



Projektna naloga pri Statistiki

Akademsko leto 2025/26

Luka Urbanč
vpisna številka 27231152
Mentor: doc. dr. Martin Raič

Ljubljana
3. avgust 2026

Kazalo

1	Prva naloga	3
1.1	Polni podatki	5
1.2	Odstranjeni osamelci	9
1.3	Stratificirano vzorčenje	10
1.4	Izboljšave in komentarji k prvi nalogi	18
2	Druga naloga	22
2.1	Ocena za θ po metodi momentov	22
2.2	Nepristranskost cenilke in standardna napaka	23
2.3	Ocena po metodi največjega verjetja	25
2.4	Nepristranskost cenilke in ocena napake	27
2.5	Bayesov pogled na θ	29
2.6	Spremenljena apriorna porazdelitev	32
2.7	Drugačen pogled na prejšnjo nalogo	34
2.8	Povzetek druge naloge	36
2.9	Izboljšave in komentarji k drugi nalogi	37
3	Tretja naloga	39
3.1	Ocenjevanje z linearno regresijo	40
3.2	Ocenjevanje s korensko regresijo	45
3.3	Logaritemska regresija	48
3.4	Testiranje regresijskih modelov s hipotezo	49
3.5	Kateri način je boljši	50
3.6	Izboljšave in komentarji k tretji nalogi	53
4	Zaključek in povzetek	55
4.1	Del izpisa za prvo nalogo	56
4.2	Del izpisa za drugo nalogo	56
4.3	Del izpisa za tretjo nalogo	57

Uvod

V današnjem svetu prevladujejo podatki. Za vsako pomembno odločitvijo, ki bi jo nekoč sprejeli na podlagi navdiha, suma ali pa (morda slabo) utemeljenega prepričanja, danes stoji veliko statistične analiza. Ker se bo to v prihodnosti še povečevalo je pravilno, da se s tem srečamo čim prej v življenju. Tako bomo tudi sami znali sprejemati boljše odločitve in iz navidez nepomembnih podatkov izpeljali koristne povezave.

V drugem semestru tretjega letnika študija matematike na Univerzi v Ljubljani je del obveznosti pri predmetu Statistika tudi izdelava projektne naloge, v kateri predstavimo rešitve za zastavljene probleme. Ti pokrivajo različna statistična področja in so namenjena preverjanju razumevanja in uporabe osvojenih znanj. Ker naloge zajemajo tudi snovi, ki jih pri predmetu nismo obravnavali, projektna naloga služi tudi kot medij za učenje novih konceptov pri statistiki.

Za projektno nalogo sem dobil tri naloge. Prva naloga pokriva vzorčenje in omenja intervale zaupanje, druga naloga je naloga iz pridobivanja cenilk in njihovi nepristranskosti, kjer je dodatna vsebina Bayesov pogled na ocenjevanje parametra preko apriornih in aposteriornih porazdelitev. Tretja naloga zajema linearno regresijo in preverjanje hipotez.

Vsaka naloga ima svoje poglavje, podpoglavja pa praviloma ustrezajo zastavljenim podnalogam. Na koncu vsakega poglavja je še podpoglavje, namenjeno komentarjem in morebitnim izboljšavam. Kljub želji po takšni strukturi, temu ni vedno tako, saj sem v poročilo vključil tudi nekatere dodatne naloge (npr. ocenjevanje s korenskim in logaritemskim modelom pri tretji nalogi), kar malenkost kvari zgoraj navedeno okvirno strukturo.

Čeprav je veliko funkcij že vgrajenih v programske jezike in knjižnice, sem vseeno poskušal večino stvari izvesti ročno. Primer tega je iskanje linearne premice, ki se podatkom najbolj prilega: lahko bi uporabil vgrajene funkcije, a sem raje sam izvedel vso linearno algebro, saj se tako sklada s povedaním na predavanjih. Kljub temu sem za nekatere teste in nekatere metode vseeno klical vgrajene funkcije, saj se mi ne zdi ključno, da sam definiram porazdelitveno funkcijo za beta porazdelitev.

1 Prva naloga

Prva naloga projektne naloge je analiza povprečnega dohodka v mestu *Kibergrad*. Za začetek naloge si najprej pogledjmo, katere podatke imamo na voljo. Izsek iz tabele podatkov je prikazan spodaj.

Indeks	Tip	Članov	Otrok	Dohodek	Četrt	Izobrazba
25351	3	2	0	20902	3	42
21635	3	2	1	28288	3	40
23013	3	2	1	17000	3	39
18265	1	2	0	25961	2	39
2188	1	2	0	24880	1	39

Tabela 1: Vzorec iz tabele `kibergrad.csv`

V tabeli se torej nahajajo podatki o družinah v mestu *Kibergrad*. Na voljo imamo podatke o tipu družine, številu članov v družini, številu otrok in dohodku. Poleg tega imamo zabeleženo tudi oceno izobraženosti družine in v kateri četrti živi družina. Pri tem so podatki zakodirani z naslednjim ključem:

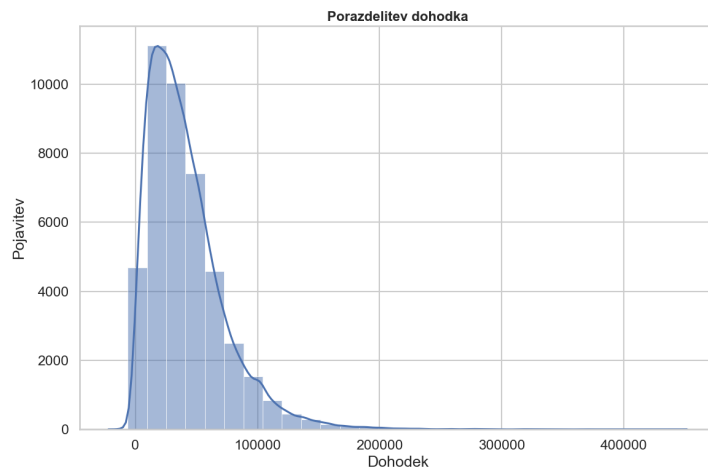
1 $\mapsto$ Severna četrt

2 $\mapsto$ Vzhodna četrt

3 $\mapsto$ Južna četrt

4 $\mapsto$ Zahodna četrt.

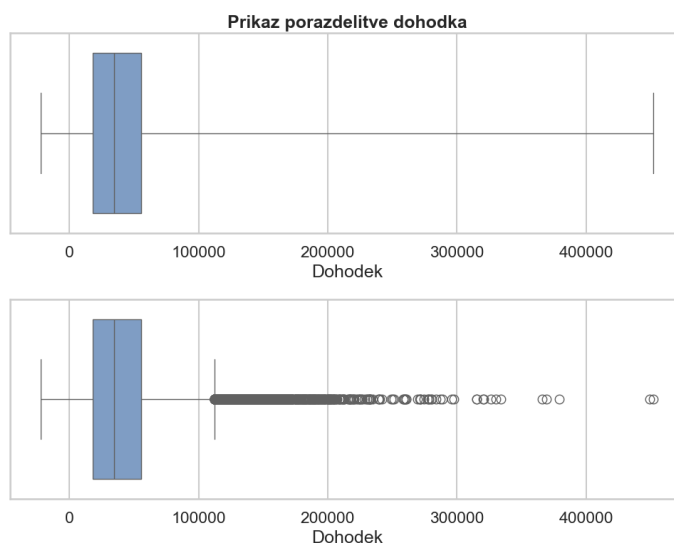
Dana naloga je, da na podlagi slučajnega vzorca ocenimo povprečen dohodek v *Kibergradu*. Ker nas zanima samo dohodek družine, si pogledjmo histogram dohodkov, ki nam bo povedal veliko o obliki porazdelitve in celotni populaciji nasploh.



Slika 1: Histogram dohodnine v *Kibergradu*

Na sliki 1 zgoraj izstopata dve dejstvi. Najprej opazimo, da je porazdelitev dohodka daleč od normalne porazdelitve. Kljub temu bomo zaradi pomanjkanja informacij o pravi porazdelitvi dohodka te naloge vseeno uporabljali normalno porazdelitev pri iskanju intervalov zaupanja. To predpostavko o normalni porazdeljenosti dohodka bi lahko zamenjali s katero drugo porazdelitvijo in bi postopek ostal relativno podoben. V primeru, da bi poznali porazdelitev dohodka, pa bi seveda uporabljali to in predpostavke ne bi bile potrebne.

Druga lastnost, ki jo opazimo, je, da obstajajo družine z neizmerno velikim dohodkom, kar vidimo na grafu kot zelo dolg desni rep. Drug dokaz za osamelce, ki bodo ovirali analizo, sta naslednja prikaza oz. "škatli z brki".



Slika 2: Škatla z brki za dohodek

Na zgornjih grafih je prisotnost osamelcev še bolj očitna. Zgornji izmed obeh prikazov ima nastavljene brke na minimum in maksimum dohodka, medtem ko jih ima spodnji nastavljena glede na vrednost $1.5 \times \text{IQR}$, torej je levi brk nastavljena na vrednost, ki je za $1.5 \times \text{IQR}$ manjša od prvega kvantila, desni brk pa na vrednost, ki je za $1.5 \times \text{IQR}$ večja od tretjega kvantila.

Drug način nam tako označi tudi vse točke, ki padejo izven tega razpona in jih zato lahko odstranimo iz analize. Vidimo tudi, da bomo na tak način odstranjevali zgolj podatke, ki so bistveno večji od večine, medtem ko bistveno manjših dohodkov ni, zato bo vpliv na povprečje velik. Lotimo se sedaj analize z osamelci in analize brez osamelcev.

Dodatna prednost uporabe škatle z brki je, da zaznamo tudi negativne dohodke, kar iz grafa na sliki 1 zaradi izbrane širine posameznega stolpca ni razvidno. Škatli z brki na zgornjem grafu povesta torej, da obstajajo ekstremno visoki dohodki, ki bodo nesorazmerno vplivali na povprečje. Meje, katere podatke obdržimo in katere ne, običajno določamo glede na primer, ki ga obravnavamo. Na

predavanjih smo sicer povedali zgornji način z interkvantilnim razponom, kar pa zdaleč ni edini način. Lahko bi na vsaki strani odvzeli recimo 10% največjih in najmanjših podatkov, ali pa bi glede na nesimetričnost porazdelitve na različnih koncih odstranili različno število podatkov. To se na nek način tudi zgodi zgoraj, ko uporabimo IQR za določanje mej, saj zaradi nesimetričnosti porazdelitve najmanjših dohodkov ne odstranjujemo.

Spodaj je prikazana še tabela povprečnega dohodka po četrtih, kar poveča preglednost reševanja naloge.

Četrto	Severna	Vzhodna	Južna	Zahodna	Vse
Dohodek	45758.54	41234.77	37473.28	42157.68	41335.51

Tabela 2: Povprečen dohodek glede na četrti v mestu.

1.1 Polni podatki

Histogram za podatke o dohodku smo si ogledali že zgoraj, zato ga na tem mestu spustimo in se lotimo ocenjevanja povprečnega dohodka na podlagi slučajnega vzorca populacije. V ta namen si najprej izberemo popolnoma naključen vzorec velikosti $n = 400$ enot. Tu smo vzorec izbrali enakomerno iz celotne populacije, torej so vsi možni vzorci brez ponavljanja imeli enako verjetnost, da so bili izbrani. Prav v temu se skriva vir slučajnosti pri slučajnih izborih. Naj X_i označuje dohodek i -te enote. Tedaj lahko za ta slučajni vzorec n enot izračunamo povprečen dohodek kot

$$\hat{\mu} = \bar{X} = \frac{X_1 + \dots + X_n}{n}.$$

Dejstvo, da je to dobra cenilka za povprečen dohodek, sledi iz tega, da je tako definiran $\hat{\mu}$ cenilka za μ po metodi momentov. V splošnem je po metodi momentov cenilka za $\mathbb{E}(X^k)$ kar

$$m_k = \frac{X_1^k + X_2^k + \dots + X_n^k}{n} = \overline{X^k}.$$

Dodatno zaradi predpostavke normalne porazdeljenosti velja, da je tak $\hat{\mu}$ tudi cenilka po metodi največjega verjetja. O cenilkah po metodi največjega verjetja in po metodi momentov bomo več povedali pri drugi nalogi, saj je to njena osrednja tema. Dodatno je taka cenila res dobra, saj ob dodatnih predpostavki neodvisnosti velja naslednji izrek.

Izrek 1.1 (Šibki zakon velikih števil). *Naj bodo $X_1, X_2, \dots, X_n$ neodvisne in enako porazdeljene realne slučajne spremenljivke z $\mathbb{E}(X_1) = \mu$ in $\mathbb{E}(X_1^2) < \infty$. Tedaj velja*

$$\frac{X_1 + X_2 + \dots + X_n}{n} \xrightarrow[n \rightarrow \infty]{d} \mu$$

Preden se lotimo reševanja naloge, se spodobi, da navedeni izrek tudi dokažemo, za njegov dokaz pa potrebujemo še dve trditvi.

Izrek 1.2 (Neenačba Markova). *Za nenegativno slučajno spremenljivo W in poljuben $\alpha > 0$ velja*

$$P(W \geq \alpha) \leq \frac{\mathbb{E}(W)}{\alpha}.$$

Dokaz. Opazimo, da velja $\mathbb{1}(w \geq \alpha) \leq \frac{w}{\alpha}$. Na tem izrazu uporabimo pričakovano vrednost in izrek sledi. $\square$

Lema 1.3. *Naj bodo $W_1, W_2, \dots$ take nenegativne slučajne spremenljivke. Če je $p > 0$ in velja*

$$\lim_{n \rightarrow \infty} \mathbb{E}(W_n^p) = 0,$$

tedaj velja tudi $W_n \xrightarrow[n \rightarrow \infty]{d} 0$.

Dokaz. Naj bo $r > 0$. Tedaj je

$$P(W_n \geq r) = P(W_n^p \geq r^p) \leq \frac{\mathbb{E}(W_n^p)}{r^p} \xrightarrow[n \rightarrow \infty]{} 0,$$

kjer smo za oceno uporabili neenačbo Markova. $\square$

Sedaj lahko dokažemo šibki zakon velikih števil.

Dokaz. Dovolj je pokazati, da gre

$$\frac{X_1 + X_2 + \dots + X_n}{n} - \mu \xrightarrow[n \rightarrow \infty]{d} 0,$$

zato izračunamo

$$\begin{aligned} \mathbb{E} \left[\left(\frac{X_1 + X_2 + \dots + X_n}{n} - \mu \right)^2 \right] &= \text{var} \left(\frac{X_1 + X_2 + \dots + X_n}{n} \right) \\ &= \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n} \xrightarrow[n \rightarrow \infty]{} 0. \end{aligned}$$

Trditev sledi iz zgornje leme. Pokazali smo, da je za $p = 2$ limita nič, kar po lemi implicira, da tudi spremenljivke konvergirajo v porazdelitvi proti ničli. $\square$

Šibki zakon velikih števil nam ob predpostavki neodvisnosti in enake porazdeljenosti, kot tudi nenegativnosti, zagotovi, da bo cenilka $\hat{\mu}$ konvergirala k pravi vrednosti, kar pa si želimo. V našem primeru dobimo vrednost

$$\hat{\mu} = 41511.67,$$

kar je precej blizu dejanskega populacijskega povprečja, ki znaša

$$\mu = 41335.51,$$

bolj natančne rezultate pa bi lahko dobili, če bi le vzeli večji vzorec. Ocenimo lahko tudi standardni odklon kot

$$\hat{\sigma}_+^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Izbira za nepristransko oceno σ^2 je naravna, saj smo že izračunali povprečje, zato je zgolj $n-1$ prostih spremenljivk. Imamo torej oceni za $\hat{\mu}$ in $\hat{\sigma}_+^2$. Pri naši oceni želimo tudi oceno, za koliko smo se največ zmotili. V ta namen ponavadi pogledamo še intervale zaupanja za ocenjeno statistiko. Želimo poiskati tako δ , da je

$$P(\mu \in (\bar{X} - \delta, \bar{X} + \delta)) \geq 1 - \alpha,$$

kjer je α naša stopnja tveganja. Spomnimo se definicije Studentove porazdelitve.

Definicija 1.4. Studentova porazdelitev s p prostostnimi stopnjami je porazdelitev slučajne spremenljivke

$$T = \frac{Z}{\sqrt{G}},$$

kjer je $Z \sim N(0, 1)$ in $G \sim \Gamma(\frac{p}{2}, \frac{p}{2})$ in sta Z in G neodvisni.

Spomnimo se tudi, da iz definicije Γ sledi, da je potem $pG \sim \Gamma(\frac{p}{2}, \frac{1}{2}) = \chi^2(p)$. Vemo že, da je

$$\frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1),$$

a ker prave σ ne poznamo, nas zanima, kaj je

$$T = \frac{\bar{X} - \mu}{\sigma_+} \sqrt{n}.$$

Ker je

$$\sum_{i=1}^n (X_i - \bar{X})^2 \sim \Gamma\left(\frac{n-1}{2}, \frac{1}{2\sigma^2}\right),$$

je potem

$$\frac{\sigma_+^2}{\sigma^2} = \frac{1}{\sigma^2(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \Gamma\left(\frac{n-1}{2}, \frac{\sigma^2(n-1)}{2\sigma^2}\right) = \Gamma\left(\frac{n-1}{2}, \frac{n-1}{2}\right).$$

Potrebujemo še neodvisnost obeh spremenljivk. Pri tem je dovolj, da sta neodvisni že $\bar{X}$ in $\sum (X_i - \bar{X})^2$, saj konstantni faktorji ne vplivajo na neodvisnost. Ker so vsi $X_i \sim N(\mu, \sigma^2)$, je potem vektor porazdeljen

$$X = (X_1, \dots, X_n)^T \sim N(\mu \mathbf{1}_n, \sigma^2 I).$$

Definiramo

$$e_1 = \frac{1}{\sqrt{n}}(1, \dots, 1)^T$$

in ga razširimo do ortonormirane baze $e_1, e_2, \dots, e_n$. Tedaj za ortogonalno matriko H , ki ima bazne vektorje po vrsticah, velja, da je $HH^T = H^TH = I_n$. Tedaj postavimo $Y = HX$. Ker je H linearna transformacija, vemo, da je

$$Y \sim N(H\mu 1_n, \sigma^2 HH^T) = N(H\mu 1_n, \sigma^2 I_n),$$

kar pa pomeni, da je kovariančna matrika spremenljivke Y ničelna izven diagonale. Z drugimi besedami je

$$\text{Cov}(Y_i, Y_j) = 0,$$

če sta le i in j različna. Ker so vsi Y_i normalni in nekorelirani, so neodvisni. Dalje lahko zaradi lastnosti H izračunamo Y_1 . Velja namreč

$$Y_1 = e_1^T X = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i = \sqrt{n}\bar{X},$$

oziroma

$$Y_1^2 = n\bar{X}^2.$$

Ker je H tudi ortogonalna, je potem tudi $\|X\|^2 = \|Y\|^2$, torej je

$$\sum_{i=1}^n X_i^2 = \sum_{i=1}^n Y_i^2.$$

Na tem mestu upoštevamo, da je

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

da dobimo

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2.$$

Dobili smo torej, da je Y_1 neodvisna od $(Y_2, \dots, Y_n)$ zaradi diagonalnosti kovariančne matrike. Ker je tudi

$$Q = \sum_{i=2}^n Y_i^2$$

izraz, ki vsebuje zgolj $Y_2, \dots, Y_n$, je Y_1 tudi neodvisna od Q . Ker pa je $Y_1 = \bar{X}\sqrt{n}$ in konstantne ne vplivajo na neodvisnost, smo pokazali torej, da sta $\bar{X}$ in $\sum_{i=1}^n (X_i - \bar{X})^2$ res neodvisna.

To pomeni, da je spremenljivka

$$T = \frac{\bar{X} - \mu}{\sigma_+} \sqrt{n}$$

porazdeljena s Studentovov porazdelitvijo z $n - 1$ prostostnimi stopnjami.

