FAKULTA INFORMAČNÍCH TECHNOLÓGIÍ ČVUT

BI-PST Domácí úkol

Patrik Jantošovič
Tomáš Zvara
Tomáš Janecký
12. prosince 2018

1 PARAMETRY A DATOVÝ SOUBOR

Reprezentant: Patrik Jantošovič
K = den narození = 16
L = počet písmen v příjmení = 10
M = ((K+L)*46)mod11 + 1 = 2
Výsledkem je tedy datový soubor: case0102, mzda dle pohlaví

1.1 VYTVOŘENÍ DATOVÉHO SOUBORU

Řešení úloh předpokladá úspěšnou instalaci knihovni Sleuth2 a vytvoření .csv souboru s příslušnými daty.
Postup uvedeme jednou na začátku abychom jsme se neopakovali.

- » install.packages("Sleuth2")
  - Instalace package Sleuth2
- » library(Sleuth2)
  - Načítaní package Sleuth2
- » write.table(case0102,"C:/data.csv",row.names=F,sep=";",dec=",")
  - Zápis dat do .csv souboru

2 Řešení úkolů

2.1 Úkol číslo 1

(1b) Načtěte datový soubor a rozdělte sledovanou proměnnou na příslušné dvě pozorované skupiny. Data stručně popište. Pro každu skupinu zvlášť odhadněte střední hodnotu, rozptyl a medián příslušného rozdělení.

- » data<-read.table("C:/data.csv",header=TRUE,sep=";")
  - Načteme data z připraveného souboru
- » female<-data[1:61,]
  - Načítaní dat pro pozorovanou skupinu: Female
- » male<-data[62:93,]
  - Načítaní dat pro pozorovanou skupinu: Male
- » female<-female[,1]
  - Odřiznutí sloupce s pohlavím pro pozorovanou skupinu: Female
- » male<-male[,1]
  - Odřiznutí sloupce s pohlavím pro pozorovanou skupinu: Male
- » length(male)
  - Velikost dat pro pozorovanou skupinu: Male
- » length(female)
  - Velikost dat pro pozorovanou skupinu: Female
- » var(male)
  - Rozptyl pro pozorovanou skupinu: Male
- » var(female)
  - Rozptyl pro pozorovanou skupinu: Female
- » mean(male)
  - Střední hodnota pro pozorovanou skupinu: Male
- » mean(female)
  - Střední hodnota pro pozorovanou skupinu: Female
- » median(male)
  - Medián pro pozorovanou skupinu: Male

- » median(female)
  - Medián pro pozorovanou skupinu: Female

Výsledky zapíšeme do následujíci tabulky:

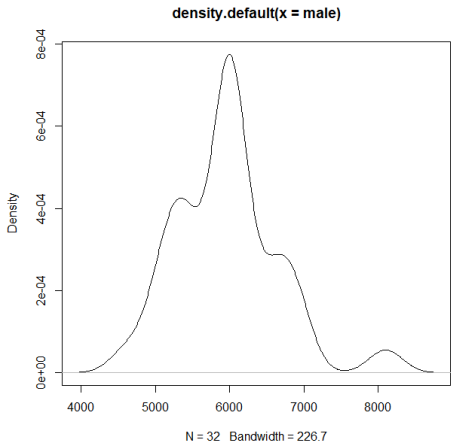
Pohlaví	Velkost dat	Střední hodnota	Rozptyl	Medián
Male	32	5956.875	477112.5	6000
Female	61	5138.852	291460.3	5220

2.2 Úkol číslo 2

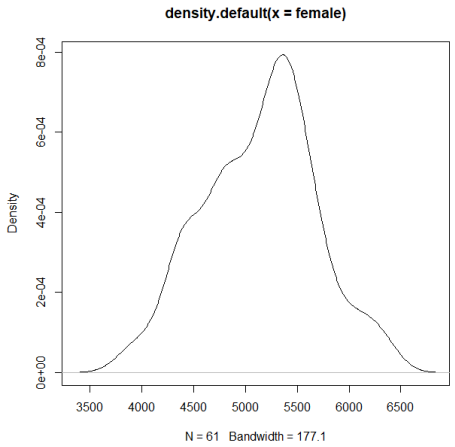
(1b) Pro každou skupinu zvlášť odhadněte hustotu a distribuční funkci pomocí histogramu a empirické distribuční funkce.

