Vícerozměrná data

Sdružená (simultánní) absolutní četnost

$$n_{jk} = N(X = x_{[j]} \wedge Y = y_{[k]})$$

Sdružená (simultánní) relativní četnost

$$p_{jk} = \frac{n_{jk}}{n}$$

Marginální absolutní četnost varianty $x_{[j]}$

$$n_{j.} = N(X = x_{[j]}) = n_{j1} + \cdots + n_{js}$$

Marginální relativní četnost varianty $x_{[j]}$

$$p_{j.} = \frac{n_{j.}}{n} = p_{j1} + \cdots + p_{js}$$

Marginální absolutní četnost varianty $y_{[k]}$

$$n_{.k} = N(X = y_{[k]}) = n_{1k} + \cdots + n_{rk}$$

Marginální relativní četnost varianty $y_{[k]}$

$$p_{.k} = \frac{n_{.k}}{n} = p_{1k} + \cdots + p_{rk}$$

Kovariance

$$s_{n,xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Výběrová kovariance

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Pearsonův korelační koeficient

$$r_{xy} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_{n,x}} \cdot \frac{y_i - \bar{y}}{s_{n,y}} = \frac{s_{n,xy}}{s_{n,x} s_{n,y}} = \frac{s_{xy}}{s_x s_y}$$

Spearmanův koeficient pořadové korelace

$$r_{xy}^S = 1 - \frac{6 \sum_{i=1}^{n} (p_i - q_i)^2}{n(n^2 - 1)}$$

Vícerozměrná náhodná veličina

Sdružená distribuční funkce vektoru $(X, Y)'$

$$F(x, y) = P(X \leqslant x, Y \leqslant y)$$

X a Y mají distribuční funkce

$$F_X(x) = F(x, \infty) \quad \text{a} \quad F_Y(y) = F(\infty, y).$$

Sdružená pravděpodobnostní funkce vektoru $(X, Y)'$

$$p(x, y) = P(X = x, Y = y)$$

Marginální pravděpodobnostní funkce X a Y jsou

$$p_X(x) = \sum_{y \in M_y} p(x, y), \quad x \in M_x,$$
$$p_Y(y) = \sum_{x \in M_x} p(x, y), \quad y \in M_y.$$

Sdružená hustota pravděpodobnosti vektoru $(X, Y)'$

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t) \, \mathrm{d}s \, \mathrm{d}t$$

Marginální hustoty X a Y jsou

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}y$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}x$$

Náhodné veličiny X a Y jsou nezávislé právě tehdy, když

$$F(x, y) = F_X(x) \cdot F_Y(y).$$

Diskrétní náhodné veličiny jsou nezávislé, právě když

$$p(x, y) = p_X(x) \cdot p_Y(y),$$

spojité náhodné veličiny jsou nezávislé právě tehdy, když

$$f(x, y) = f_X(x) \cdot f_Y(y).$$

Střední hodnota vektoru $\boldsymbol{X} = (X, Y)'$

$$\mathrm{E}(\boldsymbol{X}) = (\mathrm{E}(X), \mathrm{E}(Y))'$$

Kovariance

$$\mathrm{C}(X, Y) = \mathrm{E}[X - \mathrm{E}(X)][Y - \mathrm{E}(Y)] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$$

Korelační koeficient

$$\rho(X, Y) = \frac{\mathrm{C}(X, Y)}{\sqrt{\mathrm{D}(X)\mathrm{D}(Y)}}$$

Test významnosti korelačního koeficientu

H: $\rho = 0 \rightarrow A$: $\rho \neq 0$

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}}\sqrt{n-2} \sim t(n-2)$$

W_α: $|t| \geqslant t_{1-\alpha/2}(n-2)$

Test nezávislosti v kontingenční tabulce

H: X a Y jsou nezávislé náhodné veličiny $\rightarrow A$: X a Y jsou závislé náhodné veličiny

$$\chi^2 = \sum_{j=1}^{r}\sum_{k=1}^{s}\frac{(n_{jk} - o_{jk})^2}{o_{jk}},\ o_{jk} = \frac{n_{j.}n_{.k}}{n}$$

