Parametrické úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

Motivace: V této situaci je naším úkolem porovnat střední hodnoty či rozptyly dvou normálních rozložení na základě znalosti dvou nezávislých náhodných výběrů pořízených z 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

Máme dva nezávislé náhodné výběry, první pochází z rozložení $N(\mu_1, \sigma_1^2)$ a má rozsah $n_1 \geq 2$, druhý pochází z rozložení $N(\mu_2, \sigma_2^2)$ a má rozsah $n_2 \geq 2$. 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.

Pak platí:

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

b) $U = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$. (Pivotová statistika U slouží k řešení úloh o μ_1- μ_2, když σ_1^2 a σ_2^2 známe.)

c) Nechť $\sigma_1^2 = \sigma_2^2 =: \sigma^2$, 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 neznámém rozptylu σ^2.)

d) 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 μ_1- μ_2, když σ_1^2 a σ_2^2 neznáme, ale víme, že jsou shodné.)

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

Vysvětlení:

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

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

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

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

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:

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

$$= 1 - P\left(U \le \frac{-(0{,}28 - 0{,}25)}{\sqrt{\frac{0{,}09}{16} + \frac{0{,}04}{25}}}\right) = 1 - P(U \le -0{,}3529) = 1 - \Phi(-0{,}35) = \Phi(0{,}35) = 0{,}63683$$

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 $\mu_1 - \mu_2 = 0{,}28 - 0{,}25 = 0{,}03$, $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{0{,}09}{16} + \frac{0{,}04}{25} = 0{,}0072$, 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.

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

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

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: $(d, \infty) = \left(\frac{S_1^2/S_2^2}{F_{1-\alpha}(n_1-1, n_2-1)}, \infty \right)$

Pravostranný: $(-\infty, h) = \left(0, \frac{S_1^2/S_2^2}{F_{\alpha}(n_1-1, n_2-1)} \right)$

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 $\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$. 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 z bodu (b). Vypočteme vážený průměr výběrových rozptylů a najdeme odpovídající kvantily Studentova rozložení:

$$s_*^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{24 \cdot 1{,}7482 + 9 \cdot 1{,}7121}{33} = 1{,}738, \quad t_{0,975}(33) = 2{,}035$$

Dosadíme do vzorců pro dolní a horní mez intervalu spolehlivosti:

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

$$= 34{,}48 - 35{,}59 - \sqrt{1{,}738} \cdot \sqrt{\frac{1}{25} + \frac{1}{10}} \cdot 2{,}035 = -2{,}114$$

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

$$= 34{,}48 - 35{,}59 + \sqrt{1{,}738} \cdot \sqrt{\frac{1}{25} + \frac{1}{10}} \cdot 2{,}035 = -0{,}106$$

$-2{,}114$ 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	-2,11	-0,10

S pravděpodobností aspoň 0,95 tedy -2,114 g/l < $\mu_1 - \mu_2$ < -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 z bodu (d).

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

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

$0,28 < \frac{\sigma_1^2}{\sigma_2^2} < 2,76$ 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	0,282	2,759

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 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 testového kritéria $t_0 = \frac{m_1 - m_2 - c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

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 testového kritéria $t_0 = \dfrac{m_1 - m_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

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:
$W = \left(-\infty, -t_{1-\alpha/2}(n_1+n_2-2)\right] \cup \left[t_{1-\alpha/2}(n_1+n_2-2), \infty\right)$.

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

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

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 z bodu (c) vypočteme realizaci testové statistiky:

$$t_0 = \frac{s_1^2}{s_2^2} = \frac{6{,}307}{9{,}41} = 0{,}670.$$

Stanovíme kritický obor:

$$W = \langle 0, F_{\alpha/2}(n_1-1, n_2-1)\rangle \cup \langle F_{1-\alpha/2}(n_1-1, n_2-1), \infty) = \langle 0, F_{0,025}(19,14)\rangle \cup \langle F_{0,975}(19,14), \infty)$$
$$= \langle 0, 1/F_{0,975}(14,19)\rangle \cup \langle F_{0,975}(19,14), \infty) = \langle 0, 1/2{,}6469\rangle \cup \langle 2{,}8607, \infty) = \langle 0; 0{,}3778\rangle \cup \langle 2{,}8607, \infty)$$

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 z bodu (b) vypočteme realizaci testové statistiky:

$$t_0 = \frac{m_1 - m_2}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{8{,}25 - 8{,}13}{\sqrt{7{,}623}\sqrt{\frac{1}{20} + \frac{1}{15}}} = 0{,}127.$$

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.

