Seminarska naloga iz statistike

Klementina Pirc

Fakulteta za matematiko in fiziko
Oddelek za matematiko

julij 2020

Podrobnejša navodila nalog se nahajajo v datoteki *27_sem_nal_Klementina_Pirc.pdf*

1. naloga

V datoteki *Kibergrad.csv* so podane informacije o 43 886 družinah. Spodnje naloge se navezujejo na stopnjo izobrazbe, ki jo imajo vodje gospodinjstev. Stopnje so označene s števili od 31 do 46. Naloge sem rešila s pomočjo programa Python in ustreznih knjižnic, postopek pa se nahaja v datoteki *Kibergrad.py*

a)

Na podlagi enostavnega slučajnega vzorca 200 družin poiščemo oceno za delež družin v celotni populaciji, katerih vodja gospodinjstva nima srednješolske izobrazbe. S pomočjo opisov stopenj izobrazbe ugotovimo, da moramo prešteti družine iz vzorca s stopnjo <= 38. Naj bo d_i stopnja izobrazbe za i-to družino, definiramo

$$x_i = \begin{cases} 1 & \text{če } d_i \leq 38 \\ 0 & \text{sicer} \end{cases}$$

Sedaj lahko vzorčni delež s izračunamo s formulo

$$s = \frac{1}{n} \sum_{i=1}^{n} x_i$$

kjer je n velikost vzorca, torej 200. Dobimo rezultat $s = 0.23$.

b)

Ocenimo standardno napako $se(s)$ in določimo 95% interval zaupanja. Za izračun standardne napake moramo izračunati varianco s. Ker smo vzeli enostavni slučajno vzorec velja

$$var(s) = \frac{1}{n} \frac{N-n}{N-1} \tilde{\sigma}^2$$

kjer je $N = 43886$ velikost populacije in $\tilde{\sigma}^2$ nepristranska cenilka za populacijsko varianco. Torej

$$\begin{aligned}
\tilde{\sigma}^2 &= \frac{N-1}{N(n-1)} \sum_{i=1}^{n} (x_i - s)^2 \\
\Rightarrow var(s) &= \frac{N-n}{nN(n-1)} \sum_{i=1}^{n} (x_i - s)^2 \\
&= \frac{N-n}{nN(n-1)} \sum_{i=1}^{n} (x_i^2 - 2x_i s + s^2) \\
&= \frac{N-n}{nN(n-1)} \sum_{i=1}^{n} (x_i - 2x_i s + s^2) \\
&= \frac{N-n}{N(n-1)} (s - s^2)
\end{aligned}$$

in zato

$$se(s) = \sqrt{\frac{N-n}{N(n-1)} s(1-s)}$$

Upoštevali smo $x_i^2 = x_i$, kar velja, ker $x_i \in \{0, 1\}$. Standardna napaka za izbrani vzorec je 0.029763.

Interval zaupanja dobimo po formuli $s \mp z_\alpha se(s)$. Ker želimo 95% natančnost, je $z_\alpha = 1.96$ in tako dobimo interval $[0.171662, 0.288337]$.

c)

Izračunamo populacijski delež S in standardno napako $se(S)$ s formulama

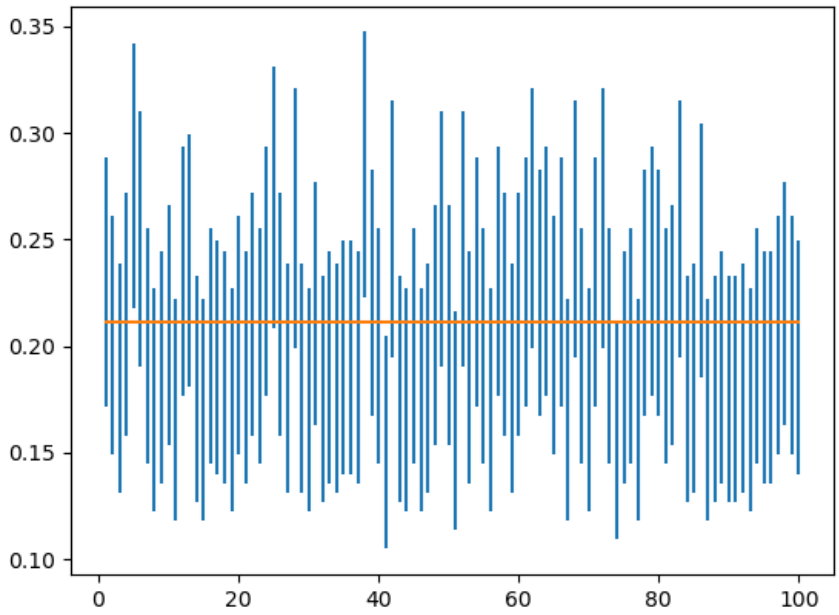
$$S = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad se(S) = \sqrt{\frac{\sigma^2}{N}} = \sqrt{\frac{1}{N^2} \sum_{i=1}^{N} (x_i - S)^2}$$

