POROČILO SEMINARSKE NALOGE

STATISTIKA

Sara Bizjak

Univerza v Ljubljani
Fakulteta za matematiko in fiziko
Oddelek za matematiko

JULIJ 2020

1. NALOGA

Podatki so vzeti iz datoteke `Kibergard`, kjer se nahajajo informacije o 43.886 družinah, ki stanujejo v mestu Kibergard. Za vsako družino so zabeleženi naslednji podatki:

- Tip družine (od 1 do 3)
- Število članov družine
- Število otrok v družini
- Skupni dohodek družine
- Mestna četrt, v kateri stanuje družina (od 1 do 4)
- Stopnja izobrazbe in vodje gospodinjstva (od 31 do 46: opisi v datoteki z navodili)

Nalogo sem reševala s pomočjo programa R. Koda, uporabljena za generiranje enostavnih slučajnih vzorcev in izračune, je dostopna v priloženi datoteki *naloga1.R*.

PRIMER A
Vzamemo enostavni slučajni vzorec 200 družin in na njegovi podlagi ocenimo delež družin v Kibergardu, v katerih vodja gospodinjstva nima srednješolske izobrazbe (niti poklicne niti splošne mature). Opisan delež znaša $p = 0.195$.

PRIMER B
Ocenimo standardno napako in postavimo 95% interval zaupanja.
Standardno napako za delež izračunamo po formuli (na strani 210 v knjigi *John A. Rice: Mathematical Statistics and Data Analysis*):

$$\hat{se}(p) = \sqrt{\frac{p \cdot (1-p)}{n} \cdot \left(1 - \frac{n-1}{N-1}\right)},$$

kjer so $p = 0.195$, $n = 200$, $N = 43.886$.
Dobimo rezultat $\hat{se}(p) = 0.02795203$.
Interval zaupanja je enak: $[0.140215, 0.249785]$.

PRIMER C
Vzorčni delež in ocenjeno standardno napako primerjamo s populacijskim deležem in pravo standardno napako.

- Vzorčni delež: 0.195
- Populacijski delež: 0.2115025
- Razlika obeh deležev: 0.01650253
- Ocenjena standardna napaka (iz vzorca): $0,02802185$
- Prava standardna napaka (iz celotne populacije): 0.02881085
- Razlika med ocenjeno in pravo standardno napako: 0.0008609634

Ker velja $0.2115025 \in [0.140215, 0.249785]$, interval zaupanja pokrije populacijski delež.

Primer d
GRAF
INTERVALI ZAUPANJA + KOLIKO JIH POKRIJE POPULACIJSKI DELEŽ

Primer e
Standardni odklon vzorčnih deležev za 100 prej dobljenih vzorcev je enak 0.02881085. Prava standardna napaka za vzorec velikosti 200 pa je 0.02888282. Razlikujeta se za $7.196778 \cdot 10^{-5}$.

Primer f
GRAF
INTERVALI ZAUPANJA + KOLIKO JIH POKRIJE POPULACIJSKI DELEŽ

2. NALOGA

3. NALOGA

Opazimo n neodvisnih realizacij zvezne porazdelitve z gostoto:

$$f(x \mid \sigma) = \begin{cases} \frac{\Gamma(2\alpha)}{(\Gamma(\alpha))^2} \left[x(1-x)\right]^{\alpha-1} & ;\ 0 < x < 1 \\ 0 & ;\ \text{sicer,} \end{cases}$$

kjer je $\alpha > 0$ neznam parameter. Če je X slučajna spremenljivka s to gostoto, se da izračunati:

$$\mathrm{E}(X) = \frac{1}{2}\,, \qquad \mathrm{var}(X) = \frac{1}{4(2\alpha+1)}.$$

Primer a
Določimo obliko porazdelitve v odvisnosti od α.

Primer b
Ocenimo α po metodi momentov.
Iz podane pričakovane vrednosti in variance lahko izračunamo vrednosti prvega in drugega momenta.

$$\mathrm{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{3},$$

$$\mathrm{E}(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathrm{var}(X) + \mathrm{E}(X)^2 = \frac{1}{4(2\alpha+1)} + \frac{1}{4}.$$

Iz enačbe drugega momenta izrazimo α.

$$\begin{aligned}
\frac{1}{4(2\alpha+1)} + \frac{1}{4} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\
\frac{1}{4(2\alpha+1)} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{4} \\
\frac{1}{2\alpha+1} &= \frac{4}{n}\sum_{i=1}^{n} X_i^2 - 1 \\
2\alpha+1 &= \frac{1}{\frac{4}{n}\sum_{i=1}^{n} X_i^2 - 1} \\
\Rightarrow \alpha &= \frac{1}{2}\left(\frac{1}{\frac{4}{n}\sum_{i=1}^{n} X_i^2 - 1} - 1\right).
\end{aligned}$$

Primer c
Poiščemo enačbo, ki določa cenilko po metodi največjega verjetja. Pogledamo, kdaj ta cenilka obstaja.

