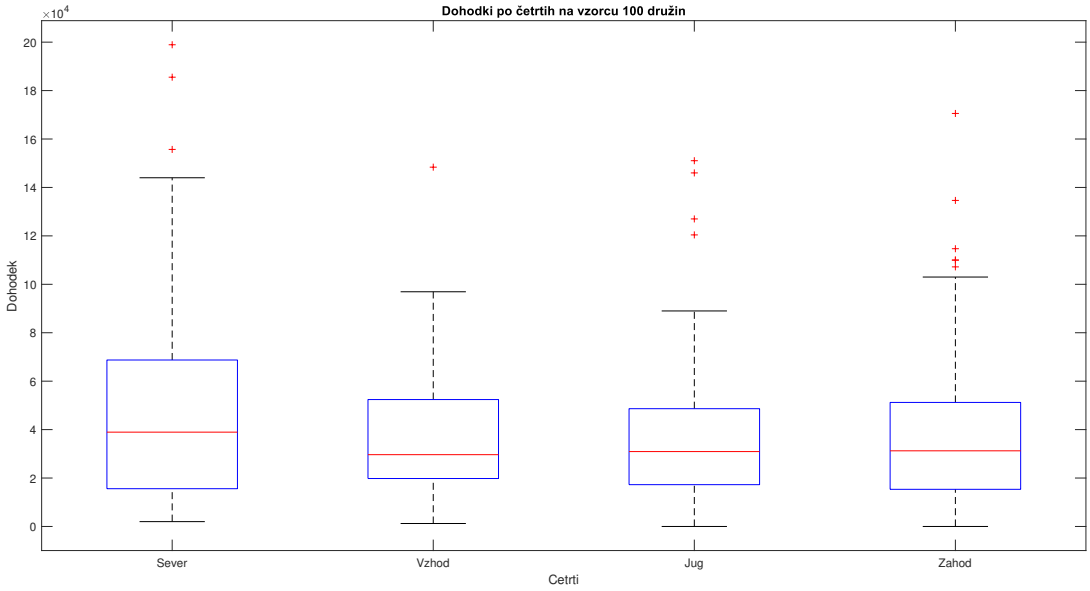
Projektna naloga pri statistiki

Timotej Stibilj
Fakulteta za matematiko in fiziko

september 2023

1 Naloga 1

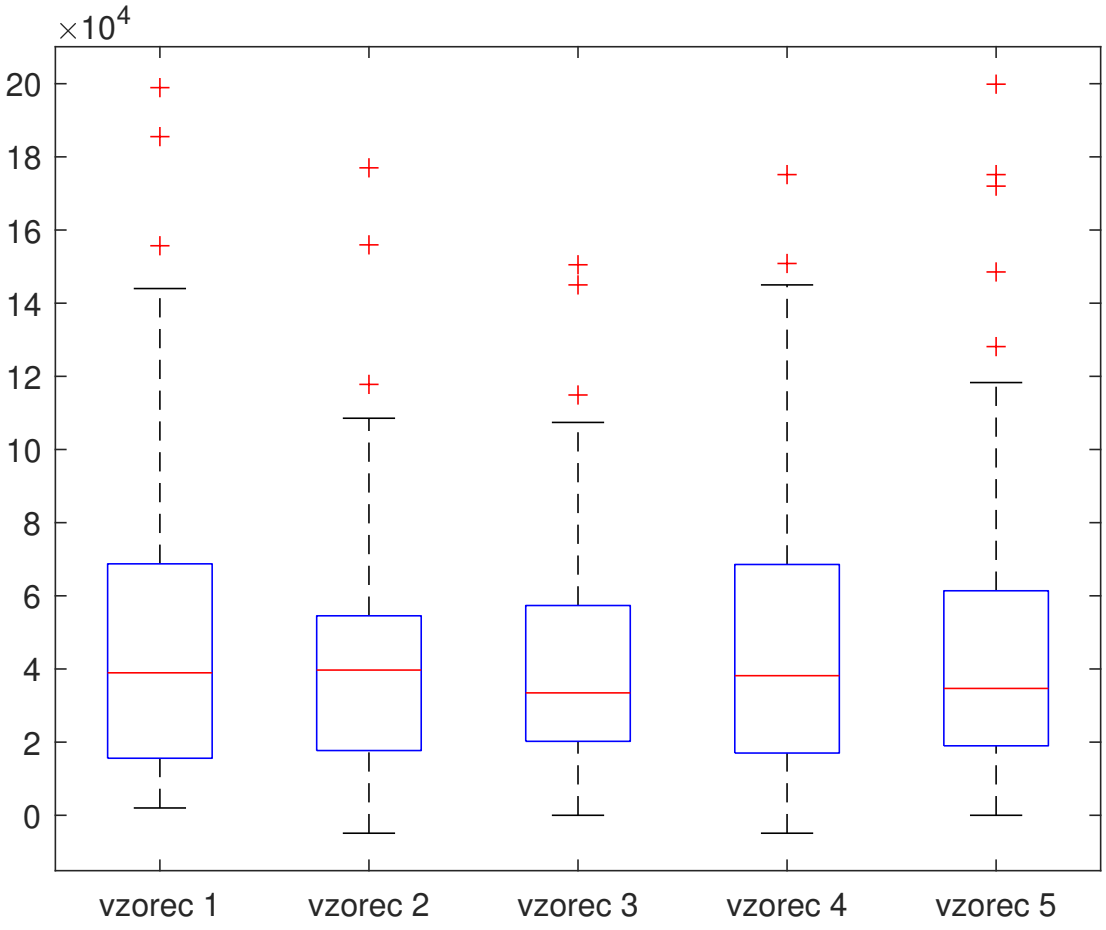
Podane imamo podatke o dohodkih 43.886 družin v mestu Kibergard. Mesto ima štiri četrti: v severni stanuje 10.149 družin, v vzhodni 10.390, v južni 13.457 in v zahodni 9.890. Iz vsake četrti vzamemo enostaven slučajni vzorec velikosti 100 in narišemo škatle z brki, ki prikazuje mediano, prvi in tretji kvartil, ter osamelce, kar nam bo omogočilo primerjavo dohodkov med četrtmi.



