Seminarska naloga statistika

Vito Založnik

August 19, 2021

1 Prva naloga

1.1

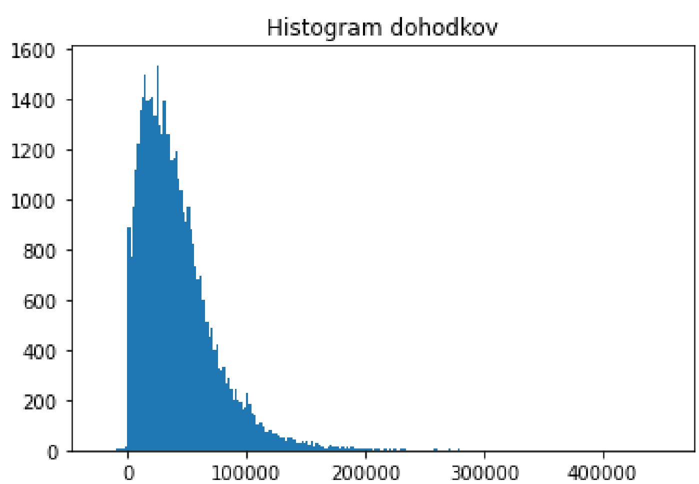
a) Narišite histogram dohodkov vseh družin v Kibergradu. Pri tem dohodke razdelite v enako široke razrede. Širino posameznega razreda določite v skladu s Freedman–Diaconisovim pravilom, Kjer sta $q1/4$ in $q3/4$ prvi in tretji kvartil, n pa je število enot. To vrednost nato smiselno zaokrožite na število 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:

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

$$IQR = q3 - q1$$

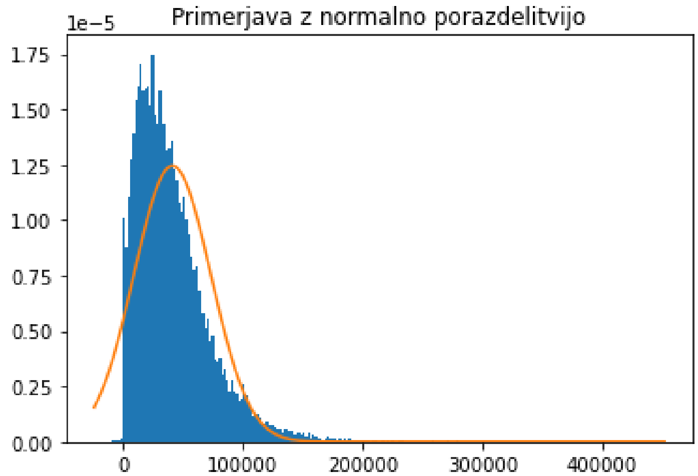
Spodnja meja osamelcev $= -37991.625$, Zgornja meja osamelcev $= 112119.375$. Med podatki ni nobenih osamelcev.



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?

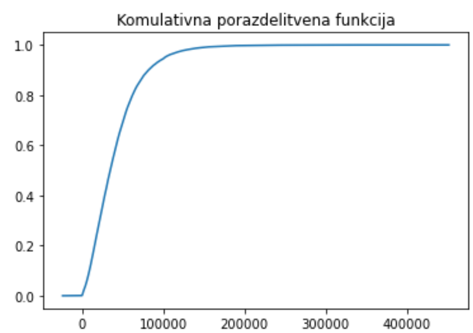
Cenilko za povprečje izračunamo po formuli: $\hat{\mu} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, kjer je X_i dohodek posmezne družine. Nepristransko cenilko za standardni odklon izračunamo po formuli: $\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{X} - X_i)^2}$. Dobimo $\bar{X} = 41335$, $\hat{\sigma}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

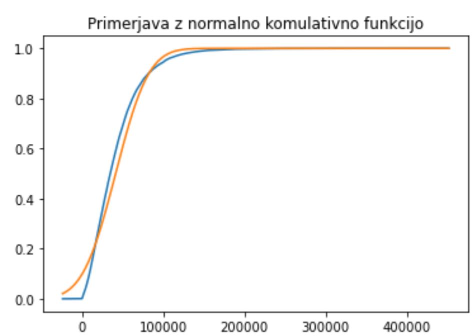


Vidimo, da se porazdelitev ne prilega dobro. Predvidevam, da ima velik vpliv na to to, da je ljudi z negativnimi prihodki veliko manj. Jasno se tudi vidi, da praktično ni gospodinjstev brez prihodkov.

