíchaný 6-ti prvkový výběr "dopadnout", jsou v tabulce 12.2.

Pořadí prvního výběru	součet pořadí T_1	Pořadí druhého výběru	součet pořadí T_2	Pořadí prvního výběru	součet pořadí T_1	Pořadí druhého výběru	součet pořadí T_2
1,2,3	6	4,5,6	15	2,3,4	9	1,5,6	12
1,2,4	7	3,5,6	14	2,3,5	10	1,4,6	11
1,2,5	8	3,4,6	13	2,3,6	11	1,4,5	10
1,2,6	9	3,4,5	12	2,4,5	11	1,3,6	10
1,3,4	8	2,5,6	13	2,4,6	12	1,3,5	9
1,3,5	9	2,4,6	12	2,5,6	13	1,3,4	8
1,3,6	10	2,4,5	11	3,4,5	12	1,2,6	9
1,4,5	10	2,3,6	11	3,4,6	13	1,2,5	8
1,4,6	11	2,3,5	10	3,5,6	14	1,2,4	7
1,5,6	12	2,3,4	9	4,5,6	15	1,2,3	6

Tabulka 12.2

Například pokud uvažované dva výběry dopadly tak, že pořadí pro první výběr jsou: 3,5,6 a pro druhý jsou: 1,2,4, potom by to mohlo znamenat, že první výběr má hustotu „vpravo" od druhého výběru; statistika T_1 se realizovala spíše velkou hodnotou a tento výsledek svědčí proti H_0. Proto statistiku T_1 můžeme použít jako testovací kritérium. Platí-li nulová hypotéza, pak všech 20 pořadí 6-ti prvkového smíchaného souboru z tabulky 12.2 jsou stejně možné. Pomocí této tabulky můžeme určit rozložení pravděpodobnosti testovacího kritéria T_1.

T_1	6	7	8	9	10	11	12	13	14	15
$p(T_1)$	1/20	1/20	2/20	3/20	3/20	3/20	3/20	1/20	1/20	1/20

Příklad 12.4
Uvažujme hypotézu
H_0 : Oba výběry pochází ze stejného rozložení $\Leftrightarrow\ x_{0,5} = y_{0,5}$
H_1 : První výběr má hustotu „vpravo" od druhého výběru $\Leftrightarrow\ x_{0,5} > y_{0,5}$
Realizace náhodných výběrů:

X	pořadí	Y	pořadí
30	5	20	1
33	6	24	2
26	3	28	4
	$T_1 = 14$		$T_2 = 7$

p-hodnota je součet pravděpodobností všech těch pořadí z tabulky 12.2 které stejně, nebo více, než pozorované pořadí svědčí proti nulové hypotéze.
Tedy $p = P(T_1 \geq 14) = P(T_1 = 14) + P(T_1 = 15) = 1/20 + 1/20 = 0,1$ Pokud by hladina testu byla $\alpha = 0,05$, pak $p > \alpha$ a H_0 nezamítáme na hladině α. Neprokázali jsme, že na hladině 5% je první výběr „posunutý doprava" vzhledem k druhému výběru. □

Pro dostatečně velké rozsahy výběrů ($n + m \geq 12$) slouží jako asymptotické testovací kritérium statistika

$$U = \frac{T_1 - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

která má při platné nulové hypotéze ($H_0 : x_{0,5} = y_{0,5}$) přibližně normální rozložení.

Poznámka 12.5
Test lze použít i když se některé hodnoty ve výběrových souborech a tedy i jejich pořadí opakují. V tomto případě se používá *průměrné pořadí* a asymptotickou testovou statistiku U je nutno korigovat.

Poznámka 12.6
Software Statistica počítá p-hodnotu pomocí ekvivalentního Mann-Whitneyova testu. Přesná p-hodnota je označená „2*1str.přesné p“ a sw. ji nabízí pro malé výběry. Vždy je na výstupu i asymptotická p-hodnota a její výpočet je popsaný v učebni *Průvodce základními statistickými metodami*, 16.4. Software poskytuje přesnou i asymptotickou p-hodnotu jen pro oboustrannou alternativu.
(Pro jednostranné alternativy je potřeba oboustrannou p-hodnotu dělit dvěma, ale pouze, je-li pozorované testové kritérium "na straně"alternativní hypotézy.)