Projetkna naloga pri statistiki

Gaja Jamnik

1 Prva naloga

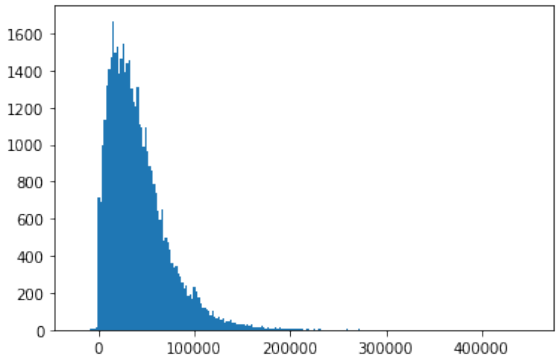
Pri prvi nalogi smo obravnavali podatke 43.886 družin iz mesta Kibergrad. Natančneje, analizirali smo dohodke družin in njihovo porazdelitev primerjali z ustrezno normalno porazdelitvijo.

1.1 A del naloge

Narisali smo histogram dohodkov vseh družin v Kibergradu. Širino in število razredov v histogramu smo določili s pomočjo formule

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

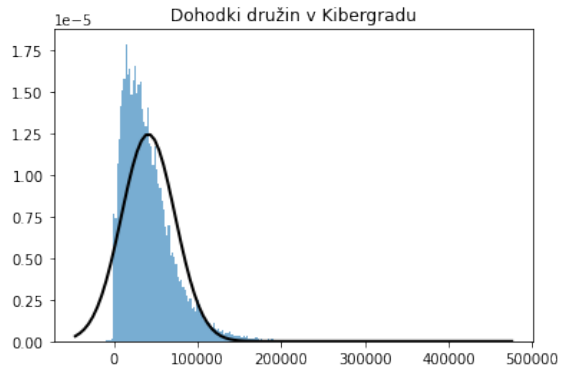
kjer sta $q_{\frac{1}{4}}$ in $q_{\frac{3}{4}}$ prvi in tretji kvartil, n pa je število enot. Tej formuli pravimo Freedman–Diaconisovo pravilo. Po formuli znaša **širina posameznega razreda** približno 2128. Glede na interval na katerem so razporejeni dohodki znaša to ravno 223 **razredov** v histogramu.



Slika 1: Histogram dohodkov družin v Kibergradu

1.2 B del naloge

V B delu naloge smo v histogram dorisali še normalno gostoto, katere pričakovana vrednost se ujema s povprečjem dohodkov družin v Kibergradu, standardni odklon pa s standardnim odklonom podatkov o dohodkih družin. Izračunano **povprečje podatkov** je 41335.50704096979, **standardni odklon** pa 32037.61941788666.

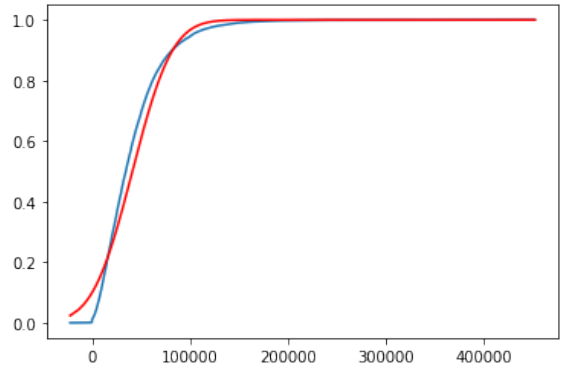


Slika 2: Histogram dohodkov družin v Kibergradu in normalna gostota

Histogram je desno asimetričen in nima značilne *zvončaste oblike*. Graf normalne gostote na desni strani sledi obliki histograma, drugod pa se mu slabo prilega. To nam da misliti, da dohodki niso porazdeljeni normalno.

1.3 C del naloge

V tem delu naloge primerjamo še kumulativni porazdelitveni funkciji podatkov o dohodkih in normalne porazdelitve s pričakovano vrednostjo in standardnim odklonom, kot v nalogi 1.2. Kumulativno porazdelitveno funkcijo porazdelitve dohodkov definiramo s pomočjo funkcije `arange`, tako da definiramo seznam deležev. Za normalno porazdelitev pa uporabimo kar že vgrajeno funkcijo `statistics.NormalDist(mean, std).cdf(x)`.

