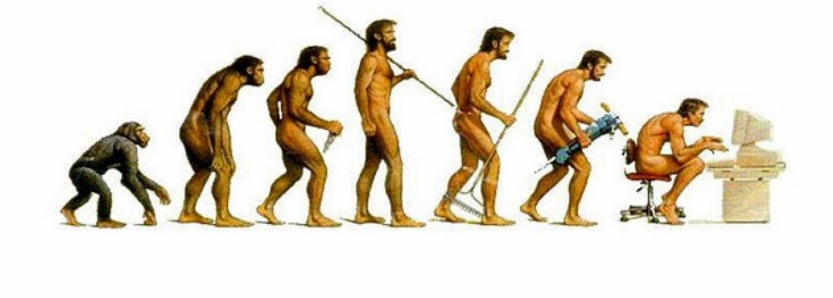
Masarykova univerzita v Brně
Přírodovědecká fakulta

SBÍRKA PŘÍKLADŮ K PŘEDMĚTU APLIKOVANÁ STATISTIKA I

Marie Budíková, Veronika Bendová



Brno, 2016

1 Bodové a intervalové rozdělení četností

Přehled použitých funkcí: read.delim, source, head, names, factor, data.frame, sum, cumsum, row.names, variacni_rada, barplot, abline, plot, lines, points, axis, table, cbind, rbind, prop.table, dim, round, range, min, max, hist, paste, text.

Bodové rozdělení četností

Bodové rozdělení četností procvičíme pomocí datového souboru znamky.txt, který obsahuje údaje o známkách z matematiky, angličtiny a pohlaví 20 studentů 1. ročníku.

Příklad 1.1. Načtěte soubor znamky.txt. Znakům X, Y, Z vytvořte návěští (X - známka z matematiky, Y - známka z angličtiny, Z - pohlaví studenta). Popište, co znamenají jednotlivé varianty (u znaků X a Y: 1 - výborně, 2 - chvalitebně, 3 - dobře, 4 - neprospěl, u znaku Z: 0 - žena, 1 - muž).

```
setwd("C:/Disk D/ND-Skola/02-Vyuka/04-Aplikovana statistika 2016/Sbirka")
source('AS-funkce.R')
data <- read.table('znamky.txt', sep='\t', dec='.')
head(data)
```

```
##   V1 V2 V3
## 1  2  2  0
## 2  1  3  1
## 3  4  3  1
## 4  1  1  0
## 5  1  2  1
## 6  4  4  1
```

```
names(data) <- c('matematika', 'anglictina', 'pohlavi')
head(data)
```

```
##   matematika anglictina pohlavi
## 1          2          2       0
## 2          1          3       1
## 3          4          3       1
## 4          1          1       0
## 5          1          2       1
## 6          4          4       1
```

```
f1    <- factor(data$matematika, levels=c(1,2,3,4),
            labels=c('vyborne','chvalitebne','dobre','nedostatecne'))
f2    <- factor(data$anglictina, levels=c(1,2,3,4),
            labels=c('vyborne','chvalitebne','dobre','nedostatecne'))
f3    <- factor(data$pohlavi, levels=c(0,1), labels=c('zena','muz'))
data2 <- data.frame(f1, f2, f3)
names(data2) = c('matematika','anglictina','pohlavi')
head(data2)
```

```
##     matematika   anglictina pohlavi
## 1  chvalitebne  chvalitebne    zena
## 2      vyborne        dobre     muz
## 3 nedostatecne        dobre     muz
## 4      vyborne      vyborne    zena
## 5      vyborne  chvalitebne     muz
## 6 nedostatecne nedostatecne     muz
```

Příklad 1.2. Vytvořte

a) variační řadu známek z matematiky a známek z angličtiny;

b) sloupkový diagram absolutních četností znaků X=Matematika a Y=Angličtina;

c) polygon absolutních četností znaků X=Matematika a Y=Angličtina.

a) Variační řada známek z matematiky

```
matematika <- data2$matematika
n1 <- sum(matematika=='vyborne')
n2 <- sum(matematika=='chvalitebne')
n3 <- sum(matematika=='dobre')
n4 <- sum(matematika=='nedostatecne')

nj <- c(n1,n2,n3,n4)
n  <- sum(nj)
pj <- nj/n
Nj <- cumsum(nj)
Fj <- cumsum(pj)

variacni.rada <- data.frame(nj=nj, Nj=Nj, pj=pj, Fj=Fj)
row.names(variacni.rada) <- c('vyborne', 'chvalitebne', 'dobre', 'nedostatecne')
variacni.rada
```

```
##              nj Nj   pj   Fj
## vyborne       7  7 0.35 0.35
## chvalitebne   3 10 0.15 0.50
## dobre         2 12 0.10 0.60
## nedostatecne  8 20 0.40 1.00
```

Variační řadu můžeme také získat použitím funkce variacni_rada(X, nazvy), která je naprogramovaná ve skriptu AS-funkce.

```
(VR.Mat <- variacni_rada(X=matematika,
                         nazvy=c('vyborne', 'chvalitebne', 'dobre', 'nedostatecne')))
```

```
##              nj Nj   pj   Fj
## vyborne       7  7 0.35 0.35
## chvalitebne   3 10 0.15 0.50
## dobre         2 12 0.10 0.60
## nedostatecne  8 20 0.40 1.00
```

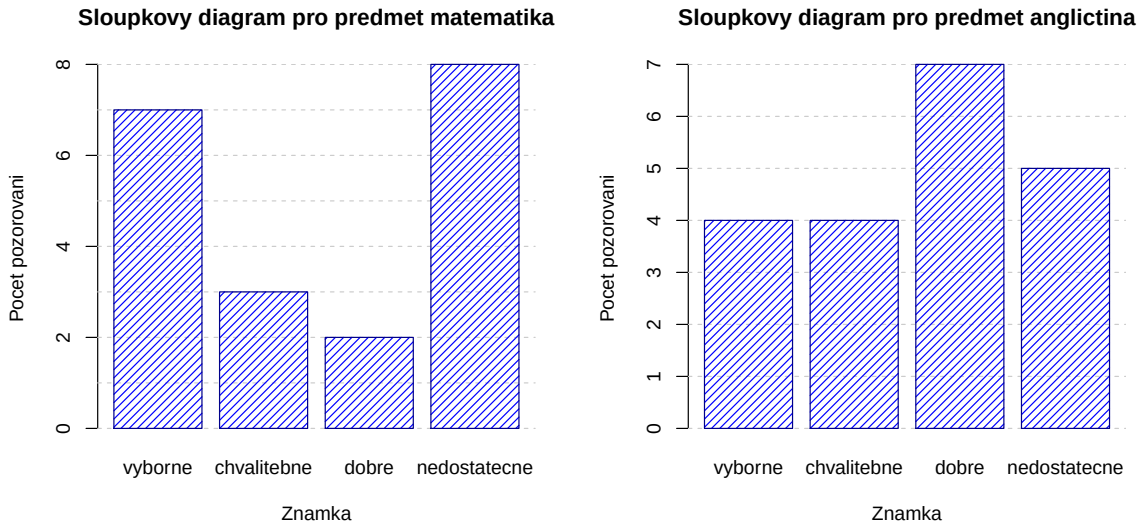
Variační řada známek z angličtiny

```
anglictina <- data2$anglictina
(VR.Ang <- variacni_rada(X=anglictina,
                         nazvy=c('vyborne', 'chvalitebne', 'dobre', 'nedostatecne')))
```

```
##              nj Nj   pj   Fj
## vyborne       4  4 0.20 0.20
## chvalitebne   4  8 0.20 0.40
## dobre         7 15 0.35 0.75
## nedostatecne  5 20 0.25 1.00
```

b) Sloupkový diagram absolutních četností znaků X=Matematika a Y=Angličtina

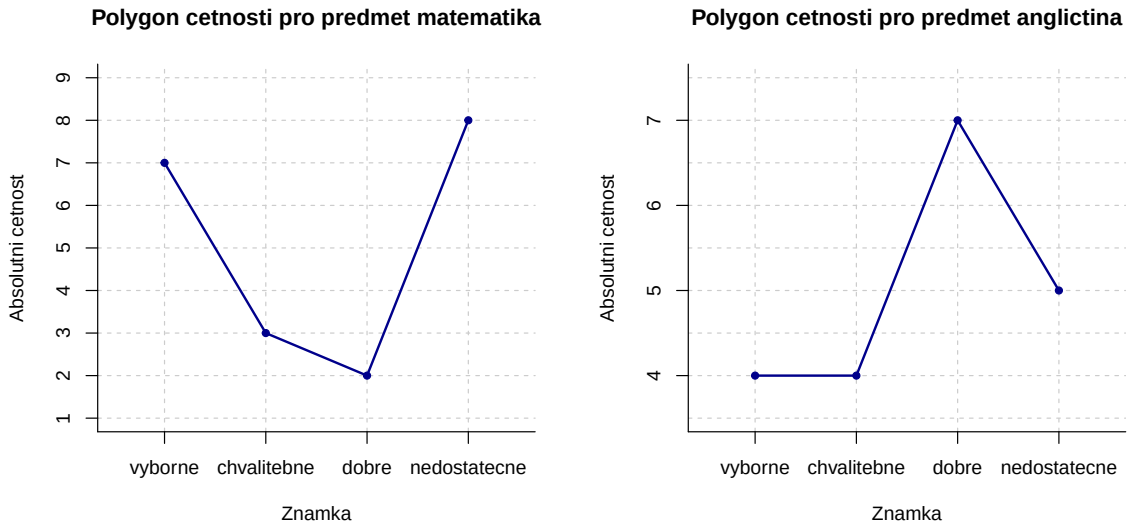
```
nazvy.znamek <- c('vyborne', 'chvalitebne', 'dobre', 'nedostatecne')

# Matematika
barplot(VR.Mat$nj, col='white', border='white', axes=T,
        xlab='Znamka', ylab='Pocet pozorovani', names=nazvy.znamek,
        main='Sloupkovy diagram pro predmet matematika')
abline(h=0:9, col='grey80', lty=2)
barplot(VR.Mat$nj, col='blue', axes=F, density=20, border='darkblue', add=T,
        names=F)

#Anglictina
barplot(VR.Ang$nj, col='white', border='white', axes=T,
        xlab='Znamka', ylab='Pocet pozorovani', names=nazvy.znamek,
        main='Sloupkovy diagram pro predmet anglictina')
abline(h=0:9, col='grey80', lty=2)
barplot(VR.Ang$nj, col='blue', axes=F, density=20, border='darkblue', add=T,
        names=F)
```



c) Polygon četností

```
# Matematika
plot(1:4, VR.Mat$nj, type='n', xlim=c(0.5,4.5), ylim=c(1,9),
     xlab='Znamka',ylab='Absolutni cetnost',
     main='Polygon cetnosti pro predmet matematika', axes=F)
abline(h=0:9, col='grey80', lty=2)
abline(v=0:9, col='grey80', lty=2)

lines(1:4, VR.Mat$nj, col='darkblue', lwd=2 )
points(1:4, VR.Mat$nj,col='darkblue',pch=20, cex=1.2)
axis(1, at=0:5, lab=c('',nazvy.znamek,''))
axis(2, at=0:10)

# Anglictina
```

```
plot(1:4, VR.Ang$nj, type='n', xlim=c(0.5,4.5), ylim=c(3.5,7.5),
     xlab='Znamka',ylab='Absolutni cetnost',
     main='Polygon cetnosti pro predmet anglictina', axes=F)
abline(h=seq(3.5, 7.5, by=0.5), col='grey80', lty=2)
abline(v=0:9, col='grey80', lty=2)

lines(1:4, VR.Ang$nj, col='darkblue', lwd=2 )
points(1:4, VR.Ang$nj,col='darkblue',pch=20, cex=1.2)
axis(1, at=0:5, lab=c('',nazvy.znamek,''))
axis(2, at=0:10)
```



Příklad 1.3. Vytvořte variační řady známek z matematiky a angličtiny pouze

a) pro ženy,

b) pro muže.

a) Variační řada známek z matematiky pro ženy

```
pohlavi<-data2$pohlavi
variacni_rada(X=matematika[pohlavi=='zena'], nazvy=nazvy.znamek)

##              nj Nj  pj  Fj
## vyborne       5  5 0.5 0.5
## chvalitebne   2  7 0.2 0.7
## dobre         1  8 0.1 0.8
## nedostatecne  2 10 0.2 1.0
```

Variační řada známek z angličtiny pro ženy

```
variacni_rada(X=anglictina[pohlavi=='zena'], nazvy=nazvy.znamek)

##              nj Nj  pj  Fj
## vyborne       4  4 0.4 0.4
## chvalitebne   2  6 0.2 0.6
## dobre         1  7 0.1 0.7
## nedostatecne  3 10 0.3 1.0
```

b) Variační řada známek z matematiky pro muže

```
variacni_rada(X=matematika[pohlavi=='muz'], nazvy=nazvy.znamek)

##              nj Nj  pj  Fj
## vyborne       2  2 0.2 0.2
## chvalitebne   1  3 0.1 0.3
## dobre         1  4 0.1 0.4
## nedostatecne  6 10 0.6 1.0
```

Variační řada známek z angličtiny pro muže

```
variacni_rada(X=anglictina[pohlavi=='muz'], nazvy=nazvy.znamek)

##              nj Nj  pj  Fj
## vyborne       0  0 0.0 0.0
## chvalitebne   2  2 0.2 0.2
## dobre         6  8 0.6 0.8
## nedostatecne  2 10 0.2 1.0
```

Příklad 1.4. Nadále budeme pracovat s celým datovým souborem. Vytvoříme kontingenční tabulku simultánních absolutních četností znaků X a Y.

```
K.Tab   <- table(matematika, anglictina)
K.Tab2  <- cbind(K.Tab, suma=apply(K.Tab, 1, sum))
(K.Tab3 <- rbind(K.Tab2, suma=apply(K.Tab2, 2, sum)))

##              vyborne chvalitebne dobre nedostatecne suma
## vyborne            4           1     2            0    7
## chvalitebne        0           2     1            0    3
## dobre              0           0     1            1    2
## nedostatecne       0           1     3            4    8
## suma               4           4     7            5   20
```

Vidíme, že ve výběrovém souboru byli 4 studenti, kteří měli z obou předmětů "výborně", jeden student, který měl z matematiky "výborně"a z angličtiny "chvalitebně"atd. až 4 studenti, kteří z obou předmětů neprospěli.

Příklad 1.5. Vytvořte kontingenční tabulku řádkově a sloupcově podmíněných relativních četností znaků X=Matematika a Y=Angličtina.

```
Tab <- table(matematika, anglictina)

# Radkove podminene relativni cetnosti
round(prop.table(Tab, margin=1), digits=3)

##               anglictina
## matematika     vyborne chvalitebne dobre nedostatecne
##   vyborne        0.571       0.143 0.286        0.000
##   chvalitebne    0.000       0.667 0.333        0.000
##   dobre          0.000       0.000 0.500        0.500
##   nedostatecne   0.000       0.125 0.375        0.500
```

Interpretace např. 2. sloupce ve 4. řádku: V souboru bylo 8 studentů, kteří neprospěli z matematiky. Mezi nimi byl jeden, který měl chvalitebně z angličtiny, což představuje $1/8 = 12.5\%$.

```r
# Sloupcove podminene relativni cetnosti
round(prop.table(Tab, margin=2), digits=3)
```

```
##               anglictina
## matematika     vyborne chvalitebne dobre nedostatecne
##   vyborne        1.000       0.250 0.286        0.000
##   chvalitebne    0.000       0.500 0.143        0.000
##   dobre          0.000       0.000 0.143        0.200
##   nedostatecne   0.000       0.250 0.429        0.800
```

Interpretace např. 4. řádku ve 2. sloupci: V souboru byli 4 studenti, kteří měli chvalitebně z angličtiny. Mezi nimi byl jeden, který neprospěl z matematiky, což představuje $1/4 = 25\%$.

Intervalové rozdělení četností

Práci s intervalovým rozdělením četností si ukážeme na datovém souboru lebky.txt.

Popis datového souboru: Máme k dispozici údaje o rozměrech lebek staroegyptské populace. Jedná se o 216 mužů a 109 žen. Znak X ... největší délka mozkovny v mm (tj. přímá vzdálenost kraniometrických bodů glabella a opisthocranion) Znak Y ... největší šířka mozkovny v mm (tj. přímá vzdálenost kraniometrických bodů euryon dx a euryon sin) Znak Z ... pohlaví osoby (1–muž, 0–žena)

Příklad 1.6. Načtěte soubor lebky.txt. Podle Sturgersova pravidla najděte optimální počet třídicích intervalů pro znaky X a Y a vhodně stanovte meze třídicích intervalů, a to zvlášť pro muže a zvlášť pro ženy.

```r
data        <- read.delim('lebky.txt', sep='\t', dec='.', header=F)
names(data) <- c('delka', 'sirka', 'pohlavi')
head(data)
```

```
##   delka sirka pohlavi
## 1   188   145     muz
## 2   172   139     muz
## 3   176   138     muz
## 4   184   128     muz
## 5   183   139     muz
## 6   177   143     muz
```

```r
# Muzi
data.M      <- data[data$pohlavi=='muz',]
n.M         <- dim(data.M)[1]
(Sturges.M  <- round(1+3.3*log10(n.M), digits=0))
```

```
## [1] 9
```

Protože mužů je 216, podle Sturgersova pravidla je optimální počet třídicích intervalů 9. Musíme zjistit minimum a maximum, abychom vhodně stanovili meze třídicích intervalů:

```r
# Znak X = Delka lebky
delka.M     <- data.M$delka
range(delka.M)
```

```
## [1] 164 199
```

```r
max(delka.M) - min(delka.M)
```

```
## [1] 35
```

```
round((max(delka.M) - min(delka.M))/Sturges.M, digits=0)

## [1] 4
```

Pro znak X = Délka lebky je minimum 164 a maximum 199, rozsah těchto hodnot je 35 a ideální délka jednoho třídicího intervalu vyšla jako $\frac{199-164}{9} \approx 4$. Jeví se vhodné dolní mez prvního třídicího intervalu zvolit 163, horní mez posledního třídicího intervalu 199. Celkem třídicí intervaly pro znak X budou: $\langle 163, 167\rangle$, $\langle 167, 171\rangle$, ..., $\langle 195, 199\rangle$.

```
# Znak Y = Sirka lebky
sirka.M     <- data.M$sirka
range(sirka.M)

## [1] 124 149

max(sirka.M) - min(sirka.M)

## [1] 25

round((max(sirka.M) - min(sirka.M))/Sturges.M, digits=0)

## [1] 3
```

Pro znak Y = šířka lebky je minimum 124 a maximum 149, rozsah těchto hodnot je 25 a ideální délka jednoho třídicího intervalu vyšla jako $\frac{149-124}{9} \approx 3$. Jeví se vhodné dolní mez prvního třídicího intervalu zvolit 123, horní mez posledního třídicího intervalu 150. Celkem třídicí intervaly pro znak X budou: $\langle 123, 126\rangle$, $\langle 126, 129\rangle$, ..., $\langle 147, 150\rangle$.

```
# Zeny
data.F     <- data[data$pohlavi=='zena',]
n.F        <- dim(data.F)[1]
(Sturges.F <- round(1+3.3*log10(n.F), digits=0))

## [1] 8
```

Protože žen je 109, podle Sturgersova pravidla je optimální počet třídicích intervalů 8. Postup je analogický jako u mužů.

```
# Znak X = Delka lebky
delka.F    <- data.F$delka
range(delka.F)

## [1] 157 188

max(delka.F) - min(delka.F)

## [1] 31

round((max(delka.F) - min(delka.F))/Sturges.F, digits=0)

## [1] 4
```

Pro znak X = Délka lebky je minimum 157 a maximum 188, rozsah těchto hodnot je 31 a ideální délka jednoho třídicího intervalu vyšla jako $\frac{188-157}{8} \approx 4$. Jeví se vhodné dolní mez prvního třídicího intervalu zvolit 156, horní mez posledního třídicího intervalu 188. Celkem třídicí intervaly pro znak X budou: $\langle 156, 160\rangle$, $\langle 160, 164\rangle$, ..., $\langle 184, 188\rangle$.

```
# Znak Y = Sirka lebky
sirka.F    <- data.F$sirka
range(sirka.F)
```

```
## [1] 118 146
```

```
max(sirka.F) - min(sirka.F)
```

```
## [1] 28
```

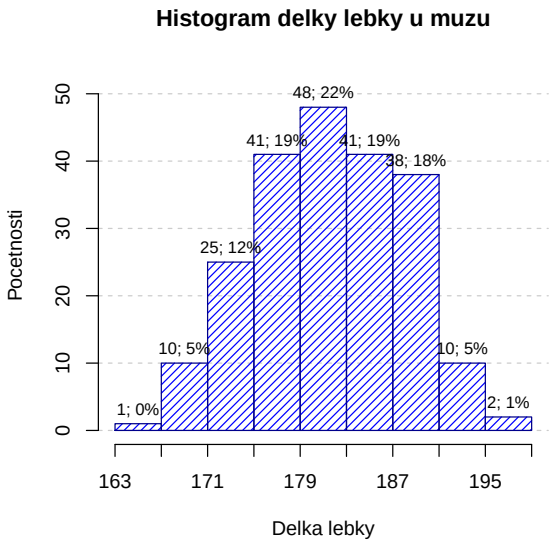
```
round((max(sirka.F) - min(sirka.F))/Sturges.F, digits=0)
```

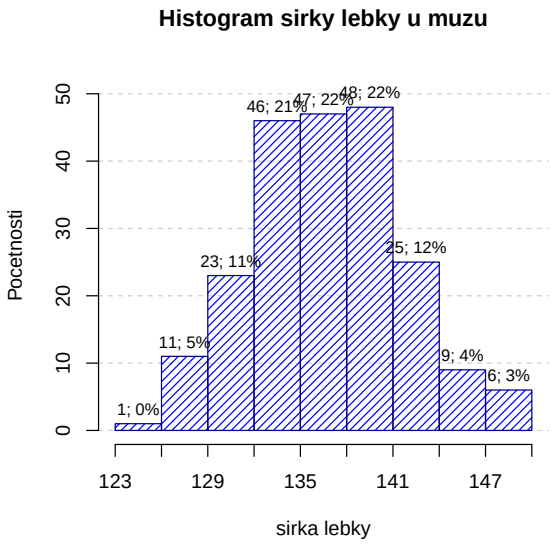
```
## [1] 4
```

Pro znak Y = šířka lebky je minimum 118 a maximum 146, rozsah těchto hodnot je 28 a ideální délka jednoho třídicího intervalu vyšla jako $\frac{146-118}{8} \approx 4$. Jeví se vhodné dolní mez prvního třídicího intervalu zvolit 116, horní mez posledního třídicího intervalu 148. Celkem třídicí intervaly pro znak X budou: $\langle 116, 120\rangle$, $\langle 120, 124\rangle$, ..., $\langle 144, 148\rangle$.

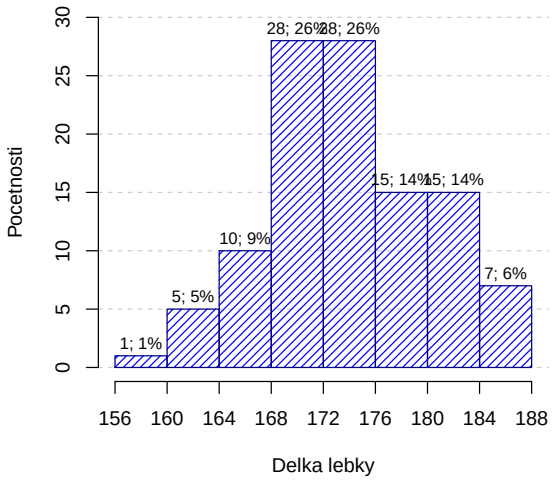
Příklad 1.7. Vytvořte histogram pro X a pro Y (s uvedenými absolutními a relativními četnostmi jednotlivých třídicích intervalů), a to zvlášť pro muže a zvlášť pro ženy.

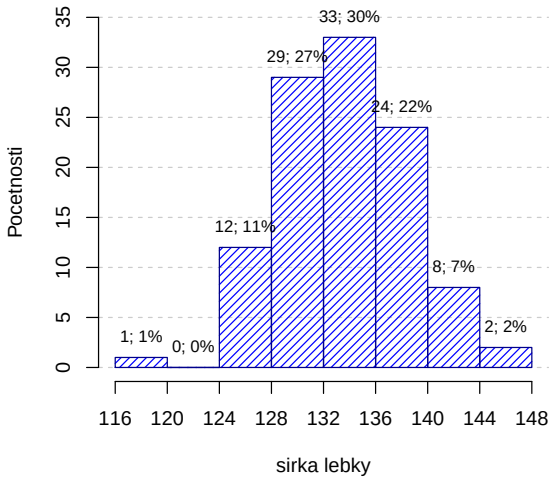
```
# Muzi
# X=Delka lebky
hist(delka.M, breaks=seq(163, 199, by=4), ylim=c(0,52),
     main='Histogram delky lebky u muzu', xlab='Delka lebky', ylab='Pocetnosti',
     col='white', border='white', density=20,  axes=F)
abline(h=seq( 0, 60, by=10), col='grey80', lty=2)
hist(delka.M, breaks=seq(163, 199, by=4),
     col='blue', border='darkblue', density=20, add=T)
axis(1, at=seq(163, 199, by=4))
axis(2, at=seq(  0,  50, by=10))

abs.c <- hist(delka.M, breaks=seq(163, 199, by=4), plot=F)$counts
stred <- hist(delka.M, breaks=seq(163, 199, by=4), plot=F)$mids
rel.c <- round(abs.c/sum(abs.c)*100, 0)

cetnosti <- paste(abs.c, '; ', rel.c, '%', sep='')
text(stred, abs.c+2, cetnosti, cex=0.8)

#-------------------------------------------------------------------------------
# Y=Sirka lebky
hranice <- seq(123, 150, by=3)
hist(sirka.M, breaks=hranice, ylim=c(0,52),
     main='Histogram sirky lebky u muzu', xlab='sirka lebky', ylab='Pocetnosti',
     col='white', border='white', density=20,  axes=F)
abline(h=seq(0, 60, by=10), col='grey80', lty=2)
hist(sirka.M, breaks=hranice,
     col='blue', border='darkblue', density=20, add=T)
axis(1, at=hranice)
axis(2, at=seq( 0, 50, by=10))

abs.c <- hist(sirka.M, breaks=hranice, plot=F)$counts
stred <- hist(sirka.M, breaks=hranice, plot=F)$mids
rel.c <- round(abs.c/sum(abs.c)*100, 0)

cetnosti <- paste(abs.c, '; ', rel.c, '%', sep='')
text(stred, abs.c+2, cetnosti, cex=0.8)
```





Pro ženy je postup analogický jako pro muže.





Příklady k samostatnému řešení

Příklad 1.8. Bodové rozdělení četností V severozápadním Skotsku byla provedena studie, která zkoumala výskyt krevních skupin. V oblasti Eskdale bylo náhodně vybráno 100 osob, v oblasti Annandale 125 osob a v oblasti Nithsdale 253 osob. Výsledky jsou uvedeny v tabulce:

oblast	Krevní skupina				$n_{.j}$
	A	B	0	AB	
Eskdale	33	6	56	5	100
Annandale	54	14	52	5	125
Nithsdale	98	35	115	5	253
$n_{.k}$	185	55	223	15	478

Jako znak X označíme oblast (má 3 varianty: Eskdale, Annandale, Nithsdale) a jako znak Y označíme krevní skupinu (má 4 varianty: A, B, AB a 0). Data jsou uložena v souboru krevni_skupiny.txt.

a) Vytvořte variační řadu znaku Y, a to pro všechny tři oblasti dohromady a pak pro každou zvlášť.

b) Nakreslete sloupkový diagram a polygon absolutních četností znaku Y.

c) Nakreslete výsečový diagram pro znak X.

d) Vytvořte kontingenční tabulku sloupcově a poté řádkově podmíněných relativních četností znaků X, Y.