1.3

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



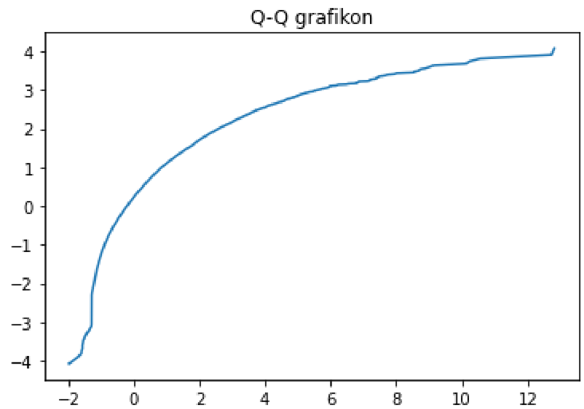


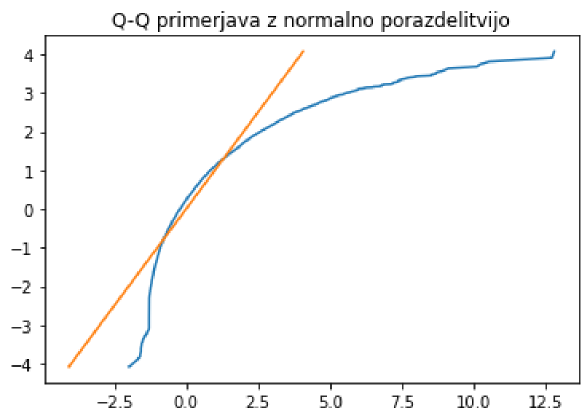
Vidimo, da se komulativna porazdelitvena funkcija dohodkov dokaj dobro ujema z komulativno funkcijo normalne porazdelitve. Razlog za to vidim v tem, da pri komulativni funkciji pretekli podatki vplivajo na sedanje. Torej se manjko na enem delu in presežek na tretjem lahko skompenzira. Sploh pa se vsaka 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.



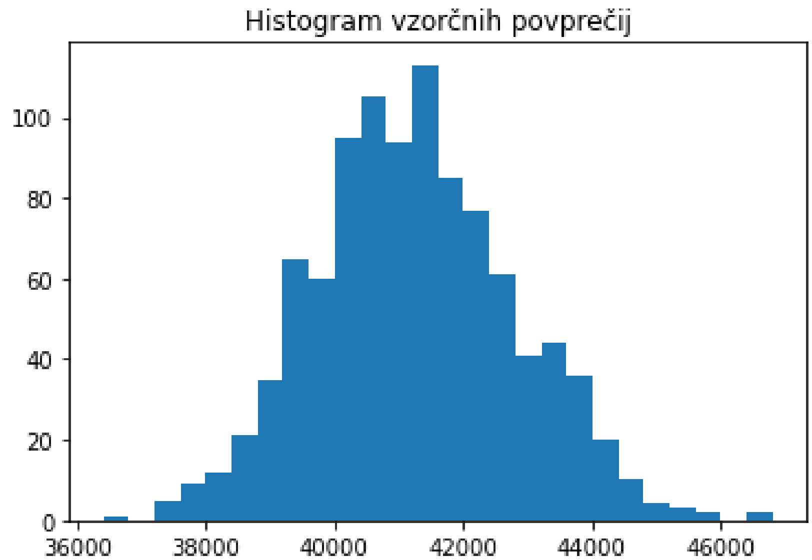


Vidimo, da se graf ne ujema s simetralo lihih kvdrantov, ki predstavlja našo normalno porazdelitev. Prav tako Podatki niso v obliki neke druge premice. To pomeni, da je porazdelitev naših podatkov daleč od kakšne normalne porazdelitve.