Primerjamo lahko mediane. Največja vzorčena mediana je na Severu, t.j. 38950€, za preostale četrti ni videti, da bi se dohodki močno razlikovali.

Dobljene mediane na Vzhodu, Jugu in Zahodu so vse okrog 30000€ in sicer 29650€, 30925€ in 31240€. Prav tako je pri teh treh četrtih mediana bližje prvemu kvartilu kakor tretjemu, kar je najvidnejše na Vzhodu. To pomeni, da je večji del dohodkov družin skoncentriran med prvim kvartilom in mediano, kakor med tretjim kvartilom in mediano. V vseh četrtih najdemo tudi osamelce. Razpon dohodkov v posameznih četrteh je velik.

Vzemimo sedaj 5 enostavnih slučajnih vzorcev iz Severne četrti in poglejmo škatle z brki. Mediane se gibljejo med 33000€ in 40000€. Dohodki



niso skoncentrirani okrog mediane oziroma razpon dohodkov je velik. To lahko vidimo tudi na velikosti interkvartilnih razmikov, npr. interkvartilni razmik v četrtem vzorcu, kjer je mediana enaka 38165€ je $IQR = Q_{\frac{3}{4}} - Q_{\frac{1}{4}} = 68566 - 17016 = 51550$. Na razlike v dohodkih na severu vpliva tudi prisotnost osamelcev, saj nekatere družine zaslužijo večkratne vrednosti mediane.

Največji osamelec je družina v petem vzorcu, ki zasluži skoraj 200000€, kar je za skoraj 4 IQR oddaljeno od mediane.

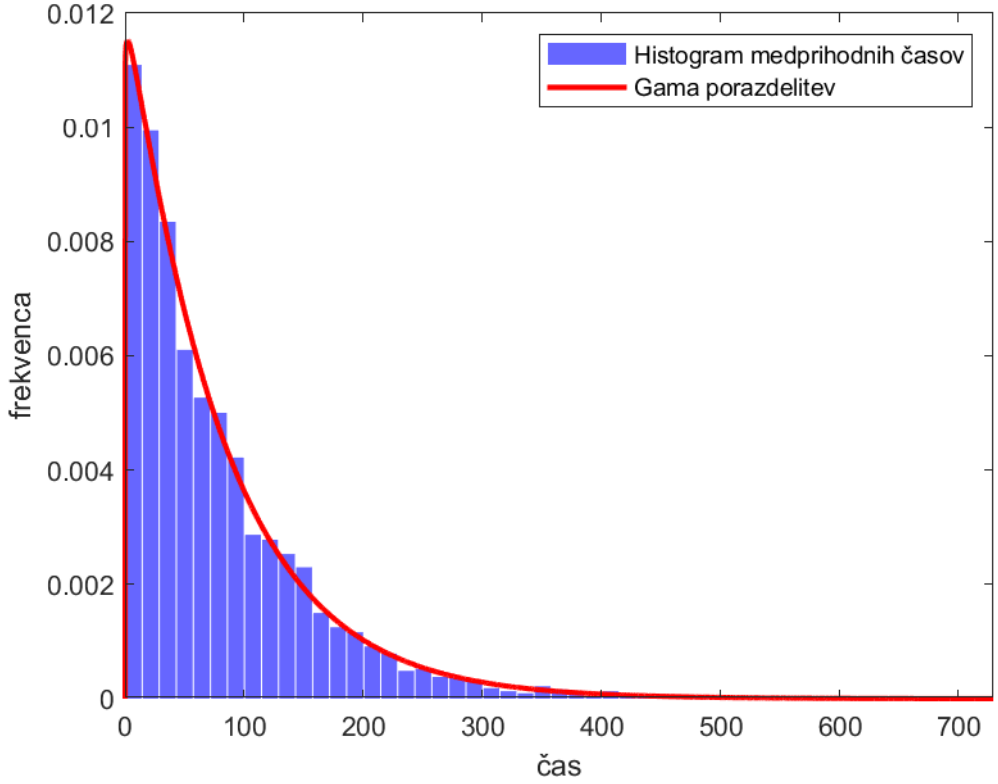
Za celotni Kibergrad (vseh 43.886 družin) izračunajmo varianco dohodka, pojasnjeno s četrtmi, in preostalo (rezidualno) varianco. Za $i = 1, \ldots, 4$ izračunamo relativne deleže w_i števila ljudi v posamičnih četrti: $w_i = \frac{n_i}{n}$, kjer je n_i število ljudi v četrti i in $n = 43886$ velikost celotne populacije. Dobimo $w_1 = 0.2313, w_2 = 0.2367, w_3 = 0.3066, w_4 = 0.2254$.
Varianco lahko zapišemo kot $\sigma^2 = \sigma_B^2 + \sigma_W^2$, kjer je $\sigma_B^2 = \sum_{i=1}^4 w_i(\mu_i - \mu)^2$ pojasnjena varianca in $\sigma_W^2 = \sum_{i=1}^4 w_i\sigma_i^2$ rezidualna varianca. Pojasnjena varianca upošteva, kako se povprečja na skupinah razlikujejo od skupnega povprečja, v našem primeru gre za razlike med povprečnimi dohodki v četrteh in skupnim povprečnim dohodkom. Intuitivno velika pojasnjena varianca pomeni, da se vrednosti(dohodki) na skupinah(četrteh) precej razlikujejo in je varianca dobro pojasnjena z stratifikacijo po teh skupinah(četrteh).