W_α: $\chi^2 \geqslant \chi^2_{1-\alpha}(\nu)$, $\nu = (r-1)(s-1)$

Pearsonův koeficient kontingence

$$C_P = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Cramerův koeficient kontingence

$$C_V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, s-1)}}$$

Čuprovův koeficient kontingence

$$C_T = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(r-1)(s-1)}}}$$

Analýza rozptylu

Jednofaktorová analýza rozptylu

Výběrový průměr pro i-tý náhodný výběr (i-tou skupinu)

$$\bar{Y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

Výběrový rozptyl pro i-tý náhodný výběr

$$S_{i.}^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2$$

Rozptyl výběrových průměrů

$$S_{\bar{Y}_{i.}}^2 = \frac{1}{n-1}\sum_{i=1}^{k}(\bar{Y}_{i.} - \bar{Y})^2 n_i$$

Průměr výběrových rozptylů

$$\bar{S}_{i.}^2 = \frac{1}{n}\sum_{i=1}^{k} S_{i.}^2 n_i$$

Celkový průměr

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

Celkový rozptyl

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$$

Variabilita vysvětlená faktorem A (meziskupinová)

$$S_A = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y})^2 = (n-1)S^2_{\bar{Y}_{i.}}$$

Variabilita reziduální (vnitřní)

$$S_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^{k}(n_i - 1)S^2_{i.}$$

Celková variabilita

$$S_c = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = (n-1)S^2$$

$$S_c = S_A + S_e$$

Index determinace

$$i^2 = \frac{S_A}{S_c}$$

Test shody středních hodnot
$H : \mu_1 = \mu_2 = \cdots = \mu_k \rightarrow A : \mu_i \neq \mu_j$ pro nějaké $i, j = 1, 2, \ldots, k, i \neq j$

$$F = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}} = \frac{(n-k)\cdot S_A}{(k-1)\cdot S_e},$$

$W_\alpha : F \geqslant F_{1-\alpha}(k-1, n-k)$.

Zdroj variability	*Součet čtverců SS*	*Stupně volnosti df*	*Podíl* $\frac{SS}{df}$	*Testová statistika*
Faktor	S_A	$df_A = k-1$	$\frac{S_A}{df_A}$	$F = \frac{S_A/df_A}{S_e/df_e}$
Reziduální	S_e	$df_e = n-k$	$\frac{S_e}{df_e}$	–
Celkový	S_c	$df_c = n-1$	–	–

Dvoufaktorová analýza rozptylu

$$\bar{Y}_{ij.} = \frac{1}{r}\sum_{k=1}^{r} Y_{ijk}, \quad \bar{Y}_{i..} = \frac{1}{br}\sum_{j=1}^{b}\sum_{k=1}^{r} Y_{ijk}, \quad \bar{Y}_{.j.} = \frac{1}{ar}\sum_{i=1}^{a}\sum_{k=1}^{r} Y_{ijk}, \quad \bar{Y}_{...} = \frac{1}{abr}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} Y_{ijk}$$

$$S_c = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(Y_{ijk} - \bar{Y}_{...})^2$$

$$S_A = br\sum_{i=1}^{a}(\bar{Y}_{i..} - \bar{Y}_{...})^2$$

$$S_B = ar\sum_{j=1}^{b}(\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

$$S_{AB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$$

$$S_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(Y_{ijk} - \bar{Y}_{ij.})^2 = S_c - S_A - S_B - S_{AB}$$

$$S_c = S_A + S_B + S_{AB} + S_e$$

Zdroj variability	*Součet čtverců SS*	*Stupně volnosti df*	*Podíl* $\frac{SS}{df}$	*Testová statistika*
Faktor A	S_A	$df_A = a-1$	$\frac{S_A}{df_A}$	$F_A = \frac{S_A/df_A}{S_e/df_e}$
Faktor B	S_B	$df_B = b-1$	$\frac{S_B}{df_B}$	$F_B = \frac{S_B/df_B}{S_e/df_e}$
Interakce	S_{AB}	$df_{AB} = (a-1)(b-1)$	$\frac{S_{AB}}{df_{AB}}$	$F_{AB} = \frac{S_{AB}/df_{AB}}{S_e/df_e}$
Reziduální	S_e	$df_e = n-ab$	$\frac{S_e}{df_e}$	–
Celkový	S_c	$f_c = n-1$	–	–

