Katedra aplikovanej matematiky a štatistiky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave

Vybrané kapitoly z počítačovej štatistiky I

Základy matematickej štatistiky a jej aplikácie použitím programovacích jazykov R a S

elektronické študijné materiály

Stanislav Katina

katina@fmph.uniba.sk

18/01/2006

Podporené grantami VEGA 1/0272/03 a 2/3203/24
Recenzenti: Doc. RNDr. František Štulajter, CSc., Mgr. Ján Somorčík

Obsah

1 Úvodné poznámky

Základné informácie o **R-ku a S-PLUS**:

- `http://www.defm.fmph.uniba.sk/~katina/katina.htm` (Teaching, Computer Statistics: texty a dáta)
- *štandardný balík*: `http://cran.r-project.org` (cesta: Windows (95 and later) → base → R-2.2.1 -win32.exe)
- *Vanables, W.N., Smith, D.M., 2005:* An Introduction to R. Notes on R: A Programing Environment for Data Analysis and Graphics. Version 2.2.1 (2005-12-20), zdroj: `http://cran.r-project.org` (cesta: Documentation → Manuals)
- *R knižnice*: `http://cran.r-project.org` (cesta: → Packages (podľa potreby))

Základné príkazy:

- **default promt**: `" "`
- **ukončenie v R**: `q()`
- **default promt ak nie je ukončený príkaz** - `''+''`
- **nepoužíva sa diakritika !!!**
- **komentár**: *hashmark* `''#''`, všetko do konca riadka je komentár
- **história príkazov:** šípky hore ↑ a dole ↓
- **separácia slov:** bodka `''.''` (napr. pri priraďovaní názvov ku objektom)
- **objekty:** sú ukladané podľa mien - zobrazenie mien existujúcich objektov `objects()` alebo `ls()`; súbor objektov sa nazýva *workspace* a R ho zapisuje do súboru `''.RData''`, odstránenie objektov `rm()`

  pozn.: treba priraďovať mená podľa nasledovných pravidiel - vektory malými písmenami, matice veľkými písmenami, objektom dávať krátke zrozumiteľné mená, radšej začať veľkým písmenom, aby sa neprepísal názov nejakej funkcie
- **separácia príkazov:** bodkočiarka `'';''` alebo nový riadok (*enter*)
- **odstránenie písaného textu**: celý riadok - šípky hore ↑ a dole ↓, časť textu - *dell, backspace*
- **funkcia a jej argumenty**: `nazov.funkcie(''character.string'')`, `args(''nazov.funkcie'')`
- **príklad a demo**: `example(topic)`, `demo(topic)` v R, `example.topic(topic)` v S-PLUS
- **help** - `help(name)`
- **základné funkcie a operácie:**

  - `''+''`, `''-''`, `''*''`, `''/''`

  - `''<''`, `''>''`, `''<=''`, `''>=''`, `''==''`... rovnosť , `''!=''`... nerovnosť

  - `sqrt(x)`, `''^''`, `abs(x)`

  Trigonometrické funkcie a ich inverzie:

- `sin(x), cos(x), tan(x), asin(x), acos(x), atan(x)`

Hyperbolické funkcie a ich inverzie

- `sinh(x), cosh(x), tanh(x), asinh(x), acosh(x), atanh(x)`

Exponenciálne a logaritmické funkcie

- `exp(`*`number`*`)`

- prirodzený (natural) logaritmus (základ e, ln) `log(`*`number`*`, base=`*`number`*`)`, dekadický (common) logaritmus `log10(`*`number`*`)`

- `log(100, 10) == log10(100)`

Gama funkcia a logaritmus gama funkcie `gamma(), lgamma()`...$\log_e \Gamma(x)$

- zaokrúhľovanie `round(`*`number,digits = pocet.miest`*`)`

- `sign(x)`vracia znamienka čísel v podobe 1, 0 alebo -1, ak bočíslo kladné, nula alebo záporné

- *a súčasne (AND)* `''c1 & c2''`

- *alebo (OR)* "`c1 | c2''`

- *negácia (NOT)* "`! c1''`

- `union` $A \cup B$, `intersect` $A \cap B$, `setdiff` $A \cap \overline{B}$ a `is.element` $x \in A$

- **základné príkazy pri čítaní dát:**

  - načítanie vložených dát `data(`*`data.frame`*`)`

  - priame čítanie stĺpcov `attach(`*`data.frame`*`)` alebo `nazov.dataframe$nazov.stlpca`

  - názvy stĺpcov `names(`*`data.frame`*`)`

  - ukončenie priameho čítania stĺpcov `detach(`*`data.frame`*`)`

  - `x` $<-$ `scan()`

  # pokiaľ máme málo dát, enter a vkladáme čísla s medzerou medzi nimi

ďalšie spôsoby čítania dát - z knižnice v Rku
`data(); data.frame()`
`data(package=''`*`name`*`'')`
`data(`*`name`*`,package=''`*`name`*`'')`
`library(name); data(); data(`*`name`*`)`
editovanie dát
`xnew_edit(`*`xold`*`)`
`xnew_edit(`*`data.fram()`*`)`

- **vloženie dát**: v S-PLUSe (pozri v helpe `importData`), v Rku `read.table(''`cesta a názov súboru`'', header=T)`

nikdy nepoužívame separátor v ceste "\", ale "\\" alebo "/"
voliteľné parametre
separáror `sep=''''`, napr. `sep=''/t''` pre tabuľku, `sep=''''` pre voľný priestor (default), alebo `sep = '','', sep = '';''`
hlavička `header=T`, bez hlavičky `header=F`
desatinná bodka alebo čiarka `dec = ''.''` alebo `dec = '','`
názvy riadkov a stĺpcov použitím `row.names=..., col.names=...`
`NAZOV<` $-$`read.table(''D:/Dokumenty/nazov.txt'',header=T)`

- **export grafu a výsledkov v tabuľke**

  # aktívne okno (GSD2)

  `export.graph(''D:\\graph.wmf'', Name=''GSD2'', ExportType = ''WMF'')`

  `exportData(meno, ''D:\\data.xls'', type=''EXCEL''`

- **desatinná bodka**: `''.''`
- **priraďovanie:** `''<-''` alebo `''_''`
- **inštalácia knižnice v R**
- **volanie knižnice** `library(`*`name`*`)`, `library(`*`help = section`*`)`

Príklad 1 *príklady na základné funkcie a operácie*

```
sqrt(-5)
sqrt(-5+0i))
```

Príklad 2 *`help(sin), example(sin), args(''sin'')`*

Príklad 3 *`data(iris), attach(iris), names(iris), detach(iris)`*

Príklad 4 *`data(geyser), attach(geyser), names(geyser)`*

Príklad 5 *načítanie dát do S-PLUS a R*

1.1 Objekty a ich typy

Objekty:

- *vektor* `vector`
- *matica* `matrix`
- *pole* `array`
- *faktor* `factor` - pre kategoriálne dáta
- *množina objektov* `list`
- *dátový rámec* `data.frame`
- *funkcia* `function`

Triedy objektov, `class(`*`object`*`)`:

- `numeric` - reálne čísla
- `integer` - celé čísla
- `complex` - komplexné čísla
- `logical` - logický, hodnoty `TRUE` alebo `FALSE`
- `character, factor` - text a premenná

2 Vektory

- **vektor**: `x_c(cislo,...,cislo)`
- **dĺžka vektora**: `length(x)`
- **súčet a násobenie členov vektora:** `sum(x), prod(x)`
- **zoradenie členov vektora (vzostupne a zostupne):** `sort(x); rev(sort(x))`
- **priradenie poradia členom vektora**: `order(x)`
- **operácie s vektormi**

Príklad 6 *atribúty vektora a ich zmeny*

```
z_0:9
charz_as.character(z)
intz_as.integer(charz)

# vytvorte numerický vektor e s dĺžkou 3 tak, aby prve dva komponenty boli NA a tretí číslo 17
# typ pridávaných komponentov súhlasí s typom predhádzajúcich komponentov
e[3]_17
length(e)
# zmena atribútu "dimenzia" = vytvorte z vektora z maticu s rozmermi 2 × 10
dim(z)_c(2,10)
```

Príklad 7 *operácie s vektormi*

```
x_c(1,3,5)
length(x)
sum(x)
prod(x)
sort(x)
rev(sort(x))
order(x)
z_c(x,2,x)
u_2*x + z + 1
v_x-1
y_1/x
```

- **jednoduchá sekvencia vpred a vzad** `c(1:n), c(n:1)`
- **zložitejšia sekvencia** `e_seq(from=value, to=value, by=value, length=value)`
- **opakované sekvencie** `g_rep(x,times=value)`

Príklad 8 *jednoduchá sekvencia vpred a vzad, zložitejšia sekvencia a opakované sekvencie*

```
n_5
c(1:n)
a_c(1:n-1)
b_c(1:(n-1))
c(n:1)
d_c(10:1)
e2_seq(from=0, to=1, by=0.01)
e1_seq(from=0, to=1, length=10)
# zopakujte trojku 6x
```

Logické vektory:

- možné hodnoty: `TRUE (T,1), FALSE (F,0), NA`
- sú generované použitím tzv. *conditions* (podmienky)

Chýbajúce hodnoty, nekonečno, čísla a 'nečísla':

- "not available", ("missing values", `NA`)
- `is.na(x)` označenie, ktoré dáva logický vektor dĺžky ako `x` s hodnotami `TRUE` ak prislúchajúca hodnota na danom mieste je `NA` v `x`
- **±nekonečno:** "`Inf`", resp. "`-Inf`"
- "number"
- "not a number" NaN `''0/0''`
- `is.na(xx)` je `TRUE` pre `NA` a aj `NaN` hodnoty
- `is.nan(xx)` je `TRUE` len pre NaN

Príklad 9 *logické vektory, chýbajúce hodnoty, nekonečno, čísla a 'nečísla'*

```
x <- 1:20
x.do15_x > 15
p_c(1:5,NA)
mv.p_is.na(p)
bezmv.p _ p[!is.na(p)]
# nekonečno
1/0
x_c(-1,0,1)/0
x
is.na(x)
x > Inf
nek_c(1:5,Inf,-Inf)
```

Vektory premenných:

- `c(''x1'',''x2'')`
- `paste(vektor,sep=`*`string`*`)`

Indexované vektory: `vector[`*`index.vector`*`]`

- **odstraňovanie zložiek:** `x[-CISLO], x[- RETAZEC]`
- **logické indexovanie**: `x[!is.na(x)]`
  # vyradenie chýbajúcich hodnôt vo vektore

Príklad 10 *vektory premenných, indexovanie*

```
fruit_c(4,8,2,9)
names(fruit)_c(''banana'',''orange'',''apple'',''peach'')
```
vybratie len tých položiek vo `fruit`, ktoré zodpovedajú prvým dvom názvom, teda `''banana'', ''orange''`

`c(''X'',''Y'')` sú opakovane 5x v sekvencii za sebou a postupne sú knim priradované čísla `1:10`, teda `c(''X1'',''Y2'',...,''X9'',''Y10'')`
vytvorte sekvenciu X1 X2 ... X50

```
x[3]
x[1:10]
x1_x[-5]
x2_x[-(1:3)]
p1_p[!is.na(p)]
```

nahradnie chýbajúcich hodnôt v `x` nulami, teda `x[is.na(x)]_0`
```
s_c(-1,-5, 1:7,NA)
```
vytvorte objekt `f`, kde vyraďte vo vektore `x+1` chýbajúce a kladné hodnoty
vytvorte vektor `zz1` absolútnych hodnôt nejakého vektora `zz` zložený z nejakých celých čísel, potom vyberte zo `zz` len záporné členy priraďte ich do vektora `zz2` a nakoniec vytvorte opačný vektor ku vektoru `zz` a označte ho `zz3`
čo tak čosi zložitejšie, sekvenciu "XYYX" 4x

Príklad 11 *vektory*

vytvorte vektor $x = (1, 2, 5, 6, 9)$
vytvorte vektor $y = (3, 1, 2, 5, 6, 9, 10, 1, 2, 5, 6, 9, 2)$
vytvorte vektor z, ktorý obsahuje len 3. až 7. prvok vektora y
vytvorte vektor z, ktorý je odvodený z vektora y tak, že ho tvoria len členy prislúchajúce poradiu odvodenému z členov vektora x
vložte na tretie, šieste a ôsme miesto vektora z číslo 7
vložte do vektora z na miesta zodpovedajúce druhému až šiestemu členu vektora y čísla $11, 12, 13, 14, 15$
vypíšte vektor z

Príklad 12 *funkcia* ***rep***

vytvorte vektor $a = (1, 2, 3, 4)$
vytvorte vektor $b = (2, 2, 2, 2)$
vytvorte vektor $c = (1, 2, 3, 4, 1, 2, 3, 4)$
vytvorte vektor $d = (1, 1, 2, 2, 3, 3, 4, 4)$
vytvorte vektor $e = (1, 2, 2, 3, 3, 3, 4, 4, 4, 4)$

Príklad 13 *rep - ďalšie*

8 x jedna, 8 x dva, 8 x tri
vytvorte 3x za sebou sekvenciu čísel 112233

3 Zoradené a nezoradené premenné

Faktor je vektorový objekt používaný na špecifikáciu diskrétnej klasifikácie komponentov iných vektorov rovnakej dĺžky. Funkcia `ordered()` zoraďuje faktory podľa abecedy, často je identická ku funkcii `factor()`, ale nie vždy.

Príklad 14 *štáty*

```
   staty_c(''sl'',''cz'',''po'',''ru'', ''ukr'', ''bi'',''sl'',''cz'',''po'',''ru'',
''ukr'', ''bi'')
   statyfak_factor(staty)
   levels(statyfak)
   attr(statyfak,''levels'')
   attr(statyfak,''class'')
   # frekvenčná tabuľka pre premenné definované ako factor (multy-way frequency array)
   table(statyfak)
```

Príklad 15 *príjmy*

```
   prijmy_ordered(c(''Mid'', ''Hi'', ''Lo'', ''Mid'', ''Lo'', ''Hi'', ''Lo''))
   # abecedné zoradenie, teda to, čo nechceme
   prijmy
   # Hi < Lo < Mid
   as.numeric(prijmy)
   pri_ordered(c(''Mid'', ''Hi'', ''Lo'', ''Mid'', ''Lo'', ''Hi'', ''Lo''),
levels = c( ''Lo'', ''Mid'', ''Hi''))
   # to, čo skutočne chceme
   pri
   # Lo < Mid < Hi
```

Príklad 16 *geyser, pozri* `help(geyser)`

```
   attach(geyser)
   names(geyser)
   # waiting časový interval medzi erupciami
   # duration trvanie príslušnej erupcie
   # rozdeľte duration na 6 skupín a potom priraďte čísla 1 až 6 týmto skupinám, teda abz išlo o
zoradené faktory
   erupt_cut(duration,breaks=0:6)
   erupt_ordered(erupt, labels=levels(erupt))
   # erupt ... 0+ thru 1 < 1+ thru 2 < 2+ thru 3 < 3+ thru 4 < 4+ thru 5 < 5+ thru 6
   # intervaly sú typu (i, i + 1⟩, teda 4 minútová erupcia je daná do kategórie 3+ thru 4
   # čo tak iné delenie
   eruptNEW1_cut(duration,4)
   eruptNEW2_cut(duration,quantile(duration, c(0,1/3,2/3,1)), include.lowest=T, labels
= c(''low'', ''mid'', ''high''))
```

3.1 Funkcie "apply", "tapply", "sapply" a "lapply"

- `apply()`, zvolená funkcia sa aplikuje na obe dimenzie matice (riadky alebo stĺpce), alebo jednotlivé dimenzie poľa
- `tapply()`, tu "t" znamená "table" a používa sa na výpočty v tzv. *"ragged arrays",* teda je kombinácia numerického vektora a vektora premenných, podľa ktorých delíme (klasifikujeme) do podskupín numerický vektor, vytvorí sa tzv. *"cross-classified table",* podľa funkcie, ktorú použijeme.