Definiramo moč učinka $\eta^2 = \frac{\sigma_B^2}{\sigma_B^2+\sigma_W^2}$, $0 \leq \eta^2 \leq 1$. Če je moč učinka blizu 1, pomeni, da pripadnost četrtim skoraj v celoti pojasne razlike v dohodkih. V mestu Kibergrad velja: $\sigma_B^2 = 9.2529 \cdot 10^6$, $\sigma_W^2 = 1.0172 \cdot 10^9$, $\eta^2 = 0.009$. Rezidualna varianca je veliko večja od pojasnjene oziroma moč učinka je blizu 0, torej zgolj pripadnost določeni četrti ne pojasni razlik v dohodkih. Že prej smo videli, da dohodki znotraj posamezne četrti niso skoncentrirani okrog povprečja(oziroma mediane), torej da je varianca znotraj posamične četrti velika, razlike med povprečji posameznih skupin pa ne toliko. Odstopala je le Severna četrt. Razlike v dohodkih bi lahko skušali pojasniti z upoštevanjem drugih faktorjev, na primer s stratifikacijo glede na stopnjo izobrazbe.

Opomba 1 *Vzorčili smo 100 ljudi iz vsake četrti, celotna populacija pa je cca. 40000. Vzorec 100 ljudi ni nujno reprezentativen, ker je pojasnjena varianca majhna oziroma je varianca znotraj posamične skupine velika v primerjavi z varianco med skupinami.*