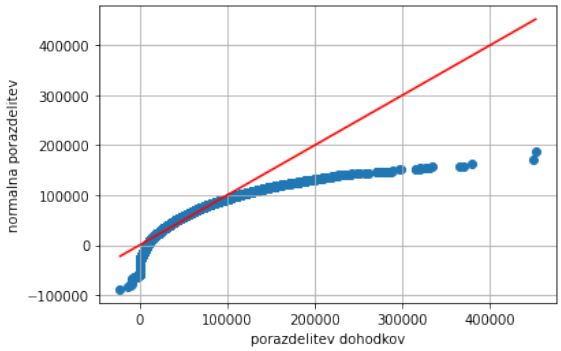


Slika 3: Primerjava kumulativnih porazdelitvenih funkcij

Grafa se od vrednosti 150000 dalje skoraj popolnoma prilegata, kar smo opazili že na grafu 2. Na levi strani pa je nekaj manjših odstopanj.

1.4 D del naloge

S primerjalnim kvantilnim grafikonom oz. Q-Q grafikonom smo primerjali kvantile normalne porazdelitve s kvantili porazdelitve dohodkov družin v Kibergradu.

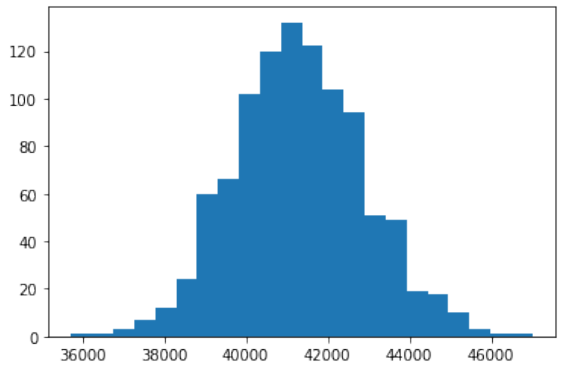


Slika 4: Primerjalni kvantilni grafikon

Če bi bili dohodki normalno porazdeljeni, bi kvantilni grafikon, narisan z modro, sovpadal s 45-stopinjsko premico, narisano z rdečo barvo. Iz grafikona razberemo, da dohodki niso porazdeljeni normalno.

1.5 E del naloge

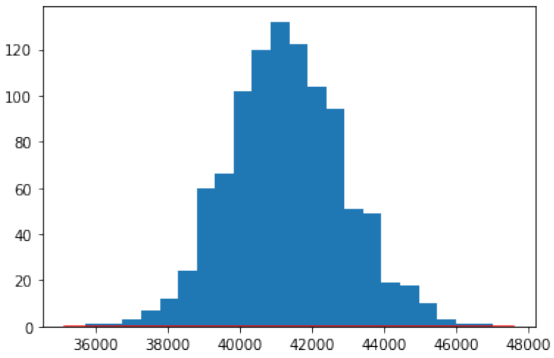
Nato smo analizirali porazdelitev dohodkov z enostavnim slučajnim vzorčenjem. Izbrali smo 1000 enostavnih slučajnih vzorcev velikosti 400. To smo naredili s pomočjo vgrajene funckije `random.randint`. Narisali smo histogram vzorčnih povprečij dohodkov.



Slika 5: Histogram vzorčnih povprečij

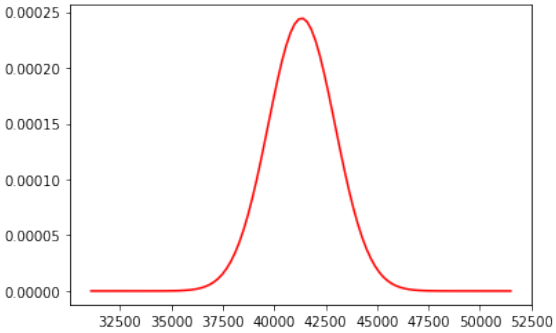
1.6 F del naloge

Na histogram smo dorisali graf normalne gostote. Za pričakovano vrednost smo vzeli povprečje dohodkov celotne populacije, za standardni odklon pa standardno napako nekega slučajnega vzorca velikosti 400. Izračunana **pričakovana vrednost** je 41335.50704096979, **standardna napaka** pa 1631.5645751757295. Vzorec smo ponovno generirali s pomočjo funkcije `random.randint()`. Standardno napako, pa smo glede na vzorec izračunali s funkcijo `sem()`.



