Zadanie

Za istých podmienok (vyhľadajte) možno hypergeometrické rozdelenie aproximovať binomickým rozdelením.

1. Majme balíček kariet, náhodne vyberieme 5 kariet. Aká je pravdepodobnosť, že sme vytiahli dve esá?
2. Majme 10 balíčkov kariet, náhodne vyberieme 5 kariet. Aká je pravdepodobnosť, že sme vytiahli dve esá?

Riešte tieto úlohy binomickým a hypergeometrickým rozdelením, porovnajte rozdiely. Zostrojte graf rozdelenia pravdepodobnosti (binomického a hypergeometrického) pre prvú a druhú úlohu. Porovnajte rozdiely.

R Notebook DU2

Code

Filip Agh

zadanie:

Zadanie Za istých podmienok (vyhľadajte) možno hypergeometrické rozdelenie aproximovať binomickým rozdelením.

1. Majme balíček kariet, náhodne vyberieme 5 kariet. Aká je pravdepodobnosť, že sme vytiahli dve esá?
2. Majme 10 balíčkov kariet, náhodne vyberieme 5 kariet. Aká je pravdepodobnosť, že sme vytiahli dve esá?

Riešte tieto úlohy binomickým a hypergeometrickým rozdelením, porovnajte rozdiely. Zostrojte graf rozdelenia pravdepodobnosti (binomického a hypergeometrického) pre prvú a druhú úlohu. Porovnajte rozdiely

info hracie karty pocet 54 (2x zolik, 4x eso)

uloha 1

Hide

```
dbinom(2, 5, 4 / 54)
```

```
[1] 0.04355732
```

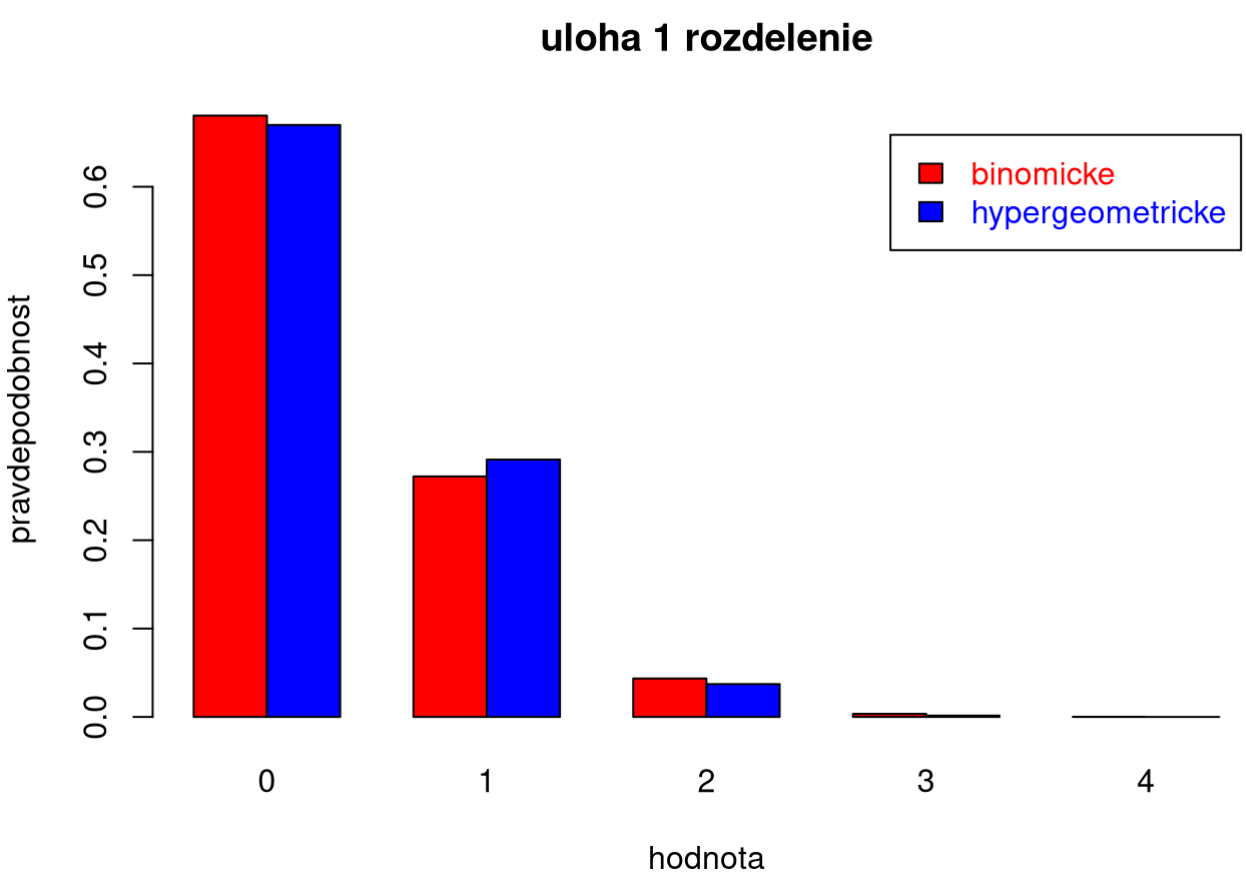
Hide

```
dhyper(2, 4, 54 - 4, 5)
```

```
[1] 0.03718565
```

Hide

```
xB <- c(0:4)
hustotaB <- dbinom(xB, 5, 4 / 54)
hustota2 <- dhyper(xB, 4, 54 - 4, 5)
tabulka <- data.frame(hodnota = xB, pravdepodobnost = hustotaB, hypgeo = hustota2)
collnames <- c('binomicke', 'hypergeometricke')
colcolor <- c('red','blue')
barplot(t(tabulka[c('pravdepodobnost', 'hypgeo')]), beside = T, main = "uloha 1 rozde
lenie", xlab = "hodnota", ylab = "pravdepodobnost", names.arg = xB, col = colcolor, l
egend.text = collnames, args.legend = list(text.col=colcolor,col=colcolor))
```



uloha 2

Hide

```
dbinom(2, 5, 40 / 540)
```

```
[1] 0.04355732
```

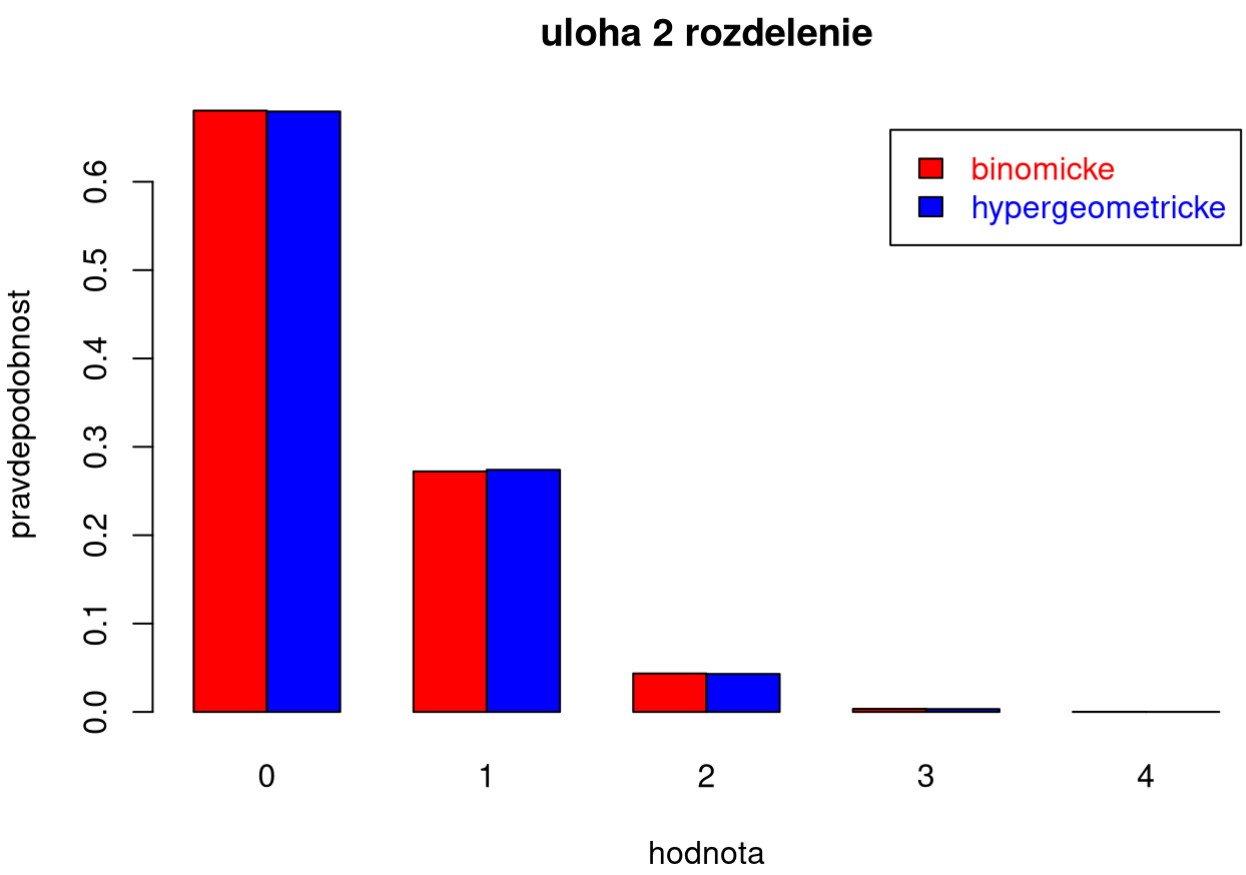
Hide

```
dhyper(2, 40, 540 - 40, 5)
```

```
[1] 0.04300516
```

Hide

```
xB <- c(0:4)
hustotaB <- dbinom(xB, 5, 40 / 540)
hustota2 <- dhyper(xB, 40, 540 - 40, 5)
tabulka <- data.frame(hodnota = xB, pravdepodobnost = hustotaB, hypgeo = hustota2)
collnames <- c('binomicke', 'hypergeometricke')
colcolor <- c('red','blue')
barplot(t(tabulka[c('pravdepodobnost', 'hypgeo')]), beside = T, main = "uloha 2 rozde
lenie", xlab = "hodnota", ylab = "pravdepodobnost", names.arg = xB, col = colcolor, l
egend.text = collnames, args.legend = list(text.col=colcolor,col=colcolor))
```



porovnanie rozdielov

v pripade malej pocetnosti sa pravdepodobnost nezhoduje, naopak v pripade velkej pocetnosti je velmi podobna az ju mozeme povazovat za zhodnu, rozdiel je sposobeny principialnym rozdielom tychto rozdeleni, kde hypergeometricke je odvisle od predoslich pokusov ale binomicke nezohladnije pokusy a ma konstantnu pravdepodobnost.

Zadanie

Dataset diamonds obsahuje vyše 50 tisíc údajov o cene a iných charakteristikách diamantov.

The descriptiveness for the documentation will vary, depending on the package author.

Variable	Description	Values
price	price in US dollars	$326-$18,823
carat	weight of the diamond	0.2-5.01
cut	quality of the cut	Fair, Good, Very Good, Premium, Ideal
color	diamond color	J (worst) to D (best)
clarity	measurement of how clear the diamond is	I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)
x	length in mm	0-10.74
y	width in mm	0-58.9
z	depth in mm	0-31.8
depth	total depth percentage	43-79
table	width of top of diamond relative to widest point	43-95

Vyberte náhodne 100 vektorov meraní. Urobte prvotnú štatistickú analýzu (výpočtovú a grafickú) pre premennú cena vzhľadom na vybraný atribút.

cv3_DU_MarekVitaz.R

Marek

2022-03-08

Domaca uloha - cvicenie c.3 Meno: Marek Vitaz AIS ID: 97103

```
library(psych)
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following object is masked from 'package:psych':
##
##     describe
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(FSA)
```

```
## ## FSA v0.9.3. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
##
## Attaching package: 'FSA'
```

```
## The following object is masked from 'package:psych':
##
##     headtail
```

```
library(pastecs)
library(moments)
library(ggplot2)
library(vioplot)
```

```
## Loading required package: sm
```

```
## Package 'sm', version 2.2-5.7: type help(sm) for summary information
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(RColorBrewer)
library(rmarkdown)
```

Import datasetu diamantov z kniznice

```
diamanty <- ggplot2::diamonds
```

Nahodny vyber 100 prvkov z mnoziny diamantov

```
#nahodna_price <- sample (diamanty$price, size = 100)
nahodny_vyber <- diamanty[sample(nrow(diamanty), 100), ]
```

Minimum, maximum, aritmeticky priemer a median ceny vybranych diamantov

```
min(nahodny_vyber$price)
```

```
## [1] 428
```

```
max(nahodny_vyber$price)
```

```
## [1] 17474
```

```
mean(nahodny_vyber$price)
```

```
## [1] 3786.07
```

```
median(nahodny_vyber$price)
```

```
## [1] 2309
```

Tabulky realitvnych a kumulativnych pocetnosti

```
table(nahodny_vyber$price)
```

```
##
##   428   441   462   489   524   536   579   609   639   666   680   709   720
##     1     1     1     1     1     1     1     1     1     1     1     1     1
##   729   745   816   847   855   872   873   885   904   923   965   995  1013
##     1     1     1     1     1     2     1     1     1     1     1     1     1
##  1033  1042  1063  1107  1114  1140  1147  1158  1179  1292  1554  1559  1617
##     1     1     1     2     1     1     1     1     1     1     1     1     1
##  1715  1732  1912  1986  2042  2131  2208  2227  2246  2372  2409  2549  2763
##     1     1     1     1     1     1     1     1     1     1     1     1     1
##  2821  2863  2914  2964  3615  3847  4072  4119  4191  4259  4378  4544  4561
##     1     1     1     1     1     1     1     1     1     1     1     1     1
##  4712  4766  4937  5014  5165  5208  5358  5418  5586  5626  5696  5750  6096
##     1     1     1     1     1     1     1     1     1     1     1     1     1
##  6267  6794  6871  7094  7191  7485  7504  8266  8429  8496  8602  8637  9025
##     1     1     1     1     1     1     1     1     1     1     1     1     1
##  9979 10761 10800 12300 16343 16629 17474
##     1     1     1     1     1     1     1
```

```
prop.table(nahodny_vyber$price)
```

```
##   [1] 0.002728423 0.002337516 0.005245545 0.007696635 0.014151878 0.002305821
##   [7] 0.014754085 0.002675598 0.028525622 0.014310353 0.004104520 0.002628055
##  [13] 0.004270919 0.019820024 0.012046793 0.001384021 0.007561931 0.014859736
##  [19] 0.002303180 0.003058581 0.001415716 0.003412510 0.006362798 0.010160932
##  [25] 0.001796058 0.005882089 0.011249132 0.012445623 0.002548817 0.021832666
##  [31] 0.026357146 0.004529763 0.018148106 0.001872654 0.016101129 0.001291577
##  [37] 0.007450998 0.002807661 0.001164796 0.002258278 0.032487513 0.013039907
##  [43] 0.012588251 0.010879355 0.011563442 0.013243284 0.009548159 0.013755689
##  [49] 0.043921533 0.006732575 0.022263191 0.001759080 0.022720129 0.004117726
##  [55] 0.046153399 0.013642114 0.018737107 0.043166133 0.003114047 0.022812573
##  [61] 0.005628528 0.001130460 0.002303180 0.001220263 0.012001891 0.002923876
##  [67] 0.016552784 0.015044624 0.007297805 0.002942365 0.002437884 0.006265072
##  [73] 0.001529290 0.001608528 0.010755216 0.002752194 0.002923876 0.007828698
##  [79] 0.002387700 0.004574664 0.003011038 0.022440156 0.005932273 0.003029527
##  [85] 0.005050092 0.023837383 0.001967740 0.017944729 0.005393455 0.011069526
##  [91] 0.005831905 0.001687766 0.028422612 0.018993310 0.002237148 0.019769840
##  [97] 0.001925479 0.002155269 0.015187252 0.001901708
```

```
cumsum(table(nahodny_vyber$price))
```

```
##   428   441   462   489   524   536   579   609   639   666   680   709   720
##     1     2     3     4     5     6     7     8     9    10    11    12    13
##   729   745   816   847   855   872   873   885   904   923   965   995  1013
##    14    15    16    17    18    20    21    22    23    24    25    26    27
##  1033  1042  1063  1107  1114  1140  1147  1158  1179  1292  1554  1559  1617
##    28    29    30    32    33    34    35    36    37    38    39    40    41
##  1715  1732  1912  1986  2042  2131  2208  2227  2246  2372  2409  2549  2763
##    42    43    44    45    46    47    48    49    50    51    52    53    54
##  2821  2863  2914  2964  3615  3847  4072  4119  4191  4259  4378  4544  4561
##    55    56    57    58    59    60    61    62    63    64    65    66    67
##  4712  4766  4937  5014  5165  5208  5358  5418  5586  5626  5696  5750  6096
##    68    69    70    71    72    73    74    75    76    77    78    79    80
##  6267  6794  6871  7094  7191  7485  7504  8266  8429  8496  8602  8637  9025
##    81    82    83    84    85    86    87    88    89    90    91    92    93
##  9979 10761 10800 12300 16343 16629 17474
##    94    95    96    97    98    99   100
```

```
cumsum(prop.table(nahodny_vyber$price))
```

```
##   [1] 0.002728423 0.005065939 0.010311484 0.018008119 0.032159997 0.034465818
##   [7] 0.049219903 0.051895501 0.080421123 0.094731476 0.098835996 0.101464051
##  [13] 0.105734971 0.125554995 0.137601788 0.138985809 0.146547739 0.161407475
##  [19] 0.163710655 0.166769236 0.168184952 0.171597461 0.177960260 0.188121192
##  [25] 0.189917249 0.195799338 0.207048470 0.219494093 0.222042910 0.243875575
##  [31] 0.270232722 0.274762485 0.292910591 0.294783245 0.310884374 0.312175950
##  [37] 0.319626948 0.322434609 0.323599405 0.325857684 0.358345197 0.371385104
##  [43] 0.383973355 0.394852710 0.406416152 0.419659436 0.429207595 0.442963284
##  [49] 0.486884817 0.493617392 0.515880583 0.517639663 0.540359793 0.544477519
##  [55] 0.590630918 0.604273032 0.623010140 0.666176272 0.669290320 0.692102893
##  [61] 0.697731421 0.698861881 0.701165060 0.702385323 0.714387214 0.717311090
##  [67] 0.733863875 0.748908499 0.756206304 0.759148669 0.761586553 0.767851625
##  [73] 0.769380915 0.770989443 0.781744659 0.784496853 0.787420729 0.795249428
##  [79] 0.797637128 0.802211792 0.805222830 0.827662986 0.833595258 0.836624785
##  [85] 0.841674877 0.865512259 0.867479999 0.885424728 0.890818183 0.901887709
##  [91] 0.907719614 0.909407380 0.937829993 0.956823302 0.959060451 0.978830291
##  [97] 0.980755770 0.982911040 0.998098292 1.000000000
```

Usporiadanie nahodneho vyberu a zistovanie uzitocnych pocetnosti

```
sort(nahodny_vyber$price)
```

```
##   [1]   428   441   462   489   524   536   579   609   639   666   680   709
##  [13]   720   729   745   816   847   855   872   872   873   885   904   923
##  [25]   965   995  1013  1033  1042  1063  1107  1107  1114  1140  1147  1158
##  [37]  1179  1292  1554  1559  1617  1715  1732  1912  1986  2042  2131  2208
##  [49]  2227  2246  2372  2409  2549  2763  2821  2863  2914  2964  3615  3847
##  [61]  4072  4119  4191  4259  4378  4544  4561  4712  4766  4937  5014  5165
##  [73]  5208  5358  5418  5586  5626  5696  5750  6096  6267  6794  6871  7094
##  [85]  7191  7485  7504  8266  8429  8496  8602  8637  9025  9979 10761 10800
##  [97] 12300 16343 16629 17474
```

```
sort(nahodny_vyber$price) [10]
```

```
## [1] 666
```

```
sum(nahodny_vyber$price==1727)
```

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

```
sum(nahodny_vyber$price>1500)
```

```
## [1] 62
```

Porovnanie rozdielov medzi aritmetickym, geometrickym a harmonickym priemerom

```
mean(nahodny_vyber$price)
```

```
## [1] 3786.07
```

```
geometric.mean(nahodny_vyber$price)
```

```
## [1] 2381.066
```

```
harmonic.mean(nahodny_vyber$price)
```

```
## [1] 1528.785
```

Tri dolezite kvantily

```
quantile(nahodny_vyber$price,0.25)
```

```
##   25%
## 987.5
```

```
quantile(nahodny_vyber$price,0.5)
```

```
##  50%
## 2309
```

```
quantile(nahodny_vyber$price,0.75)
```

```
##  75%
## 5460
```

Variacne rozpatie

```
max(nahodny_vyber$price) - min(nahodny_vyber$price)
```

```
## [1] 17046
```

Interquartile range

```
IQR(nahodny_vyber$price)
```

```
## [1] 4472.5
```

Sikmost a spicatost nahodneho vyberu ceny diamantov

```
skewness(nahodny_vyber$price)
```

```
## [1] 1.632325
```

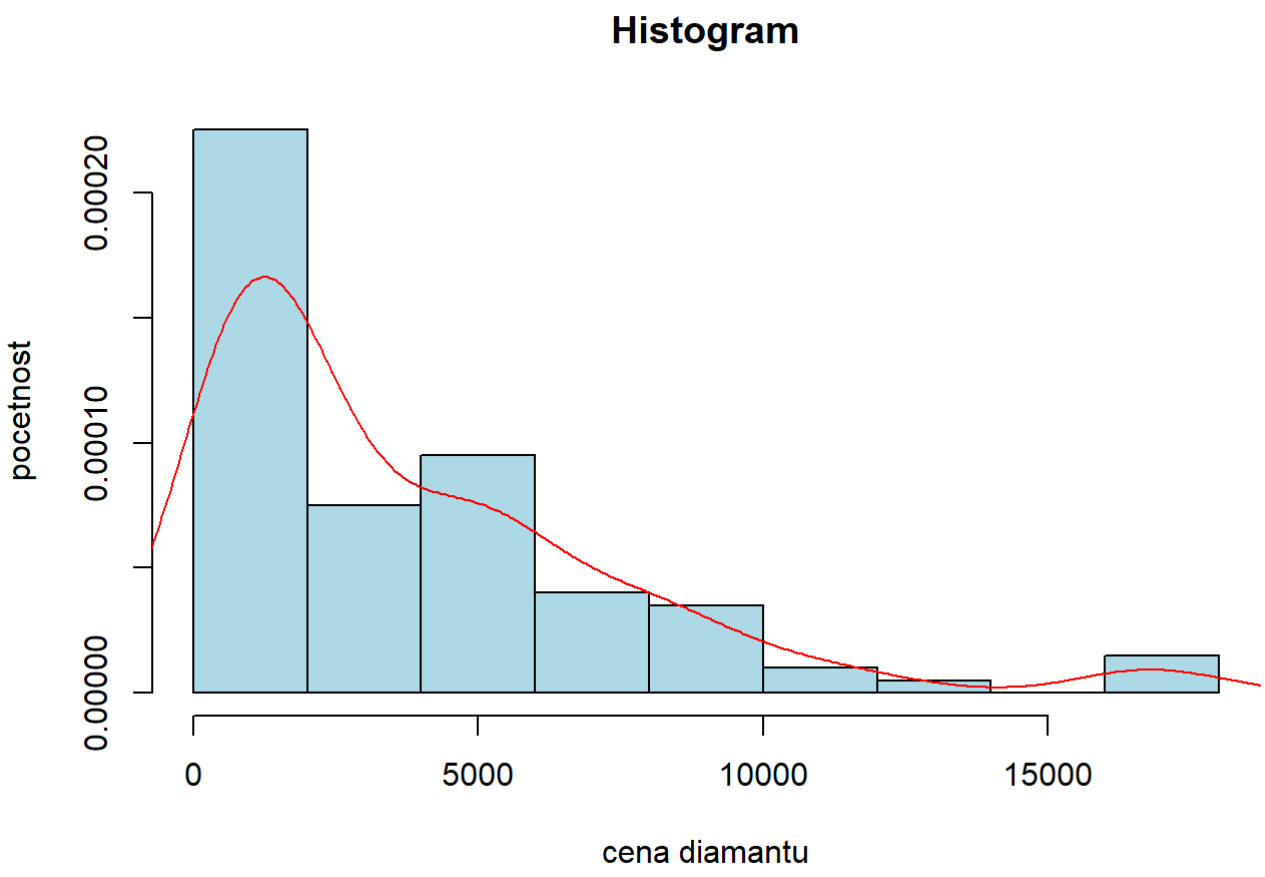
>0 - nakloneny dolava <0 - nakloneny doprava

```
kurtosis(nahodny_vyber$price)
```

```
## [1] 5.815065
```

>3 - spicaty <3 - plochy - Rychly histogram relativnych pocetnosti na potvrdenie vysledkov sikomsti a spicatosti

```
par(mfrow=c(1,1))
hist(nahodny_vyber$price, breaks = "Sturges", col = "lightblue", main = "Histogram",
xlab = "cena diamantu", ylab = "pocetnost", freq = F)
lines(density(nahodny_vyber$price), col = "red")
```



Sumar zistenych poznatkov

```
Summarize(nahodny_vyber$price)
```

```
##         n      mean        sd       min        Q1    median        Q3       max
##   100.000  3786.070  3692.602   428.000   987.500  2309.000  5460.000 17474.000
```

Porovnanie priemernych cien pre nahodny vyber a celu mnozinu diamantov v dvoch roznych atributoch

```
tapply(nahodny_vyber$price,nahodny_vyber$cut,mean)
```

```
##      Fair      Good Very Good   Premium     Ideal
##  2905.400  3297.154  4354.727  5305.591  2860.263
```

```
tapply(diamanty$price,diamanty$cut,mean)
```

```
##      Fair      Good Very Good   Premium     Ideal
##  4358.758  3928.864  3981.760  4584.258  3457.542
```

```
tapply(nahodny_vyber$price,nahodny_vyber$clarity,mean)
```

```
##       I1      SI2      SI1      VS2      VS1     VVS2     VVS1       IF
## 6514.333 4304.059 4434.625 3035.000 3995.062 1697.900 5040.833  774.500
```