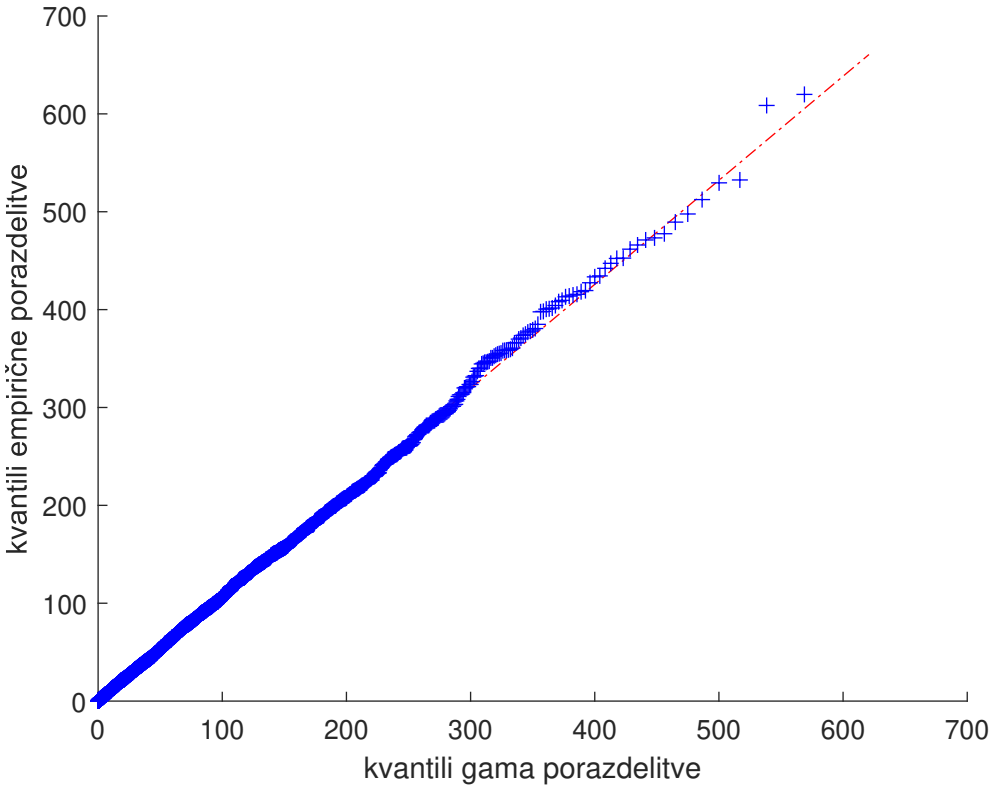
2 Naloga 2

Imamo podatke o medprihodnih časih 3935 fotonov podanih v sekundah. Narišimo histogram, v katerem širino posameznega razreda določimo v skladu z modificiranim Freedman-Diaconisovim pravilom. Izračunamo interkvartilni razmik $IQR = Q_{\frac{3}{4}} - Q_{\frac{1}{4}} = 87.59$ in določimo širino kot $\frac{2.6 \cdot IQR}{\sqrt[3]{n}}$, kjer je $n = 3935$ velikost danih podatkov (uporabljeno tudi v nadaljevanju).



Spomnimo se, da je porazdelitev gama $\Gamma(a, \lambda)$ podana z gostoto $f_x(x) = \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)}$ za $x > 0$ in $a, \lambda > 0$.

Iz histograma je razvidno, da porazdelitev gama dobro modelira medprihodne čase. V matlabu si lahko za parametre porazdelitve gama pomagamo s funkcijo $fitdist$(podatki, 'gamma'). Funkcija $fitdist$ uporablja metodo največjega verjetja. Ocenili bomo parametre porazdelitve gama še brez direktne uporabe vgrajene funkcije $fitdist$. Pred tem pa si še s pomočjo (Q-Q)-grafikona prepričajmo, da izgleda, da so medprihodni časi dobro modelirani z gama porazdelitvijo. Izgleda, da frekvence časov upadajo približno eksponentno, zato smo v grafikonu uporabili $a = 1$, kar nam da poseben primer gama porazdelitve, t.j. eksponentno porazdelitev. Parameter λ pa smo določili tako, da grafikon približno sledi simetrali lihih kvadrantov.



Sedaj ocenimo parametre porazdelitve gama. Začnimo z metodo momentov. Za porazdelitev gama poznamo pričakovano vrednost in varianco. Za $X \sim \Gamma(a, \lambda)$ velja: $E[X] = \frac{a}{\lambda}$ in $\mathrm{Var(X)} = \frac{a}{\lambda^2}$. Iz variance dobimo drugi moment: $E[X^2] = \mathrm{Var(X)} + E[X]^2 = \frac{a(1+a)}{\lambda^2}$. Vstavimo $a = \lambda E[X]$ v drugo enačbo in računamo:

$$\begin{aligned} E[X^2] &= \frac{\lambda E[X](1 + \lambda E[X])}{\lambda^2} \\ \lambda E[X^2] &= E[X] + \lambda E[X]^2 \\ \lambda(E[X^2] - E[X]^2) &= E[X] \\ \lambda &= \frac{E[X]}{E[X^2] - E[X]^2}. \end{aligned}$$

Iz $a = \lambda E[X]$ potem sledi:

$$a = \frac{E[X]^2}{E[X^2] - E[X]^2}.$$

Teoretična momenta ocenimo z empiričnima: $\widehat{E[X]} = \overline{X} = \frac{X_1 + \cdots + X_n}{n}$ in $\widehat{E[X^2]} = \overline{X^2} = \frac{X_1^2 + \cdots + X_n^2}{n}$.

Dobimo cenilki za λ in a po metodi momentov:

$$\hat{\lambda} = \frac{\overline{X}}{\overline{X^2} - \overline{X}^2} \qquad \text{in} \qquad \hat{a} = \frac{\overline{X}^2}{\overline{X^2} - \overline{X}^2}.$$

Pri konkretnih podatkih dobimo $\overline{X} = 79.9352$ in $\overline{X^2} = 1.2701 \cdot 10^4$, kar nam podaja oceni $\hat{\lambda} = 0.0127$ in $\hat{a} = 1.0124$.

Poglejmo si še oceno parametrov z metodo največjega verjetja. Verjetje je definirano kot $L(a, \lambda | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{\lambda^a x_i^{a-1} e^{-\lambda x_i}}{\Gamma(a)}$. Logaritem verjetja nam da:

$$\begin{aligned} l(a, \lambda | x_1, x_2, \ldots, x_n) &= \log L(a, \lambda | x_1, x_2, \ldots, x_n) \\ &= \sum_{i=1}^{n} \left(a \log(\lambda) + (a-1) \log(x_i) - \lambda x_i - \log(\Gamma(a)) \right) \\ &= n \cdot (a \log(\lambda) - \log(\Gamma(a))) + (a-1) \sum_{i=1}^{n} \log(x_i) - \lambda \sum_{i=1}^{n} x_i. \end{aligned}$$

Iščemo maksimum $l(a, \lambda | x_1, x_2, \ldots, x_n)$, pri čemer je $a > 0$ in $\lambda > 0$. Maksimum bo dosežen v stacionarni točki, ker funkcija l na robovih limitira proti $-\infty$. Poglejmo stacionarne točke. Za krajši zapis pišemo l namesto $l(a, \lambda | x_1, x_2, \ldots, x_n)$. Dobimo:

$$\frac{\partial l}{\partial \lambda} = \frac{na}{\lambda} - \sum_{i=1}^{n} X_i = 0.$$

