

Projekt statistika
Projekt pri predmetu Statistika

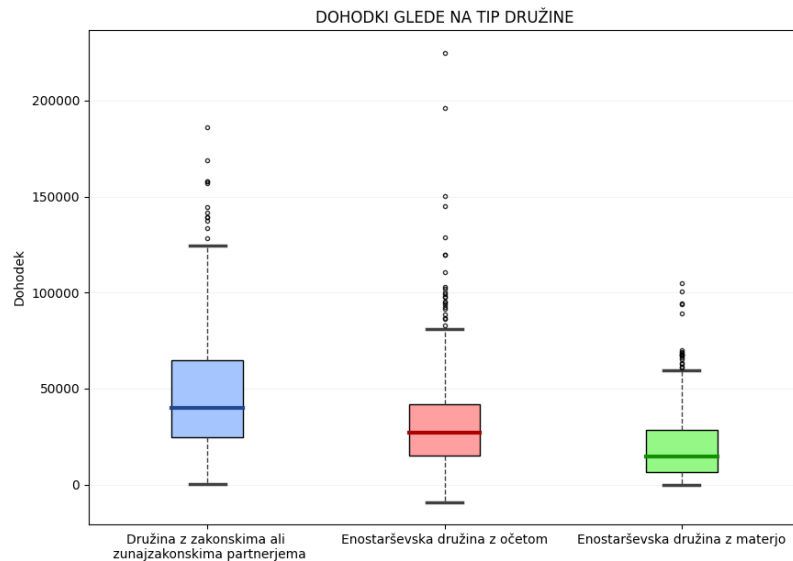
Matej Novoselec

30. junij 2023

1 Kibergrad

Podani so nam podatki o dohodkih 43886 družin iz mesta Kibergrad. Podatke o dohodkih bomo analizirali v odvisnosti od tipa družin, ki je posredovala podatke. Pri obdelavi podatkov si bomo pomagali z datoteko *statistika_naloga_1.py*, ki je namenjena predvsem izrisu grafov, ter numeričnim izračunom željenih vrednosti, prek danih podatkov.

Družine razdelimo v 3 tipe: tip 1 - družina z zakonskima ali zunajzakonskima partnerjema, tip 2 - enostarševska družina z očetom, tip 3 - enostarševska družina z materjo. Imamo 33403 podatkov od družin tipa 1, 2054 podatkov od družin tipa 2 in 8429 podatkov od družin tipa 3. Za vsak tip družine izberemo enostaven slučajen vzorec velikosti 500 in narišemo pripadajočo škatlo z brki.



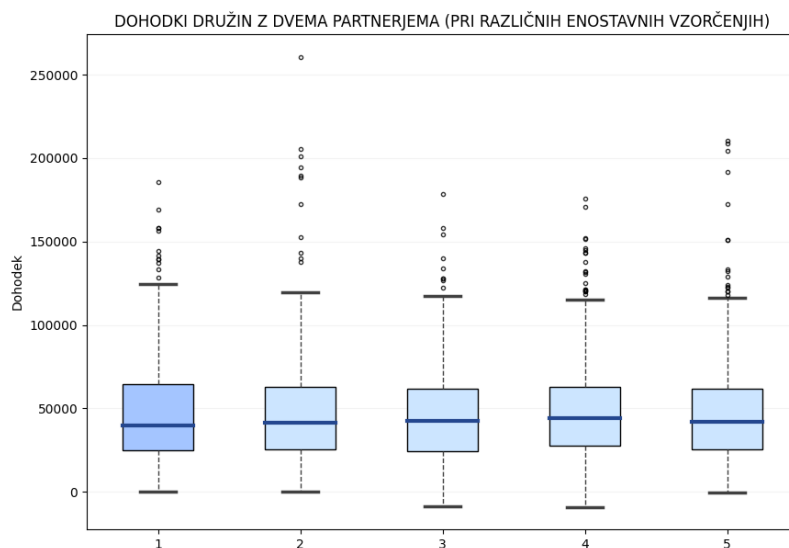
Slika 1: Škatle z brki za dohodke različnih tipov družin

