Seminarska naloga statistika

Vito Založnik

August 17, 2021

1 Prva naloga

1.1

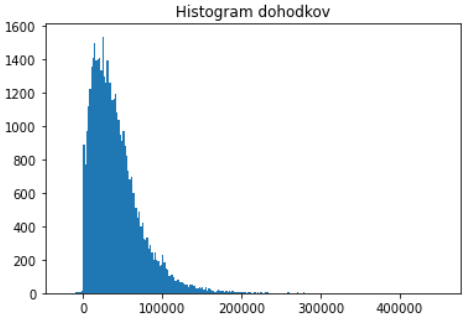
a) Narišite histogram dohodkov vseh družin v Kibergradu. Pri tem dohodke razdelite v enako široke razrede. Širino posameznega razreda določite v skladu s Freedman–Diaconisovim pravilom, Kjer sta $q1/4$ in $q3/4$ prvi in tretji kvartil, n pa je število enot. To vrednost nato smiselno zaokrožite na število oblike $k \cdot 10^r$, kjer je $k \in 1, 2, 5$ in $r \in \mathbf{Z}$.

$q1 = 18300.0$, $q3 = 55827.75$, $širina = 2127.847614822404$, zaokrožena širina $= 2000$.
Preverimo, če je kak podatek osamelec. Osamelci so elementu zunaj intervala:

$$[q1 - \frac{3}{2}IQR, q3 + \frac{3}{2}IQR]$$

$$IQR = q3 - q1$$

Spodnja meja osamelcev $= -37991.625$, Zgornja meja osamelcev $= 112119.375$. Med podatki ni nobenih osamelcev.

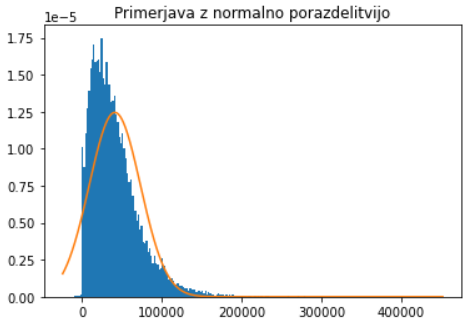


1.2

b) Dorišite normalno gostoto, katere pričakovana vrednost in standardni odklon se ujemata s povprečjem in standardnim odklonom dohodka družine v Kibergradu. Kako dobro se prilega?

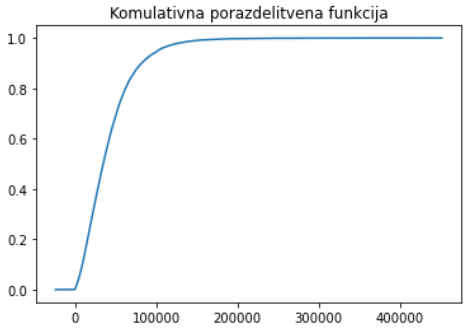
Cenilko za povprečje izračunamo po formuli: $\hat{\mu} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, kjer je X_i dohodek posmezne družine. Nepristransko cenilko za standardni odklon izračunamo po formuli: $\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{X} - X_i)^2}$. Dobimo $\bar{X} = 41335$, $\hat{\sigma} = 32037$.Poglejmo primerjavo normaliziranega histograma dohodkov z porazdelitvjio $N(\hat{\mu}, \hat{\sigma}^2)$.

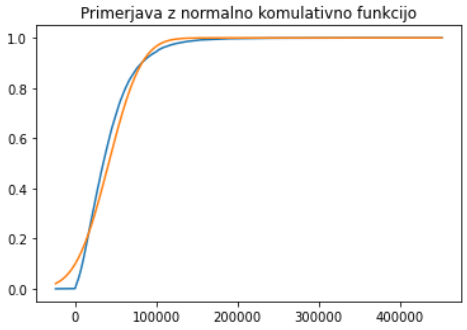
Vidimo, da se porazdelitev ne prilega dobro. Predvidevam, da ima velik vpliv na to to, da je ljudi z negativnimi prihodki veliko manj. Jasno se tudi vidi, da praktično ni gospodinjstev brez prihodkov.



1.3

c) Narišite kumulativno porazdelitveno funkcijo porazdelitve dohodkov družin v Kibergradu in primerjajte s kumulativno porazdelitveno funkcijo ustrezne normalne porazdelitve.