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:

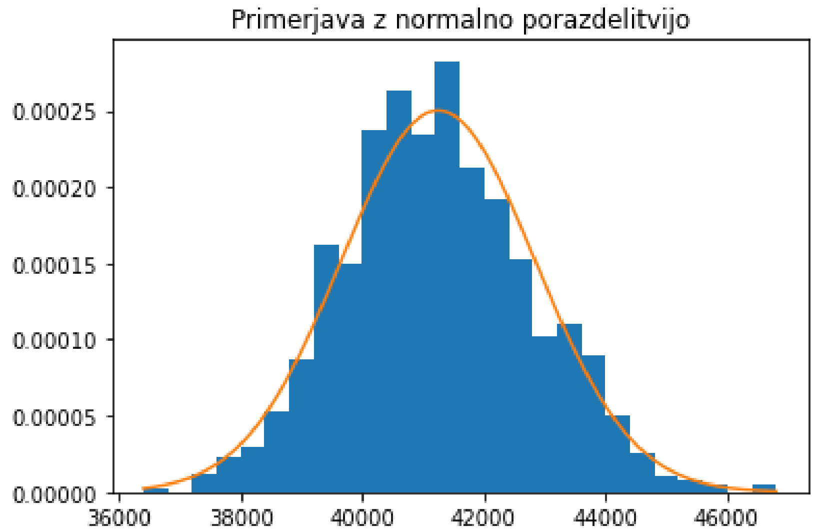
$$SE^2 = \frac{N-n}{(N-1)} \cdot \frac{\sigma^2}{n}$$

σ ne poznamo, lahko ga pa ocenimo. Izkaže se, da je:

$$\widehat{\sigma}^2 = \frac{N-1}{N} \frac{1}{n-1} \sum_{i=1}^{n} (\bar{X} - X_i)^2$$

$$\widehat{SE}^2 = \frac{N-n}{N} \frac{\sum_{i=1}^{n} (\bar{X} - X_i)^2}{n(n-1)}$$

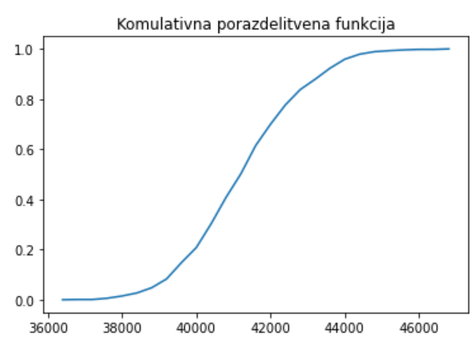
Kjer je N velikost populacije. Dobimo: $\hat{\mu}$ = Povprečje vzorčnih povprečij = 41247.3296675, $\hat{SE}$ = 1595. Standardna napaka je po definiciji ravno standardni odklon vzorčnih povprečij.

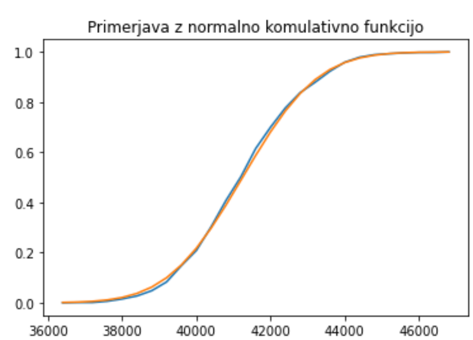


Vidimo, da se vzorčna povprečja dobro prilegajo normalni porazdelitvi kar je pričakovano, saj je standardna napaka za enostavni slučajni vzorec ravno standardni odklon vzorčnih povprečij. Ker smo za SE uporabili cenilko $\hat{SE}$ bi naša porazzdelitev morala iti proti Studentovi vendar ker je število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