Hitro vidimo, da so vrednosti prvega, drugega in tretjega kvantila pri družinah tipa 1 (družinah z zakonskima ali zunajzakonskima partnerjema) višji od prvega, drugega in tretjega kvantila pri družinah tipa 2 in tipa 3. Iz tega lahko sklepamo, da so povprečju dohodki omenjenega tipa družin večji od obeh ostalih tipov. Na podoben način lahko sklepamo tudi, da so dohodki v enostarševski družini z očetom v povprečju nekoliko višji od dohodkov enostarševskih družin z materjo. Opazimo tudi, da je varianca dohodkov pri družinah tipa 1 večja od variance podatkov o dohodkih družin preostalih tipov. Osamelcev opazimo največ pri družinah tipa 2, podatki pri družinah tipa 3 pa so najmanj razpršeni. Vredno je omeniti tudi, da je pri družinah tipa 1 in tipa 3 porazdelitev dohodkov

nekoliko nagnjena k dohodkov višje vrednosti.

Povzeli bi lahko, da nekoliko izstopajo podatki o družinah tipa 1, ki imajo povprečno najvišji dohodek, a imajo obenem tudi veliko varianco, njihova porazdelitev pa je nagnjena k višjim vrednostim. Vseeno moramo biti s sklepanjem nekoliko pazljivi, saj je vzorec velikosti 500 za 33403 podatkov relativno majhen.

Sedaj se osredotočimo na družine tipa 1 in iz podatkov izberimo še štiri enostavne slučajne vzorce. Opažanja so prikazana z vzporednimi škatlami z brki, pri čemer temnejše obarvana škatla z brki pripada podatkom, ki so bili izbrani kot enostavni slučajni vzorec za družine tega tipa iz prejšnje podnaloge.



Slika 2: Škatle z brki za dohodke družin z dvema partnerjema

Opazimo, da je bil zgoraj izbrani vzorec, kljub temu, da je bil dokaj majhen glede na število vseh podatkov, v veliki meri (vsaj iz vidika ostalih štirih vzorcev) reprezentativen. Vidno je namreč, da so vrednosti prvega in tretjega kvantila pri vzorcih dokaj usklajeni. Tudi iz vidika števila osamelcev je prvotno izbran vzorec reprezentativen. Nekoliko se od reprezentativnosti odmaknemo le pri vrednosti drugega kvantila, ki je pri našem vzorcu nekoliko nižje kot pri ostalih, posledično pa so podatki o dohodkih prvotno izbranega vzorca nekoliko nagnjeni k višjim vrednostim.

Zadnje podnaloge se za začetek lotimo nekoliko bolj teoretično, ter šele nato vstavimo podatke za izračun željenih vrednosti. Z n označimo število vseh numeričnim podatkov, t.j. skupni numerus je enak n . Podatki so razdeljeni v

skupine (glede na tip družine), n_i naj označuje število podatkov za skupino s tipom družine i (i seveda tu 1, 2 ali 3). Definirajmo še uteži za posamezno skupino, kot $w_i = n_i/n$.

Pričakovano vrednost za celoten Kibergrad lahko izrazimo prek pričakovanih vrednosti za posamezne skupine (označimo jih z μ_i) kot:

$$\mu = w_1\mu_1 + w_2\mu_2 + w_3\mu_3.$$

Če z σ_i^2 označimo varianco znotraj i -te skupine, lahko na podoben način varianco dohodka za celoten Kibergrad izrazimo kot:

$$\sigma^2 = \sigma_P^2 + \sigma_N^2, \quad \text{kjer: } \sigma_P^2 = \sum_{i=1}^3 w_i(\mu_i - \mu)^2 \quad \text{in} \quad \sigma_N^2 = \sum_{i=1}^3 w_i\sigma_i^2.$$

Izpeljava je navedena v [2]. V zgornjem zapisu σ_P^2 označuje s tipom družine pojasnjeno varianco, σ_N^2 pa nepojasnjeno varianco. Naloga nas sprašuje po obeh. Numerično v *python* datoteki poračunamo približke za pričakovane vrednosti in variance za posamezne skupini, ter po zgoraj zapisanih formulah dobimo, da s tipom družine pojasnjena varianca znaša 113781161.94, nepojasnjena varianca pa 912604507.95. Omenimo še, da delež pojasnjene variance ($\eta^2 = \frac{\sigma_P^2}{\sigma^2}$) znaša 0.11085, pojasnjeni standardni odklon pa 10666.83.

