Parametrické úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

Motivace:

Máme-li k dispozici dva nezávislé náhodné výběry z normálních rozložení, je naším úkolem porovnat střední hodnoty či rozptyly těchto rozložení. Zpravidla konstruujeme intervaly spolehlivosti pro rozdíl středních hodnot respektive hodnotíme shodu středních hodnot pomocí dvouvýběrového t-testu či dvouvýběrového z-testu a shodu rozptylů pomocí F-testu.

Rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů normálních rozložení

Předpokládáme, že

$X_{11},\ldots,X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$,

$X_{21},\ldots,X_{2n_2}$ je náhodný výběr z rozložení $N(\mu_2, \sigma_2^2)$,

přičemž $n_1 \geq 2$ a $n_2 \geq 2$ a oba výběry jsou stochasticky nezávislé.

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.

Pak platí:

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

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

(Pivotová statistika U slouží k řešení úloh o $\mu_1 - \mu_2$, když σ_1^2 a σ_2^2 známe.)

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 $\mu_1 - \mu_2$, 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.)

Vysvětlení:
ad a) Neuvádíme, viz např. J. Anděl: Matematická statistika.
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) $U = \dfrac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}} \sim N(0, 1)$, $K = \dfrac{(n_1 + n_2 - 2)S_*^{\,2}}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$ jsou stochasticky nezávislé, protože

$M_1 - M_2$ a $S_*^{\,2}$ jsou stochasticky nezávislé. $T = \dfrac{U}{\sqrt{\dfrac{K}{n_1 + n_2 - 2}}} = \dfrac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$.

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

$$F = \frac{\dfrac{K_1}{n_1 - 1}}{\dfrac{K_2}{n_2 - 1}} = \frac{S_1^{\,2} / S_2^{\,2}}{\sigma_1^{\,2} / \sigma_2^{\,2}} \sim F(n_1 - 1, n_2 - 1).$$

Příklad: Nechť jsou dány dva nezávislé náhodné výběry, první pochází z rozložení N(0,28; 0,09) a má rozsah 16, druhý pochází z rozložení N(0,25; 0,04) a má rozsah 25. Jaká je pravděpodobnost, že výběrový průměr 1. výběru bude větší než výběrový průměr 2. výběru?

Řešení:

$$P(M_1 > M_2) = P(M_1 - M_2 > 0) = 1 - P(M_1 - M_2 \leq 0) = 1 - P\left(\frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq \frac{0 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) =$$

$$= 1 - P\left(U \leq \frac{-0,28 + 0,25}{\sqrt{\frac{0,09}{16} + \frac{0,04}{25}}} \right) = 1 - P(U \leq -0,35294) = 1 - \Phi(-0,35) = \Phi(0,35) = 0,63683$$

S pravděpodobností přibližně 63,7% je výběrový průměr 1. výběru větší než výběrový průměr 2. výběru.

Výpočet pomocí systému STATISTICA:

Statistika $M_1 - M_2$ se podle bodu (a) řídí rozložením $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$,

kde $\mu_1 - \mu_2 = 0{,}28 - 0{,}25 = 0{,}03$, $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{0{,}09}{16} + \frac{0{,}04}{25} = 0{,}007225$, tj. statistika

$M_1 - M_2 \sim N(0{,}03;0{,}007225)$.

Otevřeme nový datový soubor o jedné proměnné a jednom případu. Do Dlouhého jména této proměnné napíšeme

= 1-INormal(0;0,03;sqrt(0,007225)). V proměnné Prom1 se objeví hodnota 0,637934:

	1 Prom1
1	0,637934

Intervaly spolehlivosti pro parametrické funkce $\mu_1 - \mu_2$, 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 (využití pivotové statistiky U)

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

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

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

b) Interval spolehlivosti pro μ_1-μ_2, když σ_1^2, σ_2^2 neznáme, ale víme, že jsou shodné (využití pivotové statistiky T)

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 (využití pivotové statistiky K)

