1 8-Parametrické úlohy o dvou nezávislých náhodných výběrech z normálního rozložení

Vyjádření k domácímu úkolu

- interpretace koeficientu korelace: $r = 0.8 \rightarrow$ vysoký stupeň PŘÍMÉ LINEÁRNÍ závislosti
- u testování hypotéz nezapomínejte stanovit H_0, H_1, a psát celkový závěr příkladu!!!
- testování pomocí IS: $H_0 : \mu = c \rightarrow H_0$ nezamítáme, pokud $c \in IS$.

Stará látka

- přehled testů - rozdat
- Testování hypotéz
  - datový soubor reálných dat
  - máme předpoklady o datovém souboru: o charakteristikách, o rozložení náh. veličiny, o nezávislosti dvou náh.veličin
  - zatím jsme ověřovali *předpoklady o charakteristikách* jednovýběrových dat (μ, σ) z normálního rozložení
  - postup testování hypotéz
    - Formulace problému
    - stanovení H_0
    - stanovení H_1
      - oboustranná
      - pravostranná
      - levostranná
    - volba hl.významnosti $\alpha =$ pst, že H_0 zamítáme, i když platí.
    - provedení měření
    - testování hypotéz
      - Kritický obor
      - IS
      - p-hodnota
    - rozhodnutí o zamítnutí H_0
    - INTERPRETACE VÝSLEDKŮ
  - chyba 1.druhu: H_0 platí a přesto ji zamítneme
  - chyba 2.druhu: H_0 neplatí a přesto ji nezamítneme
- kritický obor
  - stanovíme T_0
  - stanovíme kritický obor W - tvar podle typu alternativy

- H_0 zamítáme, pokud $T_0 \in W$.

- IS:
    - z H_0 známe konstantu c
    - stanovíme IS - tvar podle alternativy: A/IS: O/O, L/P, P/L
    - H_0 zamítáme, pokud $c \notin IS$

- p-hodnota
    - stanovíme T_0
    - stanovíme p-hodnotu podle typu alternativy
    - H_0 zamítáme, pokud $p < \alpha$.

Nová látka

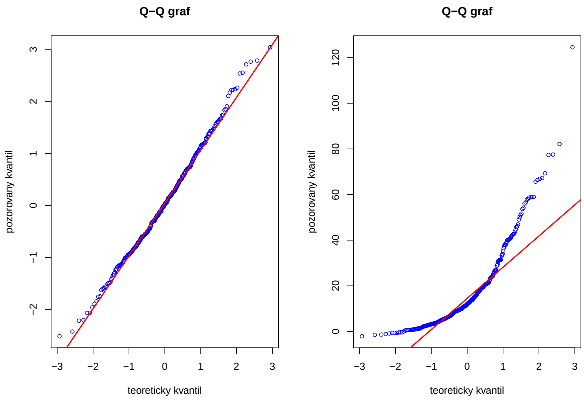
Párové testy:

- data z dvourozměrného normálního rozložení
- porovnání rozdílů párových součástí objektu, párových orgánů člověka
- porovnání délky uší, výšky/šířky očí, nadočnicového oblouku, sjetost pneumatik, zkoumání podobných rysů dvojčat atp.
- Nechť $(X_1, Y_1) \dots (X_n, Y_n)$ je náh. výběr z dvourozměrného normálního rozložení, přičemž $n \geq 2$. Střední hodnota znaku X je μ_1, střední hodnota znaku Y je μ_2.
- $H_0 : \mu_1 = \mu_2 : \mu_1 - \mu_2 = 0$
- $H_1 : \mu_1 \neq \mu_2 : \mu_1 - \mu_2 \neq 0$
- utvoříme rozdíly $Z_1 = X_1 - Y_1 \dots Z_n = X_n - Y_n$.
- $Z_1, \dots Z_n$ je datový soubor z normálního rozložení $\rightarrow$ získáváme jednovýběrový datový soubor
- aplikujeme jednovýběrový test o střední hodnotě μ, když σ^2 neznáme.
- Příklad z domácího úkolu (*viz příloha sken1)*

