1. domača naloga

Statistika 2 in Matematična statistika

Zala Jamšek, 27122040

1. avgust 2014

1. naloga

Za analizo v nalogi sem si izbrala podatke iz datoteke 2010.dohodek.eur. Odvisna spremenljivka v teh podatkih je minimalni dohodek v EUR/mesec, odvisna spremenljivka pa je BDP na prebivalca v EUR.

```
> #prebrani podatki iz .txt datoteke
> setwd('/Users/zala/Documents/Statistics_2/')
> data_BDP <- read.table('2010.dohodek.eur.txt', header=TRUE, sep = ' ')
> data_BDP[1:4,]
```

```
          dohodek BDP.eur
Avstrija       NA   31100
Belgija   1387.50   29200
Bolgarija  122.71   10700
Ciper          NA   23600
```

Ker je nekaj nepopolnih podatkov, sem jih odstranila.

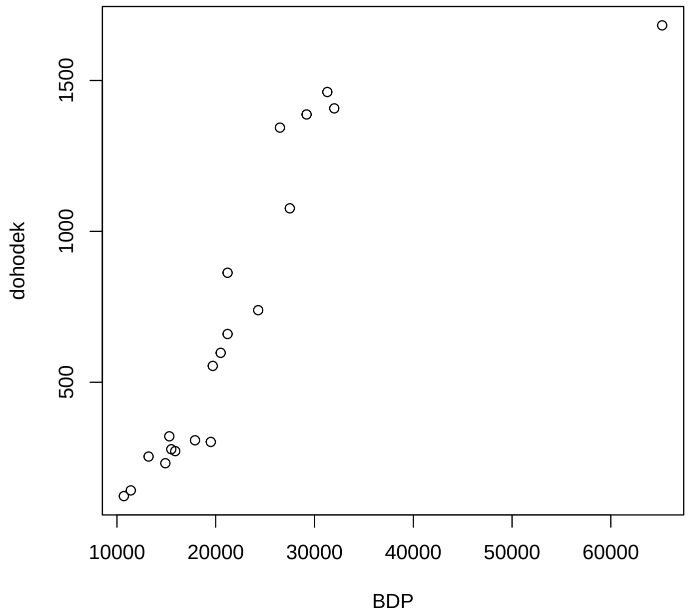
```
> #ostranjeni nepopolni podatki
> data_BDP <- na.omit(data_BDP)
> summary(data_BDP)
```

```
    dohodek          BDP.eur
 Min.   : 122.7   Min.   :10700
 1st Qu.: 276.5   1st Qu.:15450
 Median : 575.8   Median :20100
 Mean   : 700.2   Mean   :22645
 3rd Qu.:1143.3   3rd Qu.:26750
 Max.   :1682.8   Max.   :65200
```

1.

Razsevni grafikon ima na osi x neodvisno spremenljivko BDP na prebivalca v EUR in na y osi odvisno spremenljivko dohodek.

```
> #podatki kot vektor
> BDP <- data_BDP$BDP.eur
> dohodek <- data_BDP$dohodek
> #RAZSEVNI DIAGRAM
> plot(x=BDP, y=dohodek)
```



Na grafikonu vsaka točka predstavlja eno državo. Ena država močno izstopa predvsem po velikosti BDP na prebivalca in ima poleg tega tudi največji dohodek. Sicer pa je iz grafikona razvidno, da imajo države z višjim BDP na prebivalca tudi višji dohodek. Bolj natančna mera za povezanost med spremenljivkama pa je korelacijski koeficient. Prav tako točke ležijo skoraj na premici.

Pearsonov koeficient korelacije meri korelacijo (linearno povezanost) med dvema spremenljivkama in je izračunan kot razmerje med kovarianco in standardnima odklonoma:

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_i)(y_i - \mu_i)$$

Funkcija za izračun koeficienta je že vgrajena v R.

```
> #Pearsonov koeficient
> Pearson<- function(d){
+   pearson <-cor(d, method = 'pearson')
+   return(pearson[2,1])
+ }
> Pearson(data_BDP)
```

```
[1] 0.8388114
```

Pearsonov korelacijski koeficient je 0.8388114. Kot sem opazila že iz same oblike razsevnega grafikona je korelacija med spremenljivkama pozitivna. Pearsonov koeficient je blizu 1, kar pomeni, da je korelacija (povezanost) med spremenljvkama skoraj linearna, kar je videti tudi iz zame oblike grafikona, saj točke ležijo skoraj na premici.

