7. Parametrické úlohy o dvou nezávislých náhodných výběrech

7.1. Dva nezávislé náhodné výběry z normálních rozložení

7.1.1. Rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů

Nechť $X_{11},\dots,X_{n_1 1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{12},\dots,X_{n_2 2}$ je na něm nezávislý náhodný výběr z rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$. Označme M_1, M_2 výběrové průměry a S_1^2, S_2^2 výběrové rozptyly. Pak platí:

a) Statistiky $M_1 - M_2$ a $S_*^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ jsou stochasticky nezávislé.

b) $M_1 - M_2 \sim N(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2})$, tedy $U = \dfrac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0, 1)$.

(Pivotová statistika U slouží k řešení úloh o μ_1- μ_2, když σ_1^2 a σ_2^2známe.)

c) Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak $K = \dfrac{(n_1 + n_2 - 2)S_*^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$.

(Pivotová statistika K slouží k řešení úloh o neznámém společném rozptylu σ^2.)

d) Jestliže $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, pak $T = \dfrac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$.

(Pivotová statistika T slouží k řešení úloh o μ_1- μ_2, když σ_1^2 a σ_2^2 neznáme, ale víme, že jsou shodné.)

e) $F = \dfrac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$.

(Pivotová statistika F slouží k řešení úloh o σ_1^2/ σ_2^2.)

Důkaz:

ad a) Nebudeme provádět.

ad b) M_1-M_2 je lineární kombinace náhodných veličin s normálním rozložením, má tedy normální rozložení s parametry $E(M_1-M_2) = \mu_1- \mu_2$, $D(M_1-M_2) = \sigma_1^2/n_1+ \sigma_2^2/n_2$. U se získá standardizací M_1-M_2.

ad c) $K_1 = \dfrac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1-1)$ a $K_2 = \dfrac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2-1)$ jsou stochasticky nezávislé náhodné veličiny, tedy $K = K_1+K_2 \sim \chi^2(n_1 + n_2 - 2)$.

ad d) $U = \dfrac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}}$ ~ N(0, 1), $K = \dfrac{(n_1 + n_2 - 2)S_*^2}{\sigma^2}$ ~ $\chi^2(n_1 + n_2 - 2)$ jsou stochasticky nezávislé, protože $M_1 - M_2$ a S_*^2 jsou stochasticky nezávislé. $T = \dfrac{U}{\sqrt{\dfrac{K}{n_1 + n_2 - 2}}} =$

$$= \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

ad e) $K_1 = \dfrac{(n_1 - 1)S_1^2}{\sigma_1^2}$ ~ $\chi^2(n_1-1)$ a $K_2 = \dfrac{(n_2 - 1)S_2^2}{\sigma_2^2}$ ~ $\chi^2(n_2-1)$ jsou stochasticky nezávislé náhodné veličiny, tedy $F = \dfrac{\frac{K_1}{n_1-1}}{\frac{K_2}{n_2-1}} = \dfrac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$.

Příklad: Nechť jsou dány dva nezávislé náhodné výběry, první pochází z rozložení N(2, 1,5) a má rozsah 10, druhý pochází z rozložení N(3, 4) a má rozsah 5. Jaká je pravděpodobnost, že výběrový průměr 1. výběru bude menší než výběrový průměr 2. výběru?

Řešení: $P(M_1 < M_2) = P(M_1 - M_2 < 0) =$

$$P\left(\frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} < \frac{0 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}\right) = P\left(U < \frac{-2+3}{\sqrt{\dfrac{1,5}{10} + \dfrac{4}{5}}}\right) = P(U < 1,026) = \Phi(1,026) = 0,8475.$$

S pravděpodobností přibližně 84,8% je výběrový průměr 1. výběru menší než výběrový průměr 2. výběru.