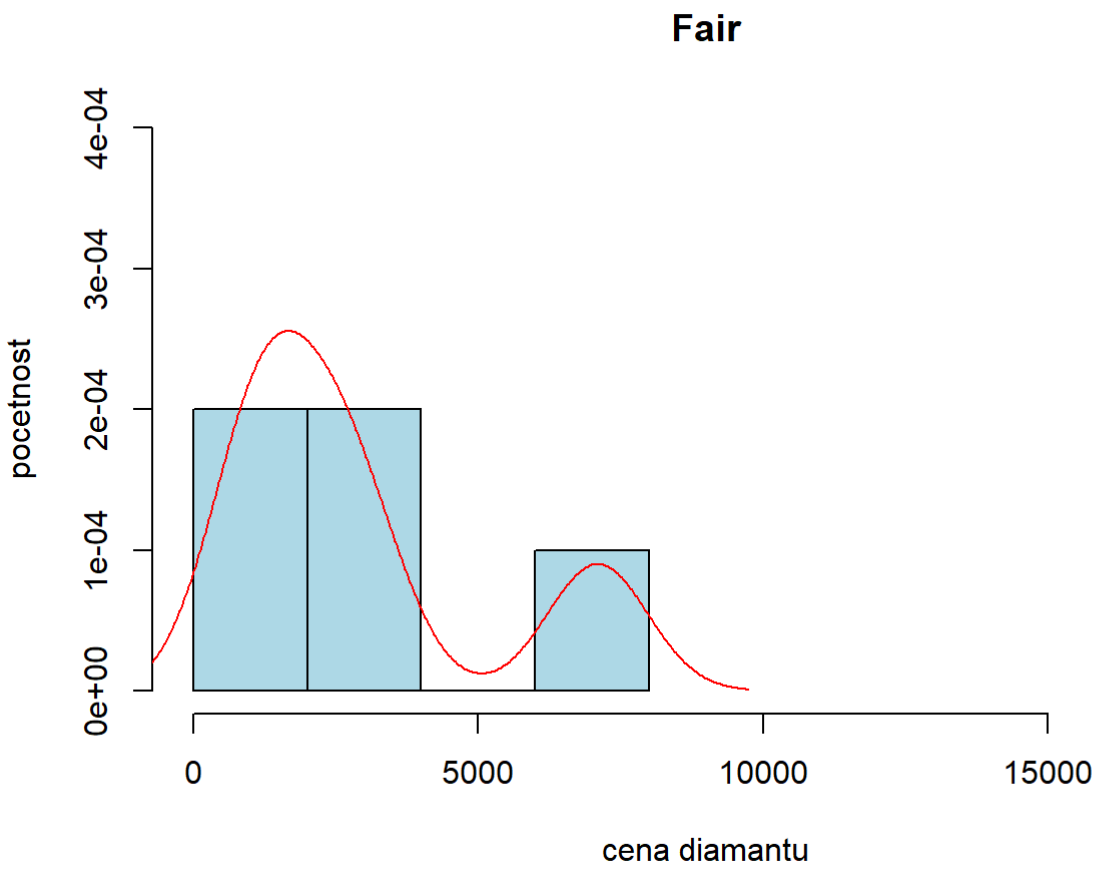
```
tapply(diamanty$price,diamanty$clarity,mean)
```

```
##       I1      SI2      SI1      VS2      VS1     VVS2     VVS1       IF
## 3924.169 5063.029 3996.001 3924.989 3839.455 3283.737 2523.115 2864.839
```

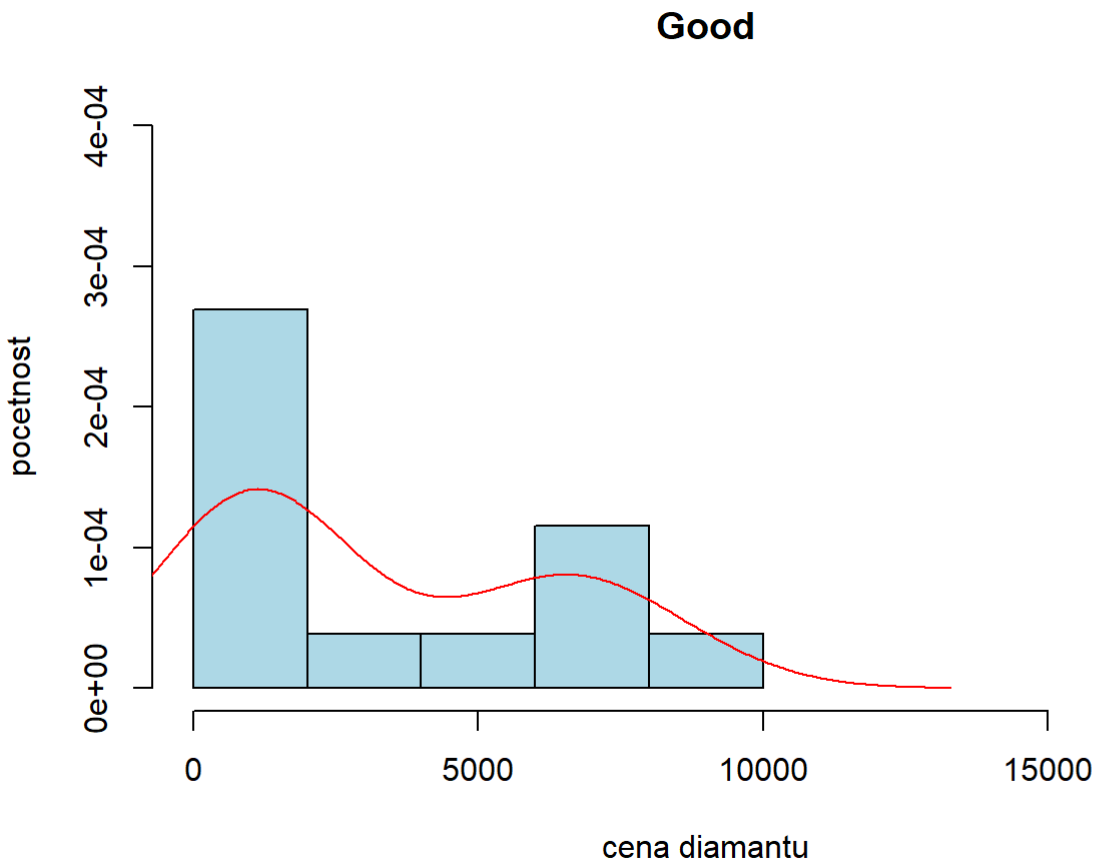
```
ceny <- nahodny_vyber$price

fair <- subset(nahodny_vyber$price,nahodny_vyber$cut == "Fair")
good <- subset(nahodny_vyber$price,nahodny_vyber$cut == "Good")
very_good <- subset(nahodny_vyber$price,nahodny_vyber$cut == "Very Good")
premium <- subset(nahodny_vyber$price,nahodny_vyber$cut == "Premium")
ideal <- subset(nahodny_vyber$price,nahodny_vyber$cut == "Ideal")

fair_all <- subset(diamanty$price,diamanty$cut == "Fair")
good_all <- subset(diamanty$price,diamanty$cut == "Good")
very_good_all <- subset(diamanty$price,diamanty$cut == "Very Good")
premium_all <- subset(diamanty$price,diamanty$cut == "Premium")
ideal_all <- subset(diamanty$price,diamanty$cut == "Ideal")

par(mfrow=c(2,5))
```

Histogramy pre pocetnost a cenu diamantov v zavislosti od ich jednotlivych rezov - Porovnanie medzi nahodnym vyberom a celou mnozinou diamantov - Je nahodny vyber dobrou reprezentaciou celeho suboru diamantov? Porovname z histogramov

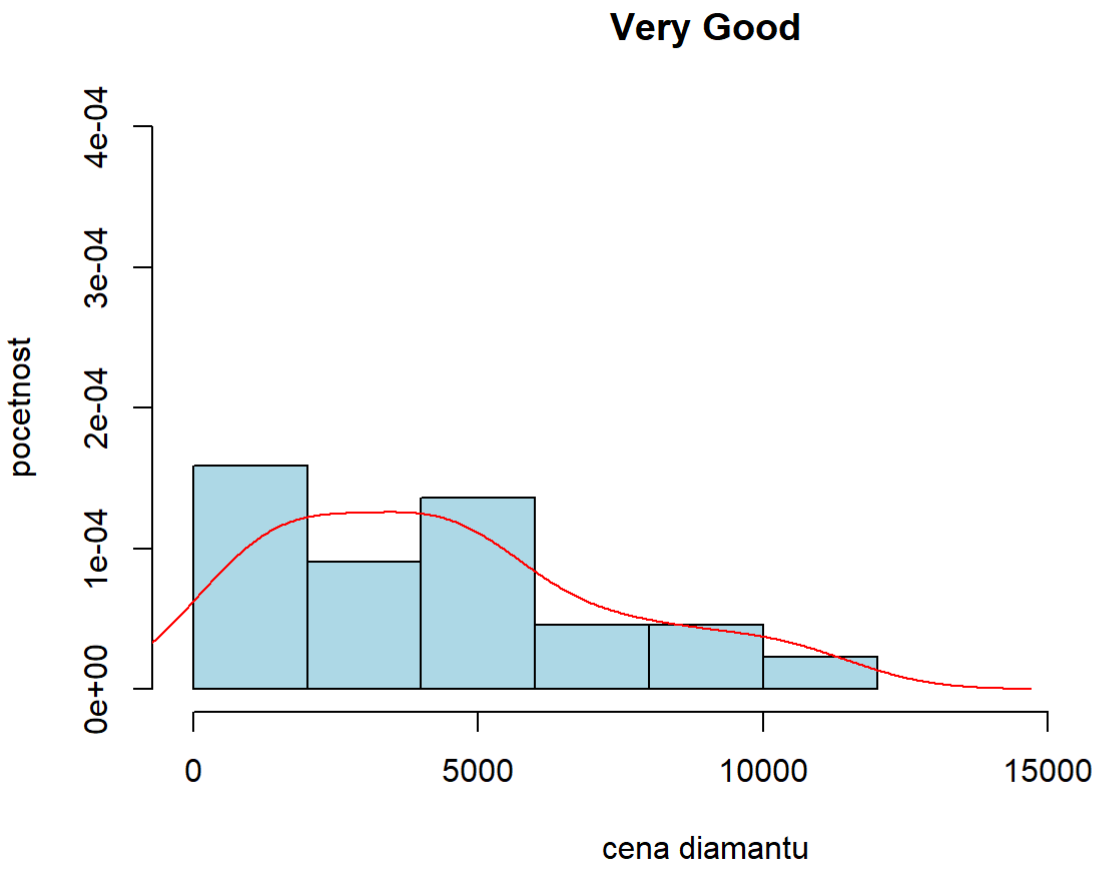
```
hist(fair, breaks = "Sturges", col = "lightblue", main = "Fair", xlab = "cena diamant
u", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histogram relat
ivnych pocetnosti
lines(density(fair), col = "red")
```



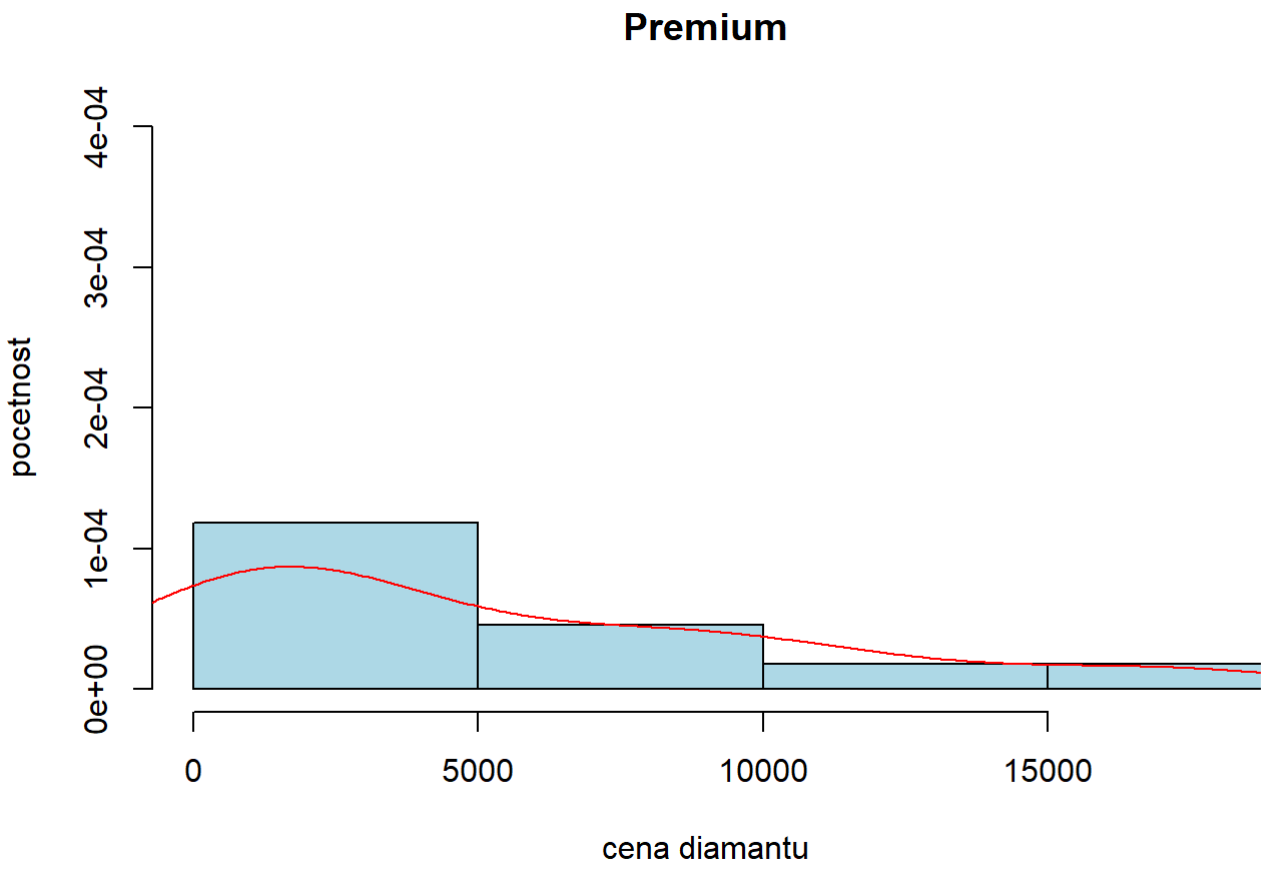
```
hist(good, breaks = "Sturges", col = "lightblue", main = "Good", xlab = "cena diamant
u", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histogram relat
ivnych pocetnosti
lines(density(good), col = "red")
```



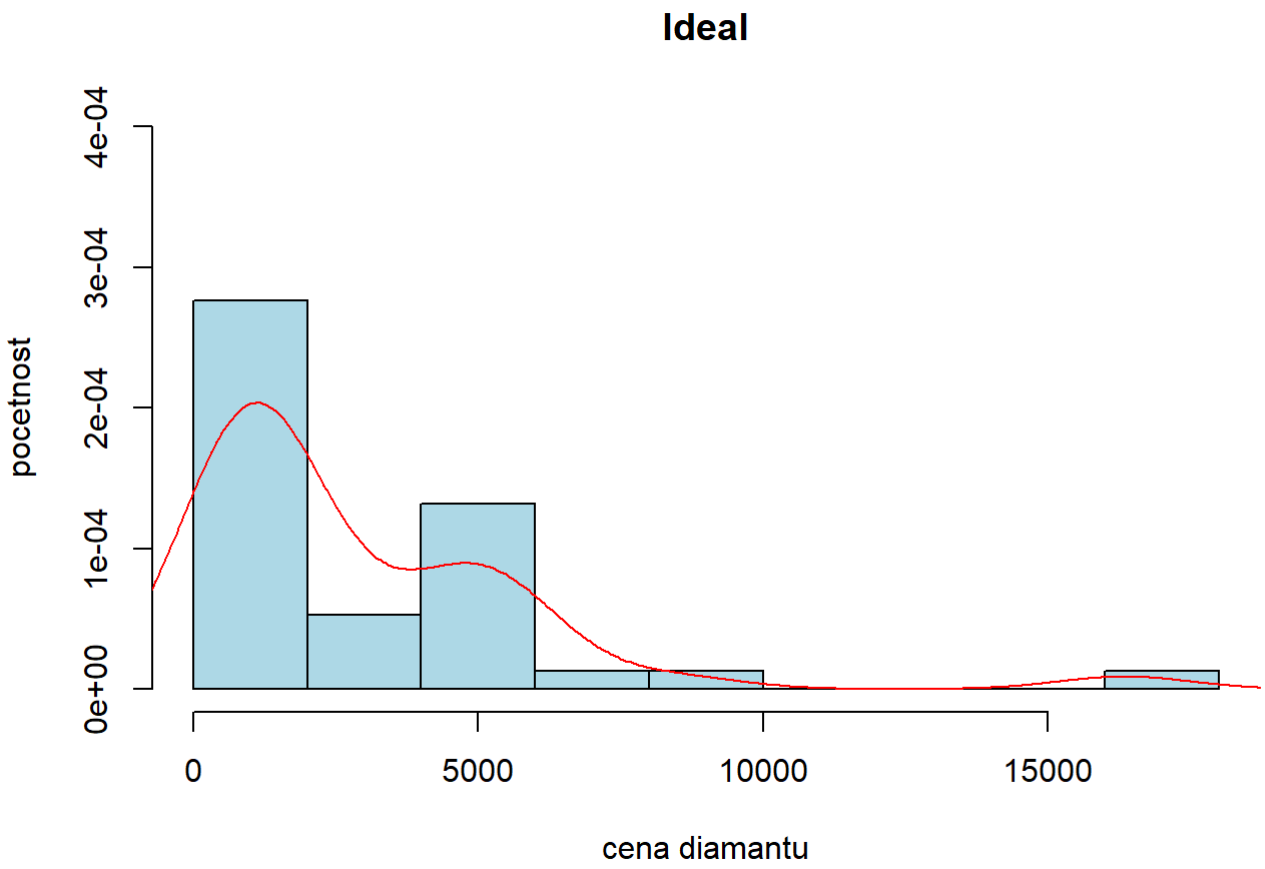

```
hist(very_good, breaks = "Sturges", col = "lightblue", main = "Very Good", xlab = "ce
na diamantu", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histo
gram relativnych pocetnosti
lines(density(very_good), col = "red")
```



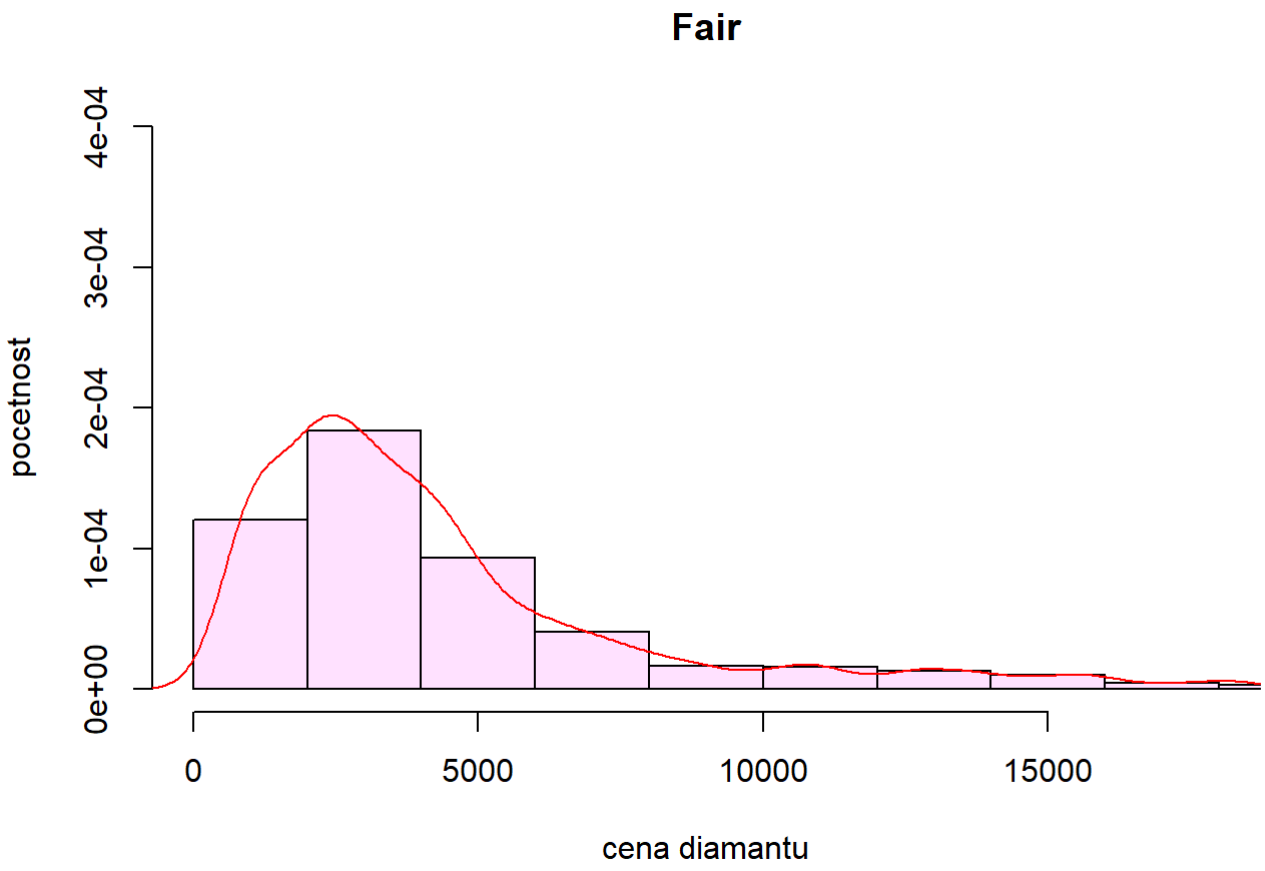

```
hist(premium, breaks = "Sturges", col = "lightblue", main = "Premium", xlab = "cena d
iamantu", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histogram
relativnych pocetnosti
lines(density(premium), col = "red")
```



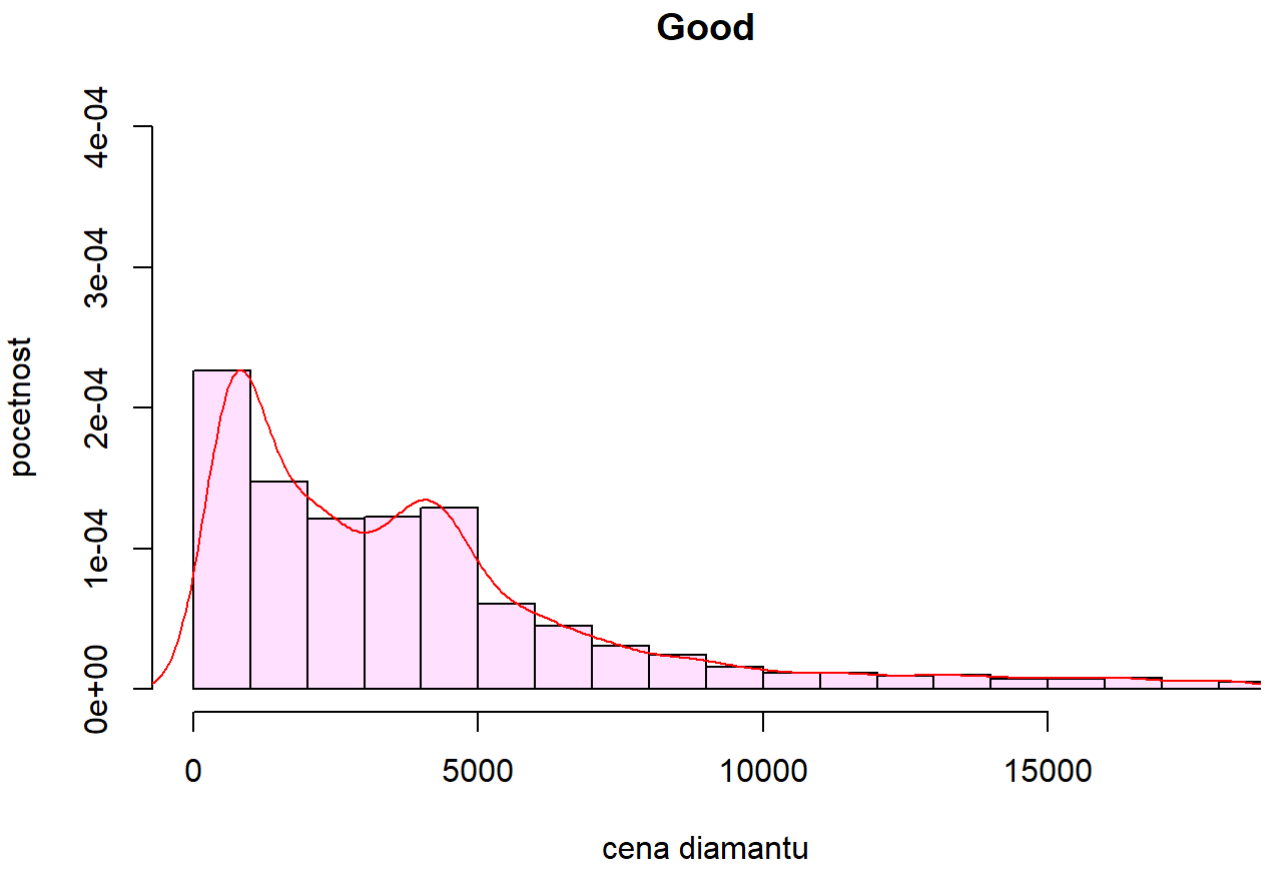

```
hist(ideal, breaks = "Sturges", col = "lightblue", main = "Ideal", xlab = "cena diama
ntu", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histogram rel
ativnych pocetnosti
lines(density(ideal), col = "red")
```



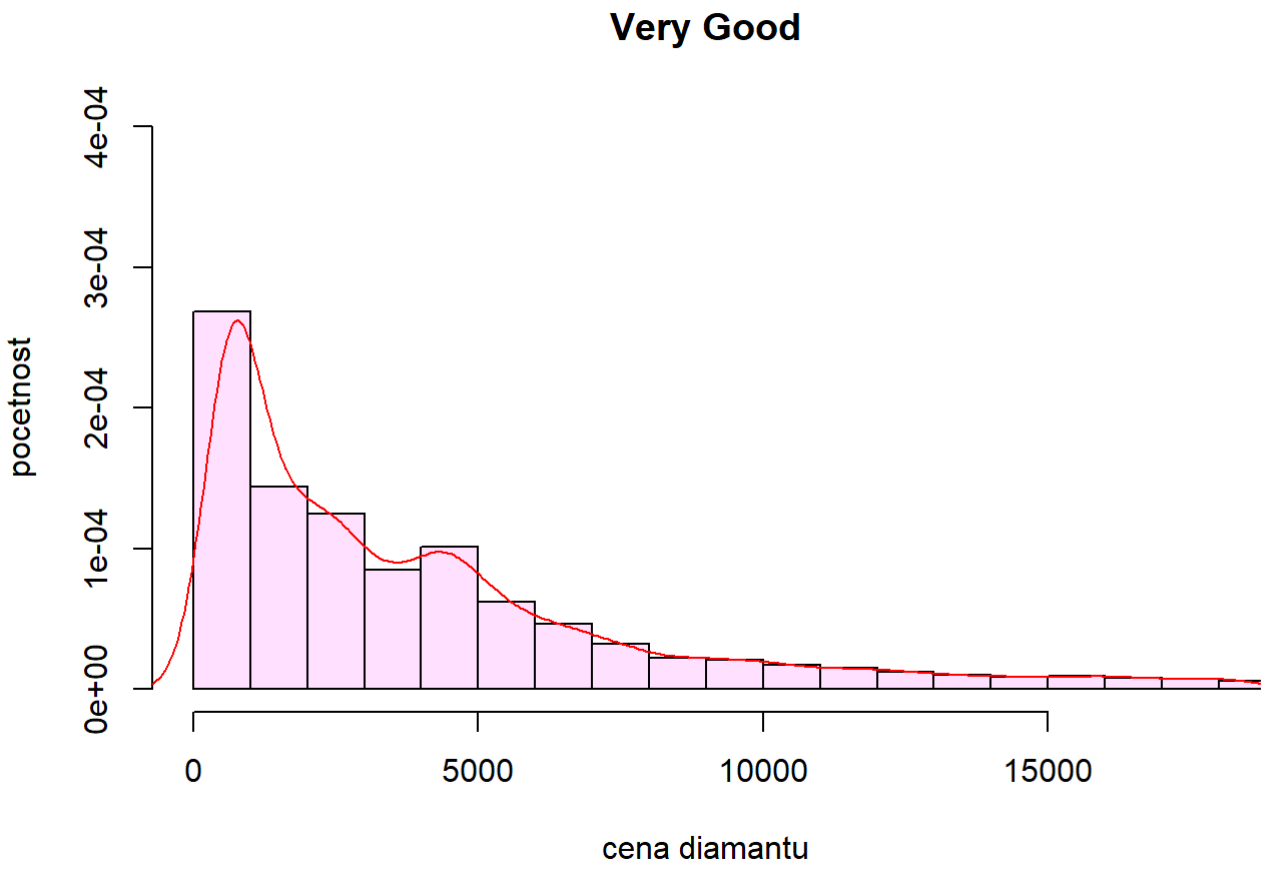

```
hist(fair_all, breaks = "Sturges", col = "thistle1", main = "Fair", xlab = "cena diam
antu", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histogram re
lativnych pocetnosti
lines(density(fair_all), col = "red")
```