Řešení:

a) Variační řady:

- Všechny tři oblasti dohromady

```
##     nj  Nj     pj     Fj
## A  185 185 0.3870 0.3870
## B   55 240 0.1151 0.5021
## 0  223 463 0.4665 0.9686
## AB  15 478 0.0314 1.0000
```

- Eskdale

```
##    nj  Nj   pj   Fj
## A  33  33 0.33 0.33
## B   6  39 0.06 0.39
## 0  56  95 0.56 0.95
## AB  5 100 0.05 1.00
```

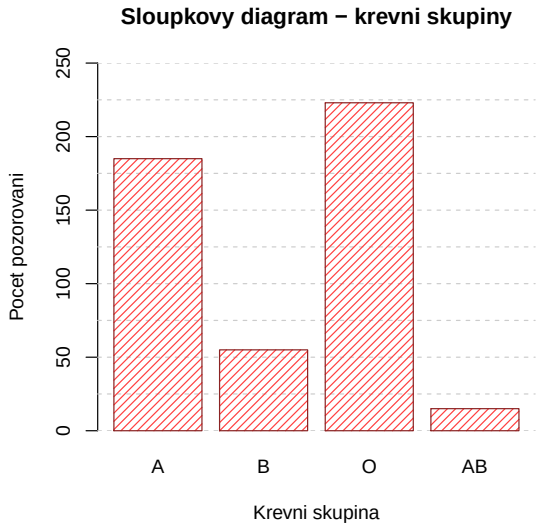
- Annandale

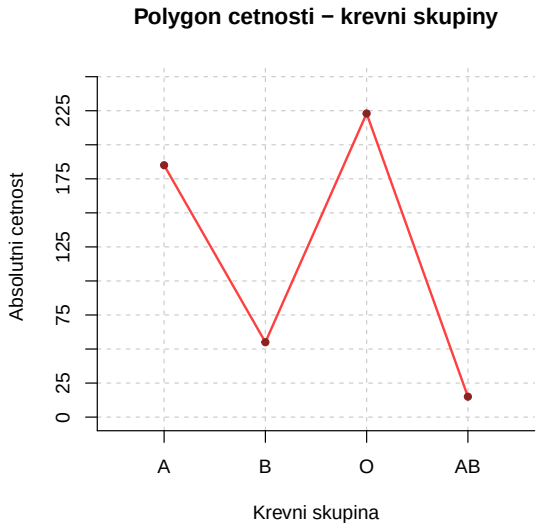
```
##    nj  Nj    pj    Fj
## A  54  54 0.432 0.432
## B  14  68 0.112 0.544
## 0  52 120 0.416 0.960
## AB  5 125 0.040 1.000
```

- Nithsdale

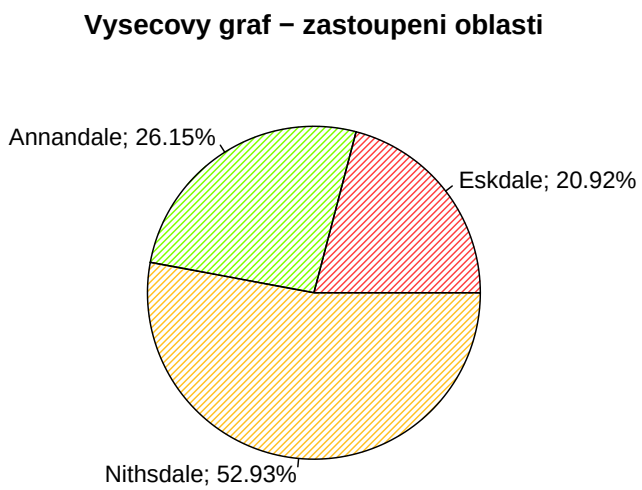
```
##     nj  Nj     pj     Fj
## A   98  98 0.3874 0.3874
## B   35 133 0.1383 0.5257
## 0  115 248 0.4545 0.9802
## AB   5 253 0.0198 1.0000
```

b) Sloupkový graf a polygon četností





c) Výsečový graf



d) Řádkově a sloupcově podmíněné četnosti

- Tabulka řádkově podmíněných četností

```
##           Krev.Skupina
## Oblast          A     B     O    AB
##   Annandale 0.432 0.112 0.416 0.040
##   Eskdale   0.330 0.060 0.560 0.050
##   Nithsdale 0.387 0.138 0.455 0.020
```

- Tabulka sloupcově podmíněných četností

```
##            Krev.Skupina
## Oblast            A     B     O    AB
##   Annandale   0.292 0.255 0.233 0.333
##   Eskdale     0.178 0.109 0.251 0.333
##   Nithsdale   0.530 0.636 0.516 0.333
```

Příklad 1.9. Intervalové rozdělení četností U 50 studentů a studentek byla zjišťována jejich hmotnost (znak X, v kg), výška (znak Y, v cm) a pohlaví (znak Z, 0... žena, 1... muž). Data jsou uložena v souboru vys_vah.txt.

a) Podle Sturgesova pravidla najděte optimální počet třídicích intervalů pro znaky X a Y a vhodně stanovte meze třídicích intervalů.

b) Vytvořte histogram pro X a pro Y (s uvedenými absolutními a relativními četnostmi jednotlivých třídicích intervalů).

Řešení

a) Protože osob je 50, podle Sturgesova pravidla je optimální počet třídicích intervalů 7. Musíme zjistit minimum a maximum, abychom vhodně stanovili meze třídicích intervalů.

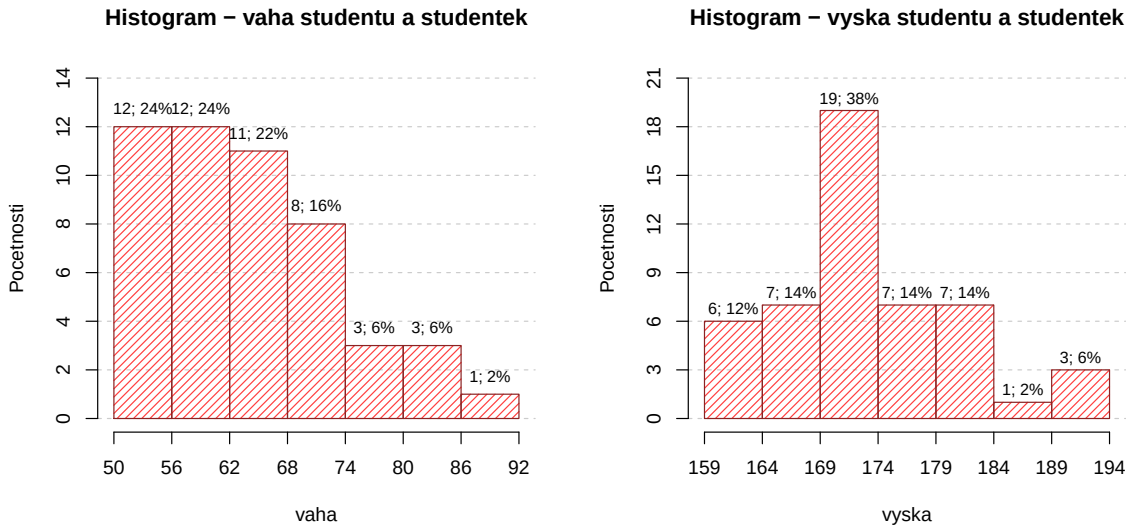
Pro znak X je minimum 51 a maximum 90. Jeví se vhodné dolní mez prvního třídicího intervalu zvolit 50, horní mez posledního třídicího intervalu 92. Délka třídicích intervalů je tedy $\frac{92-50}{7} = 6$. Celkem třídicí intervaly pro znak X budou:

$$\langle 50\,;\,56\rangle,\ \ \langle 56\,;\,62\rangle,\ \ldots,\ \langle 86\,;\,92\rangle.$$

Pro znak Y je minimum 160 a maximum 192. Jeví se vhodné dolní mez prvního třídicího intervalu zvolit 159, horní mez posledního třídicího intervalu 194. Délka třídicích intervalů je tedy $\frac{194-159}{7} = 5$. Celkem třídicí intervaly pro znak Y budou:

$$\langle 159\,;\,164\rangle,\ \ \langle 164\,;\,169\rangle,\ \ldots,\ \langle 189\,;\,194\rangle.$$

b) Histogramy pro váhu a výšku



Poznámka: Tytéž úkoly lze řešit zvlášť pro muže a zvlášť pro ženy.

2 Výpočet číselných charakteristik jednorozměrného a dvourozměrného datového souboru

Přehled použitých funkcí: data.frame, apply, library, round, cramersV, read.delim, source, head, names, factor, quantile, boxplot, cor, dotplot, abline, length, mean, var, sqrt, skewness, kurtosis, cbind.

Příklad 2.1. U 100 náhodně vybraných domácností byl zjiťován způsob zásobování bramborami (znak X, varianty 1 = vlastní sklep, 2 = jinde, 3 = nákup) a bydliště (znak Y, varianty 1 = velké město, 2 = malé město, 3 = vesnice).

	velké město	malé město	vesnice
vlastní sklep	13	15	14
jinde	11	7	2
nákup	19	9	10

a) Pro oba znaky určíme modus.

b) Vypočteme Cramérův koeficient znaků X, Y.

a) Stanovení modu

```
(data <- data.frame(velke.mesto=c(13,11,19), male.mesto=c(15,7,9), vesnice=c(14,2,10),
                    row.names=c('sklep','jinde','nakup')))

##       velke.mesto male.mesto vesnice
## sklep          13         15      14
## jinde          11          7       2
## nakup          19          9      10

apply(data,1,sum)

## sklep jinde nakup
##    42    20    38

apply(data,2,sum)

## velke.mesto  male.mesto     vesnice
##          43          31          26
```

Znak X má modus 1, tj. nejvíce domácností skladuje brambory ve vlastním sklepě a znak Y má také modus 1, tj. nejvíce domácností bydlí ve velkém městě.

b) Výpočet Cramérova koeficientu

Hodnotu Cramérova koeficientu vypočítáme pomocí funkce cramersV, která je součástí knihovny lsr. Nejprve tedy musíme nainstalovat tuto knihovnu (Packages → Install → lsr → Install) a následně ji načíst (library(lsr)). Teprve potom můžeme funcki cramersV() použít na naši datovou tabulku a Cramérův koeficient dopočítat.

```
library(lsr)
round(cramersV(data), digits=3)

## [1] 0.179
```

Cramérův koeficient nabývá hodnoty 0.179, tedy mezi způsobem zásobování bramborami a bydlištěm domácnosti existuje jen slabá závislost - viz následující tabulka:

Cramérův koeficient	interpretace
0 – 0.1	zanedbatelná závislost
0.1 – 0.3	slabá závislost
0.3 – 0.7	střední závislost
0.7 – 1	silná závislost

Příklad 2.2. Otevřeme datový soubor znamky.txt.

a) Pro známky z matematiky a angličtiny vypočteme medián, dolní a horní kvartil, kvartilovou odchylku a vytvoříme krabicový diagram.

b) Vypočteme Spearmanův korelační koeficient známek z matematiky a angličtiny pro všechny studenty, pak zvlášť pro muže a zvlášť pro ženy. Získané výsledky budeme interpretovat.

a)
```
data <- read.delim('znamky.txt', sep='\t', dec='.',header=F)
source('AS-funkce.R')
head(data)
```

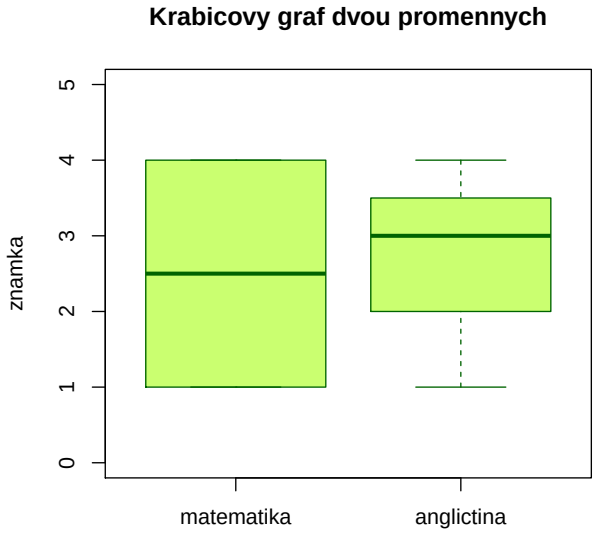
```
##   V1 V2 V3
## 1  2  2  0
## 2  1  3  1
## 3  4  3  1
## 4  1  1  0
## 5  1  2  1
## 6  4  4  1
```

```
names(data) <- c('matematika', 'anglictina', 'pohlavi')
f3    <- factor(data$pohlavi, levels=c(0,1), labels=c('zena','muz'))
data[,3] <- f3
head(data)
```

```
##   matematika anglictina pohlavi
## 1          2          2    zena
## 2          1          3     muz
## 3          4          3     muz
## 4          1          1    zena
## 5          1          2     muz
## 6          4          4     muz
```

```
matematika <- data$matematika
anglictina <- data$anglictina
pohlavi    <- data$pohlavi

q.M <- quantile(matematika, probs=c(0.5,0.25,0.75), type=2) #type=5
q.A <- quantile(anglictina, probs=c(0.5,0.25,0.75), type=2)
iqr.M <- q.M[3]-q.M[2]
iqr.A <- q.A[3]-q.A[2]

(tabulka<-data.frame(median=c(q.M[1],q.A[1]), kv1=c(q.M[2],q.A[2]), kv3=c(q.M[3],q.A[3]),
                     IQR=c(iqr.M, iqr.A), row.names=c('matematika','anglictina')))
```

```
##            median kv1 kv3 IQR
## matematika    2.5   1 4.0 3.0
## anglictina    3.0   2 3.5 1.5

boxplot(matematika, anglictina, main='Krabicovy graf dvou promennych',
        names=c('matematika','anglictina'), ylab='znamka', ylim=c(0,5),
        border='darkgreen', col='darkolivegreen1')
```

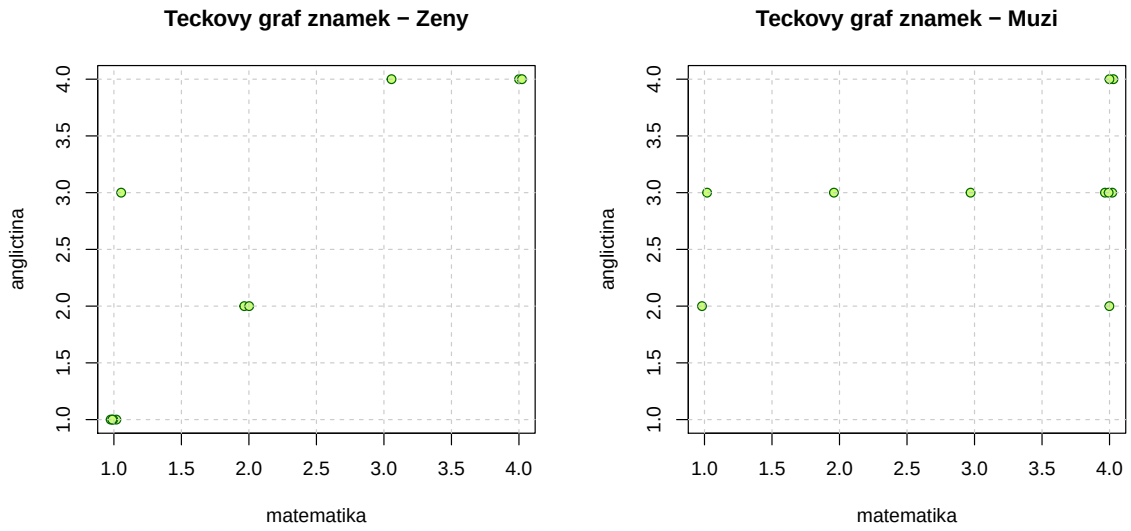


b)
```
cor(matematika, anglictina, method='spearman')

## [1] 0.6884422

cor(matematika[pohlavi=='zena'], anglictina[pohlavi=='zena'], method='spearman')

## [1] 0.8603138

cor(matematika[pohlavi=='muz'], anglictina[pohlavi=='muz'], method='spearman')

## [1] 0.3735437
```

Vidíme, že nejsilnější přímá pořadová závislost mezi známkami z matematiky a angličtiny je u žen, $r_S = 0.86$. U mužů je tato závislost mnohem slabší, $r_S = 0.37$. U žen tedy dochází k tomu, že se sdružují podobné známky z obou předmětů, zatímco u mužů se projevuje spíše tendence k různým známkám. Je to zřetelně vidět na dvourozměrných tečkových diagramech.

```
dotplot(matematika[pohlavi=='zena'], anglictina[pohlavi=='zena'],
        main='Teckovy graf znamek - Zeny', xlab='matematika', ylab='anglictina',
        col='darkgreen', bg='darkolivegreen1', xlim=c(1,4), ylim=c(1,4))
abline(v=seq(1,4,by=0.5), col='grey80', lty=2)
abline(h=seq(1,4,by=0.5), col='grey80', lty=2)
```

```r
dotplot(matematika[pohlavi=='muz'], anglictina[pohlavi=='muz'],
        main='Teckovy graf znamek - Muzi', xlab='matematika', ylab='anglictina',
        col='darkgreen', bg='darkolivegreen1', xlim=c(1,4), ylim=c(1,4))
abline(v=seq(1,4,by=0.5), col='grey80', lty=2)
abline(h=seq(1,4,by=0.5), col='grey80', lty=2)
```



Význam hodnot Spearmanova (i Pearsonova) koeficientu korelace je popsán v tabulce:

Abs.hod. korel.koef.	Interpretace hodnoty
0	pořadová (lineární) nezávislost
$(0; 0.1)$	velmi nízký stupeň závislosti
$[0.1; 0.3)$	nízký stupeň závislosti
$[0.30; 0.50)$	mírný stupeň závislosti
$[0.50; 0.70)$	význačný stupeň závislosti
$[0.70; 0.90)$	vysoký stupeň závislosti
$[0.90; 1)$	velmi vysoký stupeň závislosti
1	úplná pořadová (lineární) závislost

Podle výše uvedené tabulky existuje mezi známkami z matematiky a známkami z angličtiny význačný stupeň přímé pořadové závislosti ($r_S = 0.69$), dále v případě žen existuje mezi známkami z matematiky a z angličtiny vysoký stupeň přímé pořadové závislosti ($r_S = 0.86$), zatímco u mužů existuje mezi známkami z matematiky a z angličtiny pouze mírný stupeň přímé pořadové závislosti ($r_S = 0.37$).

Příklad 2.3. Otevřeme datový soubor lebky.txt.

a) Pro největší délku a největší šířku mozkovny mužů vypočteme aritmetický průměr, rozptyl, směrodatnou odchylku, koeficient variace, šikmost a špičatost.

b) Vypočítejte Pearsonův koeficient korelace největší délky a největší šířky mozkovny mužů. Dále vypočtěte kovarianci těchto dvou znaků a nakreslete dvourozměrný tečkový diagram.

a)
```r
library(e1071)
data        <- read.delim('lebky.txt', sep='\t', dec='.', header=F)
names(data) <- c('delka', 'sirka', 'pohlavi')
head(data)
```

```
##   delka sirka pohlavi
## 1   188   145     muz
## 2   172   139     muz
## 3   176   138     muz
## 4   184   128     muz
## 5   183   139     muz
## 6   177   143     muz
```

```
delka.M   <- data$delka[data$pohlavi=='muz']
n         <- length(delka.M)

prumer.D    <- mean(delka.M)
rozptyl.D   <- 1/n*sum((delka.M-prumer.D)^2)
sm.odch.D   <- sqrt(rozptyl.D)
koef.var.D  <- sm.odch.D/mean(delka.M)*100
sikmost.D   <- skewness(delka.M, type=2)
spicatost.D <- kurtosis(delka.M, type=2)
(tab.D      <- round(data.frame(n=n, prumer=prumer.D, rozptyl=rozptyl.D, sm.odch=sm.odch.D,
                  koef.var=koef.var.D, sikmost=sikmost.D, spicatost=spicatost.D), digits=4))
```

```
##     n   prumer rozptyl sm.odch koef.var sikmost spicatost
## 1 216 182.0324 40.5777  6.3701   3.4994 -0.0551   -0.4511
```

Analogický postup zvolíme pro výpočty základních charakteristik pro šířku mozkovny mužů. Výsledné charakteristiky pro obě proměnné sloučíme do jedné tabulky.

```
##         n   prumer rozptyl sm.odch koef.var sikmost spicatost
## delka 216 182.0324 40.5777  6.3701   3.4994 -0.0551   -0.4511
## sirka 216 137.1852 23.1694  4.8135   3.5087  0.0853   -0.2485
```

b) **Výpočet Pearsonova korelačního koeficientu**

```
cor(delka.M, sirka.M, method='pearson')
```

```
## [1] 0.168157
```

Vidíme, že mezi délkou mozkovny a šířkou mozkovky u mužů existuje nízký stupeň přímé lineární závislosti.

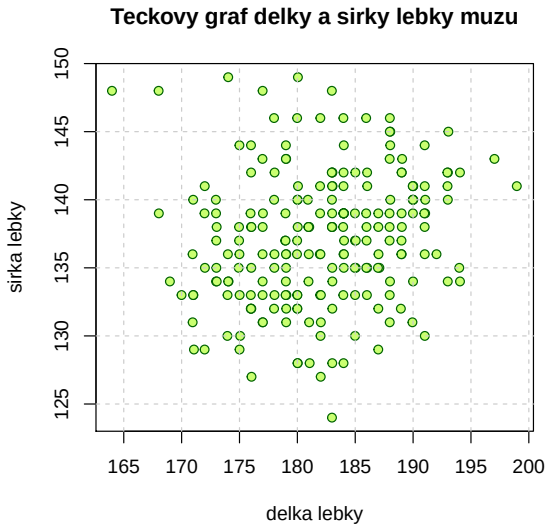
Výpočet kovariance

```
kovariance <- sum((delka.M-prumer.D)*(sirka.M-prumer.S))/n
round(kovariance, 4)
```

```
## [1] 5.156
```

Tečkový diagram

```
dotplot(delka.M, sirka.M, main='Teckovy graf delky a sirky lebky muzu',
        xlab='delka lebky', ylab='sirka lebky', col='darkgreen', bg='darkolivegreen1')
abline(v=seq(160,200,by=5), col='grey80', lty=2)
abline(h=seq(120,145,by=5), col='grey80', lty=2)
```



Vzhledu diagramu potvrzuje naše zjištění, že mezi délkou a šířkou mozkovny u mužů existuje nízká přímá lineární závislost.

3 Opakované pokusy

Přehled použitých funkcí: sum, dbinom, pbinom, dgeom, pgeom, dhyper, phyper.

3.1 Opakované nezávislé pokusy

Opakovaně nezávisle provádíme týž náhodný pokus a sledujeme nastoupení jevu, kterému říkáme **úspěch**. V každém z těchto pokusů nastává úspěch s pravděpodobností θ, $0 < \theta < 1$.

Binomické rozdělení pravděpodobnosti

Pravděpodobnost, že v prvních n pokusech úspěch nastane právě x-krát ($0 \leq x \leq n$):

$$P_n(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{1}$$

Pravděpodobnost, že v prvních n pokusech úspěch nastane nejvýše x_1-krát ($0 \leq x_1 \leq n$):

$$\sum_{x=0}^{x_1} P_n(x). \tag{2}$$

Pravděpodobnost, že v prvních n pokusech úspěch nastane aspoň x_0-krát ($0 \leq x_0 \leq n$):

$$\sum_{x=x_0}^{n} P_n(x). \tag{3}$$

Pravděpodobnost, že v prvních n pokusech úspěch nastane aspoň x_0-krát a nejvýše x_1-krát:

$$\sum_{x=x_0}^{x_1} P_n(x). \tag{4}$$

Příklad 3.1. Pojišťovna zjistila, že 12 % pojistných událostí je způsobeno vloupáním. Jaká je pravděpodobnost, že mezi 30 náhodně vybranými pojistnými událostmi bude způsobeno vloupáním

a) nejvýše 6;

b) aspoň 6;

c) právě 6;

d) od dvou do pěti?

Počet pokusů: $n = 30$, pravděpodobnost úspěchu: $\theta = 0.12$

ad a)

$$\sum_{x=0}^{x_1} P_n(x) = \sum_{x=0}^{6} P_{30}(x) = \sum_{x=0}^{6} \binom{30}{x} 0.12^x 0.88^{30-x} = 0.9394.$$

```
sum(dbinom(0:6, size=30, prob=0.12))

## [1] 0.9393926

pbinom(6, size=30, prob=0.12)

## [1] 0.9393926
```

S pravděpodobností 93.94 % bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním nejvýše 6 událostí.

ad b)

$$\sum_{x=x_0}^{n} P_n(x) = \sum_{x=6}^{30} P_{30}(x) = 1 - \sum_{x=0}^{5} P_{30}(x) = 1 - \sum_{x=0}^{5} \binom{30}{x} 0.12^x 0.88^{30-x} = 0.1431.$$

```
1-sum(dbinom(0:5, size=30, prob=0.12))

## [1] 0.1430769

1-pbinom(5, size=30, prob=0.12)

## [1] 0.1430769
```

S pravděpodobností 14.31 % bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním aspoň 6 událostí.

ad c)

$$P_n(x) = P_{30}(6) = \binom{30}{6} 0.12^6 0.88^{24} = 0.0825.$$

```
dbinom(6, size=30, prob=0.12)

## [1] 0.08246953
```

S pravděpodobností 8.25 % bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním právě 6 událostí.

ad d)

$$\sum_{x=x_0}^{x_1} P_n(x) = \sum_{x=2}^{5} P_{30}(x) = \sum_{x=0}^{5} P_{30}(x) - \sum_{x=0}^{1} P_{30}(x) = \sum_{x=0}^{5} \binom{30}{x} 0.12^x 0.88^{30-x} - \sum_{x=0}^{1} \binom{30}{x} 0.12^x 0.88^{30-x} = 0.7470.$$

```
pbinom(5, size=30, prob=0.12) - pbinom(1, 30, 0.12)

## [1] 0.7469528

sum(dbinom(2:5, size=30, prob=0.12))

## [1] 0.7469528
```

S pravděpodobností 74.70 % bude mezi 30 náhodně vybranými pojistnými událostmi způsobeno vloupáním od 2 do 5 událostí.

Příklady k samostatnému řešení

Příklad 3.2. V rodině je 10 dětí. Za předpokladu, že chlapci i dívky se rodí s pravděpodobností 0.5 a pohlaví se formuje nezávisle na sobě, určete pravděpodobnost, že v této rodině je

a) právě 5 chlapců;

b) nejméně 3 a nejvýše 8 chlapců.

$n = 10$; úspěch = narození chlapce; pravděpodobnost úspěchu $\theta = 0.5$;

ad a)
```
## [1] 0.2460938
```

ad b)
```
## [1] 0.9345703
```

Příklad 3.3. Na dvoukolejném železničním mostě se potkají během 24 hodin nejvýše dva vlaky, a to s pravděpodobností 0.2. Za předpokladu, že denní provozy jsou nezávislé, určete pravděpodobnost, že během týdne se dva vlaky na mostě potkají

a) právě třikrát;

b) nejvýše třikrát;

c) alespoň třikrát.

$n = 7$; úspěch = potkání dvou vlaků během 24 hodin; pravděpodobnost úspěchu $\theta = 0.2$;

ad a)
```
## [1] 0.114688
```

ad b)
```
## [1] 0.966656
```

ad c)
```
## [1] 0.148032
```

Příklad 3.4. Je pravděpodobnější vyhrát se stejně silným soupeřem tři partie ze čtyř nebo pět partií z osmi, když nerozhodný výsledek je vyloučen a výsledky jsou nezávislé?

Úspěch je výhra partie se stejně silným soupeřem, když remíza je vyloučena; pravděpodobnost úspěchu $\theta = 0.5$;

a) n = 4, x = 3;

b) n = 8, x = 5.

ad a)
```
## [1] 0.25
```

ad b)
```
## [1] 0.21875
```

Příklad 3.5. Dvacetkrát nezávisle na sobě házíme třemi mincemi. Jaká je pravděpodobnost, že alespoň v jednom hodu padnou tři líce?

$n = 20$; úspěch je padnutí tří líců při hodu třemi mincemi; $\theta = 1/8 = 0.125$;

```
## [1] 0.9307912
```

Geometrické rozdělení pravděpodobnosti

Pravděpodobnost, že prvnímu úspěchu bude předcházet x neúspěchů:

$$P(x) = (1-\theta)^x \theta. \tag{5}$$

Pravděpodobnost, že prvnímu úspěchu bude předcházet nejvýše x_1 neúspěchů:

$$\sum_{x=0}^{x_1} P(x). \tag{6}$$

Pravděpodobnost, že prvnímu úspěchu bude předcházet aspoň x_0 neúspěchů:

$$1 - \sum_{x=0}^{x_0-1} P(x). \tag{7}$$

Příklad 3.6. Jaká je pravděpodobnost, že při hře "Člověče, nezlob se!" nasadíme figurku nejpozději při třetím hodu?

Počet neúspěchů: $x = 0, 1, 2$; pravděpodobnost úspěchu: $\theta = \frac{1}{6}$;

$$\sum_{x=0}^{2} P(x) = \sum_{x=2}^{2} (1-\theta)^x \theta = \sum_{x=0}^{2} \left(\frac{5}{6}\right)^x \frac{1}{6} = 0.4213$$

```
sum(dgeom(0:2, prob=1/6))
```

```
## [1] 0.4212963
```

```
pgeom(2, prob=1/6)
```

```
## [1] 0.4212963
```

Pravděpodobnost, že figurku nasadíme nejpozději při třetím hodu, je 42.13 %.

Příklad k samostatnému řešení

Příklad 3.7. Studenti biologie zkoumají barvu očí octomilek. Pravděpodobnost, že octomilka má bílou barvu očí, je 0.25, pravděpodobnost, že má červenou barvu očí, je 0.75. Jaká je pravděpodobnost, že až čtvrtá zkoumaná octomilka bude mít bílou barvu očí?

Počet neúspěchů: $x = 3$; pravděpodobnost úspěchu: $\theta = 0.25$;

```
## [1] 0.1054688
```

3.2 Opakované závislé pokusy

Hypergeometrické rozdělení pravděpodobností

Máme N objektů, mezi nimi je M objektů označeno $0 \leq M \leq N$. Náhodně bez vracení vybereme k objektů $(0 \leq k \leq N)$.
Pravděpodobnost, že ve výběru je právě x označených objektů $(\max\{0, M-N+k\} \leq x \leq \min\{k, M\})$:

$$P_{N,M,k}(x) = \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}}. \tag{8}$$

Pravděpodobnost, že ve výběru je nejvýše x_1 označených objektů:

$$\sum_{x=\max\{0,M-N+k\}}^{x_1} P_{N,M,k}(x). \tag{9}$$

Pravděpodobnost, že ve výběru je aspoň x_0 označených objektů:

$$\sum_{x=x_0}^{\min\{k,M\}} P_{N,M,k}P(x). \tag{10}$$

Příklad 3.8. Koupili jsme 10 cibulek červených tulipánů a 5 cibulek žlutých tulipánů. Zasadili jsme 8 náhodně vybraných cibulek.

a) Jaká je pravděpodobnost, že žádná nebude cibulka žlutých tulipánů?

b) Jaká je pravděpodobnost, že jsme zasadili všech 5 cibulek žlutých tulipánů?

c) Jaká je pravděpodobnost, že aspoň dvě budou cibulky žlutých tulipánů?

Počet objektů: $N = 15$, počet označených objektů: $M = 5$, počet vybraných objektů: $n = 8$

ad a)

$$P_{15,5,8}(0) = \frac{\binom{5}{0}\binom{10}{8}}{\binom{15}{8}} = \frac{\binom{10}{8}}{\binom{15}{8}} = 0.007$$

```
dhyper(0, m=5, n=10, k=8)

## [1] 0.006993007
```

Mezi 8 náhodně vybranými cibulkami se s pravděpodobností 0.7 % nevyskytne žádná cibulka žlutých tulipánů.

ad b)

$$P_{15,5,8}(5) = \frac{\binom{5}{5}\binom{10}{3}}{\binom{15}{8}} = \frac{\binom{10}{3}}{\binom{15}{8}} = 0.0186$$

```
dhyper(5, m=5, n=10, k=8)

## [1] 0.01864802
```

S pravděpodobností 1.86 % bude mezi 8 náhodně vybranými cibulkami právě 5 cibulek žlutých tulipánů.

ad c)

$$1 - P_{15,5,8}(0) - P_{15,5,8}(1) = 1 - \frac{\binom{5}{0}\binom{10}{8}}{\binom{15}{8}} - \frac{\binom{5}{1}\binom{10}{7}}{\binom{15}{8}} = 1 - \frac{\binom{10}{8}}{\binom{15}{8}} - \frac{5\binom{10}{7}}{\binom{10}{8}} = 0.8998$$

```
sum(dhyper(2:5, m=5, n=10, k=8))

## [1] 0.8997669

phyper(5, m=5, n=10, k=8)-phyper(1, m=5, n=10, k=8)

## [1] 0.8997669
```

S pravděpodobností 89.98 % budou mezi 8 náhodně vybranými cibulkami aspoň dvě cibulky žlutých tulipánů.

