Domácí úkol vypracovali:

Jméno	*ČVUT-ID*
Filip Leško	**leskofil**
František Sciranka	**scirafra**
Roman Isaev	**isaevrom**

- K = den narození reprezentanta skupiny (1-31),
- L = počet písmen v příjmení reprezentanta,
- M = ((K+L)*47)mod(11)+1.

Reprezentantem skupiny je **František Sciranka**.

K=18;

L=8;

M=(26 * 47)mod(11) + 1
= 1 + 1
= **2**.

1. Načtěte datový soubor a rozdělte sledovanou proměnnou na příslušné dvě pozorované skupiny. Stručně popište data a zkoumaný problém. Pro každou skupinu zvlášť odhadněte střední hodnotu, rozptyl a medián příslušného rozdělení.

- M nám vyšlo rovné 2, proto používáme dataset **case0102** z knihovny **Sleuth2**, jehož data jsou věnována tématu "Sex Discrimination in Employment".
- Načteme si samostatně data o platech jak pro muže, tak i pro ženy.
  Jako filtr použijeme příznak pohlaví.
- Poté nad těmito daty provedeme pozorování, z nichž plyne, že, muži vydělávají více, resp. že jejich průměrné i očekávané platové ohodnocení je vyšší, než u žen.
  Nicméně jsou jejich platy více variabilní (~1.64-krát), než konzistentnější platy žen.

```
library(Sleuth2)
male <- case0102$Salary[case0102$Sex=='Male']
female <- case0102$Salary[case0102$Sex=='Female']
Hodnoty <- c("Stredná hodnota", "Rozptyl", "Medián")
Muži <- c(mean(male), var(male), median(male))
Ženy <- c(mean(female), var(female), median(female))
descDf = data.frame(Hodnoty, Muži, Ženy)
descDf
```

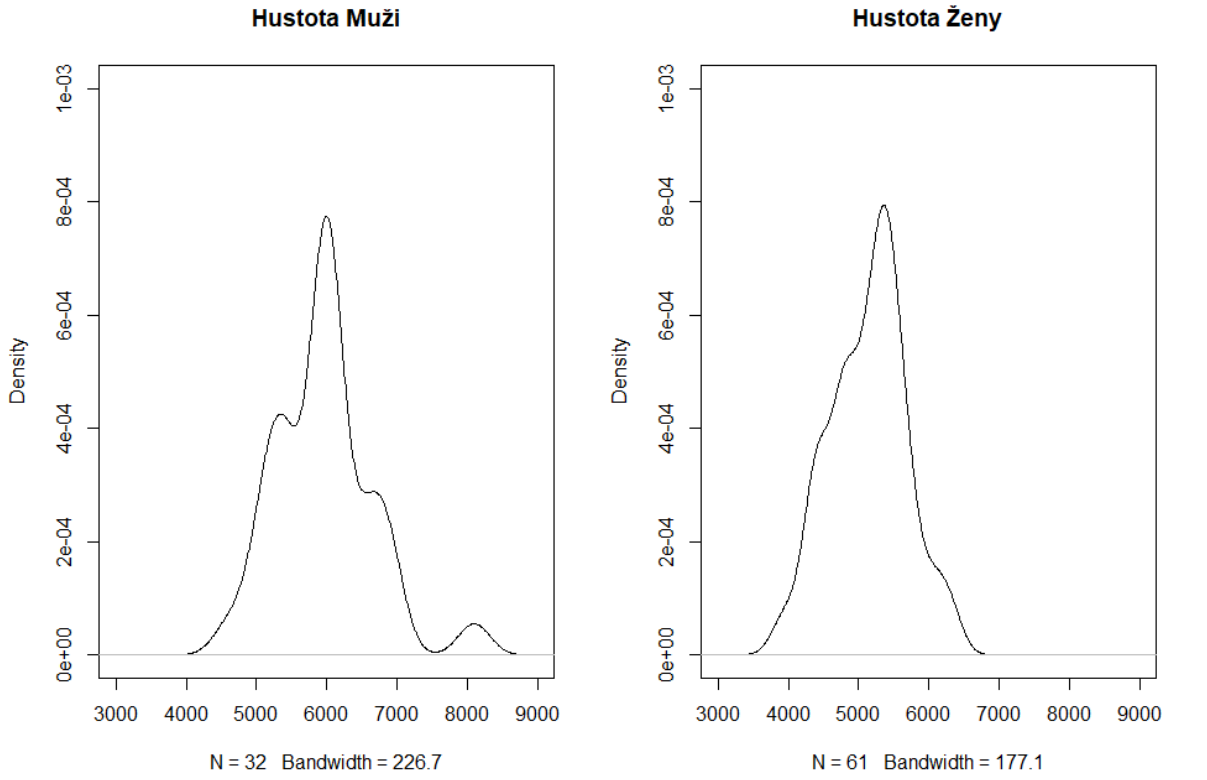
```
> Štatistika <- c("Stredná hodnota", "Rozptyl", "Medián")
> Muži <- c(mean(male), var(male), median(male))
> Ženy <- c(mean(female), var(female), median(female))
> descDf = data.frame(Štatistika, Muži, Ženy)
> descDf
       Štatistika       Muži       Ženy
1 Stredná hodnota   5956.875   5138.852
2         Rozptyl 477112.500 291460.328
3          Medián   6000.000   5220.000
> Hodnoty <- c("Stredná hodnota", "Rozptyl", "Medián")
> Muži <- c(mean(male), var(male), median(male))
> Ženy <- c(mean(female), var(female), median(female))
> descDf = data.frame(Hodnoty, Muži, Ženy)
> descDf
         Hodnoty       Muži       Ženy
1 Stredná hodnota   5956.875   5138.852
2         Rozptyl 477112.500 291460.328
3          Medián   6000.000   5220.000
```