2.
Za regresijsko premico želim določiti a in b, tako da bo

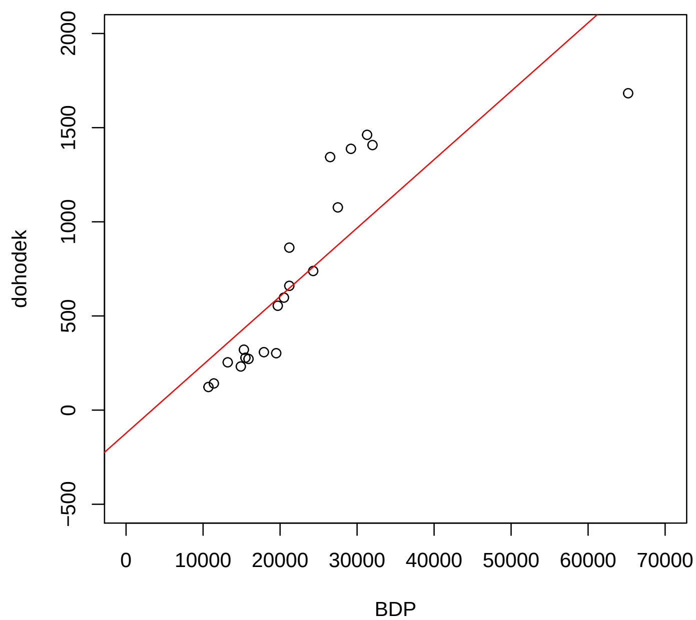
$$dohodek = a + bBDP$$

Parametra a in b sta dobljena z metodo najmanjših kvadratov. V R-u se koeficienta dobi s funkcijo lm. Funkcijo $koef_lr$, ki vrne parametra a in b sem ustvarila, ker jo bom uporabila v nadaljevanju naloge.

```
> #funkcija, ki vrne koeficiente linearne regresije
> koef_lr <- function(d){
+   lr <- lm(d$dohodek~d$BDP)
+   a <-lr$coefficients[1]
+   b <-lr$coefficients[2]
+   return(c(a,b))
+ }
> l_reg <- lm(dohodek~BDP)
> summary(l_reg)

Call:
lm(formula = dohodek ~ BDP)

Residuals:
    Min      1Q  Median      3Q     Max 
-563.10 -167.82  -71.19  203.74  503.57 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.223e+02  1.413e+02  -0.866    0.398    
BDP          3.632e-02  5.557e-03   6.537 3.83e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 287.5 on 18 degrees of freedom
Multiple R-squared:  0.7036,	Adjusted R-squared:  0.6871 
F-statistic: 42.73 on 1 and 18 DF,  p-value: 3.826e-06
```

Dobimo rezultata: $a = -122,336$, $b = 0,0362$, kar pomeni, da če se BDP na prebivalca na prebivalca poveča za eno 1 EUR, se minimalni dohodek v državi poveča za $0,0362$ EUR. In če je BDP na prebivalca enak 0, je minimalni dohodek v državi negativen in enak $-122,336$. V razsevni grafikon je vrisana regresijska premica.

```
> #linearna regresija
> plot(x=BDP, y=dohodek, xlim = c(0,70000), ylim=c(-500,2000))
> abline(l_reg, col=2)
```



3.