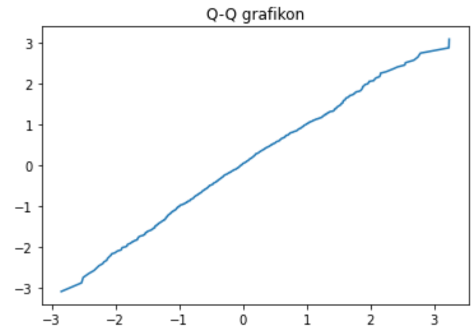
1.7

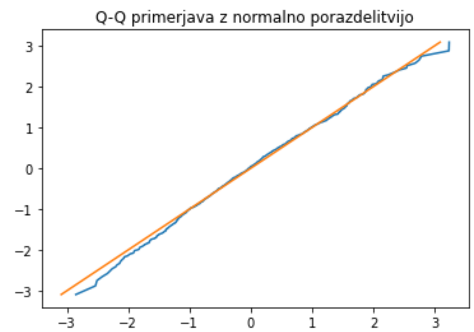
g) Za vzorčna povprečja podobno kot prej narišite še kumulativno porazdelitveno funkcijo in primerjalni kvantilni grafikon ter primerjajte z normalno porazdelitvijo. Komentirajte prileganje





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: $\mu_m = \mu_z$

Imamo stratificirano vzorčenje. Imamo stratume $i = 1, \ldots, k$ Iz vsakega stratuma dobimo vzorec X_{ij}, $j = 1, \ldots, n_i$. Označimo teoretično povprečje (pričakovana vrednost) i-tega stratuma z μ_i in varianco s σ_i^2. Dodatno bomo predpostavili, da vsak X_{ij} prihaja iz normalne porazdelitve (torej iz $N(\mu_i, \sigma_i^2)$). Predpostavljamo tudi, da so X_{ij} med sabo neodvisni za vse i, j. Ozančimo še: $n = n_1 + \cdots + n_k$ $w_i =$ delež stratuma v populaciji, $\mu := \sum_i w_i \mu_i$, $\bar{X} := \sum_i w_i \bar{X}_i = \frac{1}{n} \sum_{ij} X_{ij}$, kjer je $\bar{X}_i = \frac{1}{n_i} \sum_j X_{ij}$. $\bar{X}$ je nepristranska cenilka za μ. $\widehat{SE}^2 := \sum_i \frac{w_i^2 \hat{\sigma}_i^2}{n_i}$ je cenilka za standardno napako od $\bar{X}$. (Tu je $\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_j (X_{ij} - \bar{X}_i)^2$.) Izrek iz predavanj:

$$\frac{\bar{X} - \mu}{\widehat{SE}} \approx Student(v)$$

kjer število v lahko ocenimo z

$$\hat{v} = \frac{\widehat{SE}^4}{\sum_i \frac{w_i^4 \hat{\sigma}_i^4}{n_i^2 (n_i - 1)}}$$

Vzemimo zdaj podatke in ugotovimo, ali so dejanska povprečja stratumov med sabo vsa ista (in enaka znani konstanti μ_0) ali pa so opažene razlike morda samo posledica naklučja. Radi bi preizkusili $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k =: \mu_0$ Če je H_0 resnična, je $\mu = \sum_i w_i \mu_i = \mu_0$. Potem je po izreku

$$T := \frac{\bar{X} - \mu_0}{\widehat{SE}} \approx Student(v)$$

Če H_0 ne bi veljala, potem bi v števcu odštevali neko drugo število in zato porazdelitev T-ja ne bi bila skoncentrirana okoli ničle. Ideja je torej: če H_0 drži, bodo vrednosti T-ja blizu ničle, če ne, pa stran od ničle.