Izrazimo lahko

$$\frac{a}{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}.$$

Izračunamo še parcialni odvod po a:

$$\frac{\partial l}{\partial a} = n \left(\log(\lambda) - \frac{d}{da} (\log(\Gamma(a))) \right) + \sum_{i=1}^{n} \log(X_i) = 0.$$

Če iz prve enačbe izrazimo $\lambda = \frac{a}{\overline{X}}$ problem iskanja stacionarne točke funkcije l prevedemo na iskanje ničle neelementarne funkcije:

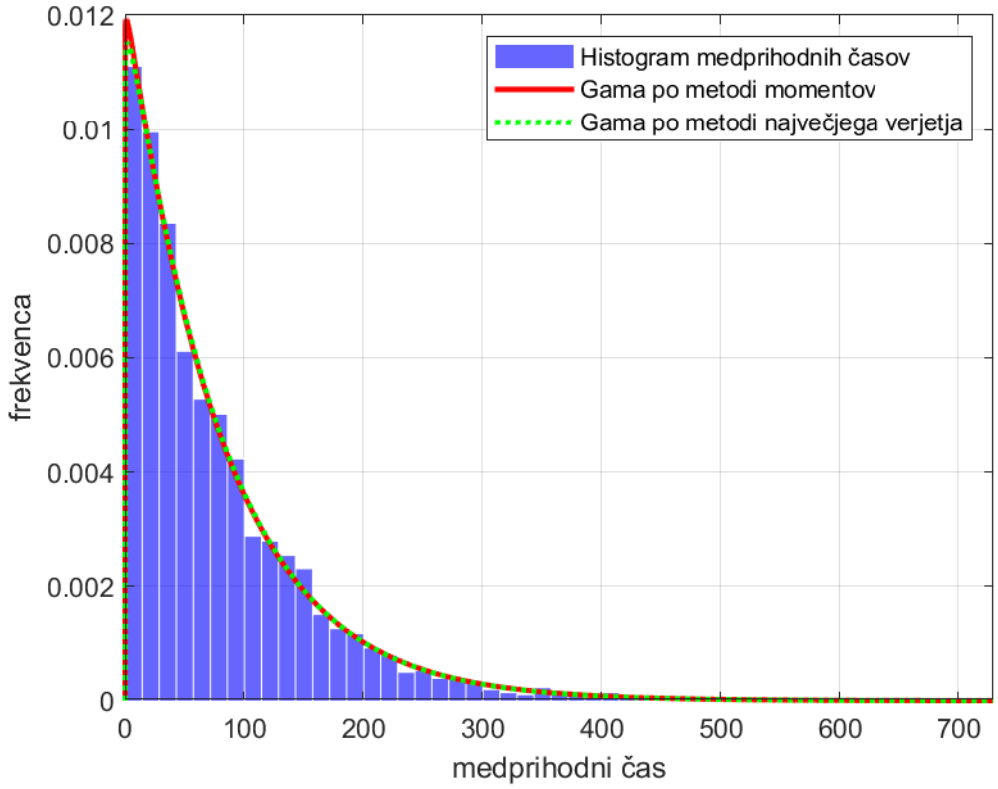
$$h(a) = n \left(\log(\frac{a}{\overline{X}}) - \frac{d}{da} (\log(\Gamma(a))) \right) + \sum_{i=1}^{n} \log(X_i). \tag{1}$$

Logaritemskemu odvodu funkcije gama pravimo funkcija digama. V matlabu lahko za funkcijo digama uporabimo vgrajeno funkcijo $\Psi(a)$. Za iskanje ničle lahko nato uporabimo funkcijo $fzero$. Ničla določa stacionarno točko in