7.1.2. Intervaly spolehlivosti pro parametrické funkce μ_1-μ_2, σ_1^2/σ_2^2

Uvedeme přehled vzorců pro meze 100(1-α)% empirických intervalů spolehlivosti pro parametrické funkce $\mu_1 - \mu_2$, σ_1^2 / σ_2^2

a) Interval spolehlivosti pro μ_1-μ_2, když σ_1^2, σ_2^2 známe

Oboustranný: $(d, h) = \left(m_1 - m_2 - \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\, u_{1-\alpha/2},\ m_1 - m_2 + \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\, u_{1-\alpha/2}\right)$

Levostranný: $(d, \infty) = \left(m_1 - m_2 - \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\, u_{1-\alpha},\ \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, m_1 - m_2 + \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}\, u_{1-\alpha}\right)$

b) Interval spolehlivosti pro μ_1-μ_2, když σ_1^2, σ_2^2 neznáme, ale víme, že jsou shodné
Oboustranný:

$$(d, h) = \left(m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{1-\alpha/2}(n_1+n_2-2),\; m_1 - m_2 + s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{1-\alpha/2}(n_1+n_2-2)\right)$$

Levostranný: $(d, \infty) = \left(m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{1-\alpha}(n_1+n_2-2),\; \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty,\; m_1 - m_2 + s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{1-\alpha}(n_1+n_2-2)\right)$

c) Interval spolehlivosti pro společný neznámý rozptyl σ^2

Oboustranný: $(d, h) = \left(\frac{(n_1 + n_2 - 2)s_*^2}{\chi^2_{1-\alpha/2}(n_1 + n_2 - 2)}, \frac{(n_1 + n_2 - 2)s_*^2}{\chi^2_{\alpha/2}(n_1 + n_2 - 2)}\right)$

Levostranný: $(d, \infty) = \left(\frac{(n_1 + n_2 - 2)s_*^2}{\chi^2_{1-\alpha}(n_1 + n_2 - 2)}, \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, \frac{(n_1 + n_2 - 2)s_*^2}{\chi^2_{\alpha}(n_1 + n_2 - 2)}\right)$

d) Interval spolehlivosti pro podíl rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$

Oboustranný: $(d, h) = \left(\frac{s_1^2 / s_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{s_1^2 / s_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}\right)$

Levostranný: $(d, \infty) = \left(\frac{s_1^2 / s_2^2}{F_{1-\alpha}(n_1 - 1, n_2 - 1)}, \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, \frac{s_1^2 / s_2^2}{F_{\alpha}(n_1 - 1, n_2 - 1)}\right)$

Upozornění: Není-li v 7.1.2. (b) splněn předpoklad o shodě rozptylů, lze sestrojit aspoň přibližný 100(1-α)% interval spolehlivosti pro μ_1-μ_2. V tomto případě má statistika T přibližně rozložení t(ν), kde počet stupňů volnosti $\nu = \frac{\left(s_1^2 / n_1 + s_2^2 / n_2\right)^2}{\frac{\left(s_1^2 / n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2 / n_2\right)^2}{n_2 - 1}}$. Není-li ν celé číslo, použijeme v tabulkách kvantilů Studentova rozložení lineární interpolaci.

Příklad: 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: $m_1 = 34{,}48$, $m_2 = 35{,}59$, $s_1^2 = 1{,}7482$, $s_2^2 = 1{,}7121$. Hodnoty zjištěné z odebraných vzorků považujeme za realizace dvou nezávislých náhodných výběrů z rozložení N(μ_1, σ^2) a N(μ_2, σ^2). Sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot μ_1 - μ_2.

Řešení: $s_*^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{24\cdot 1{,}7482 + 9\cdot 1{,}7121}{33} = 1{,}7384$, $t_{0,975}(33) = 2{,}035$

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

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

$-2{,}114$ g/l $< \mu_1 - \mu_2 < -0{,}106$ g/l s pravděpodobností aspoň 0,95.