2. Pro každou skupinu zvlášť odhadněte hustotu a distribuční funkci pomocí histogramu a empirické distribuční funkce.

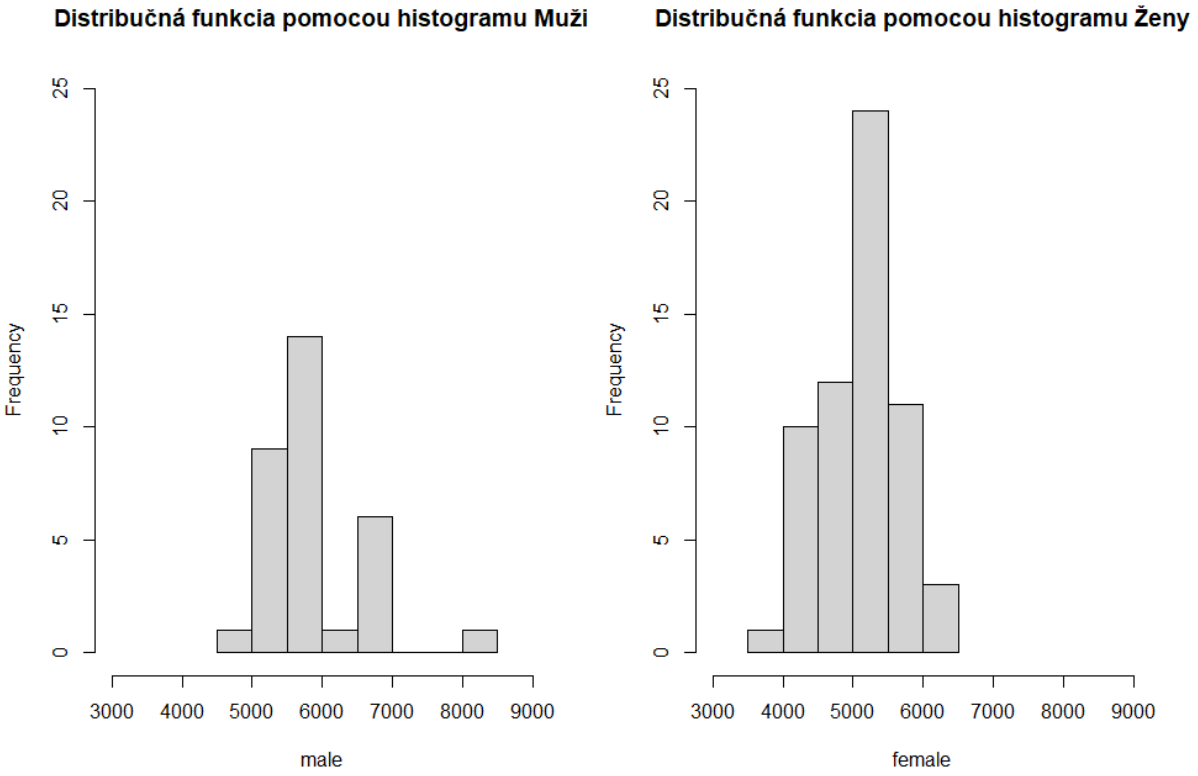
```
par(mfrow=c(1,2))
```

- Funkce **par** nám umožní vykreslit si vedle sebe dva grafy - pro muže i pro ženy.

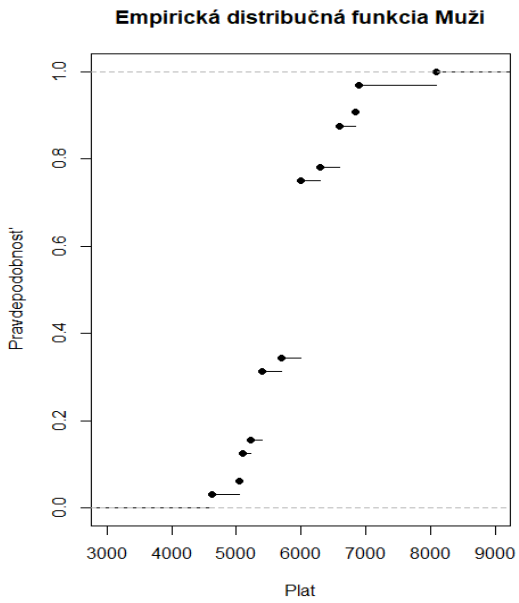
```
plot(density(male),    main="Hustota Muži", xlim=c(3000, 9000), ylim=c(0, 0.001))
plot(density(female), main="Hustota Ženy", xlim=c(3000, 9000), ylim=c(0, 0.001))
```

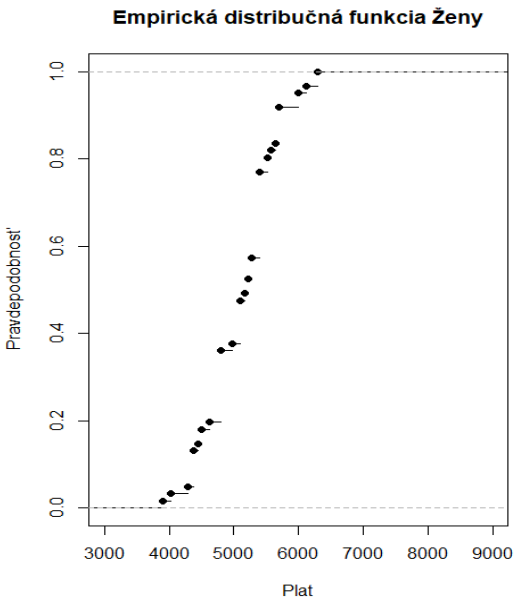


```
hist(male, breaks=8,  main="Distribučná funkcia pomocou histogramu Muži", xlim=c(3000, 9000), ylim=c(0, 25))
hist(female, breaks=8, main="Distribučná funkcia pomocou histogramu Ženy", xlim=c(3000, 9000), ylim=c(0, 25))
```



```
plot(ecdf(male),    main="Empirická distribučná funkcia Muži", xlab="Plat",
ylab="Pravdepodobnosť", xlim=c(3000, 9000), ylim=c(0, 1))

plot(ecdf(female), main="Empirická distribučná funkcia Ženy", xlab="Plat",
ylab="Pravdepodobnosť", xlim=c(3000, 9000), ylim=c(0, 1))
```