- `lapply()`, tu "l" znamená "list", teda vytvorí sa `list` ako výsledok aplikovania nejakej funkcie
- `sapply()`, tu "s" znamená "simplify", teda vytvorí sa `vector` ako výsledok aplikovania nejakej funkcie

Príklad 17 *štáty*

```
prijem_c(10 000, 13 000, 12 500, 9 400, 9 100, 9 300, 10 100, 13 100, 12 600, 9 500,
9 200, 9 400)
prijemMean_tapply(prijem, statyfak, mean)
prijemMean
```

pozn.:

```
# v Rku
library(MASS)
library(help=MASS)
data(quine)
tapply(Days,Age,mean)
tapply(Days,list(Sex,Age),mean)
```

Príklad 18 *štáty opäť*

```
# rozdelenie príjmu na pravidelné skupiny podľa presne stanoveného kritéria
prijemfak_cut(prijem,breaks=9000+1000*(0:6))
# frekvenčná tabuľka
table(prijemfak,statyfak)
```

Príklad 19 *iris (pole* $50 \times 4 \times 3$ *)*

```
# priemer podľa stĺpcov
apply(iris, 2, mean)
# priemer podľa druhej a zároveň tretej dimenzie poľa s odseknutými okrajmi (10%)
apply(iris, c(2,3), mean,trim=0.1)
```

Príklad 20 *ISwR knižnica pre Rko, inštalovanie knižnice, práca s chýbajúcimi pozorovaniami, data* `thuesan`

```
library(ISwR)
library(help=ISwR)
data(tuesan)
help(tuesan)
lapply(thuesen, mean, na.rm=T)
sapply(thuesen, mean, na.rm=T)
```

4 Matice

Príkazy súvisiace s maticami:

- **spájanie vektorov do matice**: `rbind(...)` ... horizontálne, po riadkoch, `cbind(...)` ... vertikálne, po stĺpcoch
- **vytvorenie matice**: `matrix(name, nrow=...,ncol=...)`
- **vytvorenie matice inak**: `matrix(name, dim=c(...,...))`
- **priradenie názvov riadkom:** `row.names(...)_c(''...'',...,''...'')`
- **vypisovanie:** `meno[,...]`, `meno[...,]`, `meno[...,...]`, `meno[,...:...]`, `meno[...:...,]`
- **priradenie názvov riadkom a stĺpcom**: `dimnames(...)_list(c(`*row*`),c(`*column*`))`
- **priradenie názvov stĺpcom**: `dimnames(...)_list(NULL,c(`*column*`))`

Príklad 21 *matice*

```
# vytvorte maticu pozostávajúcu z prvkov 1 až 12 nasledovne
# po stĺpcoch bycol=T (default)
# po riadkoch byrow=T
# priraďte mená riadkom a aj stĺpcom
# ešte inak
matrix(1:12,nrow=3)
```

Príklad 22 *štáty znova,* `dim` *atribút, voľby* `bycol` *a* `byrow`

```
names(staty)_statyfak
names(staty)_NULL
dim(staty)_c(2,6)
staty
dim(staty)_NULL
# bycol=T (default)
S1_matrix(staty,2,6)
S2_matrix(staty,2,6,byrow=T)
```

Príklad 23 *geyser opäť, funkcia* `cbind`

```
cbind(waiting,erupt)
```

Príklad 24 *dolná trojuholníková matica (*`lower.tri, upper.tri`*)*

```
A_matrix(1:16,dim=c(4,4))
Aut_A[col(A)>= row(A)]
Alt_A[col(A)<= row(A)]
```

4.1 Polia

Príklad 25 *práca s dimenziami, parameter a funkcia* `dim()`

```
# pole 2x2x5 z čísel 1 až 20
# pole 4x5 z čísel 1 až 20
# čosi naviac
Z_array(c(1:10,11:20), dim=c(2,10))
I_array(c(1:3,3:1), dim=c(3,2))
X[5]
X[5]_0
```

Príklad 26 *pole ešte raz*

```
# pole 2x2x5 z čísel 1 až 20 inak
# pole núl 2x2x5
```

4.2 Objekt "list"

List - objekt pozostávajúci zo zoradeneného zoznamu objektov, teda z komponentov

- použitie napr. pri programovaní funkcií a hľadaní komponentov výsledkov nejakej funkcie
- `List_list(name=''Fred'', wife=''Mary'', no.children=''3'', child.age=c(4,7,9))`
  `List_list(name.1=object.1, ..., name.m=object.m)`
- komponenty sú očíslované: `List[[1]], List[[2]],...,`
- 4. komponent a jeho prvá položka: `List[[4]][1]`
- počet hierarchicky najvyšších komponentov: `length(List)`
- pomenovania - `name$component.name` (`List$wife` je ako `List[[2]], List[[''wife'']]` je `''Mary''); x_''wife''; List[[x]]`
- pozor na hierarchiu `''[[]]''` a potom `''[]''`
- ak objekt pozostáva z jednotlivých listov, tak aj celok je list `List.ABC_c(List.A, List.B, List.C)`

4.3 Dátové rámce ''data.frame''

- už bolo: `attach(name.dataframe), detach(name.dataframe)`
- *indexácia*
  name`[1:3, c(1,5,7)]`
- konverzia dvoch rámcov z a do matice
  `as.data.frame()`
  `as.martix()`

pozri aj `data.matrix()` a `column.levels`

Príklad 27 *podmnožiny, funkcia* `subset()`

```
# v Rku
library(ISwR)
library(help=ISwR)
data(tuesan)
help(tuesan)
thuesanNEW <- subset(thuesan, blood.glucose < 7)
# v S-PLUS
help(painters)
attach(painters)
painters[Colour >= 17, ]
painters[Colour >= 15 & Composition > 10,]
painters[Colour >= 17 & School != ''D'', ]
painters[is.element(Shool, c(''A'',''B'',''D'')), ]
# v R a S-PLUS 6.1 a 6.2
painters[School %in% c(''A'',''B'',''D'')), ]
```

Príklad 28 *podmnožiny, funkcia* `transform(), scale()`

```
# pokračujeme s dátami tuesan
thuesanNEW1 _ transform(thuesan, log.gluc = log(blood.glucose))

# pokračujeme s dátami crabs
crabsNEW _ crabs
# centrovanie a preškálovanie
crabsNEW[, 4:8] _ lapply(crabsNEW[, 4:8], scale)
# centrovanie
crabsNEW[, 4:8] _ lapply(crabsNEW[, 4:8], scale, scale=F)
# ktoré premenné sú numerické?
sapply(crabs,is.numeric)
crabsNEW[ ] _ lapply(crabsNEW, function(x){if(is.numeric(x)) scale(x) else x})
```

Príklad 29 *chýbajúce pozorovania, funkcia* `na.omit()`

```
help(claims)
names(claims)
attach(claims)
claimsNEW _ na.exclude(claims)
claimsNEW _ na.omit(claims)
```

Príklad 30 *funkcia* `aggregate`

- použitie `aggregate` je ekvivalentné `tapply` na každom stĺpci dátového rámca

  ```
  attach(crabs)
  aggregare(crabs[, 4:8], list(sp=crabs$sp, sex=crabs$sex), median)
  ```

Príklad 31 *funkcia* `by()`

- funkcia `by` vyrába dátový rámec a delí ho (`split`) podľa druhého argumentu funkcie, podobné funkcii `tapply`

  ```
  by(crabs[, 4:8], list(crabs$sp, crabs$sex),summary)
  ```

Príklad 32 *funkcia* `merge()`

- funkcia `merge` spája dva dátové rámce ako databázy - kombinuje napr. príslušné riadky

```
help(merge)
# vytvorte dátový rámec authors a potom books v S-PLUS 6.2 podľa stĺpcov 1 a 2, kde sú mená a priezviská
merge(authors, books, by=1:2)
# v S-PLUS 4.5, podľa stĺpca "row.names"
merge(Animals, mamals, by=''row.names'')
```

Príklad 33 *sortovanie a náhodný výber*

```
library(car)
library(help=car)
data(Womenlf)
help(Womenlf)
attach(Womenlf)
# 20 náhodne vybratých riadkov
sample.20 <- sort(sample(nrow(Womenlf),20))
# podľa tohoto náhodného výberu vyberte prislúchajúce riadky
Womenlf[sample.20, ]
```

Príklad 34 *prekódovávanie* `recode()`

```
# pokračujeme s dátami Womenlf
working <- recode(partic,c('partime','fulltime')='yes'; 'notwork'='no'')
working[sample.20]
# alternatívne
working.alt <- recode(partic,c('partime','fulltime')='yes'; else='no'')
# ponechajte len pracujúcich, ostatní budú NA
fulltime <- recode(partic,'fulltime'='yes';'partime'='no';'notwork'= NA'')
fulltime[sample.20]
# prekódujte region
region.4 <- recode(region,''c('Prairie','BC')='West'')
region.4[sample.20]
```

4.4 Operácie - vektory, matice, polia, dátové rámce

- operácie *člen-po-člene* sú ''+'', ''-'', ''*'', ''/''.
- *rozšírené operácie pozri v* - `library(Matrix)`

Príklad 35 *operácie - príklady*

```
# vytvorte maticu z čísel 1 až 9 po stĺpcoch a pripočítajte k nej vektor 1:3 po stĺpcoch
A _ matrix(1:9,ncol=3) + 1:3
# alternatívne
sweep(A ,1,1:3,''+'')
# pripočítaj k matici A vektor 1:3 po riadkoch
matrix(1:9,ncol=3) + t(1:3)
# alternatívne
sweep(matrix(1:9,ncol=3),2,1:3,''+'')
# vynásobte maticu A s maticou A
B_A%*%A
# vynásobte maticu A s maticou B a potom maticu B s maticou A, porovnajte
```

```
C1_A%*%B
C2_B%*%A
```
vynásobte maticu `A` s vektorom `1:3` z ľava a potom vektorom `1:3` transponovaným, porovnajte
```
D1_1:3%*%A
D2_t(1:3)%*%A
```
jednotková matica (identity matrix)
```
diag(4)
```
matica s číslami 1, 2, 3, 4 na diagonále
`diag(1:4)` ... matica s číslami 1, 2, 3, 4 na diagonále
vektor 3, 5, 7 na diagonále
```
x_c(3,5,7); diag(x)
```
A' transpozícia matice
```
At_t(A)
```
A^{-1} inverzia matice
```
As_solve(A)
```

Príklad 36 *Riešenie systému lineárnych rovníc*

```
solve(A, c(9,5,14))
```

Príklad 37 *stopa matice a jej determinant*

stopa (*trace*) štvorcovej matice `tr(A); sum(diag(A))` # ak A je štvorcová matica a $Ax = \lambda x$, kde λ je skalár a x je vektor, potom λ je vlastné číslo (*eigenvalue*) matice A a x je vlastný vektor (*eigenvector*) matice A

`eigen()` ... (`...$values, ...$vectors`)
determinant matice
```
detA_function(x) prod(eigen(x)$values)
```

Príklad 38 *zoradovanie, poradia vo vektoroch, maticiach a dátových rámcoch*

zoraď zostupne 10 štátov USA (`state.name`) podľa oblastí
```
state.name[rev(order(state.x77[,''Area'']))][1:10]
```
vytvorte maticu a potom ju uprav podľa prvého stĺpca a následne podla zoradených prvých dvoch stĺcov
```
MAT <- matrix(c(1, 3, 2, 2, 5, 4, 6, 5, 6, 9, 7, 8, 10, 11, 11, 12, 12, 16, 14,
14), 5, 4)
MAT1 <- MAT[order(MAT[,1]),1:4]
MAT1 <- MAT[order(MAT[,1],MAT[,2]),]
```
priraďte členom vektora ich poradia
```
rank(c(20:1,1:5))
```

5 Základy štatistického uvažovania

V nasledovnom texte budeme pracovať s pojmami, ktoré treba dôsledne rozlišovať, ako

- základný súbor → výberový súbor,
- parameter → odhad,
- populačný priemer → výberový priemer,
- populačná disperzia → výberová disperzia,
- populačná smerodajná odchýlka → výberová smerodajná odchýlka
- populačná štandardná chyba → výberová štandardná chyba,
- pravdepodobnosť → relatívna početnosť.

Štatistický (základný) súbor je konečná množina prvkov (napr. osôb), na ktorej sledujeme určité znaky, veličiny, vlastnosti a pod. Ak uskutočníme výber zo štatistického súboru (pozri nižšie), hovoríme o *výberovom súbore*, ktorý reprezentuje štatistický súbor výberom určitého počtu štatistických jednotiek. *Štatistické jednotky* tvoria napr. respondenti dotazníka, pričom po uskutočnení výberu hovoríme o *výberových jednotkách*. Na štatistických jednotkách skúmame *štatistické znaky* (premenné, faktory), ktoré môžeme rozdeliť do nasledovných kategórií:

1. *kvantitatívne (metrické, kardinálne)* - diskrétne, spojité a intervalové (jav, kt. nevieme presne pozorovať)
2. *kvalitatívne*

   - *nominálne znaky (neusporiadaná kategorizácia, natural ordering)* - ide o príslušnosť sledovaného objektu k určitej triede objektov, napr. vyučovacie predmety; náboženstvá - Katolíci, Protestanti, Židia, Moslimovia; typ sídla - dom, apartmán, kondo

   - *ordinálne znaky (usporiadaná kategorizácia, ordered categories)* - ide o znak, ktorého hodnoty môžeme prirodzene usporiadať, napr. stupnica známok; sociálne skupiny - horná, stredná, nízka; politická filozofia - liberálna, stredná, konzervatívna

   - *alternatívne (binárne)* - typ áno/nie.
3. *frekvencie výskytu nejakej udalosti* - budeme striktne rozlišovať nasledovné:

- *frekvencie (frequencies)* - ak sa udalosti vyskytujú nezávisle na rôznych štatistických jednotkách (*units*) alebo jedincoch (*individuals*),
- *početnosti (counts)* - ak máme rôzne udalosti na tej istej premennej na rovnakej štatistickej jednotke - tu môžeme očakávať určitý typ závislostí medzi opakovanými udalosťami

Výber môže byť

1. *náhodný (pravdepodobnostný)* - vyberanie šatistických jednotiek z populácie celkom náhodne a nezávisle na našom úsudku -

   · *jednoduchý náhodný* - priamy výber štatistických jednotiek, pričom každá má rovnakú pravdepodobnosť, že bude vybraná (napr. losovanie, pričom je výhodné, keď sú štatistické jednotky očíslované a je možné použiť matematickú teóriu náhodných čísel);

   · *mechanický (systematický)* - je založený na určitom, dopredu stanovenom, usporiadaní prvkov populácie - do výberového súboru zaradíme všetky prvky, ktoré sú od seba vzdialené o zvolený výberový krok, kedy prvý prvok vyberieme jednoduchým náhodným výberom; pri tomto výbere musíme dať pozor, aby usporiadanie prvkov nesúviselo so sledovaným znakom;

PRÍKLAD: z abecedne usporiadanej kartotéky (databázy) pacientov u praktického lekára, vyberáme s krokom 10 a prvú kartu vylosujeme (napr. to bude deviata karta) a potom bude výber tvorený pacientami s poradiami kariet 9, 19, 29, 39, 39 atď.;

PRÍKLAD: zisťujeme znak zamestnanie u stomatologického pracoviska, pacienti prichádzajú v určitom časovom slede a ich záznamy sú potom takto aj usporiadané; urobíme výber s krokom rovným dennému počtu pacientov, kedy predpokladáme, že každý deň príde do ambulancie rovnaký počet pacientov; teto postup vedie potom k značne selektívnemu výberu, kedy vyberáme pacientov, ktorí prichádzajú v určitý čas dňa, čo môže súvisieť s typom ich zamestnania;

· *oblastný (stratifikovaný)* - základný súbor je rozdelený na oblasti podľa určitého hľadiska, sú vytvorené tak, aby boli vnútri homogénne (v sledovaných znakoch sa príliž neodlišujú) a medzi sebou heterogénne (v sledovaných znakoch sa značne odlišujú); v jednotlivých oblastiach sa uskutoční jednoduchý náhodný výber alebo mechanický výber; percento vybraných jednotiek môže byť buď vo všetkých oblastiach rovnaké, alebo sa medzi oblasťami môže aj líšiť (tu napr. máme ekonomické dôvody na vyberanie menšieho počtu jednotiek alebo aj náročnosť výberu jednotiek); konečný výberový súbor vytvoríme spojením všetkých oblastí;

PRÍKLAD: ak robíme výber na obyvateľstve SR, oblasťami sú územné celky, vekové skupiny alebo socioekonomický status;

· *skupinový* - pokiaľ je štatistický (základný) súbor pomerne rozsiahly (stotisíce alebo milióny osôb), môžeme uskutočniť jedoduchý NV veľmi ťažko; najprv sa náhodne vyberú skupiny jednotiek (nie jednotlivé jednotky), ktoré tvoria buď prirodzené alebo umelé agregáty; je žiadúce, aby boli jednotlivé agregáty, pokiaľ je to možné, rovnako veľké a vnútri každej skupiny rôznorodé; variabilita medzi skupinami musí byť čo najmenšia (obrátená podoba, ako pri oblastnom výbere);

PRÍKLAD: agregát môže byť malý - rodina, škola, podnik, zdravotný obvod, ale i väčší - obec, okres;

tu nastupujú dve možnosti:

– *výber všetkých jednotiek v agregáte (skupine)*

– *viacstupňový výber* (dvoj a viacstupňový výber) - je založený na hierarchickom popise prvkov základného súboru, ku ktorým sa dostávame postupne cez vyššie výberové jednotky; každá výberová jednotka je skupinou výberových jednotiek nižšieho rádu; rozlišujeme *jednotky prvého stupňa (primárne jednotky)*, potom *jednotky druhého stupňa (sekundárne jednotky)*, atď. až nakoniec máme *základné jednotky* štatistického súboru; postupné výbery prebiehajú často jednoduchým náhodným výberom, alebo sa môže použiť aj mechanický alebo oblastný výber; tento výber má hlavne ekonomické výhody a naviac sa používa aj vtedy, keď nie je v danej situácii (pred začiatkom výberového postupu) úplná opora výberu;

PRÍKLAD: vyššie výberové jednotky → prvkov základného súboru = mestá - bloky - domy - domácnosti, okresy - podniky - dielne - zamestnanci;

2. *selektívny* - dáva skreslený obraz o študovanej populácii;

   PRÍKLAD: vzorka $15-16$ ročných chlapcov, prvoligových basketbalistov, z ktorého by sme chceli robiť inferenciu o výške celej slovenskej populácie $15-16$ ročných chlapcov;

3. *zámerný* - o výbere jednotky rozhodujú okrem náhody častokrát nekontrolovateľné činitele (subjektívny názor vyberajúceho, ochota, resp. neochota odpovedať na kladené otázky a pod.); všeobecne sa tvrdí, že tento typ výberu sa opiera o "expertné" stanovisko a rôzne "odhady", ako získať reprezentatívny výber - takto získané výberové súbory sú často ovplyvnené subjektívnym pohľadom "experta" ale aj ďalšími faktormi ovplyvňujúcimi tvorbu výberu a presnosť zovšeobecňujúcich záverov sa skôr opiera o "expertný" pohľad než o metodológiu.