Slika 6: Histogram vzorčnih povprečij z normalno gostoto

Graf gostote normalne porazdelitve je na sliki 6 narisan z rdečo barvo. Graf je v primerjavi z histogramom nizek, zato iz slike ni razvidno ali se graf histogramu prilega. Zato sem za primerjavo še posebej narisala graf dane normalne porazdelitve.

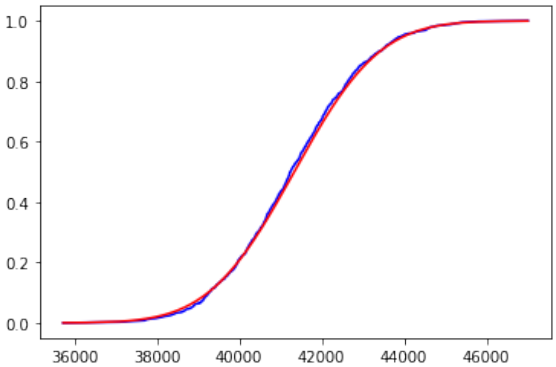


Slika 7: Gostota normalne porazdelitve

Opazimo lahko, da imata graf in histogram vrh pri približno enaki vrednosti abscise, vendar je vrh normalne gostote znatno nižji. Sklepamo lahko, da se krivulja histogramu prilega.

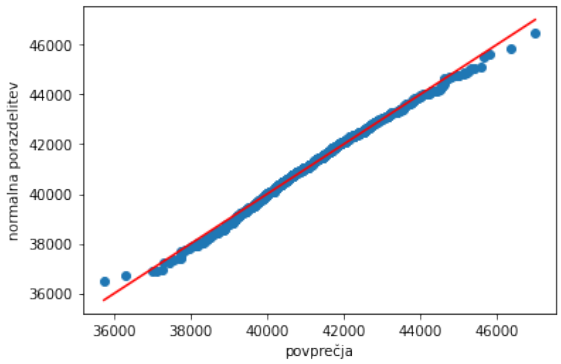
1.7 G del naloge

Za boljšo primerjavo porazdelitve povprečij slučajnih vzorcev in normalne porazdelitve iz naloge F smo primerjali še kumulativni porazdelitveni funkciji obeh porazdelitev. Iz grafa 8 opazimo, da se funkciji dobro prilegata.



Slika 8: Kumulativna porazdelitvena funkcija za vzorčna povprečja (z modro) in kumulativna porazdelitvena funkcija normalne porazdelitve (z rdečo)

Narisali smo še primerjalni kvantilni grafikon s katerim smo primerjali porazdelitev vzorčnih povprečij z normalno porazdelitvijo.



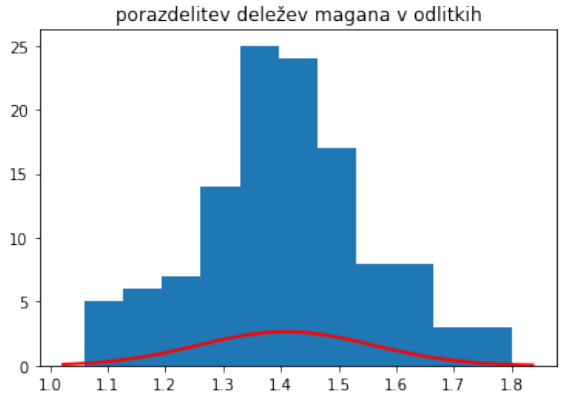
Slika 9: Kvantilni grafikon vzorčnega povprečja v primerjavi z normalno porazdelitvijo

Na grafu 9 se kvantilni grafikon, narisan z modro, dobro prilega rdeči 45-stopinjski premici. To pomeni, da ima vzorčno povprečje normalno porazdelitev.

2 Naloga 2

V drugi nalogi smo analizirali podatke o deležu mangana v železu, pridobljenem v plavžu. Podatki so vsebovali deleže mangana v odlitkih, ki so jih jemali skozi 24 dni, petkrat dnevno. Preučiti smo želeli normalnost dobljene empirične porazdelitve. Porazdelitev deležev mangana smo primerjali z normalno porazdelitvijo, katere pričakovana vrednost in standardni odklon se ujemata s povprečjem in standardnim odklonom deležev mangana iz raziskave.

Najprej smo narisali histogram vseh deležev z dorisano normalno gostoto. Širino posameznega razreda v histogramu smo kot v nalogi 1 izračunali s pomočjo Freedman-Diaconisovega pravila. **Širina razreda** je znašala 0.07298642394688079, kar je ravno 11 **razredov**.