Oglejmo si novo pridobljene podatke v luči preteklih opažanj, predvsem iz vidika povprečnih dohodkov različnim tipov družin. Že v komentarju prve podnaloge smo omenili, da so med povprečnimi dohodki tipov družin opazne razlike. Sedaj opazimo, da je tudi vrednost pojasnjenega standardnega odklona visoka glede na povprečne dohodke posameznih tipov družin (te znašajo 47187.48 za tip 1, 31637.36 za tip 2 in 20508.19 za tip 3), kar dodatno nakazuje na razlike med povprečnimi dohodki posameznih tipov družin. Vrednost deleža pojasnjene variance le dodatno potrdi že zapisano.

2 Kromatin

Vredno je zapisati nekoliko bolj matematično interpretacijo problema/navodila. Podatki pripadajo trem različnim eksperimentom (podatki so ločeni glede na dolžino opazovanega Kromatina), ki bodo določali končne vrednosti in rezultate, a so problemi, ki jih podajajo, teoretično iste narave. Za podan problem sedaj razvijemo teoretični pristop in se lotimo reševanja podnaloge a) in b).

Imamo n neodvisnih, enako porazdeljenih slučajnih spremenljivk, označimo jih z R_1, R_2, \dots, R_n . Porazdeljene naj bodo Rayleighovo, t.j. z gostoto:

$$f_{R_i}(r_i | \theta) = \begin{cases} \frac{r_i}{\theta^2} \exp\left(-\frac{r_i^2}{2\theta^2}\right) & ; \quad r > 0 \\ 0 & ; \quad \text{sicer} \end{cases}.$$

Zaradi predpostavljene neodvisnosti, je potem (R_1, R_2, \dots, R_n) porazdeljen z gostoto:

$$\prod_{i=1}^n f_{R_i}(r_i | \theta) = \left(\frac{1}{\theta^{2n}} \prod_{i=1}^n r_i \right) \exp\left(-\frac{1}{2\theta^2} \sum_{i=1}^n r_i^2\right); \quad r_i > 0.$$

Lotimo se podnaloge a). Velja:

$$L(\theta | (r_1, r_2, \dots, r_n)) = \prod_{i=1}^n L_i(\theta | r_i) = \prod_{i=1}^n f_{R_i}(r_i | \theta)$$

in

$$l(\theta | (r_1, r_2, \dots, r_n)) = \ln L(\theta | (r_1, r_2, \dots, r_n)) = \sum_{i=1}^n \ln(f_{R_i}(r_i | \theta)),$$

ter zato:

$$l(\theta | (r_1, r_2, \dots, r_n)) = -2n \ln(\theta) + \sum_{i=1}^n \ln(r_i) - \frac{1}{2\theta^2} \sum_{i=1}^n r_i^2.$$

Iščemo cenilko za θ po metodi največjega verjetja, zato si ogledamo enakost:

$$0 = \frac{\partial l(\theta | (r_1, \dots, r_n))}{\partial \theta} = -\frac{2n}{\theta} + \frac{1}{\theta^3} \sum_{i=1}^n r_i^2.$$

Za cenilko po metodi največjega verjetja tako dobimo:

$$\hat{\theta}_{MNV} = \sqrt{\frac{1}{2n} \sum_{i=1}^n r_i^2}.$$

Za rešitev podnaloge b) si oglejmo pričakovano vrednost (Rayleighove) slučajne spremenljivke R :

$$\mathbb{E}(R) = \int_0^\infty r f(r | \theta) dr = \int_0^\infty \frac{r^2}{\theta^2} \exp\left(-\frac{r^2}{2\theta^2}\right) dr.$$

V integral uvedemo $\tau = \frac{r^2}{2\theta^2}$ in dobimo

$$\mathbb{E}(R) = \int_0^\infty \sqrt{2\tau} \theta e^{-\tau} d\tau = \theta\sqrt{2} \int_0^\infty \tau^{1/2} e^{-\tau} d\tau = \theta\sqrt{2} \Gamma(3/2) = \theta\sqrt{\frac{\pi}{2}}.$$

Cenilka po metodi momentov je zato podano s predpisom:

$$\hat{\theta}_{MM} = \bar{R}\sqrt{\frac{2}{\pi}}.$$

Iz izpeljave cenilke, dobljene po metodi momentov, je jasno vidno, da je cenilka nepristranska (hitro bi lahko tudi direktno preverili, da res velja $\mathbb{E}(\hat{\theta}_{MM}) = \theta$). Na podoben način, kot smo to storili zgoraj, se dokopljemo do pomožnega rezultata, ki nam bo prav prišel kasneje. Velja:

$$\mathbb{E}(R^2) = \int_0^\infty r^2 f(r | \theta) dr = 2\theta^2.$$

Sedaj se lotimo reševanja podnaloge c). Ker je cenilka za θ po metodi momentov nepristranska, velja $MSE(\hat{\theta}_{MM}) = Var(\hat{\theta}_{MM})$. Sedaj varianco tudi poračunajmo.