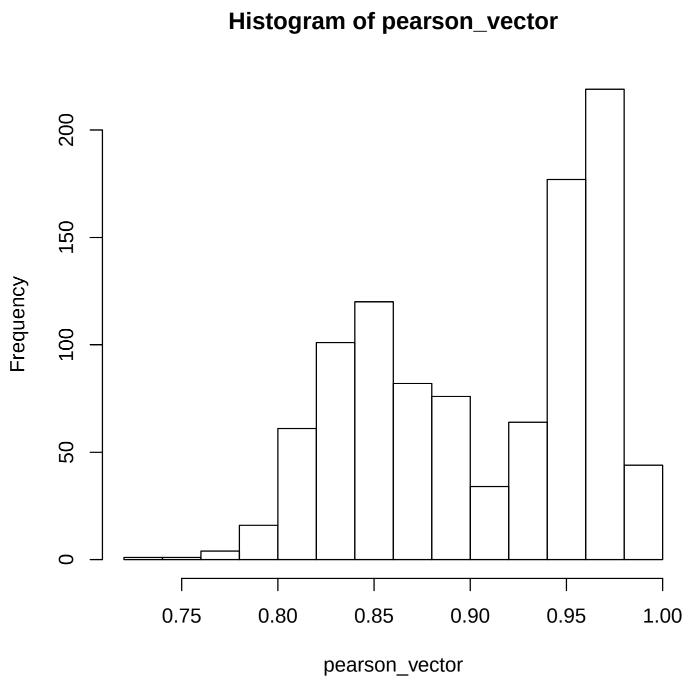
Za simulacijo v vsakem od 1000 korakov je naključno izberan vzorec velikosti 10 in na tem vzorcu je izračunan Pearsonov koeficient korelacije. Koeficienti so predstavljeni v histogramu.

```
> #Simulacija za Pearsonov koeficient
> pearson_vector <- numeric(0)
> for (i in 1:1000){
+   data_BDP_pearson <- data_BDP[sample(nrow(data_BDP),10),]
+   pearson_vector[i]<-Pearson(data_BDP_pearson)
+ }
> hist(pearson_vector)
> mean(pearson_vector)
```

```
[1] 0.906198
```

```
> sd(pearson_vector)
```

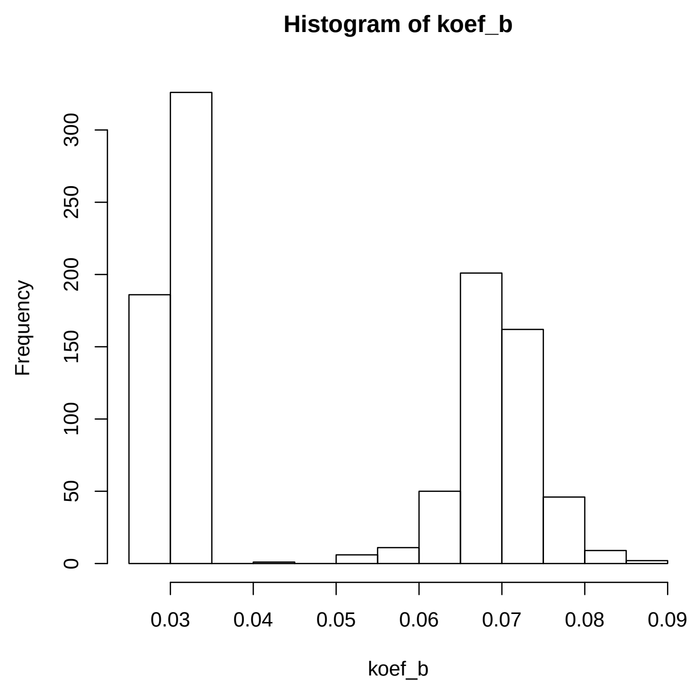
```
[1] 0.05930891
```



Povpre??je vzor??ne porazdelitve je $0,9017269$, standardni odkolon pa 0.06092566

4.
V vsakem koraku simulacije je naklju??no izberan vzorec velikosti 10 in na tem vzorcu izra??unan regresijski koeficient. Regresijski koeficienti so predstavljeni v histogramu.

```
> #Simulacija za regresijski koeficient
> koef_b <- numeric(0)
> for (i in 1:1000){
+   data_BDP_koef <- data_BDP[sample(nrow(data_BDP),10),]
+   koef_b[i] <- koef_lr(data_BDP_koef)[2]
+ }
> hist(koef_b)
```