Příklad k samostatnému řešení:

Příklad 3.9. Dítě dostalo sáček, v němž bylo 5 červených a 5 žlutých bonbónů. Dítě náhodně vybralo ze sáčku 6 bonbónů. Jaká je pravděpodobnost, že mezi vybranými bonbóny budou právě 2 červené?

```
## [1] 0.2380952
```

4 Pravděpodobnostní funkce, hustoty a distribuční funkce, výpočet pravděpodobností pomocí distribučních funkcí

V této kapitole se zaměříme zejména na binomické rozdělení, Poissonovo rozdělení, exponenciální rozdělení a normální rozdělení.

4.1 Binomické rozdělení $\mathrm{Bin}(n, \theta)$

Náhodná veličina X udává počet úspěchů v posloupnosti n nezávislých opakovaných pokusů, přičemž pravděpodobnost úspěchu v každém pokusu je vyjádřena pomocí parametru θ. Píšeme $X \sim \mathrm{Bin}(n, \theta)$.

Pravděpodobnostní funkce binomického rozdělení má tvar

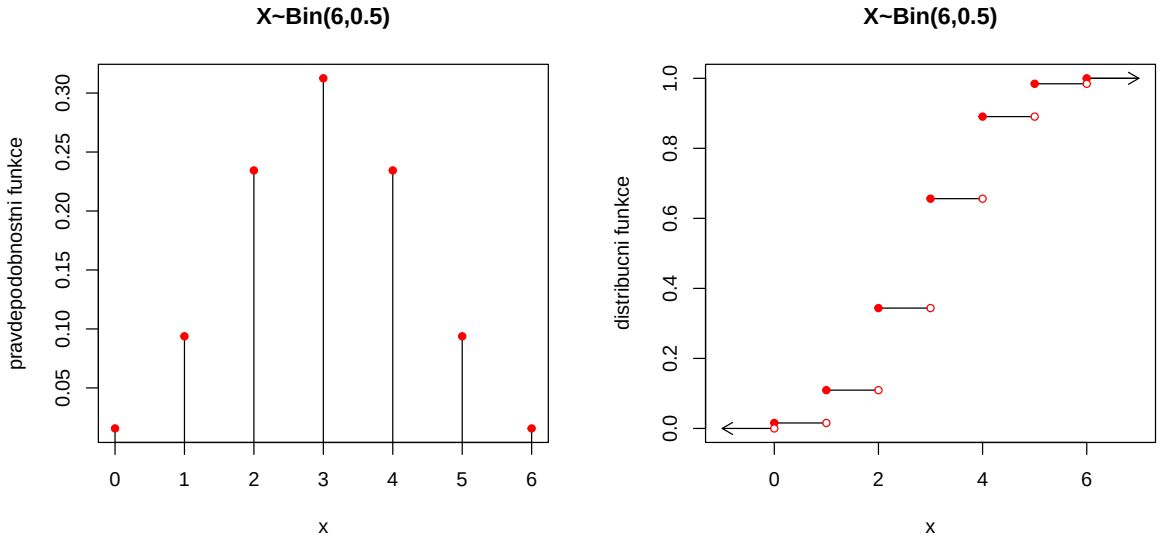
$$p(x) = \begin{cases} \binom{n}{x}\theta^x(1-\theta)^{n-x} & \text{pro } x = 0, \ldots, n; \\ 0 & \text{jinak.} \end{cases} \tag{11}$$

Distribuční funkce binomického rozdělení má tvar

$$F(x) = \sum_{t=0}^{x} \binom{n}{t} \theta^t (1-\theta)^{n-t}. \tag{12}$$

Příklad 4.1. Nakreslete graf pravděpodobnostní funkce a distribuční funkce náhodné veličiny $X \sim \mathrm{Bin}(6, 0.5)$.

```
#graf pravdepodobnostni funkce
x  <- 0:6
px <- dbinom(x, size=6, prob=0.5)
plot(x, px, type='h', main='X~Bin(6,0.5)', ylab='pravdepodobnostni funkce', xlab='x')
points(x, px, col='red', pch=19, cex=0.8)

#graf distribucni funkce
Fx <- pbinom(x, size=6, prob=0.5)
n  <- length(Fx)

plot(x, Fx, type='n', main='X~Bin(6,0.5)', ylab='distribucni funkce',
     xlab='x', xlim=c(-1,n), ylim=c(0,1))
segments(x, Fx, x+1, Fx)
arrows(0,  0, -1, 0, length=0.1)
arrows(n-1,1,  n, 1, length=0.1)
points(x, Fx, col='red', pch=19, cex=0.8)
points(x, c(0, Fx[-n]), col='red', bg='white', pch=21, cex=0.8)
```



Analogickým způsobem můžeme získat grafy pravděpodobnostních a distribučních funkcí binomického rozdělení pro různá n a θ a sledovat vliv těchto parametrů na vzhled grafů.

4.2 Poissonovo rozdělení $\mathrm{Po}(\lambda)$

Náhodná veličina X udává počet událostí, které nastanou v jednotkovém časovém intervalu (resp. v jednotkové oblasti), přičemž k událostem dochází náhodně, jednotlivě a vzájemně nezávisle. Parametr $\lambda > 0$ je střední počet těchto událostí. Píšeme $X \sim \mathrm{Po}(\lambda)$.

Pravděpodobnostní funkce Poissonova rozdělení má tvar

$$p(x) = \begin{cases} \dfrac{\lambda^x}{x!} e^{-\lambda} & \text{pro x=0,1,...;} \\ 0 & \text{jinak.} \end{cases} \tag{13}$$

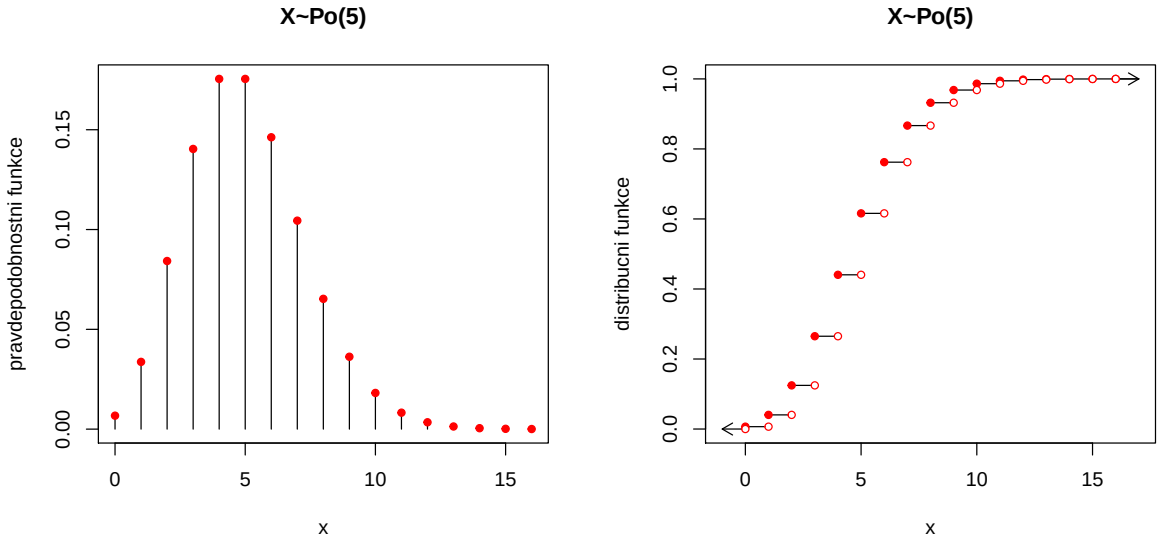
Distribuční funkce Poissonova rozdělení má tvar

$$F(x) = \sum_{t=0}^{x} \frac{\lambda^t}{t!} e^{-\lambda}. \tag{14}$$

Příklad 4.2. Nakreslete graf pravděpodobnostní funkce a distribuční funkce náhodné veličiny $X \sim \mathrm{Po}(5)$.

```
#graf pravdepodobnostni funkce
x  <- 0:16
px <- dpois(x, lambda=5)
plot(x,   px, type='h', main='X~Po(5)', ylab='pravdepodobnostni funkce', xlab='x')
points(x, px, col='red', pch=19, cex=0.8)

#graf distribucni funkce
Fx <- ppois(x, lambda=5)
n  <- length(Fx)
plot(x, Fx, type='n', main='X~Po(5)', ylab='distribucni funkce', xlab='x',
     xlim=c(-1,n), ylim=c(0,1))
segments(x, Fx, x+1, Fx)
arrows(0,   0, -1, 0, length=0.1)
```

```r
arrows(n-1, 1,  n, 1, length=0.1)
points(x, Fx, col='red', pch=19, cex=0.8)
points(x, c(0, Fx[-n]), col='red', bg='white', pch=21, cex=0.8)
```



Příklad 4.3. Při provozu balicího automatu vznikají během směny náhodné poruchy, které se řídí rozdělením Po(2). Jaká je pravděpodobnost, že během směny dojde k alespoň jedné poruše?

X ... počet poruch během směny; $X \sim \text{Po}(2)$;

$$\text{P}(X \geq 1) = 1\text{–P}(X < 1) = 1\text{–P}(X \leq 0) = 1 - \text{P}(X = 0) = 1 - \frac{2^0}{0!}e^{-2} = 0.8647.$$

```r
1-dpois(0, lambda=2)
```

```
## [1] 0.8646647
```

```r
1-ppois(0, lambda=2)
```

```
## [1] 0.8646647
```

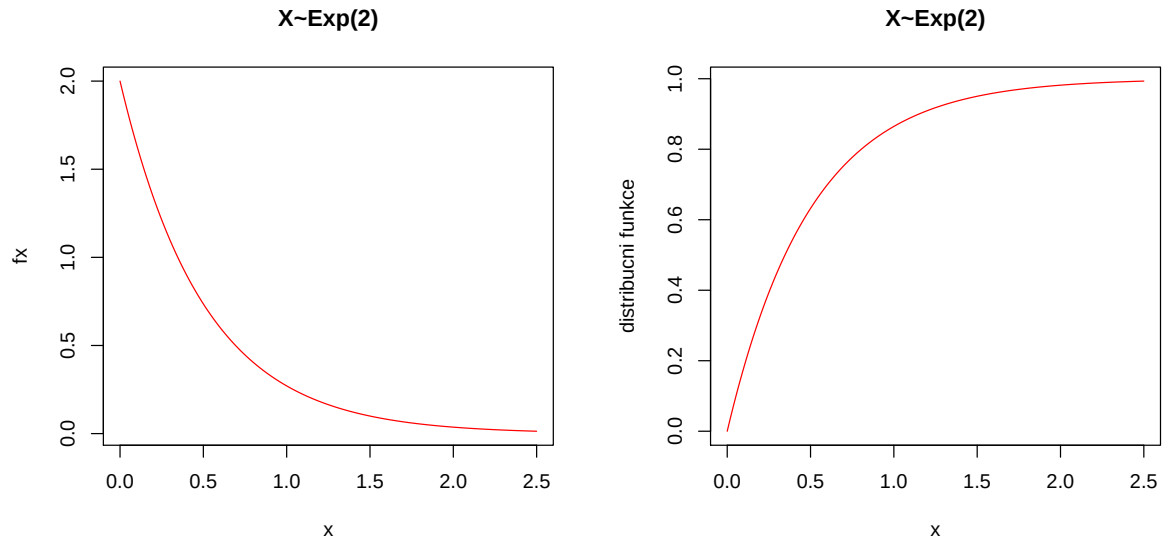
4.3 Exponenciální rozdělení Exp(λ)

Náhodná veličina X udává dobu čekání na příchod nějaké události, která se může dostavit každým okamžikem se stejnou šancí bez ohledu na dosud pročekanou dobu. (Jde o tzv. čekání bez paměti.) Přitom $\frac{1}{\lambda}$ vyjadřuje střední dobu čekání. Náhodná veličina $X \sim \text{Exp}(\lambda)$ má hustotu

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x > 0; \\ 0 & \text{jinak.} \end{cases} \tag{15}$$

Příklad 4.4. Nakreslete graf hustoty a distribuční funkce náhodné veličiny $X \sim \text{Exp}(2)$.

```
#graf hustoty
x  <- seq(from=0, to=2.5, length=512)
fx <- dexp(x, rate=2)
plot(x, fx, type='l', main='X~Exp(2)', xlab='x', col='red')

#graf distribucni funkce
Fx <- pexp(x, rate=2)
plot(x, Fx, main='X~Exp(2)', xlab='x', ylab='distribucni funkce', type='l', col='red')
```



Příklad 4.5. Doba do ukončení opravy v opravně obuvi je náhodná veličina, která se řídí exponenciálním rozdělením se střední dobou opravy 3 dny. Jaká je pravděpodobnost, že oprava bude ukončena do dvou dnů?

$X \sim \text{Exp}(1/3)$;

$$\text{P}(X \leq 2) = \int_0^2 \frac{1}{3} e^{-\frac{x}{3}} \mathrm{d}x = \left[-e^{-\frac{x}{3}}\right]_0^2 = 1 - e^{-\frac{2}{3}} = 0.4866.$$

```
pexp(2, rate=1/3)

## [1] 0.4865829
```

Příklad 4.6. Doba (v hodinách), která uplyne mezi dvěma naléhavými příjmy v jisté nemocnici, se řídí exponenciálním rozdělením se střední dobou čekání 2 h. Jaká je pravděpodobnost, že uplyne více než 5 h bez naléhavého příjmu?

$X \sim \text{Exp}(1/2)$;

$$\text{P}(X > 5) = \text{P}(X \geq 5) = \int_5^\infty \frac{1}{2} e^{-\frac{x}{2}} \text{d}x = \left[-e^{-\frac{x}{2}}\right]_0^2 = e^{-\frac{5}{2}} = 0.0821.$$

```
1-pexp(5, rate=1/2)

## [1] 0.082085
```

4.4 Normální rozdělení $N(\mu, \sigma^2)$

Náhodná veličina $X \sim (\mu, \sigma^2)$ má hustotu

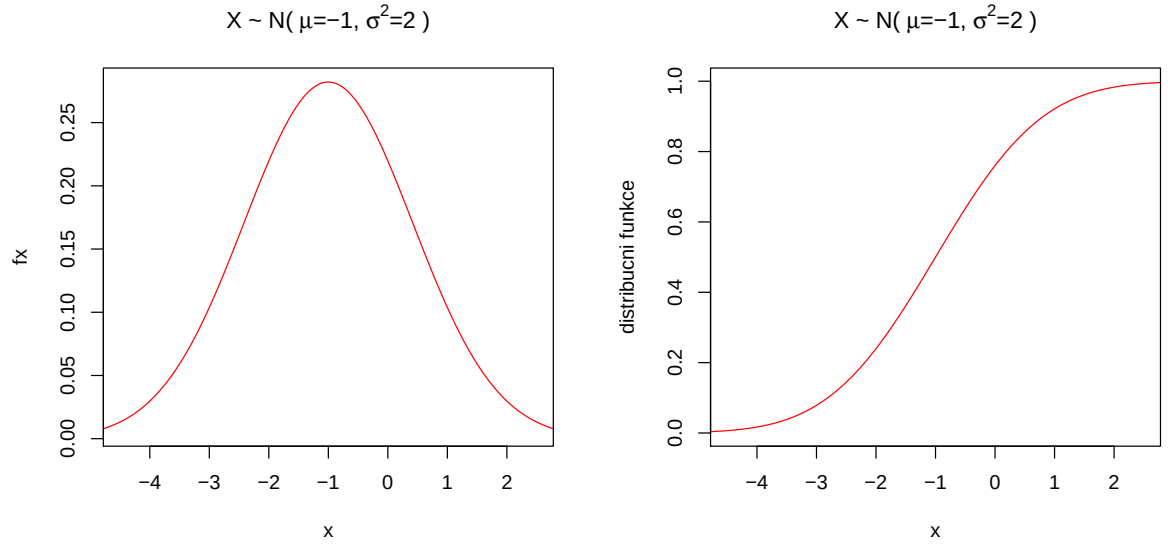
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{16}$$

Pro $\mu = 0$ a $\sigma^2 = 1$ se jedná o standardizované normální rozdělení, píšeme $U \sim \text{N}(0, 1)$. Hustota pravděpodobnosti má v tomto případě tvar

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \tag{17}$$

Příklad 4.7. Nakreslete graf hustoty a distribuční funkce náhodné veličiny $X \sim N(-1, 2)$.

```
#graf hustoty
x  <- seq(from=-5, to=3, length=512)
fx <- dnorm(x,  mean=-1, sd=sqrt(2))
plot(x, fx, type='l', main=bquote(paste('X ~ N( ',mu,'=-1, ',sigma^2,'=2 )')), xlim=c(-4.5,2.5),
     col='red')

#graf distribucni funkce
Fx <- pnorm(x, mean=-1, sd=sqrt(2))
plot(x, Fx, type='l', main=bquote(paste('X ~ N( ',mu,'=-1, ',sigma^2,'=2 )')), xlim=c(-4.5,2.5),
     ylab='distribucni funkce', col='red')
```



Příklad 4.8. Výsledky u přijímacích zkoušek na jistou VŠ jsou normálně rozděleny s parametry $\mu = 550$ bodů, $\sigma = 100$ bodů. S jakou pravděpodobností bude mít náhodně vybraný uchazeč alespoň 600 bodů?

X ... výsledek náhodně vybraného uchazeče; $X \sim N(550, 100^2)$.

$$\mathrm{P}(X \geq 600) = 1\text{–}\mathrm{P}(X \leq 600) = 1 - \mathrm{P}\left(\frac{X-\mu}{\sigma} \leq \frac{600-\mu}{\sigma}\right) = \mathrm{P}\left(U \leq \frac{600-550}{100}\right)$$
$$= 1 - \mathrm{P}(U \leq 0.5) = 1 - \Phi(0.5) = 1\text{–}0.69146 = 0.3085.$$

```
1-pnorm(600, mean=550, sd=100)
```

```
## [1] 0.3085375
```

```
1-pnorm(0.5, mean=0,   sd=1)
```

```
## [1] 0.3085375
```

Příklad 4.9. Životnost baterie v hodinách je náhodná veličina, která má normální rozdělení se střední hodnotou 300 hodin a směrodatnou odchylkou 35 hodin. Jaká je pravděpodobnost, že náhodně vybraná baterie bude mít životnost

a) alespoň 320 hodin?

b) nejvýše 310 hodin?

a) `## [1] 0.2838546`

b) `## [1] 0.6124515`

Příklad 4.10. Na výrobní lince jsou automaticky baleny balíčky rýže o deklarované hmotnosti 1000 g. Působením náhodných vlivů hmotnost balíčků kolísá. Lze ji považovat za náhodnou veličinu, která se řídí normálním rozdělením se střední hodnotou 996 g a směrodatnou odchylkou 18 g. Jaká je pravděpodobnost, že náhodně vybraný balíček rýže neprojde výstupní kontrolou, jestliže je povolená tolerance ±30 g od deklarované hmotnosti 1000 g?

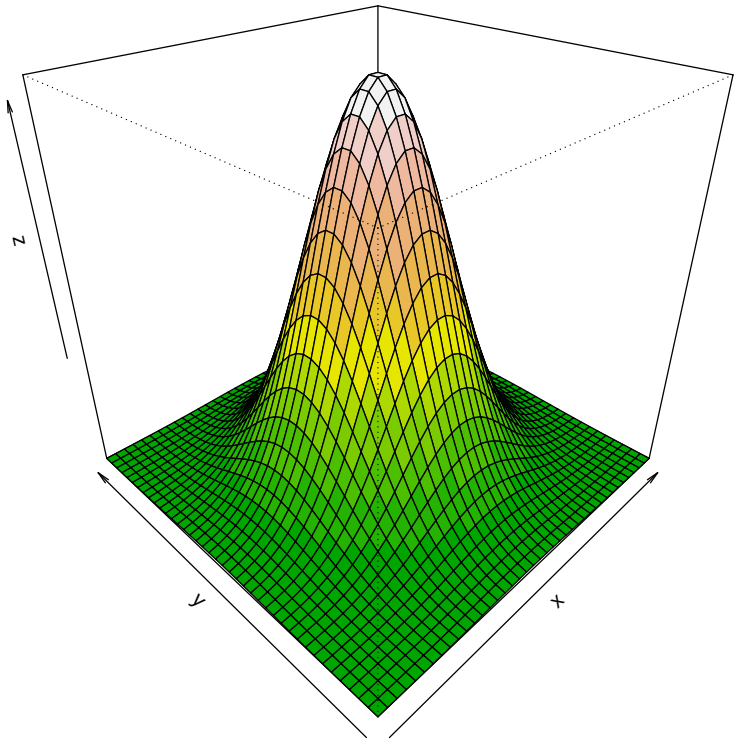
$$\mathrm{P}(X \notin \langle 970; 1030 \rangle) = 1 - \mathrm{P}(970 \leq X \leq 1030) = 1 - \mathrm{P}(970 < X < 1030)$$

```
## [1] 0.1037604
```

Příklad 4.11. Nakreslete graf hustoty dvourozměrného standardizovaného normálního rozdělení.

```
source('AS-funkce.R')
x  <- seq(from=-3, to=3, length=40)
y  <- seq(from=-3, to=3, length=40)
nx <- length(x)
ny <- length(y)
z  <- matrix(NA, nrow=nx, ncol=ny)
for(i in 1:nx){
  for(j in 1:ny){
```

```
    z[i,j]  <- norm2(x[i], y[j], mu1=0, mu2=0, sigma1=1, sigma2=1)
  }
}

color      <- terrain.colors(12)
stredy     <- (z[-1, -1] + z[-1, -ncol(z)] + z[-nrow(z), -1] + z[-nrow(z), -ncol(z)])/4
stredy.col <- cut(stredy, 12)
persp(x, y, z, col = color[stredy.col], phi = 30, theta = -45)
```



5 Výpočet číselných charakteristik náhodných veličin, aplikace Moivreovy – Laplaceovy věty

5.1 Kvantily vybraných spojitých rozdělení

α-kvantil náhodné veličiny X značíme x_α.

Normální rozdělení $N(\mu, \sigma^2)$

Náhodná veličina $X \sim N(\mu, \sigma^2)$ má hustotu

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Pro $\mu = 0$, $\sigma^2 = 1$ se jedná o standardizované normální rozdělení, píšeme $U \sim N(0, 1)$. Hustota standardizovaného normálního rozdělení má v tomto případě tvar

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}.$$

α-kvantil standardizovaného normálního rozdělení značíme u_α. Standardizované normální rozdělení je symetrické okolo nuly, proto pro kvantily tohoto rozdělení platí vztah

$$u_\alpha = -u_{1-\alpha}.$$

Poznámka: Vyjádření hustot následujících tří rozdělení je příliš složité, lze ho najít např. v příloze A skript Marie Budíková, Pavel Osecký, Štěpán Mikoláš: Teorie pravděpodobnosti a matematická statistika. Sbírka příkladů. MU Brno 2007.

χ^2 rozdělení s n stupni volnosti $\chi^2(n)$

Nechť $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, $X_i \sim N(0, 1)$, $i = 1, \ldots, n$. Pak náhodná veličina

$$X = X_1^2 + ... + X_n^2$$

má χ^2 rozdělení s n stupni volnosti

$$X \sim \chi^2(n).$$

α-kvantil χ^2 rozdělení s n stupni volnosti značíme $\chi^2_\alpha(n)$.

Studentovo rozdělení s n stupni volnosti $t(n)$

Nechť X_1, X_2 jsou stochasticky nezávislé náhodné veličiny, $X_1 \sim N(0, 1)$, $X_2 \sim \chi^2(n)$. Pak náhodná veličina

$$X = \frac{X_1}{\sqrt{\frac{X_2}{n}}}$$

má Studentovo rozdělení s n stupni volnosti

$$X \sim t(n).$$

α-kvantil Studentova rozdělení s n stupni volnosti značíme $t_\alpha(n)$. Studentovo rozdělení je symetrické okolo nuly, proto pro kvantily tohoto rozdělení platí vztah

$$t_\alpha(n) = -t_{1-\alpha}(n).$$

Fisherovo-Snedecorovo rozdělení s n_1 a n_2 stupni volnosti $F(n_1, n_2)$

Nechť $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, $X_i \sim \chi^2(n_i)$, $i = 1, 2$. Pak náhodná veličina

$$X = \frac{X_1/n_1}{X_2/n_2}$$

má Fishorovo rozdělení s n_1 a n_2 stupni volnosti

$$X \sim F(n_1, n_2).$$

α-kvantil Fisherova rozdělení s n_1 a n_2 stupni volnosti značíme $F_\alpha(n_1, n_2)$ Pro kvantily Fisherova rozdělení platí následující vztah

$$F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_1, n_2)}.$$

Příklad 5.1. Najděte medián a horní a dolní kvartil náhodné veličiny $U \sim N(0, 1)$.

```
qnorm(0.5)
```

```
## [1] 0
```

```
qnorm(0.25)
```

```
## [1] -0.6744898
```

```
qnorm(0.75)
```

```
## [1] 0.6744898
```

Příklad 5.2. Najděte dolní kvartil náhodné veličiny $X \sim N(3, 5)$.

```
qnorm(0.25, 3, sqrt(5))
```

```
## [1] 1.491795
```

Příklad 5.3. Určete kvantil $\chi^2_{0.025}(25)$.

```
qchisq(0.025, 25)
```

```
## [1] 13.11972
```

Příklad 5.4. Určete kvantily $t_{0.99}(30)$ a $t_{0.05}(14)$.

```
qt(0.99, 30)
```

```
## [1] 2.457262
```

```
qt(0.05, 14)
```

```
## [1] -1.76131
```

Příklad 5.5. Určete kvantily $F_{0.975}(5, 20)$ a $F_{0.05}(2, 10)$.

```r
qf(0.975, 5,20)

## [1] 3.289056

qf(0.05,  2,10)

## [1] 0.0515573
```

5.2 Výpočet střední hodnoty a rozptylu diskrétních náhodných veličin

Příklad 5.6. Postupně se zkouší spolehlivost čtyř přístrojů. Další se zkouší jen tehdy, když předchozí je spolehlivý. Každý z přístrojů vydrží zkoušku s pravděpodobností 0.8. Náhodná veličina X udává počet zkoušených přístrojů. Vypočtěte střední hodnotu a rozptyl náhodné veličiny X.

X nabývá hodnot $1, 2, 3, 4$ a její pravděpodobnostní funkce je $\pi(1) = 0.2$, $\pi(2) = 0.8 * 0.2 = 0.16$, $\pi(3) = 0.8^2 * 0.2 = 0.128$, $\pi(4) = 0.8^3 * 0.2 + 0.8^4 = 0.512$, $\pi(0) = 0$ jinak.

$E(X) = 1 * 0.2 + 2 * 0.16 + 3 * 0.128 + 4 * 0.512 = 2.952$

$D(X) = 1^2 * 0.2 + 2^2 * 0.16 + 3^2 * 0.128 + 4^2 * 0.512–2.9522^2 = 1.4697$

```r
x <- 1:4
pi <- c(0.2, 0.8*0.2, 0.8^2*0.2, 0.8^3*0.2+0.8^4)
(EX <- sum(x*pi))

## [1] 2.952

(DX <- sum(x^2*pi)-EX^2)

## [1] 1.469696
```

Příklad k samostatnému řešení

Příklad 5.7. Náhodná veličina X udává počet ok při hodu kostkou. Vypočtěte její střední hodnotu a rozptyl.

```r
x <- 1:6
pi <- rep(1/6, 6)
(EX <- sum(x*pi))

## [1] 3.5

(DX <- sum(x^2*pi)-EX^2)

## [1] 2.916667
```

5.3 Výpočet koeficientu korelace diskrétních náhodných veličin

Příklad 5.8. Náhodná veličina X udává příjem manžela (v tisících dolarů) a náhodná veličina Y příjem manželky (v tisících dolarů). Je známa simultánní pravděpodobnostní funkce $\pi(x, y)$ diskrétního náhodného vektoru (X, Y): $\pi(10, 10) = 0.2$, $\pi(10, 20) = 0.04$, $\pi(10, 30) = 0.01$, $\pi(10, 40) = 0$, $\pi(20, 10) = 0.1$, $\pi(20, 20) = 0.36$, $\pi(20, 30) = 0.09$, $\pi(20, 40) = 0$, $\pi(30, 10) = 0$, $\pi(30, 20) = 0.05$, $\pi(30, 30) = 0.1$, $\pi(30, 40) = 0$, $\pi(40, 10) = 0$, $\pi(40, 20) = 0$,

$\pi(40,30) = 0$, $\pi(40,40) = 0.05$, $\pi(x,y) = 0$ jinak. Vypočtěte koeficient korelace příjmů manžela a manželky.

Náhodná veličina X i náhodná veličina Y nabývají hodnot $10, 20, 30, 40$.
Stanovíme hodnoty marginálních pravděpodobnostních funkcí: $\pi_1(10) = 0.25$, $\pi_1(20) = 0.55$, $\pi_1(30) = 0.15$, $\pi_1(40) = 0.05$, $\pi_1(x) = 0$ jinak, $\pi_2(10) = 0.3$, $\pi_2(20) = 0.45$, $\pi_2(30) = 0.2$, $\pi_2(10) = 0.05$, $\pi_2(y) = 0$ jinak. Všechny hodnoty si zapíšeme do tabulky simultánních a marginálních pravděpodobností.

