A6M33SSL: Vzorce ke zkoušce

25. května 2016

1 Vlastnosti podmíněných rozdělení

$$\mathrm{E}Y = \mathrm{E}(\mathrm{E}(Y|X)),$$
$$\mathrm{D}Y = \mathrm{D}(\mathrm{E}(Y|X)) + \mathrm{E}(\mathrm{D}(Y|X)).$$

2 Intervalové odhady parametrů **normálního** rozdělení $\mathrm{N}(\mu, \sigma^2)$

(Uvedeny pouze symetrické oboustranné odhady.)

Odhad střední hodnoty při **známém** rozptylu σ^2: $\bar{\boldsymbol{x}} \pm \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

Odhad střední hodnoty při **neznámém** rozptylu: $\bar{\boldsymbol{x}} \pm \frac{\sigma}{\sqrt{n}} q_{\mathrm{t}(n-1)}\left(1 - \frac{\alpha}{2}\right)$.

Odhad rozptylu: $\left\langle \frac{(n-1)s_{\boldsymbol{x}}^2}{q_{\chi^2(n-1)}\left(1-\frac{\alpha}{2}\right)}, \frac{(n-1)s_{\boldsymbol{x}}^2}{q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right)} \right\rangle$.

3 Testování hypotéz

Typický tvar testu: Testovací statistiku T (přesněji její relaizaci t), která roste s parametrem θ, porovnáme s kvantily příslušného rozdělení za předpokladu $\theta = c$:

H_0	H_A	H_0 zamítáme, když	dosažená významnost P
$\theta \le c$	$\theta > c$	$t > q_T(1-\alpha)$	$1 - F_T(t)$
$\theta \ge c$	$\theta < c$	$t < q_T(\alpha)$	$F_T(t)$
$\theta = c$	$\theta \ne c$	$t > q_T(1-\frac{\alpha}{2})$ nebo $t < q_T(\frac{\alpha}{2})$	$2\min(F_T(t), 1 - F_T(t))$

Test střední hodnoty normálního rozdělení při **známém** rozptylu σ^2: $\frac{\bar{\boldsymbol{x}} - c}{\sigma}\sqrt{n}$ testujeme na rozdělení $\mathrm{N}(0,1)$.

Test střední hodnoty normálního rozdělení při **neznámém** rozptylu: $\frac{\bar{\boldsymbol{x}} - c}{s_{\boldsymbol{x}}}\sqrt{n}$ testujeme na rozdělení $\mathrm{t}(n-1)$.

Test rozptylu normálního rozdělení: $\frac{(n-1)s_{\boldsymbol{x}}^2}{c}$ testujeme na rozdělení $\chi^2(n-1)$.

Test rovnosti rozptylů dvou normálních rozdělení (Fisherův): $\frac{s_{\boldsymbol{x}}^2}{s_{\boldsymbol{y}}^2}$ testujeme na rozdělení $\mathrm{F}(m-1, n-1)$.

Test rovnosti středních hodnot dvou normálních rozdělení se **známým** rozptylem σ^2: $\frac{\bar{\boldsymbol{x}}_m - \bar{\boldsymbol{y}}_n}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}$ testujeme na rozdělení $\mathrm{N}(0,1)$.

Test rovnosti středních hodnot dvou normálních rozdělení s **neznámým** rozptylem: $\frac{\bar{\boldsymbol{x}}_m - \bar{\boldsymbol{y}}_n}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$, kde $s^2 = \frac{(m-1)s_{\boldsymbol{x}}^2 + (n-1)s_{\boldsymbol{y}}^2}{m+n-2}$, testujeme na rozdělení $\mathrm{t}(m+n-2)$.

Párový pokus - test rovnosti středních hodnot dvou normálních rozdělení se **známým** rozptylem σ^2: $\frac{\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}}{\sigma}\sqrt{\frac{n}{2}}$ testujeme na rozdělení $\mathrm{N}(0,1)$.

Párový pokus - test rovnosti středních hodnot dvou normálních rozdělení se **neznámým** rozptylem: $\frac{\bar{\boldsymbol{\delta}}}{s_{\boldsymbol{\delta}}}\sqrt{n}$, kde $\delta_j = x_j - y_j$, testujeme na rozdělení $\mathrm{t}(n-1)$.