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

Pravostranný: (0, h) = $\left(0, \frac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha}(n_1+n_2-2)} \right)$

d) Interval spolehlivosti pro podíl rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$ (využití pivotové statistiky F)

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

Pravostranný: (0, h) = $\left(0, \frac{s_1^2/s_2^2}{F_{\alpha}(n_1-1, n_2-1)} \right)$

Upozornění: Není-li v bodě (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 $t(\nu)$, 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 $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$. Sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Řešení:

Úloha vede na vzorec (b) s využitím statistiky T. Vypočteme vážený průměr výběrových rozptylů a najdeme odpovídající kvantily Studentova rozložení:

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

Dosadíme do vzorců pro dolní a horní mez intervalu spolehlivosti:

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

$$= 34{,}48 - 35{,}59 - \sqrt{1{,}7384}\cdot\sqrt{\frac{1}{25}+\frac{1}{10}}\cdot 2{,}035 = -2{,}114$$

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

$$= 34{,}48 - 35{,}59 + \sqrt{1{,}7384}\cdot\sqrt{\frac{1}{25}+\frac{1}{10}}\cdot 2{,}035 = -0{,}106$$

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

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných d a h a jednom případu.

Do Dlouhého jména proměnné d napíšeme

=34,48-35,59-sqrt((24*1,7482+9*1,7121)/33)*sqrt((1/25)+(1/10))*VStudent(0,975;33)

Do Dlouhého jména proměnné h napíšeme

=34,48-35,59+

sqrt((24*1,7482+9*1,7121)/33)*sqrt((1/25)+(1/10))*VStudent(0,975;33)

	1 d	2 h
1	-2,11368	-0,10632

S pravděpodobností aspoň 0,95 tedy -2,114 g/l < μ_1 - μ_2 < -0,106 g/l.

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:

Úloha vede na vzorec (d) s využitím statistiky F.

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

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

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

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných d a h a jednom případu.

Do Dlouhého jména proměnné d napíšeme

=(1,7482/1,7121)/VF(0,975;24;9)

(Funkce VF(x;ný;omega) počítá x-kvantil Fisherova – Snedecorova rozložení F(ný, omega).)

Do Dlouhého jména proměnné h napíšeme

=(1,7482/1,7121)/VF(0,025;24;9)

	1 d	2 h
1	0,282521	2,759698

S pravděpodobností aspoň 0,95 tedy platí: $0{,}28 < \sigma_1^2 / \sigma_2^2 < 2{,}76$.

Jednotlivé typy testů o parametrických funkcích $\mu_1 - \mu_2$, σ_1^2/σ_2^2

a) Nechť $X_{11},\ldots,X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21},\ldots,X_{2n_2}$ 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 $X_{11},\ldots,X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma^2)$ a $X_{21},\ldots,X_{2n_2}$ je na něm nezávislý náhodný výběr z 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 $X_{11},\ldots,X_{1n_1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{21},\ldots,X_{2n_2}$ 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.

Provedení testů o parametrických funkcích $\mu_1 - \mu_2$, σ_1^2/σ_2^2 pomocí kritického oboru

a) Provedení dvouvýběrového z-testu

Vypočteme realizaci t_0 testového kritéria $T_0 = \dfrac{(M_1 - M_2) - c}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$. Kritický obor má tvar: $W = (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$.

Levostranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 < c$. Kritický obor má tvar: $W = (-\infty, -u_{1-\alpha}\rangle$.

Pravostranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 > c$. Kritický obor má tvar: $W = \langle u_{1-\alpha}, \infty)$.

b) Provedení dvouvýběrového t-testu

Vypočteme realizaci t_0 testového kritéria $T_0 = \dfrac{(M_1 - M_2) - c}{S_* \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 \neq c$. Kritický obor má tvar:
$W = (-\infty, -t_{1-\alpha/2}(n_1 + n_2 - 2)\rangle \cup \langle t_{1-\alpha/2}(n_1 + n_2 - 2), \infty)$.

Levostranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 < c$. Kritický obor má tvar: $W = (-\infty, -t_{1-\alpha}(n_1 + n_2 - 2)\rangle$.