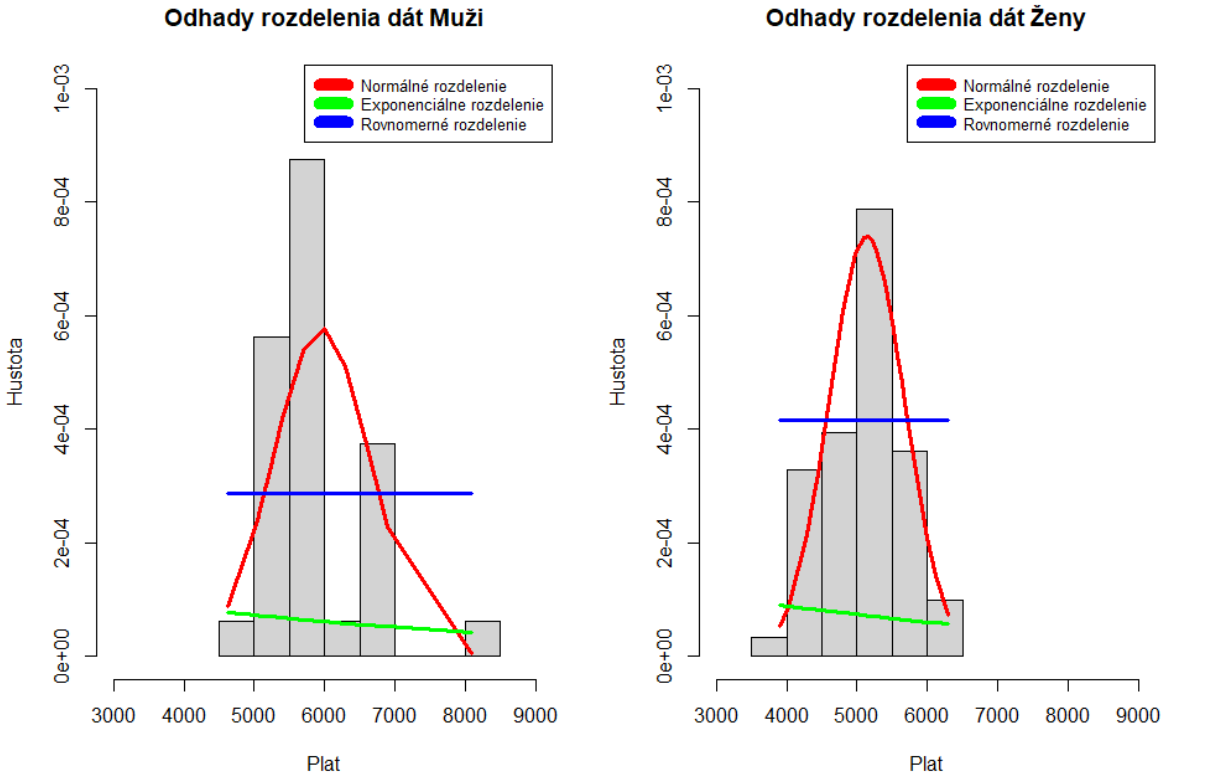
3. Pro každou skupinu zvlášť najděte nejbližší rozdělení: Odhadněte parametry normálního, exponenciálního a rovnoměrného rozdělení. Zaneste příslušné hustoty s odhadnutými parametry do grafů histogramu. Diskutujte, které z rozdělení odpovídá pozorovaným datům nejlépe.