$$\begin{aligned} Var(\hat{\theta}_{MM}) &= Var\left(\bar{R}\sqrt{\frac{2}{\pi}}\right) = \frac{2}{\pi} \frac{Var(R)}{n} = \frac{2}{\pi} \frac{\mathbb{E}(R^2) - (\mathbb{E}(R))^2}{n} = \\ &= \frac{2}{\pi} \left(\frac{2\theta^2 - \left(\frac{\pi}{2}\theta\right)^2}{n} \right) = \frac{\theta^2}{n} \left(\frac{4 - \pi}{\pi} \right) \end{aligned}$$

Po zgornjem komentarju torej

$$MSE(\hat{\theta}_{MM}) = \frac{\theta^2}{n} \left(\frac{4 - \pi}{\pi} \right) \approx \frac{\theta^2}{n} \cdot 0,273.$$

Pri izračunu asimptotične srednje kvadratične napake pri cenilki za θ po metodi največjega verjetja, nam na pomoč priskoči Fischerjeva informacija. Velja:

$$\begin{aligned} FI(\theta) &= -\mathbb{E}\left(\frac{\partial^2 l(\theta | (r_1, \dots, r_n))}{\partial \theta^2}\right) = -\mathbb{E}\left(\frac{1}{\theta^4} \left(2n\theta^2 - 3 \sum_{i=1}^n r_i^2\right)\right) = \\ &= -\left(\frac{1}{\theta^4} \left(2n\theta^2 - 3n \mathbb{E}(R^2)\right)\right) = -\mathbb{E}\left(\frac{1}{\theta^4} \left(2n\theta^2 - 3n\left(\frac{\pi}{2}\theta^2 + \frac{4 - \pi}{2}\theta^2\right)\right)\right) = \frac{4n}{\theta^2}, \end{aligned}$$

oziroma:

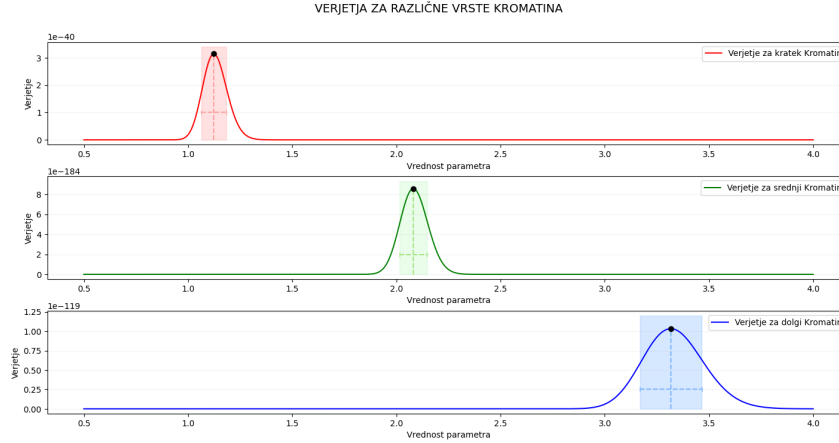
$$FI^{-1}(\theta) = \frac{1}{4} \frac{\theta^2}{n}, \text{ zato asimptotično velja: } MSE(\hat{\theta}_{MNV}) \approx \frac{\theta^2}{n} \cdot 0,25.$$

Opazimo, da je faktor pred asimptotično srednje kvadratično napako za cenilko, dobljeno po metodi največjega verjetja nekoliko manjši, zato je vsaj asimptotično cenilka $\hat{\theta}_{MNV}$ nekoliko boljša od cenilke $\hat{\theta}_{MM}$.