Pravostranný test: Testujeme H_0: $\mu_1 - \mu_2 = c$ proti H_1: $\mu_1 - \mu_2 > c$. Kritický obor má tvar: $W = \langle t_{1-\alpha}(n_1 + n_2 - 2), \infty)$.

c) Provedení F-testu

Vypočteme realizaci testového kritéria $t_0 = \frac{s_1^2}{s_2^2}$. Stanovíme kritický obor W. Pokud $t_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.

Oboustranný test: Testujeme H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$. Kritický obor má tvar:

$$W = \left(0, F_{\alpha/2}(n_1 - 1, n_2 - 1)\right\rangle \cup \left\langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \infty\right).$$

Levostranný test: Testujeme H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\frac{\sigma_1^2}{\sigma_2^2} < 1$. Kritický obor má tvar: $W = \left(0, F_{\alpha}(n_1 - 1, n_2 - 1)\right\rangle$.

Pravostranný test: Testujeme H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\frac{\sigma_1^2}{\sigma_2^2} > 1$. Kritický obor má tvar: $W = \left\langle F_{1-\alpha}(n_1 - 1, n_2 - 1), \infty\right)$.

Příklad: V restauraci "U bílého koníčka" měřili ve 20 případech čas obsluhy zákazníka. Výsledky 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:

Na hladině významnosti 0,05 testujeme nulovou hypotézu H_0: $\mu_1 - \mu_2 = 0$ proti oboustranné alternativě H_1: $\mu_1 - \mu_2 \neq 0$. Je to úloha na dvouvýběrový t-test. Před provedením tohoto testu je však nutné pomocí F-testu ověřit shodu rozptylů. Na hladině významnosti 0,05 tedy testujeme H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ proti H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$. Nejprve vypočteme $m_1 = 8{,}25$, $m_2 = 8{,}13$, $s_1^2 = 6{,}307$, $s_2^2 = 9{,}41$, $s_*^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{19 \cdot 6{,}307 + 14 \cdot 9{,}41}{33} = 7{,}623$. Podle vzorce (c) vypočteme realizaci testové statistiky:

$t_0 = \frac{s_1^2}{s_2^2} = \frac{6{,}307}{9{,}41} = 0{,}6702$. Stanovíme kritický obor:

$$W = \langle 0, F_{\alpha/2}(n_1-1, n_2-1)\rangle \cup \langle F_{1-\alpha/2}(n_1-1, n_2-1), \infty) = \langle 0, F_{0,025}(19,14)\rangle \cup \langle F_{0,975}(19,14), \infty) =$$
$$= \langle 0, 1/F_{0,975}(14,19)\rangle \cup \langle F_{0,975}(19,14), \infty) = \langle 0, 1/2{,}649\rangle \cup \langle 2{,}8607, \infty) = \langle 0; 0{,}3778\rangle \cup \langle 2{,}8607, \infty)$$

Protože se testová statistika nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05. Rozptyly tedy můžeme považovat za shodné.

Nyní se vrátíme k dvouvýběrovému t-testu. Podle vzorce (b) vypočteme realizaci testové statistiky:

$t_0 = \frac{m_1 - m_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{8{,}25 - 8{,}13}{\sqrt{7{,}623}\sqrt{\frac{1}{20} + \frac{1}{15}}} = 0{,}124$. Stanovíme kritický obor:

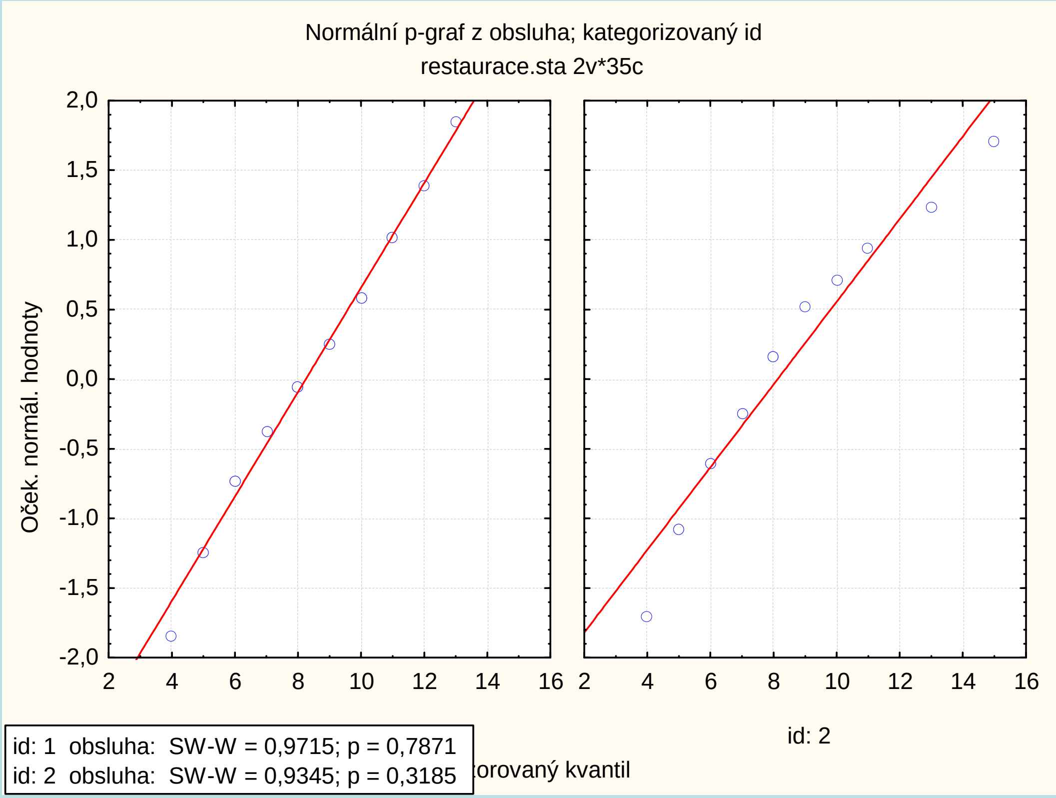
$$W = (-\infty, -t_{1-\alpha/2}(n_1+n_2-2)\rangle \cup \langle t_{1-\alpha/2}(n_1+n_2-2), \infty) = (-\infty, -t_{0,975}(33)\rangle \cup \langle t_{0,975}(33), \infty) = (-\infty, -2{,}035\rangle \cup \langle 2{,}035, \infty)$$

Protože testová statistika se nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných a 35 případech. První proměnnou nazveme OBSLUHA, druhou ID. Do proměnné OBSLUHA napíšeme nejprve doby obsluhy v první restauraci a poté doby obsluhy ve druhé restauraci. Do proměnné ID, která slouží k rozlišení první a druhé restaurace, napíšeme 20 krát jedničku a 15 krát dvojku.

Pomocí NP-grafu ověříme normalitu dat v obou skupinách. Grafy – 2D Grafy – Normální pravděpodobnostní grafy – zaškrtneme S-W test - Proměnné OBSLUHA, OK, Kategorizovaný – Kategorie X, zaškrtneme Zapnuto, Změnit proměnnou – ID, OK. Dostaneme graf



V obou případech se tečky odchylují od přímky jenom málo a p-hodnoty S-W testu převyšují 0,05. Předpoklad o normálním rozložení dat v obou skupinách je oprávněný.

Nyní provedeme dvouvýběrový t-test současně s testem o shodě rozptylů:

Statistika – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK, Proměnné –Závislé proměnné OBSLUHA, Grupovací proměnná ID – OK.