- Provedeme odhadnutí parametrů a posléze si vykreslíme histogramy, ze kterých posoudíme o tom, které rozdělení se nejvíce podobá platům.

```
hist(male, prob=T, main="Odhady rozdelenia dát Muži", xlab="Plat", ylab="Hustota", breaks=8, xlim=c(3000, 9000), ylim=c(0, 0.001))

lines(sort(male), dnorm(sort(male), mean=mean(male), sd=sd(male)), col="red", lwd=3)

lines(sort(male), (dexp(sort(male), (1 / mean(male)))), col="green", lwd=3)

lines(sort(male), dunif(sort(male), min(male), max(male)), col="blue", lwd=3)

legend("topright", c("Normálné rozdelenie", "Exponenciálne rozdelenie", "Rovnomerné rozdelenie"), col=c("red", "green", "blue"), lwd=10, cex = 0.8)
```

```
hist(female, prob=T, main="Odhady rozdelenia dát Ženy", xlab="Plat", ylab="Hustota", breaks=8, xlim=c(3000, 9000), ylim=c(0, 0.001))

lines(sort(female), dnorm(sort(female), mean=mean(female), sd=sd(female)), col="red", lwd=3)

lines(sort(female), (dexp(sort(female), (1 / mean(female)))), col="green", lwd=3)

lines(sort(female), dunif(sort(female), min(female), max(female)), col="blue", lwd=3)

legend("topright",  c("Normálné rozdelenie", "Exponenciálne rozdelenie", "Rovnomerné rozdelenie"), col=c("red", "green", "blue"), lwd=10, cex = 0.8)
```

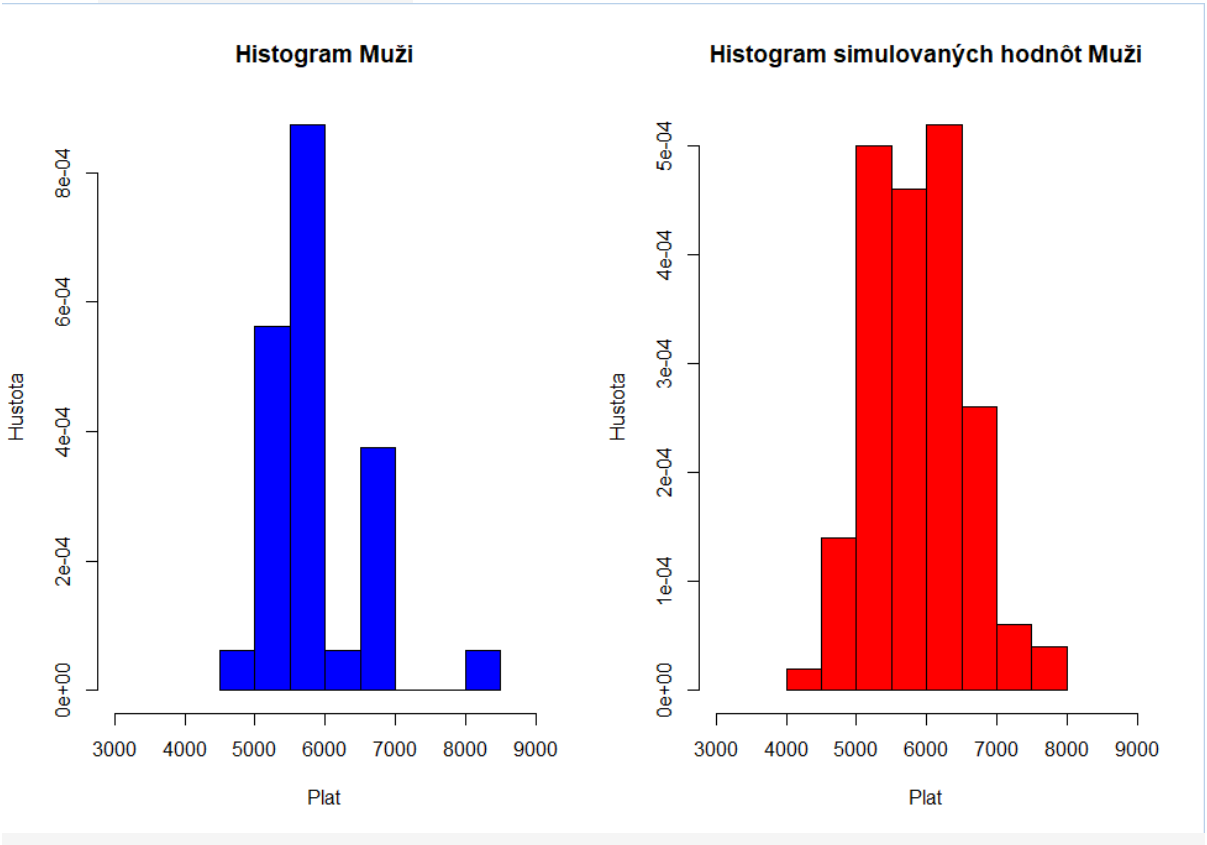
- Je patrné, že jak u platů mužů, tak i u platů žen rozdělení platů nejvíce odpovídá odhad typu “Normální rozdělení”, neboť jak u žen, tak i u mužů sledujeme výskyt nejčastějšího platu uprostřed.