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.

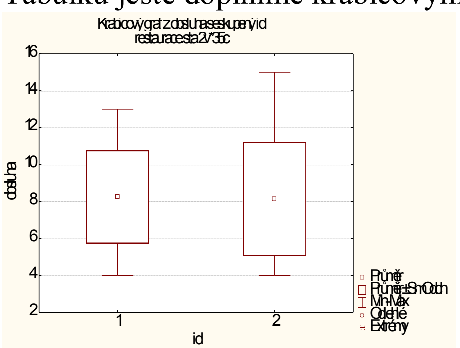
Po kliknutí na tlačítko Souhrn dostaneme tabulku

	t-testy; grupováno: ID (restaurace) Skup. 1: 1 Skup. 2: 2										
Proměr	Průmi 1	Průmi 2	t	sv	p	Poč.p 1	Poč.pl 2	Sm.od 1	Sm.od 2	F-pom rozpty	p rozpty
OBSLU	8,250	8,133	0,123	3:	0,902	2(	1:	2,510	3,067	1,492	0,410

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 s rizikem omylu nejvýše 5% 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.

Cohenův koeficient věcného účinku – doplnění významu dvouvýběrového t-testu:

Nechť $X_{11}, \ldots, X_{1n_1}$ je náhodný výběr z rozložení N(μ_1, σ^2) a $X_{21}, \ldots, X_{2n_2}$ je na něm nezávislý náhodný výběr rozložení N(μ_2, σ^2), přičemž $n_1 \geq 2$ a $n_2 \geq 2$ a σ^2 neznáme. Nechť c je konstanta.
Testujeme H_0: $\mu_1 - \mu_2 = 0$ proti H_1: $\mu_1 - \mu_2 \neq 0$. Označme m_1, m_2 realizace výběrových průměrů hodnot dané veličiny v těchto dvou skupinách, s_1^2, s_2^2 realizace výběrových rozptylů a $s_*^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ realizaci váženého průměru výběrových rozptylů.

Cohenův koeficient d vypočteme podle vzorce: $d = \frac{|m_1 - m_2|}{s_*}$.

Tento koeficient slouží k posouzení velikosti rozdílu průměrů, který je standardizován pomocí odmocniny z váženého průměru výběrových rozptylů. Jedná se o tzv. věcnou významnost neboli velikost účinku skupiny na variabilitu hodnot sledované náhodné veličiny. Velikost účinku hodnotíme podle následující tabulky:

Hodnota d	účinek
aspoň 0,8	velký
mezi 0,5 až 0,8	střední
mezi 0,2 až 0,5	malý
pod 0,2	zanedbatelný

(Uvedené hodnoty nemají samozřejmě absolutní platnost, posouzení, jaký účinek považujeme za velký či malý, závisí na kontextu.)

Je zapotřebí si uvědomit, že při dostatečně velkých rozsazích náhodných výběrů i malý rozdíl ve výběrových průměrech způsobí zamítnutí nulové hypotézy na hladině významnosti α, i když z věcného hlediska tak malý rozdíl nemá význam. Naopak, máme-li výběry malých rozsahů, pak i značně velký rozdíl ve výběrových průměrech nemusí vést k zamítnutí nulové hypotézy na hladině významnosti α.