Příklad: V předešlém příkladě nyní předpokládáme, 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: $d = \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} = \frac{1{,}7482/1{,}7121}{F_{0,975}(24,9)} = \frac{1{,}7482/1{,}7121}{3{,}6142} = 0{,}28$

$$h = \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)} = \frac{1{,}7482/1{,}7121}{F_{0,025}(24,9)} = \frac{1{,}7482/1{,}7121}{1/F_{0,975}(9,24)} = \frac{1{,}7482/1{,}7121}{1/2{,}7027} = 2{,}76$$

$0{,}28 < \frac{\sigma_1^2}{\sigma_2^2} < 2{,}76$ s pravděpodobností aspoň 0,95.

7.1.3. Testování hypotéz o parametrických funkcích $\mu_1-\mu_2$, σ_1^2/σ_2^2

a) Nechť $X_{11},\ldots,X_{n_1 1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{12},\ldots,X_{n_2 2}$ je na něm nezávislý náhodný výběr z rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$, $n_2 \geq 2$ a σ_1^2, σ_2^2 známe. Nechť c je konstanta. Test H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$ se nazývá dvouvýběrový z-test.

b) Nechť $X_{11},\ldots,X_{n_1 1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma^2)$ a $X_{12},\ldots,X_{n_2 2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$ a σ^2 neznáme. Nechť c je konstanta. Test H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$ se nazývá dvouvýběrový t-test.

c) Nechť $X_{11},\ldots,X_{n_1 1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{12},\ldots,X_{n_2 2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$. Test H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$ se nazývá F-test.

7.1.4. Provedení testů o parametrických funkcích μ_1-μ_2, σ_1^2/σ_2^2 pomocí kritického oboru

a) Provedení dvouvýběrového z-testu

Hypotézu 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: $\mu_1 - \mu_2 > c$) zamítáme na hladině významnosti α, jestliže $\left|\dfrac{m_1 - m_2 - c}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}\right| \geq u_{1-\alpha/2}$

(resp. $\dfrac{m_1 - m_2 - c}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \leq -u_{1-\alpha}$ resp. $\dfrac{m_1 - m_2 - c}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \geq u_{1-\alpha}$).

b) Provedení dvouvýběrového t-testu

Hypotézu 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: $\mu_1 - \mu_2 > c$) zamítáme na hladině významnosti α, jestliže $\left|\dfrac{m_1 - m_2 - c}{s_*\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}\right| \geq t_{1-\alpha/2}(n_1 + n_2 - 2)$

(resp. $\dfrac{m_1 - m_2 - c}{s_*\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \leq -t_{1-\alpha}(n_1 + n_2 - 2)$ resp. $\dfrac{m_1 - m_2 - c}{s_*\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \geq t_{1-\alpha}(n_1 + n_2 - 2)$)

.

c) Provedení F-testu

Hypotézu H_0: $\dfrac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$ (resp. H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} < 1$ resp. H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} > 1$) zamítáme na hladině významnosti α, jestliže $\dfrac{s_1^2}{s_2^2} \leq F_{\alpha/2}(n_1+n_2\text{-}2)$ nebo $\dfrac{s_1^2}{s_2^2} \geq F_{1-\alpha/2}(n_1+n_2\text{-}2)$

(resp. $\dfrac{s_1^2}{s_2^2} \leq F_{\alpha}(n_1+n_2\text{-}2)$ resp. $\dfrac{s_1^2}{s_2^2} \geq F_{1-\alpha}(n_1+n_2\text{-}2)$).

Příklad: Výrobce limonád chtěl zjistit, zda změna technologie výroby se projeví v prodeji limonád. Proto sledoval po 14 náhodně vybraných dnů před zavedením nových limonád tržby v určitém regionu a zjistil, že za den utržil v průměru 39 600 Kč se směrodatnou odchylkou 5 060 Kč. Po zavedení nových limonád prověřil stejným způsobem tržby v 11 náhodně vybraných dnech v témž regionu a zjistil průměrný příjem 41 200 Kč se směrodatnou odchylkou 4 310 Kč. Předpokládejte, že tržby za starý typ limonád se řídí rozložením $N(\mu_1, \sigma_1^2)$ a tržby za nový typ limonád se řídí rozložením $N(\mu_2, \sigma_2^2)$.

a) Na hladině významnosti 0,05 testujte hypotézu H_0: $\dfrac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$.

b) Na hladině významnosti 0,05 testujte hypotézu H_0: $\mu_1 - \mu_2 = 0$ proti H_1: $\mu_1 - \mu_2 \neq 0$.

