Zadanie testy dobrej zhody, vypracované

- Príklad 1
- Príklad 2
- Príklad 3
- Príklad 4
- Príklad 5
- Príklad 6
- Príklad 7
- Príklad 8
- Príklad 9

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.

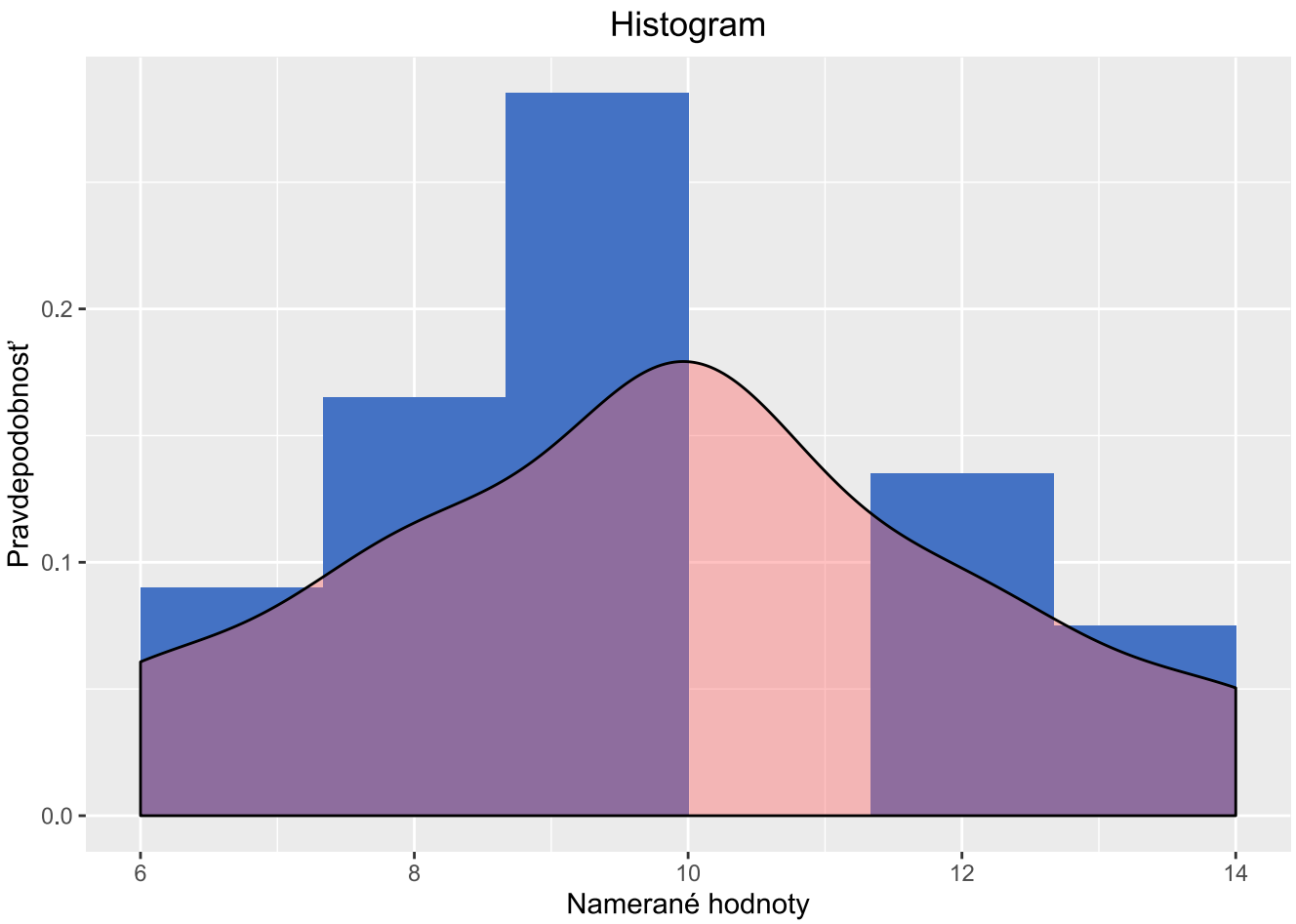
H_0 : Výber pochádza z normálneho rozdelenia.

