Projektna naloga pri Statistiki

Jimmy Zakeršnik

17.7.2022

Povzetek

V tej nalogi pri predmetu Statistika, so obravnavane tri naloge, vsaka v svojem lastnem poglavju. V prvem poglavju se obravnava dohodke družin mesta Kibergrad s pomočjo enostavnega slučajnega vzorčenja, škatel z brki ter analize s tipom družine pojasnjene variance dohodka. V drugem poglavju se s pomočjo Rayleighove porazdelitve $Rayleigh(\theta)$ obravnavajo izmerjene dolžine kromatina. Za parameter θ se določita cenilki po metodi največjega verjetja ter po metodi momentov. Obe cenilki sta tudi primerjani glede na asimptotsko MSE in nepristranskost, nato pa se tudi izračuna njuno numerično vrednost na podlagi priloženih podatkov in primerja prileganje pripadajočih verjetij s histogrami podatkov. V zadnjem poglavju, se obravnavata regresijska modela A ter B. Model A je zavrnjen znotraj B pri stopnji tveganja 0,05, sprejet pa pri stopnji tveganja 0,01. S pomočjo Akaikejeve informacije se pokaže, da je model B bolj optimalen za obravnavo povprečnih mesečnih temperatur, kot pa model A.

Kazalo

1 Kibergrad **5**
- 1.1 Primerjava dohodkov med tipi družin prvega vzorca 5
- 1.2 Primerjava dohodkov družin tipa 1 v petih vzorcih 7
- 1.3 S tipom pojasnjena varianca populacije 9

2 Slučajni sprehod **10**
- 2.1 Cenilka za θ po metodi največjega verjetja 11
- 2.2 Cenilka za θ po metodi momentov 12
- 2.3 Asimptotični MSE cenilk 12
  - 2.3.1 MSE cenilke po metodi največjega verjetja 13
  - 2.3.2 MSE cenilke po metodi momentov 14
- 2.4 Ševilska ocena prve cenilke 14
- 2.5 Številska ocena druge cenilke 16
- 2.6 Histogram meritev in grafi gostot cenilk 18

3 Temperature **20**
- 3.1 Preizkus modela A znotraj modela B 21
  - 3.1.1 Cenilki β in $\bar{\beta}$ po Metodi najmanjših kvadratov 22
  - 3.1.2 Residuali in RSS obeh modelov 22
  - 3.1.3 Preizkus modela A v B s stopnjama tveganja 0,01 in 0,05 23
- 3.2 Akaikejeva informacija modelov 23

Slike

1 Vzporedno narisane škatle z brki tipov družin 1, 2 in 3 6
2 Vzporedno narisane škatle z brki družin tipa 1 po vzorcih 7
3 Popravljene vzporedno narisane škatle z brki družin tipa 1 po vzorcih 8
4 Primerjalni kvartilni grafikon dohodkov družin tipa 1 v vzorcu 1 9
5 Verjetje za oceno cenilke θ, pridobljene po metodi največjega verjetja, v prvem poskusu 15
6 Verjetje za oceno cenilke θ, pridobljene po metodi največjega verjetja, v drugem poskusu 15
7 Verjetje za oceno cenilke θ, pridobljene po metodi največjega verjetja, v tretjem poskusu 16
8 Verjetje za oceno cenilke θ, pridobljene po metodi momentov, v prvem poskusu 17
9 Verjetje za oceno cenilke θ, pridobljene po metodi momentov, v drugem poskusu 17
10 Verjetje za oceno cenilke θ, pridobljene po metodi momentov, v tretjem poskusu 17
11 Histogram meritev v prvem poskusu skupaj z gostotama za pripadajoči cenilki. 18
12 Histogram meritev v drugem poskusu skupaj z gostotama za pripadajoči cenilki. 19
13 Histogram meritev v tretjem poskusu skupaj z gostotama za pripadajoči cenilki. 19

14 Razpršen graf povprečnih mesečnih temperatur glede na mesec. . 20
15 Primerjava prileganja modelov A in B s podatki iz *Temp_LJ.csv* 24

Tabele

1 Tabela vrednosti, ki so potrebne za risanje škatel z brki za vsak tip družine 5
2 Povprečne vrednosti in standardni odkloni dohodkov po tipih . . 5
3 Interkvartilni razmiki dohodkov po tipih 6
4 Ekstremi, kvartili, povprečja, standardni odkloni in interkvartilni razmiki dohodkov družin tipa 1 po vzorcih 8
5 Populacijska, s tipi pojasnjena in nepojasnjena varianca 10
6 Numerične vrednosti MNV cenilke in standardne napake po eksperimentih 14
7 Numerične vrednosti cenilke pridobljene po metodi momentov in standardne napake po eksperimentih 16
8 Tabela vrednosti RSS za modela A in B 23
9 Tabela vrednosti $F^{-1}_{Fisher(9,407)}(1-\alpha)$ za modela $\alpha = 0{,}01$ in $\alpha = 0{,}05$ 23
10 Tabela vrednosti AIC modelov A in B 23

1 Kibergrad

Priložena datoteka *Kibergrad.csv*, ki vsebuje podatke o dani populaciji (prebivalci mesta Kibergrad), je bila odprta v programu LibreOffice Calc. S pomočjo vgrajenega orodja so nato, bili zbrani vzorci velikosti 500 po postopku enostavnega vzorčenja. V priloženi datoteki *Kibergradwork.ods* so na prvi strani izpisani vsi podatki ter vseh pet pridobljenih vzorcev, na drugi strani je posebej obravnavan prvi vzorec v smislu kvartilov in extremnih vrednosti, na tretji strani pa se na enak način obravnavajo vsi vzorci glede na dohodek družin tipa 1. Obe primerjavi sta dodatno podprti s pomočjo škatel z brki in na koncu sta izračunani še s tipi pojasnjena varianca in nepojasnjena varianca dohodkov. Pri tem je v veliko pomoč programski jezik **R**.

1.1 Primerjava dohodkov med tipi družin prvega vzorca

Poglejmo si najprej prvi vzorec in primerjajmo dohodke družin glede na tip družine. Za pravilno delovanje vsake od skript, ki so omenjene v tej nalogi, je ključno to, da je delovno okolje pravilno nastavljeno na mapo *Podatki_in_skripte*. Za delovno okolje torej nastavimo to mapo ter odpremo in poženemo skripto *Kibergrad_a.R*, napisano v jeziku **R**. Ob pogonu se v konzoli izpišejo vrednosti o dohodku, ki so navedene v spodnji tabeli.

	Tip 1	Tip 2	Tip 3
Max	228727	181696	168926
Q3	55360	54859	57631
Med	32975	38883	33310
Q1	18700	22011	16071
Min	0	5184	0

Tabela 1: *Tabela vrednosti, ki so potrebne za risanje škatel z brki za vsak tip družine*

Istočasno skripta izriše vzporedne škatle z brki, kot lahko vidimo na spodnji sliki, s pomočjo katerih lahko grafično primerjamo dohodke družin različnih tipov.

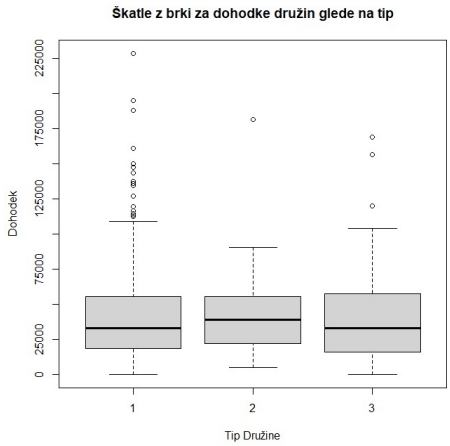
Škatle z brki nam ponudijo nekaj zanimivih ugotovitev. V minimalnih dohodkih ni velikih odstopanj, razen pri tipu 2, ki ima za razliko od ostalih pozitiven minimalen dohodek. Če ignoriramo ostale vzorce in tem rezultatom naivno verjamemo, so enostarševske družine z očetom (torej družine tipa 2) relativno bolj premožne od ostalih tipov. To domnevo podpira tudi opazka, da je povprečna vrednost tipa 2 višja kot povprečni vrednosti ostalih dveh tipov, kar lahko preberemo iz izpisa na konzoli.

Vrednosti so hkrati tudi dostopne v spodnji tabeli.

Tip	1	2	3
Povprečje	41655	46301	39931
SD	32974,23	39500,3	31069,71

Tabela 2: *Povprečne vrednosti in standardni odkloni dohodkov po tipih*

Če ne bi imeli že izračunanih povprečij, bi lahko še vedno sklepali o njihovih velikostih s pomoćjo škatel z brki. Opazimo namreč, da se prvi kvartili nahajajo



Slika 1: *Vzporedno narisane škatle z brki tipov družin* 1, 2 *in* 3

na približno enaki višini z maksimalno razliko v okolici 6000 v prid družinam tipa 2, kar velja tudi za mediane. Šele pri tretjem kvartilu ostala dva tipa premagata tip 2, a tudi tu je največja razlika v rangu 4000. V tem primeru tipu 2 pomaga to, da ima izmed vseh tipov družin najmanjšo razdaljo med tretjim kvartilom in mediano, kar pomeni, da so vrednosti, ki pripadajo temu intervalu, bolj gosto porazdeljene.