Tabulka pravděpodobnostních funkcí $\pi(X,Y)$					
X - příjem manžela	Y - příjem manželky				
	10	20	30	40	$\sum$
10	0.2	0.04	0.01	0	0.25
20	0.1	0.36	0.09	0	0.55
30	0	0.05	0.1	0	0.15
40	0	0	0	0.05	0.05
$\sum$	0.3	0.45	0.2	0.05	1

Spočteme $E(X) = 20$, $E(Y) = 20$, $D(X) = 60$, $D(Y) = 70$. Dosazením do vzorce pro výpočet kovariance zjistíme, že $C(X,Y) = 49$, tedy koeficient korelace $R(X,Y) = \frac{49}{\sqrt{60}\sqrt{70}} = 0.76$.

```
x <- c(10, 20, 30, 40)
y <- c(10, 20, 30, 40)
n <- length(x)
pi <- data.frame(Deset=   c(0.2,  0.1,  0,    0),
                 Dvacet=  c(0.04, 0.36, 0.05, 0),
                 Tricet=  c(0.01, 0.09, 0.1,  0),
                 Ctyricet=c(0,    0,    0,    0.05),
                 row.names=c('Deset', 'Dvacet', 'Tricet', 'Ctyricet'))
pix <- apply(pi, 1, sum)
piy <- apply(pi, 2, sum)

(EX <- sum(x*pix))

## [1] 20

(EY <- sum(y*piy))

## [1] 20

(DX <- sum(x^2*pix)-EX^2)

## [1] 60

(DY <- sum(y^2*piy)-EY^2)

## [1] 70

(CXY <- sum(c((x-EX)*(y-EY)[1], (x-EX)*(y-EY)[2], (x-EX)*(y-EY)[3], (x-EX)*(y-EY)[4])
            * c(as.matrix(pi))))

## [1] 49

(RXY <- CXY/(sqrt(DX)*sqrt(DY)))

## [1] 0.7560864
```

Příklady k samostatnému řešení

Příklad 5.9. Objektem zájmu rozsáhlé studie bylo sledování pohřebního rituálu dnes již vymřelého, ale v minulosti velmi dlouho přetrvávajícího a rozsáhlého jihoamerického kmene. Součástí pohřebního rituálu tohoto kmene bylo odsekávání článků prstů na rukou a nohou zemřelého a jejich následné obětování bohům jako dar, aby zemřelého přijali mezi sebe. Zemřelému tak byl na ruce odetnut buď jeden nebo dva prsty a na noze tři nebo čtyři prsty.

Dále bylo zjištěno, že domorodci odtínali jeden prst na ruce a tři prsty na noze zemřelého s pravděpodobností 0.1, dva prsty na ruce a tři prsty na noze s pravděpodobností 0.3, jeden prst na ruce a čtyři prsty na noze s pravděpodobností 0.35 a dva prsty na ruce a čtyři prsty na noze s pravděpodobností 0.25. Určete korelaci znaků X – počet odetnutých prstů na rukou a Y – počet odetnutých prstů na nohou.

```
##                   EX   EY   DX     DY   CXY     RXY
## Charakteristiky 1.6 3.55 0.24 0.2475 -0.08 -0.3282
```

Z tabulky výsledků vidíme, že střední hodnota počtu prstů odetnutých na rukou je 1.6, zatímco střední hodnota počtu prstů odetnutých na nohou je 3.55. Rozptyl počtu prstů odetnutých na rukou je 0.24 a rozptyl počtu prstů odetnutých na nohou je 0.2475. Kovarince mezi znaky X a Y nabývá hodnoty -0.08. Hodnota korelačního koeficientu vyšla -0.3282, což značí, že mezi počtem prstů odetnutých na rukou a počtem prstů odetnutých na nohou existuje mírný stupeň nepřímé lineární závislosti.

Příklad 5.10. Diskrétní náhodný vektor (X, Y) má simultánní pravděpodobnostní funkci s hodnotami $\pi(0,-1) = c$, $\pi(0,0) = \pi(0,1) = \pi(1,-1) = \pi(2,-1) = 0$, $\pi(1,0) = \pi(1,1) = \pi(2,1) = 2c$, $\pi(2,0) = 3c$, $\pi(x,y) = 0$ jinak. Určete konstantu c a vypočtěte $R(X,Y)$.

Tabulka pr. funkcí $\pi(X,Y)$				
X	Y			
	-1	0	1	$\sum$
0	c	0	0	c
1	0	$2c$	$2c$	$4c$
2	0	$3c$	$2c$	$5c$
$\sum$	c	$5c$	$4c$	$10c$=1

```
##                   EX  EY   DX   DY  CXY    RXY
## Charakteristiky 1.4 0.3 0.44 0.41 0.18 0.4238
```

Střední hodnota náhodné veličiny X je 1.4, střední hodnota náhodné veličiny Y je 0.3. Rozptyl náhodné veličiny X nabývá hodnoty 0.44, rozptyl náhodné veličiny Y nabývá hodnoty 0.41. Kovariance mezi veličinami X a Y je 0.18 a korelační koeficient nabývá hodnoty 0.4238, což značí, že mezi znaky X a Y existuje mírný stupeň přímé lineární závislosti.

Příklad 5.11. Zkoumali jsme potomky kosmanů. Náhodná veličina X udává počet manželských potomků, které samice porodila a náhodná veličina Y počet nemanželských potomků, které samice porodila. Je známa simultánní pravděpodobnostní funkce $\pi(x,y)$ diskrétního náhodného vektoru (X,Y):

Tabulka simultánní pstní fce $\pi(X,Y)$			
X - počet manž.p.	Y - počet nemanž.p.		
	1	2	3
1	0.2	0.04	0.01
2	0.15	0.36	0.09
3	0.05	0.1	0.0

Vypočtěte koeficient korelace manželských a nemanželských potomků.

```
##                   EX  EY   DX   DY  CXY    RXY
## Charakteristiky 1.9 1.7 0.39 0.41 0.11 0.2751
```

Střední hodnota počtu manželských potomků kosmanů je 1.9, střední hodnota počtu nemanželských potomků je 1.6. Rozptyl manželských potomků je 0.39, rozptyl nemanželských potomků je 0.41. Kovariance mezi počtem manželských a nemanželských potomků je 0.11. Korelační koeficient nabývá hodnoty 0.2751, což znamená, že mezi počtem manželských a nemanželských potomků kosmanů existuje nízký stupeň přímé lineární závislosti.

5.4 Aplikace Moivreovy-Laplaceovy věty

Příklad 5.12. Pravděpodobnost úspěchu při jednom pokusu je 0.3. S jakou pravděpodobností lze tvrdit, že počet úspěchů ve 100 pokusech bude v mezích od 20 do 40? Výpočet proveďte

a) přesně;

b) pomocí aproximace normálním rozdělením.

```
# a)
sum(dbinom(20:40, 100, 0.3))

## [1] 0.9786144

pbinom(40, 100, 0.3)-pbinom(19, 100, 0.3)

## [1] 0.9786144

# b)
pnorm(40, 100*0.3, sqrt(100*0.3*0.7))-pnorm(19, 100*0.3, sqrt(100*0.3*0.7))

## [1] 0.9772632
```

Příklad 5.13. Pravděpodobnost, že zakoupený elektrospotřebič bude vyžadovat opravu během záruční doby, je rovna 0.2. Jaká je pravděpodobnost, že během záruční doby bude nutno ze 400 prodaných spotřebičů opravit více než 96? Výpočet proveďte

a) přesně;

b) pomocí aproximace normálním rozdělením.

```
# a)
1-pbinom(96, 400, 0.2)

## [1] 0.02138855

1-pnorm(96, 400*0.2, sqrt(400*0.2*0.8))

## [1] 0.02275013
```

Výsledek: ad a) 0.0246, ad b) 0.0228

Příklad k samostatnému řešení

Příklad 5.14. Pravděpodobnost, že určitý typ výrobku má výrobní vadu, je 0.05. Jaká je pravděpodobnost, že ze série 1000 výrobků bude mít výrobní vadu nejvýše 70? Výpočet proveďte

a) přesně;

b) pomocí aproximace normálním rozdělením.

```
# a)
pbinom(70, 1000, 0.05)

## [1] 0.9976697

# b)
pnorm(70, 1000*0.05, sqrt(1000*0.05*0.95))

## [1] 0.9981455
```

6 Základní pojmy matematické statistiky

6.1 Bodové odhady parametrů

Příklad 6.1. Ve 12-ti náhodně vybraných internetových obchodech byly zjištěny následující ceny deskriptoru artefaktů (v Kč): 102, 99, 106, 103, 96, 98, 100, 105, 103, 98, 104, 107. Těchto 12 hodnot považujeme za realizace náhodného výběru $X_1, \ldots, X_{12}$ z rozdělení, které má střední hodnotu μ a rozptyl σ^2.

a) Určete nestranné bodové odhady neznámé střední hodnoty μ a neznámého rozptylu σ^2.

b) Najděte výběrovou distribuční funkci $F_{12}(x)$ a nakreslete její graf.

ad a) Vypočteme realizaci výběrového průměru

$$m = \frac{1}{12}(102 + 99 + \cdots + 107) = 101.75\,\text{Kč}$$

Vypočteme realizaci výběrového rozptylu:

$$s^2 = \frac{1}{11}\left[(102 - 101.75)^2 + (99 - 101.75)^2 + \cdots + (107 - 101.75)^2\right] = 12.39\,\text{Kč}^2$$

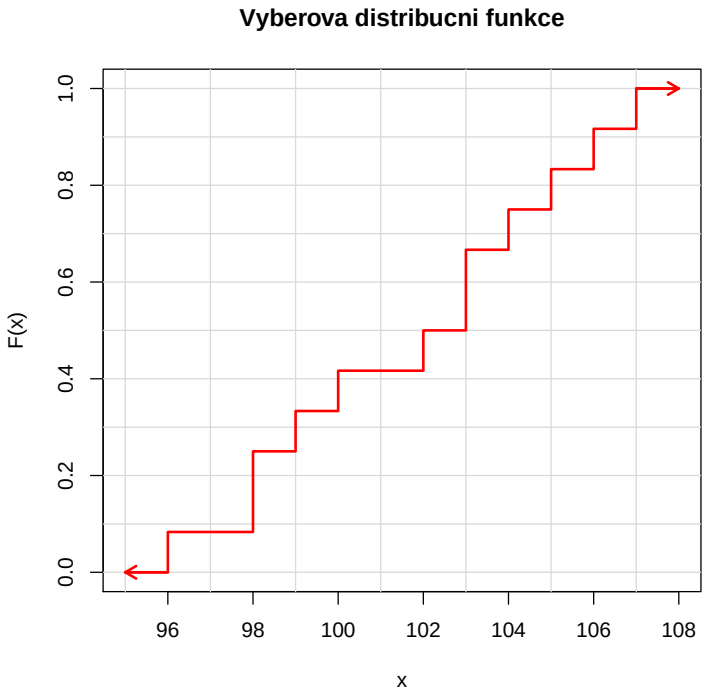
```
x <- c(96, 98, 98, 99, 100, 102, 103, 103, 104, 105, 106, 107)
n <- length(x)
(m <- mean(x))

## [1] 101.75

(s2 <- var(x))

## [1] 12.38636
```

ad b) Pro usnadnění výpočtu hodnot výběrové distribuční funkce $F_{12}(x)$ uspořádáme ceny podle velikosti: 96, 98, 98, 99, 100, 102, 103, 103, 104, 105, 106, 107. Číselnou osu rozdělíme na 11 intervalů a v každém intervalu

stanovíme hodnotu výběrové distribuční funkce:

$$
\begin{aligned}
&x < 96 : F_{12}(x) = 0 \\
&96 \leq x < 98 : F_{12}(x) = \frac{1}{12} = 0.08\bar{3} \\
&98 \leq x < 99 : F_{12}(x) = \frac{3}{12} = 0.25 \\
&99 \leq x < 100 : F_{12}(x) = \frac{4}{12} = 0.\bar{3} \\
&100 \leq x < 102 : F_{12}(x) = \frac{5}{12} = 0.41\bar{6} \\
&102 \leq x < 103 : F_{12}(x) = \frac{6}{12} = 0.5 \\
&103 \leq x < 104 : F_{12}(x) = \frac{8}{12} = 0.\bar{6} \\
&104 \leq x < 105 : F_{12}(x) = \frac{9}{12} = 0.75 \\
&105 \leq x < 106 : F_{12}(x) = \frac{10}{12} = 0.8\bar{3} \\
&106 \leq x < 107 : F_{12}(x) = \frac{11}{12} = 0.91\bar{6} \\
&x \geq 107 : F_{12}(x) = 1
\end{aligned}
$$

```
# Vyberova distribucni funkce
t  <- unique(sort(x))
y  <- sort(x)
nt <- length(t)

cetnost <- NULL
for(i in 1:nt){
  cetnost[i] <- sum(y<=t[i])}
Fx <- cetnost/n
round(Fx, digits=4)

##  [1] 0.0833 0.2500 0.3333 0.4167 0.5000 0.6667 0.7500 0.8333 0.9167 1.0000

# graf výběrové distribuční funkce F(x)
x <- c(min(t)-1,t, max(t)+1)
y <- c(0,Fx,1)
plot(x, y, type='n', xlab='x', ylab='F(x)',
     main='Vyberova distribucni funkce')
abline(h=seq(0,1,by=0.1), col='grey85')
abline(v=seq(95, 108,by=2), col='grey85')
lines(x,y, type='s', col='red', lwd=2)
arrows(96,0,95,0, col='red', lwd=2, length=0.1)
arrows(107,1,108,1, col='red', lwd=2, length=0.1)
```



Příklad 6.2. Přírůstky cen akcií v % na burze v New Yorku u 10 náhodně vybraných společností dosáhly těchto hodnot: $10, 16, 5, 10, 12, 8, 4, 6, 5, 4$.

a) Odhadněte střední hodnotu, rozptyl a směrodatnou odchylku růstu cen akcií.

b) Odhadněte pravděpodobnost růstu cen akcií aspoň o 8.5 %.

ad a)
```
x <- c(10, 16, 5, 10, 12, 8, 4, 6, 5, 4)
x <- sort(x)
n <- length(x)
s2 <- var(x)
s  <- sd(x)
Tab <- data.frame(m=m, s2=s2, s=s, row.names='akcie')
round(Tab, digits=2)

##            m    s2    s
## akcie 101.75 15.78 3.97
```

ad b)

$$P(X \geq 8.5) = \frac{n_{x \geq 8.5}}{n}$$

$$P(X \geq 8.5) = 1 - \frac{n_{x < 8.5}}{n}$$

```
# P(X>=8.5)
pst <- sum(x>=8.5)/length(x)
pst <- 1-sum(x<8.5)/length(x)
round(pst,4)

## [1] 0.4
```

Průměrný růst cen akcií odhadujeme na 8 % se směrodatnou odchylkou 3.97 %. Dále, u 40 % akcií vzrostla cena alespoň o 8.5 %.

Příklad 6.3. Bylo zkoumáno 9 vzorků půdy s různým obsahem fosforu (veličina X). Hodnoty veličiny Y označují obsah fosforu v obilných klíčcích (po 38 dnech), jež vyrostly na těchto vzorcích půdy.

číslo vzorku	1	2	3	4	5	6	7	8	9
X	1	4	5	9	11	13	23	23	28
Y	64	71	54	81	76	93	77	95	109

Těchto 9 dvojic hodnot považujeme za realizace náhodného výběru $(X_1, Y_1), \ldots, (X_9, Y_9)$ z dvourozměrného rozdělení s kovariancí σ_{12} a koeficientem korelace ρ. Najděte bodové odhady kovariance σ_{12} a koeficientu korelace ρ.

```
x <- c(1, 4, 5, 9, 11, 13, 23, 23, 28)
y <- c(64, 71, 54, 81, 76, 93, 77, 95, 109)
cov(x,y)

## [1] 130

cor(x,y)

## [1] 0.8049892
```

Výběrová kovariance veličin X, Y se realizuje hodnotou 130. Výběrový koeficient korelace veličin X, Y nabyl hodnoty 0.805, tedy mezi veličinami X, Y existuje silná přímá lineární závislost.

6.2 Intervalové odhady parametrů

Nechť $X_1, \ldots, X_n$ je náhodný výber z rozdělení $L(\theta)$, θ je sledovaný parametr a $\alpha \in (0, 1)$.

Interval (D, H) se nazývá $100(1-\alpha)\%$ *oboustranný interval spolehlivosti* parametru θ, pokud pro každé $\theta \in \Theta$ platí

$$\mathrm{P}(D < \theta < H) = 1 - \alpha.$$

Interval (D, ∞) se nazývá $100(1-\alpha)\%$ *levostranný interval spolehlivosti* parametru θ, pokud pro každé $\theta \in \Theta$ platí

$$\mathrm{P}(D < \theta) = 1 - \alpha.$$

Interval $(-\infty, H)$ se nazývá $100(1-\alpha)\%$ *pravostranný interval spolehlivosti* parametru θ, pokud pro každé $\theta \in \Theta$ platí

$$\mathrm{P}(\theta < H) = 1 - \alpha.$$

Číslo α se nazývá *riziko*, číslo $1 - \alpha$ se nazývá *spolehlivost*.

Příklad 6.4. Při kontrolních zkouškách životnosti 16 žárovek byl stanoven odhad $m = 3000\,h$ střední hodnoty jejich životnosti. Z dřívějších zkoušek je známo, že životnost žárovky se řídí normálním rozdělením se směrodatnou odchylkou $\sigma = 20\,h$. Vypočtěte

a) 99% empirický interval spolehlivosti pro střední hodnotu životnosti;

b) 90% levostranný empirický interval spolehlivosti pro střední hodnotu životnosti;

c) 95% pravostranný empirický interval spolehlivosti pro střední hodnotu životnosti.

Výsledek zaokrouhlete na jedno desetinné místo a vyjádřete v hodinách a minutách.

ad a)

$$d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 3000 - \frac{20}{\sqrt{16}} 2.57583 = 2987.1$$
$$h = m - \frac{\sigma}{\sqrt{n}} u_{\alpha} = 3000 + \frac{20}{\sqrt{16}} 2.57583 = 3012.9$$

```
m <- 3000
s <- 20
n <- 16

# a)
alpha <- 0.01
(dh <- m-s/sqrt(n)*qnorm(1-alpha/2))

## [1] 2987.121

(hh <- m-s/sqrt(n)*qnorm(alpha/2))

## [1] 3012.879
```

2987 h a 6 min $< \mu <$ 3012 h a 54 min s pravděpodobnostní 0.99.

ad b)

$$d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 3000 - \frac{20}{\sqrt{16}} 1.28155 = 2993.6$$

```
alpha <- 0.1
(dh <- m-s/sqrt(n)*qnorm(1-alpha))

## [1] 2993.592
```

2993 h a 36 min $< \mu$ s pravděpodobnostní 0.9.

ad c)

$$h = m - \frac{\sigma}{\sqrt{n}} u_{\alpha} = 3000 + \frac{20}{\sqrt{16}} 1.95996 = 3008.2$$

```
alpha <- 0.05
(hh <- m-s/sqrt(n)*qnorm(alpha))

## [1] 3008.224
```

3008 h a 13 min $> \mu$ s pravděpodobnostní 0.95.

Užitečný odkaz: na adrese http://www.prevody-jednotek.cz je program, s jehož pomocí lze převádět různé fyzikální jednotky, v našem případě hodiny na minuty.

6.3 Testování hypotézy

Příklad 6.5. Víme, že výška hochů ve věku 9.5 až 10 let má normální rozdělení s neznámou střední hodnotou μ a známým rozptylem $\sigma^2 = 39.112$ cm^2. Dětský lékař náhodně vybral 15 hochů uvedeného věku, změřil je a vypočítal realizaci výběrového průměru $m = 139.13$ cm. Podle jeho názoru by výška hochů v tomto věku neměla přesáhnout 142 cm s pravděpodobností 0.95. Lze tvrzení lékaře akceptovat?

Testujeme $H_0 : \mu \geq 142$ proti $H_1 : \mu < 142$ na hladině významnosti $\alpha = 0.05$.

a) Test provedeme pomocí kritického oboru.
Pro úlohy o střední hodnotě normálního rozdělení při známém rozptylu používáme pivotovou statistiku

$$U = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

Testovací statistika bude mít tedy tvar

$$T_0 = \frac{M - c}{\frac{\sigma}{\sqrt{n}}}$$

a za platnosti nulové hypotézy H_0 se tato statistika bude řídit standardizovaným normálním rozdělením

$$T_0 \sim N(0, 1).$$

Vypočítáme realizaci testového kritéria:

$$t_0 = \frac{139.13 - 142}{\frac{\sqrt{39.112}}{\sqrt{15}}} = -1.7773.$$

```
sigma <- sqrt(39.112)
n <- 15
m <- 139.13
c <- 142
alpha <- 0.05

# ad a)
t0 <- (m-c)/(sigma/sqrt(n))
qnorm(alpha)

## [1] -1.644854
```

Stanovíme kritický obor: $W \in (-\infty, u_\alpha\rangle = (-\infty, u_{0.05}\rangle = (-\infty, -u_{0.95}\rangle = (-\infty, -1.6449\rangle$. Protože $-1.7773 \in W$, nulovou hypotézu H_0 zamítáme na hladině významnosti $\alpha = 0.05$. Tvrzení lékaře lze tedy akceptovat s rizikem omylu 5 %.

b) Test provedeme pomocí intervalu spolehlivosti.
Meze $100(1-\alpha)\%$ empirického pravostranného intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2 jsou

$$(-\infty, h) = (-\infty, m - \frac{\sigma}{\sqrt{n}} u_\alpha).$$

V našem případě dostáváme:

$$h = 139.13 - \frac{\sqrt{39.112}}{\sqrt{15}} u_{0.05} = 139.13 + \frac{\sqrt{39.112}}{\sqrt{15}} 1.645 = 141.79.$$

```
(hh <- m-(sigma/sqrt(n))*qnorm(alpha))

## [1] 141.7861
```

Protože $142 \notin (-\infty; 141.79)$, H_0 zamítáme na hladině významnosti 0.05.

c) Test provedeme pomocí p-hodnoty.

$$p = \mathrm{P}(T_0 \leq t_0) = \Phi(-1.7773) = 0.0378$$

```
(pval <- pnorm(t0))

## [1] 0.03775549
```

Jelikož $0.0378 \leq 0.05$, nulovou hypotézu zamítáme na hladině významnosti 0.05.

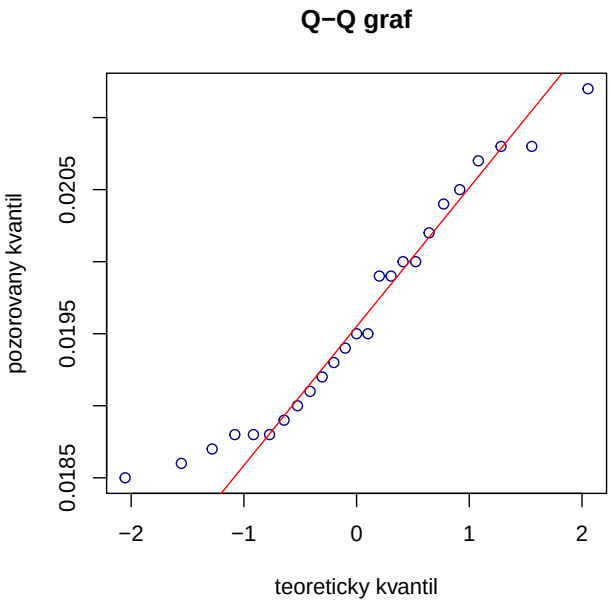
7 Ověřování normality a parametrické úlohy o jednom náhodném výběru z normálního rozdělení a dvourozměrného rozdělení

7.1 Grafické ověřování normality

Příklad 7.1. Při nanášení tenkých kovových vrstev stříbra na polymerní materiál se vyžaduje, aby tloušťka vrstvy byla 0.020 μm. Pomocí atomové absorpční spektroskopie se zjistily hodnoty, jež jsou uvedeny v tabulce a uloženy v souboru vrstva_stribra.txt. Posuďte Q-Q grafem, zda se výsledky měření řídí normálním rozdělením.

tloušťka				
0.0212	0.0186	0.0192	0.0207	0.0200
0.0200	0.0190	0.0188	0.0208	0.0194
0.0188	0.0193	0.0204	0.0185	0.0187
0.0195	0.0191	0.0195	0.0199	0.0205
0.0189	0.0188	0.0199	0.0202	0.0208

```
data <- read.delim('vrstva_stribra.txt')
stribro <- data$tloustka_vrstvy
qqnorm(stribro, col='darkblue', xlab='teoreticky kvantil',
       ylab='pozorovany kvantil', main='Q-Q graf')
qqline(stribro, col='red')
```



Dle vzhledu Q-Q grafu lze soudit, že data vykazují jen lehké odchylky od normality.

7.2 Testy normality

Příklad 7.2. U 48 studentek VŠE v Praze byla zjišťována výška a obor studia (1 – národní hospodářství, 2 -– informatika). Hodnoty jsou uloženy v souboru vyska.txt. Pomocí Lilieforsovy modifikace K-S testu, pomocí S-W testu a pomocí A-D testu testujte na hladině významnosti $\alpha = 0.05$ hypotézu, že data pochází z normálního rozdělení. Pomocí Q-Q grafu posuďte vizuálně předpoklad normality.

```
library(nortest)
data <- read.table('vyska.txt', header=T, row.names=NULL)
head(data)

##   vyska obor
## 1   165    1
## 2   170    1
## 3   170    1
## 4   179    1
## 5   170    1
## 6   168    1

vyska <- data$vyska
shapiro.test(vyska)

##
##  Shapiro-Wilk normality test
##
## data:  vyska
## W = 0.966, p-value = 0.176

lillie.test(vyska)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

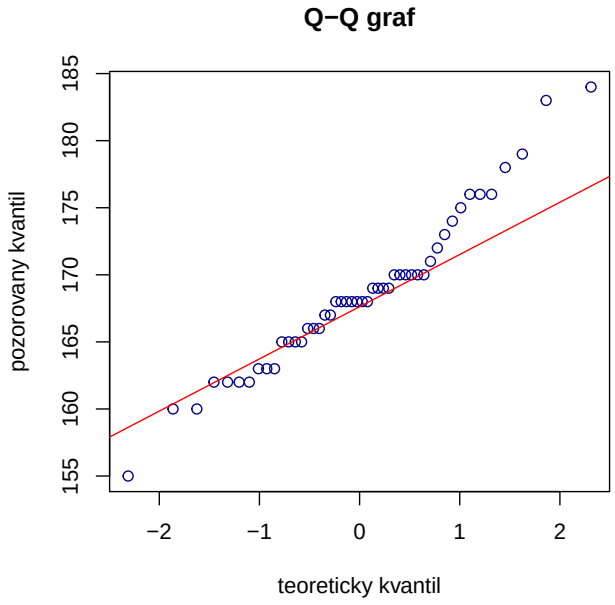
```
##
## data:  vyska
## D = 0.15562, p-value = 0.005258

ad.test(vyska)

##
##  Anderson-Darling normality test
##
## data:  vyska
## A = 0.66099, p-value = 0.07933
```

Shrnutí výsledků

```
st  <- shapiro.test(vyska)
lt  <- lillie.test(vyska)
at  <- ad.test(vyska)
(tab <- data.frame(test.statistika=round(c(lt$statistic, st$statistic, at$statistic),6),
                   p.value=round(c(lt$p.value, st$p.value, at$p.value),6),
                   rozhodnuti=c('zamitame', 'nezamitame', 'nezamitame'),
                   row.names=c('Lillie.test', 'S-W.test', 'A-D.test')))

##             test.statistika  p.value rozhodnuti
## Lillie.test        0.155621 0.005258   zamitame
## S-W.test           0.965996 0.176031 nezamitame
## A-D.test           0.660990 0.079330 nezamitame
```

Q-Q graf

```
qqnorm(vyska, col='darkblue', xlab='teoreticky kvantil', ylab='pozorovany kvantil', main='Q-Q graf')
qqline(vyska, col='red')
```



Tečky se řadí podél ideální přímky, normalita je jen lehce porušena.

Příklad k samostatnému řešení

Příklad 7.3. Testy normality a grafické ověření normality proveďte

a) pro výšku studentek oboru národní hospodářství;

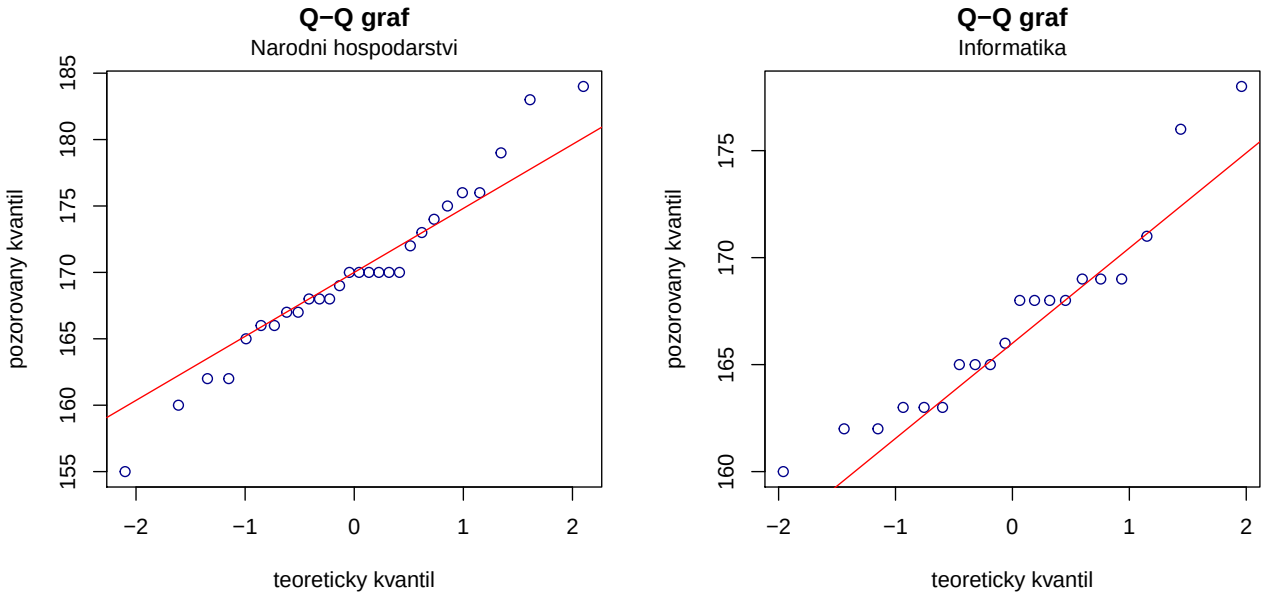
b) pro výšku studentek oboru informatika.

ad a)
```
##             test.statistika  p.value rozhodnuti
## Lillie.test        0.167473 0.042926   zamitame
## S-W.test           0.970969 0.606793 nezamitame
## A-D.test           0.419238 0.305268 nezamitame
```

Vidíme, že Lilieforsova varianta K-S testu zamítá hypotézu o normalitě na hladině významnosti $\alpha = 0.05$ (p-hodnota je menší než 0.05), zatímco S-W test hypotézu o normalitě nezamítá (p-hodnota je větší než 0.05). A-D test poskytne hodnotu testové statistiky 0.4192, odpovídající p-hodnota je 0.3053, tedy A-D test nezamítá hypotézu o normalitě na hladině významnosti $\alpha = 0.05$.

ad b)
```
##             test.statistika  p.value rozhodnuti
## Lillie.test        0.172301 0.123974 nezamitame
## S-W.test           0.922747 0.111924 nezamitame
## A-D.test           0.566019 0.123709 nezamitame
```

V případě b) ani jeden z testů hypotézu o normalitě nezamítá na hladině významnosti $\alpha = 0.05$.



7.3 Parametrické úlohy o jednom náhodném výběru z normálního rozdělení

Upozornění: Pokud to povaha úlohy vyžaduje, proveďte test normality dat.

Příklad 7.4. Vlastnosti výběrového průměru z normálního rozdělení: Předpokládejme, že velký ročník na vysoké škole má výsledky ze statistiky normálně rozděleny kolem střední hodnoty 72 bodů se směrodatnou odchylkou 9 bodů. Najděte pravděpodobnost, že průměr výsledků náhodného výběru 10 studentů bude větší než 80 bodů.

Návod:
$X_1, \ldots, X_{10}$ je náhodný výběr z $N(72, 81)$. Počítáme $\mathrm{P}(M > 80)$, přičemž výběrový průměr M má normální rozdělení se střední hodnotou $E(M) = \mu = 72$ a rozptylem $D(M) = \frac{\sigma^2}{n} = \frac{81}{10} = 8.1$. Tedy $\mathrm{P}(M > 80) = 1 - P(M \leq 80) = 1–\Phi(80)$, kde $\Phi(80)$ je hodnota distribuční funkce rozdělení $N(72, 8.1)$ v bodě 80.