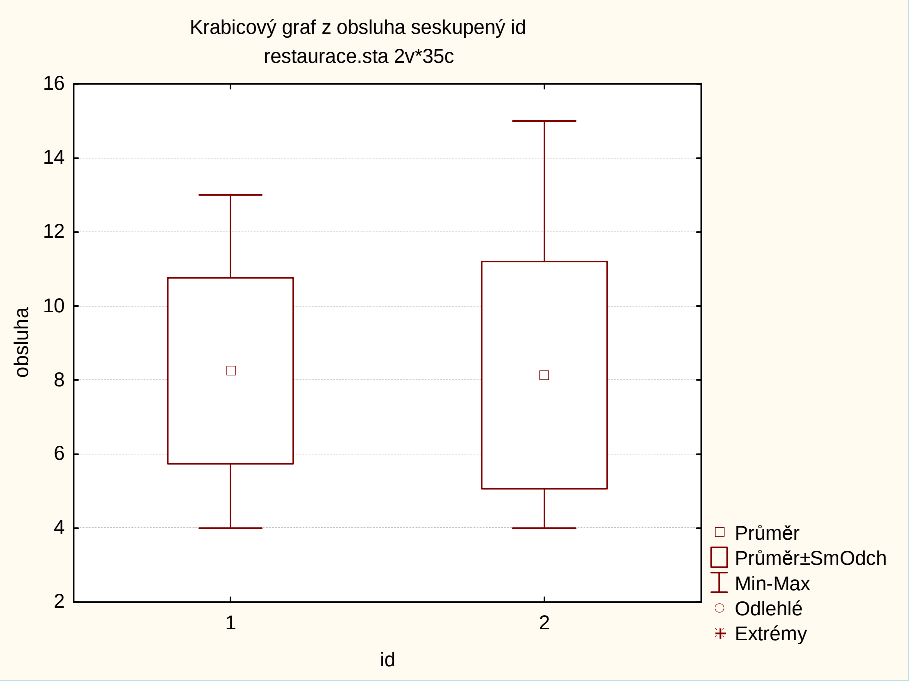
Po kliknutí na tlačítko Souhrn dostaneme tabulku

	t-testy; grupováno: ID (restaurace) Skup. 1: 1 Skup. 2: 2										
Proměnná	Průměr 1	Průměr 2	t	sv	p	Poč.plat 1	Poč.plat. 2	Sm.odch. 1	Sm.odch. 2	F-poměr rozptyly	p rozptyly
OBSLUHA	8,250000	8,133333	0,123730	33	0,902279	20	15	2,510504	3,067495	1,492952	0,410440

Vidíme, že testová statistika pro test shody rozptylů se realizuje hodnotou 1,492952 (je to převrácená hodnota k číslu 0,6702, které jsme vypočítali při ručním postupu), odpovídající p-hodnota je 0,41044, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě rozptylů. (Upozornění: v případě zamítnutí hypotézy o shodě rozptylů je zapotřebí v tabulce t-testu pro nezávislé vzorky dle skupin zaškrtnout volbu Test se samostatnými odhady rozptylu.)

Dále z tabulky plyne, že testová statistika pro test shody středních hodnot se realizuje hodnotou 0,12373, počet stupňů volnosti je 33, odpovídající p-hodnota 0,902279, tedy hypotézu o shodě středních hodnot nezamítáme na hladině významnosti 0,05. Znamená to, že se neprokázal rozdíl ve středních hodnotách dob obsluhy v restauracích "U bílého koníčka" a „Zlatý lev“.

Tabulku ještě doplníme krabicovými diagramy. Na záložce Detaily zaškrtneme krabicový graf a vybereme volbu Průměr/SmOdch/Min-Max.