Višje povprečje dohodka družin tipa 2 ni edino, kar izstopa pri škatlah z brki. Družine tipa 1 izstopajo v tem, da imajo, v primerjavi z ostalimi tipi, veliko število osamelcev (torej tistih vrednosti, ki so na sliki označene s krogi izven škatel).

Družine tipa 3 se odlikujejo po tem, da imajo najširši interkvartilni razmik. Za lažji pregled in primerjavo, so vsi interkvartilni razmiki navedeni v konzoli (po pogonu skripte) ter v tabeli spodaj:

Tip	1	2	3
IQR	36660	32848	41560

Tabela 3: *Interkvartilni razmiki dohodkov po tipih*

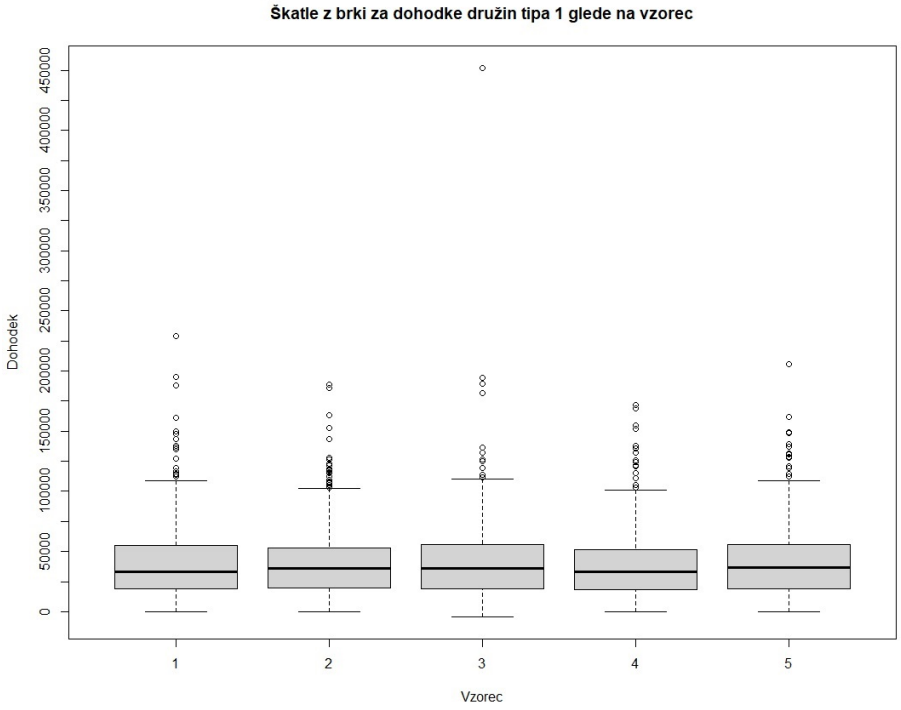
Družine tipa 2 so torej v povprečju bolj premožne od družin ostalih tipov in družine tipa 3 so v povprečju najmanj premožne. Vrednosti družin tipa 3 so hkrati tudi najbolj razpršene v »srednji polovici«, kar nam pove velikost interkvartilnega razmika. Družine tipa 1 se v obeh primerih nahajajo v sredini med družinami tipa 2 in 3. Hkrati imajo tudi razmeroma več osamelcev od ostalih tipov. V vsakem primeru nam rezultati namigujejo, da obstaja povezava med tipom družine in njenim dohodkom.

1.2 Primerjava dohodkov družin tipa 1 v petih vzorcih

Da nadaljujemo analizo podatkov si oglejmo porazdelitev dohodkov družin nekega tipa preko 5 neodvisno izbranih vzorcev. Pri tem za prvega vzamemo kar vzorec, ki smo ga obravnavali v prejšnjem podpoglavju, ostale štiri pa pridobimo s pomočjo orodij v LibreOffice Calc. Vsi vzorci so posebej shranjeni v lastni datoteki tipa *.csv* z imeni tipa *KibergradVzorec#.csv*.

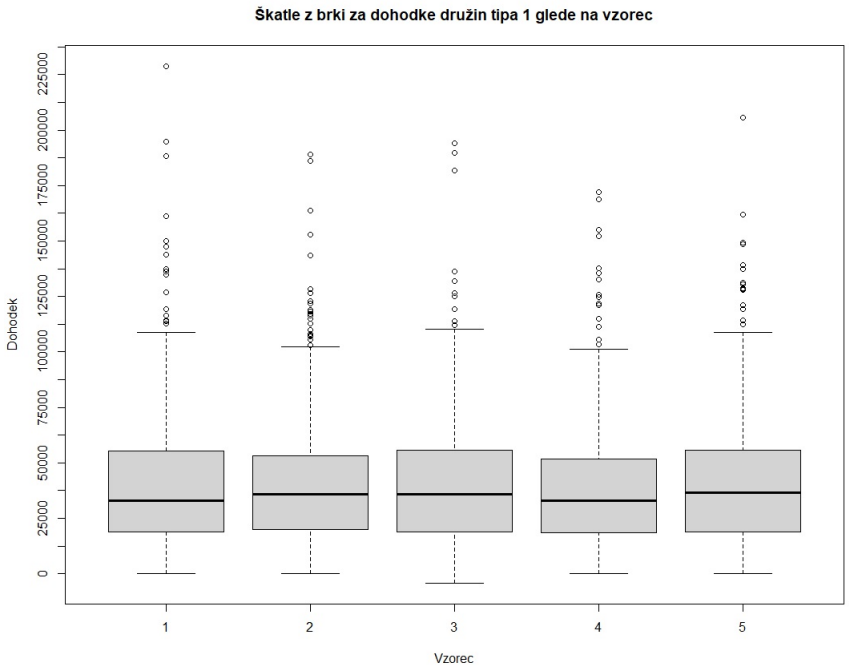
Da izrišemo škatle z brki, s pomočjo katerih bomo primerjali dohodke preko vzorcev, poženemo skripto *Kibergrad_b.R*. V njej najprej naložimo vzorce, nato vsakemu vzorcu dodamo stolpec vrednosti, ki nam pove kateremu vzorcu pripada dan podatek. Torej vzorcu 1 dodamo stolpec samih enk, vzorcu 2 stolpec samih dvojk itd. Te tabele nato združimo v eno samo tabelo, iz nje prefiltriramo družine vseh tipov razen 1 in nato s pomočjo te nove tabele po enakem postopku kot v prejšnjem podpoglavju primerjamo dohodke družin glede na vzorec.

Dobljene škatle z brki so prikazane spodaj.



Slika 2: *Vzporedno narisane škatle z brki družin tipa* 1 *po vzorcih*

Takoj opazimo, da je prikazan graf razpotegnjen, v glavnem na račun enega osamelca iz tretjega vzorca. Da dobimo bolj pregleden graf, odstranimo vse vrednosti, ki so večje od 250000 (v resnici je taka zgolj ena). Škatle z brki, ki jih dobimo po tem popravku in so prikazane spodaj, skripta samostojno izriše.



Slika 3: *Popravljene vzporedno narisane škatle z brki družin tipa* 1 *po vzorcih*

Ker smo pri »popravku« zanemarili zgolj eno vrednost, nam to bistveno ne pokvari primerjave. Izoliran osamelec bi lahko na našo obravnavo vplival kvečjemu negativno. Zato bomo pri izračunu povprečja za vsak vzorec uporabili tabelo, ki tega osamelca ne vsebuje.