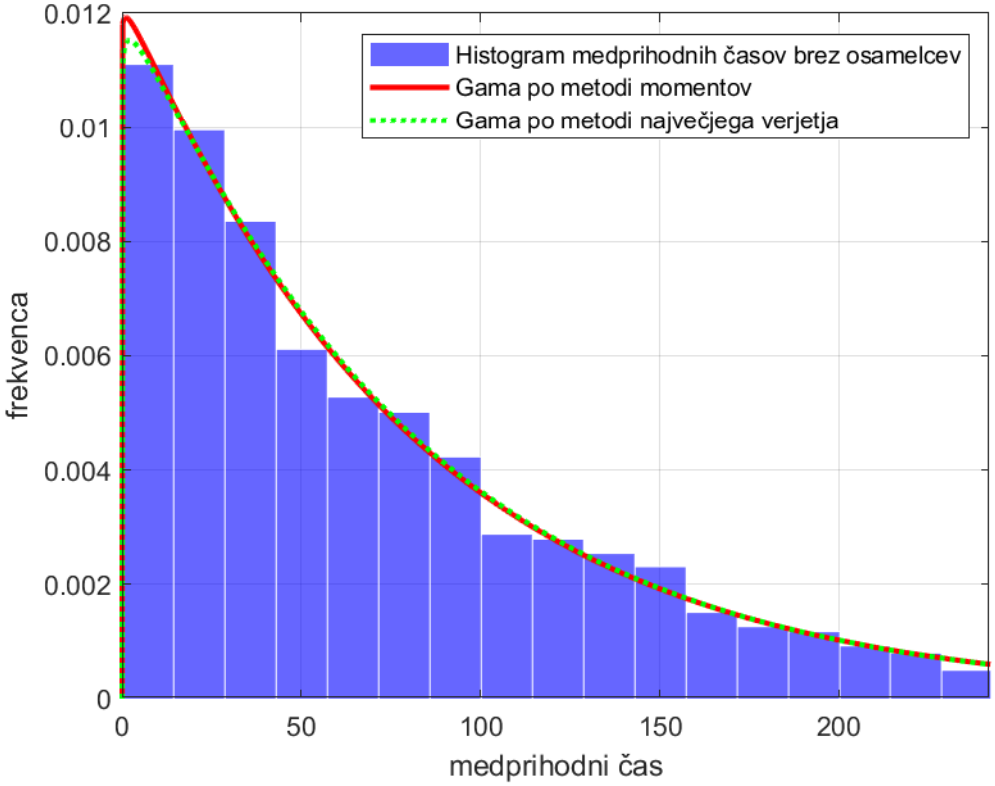
posledično oceno parametrov a in λ po metodi največjega verjetja.
Z upoštevanjem podatkov o medprihodnih časih dobimo:

$$\hat{a}_2 = 1.0263 \qquad \text{in} \qquad \hat{\lambda}_2 = \frac{\hat{a}_2}{\overline{X}} = 0.0128.$$

Vidimo, da tako metoda momentov kot metoda največjega verjetja dasta primerljivi oceni, kar lahko vidimo tudi grafično.

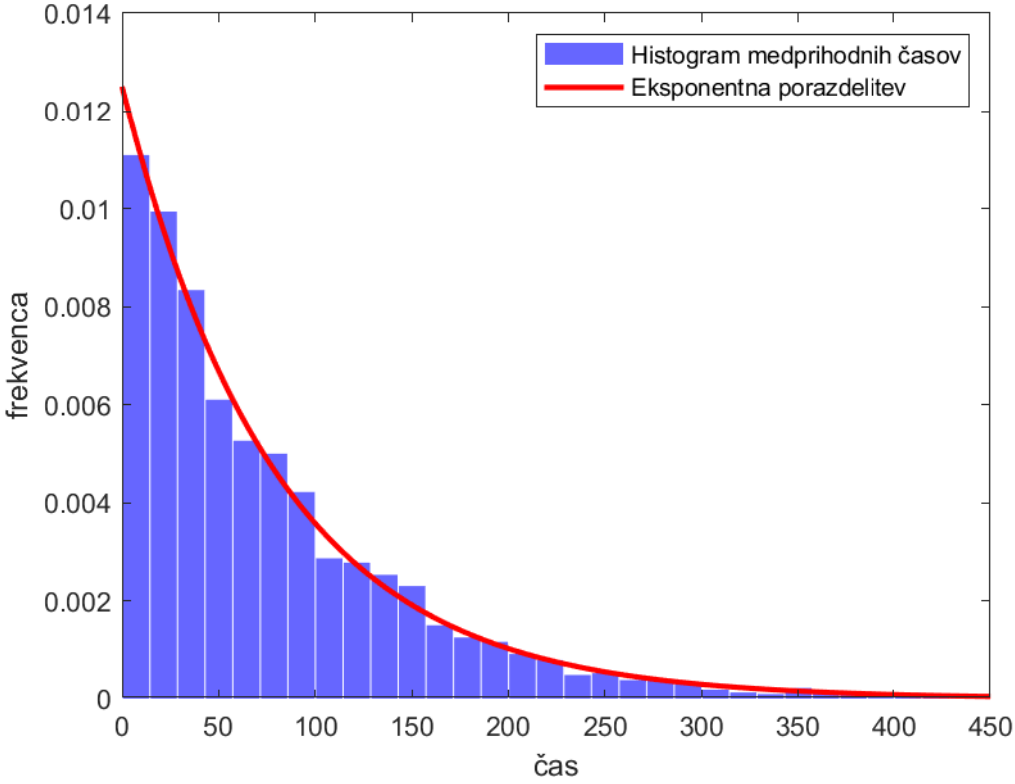


Prikažimo to še na grafu brez osamelcev, t.j. do vrednosti $Q_{\frac{3}{4}} + 1.5 \cdot IQR$.



Histogram se dobro prilega dobljenima porazdelitvama. Edino morebitno odstopanje se pojavi za čase blizu 0. Obe porazdelitvi gresta namreč za majhne čase proti 0, medtem ko na grafikonu vidimo, da imamo tudi kar nekaj primerov časov blizu 0. Frekvenca je v najmanjšem razredu pravzaprav največja, a to bi lahko posledica premajhnega števila podatkov oziroma prevelike širine razreda, saj ta vsebuje tudi maksimum obeh ocenjenih porazdelitev gama.

