Analýza a vizualizácia dát

8. apríla 2024

cvičenie

- pre *HRdata* testujte hypotézu o tom, že *Attrition* nastáva pre v priemere mladších
- testujte hypotézu o tom, že muži majú v priemere vyšší *MonthlyIncome* ako ženy

KS test

- motivácia: je možné, že dáta sú z normálneho rozdelenia? (najčastejšie pre spojité rozdelenia alebo dva náhodné výbery)
- idea: porovnať empirickú distribučnú funkciu (získanú z dát) s teoretickou (testované rozdelenie)

$$H_0 : F_n = F \quad vs \quad H_1 : F_n \neq F$$

- testová štatistika:

$$D = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

- kritické hodnoty sú tabelované

χ^2 test dobrej zhody

- motivácia:
  - je možné, že diskrétne dáta sú z tohto rozdelenia?
  - porovnanie pozorovaných frekvencií s teoretickými
  - test nezávislosti v kontingenčných tabuľkách
- idea: porovnať pozorovaných frekvencie s teoretickými
- testová štatistika:

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i} = n \sum_{i=1}^{r} \frac{\frac{O_i}{n} - p_i}{p_i}$$

- testová štatistika má asymptoticky χ^2 rozdelenie

$$\chi^2 \approx \chi^2(r - p)$$

cvičenie

- simulujte 100 hodov nevyváženou kockou a následne testujte chí-kvadrát testom dobrej zhody, či je možné, že je kocka vyvážená
- simulujte 100 realizácií z geometrického rozdelenia s parametrom $p = 0{,}3$ a testujte pomocou chí-kvadrát testu dobrej zhody, či je možné, že dáta sú z geometrického rozdelenia

korelačná analýza

- teoretický koeficient korelácie

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{D(X)D(Y)}}$$

- výberový koeficient korelácie (Pearsonov)

$$R_{X,Y} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

test pre korelačný koeficient

- hypotéza

$$H_0 : \rho = 0 \quad vs. \quad H_1 : \rho \neq 0$$

- testová štatistika

$$T = R\sqrt{\frac{n-2}{1-R^2}} \sim_{H_0} t(n-2)$$

alebo

$$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \approx_{H_0} N\left(0, \frac{1}{n-3}\right)$$

cvičenie

- pre dáta z tabuľky *TripleVertical* spravte korelačnú analýzu a skúmajte lineárnu regresiu

lineárna regresia

- motivácia: jednoduchý model, ktorý vysvetľuje vzťah medzi dvoma premennými (prípadne chceme robiť predikcie)
- lineárny model

$$y_i \approx \beta_0 + \beta_1 x_i$$

- je tam ešte niečo, čo nepoznáme

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- náš odhad

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- β_0, β_1 odhadujeme metódou najmenších štvorcov

testovanie hypotéz v lineárnej regresii

- hypotézy o β_0, β_1

$$H_0 : \beta_1 = 0 \quad vs. H_1 : \beta_1 \neq 0$$

- či má vôbec takýto model zmysel - porovnáva sa s modelom bez x-ov

cvičenie

- pre dáta z tabuľky *beerhall* spravte korelačnú analýzu
- pomocou lineárnej regresie modelujte *Criminals per 100k population*