Slika 10: Histogram deležev mangana v plavžu z dorisano normalno gostoto (z rdečo)

Graf normalne gostote se lepo prilega histogramu. Vrh histograma in grafa sta dosežena pri približno enaki vrednosti na abscisi. Histogram je tudi zvončaste oblike, kar namiguje na normalno porazdelitev.

Vendar to opažanje ni dovolj, saj normalna porazdelitev ni edina z zvončasto obliko histograma. Zato smo za primerjavo uporabili še viseči histogram razlik korenov frekvenc. Klasičen histogram ponazori frekvence posameznih vrednosti iz podatkov. Viseči histogram pa primerja korene frekvenc dane porazdelitve s koreni frekvenc ustrezne normalne porazdelitve. To naredimo tako, da najprej poračunamo verjetnosti, da ima normalno porazdeljena spremenljivka vrednost na danih inetrvalih. Ocena verjetnosti na intervalu $[x_{j-1}, x_j]$ bo

$$\hat{p_j} = \Phi\left(\frac{x_j - \overline{x}}{\hat{\sigma}}\right) - \Phi\left(\frac{x_{j-1} - \overline{x}}{\hat{\sigma}}\right), \tag{2}$$

kjer je $\overline{x}$ povprečje naših podatkov, $\hat{\sigma}$ pa cenilka za standardni odklon.

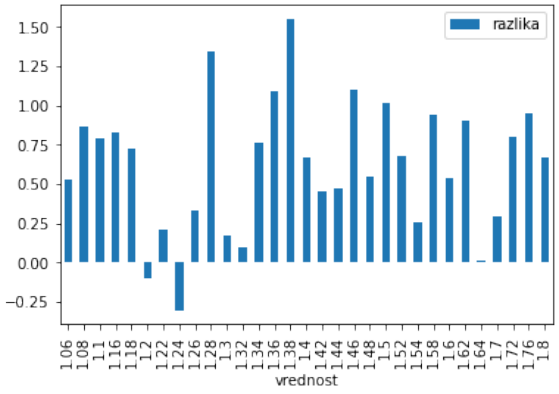
Nato določimo predvideno število pojavitev na j-tem intervalu kot

$$\hat{n_j} = n\hat{p_j}, \tag{3}$$

kjer je n število podatkov o deležih mangana. V zadnjem koraku poračunamo razliko korena frekvenc deležev mangana na j-tem intervalu, kar označimo z n_j, s korenom ocenjenih ponovitev normalne porazdelitve:

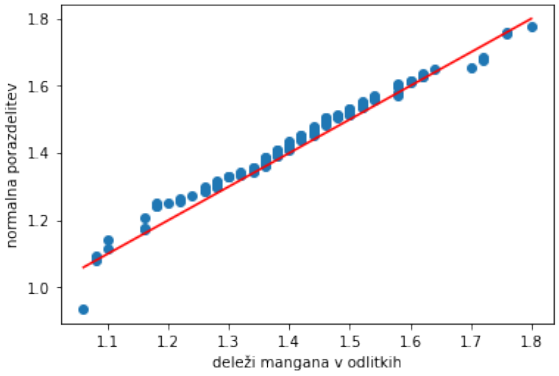
$$\sqrt{n_j} - \sqrt{\hat{n_j}}. \tag{4}$$

Poračunane vrednosti narišemo v histogram. Iz grafa 11 vidimo, da imamo vzdolž vseh vrednosti velika odstopanja v frekvencah.



Slika 11: Viseči histogram razlik korenov frekvenc

Za nadaljno primerjavo uporabimo še kvantilni (Q-Q) grafikon.



Slika 12: Primerjalni kvantilni grafikon

Iz grafikona 12 opazimo, da se kvantili prilegajo 45-stopinjski premici, na robovih pa pride do rahlih odstopanj. Kvantilni grafikon kaže, da so deleži mangana porazdeljeni približno normalno.

3 Naloga 3

V tretji nalogi smo analizirali podatke o dolžini zob morskih prašičkov, ki so jim dodajali vitamin C v različnih količinah na dva različna načina: bodisi neposredno bodisi s pomarančnim sokom.