H_1 : Výber nepochádza z normálneho rozdelenia.

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

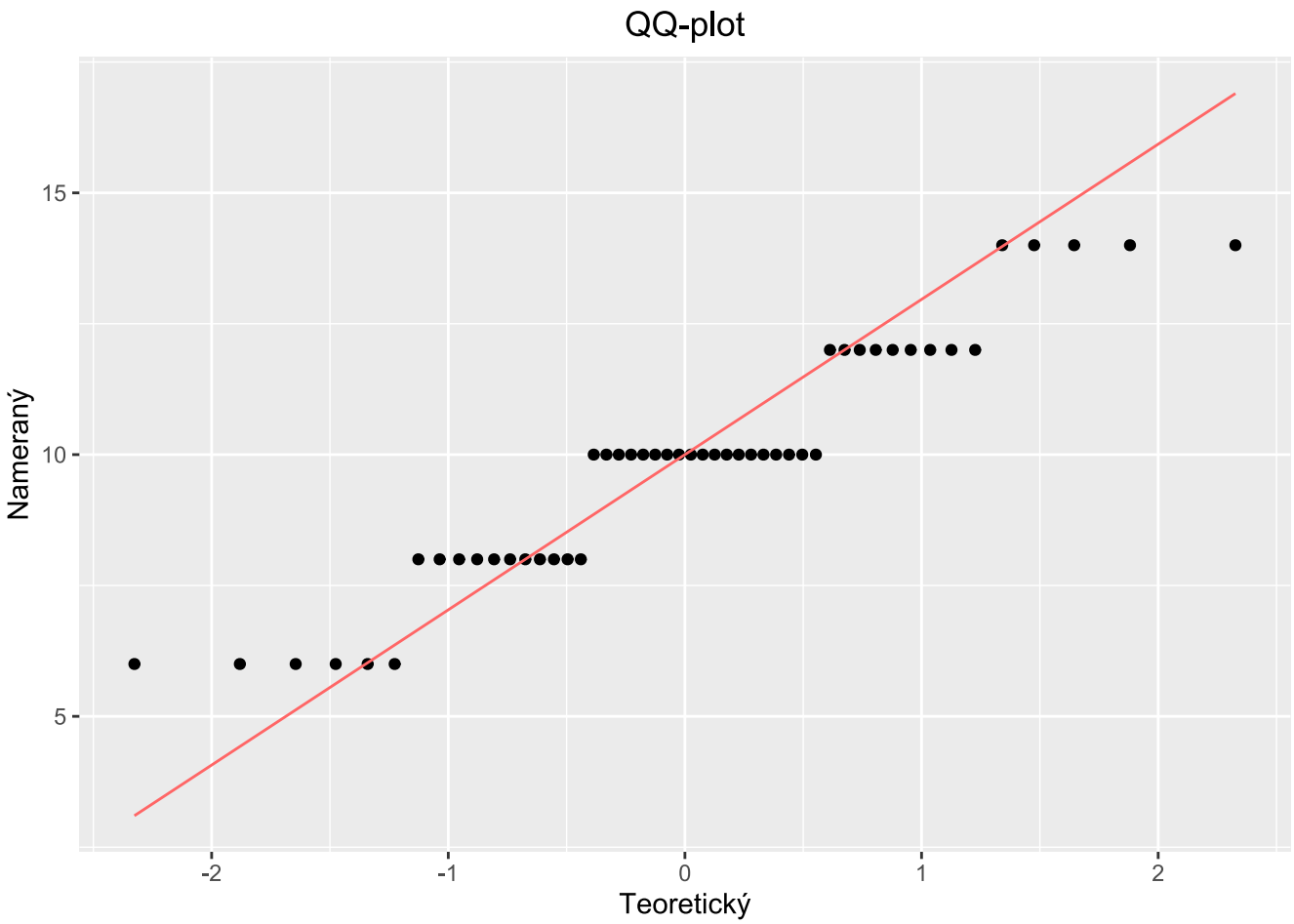
```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
