Statistička analiza podataka: Podsjetnik na formule

UNIZG FER, ak. god. 2019./2020.

1 Razni testovi

Z-test

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma}\sqrt{n}$$

T-test za jedan uzorak

$$T = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$$

T-test za dva uzorka (uz pretpostavku jednakosti varijanci)

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_X\sqrt{1/n_1 + 1/n_2}}$$

$$S_X^2 = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2]$$

χ^2-test o varijanci

$$\chi^2 = \frac{(n-1)S_n^2}{\sigma^2}$$

F-test

$$F = \frac{S_{X_1}^2}{S_{X_2}^2}$$

χ^2-test prilagodbe modela podacima

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

χ^2-test nezavisnosti/homogenosti

$$\chi^2 = \sum_{i,j} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}$$

2 Jackknife

$$\begin{aligned}
\hat{bias}(\hat{\theta})_{\text{jack}} &= (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) \\
ps_i &= n\hat{\theta} - (n-1)\hat{\theta}_{(i)} \\
SE(\hat{\theta})_{\text{jack}} &= \Big(\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2\Big)^{1/2} \\
&= \Big(\frac{1}{n(n-1)}\sum_{i=1}^{n}(ps_i - \overline{ps})^2\Big)^{1/2}
\end{aligned}$$

3 Jednostavna regresija

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Procjena koeficijenata

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1\bar{x}$$

$$s^2 = \frac{S_{yy} - b_1 S_{xy}}{n-2}$$

Testovi o regresijskim koeficijentima

$$T = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}} \sim t(n-2)$$

$$T = \frac{B_0 - \beta_0}{S\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}} \sim t(n-2)$$

Predikcija srednje vrijednosti

$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S\sqrt{1/n + (x_0 - \bar{x})^2/S_{xx}}} \sim t(n-2)$$

Predikcija vrijednosti Y za dani x_0

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{xx}}} \sim t(n-2)$$

Koeficijent determinacije

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Pearsonov koeficijent korelacije

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

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

$$z = \frac{\sqrt{n-3}}{2}\left[\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right] \sim AN(0,1)$$

Standardizirani reziduali

$$t_i = \frac{e_i}{s\sqrt{1-h_{ii}}}, \quad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

4 Višestruka regresija

$$\mathbf{A} = \mathbf{X}^t\mathbf{X}$$

Procjena koeficijenata

$$\mathbf{Ab} = \mathbf{X}^t\mathbf{y}$$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$$

Testovi o regresijskim koeficijentima

$$T = \frac{B_j - \beta_j}{s\sqrt{c_{jj}}} \sim t(n-k-1), \quad c_{jj} = (\mathbf{A}^{-1})_{jj}$$

Predikcija srednje vrijednosti

$$T = \frac{\hat{Y}_0 - \mu_{Y|\mathbf{x}_0}}{s\sqrt{\mathbf{x}_0^t \mathbf{A}^{-1} \mathbf{x}_0}} \sim t(n-k-1)$$

Predikcija vrijednosti Y za dani x_0

$$T = \frac{\hat{Y}_0 - Y_0}{s\sqrt{1 + \mathbf{x}_0^t \mathbf{A}^{-1} \mathbf{x}_0}} \sim t(n-k-1)$$

Dekompozicija kvadratnih odstupanja

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SST = SSE + SSR$$

Koeficijent determinacije

$$R^2 = 1 - \frac{SSE}{SST}$$

Prilagođeni koeficijent determinacije

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

Standardizirani reziduali

$$t_i = \frac{e_i}{s\sqrt{1 - \mathbf{H}_{ii}}}, \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$$

5 Jednofaktorska ANOVA

Formule za sume kvadrata kada imamo uzorke jednakih veličina:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$$

$$SSA = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$$
$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed f**
Treatments	SSA	$k-1$	$s_1^2 = \frac{SSA}{k-1}$	$\frac{s_1^2}{s^2}$
Error	SSE	$k(n-1)$	$s^2 = \frac{SSE}{k(n-1)}$	
Total	SST	$kn-1$		

Tablica 1: Jednofaktorska ANOVA

Formule za sume kvadrata kada imamo uzorke različitih veličina:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$$
$$SSA = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$
$$SSE = SST - SSA$$

Bartlettov test:

$$b = \frac{\left[(s_1^2)^{n_1-1}(s_2^2)^{n_2-1}\cdots(s_k^2)^{n_k-1}\right]^{1/(N-k)}}{s_p^2}$$

$$s_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)s_i^2$$

6 Dvofaktorska ANOVA

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$+ n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$
$$+ \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{ij.})^2$$

$$SST = SSA + SSB + SS(AB) + SST$$

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed** f
Main effect:				
A	SSA	$a-1$	$s_1^2 = \frac{SSA}{a-1}$	$f_1 = \frac{s_1^2}{s^2}$
B	SSB	$b-1$	$s_2^2 = \frac{SSB}{b-1}$	$f_2 = \frac{s_2^2}{s^2}$
Two-factor interactions:				
AB	$SS(AB)$	$(a-1)(b-1)$	$s_3^2 = \frac{SS(AB)}{(a-1)(b-1)}$	$f_3 = \frac{s_3^2}{s^2}$
Error	SSE	$ab(n-1)$	$s^2 = \frac{SSE}{ab(n-1)}$	
Total	SST	$abn-1$		

Tablica 2: Dvofaktorska ANOVA

7 Neparametarski postupci

Mann-Whitney-Wilcoxonov test (Wilcoxon rank-sum test)

$$u_1 = w_1 - \frac{n_1(n_1+1)}{2}$$
$$u_2 = w_2 - \frac{n_2(n_2+1)}{2}$$
$$u = \min(u_1, u_2)$$

Kruskal-Wallisov test

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \ \sim \ \chi_\alpha^2(k-1)$$

Spearmanov koeficijent korelacije

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2$$

Za $n > 30$:

$$z = \frac{r_s - 0}{1/\sqrt{n-1}} = r_s\sqrt{n-1}$$

8 Bayesovska statistika

Bayesova formula

$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\pi(\theta)}{g(\mathbf{x})}$$

Marginalna distribucija

$$g(\mathbf{x}) = \begin{cases} \sum\limits_{\theta} f(\mathbf{x} \mid \theta)\pi(\theta), & \theta \text{ je diskretan} \\ \int_{-\infty}^{\infty} f(\mathbf{x} \mid \theta)\pi(\theta)d\theta, & \theta \text{ je kontinuiran} \end{cases}$$

Beta funkcija

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \alpha, \beta > 0,$$

gdje je $\Gamma(\alpha)$ gamma funkcija.

Funkcija gustoće Beta distribucije

$$f(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1 \\ 0, & \text{inače} \end{cases}$$

$\alpha > 0$, $\beta > 0$

Očekivanje i varijanca Beta distribucije

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Bayesov interval vjerodostojnosti

Interval $a < \theta < b$ nazivamo $100(1-\alpha)\%$ Bayesov interval za θ ako

$$\int_{-\infty}^{a} \pi(\theta \mid \mathbf{x})d\theta = \int_{b}^{\infty} \pi(\theta \mid \mathbf{x})d\theta = \frac{a}{2}.$$

Aposteriorna prediktivna distribucija

Za novi podatak x_{novi} i poznati uzorak $\mathbf{x}$ je:

$$p(x_{novi} \mid x) = \int_{\Theta} p(x_{novi} \mid \theta, \mathbf{x})p(\theta \mid \mathbf{x})d\theta = \int_{\Theta} p(x_{novi} \mid \theta)p(\theta \mid \mathbf{x})d\theta$$