3.1 A del naloge

Najprej nas je zanimalo ali dodajanje vitamina C sploh vpliva na rast zob. Za začetek smo izračunali povprečno dolžino zoba glede na količino dodatka vitamina C. Za količino 0.5 povprečna dolžina zoba znaša 10.605, za količino 1.0 19.735, za količino 2.0 pa 26.100. To nam da misliti, da

dolžina zoba raste s količino dodatka vitamina C. S preizkusom dokažimo, da pričakovana vrednost dolžine zoba raste z dodajanjem vitamina C. Definirajmo ničelno in alternativno hipotezo:

$$H0: \ \mu_{0.5} = \mu_{1.0} \tag{5}$$
$$H1: \ \mu_{0.5} < \mu_{1.0}, \tag{6}$$

kjer je $\mu_{0.5}$ pričakovana vrednost dolžine zoba pri dodajanju 0.5 vitamina C, $\mu_{1.0}$ pa pri dodajanju količine 1.0. Hipotezo preizkusimo s Studentovim t-testom. Izberimo stopnjo tveganja $\alpha = 0.05$. Naj slučajna spremenljivka X označuje porazdelitev velikosti zob pri dodajanju količine 0.5 vitamina C, spremenljivka Y pa pri dodajanju količine 1.0.

Na vajah smo pokazali, da za spremenljivko T velja

$$T \coloneqq \frac{\overline{Y} - \overline{X}}{S}\sqrt{\frac{mn}{m+n}} \sim Student(n+m-2),$$

kjer je

$$S \coloneqq \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2}{m+n-2}}.$$

Število n predstavlja število podatkov pri dodajanju količine 0.5, m pa pri 1.0. V našem primeru je $n = m = 20$. Iz formule

$$P(T \geq \mathrm{C}_\alpha) \leq \alpha$$

moramo nato poračunati C_α. To storimo tako, da iz razpredelnice odčitamo inverz Studentove porazdelitve, pri stopnji tveganja $\alpha = 0.00$ in pri stopnji prostosti $n + m - 2 = 38$. Dobljena vrednost je $C_{0.05} = 1.6860$.

Iz podatkov iz našega vzorca moramo sedaj poračunati vrednost spremenljivke T. To lahko naredimo neposredno iz formul ali pa uporabimo že vgrajeno funkcijo `stats.ttest_ind`, ki nam vrne vrednost T kot `statistic`. Dobljena vrednost je $T = 6.476647726589102$, kar je večje kot $C_{0.05}$. Torej **ničelno hipotezo zavržemo**.

Podobno storimo še s primerjavo pričakovane vrednosti dolžin zob z dodajanjem količine 1.0 in 2.0. Z enakim postopkom dobimo $T = 4.90048431719355$, kar je še vedno več kot $C_{0.05}$. Tudi tu **ničelno hipotezo zavržemo**. Od tod sledi, da **dodajanje vitamina C res pozitivno vpliva na rast zob**.

3.2 B del naloge

Dalje nas je zanimalo, kateri način dodajanja vitamina C je učinkovitejši. Ponovno naredimo preizkus hipoteze s Studentovim t-testom. Naj P označuje slučajno spremenljivko za dolžino zob pri dodajanju s pomarančnim sokom, N pa pri neposrednem dodajanju. Zapišimo hipotezi:

$$H0 : \mu_P = \mu_N \tag{7}$$
$$H1 : \mu_P > \mu_N \tag{8}$$

Ponovno uporabimo vgrajeno funkcijo in dobimo $T = 1.91526826869527$. Iz tabele ponovno odčitamo vrednost $C_{0.05}$, tokrat pri stopnji prostosti $n+m-2 = 30+30-2 = 58$. Dobimo $C_{0.05} = 1.6716$. Torej **ničelno hipotezo ponovno zavržemo**. Sledi, da je **dodajanje s pomarančnim sokom učinkovitejše**. Vendar bi za stopnjo tveganja $\alpha = 0.01$ ničelno hipotezo obdržali, saj bi iz tabele odčitali vrednost $C_{0.01} = 2.3924$, ki je v tem primeru večja od T.

Poračunajmo še ali je razlika statistično značilna. Za izračunan T iz danih podatkov moramo poračunati $P(T \geq \mathrm{C}_\alpha)$. Izberimo stopnjo tveganosti $\alpha = 0.05$. Verjetnost lahko poračunamo z vgrajeno funkcijo `stats.ttest_ind`. Dobimo vrednost $p = 0.03019668561206424$. To vrednost sedaj primerjamo z α. Ker je $p < \alpha$ je razlika **statistično značilna**.