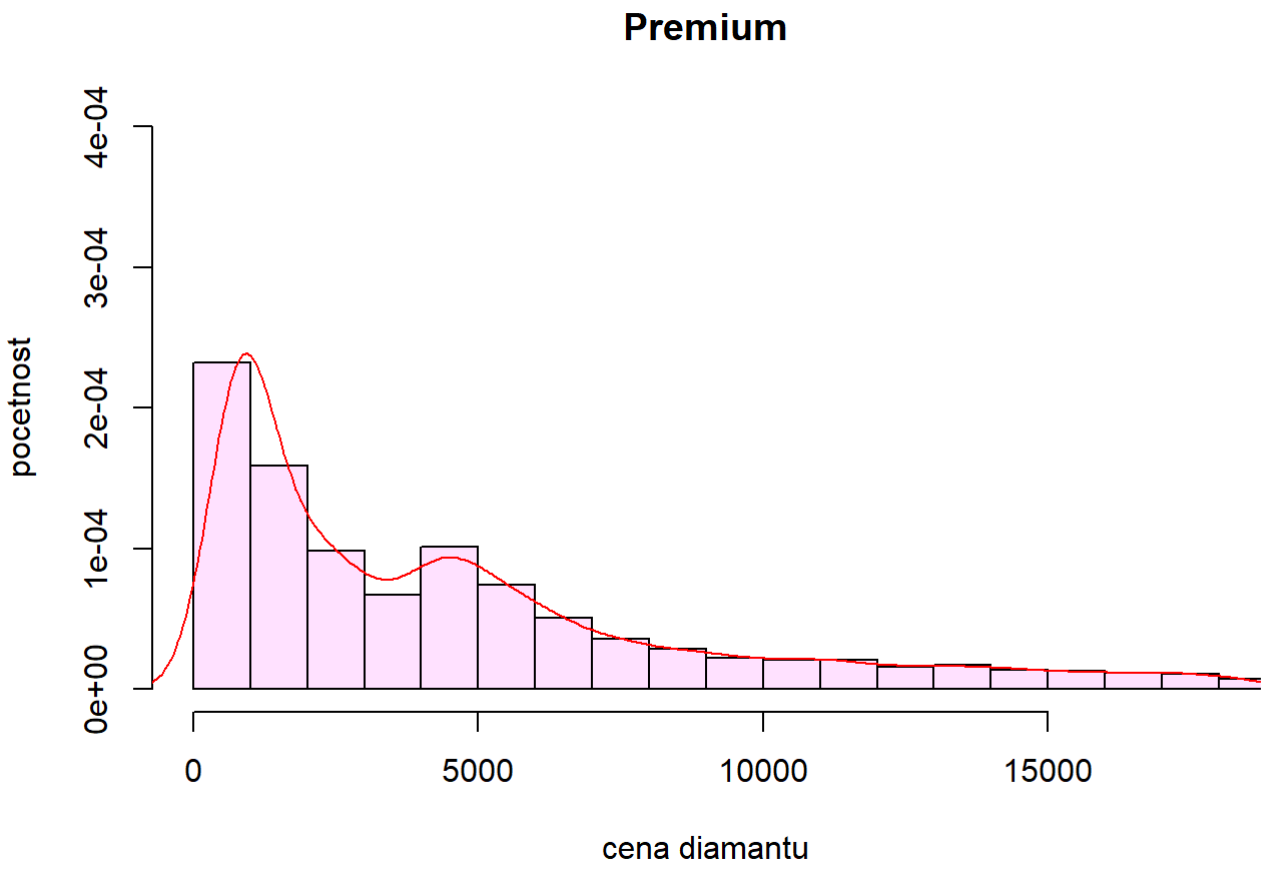

```
hist(good_all, breaks = "Sturges", col = "thistle1", main = "Good", xlab = "cena diam
antu", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histogram re
lativnych pocetnosti
lines(density(good_all), col = "red")
```




```
hist(very_good_all, breaks = "Sturges", col = "thistle1", main = "Very Good", xlab =
"cena diamantu", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #hi
stogram relativnych pocetnosti
lines(density(very_good_all), col = "red")
```

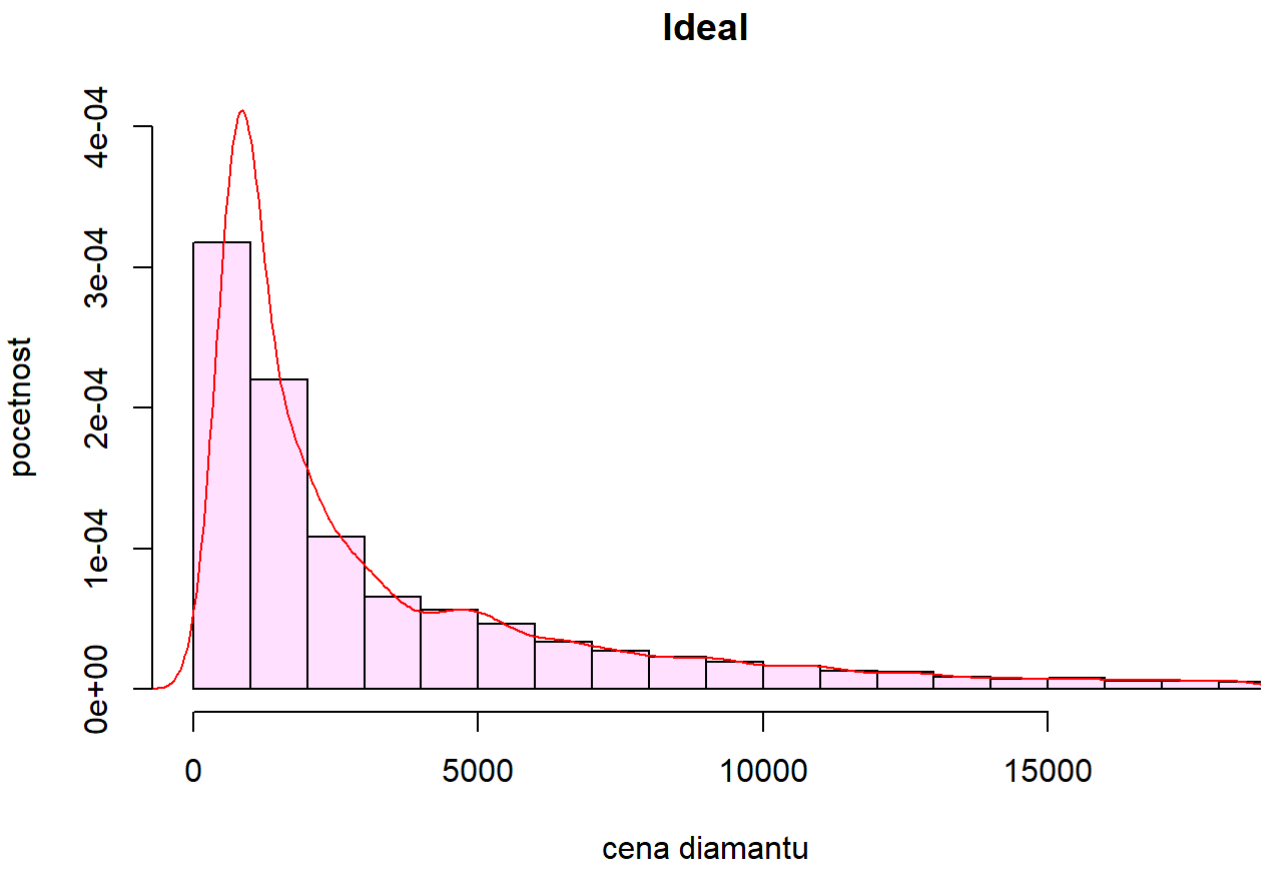



```
hist(premium_all, breaks = "Sturges", col = "thistle1", main = "Premium", xlab = "cen
a diamantu", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histog
ram relativnych pocetnosti
lines(density(premium_all), col = "red")
```



Vidime, ze 100 je pomerne mala vzorka na to, aby dostatocne dobre rezprezntovala rozdelenia pri roznych hodnotach atributu rezu diamantov

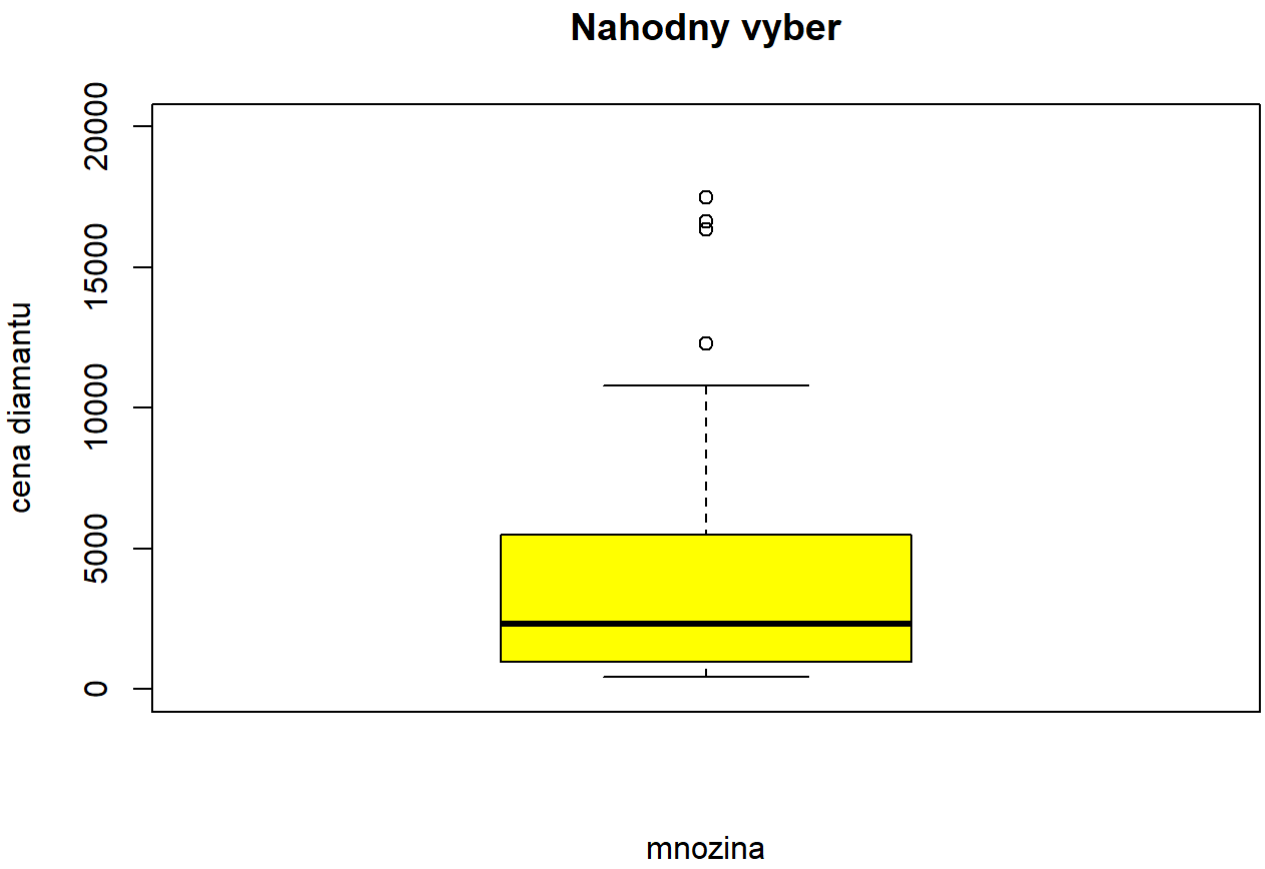
```
hist(ideal_all, breaks = "Sturges", col = "thistle1", main = "Ideal", xlab = "cena di
amantu", ylab = "pocetnost", freq = F, ylim=c(0,0.0004), xlim=c(0,18000)) #histogram
relativnych pocetnosti
lines(density(ideal_all), col = "red")
```



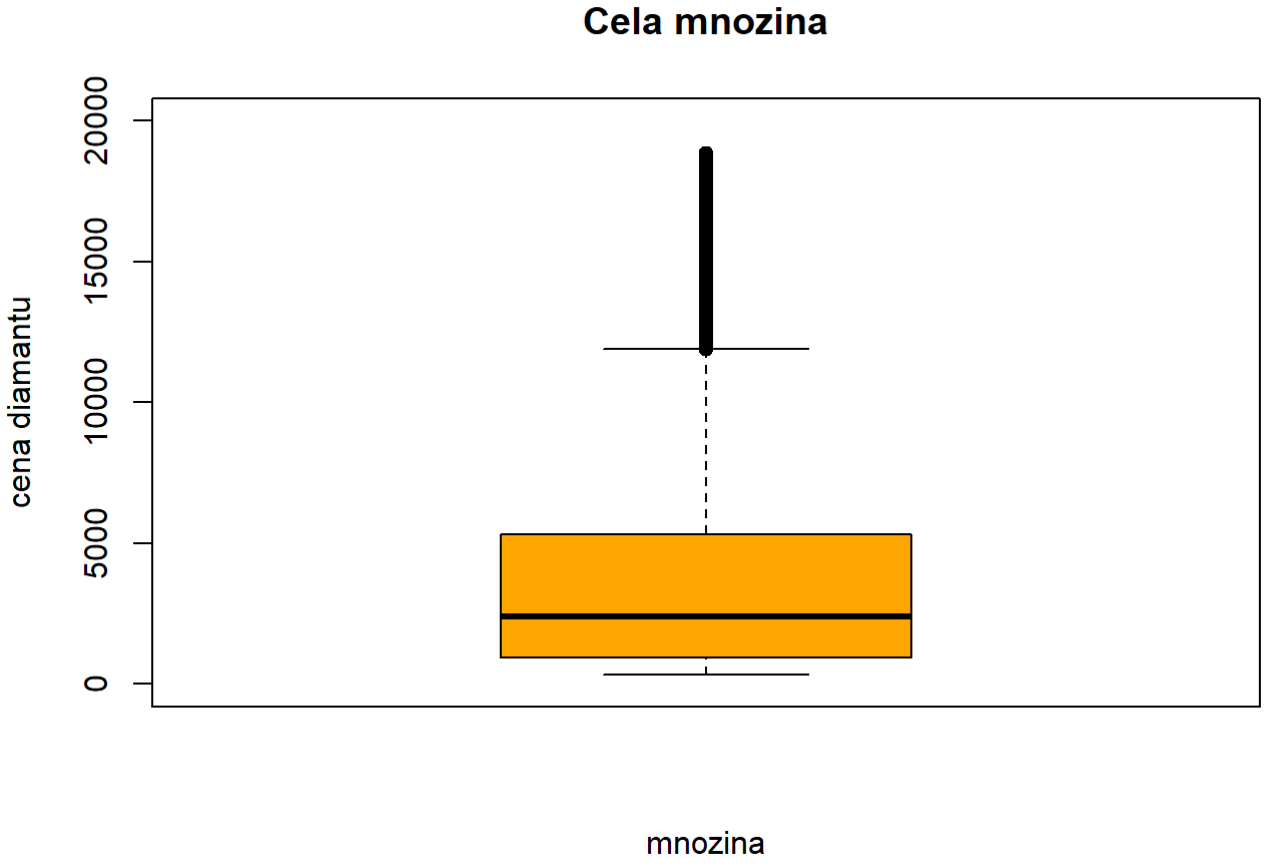

```
par(mfrow=c(1,2))
```

Boxplot pre ceny diamantov nahodneho vyberu a celej mnoziny

```
boxplot(nahodny_vyber$price, col = "yellow", xlab = "mnozina", ylab = "cena diamant
u", main="Nahodny vyber", ylim=c(0,20000))
```



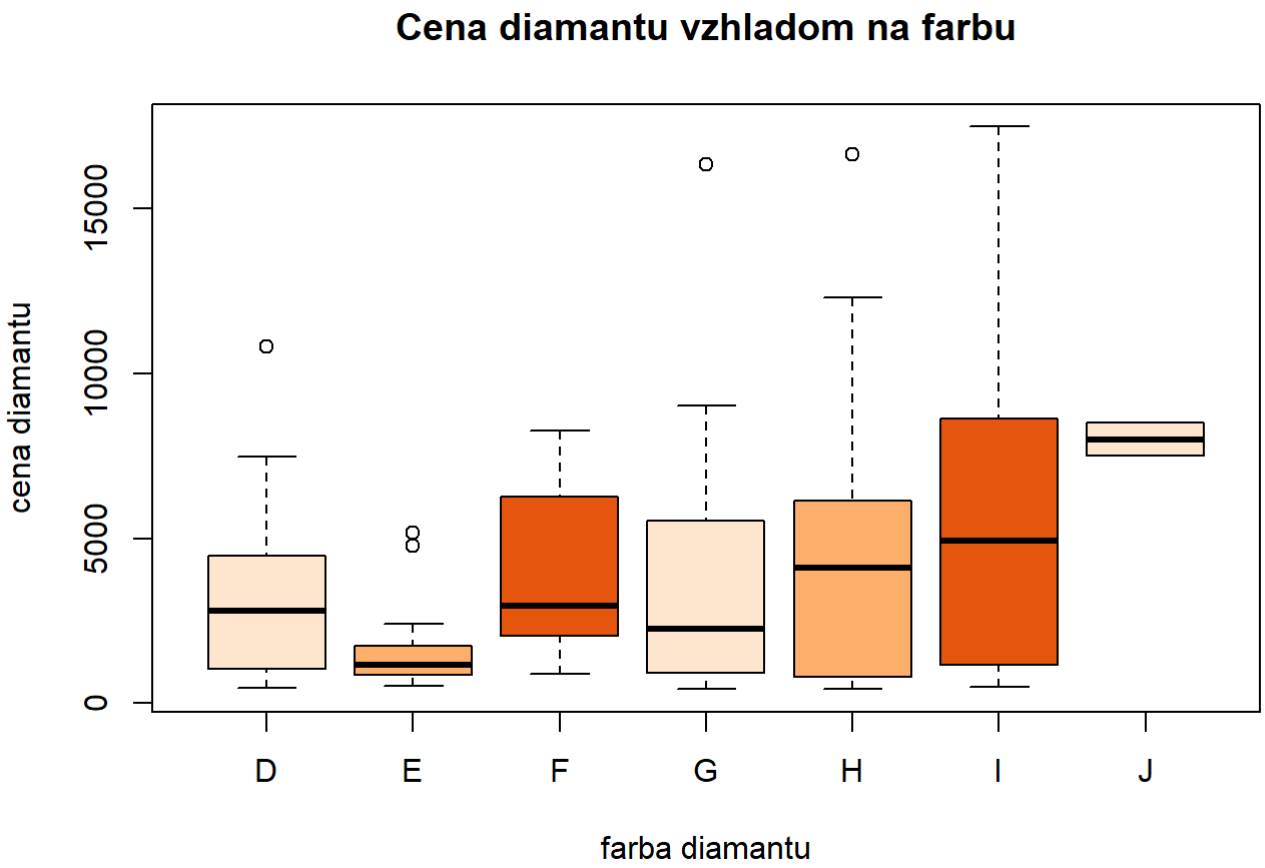

```
boxplot(diamanty$price, col = "orange", xlab = "mnozina", ylab = "cena diamantu", mai
n="Cela mnozina", ylim=c(0,20000))
```



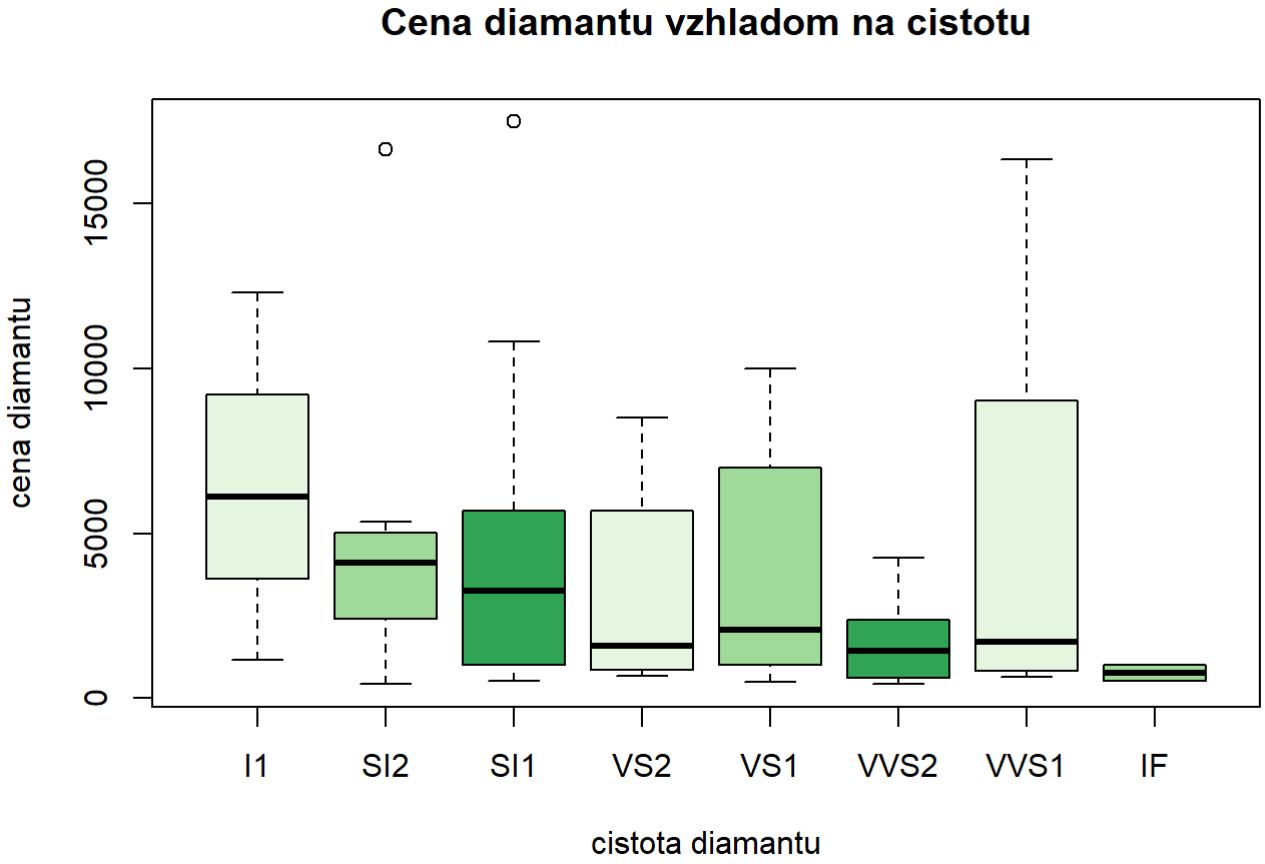
```
par(mfrow=c(2,3))
```

Boxploty a violoty pre ceny diamantov nahodneho vyberu podla farby, cistoty a rezu diamantov

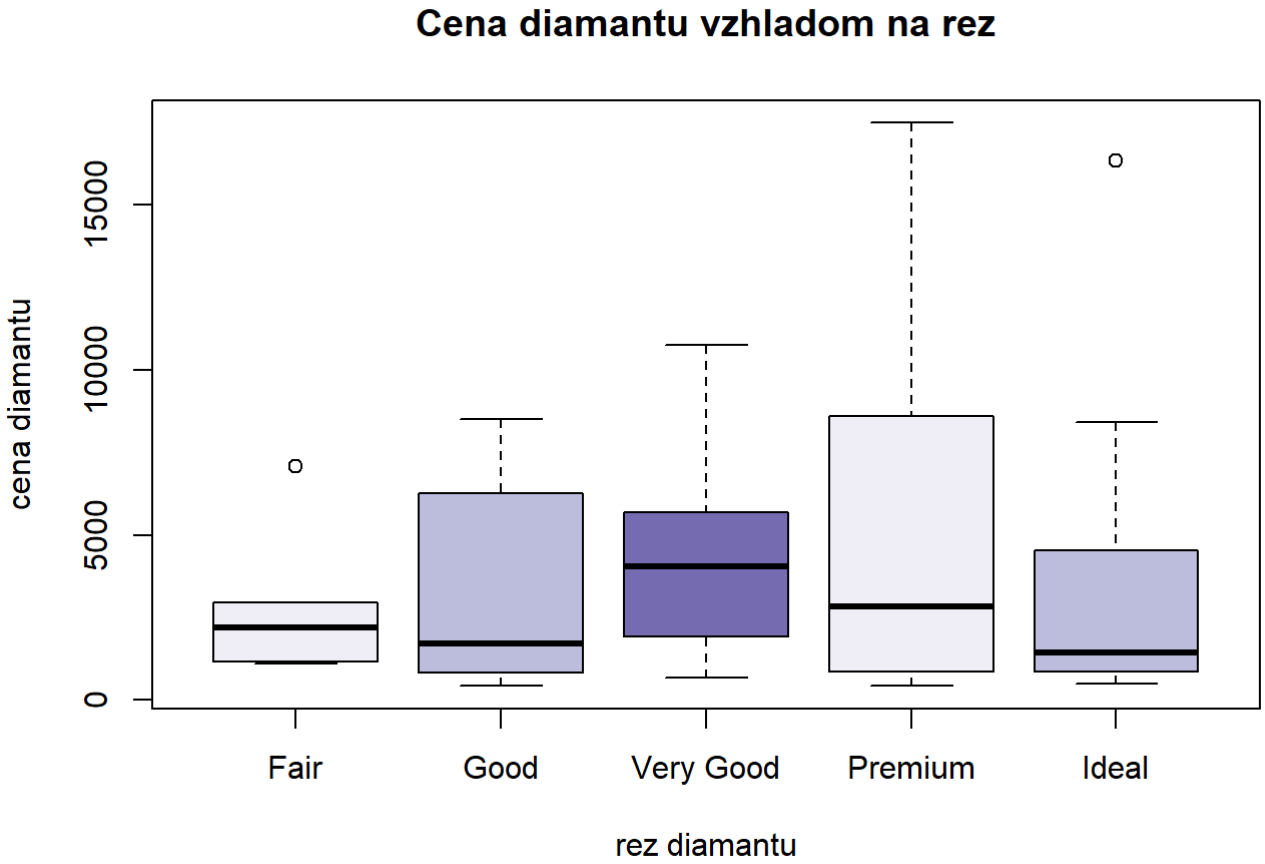
```
boxplot(nahodny_vyber$price~nahodny_vyber$color, col = brewer.pal(3,"Oranges"), xlab
= "farba diamantu", ylab = "cena diamantu", main="Cena diamantu vzhladom na farbu")
```



