

BI-PST Domácí úkol

Patrik Jantošovič

Tomáš Zvara

Tomáš Janecký

16. prosince 2018

1 PARAMETRY A DATOVÝ SOUBOR

Reprezentant: Patrik Jantošovič

$K = \text{den narození} = 16$

$L = \text{počet písmen v příjmení} = 10$

$M = ((K+L)*46) \bmod 11 + 1 = 2$

Výsledkem je tedy datový soubor: case0102, mzda dle pohlaví

1.1 VYTVOŘENÍ DATOVÉHO SOUBORU

Řešení úloh předpokládá úspěšnou instalaci knihovny 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)`
 - Načítání package Sleuth2
- » `write.table(case0102, "C:/data.csv", row.names=F, sep=";", dec=",")`

- Zápis dat do .csv souboru
- » `data<-read.table("C:/data.csv",header=TRUE,sep=";")`
 - Načteme data z připraveného souboru
- » `female<-data[1:61,1]`
 - Načítání dat pro pozorovanou skupinu: Female
- » `male<-data[62:93,1]`
 - Načítání dat pro pozorovanou skupinu: Male

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 každou skupinu zvlášť odhadněte střední hodnotu, rozptyl a medián příslušného rozdělení.

- » `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 následující tabulky:

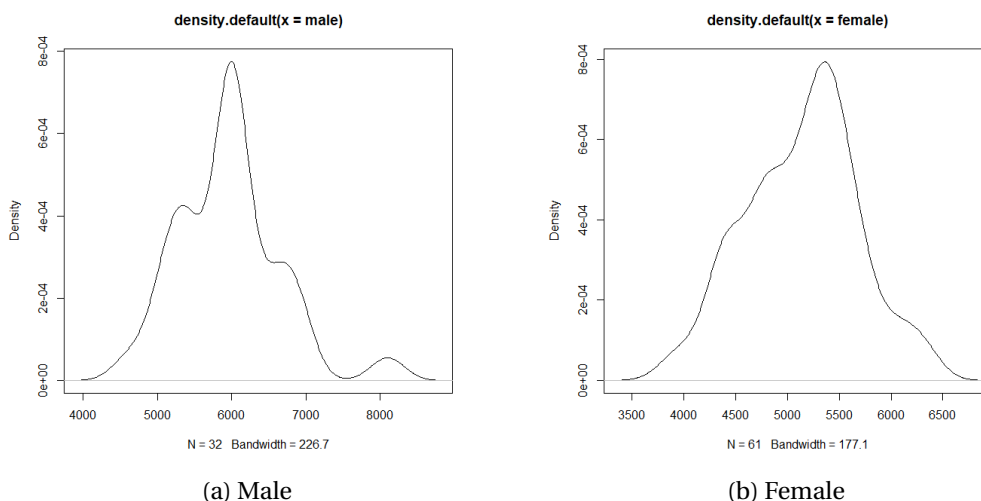
Pohlaví	Velikost dat	Střední hodnota	Rozptyl	Medián
Male	32	5956.875	477112.5	6000
Female	61	5138.852	291460.3	5220