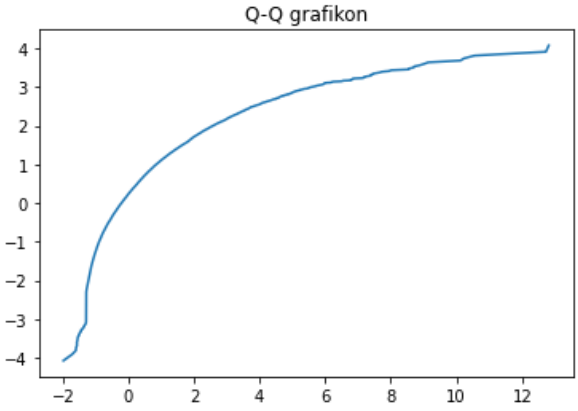
Vidimo, da se komulativna porazdelitvena funkcija dohodkov dokaj dobro ujema z komulativno funkcijo normalne porazdelitve. Razlok za to vidim v tem, da pri komulativni funkciji pretekli podatki vplivajo na sedanje. Torej se manjko na enem delu in presežek na tretjem lahko skompenzira. Sploh pa se vsaka komulativna funkcija zaključi pri 1 in ker je glavnina podatkov že mimo se v poznejšem delu ne prišteva več dosti podatkov in zato komulativna funkciji izgledata popolnoma enako v drugi polovici.

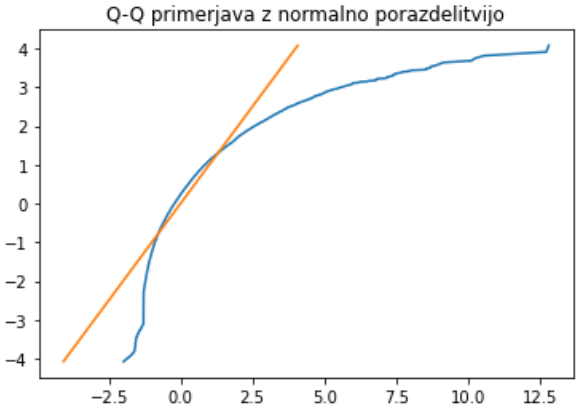
1.4

d) Narišite še primerjalni kvantilni (Q–Q) grafikon, ki porazdelitev dohodkov družin v Kibergradu primerja z normalno porazdelitvijo.

Primerjalni kvantilni grafikon je grafična metoda za določanje, če imata dva vzorca isto porazdelitev. V mojem primeru bom porazdelitev prihodkov primerjal z normalno porazdelitvijo. Podatke najprej uredimo po velikosti naraščajoče. Nato si izračunamo teoretične vrednosti normalne porazdelitve

razdeljene na n+1 kvantilov. Nato narišemo na x-os teoretične izračunane vrednosti, na y os pa naše urejene podatke. Če se naši podatki ujemajo s simetralo lihih kvadrantov se porazdelitev podatkov ujema s to teoretično porazdelitvijo. Če so podatki v obliki neke druge premice se porazdelitev podatkov ujema z neko drugo normalno porazdelitvijo.





Vidimo, da se graf ne ujema s simetralo lihih kvdrantov, ki predstavlja našo normalno porazdelitev. Prav tako Podatki niso v obliki neke druge premice. To pomeni, da je porazdelitev naših podatkov daleč od kakšne normalne porazdelitve.

1.5

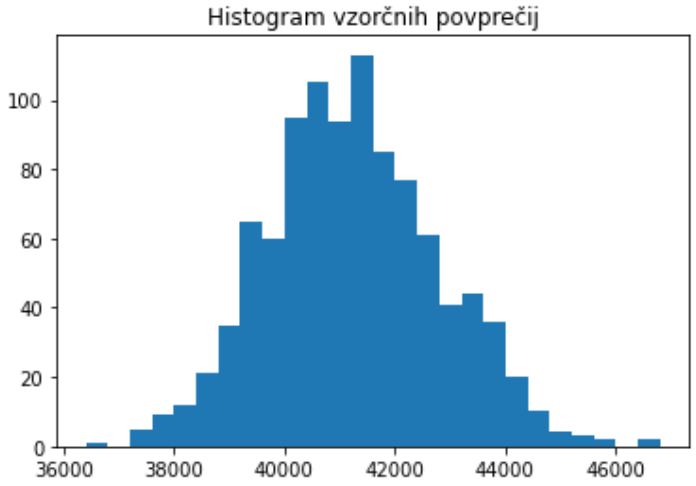
e)Vzemite 1000 enostavnih slučajnih vzorcev velikosti 400 in narišite histogram vzorčnih povprečij dohodkov družin.