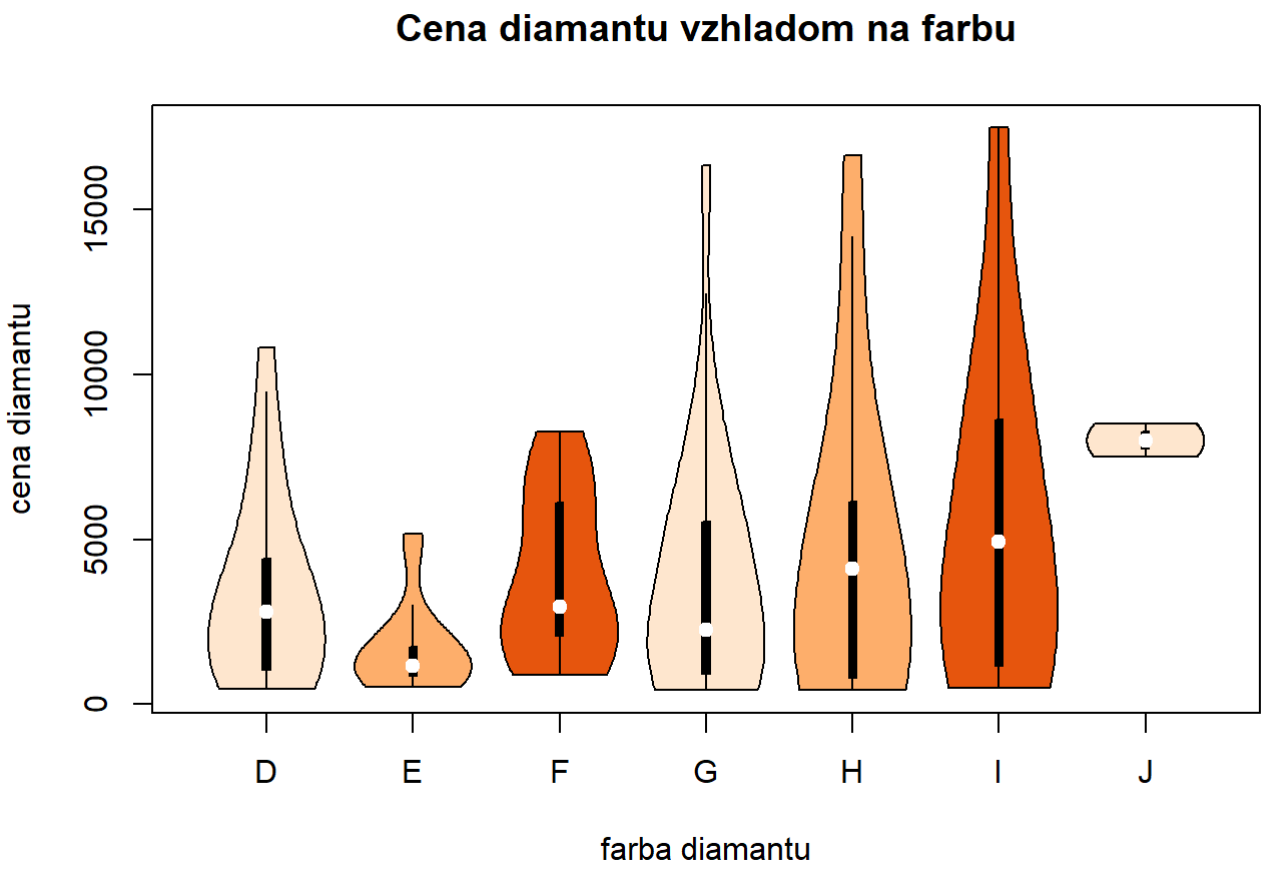

```
boxplot(nahodny_vyber$price~nahodny_vyber$clarity, col = brewer.pal(3,"Greens"), xlab
= "cistota diamantu", ylab = "cena diamantu", main="Cena diamantu vzhladom na cistot
u")
```



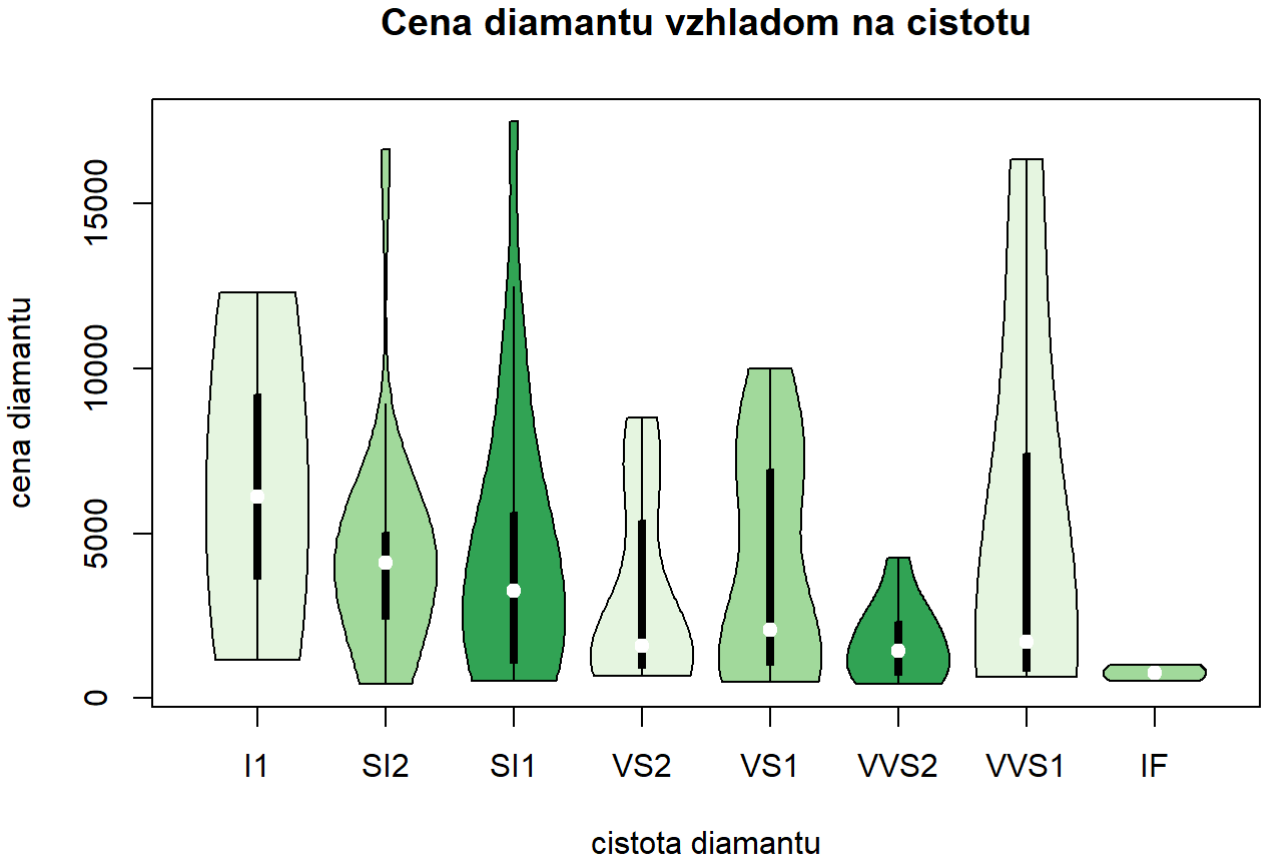

```
boxplot(nahodny_vyber$price~nahodny_vyber$cut, col = brewer.pal(3,"Purples"), xlab =
"rez diamantu", ylab = "cena diamantu", main="Cena diamantu vzhladom na rez")
```



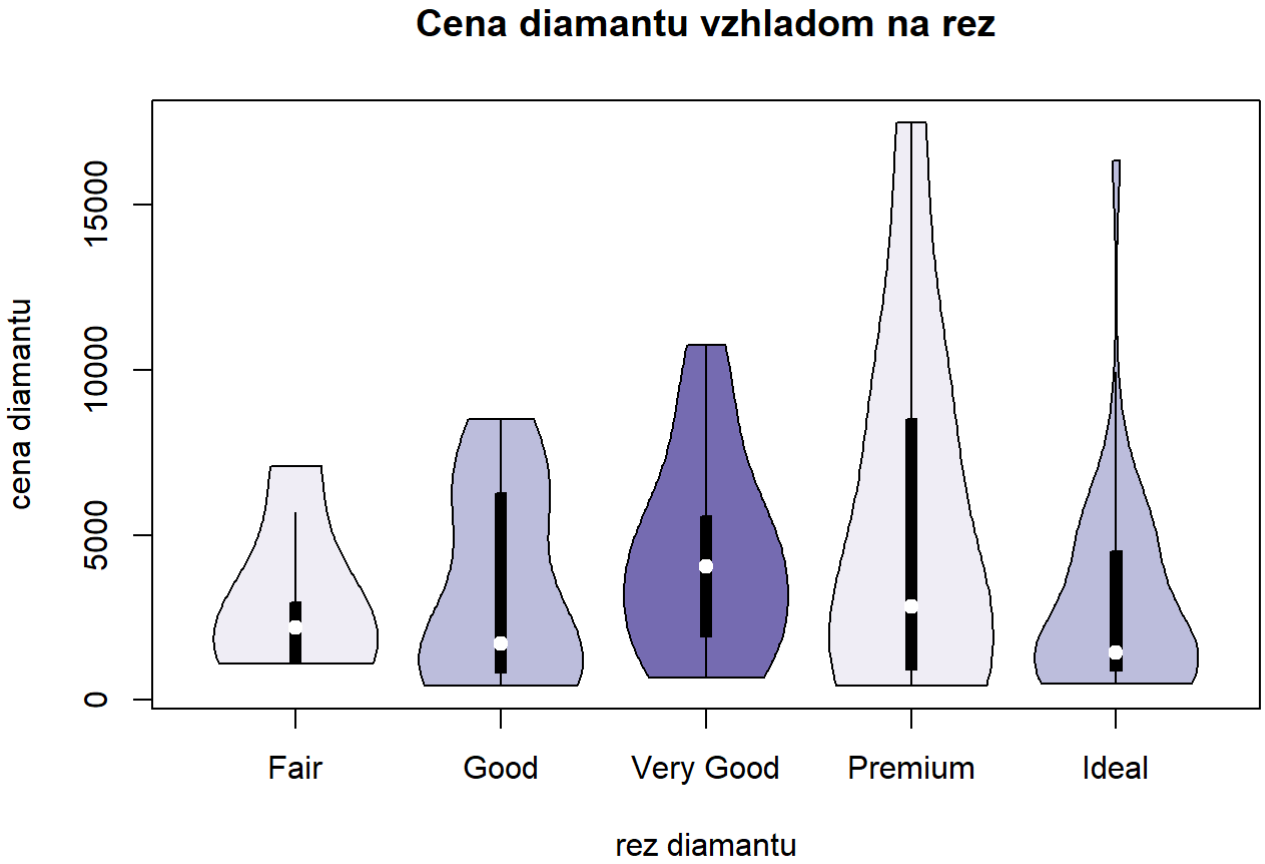
```
vioplot(nahodny_vyber$price~nahodny_vyber$color, col = brewer.pal(3,"Oranges"), xlab
= "farba diamantu", ylab = "cena diamantu", main="Cena diamantu vzhladom na farbu")
```




```
vioplot(nahodny_vyber$price~nahodny_vyber$clarity, col = brewer.pal(3,"Greens"), xlab
= "cistota diamantu", ylab = "cena diamantu", main="Cena diamantu vzhladom na cistot
u")
```




```
vioplot(nahodny_vyber$price~nahodny_vyber$cut, col = brewer.pal(3,"Purples"), xlab =
"rez diamantu", ylab = "cena diamantu", main="Cena diamantu vzhladom na rez")
```



Príklad 1 (podiel konzerv)

Pri kontrole dátumu spotreby určitého druhu mäsovej konzervy v skladoch bolo náhodne vybraných 320 z 20 000 konzerv a zistené, že 59 z nich má expirovanú záručnú dobu. Stanovte so spoľahlivosťou 95% intervalový odhad podielu expirovaných mäsových konzerv.

Príklad 2 (obsah hnojiva)

Urobilo sa šesť paralelných stanovení obsahu P_2O_5 vo vzorke hnojiva s nasledujúcimi výsledkami: 16.5 , 15.9, 16.6, 15.8, 16.4, 16.0, 15. Predpokladajme, že ide o výber z normálneho rozdelenia N(μ,σ^2). Vypočítajte

a) obojstranný 95% a 90% interval spoľahlivosti pre strednú hodnotu μ obsahu P_2O_5
b) dolný (ľavostranný) 99%-ný interval spoľahlivosti pre smerodajnú odchýlku σ.

Na základe vhodných intervalov spoľahlivosti odpovedzte na otázky:
c) Dá sa spoľahlivo (s 95%-nou spoľahlivosťou) tvrdiť, že stredná hodnota obsahu P_2O_5 v hnojive nie je rovná 16.4 (inak povedané: líši sa stredná hodnota štatisticky významne od 16.4)?

is_vypracovanie.R

Janka

2023-04-14

Podiel expirovanych konzerv

```
binom.test(59, n=320, conf.level=0.95)$conf.int
```

```
## [1] 0.1434193 0.2312853
## attr(,"conf.level")
## [1] 0.95
```

Obsah hnojiva a)

```
x <- c(16.5, 15.9, 16.6, 15.8, 16.4, 16.0, 15)
cat("a) 90% a 95% IS pre strednú hodnotu: ")
```

```
## a) 90% a 95% IS pre strednú hodnotu:
```

```
t.test(x, conf.level = 0.9)$conf.int
```

```
## [1] 15.62470 16.43244
## attr(,"conf.level")
## [1] 0.9
```

```
ISmu <- t.test(x)$conf.int;ISmu
```

```
## [1] 15.52001 16.53714
## attr(,"conf.level")
## [1] 0.95
```

```
cat("b) 99% L-IS pre smerodajnú odchýlku: ")
```

```
## b) 99% L-IS pre smerodajnú odchýlku:
```

```
is <- EnvStats::varTest(x, alternative = "greater", conf.level = 0.99)$conf.int
c(sqrt(is[1]),sqrt(is[2]))
```

```
##       LCL       UCL
## 0.3285069       Inf
```

```
cat("c) Tvrdenie o strednej hodnote")
```

```
## c) Tvrdenie o strednej hodnote
```

```
is.in <- dplyr::between(16.4, ISmu[1], ISmu[2])
is.in
```

```
## [1] TRUE
```

```
cat("Nelíši sa.")
```

```
## Nelíši sa.
```

Testy hypotéz o parametroch normálneho rozdelenia, dvojvýberove testy

(predpokladáme, že dáta príkladov sú normálne rozdelené, ak nie je uvedené inak, tak $\alpha=0.05$)

Na určenie poistnej zásoby tovaru, treba poznať, ako dlho trvá vybavenie objednávky u dodávateľa. V n=20 sa zistilo, že vybavenie objednávky trvá nasledujúce počty dní.

18	12	18	13	13	16	11	17	20	13
14	15	16	14	15	15	17	13	14	16

Nech $\sigma=1.5$

- Testujte hypotézu, že čakacia doba je dva týždne.
- Testujte hypotézu, že čakacia doba je väčšia ako dva týždne, hladina významnosti je 0.05.

Nech σ nie je známe.

- Testujte hypotézu, že čakacia doba je dva týždne.
- Testujte hypotézu, že čakacia doba je väčšia ako dva týždne, hladina významnosti je 0.05.
- Testujte hypotézu, že disperzia je rovná 4.

Spoločnosť CTC nakupuje baterky do elektronických. hračiek. Dodávateľ garantuje životnosť bateriek minimálne 19 hodín. Kontrolór náhodne vyberie 10 bateriek a skúša životnosť. Výsledky sú nasledovné

20.2	19.6	18.6	19.4	17	18.5	18	18.4	19	18

Otestujte, či je tvrdenie dodávateľa o životnosti bateriek pravdivé, hladina významnosti 0.05. Vyberte vhodný test a správnu alternatívu.

Výrobca motoriek testoval spotrebu na 100 km u jedného typu motoriek. Boli namerané tieto hodnoty

7.7	6.8	5	9.8	7.4	8.7	6.3	8	8.6	6.4

- Testujte hypotézu, že priemerná spotreba na 100 km je 7 litrov , hladina významnosti je 0.1.
- Testujte hypotézu, že priemerná spotreba na 100 km nie je viac ako 8 litrov, hladina významnosti je 0.05

Pizzeria ABC , ktorá robí rozvoz pizze, má na letáku uvedené, že pizzu dovezú do 30 min. Overte toto tvrdenie, ak máte k dispozícii tieto záznamy o dodávkach:

27	28	35	36	29	24	30	26	28	32
24	32	31	29	28	29	35	34	30	31

Testujte hypotézu, že $\sigma=4$.

Dvojvýberové testy

Zamestnanci istej firmy absolvovali kurz optimalizácie spracovania bežnej agendy. V tabuľke sú uvedené časy v minútach, ktoré jeden zamestnanec venuje bežnej agende denne (pred a po kurze). Na hladine významnosti $\alpha=0.05$ testujte hypotézu, že kurz nemá vplyv na dĺžku času spracovania agendy

pred	34	35	27	29	29	30	31	30	28	32
po	28	30	26	24	26	26	26	31	17	26

Testujeme, špeciálnu diétu na zníženie hmotnosti. Testu sa zúčastnilo 12 dobrovoľníkov. Ich hmotnosti v kg pred a po diétnej kúre sú dané tabuľkou. Na hladine významnosti $\alpha=0.05$ testujte hypotézu, že daná diéta zníži hmotnosť v priemere o 5 kg.

pred	85	75	65	150	80	110	65	88	74	67	110	90
po	79	72	61	140	75	100	65	85	75	65	102	91

Aby sme zistili, aký vplyv má vonkajšia teplota na systematickú chybu uhlomerného prístroja, merali sme horizontálny uhol zvoleného objektu ráno pri teplote 10^0C a napoludnie pri teplote 26^0C. Výsledku sú v tabuľke. Možno tvrdiť, že teplota má vplyv na systematickú odchýlku? $\alpha=0.05$

ráno	38,2	36,4	37,7	36,1	37,9	37,8		
obed	39,5	38,7	37,8	38,6	39,2	39,1	38,9	39,2

Študenti medicíny dostali na laboratórnych cvičeniach za úlohu naočkovať potkany látkou nazvanou aloxan. Aloxan je látka, ktorá deštruuje bunky pankreasu, ktoré sú zodpovedné za tvorbu inzulínu (odbúrava cukor). Bolo sledovaných 10 potkanov. Aloxan bol naočkovaný 4 z nich, pričom zvyšným 6 bolo podávané len placebo. Po čase bol všetkým potkanom podaný roztok s cukrom. Po určitej dobe bola v krvi potkanov zmeraná hladina cukru (tabuľka). Má aloxan vplyv na odburávanie cukru v krvi? $\alpha=0.05$

aloxín	23.5	28.8	28.3	21.8		
placebo	22	18.3	15.2	15.6	15.6	19.1

13 polí rovnakej kvality bolo rozdelených na dve skupiny. Na 8 z nich sa skúšal nový spôsob hnojenia a 5 bolo hnojených bežným spôsobom. Výnosy pšenice (t/ha) z polí hnojených novým spôsobom sú označené x_i a výnosy pšenice z polí hnojených bežným spôsobom sú označené y_i. Zistite, či spôsob hnojenia má vplyv na výnosy pšenice (testujte na hladine významnosti $\alpha=0.05$)

nový spôsob hnojenia	5.7	5.5	4.3	5.9	5.2	5.6	5.8	5.1
starý spôsob hnojenia	5	4.5	4.2	5.4	4.4			

Počas pracovného dňa boli zaznamenávané dĺžky dodávok objednaného jedla(v min) dvoch zariadení rýchleho občerstvenia AA a BB. Overte hypotézu, že dĺžky dodávok jedla sú rovnaké (testujte na hladine významnosti $\alpha=0,05$)

AA	10	12	15	25	18	20	15	25
BB	15	15	18	10	16	12	15	

testy_vypracovanie.R

Jana Kalická

2023-04-10

Testy, vypracovanie

```
tovar <- c(18,12,13,13,16,11,17,20,13,14,15,16,14,15,17,13,14,16)
```

sigma zname, prva uloha

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.2.3
```

```
ZTest(tovar,mu=14,sd_pop = 1.5)
```

```
##
##  One Sample z-test
##
## data:  tovar
## z = 2.357, Std. Dev. Population = 1.5, p-value = 0.01842
## alternative hypothesis: true mean is not equal to 14
## 95 percent confidence interval:
##  14.14038 15.52629
## sample estimates:
## mean of x
##  14.83333
```

```
ZTest(tovar,mu=14,sd_pop = 1.5)$p.value
```

```
## [1] 0.01842213
```

P hodnota <0.05, zamietam H0, ze cakacia doba je 2 tyzdne

Druha uloha Vacsia ako 14 dni, alternativa mensia, less

```
ZTest(tovar,mu=14,sd_pop = 1.5,alternative = "less")
```

```
##
##  One Sample z-test
##
## data:  tovar
## z = 2.357, Std. Dev. Population = 1.5, p-value = 0.9908
## alternative hypothesis: true mean is less than 14
## 95 percent confidence interval:
##      -Inf 15.41488
## sample estimates:
## mean of x
##  14.83333
```

P hodnota >0.05, nezamietam H0, ze cakacia doba je viac ako 2 tyzdne.

Testujeme to iste, ale sigma nepozname, prva uloha

```
t.test(tovar,mu=14)
```

```
##
##  One Sample t-test
##
## data:  tovar
## t = 1.5496, df = 17, p-value = 0.1397
## alternative hypothesis: true mean is not equal to 14
## 95 percent confidence interval:
##  13.69870 15.96797
## sample estimates:
## mean of x
##  14.83333
```

```
t.test(tovar,mu=14)$p.value
```

```
## [1] 0.1396628
```

v tomto pripade nezamietam H0, cakacia doba je 2 tyzdne.

Druha uloha:Vacsia ako 14 dni, teda alternativa mensia, less

```
t.test(tovar,mu=14,alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  tovar
## t = 1.5496, df = 17, p-value = 0.9302
## alternative hypothesis: true mean is less than 14
## 95 percent confidence interval:
##      -Inf 15.76887
## sample estimates:
## mean of x
##  14.83333
```

v tomto pripade nezamietam H0, cakacia doba je aspon 2 tyzdne ale nie menej.

Test pre samotne sigma

```
library(EnvStats)
```

```
## Warning: package 'EnvStats' was built under R version 4.2.3
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```
varTest(tovar,sigma.squared = 4)
```

```
## $statistic
## Chi-Squared
##      22.125
##
## $parameters
## df
## 17
##
## $p.value
## [1] 0.359919
##
## $estimate
## variance
## 5.205882
##
## $null.value
## variance
##        4
##
## $alternative
## [1] "two.sided"
##
## $method
## [1] "Chi-Squared Test on Variance"
##
## $data.name
## [1] "tovar"
##
## $conf.int
##       LCL       UCL
##  2.931336 11.699870
## attr(,"conf.level")
## [1] 0.95
##
## attr(,"class")
## [1] "htestEnvStats"
```

P hodnota>0.05, nezamietam H0, disperzia je rovna 4.

```
############################################################
```

CTC baterky, alternativa: vydrzia menej, nezamietam H0.

```
baterky <- c(20.2,19.6,18.6,19.4,17,18.5,18,18.4,19,18)
t.test(baterky,mu=19,alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  baterky
## t = -1.1326, df = 9, p-value = 0.1433
## alternative hypothesis: true mean is less than 19
## 95 percent confidence interval:
##      -Inf 19.20413
## sample estimates:
## mean of x
##     18.67
```

```
############################################################
```

Motorky, prva uloha alfa=0.1

```
motorky <- c(7.7,6.8,5,9.8,7.4,8.7,6.3,8,8.6,6.4)
t.test(motorky,mu=7)
```

```
##
##  One Sample t-test
##
## data:  motorky
## t = 1.0622, df = 9, p-value = 0.3158
## alternative hypothesis: true mean is not equal to 7
## 95 percent confidence interval:
##  6.46904 8.47096
## sample estimates:
## mean of x
##      7.47
```

P hodnota >0.1 , nezamietam H0.

Druha uloha, viac ako 8, alternativa je menej, less.

```
t.test(motorky,mu=8,alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  motorky
## t = -1.1978, df = 9, p-value = 0.1308
## alternative hypothesis: true mean is less than 8
## 95 percent confidence interval:
##      -Inf 8.281117
## sample estimates:
## mean of x
##      7.47
```

Aj toto tvrdenie nezamietam.

```
############################################################
```

Pizza

Do 30 min, alternativa viac ako 30, greater.

```
pizza <- c(27,28,35,36,29,24,30,26,28,32,24,32,31,29,29,35,34,30,31)
t.test(pizza,mu=30,alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  pizza
## t = 0, df = 18, p-value = 0.5
## alternative hypothesis: true mean is greater than 30
## 95 percent confidence interval:
##  28.6092     Inf
## sample estimates:
## mean of x
##        30
```

Nezamietam, dovezu do 30 min.

Test pre disperziu, nezamietam H0.

```
VarTest(pizza,sigma.squared = 4^2)
```

```
##
##  One Sample Chi-Square test on variance
##
## data:  pizza
## X-squared = 13.75, df = 18, p-value = 0.5095
## alternative hypothesis: true variance is not equal to 16
## 95 percent confidence interval:
##   6.978283 26.729047
## sample estimates:
## variance of x
##      12.22222
```

```
############################################################
```

Parove testy: Zamestnanci (vzor, nema vplyv testujeme tozdiel je 0) a dieta(samostatna praca, rozdiel testujeme na hodnotu 5=pred-po).

```
pred <- c(34,35,27,29,29,30,31,30,28,32)
po <- c(28,30,26,24,26,26,26,31,17,26)
t.test(pred,po,mu=0,paired = T)
```

```
##
##  Paired t-test
##
## data:  pred and po
## t = 4.4388, df = 9, p-value = 0.001626
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  2.206639 6.793361
## sample estimates:
## mean difference
##             4.5
```

Zamietam H0, kurz ma vplyv. Lepsie by bolo testovat, ze im to teraz trva menej, teda alternativa greater.

```
############################################################
```

Posledne 4 ulohy su neparove testy, uvedieme len prvy. Najprv otestovat rovnost disperzii a vybrat spravny test.