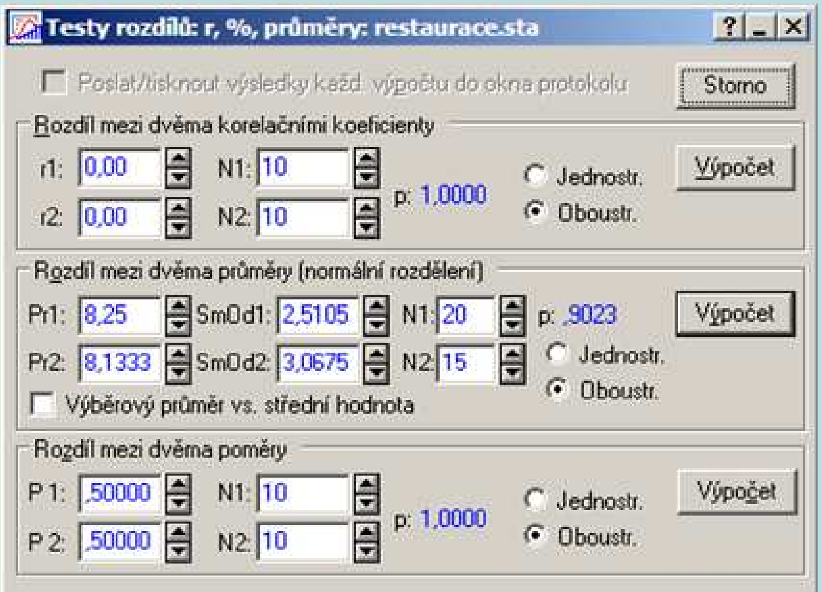


Z grafu je vidět, že průměrná doba obsluhy v první restauraci je nepatrně delší a má menší variabilitu než ve druhé restauraci. Extrémní ani odlehlé hodnoty se zde nevyskytují.

Upozornění:

V případě, že známe realizace obou výběrových průměrů a směrodatných odchylek, můžeme pro provedení dvouvýběrového t-testu v systému STATISTICA použít aplikaci Tesy rozdílů. Postup si ukážeme na příkaldě s dobou obsluhy ve dvou restauracích

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma průměry (normální rozdělení) – do políčka Pr1 napíšeme 8,25, do políčka SmOd1 napíšeme 2,5105, do políčka N1 napíšeme 20, do políčka Pr1 napíšeme 8,25, do políčka SmOd1 napíšeme 3,0675, do políčka N1 napíšeme 15 – Výpočet. Dostaneme p-hodnotu 0,9023, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.



Parametrické úlohy o dvou nezávislých náhodných výběrech z alternativních rozložení

Motivace:

Provádíme opakovaně nezávisle n_1-krát jeden náhodný pokus a nezávisle na tom n_2-krát druhý náhodný pokus. V první sérii pokusů sledujeme nějaký jev, který v každém pokusu může nastat s pravděpodobností ϑ_1 a ve druhé sérii pokusů sledujeme nějaký jiný jev, jehož pravděpodobnost nastoupení je ϑ_2. Parametry ϑ_1, ϑ_2 neznáme. Naším úkolem bude konstruovat interval spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$ nebo testovat hypotézu o této parametrické funkci, a to pomocí dvou nezávislých náhodných výběrů z alternativních rozložení $A(\vartheta_1)$, $A(\vartheta_2)$.

Asymptotické rozložení statistiky odvozené ze dvou výběrových průměrů alternativních rozložení

Nechť $X_{11},\dots,X_{1n_1}$ je náhodný výběr z alternativního rozložení A(ϑ_1) a $X_{21},\dots,X_{2n_2}$ je na něm nezávislý náhodný výběr alternativního 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 $U = \dfrac{M_1 - M_2 - (\vartheta_1 - \vartheta_2)}{\sqrt{\dfrac{\vartheta_1(1-\vartheta_1)}{n_1} + \dfrac{\vartheta_2(1-\vartheta_2)}{n_2}}} \approx N(0,1)$.

Vysvětlení: Analogicky jako v případě jednoho náhodného výběru z alternativního rozložení.

Vzorec pro meze 100(1-α)% asymptotického empirického intervalu spolehlivosti pro parametrickou funkci $\vartheta_1 - \vartheta_2$.

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

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

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

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

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

Příklad: 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_{1i}, která bude nabývat hodnoty 1, když v týdnu bez slevy i-tý náhodně vybraný zákazník uskuteční větší nákup a hodnoty 0 jinak, i = 1, …, 200. Náhodné veličiny $X_{1,1}, \dots, X_{1,200}$ 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, …, 300. Náhodné veličiny $X_{2,1}, \dots, X_{2,300}$ tvoří náhodný výběr z rozložení $A(\vartheta_2)$.