Podnalog a d) nam narekuje oddaljitev od teoretičnih zapisov in izračun številskih ocen. Pomagajmo si z datoteko *statistika_naloga_2.py*. Ta nam po obdelavi podatkov za posamezno vrsto Kromatina pove, da številska ocena za θ po metodi največjega verjetja pri kratkem Kromatinu znaša 1.1238, pri srednjem Kromatinu znaša 2.0812, pri dolgem Kromatinu pa 3.4045. Standardno napako lahko sedaj številsko ocenimo s pomočjo številske ocene za θ (po metodi največjega verjetja). Natančneje, za oceno uporabimo:

$$\hat{SE}(\hat{\theta}_{MNV}) = \sqrt{\frac{1}{4n}} \cdot \hat{\theta}_{MNV}.$$

Številska ocena za standardno napako tako pri kratkem Kromatinu znaša 0.059735, pri srednjem Kromatinu 0.065743, ter pri dolgem Kromatinu 0.148161. Zapisane številske ocene si oglejmo na spodnji sliki.



Slika 3: Grafi funkcij verjetja pri različnih vrstah Kromatina

Na sliki so prikazani grafi funkcij verjetja posamezne vrste Kromatina, v odvisnosti od parametra modela θ . Kot je zapisano v legendi, so grafi funkcij verjetja za različne vrste Kromatina različno obarvani. Navpična črtkasta črta pri vsaki vrsti Kromatina povezuje vrednost številske ocene za θ po metodi največjega verjetja (tiste vrste Kromatina) z vrednostjo funkcije verjetja pri dani številski oceni (točka funkcijske vrednosti označena s črno piko). Na grafu je dobro vidno tudi analitično ozadje metode največjega verjetja, za vsako vrsto Kromatina namreč grafično opazimo, da ima funkcija verjetja (za tisto vrsto Kromatina) v (oziroma blizu) številski oceni za θ stacionarno točko. Natančneje, opazimo, da v številski oceni zavzamemo maksimum funkcije verjetja. Komentirajmo še pomen velikosti (pri nas ocene) standardne napake. Funkcije verjetja določajo (za vsako vrsto Kromatina) dani podatki (o tisti vrsti Kromatina), ki obenem v model vnašajo slučaj. Zaradi slučajnosti, ki jih vnašajo

podatki, bi radi ocenili napako, ki se pojavi pri določitvi argument maksimuma funkcije verjetja (torej cenilki po metodi največjega verjetja za parameter našega modela, t.j. $\hat{\theta}_{MNV}$). Vrednost standardne napake, oziroma v našem primeru njena ocena, nam torej določata velikost intervala okoli naše poračunane ocene za θ po metodi največjega verjetja. Na sliki je za vsako vrsto Kromatina to tudi označeno s prosojno obarvanim območjem, pri čemer meje intervala dodatno nakazuje vodoravna črtkasta daljica.

Vredno je omeniti tudi grafično opaženo korelacijo med položnostjo zavzetja maksimuma funkcije verjetja in velikostjo (ocene) standardne napake za pripadajočo vrsto Kromatina. Vidno je namreč, da sta številske oceni za standardni napaki kratkega in srednjega Kromatina skoraj enaki, kar se ujema z skoraj enakima širinama "hriba" na vrhu katerega funkcija zavzame maksimum. V primerjavi s kratkim in srednjim Kromatinom, je številska ocena za standardno napako pri dolgem Kromatinu večja. Prav tako hitro vidimo, da pri grafu verjetja za dolgi Kromatin maksimum zavzamemo nekoliko bolj položno oziroma je širini že omenjenega "hriba" vidno večja.

Sedaj se lotimo podnaloge e). Že zgoraj omenjena *python* datoteka nam poda številske ocene tudi za θ po metodi momentov. Te znašajo 1.1706 za kratek Kromatin, 2.0748 za srednji Kromatin in 3.4045 za dolgi Kromatin. Oceno za napako iščemo na podoben način, kot smo to storili pri metodi največjega verjetja. Tokrat bo številska ocena za θ tista iz metode momentov, faktor pred napako pa ustrezno drugačen. Natančneje, za oceno uporabimo:

$$\hat{SE}(\hat{\theta}_{MM}) = \sqrt{\frac{4 - \pi}{\pi n}} \cdot \hat{\theta}_{MM}.$$