Náhodný výber (NV) z daného rozdelenia F je usporiadaná n-tica $X_1, X_2, ..., X_n$ rovnako rozdelených nezávislých premenných, ktorých realizácie sú $x_1, x_2, ..., x_n$.

Typy štatistického triedenia dát:

- slúži na rozdelenie výberových jednotiek do skupín (tried) podľa dopredu určených triediacich znakov;
- podľa počtu triediacich znakov rozlišujeme

  - *jednostupňové* - len jeden triediaci znak, napr. triedenie novorodencov podľa pohlavia,

  - *viacstupňové (kombinačné)* - dva a viac triediacich znakov, napr. triedenie zomretých osôb podľa veku, pohlavia a zamestnania;
- *triedy* - pri triedení podľa kvantitatívnych znakov sú určené pomocou *triednych intervalov (TI)*, ktoré musia pokryť všetky hodnoty sledovaného kvantitatívneho znaku a musia byť vzájomne sa neprekrývajúce (disjunktné); hranice TI sú dané číselne a ich rozdiel je *dĺžka TI*; *stred TI* - aritmetický priemer hraníc TI; rovnako dlhé TI - *ekvidištantné*; *počet TI* - najčastejšie od 5 do 20, je určený s ohľadom na rozsah súboru a jeho rozpätie.

6 Poznámky ku rozdeleniam pravdepodobnosti náhodných premenných

V nasledovnej kapitole sa budeme venovať vybraným rozdeleniam pravdepodobnosti bezprostredne súvisiacich so štatistickou inferenciou.

Základné pojmy:
Diskrétne rozdelenie pravdepodobnosti $\{x_i, p_i\}_{i=1}^{n(\infty)}$.
Distribučná funkcia $F^X(x) = \Pr(X < x)$
$F^X(x) = \sum_{i:x_i<x} P(X = x_i) = \sum_{i:x_i<x} p_i, \sum_{i=1}^{\infty} p_i = 1$
$F^X(x) = \int_{-\infty}^{x} f(t)\,dt, f(x) > 0, \int_{-\infty}^{\infty} f(x)\,dx = 1$
Vyjadrenie: vzorec, tabuľka, graf
Stredná hodnota $E(X) = \sum_{i=1}^{\infty} x_i p_i$, $E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$
Disperzia
$D[X] = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E^2[X] = \sum_{i=1}^{\infty} [x_i - E(X)]^2 p_i = \sum_{i=1}^{\infty} x_i^2 p_i - \left[\sum_{i=1}^{\infty} x_i p_i\right]^2$,
$D(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)\,dx, D(X) = E\left[(X - E(X))^2\right] = E\left(X^2\right) - [E(X)]^2$
Vlastnosti:
$E[X + Y] = E[X] + E[Y]$
$E[cX] = cE[X]$
$E\left[\sum_{k=1}^{n} c_k X_k\right] = \sum_{k=1}^{n} c_k E[X_k]$
v prípade nezávislosti $E[XY] = E[X]E[Y]$
$D[cX] = c^2 D[X]$
v prípade nezávislosti $D[X + Y] = D[X] + D[Y]$

6.1 Normálne rozdelenie pravdepodobnosti

Všeobecné normálne rozdelenie $N(\mu, \sigma^2)$
hustota $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in R$
Štandardizované normálne rozdelenie $N(0, 1)$
hustota $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
distribučná funkcia $F^X(x) = P(X < x) = \int_{-\infty}^{x} f(t)\,dt$
Vlastnosti:
- $f(x)$ - symterická okolo osi y, t.j. priamky $x = 0, f(-x) = f(x)$
- $F^X(-x) = 1 - F^X(x)$,
- $\Pr(X > x) = 1 - \Pr(X < x)$
- $f(x) \mapsto 0$ veľmi rýchlo pre $x \mapsto \pm\infty$
- $f(\pm 3) \approx 0$
- μ =*modus=medián*
- inflexné body v $x = \pm 1$
- $\Pr(a \le X \le b) = \Pr(X \le b) - \Pr(X < a) = F^X(b) - F^X(a)$
- $\Pr(|X - \mu| > \sigma) = 0.3173, \Pr(|X - \mu| < \sigma) = 1 - 0.3173 = 0.6827$
- $\Pr(|X - \mu| > 2\sigma) = 0.0455, \Pr(|X - \mu| < 2\sigma) = 1 - 0.0455 = 0.9545$
- $\Pr(|X - \mu| > 3\sigma) = 0.0027, \Pr(|X - \mu| < 3\sigma) = 1 - 0.0027 = 0.9973$
- pravidlo $68.27 - 95.45 - 99.73$ ("miery normálneho rozdelenia")

Štandardizačná rovnica, štandardizácia.

Veta 39 *Ak* $X \sim N(\mu, \sigma^2)$ *a* $Z = \frac{X-\mu}{\sigma}$, *potom* $Z \sim N(0, 1)$.

Definícia 40 95% *predikčný interval* (a,b) *nazývame interval, pre kt. platí* $P(a \leq X \leq b) = 0.95$.

Veta 41 *Nech* $X_1, ..., X_n$ *je náhodný výber z* $N\left(\mu, \sigma^2\right)$, *potom platí*

1. $\overline{X} \sim N\left(\mu, \sigma^2/n\right) \Longleftrightarrow \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N\left(0,1\right)$,
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$,
3. $\frac{\overline{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$.

6.2 Rozdelenia odvodené od normálneho

Chí-kvadrát χ^2_n-rozdelenie s n stupňami voľnosti
Ak sú $Z_1, ..., Z_n \sim N\left(0,1\right)$ nezávislé, potom platí

$$X^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.$$

Studentovo t_n-rozdelenie s n stupňami voľnosti
Ak $Y \sim \chi^2_n$ a $Z \sim N\left(0,1\right)$ sú nezávislé, potom

$$T = \frac{Z}{\sqrt{Y/n}} \sim t_n.$$

Fisherovo F_{n_1,n_2}-rozdelenie
Nech $X_1 \sim \chi^2_{n_1}, X_2 \sim \chi^2_{n_2}, X_1$a X_2 sú nezávislé, potom

$$F = \frac{X_1/n_1}{X_2/n_2} \sim F_{n_1,n_2}.$$

Pozn.: Platí $Z^2 \sim \chi^2_1$, $T^2 \sim F_{1,n_2}$. Ďalej $F_{n_1,n_2}\left(\alpha\right) = \frac{1}{F_{n_1,n_2}(1-\alpha)}$ (aproximácia pre tabuľky).

Logaritmicko - normálne (lognormálne) rozdelenie
Náhodná premenná $X \sim LN\left(\mu_x, \sigma_x^2\right)$ ak veličina $Y = \ln X \sim N\left(\mu_y, \sigma_y^2\right)$. Zo skúseností vieme, že LN rozdelenie máva napr. telesná hmotnosť, čas dožívania po jednej dávke ožiarenia, minimálna smrtná dávka prípravku v homogénnej skupine pokusných zvierat. Z dát sa dá ľahko usúdiť, či je takýto model vhodný. Veľmi často sú zošikmené kladne alebo záporne, napr. pri hematologických, hormonálnych alebo iných biologických veličinách.

Potom
- *aritmetický priemer* $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i \rightarrow \overline{x} = \exp\left(\overline{y}\right)$
- *medián* $\widetilde{y}_{0.5} = \ln \widetilde{x}_{0.5} \rightarrow \widetilde{x}_{0.5} = \exp\left(\widetilde{y}_{0.5}\right)$
- *výberová smerodajná odchýlka* $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\ln x_i - \overline{y}\right)^2} \rightarrow s_x = \exp\left(s_y\right)$

Pozn.: V prípade potreby môžeme použiť iný typ transformácie, $\sqrt{x}$, kde $x \in \mathcal{R}^+ 1/x$, kde $x \neq 0$. Pokiaľ nenájdeme vhodnú transformáciu, použijeme alternatívne prístupy (pozri ďalej).

6.3 Binomické rozdelenie

Náhodná premenná X má *binomické rozdelenie pravdepodobnosti* s parametrami $n \in N, p \in (0,1)$, ak nadobúda hodnoty $0, 1, 2, ..., n$ s pravdepodobnostami $\Pr\left(X = k\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k}$, $k = 0, 1, 2, ..., n$. Ozn.: $X \sim Bin\left(n,p\right)$, distribučná funkcia $F^X\left(x\right) = \Pr\left(X < x\right) = \sum_{k<x}\binom{n}{k} p^k \left(1-p\right)^{n-k}$, $E\left[X\right] = np$, $D\left[X\right] = np\left(1-p\right)$, kde ide o postupnosť $n-$ nezávislých pokusov, k je počet nastatí úspešných pokusov a p je pravdepodobnosť nastatia úspešného pokusu.

Veta 42 ***Moivre - Laplace.*** *Nech X_n je počet úspechov pri n-násobnom nezávislom opakovaní pokusu, pričom pravdepodobnosť úspechu v jednom pokuse je $p \in (0,1)$. Potom pre nahodnu premennú Z_n plati*

$$Z_n = \frac{X_n - np}{\sqrt{np(1-p)}} \sim N(0,1), \lim_{n\to\infty} P(Z_n < x) = \Phi(x).$$

6.4 Kvantily a kritické hodnoty

Definícia 43 *Nech F je spojitá a monotónna funkcia, nech $\beta \in (0,1)$. Potom číslo $x_\beta = F^{-1}(\beta)$ $(F(x_\beta) = \beta)$ sa nazýva β-**kvantil** príslušného rozdelenia a platí:*

- $\Pr\left(x_{\alpha/2} < X < x_{1-\alpha/2}\right) = F\left(x_{1-\alpha/2}\right) - F\left(x_{\alpha/2}\right) = 1-\alpha$, *kde* $\alpha \in (0, 1/2)$
- $Q_1 = F^{-1}(1/4)$, *1. kvartil (dolný)*
- $Q_2 = F^{-1}(1/2)$, *2. kvartil, medián*
- $Q_3 = F^{-1}(3/4)$, *3. kvartil (horný)*
- $F^{-1}(k/10)$, *k-ty decil,* $F^{-1}(k/100)$, *k-ty percentil*

Definícia 44 ***Kritická hodnota*** *príslušného rozdelenia je hodnota, ktorú náhodná premenná X prekročí s pravdepodobnosťou α.*

- $\Pr(X > u(\alpha)) = \alpha$ pre $X \sim N(0,1)$
- $\Pr\left(X > \chi^2_n(\alpha)\right) = \int_{\chi^2(n,\alpha)}^{\infty} f_n(x)\,dx = \alpha$ pre $X \sim \chi^2_n$
- $\Pr(X > t_n(\alpha)) = \alpha$ pre $X \sim t(n)$
- $\Pr(X > F_{n_1,n_2}(\alpha)) = \alpha$ pre $X \sim F(n_1, n_2)$

Pozn.: pravidlo $90.00 - 95.00 - 99.00$ ("upravené miery normálneho rozdelenia")
- 90.00% dát leží v intervale $\mu \pm 1.64\sigma$, kde $u(0.1) = 1.64$,
- 95.00% dát leží v intervale $\mu \pm 1.96\sigma$, kde $u(0.05) = 1.96$,
- 99.00% dát leží v intervale $\mu \pm 2.57\sigma$, kde $u(0.01) = 2.57$.

6.5 Príklady v S-PLUS a R

Popis argumentov funkcií

Koncová časť funkcie po `d`..., `p`..., `q`... hovorí o type rozdelenia. V prvom argumente funkcie je vždy hodnota, ktorá vstupuje do výpočtu, teda `q` (pre kvantilovú funkciu), `d` pre hustotu a `p` pre CDF funkciu. Ďalšie argumenty špecifikujú parametre s "defaultom" pre štandardnú verziu rozdelenia, napr pri normálnom rozdelení ide o parametre rozdelenia $N(0,1)$. Prvým argumentom funkcie može byť aj vektor.

Niektoré rozdelenia a ich parametre

rozdelenie	názov v **S-PLUS a R**	**parametre**
binomické	`binom`	`size, prob`
chi-kvadrát	`chisq`	`df`
exponenciálne	`exp`	`rate`
Fisherovo F	`f`	`df1, df2`
gama	`gamma`	`shape, rate`
geometrické	`geom`	`prob`
hypergeometrické	`hyper`	`m, n, k`
lognormálne	`lnorm`	`meanlog, sdlog`
normálne	`norm`	`mean, sd`
Poissonovo	`pois`	`lambda`
Studentovo t	`t`	`df`
Wilcoxonovo	`wilcox`	`m, n`
mnohorozmerné normálne	`mvnorm`	`mean, cov, ...`

Príklad 45 *normálne rozdelenie, t-rozdelenie, F-rozdelenie (R a S-PLUS)*

- *dvojstranný test al. dvojstranný interval spoľahlivosti (IS)*

`qt(0.975, df=11)...2.200985`, kde `df` je počet stupňov voľnosti, 97.5 percentil, 2.5% kritická hodnota

`qf(0.995,2,7)...12.40396`, kde $n_1 = 2, n_2 = 7$, 99.5 percentil pre $F(2,7)$ alebo 0.5% kritická hodnota

`1-pnorm(1.96)...0.025`, p-hodnota pre $N(0,1)$

`1-pt(1.96,100000) ... 0.025`, p-hodnota pre t_n

- *jednostranný test al. jednostranný IS*

`qt(0.95, df=11)...1.795885`, kde `df` je počet stupňov voľnosti, 95 percentil, 5% kritická hodnota

`qf(0.99,2,7)...9.546578`, kde $n_1 = 2, n_2 = 7$, 99 percentil pre $F(2,7)$, 1% kritická hodnota

`1-pnorm(1.644854)...0.05`, p-hodnota pre $N(0,1)$

`1-pt(1.644869,100000) ... 0.05`, p-hodnota pre t_n

- ak predpokladáme, že pre systolický krvný tlak existuje model $N\left(132, 13^2\right)$, potom aká časť populácie bude mať hodnoty väčšie ako 160?

```
1-pnorm(160,mean=132,sd=13)
```
teda asi 1.6% ppopulácie z $N\left(132, 13^2\right)$

- *binomické rozdelenie* = predpokladajme, že počet ľudí uprednostňujúcich liečbu A pred liečbou B sa správa podľa modelu $Bin\,(n = 20, p = 0.5)$., teda ľudia preferujú oba typy liečby rovnako. Aká je pravdepodobnosť, že budem mať 16 a viac pacientov uprednostňujúcich liečbu A pred liečbou B ?

```
pbinom(16,size=20,prob=0.5)
```
potom
```
1-pbinom(16,size=20,prob=0.5)
```
Pozor! To znamená, že ide o pravdepodobnosť ≤ 16, ale my potrebujeme
```
1-pbinom(15,size=20,prob=0.5)
```
čo ak hľadám pravdepodonosť, že budem mať 16 a viac a zároveň 4 alebo menej pacientov uprednostňujúcich liečbu A pred liečbou B?
```
1-pbinom(15,size=20,prob=0.5) + pbinom(4,size=20,prob=0.5)
```

```
# to je v skutočnosti 2× predchádzajúca pravdepodobnosť
```

Generovanie presudonáhodných čísel z daného rozdelenia a permutácie
- predpona funkcie je `r`, čo znamená "random", napr. `rnorm` je funkcia na generovanie presudonáhodných čísel z normálneho rozdelenia

- prvy argument `n` je rozsah náhodného výberu a potom nasledujú parametre rozdelenia

Príklad 46 *generovanie presudonáhodných čísel z normálneho rozdelenia z rozsahom* 100*, so strednou hodnotou rovnou* 0*, smerodajnou odchýlkou rovnou* 1*. Potom zamente parametre na strednú hodnotu rovnú* 50*, smerodajnú odchýlku rovnú* 0.4*.*

```
rnorm(100)
rnorm(100, mean=50, sd=0.4)
```

Príklad 47 *Vygenerujte* 100 *pseudonahodnych vyberov z kontaminovaného normálneho rozdelenia, v ktorom máme* $N(0,1)$ *s pravdepodobnosťou* 0.95 *a inak* $N(0,9)$

```
rnorm(100,0,1+2*rbinom(100,1,0.05))
```

Príklad 48 *funkcia* `sample`

Funkcia sample prevzorkuje (resamples) z vektora dát, s alebo bez nahradenia. Parameter `n` je celé číslo, `x` je vektor pozorovní, `p` je rozdelenie pravdepodobnosti na `1, ..., length(x)`

`sample(n)`	výber náhodnej permutácie z 1, ... n
`sample(x)`	náhodne permutovaný vektor `x`
`sample(x, replace=T)`	bootstrapový výber
`sample(x, n)`	výber `n` jednotiek z `x` bez nahradenia
`sample(x, n, replace=T)`	výber `n` jednotiek z `x` s nahradením s rovnakou pravdep.
`sample(x, n, replace=T, prob=T)`	výber `n` jednotiek z `x` s nahradením s rôznou pravdep.

```
# zoberie náhodne 10 štátov zo súboru state.name (štáty USA)
sample(state.name, 10)
# zoberie náhodne 75 čísel medzi jedna a jeden milión
sample(1e6, 75)
# zoberie náhodné permutácie čísel 1:50
sample(50)
# Bernuliho rozdelenie s p = 0.3 pre jednotky a p = 0.7 pre nuly s rozsahom 100
sample(0:1,100,replace=T,prob=c(0.3,0.7))
# 20 rovnomerne (s rovnakou pravdepodobnosťou) rozdelených čísel z vektora 1:10 s nahradením
sample(10, 20, T)
# 24 nerovnomerne (s rôznou pravdepodobnosťou, tu c(.3,.4,.1,.1,.1)) rozdelených čísel z vektora
1:5 s nahradením
sample(5, 24, replace=T, prob=c(.3,.4,.1,.1,.1))
```

7 Základná (deskriptívna) štatistika

Hodnoty sledovaných znakov pre jednotlivé premenné zapisujeme do tabuliek. Najprv zapíšeme hodnoty sledovaných znakov v poradí, v ktorom ich dostávame, t.j. do *primárnej tabuľky*. Potom túto tabuľku kvôli prehľadnosti prepíšeme do *sekundárnej tabuľky* (tabuľka rozdelenia početnosti), ktorá nám usporiada sledované znaky podľa veľkosti.