Vrnimo se nazaj na interval zaupanja. Poiščemo ga lahko tako, da nastavimo

$$P\left(F_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right) \leq \frac{\bar{X} - \mu}{\sigma_+} \sqrt{n} \leq F_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha,$$

kar pa z malo preurejanja vrne neenakost

$$\bar{X} - \frac{\hat{\sigma}_+}{\sqrt{n}} F_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right) \leq \mu \leq \bar{X} + \frac{\hat{\sigma}_+}{\sqrt{n}} F_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Ko za naš primer vstavimo konkretne številke, dobimo, da za povprečni dohodek v mestu *Kibergrad* s stopnjo tveganja $\alpha = 0.05$ in vzorcem $n = 400$ enot velja:

$$38123.68 \leq \mu \leq 44899.65,$$

kar se sklada tudi z dejanskim povprečjem celotne populacije. Določimo še standardno napako za oceno povprečja. Ker so vsi X_i porazdeljeni enako in so po predpostavki med sabo neodvisni, je

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{\sigma^2}{n}.$$

Težava pri nas je, da dejanske vrednosti σ^2 ne poznamo, lahko pa jo nadomestimo z nepristransko cenilko za σ^2 , ki je oblike

$$\hat{\sigma}_+^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Tedaj je naša standardna napaka

$$\text{SE}(\bar{X}) = \text{Var}(\bar{X}) = \frac{\sigma_+^2}{n}.$$

Ko vstavimo še to vrednost, dobimo da je

$$\text{SE}(\bar{X}) = 1723.35.$$

1.2 Odstranjeni osamelci

Histogram za dohodek po odstranjenih osamelcih pa zgleda že precej drugače. Še vedno imamo asimetrično porazdelitev okoli povprečja, zato predpostavka o normalni porazdelitvi še vedno ni točna, je pa porazdelitev mnogo bolj enakomerna. Vidimo tudi, da sta se modus in mediana ustrezno spremenila, kar je seveda pomembno.

Pri podatkih, kjer smo odstranili osamelce, sledimo enakemu postopku kot zgoraj in dobimo, da je

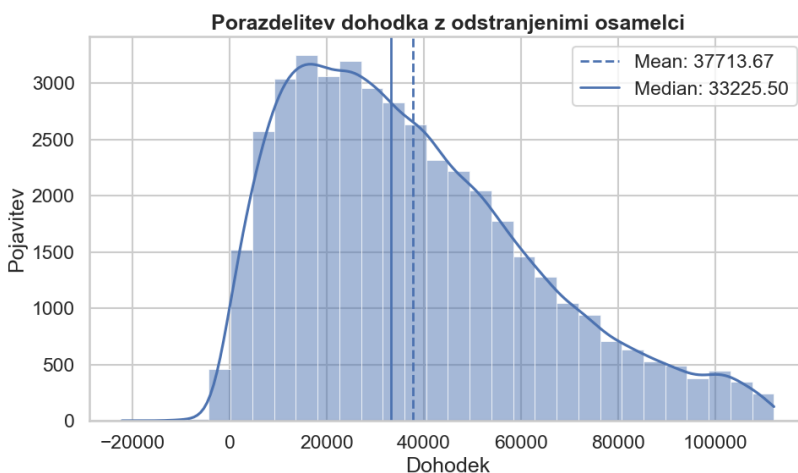
$$\hat{\mu} = 36035.58,$$

za interval zaupanja pri enakih pogojih pa za povprečen dohodek velja, da je

$$33559.87 \leq \mu \leq 38511.28,$$

ocena za standardno napako pa je

$$SE(\bar{X}) = 1259.31.$$



Slika 3: Histogram dohodnine z odstranjeni osamelci

1.3 Stratificirano vzorčenje

Do sedaj smo ocenjevali povprečni dohodek tako, da smo iz celotne populacije vzeli vzorec družin brez omejitev. Sedaj nas zanima, kako se rezultat spremeni, če izbiramo stratificirano glede na četrti, v katerih družine živijo.

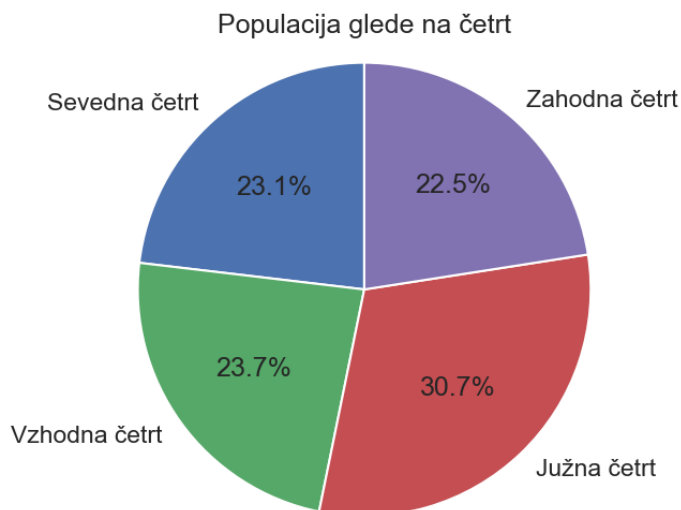
V splošnem je stratificiranje postopek deljenja enote neke populacije v homogeno skupino, na katerih potem posebej izbiramo vzorce. Strata, torej razdelitev enot v skupine, mora tvoriti particijo celotne populacije, kar pomeni, da vsaka enota spada v natanko eno skupino. Načinov, da populacijo razdelimo v skupine, je seveda več, a sta dva prevladujoče.

Prvi, *proporcionalna alokacija*, je določanje števila izbranih enot iz skupine glede na delež celotne populacije, ki jo skupina predstavlja. Tako imajo skupine, v katere spada več ljudi, večji vpliv na končno oceno. Poglejmo si ta pojem še na primeru. Recimo, da nas zanima ocena glasov pri nekih volitvah in vzorce izbiramo v treh različnih mestih. Prvo mesto ima 10000 gradbenikov, drugo mesto ima 20000 poslovnih, tretje mesto pa ima 30000 upokojencev. Ker bi se pri naključnem vzorcu iz celotne populacije lahko zgodilo, da bi bilo neko mesto bolje zastopano kot ostali dve, izberemo iz vsakega mesta ustrezen vzorec glede na delež, ki ga to mesto predstavlja. Torej če bi želeli izbrati 600 enot za oceno, bi izbrali 100 enot iz prvega mesta, 200 enot iz tretjega mesta in

300 enot iz zadnjega mesta. Tako so populacije pravično zastopane v izbranem vzorcu, glede na njihovo velikost, česar pri naključnem izboru 600 oseb ne bi mogli trditi.

Drug način stratifikacije pa je *optimalna ali disproporcionalna alokacija*. Tudi pri tem načinu izbiramo vzorce iz posameznih skupin glede na delež celote, ki spada v to skupino, toda dodatno upoštevamo še standardni odklon pri izbiri. Večje izbore namreč izberemo tam, kjer je varianca največja, saj tako dobimo najmanjšo skupno varianco.

V našem primeru bomo poskusili s pomočjo proporcionalne alokacije izboljšati oceno za povprečen dohodek tako, da stratificiramo po četrtih. Vsako družino bomo torej razvrstili glede na četrt, v kateri živi, nato pa bomo glede na velikosti četrti iz vsake skupine izbrali ustrezno število enot. Iz tortnega grafikona na sliki 4 razberemo, da so četrti razmeroma enako velike, zgolj južna četrt izstopa kot opazno večja.



Slika 4: Velikosti četrti

Ker izbiramo stratificirano po četrtih, bomo torej v južni četrti izbrali nekoliko več družin, iz ostalih pa bomo izbrali približno enako enot. Rezultate izbora vidimo v spodnji tabeli.

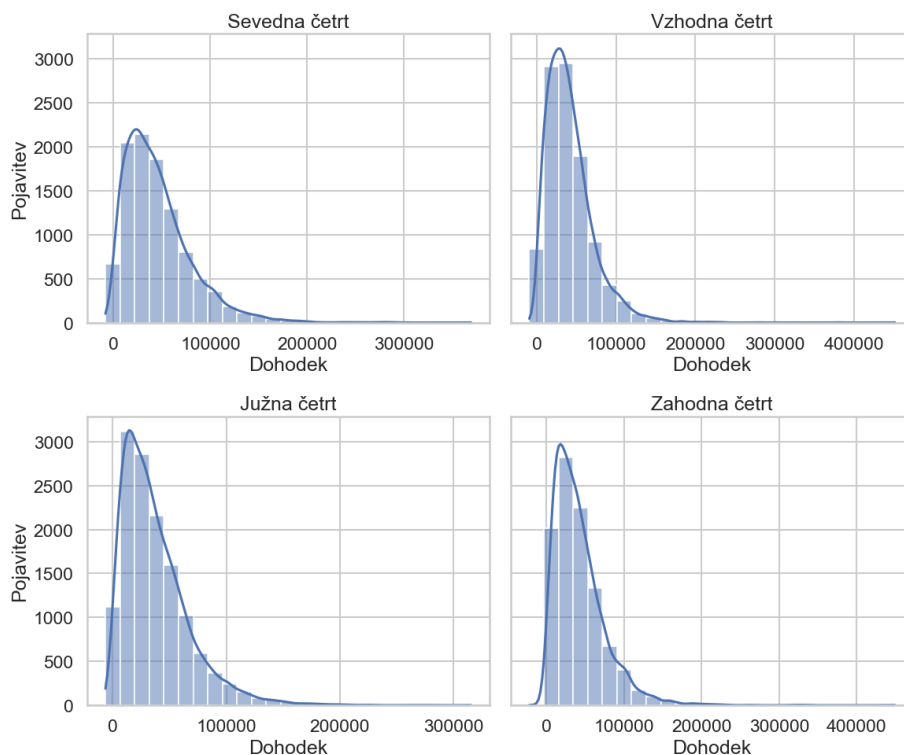
Četrt	Severna	Vzhodna	Južna	Zahodna	Vsota
Število družin	10149	10390	13457	9890	43886
Delež celote	23.1%	23.7 %	30.7 %	22.5%	1
Izbranih enot	93	95	123	90	401

Tabela 3: Razporeditev družin v četrti

V zgornji tabeli smo število izbranih enot izračunali po formuli

$$\frac{\text{Velikost četrti}}{\text{Celotna populacija}} \cdot n,$$

ker pa smo število izbranih enot zaokroževali na cela števila, smo na koncu morali izbrati 401 enot namesto 400. Število izbranih družin se torej res ujema z našim premislekom zgoraj. Preden izračunamo ocene in dejanske vrednosti, si pogledjmo še histogram dohodka znotraj vsake četrti.



Slika 5: Histogram dohodkov v četrtih

Na zgornjem grafu vidimo, da so, z izjemo severne četrti, dohodki razmeroma enako porazdeljeni. Pri severni četrti so pojavitve nekoliko nižje, je pa zato histogram nekoliko bolj širok oziroma debel, kar pomeni, da je več pojavitvev pri višjih dohodkih. To lahko sklepamo tudi iz tega, da so pojavitve nižje kot pri vzhodni in zahodni četrti, čeprav so velikosti primerljive. Res pa je, da to histogram za severno četrt zaradi svoje drugačne skale v x smeri malenkost prekrije.

Vredno je omeniti tudi, da trenutno izbiramo na celotni populaciji, ki vsebuje vse ekstremne dohodke, ki lahko precej vplivajo na povprečja.

Sedaj ostane zgolj še izračunati povprečen dohodek vzorca in njegovo varianco. Ker sedaj gledamo stratificirano vzorčenje, se formuli spremenita. Velja torej

$$\bar{X} = \frac{1}{N} \sum_i N_i \bar{X}_i$$

in

$$\sigma_X^2 = \sum_i \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h - 1} \right) \frac{\sigma_i^2}{n_h},$$

kjer pa pogosto spustimo popravek zaradi končnih populacij. Tako je formula za varianco

$$\sigma_X^2 = \sum_i \left(\frac{N_h}{N} \right)^2 \frac{\sigma_i^2}{n_h}, \quad (1.1)$$

kjer vsote tečejo po vseh skupinah, N označuje velikost celotnega vzorca, N_i predstavlja velikost iz i -te skupina, $\bar{X}_i$ predstavlja povprečje znotraj ene skupine, n_k predstavlja število izbranih enot iz vsake skupine in σ_i^2 označuje varianco znotraj vsake skupine.

Zgornje formule odražajo dejstvo, da pred izborom enote sortiramo v skupine, iz katerih nato izbiramo. Tako povprečje potem postane uteženo povprečje, kar je smiselno: iz vsake skupine vzamemo vzorce, pogledamo njegovo povprečno vrednost, kar bo cenilka za povprečno vrednost celotne skupine. Nato vsakemu vzorčnemu povprečju dodamo še utež, ki predstavlja delež populacije, ki spada v izbrano skupino. S tem odražamo dejstvo, da izbiramo glede na skupine, znotraj skupine pa izbiramo enakomerno, kjer imajo vse enote enako možnost, da so izbrane.

Formulo za varianco za stratificirano vzorčenje lahko izpeljemo podobno kot smo izpeljali formulo za varianco pri slučajnem vzorčenju. Ker je $\hat{\mu}_{strat} = W_1 \bar{X}_1 + W_2 \bar{X}_2 + \dots + W_k \bar{X}_k$, je potem po pravilih za računanje z varianco

$$\text{var}(\mu_{strat}) = W_1^2 \text{var}(\bar{X}_1) + W_2^2 \text{var}(\bar{X}_2) + \dots + W_k^2 \text{var}(\bar{X}_k).$$

Vsaka varianca znotraj skupine je zgolj varianca, ki bi jo dobili pri enakomernem vzorčenju:

$$\text{var}(X_i) = \frac{\sigma_i^2}{n_i},$$

kjer nismo upoštevali popravka za končne populacije. Ker izbiramo resnično majhen del velike populacije, to smemo storiti. Popravek za končne populacije bi morali upoštevati, če bi izbirali velik delež populacije. Če izbiramo 40 ljudi iz populacije velikosti $n = 45$, je v tem res malo naključja in poznamo praktično celotno populacijo. V takem primeru bi bil potem

$$\text{var}(X_i) = \frac{\sigma_i^2}{n_i} \frac{N_i - n_i}{N_i - 1},$$

kar pa izpelje tudi formulo za varianco pri stratificiranem vzorčenju z dodatkom za končne populacije.

Smiselno se je vprašati tudi o varianci stratificiranega vzorčenja glede na varianco pri slučajnem vzorčenju. Poglejmo si to zvezo natančneje. Pri slučajnem vzorčenju je

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Ker bi to radi primerjali z varianco znotraj vsake skupine, potrebujemo neko zvezo med σ_l^2 in pa σ^2 . Velja

$$\sigma^2 = \frac{1}{N} \sum_l^L \sum_i^{N_l} (x_{il} - \mu)^2.$$

Dalje je tudi

$$(x_{il} - \mu)^2 = (x_{il} - \mu_l)^2 + 2(x_{il} - \mu_l)(\mu_l - \mu) + (\mu_l - \mu)^2,$$

kjer sredinski člen izgine, ko seštevamo po vseh l , saj je $N_l \mu_l = \sum_{i=1}^{N_l} x_{il}$. Tako dobimo

$$\begin{aligned} \sum_{i=1}^{N_l} (x_{il} - \mu)^2 &= \sum_{i=1}^{n_l} (x_{il} - \mu_l)^2 + N_l(\mu_l - \mu)^2 \\ &= N_l \sigma_l^2 + N_l(\mu_l - \mu)^2. \end{aligned}$$

Tedaj lahko obe strani enačbe delimo še z N in seštejemo po vseh l , da dobimo

$$\sigma^2 = \sum_{l=1}^L W_l \sigma_l^2 + \sum_{l=1}^L W_l (\mu_l - \mu)^2.$$

Ko to uporabimo v izrazu za varianco povprečja in združimo skupaj z enačbo (1.1), smo dokazali naslednji izrek.

Izrek 1.5. *Razlika variance povprečja pri slučajnem vzorčenju in variance povprečja pri stratificiranem vzorčenju s proporcionalno alokacijo je*

$$\text{var}(\bar{X}) - \text{var}(\bar{X}_{sp}) = \frac{1}{n} \sum_{l=1}^L W_l (\mu_l - \mu)^2,$$

če zanemarimo popravek zaradi končnih populacij.

Zgornji izrek torej pove, da je varianca pri slučajnem vzorčenju vedno večja od variance pri stratificiranem vzorčenju, kar je ena izmed prednosti stratificiranega vzorčenja s proporcionalno alokacijo.

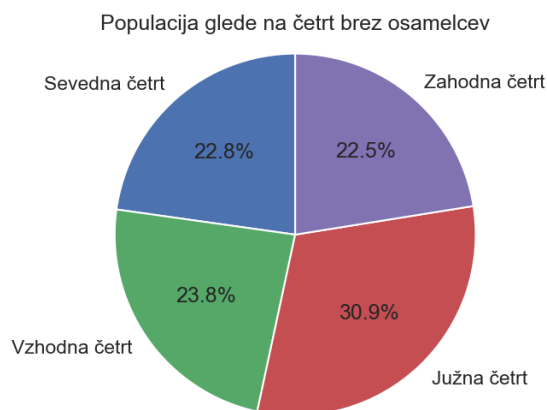
Težave pri stratificiranem vzorčenju so v praktični uporabi pogosto pri temu, da je potrebno vsako enoto enolično in z gotovostjo razdeliti v natanko eno skupino. Rado se zgodi, da podatkov nimamo na voljo, ali pa da posamezna enota ustreza več skupinam. Kljub temu pa s takšnim vzorčenjem pogosto bolje opišemo populacijo, saj je bolje predstavimo. Pri izbiri skupin je potrebno

tudi izbrati tisto, za katero sklepamo, da ima največji vpliv. Če analiziramo glasove na parlamentarnih volitvah, je smiselno deliti, recimo, glede na dohodek, malo manj smiselno pa je deliti glede na število prehojenih korakov vsakega posameznika. Primer je sicer pretiran in nerealističen, a kljub temu izpostavi potrebo po dobrem in premišljenem izboru delilnih skupin.

Sedaj ko poznamo izpeljavo formul, lahko vstavimo konkretne podatke in dobimo

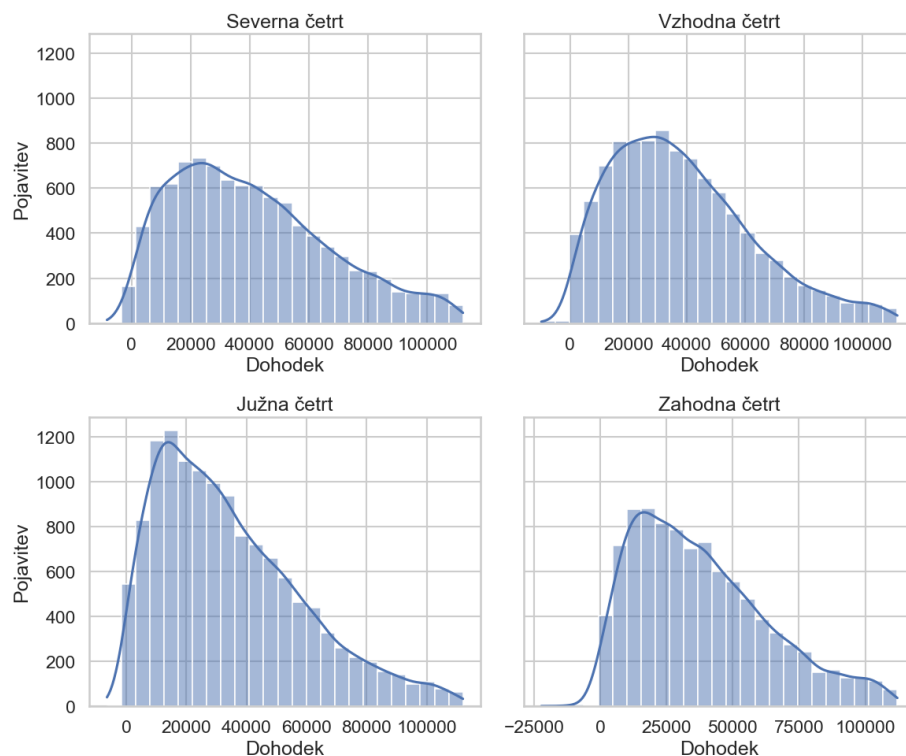
- $\bar{X} = 41553.36$,
- $38331.2 \leq \mu \leq 44775.51$,
- $SE(\bar{X}) = 1639.00$.