Vrednosti (kvartili, povprečja, maksimalna in minimalna vrednost, IQR), ki jih skripta izpiše v konzolo, so prikazane v spodnji tabeli:

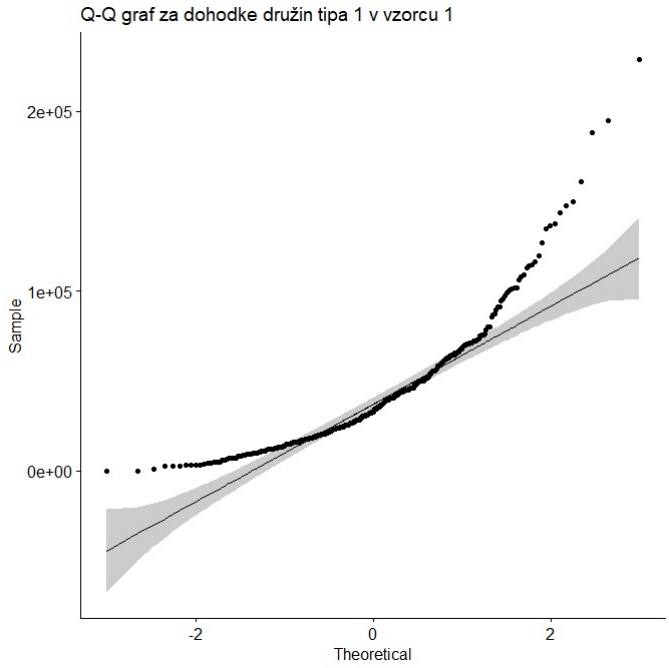
	Vzorec 1	Vzorec 2	Vzorec 3	Vzorec 4	Vzorec 5
Max	228727	188899	194230	171999	205712
Q3	55360	53121	55700	51700	55863
Med	32975	36000	35850	33006	36527
Q1	18700	20008	18800	18327	18668
Min	0	0	−4198	0	0
Povprečje	41655	41822	41337	39344	41697
SD	32974,23	31037,97	30517,25	29867,4	31099,04
IQR	36660	33113	36900	33373	37195

Tabela 4: *Ekstremi, kvartili, povprečja, standardni odkloni in interkvartilni razmiki dohodkov družin tipa* 1 *po vzorcih*

Sedaj, ko imamo narisane škatle z brki in zraven napisano tabelo, lahko komentiramo rezultate. V prvi vrsti opazimo, da so si povprečja dokaj blizu. Vsa povprečja razen povprečje četrtega vzorca se nahajajo v okolici 41500 ± 500, povprečje vzorca 4 pa se od 41500 razlikuje za manj kot 2500. Če bi si izbrali še več vzorcev, bi se po vsej verjetnosti njihova povprečja tudi nahajala v neki

bližnji okolici 41500. Podobno obnašanje standardnih odklonov, ki se nabirajo v okolici 31000 ± 2000 nas privede do nepresenetljivega sklepa, da je porazdelitev dohodka družin tipa 1 neodvisna od vzorca. To potrjuje tudi relativna bližina kvartilov v tabeli (npr. tretji kvartili se zbirajo v okolici 53000 ± 3000).

Na tej točki bi želeli preveriti, ali je porazdelitev slučajno normalna. Test s primerjalnim kvartilnim grafikonom na vzorcu 1 nam pove, da to ne drži. Enak sklep seveda velja tudi za ostale vzorce. Zakomentiran ukaz za izris spodnjega grafikona se nahaja na koncu skripte *Kibergrad_b.R*. Ukaz je zakomentiran zato, da program raje izriše grafikon škatel z brki, na katerih je povdarek te naloge.



Slika 4: *Primerjalni kvartilni grafikon dohodkov družin tipa* 1 *v vzorcu 1*

Četudi dohodki niso porazdeljeni normalno, so še vedno razmeroma konsistentni preko obravnavanih vzorcev. To je v kontrastu z razlikami in odstopanji, ki smo jih opazili, ko smo v prvem vzorcu primerjali dohodke glede na tip družine. Ta kontrast dodatno potrjuje sklep, da tip družine netrivialno vpliva na dohodek družine.

1.3 S tipom pojasnjena varianca populacije

Na koncu prejšnjega podpoglavja smo prišli do sklepa, da ima tip družine netrivialen vpliv na dohodek družine. Če to drži ali ne, lahko preverimo z izračunom s tipom družine pojesnjene variance. Čim smo poračunali to, se lahko skličemo na zvezo med varianco in pojasnjeno ter nepojasnjeno varianco ($Celotna\_varianca = Pojasnjena\_varianca + Nepojasnjena\_varianca$) in poračunamo še slednjo. Vsi računi in primerjave se prikažejo ob pogonu skripte *Kibergrad_c.R*.

Z n_i označimo število družin tipa i ter z N velikost naše populacije. Z X

označimo dohodke družin, z Y pa slučajno spremenljivko tipov družin, ki ima porazdelitev $P(Y = i) = n_i/N$. Predpostavimo, da so dohodki tipov družin $X_i = X_{|Y=i}$ medseboj neodvisni. Predpostavko upravičimo z argumentom, da v splošnem dohodek soseda ne vpliva na naš dohodek. Pomnimo tudi, da je $E[X|Y]$ neka funkcija spremenljivke Y, recimo $\Phi(Y)$. Z $\bar{X}_i$ še označimo pričakovano vrednost dohodka v družini tipa i, torej $\bar{X}_i = E[X|Y = i]$. S tipom pojasnjeno varianco potem izračunamo po formuli:

$$\begin{aligned}
&Var(E[X|Y]) = Var(\Phi(Y)) = E\left[(\Phi(Y) - E[\Phi(Y)])^2\right] = \\
&= \sum_{i=1}^{3} (\Phi(Y = i) - E[\Phi(Y)])^2 * P(Y = i) = \\
&= 1/N * \sum_{i=1}^{3} n_i * (E[X|Y = i] - E[E[X|Y]])^2 = \\
&= 1/N * \sum_{i=1}^{3} n_i * (\bar{X}_i - E[X])^2 = 1/N * \sum_{i=1}^{3} n_i * (\bar{X}_i - \bar{X})^2
\end{aligned}$$

S pomočjo zgoraj pridobljene formule v *Kibergrad_c.R* poračunamo pojasnjeno varianco. Nepojasnjeno varianco nato poračunamo kot razliko populacijske variance in pojasnjene variance. Vrednosti skripta izpiše v konzolo, dostopne pa so tudi v spodnji tabeli.

Varianca	1026385670
Pojasnjena	113781162
Nepojasnjena	912604508
SD	32037,2544062437

Tabela 5: *Populacijska, s tipi pojasnjena in nepojasnjena varianca*

Opazimo, da je nepojasnjena varianca bistveno višja od pojasnjene variance. Če pogledamo delež, ki ga varianci zavzemata, nam s tipi družin pojasnjena varianca predstavlja le približno 11,09%. To nam pove, da je tip družine netrivialen faktor pri napovedi dohodka družine, ni pa glavni faktor. To se ujema s tem, kar smo razbrali v prejšnjih podpoglavjih. Že zgolj za družine tipa 1 v podpoglavju 1.2 je bil standardni odklon, torej koren variance, razmeroma visok v vsakem vzorcu. To se ujema s tem, da večino variance pridobimo od faktorjev, ki niso tip družine. Če bi tip družine bil odgovoren za večji delež celotne variance dohodkov, bi bila varianca znotaj tipov manjša.

2 Slučajni sprehod

Za začetek omenimo, da se podatki, ki so uporabljeni v tem delu naloge, vsebovani v datotekah *Kromatin_kratki.csv*, *Kromatin_srednji.csv* in *Kromatin_dolgi.csv*. Ti podatki so razdalje med pari zaporedij nukleotidov, ki so bile izmerjene v treh različnih eksperimentih. Spomnimo se tudi, da imajo te razdalje Rayleighovo porazdelitev, ki je podana z gostoto

$$f(r|\theta) = \begin{cases} \frac{r}{\theta^2} \exp\left(-\frac{r^2}{2\theta^2}\right) \ ; \ r > 0 \\ 0 \ ; \ \text{sicer} \end{cases}$$

V prvih dveh podpoglavjih, torej 2.1 in 2.2, določili cenilki za θ po metodi največjega verjetja in po metodi momentov. Nato bomo za obe pridobljeni cenilki ugotovili, ali sta nepristranski oz. katera je vsaj asimptotično bolj nepristranska. Pri tem bomo uporabili izračun asimptotične srednje kvadratične napake oz. MSE. V preostalih delih bomo konkretno uporabili priložene datoteke z meritvami, najprej da določimo numerične ocene cenilk za vsak eksperiment (torej vsako datoteko) posebej. Za vse izračune bomo tudi ocenili standardno napako in rezultate grafično prikazali. Na koncu bomo pridobljeni gostoti primerjali s histogramom meritev.