ter dobimo $S = 0.211502$ in $se(S) = 0.001949$. Interval zaupanja torej pokrije populacijski delež, saj $S \in [0.171662, 0.288337]$.

Razlika med vzorčnim in populacijskim deležem znaša 0.018497, razlika med vzorčno in populacijsko standardno napako pa 0.027814.

d)

Izberemo še 99 vzorcev, določimo 95% intervale zaupanja, ter jih narišemo skupaj z intervalom za prvi vzorec. S horizontalno črto označimo vrednost S. Vidimo, da populacijski delež pokrije 96 intervalov.



e)

Izračunamo standardni odklon vzorčnih deležev iz 100 prej izbranih vzorcev. Uporabimo formulo

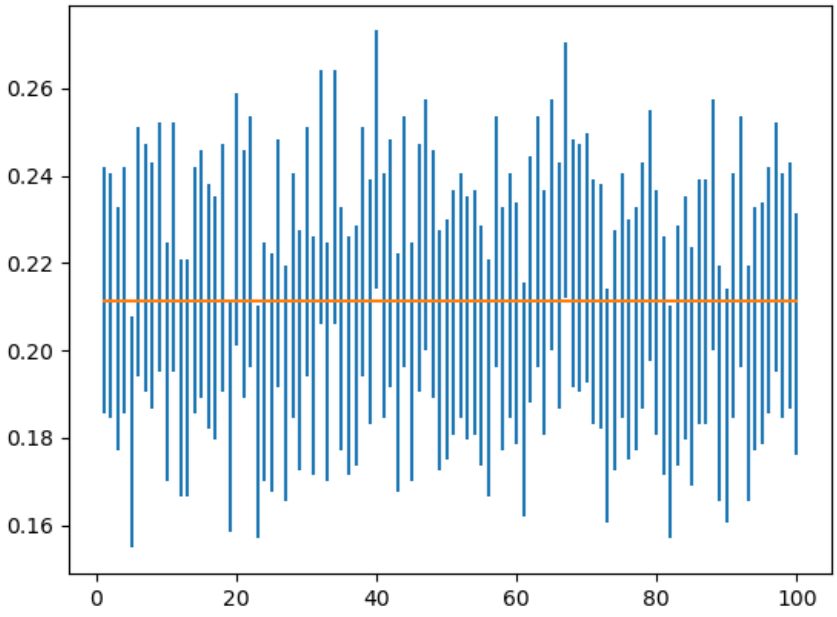
$$\sigma_s = \sqrt{\frac{1}{m}\sum_{i=1} m(s_i - \bar{s})^2}$$

kjer je $m = 100$ število vzorcev, s_i vzorčni delež i-tega vzorca in $\bar{s}$ povprečje vzorčnih deležev. $\sigma_s = 0.028618$, razlika med populacijsko standardno napako $se(S)$ in σ_s pa je 0.026668.

f)

Sedaj izberemo 100 vzorcev velikosti 800 in ponovimo postopke iz točk d) in e).

- $\sigma_s = 0.013002$
- $|\sigma_s - se(S)| = 0.011053$
- populacijski delež S pokrije 95 intervalov zaupanja



2. naloga

Želimo oceniti skupno vrednost inventarja z N enotami, kjer ima vsaka knjigovodsko vrednost X in dejansko vrednost Y. Poznamo X za vse enote in želimo oceniti vsoto dejanskih vrednosti vseh enot τ_Y.

$$\tau_Y = \sum_{i=1}^{N} Y_i$$

Vzamemo enostavni slučajni vzorec z n enotami in predlagamo 3 možne cenilke

$$\hat{\tau}_Y^{(0)} := N\bar{Y} \qquad \hat{\tau}_Y^{(1)} := \tau_X + N(\bar{Y} - \bar{X}) \qquad \hat{\tau}_Y^{(2)} := \frac{\bar{Y}}{\bar{X}}\tau_X$$

a)

Dokažimo, da je $\hat{\tau_Y}^{(1)}$ nepristranska: $E(\hat{\tau}_Y^{(1)}) = \tau_Y$ in izračunajmo njeno varianco. Uporabili bomo naslednje oznake.