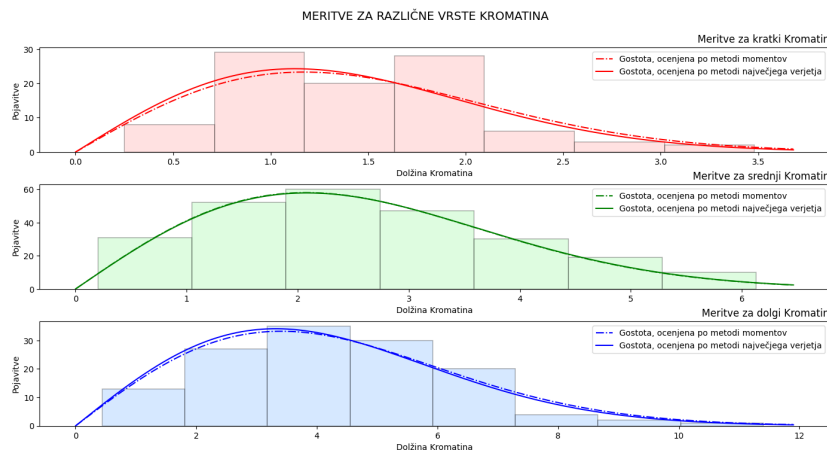
Številske ocene znašajo 0.06245 za kratek Kromatin, 0.068731 za srednji Kromatin in 0.154895 za dolgi Kromatin.

Opazimo, da so številske vrednosti ocen standardnih napak (za θ po metodi momentov in po metodi največjega verjetja) pri enaki vrsti Kromatina skoraj enake. Slednje je seveda razumljivo, saj smo za oceno standardne napake vzeli koren asimptotične ocene srednje kvadratične napake, pri čemer smo θ nadomestili z glede na metodo določeno cenilko. Ker nam metoda največjega verjetja asimptotične daje boljšo cenilko, je tudi ocena za standardno napako, pridobljena prek te metode manjša. Pomen (ocene) standardne napake je enak tistemu, ki smo ga komentirali na začetku obravnave metode največjega verjetja.

KOMENTAR STANDARDNE NAPAKE, ZAKAJ JE RELEVANTNA??

Podatke znotraj vrste Kromatina sedaj predstavimo še s histogramom. Pri tem določimo širino bloka oziroma število blokov histograma s pomočjo modificiranega Freedman–Diaconisovega pravila. Ta nam prek medkvartilnega razmika določi optimalno širino blokov histograma. Omenjena optimalna širina bloka pri kratkem Kromatinu znaša 0.4685, za srednji Kromatin 0.7976, za dolgi Kromatin pa 1.4145. Zapisane vrednosti širine blokov, se glede na razpon podatkov

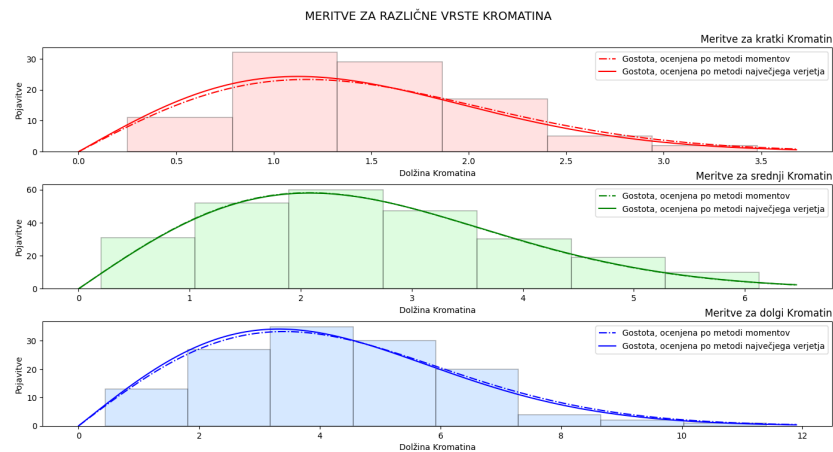
najbolje ujemajo (najbližje naravno število) s 7 bloki pri kratkem Kromatinu, s 7 bloki pri srednjem Kromatinu in z 8 bloki pri dolgem Kromatinu. Za vsako vrsto Kromatina, na pripadajoč histogram (določen s zgoraj zapisanim številom blokov) dorišimo še pripadajoči oceni za gostoti, ki smo jo dobili iz ocene za θ po metodi momentov in po metodi največjega verjetja.



Slika 4: Histogrami in ocenjene gostote za različne vrste Kromatina

Kot je to zapisano že v legendi, sta gostoti na pripadajočem histogramu dorisani z barvno usklajeno črto, tip katere pa se razlikuje glede na metodo, ki nam je omogočala izračun ocene gostote. Vredno je omeniti, da gre za prikaz skaliranih (množili smo s skupno površino histograma) gostot, saj bomo le na ta način lahko smiselno komentirali prileganje.