2.1 Cenilka za θ po metodi največjega verjetja

Denimo, da imamo n medseboj neodvisnih in enako porazdeljenih spremenljivk X_i ; $i \in \{1, 2, \ldots, n\}$, ki so vse porazdeljene z Rayleighovo porazdelitvijo 2. Verjetje definiramo s predpisom

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta) = \begin{cases} \prod_{i=1}^{n} \frac{x_i}{\theta^2} \exp\left(-\frac{x_i^2}{2\theta^2}\right) \ ; \ x_1, \ldots, x_n > 0 \\ 0 \ ; \ \text{sicer} \end{cases}$$

Ker je lahko delo s tem produktom zahtevno, raje vse skupaj logaritmiramo in dobimo

$$l(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln(\frac{x_i}{\theta^2}) - \frac{x_i^2}{2\theta^2} \ ; \ za \ x_1, \ldots, x_n > 0$$

Cenilka θ po metodi največjega verjetja je neka funkcija $h(X_1, \ldots, X_n)$ pri kateri $L(\theta|x_1, \ldots, x_n)$ doseže maksimum za vse $X_1, \ldots, X_n$. Logaritmiranje L ohrani ta ekstrem v smislu, da če bo L dosegel svoj maksimum v θ, bo tam svoj maksimum dosegel tudi l in obratno. Sedaj odvajamo l po θ in dobimo:

$$\frac{\partial l}{\partial \theta}(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} \left(\frac{-2 * \theta^2 * x_i}{\theta^3 * x_i} - \frac{-2 * x_i^2}{2 * \theta^3}\right) = \sum_{i=1}^{n} \left(\frac{x_i^2}{\theta^3} - \frac{2}{\theta}\right) = \sum_{i=1}^{n} \left(\frac{x_i^2}{\theta^3}\right) - \frac{2 * n}{\theta}$$

Da najdemo ekstrem moramo rešiti enačbo $\frac{\partial l}{\partial \theta}(\theta|x_1, \ldots, x_n) = 0$. Ko vanjo vstavimo, kar smo ravnokar poračunali zgoraj, dobimo

$$\sum_{i=1}^{n} \frac{x_i^2}{\theta^3} = \frac{2 * n}{\theta}$$

oziroma

$$\sum_{i=1}^{n} x_i^2 = 2 * n * \theta^2$$

Od tod izrazimo θ^2 iz zgornje enakosti in rezultat korenimo, da dobimo θ. Velja:

$$\theta = \pm\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{2 * n}}$$

Preveriti moramo še, da l v θ res doseže maksimum. Za to poračunamo drugi odvod l po θ:

$$\frac{\partial^2 l}{\partial \theta^2}(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} (-3)\frac{x_i^2}{\theta^4} + \frac{2n}{\theta^2}$$

Sedaj v izraz vstavimo $\theta = \pm\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{2*n}}$ in tako dobimo

$$\sum_{i=1}^{n}(-3)\frac{4n^2x_i^2}{(\sum_{i=1}^{n}x_i^2)^2}+\frac{4n^2}{\sum_{i=1}^{n}x_i^2} = \frac{4n^2}{\sum_{i=1}^{n}x_i^2}(\frac{(-3)\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}x_i^2}+1) = (-2)\frac{4n^2}{\sum_{i=1}^{n}x_i^2} < 0$$

Ugotovili smo že, da ima verjetje l ekstrema v $+\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{2n}}$ in $-\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{2n}}$, sedaj pa vemo tudi, da sta oba ekstrema maksimuma. Za cenilko θ po metodi največjega verjetja izberemo koren s pozitivnim predznakom. Cenilka θ po metodi največjega verjetja je torej $\widehat{\theta} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{2*n}}$. Vrednost cenilke je odvisna od števila spremenljivk. Bolj primerna je torej oznaka $\widehat{\theta}_n = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{2*n}}$.

2.2 Cenilka za θ po metodi momentov

Sedaj se obrnemo na metodo momentov. Denimo, da so $X_1, \ldots, X_n$ neodvisne enako porazdeljene slučajne spremenljivke. Da določimo cenilko za izbrane parametre po tej metodi, moramo najprej poračunati momente nizkih stopenj, torej $E\left[X^k\right]$, v odvisnosti od parametrov, ki jih želimo oceniti. Nato iz dobljenih enačb izrazimo parametre v odvisnosti od momentov. Cenilko za parametre dobimo tako, da v enačbi k-ti moment zamenjamo s povprečjem k-tih potenc $\frac{1}{n}\sum_{i=1}^{n} X_i^k$. Ker v našem primeru skušamo oceniti samo en paramater, θ, načeloma zadošča če izračunamo samo prvi moment.

$$E\left[X\right] = \int_{-\infty}^{\infty} x f(x|\theta)\, dx = \int_{0}^{\infty} \frac{x^2}{\theta^2} e^{-\frac{x^2}{2\theta^2}}\, dx$$

Uvedemo novo spremenljivko $u = \frac{x^2}{2\theta^2}$; $du = \frac{x}{\theta^2}\, dx$, torej je $dx = \frac{\theta^2}{x} du$. Naš integral nato postane:

$$\int_{0}^{\infty} x e^{-u}\, du = \int_{0}^{\infty} \sqrt{2u\theta^2} e^{-u}\, du = \theta\sqrt{2}\int_{0}^{\infty} u^{\frac{1}{2}} e^{-u}\, du$$

V integralu prepoznamo obliko gama funkcije in hitro ugotovimo, da je integral enak $\Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{\pi}$.

Sledi, da je pričakovana vrednost Rayleighove porazdelitve s parametrom θ enaka $\sqrt{\frac{\pi}{2}}\theta$. Od tod izrazimo θ kot $\theta = \sqrt{\frac{2}{\pi}} E\left[X\right]$. Cenilka θ po metodi momentov je torej $\widehat{\theta} = \sqrt{\frac{2}{\pi}}\bar{X} = \frac{\sqrt{2}}{n\sqrt{\pi}}\sum_{i=1}^{n} X_i$. Tudi v tem primeru nam število spremenljivk vpliva na vrednost cenilke. Zato smo v resnici pridobili celo zaporedje cenilk $\widehat{\theta}_n$, tako kot v prejšnjem podpoglavju.

2.3 Asimptotični MSE cenilk

Srednja kvadratična napaka cenilke definirana s formulo $MSE(\widehat{\theta}|\theta) = E[(\widehat{\theta} - \theta)^2]$. Da izračunamo asimptotično MSE, najprej za fiksen n poračunamo MSE, nato pa dobljeno limitiramo: $\lim_{n\mapsto\infty} MSE(\widehat{\theta}_n|\theta)$. Pri računanju nam tudi pomaga enakost $MSE(\widehat{\theta}|\theta) = Var(\widehat{\theta}) + Bias(\widehat{\theta}|\theta)^2$. Pri tem se pristranskost oz »Bias« izračuna po formuli $Bias(\widehat{\theta}|\theta) = E[\widehat{\theta}] - \theta$.

2.3.1 MSE cenilke po metodi največjega verjetja

Začnimo s cenilko, ki smo jo pridobili po metodi največjega verjetja. Najprej za fiksen n poračunamo pristranskost. Pri tem nam pomaga informacija, ki smo jo pridobili iz vira [2], da ima $Y = \sum_{i=1}^{n} X_i^2$ gama porazdelitev $\Gamma(n, \frac{1}{2*\theta^2})$.

$$E[\widehat{\theta}_n] = E[\frac{1}{\sqrt{2n}}\sqrt{Y}] = \frac{1}{\sqrt{2n}}E[\sqrt{Y}]$$

Poračunajmo sedaj $E[\sqrt{Y}]$.

$$E[\sqrt{Y}] = \int_0^\infty \sqrt{y}\frac{(\frac{1}{2\theta^2})^n}{\Gamma(n)}y^{n-1}e^{-\frac{y}{2\theta^2}}\,dy = \int_0^\infty (\frac{y}{2\theta^2})^n\frac{1}{\Gamma(n)\sqrt{y}}e^{-\frac{y}{2\theta^2}}\,dy$$

Vstavimo novo spremenljivko $u = \frac{y}{2\theta^2}$; $du = \frac{1}{2\theta^2}\,dy$. Pri tem še opazimo, da je $\sqrt{y} = \sqrt{2\theta^2 u}$. Sedaj vstavimo to v integral.

$$E[\sqrt{Y}] = \int_0^\infty u^n\frac{1}{\Gamma(n)\sqrt{2\theta^2}\sqrt{u}}e^{-u}2\theta^2\,du = \frac{\sqrt{2\theta^2}}{\Gamma(n)}\int_0^\infty u^{n-\frac{1}{2}}e^{-u}\,du = \sqrt{2\theta^2}\frac{\Gamma(n+\frac{1}{2})}{\Gamma(n)}$$

Sledi, da je $E[\widehat{\theta}_n] = \frac{\theta\Gamma(n+\frac{1}{2})}{\sqrt{n}\Gamma(n)}$ in potem je $\widehat{\theta}_n$ pristranska cenilka, saj je

$$E[\widehat{\theta}_n] - \theta = \theta(\frac{\Gamma(n+\frac{1}{2})}{\sqrt{n}\Gamma(n)} - 1) \neq 0$$

za $\theta \neq 0$. Cenilko lahko popravimo na nepristransko. Nepristranska cenilka je tako

$$\theta_n^+ = \frac{\Gamma(n)\sqrt{n}}{\Gamma(n+\frac{1}{2})}\widehat{\theta}_n = \frac{\Gamma(n)}{\Gamma(n+\frac{1}{2})\sqrt{2}}\sqrt{Y}$$

Uspelo nam je torej izračunati pristranskost naše cenilke $\widehat{\theta}_n$:

$$Bias(\widehat{\theta}_n) = \theta(\frac{\Gamma(n+\frac{1}{2})}{\sqrt{n}\Gamma(n)} - 1)$$

