Porovnání 2 výběrů

máme $X_1, \dots, X_n$ výběr z $N(\mu_1, \sigma_1^2)$

$Y_1, \dots, Y_m$ výběr z $N(\mu_2, \sigma_2^2)$

možné hypotézy: $H: \mu_1 = \mu_2$ proti $A: \mu_1 > \mu_2$ ($\mu_1 < \mu_2$, $\mu_1 \neq \mu_2$)

$H: \sigma_1^2 = \sigma_2^2$ proti $A: \sigma_1^2 > \sigma_2^2$ ($\sigma_1^2 < \sigma_2^2$, $\sigma_1^2 \neq \sigma_2^2$)

používá se hlavně:

párový t-test pro $H: \mu_1 = \mu_2$ v případě, že $\sigma_1^2 \neq \sigma_2^2$ nebo výběry nejsou na sobě nezávislé

dvouvýběrový t-test pro $H: \mu_1 = \mu_2$ v případě, že výběry jsou nezávislé a $\sigma_1^2 = \sigma_2^2$

případně jeho modifikace pro nestejné rozptyly

předpoklad nezávislosti musí platit vždy

F-test pro $H: \sigma_1^2 = \sigma_2^2$ za předpokladu nezávislosti výběrů

Párový t-test

vytvoří se rozdíly $Z_1 = X_1 - Y_1, \dots, Z_n = X_n - Y_n$, kde $n = \min(n, m)$

$Z_1, \dots, Z_n$ jsou nezávislé, mají rozdělení $N(\mu_1 - \mu_2, \sigma^2)$

(σ^2 nás nezajímá)

označíme $\Delta = \mu_1 - \mu_2$

test $H: \mu_1 = \mu_2$ proti $A: \mu_1 > \mu_2$ ($\mu_1 < \mu_2$, $\mu_1 \neq \mu_2$)

je totožný s testem (jednovýběrovým)

$H: \Delta = 0$ proti $A: \Delta > 0$ ($\Delta < 0$, $\Delta \neq 0$)

Dvouvýběrový t-test při stejných rozptylech

je založený na veličině $\bar{X}-\bar{Y} \sim N\left(\mu_1-\mu_2,\ \frac{\sigma^2}{n}+\frac{\sigma^2}{m}\right)$

platí $\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$

z toho
$$\frac{\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma^2/n+\sigma^2/m}}}{\sqrt{\dfrac{\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2}}{n+m-2}}} \sim t_{n+m-2}$$

to se upraví
$$\frac{\dfrac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(n+m)/nm}}}{\sqrt{\dfrac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(n-1)S_X^2+(m-1)S_Y^2}}\cdot\sqrt{\frac{nm(n+m-2)}{n+m}}$$

to se porovná s kritickou hodnotou t_{n+m-2}-rozdělení na hl. α

Test shodnosti rozptylů

je založen na podílu $\frac{S_X^2}{S_Y^2}$ – v případě shodnosti rozptylů by to měla být přibližně 1

platí $\frac{S_X^2}{S_Y^2} \sim F_{n-1,\,m-1}$

$H: \sigma_1^2=\sigma_2^2$ proti $A: \sigma_1^2 \neq \sigma_2^2$ se zamítá, když
$$\frac{S_X^2}{S_Y^2} \le k_1 \text{ nebo } \frac{S_X^2}{S_Y^2} \ge k_2$$

kde k_1, k_2 se určí z podmínek
$$P\left(\frac{S_X^2}{S_Y^2} \le k_1\right)=\frac{\alpha}{2},\quad P\left(\frac{S_X^2}{S_Y^2} \ge k_2\right)=\frac{\alpha}{2}$$

$$\frac{\alpha}{2} = P\left(\frac{S_x^2}{S_y^2} \le k_1\right) = 1 - P\left(\frac{S_x^2}{S_y^2} \geq k_1\right)$$

$$\Rightarrow P\left(\frac{S_x^2}{S_y^2} \geq k_1\right) = 1 - \frac{\alpha}{2} \Rightarrow k_1 \text{ je kritická hodnota}$$

$$f'_{n-1, m-1, 1-\alpha/2} = \frac{1}{f'_{m-1, n-1, \alpha/2}}$$

$$\frac{\alpha}{2} = P\left(\frac{S_x^2}{S_y^2} \geq k_2\right) \Rightarrow k_2 = f'_{n-1, m-1, \alpha/2}$$

v případě, že $S_x^2 \geq S_y^2$, stačí porovnat s k_2

(v opačném případě pohodíme pořadí výběrů)

test nezávislosti výběrů

máme náhodný výběr $X_1, \dots, X_n$, který realizuje veličinu $X \sim N(\mu_1, \sigma_1^2)$

— " — $Y_1, \dots, Y_n$ — " — $Y \sim N(\mu_2, \sigma_2^2)$

předpokládá se, že veličiny X, Y mají sdružené 2-rozměrné normální rozdělení

mírou závislosti mezi X a Y je jejich kovariance

def: $\operatorname{cov}(X, Y) = E(X - EX)(Y - EY)$

jsou-li X, Y nezávislé, je $\operatorname{cov}(X, Y) = 0$ (v případě normálního rozdělení platí i naopak)

$$\operatorname{cov}(X, Y) = E(X - EX)(Y - EY) = EXY - 2EXEY + EXEY =$$

$$= EXY - EXEY = 0$$

normovaná kovariance se nazývá korelace

def. korelační koeficient $\rho = \dfrac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var} X \operatorname{var} Y}}$

zřejmě $\rho = 0 \iff cov(X,Y) = 0$ pro nezávislé náh. veličiny

jinak platí $-1 \le \rho \le 1$, neboť $|cov(X,Y)| = |E(X-EX)(Y-EY)| \le$

$$\le \sqrt{E(X-EX)^2 E(Y-EY)^2} = \sqrt{var X var Y}$$ (Schwarzova nerovnost)