Podoben postopek lahko sedaj ponovimo še z odstranjenimi osamelci. Tedaj so deleži četrti malenkost spremenjeni, histogrami porazdelitve pa so mnogo bolj pregledni, kar bosta prikazala naslednja dva grafa.



Slika 6: Velikost četrti brez osamelcev

Na sliki 6 vidimo, da so deleži približno enaki kot prej. Splošno velja, da večji delež med filtriranimi podatki ustreza temu, da smo jih iz te skupine manj odstranili, manjši delež pa pomeni, da smo odstranili več podatkov. Velikost severne četrti se je zmanjšala za 0.3%, zato sklepamo, da smo v njej odstranili več podatkov, kot v zahodni četrti, kjer je delež ostal enak. Podoben razmislek lahko naredimo tudi pri ostalih četrtih.



Slika 7: Histogram dohodkov po četrtih brez osamelcev

Dalje pri histogramih lahko opazimo podobno kot v primeru, ko nimamo odstranjenih podatkov. Severna četrt izstopa po tem, da je precej bolj širok, kar nakazuje na več ljudi z višjim dohodkom. Res pa je, da smo z odstranjevanjem osamelcev opazno izboljšali preglednost histogramov.

Podobno kot v primeru z vsemi podatki, lahko sedaj oblikujemo tabelo, kjer predstavimo podatke proporcionalnega izbiranja.

Četrt	Severna	Vzhodna	Južna	Zahodna	Vsota
Število družin	9655	10108	13099	9526	42388
Delež celote	22.8%	23.8 %	30.9 %	22.5%	1
Izbranih enot	91	95	124	90	401

Tabela 4: Razporeditev družin v četrti po odstranjenih osamelcih

Iz zgornje tabele ob prebiranju števila izbranih enot znotraj vsake četrti vidimo, da se vrednosti praktično niso spremenile. Sklepamo torej, da odstranjevanje osamelcev ni imelo pretiranega vpliva pri takšni velikosti vzorca, res pa je, da so možnosti, da bi nenavadno veliki dohodki kvarili rezultate, sedaj manjše.

Ko vstavimo dejanske podatke, lahko izračunamo nove ocene za povprečen dohodek v mestu *Kibegrad*, kar navedemo spodaj

- $\bar{X} = 38422.11$,
- $35989.79 \leq \mu \leq 40854.42$,
- $SE(\bar{X}) = 1237.24$,

celotno nalogo pa povzamemo v spodnji tabeli:

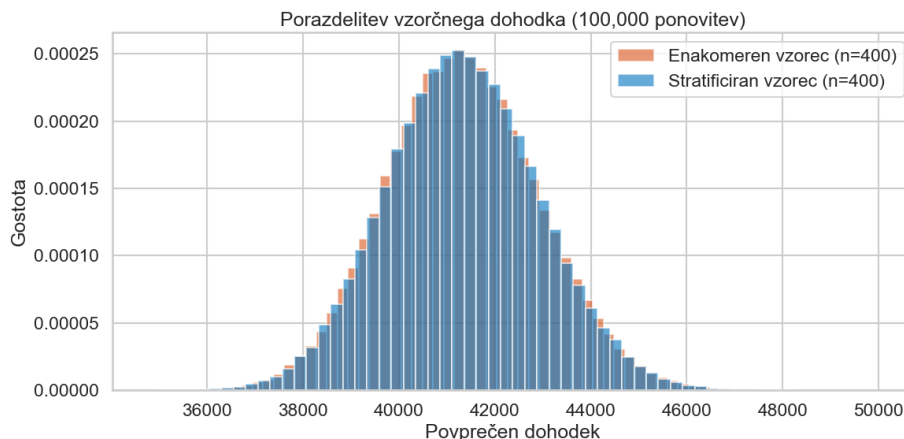
Način dela	Ocena za $\bar{X}$	Leva meja	Desna meja	$SE(\bar{X})$
Slučajen vzorec	41511.67	38123.68	44899.65	1723.35
Slučajen vzorec*	36035.58	33559.87	38511.28	1259.31
Proporcionalna alok.	41553.36	38331.2	44775.51	1639.00
Proporcionalna alok.*	38422.11	35989.79	40854.42	1237.24

Tabela 5: Rezultati ocen, kjer (*) označuje metodo, kjer smo odstranili osamelce.

V tabeli lahko opazimo kar nekaj stvari. Proporcionalna alokacija je v obeh primerih izboljšala standardno napako ocene. Pri odstranjenih osamelcih je proporcionalna alokacija tudi izboljšala oceno za $\bar{X}$, saj je dejanska vrednost okoli 37714. Drugače je bilo, ko nismo odstranili osamelcev, saj se je ocena, in s tem odstopanje od dejanske vrednosti, povečala. Prava vrednost je namreč okoli 41335. Je pa pri teh podatkih proporcionalna alokacija malenkostno izboljšala interval zaupanja v smislu, da je razpon manjši. Drugače je bilo, ko smo odstranili osamelce, saj je širina intervala ostala približno enaka, a je pri proporcionalni alokaciji sredina intervala blizu dejanske vrednosti, medtem ko je pri naključnem vzorcu prava vrednost že blizu roba intervala zaupanja.

Pri interpretaciji rezultatov je zelo pomembno, da se zavedamo, da je vse to lahko zgolj slučaj. Lahko bi namreč izbrali drugačen vzorec pri kakšnem ocenjevanju in dobili popolnoma drugačne rezultate. Lahko bi za naključen vzorec izbrali družine, ki so vse zelo blizu poprečnemu dohodku, pri proporcionalni alokaciji pa bi izbrali samo take družine, ki imajo dohodek nad 100000, kar bi seveda kvarilo rezultate. Poglejmo si, kako bi preverili to.

Izvedimo simulacijo, ki bo 100000 krat izbrala tako stratificiran vzorec kot tudi naključen vzorec, oba velikosti 400 družin. Nato za vsak vzorec ocenimo povprečen dohodek kot vzorčno povprečje. Nato narišemo histogram, ki ga normaliziramo na gostoto in ne na pojavitve, da lahko primerjamo porazdelitev povprečja pri obeh metodah. Rezultat simulacije je prikazan na spodnji sliki.



Slika 8: Rezultati simulacije za 10^5 ponovitev

Graf nam pove, da sta porazdelitvi zelo blizu, torej da ni velike prednosti pri uporabi stratificiranega vzorčenja. Obe sta centrirani okoli iste vrednosti, saj sta obe cenilki, tako s stratificiranim vzorčenjem kot z naključnim izborom, nepristranski cenilki. Zanimivo pa je, da je porazdelitev stratificiranega vzorčenja (modre barve) malenkost ožja od naključnega izbora (oranžne barve). Sklepamo torej, da je varianca manjša pri stratificiranem vzorčenju, a je razlika komaj opazna. To se sklada tudi z ugotovitvami pod zgornjo tabelo, kjer smo dobili ožje intervale zaupanja.

Izračunamo lahko tudi povprečni varianci za oba načina izbiranja. Za obe varianci dobimo vrednosti

$$\hat{\sigma}_{uni} = 2530797.32, \quad \hat{\sigma}_{strat} = 2512489.48, \quad \frac{\hat{\sigma}_{uni}}{\hat{\sigma}_{strat}} = 1.0073,$$

zaradi česar sklepamo, da je varianca stratificiranega vzorčenja manjša, kar razloži ugotovitve iz prejšnjega odstavka. S tem tudi numerično potrdimo, da je razlika res majhna, a vendar prisotna.

Razlog za take rezultate je v tem, da je razporeditev dohodka v vseh četrtih približno enaka, prav tako pa je razdelitev populacije v četrti razmeroma enakomerna. S tem mislim, da so četrti približno enako velike, zaradi česar je učinek stratifikacije manj opazen. Ker je tudi dohodek približno enako razporejen znotraj vsake četrti, pa je že tako majhen vpliv stratifikacije še manjši.

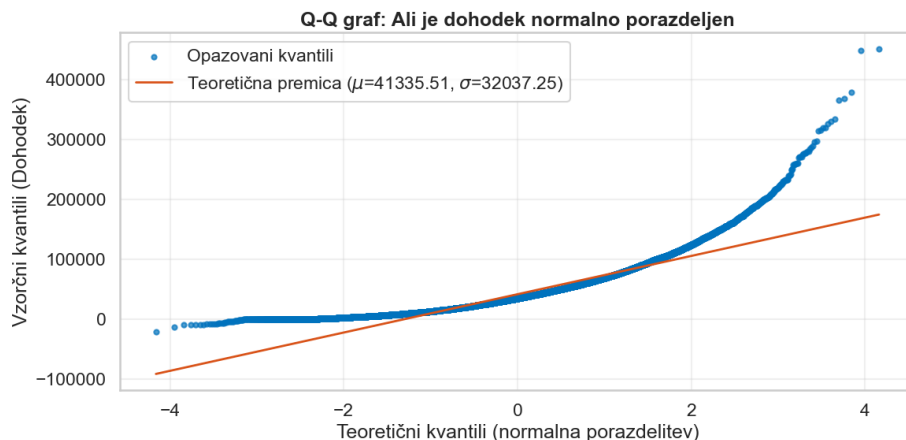
Mimogrede pa je graf simulacije lep primer delovanja Centralnega limitnega izreka, saj opazimo da imata obe porazdelitvi vzorčnih povprečji obliko normalne porazdelitve.

1.4 Izboljšave in komentarji k prvi nalogi

Tekom celotne naloge smo si pomagali s številnimi predpostavkami o podatkih. Predpostavili smo namreč, da so podatki normalno porazdeljeni in med sabo

neodvisni, kar seveda ni nujno res. Dejstvo, da niso normalno porazdeljeni, smo pokazali že na sliki 1, dodatno pa predpostavka o neodvisnosti ni nujno resnična. Prihodki so verjetno večinoma ustvarjeni s plačo, torej je odvisna od opravljenega dela, delovna mesta pa verjetno niso neodvisna med sabo. Dodano smo uporabljali izreke, ki zahtevajo nenegativnost slučajnih spremenljivk. Tega spet ne smemo trditi v našem primeru, saj imamo družine v mestu, ki imajo negativne dohodke.

Glede nenormalnosti porazdelitve poznamo tudi druge, bolj formalne načine, da to pokažemo. Tekom predavanj smo omenili tako imenovane Q-Q grafikone, kjer lahko primerjamo dve porazdelitvi med sabo. Če primerjamo porazdelitev dohodka v mestu *Kibegrad* z normalno porazdelitvijo, ki ima parametra $\mu = \bar{X}$ in $\sigma^2 = 1/n \sum_i (X_i - \bar{X})^2$, bi normalno porazdeljen dohodek ležal na premici na grafu. Na spodnji sliki je prikazan takšen Q-Q grafikon, kjer vidimo, da temu ni tako.



Slika 9: Q-Q grafikon za dohodek

Iz zgornje slike je ponovno razvidno, zakaj dohodek ne sledi normalni porazdelitvi. V tem primeru bi morala biti modra krivulja približno poravnana z oranžno premico, kar pa vidimo, da ni res. Tako na levem kot tudi na desnem koncu se točke oddaljijo, kar je pričakovano za dohodke. To se zgodi, saj negativnih dohodkov praktično ni, medtem ko negativne normalne porazdelitve seveda obstajajo. Ponavadi se tudi zgodi, da podatki o dohodkih vsebujejo tudi malo zelo visokih dohodkov, kar povzroči odklon navzgor v primerjavi z normalno porazdelitvijo.

Poleg tega lahko izvedemo tudi enega izmed številnih testov normalnosti. Po hitrem iskanju po internetu sem našel na ogromno testov, ki so namenjeni odkrivanju tega, ali je določena porazdelitev blizu normalne porazdelitve. Ker nobenih nisimo omenjal na predavanjih, o njih ne bom podrobno razlagal. Kljub temu bi rad omenil Shapiro-Wilksov (ki je močno povezan z odstopanjem v Q-Q grafikonih, kot vidimo zgoraj) in pa D'Agostin-Pearsonov

test, ki sta vgrajena v statistično knjižnico. V primeru prvega dobimo izračunano p -vrednost kar $8.6703 \cdot 10^{-100}$, a nam program javlja naslednjo opozorilo: `UserWarning: scipy.stats.shapiro: For N > 5000, computed p-value may not be accurate. Current N is 43886`. V ta namen temu rezultatu, čeprav se sklada z našim pričakovanjem, ne zaupamo popolnoma. Izvedemo lahko še drugi test, kjer pa dobimo p -vrednost 0.000, kar je dovolj, da tudi s tem testom potrdimo hipotezo, da dohodek ni porazdeljen normalno.

Ker smo pri zgornjih testih uporabili pojem p -vrednosti, je smiselno, da ga tudi razložimo. Ta vrednost je verjetnost, da bi testna statistika pokazala vsaj tako ekstremne vrednosti, kot smo jih opazili. Če je $p = 0.05$, to pomeni, da je vrednost naše testne statistike tako velika ali majhna, da jih s stopnjo tveganja 0.05 nismo dobili zaradi naključja. Dejansko gre tudi za identičen pristop, kot smo ga spoznali na predavanjih za potrjevanje in zavračanje hipotez, le da smo v našem primeru poiskali robno vrednost za dano stopnjo tveganja, pri p -vrednosti pa glede na opažene podatke pridobimo mero ekstremnosti opažanj. V našem primeru imamo predpisano stopnjo tveganja in računamo vrednost naših opažanj, v primeru p -vrednosti pa glede na naša opažanja določimo njihovo ekstremnost. Dogajanje nas spomni na pregovor, da gledamo zgolj dve plati istega kovanca.

Zanimivo bi bilo pogledati tudi, kako se povprečja spremenijo z odstranjevanjem osamelcev. Zaradi veliko večjih dohodkov bi bil vpliv lahko opazen. Dodatno bi se splačalo raziskovati, katere dohodke izločiti, saj smo s trenutnim pristopom izločili kar precej podatkov. Naloge bi lahko torej reševali tako, da bi odstranili recimo 1% največjih dohodkov, obdržali pa bi ostale.

Prav tako so na mestu številni komentarji glede histogramov. Pri interpretaciji velikosti v histogramih je pomembno imeti v mislih, da vtis velikosti tvori njegova ploščina in ne njegova višina. To smo videli in komentirali pri histogramu na sliki 5, kjer je histogram za severno razmeroma nizek, a je povprečni dohodek znotraj četrti znatno večji od ostalih. Poleg tega sem pri risanju histogramov uporabljal fiksno število stolpcev. To ni idealno za poljubne podatke, saj bi lahko dobili nepregleden graf. Kljub temu da sem uporabljal vrednosti 25, 30 in ponekod 60, so stolpci zaradi velikega števila družin še vedno dovolj visoki, grafi pa informativni in pregledni. Poleg tega s tem omogočamo večjo konsistenco, ko na isti sliki prikazujemo številne histograme.

To bi lahko spremenili in upoštevali napotke s predavanj. Spoznali smo namreč Scottovo pravilo za širino stolpcev histograma. Naj bo $\hat{f}(x)$ histogramska, torej odsekoma linearna, aproksimacija neke funkcije $f(x)$. Ker pri takem prikazu želimo sklepati o porazdelitvi oziroma osnovni funkciji $f(x)$, je smiselno ciljati tako širino, in s tem število stolpcev, da minimiziramo integralsko povprečno kvadratično napako (IMSE), ki je definirana kot

$$\text{IMSE} = \mathbb{E} \left[\int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx \right].$$

S Taylorjevim razvojem izraza lahko pokažemo, da je optimalna izbira širine

$$h^* = \left(6 / \int_{-\infty}^{\infty} f'(x)^2 dx \right)^{1/3} n^{-1/3},$$

pri predpostavki, da je $f(x)$ porazdelitvena funkcija normalne porazdelitve, pa ima ta izraz obliko

$$h^* = (24\sqrt{\pi})^{1/3} \sigma n^{-1/3} \sim 3.5 \sigma n^{-1/3}.$$

Pravilo je uveljavljeno in ga imajo številni programski jeziki vgrajenega kot privzeto število stolpcev. Žal pa med njimi ni **Pythona** in njegove knjižnice **Pandas**, ki ima privzeto število stolpec nastavljeno na 10.

2 Druga naloga

Pri drugi nalogi imamo podano diskretno slučajno spremenljivko X , ki ima naslednjo porazdelitev:

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{2}{3}\theta & \frac{\theta}{3} & \frac{2}{3}(1-\theta) & \frac{1-\theta}{3} \end{pmatrix},$$

kjer je $0 \leq \theta \leq 1$ nek neznani parameter. Poleg tega imamo na voljo tudi deset opazanj te slučajne spremenljivke, in sicer:

$$3, 0, 2, 1, 3, 2, 1, 0, 2, 2. \quad (2.1)$$

2.1 Ocena za θ po metodi momentov

Preden se lotimo reševanja naloge in cinitve θ po metodi momentov, si najprej pogledajmo idejo pri tem načinu ocenjevanja. Spomnimo se najprej definicije momenta.

Definicija 2.1. Naj bo X slučajna spremenljivka. Količini $m_k = \mathbb{E}(X^k)$ pravimo *k-ti moment slučajne spremenljivke X* . Če označimo $\mu = \mathbb{E}(X)$, definiramo še $c_k = \mathbb{E}[(X - \mu)^k]$ kot *k-ti centralni moment slučajne spremenljivke X* .

Denimo torej, da imamo slučajno spremenljivko X , katere porazdelitev je na nek način odvisen od parametrov $p_1, p_2, \dots, p_n$. Pri metodi momentov želimo oceniti vrednost parametrov $p_1, p_2, \dots, p_n$ tako, da izrazimo momente slučajne spremenljivke X s parametri, ki jih iščemo, namreč $p_1, p_2, \dots, p_n$, ter nato ocenimo parametre z ocenami za uporabljene momente. Ponavadi potrebujemo toliko momentov, kot je število parametrov, ki jih želimo oceniti, zato lahko hitro vidimo, da bo za reševanje dane naloge dovolj že, da poiščemo zvezo med θ in $\mathbb{E}(X)$. Računajmo torej

$$\begin{aligned} \mathbb{E}(X) &= \sum_{i=0}^3 i P(X = i) = \\ &= 0 \cdot \frac{2\theta}{3} + 1 \cdot \frac{\theta}{3} + 2 \cdot \frac{2}{3}(1-\theta) + 3 \cdot \frac{1-\theta}{3} = \\ &= \frac{\theta}{3} + \frac{4}{3} - \frac{4\theta}{3} + (1-\theta) = \\ &= \frac{7}{3} - 2\theta. \end{aligned} \quad (2.2)$$

Kljub temu da za oceno θ potrebujemo zgolj prvi moment, se bo kasneje izkazalo, da je koristno izračunati tudi drugi moment slučajne spremenljivke X .

Zapišemo

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{k=0}^3 k^2 P(X = k) = \\
&= 0 \cdot \frac{2\theta}{3} + 1^2 \cdot \frac{\theta}{3} + 2^2 \frac{2}{3}(1 - \theta) + 3^2 \cdot \frac{1 - \theta}{3} = \\
&= \frac{\theta}{3} + \frac{8}{3} - \frac{8}{3}\theta + 3 - 3\theta = \\
&= \frac{17 - 16\theta}{3}.
\end{aligned} \tag{2.3}$$

Ker je najboljša cenilka za $\mathbb{E}(X)$ kar $\overline{X}$, lahko iz opažanj θ ocenimo kot

$$\hat{\theta} = \frac{1}{2} \left(\frac{7}{3} - \mathbb{E}(X) \right).$$

Sedaj lahko upoštevamo dejstvo, da je

$$\overline{X} = \frac{3 + 0 + 2 + 1 + 3 + 2 + 1 + 0 + 2 + 2}{10} = \frac{16}{10} = \frac{8}{5},$$

zaradi česar je

$$\hat{\theta} = \frac{1}{2} \left(\frac{7}{3} - \frac{8}{5} \right) = \frac{7}{6} - \frac{4}{5} = \frac{11}{30} \approx 0.367.$$

S tem je prvi del naloge rešen.