Sedaj zapišimo nekaj splošnih opazk. Pri histogramu za srednji Kromatin opazimo, da se grafa ocen gostot skoraj ujemata, obenem pa se (izmed vseh vrst Kromatina) oceni gostot najbolj prilegata podatkom, predstavljenim s histogramom. Skorajšno prekrivanje ocen gostot na omenjenem grafu, je posledica bližine vrednosti za oceno cenilke obeh metod (2.0748 pri metodi momentov, in 2.0812 pri metodi največjega verjetja). Oceni za gostoti se pri dolgem Kromatinu dokaj dobro ujemata z obliko histograma okoli vrednosti 4, ter v desnem robu prikaza, drugod pa se obe gostoti minimalno odmakneta od histograma. Nekoliko manj smo lahko navdušeni nad ujemanjem grafov gostot s histogramom pri podatkih kratkega Kromatina. Tu je pomanjkanje ujemanja nekoliko posledica tudi tega, da pri izbiri števila blokov za ta primer število zaokrožimo navzgor. Če število blokov za ta primer ročno zmanjšamo za 1, pridemo do nekoliko lepšega ujemanja, vidnega na spodnji sliki.



Slika 5: Popravljen histogram za kratek Kromatin

Opazimo, da je tako nekoliko manj neujemanja, moteče je le pomanjkanje podatkov histograma na levem delu grafa, ter zato presežek podatkov na sredini histograma. Kljub temu smo lahko nad ujemanjem gostot (med seboj) in histograma pri vseh vrstah Kromatina zadovoljni in lahko trdimo, da so rezultati razumni.

3 Naloga 3

Naloga nam pove, da imamo podane mesečne temperature (torej 12 podatkov letno) za n zaporednih let (v našem primeru podatki med letoma 1986 in 2020, torej $n = 35$). Podnaloga a) nam narekuje, da podatke o meritvah zapišemo kot urejene pare, da se bomo lahko problema lotili z enostavno linearno regresijo. Opažene meritve so smiselno časovne urejene (leto-mesec), zato si podatke lahko predstavljamo kot pare $(x_i, y_i)_{i=1, 2, \dots, 12n}$, kjer x_i predstavlja časovno komponento problema, y_i pa opaženo temperaturo pri časovno i -ti meritvi. V resnici lahko le pretvorimo podatek oblike: $(leto, mesec, temperatura)$ v par $(leto + mesec/12 - 1986, temperatura)$ (tu je -1986 le zato, ker je to leto našega prvega podatka, lahko bi brez tega). Sedaj označimo:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{12n} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{12n} \end{bmatrix} \text{ in } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

V duhu enostavne linearne regresije iščemo β_0 in β_1 , tako da se bo $X\beta$ po metodi najmanjših kvadratov najboljše prilegal Y . Teorija nam pove, da lahko β_0 in β_1 ocenimo s cenilkama:

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^{12n} x_i^2\right) \left(\sum_{i=1}^{12n} y_i\right) - \left(\sum_{i=1}^{12n} x_i\right) \left(\sum_{i=1}^{12n} x_i y_i\right)}{12n \sum_{i=1}^{12n} x_i^2 - \left(\sum_{i=1}^{12n} x_i\right)^2}$$

$$\hat{\beta}_1 = \frac{12n \sum_{i=1}^{12n} x_i y_i - \left(\sum_{i=1}^{12n} x_i\right) \left(\sum_{i=1}^{12n} y_i\right)}{12n \sum_{i=1}^{12n} x_i^2 - \left(\sum_{i=1}^{12n} x_i\right)^2}$$

Literatura

- [1] E. Zakrajšek, *Verižnica*, [ogled 30. 6. 2023], dostopno na https://ucilnica.fmf.uni-lj.si/pluginfile.php/8283/mod_resource/content4/predavanja/veriznica/veriznica.pdf.
- [2] M. Raič *Zapiski predavanj- Izražava povprečja in variance pri podatkih, organiziranih po skupinah* [ogled 3. 7. 2023], dostopno na https://ucilnica.fmf.uni-lj.si/pluginfile.php/135261/mod_resource/content/2/Razcep_var.pdf.