Ponovno si izračunajmo širino intervala po Freedman–Diaconisovem pravilu. Za naša vzorčna povprečja dobimo: $q1 = 40210.5$, $q3 = 42351.674375$, širina posameznega razreda= 413.4125, Zaokrožena širina = 400.

1.6

f) Dorišite normalno gostoto, katere pričakovana vrednost se ujema s povprečnim dohodkom na družino v Kibergradu, standardni odklon pa s standardno napako za enostavni slučajni vzorec velikosti 400. Komentirajte, kako dobro se prilega.

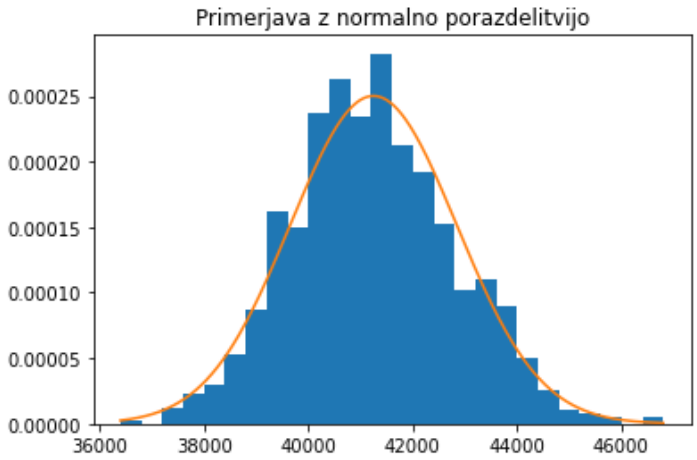
Nepristranska cenilka za povprečni dohodek je enaka povprečju vzorčnih povprečij. Standardno napako za enostavni slučajni vzorec izračunamo tako, da iz populacije izberemo enostavni slučajni



vzorec velikost $n = 400$ in po formuli:

$$\hat{SE}^2 = \frac{N-n}{N \cdot n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Kjer je N velikost populacije. Dobimo: $\hat{\mu}$ = Povprecje vzorcnih pvprecij = 41247.3296675, $\hat{SE} = 1595$. Standardna napaka je po definiciji ravno standardni odklon vzorčnih povprečij.



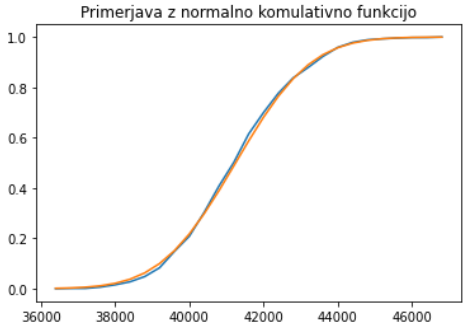
Vidimo, da se vzorčna povprečja dobro prilegajo normalni porazdelitvi kar je pričakovano, saj je standardna napaka za enostavni slučajni vzorec ravno standardni odklon vzorčnih povprečij. Ker smo za SE uporabili cenilko $\hat{SE}$ bi naša porazzdelitev morala iti proti Studentovi vendar ker je število vzočnih povprečij veliko gre Studentova porazdelitev proti normalni.

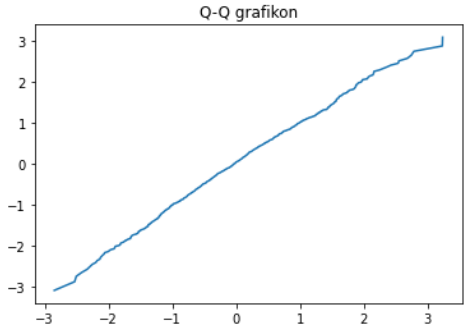
1.7

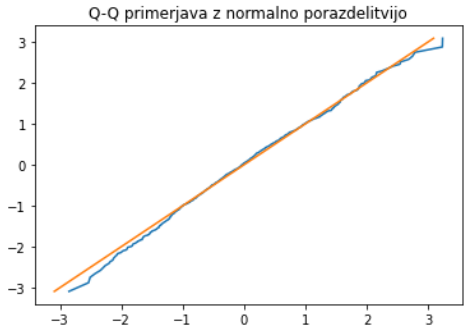
g) Za vzorčna povprečja podobno kot prej narišite še kumulativno porazdelitveno funkcijo in primerjalni kvantilni grafikon ter primerjajte z normalno porazdelitvijo. Komentirajte prileganje

Komulativna funkcija se skoraj popolnoma prilega komulativni funkciji normalne porazdelitve.

