Statistika: popisná analýza, intervalové odhady

KMDG, FS, VŠB-TUO

Výběrové charakteristiky

průměr

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ aritmetický

$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$ geometrický

$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$ harmonický

$100p\%$ kvantil - hodnota x_p dělící setříděný soubor na 2 části, poměr dělení je $p : (1-p)$, $p \in \langle 0, 1 \rangle$

$x_{0.25}$ dolní kvartil

$x_{0.5} = \tilde{x}$ medián, prostřední hodnota

$x_{0.75}$ horní kvartil

modus

$\hat{x}$ nejčastěji se vyskytující hodnota proměnné

variační rozpětí

$R = x_{max} - x_{min}$

interkvartilové rozpětí

$IQR = x_{0.75} - x_{0.25}$

výběrový rozptyl

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

výběrová směrodatná odchylka

$s = \sqrt{s^2}$

variační koeficient

$V = \frac{s}{\bar{x}}$ vyjadřuje se v procentech $(\cdot 100\%)$

výběrová střední chyba průměru

$s_X = \frac{s}{\sqrt{n}}$

výběrová šikmost

$$a = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

výběrová špičatost

$$e = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Identifikace odlehlých pozorování

z-souřadnice

$z_i = \frac{x_i - x}{s}$ $|z_i| > 3 \Rightarrow x_i$ je odlehlá

mediánová souřadnice

$m_i = \frac{x_i - x_{0.5}}{1.483\, MAD}$ $|m_i| > 3 \Rightarrow x_i$ je odlehlá

$MAD \ldots$ medián absolutních odchylek od mediánu

vnitřní hradby

$[(x_i < x_{0.25} - 1.5\, IQR) \vee (x_i > x_{0.75} + 1.5\, IQR)] \Rightarrow x_i$ je odlehlá

vnější hradby

$[(x_i < x_{0.25} - 3\, IQR) \vee (x_i > x_{0.75} + 3\, IQR)] \Rightarrow x_i$ je extrémní

Intervalové odhady

střední hodnota

$$P\left(\bar{x} - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \le \mu \le \bar{x} + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$ normalita nebo $n > 30$, známe σ^2

Excel: $u_{1-\frac{\alpha}{2}} =$ NORM.S.INV$(1 - \frac{\alpha}{2})$

$$P\left(\bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2};n-1} \le \mu \le \bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2};n-1}\right) = 1 - \alpha$$ normalita, neznáme σ^2

Excel: $t_{1-\frac{\alpha}{2};n-1} =$ T.INV$(1 - \frac{\alpha}{2}; n-1)$

rozptyl

$$P\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1 - \alpha$$ normalita

Excel: $\chi^2_{1-\frac{\alpha}{2};n-1} =$ CHISQ.INV$(1 - \frac{\alpha}{2}; n-1)$

parametr binomického rozdělení

$$P\left(p - u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le \pi \le p + u_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$ $n > \frac{9}{p(1-p)}$

Excel: $u_{1-\frac{\alpha}{2}} =$ NORM.S.INV$(1 - \frac{\alpha}{2})$

Statistika: testování hypotéz

KMDG, FS, VŠB-TUO

Jednovýběrové testy

z-test $\mu = \mu_0$

testovací kritérium $T(X) : \dfrac{\bar{x} - \mu_0}{\sigma}\sqrt{n} \sim N(0,1)$ normalita nebo $n > 30$

t-test $\mu = \mu_0$

testovací kritérium $T(X) : \dfrac{\bar{x} - \mu_0}{s}\sqrt{n} \sim t_{n-1}$ normalita

test o rozptylu $\sigma^2 = \sigma_0^2$

testovací kritérium $T(X) : \dfrac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$ normalita

test o parametru binomického rozdělení $\pi = \pi_0$

testovací kritérium $T(X) : \dfrac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n} \sim N(0,1)$ normalita

znaménkový test $x_{0.5} = c$

testovací kritérium $T(X)$: počet kladných odchylek od c $n > 10$
nulové rozdělení: $Bi(n, \pi = 0.5)$

Wilcoxonův test $x_{0.5} = c$

testovací kritérium $T(X) : \min(R^+, R^-)$

$R^+ = \sum\limits_{x_i - c \geq 0} R_i, \quad R^- = \sum\limits_{x_i - c < 0} R_i$, kde R_i je pořadí seřazených odchylek od c

kritická hodnota: $w_n(\alpha)$ je tabelována

Dvouvýběrové testy – dva závislé výběry

párový t-test $\mu_1 = \mu_2$

platí: $n_1 = n_2 = n$

testovací kritérium $T(X) : \dfrac{\bar{d}\sqrt{n-1}}{s_d} \sim t(n-1)$, kde

$\bar{d}$ je průměr hodnot $d_i = x_{1i} - x_{2i}$
s_d je směrodatná odchylka hodnot d_i

párový Wilcoxonův test $x_{0.5} = y_{0.5}$

platí: $n_x = n_y = n$

určíme $z_i = x_i - y_i$
převedeme na jednovýběrový Wilcoxonův test s $H_0 : z_{0.5} = 0$

Dvouvýběrové testy – dva nezávislé výběry

F-test $\sigma_1^2 = \sigma_2^2$

testovací kritérium $T(X) : \dfrac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$, $s_1^2 > s_2^2$

nepárový t-test $\mu_1 = \mu_2$

1) platí $\sigma_1^2 = \sigma_2^2$

testovací kritérium $T(X) : \dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\,\frac{n_1+n_2}{n_1 \cdot n_2}}} \sim t_{n_1+n_2-2}$

2) neplatí $\sigma_1^2 = \sigma_2^2$

testovací kritérium $T(X) : \dfrac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$, kde $\nu = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$

Mannův-Whitneyho test $x_{0.5}^1 = x_{0.5}^2$

pozorování dohromady seřadíme, T_i značí součet pořadí hodnot i-tého souboru

testovací kritérium $T(X) : \min(U_1, U_2)$, kde

$U_1 = n_1 n_2 + \dfrac{n_1(n_1+1)}{2} - T_1$, $U_2 = n_1 n_2 + \dfrac{n_2(n_2+1)}{2} - T_2$

kritické hodnoty tabelovány

ANOVA

nulová hypotéza $\mu_1 = \mu_2 = \cdots = \mu_k$

předpoklad: normalita a homoskedasticita ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$)

testovací kritérium $T(X) : \dfrac{M_G}{M_R} \sim F_{k-1,n-k}$, kde $M_G = \dfrac{S_G}{df_G}$ a $M_R = \dfrac{S_R}{df_R}$

součet čtverců, stupně volnosti

$S_G = \sum\limits_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$, $df_G = k - 1$ mezi skupinami

$S_R = \sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$, $df_R = n - k$ reziduální

$S_T = \sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}(x_{ij} - \bar{x})^2$, $df_T = n - 1$ celkový