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))
```



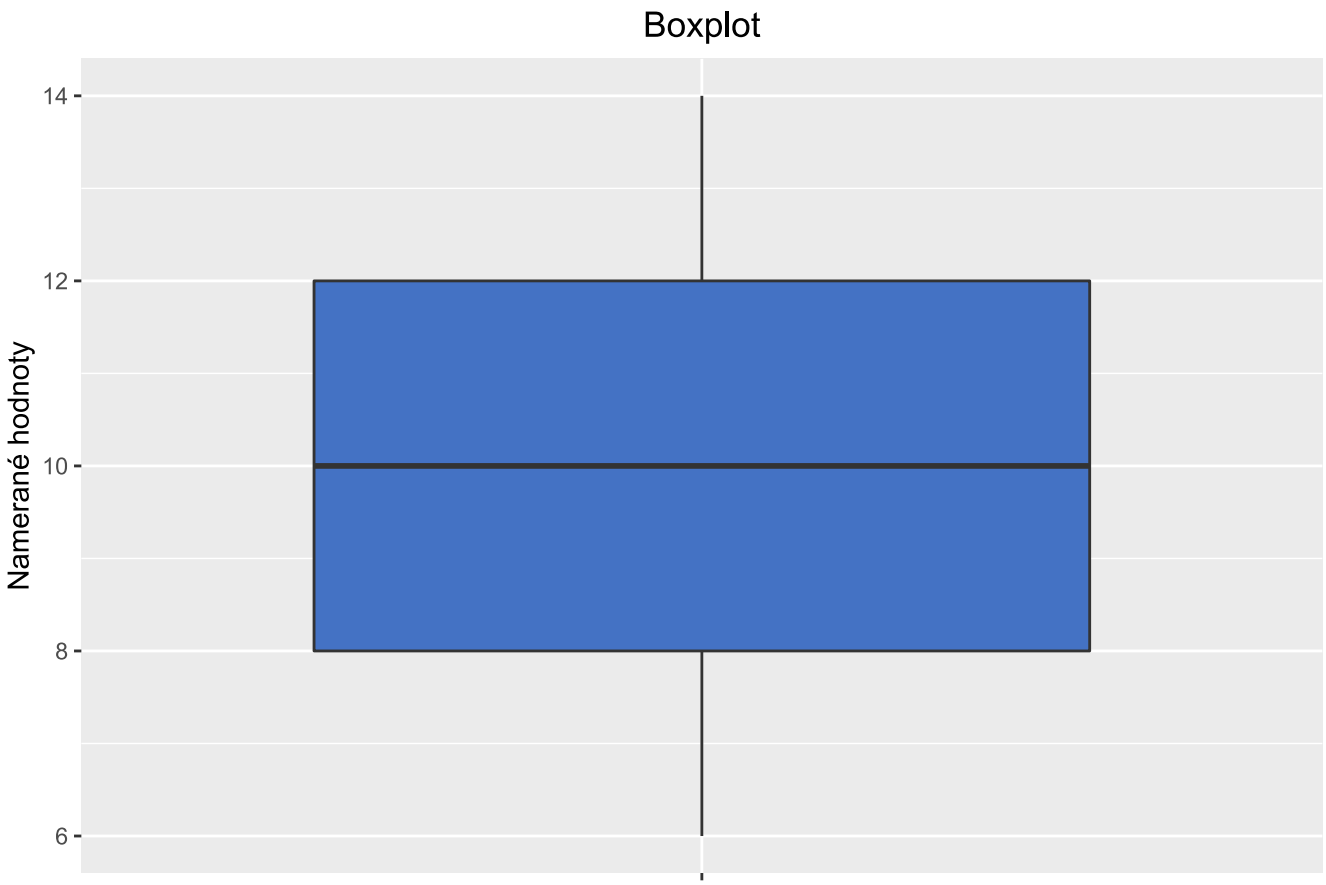
Z histogramu, ktorý je "prerušený" blízko strednej hodnoty, aj keď okrem toho je celkom symetrický, súdime, že dáta nebudú normálne rozdelené.