Vidimo, da se tudi q-q grafikon dobro prilega. $\hat{SE}$ bi lahko tudi odstopal od dejanske standardne napake saj je izračunan glede na izbran slučajni vzorec ki pa je lahko različen. V tem primeru bi na q-q grafikonu videli naše podatke v obliki neke druge premice z drugačnim smernim koeficientom.







2 Druga naloga

2.1 Ocenite povprečje in standardni odklon za telesno temperaturo posebej pri moških in posebej pri ženskah. Po navodilih privzamemo, da je temperatura pri moških in ženskah razdeljena normalno.

- Povprečje temperature moških je: 98.1046153846154
- Povprečje temperature žensk je: 98.39384615384616
- Standardni odklon temperature moških je: 0.6987557623265904

- Standardni odklon temperature žensk je: 0.7434877527313662

2.2 Za povprečji iz prejšnje točke določite 0.95 intervala zaupanja.

Interval izračunamo po formuli:

$$\left[\bar{X} - \frac{\hat{\sigma}}{\sqrt{n}} F^{-1}_{Student(n-1)(1-\frac{\alpha}{2})}, \bar{X} + \frac{\hat{\sigma}}{\sqrt{n}} F^{-1}_{Student(n-1)(1-\frac{\alpha}{2})}\right]$$

- Interval zaupanja za povprečje temperature moških je: [97.93147218445705, 98.27775858477375]
- Interval zaupanja za povprečje temperature žensk je: [98.20961890918225, 98.57807339851006]

2.3 Preizkusite domnevo, da imajo moški in ženske v povprečju enako telesno temperaturo.

Računali bomo pri stopnji tveganja $\alpha = 0.05$ in $\alpha = 0.01$.

- H_0: Moški in ženske imajo v povprečju enako telesno temperturo.
- H_1: Moški in ženske v povprečju nimajo iste temperature.
- H_0: $\mu_m = \mu_z$

  Vemo:

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

Razlika normalnih porazdelitev je ponovno normalna porazdelitev.

$$\frac{\bar{X}_m - \bar{X}_z - (\mu_m - \mu_z)}{\sqrt{\frac{\sigma_m^2}{N_m} + \frac{\sigma_z^2}{N_z}}} \to N(0,1)$$

V našem primeru zamenjamo σ_i s $\hat{\sigma}_i = \sqrt{\frac{1}{N_i - 1}\sum_j (X_j - \bar{X}_i)^2}$. Testna statistika:

$$T = \frac{\bar{X}_m - \bar{X}_z - (\mu_m - \mu_z)}{\sqrt{\frac{\hat{\sigma}_m^2}{N_m} + \frac{\hat{\sigma}_z^2}{N_z}}} \sim F_{student}(n-2)$$

Dobimo $df = 127$. Če H_0 drži, je $\mu_m - \mu_z = 0$. Dobimo:

$$T = \frac{\bar{X}_m - \bar{X}_z}{\sqrt{\frac{\hat{\sigma}_m^2}{N_m} + \frac{\hat{\sigma}_z^2}{N_z}}} \sim F_{student}(127)$$

Studentova porazdelitev gre proti $N(0,1)$ ko gre $n \to \infty$. V tabeli imamo studentovo porazdelitev samo do 120 prostorsek stopnje tako, da bomo za studentovo porazdelitev z 127 prostorskimi stopnjami vzeli kar normalno porzdelitev $N(0,1)$.

$$T = \frac{\bar{X}_m - \bar{X}_z}{\sqrt{\frac{\hat{\sigma}_m^2}{N_m} + \frac{\hat{\sigma}_z^2}{N_z}}} \sim N(0,1)$$

Imamo dvostranski test. Hipotezo, da sta povprečja enaka bomo zavrnili če bo $|T| > \Phi^{-1}(1 - \alpha/2)$.

Pri $\alpha = 0.05 \Rightarrow \Phi^{-1}(0.975) = 1.96$
$|T| = 2.285 > 1.96 \Rightarrow$ Hipotezo zavrnemo.
Pri $\alpha = 0.01 \Rightarrow \Phi^{-1}(0.995) = 2.85$
$|T| = 2.2852.85 \Rightarrow$ Ni dovolj dokazov, da bi lahko hipotezo zavrnili.

Kje vmes pa se nahajamo? Izračunajmo p vrednost.

$$\begin{aligned} P(T \in (-\infty, -2.285) \bigcup (2.285, \infty), \text{ če } H_0 \text{ res.}) &= 2 \cdot \Phi(-2.285) \\ &= 2(1 - \Phi(2.285)) \\ &= 2 - 2 \cdot 0.98885 \ = 0.0223 = p \end{aligned}$$

$0.01 < 0.0223 < 0.05$. Ponovno vidimo, da pri $\alpha = 0.05$ domnevo zavrnemo, pri $\alpha = 0.01$ pa je ne moremo.