Na histogramu je opaziti dvojni vrh. Pri v drugem stolpcu (vrednosti 0,03-0,035) in drugi vrh v devetem okencu (vrednosti 0,065/0,07).

```
> perm_data <- data_BDP[sample(nrow(data_BDP),10),]
> perm_data

                             dohodek BDP.eur
\304\214e\305\241ka.republika  302.19   19500
Belgija                      1387.50   29200
Malta                         659.92   21200
Velika.Britanija             1076.46   27500
\305\240panija                738.85   24300
Poljska                       320.87   15300
Nizozemska                   1407.60   32000
Litva                         231.70   14900
Luksemburg                   1682.76   65200
Slova\305\241ka               307.70   17900

> #t je funkcija, ki izracuna testno statistiko t
> t <- function(data){
+   r <- Pearson(data)
+   n <- length(data[,1])
+   t <-sqrt(n-2)*(r/sqrt(1-r^2))
```

```
+   return(t)
+ }
> t_star <- t(perm_data)
> t(data_BDP)
```

```
[1] 6.536794
```

```
> p_star <- Pearson(perm_data)
> #PERMUTACIJSKI TEST
> #Monte Carlo Simulacija za
>
> permutation_test <- function(data, B, test){
+   t_vect <- numeric(0)
+
+   BDP <- data$BDP
+   dohodek <- data$dohodek
+
+   for (i in 1:B){
+      X<-data.frame(sample(BDP),sample(dohodek))
+      t_vect[i] <- test(X)
+   }
+   hist(t_vect)
+  return(t_vect)
+ }
>
> #permutation_test(perm_data,1000,t)
> #permutation_test(perm_data,1000,Pearson)
>
> #permutation_test(data_BDP,2000,t)>t(data_BDP)
> #sum(permutation_test(data_BDP,1000,Pearson)>=p_star)
>
> #permutation_test(perm_data,1000,t)>t(data_BDP)
> #permutation_test(perm_data,1000,Pearson)>p_star
```

2. naloga

1.
Izra??un parametrov pareto porazdelitve z metodo najve??jega verjetja:

$$f(x_i) = \frac{\alpha\lambda^\alpha}{x_i^{\alpha+1}}$$

$$L = \prod_{i=1}^{n} \frac{\alpha\lambda^\alpha}{x_i^{\alpha+1}} = \alpha^n \lambda^{\alpha n} \prod_{i=1}^{n} x_i^{-(\alpha+1)}$$

$$l = log(L) = nlog(\alpha) + n\alpha log(\lambda) - (\alpha+1)\sum_{i=1}^{n} log(x_i)$$

$$\frac{\partial u}{\partial \lambda} = \frac{n\alpha}{\lambda} = 0$$

Za parameter λ iskanje maksimuma z odvajanjem ni ustrezna metoda. Iz definicije Pareto porazdelitve pa je znano, da je $\lambda \leq x_i$ za vsak i. Torej lahko je parameter ocenjen iz tega pogoja:

$$\lambda \leq x_i \Rightarrow \hat{\lambda} = min(x_i)$$

$$\frac{\partial l}{\partial \alpha} = \frac{n}{\alpha} + nlog(\lambda) - \sum_{i=1}^{n} log(x_i) = 0$$

$$\frac{n}{\alpha} = \sum_{i=1}^{n} log(x_i) - log(\lambda)^n$$

$$\frac{\alpha}{n} = \frac{1}{\sum_{i=1}^{n} log(\frac{x_i}{\lambda})}$$

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} log(\frac{x_i}{\hat{\lambda}})}$$

2.

