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

Naredimo raziskavo na populaciji, ki ima $K$ stratumov z velikostmi $N_1, N_2, \ldots, N_K$. Denimo, da lahko izberemo vzorec velikosti $n$.

Primer a
Denimo, da so stroški raziskave enaki $C = C_0 + nC_1$, kjer je $n$ število enot v vzorcu ($C_0$ je začetni stršek, $C_1$ pa je nadaljnji strošek na enoto). Pri danih sredstvih za raziskavo v višini $C$ poiščemo velikosti podvzorcev $n_1, n_2, \ldots, n_K$, pri katerih je varianca standardne cenilke populacijskega povprečja minimalna.
Označimo z $\overline{X}$ populacijsko povprečje.
Če imamo stratume velikosti $N_1, N_2, \ldots, N_K$, se moramo odločiti, kako velike

vzorce bomo izbrali iz posameznih stratumov. Izbiramo tako, da bo standardna napaka cenilke $\overline{X}$ čim manjša. Poiščemo torej take $n_1, n_2, \ldots, n_K$, kjer $n = n_1 + n_2 + \ldots + n_K$, da bo

$$\mathrm{var}(\overline{X}) = \sum_{i=k}^{K} W_i^2 \left(\frac{\sigma_k^2}{n_k}\right)\left(\frac{N_k - n_k}{N_k - 1}\right)$$

čim manjša. Ker so v večini praktičnih situacij korekturni faktorji $\frac{N_k - n_k}{N_k - 1} \approx 1$, jih lahko zanemarimo. Rešujemo torej problem vezanega ekstrema:

$$f(n_1, n_2, \ldots, n_K) = \sum_{k=1}^{K} \frac{\sigma_k^2}{n_k} W_k^2$$

$$\text{z vezjo } C = C_0 + nC_1.$$

Sestavimo Lagrangeovo funkcijo:

$$F(n_1, n_2, \ldots, n_K, \lambda) = f(n_1, n_2, \ldots, n_K) + \lambda(C_0 + \sum_{k=1}^{K} n_k C_1 - C).$$

Zapišemo parcialne odvode funkcije $F$ in jih enačimo z 0.

$$\frac{\partial F}{\partial n_i} = -\frac{\sigma_i^2}{n_i^2}\, W_i^2 + \lambda\, C_1 = 0 \quad \text{za } i = 1, 2, \ldots, K.$$

Izrazimo $n_i$ in dobimo sistem enačb

$$n_i = \frac{W_k \sigma_i}{\sqrt{\lambda\, C_1}} \tag{1}$$

Da določimo $\lambda$, naredimo vsoto enačbe (1) po $k$, $k = 1, 2, \ldots, K$.

$$n = \frac{1}{\sqrt{\lambda\, C_1}} \sum_{k=1}^{K} W_k \sigma_k$$

$$\Rightarrow \sqrt{\lambda} = \frac{\sum_{k=1}^{K} W_k \sigma_k}{n\sqrt{C_1}} \tag{2}$$

Če združimo (1) in (2), dobimo:

$$n_i = n\, \frac{W_i \sigma_i}{\sum_{k=1}^{K} W_k \sigma_k}.$$

Rezultat je enak, kot če bi iskali minimalno varianco brez danega začetnega pogoja – ni pomembno, kakšne $n_1, n_2, \ldots, n_K$ izberemo. Stroški raziskave $C$

bodo vedno enaki, ker je cena $C_1$ vedno enaka.

Primer b

Sedaj se lahko stroški opažanja spreminjajo od stratuma do strauma. Če je $n_k$ število enot iz $k-$tega stratuma, ki so zajete v vzorec, naj bodo stroški raziskave enaki:

$$C = C_0 + \sum_{k=1}^{K} n_k C_k.$$

Pri danih sredstvih za raziskavo v višini $C$ poiščemo tiste velikosti podvzorcev, pri katerih je varianca cenilke populacijskega povprečja minimalna. Rešujemo podoben primer kot prej, le da sedaj nimamo več fiksne $C_1$, ampak se spreminja z vsakim stratumom. Z enakim razmislekom kot prej sestavimo Lagrangeovo funkcijo:

$$F(n_1, n_2, \dots, n_K, \lambda) = f(n_1, n_2, \dots, n_K) + \lambda(C_0 + \sum_{k=1}^{K} n_k C_k - C).$$