Testování normality

- normalita dat je velmi důležitá vlastnosti datového souboru
- pro mnoho parametrických testů je předpoklad normality důležitým základem
    - jednovýběrové testy
    - párové testy
    - dvouvýběrové testy, ...
- testování normality datového souboru

– Testujeme nulovu hypotézu: H_0: Datový soubor/data pochází z normálního rozložení
– alternativní hypotéza: H_0 : Data nepochází z normálního rozložení
– testování provádíme
  1. graficky
     (a) histogram + křivka hustoty teoretického normálního rozložení s odhadem střední hodnoty $\mu = mean(data)$ a odhadem rozptylu $\sigma^2 = (sd(data))^2$ + křivka hustoty odhadnutá z dat density(vektor_dat).
     (b) krabicový graf - více než normalitu testuje vyšikmenost dat boxplot()
     (c) Q-Q graf příkazy qqnorm(vektor_dat) a qqline(vektor_dat)



Obrázek 1: a)vlevo-body leží na přímce $\rightarrow$ data jsou z normálního rozložení; b)-vpravo - body neleží na přímce $\rightarrow$ data nejsou z normálního rozložení

  2. početně - testováním
     (a) Shairo-Wilkův test
        * je vhodný pro testování souborů o menších rozsahů ($n \leq 30$)
        * shapiro.test() knihovna stat
     (b) Kolmogorův-Smirnovův test
     (c) Lillie-Forsův test
        * modifikace K-S testu
        * lillie.test() knihovna nortest
     (d) Anderson-Darlingův test
        * ad.test() knihovna nortest
     (e) Pearsonův χ^2 test
        * pearson.test knihovna nortest
– vždy je vhodné provést alespoň dva testy normality

- ODTEĎ PŘED KAŽDÝM TESTOVÁNÍM POMOCÍ PARAMETRICKÝCH TESTŮ MUSÍME OVĚŘIT NORMALITU DAT

Testy o dvou nezávislých náhodných výběrech

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

$$S_*^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

je vážený průměr výběrových rozptylů.

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

2. Pokud $\sigma_1^2 = \sigma_2^2 = \sigma^2$, pak pivotová statistika
$$K = \frac{(n_1 + n_2 - 2)S_*^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$
slouží k řešení úloh o neznámém společném roztylu σ^2.

3. Pokud $\sigma_1^2 = \sigma_2^2 = \sigma^2$, pak pivotová statistika
$$T = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$
slouží k řešení úloh o $\mu_1 - \mu_2$, když σ_1^2, σ_2^2 neznáme, ale víme, že jsou shodné.

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

Intervaly spolehlivosti

1. IS pro $\mu_1 - \mu_2$, když σ_1^2,σ_2^2 známe (využití statistiky U)
   (a) Oboustranný IS (d, h):
$$\left(m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha/2} \, ; \, m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{\alpha/2} \right)$$
   (b) Levostranný IS $(d; \infty)$:
$$\left(m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha} ; \infty \right)$$
   (c) Pravostranný IS $(-\infty; h)$:
$$\left(-\infty ; \, m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{\alpha} \right)$$

2. IS pro $\mu_1 - \mu_2$, když σ_1^2,σ_2^2 neznáme, ale víme, že jsou shodné (využití statistiky T)
   (a) Oboustranný IS (d, h):
$$\left(m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2) \, ; m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{\alpha/2}(n_1 + n_2 - 2) \right)$$
   (b) Levostranný IS $(d; \infty)$:
$$\left(m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha}(n_1 + n_2 - 2) \, ; \, \infty \right)$$
   (c) Pravostranný IS $(-\infty; h)$:
$$\left(-\infty \, ; \, m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{\alpha}(n_1 + n_2 - 2) \right)$$

3. IS pro společný neznámý rozptyl σ^2 (využití pivotové statistiky K)
   (a) Oboustranný IS (d, h):
$$\left(\frac{(n_1 + n_2 - 2)s_*^2}{\chi^2_{1-\alpha/2}(n_1 + n_2 - 2)} \, ; \, \frac{(n_1 + n_2 - 2)s_*^2}{\chi^2_{\alpha/2}(n_1 + n_2 - 2)} \right)$$
   (b) Levostranný IS (d, ∞)
$$\left(\frac{(n_1 + n_2 - 2)s_*^2}{\chi^2_{1-\alpha}(n_1 + n_2 - 2)} \, ; \, \infty \right)$$
   (c) Pravostranný IS $(-\infty; h)$
$$\left(-\infty \, ; \, \frac{(n_1 + n_2 - 2)s_*^2}{\chi^2_{\alpha}(n_1 + n_2 - 2)} \right)$$