Za Fisherjevo informacijsko matriko sem najprej izra??unala druge odvode log porazdelitvene funkcije:

$$log(f(x)) = log(\alpha) + \alpha log(\lambda) - (\alpha + 1)log(x)$$

$$\frac{\partial log(f(x))}{\partial \alpha} = \frac{1}{\alpha} + log(\lambda) - log(x)$$

$$\frac{\partial log(f(x))}{\partial \lambda} = \frac{\alpha}{\lambda}$$

$$\frac{\partial^2 log(f(x))}{\partial \alpha^2} = -\frac{1}{\alpha^2}$$

$$\frac{\partial^2 log(f(x))}{\partial \lambda^2} = -\frac{\alpha}{\lambda^2}$$

$$\frac{\partial^2 log(f(x))}{\partial \alpha \partial \lambda} = \frac{1}{\lambda}$$

Fisherjevo informacijsko matriko je:

$$I(\theta) = -E[\frac{\partial^2}{\partial \theta^2} log f(x;\theta)|\theta]$$

$$I(\theta) = \begin{bmatrix} \frac{1}{\alpha^2} & -\frac{1}{\lambda} \\ -\frac{1}{\lambda} & \frac{\alpha}{\lambda^2} \end{bmatrix}$$

3.
Za metodo momentov uporabimo:

$$\mu_1 = E(X)$$

$$\mu_2 = E(X^2)$$

Momenti Pareto porazdelitve so:

$$\begin{aligned}
E(X) &= \int_{\lambda}^{\infty} x f(x) dx = \int_{\lambda}^{\infty} x \frac{\alpha \lambda^{\alpha}}{x^{\alpha+1}} dx \\
&= \alpha \lambda^{\alpha} \int_{\lambda}^{\infty} \frac{1}{x^{\alpha}} dx \\
&= \frac{\alpha \lambda^{\alpha}}{-(\alpha-1)} \frac{1}{x^{\alpha-1}} \Big|_{\lambda}^{\infty} \\
&= \frac{\alpha \lambda^{\alpha}}{-(\alpha-1)} \left(-\frac{1}{\lambda^{\alpha-1}}\right) \\
&= \frac{\alpha \lambda}{\alpha-1}
\end{aligned}$$

$$\begin{aligned}
E(X^2) &= \int_{\lambda}^{\infty} x^2 f(x) dx = \int_{\lambda}^{\infty} x^2 \frac{\alpha \lambda^{\alpha}}{x^{\alpha+1}} dx \\
&= \alpha \lambda^{\alpha} \int_{\lambda}^{\infty} \frac{1}{x^{\alpha-1}} dx \\
&= \frac{\alpha \lambda^{\alpha}}{-(\alpha-2)} \frac{1}{x^{\alpha-2}} \Big|_{\lambda}^{\infty} \\
&= \frac{\alpha \lambda^{\alpha}}{-(\alpha-2)} \left(-\frac{1}{\lambda^{\alpha-2}}\right) \\
&= \frac{\alpha \lambda^2}{\alpha-2}
\end{aligned}$$

Potreben pogoj za obstoj momentov je torej, da je $\alpha > 2$. V obeh momentih nastopata oba parametra, zato z deljenjem eliminiramo λ

$$\begin{aligned}
\frac{\mu_1^2}{\mu_2} &= \frac{\alpha^2 \lambda^2}{(\alpha-1)^2} \frac{\alpha-2}{\alpha \lambda^2} \\
&= \frac{\alpha(\alpha-2)}{(\alpha-1)^2}
\end{aligned}$$

$$
\begin{aligned}
\frac{\mu_1^2}{\mu_2} - 1 &= \frac{\alpha^2 - 2\alpha - \alpha^2 + 2\alpha - 1}{\alpha^2 - 2\alpha + 1} \\
(\alpha - 1)^2 &= \frac{\mu_2}{\mu_2 - \mu_1^2} \\
\hat{\alpha} &= 1 + \sqrt{\frac{\mu_2}{\mu_2 - \mu_1^2}} \\
&= 1 + \sqrt{1 + \frac{\mu_1^2}{\mu_2 - \mu_1^2}}
\end{aligned}
$$

Ker mora biti parameter α pozitiven, je upo??tevan pozitivni koren.

$$
\begin{aligned}
\mu_1 &= \frac{\alpha\lambda}{\alpha - 1} \\
\lambda &= \frac{\mu_1(\alpha - 1)}{\alpha} \\
\hat{\lambda} &= \frac{\mu_1(\hat{\alpha} - 1)}{\hat{\alpha}}
\end{aligned}
$$