τ_X, τ_Y ... vsota vrednosti spremenljivke X oz. Y na celotnem inventarju,
μ_X, μ_Y ... povprečna vrednost X oz. Y na celotnem inventarju,
σ_X, σ_Y ... standardni odklon X oz. Y na vseh enotah inventarja,
ρ_{XY} ... korelacijski koeficient med X in Y na celotnem inventarju,
$\bar{X}$, $\bar{Y}$ ... vzorčno povprečje X oz. Y.

$$\begin{aligned} E(\hat{\tau}_Y^{(1)}) &= E(\tau_X + N(\bar{Y} - \bar{X})) \\ &= \tau_X + N(E(\bar{Y}) - E(\bar{X})) \\ &= \tau_X + N\mu_Y - N\mu_X \\ &= \tau_X + \tau_Y - \tau_X \\ &= \tau_Y \end{aligned}$$

$$\begin{aligned} var(\hat{\tau}_Y^{(1)}) &= var(\tau_X + N(\bar{Y} - \bar{X})) \\ &= N^2 var(\bar{Y} - \bar{X}) \\ &= N^2 var(\bar{Y} + \bar{X} - 2cov(\bar{Y}, \bar{X})) \\ &= N^2 (\frac{\sigma_Y^2}{n}\frac{N-n}{N-1} + \frac{\sigma_X^2}{n}\frac{N-n}{N-1} - 2\rho_{XY}) \\ &= \frac{N^2}{n}\frac{N-n}{N-1}(\sigma_Y^2 + \sigma_X^2) - 2N^2\rho_{XY} \end{aligned}$$

b)

Privzamemo, da sta X in $Y - X$ na populacijii neodvisni. Zanima nas, kdaj ima cenilka $\hat{\tau}_Y^{(1)}$ nižjo varianco kot cenilka $\hat{\tau}_Y^{(0)}$.

$$\begin{aligned} var(\hat{\tau}_Y^{(0)}) &= var(N\bar{Y}) \\ &= \frac{N^2}{n}\frac{N-n}{N-1}\sigma_Y^2 \end{aligned}$$

$$
\begin{aligned}
var(\hat{\tau}_Y^{(1)}) &< var(\hat{\tau}_Y^{(0)}) \\
\frac{N^2}{n}\frac{N-n}{N-1}(\sigma_Y^2+\sigma_X^2) - 2N^2\rho_{XY} &< \frac{N^2}{n}\frac{N-n}{N-1}\sigma_Y^2 \\
\frac{N^2}{n}\frac{N-n}{N-1}\sigma_X^2 - 2N^2\rho_{XY} &< 0 \\
\frac{N^2}{n}\frac{N-n}{N-1}\sigma_X^2 - 2\frac{N^2}{n}\frac{N-n}{N-1}\sigma_X^2 &< 0 \\
-\frac{N^2}{n}\frac{N-n}{N-1}\sigma_X^2 &< 0
\end{aligned}
$$

To je očitno res, saj je $\sigma_X^2 > 0$ in zato $var(\hat{\tau}_Y^{(1)}) < var(\hat{\tau}_Y^{(0)})$ velja vedno. Zgoraj smo upoštevali še

$$
\begin{aligned}
cov(\bar{X},\bar{Y}) &= cov(\bar{X},\bar{Y}-\bar{X}+\bar{X}) \\
&= cov(\bar{X},\bar{Y}-\bar{X}) + cov(\bar{X},\bar{X}) \\
&= 0 + var(\bar{X}) \\
&= \frac{\sigma_X^2}{n}\frac{N-n}{N-1}
\end{aligned}
$$