Delež moških in žensk je v populaciji približno enak: $w_m = w_z = \frac{1}{2}$.

$$\bar{X} = \frac{\bar{X_m}}{2} + \frac{\bar{X_z}}{2}$$

$$\bar{\mu} = \frac{\mu_m}{2} + \frac{\mu_z}{2}$$

Če H_0:

$$\frac{\bar{X} - \mu}{\widehat{SE}} = \frac{\frac{\bar{X_m}}{2} + \frac{\bar{X_z}}{2} - \frac{\mu_m}{2} - \frac{\mu_z}{2}}{\frac{1}{2}\sqrt{\frac{\hat{\sigma}_m^2}{N_m} + \frac{\hat{\sigma}_z^2}{N_z}}} =\approx Student(v) \approx \frac{\bar{x}_m - \bar{x}_z - (\mu_m - \mu_z)}{\sqrt{\frac{\hat{\sigma}_m^2}{N_m} + \frac{\hat{\sigma}_z^2}{N_z}}} = \frac{\bar{x}_m - \bar{x}_z}{\sqrt{\frac{\hat{\sigma}_m^2}{N_m} + \frac{\hat{\sigma}_z^2}{N_z}}}$$

Testna statistika:

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

Dobimo $v = 63.7$, $df = 64$. Če H_0 drži, je $\mu_m - \mu_z = 0$. Dobimo:

$$T = \frac{\bar{X}_m - \bar{X}_z}{\sqrt{\frac{\hat{\sigma}_m^2}{N_m} + \frac{\hat{\sigma}_z^2}{N_z}}} \sim F_{student}(64)$$

Studentova porazdelitev gre proti $N(0,1)$ ko gre $n \to \infty$. V tabeli nimamo kvantilov Studentove porazdelitve z 64 prostorskimi topnjami imamo pa s 60. Ker je vzorec velik bi lahko vzeli kar normalno porazdelitev.

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

Imamo dvostranski test. Hipotezo H_0, da sta povprečja enaka bomo zavrnili če bo

$$|T| > F^{-1}_{Student(60)}(1-\alpha/2)$$

.

$$\begin{aligned}
\text{Pri } \alpha = 0.05 &\Rightarrow F^{-1}_{Student(60)}(0.975) = 2 \\
|T| &= 2.285 > 2 \Rightarrow \text{Hipotezo zavrnemo.} \\
\text{Pri } \alpha = 0.01 &\Rightarrow F^{-1}_{Student(60)}(0.995) = 2.66 \\
|T| &= 2.285 < 2.66 \Rightarrow \text{Ni dovolj dokazov, da bi lahko hipotezo zavrnili.}
\end{aligned}$$

Kje vmes pa se nahajamo? Predpostvimo da je vzorec dovolj velik, da bo porazdeljen približno normalno. 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 Fischerjevo porazdelitvijo.
Fisherjeva porazdelitev s (v_1, v_2) prostorskimi 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.