```
rano <- c(38.2,36.4,37.7,36.1,37.9,37.8)
obed <- c(39.5,38.7,37.8,38.6,39.2,39.1,38.9,39.2)
var.test(rano,obed)
```

```
##
##  F test to compare two variances
##
## data:  rano and obed
## F = 2.789, num df = 5, denom df = 7, p-value = 0.2138
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.5277027 19.1134605
## sample estimates:
## ratio of variances
##           2.789034
```

Nezamietam hypotezu o rovnosti disperzii, var.equal=T

```
t.test(rano,obed,paired = F,var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  rano and obed
## t = -4.0864, df = 12, p-value = 0.001509
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.3381113 -0.7118887
## sample estimates:
## mean of x mean of y
##    37.350    38.875
```

Zamietam H0, teplota ma vplyv na merania.

Neparametrické testy

- Pokiaľ nie je dané inak, $\alpha = 0.05$.
- Nulová a alternatívna hypotéza musia byť súčasťou riešenia.
- Výsledky slovne popíšte- interpretujte.
- Všetko čo sa dá a má zmysel vizualizujte a popíšte, interpretujte, čo výsledok testu znamená v praxi.

Príklad 1

(1b) Sledoval sa účinok troch rôznych typov liečby depresie. 17 náhodne vybraných pacientov bolo rozdelených do troch skupín, pričom v každej skupine bol na liečbu depresie použitý iný typ liečby. Prvá skupina pacientov sa liečila kognitívne behaviorálnou terapiou, druhá skupina psychoanalýzou a tretia skupina terapiou zameranou na riešenie. Účinnosť liečby sa posudzovala na základe doby trvania príznakov od jej začiatku. Pacienti s diagnostikovanou depresiou podstúpili liečbu raz za týždeň. Doba trvania príznakov (v dňoch) sú dané:

skupina 1	100	88	75	115	45	
skupina 2	56	43	24	96	59	80
skupina 3	106	113	125	63+k	98	100

Sú všetky tri typy liečby rovnako účinné? Ak nie, kde sú štatisticky významné rozdiely?

Príklad 2

(1b) Posudzovala sa účinnosť nového lieku určeného na redukciu opakovaného správania u detí postihnutých autizmom. Osem detí bolo pozorovaných psychológom pred užitím lieku a znovu po jeho užití po dobu jedného týždňa. Sleduje sa čas, ktorý dieťa strávilo opakovaným správaním. Výsledky sa zaznamenávajú na škále 0-100, pričom skóre vyjadruje koľko percent času sa zaoberali opakovaným správaním. Napr. 0 znamená, že počas celej doby pozorovania dieťa nevykonávalo opakované správanie, zatiaľ čo 100 znamená, že dieťa bolo stále v opakovanom správaní. Na hladine významnosti 0.05 posúďte, či má nový liek vplyv na opakované správanie u detí s autizmom.

pred	85	70	40+k	65	80	75	55	20
po	75	50	50	40	20	65	40	30

Príklad 3

(2b) Obchodník si náhodne vybral 10 nákupov platených kreditnou kartou a 8 nákupov platených hotovosťou.

1. Zistite, či veľkosti nákupov platených kreditnou kartou a hotovosťou sú odlišné.

2. Overte, či výber je naozaj náhodný pre obidva spôsoby platby testujte Wald Wolfovitz testom a Turning point testom

kreditka	34.32	8.8	104.72	59.84	66.88	110.88	46.64	89.8	69.52	90.55
hotovosť	36.96,	67.76	40.48	64.24	68.64	29.04	32.56	30.4		

Príklad 4

(1b) Automat plní plechovky náterovou látkou, pričom v každej plechovke majú byť 2 kg látky. Z produkcie sme náhodne vybrali 8 plechoviek a ich obsah bol prevážený. Zistili sme tieto odchýlky v gramoch

-50	10	-10	-80	70	-20	20	-60

Na hladine významnosti α=0.1 overte , či plniaci automat je správne nastavený

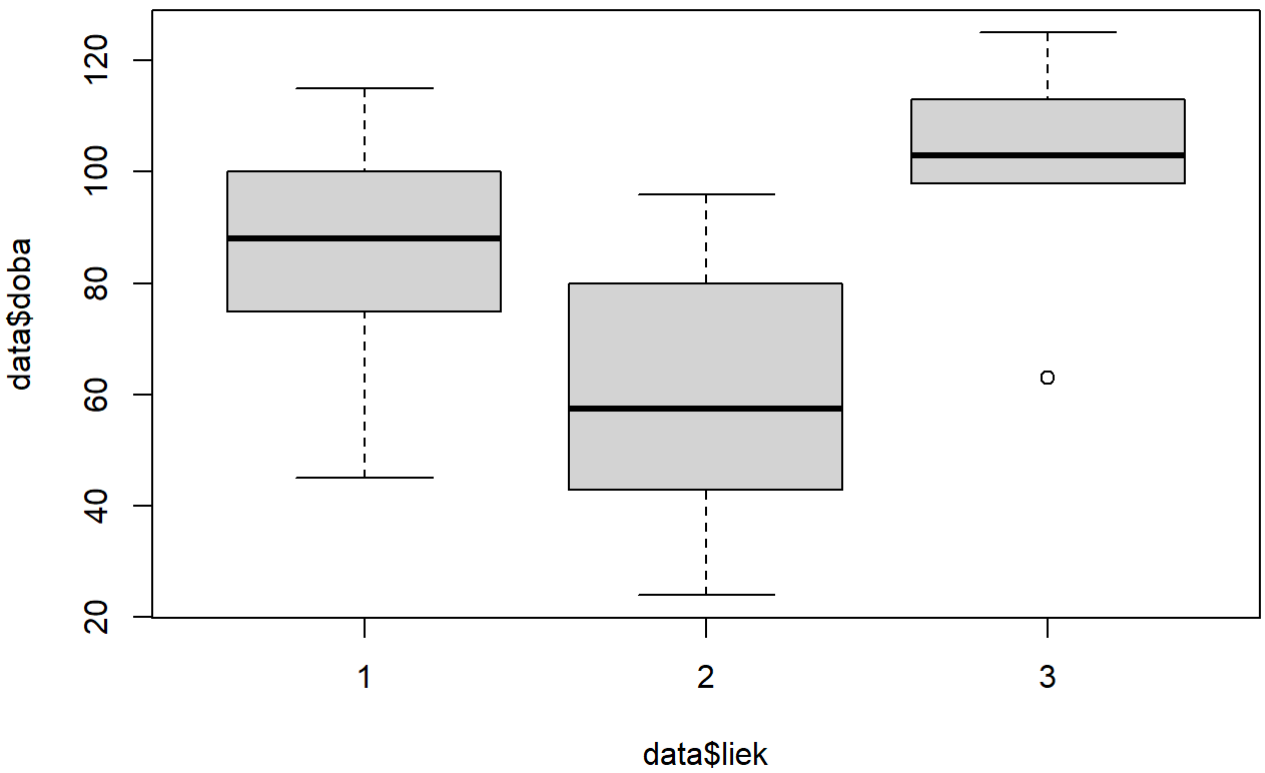
neparametricke_vypracovanie.R

Jana Kalická

2023-04-10

Lieciva, Kruskal Wallis, nezabudnut faktorizovat a dobre zadat data.

```
s1 <- c(100,88,75,115,45)
s2 <- c(56,43,24,96,59,80)
s3 <- c(106,113,125,63,98,100)
druh <- rep(c(1,2,3),times=c(length(s1),length(s2),length(s3)))
data <- data.frame("doba"=c(s1,s2,s3),"liek"=druh)
boxplot((data$doba~data$liek))
```




```
data$liek <- factor(data$liek)
kruskal.test(data$doba,data$liek)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$doba and data$liek
## Kruskal-Wallis chi-squared = 6.8434, df = 2, p-value = 0.03266
```

Zamietam hypotezu o rovnakej ucinnosti troch druhov liecenia Zistime, ktore triedy sa lisia.

```
library(dunn.test)
dunn.test(data$doba,data$liek,altp=T,list=T)
```

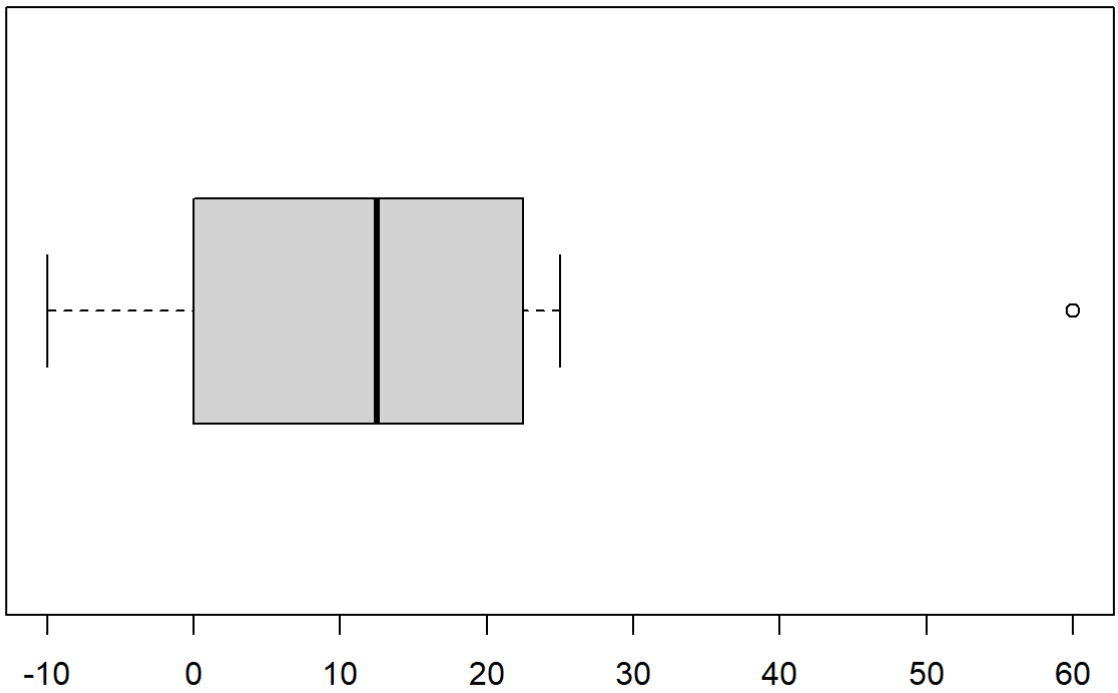
```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 6.8434, df = 2, p-value = 0.03
##
##
##                             Comparison of x by group
##                                   (No adjustment)
## Col Mean-|
## Row Mean |          1          2
## ---------+----------------------
##        2 |   1.472559
##          |     0.1409
##          |
##        3 |  -1.008976  -2.602657
##          |     0.3130    0.0093*
##
##
## List of pairwise comparisons: Z statistic (p-value)
## ---------------------------
## 1 - 2 :  1.472559 (0.1409)
## 1 - 3 : -1.008976 (0.3130)
## 2 - 3 : -2.602657 (0.0093)*
##
## alpha = 0.05
## Reject Ho if p <= alpha
```

Lisia sa triedy 2 a 3.

```
############################################################
```

Autizmus, jednovyberovy Wilcoxonov (parovy , testujeme rozdiely,), overit symetriu rozdielov, ak nie su symetricke, tak znamienkovy test, parovy.

```
pred <- c(85,70,40,65,80,75,55,20)
po <- c(75,50,50,40,20,65,40,30)
rozdiel <- pred-po
boxplot(rozdiel,horizontal = T)#zdaju s abyt asymetricke, sikmost a test
```




```
library(moments)
skewness(rozdiel)
```

```
## [1] 0.8479351
```

```
library(lawstat)
```

```
## Warning: package 'lawstat' was built under R version 4.2.3
```

```
symmetry.test(rozdiel)#su symetricke, Wilcoxonov test
```

```
##
##  m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)
##
## data:  rozdiel
## Test statistic = 0.49784, p-value = 0.554
## alternative hypothesis: the distribution is asymmetric.
## sample estimates:
## bootstrap optimal m
##                   8
```

```
wilcox.test(rozdiel,mu=0)
```

```
## Warning in wilcox.test.default(rozdiel, mu = 0): cannot compute exact p-value
## with ties
```

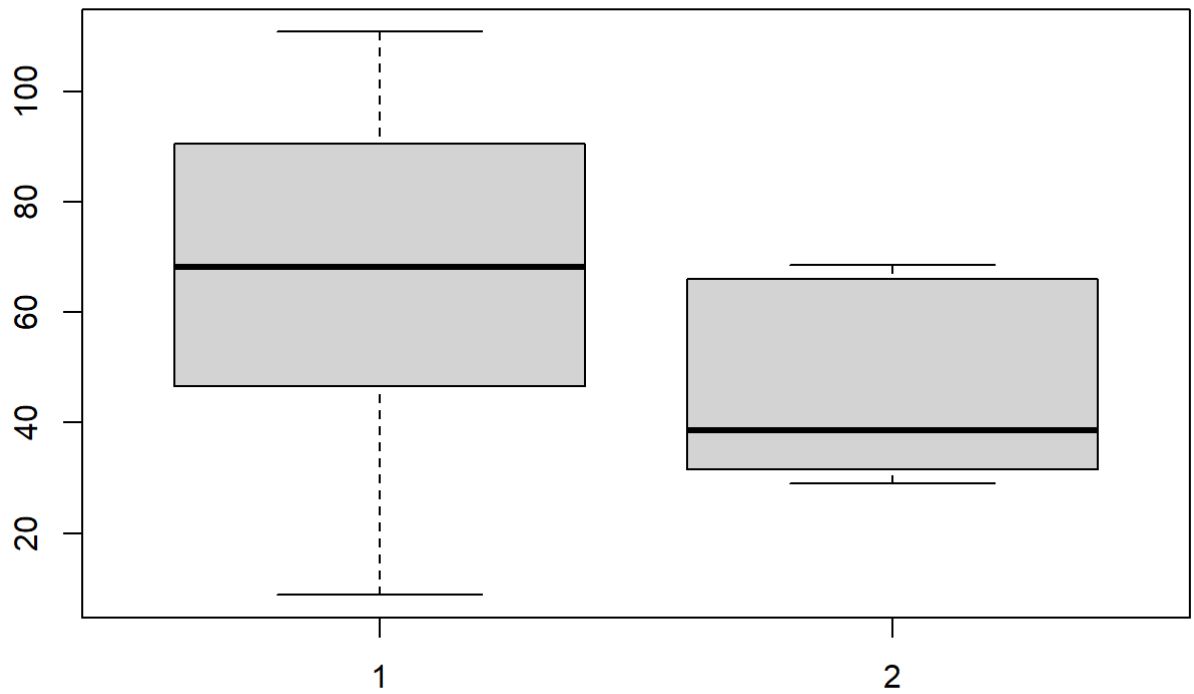
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  rozdiel
## V = 31, p-value = 0.07636
## alternative hypothesis: true location is not equal to 0
```

Nezamietam H0, nema vplyv na spravanie.

```
#####################################################################
```

Obchodnik, dvojvyberovy Wilcoxonov, ale ak su odlisne s roznymi disperziami, tak radsej Kolmogorov Smirnov test.

```
kreditka <- c(34.32,8.8,104.72,59.84,66.88,110.88,46.64,89.8,69.52,90.55)
hotovost <- c(36.96,67.76,40.48,64.24,68.64,29.04,32.56,30.4)
boxplot(kreditka,hotovost)# zdaju sa odlisne
```



Uprava dat, aby sme mohli pouzit neparametricky test pre rovnost disperzii

```
data <- data.frame("platba"=c(kreditka,hotovost),"sposob"=rep(c(1,2),
                      times=c(length(kreditka),length(hotovost))))
levene.test(data$platba,data$sposob)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  data$platba
## Test Statistic = 2.0228, p-value = 0.1742
```

Nezamietam hypotezu o rovnosti disperzii, dvojvyberovy WT.

```
wilcox.test(kreditka,hotovost)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  kreditka and hotovost
## W = 59, p-value = 0.1011
## alternative hypothesis: true location shift is not equal to 0
```

Nezamietam H0, nie su odlisnosti. Teraz aspon jeden datovy subor a testy nahodnosti.

```
library(randtests)
```

```
##
## Attaching package: 'randtests'
```

```
## The following object is masked from 'package:lawstat':
##
##     runs.test
```

```
runs.test(kreditka)
```

```
##
##  Runs Test
##
## data:  kreditka
## statistic = 0, runs = 6, n1 = 5, n2 = 5, n = 10, p-value = 1
## alternative hypothesis: nonrandomness
```

```
turning.point.test(kreditka)
```

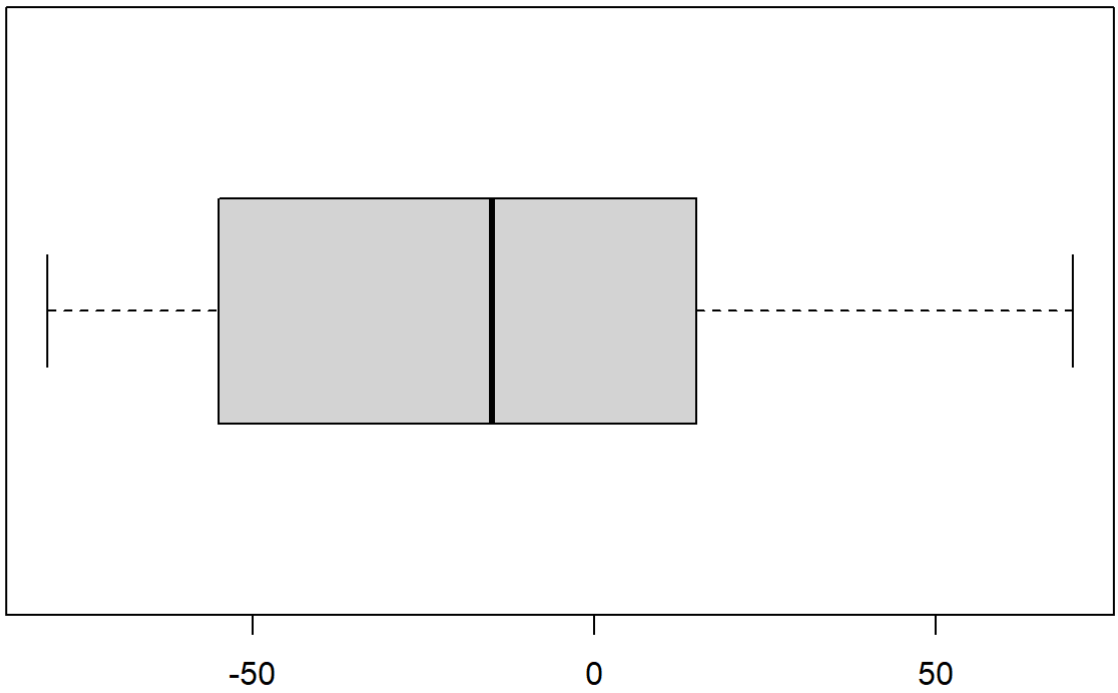
```
##
##  Turning Point Test
##
## data:  kreditka
## statistic = 1.3814, n = 10, p-value = 0.1671
## alternative hypothesis: non randomness
```

Merania su nahodne.

```
#################################################################
```

Plnicka- ak symetria, potom WT, ak nie su symetricke sign test.

```
farba <- c(-50, 10, -10,-80, 70, -20, 20, -60)
boxplot(farba,horizontal = T)
```




```
skewness(farba)
```

```
## [1] 0.3491297
```

```
symmetry.test(farba)# su symetricke
```

```
##
##  m-out-of-n bootstrap symmetry test by Miao, Gel, and Gastwirth (2006)
##
## data:  farba
## Test statistic = 0, p-value = 0.994
## alternative hypothesis: the distribution is asymmetric.
## sample estimates:
## bootstrap optimal m
##                   8
```

```
wilcox.test(farba,mu=0)
```

```
## Warning in wilcox.test.default(farba, mu = 0): cannot compute exact p-value
## with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  farba
## V = 12, p-value = 0.4401
## alternative hypothesis: true location is not equal to 0
```

Nezamietam H0, plnicka plni spravne.

Zadanie testy dobrej zhody, vypracované

V príkladoch, kde to má zmysel testujte aj prítomnosť extremálnych hodnôt. Teda v prípade normálneho rozdelenia použijeme Dixonov alebo Grubbsov test. Inak použijeme metódu založenú na medzikvartilovom rozpätí (viď prednáška Popisná štatistika)

Príklad 1

Náhodným výberom, ktorý je daný v tabuľke bola vybratá vzorka rozsahu n = 50. Overte na hladine významnosti 0.05, či empirické rozdelenie početností zodpovedá normálnemu rozdeleniu.

```
zi<-c(6,8,10,12,14) #hodnota
ni<-c(6,11,19,9,5) #početnosť jednotlivých hodnôt
data<-rep(zi,ni)
```

Testujeme ako v praxi často, zhodu vo všeobecnosti s normálnym rozdelením bez toho, aby sme poznali parametre, najčastejšie ako predpoklad použitia inej metódy. Keďže parametre nepoznáme, použiť môžeme Lillieforsov a/alebo Shapiro-Wilkov test. Shapiro-Wilkov test je vhodný test,keďže máme menší rozsah, ale výsledky môžeme porovnať.

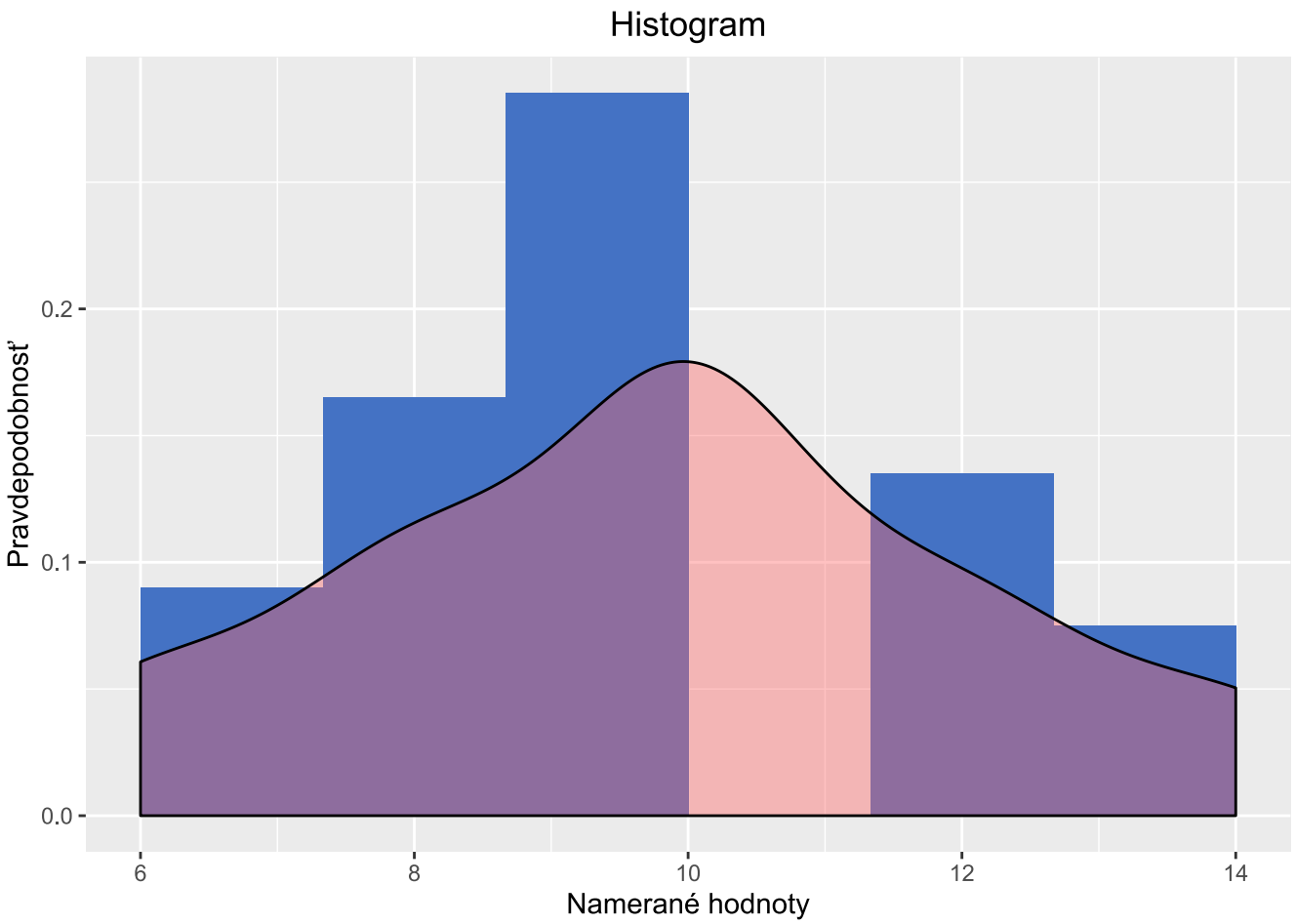
H_0 : Výber pochádza z normálneho rozdelenia.