7.1 Charakteristiky polohy

Majme *náhodný výber* (NV) $X_1, X_2, ..., X_n$ z nejakej populácie, kde n je jeho rozsah, *realizácie* tohoto NV budeme označovať ako $x_1, x_2, ..., x_n$, uporiadané realizácie budú $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. Potom môžeme definovať

- *minimum* $x_{\min} = x_{(1)}$,
- *maximum* $x_{\max} = x_{(n)}$,
- *aritmetický priemer* $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
- *medián (prostredná hodnota, robustný odhad polohy)*

$$\widetilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{ak } n \text{ je nepárne} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{ak } n \text{ je párne} \end{cases}$$

Kvantily sú hodnoty skúmanej veličiny, ktoré delia súbor na dve časti. *Kvartily* poznáme tri

- *prvý kvartil* predstavuje hodnotu, od ktorej je 1/4 dát menšia a 3/4 dát je väčšia,

$$Q_1 = P\left[-\infty, \widetilde{x}_{0.25}\right] = \Pr\left[X \leq \widetilde{x}_{0.25}\right] = \frac{1}{4}, Q_1 = \Pr\left[\widetilde{x}_{0.25}, +\infty\right] = \Pr\left[X \geq \widetilde{x}_{0.25}\right] = \frac{3}{4}$$

- *medián (druhý kvartil)* je hodnota, od ktorej je 1/2 dát menšia a 1/2 dát je väčšia,

$$Q_2 = \Pr\left[-\infty, \widetilde{x}_{0.5}\right] = \Pr\left[X \leq \widetilde{x}_{0.5}\right] = \frac{1}{2}, Q_2 = \Pr\left[\widetilde{x}_{0.5}, +\infty\right] = \Pr\left[X \geq \widetilde{x}_{0.5}\right] = \frac{1}{2}$$

- *tretí kvartil* predstavuje hodnotu, od ktorej je 1/4 dát je väčšia a 3/4 dát je menších

$$Q_3 = \Pr\left[-\infty, x_{0.75}\right] = \Pr\left[X \leq x_{0.75}\right] = \frac{3}{4}, Q_3 = \Pr\left[x_{0.75}, +\infty\right] = \Pr\left[X \geq x_{0.75}\right] = \frac{1}{4}$$

- *decily* - delia súbor na desatiny
- *percentily* $p \in (0,1)$ - delia súbor na stotiny a ich všeobecná definícia je nasledovná - $100p$-percentil

$$\widetilde{x}_p = \begin{cases} x_{(k+1)} & \text{pre } k \neq np \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{pre } k = np \end{cases},$$

kde $k = [np]$, čo je celá časť čísla np.

Príklad 49 *Majme výšky $n = 12$ náhodne vybraných 10 ročných dievčat v cm usporiadaných podľa veľkosti (r_i poradia (ranks) pre $x_{(i)}$, pri rovnakých pozorovaniach hovoríme o* ***strednoporadiach****)*

i	1	2	3	4	5	6	7	8	9	10	11	12
$x_{(i)}$	131	132	135	141	141	141	141	142	143	146	146	151
r_i	1	2	3	5.5	5.5	5.5	5.5	8	9	10.5	10.5	12

$\overline{x} \doteq 140.83$, $\widetilde{x} = \frac{1}{2}\left(x_{(6)} + x_{(7)}\right) = 141$, $Q_1 = \widetilde{x}_{0.25} = \frac{1}{2}\left(x_{(3)} + x_{(4)}\right) = 138$, kde $k = [12 \times 0.25] = 3$, $Q_3 = \widetilde{x}_{0.75} = \frac{1}{2}\left(x_{(9)} + x_{(10)}\right) = 144.5$, kde $k = [12 \times 0.75] = 9$

Existujú aj iné možnosti výpočtu kvantilov (pozri Zvára 2003, str. 20).

Čo sa stane so spomínanými charakteristikami polohy, pokiaľ zmeníme merítko (škálu)? Napr. gramy na kilogramy, alebo namiesto hmotnosti, použijeme logaritmus hmotnosti.

Nech $a, b \in \mathcal{R}$ sú nejaké dané konštanty. Potom $y_i = a + bx_i$ a pre priemer

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\left(na + b\sum_{i=1}^{n}x_i\right) = a + b\overline{y} = \overline{a + by}.$$

Pokiaľ $b \in \mathcal{R}^+$, sa usporiadanie hodnôt x_i pri prechode na $y_i = a + bx_i$ nezmení, teda

$$a + bx_{(1)} \leq a + bx_{(2)} \leq ... \leq a + bx_{(n)}.$$

Teda pre medián v tomto prípade platí

$$\widetilde{y} = a + b\widetilde{x} = \widetilde{a + bx}.$$

Ľahko sa dá nahliadnuť, že usporiadanie zachová každá rastúca fcia $g(x)$, teda platí

$$g\left(x_{(1)}\right) \leq g\left(x_{(2)}\right) \leq ... \leq g\left(x_{(n)}\right)$$

a pre medián bude platiť $g(\widetilde{x}) \doteq \widetilde{g(x)}$. Pre nepárne n platí predchádzajúci vzťah presne, označenie "približnosti" potrebujeme pre párne n, kde $x_{\left(\frac{n}{2}\right)} < x_{\left(\frac{n}{2}+1\right)}$. V tomto prípade je však 1/2 hodnôt $g(x_i)$ menšia ako $\widetilde{g(x)}$. Teda špeciálne môžeme medián logaritmu (napr. hmotnosti) spočítať ako logaritmus mediánu (napr. hmotnosti).

Pozn.: Pokiaľ teda dôjde v pozorovaniach k posunutiu, dôjde k rovnakému posunutiu ak u charakteritiky polohy. Ak zmeníme merítko, potom stačí urobiť rovnakú úpravu aj u charakteristiky polohy.

7.2 Charakteristiky variability

Definujeme nasledovné základné pojmy

- *výberový rozptyl*

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Pri lineárnej transformácii sa rozptyl mení nasledovne

$$s_y^2 = s_{a+bx}^2 = b^2 s_x^2,$$

prečo?

$$\begin{aligned} s_y^2 &= s_{a+bx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(a + bx_i - \overline{a + bx}\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(a + bx_i - (a + b\overline{x})\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(b\left(x_i - \overline{x}\right)\right)^2 = b^2 s_x^2 \end{aligned}$$

- *smerodajná odchýlka*
$$s_x = \sqrt{s_x^2}.$$
Pri lineárnej transformácii sa smerodajná odchýlka mení nasledovne
$$s_y = s_{a+bx} = |b|\, s_x^2,$$
teda, ak pripočítame ku všetkým pozorovaniam rovnakú konštantu, miera variability sa nezmení. Zmena merítka (u pomerového merítka zmena jednotiek) má za následok rovnakú úpravu jednotlivých pozorovaní i miery variability (s výnimkou rozptylu).

- *výpočtová podoba rozptylu*
$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{\sum_{i=1}^{n} x_i^2 f_i}{\sum_{i=1}^{n} f_i} - \frac{n}{n-1}\overline{x}^2,$$
kde f_i sú frekvencie (počty) prislúchajúcich x_i a $n = \sum_{i=1}^{n} f_i$.

- *výberový koeficient variácie* - podiel variability na priemere
$$V_k = \frac{s_x}{\overline{x}}.$$
Používa sa pri porovnávaní variability súborov s nerovnakými priemermi (napr. pri porovnaní variability výšky detí určitého veku s výškou dospelých určitého veku alebo pri porovnaní variability premenných meraných v rôznych jednotkách), je bezrozmerný a zvyčajne sa vyjadruje v percentách, t.j. $100\,(\sigma/\mu)\,\%$.

- *odhad rozptyl priemeru*
$$s_{\overline{x}}^2 = \frac{s_x^2}{n},$$

- *odhad strednej chyby priemeru (štandardná chyba, standard error)*
$$s_{\overline{x}} = \frac{s_x}{\sqrt{n}},$$

- *výberový koeficient šikmosti (skewness)*
$$b_1 = \frac{n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left[n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{3/2}},$$
kde rozdelenie je pozitívne zošikmené (rozdelenie pravdepodobnosti na ľavej strane stúpa strmejšie, $b_1 > 0$), negatívne ($b_1 < 0$),

- *výberový koeficient špicatosti (kurtosis)*
$$b_2 = \frac{n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left[n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2} - 3 = \frac{n\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2} - 3,$$
kde rozdelenie je špicaté (leptokurtické, $b_2 > 0$), normálne (mezokurtické, $b_2 = 0$) a ploché (platykurtické, $b_2 < 0$),

- *suma štvorcov (sum of squares)*
$$SS = \sum_{i=1}^{n}(x_i-\overline{x})^2$$

– čitateľ rozptylu, používa sa napr. v regresnej analýze, v modeli ANOVA.

- *rozpätie (distance)* $D = x_{\max} - x_{\min}$,
- *medzikvartilové rozpätie* $D_Q = Q_3 - Q_1$,
- *decilové rozpätie,*
- *percentilové rozpätie,*
- *5 číselné summary:* $x_{\min}$,$Q_1, Q_2, Q_3, x_{\max}$,
- *symetria:* $Q_3 - Q_2 = Q_2 - Q_1$
- *pozitívna šikmosť:* $Q_3 - Q_2 > Q_2 - Q_1$
- *negatívne šikmosť:* $Q_3 - Q_2 < Q_2 - Q_1$
- robustný výpočet minima a maxima ("*vnútorné hradby*", prvky vybočujúce z hradieb sa považujú za *podozrivé*, potencionálne outliery - pozri ďalej)
  - $x^*_{\min} = B_D = Q_1 - 1.5\,(Q_3 - Q_1) = Q_1 - 1.5D_Q$
  - $x^*_{\max} = B_H = Q_3 + 1.5\,(Q_3 - Q_1) = Q_3 + 1.5D_Q$

"*Vonkajšie hradby*" $B^*_D = Q_1 - 3D_Q$, $B^*_H = Q_3 + 3D_Q$. Pokiaľ sú nejaké $x_i < B^*_D \vee x_i > B^*_H$, hovoríme, že ide o *vzdialené body*, ak $x_i \in \langle B^*_D, B_D) \vee (B_H, B^*_H\rangle$, ide o *body vonkajšie*, ak $x_i \in \langle B_D, B_H\rangle$, ide o *body vnútorné*, alebo *body priľahlé mediánu*. Pre normálne rozdelenie platí $B_H - B_D = Q_3 + 1.5D_Q - Q_1 + 1.5D_Q = 4D_Q \doteq 4.2$. Pravdepodobnosť, že $x_i \notin \langle B_D, B_H\rangle$ je potom 0.04.

Niekedy nás zaujímajú normované proťajšky realizácií, a to *z-skóre* (často používané v antropológii)

$$z_i = \frac{x_i - \overline{x}}{s_x},$$

ktoré dostaneme ako špeciálny prípad lineárnej transformácie $y = a + bx$, kde voľbou $b = 1/s_x$ a $a = -\overline{x}/s_x$. Potom dostaneme

$$\overline{z} = -\frac{\overline{x}}{s_x} + \frac{1}{s_x}\overline{x} = 0$$

a

$$s_z^2 = \left(\frac{1}{s_x}\right)^2 s_x^2 = 1$$

a preto nazývame z-skóre aj normované veličiny. Pomocou z-skóre môžeme vyjadriť aj koeficient šikmosti a špicatosti

$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^3 = \frac{1}{n}\sum_{i=1}^{n} z_i^3,$$

kde ak dáta pochádzajú z normálneho rozdelenia

$$E\,[b_1] = 0, Var\,[b_1] = \frac{n-2}{(n+1)\,(n+3)},$$

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^4 = \frac{1}{n}\sum_{i=1}^{n} z_i^4,$$

kde ak dáta pochádzajú z normálneho rozdelenia

$$E\,[b_2] = 3 - \frac{6}{n+1}, Var\,[b_2] = \frac{24n\,(n-2)\,(n-3)}{(n+1)^2\,(n+3)\,(n+5)}.$$

Pokiaľ dáta pochádzajú z normálneho rozdelenia, budú mať oba koeficienty hodnoty blízko nuly (pri b_2, ak od neho odčítame konštantu 3).

7.3 Príklady v S-PLUS a R

Príklad 50 *charakteristiky polohy a variability*

- `x _ 1:100`, funkcie:

minimum `min(x)`
maximum `max(x)`
rozsah súboru `c(max(x),min(x))`; `range(x)`
medián `median(x)`
aritmetický priemer `mean(x); sum(x)/length(x)`
prvý kvartil `q1 _ quantile(x,0.25)`
druhý kvartil `q2 _ quantile(x,0.50)`
tretí kvartil `q3 _ quantile(x,0.75)`
kvartily `quantile(x,c(0.25,0.5,0.75))`
5 číselné summary `quantile(x,c(0,0.25,0.5,0.75,1))`
medzikvartilové rozpätie `Dq _ quantile(x,0.75)-quantile(x,0.25)`
zistovanie symetrie `c(q3 - q2, q2 - q1)`
robustný výpočet minima a maxima ("*vnútorné hradby*")
`Bd_q1-1.5Dq; Bh_q3+1.5Dq`
robustný rozsah súboru `c(Bd,Bh)`
rozptyl `var(x)`
`sum((x-mean(x))^2)/(length(x)-1)`
štandardná chyba
`stderr_function(x) sqrt(var(x) / length(x))`
`stderr(x)`
`prijemstderr_tapply(prijem, statyfak,stderr)`

- posun o konštatnu `a <- 900`:

```
x_c(907,908,898,902,897)
x1 _ x-900
mean(x)
var(x)
```

- vytvorte funkciu na výpočet sumy štvorcov
- vytvorte funkciu na výpočet výberového koeficientu šikmosti a špicatosti
- vytvorte funkciu na výpočet 5 číselného summary, priemeru a smerodajnej odchýlky tak

```
CHAR.SAMPLE<-function(x)
{
 RESULTS <- list( QUANTILE = 0, MEAN = 0, SD = 0)
 QUANTILE <- quantile(x, c(0, 0.25, 0.5, 0.75, 1))
 MEAN <- mean(x)
 SD <- sqrt(var(x))
 RESULTS <- rbind(QUANTILE, MEAN, SD)
 RESULTS$QUANTILE <- QUANTILE
 RESULTS$MEAN <- MEAN
 RESULTS$SD <- SD
 return(RESULTS)
}
```

- vytvorte funkciu na výpočet aritmetického priemeru tak, aby bol ošetrená na chýbajéce pozorovania použitím nejakej základnej funkcie

```
Mmean_function(x) mean(x[!is.na (x)])
```

8 Grafická interpretácia výberového súboru

V modernej počítačovej štatistike hovoríme o *exploratórnej analýze dát (EDA).*
Grafická interpretácia výberového súboru je možná pomocou nasledovných grafov.

- *Stĺpcový diagram (barplot)* - číselné hodnoty sú vyjadrené pomocou obdĺžnikových stĺpcov (obyčajne v zvislej polohe), často škálovaný (v *relatívnej škále*, ak sledujeme viac súborov, ide potom o lepšie porovnanie týchto súborov; neškálovaný - *absolútna škála*) - jeho špeciálnymi prípadmi sú *veková pyramída* (strom života, znázorňuje vekové zloženie obyvateľstva) a histogram.