- » hist(female, freq=FALSE)
  - Vykreslení histogramu female. freq=FALSE používame jako přepínač pro hustotu
- »hist(male, freq=FALSE)
  - Vykreslení histogramu female. freq=FALSE používame jako přepínač pro hustotu
- »plot(density(male))
  - Vykreslení hustoty Male
- »plot(density(female))
  - Vykreslení hustoty Female
- »plot(ecdf(male))
  - Vykreslení empirické distribuční funkce pro Male
- »plot(ecdf(female))
  - Vykreslení empirické distribuční funkce pro Female

Výsledkem jsou grafy přiložené na následující stránce.

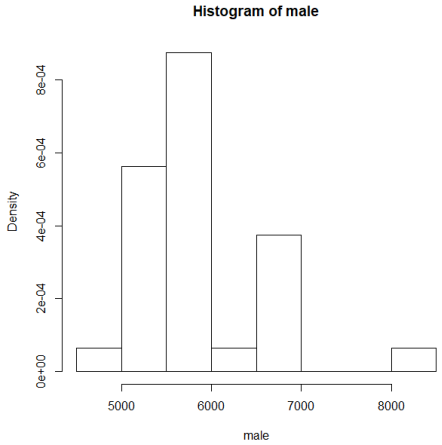


(a) Male

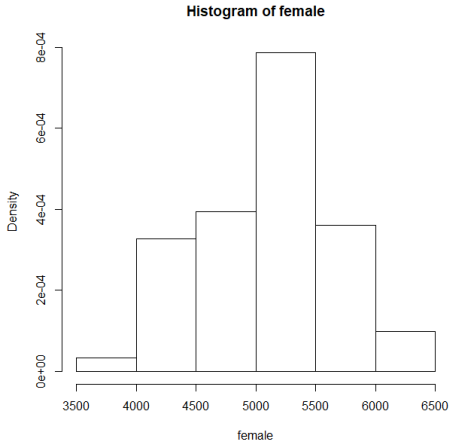


(b) Female

Obrázek 2.1: Hustota

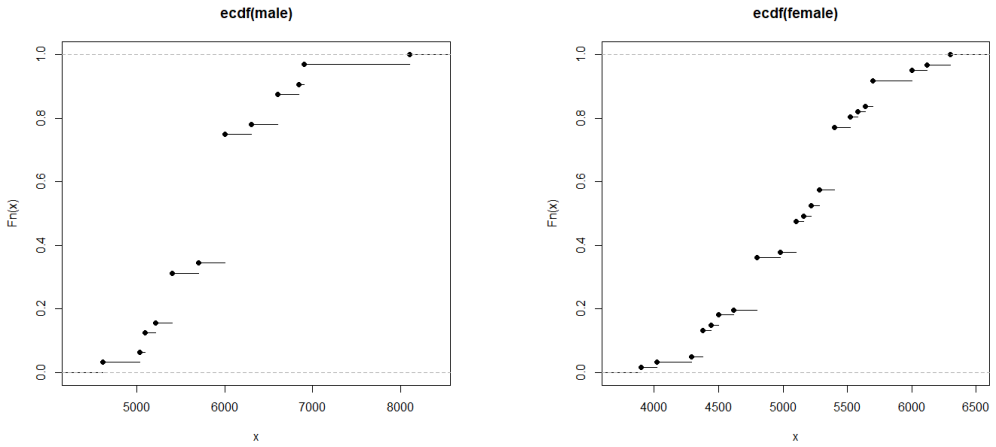


(a) Male



(b) Female

Obrázek 2.2: Histogram



(a) Male (b) Female

Obrázek 2.3: Empirická distribuční funkce

2.3 Úkol číslo 3

(3b) Pro každou skupinu zvlášť najděte nejbližší rozdělení: Odhadněte parametry normálního, exponenciálního a rovnoměrného rozdělení. Zaneste příslušné hustoty s odhadnutými parametry do grafů histogramu. Diskutujte, které z rozdělení odpovídá pozorovaným datům nejlépe.