Primer d
Poiščemo asimptotično varianco cenilke po metodi največjega verjetja.

4. NALOGA
5. NALOGA

X in Y sta slučajni spremenljivki, za kateri velja:

- $\mathrm{E}(X) = \mu_x$,
- $\mathrm{E}(Y) = \mu_y$,

- $\mathrm{var}(X) = \sigma_x^2$,
- $\mathrm{var}(Y) = \sigma_y^2$,
- $\mathrm{cov}(X, Y) = \sigma_{x,y}$.

Opazimo X in želimo napovedati Y.

Primer a

Poiščemo napoved, ki je oblike $\hat{Y} = \alpha + \beta \cdot X$, kjer α in β izberemo tako, da je srednja kvadratična napaka $\mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right]$ minimalna.
Uporabimo namig

$$\mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right] = \left[\mathrm{E}(Y) - \mathrm{E}(\hat{Y})\right]^2 + \mathrm{var}(Y - \hat{Y}). \tag{1}$$

Ker sta oba člena desne strani enačbe večja od 0, je dovolj, da poiščemo vrednosti α in β, ki minimizirata ta dva člena – potem bo najmanjša možna tudi njuna vsota.
Poglejmo si najprej prvi člen enačbe. Vrednost enačbe

$$\left[\mathrm{E}(Y) - \mathrm{E}(\hat{Y})\right]^2 = [\mu_y - \alpha - \beta\mu_x]^2$$

bo najmanjša, ko bo $\mu_y - \alpha - \beta \cdot \mu_x = 0$. To pa bo res natanko takrat, ko bo $\alpha = \mu_y - \beta\mu_x$.
Poglejmo si še drugi člen enačbe. Vidimo, da je

$$\begin{aligned}\mathrm{var}(Y - \hat{Y}) &= \mathrm{var}(Y - \alpha - \beta X) = \mathrm{var}(Y - \beta X) = \mathrm{var}(Y) - 2\beta\mathrm{cov}(X, Y) + \beta^2\mathrm{var}(X)\\ &= \sigma_y^2 - 2\beta\sigma_{x,y} + \beta^2\sigma_x^2\end{aligned}$$

funkcija spremenljivke β. Za izračun minimuma enačbo najprej odvajamo po β in enačimo z 0.

$$\frac{\partial}{\partial\beta}(\mathrm{var}(Y - \hat{Y})) = -2\sigma_{x,y} + 2\beta\sigma_x^2 = 0$$

To bo res, ko bo $\beta = \frac{\sigma_{x,y}}{\sigma_x^2}$.
Vrednosti α in β, pri katerih bo izraz $\mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right]$ minimalen, sta

$$\alpha = \mu_y - \mu_x\frac{\sigma_{x,y}}{\sigma_x^2} \qquad \text{in} \qquad \beta = \frac{\sigma_{x,y}}{\sigma_x^2}$$

Primer b

Pokažemo, da se pri tako izbranih koeficientih *determinacijski koeficient* (kvadrat korelacijskega koeficienta) izraža v obliki

$$r_{x,y}^2 = 1 - \frac{\mathrm{var}(Y - \hat{Y})}{\mathrm{var}(Y)}.$$

Spomnimo se najprej formule za *korelacijski koeficient*:

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = r_{x,y}.$$

Ker je determinacijski koeficient enak kvadratu korelacije, je enak

$$r_{x,y}^2 = \frac{\mathrm{cov}(X, Y)^2}{\mathrm{var}(X)\mathrm{var}(Y)}.$$

Dobimo enakost

$$1 - \frac{\mathrm{var}(Y - \hat{Y})}{\mathrm{var}(Y)} = \frac{\mathrm{cov}(X, Y)^2}{\mathrm{var}(X)\mathrm{var}(Y)}$$

in če upoštevamo še vrednosti α in β iz prvega dela naloge:

$$\begin{aligned} 1 - \frac{\mathrm{var}(Y - \hat{Y})}{\mathrm{var}(Y)} &= \frac{\mathrm{var}(Y) - \mathrm{var}(Y - \hat{Y})}{\mathrm{var}(Y)} = \frac{\sigma_y^2 - \sigma_y^2 + \frac{\sigma_{x,y}^2}{\sigma_x^2}}{\sigma_y^2} \\ &= \frac{\sigma_{x,y}^2}{\sigma_x^2 \sigma_y^2} = \frac{\mathrm{cov}(X, Y)^2}{\mathrm{var}(X)\mathrm{var}(Y)} = r_{x,y}^2. \end{aligned}$$

Pokazali smo, da je determinacijski koeficient res enak $1 - \frac{\mathrm{var}(Y - \hat{Y})}{\mathrm{var}(Y)}$.