```
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

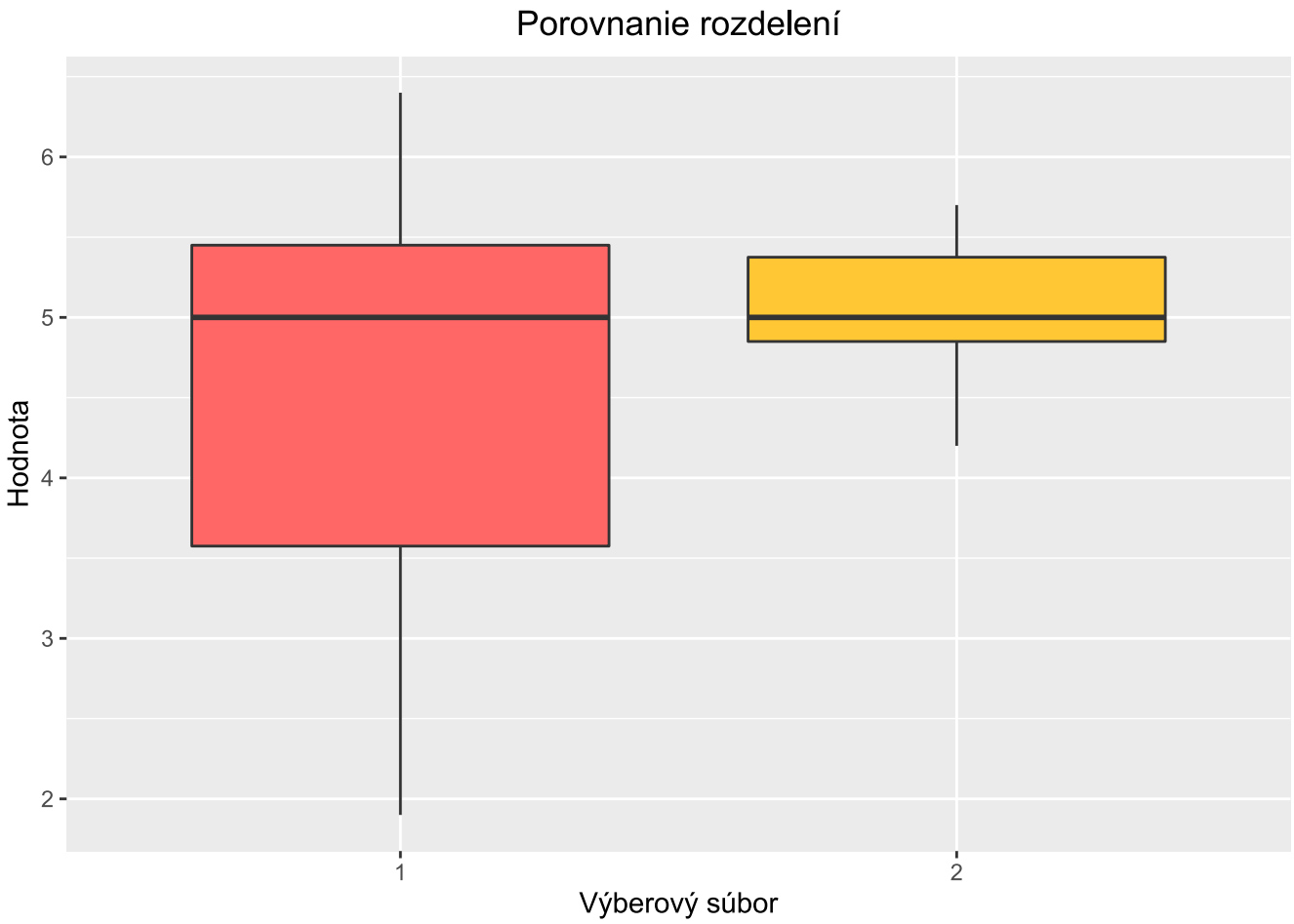
H_1: F≠G

```
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