1. naloga

```
# -*- coding: utf-8 -*-
"""
Created on Sun Aug  8 23:09:49 2021

@author: Vito ZaloĹžnik
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
from scipy.special import ndtri

pot = "kibergrad.csv"
data = pd.read_csv(pot)
lastnosti = data.columns
#print(data.head())

dohodki = data.iloc[:,[3]] #stolpec dohodkov

q1 = np.percentile(dohodki, 25) #Q1
q3 = np.percentile(dohodki, 75) #Q3

povprecje = int(np.mean(dohodki))
std = int(np.std(dohodki, ddof=1))

m = len(dohodki)
print("q1 je: ", q1)

print("q3 je: ", q3)

sirina = 2*(q3-q1)/np.cbrt(m)


iqr = q3 - q1
spodnja_meja = q1 - 3/2*iqr
zgornja_meja = q3 + 3/2*iqr
print("Spodnja meja osamelcev je: ", spodnja_meja)
print("Zgornja meja osamelcev je: ", zgornja_meja)
#ni osamelcev

print("sirina je: ", sirina) #zaokrozimo na 400
sirina = 2000
print("popravljena sirina je: ", sirina)
zacetek =  (int(dohodki.min())//sirina)*sirina
konec = (int(dohodki.max())//sirina+1)*sirina

print("std je: ", std)
print("povprecje je: ", povprecje)


print("zacetek je: ", zacetek)
print("konec je: ", konec)

dohodki = dohodki.values.tolist()

flat_list = []
for sublist in dohodki:
   for item in sublist:
       flat_list.append(item)
dohodki = flat_list


"""histogram dohodkov"""
plt.figure()
plt.hist(dohodki, bins=int(((konec - zacetek)//sirina )), range=(zacetek, konec))
```

```python
plt.title("Histogram dohodkov")
plt.show()

"""primerjava histograma z normalno porazdelitvijo"""
x = np.linspace(zacetek, konec, m)
plt.figure()
plt.hist(dohodki, bins=int(((konec - zacetek)//sirina )), range=(zacetek, konec), density=True)
plt.plot(x, stats.norm.pdf(x, povprecje, std))
plt.title("Primerjava z normalno porazdelitvijo")
plt.show()



"""komulativni histogram"""
st = int(((konec - zacetek)//sirina ))
res = stats.cumfreq(dohodki, numbins=st, defaultreallimits=(zacetek, konec)) #sesteje komulativno po intervalih
res = res[0]
res = np.insert(res, 0, 0., axis=0)
res = res/res[-1] #normiramo

x = np.linspace(zacetek,konec, st +1)
y_cdf = stats.norm.cdf(x, povprecje, std)

plt.figure()
plt.plot(x, res)
plt.title("Komulativna porazdelitvena funkcija")
plt.show()


"""primerjava z normalno komulativno"""
plt.figure()
plt.plot(x, res)
plt.plot(x, stats.norm.cdf(x, povprecje, std))
plt.title("Primerjava z normalno komulativno funkcijo")
plt.show()



"""Q-Q"""

urejeno = np.sort(dohodki) #uredimo po vrsti

#normalno porazdelitev razdelimo na n+1 delov.
delcki = np.arange(1,m+1)/(m+1) #range ne vkluci zadnjega

#izracunamo teoreticne vrednosti porazdelitve
teoreticne_vrednosti = ndtri(delcki)


#normaliziramo nase vrednosti
norm_podatki = (urejeno - povprecje)/std
"""q-q"""
plt.figure()
plt.plot(norm_podatki, teoreticne_vrednosti) #scatter da v tocke
plt.title("Q-Q grafikon")
plt.show()

plt.figure()
plt.plot(norm_podatki, teoreticne_vrednosti) #scatter da v tocke
plt.plot(teoreticne_vrednosti, teoreticne_vrednosti) #primerjava z normalno porazdelitvijo
plt.title("Q-Q primerjava z normalno porazdelitvijo")
plt.show()
```

```python
# -*- coding: utf-8 -*-
"""
Created on Sun Aug  8 21:40:42 2021

@author: Vito ZaloĹžnik
"""



import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
from scipy.special import ndtri

pot = "kibergrad.csv"
data = pd.read_csv(pot)
lastnosti = data.columns
#print(data.head())

dohodki = data.iloc[:,[3]] #stolpec dohodkov

vzorci_povprecja = []
n = 400 #elikost vzorca
m = 1000 #st vzorcev
for i in range(m):
    vzorec = dohodki.sample(n=400)
    vzorec_mean = vzorec.describe().loc['mean'][0] #povprecje vzorca
    vzorci_povprecja.append(vzorec_mean)



q1 = np.percentile(vzorci_povprecja, 25) #Q1
q3 = np.percentile(vzorci_povprecja, 75) #Q3

print("q1 je: ", q1)

print("q3 je: ", q3)

sirina = 2*(q3-q1)/np.cbrt(m)

zacetek =  (min(vzorci_povprecja)//400)*400
konec = (max(vzorci_povprecja)//400+1)*400
print("sirina je: ", sirina) #zaokrozimo na 400

sirina = 400

"""prvi graf"""
plt.figure()
plt.hist(vzorci_povprecja, bins=int(((konec - zacetek)//sirina )), range=(zacetek, konec))
plt.title("Histogram vzorÄnih povpreÄij")
plt.show()


print("Popravljena sirina je: ", sirina)
povprecje = np.mean(vzorci_povprecja)
print("Povprecje vzorcnih pvprecij je: ", povprecje)

std = np.std(dohodki, ddof=1)

vzorec = dohodki.sample(n=400)
std = np.std(vzorec, ddof=1)


povprecje_vseh = np.mean(dohodki)
print("std je: ", std)
print("povprecje vseh je: ", povprecje_vseh)
```

```python
N = len(dohodki) #st vseh dohodkov
SE = int(np.sqrt((N - n)/(N)* (std*std / n)))
print("Standardna napaka za vorec velikosti 400 je: ", SE)
"""histogram z dodano nomrlano porazdelitvijo"""
x = np.linspace(zacetek, konec, n)
plt.figure()
plt.hist(vzorci_povprecja, bins=int(((konec - zacetek)//sirina )), range=(zacetek, konec), density=True)
plt.plot(x, stats.norm.pdf(x, povprecje, SE))
plt.title("Primerjava z normalno porazdelitvijo")
plt.show()



#f)
"""komulativni histogram"""
st = int(((konec - zacetek)//sirina ))
res = stats.cumfreq(vzorci_povprecja, numbins=st, defaultreallimits=(zacetek, konec)) #sesteje komulativno po intervalih
res = res[0]
res = np.insert(res, 0, 0., axis=0)
res = res/res[-1] #normiramo

x = np.linspace(zacetek,konec, st +1)
y_cdf = stats.norm.cdf(x, povprecje, SE)

plt.figure()
plt.plot(x, res)
plt.title("Komulativna porazdelitvena funkcija")
plt.show()

"""primerjava z normalno komulativno"""
plt.figure()
plt.plot(x, res)
plt.plot(x, stats.norm.cdf(x, povprecje, SE))
plt.title("Primerjava z normalno komulativno funkcijo")
plt.show()


"""Q-Q"""

urejeno = np.sort(vzorci_povprecja) #uredimo po vrsti

#normalno porazdelitev razdelimo na n+1 delov.
delcki = np.arange(1,m+1)/(m+1) #range ne vkluci zadnjega

#izracunamo teoreticne vrednosti porazdelitve
teoreticne_vrednosti = ndtri(delcki)


#normaliziramo nase vrednosti
norm_podatki = (urejeno - povprecje)/SE
"""q-q"""
plt.figure()
plt.plot(norm_podatki, teoreticne_vrednosti) #scatter da v tocke
plt.title("Q-Q grafikon")
plt.show()

plt.figure()
plt.plot(norm_podatki, teoreticne_vrednosti) #scatter da v tocke
plt.plot(teoreticne_vrednosti, teoreticne_vrednosti) #primerjava z normalno porazdelitvijo
plt.title("Q-Q primerjava z normalno porazdelitvijo")
plt.show()
```