Medprihodne čase bi lahko modelirali tudi z eksponentno porazdelitvijo, ki je poseben primer gama porazdelitve pri $a = 1$ (pri obeh gama porazdelitvah do zdaj je bil a le malenkost večji od 1). Razlika z do zdaj uporabljenima gama porazdelitvama je v tem, da eksponentna porazdelitvena funkcija ne gre proti 0 za majhne čase. Dobimo:



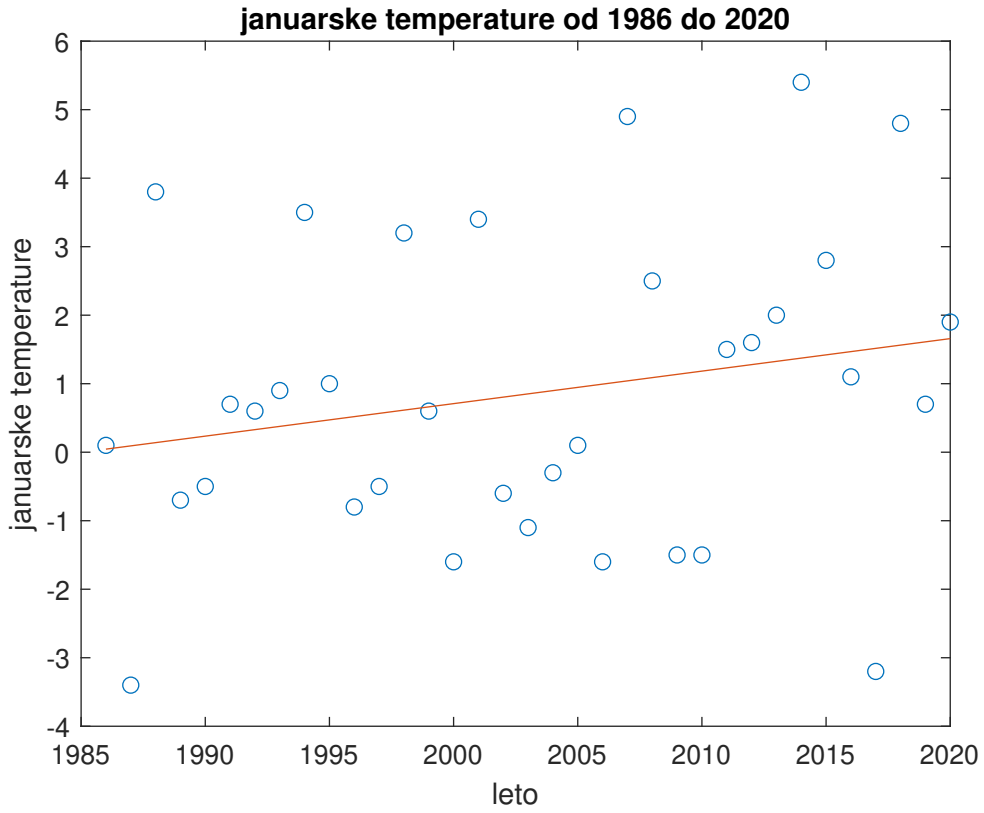
3 Naloga 3

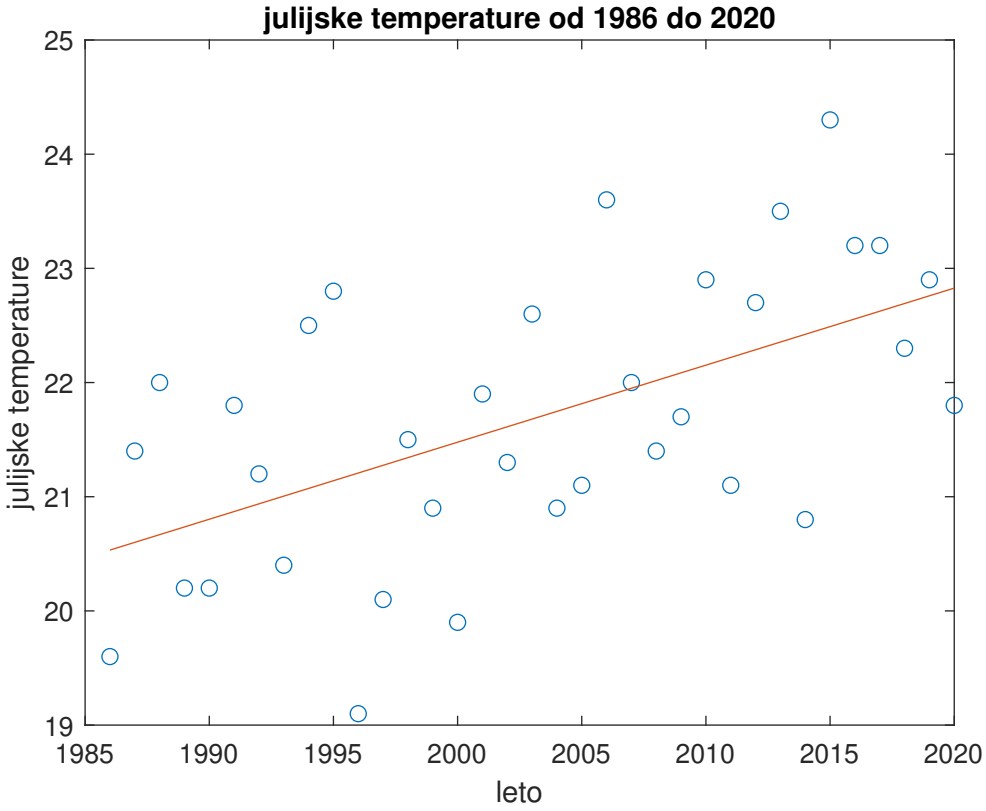
Podane imamo izmerjene mesečne temperature v Ljubljani v letih od 1986 do 2020. Z linearno regresijo preizkusimo, ali januarske in julijske temperature naraščajo enako hitro ali z različno stopnjo. Za januar postavimo model $Y_i = a_{jan} + k_{jan}X_i + \varepsilon_i$ za $i = 1986, \ldots, 2020$. Pri tem je Y_i izmerjena januarska temperatura po v letu i, X_i predstavlja leto i, a_{jan} in k_{jan} sta parametra, ki ju želimo oceniti, ε_i pa je šum in predstavlja odstopanje od regresijske premice. Pri tem za ϵ predpostavimo $\epsilon \sim N(0, \sigma^2 I)$. Podobno naredimo za julij: $Y_{2,i} = a_{jul} + k_{jul}X_i + \epsilon_{2,i}$. Želeli bi preizkusiti hipotezo $k_{jan} = k_{jul}$, torej enakost strmine regresijskih premic. Regresijska modela