2.2 ÚKOL ČÍSLO 2

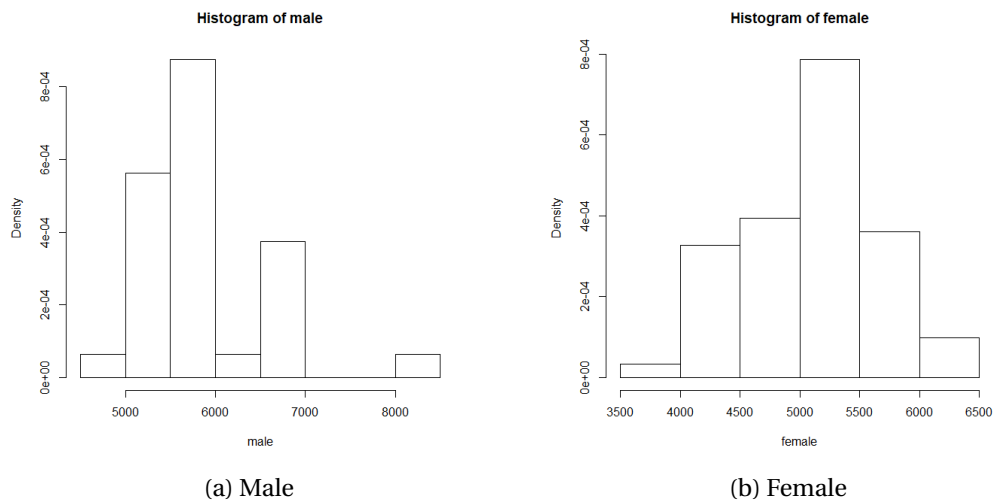
(1b) Pro každou skupinu zvlášť odhadněte hustotu a distribuční funkci pomocí histogramu a empirické distribuční funkce.

- » `hist(female, freq=FALSE)`
 - Vykreslení histogramu female. `freq=FALSE` používáme jako přepínač pro hustotu
- » `hist(male, freq=FALSE)`
 - Vykreslení histogramu female. `freq=FALSE` používáme 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

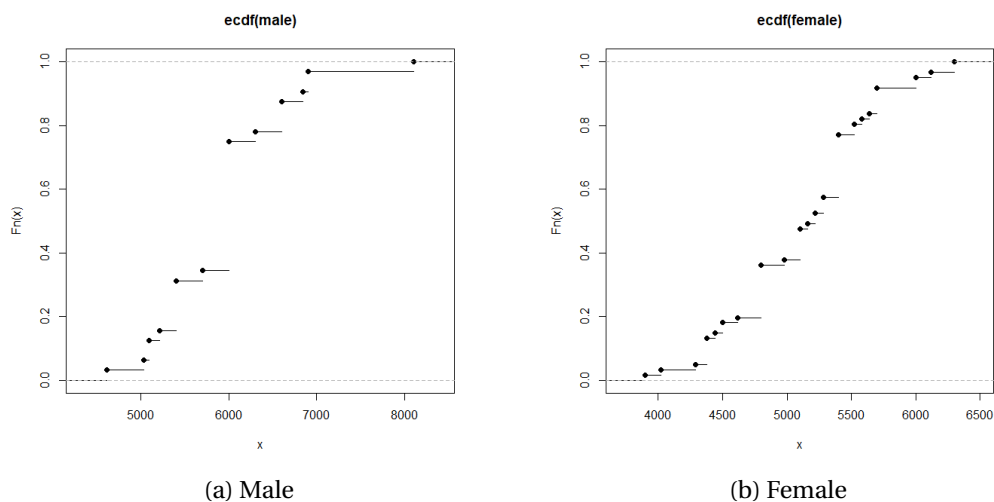
Výsledkem jsou grafy přiložené na následující stránce.



Obrázek 2.1: Hustota



Obrázek 2.2: Histogram



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)`
 - * Porovnáváme 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 $\frac{1}{\text{střední hodnota}}$
 - * » `maleExp<-dexp(maleV, lambdaMale)`
 - využijeme funkci `dexp` na výpočet bodu exponenciálního rozdělení
 - * » `lines(maleV,maleExp, col="red")`
 - vykreslíme exponenciální 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 hodnotě 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)`
 - * Porovnáváme distribuční funkce různých rozdělení na histogramu.
 - Normální 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