- *Spojnicový graf* - znázorňuje priebeh časového radu a jeho špecálnymi prípadmi sú polygón početnosti, frekvenčná krivka, *polygón kumulatívnych početností.*

- *Bodový graf (scatterplot)* - zobrazuje namerané hodnoty v pravouhlej súradnicovej sústave (2D, 3D), na odlíšenie rôznych kategórií použijeme rôzne znaky, farby a pod.; často používaný na zobrazenie závislosti dvoch znakov.

- *Kruhový diagram (výsečový, koláčový, piechard)* - zachytáva štruktúru súboru, kde plocha kruhu predstavuje celý súbor a jednotlivé časti sú znázornené kruhovými výsečami, tu 360st. zodpovedá 100% plochy kruhu, výseč 3.6 stupňa je 1%.

- *Histogram* - stĺpcový diagram s k stĺpcami, ktorých základňa sa rovná šírke intervalu $I_i = (x_i, x_{i+1}\rangle$ a výška $i-$teho stĺpca jeho početnosti ($i = 1, 2, ..., k$); vystihuje celkom presne počet pozorovaní v jednotlivých intervaloch; množstvo intervalov volí príslušný software alebo aj sám užívateľ. Histogram (a mnohé ďalšie zobrazovacie metódy) môžeme použiť aj v škále relatívnych početností. Potrebujeme aspoň 12 triednych intervalov (ich počet nesmie klesnuť pod 6). Šírka jedného je minimálne $h_{\min} = 0.08(x_{\max} - x_{\min})$. Musí obsahovať minimálne 5 meraní. *Frekvenčná tabuľka*: znak, početnosť, relatívna početnosť. *Počet tried* je $k = \log_2 n + 1 \doteq 2 + 3.3 \log_{10} n$ (*Sturgesova formula*), intervaly sú definované ako $\langle x_0, x_1\rangle$, $(x_1, x_2\rangle$, ...; šírka intervalov je teda $h = D/\left(\log_2 n + 1\right)$ *pre realizácie z normálneho rozdelenia* (Scott, 1992). Teraz už vlastne nepracujeme s realizáciami x_i, ale so stredmi intervalov $x_i^* = \left(x_i + x_{i+1}\right)/2$. Počty hodnôt n_i, ktoré sa v intervale I_i nachádzajú, sa nazývajú *triedne početnosti.* Pokiaľ realizácie nemajú normálne rozdelenie, treba použiť robustné algoritmy. Taktiež outliery môžu dramaticky nafúknuť (inflate) rozpätie súboru, čo môže spôsobiť nárast šírky intervalov. Preto sa využívajú dva algoritmy ako kompromis medzi výchylkou (biasom) a rozptylom realizácií pochádzajúcimi z normálneho rozdelenia. Potom šírky triednych intervalov budú $h_1 = 3.49\widehat{\sigma}n^{-1/3}$, $\widehat{\sigma} = s$ (pre výbery z normálneho rozdelenia, *Scottova formula,* Scott, 1979), $h_2 = 2D_Q n^{-1/3}$ (robustnejšia, *Freedman - Diaconisova formula,* ktorá je nezavislá od outlierov a vyberá menšie intervaly ako Scottova formula *(*Freedman, Diaconis, 1981). Pre symetrické rozdelenia platí $h_3 = [2\sqrt{n}]$ alebo $h_4 = \left[2.46 \times (n-1)^{0.4}\right]$, kde $[x]$ je celá časť čísla x. Pokiaľ sa neočakáva príliž zošikmené rozdelenie, šírka triednych intervalov h je konštantná. Pre komplikovanejšie tvary výberových rozdelení treba zväčšiť počet triednych intervalov, alebo použiť špeciálne postupy na hľadanie nekonštantne dlhých triednych intervalov (pozri Meloun a Militký, 2004).

- *Polygón početnosti* - spojnicový diagram, kde nad stredmi triednych intervalov I_i vztýčime kolmice, ktorých výška je úmerná príslušným triednym početnostiam a koncové body kolmíc pospájame, ich súradnice sú $[x_i^*, n_i]$; ide o dosť dobré podanie priebehu početnosti aj vnútri jedného intervalu, ale plocha uzavretá spojnicou polygónu nie je úmerná počtu pozorovaní v intervale.

- *Frekvenčná krivka* - vznikne ak koncové body polygónu pospájame hladkou krivkou; vystihuje celkom presne priebeh rozdelenia početnosti a plocha v každom mieste ohraničená krivkou je priamo úmerná počtu pozorovaní.

- *Histogram kumulatívnych početností (súčtový histogram)* - namiesto početnosti bude- me nad jednotlivými intervalmi I_i zakreslovať obdĺžniky s výškou rovnajúcou sa príslušným kumulatívnym

početnostiam, $N_i = \sum_{j=1}^{i} n_j$. Kumulatívne relatívne početnosti definujeme ako N_i/n, čomu zodpovedá *empirická distribučná funkcia*, definovaná pre zvolené číslo x, ako relatívna početnosť v intervale $(-\infty, x\rangle$, teda ako N_i-tina hodnôt x_i menších alebo rovných ako x

$$\widehat{F}^X(x) = \frac{\#x_i < x}{n}.$$

- *Boxploty - "krabicové diagramy"* (dobre identifikujú symetriu rozdelenia medzi kvartilmi, symetriu rozdelenia v koncoch rozdeleniaoutliery, znázorňujú robustný odhad polohy - medián - naviac sa používajú na grafické porovnanie dvoch súborov), ktorých šírka $\sqrt{n}$ predstavuje odmocninu z rozsahu výberového súboru a vidieť na nich po poradí $x_{\min}$,Q_1, Q_2, Q_3 a $x_{\max}$ a je možné do nich vkresliť aj aritmetický priemer, čo zvýrazení prípadné odchýlky od normality. Ak $Q_2 < \overline{x}$, ide o pravostranne zošikmené rozdelenie, ak $Q_2 > \overline{x}$, ide o ľavostranne zošikmené rozdelenie. Pokiaľ hovoríme o *"krabicových diagramoch so zárezom"*, ide o zárez charakterizujúci *robustný interval spoľahlivosti* (pozri nižšie) mediánu $\widetilde{x} \in (I_{DH}, I_{HH})$, kde

$$I_{DH} = \widetilde{x} - 1.57\frac{D_Q}{\sqrt{n}}, I_{HH} = \widetilde{x} + 1.57\frac{D_Q}{\sqrt{n}},$$

naviac odhadom teoretického mediánu je $\widetilde{x}$ (ako už vieme) a odhadom jeho rozptylu je $\sigma^2_{\widetilde{x}} = D_Q/1.349$, kde vo všeobecnosti platí (pre akékoľvek rozdelenie pravdepodonosti)

$$\sigma^2_{\widetilde{x}}\left(f\left(\widetilde{x}\right)\right) = 1/\left(4f^2\left(\widetilde{x}\right)\right),$$

kde f je hustota rozdelenia pravdepodobnosti. Pre normálne rozdelenie bude platiť $\sigma^2_{\widetilde{x}} = \sigma^2_x \frac{\pi}{2n}$, kde $\widetilde{x} \sim N\left(\widetilde{\mu}, \sigma^2_x\right)$.

- *Kvantilové diagramy* (qq-diagramy, normal probability plot) - zobrazujú body so súradnicami $\left[\Phi^{-1}\left(i/\left(n+1\right)\right), x_{(i)}\right]$, kde $\Phi^{-1}(p)$ je kvantilová funkcia normovaného normálneho rozdelenia definovaná nasledovne

$$\Pr\left(Z \leq \Phi^{-1}(p)\right) = p.$$

Šikmosť $b_1 > 0$ sa prejaví v podobe konvexného usporiadania hodnôt, $b_1 < 0$ sa prejaví v podobe konkávneho usporiadania hodnôt. Taktiež je tu viditeľná dĺžka "chvostov" rozdelenia, krátke "chvosty" sú prejavom esovitého usporiadania bodov, dlhé "chvosty" naopak hovoria o inverzne esovitom usporiadaní bodov. Tiež je zreteľná prípadná bimodalita rozdelenia. Štatisticky je možné testovať normalitu rozdelenia pomocou simulácií z $N(0,1)$ a vytvorením Atkinsonovej obálky (e.g. Katina, 2003). Pre normálne rozdelenie bude platiť $\sigma^2_{\widetilde{x}_p} = \sigma^2_x \frac{\pi^2}{24 \ln n}$, kde $\widetilde{x}_p \sim N\left(\widetilde{\mu}_p, \sigma^2_{\widetilde{x}_p}\right)$.

Pozn.: Ako jednoducho identifikovať hodnoty vybočujúce (pre normálne rozdelenie)? Ako mieru rozptylu definujme *kvantilovú odchýlku* $D_{Q^*} = 2D_Q$. Pokiaľ urobíme štandardi- záciu (pozri Meloun a Militký, 2004, str. 86; Parsen, 1988), dostaneme $D_{Q^*_{st}} = 1$ a štandardizovaný medián bude $\widetilde{x}_{st} = 0$ a štandardizovaná kvantilová funkcia indikujúca tvar bude

$$Q_{st}(p) = \frac{\widetilde{x}_p - \widetilde{x}_{0.5}}{D_{Q^*}}.$$

Hodnoty kvantilov, pre ktoré platí $|Q_{st}(p)| \geq 1$, sú považované za vybočujúce (pre normálne rozdelenie) a hovoríme potom o identifikátoroch dlhých chvostov (koncov). Hodnoty $Q_{st}(p)$ môžeme použiť na

- identifikáciu miery šikmosti $SQ = Q_{st}(0.25) + Q_{st}(0.75)$, kedy je rozdelenie pravdepodobnosti symetrické, ak SQ je rovné nule,

- identifikáciu dĺžky koncov, kedy

$Q_{st}(0.95) < 0.5$ hovorí o krátkych koncoch,

$Q_{st}(0.95) > 1$ hovorí o dlhých koncoch a

pre stredne dlhé konce bude platiť $Q_{st}(0.95) \in \langle 0.5, 1.0\rangle$.

Čo je *homogénny NV*? všetky prvky NV $x_i, i = 1, 2, ..., n$, pochádzajú z rovnakého rozdelenia pravdepodobnosti s konštantným rozptylom σ^2, ako sme už spomínali. K nehomogenitám dát dochádza všade tam, kde sa vyskytujú výrazné nerovnomernosti na- meraných premenných, náhle sa menia podmienky experimentu a pod. Nehomogenita môže byť spôsobená aj nevhodnou špecifikáciou súboru. Pokiaľ ide NV rozdeliť podľa nejakých logických kritérií do niekoľkých podskupín, je možné štatisticky spracovať každú takúto podskupinu zvlášť a potom ich porovnať napr. na základe testov stredných hodnôt.

Špeciálnym prípadom sú odľahlé pozorovania. Takéto pozorovanie skresľujú odhady polohy a hlavne rozptylu σ^2, takže môžu znehodnotiť ďalšiu štatistickú analýzu. Problém takýchto pozorovaní je značne komplikovaný. Pri ich overovaní sa používa mnoho ideali- zovaných predpokladov. Je nutné poznať ich predpokladaný počet, ich rozdelenie a tiež rozdelenie ostatných prvkov NV. Naviac je nutné zostrojiť model, podľa ktorého sa odľahlé pozorovania chovajú. Testovanie odľahlých pozorovaní bez doplnkových informácií je teda málo spoľahlivé.

Jednoduchou technikou, kedy sa predpokladá, že dáta majú normálne rozdelenie, je modifikácia vnútorných hradieb B_D a B_H na

$$B_D^{mod} = Q_1 - kD_Q, B_H^{mod} = Q_3 + kD_Q,$$

kde sa parameter k sa volí tak, aby pravdepodobnosť $\Pr(n, k)$, že z NV s rozsahom n pochádzajúceho z normálneho rozdelenia, nebude žiaden prvok mimo intervalu $I = \left\langle B_D^{mod}, B_H^{mod} \right\rangle$, bola dostatočne vysoká, napr. 0.95. Ak $\Pr(n, k) = 0.95$ a $n \in \langle 8, 100 \rangle$ použijeme aproximáciu $k \approx 2.25 - 3.6/n$. Teda, prvky mimo I sa považujú za odľahlé. Postup spomenutý vyššie je robustný.

8.1 Príklady v S-PLUS a R

- **histogram, hustota rozdelenia pravdepodobnosti a kumulatívna distribučná funkcia**

Príklad 51 ***histogram*** *`hist(variable,probability = T, nclass = number, plot = T)`*

argumenty: `probability = T` je pravdepodobnostná škála, `nclass = number` počet tried, `plot = F` slúži na vypísanie hraníc intervalov

default pre `nclass` je $\log_2 n + 1$ (*Sturgesova formula*), intervaly sú definovane ako $\langle x_0, x_1 \rangle, (x_1, x_2 \rangle, ...$; špeciálne môžeme vylúčiť ľavý koncový bod x_0 prostredníctvom argumentu `include.lowest=F` (s defaultom `T`); šírka intervalov je teda `range(x)/` $\log_2 n + 1$, pre normálne rozdelené dáta (Scott, 1992)

```
help(lottery.payoff)
hist(lottery.payoff)
hist(lottery.payoff, nclass=15)
hist(lottery.payoff, nclass=15, probability=T)
x <- hist(lottery.payoff, nclass=15, probability=T, plot = F)
x
attach(geyser)
names(geyser)
hist(waiting,probability=T)
# čiernobiely štýl s "kobercom"
hist(waiting,probability=T,nclass=20,style=''old'')
rug(waiting)
```

Pozn.: Outliery môžu dramaticky umelo nafúknuť rozsah súboru, čo môže spôsobiť nárast širky intervalov v centre rozdelenia. Preto sa najčastejšie využívajú dve pravidlá ako kompromis medzi výchýlkou a rozptylom pre dáta z normalnho rozdelenia s outliermi , ktoré využívajú nasledovné šírky intervalov definované ako *Scottova formula* (Scott, 1979) $h_1 = 3.49\widehat{\sigma} n^{-1/3}$ a *Freedman-Diaconisova formula* (Freedman a Diaconis, 1981) $h_2 = 2D_Q n^{-1/3}$, kde $r = Q_3 - Q_1$

```
# Scottova formula
nclas.scott_function(x)
{
     h_3.5*(sqrt(var(x))*length(x)^(-1/3)
```

```
     ceiling(diff(range(x))/h)
}
hist.scott_function(x, prob=T, xlab=deparse(substitute(x)), ...)
     invisible(hist(x, nclass=nclass.scott(x), prob=prob,xlab=xlab,...))
# Freedman-Diaconisova formula
nclass.FD_function(x)
{
     r_as.vector(quantile(x,c(0.25,0.75)))
     h_2*(r[2]-r[1])*length(x)^(-1/3)
     ceiling(diff(range(x))/h)
}
hist.FD_function(x, prob=T, xlab=deparse(substitute(x)), ...)
     invisible(hist(x, nclass=nclass.FD(x), prob=prob,xlab=xlab,...))
```

Príklad 52 ***hustota rozdelenia pravdepodobnosti****, funkcia* `plot()` *a jej argumenty, tu zakreslenie čiary* `type=''l''`

```
x <- seq(-4,4,by=0.01)
plot(x,dnorm(x),type=''l'')
```

`type= ''p''` pre body (points), `type=''l''` pre čiary (lines), `type=''b''` pre oboje (both), `type=''n''` pre prázdny obrázok

```
# V Rku inak, funkcia curve(), odkiaľ, kam a koľko bodov
curve(dnorm(x),from=-4,to=4,n=10000)
# vkreslenie hustoty do histogramu
x <- rnorm(10000)
hist(x,prob=T)
lines(density(x))
# v Rku
x <- rnorm(1000)
hist(x,freq=F)
curve(dnorm(x),from=-4,to=4,n=1000,add=T)
```

čo tak v Rku hustotu binomického rozdelenia s parametrami $n = 50$, $p = 0.33$

```
x <- 0:50
plot(x,dbinom(x, size=50, prob=0.33),type=''h'')
```

Príklad 53 ***kumulatívna distribučná funkcia****, funkcia* `cdf.compare()`

- *jednovýberový problém:* grafické porovnanie empirickej a hypotetickej (teoretickej) kumulatívnej distribučnej funkcie (ecdf a tcdf)

- *dvojvýberový problém:* grafické porovnanie dvoch empirických kumulatívnych distribučných funkcií

pozn.: použitie pred Kolmogorov-Smirnov testom

- argumenty funkcie:

`cdf.compare(x, y = NULL, distribution = ''normal'')`

`x` a `y` sú numerické vektory (`y` len pri porovnávaní dvoch ecdf)

typ rozdelenia pravdepodobnosti: `distribution = ''normal'', ''chisquare'', ''f'', ''gamma'', ''lognormal'', ''logistic'', ''t'', ''binomial'', ''geometric'', ''hypergeometric'', ''poisson'', ''wilcoxon''`

```
# jednovýberový problém
# nasimulujte dáta z normálneho rozdelenia s rozsahom 100
z <- rnorm(100)
# porovnajte ich s tcdf normálneho a potom chi-kvadrát rozdelenia
cdf.compare(z,dist=''normal'')
```

```
cdf.compare(z,dist=''chisquare'',df=2)

# dvojvýberový problém
# nasimulujte dáta z normálneho rozdelenia s rozsahom 25 a normálneho rozdelenia s rozsahom 100
x <- rnorm(25)
y <- rexp(100)
# porovnajte ich
cdf.compare(x,y)
# reálne dáta waiting
cdf.compare(waiting,dist=''norm'')
cdf.compare(waiting,dist=''norm'',mean=mean(waiting),sd=sqrt(var(waiting)))

# v Rku
# funkcia ecdf(), parameter do.points=F nekreslí body a verticals=T kreslí schodovitú funkciu
library(ISwR)
library(help=ISwR)
data(vitcap2)
attach(vitcap2)
names(vitcap2)
plot(ecdf(vital.capacity),do.points=F, verticals=T)
x <- seq(min(vital.capacity), max(vital.capacity), by=0.01)
lines(x,pnorm(x,mean=mean(vital.capacity),sd=sqrt(var(vital.capacity))))
```

- **krabicový diagram (boxplot)**

```
# argumenty funkcie boxplot()
```

`varwidth` je argument relatívnej šírky jednotlivých krabičiek. Default je `F`, čo znamená rovnaká šírka pre všetky krabičky; ak ho zmeníme na `T`, šírka krabičiek bude proporcionálna ku druhej odmocnine z počtu pozorovaní

`names` je vektor pomenovaní pre jednotlivé zobrazované skupiny, ak ho vynecháme, použijú sa názvy z atribútu `names` z dátového rámca

`plot` ak je `T`, krabičky sa zobrazia, ak je `F`, budú počítané charakeristiky krabičiek v číselnej podobe

`notch` ak je `T`, zobrazia sa zárezy krabičiek, ktoré zodpovedajú 95% intervalom spoľahlivosti pre medián

`boxcol` zodpovedá farbe vnútri krabičiek, uvedie sa číslo farby; ak `boxcol=-1`, nepoužije sa žiadna farba

`medchar` logický argument indikukúci, či zobraziť medián ako bod alebo nie. Implicitne je nastavené `T`, ak je použitý argument `medpch`, inak default je `F`.