4. IS pro podíl rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$ (využití pivotové statistiky F)

(a) Oboustranný IS (d, h):

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

(b) Levostranný IS $(d; \infty)$

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

(c) Pravostranný IS $(-\infty; h)$

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

Dvouvýběrové testy - Kritické obory

1. Nechť $X_{11}, \dots X_{1n_1}$ je náhodný výběr z $N(\mu_1, \sigma_1^2)$, a $X_{21}, \dots X_{2n_2}$ 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.

   - Testujeme $H_0 : \mu_1 - \mu_2 = c$ oproti $H_{11} : \mu_1 - \mu_2 \neq c$, případně $H_{12} : \mu_1 - \mu_2 < c$, či $H_{13} : \mu_1 - \mu_2 > c$.
   - Takovýto test se nazývá *dvouvýběrový z-test*
   - Realizace testové statistiky:
   $$t_0 = \frac{(m_1 - m_2) - c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$
   - kritický obor pro oboustrannou alternativu H_{11}: $W = (-\infty; u_{\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty)$
   - kritický obor pro levostrannou alternativu H_{12}: $W = (-\infty; u_\alpha\rangle$
   - kritický obor pro pravostrannou alternativu H_{13}: $W = \langle u_{1-\alpha}; \infty)$

2. Nechť $X_{11}, \dots X_{1n_1}$ je náhodný výběr z $N(\mu_1, \sigma^2)$, a $X_{21}, \dots X_{2n_2}$ je na něm nezávislý náhodný výběr z rozložení $N(\mu_2, \sigma^2)$, přičemž $n_1 \geq 2$, $n_2 \geq 2$ a σ^2 neznáme. Nechť c je konstanta.

   - Testujeme $H_0 : \mu_1 - \mu_2 = c$ oproti $H_{11} : \mu_1 - \mu_2 \neq c$, případně $H_{12} : \mu_1 - \mu_2 < c$, či $H_{13} : \mu_1 - \mu_2 > c$.
   - Takovýto test se nazývá *dvouvýběrový t-test*
   - Realizace testové statistiky:
   $$t_0 = \frac{(m_1 - m_2) - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$
   - kritický obor pro oboustrannou alternativu H_{11}: $W = (-\infty; t_{\alpha/2}(n_1+n_2-2)\rangle \cup \langle t_{1-\alpha/2}(n_1+n_2-2), \infty)$
   - kritický obor pro levostrannou alternativu H_{12}: $W = (-\infty; t_\alpha(n_1 + n_2 - 2)\rangle$
   - kritický obor pro pravostrannou alternativu H_{13}: $W = \langle t_{1-\alpha}(n_1 + n_2 - 2); \infty)$

3. Nechť $X_{11}, \dots X_{1n_1}$ je náhodný výběr z $N(\mu_1, \sigma^2)$, a $X_{21}, \dots X_{2n_2}$ je na něm nezávislý náhodný výběr z rozložení $N(\mu_2, \sigma^2)$, přičemž $n_1 \geq 2$, $n_2 \geq 2$.

   - Testujeme $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ oproti $H_{11} : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$.
   - Takovýto test se nazývá *F-test.*
   - Realizace testové statistiky:
   $$t_0 = \frac{s_1^2}{s_2^2}$$
   - kritický obor pro oboustrannou alternativu H_{11}: $W = (0; F_{\alpha/2}(n_1-1, n_2-1)\rangle \cup \langle F_{1-\alpha/2}(n_1-1, n_2-1), \infty)$

- kritický obor pro levostrannou alternativu H_{12}: $W = (0; F_{\alpha}(n_1 - 1, n_2 - 1)\rangle$
- kritický obor pro pravostrannou alternativu H_{13}: $W = \langle F_{1-\alpha}(n_1 - 1, n_2 - 1); \infty)$