3. naloga

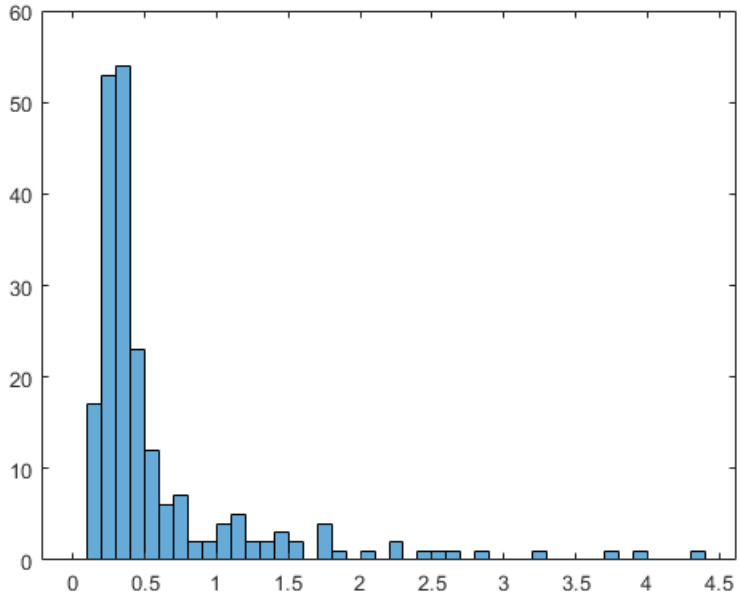
V datoteki *Kiti.csv* so podani časi $t_1, \ldots, t_{210}$ za 210 kitov. i-ti kit v času t_i preplava 1 km. Nalogo sem reševala s pomočjo programa Matlab, postopek pa se nahaja v datoteki *Kiti.m*.

a)

Narišemo histogram za dane podatke s pomočjo vgrajene funkcije *histogram*. Širino razredov določimo glede na navodila in dobimo

$$
l = \frac{2(q_{3/4} - q_{1/4})}{\sqrt[3]{n}} = 0.1141 \doteq 0.1
$$

kjer sta $q_{1/4}$, $q_{3/4}$ prvi in tretji kvartil, n pa število podatkov.



Glede na dobljeni histogram, se porazdelitev gama zdi ustrezna izbira za statistični model.

b)

Radi bi ocenili parametra λ in α porazdelitve γ po metodi momentov. Pomagamo si s poglavjem 8.4 v knjigi. Na strani 260 izvemo, da je k-ti moment enak

$$\mu_k = E(t^k) \quad k \in \mathbb{N}$$

kjer je t slučajna spremenljivka. Oceno za μ_k pa lahko dobimo s pomočjo podanih podatkov kot

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} t_i^k$$

Ker imamo dva neznana parametra bomo potrebovali μ_1 in μ_2 za γ porazdelitev. Najdemo ju na strani 263.

$$\mu_1 = E(t) = \frac{\alpha}{\lambda} \qquad \mu_2 = E(t^2) = \frac{\alpha(\alpha+1)}{\lambda^2}$$

Enačbi preoblikujemo in dobimo predpisa za λ in α

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2} \qquad \alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$$

Z ocenami za μ_1 in μ_2 dobimo oceni $\hat{\lambda} = 1.3188$ in $\hat{\alpha} = 0.7992$.

c)

Ocenimo sedaj λ in α še po metodi največjega verjetja. Logaritemska funkcija verjetja je

$$\begin{aligned}\ell(\lambda, \alpha | t_1, \ldots, t_n) &= \sum_{i=1}^{n} \log \gamma(t_i) \\ &= \sum_{i=1}^{n} (\alpha \log \lambda + (\alpha - 1) \log t_i - \lambda t_i - \log \Gamma(\alpha))\end{aligned}$$

Odvajamo ℓ in enačimo z 0, da dobimo cenilki za λ in α

$$\begin{aligned}&\frac{\partial \ell}{\partial \lambda} = \frac{n\alpha}{\lambda} + 0 - \sum_{i=1}^{n} t_i - 0 = 0 \\ \Rightarrow \qquad &\hat{\lambda} = \frac{n\alpha}{\sum_{i=1}^{n} t_i}\end{aligned}$$