Řešení:
ad a)
$\frac{s_1^2}{s_2^2} = \frac{5060^2}{4310^2} = 1{,}3783$, $F_{\alpha/2}(n_1-1, n_2-1) = F_{0{,}025}(13{,}10) = 0{,}3077$, $F_{1-\alpha/2}(n_1-1, n_2-1) = F_{0{,}975}(13{,}10) =$
$= 3{,}5832$

Protože testové kritérium $\frac{s_1^2}{s_2^2} = 1{,}3783$ se nerealizuje v kritickém oboru
$W = (0; 0{,}3077\rangle \cup \langle 3{,}5832; \infty)$, nelze na hladině významnosti 0,05 zamítnout hypotézu o shodě rozptylů.
ad b) Protože jsme na hladině významnosti 0,05 nezamítli hypotézu o shodě rozptylů, můžeme rozptyly σ_1^2, σ_2^2 považovat za shodné a za jejich odhad vezmeme vážený průměr výběrových rozptylů

$$s_*^2 = \frac{13 \cdot 5060^2 + 10 \cdot 4310^2}{23} = 22548165{,}217 .$$

$$\frac{m_1 - m_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{39600 - 41200}{\sqrt{22548165{,}217}\sqrt{\frac{1}{14} + \frac{1}{11}}} = -0{,}8363,\ t_{1-\alpha/2}(n_1 + n_2 - 2) = t_{0{,}975}(23) = 2{,}0687$$

Protože testové kritérium -0,8363 se nerealizuje v kritickém oboru
$W = (-\infty; -2{,}0687\rangle \cup \langle 2{,}0687; \infty)$, na hladině významnosti 0,05 nelze zamítnout hypotézu o shodě středních hodnot.

7.2. Dva nezávislé náhodné výběry z alternativních rozložení

7.2.1. Rozložení statistiky odvozené z výběrových průměrů

Nechť $X_{11}, \ldots, X_{n_1 1}$ je náhodný výběr z rozložení A(ϑ_1) a $X_{12}, \ldots, X_{n_2 2}$ je na něm nezávislý náhodný výběr rozložení A(ϑ_2) a nechť jsou splněny podmínky $n_1\vartheta_1(1-\vartheta_1) > 9$ a $n_2\vartheta_2(1-\vartheta_2) > 9$. Označme M_1, M_2 výběrové průměry. Pak statistika

$$U = \frac{M_1 - M_2 - (\vartheta_1 - \vartheta_2)}{\sqrt{\frac{\vartheta_1(1-\vartheta_1)}{n_1} + \frac{\vartheta_2(1-\vartheta_2)}{n_2}}}$$

konverguje v distribuci k náhodné veličině se standardizovaným normálním rozložením.
Důkaz: Analogicky jako v 6.3.1.

7.2.2. Asymptotický interval spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$

Pokud rozptyl $D(M_i) = \vartheta_i(1-\vartheta_i)/n_i$ nahradíme odhadem $M_i(1-M_i)/n_i$, i = 1, 2, konvergence náhodné veličiny U k veličině s rozložením N(0,1) se neporuší. Tedy

$$\forall \vartheta_1 - \vartheta_2 \in \Xi : 1-\alpha \le P\left(-u_{1-\alpha/2} < \frac{M_1 - M_2 - (\vartheta_1 - \vartheta_2)}{\sqrt{\frac{M_1(1-M_1)}{n_1} + \frac{M_2(1-M_2)}{n_2}}} < u_{1-\alpha/2} \right) =$$

$$= P(M_1 - M_2 - \sqrt{\frac{M_1(1-M_1)}{n_1} + \frac{M_2(1-M_2)}{n_2}}\, u_{1-\alpha/2} < \vartheta_1 - \vartheta_2 <$$

$$< M_1 - M_2 + \sqrt{\frac{M_1(1-M_1)}{n_1} + \frac{M_2(1-M_2)}{n_2}}\, u_{1-\alpha/2})$$

Meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro $\vartheta_1 - \vartheta_2$ jsou:

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

7.2.3. Testování hypotézy o parametrické funkci $\vartheta_1 - \vartheta_2$

Nechť $X_{11},\ldots,X_{n_1 1}$ je náhodný výběr z rozložení A(ϑ_1) a $X_{12},\ldots,X_{n_2 2}$ je na něm nezávislý náhodný výběr rozložení A(ϑ_2) a nechť jsou splněny podmínky $n_1 \vartheta_1$ (1- ϑ_1) > 9 a $n_2 \vartheta_2$ (1- ϑ_2) > 9. Na asymptotické hladině významnosti α testujeme nulovou hypotézu H_0: $\vartheta_1 - \vartheta_2 = c$ proti alternativě H_1: $\vartheta_1 - \vartheta_2 \neq c$ (resp. H_1: $\vartheta_1 - \vartheta_2 < c$ resp. H_1: $\vartheta_1 - \vartheta_2 > c$). Testovým kritériem je statistika

$$T_0 = \frac{M_1 - M_2 - c}{\sqrt{\frac{M_1(1-M_1)}{n_1} + \frac{M_2(1-M_2)}{n_2}}},$$ která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1). Kritický obor má tvar $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$ (resp. $W = \left(-\infty, -u_{1-\alpha}\right\rangle$ resp. $W = \left\langle u_{1-\alpha}, \infty\right)$).

Testování hypotézy o parametrické funkci $\vartheta_1 - \vartheta_2$ lze samozřejmě provést i pomocí 100(1-α)% asymptotického intervalu spolehlivosti nebo pomocí p-hodnoty.

Upozornění: Je-li c = 0, pak označme $M_* = \frac{n_1 M_1 + n_2 M_2}{n_1 + n_2}$ vážený průměr výběrových rozptylů. Jako testová statistika slouží $T_0 = \frac{M_1 - M_2}{\sqrt{M_*(1-M_*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, která v případě platnosti nulové hypotézy má asymptoticky rozložení N(0,1). Kritický obor má tvar $W = \left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right)$ (resp. $W = \left(-\infty, -u_{1-\alpha}\right\rangle$ resp. $W = \left\langle u_{1-\alpha}, \infty\right)$). Testová statistika T_0 vznikne standardizací statistiky $M_1 - M_2$, kde neznámé parametry ϑ_1, ϑ_2 nahradíme společným odhadem M_*.

Příklad: Management supermarketu chce se spolehlivostí 95% zjistit, zda týden slev 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.

Řešení: Testujeme hypotézu $\vartheta_1 - \vartheta_2 = 0$ proti levostranné alternativě H_1: $\vartheta_1 - \vartheta_2 < 0$ na asymptotické hladině významnosti 0,05.
$n_1 = 200$, $n_2 = 300$, $m_1 = 97/200$, $m_2 = 162/300$, $m_* = (97 + 162)/500 = 0{,}518$, $u_{0,95} = 1{,}645$.
Ověření podmínek $n_1 \vartheta_1 (1-\vartheta_1) > 9$ a $n_2 \vartheta_2 (1-\vartheta_2) > 9$: Parametry ϑ_1 a ϑ_2 neznáme, nahradíme je odhady m_1 a m_2. 97.(1-97/200) = 49,955 > 9, 162.(1-162/300) = 74,52 > 9.
Realizace testového kritéria:
$$t_0 = \frac{m_1 - m_2}{\sqrt{m_*(1-m_*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\frac{97}{200} - \frac{162}{300}}{\sqrt{0{,}518(1-0{,}518)\left(\frac{1}{200}+\frac{1}{300}\right)}} = -1{,}2058.$$ Kritický obor je
$W = \left(-\infty, -u_{1-\alpha}\right\rangle = \left(-\infty, -1{,}645\right\rangle$. Protože testové kritérium nepatří do kritického oboru, nelze na asymptotické hladině významnosti 0,05 zamítnout hypotézu, že slevy nemají vliv na podíl velkých nákupů v daném supermarketu.