3.1 Testy korelace

$$r_{\boldsymbol{x},\boldsymbol{y}} = \frac{\sum\limits_{j=1}^{n}(x_j-\bar{\boldsymbol{x}})(y_j-\bar{\boldsymbol{y}})}{\sqrt{\left(\sum\limits_{j=1}^{n}(x_j-\bar{\boldsymbol{x}})^2\right)\left(\sum\limits_{j=1}^{n}(y_j-\bar{\boldsymbol{y}})^2\right)}} = \frac{n\sum\limits_{j=1}^{n}x_jy_j-\left(\sum\limits_{j=1}^{n}x_j\right)\left(\sum\limits_{j=1}^{n}y_j\right)}{\sqrt{\left(n\sum\limits_{j=1}^{n}x_j^2-\left(\sum\limits_{j=1}^{n}x_j\right)^2\right)\left(n\sum\limits_{j=1}^{n}y_j^2-\left(\sum\limits_{j=1}^{n}y_j\right)^2\right)}} = \rho(\mathrm{Emp}(\boldsymbol{x},\boldsymbol{y})),$$

$z_{\boldsymbol{x},\boldsymbol{y}} = \frac{1}{2}\ln\frac{1+r_{\boldsymbol{x},\boldsymbol{y}}}{1-r_{\boldsymbol{x},\boldsymbol{y}}} = h(r_{\boldsymbol{x},\boldsymbol{y}})$, kde $h(t) = \frac{1}{2}\ln\frac{1+t}{1-t}$.

Test nekorelovanosti dvou normál. rozdělení, $H_0 : \rho(X,Y)=0$: $\frac{r_{\boldsymbol{x},\boldsymbol{y}}}{\sqrt{1-r^2_{\boldsymbol{x},\boldsymbol{y}}}}\sqrt{n-2}$ testujeme na rozdělení $\mathrm{t}(n-2)$.

Test **nenulové** hodnoty korelace dvou normálních rozdělení, $H_0 : \rho(X,Y)=c$, kde $c\neq 0$:
$(h(r_{\boldsymbol{x},\boldsymbol{y}})-h(c))\sqrt{n-3} = \frac{z_{\boldsymbol{x},\boldsymbol{y}}-\mu_{z_{\boldsymbol{x},\boldsymbol{y}}}}{\sigma_{z_{\boldsymbol{x},\boldsymbol{y}}}}$, kde $\mu_{z_{\boldsymbol{x},\boldsymbol{y}}} = h(c) = \frac{1}{2}\ln\frac{1+c}{1-c}$ a $\sigma_{r_{\boldsymbol{x},\boldsymbol{y}}} = \sqrt{\frac{1}{n-3}}$, testujeme na rozd. $\mathrm{N}(0,1)$.

Test rovnosti korelací dvou normálních rozdělení, $H_0 : \rho(X,Y)=\rho(U,V)$: $\frac{h(r_{\boldsymbol{x},\boldsymbol{y}})-h(r_{\boldsymbol{u},\boldsymbol{v}})}{\sqrt{\frac{1}{n-3}+\frac{1}{m-3}}} = \frac{z_{\boldsymbol{x},\boldsymbol{y}}-z_{\boldsymbol{u},\boldsymbol{v}}}{\sigma_Z}$, kde

$\sigma_Z = \sqrt{\frac{1}{n-3}+\frac{1}{m-3}}$, testujeme na rozdělení $\mathrm{N}(0,1)$.

4 Regrese

4.1 Lineární model dimenze 1

$$Y = \theta_0+\theta_1X+\mathcal{E}, \qquad y_j = \theta_0+\theta_1x_j+e_j.$$

Regresní přímka 1:

$$y = \hat\theta_0+\hat\theta_1x, \qquad y-\bar y = \hat\theta_1(x-\bar x),$$
$$\hat\theta_0 = \bar y-\hat\theta_1\bar x, \qquad \hat\theta_1 = \frac{\sum_j x_jy_j-n\bar x\bar y}{\sum_j x_j^2-n\bar x^2} = \frac{\sum_j(x_j-\bar x)(y_j-\bar y)}{\sum_j(x_j-\bar x)^2}.$$

Odhad hodnot nezávisle proměnné v jednotlivých realizacích a chyby (**rezidua**): $y_j = \hat\theta_0+\hat\theta_1x_j, \quad \hat e_j = y_j-\hat y_j$.
Regresní přímka 2: $x = \hat\theta_0^*+\hat\theta_1^*y, \quad x-\bar x = \hat\theta_1^*(y-\bar y)$.