$n_1 = 200$, $n_2 = 300$, $m_1 = 97/200 = 0{,}485$, $m_2 = 162/300 = 0{,}54$.

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, tedy

$97.(1-97/200) = 49{,}955 > 9$, $162.(1-162/300) = 74{,}52 > 9$.

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

$$d = m_1 - m_2 - \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\,u_{1-\alpha/2} = \frac{97}{200} - \frac{162}{300} - \sqrt{\frac{\frac{97}{200}(1-\frac{97}{200})}{200} + \frac{\frac{162}{300}(1-\frac{162}{300})}{300}}\,1{,}96 = -0{,}1443$$

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\,u_{1-\alpha/2} = \frac{97}{200} - \frac{162}{300} + \sqrt{\frac{\frac{97}{200}(1-\frac{97}{200})}{200} + \frac{\frac{162}{300}(1-\frac{162}{300})}{300}}\,1{,}96 = 0{,}0343$$

Zjistili jsme tedy, že s pravděpodobností přibližně 0,95: $-0{,}1443 < \vartheta_1 - \vartheta_2 < 0{,}0343$.

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

Nechť $X_{11},\dots,X_{1n_1}$ je náhodný výběr z alternativního rozložení A(ϑ_1) a $X_{21},\dots,X_{2n_2}$ je na něm nezávislý náhodný výběr alternativního rozložení A(ϑ_2) a nechť jsou splněny podmínky $n_1\vartheta_1(1-\vartheta_1) > 9$ a $n_2\vartheta_2(1-\vartheta_2) > 9$. 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 H_0 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 provést též pomocí 100(1-α)% asymptotického intervalu spolehlivosti nebo pomocí p-hodnoty.)

Poznámka: Postup při testování hypotézy $\vartheta_1 - \vartheta_2 = 0$

Je-li c = 0, pak označme $M_* = \dfrac{n_1 M_1 + n_2 M_2}{n_1 + n_2}$ vážený průměr výběrových průměrů. Jako testová statistika slouží $T_0 = \dfrac{M_1 - M_2}{\sqrt{M_*(1-M_*)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, 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: Pro údaje z příkladu o slevách v supermarketu 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.

$n_1 = 200$, $n_2 = 300$, $m_1 = 97/200$, $m_2 = 162/300$, $m_* = (97 + 162)/500 = 0{,}518$.

Podmínky dobré aproximace byly ověřeny v předešlém příkladu.

Testování pomocí intervalu spolehlivosti:

Pro levostrannou alternativu používáme pravostranný interval spolehlivosti:

$$h = m_1 - m_2 + \sqrt{\frac{m_1(1-m_1)}{n_1} + \frac{m_2(1-m_2)}{n_2}}\,u_{1-\alpha} = \frac{97}{200} - \frac{162}{300} + \sqrt{\frac{\frac{97}{200}(1-\frac{97}{200})}{200} + \frac{\frac{162}{300}(1-\frac{162}{300})}{300}}\,1{,}645 = 0{,}02$$

Protože číslo c = 0 je obsaženo v intervalu $(-\infty; 0{,}02\rangle$, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Testování pomocí kritického oboru:

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 $W = (-\infty, -u_{1-\alpha}\rangle = (-\infty, -u_{0,95}\rangle = (-\infty, -1{,}645\rangle$. Protože testové kritérium nepatří do kritického oboru, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Testování pomocí p-hodnoty:

Pro levostrannou alternativu se p-hodnota počítá podle vzorce $p = P(T_0 \le t_0)$:

$$p = P(T_0 \le -1{,}2058) = \Phi(-1{,}2058) = 1 - \Phi(1{,}2058) = 1 - 0{,}8861 = 0{,}1139$$

Protože p-hodnota je větší než 0,05, H_0 nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,485, do políčka N1 napíšeme 200, do políčka P 2 napíšeme 0,54, do políčka N2 napíšeme 300 – zaškrtneme Jednostr. - Výpočet. Dostaneme p-hodnotu 0,1142, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.