```

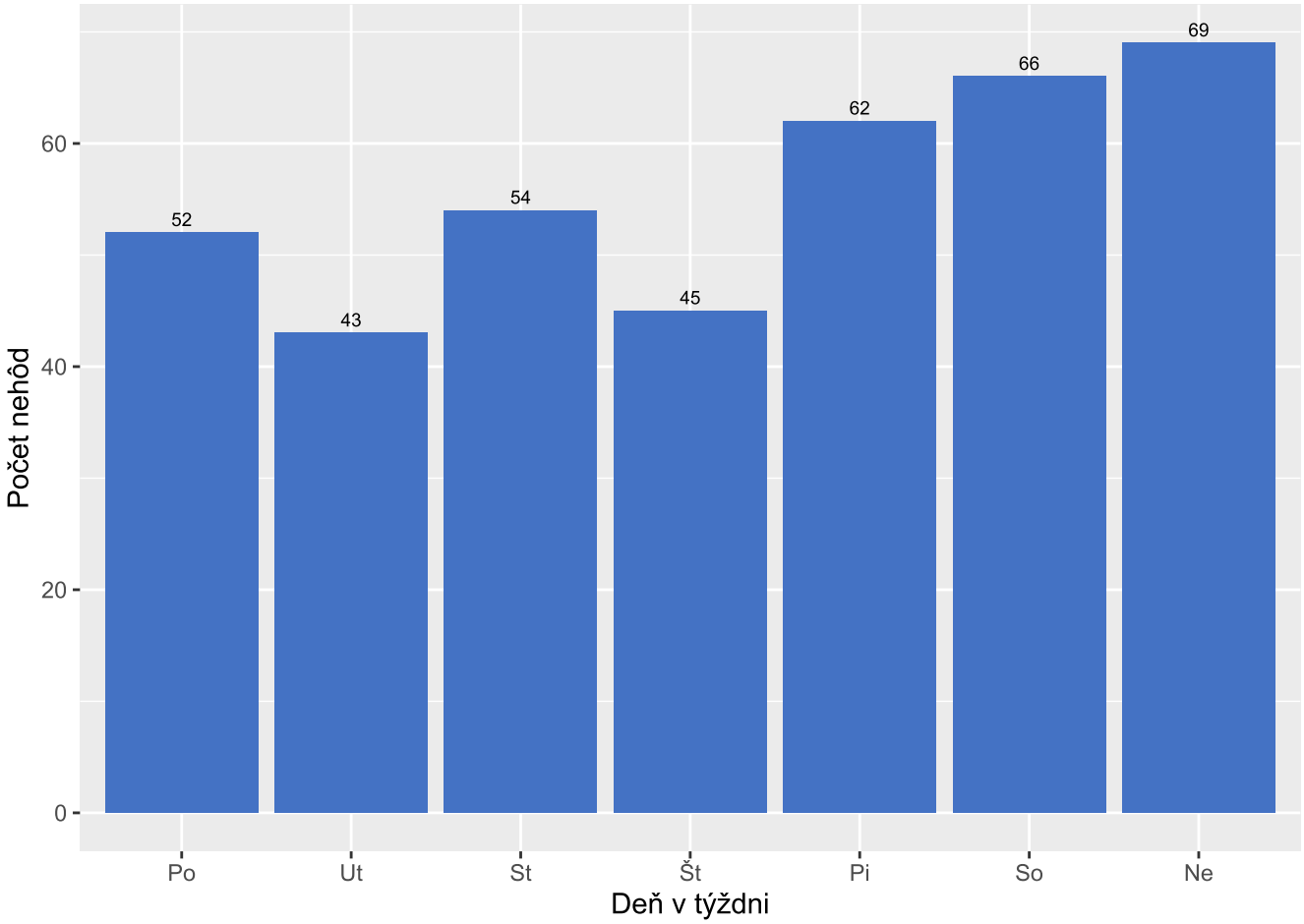
```
prob <- rep(1/7, 7)

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