Příklady k 7. kapitole

Příklad 1: Bylo vylosováno 11 stejně starých selat téhož plemene. Šesti z nich byla předepsána výkrmná dieta č. 1 a zbylým pěti výkrmná dieta č. 2. Průměrné denní přírůstky v Dg za dobu půl roku jsou následující:
dieta č. 1: 62, 54, 55, 60, 53, 58
dieta č. 2: 52, 56, 49, 50, 51.
Zjištěné hodnoty považujeme za realizace dvou nezávislých náhodných výběrů pocházejících z rozložení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$. Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů a 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

($0{,}1872\ Dg^2 < \frac{\sigma_1^2}{\sigma_2^2} < 12{,}9541\ Dg^2$ s pravděpodobností aspoň 0,95.
$0{,}99\ Dg < \mu_1 - \mu_2 < 9{,}81\ Dg$ s pravděpodobností aspoň 0,95.)

Příklad 2.: Pro údaje z příkladu 1. testujte na hladině významnosti 0,05 hypotézu, že obě výkrmné diety mají stejný vliv na hmotnostní přírůstky selat.
(1. způsob – pomocí intervalu spolehlivosti. 95% empirický interval spolehlivosti pro $\mu_1-\mu_2$ je interval (0,99; 9,81). Neobsahuje nulu, proto H_0 zamítáme na hladině významnosti 0,05.
2. způsob – pomocí kritického oboru. Protože testové kritérium se realizuje hodnotou 2,771, která patří do kritického oboru $\left(-\infty; -2{,}2622\right\rangle \cup \left\langle 2{,}2622; \infty\right)$, H_0 zamítáme na hladině významnosti 0,05.)

Příklad 3.: Při výstupní kontrole bylo náhodně vybráno 150 výrobků vyrobených na ranní směně a rovněž 150 výrobků vyrobených na odpolední směně. U ranní směny bylo zjištěno 16 zmetků a u odpolední 12 zmetků. Sestrojte 95% asymptotického interval spolehlivosti pro rozdíl podílu zmetků v obou směnách.
($-0{,}039 < \vartheta_1 - \vartheta_2 < 0{,}092$ s pravděpodobností přibližně 0,95.)

Příklad 4.: Pro údaje z příkladu 3. testujte na asymptotické hladině významnosti 0,05 hypotézu, že podíl zmetků v obou směnách je týž.
(1. způsob – pomocí intervalu spolehlivosti. 95% empirický asymptotický interval spolehlivosti pro $\vartheta_1 - \vartheta_2$ je interval (-0,039; 0,092). Obsahuje nulu, proto H_0 nezamítáme na asymptotické hladině významnosti 0,05.
2. způsob: pomocí kritického oboru. Protože testové kritérium se realizuje hodnotou 0,794, která nepatří do kritického oboru $\left(-\infty, -1{,}96\right\rangle \cup \left\langle 1{,}96, \infty\right)$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.)

Práce se systémem STATISTICA

Téma: Parametrické úlohy o dvou nezávislých náhodných výběrech

Do programu STATISTICA načtěte ASCII soubor studentky.dat, který obsahuje údaje o 48 náhodně vybraných studentkách VŠE v Praze. 1. sloupec – výška, 2. sloupec – známka z matematiky v 1. semestru, 3. sloupec – obor studia (1 – národní hospodářství, 2 – informatika). Tyto tři proměnné nazvěte X,Y,Z a vytvořte jim návěští.

Úkoly:

1. Sestrojte 95% interval spolehlivosti pro střední hodnotu výšky
   a) studentek oboru nh ($167,3 < \mu < 172,3$)
   b) studentek oboru inf ($164,8 < \mu < 169,0$).
   (Pro výběr případů splňujících danou podmínku použijte postup z úkolu č. 8 cvičení č. 1.)
   Návod: Meze 95% intervalu spolehlivosti pro střední hodnotu proměnné X zjistíme pomocí Descriptive Statistics, kde zaškrtneme Conf. limits for mean.