4.2 Interpretace regresních koeficientů

$$\widehat\sigma^2_{\boldsymbol{x}} = \frac{1}{n}\sum_j(x_j-\bar x)^2 = \frac{n-1}{n}s^2_{\boldsymbol{x}} = \mathrm{D}\,\mathrm{Emp}(\boldsymbol{x}),$$
$$\widehat\sigma^2_{\boldsymbol{y}} = \frac{1}{n}\sum_j(y_j-\bar y)^2 = \frac{n-1}{n}s^2_{\boldsymbol{y}} = \mathrm{D}\,\mathrm{Emp}(\boldsymbol{y}),$$
$$c_{\boldsymbol{x},\boldsymbol{y}} = \frac{1}{n}\sum_j(x_j-\bar x)(y_j-\bar y) = \mathrm{cov}(\mathrm{Emp}(\boldsymbol{x},\boldsymbol{y})),$$
$$r_{\boldsymbol{x},\boldsymbol{y}} = \frac{c_{\boldsymbol{x},\boldsymbol{y}}}{\widehat\sigma_{\boldsymbol{x}}\widehat\sigma_{\boldsymbol{y}}} = \rho(\mathrm{Emp}(\boldsymbol{x},\boldsymbol{y})), \qquad \hat\theta_1 = \frac{c_{\boldsymbol{x},\boldsymbol{y}}}{\widehat\sigma^2_{\boldsymbol{x}}}, \qquad \hat\theta_1\hat\theta_1^* = r^2_{\boldsymbol{x},\boldsymbol{y}}.$$

Regresní přímka 1: $y-\bar y = \frac{c_{\boldsymbol{x},\boldsymbol{y}}}{\widehat\sigma^2_{\boldsymbol{x}}}(x-\bar x), \qquad \frac{y-\bar y}{\widehat\sigma_{\boldsymbol{y}}} = r_{\boldsymbol{x},\boldsymbol{y}}\frac{x-\bar x}{\widehat\sigma_{\boldsymbol{x}}}$.

4.3 Složky rozptylu regresního odhadu

Rozptyl modelu (vysvětlený rozptyl): $\widehat{\sigma}^2_{\hat{\boldsymbol{y}}} = \frac{1}{n}\sum_j (\hat{y}_j - \bar{y})^2$.

Reziduální rozptyl (nevysvětlený rozptyl): $\widehat{\sigma}^2_{\hat{\boldsymbol{e}}} = \frac{1}{n}\sum_j \hat{e}_j^2 = \frac{1}{n}\sum_j (y_j - \hat{y}_j)^2$.

Celkový rozptyl: $\widehat{\sigma}^2_{\boldsymbol{y}} = \frac{1}{n}\sum_j (y_j - \bar{y})^2 = \widehat{\sigma}^2_{\hat{\boldsymbol{y}}} + \widehat{\sigma}^2_{\hat{\boldsymbol{e}}}$. **Koeficient determinace**: $r^2_{\boldsymbol{x},\boldsymbol{y}} = \frac{\widehat{\sigma}^2_{\hat{\boldsymbol{y}}}}{\widehat{\sigma}^2_{\boldsymbol{y}}} = 1 - \frac{\widehat{\sigma}^2_{\hat{\boldsymbol{e}}}}{\widehat{\sigma}^2_{\boldsymbol{y}}}$.

Odhady rozptylu σ^2 *původního* rozdělení:

- max. věrohodný: $\widehat{\sigma}^2_{\hat{\boldsymbol{e}}} = \frac{1}{n}\sum_j \hat{e}_j^2 = \mathrm{D}\,\mathrm{Emp}(\hat{\boldsymbol{e}})$,
- nestranný: $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_j \hat{e}_j^2$.

4.4 Rozdělení odhadů a testy hypotéz o nich

Odhady rozptylů regresních koeficientů:

$$\widehat{\sigma}^2_{\widehat{\theta}_0} = \frac{\widehat{\sigma}^2}{n^2\widehat{\sigma}^2_{\boldsymbol{x}}}\sum_j x_j^2 = \frac{\widehat{\sigma}^2_{\hat{\boldsymbol{e}}}}{n(n-2)\widehat{\sigma}^2_{\boldsymbol{x}}}\sum_j x_j^2, \qquad \widehat{\sigma}^2_{\widehat{\theta}_1} = \frac{\widehat{\sigma}^2}{n\widehat{\sigma}^2_{\boldsymbol{x}}} = \frac{\widehat{\sigma}^2_{\hat{\boldsymbol{e}}}}{(n-2)\widehat{\sigma}^2_{\boldsymbol{x}}}.$$

Test absolutního členu, $H_0 : \theta_0 = c$: $\dfrac{\widehat{\theta}_0 - c}{\frac{\widehat{\sigma}_{\hat{\boldsymbol{e}}}}{\widehat{\sigma}_{\boldsymbol{x}}}\sqrt{\frac{1}{n}\sum_j x_j^2}}\sqrt{n-2}$ testujeme na rozdělení $\mathrm{t}(n-2)$.