```
1-pnorm(80, 72, sqrt(8.1))

## [1] 0.002470053
```

Příklad 7.5. Intervaly spolehlivosti pro parametry μ, σ^2 normálního rozdělení: Z populace stejně starých selat téhož plemene bylo vylosováno šest selat a po dobu půl roku jim byla podávána táž výkrmná dieta. Byly zaznamenávány průměrné denní přírůstky hmotnosti v dkg. Z dřívějších pokusů je známo, že v populaci mívají takové přírůstky normální rozdělení, avšak střední hodnota i rozptyl se měnívají. Přírůstky v dkg: 62, 54, 55, 60, 53, 58.

a) Najděte 95% empirický levostranný interval spolehlivosti pro neznámou střední hodnotu μ při neznámé směrodatné odchylce σ.

b) Najděte 95% empirický interval spolehlivosti pro směrodatnou odchylku σ.

Shapirův - Wilkův test normality:

```
x <- c(62, 54, 55, 60, 53, 58)
shapiro.test(x)$p.val

## [1] 0.6194994
```

P-hodnota S-W testu je $0.6195 > 0.05$, tedy nulovou hypotézu o normálním rozdělení náhodného výběru nezamítáme na hladině významnosti $\alpha = 0.05$.

Výpočet intervalů spolehlivosti:

ad a)
```
x <- c(62, 54, 55, 60, 53, 58)
m <- mean(x)
s <- sd(x)
n <- length(x)
alpha <- 0.05
dd <- m-s/sqrt(n)*qt(1-alpha, n-1)
print(paste('dd =', round(dd, 4)))

## [1] "dd = 54.0568"
```

$$IS = (54.0568\,; \infty)$$

$\mu > 54.06$ dkg s pravděpodobností 0.95.

ad b)
```
dh <- (n-1)*s^2/qchisq(1-alpha/2, n-1)
hh <- (n-1)*s^2/qchisq(alpha/2, n-1)
print(paste('dh =', round(sqrt(dh), 4)))

## [1] "dh = 2.2332"

print(paste('hh =', round(sqrt(hh), 4)))

## [1] "hh = 8.7747"
```

$$IS = (2.2332\,;\,8.7747)$$

$2.23\,\text{dkg} < \sigma < 8.77\,\text{dkg}$ s pravděpodobností 0.95.

Příklad 7.6. Testování hypotézy o střední hodnotě μ: Systematická chyba měřicího přístroje se eliminuje nastavením přístroje a měřením etalonu, jehož správná hodnota je $\mu = 10.00$. Nezávislými měřeními za stejných podmínek byly získány hodnoty: 10.24, 10.12, 9.91, 10.19, 9.78, 10.14, 9.86, 10.17, 10.05, které považujeme za realizace náhodného výběru rozsahu 9 z rozdělení $N(\mu, \sigma^2)$. Je možné při riziku 0.05 vysvětlit odchylky od hodnoty 10.00 působením náhodných vlivů?

Shapirův - Wilkův test normality:

```
x <- c(10.24,  10.12,  9.91,  10.19,  9.78,  10.14,  9.86,  10.17,  10.05)
shapiro.test(x)$p.val

## [1] 0.2873252
```

P-hodnota S-W testu je $0.2873 > 0.05$, tedy nulovou hypotézu o normálním rozdělení náhodného výběru nezamítáme na hladině významnosti $\alpha = 0.05$.

Testování nulové hypotézy:
Na hladině významnosti $\alpha = 0.05$ testujeme hypotézu $H_0 : \mu = 10$ proti oboustranné alternativě $H_1 : \mu \neq 10$. K testování použijeme jednovýběrový t-test.

a) Testování pomocí kritického oboru

```
x <- c(10.24,  10.12,  9.91,  10.19,  9.78,  10.14,  9.86,  10.17,  10.05)
alpha <- 0.05
m <- mean(x)
s <- sd(x)
n <- length(x)
c <- 10

t0 <- (m-c)/(s/sqrt(n))
w1 <- qt(1-alpha/2, n-1)
w2 <- qt(alpha/2,   n-1)

print(paste('t0 =', round(t0, 4)))

## [1] "t0 = 0.9426"

print(paste('w1 =', round(w1, 4)))
```

```
## [1] "w1 = 2.306"

print(paste('w2 =', round(w2, 4)))

## [1] "w2 = -2.306"
```

Testovací statistika t_0 nabývá hodnoty 0.9426, kritický obor má tvar

$$W = (-\infty\,;\,-2.3060\rangle \cup \langle 2.3060\,;\,\infty)$$

Protože $t_0 \notin W$, H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.

b) Testování pomocí intervalu spolehlivosti

```
dh <- m-s/sqrt(n)*qt(1-alpha/2, n-1)
hh <- m-s/sqrt(n)*qt(alpha/2, n-1)
print(paste('dh =', round(dh, digits=4)))

## [1] "dh = 9.9261"

print(paste('hh =', round(hh, digits=4)))

## [1] "hh = 10.1761"
```

95% empirický interval spolehlivosti pro střední hodnotu μ má tvar

$$IS = (9.9261\,;\,10.1761).$$

Protože $c = 10 \in IS$, nulovou hypotézu H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.

c) Testování pomocí p-hodnoty

```
p.val <- 2*min(pt(t0, n-1),1-pt(t0, n-1))
print(paste('p-hodnota =', round(p.val, 4)))

## [1] "p-hodnota = 0.3735"
```

Protože p-hodnota $= 0.3735 > 0.05$, nulovou hypotézu H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.

Poznámka: K otestování nulové hypotézy o střední hodnotě μ můžeme použít funkci t.test(x) s argumentem mu=10 (hodnota c z nulové hypotézy) a argumentem alternative='two.sided' (oboustranná alternativa).

```
x <- c(10.24,  10.12,  9.91,  10.19,  9.78,  10.14,  9.86,  10.17,  10.05)
t.test(x, mu=10, alternative='two.sided')

##
##  One Sample t-test
##
## data:  x
## t = 0.94261, df = 8, p-value = 0.3735
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.926073 10.176149
## sample estimates:
## mean of x
##  10.05111
```

Příklad 7.7. Testování hypotézy o směrodatné odchylce σ: U 25 náhodně vybraných dvoulitrových lahví s nealkoholickým nápojem byl zjištěn přesný objem nápoje. Výběrový průměr činil $m = 1.99\,\text{l}$ a výběrová směrodatná odchylka $s = 0.1\,\text{l}$. Předpokládejme, že objem nápoje v láhvi je náhodná veličina s normálním rozdělením. Na hladině významnosti $\alpha = 0.05$ ověřte tvrzení výrobce, že směrodatná odchylka je $0.08\,\text{l}$.

Na hladině významnosti $\alpha = 0.05$ testujeme hypotézu $H_0 : \sigma = 0.08$ proti oboustranné alternativě $H_1 : \sigma \neq 0.08$ neboli $H_0 : \sigma^2 = 0.0064$ proti oboustranné alternativě $H_1 : \sigma^2 \neq 0.0064$. K testování nulové hypotézy použijeme test o rozptylu.

a) Testování pomocí kritického oboru
Vypočteme realizaci testového kritéria

$$t_0 = \frac{(n-1)s^2}{c} = \frac{24 * 0.1^2}{0.08^2} = 37.5$$

```
alpha <- 0.05
m <- 1.99
s <- 0.1
c <- 0.0064
n <- 25
(t0 <- (n-1)*s^2/c)

## [1] 37.5

(w1 <- qchisq(alpha/2, n-1))

## [1] 12.40115

(w2 <- qchisq(1-alpha/2, n-1))

## [1] 39.36408
```

Testovací statistika t_0 nabývá hodnoty 37.5, kritický obor má tvar

$$W = (-\infty\,;\,12.4012\rangle \cup \langle 39.3640\,;\,\infty)$$

Protože $t_0 \notin W$, nejsme oprávněni na hladině významnosti $\alpha = 0.05$ zamítnout tvrzení výrobce.

b) Testování pomocí intervalu spolehlivosti

```
(dh <- (n-1)*s^2/qchisq(1-alpha/2, n-1))

## [1] 0.006096929

(hh <- (n-1)*s^2/qchisq(alpha/2, n-1))

## [1] 0.01935304
```

95% empirický interval spolehlivosti pro σ má tvar

$$IS = (0.0781\,;\,0.1391).$$

Protože $c = 0.08 \in IS$, nejsme oprávněni na hladině významnosti $\alpha = 0.05$ zamítnout tvrzení výrobce.

c) Testování pomocí p-hodnoty

```
(p.val <- 2*min(pchisq(t0, n-1), 1-pchisq(t0, n-1)))

## [1] 0.0779636
```

Protože p-hodnota $= 0.078 > 0.05$, nejsme oprávněni na hladině významnosti $\alpha = 0.05$ zamítnout tvrzení výrobce.

Příklad 7.8. Interval spolehlivosti pro rozdíl parametrů $\mu_1 - \mu_2$ dvourozměrného rozdělení: Bylo vylosováno 6 vrhů selat a z nich vždy dva sourozenci. Jeden z nich vždy dostal náhodně dietu č. 1 a druhý dietu č. 2. Přírůstky v dkg jsou následující: $(62, 52)$, $(54, 56)$, $(55, 49)$, $(60, 50)$, $(53, 51)$, $(58, 50)$. Za předpokladu, že uvedené dvojice tvoří náhodný výběr z dvourozměrného rozdělení s vektorem středních hodnot (μ_1, μ_2) a jejich rozdíly se řídí normálním rozdělením, sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot.

```
d1 <- c(62, 54, 55, 60, 53, 58)
d2 <- c(52, 56, 49, 50, 51, 50)
x <- d1-d2                          # rozdil strednich hodnot prirustku u diety 1 a diety 2
```

Shapirův - Wilkův test normality:

```
shapiro.test(x)$p.val

## [1] 0.3241142
```

P-hodnota S-W testu je $0.3241 > 0.05$, tedy nulovou hypotézu o normálním rozdělení náhodného výběru nezamítáme na hladině významnosti $\alpha = 0.05$.

Výpočet intervalů spolehlivosti:

```
## [1] "dh = 0.6265"
## [1] "hh = 10.7069"
```

95% interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$ má tvar:

$$(0.6265\,;\,10.7069).$$

$0.63\,\text{dkg} < \mu < 10.71\,\text{dkg}$ s pravděpodobností 0.95.

Poznámka: K nalezení hranic 95% empirického intervalu spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$ dvourozměrného rozdělení můžeme použít funkci t.test(x,y) s argumentem paired=T (párová test) a argumentem alternative='two.sided' (oboustranná alternativa).

```
x <- c(62, 54, 55, 60, 53, 58)
y <- c(52, 56, 49, 50, 51, 50)
t.test(x, y, paired=T, alternative='two.sided')$conf.int

## [1]  0.6264613 10.7068720
## attr(,"conf.level")
## [1] 0.95
```

Příklad 7.9. Testování hypotézy o rozdílu parametrů $\mu_1 - \mu_2$ dvourozměrného rozdělení: Bylo vybráno šest nových vozů téže značky a po určité době bylo zjištěno, o kolik mm se sjely jejich levé a pravé přední pneumatiky. Výsledky: $(1.8, 1.5)$, $(1.0, 1.1)$, $(2.2, 2.0)$, $(0.9, 1.1)$, $(1.5, 1.4)$, $(1.6, 1.4)$. Za předpokladu, že uvedené dvojice tvoří náhodný výběr z dvourozměrného rozdělení s vektorem středních hodnot (μ_1, μ_2) a jejich rozdíly se řídí normálním

rozdělením, testujte na hladině významnosti $\alpha = 0.05$ hypotézu, že obě pneumatiky se sjíždí stejně rychle.

Označme $\mu = \mu_1 - \mu_2$. Na hladině významnosti $\alpha = 0.05$ testujeme hypotézu $H_0 : \mu = 0$ proti oboustranné alternativě $H_1 : \mu \neq 0$. Testování provedeme párovým t-testem.

a) Testování pomocí kritického oboru

Shapirův - Wilkův test normality:

```
shapiro.test(x)$p.val

## [1] 0.4522054
```

P-hodnota S-W testu je $0.4522 > 0.05$, tedy nulovou hypotézu o normálním rozdělení náhodného výběru nezamítáme na hladině významnosti $\alpha = 0.05$.

Testování nulové hypotézy:

```
m <- mean(x)
n <- length(x)
s <- sd(x)
c <- 0
(t0 <- (m-c)/(s/sqrt(n)))

## [1] 1.051758

(w1 <- qt(1-alpha/2, n-1))

## [1] 2.570582

(w2 <- qt(alpha/2, n-1))

## [1] -2.570582
```

Testovací statistika t_0 nabývá hodnoty 1.0518, kritický obor má tvar

$$W = (-\infty\,;\, -2.5706\rangle \cup \langle 2.5706\,;\, \infty)$$

Protože $t_0 \notin W$, H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.

b) Testování pomocí intervalu spolehlivosti

```
(dh <- m-s/sqrt(n)*qt(1-alpha/2, n-1))

## [1] -0.1203401

(hh <- m-s/sqrt(n)*qt(alpha/2, n-1))

## [1] 0.2870068
```

95% empirický interval spolehlivosti pro σ má tvar

$$IS = (-0.1203\,;\, 0.2870).$$

Protože $c = 0 \in IS$, H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.

c) Testování pomocí p-hodnoty

```
(p.val <- 2*min(pt(t0, n-1), 1-pt(t0, n-1)))

## [1] 0.341062
```

Protože p-hodnota $= 0.3411 > 0.05$, H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.

Poznámka: K otestování nulové hypotézy o rozdílu parametrů $\mu_1 - \mu_2$ dvourozměrného rozdělení můžeme použít funkci t.test(x,y) s argumentem paired=T (párový test) a argumentem alternative='two.sided' (oboustranná alternativa).

```
x <- c(1.8, 1.0, 2.2, 0.9, 1.5, 1.6)
y <- c(1.5, 1.1, 2.0, 1.1, 1.4, 1.4)
t.test(x, y, paired=T, alternative='two.sided')

##
##  Paired t-test
##
## data:  x and y
## t = 1.0518, df = 5, p-value = 0.3411
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1203401  0.2870068
## sample estimates:
## mean of the differences
##              0.08333333
```

8 Parametrické úlohy o dvou nezávislých náhodných výběrech z normálních rozdělení a jednom náhodném výběru z alternativního rozdělení

Parametrické úlohy o dvou nezávislých náhodných výběrech z normálních rozdělení

Příklad 8.1. Interval spolehlivosti pro parametrickou funkci $\mu_1 - \mu_2$: Bylo vylosováno 11 stejně starých selat téhož plemene. Šesti z nich byla předepsána výkrmná dieta č. 1 a zbylým pěti výkrmná dieta č. 2. Průměrné denní přírůstky v dkg za dobu půl roku jsou následující:

dieta č. 1	62	54	55	60	53	58
dieta č. 2	52	56	49	50	51	

Zjištěné hodnoty považujeme za realizace dvou nezávislých náhodných výběrů pocházejících z rozdělení $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$. Sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

```
x <- c(62, 54, 55, 60, 53, 58)
y <- c(52, 56, 49, 50, 51)
m1 <- mean(x)
m2 <- mean(y)
s1 <- sd(x)
s2 <- sd(y)
n1 <- length(x)
n2 <- length(y)
alpha <- 0.05

sh <- sqrt(((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2))
(dh <- m1-m2-sh*sqrt(1/n1+1/n2)*qt(1-alpha/2, n1+n2-2))

## [1] 0.9919634

(hh <- m1-m2-sh*sqrt(1/n1+1/n2)*qt(alpha/2, n1+n2-2))

## [1] 9.808037
```

$$IS = (0.9920\,;\, 9.8080)$$

S pravděpodobností alespoň 0.95 platí, že $0.99\,\text{dkg} < \mu_1 - \mu_2 < 9.81\,\text{dkg}$.

Příklad 8.2. Testování hypotéz o parametrických funkcích $\mu_1 - \mu_2$, σ_1^2/σ_2^2:

1. Pro datový soubor z příkladu 8.1 testujte na hladině významnosti $\alpha = 0.05$ hypotézu, že
   a) rozptyly hmotnostních přírůstků selat při obou výkrmných dietách jsou shodné;
   b) obě výkrmné diety mají stejný vliv na hmotnostní přírůstky selat.
2. Výsledek testování podpořte krabicovým diagramem.

Shapirův - Wilkův test normality

Nejprve je potřeba otestovat normalitu obou náhodných výběrů.

```
x <- c(62, 54, 55, 60, 53, 58)
y <- c(52, 56, 49, 50, 51)
shapiro.test(x)$p.value

## [1] 0.6194994
```

```
shapiro.test(y)$p.value

## [1] 0.4271986
```

P-hodnota S-W testu pro přírůstky selat krmených dietou č. 1 je $0.6195 > 0.05$, p-hodnota S-W testu pro přírůstky selat krmených dietou č. 2 je $0.4272 > 0.05$. V obou případech tedy nulovou hypotézu o normalitě dat nezamítáme na hladině významnosti $\alpha = 0.05$.

ad a) Testování hypotézy o shodě rozptylů.
Na hladině významnosti $\alpha = 0.05$ testujeme nulovou hypotézu $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ oproti alternativní hypotéze $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$. K testování použijeme dvouvýběrový F-test.

Testování nulové hypotézy:

i. Testování pomocí kritického oboru

```
m1 <- mean(x)
m2 <- mean(y)
s1 <- sd(x)
s2 <- sd(y)
n1 <- length(x)
n2 <- length(y)
alpha <- 0.05
sh <- sqrt(((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2))

(t0 <- s1^2/s2^2)

## [1] 1.753425

(w1 <- qf (alpha/2, n1-1, n2-1))

## [1] 0.1353567

(w2 <- qf(1-alpha/2, n1-1, n2-1))

## [1] 9.364471
```

Testovací statistika t_0 nabývá hodnoty 1.7534, kritický obor má tvar

$$W = (-\infty\,;\, 0.1354\rangle \cup \langle 9.3645\,;\, \infty)$$

Protože $t_0 \notin W$, H_0 o shodě rozptylů σ_1^2 a σ_2^2 nezamítáme na hladině významnosti $\alpha = 0.05$.

ii. Testování pomocí intervalu spolehlivosti

```
(dh <- (s1^2/s2^2)/(qf(1-alpha/2, n1-1, n2-1)))

## [1] 0.1872423

(hh <- (s1^2/s2^2)/(qf(alpha/2, n1-1, n2-1)))

## [1] 12.9541
```

95% empirický interval spolehlivosti pro podíl σ_1^2/σ_2^2 má tvar

$$IS = (0.1872\,;\, 12.9541).$$

Protože $c = 1 \in IS$, H_0 o shodě rozptylů σ_1^2 a σ_2^2 nezamítáme na hladině významnosti $\alpha = 0.05$.

iii. Testování pomocí p-hodnoty

```
(p.val <- 2*min(pf(t0, n1-1, n2-1), 1-pf(t0, n1-1, n2-1)))

## [1] 0.6063451
```

Protože p-hodnota $= 0.6063 > 0.05$, H_0 o shodě rozptylů σ_1^2 a σ_2^2 nezamítáme na hladině významnosti $\alpha = 0.05$.

Upozornění: V případě zamítnutí hypotézy o shodě rozptylů je zapotřebí použít test se samostatnými odhady rozptylu.

ad b) Testování hypotézy o shodě středních hodnot
Na hladině významnosti $\alpha = 0.05$ testujeme nulovou hypotézu $H_0 : \mu_1 - \mu_2 = 0$ oproti alternativní hypotéze $H_1 : \mu_1 - \mu_2 \neq 0$. K testování použijeme dvouvýběrový test o rozdílu středních hodnot.

Testování nulové hypotézy:

i. Testování pomocí kritického oboru

```
c <- 0
(t0 <- ((m1-m2)-c)/(sh*sqrt(1/n1+1/n2)))

## [1] 2.771222

(w1 <- qt (alpha/2, n1+n2-2))

## [1] -2.262157

(w2 <- qt(1-alpha/2, n1+n2-2))

## [1] 2.262157
```

Testovací statistika t_0 nabývá hodnoty 2.7712, kritický obor má tvar

$$W = (-\infty\,;\,-2.2622\rangle \cup \langle 2.2622\,;\,\infty)$$

Protože $t_0 \in W$, H_0 o shodě středních hodnot μ_1 a μ_2 zamítáme na hladině významnosti $\alpha = 0.05$.

ii. Testování pomocí intervalu spolehlivosti
V příkladu 8.1 jsme zjistili, že 95% oboustranný interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$ má tvar

$$IS = (0.9920\,;\,9.8080).$$

Protože $c = 0 \notin IS$, H_0 o shodě středních hodnot μ_1 a μ_2 zamítáme na hladině významnosti $\alpha = 0.05$.

iii. Testování pomocí p-hodnoty

```
(p.val <- 2*min(pt(t0, n1+n2-2), 1-pt(t0, n1+n2-2)))

## [1] 0.02171008
```

Protože p-hodnota $= 0.0217 < 0.05$, H_0 o shodě středních hodnot μ_1 a μ_2 zamítáme na hladině významnosti $\alpha = 0.05$. Znamená to, že s rizikem omylu nejvýše 5 % se prokázalo, že obě výkrmné diety se liší účinností.

Poznámka: K otestování nulové hypotézy o rozdílu středních hodnot $\mu_1 - \mu_2$ dvou nezávislých náhodných výběrů z normálních rozdělení můžeme použít funkci t.test(x,y) s argumentem alternative='two.sided' (oboustranná alternativa) a argumentem var.equal=T (rozptyly obou náhodných výběrů si jsou rovné).

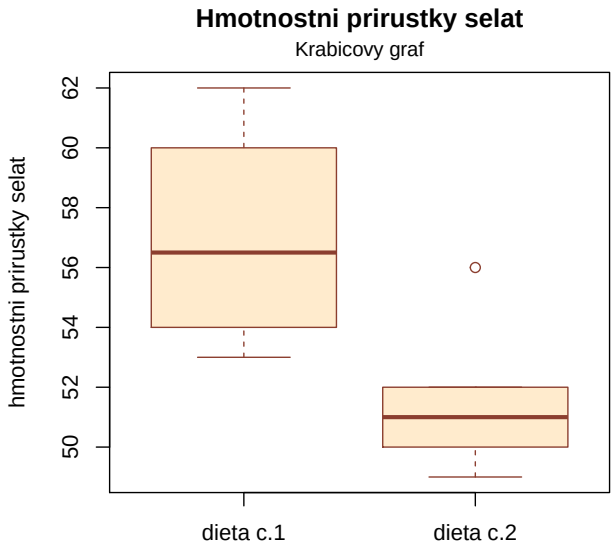
```r
x <- c(62, 54, 55, 60, 53, 58)
y <- c(52, 56, 49, 50, 51)
t.test(x, y, alternative='two.sided', var.equal=T)

##
##  Two Sample t-test
##
## data:  x and y
## t = 2.7712, df = 9, p-value = 0.02171
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9919634 9.8080366
## sample estimates:
## mean of x mean of y
##      57.0      51.6
```

Upozornění: Pokud bychom na hladině významnosti $\alpha = 0.05$ zamítli nulovou hypotézu o shodě rozptylů σ_1^2 a σ_2^2, mohli bychom k otestování nulové hypotézy o shodě středních hodnot μ_1 a μ_2 použít opět funkci t.test s argumentem alternative='two.sided' (oboustranná alternativa) a argumentem var.equal=F. Tento argument modifikuje klasický t-test na t-test s Welschovou aproximací stupňů volnosti, která se používá v případě, že rozptyly obou náhodných výběrů nejsou shodné.

Krabicový diagram

Řešení příkladu doplníme ještě o krabicový diagram.

```r
boxplot(x, y, names=c('dieta c.1', 'dieta c.2'), ylab='hmotnostni prirustky selat',
        main='Hmotnostni prirustky selat', col='blanchedalmond', border='coral4')
mtext('Krabicovy graf', line=0.4, cex=0.9)
```



Příklad k samostatnému řešení

Příklad 8.3. Načtěte datový soubor vyska.txt, který obsahuje údaje o výšce 48 studentek VŠE v Praze (proměnná vyska) a obor jejich studia (1 – národní hospodářství, 2 – informatika).

a) Pomocí S-W testu ověřte na hladině významnosti $\alpha = 0.1$ předpoklad o normalitě výšek v obou skupinách studentek.

b) Na hladině významnosti $\alpha = 0.1$ testujte hypotézu o shodě rozptylů výšek studentek v daných dvou oborech studia.

c) Na hladině významnosti $\alpha = 0.1$ testujte hypotézu o shodě středních hodnot výšek studentek v daných dvou oborech studia.

d) Výpočet doplňte krabicovými diagramy.

Shapirův - Wilkův test normality

```
data <- read.table('vyska.txt', header=T)
x <- data$vyska[data$obor==1]
y <- data$vyska[data$obor==2]
shapiro.test(x)$p.value

## [1] 0.6067928

shapiro.test(y)$p.value

## [1] 0.1119235
```

P-hodnota S-W testu pro výšku studentek oboru národní hospodářství je $0.6068 > 0.1$, p-hodnota S-W testu pro výšku studentek oboru informatika je $0.1119 > 0.1$. V obou případech tedy nulovou hypotézu o normalitě dat nezamítáme na hladině významnosti $\alpha = 0.1$.

Testování hypotézy o shodě rozptylů

i. Testování pomocí kritického oboru

```
## [1] "t0 = 1.9873"
## [1] "w1 = 0.5033"
## [1] "w2 = 2.0905"
```

Protože $t_0 \notin W$, H_0 o shodě rozptylů σ_1^2 a σ_2^2 nezamítáme na hladině významnosti $\alpha = 0.1$.

ii. Testování pomocí intervalu spolehlivosti

```
## [1] "dh = 0.9506"
## [1] "hh = 3.9487"
```

Protože $c = 1 \in IS$, H_0 o shodě rozptylů σ_1^2 a σ_2^2 nezamítáme na hladině významnosti $\alpha = 0.1$.

iii. Testování pomocí p-hodnoty

```
## [1] "p.val = 0.1249"
```

Protože p-hodnota $= 0.1249 > 0.05$, H_0 o shodě rozptylů σ_1^2 a σ_2^2 nezamítáme na hladině významnosti $\alpha = 0.1$.

Testování hypotézy o shodě středních hodnot

i. Testování pomocí kritického oboru

```
## [1] "t0 = 1.744"
## [1] "w1 = -1.6787"
## [1] "w2 = 1.6787"
```

Protože $t_0 \in W$, H_0 o shodě středních hodnot μ_1 a μ_2 zamítáme na hladině významnosti $\alpha = 0.1$.

ii. Testování pomocí intervalu spolehlivosti

```
## [1] "dh = 0.1095"
## [1] "hh = 5.7334"
```

Protože $c = 0 \notin IS$, H_0 o shodě středních hodnot μ_1 a μ_2 zamítáme na hladině významnosti $\alpha = 0.1$.

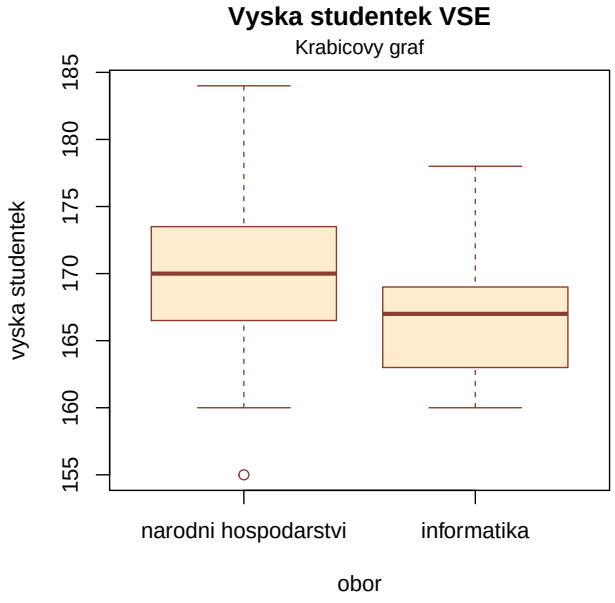
iii. Testování pomocí p-hodnoty

```
## [1] "p.val = 0.0878"
```

Protože p-hodnota $= 0.0878 < 0.1$, H_0 o shodě středních hodnot μ_1 a μ_2 zamítáme na hladině významnosti $\alpha = 0.1$.

Krabicový diagram

```
boxplot(x, y, names=c('narodni hospodarstvi', 'informatika'), ylab='vyska studentek',
        xlab='obor', main='Vyska studentek VSE',
        col='blanchedalmond', border='coral4')
mtext('Krabicovy graf', line=0.4, cex=0.9)
```



8.1 Parametrické úlohy o jednom náhodném výběru z alternativního rozdělení

Příklad 8.4. Asymptotický interval spolehlivosti pro parametr θ alternativního rozdělení: Může politická strana, pro niž se v předvolebním průzkumu vyslovilo 60 z 1000 dotázaných osob, očekávat se spolehlivostí 0.95, že by v této době ve volbách překročila 5 % hranici pro vstup do parlamentu?

Zavedeme náhodné veličiny $X_1, \ldots, X_{1000}$, přičemž $X_i = 1$, když se i-tá osoba vysloví pro danou politickou stranu, a $X_i = 0$ jinak; $i = 1, \ldots, 1000$. Tyto náhodné veličiny tvoří náhodný výběr z rozdělení $A(\theta)$. V tomto případě $n = 1000$, $m = 60/1000 = 0.06$, $\alpha = 0.05$, $u_{1-\alpha} = u_{0.95} = 1.645$.

Ověření podmínky $n\theta(1-\theta) > 9$: parametr θ neznáme, musíme ho tedy nahradit výběrovým průměrem. Pak $1000 * 0.06 * 0.94 = 56.4 > 9$.

95% levostranný interval spolehlivosti pro parametr θ má potom tvar

$$\left(m - \sqrt{\frac{m(1-m)}{n}} u_{1-\alpha}\,;\, \infty\right) = \left(0.06 - \sqrt{\frac{0.06(1-0.06)}{1000}} u_{0.95}\,;\, \infty\right)$$

V našem případě

$$d = 0.06 - \sqrt{\frac{0.06 * 0.94}{1000}} 1.645 = 0.0476.$$

S pravděpodobností přibližně 0.95 je tedy $\theta > 0.0476$. Protože tento interval zahrnuje i hodnoty nižší než 0.05, nelze vyloučit, že strana získá méně než 5 % hlasů.