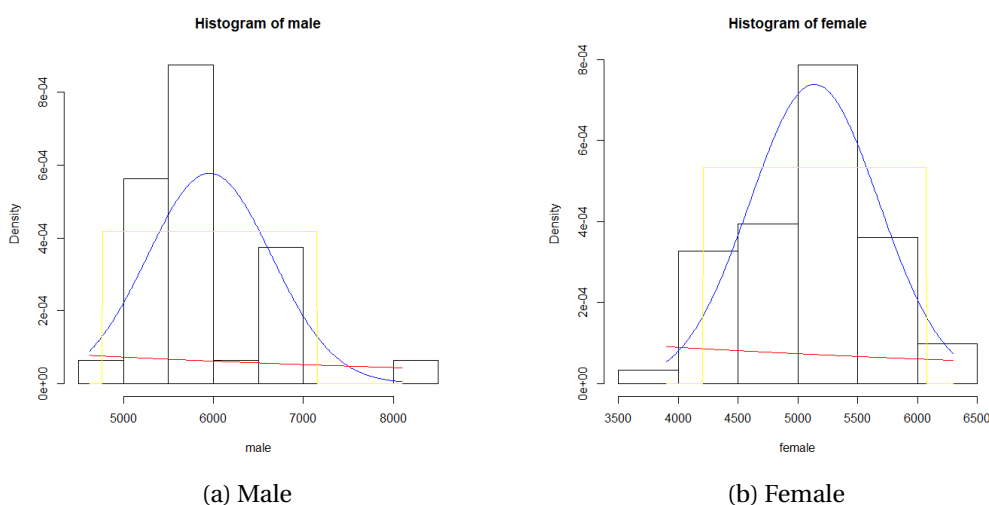
– Exponenciální rozdělení

- * » `lambdaFemale<-1/mean(female)`
 - vypočteme si parametr pro exponenciální rozdělení jako $\frac{1}{\text{střední hodnota}}$
- * » `femaleExp<-dexp(femaleV, lambdaFemale)`
 - využijeme funkci `dexp` na výpočet bodu exponenciálního rozdělení
- * » `lines(femaleV,femaleExp, col="red")`
 - vykreslíme exponenciální 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:



Obrázek 2.4: Porovnání rozdělení

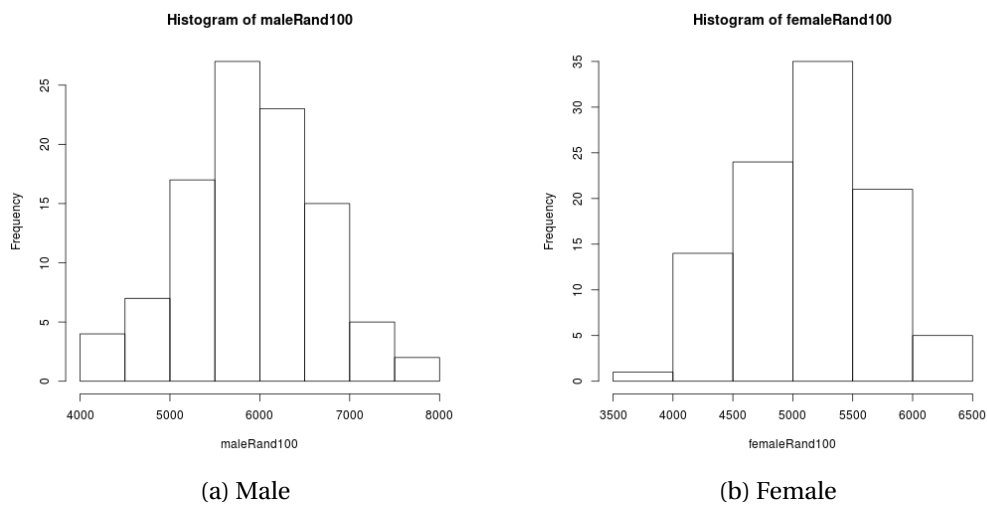
Došli jsme k závěru, že se u obou datasetů jedná o Normální rozdělení.