2.2 Nepristranskost cenilke in standardna napaka

Spomnimo se definicije nepristranskosti.

Definicija 2.2. Naj bo X slučajna spremenljiva in $\hat{y} = h(X)$ cenilka za y . *Pristranskost cenilke $\hat{y}$ za y je* $\text{Bias}_{\theta}(\hat{y} \mid y) = \mathbb{E}(\hat{y}) - y = \mathbb{E}(\hat{y} - y)$.

Ko smo definirali pristranskost cenilke, je naslednja definicija intuitivna.

Definicija 2.3. Cenilka $\hat{y}$ za y je *nepristranska*, če za vsak $\theta \in \Theta$ velja, da je $\text{Bias}_{\theta}(\hat{y} \mid y) = 0$.

Za nepristranske cenilke velja tudi, da se njihova standardna napaka izraža kar z varianco cenilke. Naj bo $\hat{y}$ nepristranska cenilka za y . Tedaj vrednosti

$$\text{SE} = \sqrt{\text{var}(\hat{y})}$$

pravimo *standardna napaka*.

Pokazati moramo torej, da je $\text{Bias}(\hat{\theta} \mid \theta) = 0$ in izračunati $\text{var}(\hat{\theta})$. Najprej se lotimo nepristranskosti. Računamo

$$\begin{aligned}
\mathbb{E}(\hat{\theta}) &= \mathbb{E} \left(\frac{7}{6} - \frac{1}{2} \overline{X} \right) = \frac{7}{6} - \frac{1}{2} \mathbb{E} \left(\frac{X_1 + \cdots + X_n}{n} \right) \\
&= \frac{7}{6} - \frac{1}{2n} \sum_{i=1}^n \mathbb{E}(X_i),
\end{aligned}$$

kjer smo upoštevali linearnost pričakovane vrednosti. Ker so spremenljivke X_i realizacije iste porazdelitve so vsi X_i enako porazdeljeni, njihova pričakovana vrednost pa je po enačbi 2.2 enaka $\frac{7}{3} - 2\theta$. Iz enake porazdelitve sledi, da se n v zgornji enačbi krajša in ostane le

$$\mathbb{E}(\hat{\theta}) = \frac{7}{6} - \frac{1}{2} \left(\frac{7}{3} - 2\theta \right) = \theta.$$

Ko to vstavimo v definicijo nepristranskosti sledi, da je

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = \theta - \theta = 0,$$

torej je cenilka za θ res nepristranska. Za standardno napako potrebujemo varianco cenilke, zato najprej pogledimo naslednjo trditev.

Trditev 2.4. *Naj bodo $X_1, X_2, \dots, X_n$ neodvisne in enako porazdeljene slučajne spremenljivke in označimo $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, tedaj velja*

$$\text{var}(\bar{X}) = \frac{1}{n} \text{var}(X_1).$$

Dokaz. Računamo

$$\begin{aligned} \text{var}(\bar{X}) &= \text{var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \text{var} \left(\sum_{i=1}^n X_i \right) = \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) = \frac{n}{n^2} \text{var}(X_1) = \frac{1}{n} \text{var}(X_1), \end{aligned}$$

kjer smo upoštevali enako porazdeljenost in formulo za varianco vsote. Iz neodvisnosti sledi, da so kovariance med X_i in X_j enake nič. $\square$

Računajmo sedaj

$$\text{var}(\hat{\theta}) = \text{var} \left(\frac{7}{6} - \frac{1}{2} \bar{X} \right) = \frac{1}{4} \text{var}(\bar{X}) = \frac{1}{4n} \text{var}(X),$$

kjer smo spet upoštevali formulo za varianco vsote ter zgornjo trditev. Vidimo, da je dovolj torej poznati $\text{var}(X)$, zato iz enačb 2.2 in 2.3 dobimo, da je

$$\text{var}(\hat{\theta}) = \frac{1}{4n} (\mathbb{E}(X^2) - \mathbb{E}(X)^2) = \frac{1}{4n} \left[\frac{17 - 16\theta}{30} - \left(\frac{7}{3} - 2\theta \right)^2 \right].$$

V našem primeru upoštevamo oceno za θ in to, da imamo 10 opazanj, da dobimo, da je

$$\text{var}(\hat{\theta}) = \frac{2 + 36\theta - 36\theta^2}{360} = \text{SE}(\hat{\theta})^2,$$

izračunana vrednost za $\text{SE}(\hat{\theta})^2$ pa je potem približno 0.02877.

Na tej točki se splača omeniti, da cenilke po metodi momentov niso vedno nepristranske. Če so $X_i \sim \text{Exp}(\lambda)$ in so dodatno med sabo slučajne spremenljivke neodvisne, velja

$$\mathbb{E}(X) = \frac{1}{\lambda},$$

torej je cenilka za λ po metodi momentov kar

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

Sedaj lahko računamo še $\mathbb{E}\left(\bar{X}^{-1}\right)$. V ta namen označimo $S = X_1 + X_2 + \dots + X_n$, ker pa je vsota n neodvisnih geometrijskih porazdelitev, je porazdeljena $S \sim \Gamma(n, \lambda)$, kar je splošno znano dejstvo iz verjetnosti. Tedaj je zaradi linearnosti

$$\mathbb{E}\left(\frac{1}{\bar{X}}\right) = n\mathbb{E}\left(\frac{1}{S}\right),$$

kar pa izračunamo po definiciji kot

$$\begin{aligned} \mathbb{E}\left(\frac{1}{S}\right) &= \int_0^\infty \frac{1}{s} \frac{\lambda^n}{\Gamma(n)} s^{n-1} e^{-\lambda s} ds = \frac{\lambda^n}{\Gamma(n)} \int_0^\infty s^{n-2} e^{-\lambda s} ds \\ &= \frac{\lambda^n}{\Gamma(n)} \frac{\Gamma(n-1)}{\lambda^{n-1}} = \frac{\lambda}{n-1}, \end{aligned}$$

torej je potem

$$\mathbb{E}\left(\frac{1}{\bar{X}}\right) = \frac{n\lambda}{n-1},$$

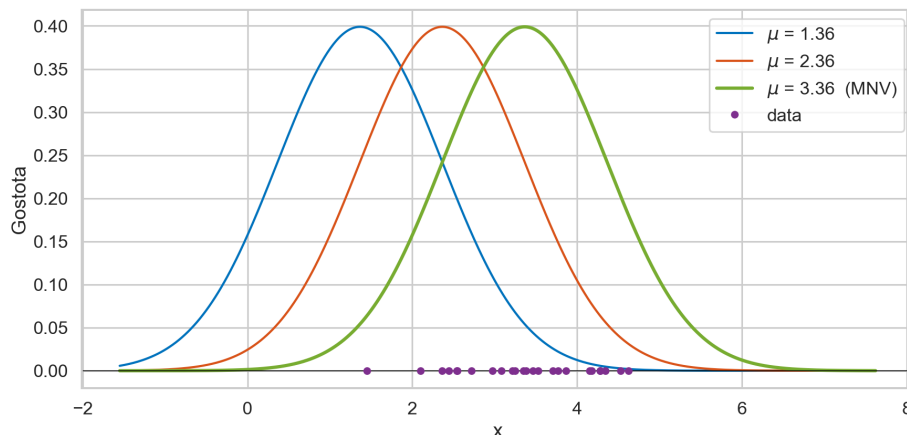
kar pa pokaže, da to ni nepristranska cenilka za λ . V resnici je vzrok linearne pričakovane vrednosti. Zgornji primer pokaže, da nelinearne transformacije pokvarijo nepristranskost, medtem ko linearne transformacije ohranjajo nepristranskost. Zadosten odgovor glede nepristranskosti cenilke $\hat{\theta}$ bi lahko bil tudi, da je ta linearno transformacija prvega momenta, zato je nepristranska.

Kljub temu pa so cenilke po metodi momentov z dodatkom relativno šibke predpostavke neodvisnosti dosledne. Seveda morajo tudi momenti, s katerimi računamo, obstajati. Doslednosti pomeni, da cenilke $\hat{y}_n$ konvergirajo v porazdelitvi proti dejanski vrednosti y , ko pošljemo n proti neskončno. To dejstvo sledi iz zakona velikih števil (izrek 1.1), saj ta zagotovi konvergenco v porazdelitvi proti pravi vrednosti. Seveda to vse nujno velja pod pogoje, da uporabimo pravi model za cenitev. Če bomo za cenitev μ uporabljali $\mathbb{E}(X^2) + 3$, gotovo ne bomo dobili najboljših možnih rezultatov. Kljub temu je pogosto težko vedeti, kaj je pravi model, zato se kaj takšnega vselej lahko zgodi.

2.3 Ocena po metodi največjega verjetja

Tako kot metoda momentov je tudi metoda največjega verjetja pogost način za pridobivanje cenilk. Ideja metode je, da bi radi maksimizirali funkcijo verjetja tako, da bodo opaženi dogodki najbolj verjetni. Spodnja slika prikazuje

kako premikamo parameter verjetnostne funkcije tako, da se najboljše prilega podatkom.



Slika 10: Grafični prikaz metode največjega verjetja

Zgornja slika prikazuje iskanje cenilk po metodi največjega verjetja. Na abscisni osi imamo označene točke, nad njimi pa narisane porazdelitvene funkcije pri različnih parametrih. Med vsemi možnostmi iščemo tudi vrednost parametrov, da bo verjetnost opažanja označenih točk čim večja. V zgornjem primeru je to porazdelitev, ki ima $\mu = 3.36$, enako idejo pa lahko uporabimo tudi za več parametrov. Sedaj ko imamo boljše razumevanje metode največjega verjetja, navedimo ustrezne definicije.

Definicija 2.5. Verjetje je verjetnostna funkcija ali gostota opažanja, ki jo gledamo kot funkcijo parametra modela. Naj opažanje zavzame vrednosti v neki množici A , neodvisno od parametra θ . Naj za diskretne spremenljivke velja, da je A števna in je verjetje $L(\theta | X) = P(X = x) > 0$. V primeru zveznih spremenljivk denimo, da je A odprta podmnožica $\mathbb{R}^n$ in je verjetje $L(\theta | X) = f_X(x | \theta) > 0$ za vsak $x \in A$.

Cenilka $\hat{\theta} = h(X)$ za θ po metodi največjega verjetja je tista, pri kateri $L(\theta | X)$ pri vsakem $x \in A$ doseže maksimum pri $\theta = h(x)$.

Poglejmo si, kako ocenimo θ po metodi največjega verjetja v našem primeru. Splošno verjetje je funkcija

$$L(\theta) = \prod_{i=1}^n P(X = i)^{n_i},$$

kjer n_i označuje število pojavitev izida i . Iz naših opažanj dobimo, da je

$$L(\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^2 \left(\frac{2(1-\theta)}{3}\right)^4 \left(\frac{1-\theta}{3}\right)^2.$$

Ker bomo funkcijo $L(\theta)$ maksimizirali, je dovolj, da gledamo

$$L(\theta) = \theta^4(1 - \theta)^6,$$

saj konstante ne vplivajo na položaj maksimuma, zgolj na njegovo vrednost. Kot se spomnimo iz vaj pri tem predmetu, je pogosto enostavneje, da namesto funkcije $L(\theta)$ maksimiziramo funkcijo

$$l(\theta) = \ln(L(\theta)),$$

saj nam s tem produkti preidejo v vsote, eksponenti pa postanejo koeficienti, nismo pa vplivali na ekstreme, saj je logaritem strogo naraščajoča funkcija. Zapišemo lahko torej

$$l(\theta) = \ln(L(\theta)) = \ln(\theta^4(1 - \theta)^6) = 4 \ln(\theta) + 6 \ln(1 - \theta),$$

kar potem odvajamo in dobimo

$$\frac{\partial l}{\partial \theta} = \frac{4}{\theta} - \frac{6}{1 - \theta}.$$

Iščemo maksimum funkcije $l(\theta)$, zato odvod enačimo z nič in izračunamo ustrezen θ :

$$\frac{4}{\theta} - \frac{6}{1 - \theta} = 0 \implies 4(1 - \theta) = 6\theta \implies \theta = \frac{2}{5}.$$

Pokažimo še, da funkcija $l(\theta)$, in s tem tudi $L(\theta)$, res doseže maksimum pri tej vrednosti θ . Drugi odvod funkcije l se glasi

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{4}{\theta^2} - \frac{6}{(1 - \theta)^2}.$$

Drugi odvod je povsod negativen, zato je pri $\theta = \frac{2}{5}$ res dosežen maksimum. Vidimo tudi, da sta oceni po metodi momentov in metodi največjega verjetja blizu, kar je pričakovano, saj bi se sicer vprašali o pravilnosti metod.

2.4 Nepristranskost cenilke in ocena napake

Podobno kot smo to storili pri metodi momentov, lahko tudi sedaj utemeljimo nepristranskost in ocenimo standardno napako. V namen ocene nepristranskosti zapišemo

$$L(\theta) = \theta^{n_0+n_1}(1 - \theta)^{n_2+n_3},$$

kjer n_i spet označuje število opažanj določenega izida. Za $l(\theta)$ potem zapišemo

$$l(\theta) = (n_0 + n_1) \ln(\theta) + (n_2 + n_3) \ln(1 - \theta),$$

za odvod pa zapišemo

$$\frac{\partial l}{\partial \theta} = \frac{n_0 + n_1}{\theta} - \frac{n_2 + n_3}{1 - \theta}.$$

Ponovno odvod enačimo z nič in dobimo

$$n_0 - n_0\theta + n_1 - n_1\theta = (n_2 + n_3)\theta,$$

torej je

$$\hat{\theta} = \frac{n_0 + n_1}{n_0 + n_1 + n_2 + n_3} = \frac{n_0 + n_1}{n}. \quad (2.4)$$

Ko vstavimo naša opažanja, se to sklada z ugotovitvijo prejšnjega podpoglavja, utemeljitev, da je to maksimum pa je enaka kot pri zgornjem primeru. Zaradi zgornje enačbe smemo zapisati

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}\left(\frac{n_0 + n_1}{n}\right) = \frac{1}{n}\mathbb{E}(n_0 + n_1).$$

Spremenljivka n_0 meri število ničel pri n realizacijah spremenljivke, zato je porazdeljena kot $n \sim \text{Bin}(n, \frac{2\theta}{3})$, po istem argumentu pa je tudi n_1 porazdeljen $n_1 \sim \text{Bin}(n, \frac{\theta}{3})$. Za binomsko porazdelitev $\text{Bin}(n, p)$ vemo, da je njena pričakovana vrednost kar np , varianca pa je $n\theta(1 - \theta)$. Zaradi tega je

$$\mathbb{E}(n_0) = \frac{2n\theta}{3} \quad \text{in} \quad \mathbb{E}(n_1) = \frac{n\theta}{3}.$$

Sledi potem, da je

$$\mathbb{E}(\hat{\theta}) = \frac{1}{n} \left(\frac{2n\theta}{3} + \frac{n\theta}{3} \right) = \theta,$$

in zato

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = \theta - \theta = 0,$$

torej je tudi ta cenila nepristranska. Poglejmo si še standardno napako cenilke. Zaradi nepristranskosti moramo izračunati le $\text{var}(\hat{\theta})$. Zaradi binomske porazdelitve tako n_0 kot tudi n_1 , je nujna vsota porazdeljena $\text{Bin}(n, \theta)$, zaradi česar lahko zapišemo

$$\begin{aligned} \text{var}(\hat{\theta}) &= \text{var}\left(\frac{n_0 + n_1}{n}\right) = \frac{1}{n^2}(\text{var}(n_0) + \text{var}(n_1)) = \\ &= \frac{1}{n^2}n\theta(1 - \theta) = \frac{\theta(1 - \theta)}{n}. \end{aligned} \quad (2.5)$$

Tedaj lahko upoštevamo oceno za θ in dobimo, da je

$$\text{var}(\hat{\theta}) = \frac{\frac{2}{2}(1 - \frac{2}{5})}{10} = \frac{6}{250},$$

torej je

$$\text{SE}(\hat{\theta})^2 = \frac{6}{260} = 0.0231.$$

Podobno kot cenilke po metodi momentov, tudi cenilke po metodi največjega verjetna niso nujno nepristranske. Pod določenimi regularnimi pogoji imajo sicer cenilke po MNV določene lepe lastnosti. Med drugimi so doslene, torej velja $\hat{\theta} \xrightarrow[n \rightarrow \infty]{} \theta$, hkrati pa so tudi asimptotsko nepristranske, kar pomeni, da $\sqrt{n} \text{Bias}(\hat{\theta} \mid \theta) \xrightarrow[n \rightarrow \infty]{} 0$.

2.5 Bayesov pogled na θ

V Bayesovi statistiki lahko tudi parametre, kot je recimo θ , obravnavamo kot slučajne spremenljivke. Te imajo neko dano apriorno porazdelitev, recimo $f_{\Theta}(\theta)$. Ta apriorna porazdelitev je posledica tega, kar vemo o tej slučajni spremenljivki preden opazimo kakršnekoli podatke, ki jih označimo z X . Tako imajo pri dani vrednosti slučajne spremenljivke $\Theta = \theta$ opažene vrednosti gostoto $f_{X|\Theta}(x | \theta)$. Skupna gostota X in Θ je potem

$$f_{X,\Theta}(x, \theta) = f_{X|\Theta}(x | \theta)f_{\Theta}(\theta), \quad (2.6)$$

robno porazdelitev za X pa seveda dobimo kot

$$f_X(x) = \int f_{X,\Theta}(x, \theta) d\theta = \int f_{X|\Theta}(x | \theta)f_{\Theta}(\theta) d\theta.$$

iz katere dobimo lahko pogojno porazdelitev te spremenljivke glede na opažene podatke. Izrazimo lahko tudi porazdelitev Θ pogojno na opažene vrednosti X , saj je

$$f_{X,\Theta}(x, \theta) = f_{\Theta|X}(\theta | x)f_X(x),$$

kar pa pove, da je

$$\begin{aligned} f_{\Theta|X}(\theta | x) &= \frac{f_{X,\Theta}(x, \theta)}{f_X(x)} \\ &= \frac{f_{X|\Theta}(x | \theta)f_{\Theta}(\theta)}{\int f_{X|\Theta}(x | \theta)f_{\Theta}(\theta) d\theta} \end{aligned} \quad (2.7)$$

Tej porazdelitvi pravimo *aposteriorna porazdelitev*, saj je to porazdelitev, ki smo jo dobili potem, ko smo že opazili neke vrednosti X , iz katerih dobimo informacije o Θ . Enačba 2.6 bo torej ključna pri nalogi. Določiti moramo aposteriorno porazdelitev za Θ glede na opažanja X_i .

Vredno je omeniti, da zgoraj napisano velja sicer v primeru zveznih slučajnih spremenljivk, a so formule enake, če imamo opravka z diskretnimi slučajnimi spremenljivkami.