H_1 : Výber nepochádza z normálneho rozdelenia.

```
df1<-data.frame(data)
library(ggplot2)
```

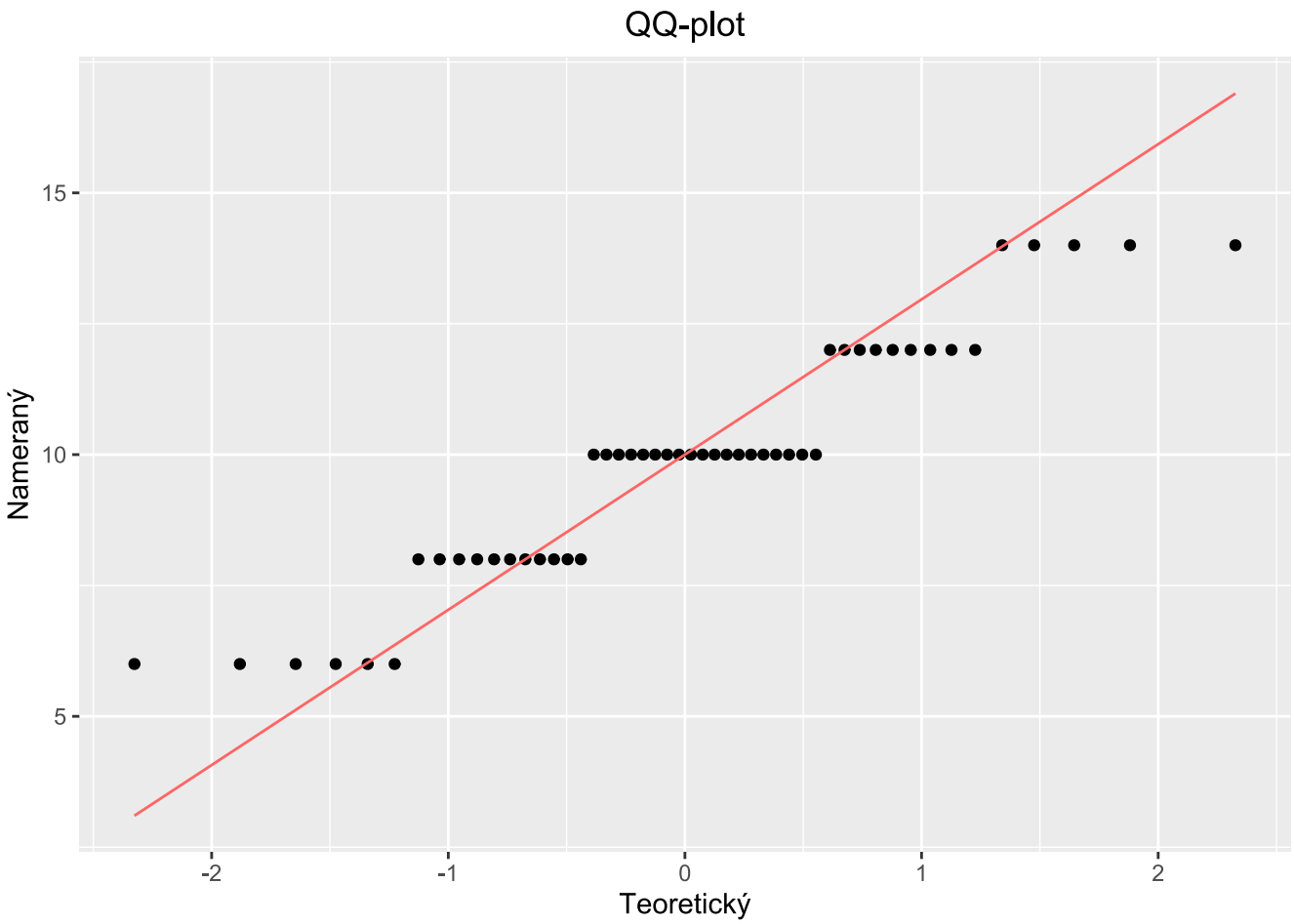
```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
ggplot(df1, aes(data)) +
  geom_histogram(aes(y= ..density..), bins=7, fill="#4472c4")+
  labs(x="Namerané hodnoty", y="Pravdepodobnosť", title = "Histogram")+
  geom_density(alpha=.4, fill="#FF6666")+
  theme(plot.title = element_text(hjust=0.5))
```



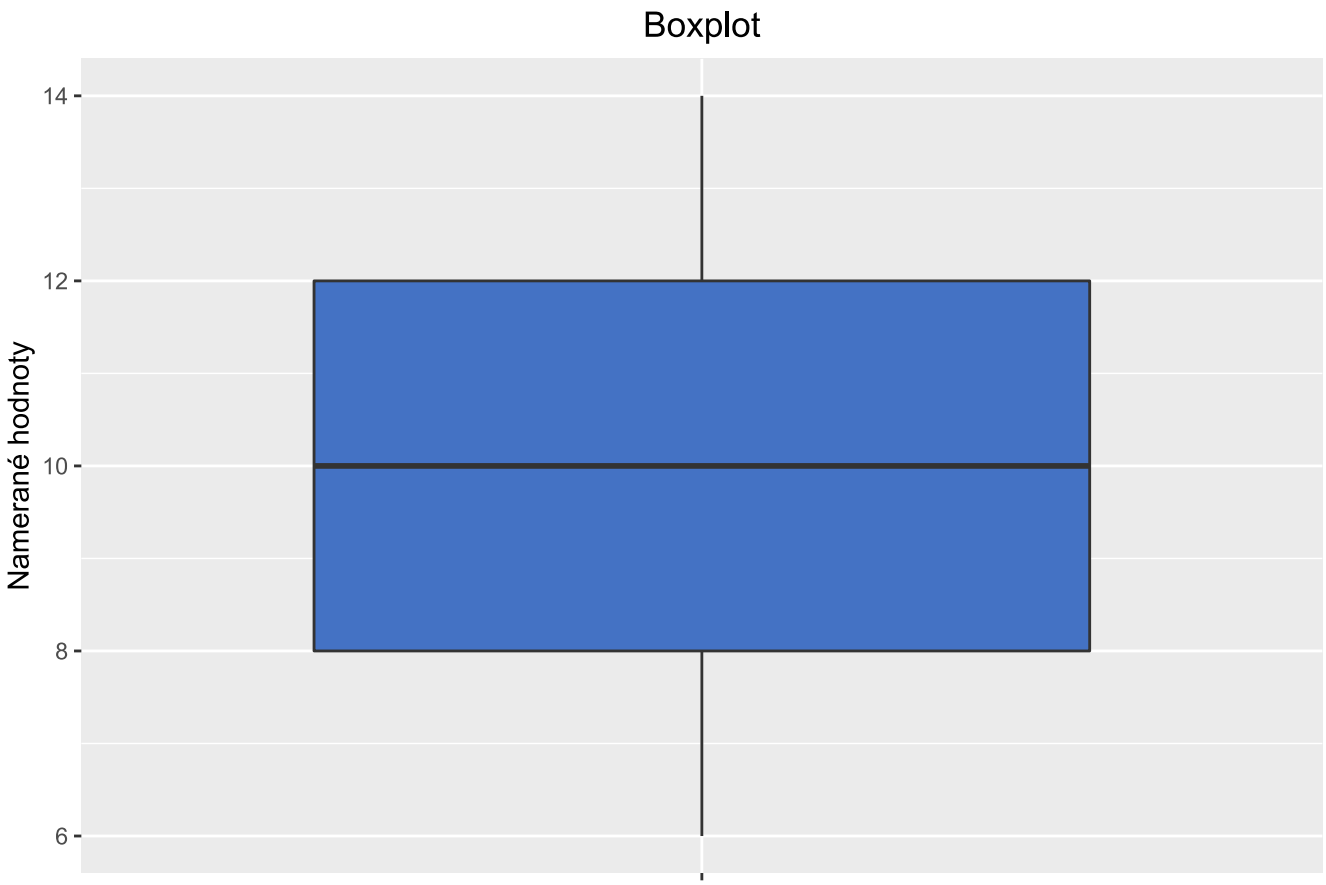
Z histogramu, ktorý je "prerušený" blízko strednej hodnoty, aj keď okrem toho je celkom symetrický, súdime, že dáta nebudú normálne rozdelené.

```
ggplot(df1, aes(sample=data)) +
  stat_qq() + stat_qq_line(colour= "#FF6666")+
  labs(x= "Teoretický", y = "Nameraný", title = "QQ-plot")+
  theme(plot.title = element_text(hjust=0.5))
```



Aj z kvantil-kvantilového grafu môžeme vidieť, že dáta zrejme nebudú normálne rozdelené. vidíme, čo sme mohli vidieť aj pri načítaní dát, opakuje sa len nieľko (celkom málo) hodnôt, to často značí, že dáta nebudú normálne rozdelené, navyše hodnoty sú diskrétne. Podľa qq-grafu by sme hovorili o normálnom rozdelení, ak by body ležali približne na priamke y=x.

```
ggplot(df1, aes(x="", y=data))+
  geom_boxplot(fill="#4472c4")+
  labs(x="", y="Namerané hodnoty", title = "Boxplot")+
  theme(plot.title = element_text(hjust=0.5))
```



Toto je príklad, kedy by sme na základe iba boxplotu zrejme povedali, že dáta sú normálne rozdelené, vzhľadom na "dokonalú" symetriu boxplotu. To je ale dôsledok diskrétnych hodnôt.

```
library(nortest)
```

```
## Warning: package 'nortest' was built under R version 3.5.2
```

```
lillie.test(data)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data
## D = 0.19203, p-value = 8.114e-05
```

```
shapiro.test(data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data
## W = 0.91493, p-value = 0.001554
```

Na základe p-hodnôt (menšie ako hladina významnosti) sme oboma testami zamietli H_0 o normalite výberu. Dáta nepochádzajú z normálneho rozdelenia.

Keďže dáta nie sú normálne rozdelené, odľahlé hodnoty hľadáme pomocou metódy popísanej ešte v prednáške k popisnej štatistike. Na základe nej vidíme, že náhodný výber neobsahuje vybočujúce ani extremálne hodnoty.

```
Q1<-quantile(data, probs=0.25)   # dolný kvartil
Q3<-quantile(data, probs=0.75)  # horný kvartil
IQR<-IQR(data)
k1<-1.5 # pre vybočujúce hodnoty
k2<-3 # pre extremálne hodnoty
data[data<(Q1-k1*IQR)]
```

```
## numeric(0)
```

```
data[data>(Q3+k1*IQR)]
```

```
## numeric(0)
```

Príklad 2

Skúmalo sa dodržiavanie šiestich pravidiel domáceho poriadku nájomníkmi. Jednoduchý náhodný výber 200 bytov odhalil nasledujúce skutočnosti. Na hladine významnosti 0.05 testujte a určte, či vzorka pochádza z rozdelenia, v ktorom počet priestupkov (zo šiestich možných priestupkov = n) na 1 byt je binomicky rozdelená náhodná premenná.

Náhodná premenná X je diskrétna- počet priestupkov na 1 byt.

H_0: X je z binomického rozdelenia.

H_1: X nie je z binomického rozdelenia.

```
xi<-c(0,1,2,3,4,5,6) #počet možných priest. na byt
poc<-c(31,51,70,32,9,5,2) #početnosť xi
```

Overujeme zhodu s diskrétnym, binomickým, rozdelením. Na overenie zhody s diskrétnym rozdelením použijeme Pearsonov χ^2-test. Najprv odhadneme parametre binomického rozdelenia, sú to n=6 zo zadania a p odhadneme.

```
mu<-mean(rep(xi,poc)) # odhad strednej hodnoty počtu priestupkov na domácnosť, mu= n*p
mu # vychádza to na približne 2 priestupky na domácnosť
```

```
## [1] 1.8
```

```
n<-6
p<-mu/n # odhad pravdepodobnosti výskytu priestupku na 1 byt, zo strednej hodnoty a n. Uvažuj
eme zatiaľ teda len či bude priestupok a nie koľko.
prob<-dbinom(xi, n, p) # rozdelenie pravdepodobnosti počtu priestupkov xi v domácnostiach, až
toto je pravdepodobnosť jednotlivých počtov priestupkov
```

Testovaním, na základe p>alfa, H_0 nemôžeme zamietnuť a teda počet priestupkov na 1 byt je NP s binomickým rozdelením.

```
chisq.test(table(rep(xi, poc)), p=prob)
```

```
## Warning in chisq.test(table(rep(xi, poc)), p = prob): Chi-squared approximation
## may be incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(rep(xi, poc))
## X-squared = 33.544, df = 6, p-value = 8.239e-06
```

Príklad 3

V genetickom laboratóriu sa sledovalo 240 potomkov dvoch heterozygotov Aa, Aa. Potomkov typu AA bolo 58, potomkov typu Aa bolo 111 a typu aa bolo 71. Podľa mendelovských zákonov sa očakáva pomer rozdelenia početností 1:2:1. Na 5 percentnej hladine významnosti treba posúdiť zhodu medzi empirickým a teoretickým rozdelením početností.

Náhodná premenná X je počet potomkov daného genotypu. Keďže ide o počet, je to určite diskrétna náhodná premenná. Použijeme teda opäť Pearsonov χ^2-test na zhodu teoretického rozdelenia s empirickým (pozorovaným).

H_0: Teoretické a empirické rozdelenie sú zhodné.

H_1: Teoretické a empirické rozdelenie nie sú zhodné.

```
data3 <- c(rep(c("AA", "Aa", "aa"), c(58, 111, 71)))
head(data3, 60)
```

```
##  [1] "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA"
## [16] "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA"
## [31] "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA"
## [46] "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "AA" "Aa" "Aa"
```

```
prob <- c(1/4, 2/4, 1/4)
chisq.test(table(data3), p=prob)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(data3)
## X-squared = 2.7583, df = 2, p-value = 0.2518
```

Na základe p-hodnoty testu, ktorá je väčšia ako hladina významnosti α=0.05, nulovú hypotézu nemôžeme zamietnuť a teoretické a empirické rozdelenie nie je štatisticky významne odlišné. Na 5% hladine významnosti môžeme povedať, že Mendelovské zákony sú zachované.

Príklad 4

Máme k dispozícii údaje o popolnatosti vzoriek uhlia z dodávok dvoch banských závodov (v % popola). Pomocou Kolmogorovovho-Smirnovho testu overte na hladine významnosti 0.05 hypotézu, že obidva výberové súbory pochádzajú z toho istého základného súboru.

Náhodná premenná je popolnatosť vzoriek uhlia, pre ktorú máme údaje z 2 banských závodov, teda ide o dva výberové súbory. Overujeme, či dva výbery s distribučnou funkciou F a G pochádzajú z toho istého teoretického rozdelenia. Použijeme teda dvojvýberový test pre spojité NP, Kolmogorov Smirnov dvojvýberový test.

H_0: F=G

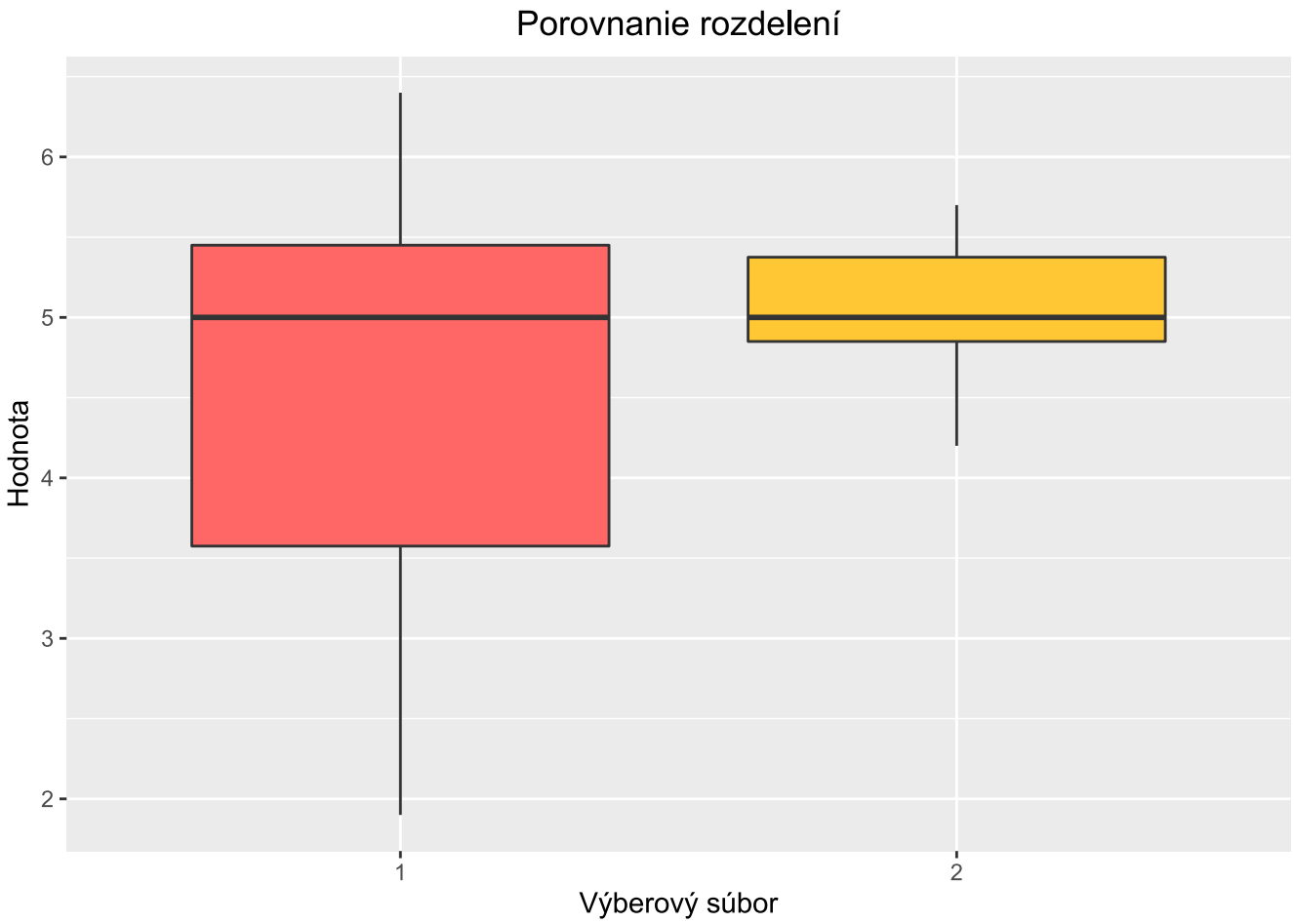
H_1: F≠G

```
x1<- c(5.2, 4.8, 1.9, 5.6, 5.5, 3.4, 5.3, 6.4, 3.5, 3.8)
x2<-c(4.8, 5.0, 5.7, 5.4, 5.5, 4.4, 4.2, 5.0, 5.3, 5.0)
df4<- data.frame(hodnota= c(x1, x2),
                 vyber= rep(c(1,2), c(length(x1), length(x2) )))
df4$vyber<-as.factor(df4$vyber)
head(df4)
```

```
##   hodnota vyber
## 1     5.2     1
## 2     4.8     1
## 3     1.9     1
## 4     5.6     1
## 5     5.5     1
## 6     3.4     1
```

Zdá sa že rozdelenia sa nelíšia v mediáne, 1. je ale plochšie ako 2.

```
ggplot(df4, aes(x=vyber, y= hodnota))+
  geom_boxplot(fill=c("#FF6666", "#FFC733"))+
  labs(x= "Výberový súbor", y= "Hodnota", title = "Porovnanie rozdelení")+
      theme(plot.title = element_text(hjust=0.5))
```




```
ks.test(x1, x2)
```

```
## Warning in ks.test(x1, x2): cannot compute exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x1 and x2
## D = 0.4, p-value = 0.4005
## alternative hypothesis: two-sided
```

Na základe p-hodnoty (viac ako alfa) vidíme, že na hladine významnosti α=0.05 neexistuje medzi rozdeleniami štatisticky významný rozdiel, môžeme povedať, že pochádzajú z toho istého teoretického rozdelenia. H_0 sme teda nezamietli na danej hladine významnosti.

Z boxplotov sa zdá, že ani jeden výberový súbor neobsahuje odľahlé hodnoty. Ešte to otestujeme. Na

overenie prítomnosti outlierov vo výberoch v prípade, že sú normálne rozdelené môžeme použiť Grubbsov alebo Dixonov test. Oba výbery sú normálne rozdelené, s menším počtom pozorovaní, preto použijeme Dixonov test.

H_0: Minimálna (maximálna) hodnota je outlier.

H_1: nie je outlier.

```
shapiro.test(x1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x1
## W = 0.93905, p-value = 0.5425
```

```
shapiro.test(x2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x2
## W = 0.95311, p-value = 0.7053
```

```
library(outliers)
```

```
## Warning: package 'outliers' was built under R version 3.5.2
```

```
dixon.test(x1)
```

```
##
##  Dixon test for outliers
##
## data:  x1
## Q = 0.40541, p-value = 0.2067
## alternative hypothesis: lowest value 1.9 is an outlier
```

```
dixon.test(x1, opposite = T)
```

```
##
##  Dixon test for outliers
##
## data:  x1
## Q = 0.26667, p-value = 0.5938
## alternative hypothesis: highest value 6.4 is an outlier
```

```
dixon.test(x2)
```

```
##
##  Dixon test for outliers
##
## data:  x2
## Q = 0.15385, p-value = 0.904
## alternative hypothesis: lowest value 4.2 is an outlier
```

```
dixon.test(x2, opposite = T)
```

```
##
##  Dixon test for outliers
##
## data:  x2
## Q = 0.15385, p-value = 0.904
## alternative hypothesis: highest value 5.7 is an outlier
```

Ani jeden z výberových súbor neobsahuje outliers.

Príklad 5

V istom časovom období bolo zaznamenaných 391 dopravných nehôd, pričom v pondelok ich bolo 52, v utorok 43, v stredu 54, vo štvrtok 45, v piatok 62, v sobotu 66 a v nedeľu 69. Treba zistiť, či sa dopravné nehody vyskytujú pravidelne vo všetkých dňoch týždňa alebo či sú v niektorých dňoch týždňa štatisticky významne častejšie.

Pozorovaná NP je počet nehôd v rôznych dňoch v týždni. Opäť ide o diskrétnu NP. Keďže chceme zistiť, či sa vyskytujú pravidelne vo všetkých dňoch týždňa, teda, či je rozdelenie rovnomerné (diskrétne). Použijeme Pearsonov χ^2-test.

H_0: X~Ro(7), teda, že nehody sa vyskytujú rovnomerne počas týždňa (7 dní).

H_1: $\neg H_0$, sú dni, kedy je počet nehôd štatisticky významne vyšší ako iné dni.

```
data5 <- table(c(
  rep("Po", 52), rep("Ut", 43), rep("St", 54),
  rep("Št", 45), rep("Pi", 62), rep("So", 66), rep("Ne", 69)
))
data5
```

```
##
## Ne Pi Po So St Št Ut
## 69 62 52 66 54 45 43
```

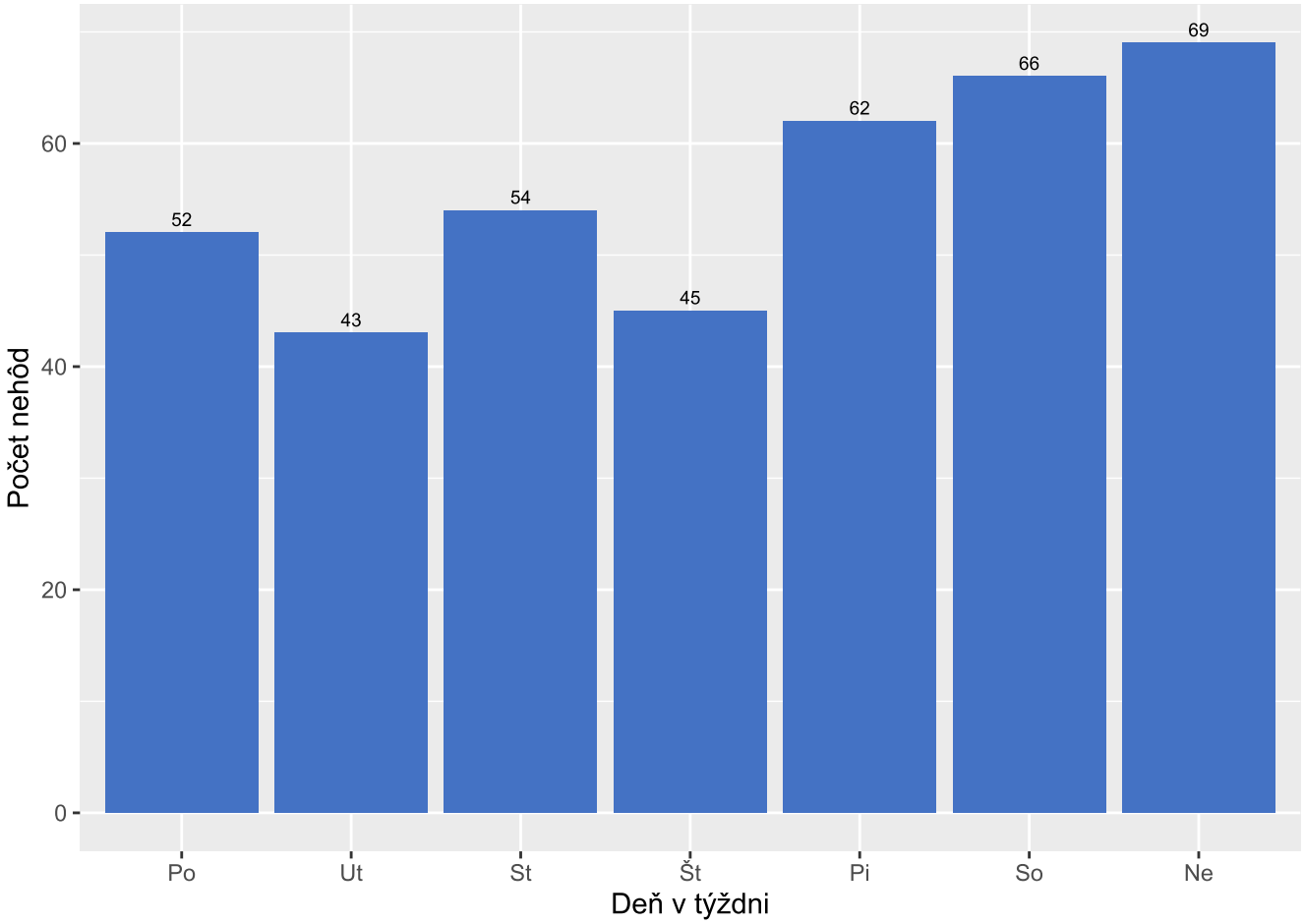
```
prob <- rep(1/7, 7)

df5<-data.frame(data5)
```

Definujeme úrovne faktora a ich poradie, aby v grafe neboli usporiadane podľa abecedy.

```
df5$Var1<-factor(df5$Var1, levels=c("Po", "Ut", "St", "Št", "Pi", "So", "Ne"))
```

Z grafu ale už aj zo zadania sme videli, že počas víkendových dní a v piatok bol počet nehôd vyšší ako 60. V ostatné dni počet nehôd je maximálne 54 a to v stredu. Či je rozdiel významný, otestujeme.

```
ggplot(df5, aes(x=Var1 , y=Freq))+
  geom_bar(stat="identity",fill="#4472c4")+
  geom_text(aes(label=Freq), vjust=-0.5,color="black", size=2.5)+
  labs(x="Deň v týždni", y="Počet nehôd")
```



Na základe p-hodnoty testu, ktorá je väčšia ako hladina významnosti 0.05, nulovú hypotézu nemôžeme zamietnuť. Nepreukázal sa štatisticky významný rozdiel v počte nehôd v rámci dní v týždni.

Príklad 6

Bolo vybraných 13 polí rovnakej kvality. Na 8 z nich sa skúšal nový spôsob hnojenia, na zvyšných 5 bol použitý tradičný spôsob hnojenia. Výnosy pšenice v tonách na hektár boli pri novom spôsobe hnojenia 5.7, 5.5, 4.3,

5.9, 5.2, 5.6, 5.8, 5.1 a pri tradičnom spôsobe hnojenia 5, 4.5, 4.2, 5.4, 4.4. Treba zistiť, či nový spôsob hnojenia má vplyv úrodu pšenice.

Náhodná premenná je výnos pšenice, meraná pri dvoch nezávislých podmienkach. Použijeme dvojvýberový test zhody dvoch rozdelení. Distribučnú funkciu pre nový spôsob hnojiva označíme F a pre tradičný spôsob G.

H_0: F=G

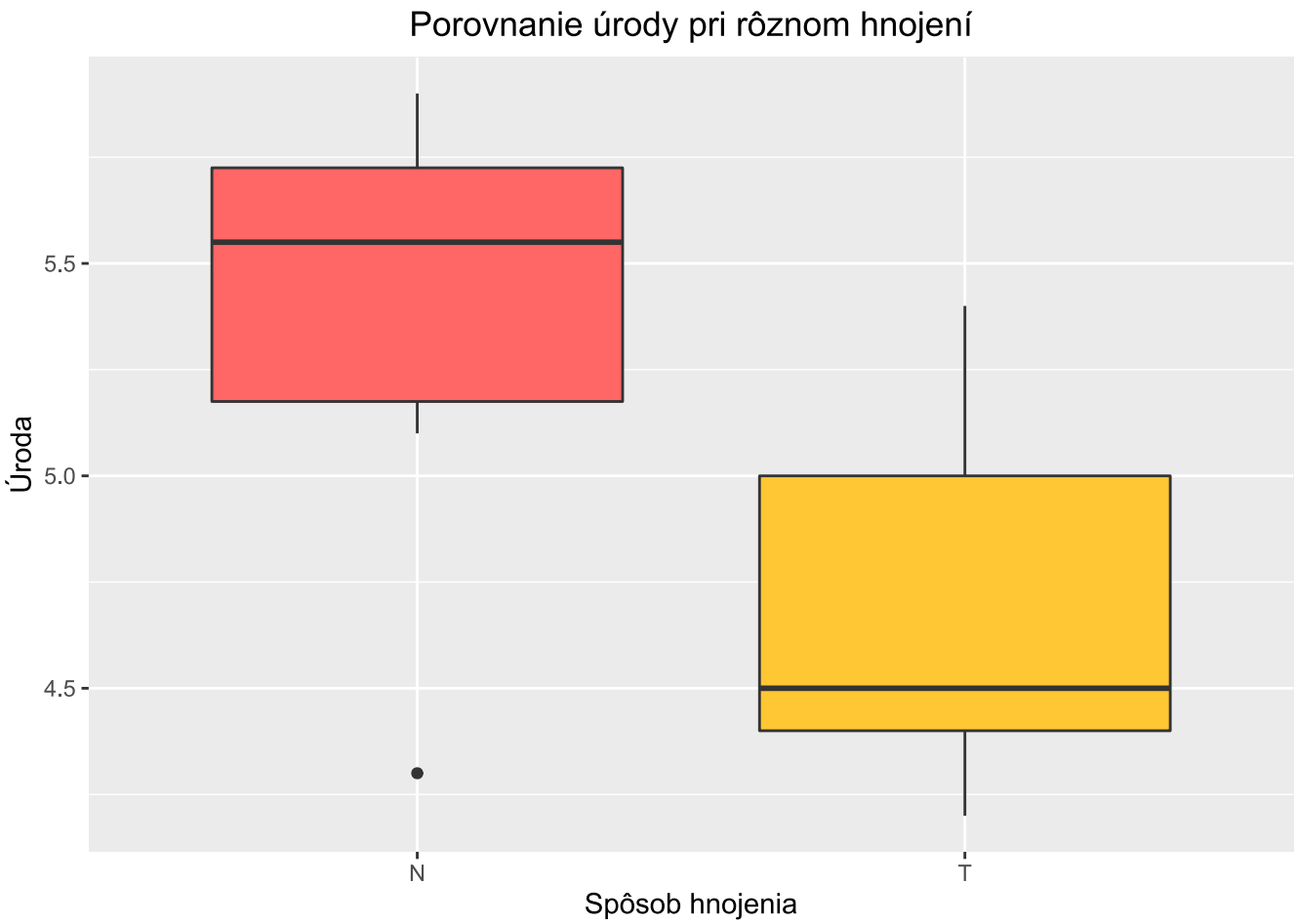
H_1: F≠G

Testujeme najprv obojstrannú alternatívnu hypotézu a ak zamietneme H_0, pozrieme sa aj na jednostrannú alternatívnu hypotézu, na určenie v akom sú vzťahu výnosy pri tradičnom a novom spôsobe hnojenia, pre zistenie, ktoré hnojenie je výhodnejšie vzhľadom na úrodu.

```
x1 <- c(5.7, 5.5, 4.3, 5.9, 5.2, 5.6, 5.8, 5.1)
x2 <- c(5, 4.5, 4.2, 5.4, 4.4)
df6<-data.frame(uroda=c(x1,x2),
                hnojenie=rep(c("N","T"), c(length(x1), length(x2))))
head(df6)
```

```
##   uroda hnojenie
## 1   5.7        N
## 2   5.5        N
## 3   4.3        N
## 4   5.9        N
## 5   5.2        N
## 6   5.6        N
```

```
ggplot(df6, aes(x=hnojenie , y= uroda))+
  geom_boxplot(fill=c("#FF6666", "#FFC733"))+
  labs(x="Spôsob hnojenia", y="Úroda", title = "Porovnanie úrody pri rôznom hnojení")+
  theme(plot.title = element_text(hjust=0.5))
```



Z boxplotov sa zdá, že pri novom type hnojenie je výnos pšenice vyšší ako pri tradičnom.

```
ks.test(x1, x2)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  x1 and x2
## D = 0.675, p-value = 0.07925
## alternative hypothesis: two-sided
```

H_0 nemôžeme na hladine významnosti α=0.05 zamietnuť, nový spôsob hnojenia nemá štatisticky významný vplyv na výnos pšenice. P-hodnota však nie je príliš veľká (na hladine významnosti 0.1 by išlo o štatisticky významný výsledok), stálo by za to experiment zopakovať, možno aj s väčšou vzorkou.