`medpch` typ bodu na zobrazenie mediánu (číslo typu bodu), default je `NA`, lebo medián je zobrazovaný ako úsečka

`medline` logický argument indikujúci, či zobraziť medián ako úsečku alebo nie. Default je `T`, ak je použitý argument `medlwd`

`medlwd` šírka úsečky charakterizujúcej medián. Implicitne je nastavená ako `T`, ak `medline` parameter je `T`, inak je hodnota `medlwd` rovná `NA`, default je rovný 5

`medcol` farba mediánu v podobe úsečky alebo bodu; default je 0

`confint` ak je rovn/'y (`T`, zobrazia sa intervaly spoľahlivosti pre medián. Ak sa intervaly dvoch krabičiek neprekrývajú, indikuje to, že neexistuje rozdiel medzi mediánmi na 5% hladine významnosti

`confnotch` ak je rovný `T`, zárezy intervalov spoľahlivosti sú zobrazené, default je `F`.

`confcol` farba intervalov spoľahlivosti, default je 2.

`outchar` logický argument na zobrazenie outlierov ako bodov; je implicitne nastavený ako `T`, ak je použitý `outpch` argument, default je `F`

`outpch` typ bodu na zobrazenie outlierov; je implicitne nastavený ako 1, ak je použitý argument `outchar`

`outline` logický argument indikujúci, či zobrazit outliery ako horizontálne úsečky alebo nie; implicitne je nastavený ako T, ak je použitý argument `outwex`

`outwex=` šírka úsečky predstavujúcej oultiery, proporčná ku šírke krabičiek; default je 1

Príklad 54 *funkcie `split()`, `boxplot()`, dátový rámec `market.frame`*

```
attach(market.frame)
# priemerný príjem podľa veku
sapply(split(income,age), mean)
# alternatívny výpočet
tapply(income,list(age), mean)
# po dekádach
split(income,age %/% 10)
# nastavte argumenty funkcie boxplot() boxplot(split(income,age))
boxplot(split(income, age), varwidth=TRUE, notch=TRUE)
boxplot(split(age, employment), notch = TRUE)
```

Príklad 55 *funkcia `split`, dátový rámec `ship`*

```
# komponent pre každý mesiac
split(ship, cycle(ship))
# nastavte argumenty funkcie boxplot()
boxplot(split(ship, cycle(ship)))
```

Príklad 56 *dátové rámce `lottery.payoff, lottery2.payoff, lottery3.payoff.`*

```
boxplot(lottery.payoff, lottery2.payoff, lottery3.payoff)
boxplot(
 split(lottery.payoff, lottery.number%/%100),
 main=''NJ Pick-it Lottery (5/22/75-3/16/76)'',
 sub=''Leading Digit of Winning Numbers'',
 ylab=''Payoff'')
```

Príklad 57 *vkreslenie aritmetického priemeru do krabicového diagramu*

```
x <- boxplot(lottery.payoff, lottery2.payoff, lottery3.payoff)
lp.mean <- mean(lottery.payoff)
lp2.mean <- mean(lottery2.payoff)
lp3.mean <- mean(lottery3.payoff)
lp1.mean <- mean(lottery.payoff)
lp.mean <- c(lp1.mean,lp2.mean,lp3.mean)
points(x,lp.mean,pch=16)
```

- **kvantilový diagram (qq-plot)**

Príklad 58 *funkcie `qqnorm()`, `qqline()`, `qqplot()`*

```
# zobrazenie kvantilov náhodného výberu generovaných pseudonáhodných čísel oproti kvantilom
normálneho normovaného rozdelenia
nc50 < − rnorm(50)
qqnorm(nc50)
qqnorm(nc50)
nc1000 < − rnorm(1000)
qqnorm(nc1000
qqnorm(nc1000)
```

```
# nagenerujte 1000 bodov z t_10 rozdelenia a porovnajte ich s normálnym rozdelením
x_rt(1000,10)
# vidime "dlhochvosté" rozdelenie
qqnorm(x);qqline(x)
# zobrazenie kvantilov náhodného výberu oproti kvantilom normálneho normovaného rozdelenia
qqnorm(lottery.payoff);qqline(lottery.payoff)
# zobrazenie kvantilov štandardizovaného náhodného výberu (z-skóre) oproti kvantilom normálneho
normovaného rozdelenia
LPscale <- scale(lottery.payoff)
qqnorm(LPscale);qqline(LPscale)
# zobrazenie kvantilov dvoch náhodných výberov a MNŠ priamky (podrobnosti neskôr)
qqplot(lottery.payoff, lottery3.payoff)
zz <- qqplot(lottery.payoff, lottery3.payoff, plot = F)
plot(zz)
abline(lm(zz$y~zz$x))
# znazornenie náhodného výberu oproti t-rozdeleniu
plot(qt(ppoints(waiting),298),sort(waiting))
```

kde funkcia `ppoints()` zobrazuje prislušnú množinu pravdepodobností pre kvantilový graf - tieto hodnoty sú $(i-1/2)/n$ pre $n \geq 11$ a $(i-3/8)/(n+1/4)$, pre $n \leq 10$ a sú generované v rastúcom poradí

Pozn.: Vizualizácia rozdelenia pravdepodobnosti náhodnej premennej *kvantil-kvantil grafom (qq graf)* nám elegantne ukáže, kde sa možné odchýlky od normality nachádzajú. Zobrazenie rozdelenia pravdepodobnosti q-q grafom nám ale dáva len čiastočnú informáciu. Vieme povedať, koľko modusov má naše rozdelenie, či má ťažšie alebo ľahšie "chvosty" ako normálne rozdelenie. Nevieme však povedať, či sú výchylky štatisticky "závažné" alebo nie. To nám povie napr. *qq graf s Atkinsonovu obálkou* (napr. Katina 2003). Tento graf zobrazuje teoretické hodnoty kvantilov z $N(0,1)$ oproti kvantilom sortovaných hodnôt náhodného výberu transformovaných na $N(0,1)$. Atkinsonova obálka slúži na ohraničenie odchýlok od normality, t.j. ide o porovnanie výberových qq-grafov získaných z množstva iných qq-grafov generovaných z náhodných výberov z normálneho rozdelenia. Toto horné a dolné ohraničenie získame napr. zo 100 simulácií z $N(0,1)$ s rozsahom rovným rozsahu hodnôt sledovaného parametra tak, že počítame minimá a maximá pre každý kvantil nasledovne. Nech $x_i, i = 1,2,...,n, n = 100$, sú pseudonáhodné čísla, $X_i \sim N(0,1)$. Označme j-te kvantily i-teho náhodného výberu q_{ij}, potom y-nové súradnice hraníc obálky budú také, že dolná hranica$= \min_i(q_{ij})$ a horná hranica$= \max_i(q_{ij})$, kde x-ové súradnice zodpovedajú príslušným teoretickým kvantilom $N(0,1)$.

Príklad 59 *qq graf s Atkinsonovu obálkou, dátový rámec* `swiss`

```
# klasický qqplot
swiss.df_data.frame(Fertility=swiss.fertility,swiss.x)
attach(swiss.df)
swiss.df[1:5,]
qqnorm(Infant.Mortality)
qqline(Infant.Mortality)
# Atkinsonovu obálka
vyberN01_cbind(Infant.Mortality,matrix(rnorm(47*19),47,19))
vyberN01_apply(scale(vyberN01),2,sort)
IM_vyberN01[,1]
XIM_qqnorm(IM,plot=F)$x
OBALKA_t(apply(vyberN01[,-1],1,range))
```

argumenty funkcie `matplot()`: `type=''pnn''` - nakreslí sa prvý stĺpec (`IM`) ako body, ďalšie stĺpce (`OBALKA`) sa nekreslia, ale hranice osí sa pre ne vytvoria; `pch=4` a `mkh=0.06` su typ a veľkosť bodov

```
matplot(XIM,cbind(IM,OBALKA),pch=c(4,18,18),mkh=0.5)
```

- **kruhový diagram (piechart)**

funkcia `pie(x)`
argumenty
`x` vektor relatívnych hodnôt (pravdepodobností), ktoré v súčte dávajú 1, teda i-ta položka bude `abs(x[i])/sum(abs(x))` kruhu (ale aj početností). Graf začína horizontálnou čiarou doprava a pokračuje proti smeru hodinových ručičiek
`names` vektor mien prislúchajúcich jednotlivým položkám grafu
`explode` logický vektor špecifikujúci položky, ktoré majú byť vysunuté
`col` vektor farieb, ktorými sú jednotlivé položky vyfarbené

Príklad 60 *dátový rámec* `testscores`

```
# vytvorte kruhový diagram skóre pre jedného študenta
st18 <- testscores[18,]
name <- attributes(st18)
pie(st18, names = name, col = c(3:7))
# vysuň skóre < 20%
pie(st18, names = name, col = c(3:7), explode = st18 < 20)
```

Príklad 61 *dátový rámec* `telsam.response`

```
SUManketa <- apply(telsam.response, 2, sum)
pie(SUManketa, dimnames(telsam.response)[[2]], explode = c(T, F, F, F),
 col = c(3:6))
# pridajte do grafu nadpis
title(main =''Kruhovy diagram'')
```

- **stĺpcový diagram (barplot)**

funkcia `barplot()`
`names` vektor pomenovaní pre jednotlivé stĺpce
`legend` vektor pomenovaní pre jednotlivé položky vrámci stĺpca
`col` vektor farieb jednotlivých položiek vrámci stĺpca

Príklad 62 *dátový rámec* `telsam.response`

použite prvých 5 riadkov dátového rámca `telsam.response`, kde tieto budú predstavovať 5 ľudí, ktorí zaradili do kategórií `poor`, `fair`, `good` and `excellent` prislúchajúce počty otátok v dotazníku a zobrazte ich prostredníctvom stĺpcového diagramu

```
barplot( t(telsam.response[1:5,]), ylim=c(0, 200))
# ohraničte y-ovú os tak, aby na nej bolo dosť miesta na legendu
barplot( t(telsam.response[1:5,]), ylim=c(0, 200))
# priraďte do grafu mená (čísla) participujpcich ľudí, popis x.ovej a y-ovej osi, názov grafu
# popisky x-ovej, y-ovej osi a nadpis
# xlab=''retazec'', ylab=''retazec'', main=''retazec''
barplot( t(telsam.response[1:5,]), ylim=c(0, 200),
 legend=dimnames(telsam.response)[[2]],
 names=as.character(dimnames(telsam.response)[[1]][1:5]),
 xlab=''respondent'', ylab=''pocet odpovedi'',
 main=''Odpovede respondentov'')
```

- funkcia `plot()` a jej argumenty (možno ich použiť aj pri mnohých iných grafických funkciách)

- `plot(x,y)` - ak sú `x` a `y` vektory, ide o `scatterplot` vektora `x` oproti vektoru `y`, ak zadáme len `plot(xy)`, potom je `xy` list obsahujúci dva argumenty alebo dvoj-stĺpcová matica
- `plot(x)` - ak x je časový rad, potom ide o "time series plot", ak `x` je numerický vektor, potom zobrazujeme hodnoty vektora oproti indexu vo vektore, ak `x` je komplexný vektor, zobrazuje sa imaginárna zlozka versus realna zlozka vektora

Pozn. ku Rku:
`plot(f)` - kde `f` je vektor faktorov, v R -ku pojde o barplot
`plot(f,y)` - zobrazí stĺpcový diagram vektora `y` pre všetky hladiny faktora `f`
`pairs(X)` - matica rozptylových grafov pre jednotlivé premenné ("pairwise scatterplot")

- `type=` argument kontrolujúci typ grafu: `type=''p''` - body (default), `type=''l''` - čiary, `type=''b''` - body pospájané čiarami, `type=''s''` - skoková funkcia, `type=''n''` - žiadne zobrazenie, len osi
- `xlab=''retazec'', ylab=''retazec'', main=''retazec''`, `sub=''sring''` - podnadpis pod x-ovou osou
- `points(x,y), lines(x,y)` - pridávanie bodov a čiar do obrázka
- `text(x, y, labels)` - pridanie textu v bodoch špecifikovaných `x`, `y`; teda `labels[i]` sa zobrazuje v bodoch (`x[i], y[i]`), default je `1:length(x)`

  ```
  plot(x,y, type=''n''); text(x, y, names)
  ```

- `title(main, sub)` - dodatočné pridanie nadpisu a podnadpisu
- `legend(x, y, legend)` - dodatočné pridanie legendy v špecifickom umiestnení

argumenty
`legend( , fill=v)` - farby výplne krabičiek
`legend( , col=v)` - farba nakreslených bodov alebo čiar
`legend( , lty=v)` - typ čiar
`legend( , lwd=v)` - šírka čiar
`legend( , pch=v)` - typ bodov

- určenie polohy špecifikovaného bodu na grafe použitím myši `locator(n, type)`

bodové (jedno kliknutie) určenie outliera

- `text(locator(1), ''outlier'')`

  # pozícia legendy

  ```
  legend(locator(1), legend=c(''Janko'', ''Marienka''), fill=2:3)
  ```

- identifikícia bodov `identify(x,y, labels)`

Príklad 63 *priklady ku vyššie uvedeným funkciám a ich argumentom*

```
# dátový rámec swiss.df
attach(swiss.df)
plot(Ferility,Infant.Mortality)
# dátový rámec corn
help(corn)
plot(corn.rain, corn.yield)
```

```
# dátový rámec ship
help(ship)
plot(ship)

# identitikácia
x <- c(1, 2, 3)
y <- c(3, 4, 5)
plot(x, y)
# kliknutie vždy raz na body (x[i],y[i])
body <- locator()
# označenie bodov
identify(x, y, pts = pts)
# použitie argumentu type, typu čiary a bodov
plot(1:10, type=''b'', lty=2, pty=7)

# vloženie textu do grafu, funkcia paste znamená vlož, argument sep separáciu názvov, argument
adj označuje vzdialenosť od bodu,
population <- state.x77[,''Population'']
area <- state.x77[,''Area'']
plot(area, population, log=''xy'', xlab=''Oblast v stvorcovzch milach'',
 ylab=''Populacia v tisicoch'')
staty.ozn <- c(''Alaska'', ''California'', ''Florida'', ''Hawaii'',
 ''New Jersey'', ''New York'', ''Rhode Island'', ''Texas'', ''Wyoming'')
text(area[staty.ozn], population[staty.ozn],
 paste('' '', staty.ozn, sep=''''), adj=0)

# legenda s použitím funkcie locator
plot(freeny.x[,1], ylim = c(1,10), pch = 15, col = 2)
points(freeny.x[,2], pch = 15, col = 3)
points(freeny.x[,3], pch = 15, col = 4)
typ.names <- c(''price index'', ''income level'', ''market potential'')
legend(locator(1), legend = typ.names, fill = 2:4)
# legenda s umiestnením, graf typu časový rad tsplot()
tsplot(bonds.yield[1:40,], lty = 1:6)
legend(13, .086, legend = as.character(bonds.coupon),
 lty = 1:6, pch = ''OXAC*I'')
# legenda s rôznymi typmi bodov
plot(testscores[,1], pch = 8)
points(testscores[,2], pch = 7)
points(testscores[,3], pch = 5)
legend(1,20,c(''DG'',''C'',''A''), marks = c(8,7,5))

# vloženie priamky do grafu, funkcia abline()
abline(a, b) - intercept-skoln typ čiary
abline(LRM$coef) - vyberanie regresných koeficientov z výsledkov lineárnej regresie (neskôr)
abline(LRM) - priame použitie zadania pre lineárny regresný model (neskôr)

x <- 1:100
y <- rnorm(100)
plot(x,y)
abline(0,0)
LRMxy <- lm(y~x)
abline(LRMxy) alternatívne abline(LRMxy$coef)
```

9 Štatistická inferencia

9.1 Štatistická inferencia pre parametre normálneho rozdelenia

V nasledovnej kapitole sa budeme venovať vybraným rozdeleniam pravdepodobnosti bezprostredne súvisiacich so štatistickou inferenciou, itervalom spoľahlivosti a testovaniu hypotéz.