Denimo torej, da je apriorna porazdelitev Θ dana kot $\Theta \sim U(0, 1)$. To seveda pomeni, da je

$$f_{\Theta}(\theta) = \begin{cases} 1; & 0 < \theta < 1, \\ 0; & \text{sicer} \end{cases} = \mathbb{1}_{(0,1)}(\theta)$$

upoštevajoč formulo iz enačbe 2.7, pa je

$$f_{\Theta|X}(\theta | x) = \frac{f_{X|\Theta}(x | \theta)f_{\Theta}(\theta)}{\int f_{X|\Theta}(x | \theta)f_{\Theta}(\theta) d\theta}. \quad (2.8)$$

Potrebujemo torej pogojno verjetje opaženih vrednosti X , glede na vrednost Θ . Velja

$$L(\theta) = f_{X|\Theta}(x | \theta) = \prod_{i=1}^n P(X_i = k_i), \quad (2.9)$$

kjer je $n = 10$ in smo s k_i označili vrednost, ki smo jo opazili pri i -ti pojavitvi spremenljivke X , torej X_i . Za ta produkt vemo že, da je

$$L(\theta) = \theta^4(1 - \theta)^6,$$

kjer smo podobno kot zgoraj zanemarili konstantne faktorje, ki se v tem primeru izničijo z imenovalcem v izrazu 2.8. Torej smo pokazali, da je

$$\begin{aligned} f_{\Theta|X}(\theta | x) &= \theta^4(1 - \theta)^6 \mathbb{1}_{(0,1)}(\theta) \left[\int_{-\infty}^{\infty} \theta^4(1 - \theta)^6 \mathbb{1}_{(0,1)}(\theta) d\theta \right]^{-1} \\ &= \theta^4(1 - \theta)^6 \mathbb{1}_{(0,1)}(\theta) \left[\int_0^1 \theta^4(1 - \theta)^6 d\theta \right]^{-1}. \end{aligned}$$

Izraz v zgornjem izrazu prepoznamo kot B funkcija. Splošna oznaka je namreč

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1 - u)^{\beta-1} du,$$

zato desni faktor prepoznamo kot $B(5, 7)$. Dalje opazimo še, da se zgoraj ponovi enaka oblika, kar je ravno definicija beta porazdelitve.

Definicija 2.6. Naj bo X zvezna slučajna spremenljivka, ki ima gostoto

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1}, & 0 < x < 1, \\ 0, & \text{sicer,} \end{cases}$$

tedaj pravimo, da ima X *beta porazdelitev s parametroma α in β* in označimo $X \sim B(\alpha, \beta)$.

Ko primerjamo definicijo beta porazdelitve z izrazom za $f_{\Theta|X}(\theta | x)$, ki smo ga dobili, vidimo, da je aposteriorna porazdelitev Θ glede na opažene vrednosti X ravno $B(5, 7)$. Ker želimo oceniti θ , pogledjmo pričakovano vrednost in modus za aposteriorno porazdeljeno Θ , ki obe lahko služita za oceno θ . Pojem pričakovane vrednosti nam je dobro znan že iz verjetnosti, manj znan pa je pojem modusa za zvezno porazdeljene slučajne spremenljivke, zato navedimo definicijo.

Definicija 2.7. Naj bo X zvezno porazdeljena slučajna spremenljivka za gostoto $f_X(x)$. *Modus* je potem tista vrednost x , kjer ima gostota $f_X(x)$ lokalni maksimum. Če ima gostota več lokalnih maksimumov, je vsak izmed teh maksimumov en *modus*.

Naslednji trditvi določata, kaj sta modus in mediana za slučajno spremenljivko $X \sim B(\alpha, \beta)$.

Trditev 2.8. *Naj bo $X \sim B(\alpha, \beta)$. Tedaj je njena pričakovana vrednost*

$$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta},$$

za $\alpha, \beta > 1$ pa je njen modus

$$\frac{\alpha - 1}{\alpha + \beta - 2}.$$

Dokaz. Pokažimo najprej izraz za pričakovano vrednost. Po definiciji pričakovane vrednosti je

$$\begin{aligned}\mathbb{E}(X) &= \int x f_X(x) dx = \frac{1}{B(\alpha, \beta)} \int_0^1 x x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 x^\alpha (1-x)^{\beta-1} dx \\ &= \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)}.\end{aligned}$$

Ker je

$$B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)},$$

je potem

$$\frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

Dalje ker je

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha),$$

je potem tudi

$$\Gamma(\alpha+\beta+1) = (\alpha+\beta)\Gamma(\alpha+\beta),$$

zato je

$$\mathbb{E}(X) = \frac{\alpha\Gamma(\alpha)\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}.$$

Za dokaz modusa pa moramo poiskati maksimum funkcije

$$g(x) = x^{\alpha-1}(1-x)^{\beta-1},$$

saj konstanten faktor jasno ne vpliva na položaj maksimuma, zgolj na njegovo vrednost. Podobno kot pri funkcijah verjetja, je tudi tukaj račun lažji, če iščemo maksimum logaritma funkcije g . Torej

$$l(x) = \ln(g(x)) = (\alpha-1)\ln(x) + (\beta-1)\ln(1-x).$$

Funkcijo $l(x)$ odvajamo in dobimo

$$l'(x) = \frac{\alpha-1}{x} - \frac{\beta-1}{1-x},$$

ter

$$l''(x) = \frac{\alpha-1}{x^2} - \frac{\beta-1}{(1-x)^2}.$$

Pri iskanju maksimuma odvod enačimo z 0 in dobimo

$$\frac{\alpha-1}{x} = \frac{\beta-1}{1-x},$$

iz česar izračunamo, da mora biti

$$x = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

To je res maksimum, saj velja

$$l''(x) = -\frac{\alpha - 1}{x^2} - \frac{\beta - 1}{(1 - x)^2} < 0$$

za vse x , ker sta α in β oba večja od 1. Torej je v tem primeru modus res

$$\frac{\alpha - 1}{\alpha + \beta - 2}.$$

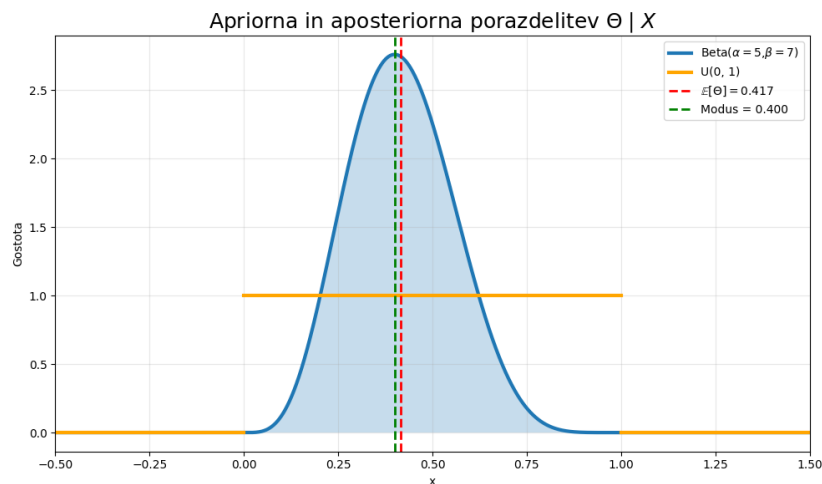
□

Sedaj lahko za $\Theta \mid X \sim B(5, 7)$ izračunamo modus in pričakovano vrednost. Dobimo

$$\mathbb{E}(\Theta \mid X) = \frac{5}{5 + 7} = \frac{5}{12} \approx 0.417,$$

za modus pa dobimo

$$\text{Modus} = \frac{5 - 1}{5 + 7 - 2} = \frac{4}{10} = 0.4.$$



Slika 11: Apriorna in aposteriora porazdelitev Θ .

2.6 Spremenljiva apriorna porazdelitev

Vpeljimo sedaj parametrizacijo

$$\Theta = \sin^2 \Phi$$

in pogledimo kako bi sedaj ocenili najprej ϕ , nato pa še θ . Ravnamo podobno kot v prejšnji točki. Naloga pravi, da je Φ apriorno porazdeljen kot $\Phi \sim U(0, \pi/2)$, kar pomeni, da je

$$f_{\Phi}(\phi) = \begin{cases} 1; & 0 < \phi < \pi/2, \\ 0; & \text{sicer} \end{cases} = \mathbb{1}_{(0, \pi/2)}(\phi).$$

Ko upoštevamo formulo iz enačbe 2.6, dobimo, da je

$$f_{\Phi|X}(\phi | x) = \frac{f_{X|\Phi}(x | \phi) f_{\Phi}(\phi)}{\int f_{X|\Phi}(x | \phi) f_{\Phi}(\phi) d\phi}. \quad (2.10)$$

Iščemo torej $f_{X|\Phi}(x | \phi)$, kar pa dobimo preko funkcije verjetja, ki smo jo uporabljali pri zgornjih podpoglavjih. Funkcija verjetja je tedaj

$$L(\theta(\phi)) = \theta^4(1 - \theta)^6,$$

ko pa upoštevamo, da je $\theta = \sin^2 \phi$ in $1 - \theta = \cos^2 \phi$, dobimo, da je

$$f_{\Phi|X}(\phi | x) = \sin^8(\phi) \cos^{12}(\phi),$$

kjer smo ponovno konstante lahko zanemarili iz enakih razlogov kot prej. Za izračun izraza (2.10) dalje upoštevamo, da je $f_{\Phi}(\phi)$ konstantna funkcija, saj je zgolj indikator. Prav tako lahko tudi izračunamo imenovalc

$$\int_0^{\pi/2} \sin^8(\phi) \cos^{12}(\phi) d\phi,$$

ki pa se integrira v neko konstanto. Pokazali smo torej, da je $f_{\Phi|X}(\phi | x)$ zgolj skalirana funkcija verjetja. Ker navodila zahtevajo, da ocenimo modus in mediano, pogledimo najprej njen modus.

Iščemo torej ekstrem funkcije, ki pa bo zaradi dejstva, da konstante na vplivajo ne položaj ekstrema, zgolj na vrednost, kar enak ekstremu funkcije verjetja. Funkcijo $L(\theta(\phi))$ lahko torej logaritmiramo in računamo

$$l = \ln L = l(\phi) = 8 \ln(\sin \phi) + 12 \ln(\cos \phi).$$

Funkcijo odvajamo in dobimo

$$l'(\phi) = 8 \cot \phi - 12 \tan \phi,$$

kar enačimo z 0, kar pove, da je

$$\tan^2(\phi) = \frac{8}{12} = \frac{2}{3}.$$

Upoštevajoč, da je

$$\sin^2 x = \frac{\tan^2 x}{1 + \tan^2 x},$$

dobimo, da je

$$\sin^2 \phi = \frac{2/3}{5/3} = \frac{2}{5},$$

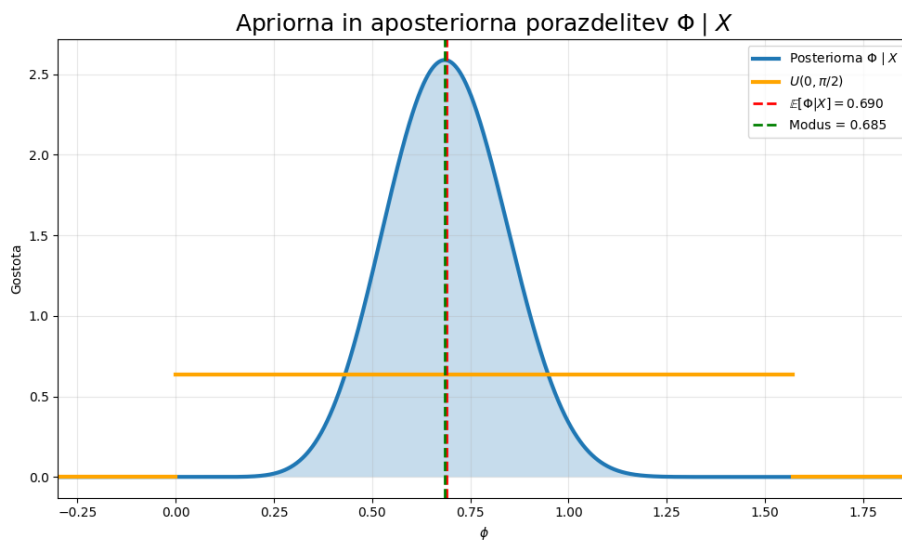
kar je že kar ocena tudi za θ . Videli smo torej, da se v tem primeru rezultat ne spremeni, saj smo maksimizirali isto funkcijo kot pri zgornjih podpoglavjih. Rezultat je torej

$$\text{Modus} = 0.4.$$

Za izračun pričakovane vrednosti si tokrat ne moremo pomagati z beta porazdelitvijo, pač pa jo izračunamo numerično. Računamo

$$\hat{\phi} = \int_0^{\pi/2} \phi f_{\Phi|X}(\phi | x) d\phi,$$

kar pove, da je $\hat{\phi} \approx 0.6898$, za θ pa dobimo ocenimo $\theta \approx 0.4050$.



Slika 12: Apriorna in aposteriorna porazdelitev za Φ

2.7 Drugačen pogled na prejšnjo nalogo

Poglejmo si sedaj še drugačen primer. Denimo spet, da je

$$\Theta = \sin^2(\Phi),$$

kjer je $\Phi \sim U(0, \frac{\pi}{2})$. Ponovno kot v prejšnjem podpoglavju nas spet zanima aposteriorna porazdelitev Θ , ocena za θ , dodatno pa moramo najprej ugotoviti še apriorno porazdelitev Θ . V ta namen se spomnimo transformacijske formule.

Izrek 2.9 (Transformacijska formula). *Naj bo X slučajna spremenljivka z gostoto $f_X(x)$ ter naj bo $P(X \in U) = 1$ za neko odprto množico U . Naj bo $Y = \Psi(X)$, V odprta množica in $\Psi : U \rightarrow V$ zvezno odvedljiva bijekcija. Potem je $P(Y \in V) = 1$ in velja*

$$f_Y(y) = f_X(\Psi^{-1}(y)) \cdot |J_{\Psi^{-1}}(y)|.$$

Uporabimo zgornji izrek v našem primeru. Za množico $U = (0, 1)$ seveda velja, da je $P(\Phi \in U) = 1$, zaradi enakomerne porazdelitve pa je gostota Φ

$$f_\Phi(\varphi) = \begin{cases} 1; & 0 < \varphi < \frac{\pi}{2}, \\ 0; & \text{sicer} \end{cases} = \frac{2}{\pi} \mathbb{1}_{(0, \frac{\pi}{2})}(\varphi).$$

Funkcija Ψ je v tem primeru definirana kot $\Psi(\varphi) = \sin^2(\varphi)$, zato je tudi $V = (0, 1)$, njen inverz pa je $\Psi^{-1}(\theta) = \arcsin(\sqrt{\theta})$. Izračunajmo še odvod inverza:

$$(\Psi^{-1}(\theta))' = \frac{1}{2\sqrt{\theta}\sqrt{1-\theta}}.$$

Po transformacijski formuli potem velja:

$$\begin{aligned} f_\Theta(\theta) &= f_\Phi(\arcsin(\sqrt{\theta})) \frac{1}{2\sqrt{\theta}\sqrt{1-\theta}} \\ &= \frac{2}{\pi} \mathbb{1}_{(0, \frac{\pi}{2})}(\arcsin(\sqrt{\theta})) \frac{1}{2\sqrt{\theta}\sqrt{1-\theta}}. \end{aligned}$$

Ker za $\theta \in (0, 1)$ vrednost $\arcsin(\sqrt{\theta})$ vedno leži na intervalu $(0, \frac{\pi}{2})$, je indikatorska funkcija vedno 1, zato je gostota Θ kar

$$f_\Theta(\theta) = \frac{1}{\pi\sqrt{\theta}\sqrt{1-\theta}}.$$

Tako smo določili apriorno porazdelitev za Θ , ki pa ni enakomerno porazdeljena. Velja, da je apriorno $\Theta \sim B(\frac{1}{2}, \frac{1}{2})$. Naslednji koraki so zelo podobni kot v predprejšnjem podpoglavju. Gostota $f_{X|\Theta}(x | \theta)$ je enaka, saj so opažanja enaka, spremeni se zgolj $f_\Theta(\theta)$. Tako računamo

$$\begin{aligned} f_{\Theta|X}(\theta | x) &= \frac{f_{X|\Theta}(x | \theta)f_\Theta(\theta)}{\int f_{X|\Theta}(x | \theta)f_\Theta(\theta) d\theta} \\ &= \frac{\theta^4(1-\theta)^6}{\pi\sqrt{\theta}\sqrt{1-\theta}} \cdot \left[\int_0^1 \frac{\theta^4(1-\theta)^6}{\pi\sqrt{\theta}\sqrt{1-\theta}} d\theta \right]^{-1} \\ &= \theta^{7/2}(1-\theta)^{11/2} \cdot \left[\int_0^1 \theta^{7/2}(1-\theta)^{11/2} d\theta \right]^{-1}. \end{aligned}$$

Zgornji izraz pove, da je $\Theta | X$ spet porazdeljen z beta porazdelitvijo. Natančneje je aposteriona porazdelitev Θ kar $B(\frac{9}{2}, \frac{13}{2})$. Za tako porazdelitev izračunamo še pričakovano vrednost in modus

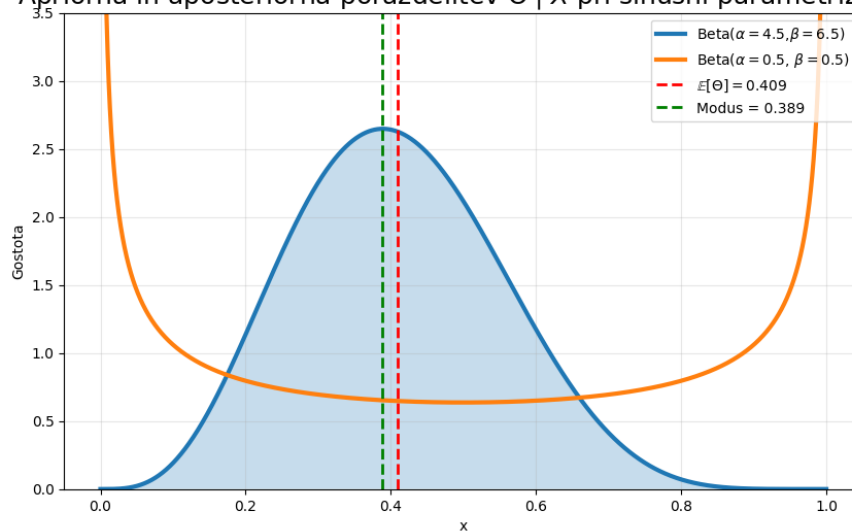
$$\mathbb{E}(\Theta | X) = \frac{\frac{9}{2}}{\frac{9}{2} + \frac{13}{2}} = \frac{9}{22} \approx 0.409,$$

za modus pa dobimo

$$\text{Modus} = \frac{\frac{9}{2} - 1}{\frac{9}{2} + \frac{13}{2} - 2} = \frac{7}{18} \approx 0.389,$$

kar izstopa po temu, da je vrednost manjša od 0.4.

Apriorna in aposteriorna porazdelitev $\Theta | X$ pri sinusni parametrizaciji



Slika 13: Apriorna in aposteriorna porazdelitev pri dani parametrizaciji

Vredno je omeniti, da sem do takšne naloge prišel zgolj zaradi svojega nepazljivega branja navodil. Napako sem pri pregledu naloge odkril in dopisal prejšnje poglavje, tega pa sem vključil zgolj zaradi zanimivih rezultatov, povezanih z beta porazdelitvijo in pa dejstva, da so rezultati vendarle blizu ostalih vrednosti. Poleg tega se mi zdi tudi tak pogled na cenatev Θ zanimiv. Razlika je zgolj, da smo izračunali apriorno porazdelitev s pomočjo transformacijske formule, nato pa smo nadaljevali kot v prvem Bayesovem primeru.

2.8 Povzetek druge naloge

Na tem mestu povzemimo še rezultate te naloge s spodnjo tabelo.

Metoda	Apriorno	Ocena za θ
Momentov	–	0.367
Največje verjetje	–	0.400
Bayesova (pričakovana vrednost)	Enakomeren	0.417
Bayesova (modus)	Enakomeren	0.400
Bayesova (pričakovana vrednost)	Ocena preko ϕ	0.405
Bayesova (modus)	Ocena preko ϕ	0.400
Bayesova (pričakovana vrednost)	Sinusni	0.409
Bayesova (modus)	Sinusni	0.389

Tabela 6: Primerjava različnih ocen parametra θ .