Ko poračunamo limito $\lim_{n\mapsto\infty}\frac{\Gamma(n+\frac{1}{2})}{\sqrt{n}\Gamma(n)} = 1$, vidimo, da je $\widehat{\theta}_n$ asimptotsko nepristranska, saj sledi $\lim_{n\mapsto\infty} Bias(\widehat{\theta}_n) = 0$. Prej omenjeno limito lahko izračunamo tako, da obravnavamo naravni logaritem Γ funkcije. Velja namreč, da je $\ln(\Gamma(n+1)) = \ln(n) + \ln(\Gamma(n))$. Za velike n postaja $\ln(\Gamma)$ vse bolj strma in »linearna«, torej lahko za velike n naredimo aproksimacijo $\ln(\Gamma(n+\frac{1}{2})) \approx \frac{1}{2}\ln(n) + \ln(\Gamma(n))$. Ko to aproksimacijo antilogaritmiramo, dobimo $\Gamma(n+\frac{1}{2}) \approx \sqrt{n}\Gamma(n)$ oz. $\frac{\Gamma(n+\frac{1}{2})}{\sqrt{n}\Gamma(n)} \approx 1$. Ko pošljemo $n \mapsto \infty$, postane rezultat točen, torej $\lim_{n\mapsto\infty}\frac{\Gamma(n+\frac{1}{2})}{\sqrt{n}\Gamma(n)} = 1$.

Poračunati moramo še varianco cenilke $Var(\widehat{\theta}_n) = E[\widehat{\theta}_n^2] - E[\widehat{\theta}_n]^2 = E[\frac{1}{2n}Y] - E[\widehat{\theta}_n]^2$. Za Porazdelitev Y vemo, da je $\Gamma(n, \frac{1}{2\theta^2})$, tako da hitro pridobimo $E[\frac{1}{2n}Y] = \frac{1}{2n}\frac{n}{\frac{1}{2\theta^2}} = \frac{1}{2n}2n\theta^2 = \theta^2$. Sledi

$$Var(\widehat{\theta}_n) = \theta^2(1 - \frac{\Gamma(n+\frac{1}{2})^2}{n\Gamma(n)^2})$$

Ko to varianco limitiramo z $n \mapsto \infty$, gre izraz v oklepaju proti 0, kar lahko vidimo s pomočjo prej opravljenega izračuna limite $\lim_{n\mapsto\infty} \frac{\Gamma(n+\frac{1}{2})}{\sqrt{n}\Gamma(n)} = 1$. Velja torej, da je $\lim_{n\mapsto\infty} Var(\widehat{\theta}_n) = 0$. Posledično je asimptotična $MSE(\widehat{\theta}_n)$ enaka 0, torej $\lim_{n\mapsto\infty} MSE(\widehat{\theta}_n) = 0$.

2.3.2 MSE cenilke po metodi momentov

Sedaj storimo enako še za cenilko, ki smo jo pridobili po metodi momentov $\widehat{\theta}_n = \sqrt{\frac{2}{\pi}}\bar{X}$. Najprej poračunajmo njeno pričakovano vrednost.

$$E[\widehat{\theta}_n] = \frac{\sqrt{2}}{n\sqrt{\pi}} \sum_{i=1}^{n} E[X_i] = \sqrt{\frac{2}{\pi}} E[X] = \theta$$

Od tod lahko že sklepamo, da je $\widehat{\theta}_n$ nepristranska cenilka. Sledi, da je $MSE(\widehat{\theta}_n) = Var(\widehat{\theta}_n)$. Poračunajmo sedaj še varianco in pri tem upoštevamo, da so X_i medseboj neodvisne in enako porazdeljene.

$$Var(\widehat{\theta}_n) = \frac{2}{n^2\pi} Var(\sum_{i=1}^{n} X_i) = \frac{2}{n^2\pi} \sum_{i=1}^{n} Var(x_i) = \frac{2}{n\pi} Var(X)$$

Na tej točki se skličemo na varianco rayleighove porazdelitve s parametrom θ^2, ki je enaka $\frac{4-\pi}{2}\theta^2$. Sledi, da je $Var(\widehat{\theta}_n) = \frac{2}{n\pi} * \frac{4-\pi}{2}\theta^2 = \frac{(4-\pi)}{n\pi}\theta^2$, ki konvergira proti 0 ko pošljemo $n \mapsto \infty$. Posledično je $\lim_{n\mapsto\infty} MSE(\widehat{\theta}_n) = \lim_{n\mapsto\infty} Var(\widehat{\theta}_n) = 0$.

Tako prva kot druga cenilka imata torej asimptotično srednjo kvadratično napako 0. Kljub temu sklepamo, da je cenilka, ki smo jo pridobili po metodi momentov, boljša od cenilke, ki smo jo pridobili po metodi največjega verjetja, saj je cenilka po metodi momentov zraven še nepristranska.

2.4 Ševilska ocena prve cenilke

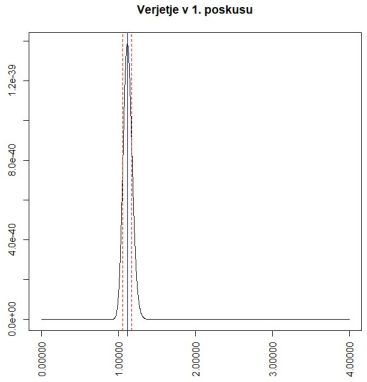
Številsko oceno cenilke za vsak eksperiment izvedemo s pomočjo skripte *SlucajniSprehodiMNV.R*. Ta ob zagonu poračuna številske vrednosti cenilke, oceni standardno napako vsake cenilke ter izriše pripadajoče grafe gostot za vsak eksperiment in graf na katerem so prikazani vsi trije grafi. Pri tem standardno napako $SE(\widehat{\theta}_n)$ cenilke pridobljene po metodi največjega verjetja izračunamo kot $\sqrt{Var(\widehat{\theta}_n)}$. Velja torej, da je $SE(\widehat{\theta}_n) = \theta\sqrt{1 - \frac{\Gamma(n+\frac{1}{2})^2}{n\Gamma(n)^2}}$. Ker parametra θ ne poznamo, izračunamo cenilko za SE po formuli $\widehat{SE}(\widehat{\theta}_n) = \widehat{\theta}_n\sqrt{1 - \frac{\Gamma(n+\frac{1}{2})^2}{n\Gamma(n)^2}}$ Prej omenjena skripta ocenjene vrednosti cenilk in standardnih napak izpiše v konzolo, so pa zbrane tudi v spodnji tabeli, ki pa vsebuje tudi na dve decimalki zaokrožene vrednosti standardnih napak.

Eksperiment	kratki	srednji	dolgi
Vrednost $\widehat{\theta}$	1,117394	2,075983	3,324422
Ocena SE	0,05728327	0,0658959	0,1451586

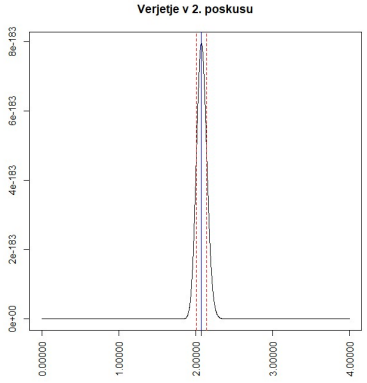
Tabela 6: *Numerične vrednosti MNV cenilke in standardne napake po eksperimentih*

Izračun standardne napake cenilke je neizogiben in pomemben del vsake statistične analize podatkov. Razlog za to je, ker nam služi kot merilo, kako dobro se naša cenilka prilega dejanskemu parametru, ki ga cenimo glede na dane podatke. Če je standardna napaka velika, vrednosti cenilke močno odstopajo od pričakovane vrednosti cenilke, ki je, v primeru ko je cenilka nepristranska, ravno enaka populacijskemu parametru, ki ga ocenjuje. Iz tega razloga, se načeloma raje dela z nepristranskimi cenilkami. Vseeno enaka intuicija velja tudi za pristranske spremenljivke - večja kot je standardna napaka na vzorcu, bolj odstopa cenilka od dejanske vrednosti parametra, ki ga ocenjujemo, za celotno populacijo in manjša kot je standardna napaka, manj cenilka odstopa od dejanske vrednosti parametra.

