Analýza a vizualizácia dát

25. marca 2024

hypotézy

- liek je efektívny vs nie je
- dáta sú z $N(0,1)$ vs nie sú
- výučbová metóda 1 je lepšia ako metóda 2 vs nie je
- minca je vyvážená vs nie je (minulý týždeň)

testovanie hypotéz o parametri

- nulová hypotéza vs alternatíva

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1$$

- medzi H_0 a H_1 sa rozhodujeme na základe pozorovaných dát = štatistika
- θ = napr. stredná hodnota, disperzia
- $\Theta_0, \Theta_1 \subset \Theta$ = parametrický priestor (prípustné hodnoty parametra)

testovanie hypotéz o strednej hodnote

- motivácia: zaujíma nás, či je priemerná výška 35 ročných mužov na Slovensku rovná 185 cm (vieme, že výška 35 r. mužov má $N(\mu, \sigma^2)$)
- náhodný výber z $N(\mu, \sigma^2)$

$$X_1, X_2, X_3, \ldots, X_n$$

- hypotéza

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

normálne rozdelenie, σ^2 známe

- testová štatistika

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \sim^{H_0} N(0,1)$$

- H_0 zamietame, ak je Z príliš veľké alebo malé

$$|Z| < z_{\frac{\alpha}{2}}$$

cvičenie

- načítajte dáta *adm_data.csv*, ktoré obsahujú informácie o študentoch hlásiacich sa na univerzitu
- testujte hypotézu o tom, že pravdepodobnosť prijatia na univerzitu je pre žiadateľov vyššia ako 50% za predpokladu, že dáta sú z normálneho rozdelenia a disperzia je 0,02

normálne rozdelenie, σ^2 neznáme

- testová štatistika

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$$

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \sim t(n-1)$$

$$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \sim^{H_0} t(n-1)$$

- H_0 zamietame, ak je T príliš veľké alebo malé

$$|T| < t_{\frac{\alpha}{2}}(n-1)$$

cvičenie

- testujte hypotézu o tom, že pravdepodobnosť prijatia na univerzitu je pre žiadateľov vyššia ako 50% za predpokladu, že dáta sú z normálneho rozdelenia a disperzia je neznáma

nie normálne rozdelenie, σ^2 neznáme

- testová štatistika

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sigma}\sqrt{n} \approx N(0,1)$$

$$\frac{\bar{X} - \mu}{S}\sqrt{n} \approx N(0,1)$$

$$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \approx^{H_0} N(0,1)$$

- H_0 zamietame, ak je T príliš veľké alebo malé

$$|T| < z_{\frac{\alpha}{2}}$$

alebo

$$|T| < t_{\frac{\alpha}{2}}(n-1)$$

dvojvýberové (parametrické) testy

- dva nezávislé dátové súbory
- muži/ženy, jedna trieda/druhá trieda, zdraví/chorí
- signifikantný rozdiel medzi dvoma skupinami

dvojvýberový z-test

- motivácia: zaujíma nás, či je priemerná výška 35 ročných mužov na Slovensku rovnaká ako žien (vieme, že výška 35 r. mužov má $N(\mu_X, \sigma^2)$ a žien $N(\mu_Y, \sigma^2)$)
- náhodný výber z $N(\mu_X, \sigma^2)$

$$X_1, X_2, X_3, \ldots, X_n$$

- náhodný výber z $N(\mu_Y, \sigma^2)$

$$Y_1, Y_2, Y_3, \ldots, Y_m$$

- normálne dáta, nezávislé, σ^2 známe, rovnaké v oboch súboroch
- hypotéza

$$H_0 : \mu_X = \mu_Y \quad \text{vs} \quad H_1 : \mu_X \neq \mu_Y$$

dvojvýberový z-test

- testová štatistika

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim^{H_0} N(0,1)$$

- H_0 zamietame, ak

$$|Z| < z_{\frac{\alpha}{2}}$$

dvojvýberový t-test

- náhodný výber z $N(\mu_X, \sigma^2)$

$$X_1, X_2, X_3, \ldots, X_n$$

- náhodný výber z $N(\mu_Y, \sigma^2)$

$$Y_1, Y_2, Y_3, \ldots, Y_m$$

- normálne dáta, nezávislé, σ^2 neznáme, rovnaké v oboch súboroch
- testová štatistika

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{n+m-2}{\frac{1}{n} + \frac{1}{m}}} \sim^{H_0} t(n+m-2)$$

Welchov t-test

- náhodný výber z $N(\mu_X, \sigma_X^2)$

$$X_1, X_2, X_3, \ldots, X_n$$

- náhodný výber z $N(\mu_Y, \sigma_Y^2)$

$$Y_1, Y_2, Y_3, \ldots, Y_m$$

- normálne dáta, nezávislé, $\sigma_X^2 \neq \sigma_Y^2$ neznáme
- testová štatistika

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \approx^{H_0} t(\nu)$$

$$\nu \approx \frac{(S_X^2/n + S_Y^2/m)^2}{S_X^4/n^2(n-1) + S_Y^4/m^2(m-1)}$$

- Welchov t-test: - asymptotický, rozumný počet dát sa očakáva aj pri normalite $(n, m > 5)$ pre $\approx t(\nu)$ - slabší ako t-test
- návod:

$$\text{normalita} \begin{cases} \text{áno} \begin{cases} \sigma_X^2 = \sigma_Y^2 & \text{známe} \rightarrow \text{z-test} \\ \sigma_X^2 = \sigma_Y^2 & \text{neznáme} \rightarrow \text{t-test} \\ \sigma_X^2 \neq \sigma_Y^2 & \text{neznáme} \rightarrow \text{Welchov t-test} \end{cases} \\ \text{nie} \rightarrow \text{Welchov t-test} \end{cases}$$

F-test rovnosti disperzií

- náhodný výber z $N(\mu_1, \sigma_1^2)$

$$X_1, X_2, X_3, \ldots, X_n$$

- náhodný výber z $N(\mu_2, \sigma_2^2)$

$$Y_1, Y_2, Y_3, \ldots, Y_m$$

- normálne dáta, nezávislé, σ_X^2, σ_Y^2 neznáme
- testová štatistika

$$\frac{S_X^2}{S_Y^2} \sim^{H_0} F(n-1, m-1)$$

$$\left(\frac{S_Y^2}{S_X^2} \sim^{H_0} F(m-1, n-1) \right)$$