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)`
 - * vykreslíme z náhodně vybraných dat histogram

Výsledkem jsou následující grafy:



Obrázek 2.5: Vygenerované histogramy

Došli jsme k závěru, že zatímco vygenerovaný graf pro male se velmi liší od původního histogramu což je způsobeno malým množstvím dat. U female kde máme $\approx 2x$ 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 neumíme přesně určit hodnotu rozptylu, proto musíme použít odhad intervalu kde se uplatňuje Studentovo rozdělení. Když předpokládáme, že naše rozdělení je normální (tedy minimálně 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%, tedy $1 - \alpha = 0.95 \Rightarrow \alpha/2 = 0.025$

- Male

- » *maleN = length(male)*
- » *maleMean = mean(male)*
 - * uděláme bodový odhad pro střední hodnotu
- » *maleSd = sd(male)*
 - * uděláme bodový odhad pro výběrovou směrodatnou 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(...) očekává jako první parametr $1 - \alpha$ (mírný rozdíl od zápisu studentova t-rozdělení v přednáškách)

(5707.839, 6205.911)

- Female

- » *femaleN = length(female)*
- » *femaleMean = mean(female)*
 - * uděláme bodový odhad pro střední hodnotu
- » *femaleSd = sd(female)*
 - * uděláme bodový odhad pro výběrovou směrodatnou 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(...) očekává jako první parametr $1 - \alpha$ (mírný rozdíl od zápisu studentova t-rozdělení v přednáškách)

(5000.585, 5277.12)

2.6 ÚKOL ČÍSLO 6

(1b) Pro každou skupinu zvlášť otestujte na hladině významnosti 5% hypotézu, zda je střední hodnota rovná hodnotě K (parametr úlohy - 16), proti oboustranné alternativě. Můžete použít buď výsledek z předešlého bodu, nebo výstup z příslušné vestavěné funkce vašeho softwaru.

H_0 : střední hodnota je rovná 16

H_A : střední hodnota není rovná 16

- Male

» *t.test(male,mu=16,alternative="two.sided")*

data: male

t = 48.654, df = 31, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 16

95 percent confidence interval:

5707.839 6205.911

sample estimates:

mean of x

5956.875

- Male

» *t.test(female,mu=16,alternative="two.sided")*

data: female

t = 74.112, df = 60, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 16

95 percent confidence interval:

5000.585 5277.120

sample estimates:

mean of x

5138.852

Je zřejmé, že testovaná hodnota $K = 16$ v konfidenčním intervalu neleží. Na hladině spolehlivosti 5% zamítneme H_0 v prospěch H_A (platí pro obě sledované skupiny).

2.7 ÚKOL ČÍSLO 7

(2b) Na hladině spolehlivosti 5% otestujte, jestli mají pozorované skupiny stejnou střední hodnotu. Typ testu a alternativy stanovte tak, aby vaše volba nejlépe korespondovala s povahou zkoumaného problému.

» `t.test(female, mu=mean(male), alternative="two.sided")`

One Sample t-test

data: female

$t = -11.834$, $df = 60$, $p\text{-value} < 2.2e-16$

alternative hypothesis: true mean is not equal to 5956.875

95 percent confidence interval:

5000.585 5277.120

sample estimates:

mean of x

5138.852

» `t.test(male, mu=mean(female), alternative="two.sided")`

One Sample t-test

data: male

$t = 6.6993$, $df = 31$, $p\text{-value} = 1.707e-07$

alternative hypothesis: true mean is not equal to 5138.852

95 percent confidence interval:

5707.839 6205.911

sample estimates:

mean of x

5956.875

Na hladině spolehlivosti 5% (která je implicitní hodnotou *conf.level*) testujeme, zda střední hodnota male \in **(5000.585, 5277.120)** konfidenčního intervalu female a obráceně, zda střední hodnota female \in **(5707.839, 6205.911)** konfidenčního intervalu male. Vyšlo nám $p = 0.00000000000000022$ pro první a $p = 0.0000001707$ pro druhou hypotézu, tudíž zamítneme obě hypotézy ve prospěch alternativní hypotézy (nerovnájí se).