Vidimo, da se vse ocene gibljejo okoli enakih vrednosti, kar je zanimivo, saj so ocene pridobljene na popolnoma različne načine. Izstopa tudi dejstvo, da se ocene po modusu za prvi dve Baysovi metodi ujemata tudi z oceno po MNV, medtem ko sta oceni preko pričakovane vrednosti sicer precej blizu, a vendar ne enaki. Ujemanje z MNV je posledica dejstva, da konstantne apriorne porazdelitve vodijo do tega, da je modus, torej maksimum apriorne porazdelitve, kar enak ekstremu, ki smo ga izračunali pri metodi največjega verjetja. Primer, kjer temu ni tako, je zadnji Baysov primer. Tedaj je Θ apriorno porazdeljen z beta porazdelitvijo, zaradi česar se spremeni modus, ki sedaj ni več enak kot ocena po MNV. Pri izračunih za pričakovano vrednost pa temu ni tako. Tam je namreč rezultat precej odvisen od tega, kakšna je apriorna porazdelitev, četudi je ta konstantna, saj to vpliva na vrednost integrala. Poleg tega zanimivo izstopa tudi to, da je ocena po metodi momentov edina ocena, ki je manjša od 0.4. Vsekakor bi si pa upal trditi, da je dobra ocena za θ res okoli 0.4, saj tako kažejo prav vse metode, ki smo jih uporabili.

Preko svojega nepazljivega branja pa sem uspel ugotoviti tudi, da vrstni red ocenjevanja igra vlogo pri končni vrednosti. Razlika med vrednostim je prisotna pri obeh ocenah, ko primerjamo metodo, ko najprej ocenimo aposteriorno porazdelitev za ϕ , nato pa iz tega ocenimo θ , z metodo ko preko apriorne porazdelitve za ϕ pridemo do apriorne porazdelitve za θ in potem ocenimo modus in mediano. Pri temu se tudi pokvari lepota metod, ki je v prejšnjih podnalogah ohranjala enako vrednost za moduse. Vsekakor zanimiva opazka.

2.9 Izboljšave in komentarji k drugi nalogi

Nekaj komentarjev smo navedli že tekom poglavja, a je kljub temu vredno omeniti določene opazke. Pri računu standardne napake (SE), smo si lahko precej pomagali s tem, da je bilo potrebno zgolj izračunati varianco. To smo smeli storiti, saj sta bili cenilki za obe metodi nepristranski, za takšni cenilki pa smo definirali standardno napako. Tega bi se lahko verjetno lotili tudi preko Fisherjeve informacije, kot smo to počeli na vajah. Fisherjevo informacijo smo

izračunali kot

$$\text{FI}(\theta) = \mathbb{E} \left[\left(\frac{\partial l(\theta | x)}{\partial \theta} \right)^2 \right] = \mathbb{E} \left[\frac{\partial^2 l(\theta | x)}{\partial \theta^2} \right].$$

Poleg tega asimptotično velja Rao-Cramerjeva ocena, ki za nepristransko cenilko pravi, da je

$$\text{SE}^2 = \text{var}(\hat{\theta}) \geq \frac{1}{\text{FI}(\theta)}.$$

Tedaj za dovolj regularne modele in cenilke po MNV velja, da $\hat{\theta} \rightarrow \theta$ in $\mathbb{E}(\hat{\theta}) \rightarrow \theta$, zato lahko pridobimo oceno za standardno napako kot

$$\text{SE}^2 \sim \mathbb{E}((\hat{\theta} - \theta)^2) \sim \frac{1}{\text{FI}(\theta)}.$$

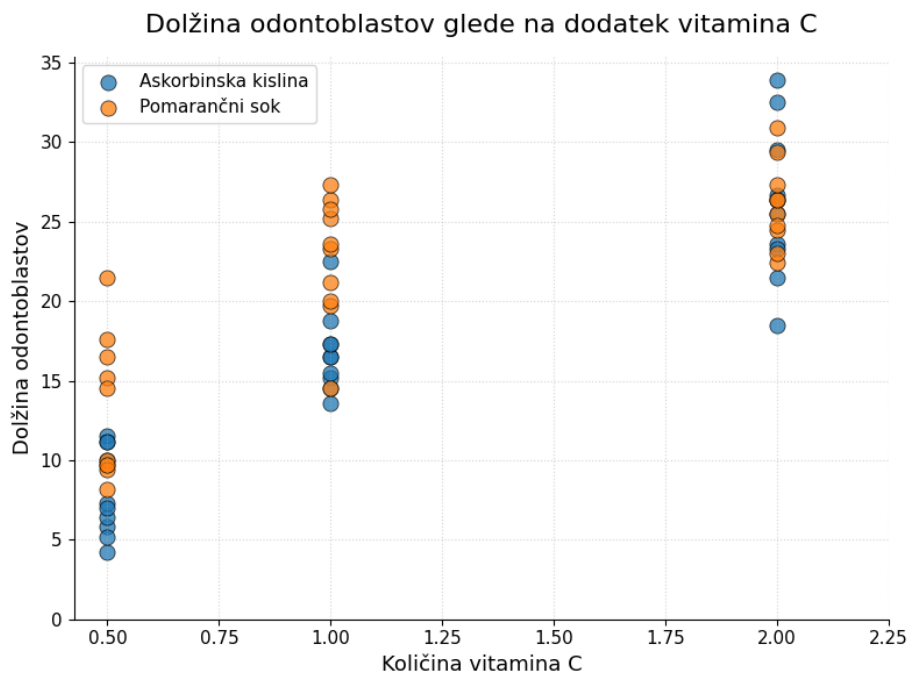
Če bi imeli opravka z večdimenzionalnimi cenilkami, bi ustrezno definirali Fisherjevo informacijo na naslednji način. Če je $\theta \in \mathbb{R}^d$ in nas zanima karakteristika $\xi = h(\theta)$ po MNV, je $\hat{\xi} = h(\hat{\theta})$ cenilka po MVN. Tedaj je Fisherjeva informacija matrika mešanih drugih odvodov, oceno za standardno napako ob nepristranskosti $\hat{\xi}$ pa dobimo kot

$$\text{SE}^2 = \text{var}(\hat{\xi}) \geq \nabla h(\theta)^T \text{FI}^{-1}(\theta) \nabla h(\theta).$$

Dalje se splača zgolj omeniti, zakaj je smiselno izbrati modus in mediano kot cenilko za θ pri reševanju naloge ob dani apriorni porazdelitvi. Velja namreč, da sta tako modus in mediana meri za centralno tendenco, medtem ko sta IQR in standardni odklon meri razpršenosti. Iz enačbe (2.2) vidimo, da je pričakovana vrednost linearna v θ , zato je izbor mer centralne tendence smiseln tudi za cenilke za θ .

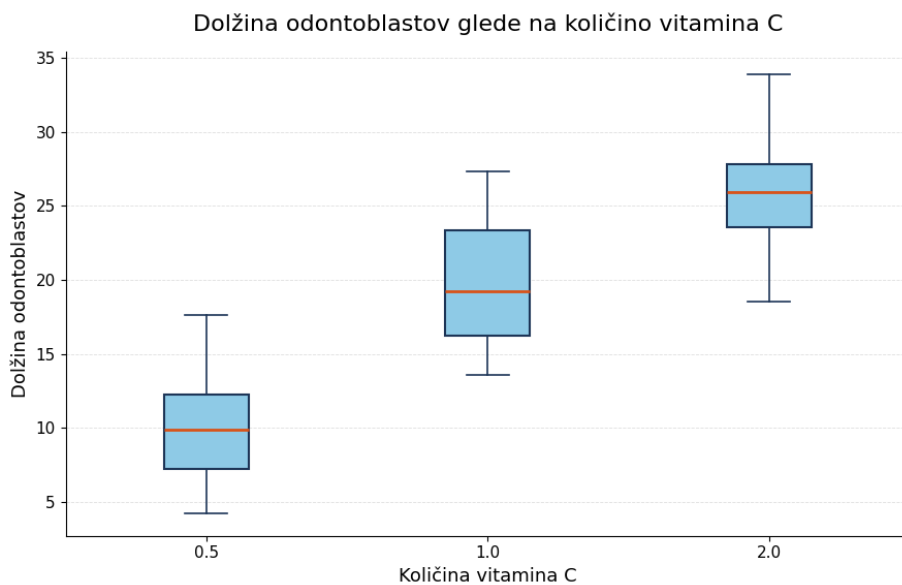
3 Tretja naloga

Naloga od nas zahteva, da moramo oceniti učinek dodajanja vitamina C na dolžino odontoblastov, torej celic, ki tvorijo dentin. Podatki, ki jih imamo na voljo, so dolžina odontoblastov glede na količino dodanega vitamina C. Ti se nahajajo v datoteki `Zobje.csv`. Datoteka vsebuje stolpce `Dolzina`, `Nacin` in `Kolicina`. V stolpcu `Dolzina` imamo navedene izmerjene dolžine odontoblastov, kar bo iskana spremenljivka. Stolpec `Nacin` vsebuje podatke o tem, na kakšen način smo dodajali vitamin C, bodisi z askorbinsko kislino, kar ustreza oznaki VC, bodisi s pomarančnim sokom, kar je zakodirano z oznako OJ. Zadnji stolpec, torej `Kolicina`, vsebuje podatke o količini dodatkov vitamina C. Za predstavbo si najprej vizualiziramo podatke, ki smo jih dobili.



Slika 14: Dolžina odontoblastov glede na dodan vitamin C

Že oblika podatkov, ki je vidna na sliki 14, nakazuje, da verjetno obstaja povezava med dodanim vitaminom C in pa dolžino odontoblastov. Zgornji sklep lahko dodatno utemeljimo, če pogledamo graf na sliki 15. Vidimo namreč, da dožine naraščajo in se škatle praktično ne prekrivajo, kar nakazuje na odvisnost dolžine od količine dodanega vitamina C. Poglejmo si, kako bi to trditev tudi resnično pokazali. Najbolj enostaven preizkus, ki ga lahko naredimo, je, da na osnovi podatkov s pomočjo linearne regresije ocenimo premico, ki se bo podatkom najboljše prilegala.



Slika 15: Škatle z brki za količine dodanega vitamina

3.1 Ocenjevanje z linearno regresijo

Označimo dolžino odontolastov z y_i in količino dodanega vitamina C z x_i . Za namene reševanja te naloge lahko predpostavimo, da bo ε_i porazdeljen $N(0, \sigma^2)$. Za regresijski model postavimo

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

kar lahko zapišemo tudi v matrični obliki kot

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

oziroma kar

$$Y = X\beta + \varepsilon,$$

za ustrezni matriki Y, X ter vektorja β in ε . Oceniti želimo vektor β , zato iščemo najboljšo nepristransko linearno cenilko za β , ki jo označimo kar z $\hat{\beta}$. Spomnimo se najprej definicije najboljše nepristranske linearne cenilke.

Definicija 3.1. *Linearna* oziroma *afina* cenilka za količino q z vrednostmi v $\mathbb{R}$ na podlagi opazanj Y z vrednostmi v $\mathbb{R}^n$ je cenilka oblike

$$\hat{q} = \psi^T Y + \rho,$$

kjer sta tako $\psi \in \mathbb{R}^n$ kot tudi $\rho \in \mathbb{R}$ konstanti. *Najboljša nepristranska linearna cenilka* za q je taka nepristranska linearna cenilka $\hat{q}$, da za poljubno alternativno nepristransko linearno cenilko Q velja

$$\text{var}(Q) \geq \text{var}(\hat{q}).$$

Ker pri linearni regresiji zahtevamo, da se premica opaženim točkam prilega glede na metodo najmanjših kvadratov, je najbolj smiselna možnost za $\hat{\beta}$ kar

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad (3.1)$$

saj je to rešitev normalnega sistema pod predpostavko, da je matrika $(X^T X)^{-1}$ nesingularna. Izrek Gauss-Markova nam zagotavlja, da je to res prava izbira ne le iz vidika numerične linearne algebre, pač pa tudi z vidika statistike.

Izrek 3.2 (Gauss-Markov). *Privzamemo model splošne linearne regresije $Y = X\beta + \varepsilon$ za $X \in \mathbb{R}^{n \times p}$ in je matika X polnega ranga, torej $\text{rang}(X) = p$. Za model naj velja, da je $\mathbb{E}(\varepsilon_i) = 0$, da so ε_i in ε_j nekorelirane in še, da je $\text{var}(\varepsilon_i) = \sigma^2 \in (0, \infty)$ za vsak i . Če je $c \in \mathbb{R}^n$ konstanten vektor, je*

$$\hat{q} = c^T \hat{\beta}$$

edina najboljša nepristranska linearna cenilka za $q = c^T \beta$.

Dokaz. Ob oznaki $V = \text{Im}(X)$, je matrika $X : \mathbb{R}^p \rightarrow V$ linearni izomorfizem, zato lahko model parametriziramo kar z vektorjem $v = X\beta$. Tako za $v \in V$ dobimo model oblike

$$Y = v + \varepsilon.$$

Zaradi Rieszovega reprezentacijskega izreka obstaja tak vektor $f \in V$, da za vse $\beta \in \mathbb{R}^p$, velja $c^T \beta = f^T X\beta = f^T v$. Če ponovno označimo

$$\hat{v} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY,$$

je $H = X(X^T X)^{-1} X^T$ ortogonalni projektor na V . Zato velja $Hv = v$ za $v \in V$ in pa $H^2 = H$. Pokazati moramo, da je $\hat{q} = f^T \hat{v}$ NNLC za $q = f^T v$.

Najprej pogledjmo splošne linearne cenilke. Velja, da so linearne cenilke za q oblike $Q = \psi^T Y + \rho$, kjer je $\psi \in \mathbb{R}^n$ in $\rho \in \mathbb{R}$. Ker smo začetni model prevedli na $Y = v + \varepsilon$, lahko to vstavimo v enačbo in izračunamo pričakovano vrednost za Q kot

$$\mathbb{E}(Q) = \mathbb{E}(\psi^T v + \psi^T \varepsilon + \rho) = \psi^T v + \rho.$$

Ker zahtevamo, da je cenilka nepristranska, mora za vsak $v \in V$ veljati

$$\psi^T v + \rho = f^T v.$$

V posebnem to velja tudi, ko za v izberemo ničelni vektor, iz česar dobimo, da mora biti $\rho = 0$. Dalje lahko tudi v prvi člen vrnemo H . Tako smo dobili

$$\mathbb{E}(Q) = \psi^T v = \psi^T H v = (H\psi)^T v = f^T v,$$

saj je H simetrična. Enačba pomeni, da mora biti $(H\psi)^T v = f^T v$ za vse $v \in V$, kar pa implicira, da je $H\psi = f$. Ker je tudi $f \in V$, je $hf = f$, zato iz $H(\psi - f) = 0$ sledi, da je $\eta = \psi - f \in \ker(H) = V^\perp$. Obratno, za poljuben $\eta \in V^\perp$ velja $H(f + \eta) = f$, torej je $\psi = f + \eta$ res rešitev. Nepristranske linearne cenilke za q so torej vse tiste, ki so oblike

$$Q = (f + \eta)^T Y,$$

kjer je $\eta \in V^\perp$.

Ker zahtevamo še, da ima cenilka še najmanjšo varianco, računamo

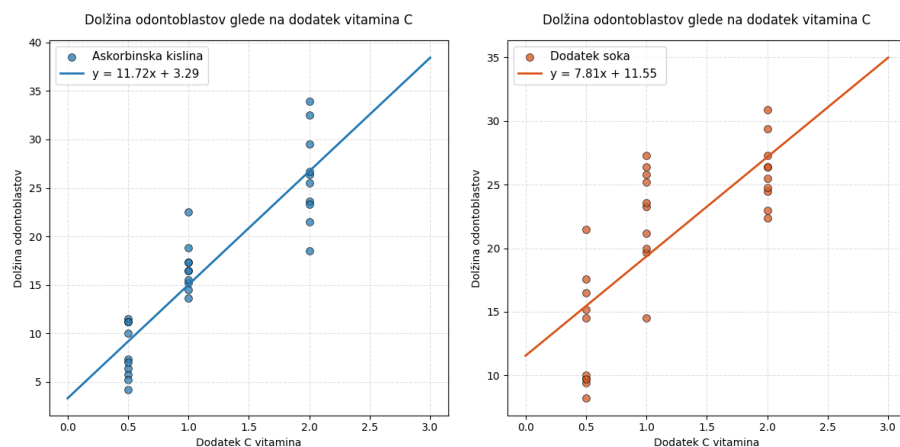
$$\begin{aligned} \text{var}(Q) &= (f + \eta)^T \text{var}(Y)(f + \eta) = \\ &= \sigma^2 (f + \eta)^T (f + \eta) \\ &= \sigma^2 \|f + \eta\|^2 \\ &= \sigma^2 (\|f\|^2 + \|\eta\|^2), \end{aligned}$$

kjer smo upoštevali, da je $\text{var}(Y) = \sigma^2 I$ in da sta f in η pravokotna. Ker je $\sigma^2 > 0$ je varianca naraščajoča funkcija v $\|\eta\|$, zato je minimalna natanko takrat, ko je $\eta = 0$. Tak minimum je zaradi neničelnosti enoličen. NNLC za q je potem $\hat{q} = f^T Y = f^T H Y = f^T \hat{v} = f^T X \hat{\beta} = c^T \hat{\beta}$. S tem je izrek dokazan. $\square$

Sklep je, da bomo β ocenili kar s formulo v enačbi (3.1), zgornji izrek pa veleva, da je to res NNLC. Za vsak način posebej, torej za dodajanje z askorbin-sko kislino in s pomarančnim sokom, glede na enačbo (3.1) izračunamo ustrezno cenilko za β in rezultata predstavimo na spodnjih grafih. Ker nas zanima tudi, kako dobro se regresijski premici prilegata danim podatkom, izračunamo še *povprečno kvadratično napako* (ang. *mean squared error*), ki jo označimo z MSE. Njeno vrednost določimo kot

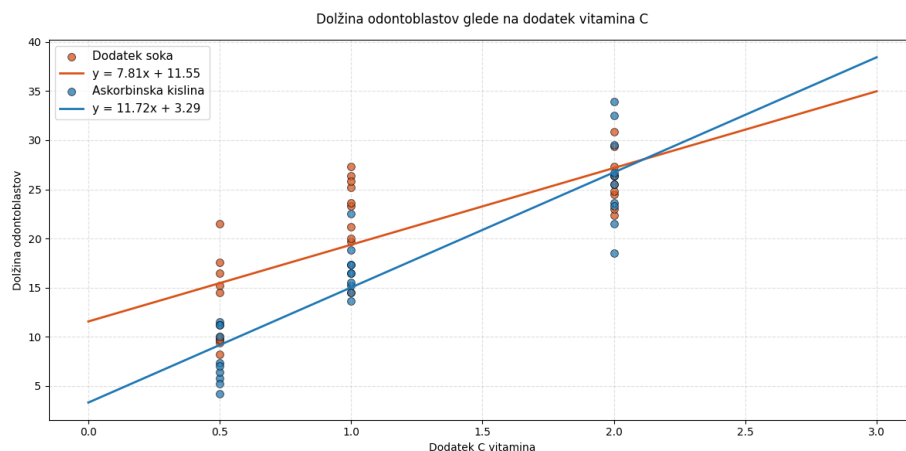
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y - \hat{Y})^2,$$

kjer je $\hat{Y}$ vektor vrednosti, ki jih dobimo z uporabo $\hat{\beta}$.



Slika 16: Regresijske premice za načine dodajanja

Poglejmo si še obe premici na istem grafu, da lažje primerjamo razlike v vplivih glede na način dodajanja.