2. Sestrojte 95% interval spolehlivosti pro podíl rozptylů výšek studentek oboru nh a inf.
   ($0,821 < \sigma_1^2/\sigma_2^2 < 4,513$)
   Návod: K datovému souboru přidáme další dvě proměnné DM a HM pro výpočet dolní a horní meze intervalu spolehlivosti. Do LongName těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro podíl rozptylů (viz 7.1.2. (d)). Výběrové rozptyly pro 1. a 2. výběr zjistíme pomocí Descriptive Statistics.

3. Na hladině významnosti 0,05 testujte hypotézu, že rozptyly výšek studentek oboru nh a inf jsou shodné.
   Návod: lze využít výsledku 2. úkolu. 95% interval spolehlivosti pro podíl rozptylů obsahuje číslo 1, tedy hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.

4. Sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot výšek studentek oboru nh a inf.
   ($-0,452 < \mu_1 - \mu_2 < 6,292$)
   Návod: K datovému souboru přidáme další dvě proměnné DM a HM pro výpočet dolní a horní meze intervalu spolehlivosti. Do LongName těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro rozdíl středních hodnot (viz 7.1.2. (b)). Výběrové průměry a výběrové rozptyly pro 1. a 2. výběr zjistíme pomocí Descriptive Statistics.

5. Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty výšek studentek oboru nh a inf jsou shodné.
   Návod:
   1. způsob: lze využít výsledku 4. úkolu. 95% interval spolehlivosti pro rozdíl středních hodnot obsahuje číslo 0, tedy hypotézu o shodě středních hodnot nezamítáme na hladině významnosti 0,05.
   2. způsob: úloha vede na dvouvýběrový t-test. Statistics – Basic Statistics – t-test, independent, by groups – OK, Variables – Dependent X, grouping Z – OK – Summary. Ve výstupní tabulce najdeme hodnotu testového kritéria a p-hodnotu. Protože p-hodnota = 0,087837 je větší než hladina významnosti 0,05, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Před prováděním těchto úkolů pomocí diagnostických grafů orientačně ověřte normalitu dat.

6. Sestrojte 95% asymptotický interval spolehlivosti pro podíl studentek, které mají z matematiky trojku, a to
   a) pro studentky oboru nh ($0{,}423 < \vartheta < 0{,}791$),
   b) pro studentky oboru inf ($0{,}094 < \vartheta < 0{,}506$).
   Návod: K datovému souboru přidáme další dvě proměnné DM a HM pro výpočet dolní a horní meze intervalu spolehlivosti. Do LongName těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro parametr ϑ alternativního rozložení (viz 6.3.2.). Výběrové průměry pro 1. a 2. výběr zjistíme pomocí Descriptive Statistics.

7. Sestrojte 95% asymptotický interval spolehlivosti pro rozdíl podílu studentek, které mají z matematiky trojku, a to pro studentky oboru nh a inf ($0{,}038 < \vartheta_1 - \vartheta_2 < 0{,}584$).
   Návod: K datovému souboru přidáme další dvě proměnné DM a HM pro výpočet dolní a horní meze intervalu spolehlivosti. Do LongName těchto proměnných zapíšeme vzorce pro dolní a horní mez intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$ (viz 7.2.2.). Výběrové průměry pro 1. a 2. výběr zjistíme pomocí Descriptive Statistics.

8. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že podíl studentek, které mají z matematiky trojku, je stejný pro studentky oboru nh a inf.
   Návod: lze využít výsledku 7. úkolu. 95% asymptotický interval spolehlivosti pro rozdíl parametrů $\vartheta_1 - \vartheta_2$ neobsahuje číslo 0, tedy hypotézu o shodě parametrů ϑ_1, ϑ_2 zamítáme na asymptotické hladině významnosti 0,05.