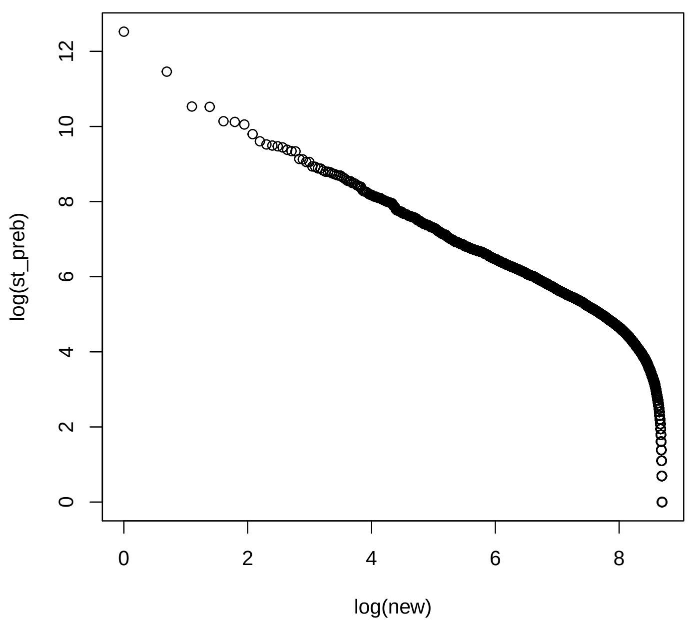
4.
Iz datoteke podatki.nasleja.txt prebrani podatki, urejeni po velikosti in na novo
izbrani samo tisti podatki, ki imajo ve?? ali enako kot enega prebivalca.

```
> data_naselja <- read.table('podatki.naselja.txt', sep='\t', header=TRUE)
> data_naselja[1:4,]

     naselje stevilo.prebivalcev
1 Ajdovscina                6597
2     Batuje                 339
3       Bela                  37
4       Brje                 390

> #urejeni podatki po padajocem st. prebivalcev
> data_naselja<-data_naselja[order(data_naselja$stevilo.prebivalcev, decreasing=TRUE),]
> sum(data_naselja$stevilo.prebivalcev>=1)

[1] 5973

> #novi podatki samo do 5973. nasleja, ki imajo st. prebivalcev vecje od 1
> new <- 1:sum(data_naselja$stevilo.prebivalcev>=1)
> data_naselja_new <- data_naselja[new,]
> data_naselja_new[1:4,]
```

```
        naselje stevilo.prebivalcev
2307 Ljubljana               274826
2556   Maribor                94809
297      Celje                37490
1832     Kranj                37151
```

Log log graf:

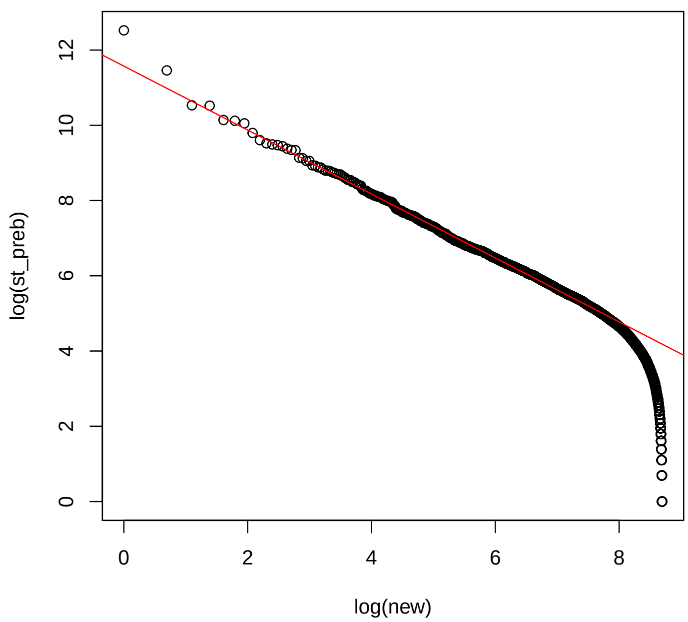
```
> st_preb <- data_naselja_new$stevilo.prebivalcev
> plot(log(new),y=log(st_preb))
```



Opaziti se da, da je na začetku log-log graf skoraj premica, kar je značilno za Pareto porazdelitev. V zadnjem delu pa točke ne ležijo več na premici.