združimo v enega samega: $Y = X\beta + \epsilon$, kjer je:

$$Y = \begin{bmatrix} \text{Jan. temp.} \\ \text{———} \\ \text{Jul. temp.} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 0 & 1986 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 2020 & 0 \\ 1 & 1 & 1986 & 1986 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 2020 & 2020 \end{bmatrix}, \quad \beta = \begin{bmatrix} a_0 \\ a_1 \\ k_0 \\ k_1 \end{bmatrix} \text{ in } \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \\ \epsilon_{n+1} \\ \vdots \\ \epsilon_{2n} \end{bmatrix}.$$

Za ϵ privzamemo $\epsilon \sim N(0, \sigma^2 I_{2n})$.
Iz modela lahko dobimo začetno vrednost in koeficient januarske in julijske regresijske premice. Za januar je začetna vrednost a_0 in koeficient $k_{jan} = k_0$, za julij pa je začetna vrednost $a_{jul} = a_0 + a_1$ in koeficient $k_{jul} = k_0 + k_1$. Dobimo $k_{jan} = 0.0475$ in $k_{jul} = 0.0675$. Izgleda, da julijske temperature naraščajo hitreje, kar vidimo tudi na spodnjem prikazu regresijskih premic.





Sedaj bi želeli testirati ali julijske temperature naraščajo z drugačno stopnjo kot januarske. Postavimo ničelno hipotezo $H_0 : k_1 = 0$ proti alternativni hipotezi $H_1 : k_1 \neq 0$. Ničelna hipoteza pravi, da naraščajo z isto stopnjo, alternativna pravi, da naraščajo z različno stopnjo.
Nepristranska cenilka za σ^2 je $\hat{\sigma_+}^2 = \frac{RSS}{2n-p}$, kjer je $2n = 70$ število podatkov o temperaturah in $p = 4$ število parametrov v β in $RSS = ||Y - X\hat{\beta}||^2$. Na danih podatkih dobimo $\hat{\sigma_+}^2 = 2.9217$.
Velja tudi:

$$\frac{c^T\hat{\beta} - c^T\beta}{\hat{\sigma_+}\sqrt{c^T(X^TX)^{-1}c}} \sim Student(2n - p) = Student(66).$$

Recimo, da velja $k_1 = 0$. Potem je za $\mathbf{c} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T$:

$$\frac{c^T\widehat{\beta}}{\hat{\sigma_+}\sqrt{c^T(X^TX)^{-1}c}} = 0.4950.$$

Če je $W \sim Student(m)$, velja:

$$\begin{aligned} P(|W| > D) &= 2P(W > D) \leq \alpha \\ P(W \leq D) &= 1 - \frac{\alpha}{2} \\ D &= F^{-1}_{Student(m)}\left(1 - \frac{\alpha}{2}\right). \end{aligned}$$

Za $\alpha = 0.05$ in $\alpha = 0.01$ dobimo: $D = 1.9966$ oziroma $D = 2.6524$. Hipoteze, da je $k_1 = 0$ torej ne zavrnemo. Drugače rečeno, ne zavrnemo hipoteze, da temperature rastejo z isto stopnjo. Na podlagi testa, ne moremo reči, da julijske temperature res rastejo z drugo stopnjo kakor januarske.
Ničelno hipotezo $H_0 : k_1 = 0$ bi lahko testirali tudi proti alternativni hipotezi, $H_1 : k_1 > 0$, torej da julijske temperature naraščajo hitreje. Dobili bi enostranski preizkus, ki niti ne zavrže ničelne hipoteze.