Príklad 7

Skupine 7 pacientov po otrase mozgu testovali reakčný čas na vizuálny podnet. Namerali sa tieto výsledky v sekundách: 5.21, 6.73, 4.31, 3.89, 2.10, 3.31, 2.86. Na hladine významnosti ɑ = 0,05 testujte, či namerané hodnoty môžeme považovať za realizáciu náhodného výberu z normálneho rozdelenia.

Náhodná premenná je dĺžka reakčného času na vizuálny podnet. (spojitá premenná)

H_0: NP je z normálneho rozdelenia.

H_1: NP nie je z normálneho rozdelenia.

```
data7 <- c(5.21, 6.73, 4.31, 3.89, 2.10, 3.31, 2.86)
shapiro.test(data7)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data7
## W = 0.97132, p-value = 0.9078
```

```
lillie.test(data7)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data7
## D = 0.14979, p-value = 0.9056
```

Na základe oboch testov, keďže p>alfa, nemôžeme na hladine významnosti 0.05 zamietnuť nulovú hypotézu o normalite reakčného času pacientov.

```
dixon.test(data7)
```

```
##
##  Dixon test for outliers
##
## data:  data7
## Q = 0.32829, p-value = 0.4451
## alternative hypothesis: highest value 6.73 is an outlier
```

```
dixon.test(data7, opposite = T)
```

```
##
##  Dixon test for outliers
##
## data:  data7
## Q = 0.16415, p-value = 0.9054
## alternative hypothesis: lowest value 2.1 is an outlier
```

```
grubbs.test(data7)
```

```
##
##  Grubbs test for one outlier
##
## data:  data7
## G = 1.72520, U = 0.42128, p-value = 0.1647
## alternative hypothesis: highest value 6.73 is an outlier
```

```
grubbs.test(data7, opposite = T)
```

```
##
##  Grubbs test for one outlier
##
## data:  data7
## G = 1.26480, U = 0.68893, p-value = 0.6764
## alternative hypothesis: lowest value 2.1 is an outlier
```

Testovaním sme zistili, že medzi nameranými hodnotami sa nenachádza vybočujúca hodnota.

Príklad 8

Firma ABC nakupuje baterky do elektronických prístrojov. Dodávateľ garantuje životnosť bateriek minimálne 19 hodín s odchýlkou 1 hodina. Kontrolór náhodne vybral 10 bateriek a sleduje ich životnosť. Testom overte, či výber pochádza z normálneho rozdelenia s danými parametrami. Namerané hodnoty: 19.2, 21.6, 17.5, 18.4, 18.8, 16.9, 20.4, 19.9, 18.1, 15.4

Tu testujeme zhodu s normálnym rozdelením s konkrétnymi parametrami μ=19 a σ^2=1. Použijeme preto Kolmogorov-Smirnov test. NP X je životnosť bateriek v hodinách.

H_0: X~N(19; 1)

H_1: $\neg H_0$

```
data8<- c(19.2, 21.6, 17.5, 18.4, 18.8, 16.9, 20.4, 19.9, 18.1, 15.4)
ks.test(data8, "pnorm", 19, 1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data8
## D = 0.23319, p-value = 0.5715
## alternative hypothesis: two-sided
```

Na základe testu H_0 nemôžeme zamietnuť na hladine významnosti α=0.05 a teda životnosť bateriek je náhodná premenná s normálnym rozdelením so strednou hodnotou 19 hodín s roztpylom (aj odchýlkou) 1 hodina. Testujeme ešte prítomnosť outlierov. Na hladine významnosti 0.05 môžeme povedať, že v dátach nie sú prítomné odľahlé pozorovania.

```
dixon.test(data8)
```

```
##
##  Dixon test for outliers
##
## data:  data8
## Q = 0.3, p-value = 0.4774
## alternative hypothesis: lowest value 15.4 is an outlier
```

```
dixon.test(data8, opposite = T)
```

```
##
##  Dixon test for outliers
##
## data:  data8
## Q = 0.25532, p-value = 0.637
## alternative hypothesis: highest value 21.6 is an outlier
```

```
grubbs.test(data8)
```

```
##
##  Grubbs test for one outlier
##
## data:  data8
## G = 1.79520, U = 0.60214, p-value = 0.2527
## alternative hypothesis: lowest value 15.4 is an outlier
```

```
grubbs.test(data8, opposite = T)
```

```
##
##  Grubbs test for one outlier
##
## data:  data8
## G = 1.66140, U = 0.65924, p-value = 0.3822
## alternative hypothesis: highest value 21.6 is an outlier
```

Príklad 9

V súbore KriminalitaEU sú reálne dáta o kriminálnej činnosti v 43 krajinách Európy za rok 2013. Dáta sú voľne dostupné na internetovej stránke UNODC- United Nations Office on Drugs and Crime. Vybrali sme 11 rôznych trestných činností: napadnutie (Npdnt), únos (Únos), krádež (Krdez), lúpež (Lúpež), vlámanie (Vlamn), vlámanie do domácnosti (VlmDo), krádež osobných motorových vozidiel (KrOMV), krádež motorových vozidiel (KrMV), celkové sexuálne násilie (CeSxN), znásilnenie (Znsln) a sexuálne trestné činy spáchané na deťoch (SxTDt). Hodnoty premenných predstavujú počty trestných činov v prepočte na 100 tisíc obyvateľov. Vyberte jednu z daných premenných a testom zitite, či je normálne rozdelená. Vyšetrite aj prítomnosť extremálnych hodnôt. Závery interpretujte.

Nemám prehľad o ostatných skupinách, ale u mňa si väčšina vybrala krádež. (náhoda?)

Náhodná premenná je počet krádeží v krajinách Európy v prepočte na 100 tisíc obyvateľov.

```
library(readxl)
```

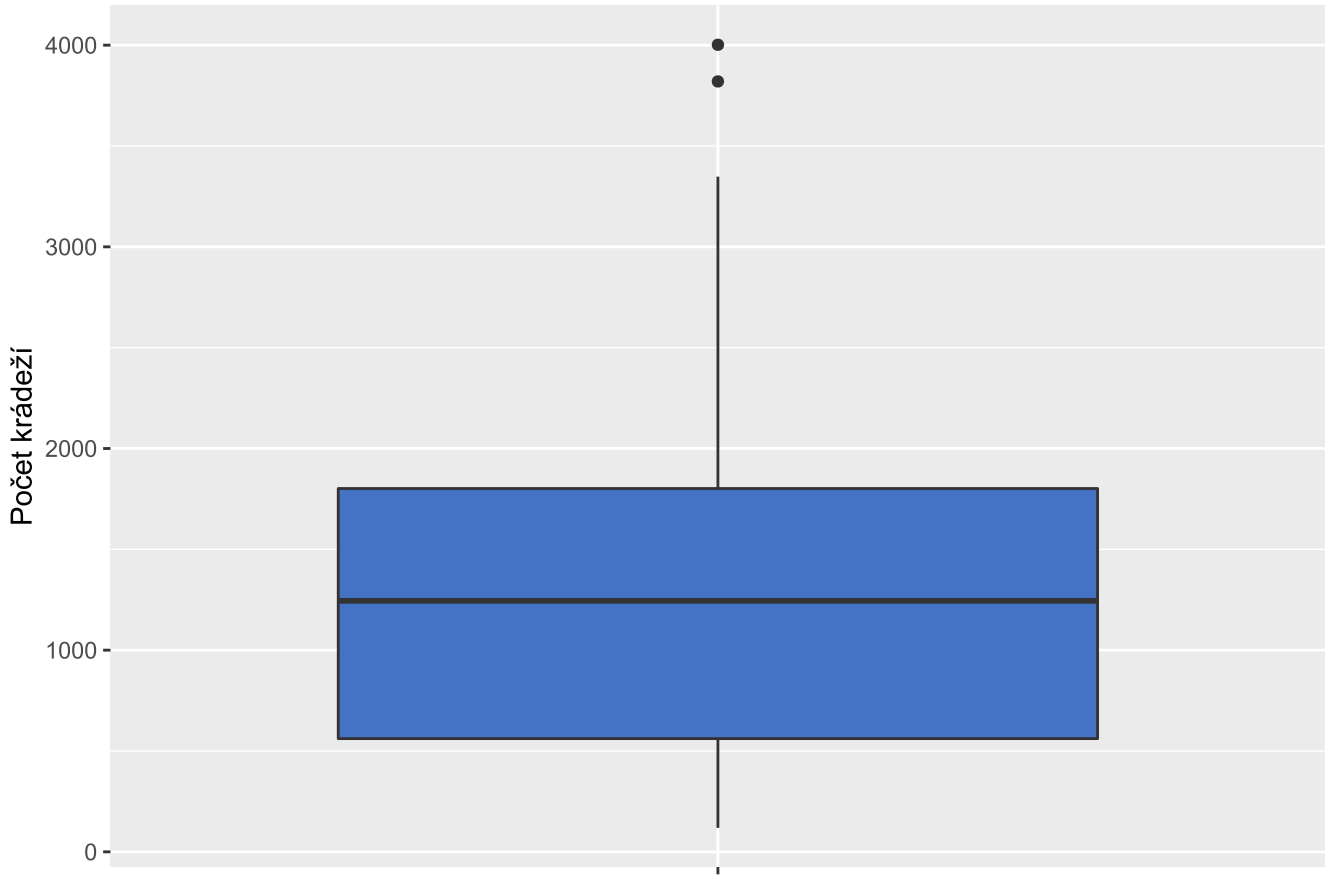
```
## Warning: package 'readxl' was built under R version 3.5.3
```

```
data9<- read_excel("KriminalitaEU.xlsx")
head(data9)
```

```
## # A tibble: 6 x 12
##   Krajina    Npdnt  Únos Krdez Lúpež Vlamn VlmDo KrOMV  KrMV CeSxN Znsln SxTDt
##   <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Bielorusko   9.6   0.1  599.  28.2  322. 177.    7.8   9.2   4.8   1     4.8
## 2 Bulharsko   34.2   1.2  627.  41.2  238.  88.6  52.8  49.6   8.7   2.3  21.5
## 3 Česko      175.    0.1 1173.  28.5  583. 104.   52.8 100.   19.7   5.5  43.6
## 4 Maďarsko   134.    0.1 1244.  23.1  382. 156.   47    57.2  59.6   2.5 308.
## 5 Poľsko       1.2   1.2  524.  32.4  310.  59.9  52.8  40.8   8.4   3.6  20.9
## 6 Moldavsko    9.3   3.3  441    4.2   65   75.8   5.3   5.6  17.4  10    23.8
```

Z boxplotu sa zdá, že rozdelenie je vyšikmené a obsahuje 2 outliery.

```
ggplot(data9, aes(x="", y=Krdez))+
  geom_boxplot(fill="#4472c4")+
  labs(x="", y="Počet krádeží")
```



Testom overíme normalitu. Podľa výsledku Shapiro-Wilkovho testu počet krádeží nie je normálne rozdelená NP. Lillie-Forsov testom by sme normalitu nezamietli, avšak vzhľadom na počet pozorovaní a silu testov sa budeme riadiť Shapiro-Wilkovým testom.

```
shapiro.test(data9$Krdez)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data9$Krdez
## W = 0.90882, p-value = 0.002343
```

```
lillie.test(data9$Krdez)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data9$Krdez
## D = 0.1091, p-value = 0.2235
```

Pozrieme sa ešte na vybočujúce hodnoty. Nízke vybočujúce hodnoty v dátach nie sú. Vysoké vybočujúce hodnoty, ktoré nie sú extremálne sú 2. Teda žiadna krajina nemá štatisticky významne nižší počet krádeží na

100tis. obyvateľov, dve maju významne viac krádeží.

```
Q1<-quantile(data9$Krdez, probs=0.25)   # dolný kvartil
Q3<-quantile(data9$Krdez, probs=0.75)  # horný kvartil
IQR<-IQR(data9$Krdez)
data9$Krdez[data9$Krdez<(Q1-k1*IQR)]
```

```
## numeric(0)
```

```
data9$Krdez[data9$Krdez>(Q3+k1*IQR)]
```

```
## [1] 4002.0 3820.2
```

```
data9$Krdez[data9$Krdez>(Q3+k2*IQR)]
```

```
## numeric(0)
```

Prirodzene nás zaujíma, ktoré sú to krajiny. Najviac krádeží na 100tis. obyvateľov je vo Švédsku a druhý najvyšší počet pripadá na Holandsko. Tento výsledok je celkom prekvapivý.

```
data9$Krajina[data9$Krdez==4002]
```

```
## [1] "Švédsko"
```

```
data9$Krajina[data9$Krdez==3820.2]
```

```
## [1] "Holandsko"
```

ANOVA (jednovýberová analýza rozptylu)

Dáta aj vizualizujte, nezabudnite overiť, či sú splnené podmienky použitia ANOVA.

1. Sledujeme priemernú spotrebu elektrickej energie v štyroch bratislavských obvodoch prepočítanú na osobu a deň, výsledky sú zaznamenané v tabuľke

1. obv	1.2	1.3	1.5	1.2	1.5	1.7
2. obv	1.4	1.5	1.8	1.6	2.1	2.5
3. obv	2.4	1.8	1.5	1.7	2	1.4
4. obv	1.5	1.4	1.6	1.9	2	

Na hladine významnosti α=0.05 testujte hypotézu o rovnosti spotreby pre jednotlivé obvody. Ak sú spotreby rozdielne, zistite, ktoré sa líšia.

2. Skúmala sa akosť obalového materiálu od štyroch rôznych výrobcov tak, že sa odobratým vzorkám stanovila pevnosť v pretrhnutí. Výsledky sú dané tabuľkou

1.	9.1	8.5	8.6	7.2	7.6	8.1	8.2	7.6
2.	8.5	9	8.1	8.5	8	7.4		
3.	9	7.9	8.2	7.6	8.7	8.6	8.8	
4.	7.7	7.2	7.7	8.2	7.7	7.5	7.9	6.7

Na hladine významnosti α=0.05 testujte hypotézu o rovnakej akosti obalového materiálu rôznych výrobcov. Ak sú akosti rozdielne, zistite, ktorí výrobcovia sa líšia.

3. Malá fabrika vyrába papierové nákupné tašky. Výskumné oddelenie tejto fabriky sa snaží zvýšiť zaťaženie tašky udané v jednotkách psí. Sila tašky závisí od koncentrácie tvrdého dreva v celulóze, ktorá sa používa na výrobu tašiek. Preto sa vybrali 4 koncentrácie tvrdého dreva na výrobu tašiek a tie sa potom testovali vzhľadom na zaťaženie. Je zaťaženie tašiek vzhľadom na koncentráciu tvrdého dreva rovnaké? Ak nie, ktoré sa líšia a ktorá koncentrácia dreva sa javí najvhodnejšou (predpokladajme, že koncentrácia dreva nemá vplyv na výrobnú cenu a prácnosť výroby)?

koncentrácia v %	zaťaženie tašky					
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

ANOVA (dvojvýberová analýza rozptylu)

Dáta aj vizualizujte, tu budeme predpokladať, že podmienky použitia ANOVA sú splnené (nemusíte overovať).

4. 6 druhov pšenice sa pestuje na štyroch typoch pôdy. V tabuľke je uvedená úroda pre každú kombináciu (vždy iba jedno meranie v triede).

pšenica	pôda			
	I.	II.	III.	IV.
A	30	27	19	25
B	28	27	22	24
C	32	30	23	21
D	18	19	21	21
E	25	24	22	22
F	17	17	19	20

Na hladine významnosti α = 0.05 zistite, či rôzne druhy pšenice a rôzne typy pôdy vplývajú na úrodu.

5. Pacientom rôznych vekových skupín (1-mládež do 18 rokov, 2-dospelí do 50 rokov, 3-dospelí nad 50 rokov boli podávané 4 rôzne lieky na podporu imunity. Pacienti boli sledovaní celý rok a zaznamenávali sa počty ochorení dýchacích ciest, Výsledky sú v tabuľke

	liek 1	liek 2	liek 3	liek 4
vek 1	1, 2, 2, 3	2, 1, 2	3, 2, 3, 3,	2, 2, 2
vek 2	2, 2, 3, 2, 1, 4	2, 3, 1, 1	3, 4, 3, 2, 1, 1, 2, 3	3, 1, 2, 2, 3
vek 3	4, 3, 3, 2	3, 3, 3, 1	5, 4, 4,3, 3, 2	3, 3, 3, 4, 3, 2, 1

Testujte, či počty ochorení závisia od veku pacienta alebo podávaného lieku. Ak áno, ktoré triedy sa líšia? Hladina významnosti $\alpha = 0.05$.

6. Sledujeme klinickú štúdiu, ktorej cieľom je posúdiť bio ekvivalenciu dvoch uroxylových prípravkov (Uroxan, Urolon). Štúdia prebieha súbežne v troch medicínskych centrách (A, B, C). V každom centre je 16 dobrovoľníkov. 8 z nich je podávaný Uroxan a 8 je podávaný Urolon. Dobrovoľníkom je, okrem iného, meraná hladina hemoglobínu [g/l] v krvi. Výsledky sú dané tabuľkou. Analýzou rozptylu zistite, či hodnoty hemoglobínu v krvi sú ovplyvnené podávaným liekom, centrom medicínskeho výskumu alebo kombináciou týchto faktorov ($\alpha = 0.05$).

prípravok	dobrovoľník	A	B	C
Uroxan	1	138	135	137
	2	126	174	123
	3	141	157	124
	4	151	136	152
	5	163	137	138
	6	139	140	144
	7	146	136	148
	8	144	144	141
Urolon	1	126	143	151
	2	132	142	142
	3	163	125	168
	4	145	155	167
	5	142	149	143
	6	159	153	139
	7	130	137	147
	8	139	133	131

anova_vypracovanie.R

Janka

2023-04-14

ANOVA (Analysis of Variance)

```
library(RColorBrewer)
library(vioplot)
```

```
## Loading required package: sm
```

```
## Warning: package 'sm' was built under R version 3.6.3
```

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.6.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggplot2)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.6.3
```

Spotreba energie

```
obv <- rep(c(1,2,3,4),times=c(6,6,6,5))
elektrina <-c(1.2, 1.3, 1.5, 1.2, 1.5, 1.7,1.4, 1.5, 1.8, 1.6, 2.1, 2.5,
              2.4, 1.8, 1.5, 1.7, 2, 1.4,1.5, 1.4, 1.6, 1.9, 2)
data <- data.frame(obv,elektrina)
```

Normalita dat

```
library(nortest)
tapply(data$elektrina,data$obv,shapiro.test)
```

```
## $`1`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.89613, p-value = 0.3515
##
##
## $`2`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.91908, p-value = 0.4988
##
##
## $`3`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95077, p-value = 0.7465
##
##
## $`4`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.91548, p-value = 0.5012
```

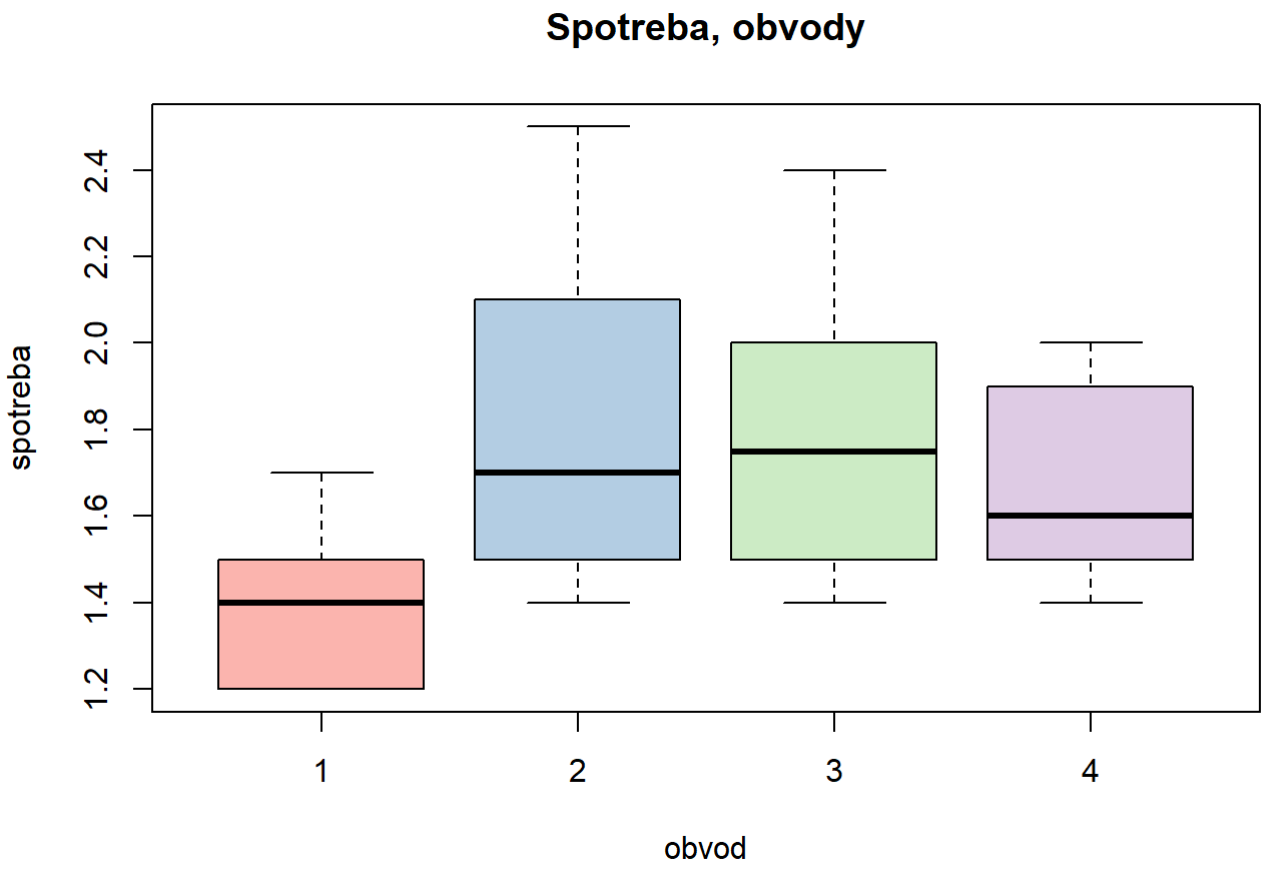
Kazda z datovych podmnozin je normalne rozdelena Test rovnosti disperzii, Bartlettov test

```
bartlett.test(data$elektrina,data$obv)
```

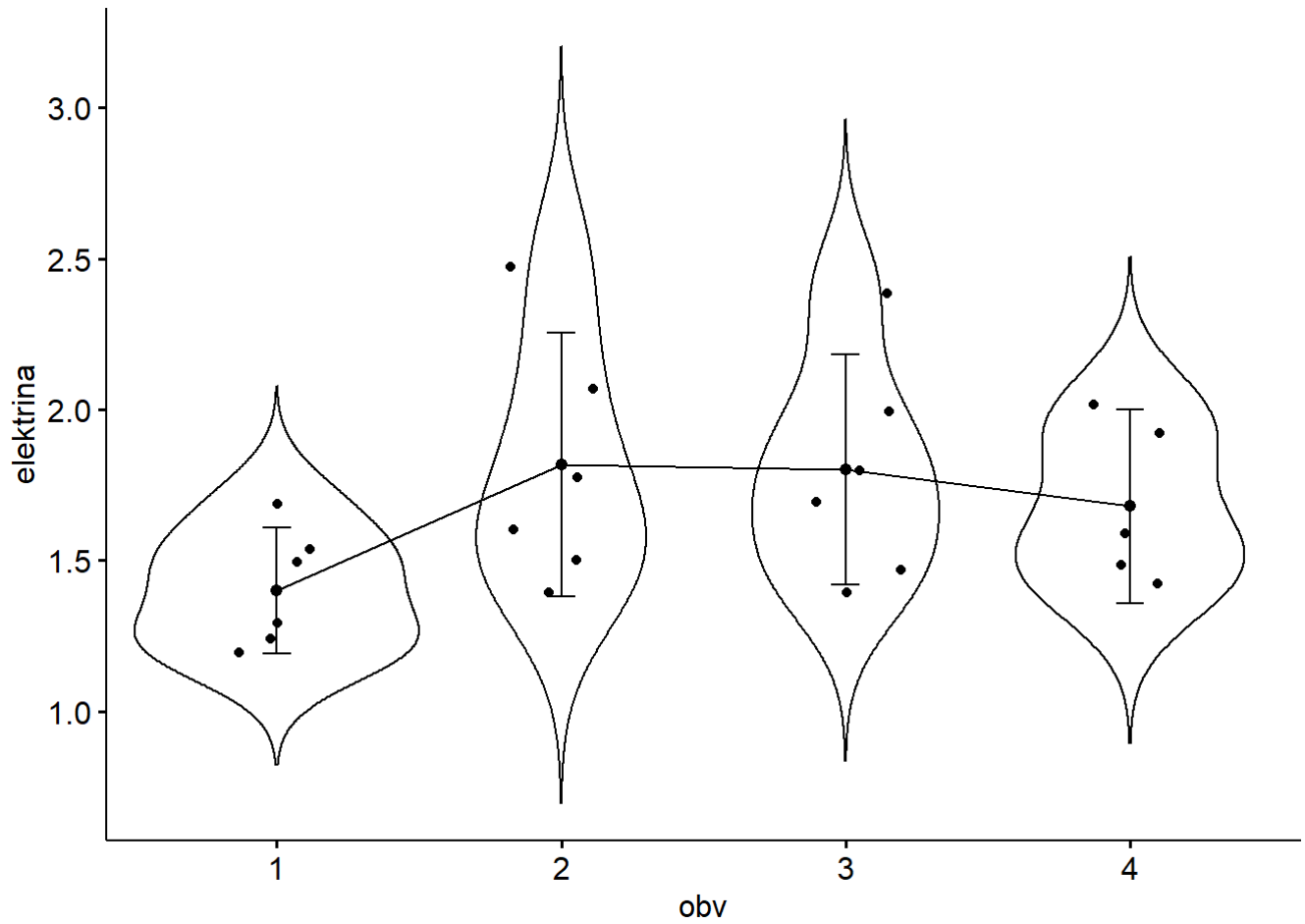
```
##
##  Bartlett test of homogeneity of variances
##
## data:  data$elektrina and data$obv
## Bartlett's K-squared = 2.7329, df = 3, p-value = 0.4347
```

0.43>0.05, nezamietam hypotezu o rovnosti disperzii, Podmienky ANOVA su splnene. Pre testom este nakreslime zopar grafov