Regresijska premica za prvih 750 naselij:

```
> n1<-750
> x<-1:n1
> st_preb_750<-(st_preb[x])
> st_preb_reg <- lm(log(st_preb_750)~log(x))
> plot(log(new),y=log(st_preb))
> abline(st_preb_reg, col=2)
```



5.

če uporabim še naslednje enačbe za momente:

$$\mu_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$$Var(X) = \mu_2 - \mu_1^2 = \frac{n-1}{n}S^2$$

Potem velja:

$$\hat{\alpha} = 1 + \sqrt{1 - \frac{n}{n-1}\frac{\hat{X}}{S^2}}$$

```
> l_mle <- min (st_preb_750)
> l_mle
```

```
[1] 406
```

```
> alpha_mle <- n1/sum(log(st_preb_750/l_mle))
> alpha_mle
```

```
[1] 1.272517
```

Parametra po metodi največjega verjetja za prvih 750 opazovanj sta:

$$\alpha = 1,272517$$

$$\lambda = 406$$

6.

V nalog 3. so izraženi parametri po metodi momentov. Za izračun teh parametrov pa je treba upoštevti še naslednje:

$$\mu_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

$$Var(X) = \mu_2 - \mu_1^2 = \frac{n-1}{n}S^2$$

Potem velja:

$$\hat{\alpha} = 1 + \sqrt{1 - \frac{n}{n-1}\frac{\bar{X}}{S^2}}$$

```
> X_bar <- mean(st_preb_750)
> S <- sd(st_preb_750)
> alpha_mm <- 1+ sqrt(1 + (n1/(n1-1))*(X_bar^2)/S^2)
> alpha_mm
```

```
[1] 2.015846
```

```
> lambda_mm <- X_bar*(alpha_mm-1)/alpha_mm
> lambda_mm
```

```
[1] 986.6094
```

7.

Novi podatki z vzorcem 100 naselij. In na teh podatki izra??unana parametra po metodi najve??jega verjetja??

```
> #7
> data_naselja_100 <- read.table('podatki.naselja.vzorec.txt', header = TRUE, sep = '\t')
> n2 <-100
> #data_naselja_vz
> st_preb_100 <- data_naselja_100$stevilo.prebivalcev
> l_mle_100 <- min (st_preb_100)
> l_mle_100
```

```
[1] 406
```

```
> alpha_mle_100 <- n2/sum(log(st_preb_100/l_mle_100))
> alpha_mle_100
```

```
[1] 1.391795
```

$$\hat{\alpha} = 1,391795$$
$$\hat{\lambda} = 406$$

8.

V to??ki 3. sem izra??unala Fisherjevo informacijo za parameter α:

$$I(\alpha) = \frac{1}{\alpha^2}$$

Od tod sledi, da je standardna napaka za parameter α:

$$Var(\alpha) = \frac{1}{nI(\alpha)}$$
$$Var(\alpha) = \frac{\alpha^2}{n}$$
$$sd(\alpha) = \sqrt{(}\frac{\alpha^2}{n})$$

```
> #8
> sd <- sqrt(alpha_mle_100^2/n2)
> sd
```

```
[1] 0.1391795
```

9.

```
> #9
> z <- qnorm(0.95)
> z
```

```
[1] 1.644854
```

```
> lb <-alpha_mle_100 - z*sqrt(sd)
> up <-alpha_mle_100 + z*sqrt(sd)
> lb
```

```
[1] 0.7781533
```

```
> up
```

```
[1] 2.005437
```

```
> alpha_mle
```

```
[1] 1.272517
```