2. naloga

```python
"""
Created on Fri Aug 5 16:21:41 2021

@author: Vito Založnik
"""


import pandas as pd
import numpy as np
import scipy.stats as stats
from scipy.stats import t


data = pd.read_csv("TempPulz.csv")
moski = data[data.SPOL == 1]
zenske = data[data.SPOL == 2]
#privzamemo da sta temperatura in spol porazdeljena normalno

m_povprecje = moski.TEMPERATURA.mean()
z_povprecje = zenske.TEMPERATURA.mean()

m_std = moski.TEMPERATURA.std(ddof=1) #ddof=1 deli z n-1 da dobimo nepristransko cenilko
z_std = zenske.TEMPERATURA.std(ddof=1)
print("Povprečje temperature moških je: ", m_povprecje)
print("Povprečje temperature žensk je: ", z_povprecje)
print("Standardni odklon temperature moških je: ", m_std)
print("Standardni odklon temperature žensk je: ", z_std)


n_m = len(moski)
n_z = len(zenske)



def interval_zaupanja_za_mi(mi, sigma, alpha, n):
    K = sigma / np.sqrt(n) * stats.t.ppf(1-alpha/2, n-1)
    a = mi - K
    b = mi + K
    return ([a,b])


interval_moski = interval_zaupanja_za_mi(m_povprecje, m_std, 0.05, n_m)
interval_zenske = interval_zaupanja_za_mi(z_povprecje, z_std, 0.05, n_z)


print("Interval zaupanja za povprečje temperature moških je: ",interval_moski)
print("Interval zaupanja za povprečje temperature žensk je: ",interval_zenske)

#iz zgornjih pdoatkov
N_m = 65
N_z = 65

u_m = 98.1046153846154
u_z =  98.39384615384616
s_m = 0.69875576232659045
s_z = 0.7434877527313662

SE_2 = 0.25*(s_m**2/N_m + s_z**2/N_z)
SE_4 = SE_2**2

SE = np.sqrt(SE_2)
print("SE je: ", SE)
v = ((8*65*65*64))*SE_4/( s_m**4 + s_z**4   )
print("Število prostorskih stopenj je: ", v)
```