```
x <- 60
n <- 1000
m <- x/n
alpha <- 0.05
(dh <- m-sqrt(m*(1-m)/n)*qnorm(1-alpha))

## [1] 0.04764716
```

Příklad k samostatnému řešení

Příklad 8.5. Přírůstky cen akcií na burze (v %) u 10 náhodně vybraných společností dosáhly těchto hodnot: 10, 16, 5, 10, 12, 8, 4, 6, 5, 4. Sestrojte 95% asymptotický empirický interval spolehlivosti pro pravděpodobnost, že přírůstek ceny akcie překročí 8.5 %.

```
## [1] "dh = 0.0964"
## [1] "hh = 0.7036"
```

$0.096 < \theta < 0.704$ s pravděpodobností aspoň 0.95. Znamená to, že pravděpodobnost, že přírůstek ceny akcie překročí 8.5 %, je alespoň 9.6 % a nanejvýš 70.4 % (při spolehlivosti 95%.)

Příklad 8.6. Testování hypotézy o parametru θ alternativního rozdělení: Určitá cestovní kancelář organizuje zahraniční zájezdy podle individuálních přání zákazníků. Z několika minulých let ví, že 30 % všech takto organizovaných zájezdů má za cíl zemi X. Po zhoršení politických podmínek v této zemi se cestovní kancelář obává, že se zájem o tuto zemi mezi zákazníky sníží. Ze 150 náhodně vybraných zákazníků v tomto roce má 38 za cíl právě zemi X. Potvrzují nejnovější data pokles zájmu o tuto zemi? Volte hladinu významnosti $\alpha = 0.05$.

Máme náhodný výběr $X_1, \ldots, X_{150}$ z rozdělení $A(0.3)$. Testujeme $H_0 : \theta = 0.3$ proti levostranné alternativě $H_1 : \theta < 0.3$. V tomto případě je testovacím kritériem statistika

$$T_0 = \frac{M - c}{\sqrt{\frac{c(1-c)}{n}}},$$

která se za platnosti nulové hypotézy asymptoticky řídí standardizovaným normálním rozdělením $N(0,1)$.

Nejprve musíme ověřit splnění podmínky $n\theta(1-\theta) > 9$: $150 * 0.3 * 0.7 = 31.5 > 9$.

a) Testování pomocí kritického oboru
Vypočteme realizaci testovacího kritéria:

$$t_0 = \frac{m-c}{\sqrt{\frac{c(1-c)}{n}}} = \frac{\frac{38}{150} - 0.3}{\sqrt{\frac{0.3(1-0.3)}{150}}} = -1.2472.$$

Kritický obor má tvar:

$$W = (-\infty\,;\, u_\alpha\rangle \cup (-\infty\,;\, -1.645\rangle.$$

```
x <- 38
n <- 150
c <- 0.3
m <- x/n
alpha <- 0.05

(t0 <- (m-c)/sqrt((c*(1-c)/n)))

## [1] -1.247219

(w1 <- qnorm(alpha))

## [1] -1.644854
```

Protože testovací kritérium nepatří do kritického oboru, H_0 nezamítáme na asymptotické hladině významnosti $\alpha = 0.05$.

b) Testování pomocí intervalu spolehlivosti
Proti levostranné alternativě H_1 postavíme 95% pravostranný interval spolehlivosti.

$$(-\infty\,;\, hh)$$

kde realizace horní hranice

$$\begin{aligned} hh &= m - \sqrt{\frac{m(1-m)}{n}} u_\alpha \\ &= 0.2533 - \sqrt{\frac{0.2533 * (1-0.2533)}{150}} \mathsf{qnorm}(0.05) \\ &= 0.2533 - \sqrt{\frac{0.2533 * 0.7476}{150}} * (-1.6448) \\ &= 0.3117 \end{aligned}$$

```
(hh <- m-sqrt(m*(1-m)/n)*qnorm(alpha))

## [1] 0.3117439
```

Protože $0.3 \in IS = (-\infty\,;\, 0.3117)$, H_0 nezamítáme na asymptotické hladině významnosti $\alpha = 0.05$.

c) Testování pomocí p-hodnoty
V části a) jsme spočítali hodnotu testovací statistiky $t_0 = -1.2472$.

Protože máme levostrannou alternativní hypotézu H_1, vypočítáme p-hodnotu podle vzorce

$$p\text{-hodnota} = P(T_0 \leq t_0) = P(T_0 \leq t_0) = \mathsf{pnorm}(\mathsf{t_0}) = 0.1062$$

```
(p.val <- pnorm(t0))

## [1] 0.1061586
```

Protože p-hodnota $= 0.1062 > 0.05$, H_0 nezamítáme na asymptotické hladině významnosti $\alpha = 0.05$.

9 Analýza rozptylu jednoduchého třídění

Příklad 9.1. V jisté továrně se měřil čas, který potřeboval každý ze tří dělníků k uskutečnění téhož pracovního úkonu. Čas v minutách:

1.dělník:	3.6	3.8	3.7	3.5		
2.dělník:	4.3	3.9	4.2	3.9	4.4	4.7
3.dělník:	4.2	4.5	4.0	4.1	4.5	4.4

Na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že výkony těchto tří dělníků jsou stejné. Zamítnete-li nulovou hypotézu, určete, výkony kterých dělníků se liší na dané hladině významnosti $\alpha = 0.05$.

Průzkumová analýza

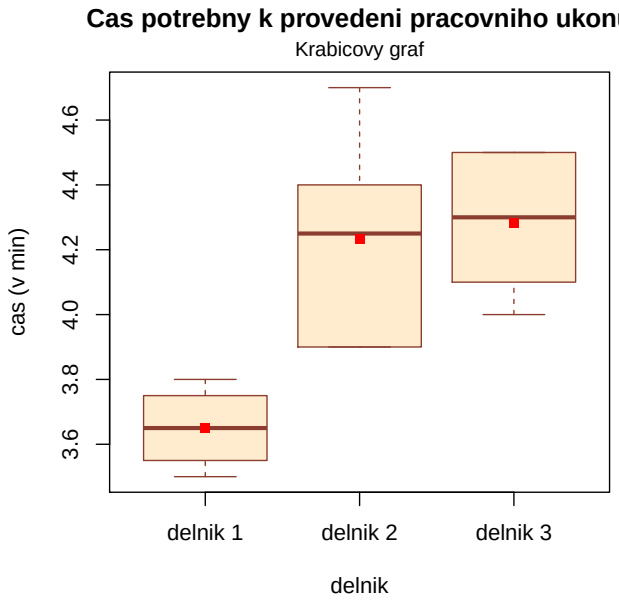
Úloha vede na analýzu rozptylu jednoduchého třídění. Načteme datový soubor cas_delniku.txt. Proměnná X obsahuje zjištěné časy, proměnná ID nabývá hodnoty 1 pro 1. dělníka, hodnoty 2 pro 2. dělníka a hodnoty 3 pro 3. dělníka.

```
data <- read.delim('cas_delniku.txt', sep='', dec=',', header=T)
nazvy <- c('delnik 1', 'delnik 2', 'delnik 3')
X <- data$X
ID <- data$ID

prum <- sm <- N <- NULL
for(i in 1:3){
  prum[i] <- mean(X[ID==i])
  sm[i] <- sd(X[ID==i])
  N[i] <- length(X[ID==i])
}
prum[4] <-mean(X)
sm[4] <- sd(X)
N[4] <- length(X)

tab <- data.frame(prumer=prum, N=N, sm.odch=sm,
                  row.names=c(nazvy, 'vsichni'))
round(tab,4)

##          prumer  N sm.odch
## delnik 1 3.6500  4  0.1291
## delnik 2 4.2333  6  0.3077
## delnik 3 4.2833  6  0.2137
## vsichni  4.1063 16  0.3530
```

Komentář: Na uskutečnění daného pracovního úkonu potřebuje nejkratší čas 1. dělník. Podává také nejvyrovnanější výkony – směrodatná odchylka proměnné X je u něj nejmenší. Naopak nejpomalejší je 3. dělník.

Krabicový graf

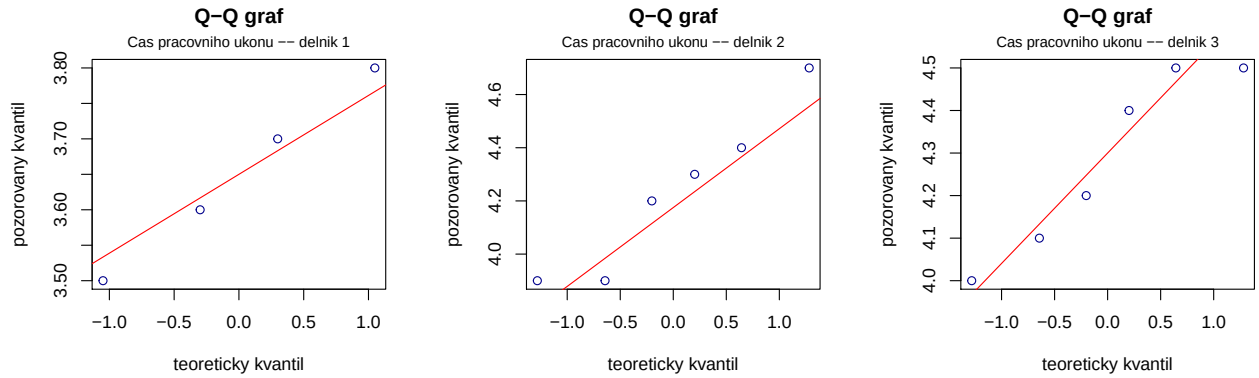
```
boxplot(X[ID==1], X[ID==2], X[ID==3], names=nazvy, ylab='cas (v min)',
        xlab='delnik', main='Cas potrebny k provedeni pracovniho ukonu',
        col='blanchedalmond', border='coral4')
mtext('Krabicovy graf', line=0.4, cex=0.9)
points(c(mean(X[ID==1]), mean(X[ID==2]), mean(X[ID==3])), pch=15, col='red')
```



Testování normality

Na testování normality všech tří výběrů použijeme kvůli jejich malým rozsahům S-W test.

```
## [1] "S-W test, delnik 1: 0.9719"
## [1] "S-W test, delnik 2: 0.5819"
## [1] "S-W test, delnik 3: 0.3313"
```



Komentář: Protože ve všech třech případech je p-hodnota S-W testu > 0.05, nulovou hypotézu o normalitě časů všech tří dělníků nezamítáme na hladině významnosti $\alpha = 0.05$. Z Q-Q grafů dále vidíme, že tečky se ve všech třech případech jen málo odchylují od přímky, což podporuje náš závěr, že všechny tři výběry pochází z normálního rozdělení.

Test homogenity rozptylů

Jelikož náhodné výběry pochází z normálního rozdělení, je vhodné na testování hypotézy o shodě rozptylů všech tří výběrů spoužít Levenův test. Na hladině významnosti $\alpha = 0.05$ testujeme nulovou hypotézu $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ oproti alternativní hypotéze $H_1 : \sigma_i^2 \neq \sigma_j^2$ pro alespoň jednu dvojici i, j.

K otestování použijeme funkci levene.test z knihovny lawstat s argumentem location='mean', čímž zvolíme klasickou formu Levenova testu. Pokud bychom zadali funkci levene.test s argumentem location='median', získali bychom robustnější modifikaci Levenova testu, která ve svém výpočtu nahrazuje aritmetický průměr mediánem. Balíček lawstat disponuje také příkazem na výpočet Bartlettova testu, který je k dispozici ve funkci bartlett.test.

```
library(lawstat)
levene.test(X, ID, location='mean')

##
##  classical Levene's test based on the absolute deviations from the mean ( none
##  not applied because the location is not set to median )
##
## data:  X
## Test Statistic = 1.5142, p-value = 0.2564

#levene.test(X, ID, location='median')
#bartlett.test(X, ID)
```

Komentář: Testovací statistika Levenova testu nabývá hodnoty 1.5142, odpovídající p-hodnota = 0.256, tedy na hladině významnosti $\alpha = 0.05$ nezamítáme hypotézu o shodě rozptylů σ_1^2, σ_2^2 a σ_3^2.

Test o shodě středních hodnot:

Na hladině významnosti $\alpha = 0.05$ testujeme nyní nulovou hypotézu

$H_0 : \mu_1 = \mu_2 = \mu_3$

oproti alternativní hypotéze

$H_1 : \mu_i \neq \mu_j$ pro alespoň jednu dvojici i, j.

```
r <- 3
n <- length(X)
Xi. <- Mi. <- ni <- NULL
for(i in 1:r){
  Xi.[i] <- sum(X[ID==i])
  ni[i]  <- length(X[ID==i])
  Mi.[i] <- sum(X[ID==i])/ni[i]
}
X.. <- sum(X)
M.. <- mean(X)

SA <- sum(ni*(Mi.-M..)^2)
fA <- r-1
ST <- sum((X-M..)^2)
SE <- ST-SA
fE <- n-r
Fa <- (SA/fA)/(SE/fE)

p1 <- pf(Fa,r-1,n-r)
```

```
p2 <- 1-p1
p.val <- min(p1, p2)

(tab <- round(data.frame(SA=SA, fA=fA, SE=SE, fE=fE, ST=ST, fT=n, Fa=Fa, p.val=p.val), digits=5))

##        SA fA      SE fE      ST fT      Fa   p.val
## 1 1.11771  2 0.75167 13 1.86938 16 9.66533 0.00268
```

Skupinový součet čtverců $S_A = 1.1177$, počet stupňů volnosti $f_A = 2$, reziduální součet čtverců $S_E = 0.7517$, počet stupňů volnosti $f_E = 13$, testovací statistika

$$F_A = \frac{S_A/f_A}{S_E/f_E}$$

nabývá hodnoty 9.6653, počet stupňů volnosti čitatele = 2, jmenovatele = 13. p-hodnota = 0.00268, tedy na hladině významnosti $\alpha = 0.05$ zamítáme nulovou hypotézu o shodě středních hodnot.

Metoda mnohonásobného porovnávání

Jelikož jsme na hladině významnosti 0.05 zamítli nulovou hypotézu o shodě středních hodnot, chceme nyní zjistit, které dvojice středních hodnot se od sebe významně liší. Stanovíme nulové a alternativní hypotézy pro dvojice středních hodnot

$H_{01} : \mu_1 = \mu_2$ oproti $H_{11} : \mu_1 \neq \mu_2$;
$H_{02} : \mu_1 = \mu_3$ oproti $H_{12} : \mu_1 \neq \mu_3$;
$H_{03} : \mu_2 = \mu_3$ oproti $H_{13} : \mu_2 \neq \mu_3$.

Protože v každé skupině máme různý počet pozorování, je vhodné použít na mnohonásobné porovnávání Scheffého metodu. Pravou a levou stranu Scheffého metody získáme použitím funkce Scheffe(X, group, names, alpha), která je k dispozici v RSkriptu AS-funkce.R. Argumenty funkce jsou X ... vektor hodnot, group ... vektor přiřazující každému pozorování skupinu, do níž náleží, names ... názvy jednotlivých skupin, alpha ... hladina významnosti.

```
source('AS-funkce.R')
Scheffe(X, ID, nazvy, alpha=0.05)

## $R
##           delnik 1  delnik 2  delnik 3
## delnik 1 0.4690839 0.4282131 0.4282131
## delnik 2 0.4282131 0.3830054 0.3830054
## delnik 3 0.4282131 0.3830054 0.3830054
##
## $L
##           delnik 1  delnik 2  delnik 3
## delnik 1 0.0000000 0.5833333 0.6333333
## delnik 2 0.5833333 0.0000000 0.0500000
## delnik 3 0.6333333 0.0500000 0.0000000

1*(Scheffe(X, ID, nazvy, alpha=0.05)$R >= Scheffe(X, ID, nazvy, alpha=0.05)$L)

##          delnik 1 delnik 2 delnik 3
## delnik 1        1        0        0
## delnik 2        0        1        1
## delnik 3        0        1        1
```

Komentář: Porovnáním pravé a levé strany Scheffého metody vidíme, že na hladině významnosti $\alpha = 0.05$ zamítáme nulovou hypotézu o shodě středních hodnot μ_1 a μ_2 a středních hodnot μ_1 a μ_3. Výsledek Scheffého metody tedy ukazuje, že na hladině významnosti $\alpha = 0.05$ se liší výkony dělníků (1,2), (1,3), naopak výkony dělníků (2,3) se neliší.

Příklad 9.2. Na střední škole byl uskutečněn experminet zjišťující efektvitu jednotlivých pedagogických metod. Studenti byli rozděleni do pěti skupin a každá skupina byla vyučována pomocí jedné z pedagogických metod: tradiční způsob, programová výuka, audiotechnika, audiovizuální technika a vizuální technika. Z každé skupiny byl potom vybrán náhodný vzorek studentů a všichni byli podrobeni témuž písemnému testu. Výsledky testu jsou uvedeny v následující tabulce a v souboru pet_metod.txt:

metoda	počet bodů							
tradicni	76.2	48.3	85.1	63.7	91.6	87.2		
programova	85.2	74.3	76.5	80.3	67.4	67.9	72.1	60.4
audio	67.3	60.1	55.4	72.3	40.0			
audiovizualni	75.8	81.6	90.3	78.0	67.8	57.6		
vizualni	50.5	70.2	88.8	67.1	77.7	73.9		

Na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že znalosti všech studentů jsou stejné a nezávisí na použité pedagogické metodě. V případě zamítnutí hypotézy zjistěte, které výběry se liší na hladině významnosti 0.05.

Průzkumová analýza

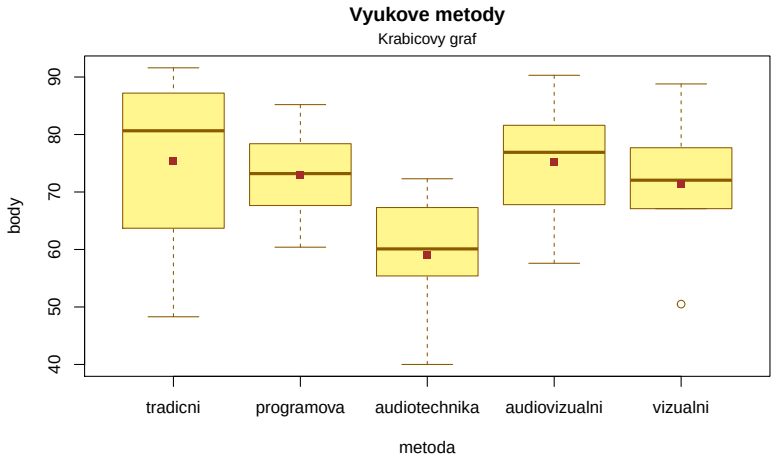
Načteme datový soubor pet_metod.txt. Proměnná body obsahuje dosažené počty bodů a proměnná metoda označení příslušné pedagogické metody. Nejprve vypočítáme průměry, směrodatné odchylky a rozsahy všech tří výběrů:

```
data <- read.delim('pet_metod.txt', sep='', dec=',', header=T)
nazvy <- c('tradicni', 'programova', 'audiotechnika', 'audiovizualni', 'vizualni')

X <- data$body
ID <- data$metoda
r <- length(unique(ID))

prum <- sm <- N <- NULL
for(i in 1:r){
  prum[i] <- mean(X[ID==i])
  sm[i] <- sd(X[ID==i])
  N[i] <- length(X[ID==i])
}
prum[r+1] <-mean(X)
sm[r+1] <- sd(X)
N[r+1] <- length(X)

tab <- data.frame(prumer=prum, N=N, sm.odch=sm,
                  row.names=c(nazvy, 'vsechny'))
round(tab,4)

##                prumer  N sm.odch
## tradicni      75.3500  6 16.5390
## programova    73.0125  8  7.8650
## audiotechnika 59.0200  5 12.4594
## audiovizualni 75.1833  6 11.3286
## vizualni      71.3667  6 12.6920
## vsechny       71.3097 31 12.6953
```

Komentář: Nejlepších výsledků dosahují studenti vyučovaní tradiční metodou, podávají však nejméně vyrovnané výkony (počty bodů v této skupině mají největší směrodatnou odchylku). Naopak nejhoršího výsledku dosáhli studenti vyučovaní audio metodou. Nejvyrovnanější výkony pozorujeme u studentů vyučovaných programovou metodou.

Krabicový graf

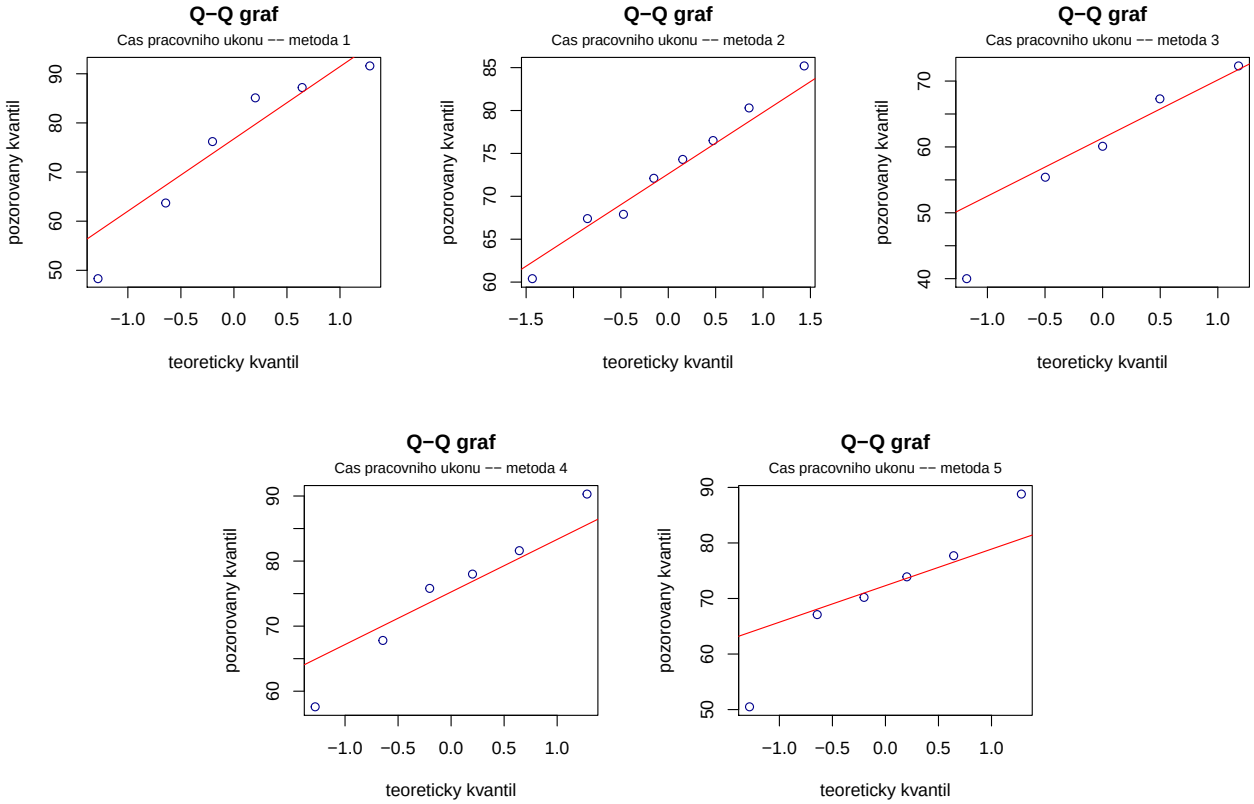
```
boxplot(X[ID==1], X[ID==2], X[ID==3], X[ID==4], X[ID==5], names=nazvy, ylab='body',
        xlab='metoda', main='Vyukove metody',
        col='khaki1', border='orange4')
mtext('Krabicovy graf', line=0.4, cex=0.9)
points(c(mean(X[ID==1]), mean(X[ID==2]), mean(X[ID==3]), mean(X[ID==4]), mean(X[ID==5])), pch=15, col='bro
```



Testování normality

Na testování normality všech tří výběrů použijeme z důvodu jejich malých rozsahů S-W test.

```
print(paste('S-W test, metoda 1:', round(shapiro.test(X[ID==1])$p.value,4)))

## [1] "S-W test, metoda 1: 0.4177"

print(paste('S-W test, metoda 2:', round(shapiro.test(X[ID==2])$p.value,4)))

## [1] "S-W test, metoda 2: 0.9966"

print(paste('S-W test, metoda 3:', round(shapiro.test(X[ID==3])$p.value,4)))

## [1] "S-W test, metoda 3: 0.7663"

print(paste('S-W test, metoda 4:', round(shapiro.test(X[ID==4])$p.value,4)))

## [1] "S-W test, metoda 4: 0.9577"

print(paste('S-W test, metoda 5:', round(shapiro.test(X[ID==5])$p.value,4)))

## [1] "S-W test, metoda 5: 0.8814"

for(i in 1:r){
qqnorm(X[ID==i], col='darkblue', xlab='teoreticky kvantil',
       ylab='pozorovany kvantil', main='Q-Q graf')
qqline(X[ID==i], col='red')
mtext(paste('Cas pracovniho ukonu -- metoda', i), line=0.4, cex=0.8)}
```



Komentář: Protože ve všech pěti případech je p-hodnota S-W testu > 0.05, nulovou hypotézu o normalitě dat u každé metody nezamítáme na hladině významnosti $\alpha = 0.05$. Vzhled Q-Q grafů potvrzuje náš závěr, že předpoklad normality je ve všech pěti případech oprávněný.

Test homogenity rozptylů

K testování hypotézy o shodě rozptylů opět použijeme klasický Levenův test. Na hladině významnosti $\alpha = 0.05$ testujeme nulovou hypotézu $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$ oproti alternativní hypotéze $H_1 : \sigma_i^2 \neq \sigma_j^2$ pro alespoň jednu dvojici i, j.

```
library(lawstat)
levene.test(X, ID, location='mean')

##
##  classical Levene's test based on the absolute deviations from the mean ( none
##  not applied because the location is not set to median )
##
## data:  X
## Test Statistic = 0.81903, p-value = 0.5248

#levene.test(X, ID, location='median')
#bartlett.test(X, ID)
```

Komentář: Testovací statistika F se realizuje hodnotou 0.8190, odpovídající p-hodnota $= 0.5248$. Na hladině významnosti $\alpha = 0.05$ tedy nezamítáme hypotézu o shodě rozptylů.

Test o shodě středních hodnot:

Na hladině významnosti $\alpha = 0.05$ testujeme nyní nulovou hypotézu

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

oproti alternativní hypotéze

$H_1 : \mu_i \neq \mu_j$ pro alespoň jednu dvojici i, j.