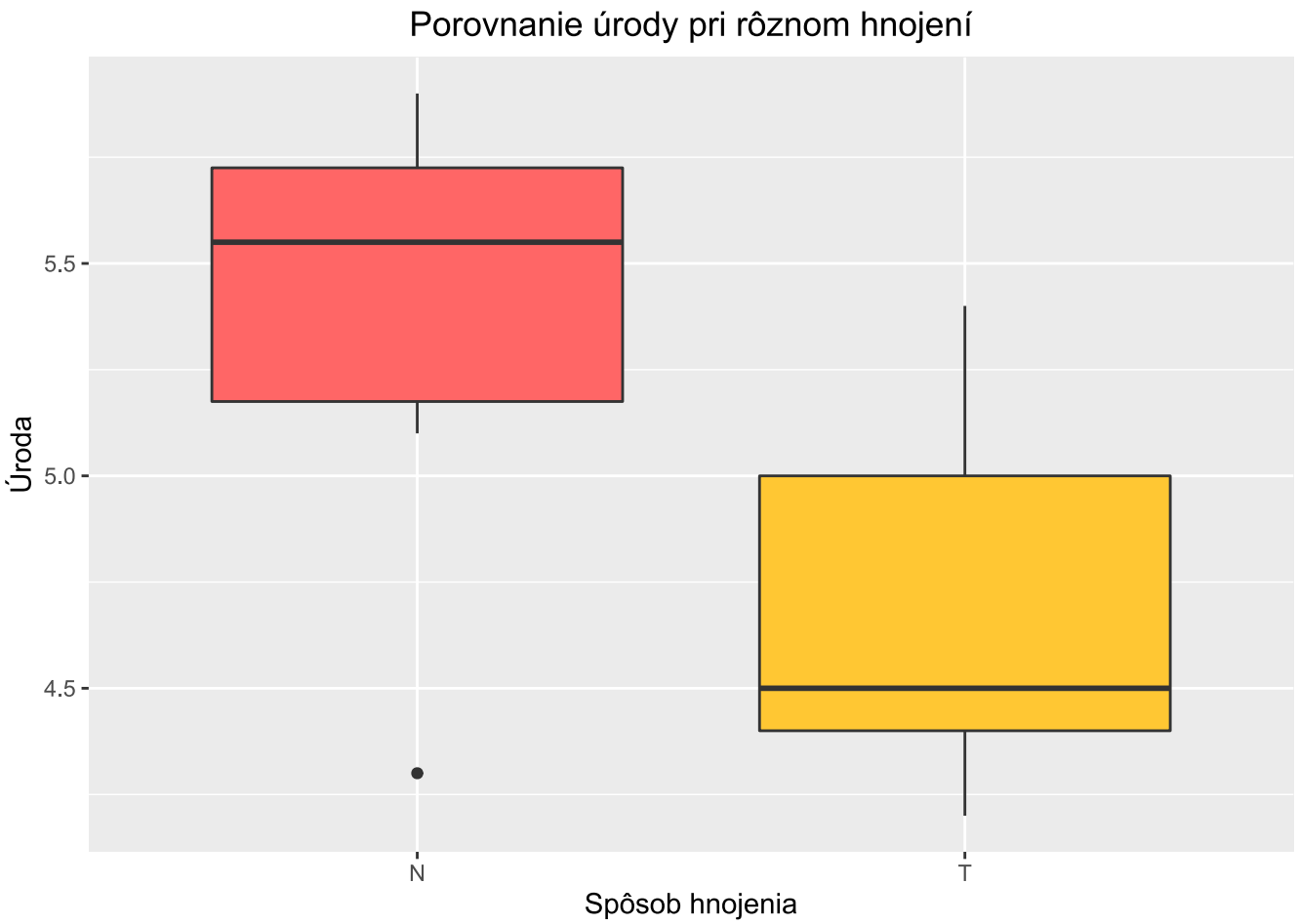
H_1: F≠G

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)
```

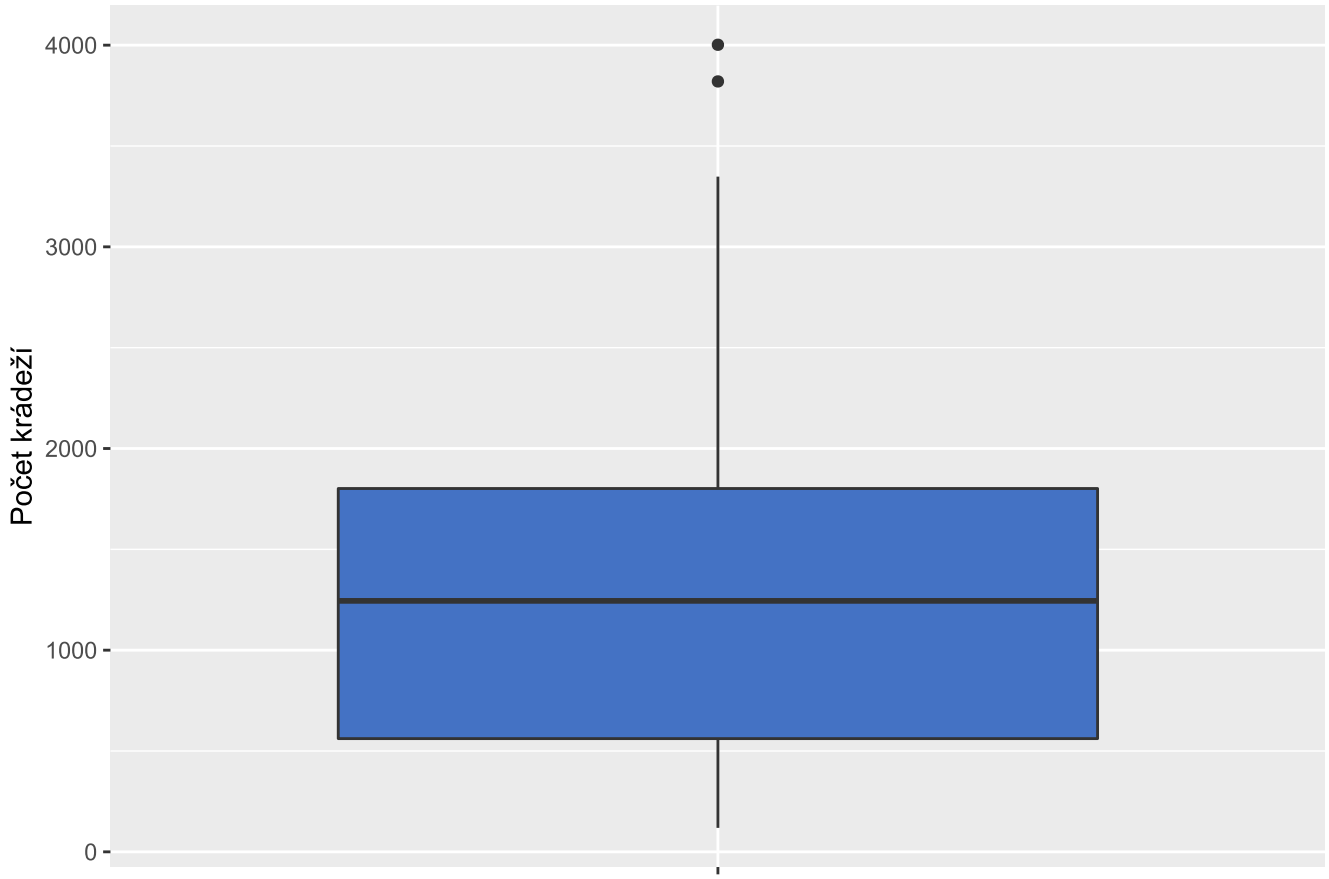
```
## Warning: package 'readxl' was built under R version 3.5.3
```

```
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"
```