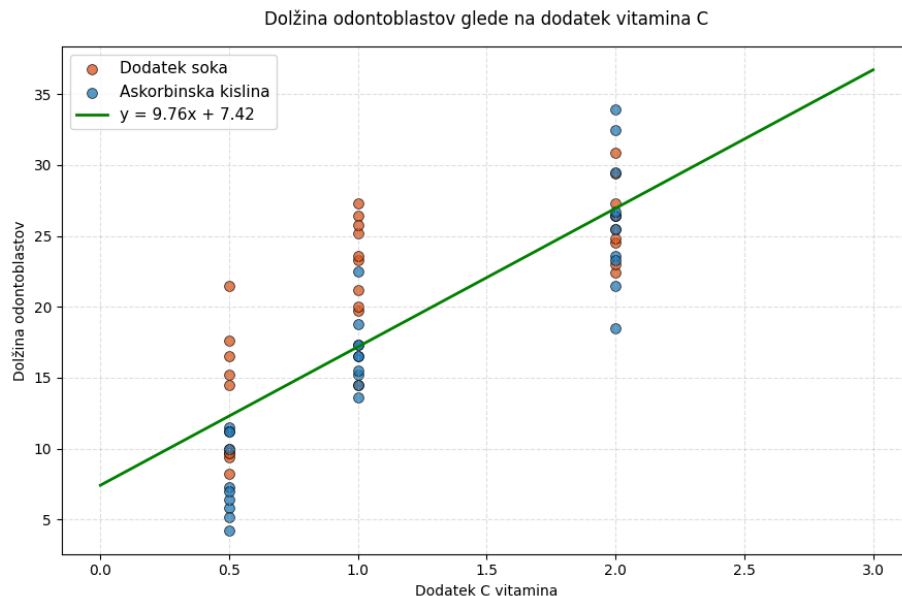


Slika 17: Primerjava obeh regresijski premic

Iz slike 17 sklepamo, da dodajanje vitamina C vpliva na dolžino odontoblastov, saj imata obe regresijski premici znaten naklon. V primeru, da vpliva ne bi bilo, bi naklon premice to odražal s tem, da bi bil blizu ničle, premica pa blizu vodoravne premice.

Do sedaj smo ocenjevali vsak način zgolj s svojimi podatki. Poskusimo oceniti regresijsko premico za vse podatke in si pogledjmo, kaj nam ta pove o povezavi med vitaminom C in dolžino. Na grafu 18 vidimo, da premica relativno

strmo narašča tudi, ko upoštevamo vse podatke in glede na njih prilagodimo zgolj eno premico.



Slika 18: Linearna premica, ki se prilega vsem podatkom

Nakloni vseh treh regresijskih premic so torej znatni, kar ni presenetljivo. Zanimivo se mi zdi, da imata različna načina dodajanja opazno različni začetni vrednosti. Iz grafa namreč sklepamo, da dodajanje vitamina C s pomarančnim sokom sprva bolj vpliva na rast odontoblastov kot dodajanje vitamina C s pomočjo askorbinske kisline. Oranžna premica, ki pripada dodajanju s pomarančnim sokom, se namreč sprva nahaja nad modro premico, ki pripada dodajanju z askorbinsko kislino. Dodajanje z askorbinsko kislino torej spodbuja več rasti odontoblastov, a je sprva učinek manjši.

Vrednosti za β_0 , β_1 in MSE vseh treh premic povzamemo v tabeli 7, iz katere sklepamo o številnih ugotovitvah. Premica, ki se prilega vsem podatkom, slabše opisuje povezave med količinama, saj je njen MSE večji od MSE krivulj, ki modelirajo zgolj svoje pripadajoče podatke. To seveda ni presenetljivo, saj se tudi premici, ki ustrezata svojem načinu dodajanja, opazno razlikujeta tako v začetni vrednosti kot tudi v naklonu.

Dalje opazimo, da so nakloni vseh treh premic razmeroma veliki, kar nakazuje, da dodajanje vitamina C pozitivno vpliva na dolžino odontoblastov.

Končno lahko sklepamo tudi, da sta zaradi različnega naklona in začetnih vrednostih premic, ki sta modelirani na pripadajočih podatkih, načina dodajanja različno učinkovita, kar pa spet ni presenetljiv sklep.

Spomnimo se, da so vrednosti za MSE v spodnji tabeli povprečne *kvadrirane* napake. Korenjene vrednosti, torej RMSE, se gibljejo med 3.5 in 4.5, kar pa

glede na razpon podatkov sploh ni slabo.

Metoda	β_0	β_1	MSE
Askorbinska kislina	3.29	11.72	12.67
Pomarančni stok	11.55	7.81	18.45
Vsi podatki	7.42	9.76	20.47

Tabela 7: Vrednosti parametrov za regresijske premice

3.2 Ocenjevanje s korensko regresijo

V prejšnjem poglavju smo se posvetili ocenjevanju z linearnim regresijskim modelom, ki je najbolj ustrezen za količine, ki so med sabo linearno odvisne. Tak model pri količinah, ki so, recimo kvadratno sorazmerne, nikakor ne bo dobro opisal te povezave. Ker podatki, ki so prikazani na sliki 14, spominjajo na točke pri konkavni funkciji, poskusimo s korensko regresijo. Omenimo pa, da je to vsekakor zgolj ugibanje in iskanje rešitve na osnovi ideje, ne pa na osnovi jasnih kriterijev in algoritmov. Nastavek je identičen kot pri linearni regresiji. Postavimo model

$$y_i = \beta_0 + \beta_1 \sqrt{x_i} + \varepsilon_i,$$

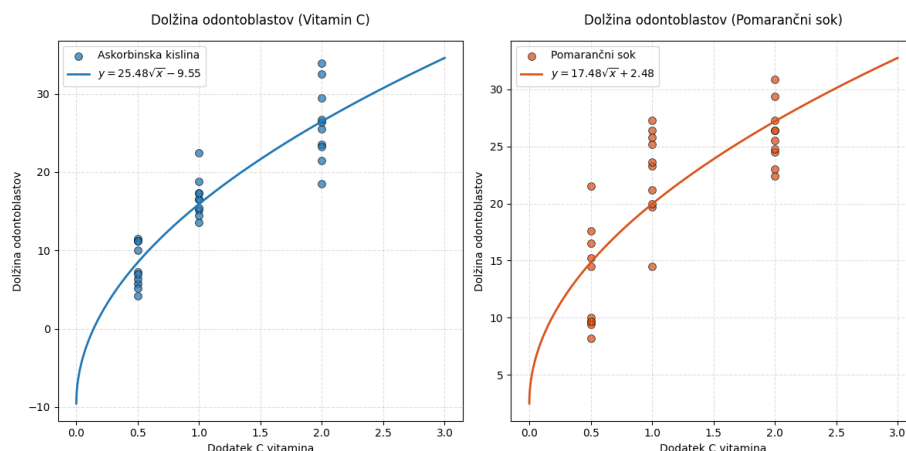
ki ga lahko ponovno zapišemo v matrični obliki kot

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \sqrt{x_1} \\ 1 & \sqrt{x_2} \\ \vdots & \vdots \\ 1 & \sqrt{x_n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \quad (3.2)$$

Tak model ni nič drugega kot običajen linearno regresijski model, kjer smo začetne x_i transformirali s preslikavo $x_i \mapsto \sqrt{x_i}$. Ravno zaradi tega, ker enačba (3.2) ustreza linearnemu modelu za ustrezno transformirane podatke, lahko najboljšo nepristransko linearno cenilko iščemo na enak način, kot smo to počeli v prejšnjem poglavju. Spet je torej

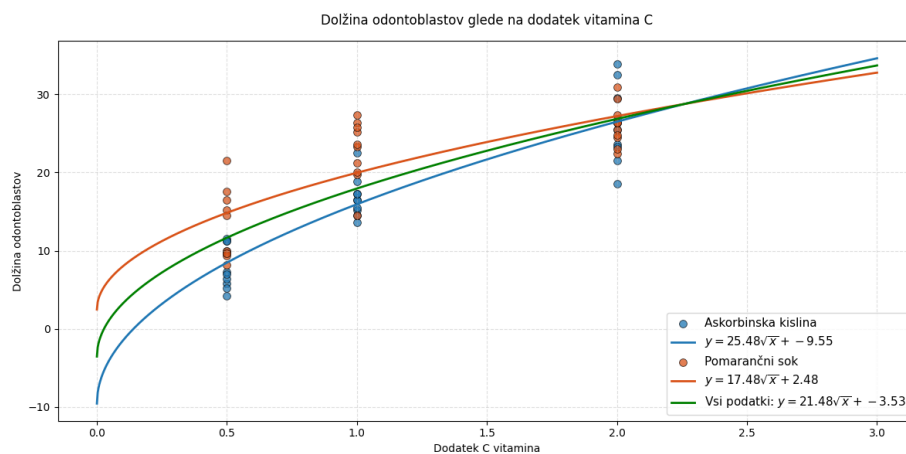
$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

za ustrezna Y in X . Z enakim postopkom kot prej, upoštevajoč transformacijo $x_i \mapsto \sqrt{x_i}$, dobimo naslednje rezultate.



Slika 19: Korenske regresijske krivulje na ustreznih podatkih

Za lažjo primerjavo si pogledjmo še obe cenliki na istem grafu, kjer dodatno narišemo še krivuljo, ki pripada celotnemu narobu podatkov.



Slika 20: Preimerjava obeh korenskih krivulj

Vrednosti za β_0 , β_1 in MSE za vse tri korenske krivulje se nahajajo v tabeli

8

Metoda	β_0	β_1	MSE
Askorbinska kislina	-9.55	25.48	11.42
Pomarančni sok	2.48	17.48	16.47
Vsi podatki	-3.53	21.48	18.71

Tabela 8: Vrednosti parametrov za splošno korensko regresijo

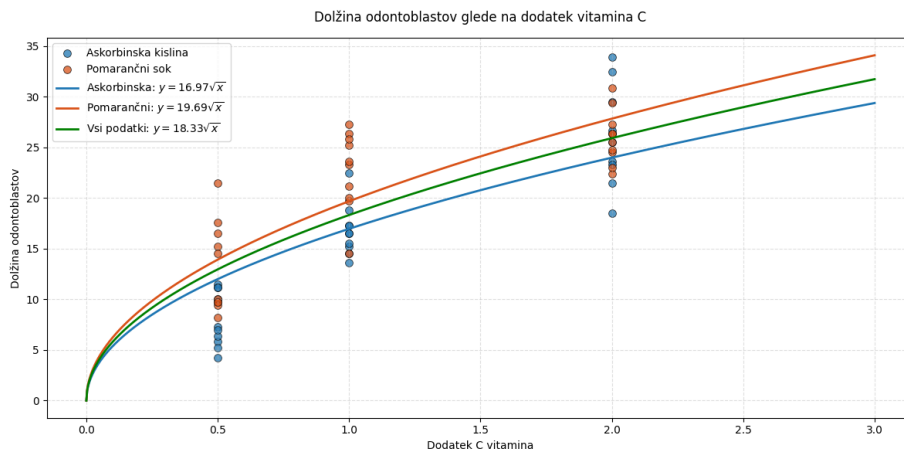
Iz vrednosti MSE v zgornji tabeli vidimo, da se korenske krivulje bolje prilegajo podatkom, saj je napaka manjša pri vseh treh krivuljah, kot je bila pri linearnih modelih. Kljub temu z modeli nismo zadovoljni. Rezultat za β_0 pri dodajanju z askorbinsko kislino in pri upoštevanju vseh podatkov je negativen, kar bi implicirano, da je dolžina odontoblastov brez dodanega vitamina C negativna. To ni smiselno, vrednost mora seveda biti pozitivna, v dobrem modelu pa bi bila enaka za oba načina dodajanja. Poskusimo ponovno uporabiti korenski model, le da odstranimo prosti člen. Nastavimo torej

$$y_i = \beta_i \sqrt{x_i} + \varepsilon_i$$

in si oglejmo rezultate. Podobno kot v zgornjih primerih sta na sliki 21 prikazani krivulji za vsako skupino, ter dodatno še krivulja celotne zbirke podatkov, v tabeli 9 pa se nahajajo tudi vrednosti za β in MSE za vse tri krivulje.

Metoda	β_0	β_1	MSE
Askorbinska kislina	–	16.97	17.99
Pomarančni sok	–	19.69	16.91
Vsi podatki	–	18.33	19.62

Tabela 9: Vrednosti parametrov za korensko regresijo brez β_0 .



Slika 21: Korenske krivulje brez prostega člena

Iz dane tabele lahko sedaj sklepamo o pravilnosti našega postopka. Napaka je praktično enaka pri zadnjih dveh metodah, torej pri dodajanju pomarančnega soka in pri upoštevanju vseh podatkov, se pa napaka pri dodajanju askorbinske kisline precej poveča. Kljub temu sem mnenja, da je model brez prostega člena vsebinsko bolj smislen od modela, ki implicira negativne dolžine.

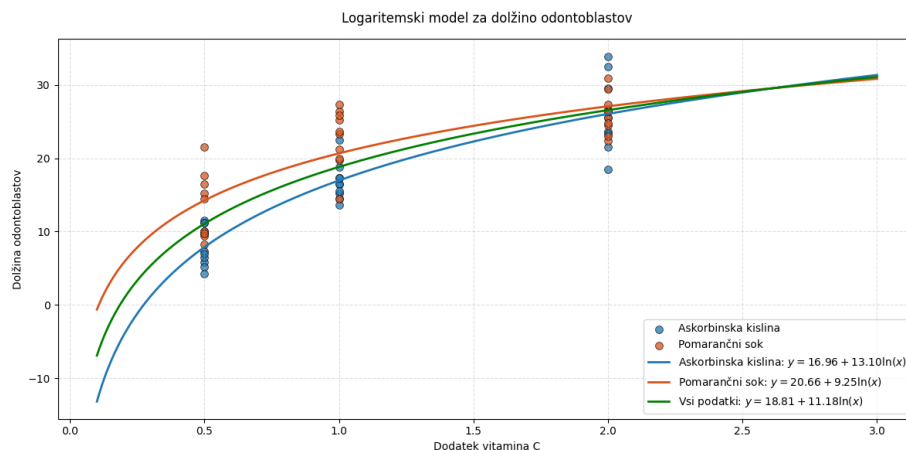
3.3 Logaritemska regresija

Enak argument kot zgoraj izloči tudi regresijski model oblike

$$y_i = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i,$$

saj bi tak model zahteval izredno negativne vrednosti za dolžino celilc, preden bi sploh dodajali vitamin C, kar spet ni v skladu z resničnostjo. Formalna pravilnost tega modela sledi iz enakega argumenta kot pravilnost uporabe korenskega modela: postavljen model je zgolj linearen regresijski model za podatke, ki jih prej transformiramo z logaritmom.

Kljub temu da imamo težave pri ničli, si vseeno pogledjmo, kako izgledajo ustrezne krivulje.



Slika 22: Logaritemski model za dolžino

Metoda	β_0	β_1	MSE
Askorbinska kislina	16.96	13.10	11.08
Pomarančni sok	20.66	9.25	14.74
Vsi podatki	18.81	11.18	17.52

Tabela 10: Vrednosti parametrov za logaritemsko regresijo.

V tabeli smo lahko presenečeni nad vrednostmi MSE, saj so napake nižje od tistih, pri korenski regresiji, in hkrati opazno nižje tudi od tistih, ki smo jih dobili pri linearnih modelih. Dejstvo je po eni strani presenetljivo, saj je model v smislu napak boljši, kljub temu da implicira nesmiselne stvari blizu ničle. Po drugi strani pa smo že na začetku opazili obliko, ki je spominjala na logaritemsko, za katero pa vemo, da je v naravi pogosto prisotna.

Vsekakor smo z različnimi regresijskimi modeli pokazali, da je trditev, da več dodanega vitamina C pomeni daljše odontoblaste, utemeljeno zastavljena.

Regresija	Metoda	β_0	β_1	MSE
Linearna	Askorbinska kislina	3.29	11.72	12.67
Linearna	Pomarančni sok	11.55	7.81	18.45
Linearna	Vsi podatki	7.42	9.76	20.47
Korenska	Askorbinska kislina	-9.55	25.48	11.42
Korenska	Pomarančni sok	2.48	17.48	16.47
Korenska	Vsi podatki	-3.53	21.48	18.71
Korenska brez β_0	Askorbinska kislina	–	16.97	17.99
Korenska brez β_0	Pomarančni sok	–	19.69	16.91
Korenska brez β_0	Vsi podatki	–	18.33	19.62
Logaritemska	Askorbinska kislina	16.96	13.10	11.08
Logaritemska	Pomarančni sok	20.66	9.25	14.74
Logaritemska	Vsi podatki	18.81	11.18	17.52

Tabela 11: Primerjava parametrov različnih regresijskih modelov.

3.4 Testiranje regresijskih modelov s hipotezo

Sedaj želimo pokazati, da je koeficient β_1 v modelu

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

dovolj velik, da je njena vrednost z dano stopnjo tveganja malo verjetna pod ničelno hipotezo. Postavimo torej ničelno in alternativno hipotezo kot

$$\begin{aligned} H_0 : \beta_1 &= 0, \\ H_1 : \beta_1 &\neq 0. \end{aligned}$$

Vemo že, da za Gaussove modele, torej modele, kjer je $\varepsilon \sim N(0, \sigma^2 I)$, velja, da je

$$Z = \frac{\hat{\beta} - \beta}{\sigma \sqrt{\text{var}(\hat{\beta})}} \sim N(0, 1)$$

in da je tudi

$$\frac{(n-2)\hat{\sigma}^2}{\sigma} \sim \chi^2(n-2).$$

Spremenljivki Z in $\hat{\sigma}^2$ sta neodvisni, zato iz njiju naredimo kvocient

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} \sqrt{\text{var}(\hat{\beta}_1)}} \sim t_{n-2}$$

Ko torej preizkušamo hipotezo H_0 pri stopnji tveganja α , bomo zaradi simetrije porazdelitve na vsaki strani odrezali $\alpha/2$, torej bomo hipotezo sprejeli, če bo

$$-t_{crit} \leq t \leq t_{crit},$$

kjer

$$t_{crit} = t \left(1 - \frac{\alpha}{2} \right)_{df}.$$

Ostane izračunati zgolj še vrednost naše statistike in jo primerjati z mejnimi vrednostmi pri stopnjah tveganja $\alpha_1 = 0.05$ in $\alpha_2 = 0.01$. Ker imamo 60 meritev in je $\beta = [\beta_0, \beta_1]^T$, je torej $n = 60$ in $m = 2$. Število prostostni stopenj je potem 58, kar za pomeni, da za α_1 kritična vrednost statistike ± 2.0017 , pri α_2 pa je kritična vrednost ± 2.6633 . Vrednost statistike izračunamo tako, da v formulo vstavimo vse opažene vrednosti in podatke in dobimo, da je vrednost naše statistike za tak primer $t = 10.2501$. To je močno nad obema mejama, zato zavrnemo ničelno hipotezo $H_0 : \beta_1 = 0$.

Podoben postopek lahko seveda izvedemo za vse ostale regresijske modele in njihove rezultate zapišemo v spodnjo tabelo.

Statistika	Linearni	Korenski	Korenski (brez β_0)	Logaritemski
$\hat{\beta}_0$	7.4225	-3.5341	–	18.8133
$\hat{\beta}_1$	9.7636	21.4788	18.3271	11.1773
SE	0.9525	1.9581	0.5338	0.9711
df	58	58	59	58
t -stat.	10.2501	10.9690	34.3316	11.5093
MSE	20.4651	18.7145	19.6154	17.5210
$t_{crit,0.05}$	2.0017	2.0017	2.0010	2.0017
Odločitev (α_1)	Zavrni H_0	Zavrni H_0	Zavrni H_0	Zavrni H_0
$t_{crit,0.01}$	2.6633	2.6633	2.6618	2.6633
Odločitev (α_2)	Zavrni H_0	Zavrni H_0	Zavrni H_0	Zavrni H_0

Tabela 12: Rezultati testiranja $H_0 : \beta_1 = 0$ za štiri regresijske modele.

V tabeli opazimo, da smo zavrnili prav vse ničelne hipoteze, vrednost testne statistike je pa bistveno večja od robnih vrednosti. To kaže na dejstvo, da dodajanje vitamina C gotovo vpliva na dolžino celic, ta sklep pa sedaj ni več ugibanje, pač pa smo s hipotezami formalno pokazali, da je učinek gotov. Kot zanimivost lahko povemo, da je verjetnost, da studentova porazdelitev s 58 prostostnimi stopnjami, zavzame vrednost manjšo od 10 kar

$$P(T_{58} \leq 10) = 0.9999999999999845,$$

če se nisem zmotil pri preštevanju devetic. Res je torej vpliv gotov.