Test směrnice, $H_0 : \theta_1 = c$: $\dfrac{\widehat{\theta}_1 - c}{\frac{\widehat{\sigma}_{\hat{\boldsymbol{e}}}}{\widehat{\sigma}_{\boldsymbol{x}}}}\sqrt{n-2}$ testujeme na rozdělení $\mathrm{t}(n-2)$.

Test chyby regrese pro dané x, $H_0 : \theta_0 + \theta_1 x = c$: $\dfrac{\widehat{\theta}_0 + \widehat{\theta}_1 x - c}{\widehat{\sigma}_{\hat{\boldsymbol{e}}}\sqrt{n+1+\frac{n(x-\bar{x})^2}{\widehat{\sigma}^2_{\boldsymbol{x}}}}}\sqrt{n-2}$ testujeme na rozdělení $\mathrm{t}(n-2)$.

4.5 Lineární model dimenze k

$$Y = \sum_{i=1}^{k}\theta_i Xi + \mathcal{E}, \qquad y_j = \sum_{i=1}^{k}\theta_i x_{ji} + e_j, \qquad \boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{e}.$$

Soustava normálních rovnic a její řešení: $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}^T\boldsymbol{y}$, $\quad \hat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
Kovarianční matice vektoru odhadů $\hat{\boldsymbol{\Theta}}$: $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Theta}}} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Reziduální součet čtverců: $R_{SS} = \sum_{j=1}^{n}\hat{e}_j^2$.

$\frac{R_{SS}}{\sigma^2}$ pochází z rozdělení $\chi^2(n-k)$.
Odhady rozptylu σ^2 původního rozdělení:

- maximálně věrohodný: $\widehat{\sigma}^2_{\hat{\boldsymbol{e}}} = \frac{1}{n}R_{SS} = \frac{1}{n}\sum_{j=1}^{n}\hat{e}_j^2$,
- nestranný: $\hat{\sigma}^2 = \frac{1}{n-k}R_{SS} = \frac{1}{n-k}\sum_{j=1}^{n}\hat{e}_j^2$.

4.6 Intervalové odhady regresních koeficientů

c_{ii} je i-tý prvek na diagonále matice $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$.
Intervalový odhad regresních koeficientů při **známém** rozptylu σ^2: $\hat{\boldsymbol{\theta}}_i \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sigma\sqrt{c_{ii}}$.
Intervalový odhad regresních koeficientů při **neznámém** rozptylu:

$$\hat{\boldsymbol{\theta}}_i \pm q_{\mathrm{t}(n-k)}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}\sqrt{c_{ii}} = \hat{\boldsymbol{\theta}}_i \pm q_{\mathrm{t}(n-k)}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{R_{SS}c_{ii}}{n-k}}.$$

5 χ^2 testy

Test dobré shody pozorovaného rozdělení s diskrétním rozdělením se **známými** parametry: $\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$ testujeme na rozdělení $\chi^2(k-1)$.

Test dobré shody pozorovaného rozdělení s diskrétním rozdělením závislém na **neznámých** parametrech: $\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$ testujeme na rozdělení $\chi^2(k-1-q)$, kde q je počet parametrů rozdělení odhadnutých na základě téhož náhodného výběru.

Test dobré shody 2 diskrétních rozdělení: $\left(\frac{1}{m} + \frac{1}{n}\right) \sum_{i=1}^{k} \frac{(m_i - mp_i)^2}{p_i}$, kde $p_i = \frac{m_i + n_i}{m+n}$, testujeme na rozdělení $\chi^2(k-1)$.

Test nezávislosti 2 diskrétních rozdělení: $\sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - np_i q_j)^2}{np_i q_j}$, kde $p_i = \frac{1}{n} \sum_{j=1}^{m} n_{ij}$ a $q_j = \frac{1}{n} \sum_{i=1}^{k} n_{ij}$, testujeme na rozdělení $\chi^2((k-1)(m-1))$.