Příklad:

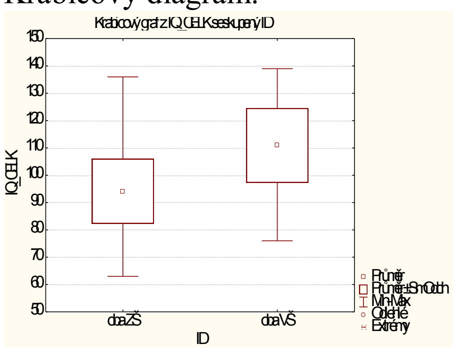
Máme k dispozici údaje o celkovém IQ 856 žáků ZŠ. Zajímáme se jednak o skupinu dětí, jejichž oba rodiče mají pouze základní vzdělání (je jich 296) a jednak o skupinu dětí, jejichž oba rodiče mají vysokoškolské vzdělání (těch je 75). Na hladině významnosti 0,05 budeme testovat hypotézu, že střední hodnota celkového IQ je v obou skupinách stejná a také vypočteme Cohenův koeficient věcného účinku.

Řešení: Provedeme dvouvýběrový t-test:

	t-testy; grupováno:ZŠ a VŠ (IQ) Skup. 1: oba ZŠ Skup. 2: oba VŠ										
Proměı	Průmı oba Z	Průmı oba V	t	sv	p	Poč.p oba Z	Poč.pl oba V	Sm.od oba Z	Sm.od oba V	F-pom Rozpt	p Rozpt
IQ_CE	94,13	110,9	-10,6	36	0,000	29	7	11,82	13,60	1,322	0,110

Hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05, protože odpovídající p-hodnota je velmi blízká 0 (hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05, p-hodnota F-testu je 0,110124, což je větší než 0,05).

Krabicový diagram:



Vidíme, že průměrné celkové IQ dětí v 1. skupině je 94,1, zatímco ve 2. skupině 110,9. Vliv skupiny na variabilitu hodnot celkového IQ posoudíme pomocí Cohenova koeficientu.

	1 n1	2 n2	3 m1	4 m2	5 s1	6 s2	7 d
1	29	7	94,13	110,9	11,82	13,60	1,374

Cohenův koeficient nabývá hodnoty 1,37, tudíž vliv skupiny na variabilitu hodnot celkového IQ lze považovat za velký.

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 $N(\mu_1, \sigma^2)$ a tržby za nový typ limonád se řídí rozložením $N(\mu_2, \sigma^2)$.
Na hladině významnosti 0,05 testujte hypotézu H_0: $\mu_1 - \mu_2 = 0$ proti oboustranné alternativě H_1: $\mu_1 - \mu_2 \neq 0$.

Řešení:
Za odhad společného neznámého rozptylu vezmeme vážený průměr výběrových rozptylů:

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

Realizace testového kritéria:

$$t_0 = \frac{m_1 - m_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{39600 - 41200}{\sqrt{22548165{,}2}\sqrt{\frac{1}{14} + \frac{1}{11}}} = -0{,}836$$

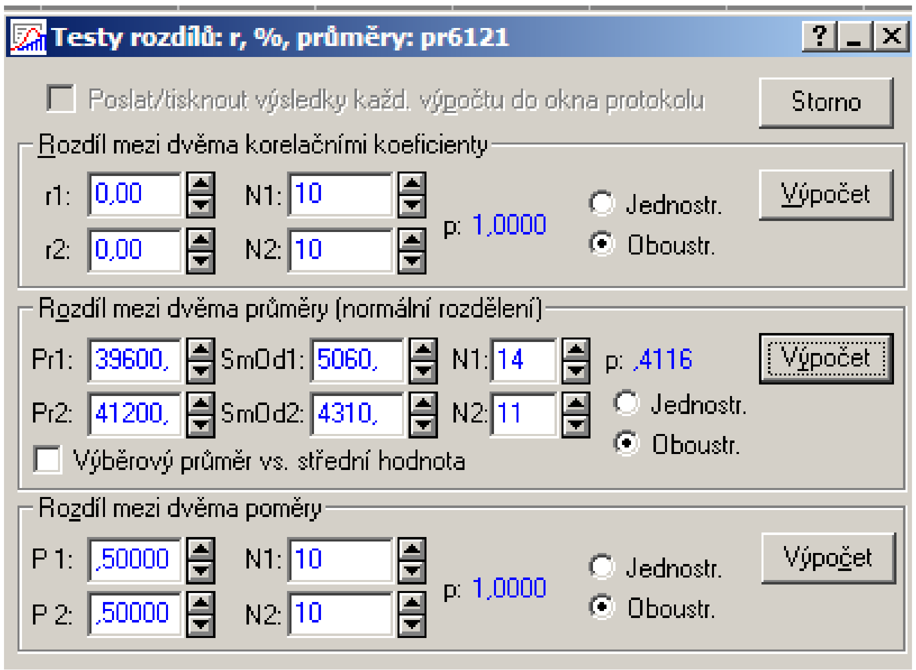
Kritický obor:

$$W = \left(-\infty, -t_{1-\alpha/2}(n_1+n_2-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n_1+n_2-2), \infty\right) = \left(-\infty, -t_{0,975}(23)\right\rangle \cup \left\langle t_{0,975}(23), \infty\right) = \left(-\infty, -2{,}0687\right\rangle \cup \left\langle 2{,}0687, \infty\right)$$

Protože testové kritérium se nerealizuje v kritickém oboru, na hladině významnosti 0,05 nelze zamítnout hypotézu o shodě středních hodnot.

Výpočet pomocí systému STATISTICA:

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma průměry (normální rozdělení) – do políčka Pr1 napíšeme 39600, do políčka SmOd1 napíšeme 5600, do políčka N1 napíšeme 14, do políčka Pr2 napíšeme 41200, do políčka SmOd1 napíšeme 4310, do políčka N1 napíšeme 14 - Výpočet. Dostaneme p-hodnotu 0,4116 tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.



Jelikož p-hodnota je větší než hladina významnosti 0,05, H_0 nezamítáme na hladině významnosti 0,05. Znamená to, že změna technologie výroby se neprojevila ve střední hodnotě tržeb.

Ověřování normality

Grafický způsob

a) Normální pravděpodobnostní graf (NP-plot)

NP-plot umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.
Způsob konstrukce: na vodorovnou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ a na svislou osu kvantily u_{α}, kde $\alpha = \frac{3j-1}{3n+1}$, přičemž j je pořadí j-té uspořádané hodnoty (jsou-li některé hodnoty stejné, pak za j bereme průměrné pořadí odpovídající takové skupince). Pocházejí-li data z normálního rozložení, pak všechny dvojice $(x_{(j)}, u_{\alpha})$ budou ležet na přímce.

b) Kvantil-kvantilový graf (Q-Q plot)

Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení (např. systém STATISTICA nabízí 8 typů rozložení: beta, exponenciální, Gumbelovo, gamma, log-normální, normální, Rayleighovo a Weibulovo). Pro nás je nejdůležitější právě normální rozložení.
Způsob konstrukce: na svislou osu vynášíme uspořádané hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ a na vodorovnou osu kvantily $K_{\alpha}(X)$ vybraného rozložení, kde $\alpha = \frac{j - r_{adj}}{n + n_{adj}}$, přičemž r_{adj} a n_{adj} jsou korigující faktory $\leq 0{,}5$, implicitně $r_{adj} = 0{,}375$ a $n_{adj} = 0{,}25$.
(Jsou-li některé hodnoty $x_{(1)} \leq \ldots \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.) Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadnou z dat nebo je může zadat uživatel. Body $(K_{\alpha}(X), x_{(j)})$ se metodou nejmenších čtverců proloží přímka. Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

c) Probability - probability plot (P-P plot)

Používá se ke stejným účelům jako Q-Q plot, ale jinak se konstruuje.

Způsob konstrukce: spočtou se standardizované hodnoty $z_{(j)} = \frac{x_{(j)} -}{s}$, j = 1, ..., n. Na vodorovnou osu se vynesou hodnoty teoretické distribuční funkce $\Phi(z_{(j)})$ a na svislou osu hodnoty empirické distribuční funkce $F(z_{(j)}) = j/n$. (Jsou-li některé hodnoty $x_{(1)} \leq ... \leq x_{(n)}$ stejné, pak za j bereme průměrné pořadí odpovídající takové skupince.)Pokud se body $(\Phi(z_{(j)}), F(z_{(j)}))$ řadí kolem hlavní diagonály čtverce [0,1] x [0,1], lze usuzovat na dobrou shodu empirického a teoretického rozložení.

d) Histogram

Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. (Ve STATISTICE je pojem histogramu širší, skrývá se za ním i sloupkový diagram.)