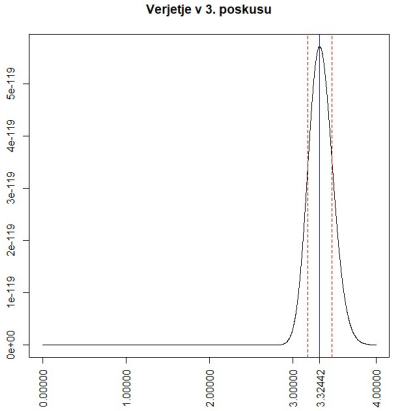
Kot je bilo že prej omenjeno, skripta *SlucajniSprehodiMNV.R* tudi nariše grafe verjetij, ki pripadajo izračunanim ocenam paramatra θ. Na teh grafih je z zeleno črto označena izračunana vrednost θ, okoli nje pa je z rdečima črtkanima premicama označena pripadajoča SE-okolica. Skripta sicer grafe izriše enega nad drugim, tukaj pa bodo priloženi vsak posebej.



Slika 5: *Verjetje za oceno cenilke θ, pridobljene po metodi največjega verjetja, v prvem poskusu*



Slika 6: *Verjetje za oceno cenilke θ, pridobljene po metodi največjega verjetja, v drugem poskusu*



Slika 7: *Verjetje za oceno cenilke θ, pridobljene po metodi največjega verjetja, v tretjem poskusu*

Na vsakem grafu verjetja vidimo, da je širina SE-okolice pripadajoče izračunane cenilke v razmerju s širino »vala«, na grafu, kjer je verjetje pozitivno. V resnici opazimo, da se te SE-okolice širijo skupaj z valovi, ti pa se širijo z večanjem meritev skozi poskuse.

Zaradi razmeroma majhne velikosti standardnih napak lahko sklepamo, da je cenilka $\widehat{\theta}$ pridobljena po metodi največjega verjetna dobra cenilka za parameter θ. Seveda jo lahko še izboljšamo, tako da jo popravimo v nepristransko cenilko $\widehat{\theta}^{+}$, ki smo jo izračunali v podpodpoglavju 2.3.1.

2.5 Številska ocena druge cenilke

Tako kot v poglavju 2.4, lahko tudi za cenilko pridobljeno po metodi momentov stnandardno napako poračunamo kar kot koren variance:

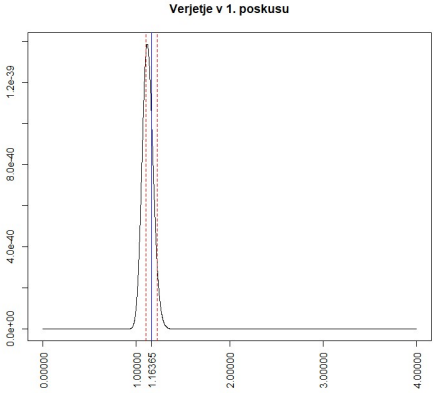
$$SE(\widehat{\theta}_n) = \sqrt{Var(\widehat{\theta}_n)} = \sqrt{\frac{(4-\pi)}{n\pi}\theta^2} = \sqrt{\frac{(4-\pi)}{n\pi}}\theta$$

Ker θ seveda ne poznamo, v izraz vstavimo cenilko $\widehat{\theta}_n$. Tako je naša formula na koncu $SE(\widehat{\theta}_n) = \sqrt{\frac{(4-\pi)}{n\pi}}\widehat{\theta}_n$. Skripta *SlucajniSprehodiMM.R* poračuna oceno cenilke, ki smo jo pridobili po metodi momentov ter standardno napako s pomočjo prej navedene formule. Vrednosti so navedene v spodnji tabeli skupaj z na dve decimalki zaokroženimi vrednostmi standardnih napak.

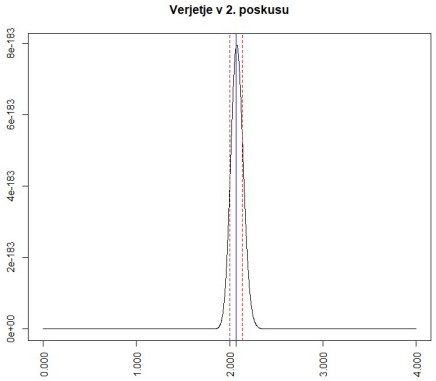
Eksperiment	kratki	srednji	dolgi
Vrednost θ	1,163652	2,068998	3,412388
Ocena SE	0,06240695	0,06867617	0,1558456

Tabela 7: *Numerične vrednosti cenilke pridobljene po metodi momentov in standardne napake po eksperimentih*

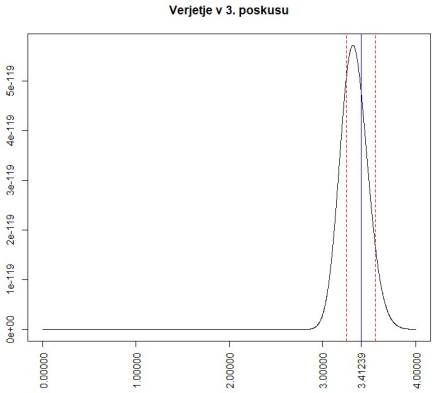
Poleg tega skripta tudi izriše graf gostote za ocenjeno cenilko za vsak poskus. Te grafe izriše enega zraven drugega, spodaj pa so grafi prikazani posebej.



Slika 8: *Verjetje za oceno cenilke θ, pridobljene po metodi momentov, v prvem poskusu*



Slika 9: *Verjetje za oceno cenilke θ, pridobljene po metodi momentov, v drugem poskusu*

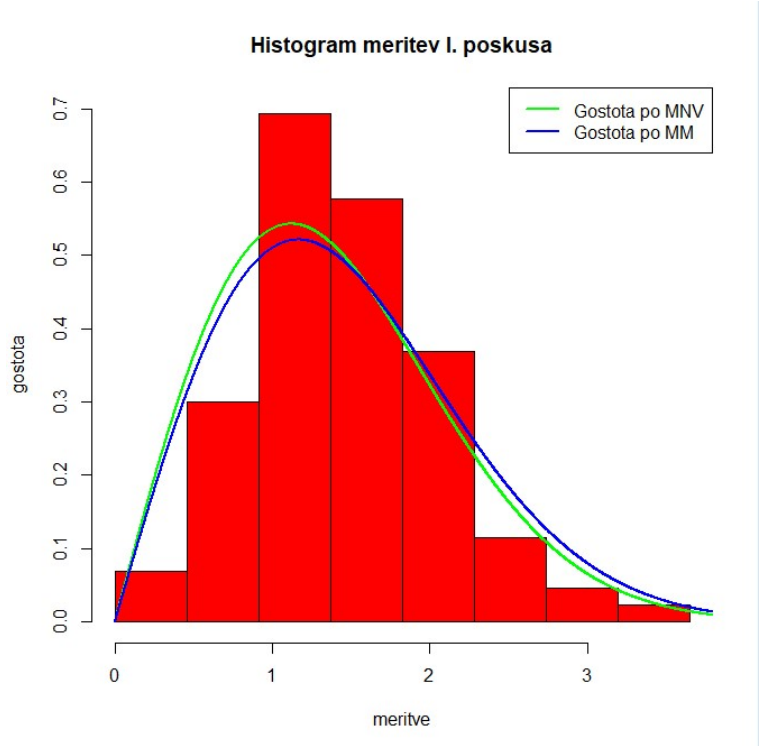


Slika 10: *Verjetje za oceno cenilke θ, pridobljene po metodi momentov, v tretjem poskusu*

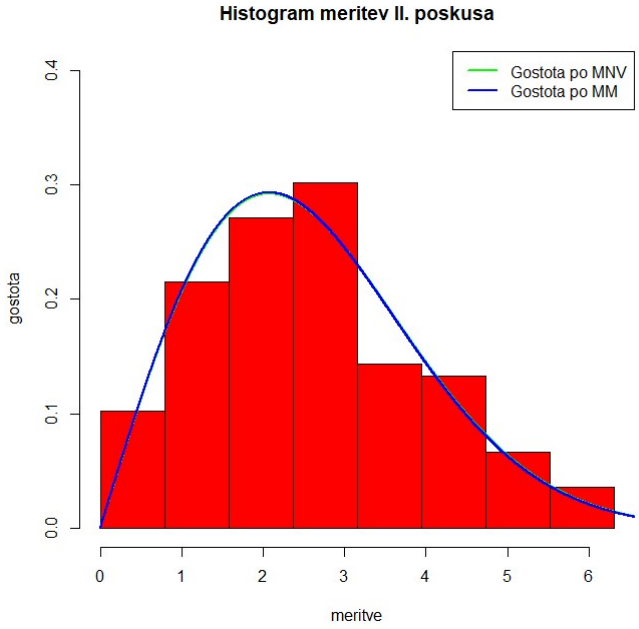
V tem primeru pa opazimo, da je situacija drugačna kot v podpoglavju 2.4, saj SE-okolice cenilke več ne vsebujejo vrha verjetij na enak, simetričen način, temveč so zamaknjene. To se zgodi kljub temu, da so standardne napake za to cenilko pridobljeno po metodi momentov razmeroma blizu vrednostim standardnih napak pri cenilki pridobljeni z metodo največjega verjetja. V primeru, ko napake zaokrožimo na dve decimalki natančno se razlika pojavi samo v zadnjem poskusu, kjer znaša 0,01. Še vedno torej velja, da je širina intervala proporcionalna s širino »vala« verjetja, toda simetrija je izgubljena. To nam pove, da je cenilka, ki smo jo pridobili po metodi momentov, sicer lažje izračunljiva od cenilke po metodi največjega verjetja, a verjetno ni enako dobra.