Zapišemo parcialne odvode funkcije $F$ in jih enačimo z 0.

$$\frac{\partial F}{\partial n_i} = -\frac{\sigma_i^2}{n_i^2}\, W_i^2 + \lambda\, C_i = 0 \quad \text{za } i = 1, 2, \dots, K.$$

Izrazimo $n_i$ in dobimo sistem enačb

$$n_i = \frac{W_i \sigma_i}{\sqrt{\lambda\, C_i}} \tag{3}$$

Da določimo $\lambda$, naredimo vsoto enačbe (3) po $k$, $k = 1, 2, \dots, K$.

$$n = \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}$$

$$\Rightarrow \sqrt{\lambda} = \frac{1}{n} \sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}} \tag{4}$$

Če združimo (3) in (4), dobimo:

$$n_i = \frac{n}{\sqrt{C_i}} \; \frac{W_i \sigma_i}{\sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}}.$$

Primer c

Naj se stroški raziskave izražajo na enak način kot v prejšnji točki, predpisano pa imamo natančnost raziskave, torej varianco cenilke. Poiščemo tiste vrednosti podvzorcev, pri katerih bodo stroški najmanjši.

Želimo torej minimizirati stroške pri pogoju, da je varianca minimalna, tj. $\text{var}()\overline{X}) = \sum_{k=1}^{K} \frac{W_k^2 \sigma_k^2}{n_k}$.

Rešujemo torej vezani ekstrem za funkcijo:

$$f(n_1, n_2, \ldots, n_K) = C_0 + \sum_{k=1}^{K} n_k C_k$$

$$\text{z vezjo } \sum_{k=1}^{K} \frac{W_k^2 \sigma_k^2}{n_k} - \text{var}(\overline{X}).$$

Sestavimo Lagrangeovo funkcijo:

$$F(n_1, n_2, \ldots, n_K, \lambda) = C_0 + \sum_{k=1}^{K} n_k C_k + \lambda \left( \sum_{k=1}^{K} \frac{W_k^2 \sigma_k^2}{n_k} - \text{var}(\overline{X}) \right)$$

Zapišemo parcialne odvode funkcije $F$ in jih enačimo z 0.

$$\frac{\partial F}{\partial n_i} = C_i - \lambda \, \frac{W_i^2 \sigma_i^2}{n_i^2} = 0 \quad \text{za } i = 1, 2, \ldots, K.$$

Izrazimo $n_i$ in dobimo sistem enačb

$$n_i = \sqrt{\frac{\lambda}{C_i}} \, W_i \sigma_i \tag{5}$$

Da določimo $\lambda$, naredimo vsoto enačbe (5) po $k$, $k = 1, 2, \ldots, K$.

$$n = \sqrt{\lambda} \, \sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}$$

$$\Rightarrow \sqrt{\lambda} = \frac{n}{\sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}} \tag{6}$$

Če združimo (5) in (6), dobimo:

$$n_i = \frac{n}{\sqrt{C_i}} \, \frac{W_i \sigma_i}{\sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}}.$$

## 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 $\alpha$.

PRIMER B

Ocenimo $\alpha$ 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 $\alpha$.

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

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 $\alpha$ in $\beta$ izberemo tako, da je srednja kvadratična napaka $\mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right]$ minimalna.
Uporabimo namig

$$\mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right] = \left[\mathrm{E}(Y) - \mathrm{E}(\hat{Y})\right]^2 + \mathrm{var}(Y - \hat{Y}). \tag{7}$$

Ker sta oba člena desne strani enačbe večja od 0, je dovolj, da poiščemo vrednosti $\alpha$ in $\beta$, 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 $\beta$. Za izračun minimuma enačbo najprej odvajamo po $\beta$ 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 $\alpha$ in $\beta$, 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 $\alpha$ in $\beta$ 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 $1 - \frac{\mathrm{var}(Y-\hat{Y})}{\mathrm{var}(Y)}$.