

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 priložena na koncu poročila pod naslovom *Dodatek A*, dostopna pa je tudi 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

Sklepi in izpeljave, ki jih bom uporabila pri vseh podnalogah, sledijo dokazu izreka v knjigi *John A. Rice: Mathematical Statistics and Data Analysis* (stran 232, 233).

Naredimo raziskavo na populaciji, ki ima K stratumov z velikostmi N_1, N_2, \dots, 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, \dots, n_K , pri katerih je varianca standardne cenilke populacijskega povprečja minimalna.

Če imamo stratumne velikosti N_1, N_2, \dots, N_K , se moramo odločiti, kako velike vzorce bomo izbrali iz posameznih stratumov. Izbiramo tako, da bo standardna napaka cenilke \bar{X} čim manjša. Poiščemo torej take n_1, n_2, \dots, n_K , kjer $n = n_1 + n_2 + \dots + n_K$, da bo

$$\text{var}(\bar{X}) = \sum_{k=1}^K W_k^2 \left(\frac{\sigma_k^2}{n_k} \right) \left(\frac{N_k - n_k}{N_k - 1} \right)$$

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:

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

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

Sestavimo Lagrangeovo funkcijo:

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

Zapišemo parcialne odvode funkcije F in jih enačimo z 0.

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

Izrazimo n_i in dobimo sistem enačb

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

Da določimo λ , naredimo vsoto enačbe (1) po i , $i = 1, 2, \dots, K$.

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

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, \dots, n_K izberemo. Stroški raziskave C bodo vedno enaki, ker je cena C_1 vedno enaka.

PRIMER B

Sedaj se lahko stroški opazanja spreminjajo od stratumu do stratumu. Če je n_k število enot iz k -tega stratumu, 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 λ , naredimo vsoto enačbe (3) po i , $i = 1, 2, \dots, K$.

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

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 $\text{var}(\bar{X}) = \sum_{k=1}^K \frac{W_k^2 \sigma_k^2}{n_k}$. Označimo $\text{var}(\bar{X}) = V$, kjer je V predpisana vrednost.

Rešujemo torej vezani ekstrem za funkcijo:

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

$$\text{z vezjo } \sum_{k=1}^K \frac{W_k^2 \sigma_k^2}{n_k} - V.$$

Sestavimo Lagrangeovo funkcijo:

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

Zapišemo parcialne odvode funkcije F in jih enačimo z 0.

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

Izrazimo n_i in dobimo sistem enačb

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

Da določimo λ , naredimo vsoto enačbe (5) po i , $i = 1, 2, \dots, K$.

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

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:

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

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

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

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.

$$E(X) = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{2},$$

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

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.

$$L_1(\alpha \mid x_1) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} x_1^{\alpha-1} (1-x_1)^{\alpha-1}$$

$$l_1(\alpha \mid x_1) = \ln(L_1(\alpha \mid x_1))$$

$$\Rightarrow l_1(\alpha \mid x_1) = \ln(\Gamma(2\alpha)) - 2\ln(\Gamma(\alpha)) + (\alpha-1)\ln(x_1) + (\alpha-1)\ln(1-x_1)$$

$$l(\alpha \mid x) = \sum_{i=1}^n l_1(\alpha \mid x_i)$$

$$\Rightarrow l(\alpha \mid x) = n \cdot \ln(\Gamma(2\alpha)) - 2n \cdot \ln(\Gamma(\alpha)) + (\alpha-1) \sum_{i=1}^n \ln(x_i) + (\alpha-1) \sum_{i=1}^n \ln(1-x_i)$$

Funkcijo $l(\alpha \mid x)$ odvajamo po α .

$$\frac{\partial l}{\partial \alpha} = n \cdot \frac{1}{\Gamma(2\alpha)} \cdot 2\Gamma'(2\alpha) - 2n \cdot \frac{1}{\Gamma(\alpha)} \cdot \Gamma'(\alpha) + \sum_{i=1}^n \ln(x_i) + \sum_{i=1}^n \ln(1-x_i)$$

Cenilka bo obstajala natanko tedaj, ko bo imela enačba

$$\frac{1}{\Gamma(2\alpha)} \cdot \Gamma'(2\alpha) - \frac{1}{\Gamma(\alpha)} \cdot \Gamma'(\alpha) = \frac{1}{2n} \sum_{i=1}^n \ln \left(\frac{1}{x_i(1-x_i)} \right)$$

rešitev.

PRIMER D

Poiščemo asimptotično varianco cenilke po metodi največjega verjetja.

Za asimptotično varianco bomo potrebovali Fischerjevo informacijo za cenilko, saj velja

$$\text{var}(\hat{\alpha}) \approx \frac{1}{n \cdot I_1(\hat{\alpha})},$$

kjer (iz predavanj)

$$I_1(\hat{\alpha}) = -E \left[\frac{\partial^2 l_1(\alpha | x_1)}{\partial \alpha^2} \right] \quad (7)$$

V $\frac{\partial l}{\partial \alpha}$ vstavimo $n = 1$ in izračunamo še drugi odvod. Dobimo:

$$\frac{\partial^2 l_1}{\partial \alpha^2} = \frac{4 \Gamma''(2\alpha) \Gamma(2\alpha) - 2\Gamma'(2\alpha)^2}{\Gamma(2\alpha)^2} - \frac{2 \Gamma''(\alpha) \Gamma(\alpha) - 2\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}$$

Pričakovana vrednost drugega odvoda je kar enaka drugemu odvodu, ker v njem ne nastopa slučajna spremenljivka X in je torej konstanta. Zato lahko takoj zapišemo varianco, ki je enaka

$$\text{var}(\hat{\alpha}) = \frac{1}{n} \cdot \frac{1}{\frac{2 \Gamma''(\alpha) \Gamma(\alpha) - 2\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} - \frac{4 \Gamma''(2\alpha) \Gamma(2\alpha) - 2\Gamma'(2\alpha)^2}{\Gamma(2\alpha)^2}}.$$

4. NALOGA

5. NALOGA

X in Y sta slučajni spremenljivki, za kateri velja:

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

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 $E \left[(Y - \hat{Y})^2 \right]$ minimalna.

Uporabimo namig

$$E \left[(Y - \hat{Y})^2 \right] = [E(Y) - E(\hat{Y})]^2 + \text{var}(Y - \hat{Y}). \quad (8)$$

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

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

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

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

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

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

To bo res, ko bo $\beta = \frac{\sigma_{x,y}}{\sigma_x^2}$.

Vrednosti α in β , pri katerih bo izraz $E \left[(Y - \hat{Y})^2 \right]$ minimalen, sta

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

PRIMER B

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

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

Spomnimo se najprej formule za *korelacijski koeficient*:

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

Ker je determinacijski koeficient enak kvadratu korelacije, je enak

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

Dobimo enakost

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

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

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

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