2.6 Histogram meritev in grafi gostot cenilk

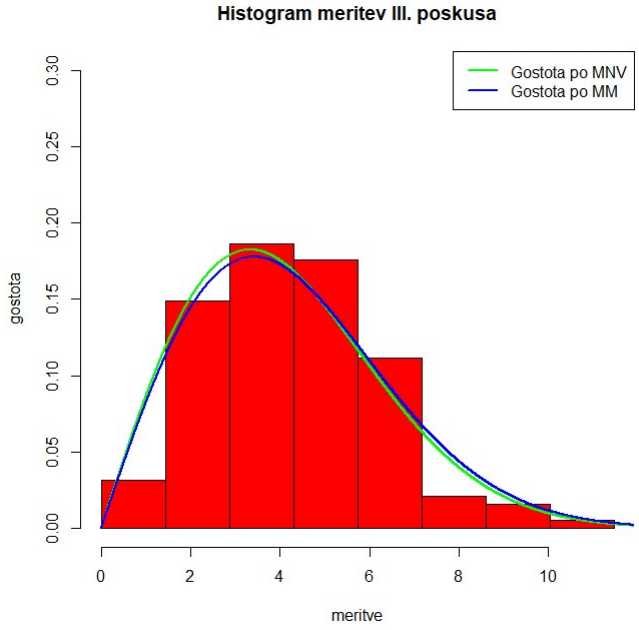
Da se dokončno prepričamo, da sta izračunani cenilki dobri, bomo za vsak poskus narisali histogram meritev in nanj dorisali gostoti za obe cenilki. Pri določanju širine razredov za histograme bomo uporabili modificirano Freedman-Diaconisovo pravilo, po katerem naj bi širina vsakega razreda bila približno $\frac{2,6IQR}{\sqrt[3]{n}}$, kjer je n število podatkov, IQR pa njihov interkvartilni razmik. Skripta *SlucajniSprehodiHIST.R* poračuna te širine za vsak eksperiment ter nato izriše pripadajoče histograme skupaj z gostotama obeh cenilki na vrhu. Skripta te grafe tudi izriše, so pa tudi prikazani spodaj.



Slika 11: *Histogram meritev v prvem poskusu skupaj z gostotama za pripadajoči cenilki.*



Slika 12: *Histogram meritev v drugem poskusu skupaj z gostotama za pripadajoči cenilki.*



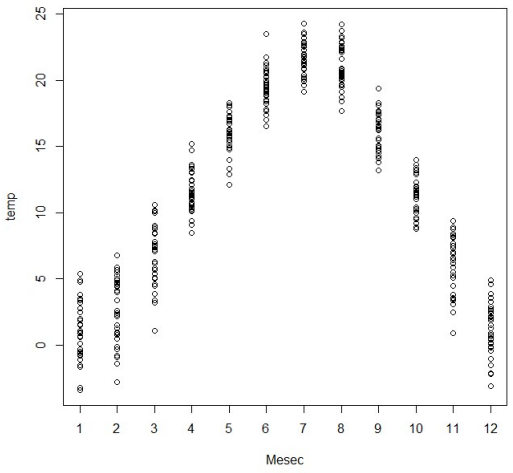
Slika 13: *Histogram meritev v tretjem poskusu skupaj z gostotama za pripadajoči cenilki.*

Vidimo lahko, da se na vsakem grafu obe gostoti medseboj, sicer z manjšimi odstopanji, razmeroma dobro prilegata. V primeru drugega poskusa, se v resnici skoraj popolnoma prekrivata. Tudi s histogrami je podobno - obe gostoti se prilegata obliki, ki jo tvorijo stolpci historgrama. Lahko torej sklepamo, da bi z večanjem vzorcev proti neskončnosti histogram »konvergiral« proti grafu gostote. Dokončno torej zatrdimo, da sta obe cenilki res dobri in predpostavka o porazdelitvi podatkov drži.

Ko primerjamo grafe verjetij s histogrami in odstopanji gostot lahko tudi sklepamo, da morda obstaja povezava med zamikom SE-okolic in razliko v prileganju gostot histogramu. V prvem in tretjem poskusu je razlika bolj opazna, kar velja tudi za zamika SE-okolic pri grafih verjetij, v drugem poskusu, kjer je zamik SE-okolice veliko manjši, pa se gostoti skorajda pokrivata.

3 Temperature

V datoteki *Temp_LJ.csv* imamo podane povprečne mesečne temperature od leta 1986 do 2020. Glede na dane podatke bi želeli uporabiti model, ki bo čim bolje napovedal temperature v prihodnosti. Predlagana sta nam dva modela. Model A vključuje linearen trend in sinusno nihanje s periodo enega leta, model B pa vključuje linearen trend in člene, od katerih vsak spreminja temperaturo za svoj mesec. S pomočjo programskega jezika **R** narišemo graf temperatur v odvisnosti od meseca. Koda za izris tega grafa se nahaja v skripti *Temperature.R*.



Slika 14: *Razpršen graf povprečnih mesečnih temperatur glede na mesec.*

Opazimo, da se v podatkih res skriva nek trend, zato je seveda smiselno, da premislimo kateri model bi bil za to najbolj primeren. Seveda sta glavna kandidata modela A in B. Le ta bosta predmet obravnave v nadaljevanju tega poglavja. V prvem podpoglavju bomo preizkusili model A znotraj modela B, v drugem pa bomo poračunali Akaikejevo informacijo za oba modela.

3.1 Preizkus modela A znotraj modela B

Da preizkusimo model A znotraj B, najprej zapišimo oba modela. Model A je podan z enačbo $Y = a + bX + c\sin(\frac{X\pi}{6}) + d\cos(\frac{X\pi}{6}) + e$, kjer so a, b, c in d parametri modela, e pa je normalno porazdeljen šum s pričakovano vrednostjo 0 in varianco σ^2. Model B je podan z enačbo $Y = m(X) + kx + f$, kjer so k ter

$$m(X) = \begin{cases} m_1; & X \mod 12 = 1 \\ m_2; & X \mod 12 = 2 \\ \vdots & \\ m_{12}; & X \mod 12 = 0 \end{cases}$$

parametri modela, f pa je normalno porazdeljen šum s pričakovano vrednostjo 0 in varianco σ^2.

V obeh primerih je X slučajna spremenljivka, ki zavzame vrednosti med 1 in 420 (slednje število je enako številu podatkov v *Temp_LJ.csv*), Y pa je povprečna temperatura za mesec, ki pripada vrednosti $X \mod 12$.

Ker bo v matrični obliki lažje delat, v njo prepišemo oba modela. Od zdaj najprej bo Y vektor velikosti 420, ki vsebuje temperature iz *Temp_LJ.csv*, z e označimo normalno porazdeljen vektor šumov $\begin{bmatrix} e_1 & e_2 & \cdots & e_{420} \end{bmatrix}^\top$ (kjer so vsi e_i medseboj neodvisni) za katerega velja, da je $E[e] = 0$ in $Var(e) = \sigma^2 I_{420}$. Poleg tega z β označimo vektor parametrov modela $\begin{bmatrix} a & b & c & d \end{bmatrix}^\top$.

Matrična oblika modela A je potem

$$Y = \begin{bmatrix} 1 & X_1 & \sin(\frac{X_1\pi}{6}) & \cos(\frac{X_1\pi}{6}) \\ 1 & X_2 & \sin(\frac{X_2\pi}{6}) & \cos(\frac{X_2\pi}{6}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{420} & \sin(\frac{X_{420}\pi}{6}) & \cos(\frac{X_{420}\pi}{6}) \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{420} \end{bmatrix} = X\beta + e$$

Pri matrični obliki modela B moramo biti malo bolj previdni, saj prosti parameter $m(X)$ v resnici ni samo en parameter, ampak skupek dvanajstih. Naivno bi morda rekli, da se matrični zapis modela glasi

$$Y = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_{420} \end{bmatrix} \begin{bmatrix} m(i) \\ k \end{bmatrix} + f = \bar{X}\bar{\beta} + f$$

Pri tem $m(i)$ sprejme vrstico, v kateri se nahaja po množenju z matriko X in vrne vrednosti, ki so navedene zgoraj. Ta zapis ni najboljši, saj se izraz močno zakomplicira, prihranimo pa samo nekaj vrstic v zapisu.