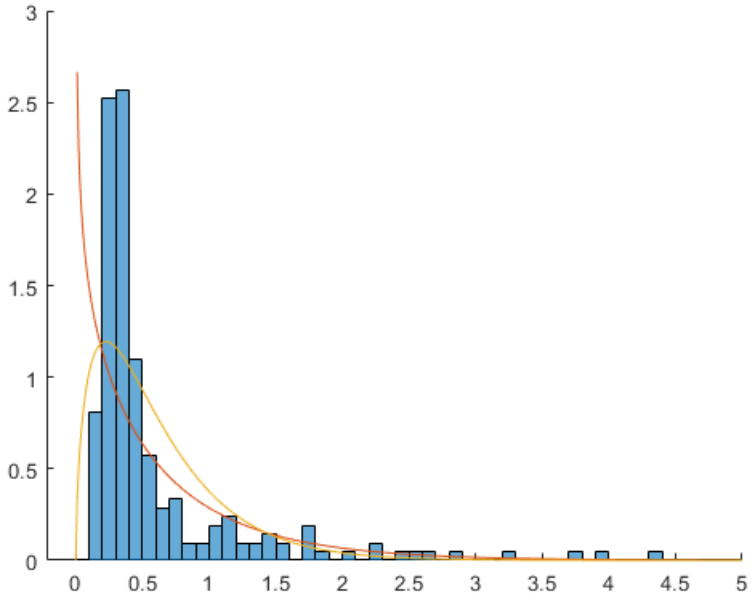
$$\begin{aligned}&\frac{\partial \ell}{\partial \alpha} = n \log \lambda + \sum_{i=1}^{n} \log t_i - 0 - n \frac{\partial \log \Gamma(\alpha)}{\partial \alpha} = 0 \\ \Rightarrow \qquad &0 = n \log \frac{n\alpha}{\sum_{i=1}^{n} t_i} + \sum_{i=1}^{n} \log t_i - n\psi(\alpha)\end{aligned}$$

kjer je ψ funkcija digama, ki je logaritemski odvod funkcije gama. Enačbo za cenilko α rešimo z vgrajeno funkcijo *fzero*, ki sprejme funkcijo, začetni približek in vrne približek za ničlo funkcije, ki je v našem primeru ravno cenilka za α. $\hat{\alpha}$ nato vstavimo v enačbo za $\hat{\lambda}$. Dobimo $\hat{\lambda} = 2.6327$ in $\hat{\alpha} = 1.5954$.

Oceni parametrov se glede na metodo ocenjevanja precej razlikujeta. Navadno je metoda največjega verjetja bolj točna.

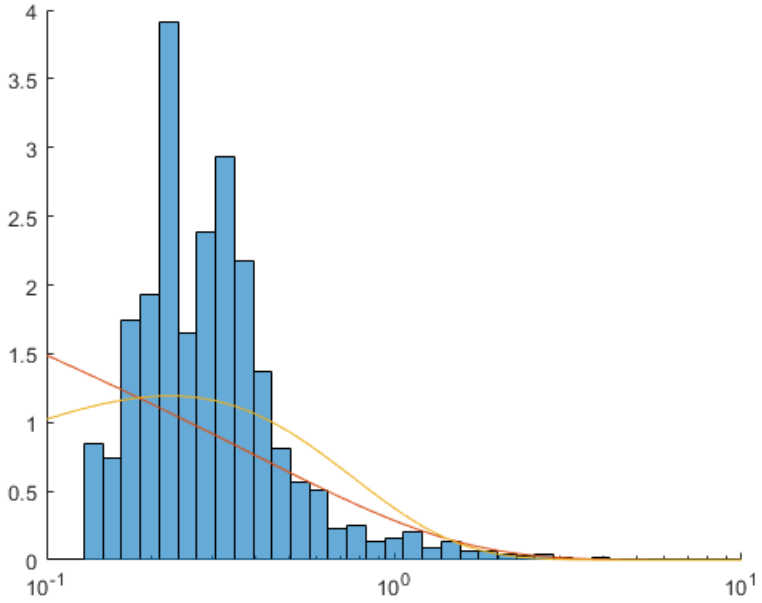
d)

Na histogram dorišemo porazdelitvi γ z dobljenimi ocenami za λ in α. Rdeč graf nam da metoda momentov, rumenega pa metoda največjega verjetja. Porazdelitvi se podatkom ne prilegata najbolje.



e)

Histogram in porazdelitvi narišemo še na logaritemski lestvici. Prileganje porazdelitev podatkom ni bistveno boljše kot prej.



Literatura

[1] E. Zakrajšek, *Verižnica* (1999) od 6 do 10. Dostopno na spletni učilnici FMF 2019/2020 predmeta Matematično modeliranje [22. 7. 2020]

[2] Zapiski s predavanj predmeta Matematično modeliranje.