3. naloga

```python
# -*- coding: utf-8 -*-
"""
Created on Fri Aug 13 16:21:41 2021

@author: Vito Založnik
"""

import pandas as pd
import numpy as np
import itertools
from itertools import *
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

data = pd.read_csv("Pulz.csv")

data.columns
my_data = data.iloc[:,[0,1,3,6]]

kombinacije = [] #vse mnozice podatkov ki jih bomo uporabili v modelu
stolpci = [0,1, 2, 3]
for L in range(0, len(stolpci)+1):
   for subset in itertools.combinations(stolpci, L):
      kombinacije.append(list(subset))
del kombinacije[0] #izbrisemo prazno mnozico


n_modelov = len(kombinacije)
AIC = []
RSS = []
reg = LinearRegression()
Y = data.PULZ1 #ocenjujemo ta podatek
n = len(Y)
#za vsako podmnozico vhodnih parametrov bomo evaluirali model
for kombinacija in kombinacije:
   m = len(kombinacija) #st parametrov
   reg.fit(my_data.iloc[: , kombinacija], Y) #regresija
   #print(reg.coef_) #koeficienti regresije

   Y_ocenjeno = reg.predict(my_data.iloc[: , kombinacija])
   mse = mean_squared_error(Y, Y_ocenjeno)
   # rss = N* MSE

   rss = n * mse
   RSS.append(rss)
   aic = 2 * m + n * np.log(rss)
   AIC.append(aic) #seznam AIC-jev za posamezni model


min_index = AIC.index(min(AIC)) #index z parametri ki dajo najmanjši AIC
optimalni_podatki = my_data.iloc[:,kombinacije[6]]
rss_od_optimalnih = RSS[min_index]


optimalni_AIC = min(AIC)
print("Najmanjši AIC je: ", optimalni_AIC, "ki ima RSS: ", rss_od_optimalnih)
optimalni_podatki = my_data.iloc[:,kombinacije[6]]
print("Najboljši podatki so: ", optimalni_podatki)

kajenje = data.iloc[:,[4]]
alkohol = data.iloc[:,[5]]

dodatno_kadi = pd.concat([optimalni_podatki, kajenje], axis=1)


dodatno_alkohol = pd.concat([optimalni_podatki, alkohol], axis=1)
```

```python
#dimenzije vseh podatkov enake tako d lahko m in n parameter ostaneta za vse
m = dodatno_alkohol.shape[1] #st parametrov
n = dodatno_alkohol.shape[0] #velikost vzorca

print("Velikost vozrca je: ", n)

AIC_2 = [optimalni_AIC]
dic = {}
dic[optimalni_AIC] = optimalni_podatki



Y = data.PULZ1 #ocenjujemo ta podatek
reg.fit(dodatno_alkohol, Y) #regresija
Y_ocenjeno = reg.predict(dodatno_alkohol)
mse = mean_squared_error(Y, Y_ocenjeno)
    # rss = N* MSE
rss = n * mse
print("RSS dodatno alkohol je: ", rss)


reg.fit(dodatno_kadi, Y) #regresija
Y_ocenjeno = reg.predict(dodatno_kadi)
mse = mean_squared_error(Y, Y_ocenjeno)
    # rss = N* MSE
rss = n * mse

print("RSS dodatno kadi je: ", rss)
```