3 Tretja naloga

3.1

Vsakemu študentu so pulz izmerili dvakrat. Določeni so imeli med obema meritvama fizično obremenitev (tek na mestu), določeni ne. Podatki so zbrani v tabeli Pulz. Raziščite, katere od dejavnikov višina, teža, spol in vadba je smiselno vključiti v linearni model, ki bo opisoval odvisnost prve meritve pulza od teh dejavnikov: poiščite model z najmanjšo Akaikejevo informacijo. Slednja nam pomaga izbrati le bistvene dejavnike. Akaikejeva informacija je sicer definirana z verjetjem, a pri linearni regresiji in Gaussovem modelu je le-ta ekvivalentna naslednji modifikaciji:

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

kjer je m število parametrov, n pa je število opažanj.

Večkrat naredimo linearno regresijo, vsakič na drugi kombinaciji podatkov. Vsakič poračunamo RSS in iz tega s pomočjo zgornje formule za AIC poiščemo model, ki ima najnižji AIC.

$$RSS = \sum_i ||\widehat{Y}_i - Y_i||^2 = n \cdot MSE = \sum_i ||Y - X\widehat{\beta}||^2 = \sum_i \widehat{\epsilon}_i^2$$

Dobili smo, da je najmanjši $AIC = 1066.5074099762016$, $RSS = 17116.092870207667$. To dobimo pri podatkih $VISINA$, $VADBA$.

3.2

Za vsakega od dejavnikov kajenje in alkohol vaš model preizkusite proti širšemu modelu, ki poleg dejavnikov, izbranih v prejšnji točki, vključi tudi še dodatni dejavnik.

Imamo V, W vektorska podprostora $p = dim(V)$, $q = dim(W)$, $W \subset V \subset \mathbf{R}^n$.
V modelu $Y = v + \epsilon$ kjer je $v \in V$, $\epsilon \sim N(0, \sigma^2 I_n)$ Preizkušamo domnevo $H_0 : v \in W$ proti alternativni domnevi $H_1 : v \notin W$. Pomagali si bomo s Ficherjevo porazdelitvijo.
Fisherjeva porazdelitev s (v_1, v_2) protorskimi stopnjami je porazdelitev slučajne premenljivke $\frac{\frac{H_1}{v_1}}{\frac{H_2}{v_2}}$ kjer sta $H_1 \sim \chi^2(v_1)$ in $H_2 \sim \chi^2(v_2)$ neodvisni slučajni spremenljivki.

Za testno statistiko uporabimo:

$$F = \frac{\frac{RSS_w - RSS_v}{p-q}}{\frac{RSS_v}{n-p}}$$

Ker je F strogo naraščajoča funkcija razmerja verjetij je H_0 pri stopnji tveganja smiselno zavrniti če je:

$$F \geq F^{-1}_{Fisher(p-q,n-p)}(1 - \alpha)$$

Naredimo regresijo na podatkih VISINA,VADBA, ALKOHOL in VISINA, VADBA, KAJENJE.

3.2.1

Na podatkih VISINA, VADBA, ALKOHOL dobimo $RSS = 16879.378$. Število parametrov je 3. Število podatkov $n = 106$ dobimo F:

$$F = \frac{\frac{17116.093-16879.378}{3-2}}{\frac{16879.378}{106-3}} = 1.444$$

$$F^{-1}_{Fisher(1,103)}(0.95) = 3.936 > 1.444 = F$$

Torej pri $\alpha = 0.05$ ni dovolj dokazov da bi lahko trditev zavrnili.
Če izjave nismo mogli zavrniti pri stopnji tveganja $\alpha = 0.05$ je zagotovo ne bomo mogli tudi pri $\alpha = 0.01$. Preverimo še računsko. Pri $alpha = 0.01$ dobimo:

$$F^{-1}_{Fisher(1,103)}(0.99) = 6.895 > 1.444 = F$$

.

3.2.2

Podobno naredimo na podatkih VISINA, VADBA, KAJENJE.

$$F = \frac{\frac{17116.093-17091.383}{3-2}}{\frac{17091.383}{106-3}} = 0.149$$

$$F^{-1}_{Fisher(1,103)}(0.95) = 3.936 > 0.149 = F$$

$$F^{-1}_{Fisher(1,103)}(0.99) = 6.895 > 0.149 = F$$

Ponovno ni dovolj dokazov, da bi lahko trditev zavrnili.