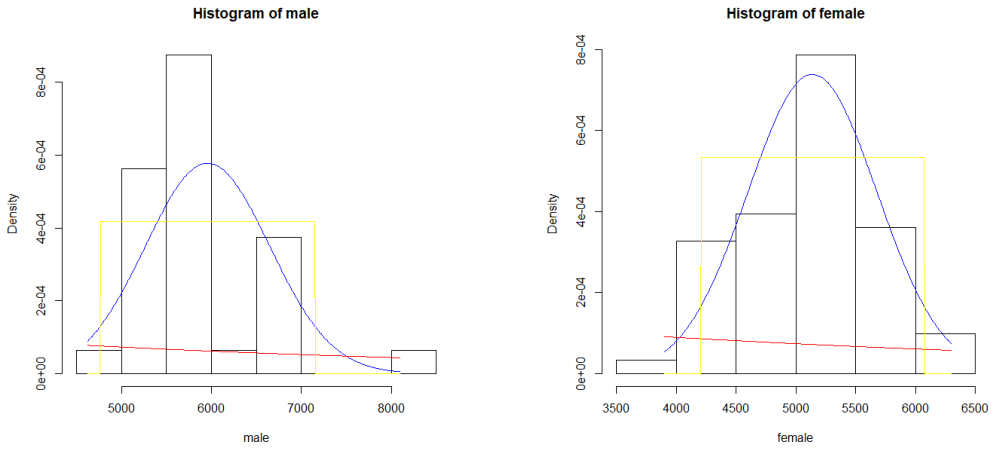
- Male
  - » hist(male, freq=FALSE)
    - Porovnávame distribuční funkce různých rozdělení na histogramu.
  - Normální rozdělení
    - » maleV<-seq(min(male),max(male),10)
      - vytvoříme si sekvenci hodnot od nejmenší po největší hodnoty
    - » maleNorm<-dnorm(maleV, mean = mean(male), sd = sd(male))
      - využijeme funkci dnorm na převod pro body normálního rozdělení
    - » lines(maleV,maleNorm, col="blue")
      - vykreslíme normální rozdělení na histogram
  - Exponenciální rozdělení
    - » lambdaMale<-1/mean(male)
      - vypočteme si parametr pro exponenciální rozdělení jako $\frac{1}{střední\ hodnota}$
    - » maleExp<-dexp(maleV, lambdaMale)
      - využijeme funkci dexp na vypočet bodu exponenciálního rozdělení
    - » lines(maleV,maleExp, col="red")

        - vykreslíme exponencionální rozdělení na histogram
    - Uniformní rozdělení
        - » aMale<-mean(male)-sqrt(3*var(male))
        - » bMale<-sqrt(3*var(male))+mean(male)
            - vypočteme si parametr ‘a‘ a ‘b‘ pro uniformní rozdělení podle vztahu k střední hodnote a rozptylu ze cvičení
        - » maleUnif<-dunif(maleV, aMale,bMale)
            - využijeme funkci dunif na výpočet bodu uniformního rozdělení
        - » lines(maleV,maleUnif, col=”yellow”)
            - vykreslíme uniformní rozdělení na histogram

- Female
    - » hist(female, freq=FALSE)
        - Porovnávame distribuční funkce různých rozdělení na histogramu.
    - Normálni rozdělení
        - » femaleV<-seq(min(female),max(female),10)
            - vytvoříme si sekvenci hodnot od nejmenší po největší hodnoty
        - » femaleNorm<-dnorm(femaleV, mean = mean(female), sd = sd(female))
            - využijeme funkci dnorm na převod pro body normálního rozdělení
        - » lines(femaleV,femaleNorm, col=”blue”)
            - vykreslíme normální rozdělení na histogram
    - Exponenciální rozdělení
        - » lambdaFemale<-1/mean(female)
            - vypočteme si parametr pro exponenciální rozdělení jako $\frac{1}{střední\ hodnota}$
        - » femaleExp<-dexp(femaleV, lambdaFemale)
            - využijeme funkci dexp na výpočet bodu exponenciálního rozdělení
        - » lines(femaleV,femaleExp, col=”red”)
            - vykreslíme exponencionální rozdělení na histogram
    - Uniformní rozdělení
        - » aFemale<-mean(female)-sqrt(3*var(female))
        - » bFemale<-sqrt(3*var(female))+mean(female)
            - vypočteme si parametr ‘a‘ a ‘b‘ pro uniformní rozdělení podle vztahu ke střední hodnotě a rozptylu ze cvičení
        - » femaleUnif<-dunif(femaleV, aFemale,bFemale)

        - využijeme funkci dunif na výpočet bodu uniformního rozdělení
    * » lines(femaleV,femaleUnif, col=”yellow”)
        - vykreslíme uniformní rozdělení na histogram

Výsledkem jsou následující grafy:



(a) Male (b) Female

Obrázek 2.4: Porovnání rozdělení

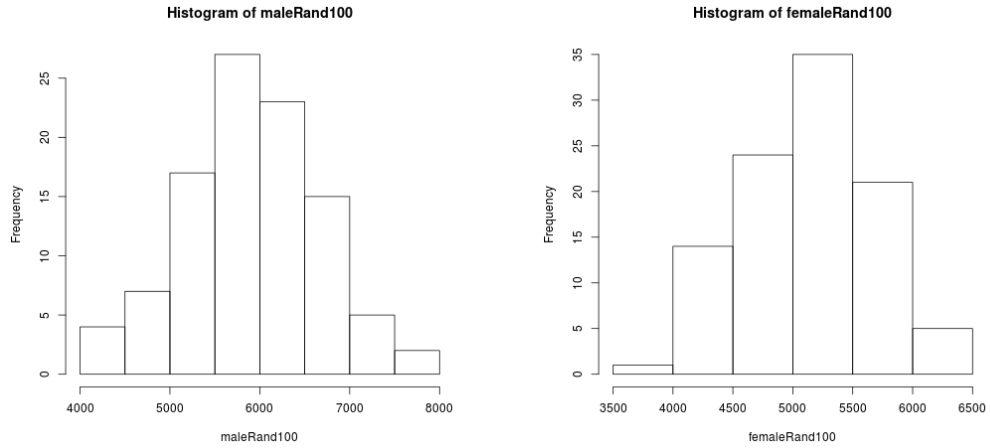
Došli jsme k závěru, že se u obou datasetu jedná o Normální rozdělení.

2.4 Úkol číslo 4

(1b) Pro každou skupinu zvlášť vygenerujte náhodný výběr o 100 hodnotách z rozdělení, které jste zvolili jako nejbližší, s parametry odhadnutými v předchozím bodě. Porovnejte histogram simulovaných hodnot s pozorovanými daty.

- Male
  - » maleRand100 = rnorm(100, mean(male), sd(male))
    * vybereme 100 náhodných hodnot použitím funkce rnorm
  - » hist(maleRand100)
    * vykreslíme z náhodně vybraných dat histogram
- Female
  - » femaleRand100 = rnorm(100, mean(female), sd(female))
    * vybereme 100 náhodných hodnot použitím funkce rnorm
  - » hist(femaleRand100)
    * vykrelslíme z náhodně vybraných dat histogram

Výsledkem jsou následujíci grafy:



(a) Male (b) Female

Obrázek 2.5: Vygenerované histogramy

Došli jsme k závěru, že zatímco vygenrovaný graf pro male se velmi liší od původního histogramu což je způsobeno malým množstvím dat. U female kde máme ≈ 2x více dat se histogramy velmi podobají i přes relativně malé množství dat.

2.5 Úkol číslo 5

(1b) Pro každou skupinu zvlášť spočítejte oboustranný 95% konfidenční interval pro střední hodnotu.

Ze zadaných dat nevíme přesně určit hodnotu rozptylu, proto musíme použit odhad intervalu kde se uplatňuje studentovo rozdelení. Když předpokladáme, že naše rozdelení je normálni (teda minimálne alespoň podobné normálnímu), výsledek bude spolehlivý i při menším počtu dat.

$$\left(\overline{X_n} - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}, \overline{X_n} + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}\right)$$

Spolehlivost intervalu je 95%, teda $1-\alpha = 0.95 \Rightarrow \alpha/2 = 0.025$

- Male
  - » *maleN = length(male)*
  - » *maleMean = mean(male)*
    - udeláme bodovej odhad pro střední hodnotu
  - » *maleSd = sd(male)*
    - udeláme bodovej odhad pro výběrovou směrodatní odchylku
  - » *maleDown = maleMean - qt(1-0.025,df=maleN-1)*maleSd/sqrt(maleN)*
  - » *maleUp = maleMean + qt(1-0.025,df=maleN-1)*maleSd/sqrt(maleN)*
    - funkce qt(...) očekáva ako první parameter $1-\alpha$ (mírny rozdíl od zápisu studentovho t-rozdělení v prednáškach)

(5707.839, 6205.911)

- Female
  - » *femaleN = length(female)*
  - » *femaleMean = mean(female)*
    - udeláme bodovej odhad pro střední hodnotu
  - » *femaleSd = sd(female)*
    - udeláme bodovej odhad pro výběrovou směrodatní odchylku
  - » *femaleDown = femaleMean - qt(1-0.025,df=femaleN-1)*femaleSd/sqrt(femaleN)*
  - » *femaleUp = femaleMean + qt(1-0.025,df=femaleN-1)*femaleSd/sqrt(femaleN)*
    - funkce qt(...) očekáva ako první parameter $1-\alpha$ (mírny rozdíl od zápisu studentovho t-rozdělení v prednáškach)

(5000.585, 5277.12)