9.1.1 Intervaly spoľahlivosti

Cieľ: zostrojiť na základe dát interval, o ktorom môžeme *dosť spoľahlivo* prehlásiť, že obsahuje skutočnú (neznámu) hodnotu parametra θ, teda tvrdíme to *s primeraným stupňom dôvery.*

Úloha: nájsť na základe náhodného výberu $X_1, ..., X_n$ taký interval, kt. bude obsahovať neznámu hodnotu parametra θ s pravdepodobnosťou $1-\alpha$. Takéto intervaly sa nazývajú $100 \times (1-\alpha)\,\%$ *intervaly spoľahlivosti* (*IS*) s *koeficientom spoľahlivosti* $1-\alpha$, $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$ - a je volená štatistikom.

Dĺžka IS:

- $\uparrow$ spoľahlivosť - dlhší IS,
- $\downarrow$ spoľahlivosť - kratší IS.

IS rozlišujeme:

- dvojstranné (*DIS*)

  pre DIS sú koncové body $DH(X_1, ..., X_n)$ a $HH(X_1, ..., X_n)$ vytvorené tak, aby platilo

$$\Pr\left(DH(X_1, ..., X_n) \leq \theta \leq HH(X_1, ..., X_n)\right) = 1-\alpha$$

- jednostranné (*JIS*)

  pre JIS volíme konkrétnu hranicu tak, aby mohlo dôjsť iba k podceneniu neznámeho parametra (pravostranný) alebo k preceneniu neznámeho parametra (ľavostranný), tj. aby platilo

  $\Pr\left(DH_*(X_1, ..., X_n) \leq \theta\right) = 1-\alpha$, pre $\forall\theta \in \Theta$ (dolný)

  $\Pr\left(\theta \leq HH^*(X_1, ..., X_n)\right) = 1-\alpha$, pre $\forall\theta \in \Theta$ (horný)

IS dostaneme použitím kvantilov, resp. kritických hodnôt rozdelenia príslušnej náhodnej premennej (v tabuľkách 1-18 sú použité kritické hodnoty).

ISs pre parametre normálneho rozdelenia:

Tab. 1

IS pre $\theta_1 = \mu, \theta_2 = \sigma^2$ - známe

$$\left(\overline{X} - u(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + u(\alpha/2)\frac{\sigma}{\sqrt{n}}\right)$$

$$\left(\overline{X} - u(\alpha)\frac{\sigma}{\sqrt{n}} \leq \mu\right)$$

$$\left(\mu \leq \overline{X} + u(\alpha)\frac{\sigma}{\sqrt{n}}\right)$$

Tab. 2

IS pre $\theta_1 = \mu, \theta_2 = \sigma^2$ - neznáme

$$\left(\overline{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}\right)$$

$$\left(\overline{X} - t_{n-1}(\alpha)\frac{S}{\sqrt{n}} \leq \mu\right)$$

$$\left(\mu \leq \overline{X} + t_{n-1}(\alpha)\frac{S}{\sqrt{n}}\right)$$

Tab. 3

$$\begin{array}{c}\text{IS pre } \theta_1=\sigma^2, \theta_2=\mu\text{- neznáme}\\ \left(\frac{(n-1)S^2}{\chi^2_{n-1}(\alpha/2)} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha/2)}\right)\\ \left(\frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha)} \leq \sigma^2\right)\\ \left(\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1}(\alpha)}\right)\end{array}$$

Označenia:

Normované dvojrozmerné normálne rozdelenie

$$N_2\left((0,0)^T, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

je rozdelenie normálneho vektora $(X,Y)^T$ s *hustotou*

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right\}, (x,y)^T \in \mathcal{R}^2,$$

kde $\rho \in (-1,1)$ je parameter; *distribučná funkcia* je

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{w^2-2\rho wz+z^2}{2(1-\rho^2)}\right\} dwdz, (x,y)^T \in \mathcal{R}^2.$$

Marginálne rozdelenie X a Y je $N(0,1)$, kde $E[X]=E[Y]=0$, $Var[X]=Var[Y]=1$, $Cov(X,Y)=\rho$ a

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right\} dy \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right\}, \end{aligned}$$

kde $\frac{1}{2\pi\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right\}, y \in \mathcal{R}^1$ je hustota $N\left(\rho x, 1-\rho^2\right)$. Stredné hodnoty a rozptyly plynú zo skutočnosti, že náhodné veličiny X a Y majú normované normálne rozdelenie. Pre kovarianciu máme

$$\begin{aligned} Cov(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right\} dxdy \\ &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \left(\int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right\} dy\right) dx \\ &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \rho x dx = \rho. \end{aligned}$$

Všeobecné dvojrozmerné normálne rozdelenie

$$N_2\left((\mu_1,\mu_2)^T, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

je rozdelenie normálneho vektora $(X,Y)^T$ s hustotou

$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left\{ \begin{array}{c} \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} \\ +\frac{(y-\mu_2)^2}{\sigma_2^2} \end{array} \right\}\right\},$$

kde $(x,y)^T \in \mathcal{R}^2$, $\mu_i \in \mathcal{R}^1$, $\sigma_i^2 > 0$, $i = 1,2$, $\rho \in (-1,1)$ sú parametre. Výraz v exponente môžeme písať ako

$$-\frac{1}{2}\begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix},$$

marginálne rozdelenia sú $N\left(\mu_1, \sigma_1^2\right)$, resp. $N\left(\mu_2, \sigma_2^2\right)$ a ρ je koeficient korelácie.

Veta 64 *Majme 2 nezávislé náhodné výbery s rozsahmi n_1, n_2 so základých súborov s rozdeleniami $N\left(\mu_1,\ \sigma_1^2\right)$, resp. $N\left(\mu_2,\ \sigma_2^2\right)$. Potom*

1. $\overline{X} = \overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$

   $U = \frac{\overline{X} - E[X]}{\sqrt{D[X]}} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma_D} \sim N(0,1)$, *kde* $\sigma_D^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2$, σ_1^2 *a* σ_2^2 *sú známe*

2. *špeciálne ak* $N\left(\mu_1, \sigma^2\right)$, *resp.* $N\left(\mu_2, \sigma^2\right)$, *kde* σ^2 *je známa, potom*

   $\overline{X} = \overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{n_1+n_2}{n_1 n_2}\sigma^2\right)$

   $U = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sigma} \sim N(0,1)$

3. *ak* σ^2 *je neznáma (teda predpokladáme rovnosť výberových disperzií), potom*

   $T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_{odh}} \sim t_{n_1+n_2-2}$, *kde* $S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$, $S_{odh} = S\sqrt{1/n_1 + 1/n_2}$, S_1^2 *a* S_2^2 *sú výberové disperzie a*

   $(n_1 + n_2 - 2)\, S^2/\sigma^2 \sim \chi^2_{n_1+n_2-2}$

4. *podiel výberových disperzií ku populačným disperziám má nasledovné rozdelenie*

   $$\frac{S_{1,n_1}^2/\sigma_1^2}{S_{2,n_2}^2/\sigma_2^2} \sim F_{n_1,n_2}$$

5. σ_1^2 *a* σ_2^2 *sú neznáme (teda predpokladáme, že výberové disperzie sú rôzne) a často* n_1, n_2 *sú malé a nevelké - je nutné použiť* ***Aspin - Welchovu aproximáciu*** *(Aspin a Welch, 1949, Wang, 1971; Bickel, Doksum, 2002)*

   $c = S_{1,n_1}^2 / \left(n_1 S_D^2\right), S_D^2 = S_1^2/n_1 + S_2^2/n_2$

   potom máme T_k, *kde k je počet stupňov voľnosti daný vzťahom*

   $k = \left[\frac{c^2}{n_1-1} + \frac{(1-c^2)}{n_2-1}\right]^{-1}$ *(ak k nie je celé číslo, použijeme lineárnu interpoláciu v t-tabuľkách).*

Veta 65 *Nech* $(X_1,Y_1), (X_2,Y_2), ..., (X_n,Y_n)$ *je náhodný výber z*

$$N_2\left((\mu_1,\mu_2)^T, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

$\mu_1,\ \mu_2 \in \mathcal{R}^1$, $\sigma_1^2, \sigma_2^2 > 0$, $-1 < \rho < 1$, *všetky parametre neznáme. Nech* $\overline{D}_n = \sum_{i=1}^n D_i/n$ *a* $S_{d,n}^2 = \sum_{i=1}^n \left(D_i - \overline{D}_n\right)/(n-1)$ *sú výberový priemer a výberový rozptyl veličín* $D_i = X_i - Y_i$, $i = 1,2,...,n$. *Potom*

$$\left(\overline{D}_n - t_{n-1}(\alpha/2)\frac{S_{d,n}}{\sqrt{n}}, \overline{D}_n + t_{n-1}(\alpha/2)\frac{S_{d,n}}{\sqrt{n}}\right)$$

je intervalový odhad parametrickej funkcie $\Delta = \mu_1 - \mu_2$ *so spoľahlivosťou* $(1-\alpha)$.

Pozn.:

- Welchova aproximácia dobre pracuje aj za rovnosti rozptylov. Problému nerovnosti rozptylov hovoríme aj **Behrens - Fisherov problém** (Behrens, 1929; Fisher, 1939)
- pre výpočet df sa uvádza aj nasledovný vzťah (Welch , 1938)

  $k = df = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(S_1^2/n_1\right)^2/(n_1-1) + \left(S_2^2/n_2\right)^2/(n_2-1)}$ (Zar, 1999), použijeme nasledujúce menšie cele číslo.

Tab. 4

$$\begin{gathered}\text{IS pre } \theta_1 = \mu_1 - \mu_2, \theta_2 = \left(\sigma_1^2, \sigma_2^2\right) \text{ - známe} \\ \left(\overline{X}_1 - \overline{X}_2 - u\left(\alpha/2\right)\sigma_D \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + u\left(\alpha/2\right)\sigma_D\right) \\ \left(\overline{X}_1 - \overline{X}_2 - u\left(\alpha\right)\sigma_D \le \mu_1 - \mu_2\right) \\ \left(\mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + u\left(\alpha\right)\sigma_D\right)\end{gathered}$$

Tab. 5

$$\begin{gathered}\text{IS pre } \theta_1 = \mu_1 - \mu_2, \theta_2 = \sigma^2 \text{ - známe} \\ \left(\overline{X}_1 - \overline{X}_2 - u\left(\alpha/2\right)\sqrt{\tfrac{n_1+n_2}{n_1 n_2}}\sigma \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + u\left(\alpha/2\right)\sqrt{\tfrac{n_1+n_2}{n_1 n_2}}\sigma\right) \\ \left(\overline{X}_1 - \overline{X}_2 - u\left(\alpha\right)\sqrt{\tfrac{n_1+n_2}{n_1 n_2}}\sigma \le \mu_1 - \mu_2\right) \\ \left(\mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + u\left(\alpha\right)\sqrt{\tfrac{n_1+n_2}{n_1 n_2}}\sigma\right)\end{gathered}$$

Tab. 6

$$\begin{gathered}\text{IS pre } \theta_1 = \mu_1 - \mu_2, \theta_2 = \sigma^2 \text{ - neznáme} \\ \overline{X}_1 - \overline{X}_2 - t_{n_1+n_2-2}\left(\alpha/2\right)S_{odh} \le \mu_1 - \mu_2 \le +t_{n_1+n_2-2}\left(\alpha/2\right)S_{odh} \\ \left(\overline{X}_1 - \overline{X}_2 - t_{n_1+n_2-2}\left(\alpha\right)S_{odh} \le \mu_1 - \mu_2\right) \\ \left(\mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + t_{n_1+n_2-2}\left(\alpha\right)S_{odh}\right)\end{gathered}$$

Tab. 7

$$\begin{gathered}\text{IS pre } \theta_1 = \mu_1 - \mu_2, \theta_2 = \left(\sigma_1^2, \sigma_2^2\right) \text{ - neznáme} \\ \left(\overline{X}_1 - \overline{X}_2 - t_k\left(\alpha/2\right)S_D \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + t_k\left(\alpha/2\right)S_D\right) \\ \left(\overline{X}_1 - \overline{X}_2 - t_k\left(\alpha\right)S_D \le \mu_1 - \mu_2\right) \\ \left(\mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + t_k\left(\alpha\right)S_D\right)\end{gathered}$$

Tab. 8

$$\begin{gathered}\text{IS pre } \theta_1 = \sigma_1^2/\sigma_2^2, \theta_2 = \mu \text{ - neznáme} \\ \left(\frac{S_{1,n_1}^2}{S_{2,n_2}^2}\frac{1}{F_{n_1-1,n_2-1}(\alpha/2)} \le \sigma_1^2/\sigma_2^2 \le \frac{S_{1,n_1}^2}{S_{2,n_2}^2}\frac{1}{F_{n_1-1,n_2-1}(1-\alpha/2)}\right) \\ \left(\frac{S_{1,n_1}^2}{S_{2,n_2}^2}\frac{1}{F_{n_1-1,n_2-1}(1-\alpha)} \le \sigma_1^2/\sigma_2^2\right) \\ \left(\sigma_1^2/\sigma_2^2 \le \frac{S_{1,n_1}^2}{S_{2,n_2}^2}\frac{1}{F_{n_1-1,n_2-1}(\alpha)}\right)\end{gathered}$$

9.1.2 Testovanie hypotéz

Základné pojmy:

Nulová hypotéza $H_0 : \theta \in \Theta_0$,
Alternatívna hypotéza $H_1 : \theta \in \Theta_1, \Theta_0 + \Theta_1 = \Theta$
Možné situácie:

A.) platí H_0 a naše rozhodutie je nezamietnuť H_0 (SPRÁVNE),

B.) platí H_0 a naše rozhodutie je zamietnuť H_0,

C.) platí H_1 a naše rozhodutie je nezamietnuť H_0,

D.) platí H_1 a naše rozhodutie je zamietnuť H_0 (SPRÁVNE).

Teda

rozhodnutie/skutočnosť	H_0 platí	H_0 neplatí
H_0 zamietnuť	chyba I. druhu	správne rozhodnutie
H_0 nezamietnuť	správne rozhodnutie	chyba II. druhu

- V prípade A, D je naše rozhodnutie správne ($\Pr(A) \geq 1-\alpha$).
- V prípade B sa dopúšťame *chyby prvého druhu (CHPD,* $\Pr(CHPD) \leq \alpha$*).*
- V prípade C sa dopúšťame *chyby druhého druhu (CHDD,* $\Pr(CHDD) = \beta$*).* Doplnok pravdepodobnosti *CHDD* $1-\beta$ sa nazýva *sila testu*, čo je pravdepodobnosť, že H_0 zamietneme, keď táto hypotéza neplatí (D), teda pravdepodobnosť, s akou neplatnosť hypotézy odhalíme. Sila testu závisí na zvolenej testovacej metóde a hlavne na tom, *aké je skutočné rozdelenie dát* (a teda použitej štatistiky) alebo aké sú skutočné hodnoty parametrov.

Kritický obor $W \in \mathcal{R}^n$. Ak $X \in W$, tak H_0 zamietame. Jeho voľba závisí na požiadavke, aby pravdepodobnosťi chýb I. druhu boli menšie alebo rovné zvolenému kladnému číslu $\alpha \in (0, 1/2)$, teda $\Pr_\theta(X \in W) \leq \alpha, \forall\theta \in \Theta_0$. Platí $sup_{\theta\in\Theta_0} \Pr_\theta(X \in W) = \alpha$, α sa nazýva *hladina významnosti.* Súčasne volíme W tak, aby pravdepodobnosti chýb I. druhu boli čo najmenšie.