Bolj primerna oblika modela B, s katero nam bo seveda tudi lažje delati, je

$$Y = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & X_1 \\ 0 & 1 & 0 & \cdots & 0 & X_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & X_{12} \\ 1 & 0 & 0 & \cdots & 0 & X_{13} \\ 0 & 1 & 0 & \cdots & 0 & X_{14} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & X_{420} \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{12} \\ k \end{bmatrix} + f = \bar{X}\bar{\beta} + f$$

Oblika nazorno pokaže, da so $m_1, \dots m_{12}$ vsi parametri, ki pa se v modelu pojavijo pod določenimi pogoji (v tem primeru je to, kateri mesec obravnavamo). Omenimo še, da je $f = \begin{bmatrix} f_1 & f_2 & \cdots & f_{420} \end{bmatrix}^\top$, tako kot e, normalno porazdeljen vektor šumov s pričakovano vrednostjo 0 in variančno matriko σI_{420}.

Sedaj, ko imamo v matrični obliki zapisana modela, se lahko lotimo preverjanja modela A znotraj modela B. Postopek, ki ga bomo ubrali, je naslednji: Najprej bomo po metodi najmanjših kvadratov poračunali cenilki za β in $\bar{\beta}$, nato s pomočjo le teh za vsak model poračunali residuale in preko njih RSS. S pomočjo slednjih, bomo izračunali kvocient F in ga primerjali z inverzom kumulativne funkcije Fisherjeve porazdelitve za stopnji tveganja 0,01 in 0,05. Če bo $F \geq F^{-1}_{Fisher(9,407)}(1-\alpha)$ za stopnjo tveganja α, bomo model A znotraj B zavrnili, sicer pa ga bomo sprejeli.

3.1.1 Cenilki β in $\bar{\beta}$ po Metodi najmanjših kvadratov

S predavanj vemo, da je cenilka po metodi najmanjših kvadratov za vektor parametrov $\gamma \in \mathbb{R}^p$ v modelu $Y = Z\gamma + \varepsilon$, kjer je $\varepsilon \sim N(0, \sigma^2 I_n)$, enaka $\widehat{\gamma} = (X^\top X)^{-1} X^\top Y$. Na ta način potem pridobimo cenilki za β in $\bar{\beta}$. Pri tem uporabimo **R** skripto *Temperature.R*, ki ob zagonu v konzolo izpiše cenilki za β in $\bar{\beta}$. ti cenilki sta navedeni spodaj.

$$\widehat{\beta} = \begin{bmatrix} 10{,}1300627498747 \\ 0{,}00535203760946302 \\ -5{,}10138050301417 \\ -9{,}04714874745839 \end{bmatrix} \text{ in } \widehat{\bar{\beta}} = \begin{bmatrix} -0{,}252767273576098 \\ 1{,}45041783380018 \\ 5{,}76503151260504 \\ 10{,}302502334267 \\ 14{,}8685445845005 \\ 18{,}5317296918767 \\ 20{,}5434862278245 \\ 19{,}9838141923436 \\ 14{,}9555707282913 \\ 10{,}2187558356676 \\ 4{,}95051237161531 \\ 0{,}156554621848738 \\ 0{,}00538632119514473 \end{bmatrix}$$

3.1.2 Residuali in RSS obeh modelov

Zdaj, ko imamo cenilki za β in $\bar{\beta}$ po metodi najmanjših kvadratov lahko poračunamo residuale. V teoretičnem modelu $Y = Z\gamma + \varepsilon$, kot je bil opisan prej v podpodpoglavju 3.1.1, residuale $\widehat{\varepsilon}_i$ poračunamo s formulo

$$\widehat{\varepsilon}_i = Y_i - \sum_{j=1}^{p} z_{ij}\widehat{\gamma}_j = Y_i - (Z\widehat{\gamma})_i = (Y - Z\widehat{\gamma})_i$$

Pri tem je $\widehat{\gamma}$ cenilka za γ po metodi najmanjših kvadratov. Kadar računamo s pomočjo računalnika, je bolj priročen vektorski zapis $\widehat{\varepsilon} = Y - Z\widehat{\gamma}$. Količina RSS tega modela je ravno enaka kvadratu 2-norme vektorja rezidualov $\widehat{\varepsilon}$ oziroma vsoti $\sum_{i=1}^{n} \widehat{\varepsilon}_i^2$. Skripta *Temperature.R* izračuna tako reziduale kot RSS za oba modela, a (v konzolo) izpiše samo RSS. Vrednosti RSS so tudi navedene v spodnji tabeli

Model	RSS
A	1260,57648627314
B	1150,50046218487

Tabela 8: *Tabela vrednosti RSS za modela A in B*

3.1.3 Preizkus modela A v B s stopnjama tveganja 0,01 in 0,05

Končno imamo vse, kar potrebujemo, da preizkusimo model A znotraj B. Skripta *Temperature.R* poleg vseh ostalih prej omenjenih vrednosti v konzolo izpiše tudi F, izračunan po formuli

$$F = \frac{\frac{(RSS_A - RSS_B)}{(p-q)}}{\frac{RSS_B}{(n-p)}}$$

Pri tem je p število parametrov modela B, q število parametrov modela A ter n število podatkov. Račun pove, da je $F = 4{,}32671049362724$. Poleg tega smo poračunali tudi $F^{-1}_{Fisher(9,407)}(1-\alpha)$ za $\alpha = 0{,}01$ ter $\alpha = 0{,}05$. Vrednosti sta predstavljeni v spodnji tabeli.

α	$F^{-1}_{Fisher(9,407)}(1-\alpha)$	$F^{-1}_{Fisher(9,407)}(1-\alpha) <= F$
0,01	2,45105980939293	Da
0,05	1,9028951359083	Da

Tabela 9: *Tabela vrednosti* $F^{-1}_{Fisher(9,407)}(1-\alpha)$ *za modela* $\alpha = 0{,}01$ *in* $\alpha = 0{,}05$

V tretjem stolpcu so tudi rezultati primerjav $F \geq F^{-1}_{Fisher(9,407)}(1-\alpha)$. Če neenakost velja, model A znotraj B zavrnemo. To se zgodi ravno pri obeh vrednostih α.

3.2 Akaikejeva informacija modelov

Preostane nam samo še izračun Akaikejeve informacije. To bomo storili z naslednjo formulo:

$$AIC = 2m + n\ln(RSS)$$

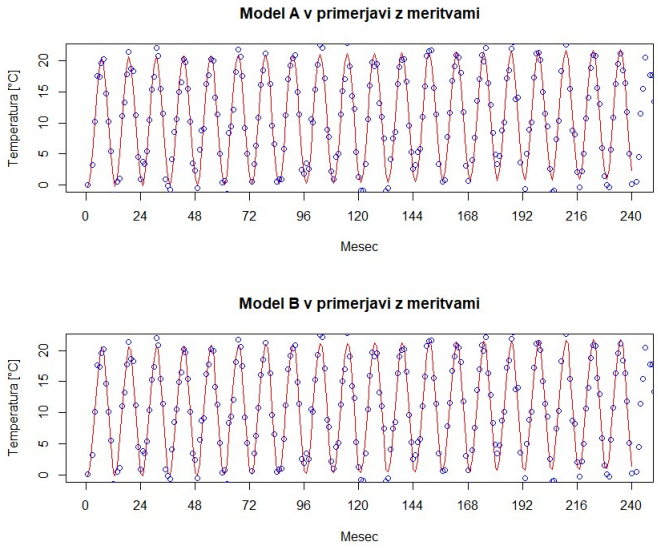
Pri tem je m število parametrov, n število podatkov oziroma opažanj ter RSS količina, ki smo jo že poračunali za oba modela. Tudi te vrednosti izračunamo v skripti *Temperature.R* in so avtomatsko izpisane v konzolo. Poleg tega so vrednosti predstavljene tudi v spodnji tabeli.

Model	AIC
A	3006,51625812173
B	2986,13997070424

Tabela 10: *Tabela vrednosti AIC modelov A in B*

Opazimo, da sta si AIC modelov A in B relativno blizu. Če oba podatka zaokrožimo na dve decimalni mesti, je razlika 20,38, kar v primerjavi z velikostjo obeh vrednosti ni veliko.

To, da sta oba modela skoraj enako dobra, lahko vidimo tudi na grafih, ki primerjata modela s priloženimi podatki. Da bo prikaz bolj pregleden, sta grafa narisana samo za prvih 20 let.



Slika 15: *Primerjava prileganja modelov A in B s podatki iz Temp_LJ.csv*

Opazimo tudi, da je $AIC(B) \leq AIC(A)$, od koder lahko sklepamo, da je model B boljši od modela A.

Literatura

[1] J. Rice, *Mathematical Statistics & Data Analysis*, 3rd ed., Duxbury, Berkeley, 2007.

[2] *Rayleighjeva porazdelitev*, v: Wikipedia, Prosta enciklopedija, [ogled 14. 7. 2022], dostopno na `https://sl.wikipedia.org/wiki/Rayleighjeva_porazdelitev`.