```
boxplot(data$elektrina~data$obv,col=brewer.pal(4,"Pastel1"),
        main="Spotreba, obvody",xlab="obvod",
        ylab="spotreba")
```




```
ggline(data,x="obv",y="elektrina",
       add=c("mean_ci","violin","jitter"))
```



```
##########################################################
```

ANOVA vyzaduje faktorizovat

```
obv <- factor(data$obv)
anova <- aov(elektrina~obv)
summary(anova)
```

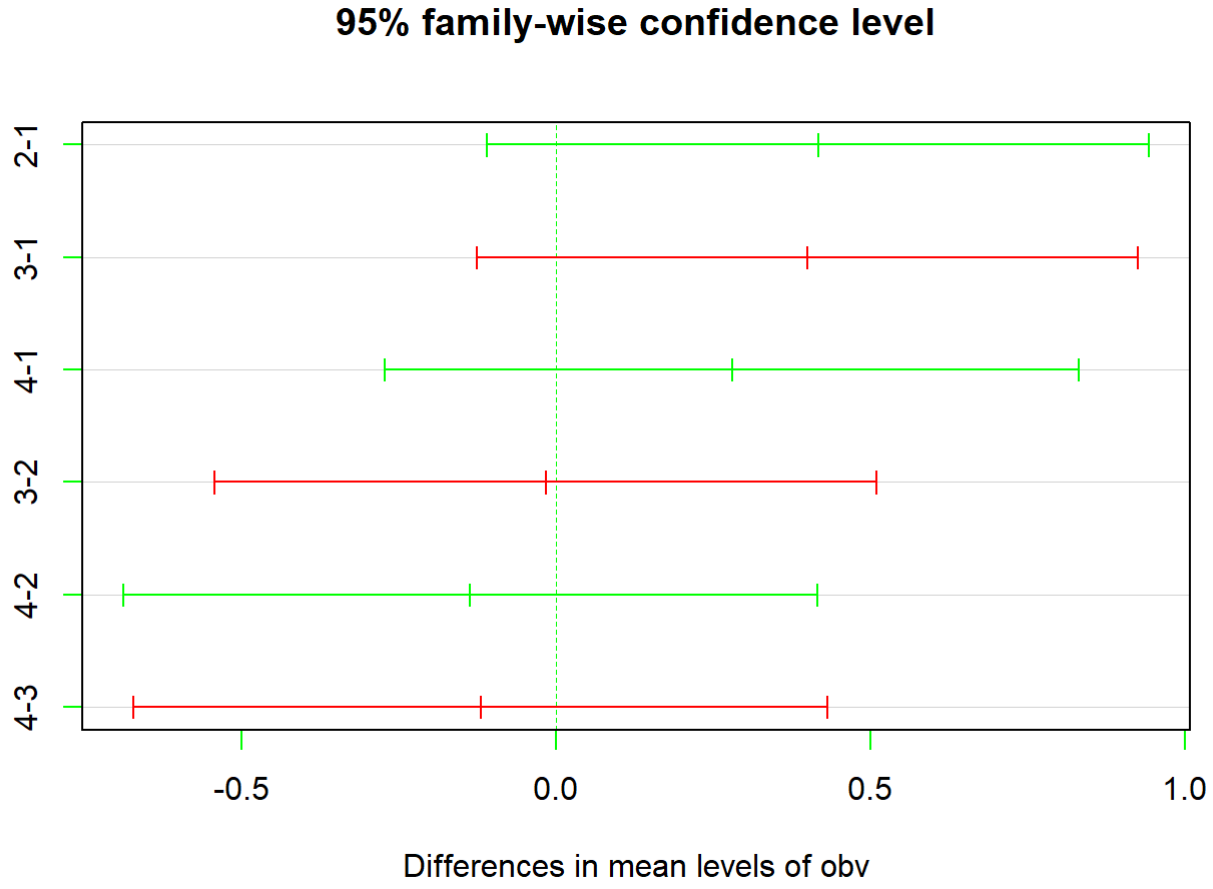
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## obv          3  0.668  0.2227   2.119  0.132
## Residuals   19  1.996  0.1051
```

P hodnota>0.05, Nezamietam hypotezu o rovnosti strednych hodnot. Nezamietam hypotezu o nulovosti triediaceho faktora, faktor obvod je statisticky nevyznamny a nema vplyv na priemerny prijem. Iba demonstracne post testy, ktorymi zistime odlisnosti pre dvojice tried, my budeme pouzivat Tukey a Scheffe test. V tomto pripade ich nie je treba robit

```
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = elektrina ~ obv)
##
## $obv
##            diff        lwr       upr     p adj
## 2-1  0.41666667 -0.1095573 0.9428906 0.1519088
## 3-1  0.40000000 -0.1262240 0.9262240 0.1772595
## 4-1  0.28000000 -0.2719084 0.8319084 0.4989999
## 3-2 -0.01666667 -0.5428906 0.5095573 0.9997373
## 4-2 -0.13666667 -0.6885750 0.4152417 0.8972352
## 4-3 -0.12000000 -0.6719084 0.4319084 0.9272019
```

```
plot(TukeyHSD(anova),col=c("green","red"))
```



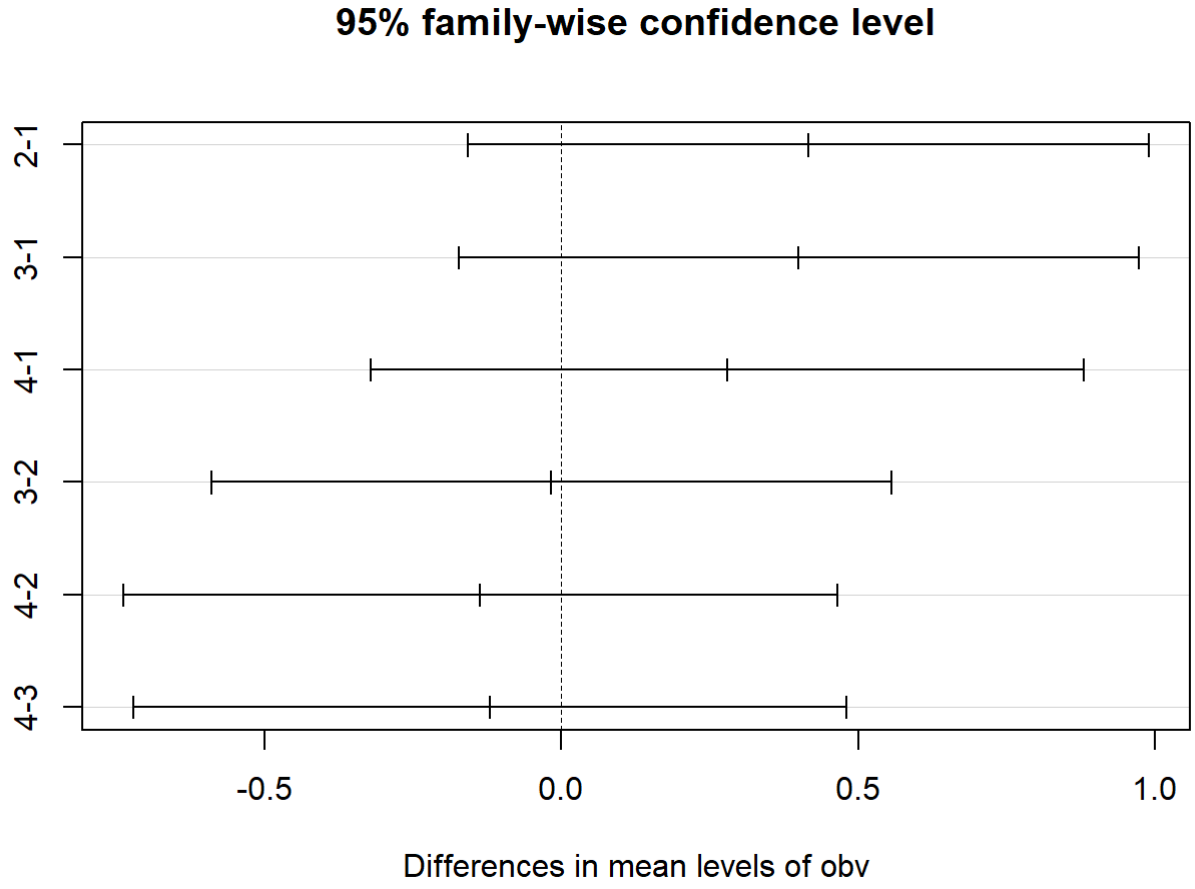

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 3.6.3
```

```
ScheffeTest(anova)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $obv
##             diff     lwr.ci    upr.ci   pval
## 2-1  0.41666667 -0.1565621 0.9898954 0.2109
## 3-1  0.40000000 -0.1732288 0.9732288 0.2409
## 4-1  0.28000000 -0.3212074 0.8812074 0.5760
## 3-2 -0.01666667 -0.5898954 0.5565621 0.9998
## 4-2 -0.13666667 -0.7378741 0.4645407 0.9209
## 4-3 -0.12000000 -0.7212074 0.4812074 0.9444
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(ScheffeTest(anova))
```



Samozrejme, kedze sme nezamietli H0 pri ANOVA, tak rozdiely nie su. Rovnakym postupom uloha 2 a 3.

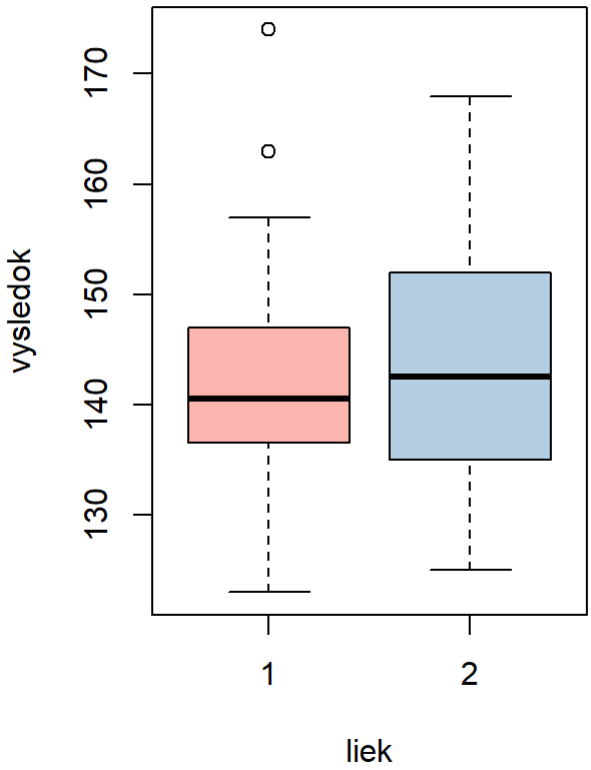
```
##############################################
```

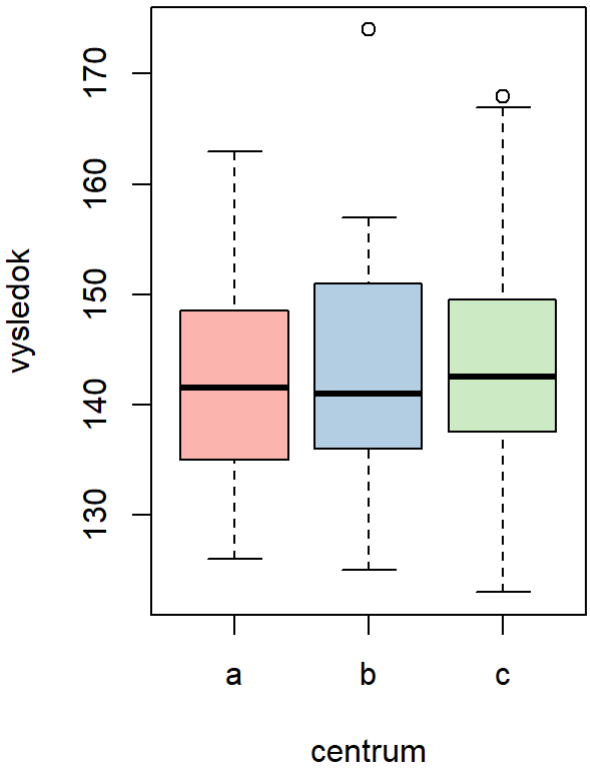
Dvojfaktorova nalyza rozptylu Iba priklad 6, lebo tam testujeme aj kombinacie faktorov

```
centrum <-rep(c("a","b","c"),times=c(16,16,16))
liek <- rep(c(rep(1:2,each=8)),times=3)
vysledok <- c(138,126,141,151,163,139,146,144,126,132,163,145,142,159,130,139,
              135,174,157,136,137,140,136,144,143,142,125,155,149,153,137,133,
              137,123,124,152,138,144,148,141,151,142,168,167,143,139,147,131)
data <- data.frame(centrum,liek,vysledok)
```

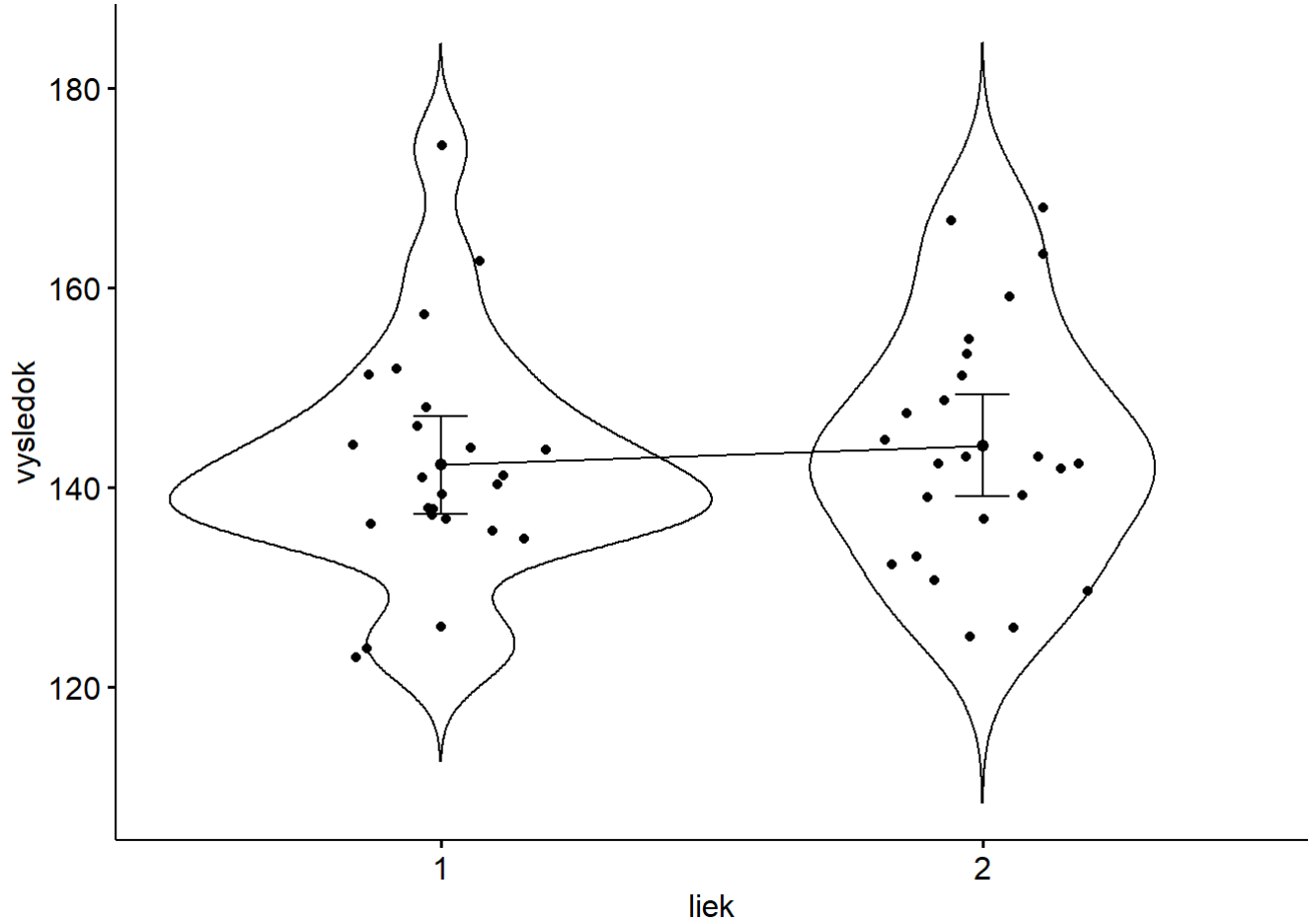
najprv faktorizujeme

```
centrum <- factor(data$centrum)
liek <- factor(data$liek)
par(mfrow=c(1,2))
boxplot(vysledok~liek,col=brewer.pal(3,"Pastel1"))
boxplot(vysledok~centrum,col=brewer.pal(3,"Pastel1"))
```

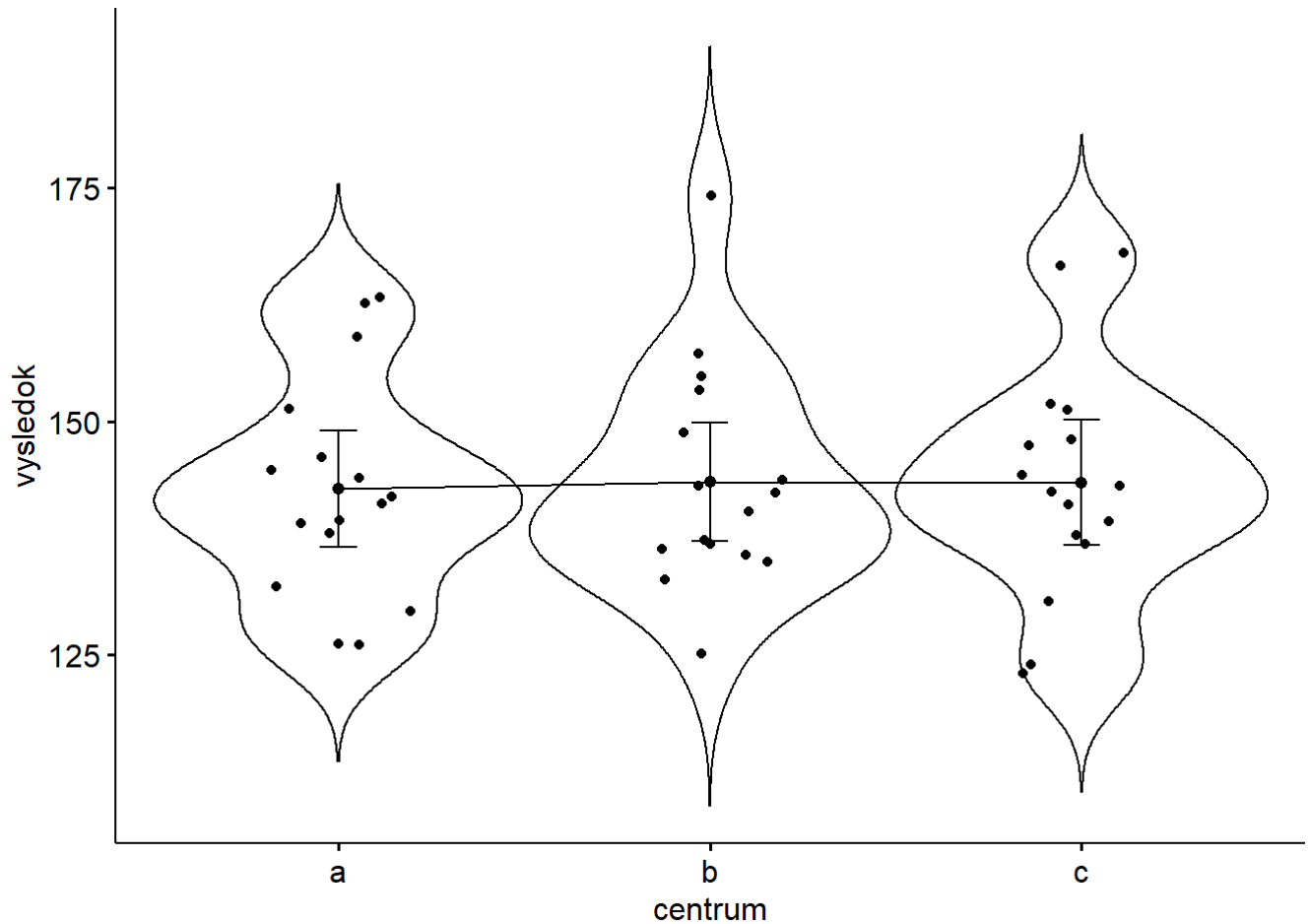





```
pomoc <- data.frame(liek,vysledok)
pomoc1 <- data.frame(centrum,vysledok)
ggline(pomoc,x="liek",y="vysledok",
       add=c("mean_ci","violin","jitter"))
```




```
ggline(pomoc1,x="centrum",y="vysledok",
       add=c("mean_ci","violin","jitter"))
```



samotna ANOVA

```
an1 <- aov(vysledok~centrum+liek)
summary(an1)
```

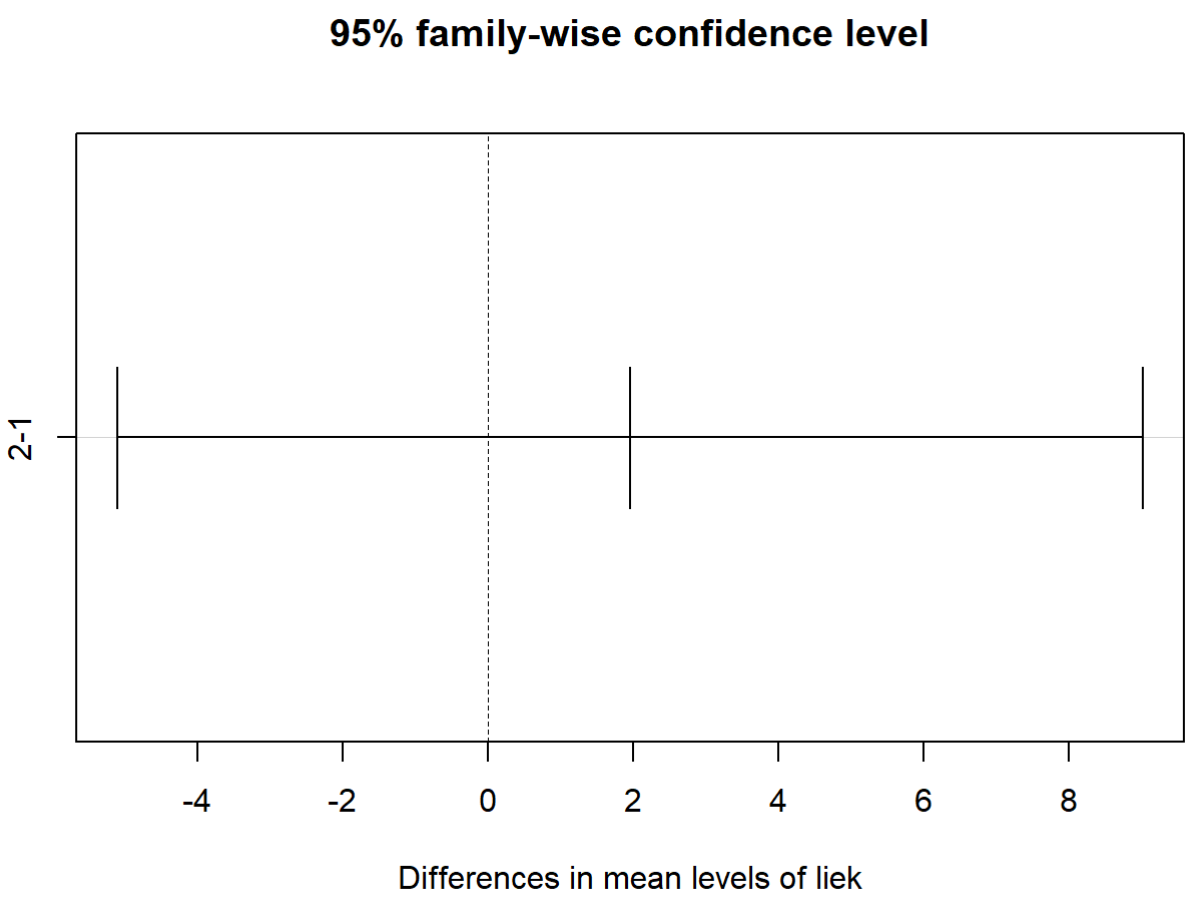
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## centrum      2      6    2.77   0.019  0.981
## liek         1     46   46.02   0.312  0.579
## Residuals   44   6491  147.52
```

nezamietame hypotezy o nulovosti oboch faktorov, nema zmysel robit post testy, iba demonstracne Post testy

```
TukeyHSD(an1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = vysledok ~ centrum + liek)
##
## $centrum
##        diff        lwr      upr     p adj
## b-a  0.7500  -9.665491 11.16549 0.9833302
## c-a  0.6875  -9.727991 11.10299 0.9859730
## c-b -0.0625 -10.477991 10.35299 0.9998832
##
## $liek
##         diff       lwr      upr     p adj
## 2-1 1.958333 -5.107938 9.024605 0.5793104
```

```
plot(TukeyHSD(an1))
```




```
ScheffeTest(an1)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $centrum
##        diff   lwr.ci  upr.ci   pval
## b-a  0.7500 -11.7323 13.2323 0.9986
## c-a  0.6875 -11.7948 13.1698 0.9989
## c-b -0.0625 -12.5448 12.4198 1.0000
##
## $liek
##         diff   lwr.ci   upr.ci   pval
## 2-1 1.958333 -8.23342 12.15009 0.9573
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

teraz aj s interakciami

```
an2 <- aov(vysledok~liek+centrum+liek*centrum)
summary(an2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## liek          1     46   46.02   0.318  0.576
## centrum       2      6    2.77   0.019  0.981
## liek:centrum  2    403  201.65   1.391  0.260
## Residuals    42   6088  144.94
```

p hodnota pre interakcie>0.05, teda interakcie nemaju vplyv na vysledky merani, vplyv je nulovy