test nezávislosti v normálním rozdělení je tedy test $H: \rho = 0$

test je založen na odhadu (tzv. výběrový korelační koeficient)

$$r = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}, \quad \text{kde } S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

po úpravě dostaneme

$$r = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{(\sum X_i^2 - n\bar{X}^2)(\sum Y_i^2 - n\bar{Y})}}$$

existují tabulky kritických hodnot pro test $H: \rho = 0$ proti $A: \rho \neq 0$

s vlastností $P(|r| \geq r_\alpha) = \alpha$

<u>ověřování normálního rozdělení</u>

t-testy obvykle nejsou na porušení normality příliš citlivé

existuje χ^2-test dobré shody

test založený na výběrových momentech (šikmosti a špičatosti)

možná se k nim dostaneme

Kolmogorov - Smirnovův test

Ověřování normality rozdělení

Předpokládejme, že $X_1, \dots, X_n$ je náhodný výběr z rozdělení se stř. hodnotou μ a rozptylem σ^2 (oba parametry známé).

Hypotéza: $X_1, \dots, X_n$ je náh. výběr z $N(\mu, \sigma^2)$

Použijeme Pearsonův χ^2-test dobré shody.

Obor hodnot náhodné veličiny X rozdělíme na k intervalů:

$$(-\infty, t_1), \langle t_1, t_2), \langle t_2, t_3) \dots, \langle t_{k-1}, \infty)$$

napočítáme, kolik pozorování padne do každého intervalu:

n_i = počet pozorování, která padnou do $\langle t_{i-1}, t_i)$

příslušné relativní četnosti budou $\frac{n_i}{n}$

teoretické pravděpodobnosti $p_i = P(t_{i-1} \leq X < t_i)$ se dají při známých parametrech přesně spočítat:

$$p_i = P(t_{i-1} \leq X < t_i) = \phi\left(\frac{t_i - \mu}{\sigma}\right) - \phi\left(\frac{t_{i-1} - \mu}{\sigma}\right),$$

kde ϕ je distr. fce rozdělení $N(0,1)$

Za platnosti hypotézy by měly být hodnoty p_i a $\frac{n_i}{n}$ přibližně stejné.

Použije se testová statistika

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n = \frac{1}{n} \sum_{i=1}^{k} \frac{n_i^2}{p_i} - n$$

která má za platnosti hypotézy rozdělení χ^2_{k-1}.

Hypotéza se zamítá, když $\chi^2 \geq \chi^2_{k-1}(\alpha)$, kde $\chi^2_{k-1}(\alpha)$ je kritická hodnota na hladině α.

Poznámky:

a) Tímto způsobem se dá ověřit shoda s jakýmkoli rozdělením – místo ϕ se použije příslušná distribuční funkce.

b) Problémem je znalost hodnot parametrů – zpravidla nejsou předem známy a musí se nějakým způsobem odhadnout (existuje test χ^2 dobré shody při neznámých parametrech, je komplikovanější).

c) Test je pouze asymptotický, aproximace χ^2-rozdělením je dobrá, pokud je $np_i \geq 5$ pro každé i.

Princip odhadu neznámých parametrů:

pravděpodobnosti p_i jsou funkcemi μ a σ^2, tj.:

$$p_i = p_i(\mu,\sigma) = \Phi\left(\frac{t_i-\mu}{\sigma}\right) - \Phi\left(\frac{t_{i-1}-\mu}{\sigma}\right) = \int_{t_{i-1}}^{t_i} f(x)\,dx,$$

$$\text{kde } f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (\text{hustota})$$

Protože ve prospěch hypotézy svědčí malé hodnoty χ^2, hledají se μ a σ tak, aby χ^2 bylo minimální, tj. řeší se rovnice

$$\frac{\partial \chi^2}{\partial \mu} = 0 \quad \text{a} \quad \frac{\partial \chi^2}{\partial \sigma} = 0, \quad \text{kde } \chi^2 = \sum_{i=1}^{k} \frac{(X_i - n p_i(\mu,\sigma))^2}{n p_i(\mu,\sigma)}.$$

Přesné řešení je ale složité, používají se různá zjednodušení těch rovnic (některé členy se zanedbají) a nakonec se stejně řeší iteračně.

Jednovýběrový Kolmogorov - Smirnovův test :

$X_1, \dots, X_n$ je náhodný výběr z nějakého rozdělení se spojitou distr. fcí

testujeme hypotézu, že tato distribuční fce je F

vypočteme F_n - empirická distr. fce :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \xi_i(x) \quad , \text{ kde } \quad \xi_i(x) = \begin{cases} 1, & \text{je-li } X_i < x \\ 0, & \text{je-li } X_i > x \end{cases}$$

platí: $F_n(x) \to F(x)$ skoro jistě pro $n \to \infty$

položíme $D_n = \sup_x |F_n(x) - F(x)|$

ve prospěch hypotézy svědčí malé hodnoty D_n

hypotéza se zamítá, když $D_n \geq D_n(\alpha)$ ... kritická hodnota

kritické hodnoty bývají tabelovány

pro velká n jsou přibližně $D_n(\alpha) \doteq \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$

Poznámky :

1) Empirická distr. fce je schodovitá, stačí tedy hodnoty D_n počítat v bodech, kde má skoky (musí se vyšetřit i limity zprava) %.

2) Potíž je, když nejsou známy parametry testované distribuční funkce. Musí se odhadovat z výběru, ale tím se mění rozdělení testové statistiky a kritické hodnoty se pak počítají na základě simulačních studií.