Lineární regresní model

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \ldots n$$

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Normální rovnice

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$$

Odhady metodou nejmenších čtverců

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Reziduální součet čtverců

$$S_e = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$$

Reziduální rozptyl

$$s_e^2 = \frac{S_e}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Směrodatné chyby odhadů

$$s(\widehat{\beta}_j) = \sqrt{s_e^2 v_{jj}},$$

kde $v_{11}, v_{22}, \ldots, v_{kk}$ jsou prvky na hlavní diagonále matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$
Oboustranný interval spolehlivosti pro odhad parametru β_j při riziku odhadu α

$$\widehat{\beta}_j - t_{1-\alpha/2}(n-k)\cdot s(\widehat{\beta}_j) < \beta_j < \widehat{\beta}_j + t_{1-\alpha/2}(n-k)\cdot s(\widehat{\beta}_j)$$

Testy významnosti parametrů β_j, $j = 1, 2, \ldots, k$
H: $\beta_j = 0 \rightarrow A$: $\beta_j \neq 0$.

$$t = \frac{\widehat{\beta}_j}{s(\widehat{\beta}_j)}$$

W_α: $|t| \geqslant t_{1-\alpha/2}(n-k)$
Odhad regresní funkce $y = y(\boldsymbol{x})$ v bodě $\boldsymbol{x} = \boldsymbol{x}_0 = (x_{01}, x_{02}, \ldots, x_{0k})'$ pak získáme ze vztahu

$$\widehat{y} = \widehat{y}(\boldsymbol{x}_0) = \boldsymbol{x}_0'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_1 x_{01} + \widehat{\beta}_2 x_{02} + \cdots + \widehat{\beta}_k x_{0k}.$$

Interval spolehlivosti pro regresní funkci v bodě $\boldsymbol{x}_0$

$$\widehat{y}(\boldsymbol{x}_0) - t_{1-\alpha/2}(n-k)\cdot s(\widehat{y}(\boldsymbol{x}_0)) < y(\boldsymbol{x}_0) < \widehat{y}(\boldsymbol{x}_0) + t_{1-\alpha/2}(n-k)\cdot s(\widehat{y}(\boldsymbol{x}_0)),$$

kde $s(\widehat{y}(\boldsymbol{x}_0)) = s_e\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}$
Interval spolehlivosti pro předpověď v bodě $\boldsymbol{x}_0$

$$\widehat{y}(\boldsymbol{x}_0) - t_{1-\alpha/2}(n-k)\cdot s_0 < Y_0 < \widehat{y}(\boldsymbol{x}_0) + t_{1-\alpha/2}(n-k)\cdot s_0,$$

kde $s_0 = s_e\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}$
Celkový součet čtverců

$$S_Y = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y}^2$$

Reziduální součet čtverců

$$S_e = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$$

Teoretický součet čtverců

$$S_T = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\overline{y}^2$$

$$S_Y = S_T + S_e$$

Koeficient determinace

$$R^2 = \frac{S_T}{S_Y} = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}$$

Koeficient determinace (korigovaný – adjusted)

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k}$$

Test významnosti regresního modelu (s konstantou)
H: $\beta_2 = \beta_3 = \cdots = \beta_k = 0 \rightarrow A$: $\beta_j \neq 0$ pro alespoň jedno $j = 2, 3, \ldots, k$.

$$F = \frac{S_T}{k-1} : \frac{S_e}{n-k}$$

W_α: $F \geqslant F_{1-\alpha}(k-1, n-k)$