Způsob konstrukce ve STATISTICE: na vodorovnou osu se vynášejí třídicí intervaly (implicitně 10, jejich počet lze změnit, stejně tak i meze třídicích intervalů) či varianty znaku a na svislou osu absolutní nebo relativní četnosti třídicích intervalů či variant. Do histogramu se zakreslí tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení. Kromě 8 typů rozložení uvedených u Q-Q plotu umožňuje STATISTICA použít ještě další 4 rozložení: Laplaceovo, logistické, geometrické, Poissonovo.

Testy normality

a) Kolmogorovův – Smirnovův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z normálního rozložení s parametry μ a σ^2. Distribuční funkci tohoto rozložení označme $\Phi_T(x)$. Nechť $F_n(x)$ je výběrová distribuční funkce. Testovou statistikou je statistika $D_n = \sup \left| F_n(x) - \Phi_T(x) \right|$. Nulovou hypotézu zamítáme na hladině významnosti α, když $D_n \geq D_n(\alpha)$, kde $D_n(\alpha)$ je tabelovaná kritická hodnota. Pro $n \geq 30$ lze $D_n(\alpha)$ aproximovat výrazem $\sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$.

V případě, že neznáme parametry μ a σ^2 normálního rozložení, musíme je odhadnout z dat (střední hodnotu odhadneme pomocí m a rozptyl pomocí s^2). Tím se změní rozložení testové statistiky D_n. Příslušné modifikované kvantily byly určeny pomocí simulačních studií. V této situaci používáme Lilieforsovu variantu Kolmogorovova – Smirnovova testu.

b) Shapirův – Wilkův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení $N(\mu, \sigma^2)$.
Testová statistika má tvar:

$$W = \frac{\left(\sum_{i=1}^{m} a_i^{(n)} \left(X_{(n-i+1)} - X_{(i)} \right) \right)^2}{\sum_{i=1}^{m} \left(X_i - M \right)^2}$$

,kde $m = n/2$ pro n sudé a $m = (n-1)/2$ pro n liché. Koeficienty $a_i^{(n)}$ jsou tabelovány.

Na testovou statistiku W lze pohlížet jako na korelační koeficient mezi uspořádanými pozorováními a jim odpovídajícími kvantily standardizovaného normálního rozložení. V případě, že data vykazují perfektní shodu s normálním rozložením, bude mít W hodnotu 1. Hypotézu o normalitě tedy zamítneme na hladině významnosti α, když se na této hladině neprokáže korelace mezi daty a jim odpovídajícími kvantily rozložení $N(0,1)$.
Lze také říci, že S – W test je založen na zjištění, zda body v Q-Q grafu jsou významně odlišné od regresní přímky proložené těmito body.
(S-W test se používá především pro výběry menších rozsahů, $n < 50$, ale v systému STATISTICA je implementováno jeho rozšíření i na výběry velkých rozsahů, kolem 2000.)

c) Test dobré shody pro normální rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, \ldots, X_n$ pochází z normálního rozložení s distribuční funkcí $\Phi(x)$.

- Data rozdělíme do r třídicích intervalů $\langle u_j, u_{j+1})$, j = 1, ..., r.
- Zjistíme absolutní četnost n_j j-tého třídicího intervalu.
- Vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.
- Vypočteme testovou statistiku: $K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}$. Platí-li nulová hypotéza, pak $K \approx \chi^2(r-1-k)$, kde k je počet odhadovaných parametrů normálního rozložení. (Obvykle z dat odhadujeme střední hodnotu i rozptyl, tedy k = 2.)
- Stanovíme kritický obor $W = \langle \chi^2_{1-\alpha}(r-1-k), \infty)$.
- Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \in W$.

(Aproximace se považuje za vyhovující, když $np_j \geq 5$, j = 1, ..., r.)

Upozornění: Hodnota testové statistiky K je silně závislá na volbě třídicích intervalů. Navíc při nesplnění podmínky $np_j \geq 5$, j = 1, ..., r je třeba některé intervaly slučovat, což vede ke ztrátě informace.