```
n <- length(X)
Xi. <- Mi. <- ni <- NULL
for(i in 1:r){
  Xi.[i] <- sum(X[ID==i])
  ni[i]  <- length(X[ID==i])
  Mi.[i] <- sum(X[ID==i])/ni[i]
}
X.. <- sum(X)
M.. <- mean(X)

SA <- sum(ni*(Mi.-M..)^2)
fA <- r-1
ST <- sum((X-M..)^2)
SE <- ST-SA
fE <- n-r
Fa <- (SA/fA)/(SE/fE)

p1 <- pf(Fa,r-1,n-r)
p2 <- 1-p1
p.val <- min(p1, p2)

(tab <- round(data.frame(SA=SA, fA=fA, SE=SE, fE=fE, ST=ST, fT=n, Fa=Fa, p.val=p.val), digits=5))

##         SA fA       SE fE       ST fT      Fa   p.val
## 1 966.3737  4 3868.773 26 4835.147 31 1.62362 0.19825
```

Komentář: Testovací statistika F se realizuje hodnotou 1.6236, počet stupňů volnosti čitatele = 4, jmenovatele = 26, odpovídající p-hodnota = 0.1983, na hladině významnosti $\alpha = 0.05$ tedy nezamítáme hypotézu o shodě středních hodnot. Znamená to, že s rizikem omylu nejvýše 5 % se neprokázal rozdíl v účinnosti jednotlivých pedagogických metod.

Příklad 9.3. Pan Novák může cestovat z místa bydliště do místa pracoviště třemi různými způsoby: tramvají, autobusem a metrem s následným přestupem na tramvaj. Máme k dispozici jeho naměřené časy cestování do práce v době ranní špičky (včetně čekání na příslušný spoj) v minutách:

autobus:	32	39	42	37	34	38	
tramvaj:	30	34	28	26	32		
metro:	40	37	31	39	38	33	34

Pro všechny tři způsoby dopravy vypočtěte průměrné časy cestování. Na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že doba cestování do práce nezávisí na způsobu dopravy. V případě zamítnutí nulové hypotézy zjistěte, které způsoby dopravy do práce se od sebe liší na hladině významnosti $\alpha = 0.05$.

Průzkumová analýza

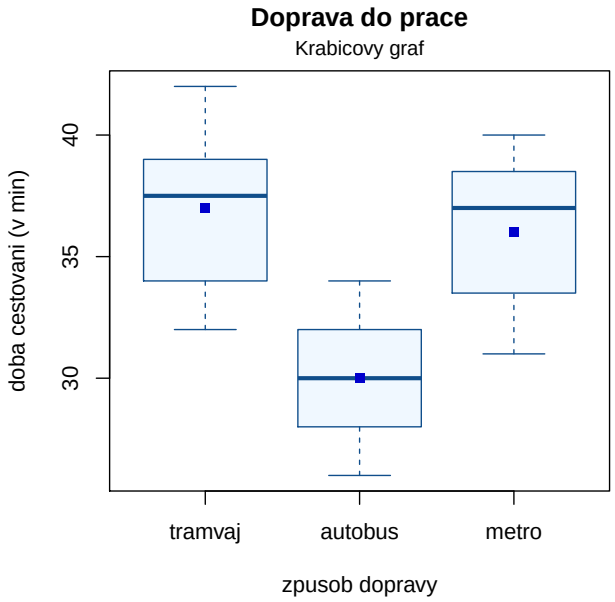
Načteme datový soubor doby_cestovani.txt. Proměnná cas obsahuje zjištěné doby cestování a proměnná ID označení příslušného způsobu dopravy. Nejprve vypočteme průměry, směrodatné odchylky a rozsahy všech tří výběrů:

```
data <- read.delim('doby_cestovani.txt', sep='', dec=',', header=T)
nazvy <- c('tramvaj', 'autobus', 'metro')

X <- data$cas
ID <- data$ID
r <- length(unique(ID))

prum <- sm <- N <- NULL
for(i in 1:r){
  prum[i] <- mean(X[ID==i])
  sm[i] <- sd(X[ID==i])
  N[i] <- length(X[ID==i])
}
prum[r+1] <-mean(X)
sm[r+1] <- sd(X)
N[r+1] <- length(X)

tab <- data.frame(prumer=prum, N=N, sm.odch=sm,
                  row.names=c(nazvy, 'vsechny'))
round(tab,4)

##          prumer  N sm.odch
## tramvaj 37.0000  6  3.5777
## autobus 30.0000  5  3.1623
## metro   36.0000  7  3.3665
## vsechny 34.6667 18  4.3791
```

Komentář: Nejkratší průměrnou dobu do zaměstnání pan Novák cestuje, když použije autobus, naopak nejdéle cestuje tramvají. Variabilita dob jednotlivých způsobů cestování je vcelku vyrovnaná.

Krabicový graf

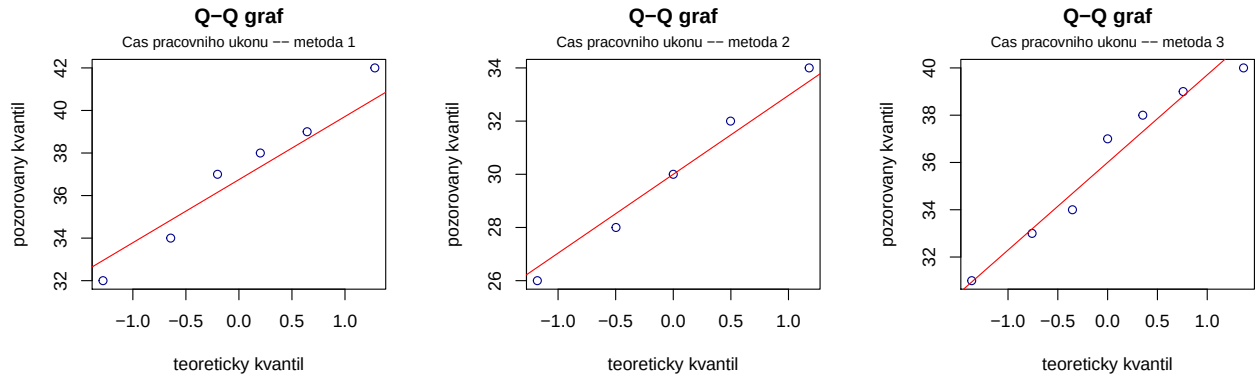
```
boxplot(X[ID==1], X[ID==2], X[ID==3], names=nazvy, ylab='doba cestovani (v min)',
        xlab='zpusob dopravy', main='Doprava do prace',
        col='aliceblue',border='dodgerblue4')
mtext('Krabicovy graf', line=0.4, cex=0.9)
points(c(mean(X[ID==1]), mean(X[ID==2]), mean(X[ID==3])), pch=15, col='blue3')
```



Testování normality

Na testování normality všech tří výběrů použijeme z důvodu jejich malých rozsahů S-W test.

```
print(paste('S-W test, autobus:', round(shapiro.test(X[ID==1])$p.value,4)))

## [1] "S-W test, autobus: 0.9539"

print(paste('S-W test, tramvaj:', round(shapiro.test(X[ID==2])$p.value,4)))

## [1] "S-W test, tramvaj: 0.9672"

print(paste('S-W test, metro:  ', round(shapiro.test(X[ID==3])$p.value,4)))

## [1] "S-W test, metro:   0.6294"

for(i in 1:r){
qqnorm(X[ID==i], col='darkblue', xlab='teoreticky kvantil',
       ylab='pozorovany kvantil', main='Q-Q graf')
qqline(X[ID==i], col='red')
mtext(paste('Cas pracovniho ukonu -- metoda', i), line=0.4, cex=0.8)}
```



Komentář: Protože ve všech třech případech je p-hodnota S-W testu > 0.05, nulovou hypotézu o normalitě dat u každé metody nezamítáme na hladině významnosti $\alpha = 0.05$. Vzhled Q-Q grafů potvrzuje náš závěr, že předpoklad normality je ve všech třech uvedených případech oprávněný.

Test homogenity rozptylů

Nulovou hypotézu o shodě rozptylů $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ ověříme Levenovým testem. Alternativní hypotéza má tvar $H_1 : \sigma_i^2 \neq \sigma_j^2$ pro alespoň jednu dvojici i, j.

```
library(lawstat)
levene.test(X, ID, location='mean')

##
##  classical Levene's test based on the absolute deviations from the mean ( none
##  not applied because the location is not set to median )
##
## data:  X
## Test Statistic = 0.10536, p-value = 0.9007

#levene.test(X, ID, location='median')
#bartlett.test(X, ID)
```

Komentář: Testovací statistika F se realizuje hodnotou 0.1054, počet stupňů volnosti čitatele $= 2$, jmenovatele $= 15$, odpovídající p-hodnota $= 0.9007$. Na hladině významnosti $\alpha = 0.05$ tedy nezamítáme hypotézu o shodě rozptylů.

Test o shodě středních hodnot:

Na hladině významnosti $\alpha = 0.05$ testujeme nyní nulovou hypotézu

$H_0 : \mu_1 = \mu_2 = \mu_3$

oproti alternativní hypotéze

$H_1 : \mu_i \neq \mu_j$ pro alespoň jednu dvojici i, j.

```
n <- length(X)
Xi. <- Mi. <- ni <- NULL
for(i in 1:r){
  Xi.[i] <- sum(X[ID==i])
  ni[i]  <- length(X[ID==i])
  Mi.[i] <- sum(X[ID==i])/ni[i]
```

```
}
X.. <- sum(X)
M.. <- mean(X)

SA <- sum(ni*(Mi.-M..)^2)
fA <- r-1
ST <- sum((X-M..)^2)
SE <- ST-SA
fE <- n-r
Fa <- (SA/fA)/(SE/fE)

p1 <- pf(Fa,r-1,n-r)
p2 <- 1-p1
p.val <- min(p1, p2)

(tab <- round(data.frame(SA=SA, fA=fA, SE=SE, fE=fE, ST=ST, fT=n, Fa=Fa, p.val=p.val), digits=5))

##    SA fA  SE fE  ST fT      Fa   p.val
## 1 154  2 172 15 326 18 6.71512 0.00827
```

Komentář: Testovací statistika F se realizuje hodnotou 6.7151, počet stupňů volnosti čitatele = 2, jmenovatele = 15, odpovídající p-hodnota = 0.0083. Na hladině významnosti $\alpha = 0.05$ tedy zamítáme hypotézu o shodě středních hodnot. Znamená to, že s rizikem omylu nejvýše 5 % se prokázal rozdíl v dobách cestování pana Nováka do zaměstnání autobusem, tramvají a metrem.

Metoda mnohonásobného porovnávání

Jelikož jsme na hladině významnosti 0.05 zamítli nulovou hypotézu o shodě středních hodnot, chceme nyní zjistit, které dvojice středních hodnot se od sebe významně liší. Stanovíme nulové a alternativní hypotézy pro dvojice středních hodnot

$H_{01} : \mu_1 = \mu_2$ oproti $H_{11} : \mu_1 \neq \mu_2$;
$H_{02} : \mu_1 = \mu_3$ oproti $H_{12} : \mu_1 \neq \mu_3$;
$H_{03} : \mu_2 = \mu_3$ oproti $H_{13} : \mu_2 \neq \mu_3$.

K porovnání středních hodnot použijeme opět Scheffého metodu.

```
source('AS-funkce.R')
#Scheffe(X, ID, alpha=0.05)
1*(Scheffe(X, ID, nazvy, alpha=0.05)$R >= Scheffe(X, ID, nazvy, alpha=0.05)$L)

##         tramvaj autobus metro
## tramvaj       1       0     1
## autobus       0       1     0
## metro         1       0     1

#Scheffe(X, ID, alpha=0.05, nazvy)
```

Komentáře: Z tabulky vyplývá, že s rizikem omylu nejvýše 5 % se liší cestování tramvají a autobusem a dále cestování autobusem a metrem.

10 Neparametrické úlohy o mediánech

Příklad 10.1. Párový znaménkový test a párový Wilcoxonův test Při zjišťování kvality jedné složky půdy se používají dvě metody označené A a B. Výsledky jsou uvedeny v následující tabulce:

Vzorek	1	2	3	4	5	6	7	8	9	10	11	12
A	0.275	0.312	0.284	0.3	0.365	0.298	0.312	0.315	0.242	0.321	0.335	0.307
B	0.28	0.312	0.288	0.298	0.361	0.307	0.319	0.315	0.242	0.323	0.341	0.315

Na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že metody A a B dávají stejné výsledky. K testování použijte jak párový znaménkový test, tak párový Wilcoxonův test. Pro lepší představu sestrojte krabicové diagramy pro obě metody.

Testujeme $H_0 : z_{0.50} = 0$ proti oboustranné alternativě $H_1 : z_{0.50} \neq 0$, kde $z_{0.50}$ je medián rozdělení, z něhož pochází rozdílový náhodný výběr $Z_1 = X_1 – Y_1, \ldots, Z_{12} = X_{12} – Y_{12}$.

Vypočteme rozdíly mezi výsledky metod A a B:

$x_i – y_i$	-0.005	0	-0.004	0.002	0.004	-0.009	-0.007	0	0	-0.002	-0.006	-0.008

Párový znaménkový test

Nenulových rozdílů je 9, testovací statistika $S_Z^+ = 2$. Ve statistických tabulkách najdeme pro $n = 9$ a $\alpha = 0.05$ kritické hodnoty $k_1 = 1$, $k_2 = 8$. Protože kritický obor $W = \langle 0; 1\rangle \cup 8$ neobsahuje hodnotu 2, nemůžeme H_0 zamítnout na hladině významnosti $\alpha = 0.05$. Neprokázalo se tedy, že by metody A a B dávaly rozdílné výsledky.

```
library(PASWR)
x1 <- c(0.275, 0.312,0.284,0.300,0.365,0.298,0.312,0.315,0.242,0.321,0.335,0.307)
x2 <- c(0.280, 0.312,0.288,0.298,0.361,0.307,0.319,0.315,0.242,0.323,0.341,0.315)
SIGN.test(x1-x2, alternative='two.sided')

##
##  One-sample Sign-Test
##
## data:  x1 - x2
## s = 2, p-value = 0.1797
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##  -0.006893636  0.000000000
## sample estimates:
## median of x
##      -0.003
##                   Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.8540 -0.0060      0
## Interpolated CI       0.9500 -0.0069      0
## Upper Achieved CI     0.9614 -0.0070      0
```

Párový Wilcoxonův test

Absolutní hodnoty nenulových rozdílů uspořádáme vzestupně podle velikosti:

Usp. abs($x_i – y_i$)	**0.002**	0.002	**0.004**	0.004	0.005	0.006	0.007	0.008	0.009
Pořadí	1	2	3	4	5	6	7	8	9
Průměrné pořadí	**1.5**	1.5	**3.5**	3.5	5	6	7	8	9

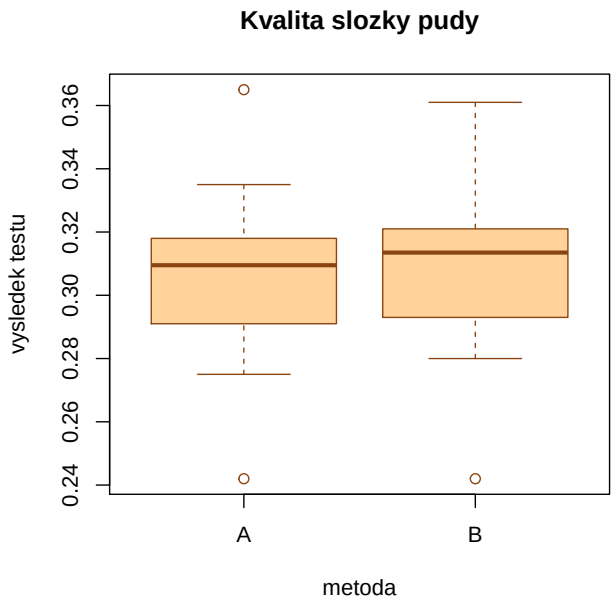
$S_W^+ = 1.5 + 3.5 = 5$,
$S_W^- = 1.5 + 3.5 + 5 + 6 + 7 + 8 + 9 = 40$,
$n = 9$, $\alpha = 0.05$, tabelovaná kritická hodnota pro $n = 9$ a $\alpha = 0.05$ je 5, testovací statistika je $= \min(S_W+, S_W^-) = \min(5, 40) = 5$. Protože $5 \leq 5$, H_0 zamítáme na hladině významnosti $\alpha = 0.05$.

```
wilcox.test(x1-x2, alternative='two.sided', correct=F, exact=F)

##
##  Wilcoxon signed rank test
##
## data:  x1 - x2
## V = 5, p-value = 0.03781
## alternative hypothesis: true location is not equal to 0
```

V tomto případě je p-hodnota 0.03781, tedy nulová hypotéza se zamítá na asymptotické hladině významnosti $\alpha = 0.05$. Nejsou však splněny předpoklady pro použití asymptotické varianty testu (příliš malý rozsah výběru), i když závěr je stejný jako při testování pomocí kritické hodnoty.

Krabicový graf

```
boxplot(x1, x2, names=c('A', 'B'),main='Kvalita slozky pudy',
        ylab='vysledek testu',xlab='metoda',col='burlywood1',border='chocolate4')
```



Z krabicových diagramů je vidět, že obě metody se poněkud liší v úrovni, ale neliší se ve variabilitě.

Příklad 10.2. Jednovýběrový znaménkový test a jednovýběrový Wilcoxonův test Vyráběné ocelové tyče mají kolísavou délku s předpokládanou hodnotou mediánu 10 m. Náhodný výběr 10-ti tyčí poskytl tyto výsledky: 9.83, 10.10, 9.72, 9.91, 10.04, 9.95, 9.82, 9.73, 9.81, 9.90. Na hladině významnosti 0.05 testujte hypotézu, že předpoklad o mediánu délky tyčí je oprávněný. K testování použijte jak jednovýběrový znaménkový test, tak jednovýběrový Wilcoxonův test. Pro lepší představu sestrojte krabicový diagram.

Testujeme $H_0 : x_{0.50} = 10$ proti oboustranné alternativě $H_1 : x_{0.50} \neq 10$. Vypočteme rozdíly mezi naměřenými délkami a konstantou 10:

x_i–10	−0.17	0.1	−0.28	−0.09	0.04	−0.05	−0.18	−0.27	−0.19	−0.1

Jednovýběrový znaménkový test

Nenulových rozdílů je 10, testovací statistika $S_Z^+ = 2$. Ve statistických tabulkách najdeme pro $n = 10$ a $\alpha = 0.05$ kritické hodnoty $k_1 = 1$, $k_2 = 9$. Protože kritický obor $W = \langle 0; 1\rangle \cup \langle 9; 10\rangle$ neobsahuje hodnotu 2, H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.

```
x <- c(9.83,  10.10,  9.72,  9.91,  10.04,  9.95,  9.82,  9.73,  9.81,  9.90)
SIGN.test(x, md=10, alternative='two.sided')

##
##  One-sample Sign-Test
##
## data:  x
## s = 2, p-value = 0.1094
## alternative hypothesis: true median is not equal to 10
## 95 percent confidence interval:
##   9.755956 10.010800
## sample estimates:
## median of x
##       9.865
##                   Conf.Level L.E.pt  U.E.pt
## Lower Achieved CI     0.8906  9.810  9.9500
## Interpolated CI       0.9500  9.756 10.0108
## Upper Achieved CI     0.9785  9.730 10.0400
```

Jednovýběrový Wilcoxonův test

Absolutní hodnoty nenulových rozdílů uspořádáme vzestupně podle velikosti:

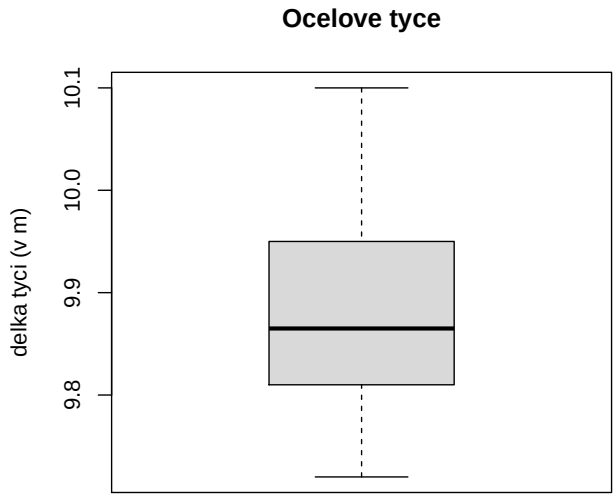
Usp. abs(xi – yi)	0.04	0.05	0.09	0.1	0.1	0.17	0.18	0.19	0.27	0.28
Pořadí	1	2	3	4	5	6	7	8	9	10
Průměrné pořadí	1	2	3	4.5	4.5	6	7	8	9	10

$S_W^+ = 1 + 4.5 = 5.5$,
$S_W^- = 2 + 3 + 4.5 + 6 + 7 + 8 + 9 + 10 = 49.5$,
$n = 10$, $\alpha = 0.05$, tabelovaná kritická hodnota pro $n = 10$ a $\alpha = 0.05$ je 8, testovací statistika $= \min(S_W^+, S_W^-) = \min(5.5; 49.5) = 5.5$. Protože $5.5 \leq 8$, H_0 zamítáme na hladině významnosti $\alpha = 0.05$. Vidíme, že Wilcoxonův test dospěl k odlišnému závěru než znaménkový test.

```
wilcox.test(x,mu=10, alternative = 'two.sided', exact=F, correct=F)

##
##  Wilcoxon signed rank test
##
## data:  x
## V = 5.5, p-value = 0.02484
## alternative hypothesis: true location is not equal to 10
```

Krabicový diagram

```
boxplot(x, main='Ocelove tyce', ylab='delka tyci (v m)', col='gray85')
```



Příklad 10.3. Dvouvýběrový Wilcoxonův test Majitel obchodu chtěl zjistit, zda velikost nákupů (v dolarech) placených kreditními kartami Master/EuroCard a Visa jsou přibližně stejné. Náhodně vybral

- 7 nákupů placených Master/EuroCard: 42, 77, 46, 73, 78, 33, 37;
- 9 nákupů placených Visou: 39, 10, 119, 68, 76, 126, 53, 79, 102.

Lze na hladině významnosti $\alpha = 0.05$ tvrdit, že velikost nákupů placených těmito dvěma typy karet se shodují? Pro lepší představu sestrojte krabicové diagramy pro oba typy platebních karet.

Na hladině významnosti $\alpha = 0.05$ testujeme $H_0 : x_{0.50} - y_{0.50} = 0$ proti oboustranné alternativě $H_1 : x_{0.50} - y_{0.50} \neq 0$.

usp. hodnoty	10	**33**	**37**	39	**42**	**46**	53	68	**73**	76	**77**	**78**	79	102	119	126
pořadí x_i		2	3		5	6			9		11	12				
pořadí y_i	1			4			7	8		10			13	14	15	16

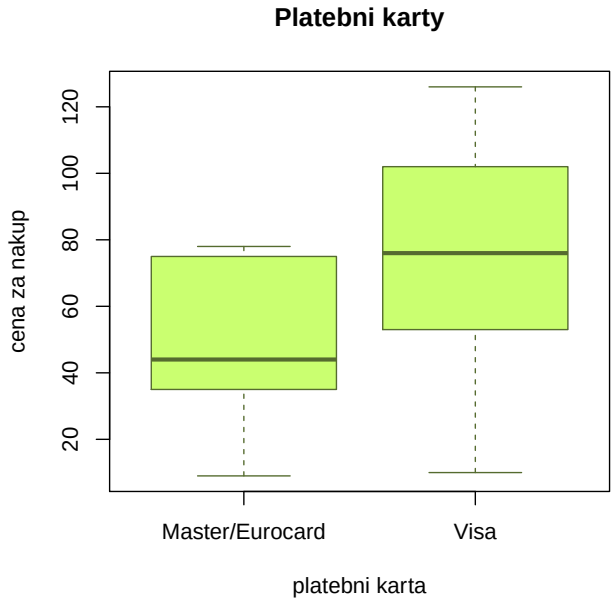
$T_1 = 2 + 3 + 5 + 6 + 9 + 11 + 12 = 48$,
$T_2 = 1 + 4 + 7 + 8 + 10 + 13 + 14 + 15 + 16 = 88$,
$U_1 = 7.9 + 7.8/2 - 48 = 43, U2 = 7.9 + 9.10/2 - 88 = 20$,
Kritická hodnota pro $\alpha = 0.05$, $\min(7, 9) = 7$, $\max(7, 9) = 9$ je 12. Protože $\min(43, 20) = 20 > 12$, nemůžeme na hladině významnosti $\alpha = 0.05$ zamítnout hypotézu, že velikosti nákupů placených kreditními kartami Master/EuroCard a Visa se shodují.

```
x <- c(42,  77,  46,  73,  78,  33,  37, 9)
y <- c(39, 10,  119,  68,  76,  126,  53,  79,  102)
wilcox.test(x, y, alternative = 'two.sided', exact=F, correct=F)

##
##  Wilcoxon rank sum test
##
```

```
## data:  x and y
## W = 20, p-value = 0.1237
## alternative hypothesis: true location shift is not equal to 0
```

Krabicový diagram

```
boxplot(x,y,main='Platebni karty',xlab='platebni karta',
        ylab='cena za nakup',names=c('Master/Eurocard','Visa'),
        col='darkolivegreen1',border='darkolivegreen')
```



Příklad 10.4. Kruskalův–Wallisův test Voda po holení jisté značky se prodává ve čtyřech různých lahvičkách stejného obsahu. Údaje o počtu prodaných lahviček za týden v různých obchodech jsou uvedeny v následující tabulce:

1.typ:	50	35	43	30	62	52	43	57	33	70	64	58	53	65	39
2.typ:	31	37	59	67	44	49	54	62	34	42	40				
3.typ:	27	19	32	20	18	23									
4.typ:	35	39	37	38	28	33									

Posuďte na 5 % hladině významnosti, zda typ lahvičky ovlivňuje úroveň prodeje. V případě zamítnutí nulové hypotézy zjistěte, prodeje kterých typů lahviček se od sebe významně liší. K testování použijte Kruskalův – Wallisův test; v případě zamítnutí nulové hypotézy použijte k zjištění významných rozdílů vhodnou metodu mnohonásobného porovnávání. Pro lepší představu sestrojte krabicové diagramy pro všechny typy lahviček.

Všech 38 hodnot uspořádáme vzestupně podle velikosti a stanovíme součet pořadí hodnot patřících do 1. až 4. výběru. $T_1 = 379$, $T_2 = 257$, $T_3 = 24$, $T4 = 81$

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^{r} \frac{T_j^2}{n_j} - 3(n+1) = \frac{12}{38 * 39} \left(\frac{379^2}{15} + \frac{257^2}{11} + \frac{24^2}{6} + \frac{81^2}{6}) \right) - 3 * 39 = 18.79$$

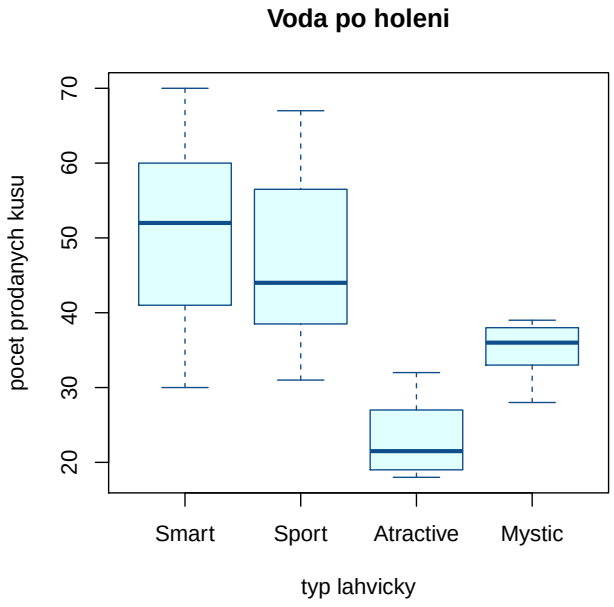
$$W = \left\langle \chi^2_{1-\alpha}(r-1)\,;\,\infty \right) = \left\langle \chi^2_{0.95}(3)\,;\,\infty \right) = \langle 7.815\,;\,\infty)$$

Protože $Q \in W$, H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0.05$.

```
x1 <- c(50, 35, 43, 30, 62, 52, 43, 57, 33, 70, 64, 58, 53, 65, 39)
x2 <- c(31, 37, 59, 67, 44, 49, 54, 62, 34, 42, 40)
x3 <- c(27, 19, 32, 20, 18, 23)
x4 <- c(35, 39, 37, 38, 28, 33)

x     <- c(x1, x2, x3, x4)
ni    <- c(length(x1),length(x2),length(x3),length(x4))
group <- c(rep(1, ni[1]), rep(2, ni[2]), rep(3, ni[3]), rep(4, ni[4]))
kruskal.test(x, group)

##
##  Kruskal-Wallis rank sum test
##
## data:  x and group
## Kruskal-Wallis chi-squared = 18.802, df = 3, p-value = 0.0003004
```

Krabicový graf

```
boxplot(x1, x2, x3, x4, main='Voda po holeni',xlab='typ lahvicky', ylab='pocet prodanych kusu',
        names=c('Smart','Sport','Atractive','Mystic'),col='lightcyan',border='dodgerblue4')
```



Je vidět, že úroveň prodeje pro 1. typ je nevyšší, zatímco pro 3. typ nejnižší.

Metoda mnohonásobného porovnávání

Nyní provedeme mnohonásobné porovnávání, abychom zjistili, které dvojice typů lahviček se liší na hladině významnosti $\alpha = 0.05$:

```
library(PMCMR)
posthoc.kruskal.nemenyi.test(x=x, g=group, method="Chisq")

##
##  Pairwise comparisons using Tukey and Kramer (Nemenyi) test
##                    with Tukey-Dist approximation for independent samples
##
## data:  x and group
##
##   1       2       3
## 2 0.97310 -       -
## 3 0.00043 0.00334 -
## 4 0.12541 0.29841 0.44923
##
## P value adjustment method: none
```

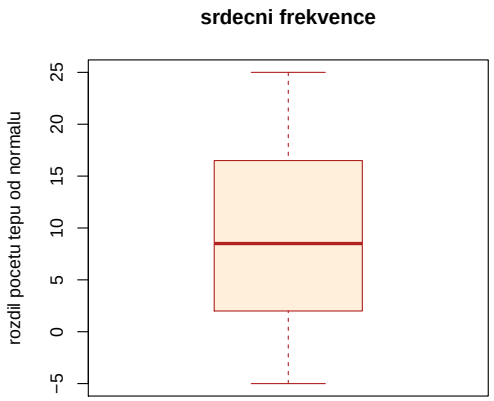
Vidíme, že se liší typy (1, 3) a (2, 3).

Příklady k samostatnému řešení

Příklad 10.5. Ve skupině 12-ti studentů se sledovala srdeční frekvence při změně polohy z lehu do stoje. Získaly se tyto rozdíly počtu tepů srdce za 1 minutu: -2, 4, 8, 25, -5, 16, 3, 1, 12, 17, 20, 9. Za předpokladu, že tyto rozdíly mají symetrické rozdělení, testujte na hladině významnosti $\alpha = 0.05$ hypotézu, že medián rozdílů obou tepových frekvencí je 15 proti oboustranné alternativě. Sestrojte krabicový diagram.