3.5 Kateri način je boljši

V prejšnjem podpoglavju smo pokazali, da dodajanje vitamina C gotovo vpliva na dolžino odontoblastov, a je pomembno izpostaviti, da pri tem preizkusu nismo ločevali med načinom dodajanja. Pogosto pa nas zanima tudi to, kateri način dodajanja je boljši. Dejstvo, da je dodajanje pozitivno povezano z dolžino

bi lahko namreč z veliko gotovostjo ugotovili že na podlagi prve slike podatkov, nekoliko težje pa bi kljub vsem predstavljenim regresijskim modelom znali sklepati, kateri način dodajanja je učinkovitejši. Res pa je, da bi lahko pozorni bralci v prejšnjih podpoglavjih opazili, da je regresijska krivulja za dodajanje s pomarančnim sokom, torej oranžna krivulja, vedno nad krivuljama, ki sta modelirali tako dodajanje z askorbinsko kislino in obe vrsti dodajanja. Ker je pri linearnem modelu krivulja za pomarančni sok bolj položna od krivulje askorbinske kisline, kar pomeni manjšo razliko v dolžini celic, ugibamo, da bo dodajanje s pomarančnim sokom manj učinkovito s časom. Poglejmo si še, kako bi takšno hipotezo formalno dokazali in pokazali, da je statistično značilna.

Idejno bomo nalogo reševali tako, da bomo ustvarili en regresijski model, ki bo uspel ločiti med danimi podatki. Postavimo torej model, ki naj ima obliko

$$y_i = \beta_0 + \beta_1 x_i + \gamma_0 D_i + \gamma_1 D_i x_i + \varepsilon_i, \quad \text{za} \quad D_i = \begin{cases} 0, & \text{askorbinska kislina} \\ 1, & \text{pomarančni sok} \end{cases},$$

kar lahko spet v matrični obliki zapišemo kot

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & D_1 & D_1 x_1 \\ 1 & x_2 & D_2 & D_2 x_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & D_n & D_n x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \gamma_0 \\ \gamma_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Tak nastavek za model ni presenetljiv. Če podatke uredimo, da so najprej vsi podatki z askorbinsko kislino, nato pa podatki za pomarančni sok, dobimo bločno matriko, kjer je desno zgornji blok ničelen. Vidimo da gre res torej za združevanje dveh regresijskih modelov v enega. Koeficienta β_0 in β_1 sta koeficienta, ki pripadata podaktom za askorbinsko kislino, vrednosti γ_0 in γ_1 interpretiramo kot popravka, ko v model vključimo še preostale podatke. To dejstvo nam pove tudi, kaj bosta naša ničelna in alternativna domena. Definiramo namreč

$$\begin{aligned} H_0 &: \gamma_1 = 0 \\ H_1 &: \gamma_1 < 0, \end{aligned}$$

kjer za H_1 izbrali tako alternativno domnevo, ker pričakujemo, da bo pomarančni sok manj učinkovit.

Pod sprejemljivo predpostavko, da je $\varepsilon \sim N(0, \sigma^2 I)$, dobimo, da je $\hat{\beta} = (X^T X)^{-1} X^T Y$, zato pa je $\text{var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$, kjer je $\hat{\sigma}^2 = \frac{1}{n-4} \sum_i (Y_i - \hat{Y}_i)^2$, saj je sedaj vektor $\hat{\beta} \in \mathbb{R}^4$. Poleg tega vemo tudi, da je pod H_0

$$t = \frac{\hat{\gamma}_1 - 0}{\text{SE}(\hat{\gamma}_1)} \sim t_{n-4},$$

zato za preverjanje hipotez uporabimo lahko identičen postopek kot v prejšnjem podpoglavju. V formulo vstavimo vse podatke, ki jih imamo na voljo in pridobimo naslednje rezultate.

Za $\alpha_1 = 0.05$ je kritična vrednost $t(\alpha_1)_{n-4} = -1.6725$. Vrednost naše statistike je $t = -2.3094$, kar je manj od kritične vrednosti. Sledi torej, da bomo zavrnili ničelno domnevo. Zanimivo pa je, da je za $\alpha_2 = 0.01$ kritična točka $t(\alpha_2)_{n-4} = -2.3948$, kar je manj od vrednosti naše statistike, zato ničelne hipoteze ne moremo zavrniti.

Seveda lahko enak pristop, ki smo ga uporabili pri linearnem modelu, uporabimo tudi za ostale predlagane modele. Splošni model z indikatorjem je potem

$$y_i = \beta_0 + \beta_1 f(x_i) + \gamma_0 D_i + \gamma_1 D_i f(x_i) + \varepsilon_i, \quad D_i = \begin{cases} 0, & \text{askorbinska kislina} \\ 1, & \text{pomarančni sok} \end{cases},$$

kjer pišemo $f(x) = x$ za linearni model, $f(x) = \sqrt{x}$ za korenski model in $f(x) = \ln x$ za logaritemski model. Tudi pri teh modelih bo γ_1 meril razliko pri naklonu, zato bomo gledali spet

$$H_0 : \gamma_1 = 0 \quad \text{proti} \quad H_1 : \gamma_1 < 0,$$

Ocene $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0, \hat{\gamma}_1)$ dobimo z enakim NNLC postopkom kot prej,

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad X = [\mathbf{1}, f(x), D, D \cdot f(x)],$$

testna statistika pa je spet

$$t = \frac{\hat{\gamma}_1}{\text{SE}(\hat{\gamma}_1)} \sim t_{n-4} \text{ pod } H_0.$$

Ko vstavimo podatke za vse tri modele, lahko oblikujemo spodnjo tabelo, v kateri opazimo, da za tri modele lahko zavrnemo H_0 pri α_1 , nobenega pa ne moremo zavrniti pri α_2 . Odgovor na vprašanje o statistični značilnosti je torej odvisen od tega, kakšno stopnjo tveganja predpišemo.

Statistika	Linearni	Korenski	Korenski (brez)	Logaritem
$\hat{\beta}_0$	3.2950	-9.5461	–	16.9633
$\hat{\beta}_1$	11.7157	25.4790	16.9658	13.0997
$\hat{\gamma}_0$	8.2550	12.0239	–	3.7000
$\hat{\gamma}_1$	-3.9043	-8.0003	2.7226	-3.8448
$SE(\hat{\gamma}_1)$	1.6906	3.4405	1.0157	1.6971
df	56	56	58	56
t -stat.	-2.3094	-2.3253	2.6804	-2.2655
MSE	15.5606	13.9455	17.4534	12.9148
$t_{krit,0.05}$	-1.6725	-1.6725	-1.6716	-1.6725
Pri (α_1)	Zavrni H_0	Zavrni H_0	Ne zavrnj H_0	Zavrni H_0
$t_{krit,0.01}$	-2.3948	-2.3948	-2.3924	-2.3948
Pri (α_2)	Ne zavrnj H_0	Ne zavrnj H_0	Ne zavrnj H_0	Ne zavrnj H_0

Tabela 13: Rezultati testiranja $H_0 : \gamma_1 = 0$ proti $H_1 : \gamma_1 > 0$.

3.6 Izboljšave in komentarji k tretji nalogi

Gotovo bi bila naloga bolj zanimiva in informativna, če bi imeli tudi podatke o kontrolni skupini morskih prašičkov. Tako bi lahko tudi nastavili pravilno vrednost za koeficient β_0 v vseh regresijskih modelih, modelirane vrednosti pa bi potem lahko primerjali tudi z dolžinami odontoblastov v kontrolni skupini. Poleg tega bi s tem pridobili več informacij tudi glede skale in velikostnega razreda učinka. Če bi bile dolžine v kontrolni skupini recimo 3 enote, bi sklepali da dodajanje dveh enot vitamina C za faktor 8 poveča dolžino. V vsakem primeru bi bile te informacije koristne in bi iz njih lahko sklepali še o drugih lastnostih dodajanja. Zanimiva možnost, ki bi se lahko zgodila, bi bila, da se vitamin C različno dobro absorbira, glede na dodano količino. To bi preverili z modelom, sestavljenim iz dveh linearnih modelov, kjer bi prvi del, torej med 0 in 1 dodano enoto vitamina, opazno hitreje naraščal, kar bi kazalo na večji vpliv pri nižjih količinah in manjši vpliv zaradi "nasičenosti" pri večjih količinah. Tak dvodelni model bi bil tudi bolj prilagodljiv kot en korenski ali logaritemski, čeprav bi bili obliki na videz lahko podobni. Z dvodelnim modelom imamo namreč možnost poljubne strmine na vsakem posameznem odseku.

Dodatno bi pri kontrolni skupini lahko raziskovali učinek placeba pri morskih prašičkih, ki verjetno ne obstaja. Kljub temu je vprašanje, ali tudi druge živali doživljajo placebo oziroma podobne učinke, zanimivo samo po sebi, čeprav ni matematično vprašanje.

Poleg dodatka kontrolne skupine bi se lahko drugače lotili tudi preverjanja, kateri način dodajanja je učinkovitejši. S trenutno rešitvijo smo se uspeli izogniti pretiranemu delu, saj smo na pameten način zakodirali ločevanje metode dodajanja. Lahko bi recimo postavili model

$$Y_{ij} = \beta_0 + \alpha_i + \delta_j + (\alpha\delta)_{ij} + \varepsilon_{ij},$$

kje bi α_i predstavljal učinek dodajanja vitamina, δ_j bi predstavljal učinek metode, $(\alpha\delta)_{ij}$ pa bi upošteval prepletenost obeh. Tedaj bi lahko za ničelno hipotezo izbrali $H_0 : \delta_{O,J} = \delta_{V,C}$ in tako preverili, ali sta učinka enaka. Lahko bi tudi ustvarili tri modele, enega za vsako količino dodajanja, nato pa bi preverili, kateri način je bolj ustrezen pri dani količini vitamina C. Lahko bi tudi pogledali zgolj razliko povprečnih vrednosti obeh načinov in s tem ocenili, kateri način v povprečju vodi v večje dolžine. Načinov je seveda veliko, a se mi je zdel tak zvit model precej zanimiv zaradi kodiranja načina dodajanja v modelu.

Omenim še, da bi lahko hipoteze preverjali tudi na drugačen način. V dodatnih zapiskih pod poglavjem o linearni regresiji je predstavljen postopek, kjer uporabimo razmerje verjeti za potrjevanje oziroma zavračanje hipotez. Ta način temelji na Fisherjevi porazditvi s prostostnimi stopnjami v_1, v_2 .

Definicija 3.3. *Fisherjeva porazdelitev* s prostostnimi stopnjami (v_1, v_2) je porazdelitev slučajne spremenljive

$$\frac{H_1/v_1}{H_2/v_2},$$

kjer sta $H_1 \sim \chi^2(v_1)$ in $H_2 \sim \chi^2(v_2)$ neodvisni slučajni spremenljivki.

V preizkusu na mestu H_1 uporabimo razliko med residuali alternativne in ničelne hipoteze, za H_2 pa zgolj rezidual pri ničlni hipotezi. Takšnemu preizkusu pravimo F -test in je zelo pogost pri preverjanju hipotez.

4 Zaključek in povzetek

V projektni nalogi za statistiko smo reševali tri naloge. Prva naloga je preverjala znanje intervalov zaupanja in vpeljala nov pojem stratificiranega vzorčenja, kot dodatek enakomernemu slučajnem vzorčenju, ki smo ga spoznali na predavanjih. Videli smo, da stratificirano vzorčenje in slučajno vzorčenje vrmeta približno enake rezultate, le da so rezultati pri stratificiranem vzorčenju manj razpršeni. Oba načina predstavljata dobro vzorčenje za določanje povprečnega dohodka v mestu *Kibegrad*. Razlog za to se skriva tudi v porazdelitvi dohodka znotraj vsake četrti in pa v razdelitvi populacije v četrti same, saj je porazdelitev približno enakomerna v vsaki četrti, odstopanja pa niso pretirano velika. Poleg tega smo pokazali, da porazdelitev dohodka v mestu gotovo ni blizu normalne porazdelitve, zato je potrebno imeti v mislih, da predpostavljene lastnosti večinoma ne veljajo.

Pri drugi nalogi smo preko metode momentov in metode največjega verjetja poiskali cenilko za parameter θ za dano diskretno slučajno spremenljivko. Utemeljili smo nepristranskost obeh tako, da smo izračunali pričakovano vrednost cenilke in pokazali, da je $\mathbb{E}(\hat{\theta}) = \theta$ za obe cenilki. Nato smo na parameter θ pogledali skozi oči Bayesove statistike, kjer smo ji najprej določili apriorno porazdelitev, nato pa na podlagi opažanj temu priredili posteriorno porazdelitev. Pri teh računih smo si precej pomagali z znano beta porazdelitvijo.

V tretji nalogi smo analizirali vpliv dodajanja vitamina C na dolžino odontoblastov preko treh regresijskih modelov, nato pa smo na podlagi zavrženih ničelnih domnev pokazali, da količina dodanega vitamina C gotovo vpliva na dolžino. Pri tem se končni rezultat ni spremenil pri različnih regresijskih modelih, smo pa zaznali različne napake pri modelih. Najboljši model za dane podatke je bil logaritemski model, če smo modele vrednotili po njihovi MSE. Tako kot smo pri prvi nalogi predpostavili lastnosti, ki niso veljale, tudi tukaj model govoto ne bo razložil zveze pretirano dobro, saj implicira negativne dolžine, če ne bi dodajali vitamina C. Kljub tej pomanjkljivosti se zdi logaritemski model verjeten zgolj zaradi dejstva, da se logaritemske zveze pogosto pojavljajo v naravnih pojavih.

Pri reševanju nalog sem si pomagal s spodaj navedenimi viri, še posebej koristen pa je bil vir [3], s katerim sem si pomagal pri reševanju druge naloge. Poleg tega sem uporabljal različne knjižnice za Python kot so `numpy`, `pandas`, `matplotlib` in druge. Pri tem sem seveda veliko uporabljal njihovo dokumentacijo, ki jo tudi spodaj navajam kot vir, a zgolj za določene metode. Tekom naloge sem tudi ogromno prebiral zapiske iz predavanj, vaj in dodatne zapiske, ki so delo mentorja doc. dr. Matrina Raiča. Primer uporabe je recimo dokaz izreka Gauss-Markova, ki sem ga poiskal v dodatnih zapiskih.

Kot že omenjeno v zadnjih podpoglavjih vsake naloge, je možnosti za izboljšavo in nadgradnjo veliko. Obstaja veliko testov, ki jih nismo omenjali na predavanjih, ki bi lahko pomagali pri razumevanju vsebine podatkov in sklepanju o povezavah med njimi. Nekatere razširitve sem vključil v svoje poročilo, nekatere pa sem zavoljo dolžine in časa izdelave ter dodatnih poglavjih, ki bi jih moral napisati, pustil zgolj kot predlog ali misel. Vsekakor pa sem tekom

pisanja užival, še posebej zaradi zavedanja, da so takšne naloge izredno pogoste in pomembne v današnjem času, kjer podjetja vsako odločitev sprejemajo na podlagi podatkov in analiz, ne pa na podlagi občutkov in ugibanja.

Okvirno sem vsako nalogo reševal v svoji datoteki v formatu interaktivne beležke, torej datoteki s formatom `.ipynb`. To mi je omogočalo pregled nad delom in pa izvajanje in reševanje naloge po korakih, brez ponovne generacije velikega števila grafov in tabel. Ker je programov in korakov veliko, delno tudi zaradi lastne želje po razširitvi obsega naloge, se mi ne zdi smiselno dodajati izpisov programov na koncu poročila, vsaj ne v celoti. V naslednji podpoglavjih so izpisani bolj pomembni rezultati programov.

4.1 Del izpisa za prvo nalogo

Slučajni vzorec 400 družin z vsemi podatki:

Povprečen odhodek v Kibergradu, ce gledamo vzorec 400 nakljucnih družin: 41511.6675

Standardni odklon v tem primeru: 34467.0421

Ocena za standardno napako povprecja je 1723.3521

Interval zaupanja je 38123.68 <= mu <= 44899.65

Slučajni vzorec 400 družin brez osamelcev:

Povprečen odhodek v Kibergradu, ce gledamo vzorec 400 nakljucnih družin brez osamelcev: 36035.575

Standardni odklon v tem primeru: 25186.1333

Ocena za standardno napako povprecja je 1259.3067

Interval zaupanja je 33559.87 <= mu <= 38511.28

Stratificirano vzorečenje 400 družin z vsemi podatki:

Povprečen dohodek, ce gledamo proporcionalno stratifikacijo: 41553.3561

Ocena za standardno napako povprecja je 1639.0000

Interval zaupanja je 38331.2 <= mu <= 44775.51

Stratificirano vzorčenje 400 družin brez osamelcev:

Povprečen dohodek, ce gledamo proporcionalno stratifikacijo brez osamelcev: 38422.1059 Ocena za standardno napako povprecja je 1237.2373

Interval zaupanja je 35989.79 <= mu <= 40854.42

4.2 Del izpisa za drugo nalogo

Pri drugi nalogi nisem uporabljal programov, razen za risanje grafov in numerično računanje integralov.

4.3 Del izpiza za tretjo nalogo

Program pri treji nalogi je izjemno ločen glede na posamezne regresijske modele, zato na tem mestu navedimo zgolj enega izmed izpisov, ki je naslednja tabela. Ta tabela je enaka kot tabela 12.

Statistika	Linearni	Korenski	Korenski (brez β_0)	Logaritemski
$\hat{\beta}_0$	7.4225	-3.5341	–	18.8133
$\hat{\beta}_1$	9.7636	21.4788	18.3271	11.1773
SE	0.9525	1.9581	0.5338	0.9711
df	58	58	59	58
t -stat.	10.2501	10.9690	34.3316	11.5093
MSE	20.4651	18.7145	19.6154	17.5210
$t_{crit,0.05}$	2.0017	2.0017	2.0010	2.0017
Odločitev (α_1)	Zavrni H_0	Zavrni H_0	Zavrni H_0	Zavrni H_0
$t_{crit,0.01}$	2.6633	2.6633	2.6618	2.6633
Odločitev (α_2)	Zavrni H_0	Zavrni H_0	Zavrni H_0	Zavrni H_0

Tabela 14: Del izpisa za program pri tretji nalogi

Literatura

- [1] NumPy Developers. *numpy.histogram_bin_edges*. NumPy v2 Manual. Dostop: 2026-07-26. URL: https://numpy.org/doc/stable/reference/generated/numpy.histogram_bin_edges.html.
- [2] Pandas Development Team. *pandas.DataFrame.hist*. pandas documentation. Dostop: 2026-07-26. URL: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.hist.html>.
- [3] John A Rice in John A Rice. *Mathematical statistics and data analysis*. Zv. 371. Thomson/Brooks/Cole Belmont, CA, 2007.
- [4] Waskom, Michael L. and Seaborn Development Team. *seaborn.histplot*. seaborn: statistical data visualization. Dostop: 2026-07-26. URL: <https://seaborn.pydata.org/generated/seaborn.histplot.html>.
- [5] Wikipedia contributors. *P-value*. Wikipedia, The Free Encyclopedia. Dostop: 2026-07-26. URL: <https://en.wikipedia.org/wiki/P-value>.
- [6] Wikipedia contributors. *Q-Q plot*. Wikipedia, The Free Encyclopedia. Dostop: 2026-07-26. URL: <https://en.wikipedia.org/wiki/Q%E2%80%93plot>.
- [7] Wikipedia contributors. *Scott's rule*. Wikipedia, The Free Encyclopedia. Dostop: 2026-07-26. URL: https://en.wikipedia.org/wiki/Scott's_rule.
- [8] Wikipedia contributors. *Stratified sampling*. Wikipedia, The Free Encyclopedia. Dostop: 2026-07-26. URL: https://en.wikipedia.org/wiki/Stratified_sampling.