4. 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.

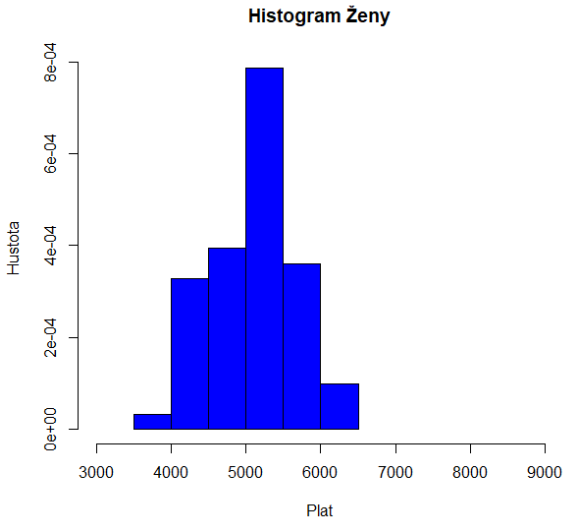
```
set.seed(18)
hist(male, prob=T, main="Histogram Muži", xlab="Plat", ylab="Hustota", col="blue", xlim=c(3000, 9000))
hist(rnorm(100, mean=mean(male), sd=sd(male)), prob=T, main="Histogram simulovaných hodnôt Muži", xlab="Plat", ylab="Hustota", col="red", xlim=c(3000, 9000))
```

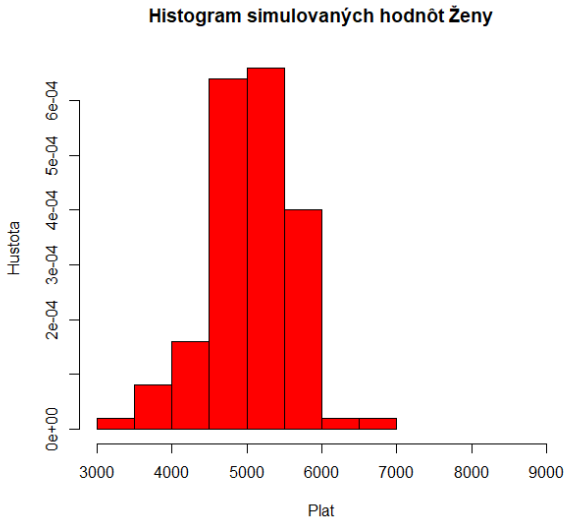
- Histogram vygenerovaných hodnot z rovnoměrného rozdělení zhruba odpovídá rozložení skutečných hodnot, přičemž některé mezery v datech jsou vyplněny vzhledem k rovnoměrnému generování.
- S jistou mírou nadsázky se při porovnání histogramů dá říct, že simulované rozdělení platů je podobné tomu reálnému. Na první pohled je vidět, že rozložení reálných dat na ose platů je posunutější doprava vůči simulovaným grafům, které začínají na 4000. Nicméně, co je důležité, je to, že tendence rozložení dat je zachována. Selsky řečeno: velké písmenko ‘A’, nakreslené na každý z těchto histogramů, bude mít v obou případech zhruba stejnou formu. Můžeme si tedy zatleskat.



```
hist(female, prob=T, main="Histogram Ženy", xlab="Plat", ylab="Hustota", col="blue", xlim=c(3000, 9000))

hist(rnorm(100, mean=mean(female), sd=sd(female)), prob=T, main="Histogram simulovaných hodnôt Ženy", xlab="Plat", ylab="Hustota", col="red", xlim=c(3000, 9000))
```