```
## [1] "Wilcoxonuv test: p-value = 0.0499"
## [1] "Znamenkovy test: p-value = 0.3877"
```

Krabicový graf



Výsledek: Zaménkový test nulovou hypotézu nezamítá na hladině významnosti $\alpha = 0.05$, avšak Wilcoxonův test ano.

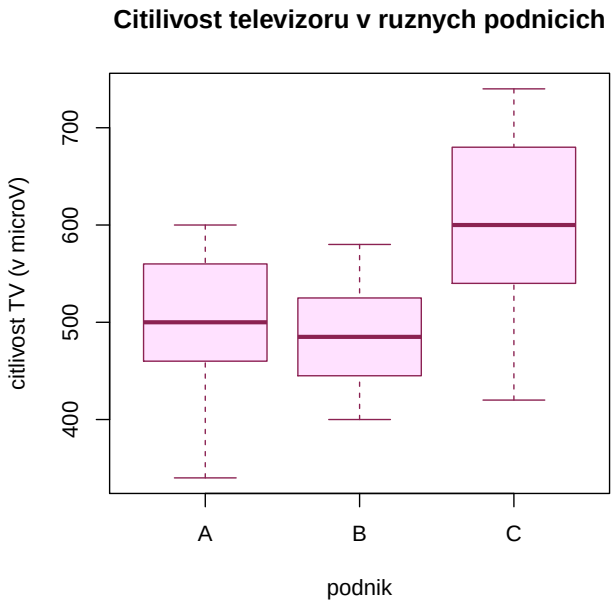
Příklad 10.6. Z produkce tří podniků vyrábějících televizory bylo vylosováno 10, 8 a 12 kusů. Byly získány následující výsledky zjišťování citlivosti těchto televizorů v mikrovoltech:

1.podnik:	420	560	600	490	550	570	340	480	510	460		
2.podnik:	400	420	580	470	470	500	520	530				
3.podnik:	450	700	630	590	420	590	610	540	740	690	540	670

Ověřte na hladině významnosti $\alpha = 0.05$ hypotézu o shodě úrovně citlivosti televizorů v jednotlivých podnicích. Sestrojte krabicové diagramy pro všechny tři podniky.

```
## [1] "Kruskaluv-Wallisuv test: p-value= 0.0157"
```

Krabicovy graf



Metoda mnohonásobného porovnávání

```
##         podnik1 podnik2
## podnik2  0.8728      NA
## podnik3  0.0672  0.0251
```

Výsledek: Na hladině významnosti $\alpha = 0.05$ se liší televizory vyráběné ve 2. a 3. podniku.

11 Hodnocení kontingenčních tabulek

Příklad 11.1. Testování hypotézy o nezávislosti, měření síly závislosti V roce 1950 zkoumali Yule a Kendall barvu očí a vlasů u 6800 mužů. Výsledky zkoumání jsou uvedeny v následující tabulce a v souboru vlasy_oci.txt.

Barva očí	Barva vlasů			
	světlá	kaštanová	černá	rezavá
modrá	1768	807	180	47
šedá/zelená	946	1387	746	53
hnědá	115	438	288	16

Na asymptotické hladině významnosti $\alpha = 0.05$ testujte hypotézu o nezávislosti barvy očí a barvy vlasů. Vypočtěte Cramérův koeficient. Simultánní četnosti znázorněte graficky.

Ověření podmínek dobré aproximace

```
library(scatterplot3d)
data <- read.delim('vlasy_oci.csv',sep=';',dec='.',header=T)
data <- data.frame(data[,2:5], row.names=data[,1])
nazvy.v <- names(data)
nazvy.o <- row.names(data)

# Overeni podmnky dobre aproximace
chisq.test(data)$expected
```

```
##                svetla kastanova    cerna   rezava
## modra       1167.2593  1085.976 500.9024 47.86217
## seda/zelena 1304.7310  1213.875 559.8952 53.49904
## hneda        357.0097   332.149 153.2025 14.63879
```

Podmínky dobré aproximace jsou splněny. Všechny teoretické četnosti jsou větší než 5. Nyní budeme testovat hypotézu o nezávislosti proměnných oci a vlasy.

```
chisq.test(data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 1088.1, df = 6, p-value < 2.2e-16
```

Ve výstupní tabulce najdeme mj. hodnotu testové statistiky $K = 1088.149$ s počtem stupňů volnosti (df= 6) a odpovídající p-hodnotou (p-value $< 2.2e-16$). Protože p-hodnota je mnohem menší než 0.05, nulovou hypotézu o nezávislosti barvy očí a barvy vlasů zamítáme na asymptotické hladině významnosti $\alpha = 0.05$.

Pro zjištění míry závislosti v kontingenční tabulce použijeme Cramérův koeficient.

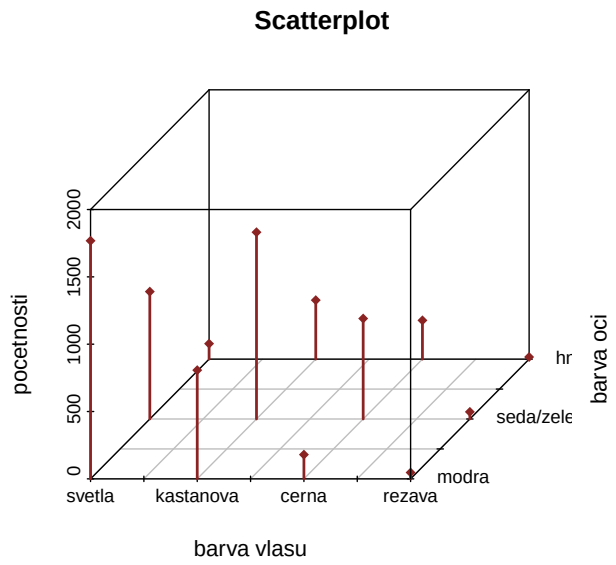
```
library(lsr)
cramersV(data)
```

```
## [1] 0.2830494
```

Hodnota Cramérova koeficientu je 0.283, což svědčí o slabé závislosti barvy očí a vlasů.

Grafické znázornění četností

```
x=rep(c(1,2,3,4),c(3,3,3,3))
y=rep(1:3,4)

z=c(as.matrix(data))
#cbind(x,y,z)
scatterplot3d(x, y , z, type='h', pch=18,
              xlab='barva vlasu', ylab='barva oci', zlab='pocetnosti', main='Scatterplot',
              x.ticklabs = c('svetla','','kastanova','','cerna','','rezava'),
              y.ticklabs = c('modra', '', 'seda/zelena', '','hneda'),color=rep('brown4',12),
              label.tick.marks=T, axis=T, tick.marks=T, angle=40, lwd=2)
```



Příklad k samostatnému řešení

Příklad 11.2. Otevřete si soubor ped_hodnost.txt. Na hladině významnosti $\alpha = 0.05$ testujte hypotézu o nezávislosti pedagogické hodnosti a pohlaví. Dále vypočtěte Cramérův koeficient vyjadřující intenzitu závislosti pedagogické hodnosti na pohlaví. Data v souboru mají následující tvar:

pohlaví	pedagogická hodnost		
	odb. asistent	docent	profesor
muž	32	15	8
žena	34	8	3

```
## [1] "Podminky dobre aproximace:"
##      odb.asistent docent profesor
## muz          36.3  12.65     6.05
## zena         29.7  10.35     4.95
```

Podmínky dobré aproximace jsou splněny, pouze jediná teoretická četnost klesne pod 5.

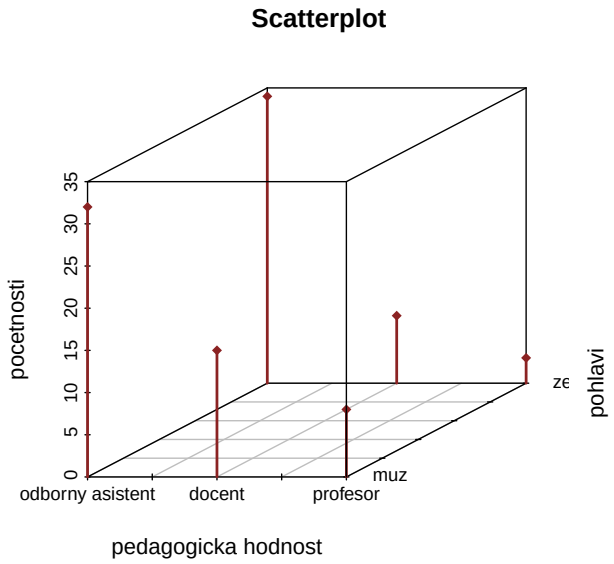
```
## [1] "Chi-kvadratovy test:"
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 3.4988, df = 2, p-value = 0.1739
```

Testovací statistika K nabývá hodnoty 3.5, p-hodnota $= 0.1739$, tedy na asymptotické hladině významnosti $\alpha = 0.05$ nezamítáme hypotézu o nezávislosti pedagogické hodnosti a pohlaví.

```
## [1] "Crameruv koeficient: V= 0.187"
```

Hodnota Cramérova koeficientu je 0.187, což svědčí o slabé závislosti mezi pedagocickou hodností a pohlavím.

Grafické znázornění četností



Příklad 11.3. Fisherův faktoriálový test 100 náhodně vybraných mužů a žen bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

pref. nápoj	pohlaví	
	muž	žena
A	20	30
B	30	20

Na hladině významnosti $\alpha = 0.05$ testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

V našem případě se jedná o oboustranný test (nevíme, zda muži více preferují nápoj A či nápoj B než ženy.

```
data <- data.frame(muz=c(20,30), zena=c(30,20), row.names=c('A','B'))
fisher.test(data, alternative='two.sided')
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data
## p-value = 0.07134
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1846933 1.0640121
## sample estimates:
## odds ratio
##  0.4481632
```

Ve výstupní zprávě funkce fisher.test() je uvedena p-hodnota = 0.07134. Protože p-hodnota je větší než 0.05, nezamítáme na hladině významnosti $\alpha = 0.05$ hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Příklad 11.4. Podíl šancí Pro údaje z příkladu č.3 vypočtěte podíl šancí a sestrojte 95 % asymptotický interval spolehlivosti pro logaritmus podílu šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti $\alpha = 0.05$ hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

Nejprve zopakujme teorii:

Ve čtyřpolních tabulkách používáme charakteristiku

$$OR = \frac{ad}{bc},$$

která se nazývá *podíl šancí* (odds ratio). Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

výsledek pokusu	okolnosti		$n_{j.}$
	I.	II.	
úspěch	a	b	a+b
neúspěch	c	d	c+d
$n_{.k}$	a+c	b+d	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za prvních okolností je $\frac{a}{c}$, za druhých okolností je $\frac{b}{d}$. Podíl šancí je $OR = \frac{ad}{bc}$. Považujeme ho za odhad skutečného podílu šancí $o\rho$. Pomocí $100(1-\alpha)\%$ asymptotického intervalu spolehlivosti pro logaritmus skutečného podílu šancí $\ln o\rho$ lze na asymptotické hladině významnosti α testovat hypotézu o nezávislosti nominálních veličin X a Y.

Upozornění: Musí být splněny podmínky dobré aproximace.

Asymptotický $100(1-\alpha)\%$ interval spolehlivosti pro přirozený logaritmus skutečného podílu šancí má tvar

$$\left(\ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{1-\alpha/2} \,;\, \ln OR - \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\, u_{\alpha/2} \right).$$

Jestliže interval spolehlivosti nezahrne 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

Výpočet příkladu

Ověříme splnění podmínek dobré aproximace a zjistíme, že všechny teoretické četnosti jsou rovny 25.

```r
# Overeni podminek dobre aproximace
a <- 20
b <- 30
c <- 30
d <- 20
alpha <- 0.05
data <- data.frame(muz=c(a,c), zena=c(b,d), row.names=c('A','B'))
chisq.test(data)$expected

##   muz zena
## A  25   25
## B  25   25
```

Podíl šancí

$$OR = \frac{ad}{bc} = \frac{20 * 20}{30 * 30} = \frac{4}{9} = 0.\bar{4}$$

Dolní a horní mez 95% intervalu spolehlivosti pro $\ln o\rho$

$$(-1,61108\,;\,-0,01078)$$

Protože tento interval spolehlivosti neobsahuje 0, na asymptotické hladině významnosti $\alpha = 0.05$ zamítáme hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

```r
# Podil sanci + IS
(OR <- (a*d)/(b*c))

## [1] 0.4444444

(dh <- log(OR)+sqrt(1/a+1/b+1/c+1/d)*qnorm(alpha/2))

## [1] -1.611082

(hh <- log(OR)+sqrt(1/a+1/b+1/c+1/d)*qnorm(1-alpha/2))

## [1] -0.01077827
```

Tento výsledek je v rozporu s výsledkem, ke kterému dospěl Fisherův přesný test. Je to způsobeno tím, že test pomocí asymptotického intervalu spolehlivosti je pouze přibližný. Ke stejnému závěru, jaký jsme dostali u testování pomocí podílu šancí, dospějeme, pokud použijeme Pearsonův chí-kvadrát test o nezávislosti.

```r
chisq.test(data, correct=F)

##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 4, df = 1, p-value = 0.0455
```

Ve funkci chisq.test() však můžeme zadat parametr correct=T, který provede korekci Pearsonova testu pro kontingenční tabulky typu 2×2. Výsledek takto provedeného testu je již v souladu s Fisherovým přesným testem.

```r
chisq.test(data, correct=T)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data
## X-squared = 3.24, df = 1, p-value = 0.07186
```

Příklad 11.5. 36 mužů onemocnělo určitou chorobou. Někteří z nich se léčili, jiní ne. Někteří se uzdravili, jiní zemřeli. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

přežití	léčení	
	ano	ne
ano	10	6
ne	12	8

Vypočtěte a interpretujte podíl šancí. Pomocí intervalu spolehlivosti pro logaritmus podílu šancí testujte na asymptotické hladině významnosti $\alpha = 0.05$ hypotézu, že přežití nezávisí na léčení, proti tvrzení, že léčení zvyšuje šance na přežití.

```
##         muz     zena
## A  9.777778 6.222222
## B 12.222222 7.777778
```

```
## [1] "OR= 1.1111"
## [1] "dolni hranice IS: -1.0283"
```

Výsledek: $OR = 1.\bar{1}$; nulovou hypotézu nezamítáme asymptotické hladině významnosti $\alpha = 0.05$, protože levostranný 95% asymptotický interval spolehlivosti pro logaritmus podílu šancí je

$$(-1.03\,;\,\infty).$$

Příklad k samostatnému řešení

Příklad 11.6. V průzkumu o kuřáctví bylo dotázáno 92 osob. Z 64 mužů jich kouří 19 a z 28 žen jich kouří 6.

a) Na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že kouření se vyskytuje stejně často u mužů a žen. Použijte Pearsonův chi-kvadrát test i Fisherův přesný test.

b) Vypočtěte a interpretujte podíl šancí a stanovte meze 95% intervalu spolehlivosti pro podíl šancí.

Výsledek

ad a) Před provedením Pearsonova chí-kvadrát testu je zapotřebí ověřit splnění podmínek dobré aproximace.

```
##             muz      zena
## kurak   17.3913  7.608696
## nekurak 46.6087 20.391304
```

Jsou splněny, všechny čtyři teoretické četnosti jsou větší než 5.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data
## X-squared = 0.31889, df = 1, p-value = 0.5723
```

Testovací statistika Pearsonova chí-kvadrát testu je $K = 0.6714$, p-hodnota je $0.5723 > 0.05$, tedy nulovou hypotézu nezamítáme na asymptotické hladině významnosti $\alpha = 0.05$.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data
## p-value = 0.4576
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.498056 5.398695
## sample estimates:
## odds ratio
##    1.54109
```

Pro Fisherův přesný test vychází p-hodnota 0.4576, což je větší než hladina významnosti 0.05, nulovou hypotézu nezamítáme na hladině významnosti $\alpha = 0.05$.

ad b)
```
## [1] "OR= 1.5481"
## [1] "dolni hranice IS: 0.5418"
## [1] "horni hranice IS: 4.4239"
```

Podíl šancí je 1.55, což znamená, že u mužů je šance na kouření 1.55× vyšší než u žen. $0.5418 < o\rho < 4.4239$ s pravděpodobností aspoň 0.95.

12 Jednoduchá korelační analýza

Příklad 12.1. Testování nezávislosti ordinálních veličin 12 různých softwarových firem nabízí speciální programové vybavení pro vedení účetnictví. Jednotlivé programy byly posouzeny odbornou komisí složenou z počítačových odborníků a komisí složenou z profesionálních účetních. Úkolem bylo doporučit vhodný program na základě stanovení pořadí jednotlivých programů. Výsledky posouzení:

Produkt firmy číslo	1	2	3	4	5	6	7	8	9	10	11	12
Pořadí dle odborníků	6	7	1	8	4	2.5	9	12	10	2.5	5	11
Pořadí dle účetních	4	5	2	10	6	1	7	11	8	3	12	9

Vypočtěte Spearmanův koeficient pořadové korelace a na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že hodnocení obou komisí jsou nezávislá. Data jsou uložena v souboru ucetnictvi.txt.

```
X <- c(6, 7, 1,  8, 4, 2.5, 9, 12, 10, 2.5,  5, 11)
Y <- c(4, 5, 2, 10, 6,   1, 7, 11,  8,   3, 12,  9)
cor.test(X, Y, method='spearman', exact=F)

##
##  Spearman's rank correlation rho
##
## data:  X and Y
## S = 81.642, p-value = 0.009024
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.714537
```

Spearmanův koeficient pořadové korelace nabývá hodnoty $r_S = 0.7145$, tedy mezi hodnocením obou komisí existuje vysoký stupeň přímé pořadové závislosti. Testovací statistika se realizuje hodnotou 81.642, odpovídající p-hodnota je 0.009024, tedy na asymptotické hladině významnosti $\alpha = 0.05$ zamítáme hypotézu o pořadové nezávislosti hodnocení dvou komisí ve prospěch oboustranné alternativy.

Upozornění: Pokud rozsah výběru nepřesáhne 20, měli bychom testování provést pomocí tabelované kritické hodnoty. V našem případě pro $n = 12$ a $\alpha = 0.05$ je kritická hodnota 0.5804. Vidíme, že nulovou hypotézu zamítáme na hladině významnosti $\alpha = 0.05$, protože $0.7145 \geq 0.5804$.

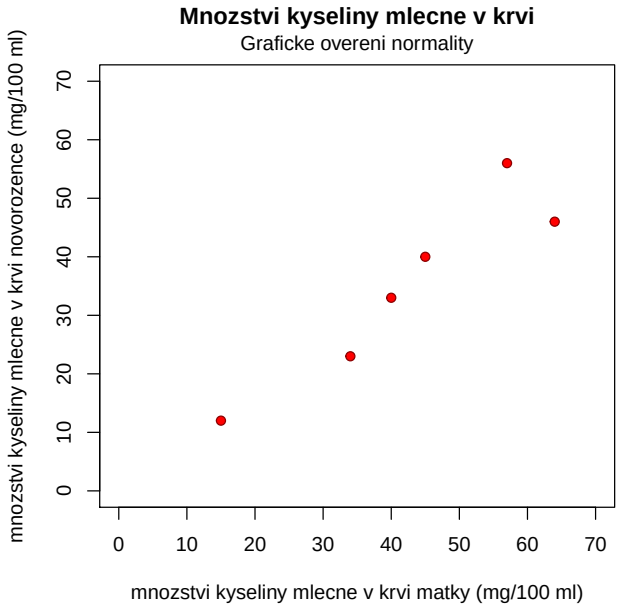
Příklad 12.2. Testování nezávislosti intervalových a poměrových veličin Zjišťovalo se, kolik mg kyseliny mléčné je ve 100 ml krve matek prvorodiček (veličina X) a u jejich novorozenců (veličina Y) těsně po porodu. Byly získány tyto výsledky:

Číslo matky	1	2	3	4	5	6
x_i	40	64	34	15	57	45
y_i	33	46	23	12	56	40

Nakreslete dvourozměrný tečkový diagram, vypočtěte výběrový korelační koeficient, sestrojte 95 % interval spolehlivosti pro korelační koeficient a na hladině významnosti $\alpha = 0.05$ testujte hypotézu o nezávislosti výsledků obou měření. Data jsou uložena v souboru kyselina_mlecna.txt.

Dvourozměrný tečkový diagram

```
data <- read.delim('kyselina_mlecna.txt', header=F)
names(data) <- c('matka', 'dite')
X <- data$matka
Y <- data$dite
plot(X, Y, xlab='mnozstvi kyseliny mlecne v krvi matky (mg/100 ml)',
     ylab='mnozstvi kyseliny mlecne v krvi novorozence (mg/100 ml)',
     main='Mnozstvi kyseliny mlecne v krvi', xlim=c(0,70), ylim=c(0,70),
     pch=21, col='darkred', bg='red')
mtext('Graficke overeni normality', line=0.4)
```



Testování hypotézy o nezávislosti:

```
cor.test(X, Y, type='pearson')

##
##  Pearson's product-moment correlation
##
## data:  X and Y
## t = 5.2653, df = 4, p-value = 0.006232
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5108072 0.9930174
## sample estimates:
##       cor
## 0.9348324
```

Ve výstupní zprávě funkce cor.test() je mj. uvedena hodnota výběrového korelačního koeficientu $r_{12} = 0.9348$, tzn. že mezi X a Y existuje velmi vysoký stupeň přímé lineární závislosti. Hodnota testovací statistiky $t = 5.2653$ a p-hodnota pro test hypotézy o nezávislosti vychází 0.006232, H_0 tedy zamítáme na hladině významnosti $\alpha = 0.05$. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že mezi oběma koncentracemi existuje závislost.

95 % interval spolehlivosti pro ρ mající meze 0.5108 a 0.9930 nepokrývá hodnotu 0, a tudíž hypotézu o nezávislosti veličin X, Y zamítáme na hladině významnosti $\alpha = 0.05$.

Příklad 12.3. Porovnání dvou korelačních koeficientů V psychologickém výzkumu bylo vyšetřeno 426 hochů a 430 dívek. Ve skupině hochů činil výběrový koeficient korelace mezi verbální a performační složkou IQ 0.6033, ve skupině dívek činil 0.5833. Za předpokladu dvourozměrné normality dat testujte na hladině významnosti $\alpha = 0.05$ hypotézu, že korelační koeficienty se neliší.

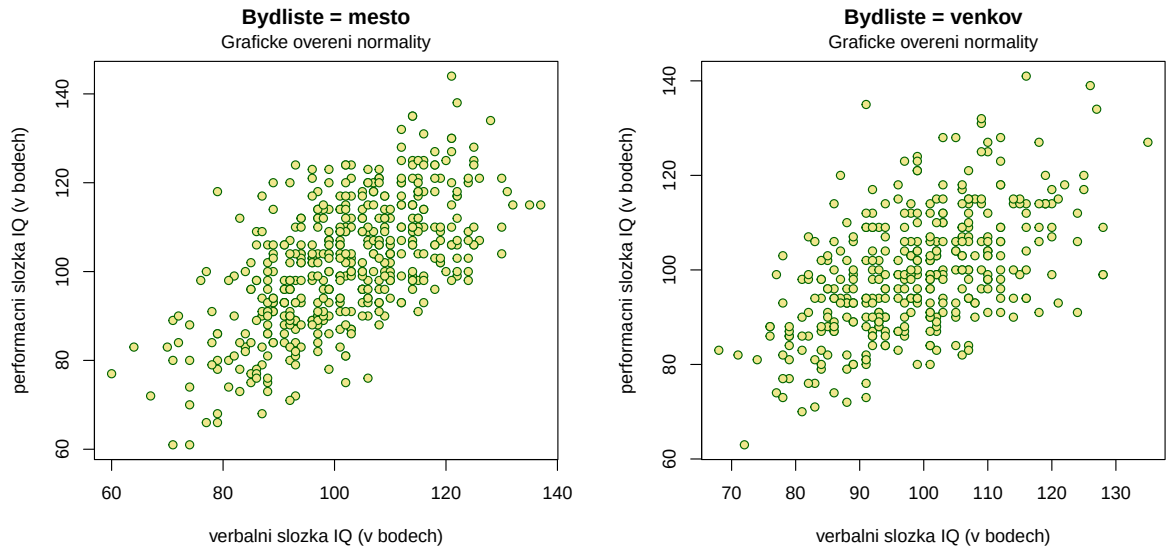
```
R1 <- 0.6033
R2 <- 0.5833
n1 <- 426
n2 <- 430
ksi <- 0
Z1 <- 1/2*log((1+R1)/(1-R1))
Z2 <- 1/2*log((1+R2)/(1-R2))
Zw <- (Z1-Z2-ksi)/sqrt(1/(n1-3)+1/(n2-3))

(p.val <- 2*min(pnorm(Zw), 1-pnorm(Zw)))

## [1] 0.6527169
```

Výsledek:
Příslušná p-hodnota je 0.6528, tedy nezamítáme nulovou hypotézu o shodě dvou koeficientů korelace na asymptotické hladině významnosti $\alpha = 0.05$.

Příklady k samostatnému řešení

Příklad 12.4. Načtěte datový soubor IQ.txt. Za předpokladu dvourozměrné normality dat (orientačně ověřte pomocí dvourozměrného tečkového diagramu) testujte na hladině významnosti $\alpha = 0.1$ hypotézu, že korelační koeficienty mezi verbální a performační složkou IQ jsou stejné u dětí z města a venkova.

```
## [1] 0.0780111
```



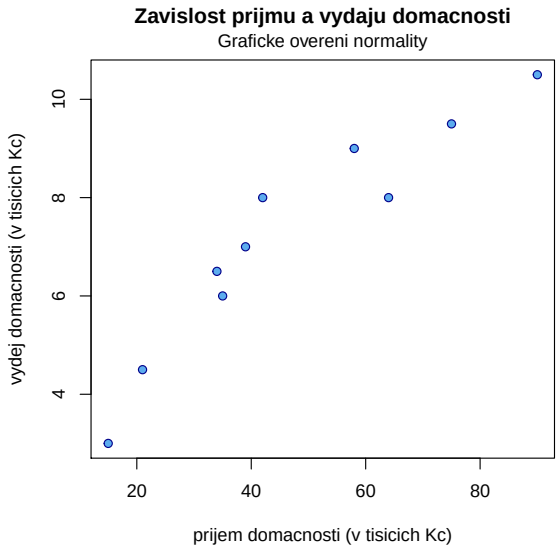
Výsledek p-hodnota= 0.07801, tedy s rizikem omylu nejvýše 10% jsme prokázali, že korelační koeficienty se liší.

Příklad 12.5. V náhodném výběru 10 dvoučlenných domácností byl zjišťován měsíční příjem (veličina X, v tisících Kč) a vydání za potraviny (veličina Y, v tisících Kč).

x_i	15	21	34	35	39	42	58	64	75	90
y_i	3	4.5	6.5	6	7	8	9	8	9.5	10.5

Vypočtěte a interpretujte výběrový koeficient korelace. Na hladině významnosti $\alpha = 0.05$ testujte hypotézu o nezávislosti veličin X, Y. Sestrojte 95 % asymptotický interval spolehlivosti pro ρ. Data jsou uložena v souboru prijem_vydani.txt.

Grafické ověření normality

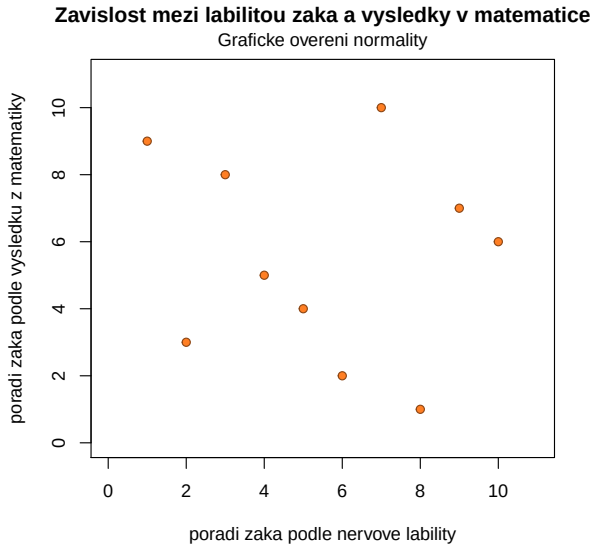


Výsledek
$r_{12} = 0.9405$, mezi měsíčními příjmy a výdaji tedy existuje velmi vysoký stupeň přímé lineární závislosti. p-hodnota$= 5.095\,\mathrm{e}-05$, tedy H_0 zamítáme na hladině významnosti $\alpha = 0.05$. S pravděpodobností alespoň 0.95 platí: $0.7623 < \rho < 0.9862$.

Příklad 12.6. Bylo sledováno 10 žáků. Na základě psychologického vyšetření byli tito žáci seřazeni podle nervové lability (čím byl žák labilnější, tím dostal vyšší pořadí R_i). Kromě toho sledování žáci dostali pořadí Q_i na základě svých výsledků v matematice (nejlepší žák v matematice dostal pořadí 1). Výsledky jsou uvedeny v tabulce:

Pořadí Ri	1	2	3	4	5	6	7	8	9	10
Pořadí Qi	9	3	8	5	4	2	10	1	7	6

Vypočtěte vhodný korelační koeficient a jeho hodnotu řádně interpretujte. Na hladině významnosti $\alpha = 0.05$ testujte hypotézu, že nervová labilita a výsledky v matematice jsou nezávislé. Data jsou uložena v souboru nervova_labilita.txt



Výsledek: Spearmanův koeficient pořadové korelace $r_S = -0.127$, tedy mezi nervovou labilitou žáka a jeho výsledky v matematice existuje nízký stupeň nepřímé pořadové závislosti. p-hodnota$= 0.7329$, a tedy H_0 nezamítáme na hladině významnosti $\alpha = 0.05$.