- Rovněž jako v případě mužů i simulovaný histogram platů žen má širší rozložení na ose platů. Rozsah se zvětšil do obou stran o 500. Tendence rozložení dat není tak ideální, jako v případě můžu, nicméně se určitá podoba dá nalézt - např. zachování hustoty v místě nejčastějších platů a hrubá aproximace toho, jak data “rostou” a “klesají”.





5. Pro každou skupinu zvlášť spočítejte oboustranný 95% konfidenční interval pro střední hodnotu.

```
sprintf("Obojstranný konfidenčný interval pre strednú hodnotu Muži: (%f ; %f)", (mean(male) - qt(0.975, df=length(male) - 1) * sd(male) / sqrt(length(male))), (mean(male) + qt(0.975, df=length(male) - 1) * sd(male) / sqrt(length(male))))
```

```
"Obojstranný konfidenčný interval pre strednú hodnotu Muži: (5707.839087 ; 6205.910913)"
```

```
sprintf("Obojstranný konfidenčný interval pre strednú hodnotu  Ženy: (%f ; %f)", (mean(female) - qt(0.975, df=length(female) - 1) * sd(female) / sqrt(length(female))), (mean(female) + qt(0.975, df=length(female) - 1) * sd(female) / sqrt(length(female))))
```

```
"Obojstranný konfidenčný interval pre strednú hodnotu  Ženy: (5000.585163 ; 5277.119755)"
```

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

K := 18

α := 0.05

t.test(male, alternative = "two.sided", mu = 18)

```
data:  male
t = 48.637, df = 31, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 18
95 percent confidence interval:
 5707.839 6205.911
sample estimates:
mean of x
 5956.875
```

t.test(female, alternative = "two.sided", mu = 18)

```
data:  female
t = 74.083, df = 60, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 18
95 percent confidence interval:
 5000.585 5277.120
sample estimates:
mean of x
 5138.852
```

H0: **“**platí, že μ **=** K”

HA: “platí, že μ **není rovno** K”

- Jak u mužů, tak i u žen platí to, že $\mu = 18 = K$ neleží v intervalu konfidence.
- V obou případech tedy zamítáme hypotézu **H0**: “platí μ = K” ve prospěch alternativní hypotézy, kterou je “**HA**: platí, že μ není rovno K”.

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

- **H0:** "Střední hodnota platů mužů rovná střední hodnotě platů žen."
- Alternativní hypotéza **HA** tedy je: "Střední hodnota platů mužů a střední hodnota platů žen se liší."
- **α** := 0.05
- **1 - α** = 0.95

```
t.test(male, female, alternative="two.sided", conf.level=0.95)
```

```
data:  male and female
t = 5.83, df = 51.329, p-value = 3.71e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  536.3758 1099.6693
sample estimates:
mean of x mean of y
 5956.875  5138.852
```

- Z přednášky víme o p-hodnotě následující věci: *"Čím je p-hodnota menší, tím významnější je zamítnutí **H0.**"* a *"Je-li p-hodnota menší, než naše požadovaná hladina významnosti **α**, zamítáme **H0.**"*
- **p-hodnota=0.000000371** je patrně menší, než naše hladina významnosti **α=0.05**.
- **H0** tedy zamítáme ve prospěch alternativní **HA**: "Střední hodnota platů mužů a střední hodnota platů žen se liší."