Pozn.:

- *hladina významnosti* α - daná (určená štatistikom),
- *testovacia štatistika* - vypočíta sa,
- *kritická hodnota* - tabuľky, PC,
- *p-hodnota, p-value, significance level* (probability), dosiahnutá hladina významnosti

Zamietnutie (napr. H_0:$\theta = \theta_0$)

- dvojstranná alternatíva je napr. pre $X \sim N(0,1)$ ak $|U| \geq u(\alpha/2)$, kde
  $H_1 : \theta \neq \theta_0$
- jednostranná alternatíva je napr. pre $X \sim N(0,1)$ ak $U \geq u(\alpha), U \leq -u(\alpha)$, kde
  pravostranná $H_1 : \theta > \theta_0$
  ľavostranná $H_1 : \theta < \theta_0$

Tab. 9

H_0	H_1	$X \in W$	predpoklad
$\mu = \mu_0$	$\mu \neq \mu_0$	$U = \frac{\lvert\overline{X}-\mu_0\rvert}{\sigma}\sqrt{n} \geq u(\alpha/2)$	σ^2 - známe
$\mu \leq \mu_0$	$\mu > \mu_0$	$U_1 = \frac{\overline{X}-\mu_0}{\sigma}\sqrt{n} \geq u(\alpha)$	σ^2 - známe
$\mu \geq \mu_0$	$\mu < \mu_0$	$U_2 = \frac{\overline{X}-\mu_0}{\sigma}\sqrt{n} \leq -u(\alpha)$	σ^2 - známe

Tab. 10

H_0	H_1	$X \in W$	predpoklad
$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\left\vert \overline{X} - \mu_0 \right\vert}{S} \sqrt{n} \geq t_{n-1}(\alpha/2)$	σ^2 - neznáme
$\mu \leq \mu_0$	$\mu > \mu_0$	$T_1 = \frac{\overline{X} - \mu_0}{S} \sqrt{n} \geq t_{n-1}(\alpha)$	σ^2 - neznáme
$\mu \geq \mu_0$	$\mu < \mu_0$	$T_2 = \frac{\overline{X} - \mu_0}{S} \sqrt{n} \leq -t_{n-1}(\alpha)$	σ^2 - neznáme

Tab. 11

H_0	H_1	$X \in W$	predpoklad
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2} \notin \left(\chi^2_{n-1}(1-\alpha/2), \chi^2_{n-1}(\alpha/2)\right)$	μ - neznáme
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2} \geq \chi^2_{n-1}(\alpha)$	μ - neznáme
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{n-1}(1-\alpha)$	μ - neznáme

Tab. 12

H_0	H_1	$X, Y \in W$	predpoklad
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$U = \frac{\left\vert \overline{X}_1 - \overline{X}_2 \right\vert}{\sigma_D} \sqrt{n} \geq u(\alpha/2)$	$\left(\sigma_1^2, \sigma_2^2\right)$ - známe
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$U_1 = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_D} \geq u(\alpha)$	$\left(\sigma_1^2, \sigma_2^2\right)$ - známe
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$U_2 = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_D} \leq -u(\alpha)$	$\left(\sigma_1^2, \sigma_2^2\right)$ - známe

Tab. 13

H_0	H_1	$X, Y \in W$	predpoklad
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$U = \frac{\left\vert \overline{X}_1 - \overline{X}_2 \right\vert}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sigma} \geq u(\alpha/2)$	σ^2 - známe
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$U_1 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sigma} \geq u(\alpha)$	σ^2 - známe
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$U_2 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sigma} \leq -u(\alpha)$	σ^2 - známe

Tab. 14

H_0	H_1	$X, Y \in W$	predpoklad
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$T = \frac{\left\vert \overline{X}_1 - \overline{X}_2 \right\vert}{S_{odh}} \geq t_{n_1+n_2-2}(\alpha/2)$	σ^2 - neznáme
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$T_1 = \frac{\overline{X}_1 - \overline{X}_2}{S_{odh}} \geq t_{n_1+n_2-2}(\alpha)$	σ^2 - neznáme
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$T_2 = \frac{\overline{X}_1 - \overline{X}_2}{S_{odh}} \leq t_{n_1+n_2-2}(\alpha)$	σ^2 - neznáme

Tab. 15

H_0	H_1	$X, Y \in W$	predpoklad
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$T = \frac{\left\vert \overline{X}_1 - \overline{X}_2 \right\vert}{S_D} \geq t_k(\alpha/2)$	$\left(\sigma_1^2, \sigma_2^2\right)$ - neznáme
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$T_1 = \frac{\overline{X}_1 - \overline{X}_2}{S_D} \geq t_k(\alpha)$	$\left(\sigma_1^2, \sigma_2^2\right)$ - neznáme
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$T_2 = \frac{\overline{X}_1 - \overline{X}_2}{S_D} \leq -t_k(\alpha)$	$\left(\sigma_1^2, \sigma_2^2\right)$ - neznáme

Tab. 16

H_0	H_1	$X, Y \in W$	predpoklad
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$\frac{S^2_{1,n_1}}{S^2_{2,n_2}} \notin \left(F_{n_1-1,n_2-1}(1-\alpha/2), F_{n_1-1,n_2-1}(\alpha/2)\right)$	μ - nezname
$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$\frac{S^2_{1,n_1}}{S^2_{2,n_2}} \geq F_{n_1-1,n_2-1}(\alpha)$	μ - nezname
$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$\frac{S^2_{1,n_1}}{S^2_{2,n_2}} \leq F_{n_1-1,n_2-1}(1-\alpha)$	μ - nezname

9.1.3 Závislé pozorovania

IS a testovacia štatistika pre párovy t-test:

- $X_1, X_2, ..., X_n$ nezávislé, $Y_1, Y_2, ..., Y_n$ nezávislé ale X_i, Y_i nemusia byť,
- $D_i = X_i - Y_i$ nezávislé,
- $X_1, X_2, ..., X_n$ a $Y_1, Y_2, ..., Y_n$ majú rovnaké rozdelenie,
- $X_i \sim N\left(\mu_1, \sigma_1^2\right), Y_i \sim N\left(\mu_2, \sigma_2^2\right)$,
- X, Y nie su nezávislé, (X_i, Y_i) normálne rozdelené,
- $X_i - Y_i \sim N\left(\mu_1 - \mu_2, \sigma^2\right)$,
- $(x_i, y_i) -$ pár pozorovaní (realizácií) na i-tej štatistickej jednotke, $i = 1, 2, ..., n$,
- $d_i = x_i - y_i$,
- $\overline{d}$ je aritmetický priemer rozdielov d_i,
- S_d - štandardná odchýlka rozdielov d_i,
- $H_0 : \mu_1 - \mu_2 = \mu_d = 0, H_1 : \mu_d \neq 0$ alebo alternatívy jednostranné

Tab. 17

IS pre $\theta_1 = \mu, \theta_2 = \sigma^2$ - neznáme

$$\left(\overline{D} - t_{n-1}(\alpha/2)\frac{S_d}{\sqrt{n}} \leq \mu \leq \overline{D} + t_{n-1}(\alpha/2)\frac{S_d}{\sqrt{n}}\right)$$

$$\left(\overline{D} - t_{n-1}(\alpha)\frac{S_d}{\sqrt{n}} \leq \mu\right)$$

$$\left(\mu \leq \overline{D} + t_{n-1}(\alpha)\frac{S_d}{\sqrt{n}}\right)$$

Tab. 18

H_0	H_1	$X \in W$	predpoklad
$\mu_d = 0$	$\mu_d \neq 0$	$T = \frac{\lvert\overline{D}\rvert}{S_d}\sqrt{n} \geq t_{n-1}(\alpha/2)$	σ^2 - neznáme
$\mu_d \leq 0$	$\mu_d > 0$	$T_1 = \frac{\overline{D}}{S_d}\sqrt{n} \geq t_{n-1}(\alpha)$	σ^2 - neznáme
$\mu_d \geq 0$	$\mu_d < 0$	$T_2 = \frac{\overline{D}}{S_d}\sqrt{n} \leq -t_{n-1}(\alpha)$	σ^2 - neznáme

9.1.4 Vzťah IS a testovania hypotéz

Nech $\mathbf{X} = (X_1, ..., X_n)^T$ je náhodný výber z rozdelenia, ktoré závisí na parametri $\theta \in \Theta$, $g(\theta)$ je parametrická funkcia. Testujme hypotézu $H_0 : g(\theta) = \gamma_0$ oproti obojstrannej alternatíve $H_1 : g(\theta) \neq \gamma_0$. Nech $(DH(\mathbf{X}), HH(\mathbf{X}))$ je intervalový odhad parametrickej funkcie $g(\theta)$ so spoľahlivostou $1 - \alpha$. Potom

$$W = \{\mathbf{X} \in \mathcal{R}^n; \gamma_0 \notin (DH(\mathbf{X}), HH(\mathbf{X}))\}$$

je kritický obor testu H_0 oproti H_1 na hladine významnosti α.

Ak máme testovat H_0 oproti jednostrannej alternatíve $H_1 : g(\theta) > \gamma_0$ a ak je $DH_*(\mathbf{X})$ dolný odhad pre $g(\theta)$ so spoľahlivostou $1 - \alpha$, potom

$$W = \{\mathbf{X} \in \mathcal{R}^n; DH_*(\mathbf{X}) > \gamma_0\}$$

je kritický obor testu H_0 oproti H_1 na hladine významnosti α. Obdobne pre alternatívu $H_1 : g(\theta) < \gamma_0$ a horný odhad $HH^*(\mathbf{X})$.

9.1.5 Príklady v S-PLUS a R

Príklad 66 *EDA graf*

```
eda.shape_function(x)
# EDA - exploratórna analýza dát
# grafy môžete vylepšiť podľa vedomostí o grafike z predchádtajúcich prednášok
{
par(mfrow=c(2,2))
hist(x)
boxplot(x)
IQD_summary(x)[5]-summary(x)[2]
plot(density(x,width=2*IQD),xlab=''x'', ylab='''', type=''l'')
qqnorm(x)
qqline(x)
}
```

Príklad 67 20 *pozorovaní rýchlosti svetla, "mich" dáta (Michelson, 1879)*

```
# zadávanie dát funkciou "scan"
mich_scan()
1:  850 740 900 1070 930
6:  850 950 980 980 880
11:  1000 980 930 650 760
16:  810 1000 1000 960 960 21
eda.shape(mich)
summary(mich)
```

Príklad 68 *funkcia ''t.test'' a jej argumenty*

```
# vstupy
# "mu" - stredná hodnota alebo rozdiel stredných hodnôt chatakteriyovaný nulovou hypotézou
# spoľahlivosť conf.level=CISLO, default je conf.level=.95
# formulácia alternatívy alternative=''two.sided'' je default, ďalšie voľby sú ''greater'', ''less''
# jednovýberový t test = zadáme "x" ako vektor dát, "mu" ako strednú hodnotu za platnosti nulovej hypotézy
# dvojvýberový t test = zadáme "x" a "y" ako vektory dát, "var.equal=T" rovnaké rozptyly a "var.equal=F" rôzne rozptyly
# párový t test = zadáme "x" a "y" ako vektory dát, stanovíme argument paired=T, defult je paired=F
# výstupy
# názov použitého testu "method"
# testovacia štatistika "t"
# počet stupňov voľnosti "df"
# p-hodnota "p-value"
# alternatívna hypotéza "alternative hypothesis"
# interval spoľahlivosti "conf.int"
# bodové odhady (aritmetické priemery) "sample estimates"
```

Príklad 69 *jednovýberový t test, "mich" dáta*

```
t.test(mich,mu=990)
t.test(mich,conf.level=.90,mu=990)
```

Príklad 70 *dvojvýberový t test, "váhový prírastok u potkanov"*

```
gain.high_scan()
134 146 104 119 124 161 107 83 113 129 97 123
gain.low_scan()
70 118 101 85 107 132 94
eda.shape(gain.high)
eda.shape(gain.low)
# var.test(gain.high,gain.low)
t.test(gain.high,gain.low)
t.test(gain.high,gain.low,alternative=''g'')
```

Príklad 71 *párový t test - "páry topánok"*

```
wear.A_scan()
14.0 8.8 11.2 14.2 11.8 6.4 9.8 11.3 9.3 13.6
wear.B_scan()
13.2 8.2 10.9 14.3 10.7 6.6 9.5 10.8 8.8 13.3
eda.shape(whear.A-whear.B)
plot(wear.A,wear.B)
```

indikuje vysokú koreláciu, ktorá signalizuje vysokú vnutro-výberovú variabilitu, väčšiu než rozdiely v priemeroch. Párovanie vedie k väčšej citlivosti testu

```
t.test(wear.A,wear.B,paired=T)
```

Príklad 72 *naprogramujte vyššie spomínané testy zo základných funkcií, pomenujte ich "jednovyberovy.ttest", "dvojvyberovy.ttest" a "parovy.ttest*

```
# pomôcky
# jednovýberový t test urobte len pre "mu=0"
Tstat_ mean(x) / ( sqrt(var(x)) / sqrt(length(x)) )
# párový t test
# najprv naprogramujte rozdiel vektorov "d" a potom pre "mu=0"
Tstat_mean(d) / ( sqrt(var(d)) / sqrt(length(d)) )
# dvojvýberový t test, tiež pre "mu=0"
Tstat_(mean(x) - mean(y)) / s1
s1 _ sp * sqrt(1/nx + 1/ny)
sp _ sqrt( ( (nx-1)*var(x) + (ny-1)*var(y) ) / (nx + ny - 2) )
nx _ length(x)
ny = length(y)
# dvojvýberový t test s Welchovou aproximáciou, tiež pre "mu=0"
Tstat_(mean(x) - mean(y)) / s2
s2 _ sqrt( var(x)/nx + var(y)/ny ),
nx _ length(x)
ny _ length(y).
# pre "df"
1 / ( (c^2)/(nx-1) + ((1-c)^2)/(ny-1) )
c _ var(x) / (nx * s2^2)
```

Príklad 73 *naprogramujte intervaly spoľahlivosti pre vyššie spomínané testy zo základných funkcií a pomenujte ich"jednovyberovy.IS", "dvojvyberovy.IS" a "parovy.IS", použite podklady z predchádzajúceho príkladu*

Príklad 74 *Monte Carlo (MC) simulácie*

Poznámky ku metodike:

Aproximácie na základe asymptotických výsledkov možu byť vytvorené MC simuláciami. Ukážeme simulácie na t-teste generovaním dát z χ^2_k rozdelenia m - krát nezávisle. Vždy spočítame hodnotu $t-$štatistiky a potom spočítame proporciu počtov z celkového počtu m (ďalej `n.sim`), ktoré prevyšujú kritickú hodnotu z t-rozdelenia. χ^2_k rozdelenie sme použili preto, aby sme ukázali, že pre malé k máme výchylky od normálneho rozdelenia.

Poznámky ku výsledkom:

Pre jednovýberovy t-test na $\alpha = 0.05$ sú asymptotické výsledky s dobrou aproximáciou, keď $n \geq 10^{1.5} \approxeq 32, k \geq 10$. χ^2_2 rozdelenie je extrémne vychýlene a pre $t_{n-1}(0.95)$ je aproximácia dobrá len s $n \geq 10^{2.5} \approxeq 316$.

Úloha:

Urobte simulácie pre `n.sim`=10 000 t-testov s použitím χ^2_k nagenerovaných dát, kde $k = 2, 10, 20, 50$. Simulácie opakujte pre rôzne rozsahy výberov a zobrazte pozorované hladiny významnosti voči $\log_{10}$ (rozsah vyberu)

```
# "kv" - kvantil t-rozdelenia
# rozsah výberu je "n"
kv <- qt(0.975, df=n)
''SIMULttest'' <- function(n.sim, n, df, kv)
{
# n.sim - pocet simulacii (# = 10000)
# n - rozsah vyberu
# df - pocet stupnov volnosti chi-kvadrat rozdelenia

T.stat <- rep(0,n.sim)
Matsim <- matrix(0,n,n.sim)
for (i in 1:  n.sim){
MATsim[,i] <- rchisq(n.sim, df)
Tstat[i]_ mean(MATsim[,i]) / ( sqrt(var(MATsim[,i])) / sqrt(n))
}
PROP <- length(Tstat[Tstat > kv])/n.sim
return(PROP)
}
```

9.2 Štatistická inferencia pre parametre binomického rozdelenia

Majme $H_0 : p = p_0$ oproti $H_0 : p \neq p_0$. Potom **Waldova štatistika** pre H_0 má tvar

$$z_W = \frac{\widehat{p} - p_0}{\sqrt{\widehat{p}\,(1-\widehat{p})\,/n}} \sim N(0,1),$$

a **Waldov približný** $(1-\alpha)\%$ **IS**

$$p_0 \in \left(\widehat{p} \pm u_{\alpha/2}\sqrt{\widehat{p}\,(1-\widehat{p})\,/n}\right).$$

Pokiaľ nie je n dosť veľké, použijeme modifikovanú z_W štatistiku v tvare

$$z_{W_{mod}} = \frac{|X - np_0| - 1/2}{\sqrt{n\widehat{p}\,(1-\widehat{p})}} \sim N(0,1),$$

kde $\hat{p} = x/n$.

Nech p_1 je pravdepodobnosť, že v pokuse nastane jav A. Nech v n_1 nezávislých pokusoch nastal jav A spolu $X-$krát. Potom, ak napr. opakujeme pokus A za iných podmienok, jav A nastáva s pravdepodobnosťou p_2. Nech v n_2 týchto ďalších nezávislých pokusoch nastal jav A spolu $Y-$krát. Na základe týchto údajov chceme testovať $H_0 : p_1 = p_2$ oproti $H_1 : p_1 \neq p_2$. Túto hypotézu nazývame niekedy *hypotéza homogenity dvoch binomických rozdelení*, pretože $X \sim Bin(n_1, p_1)$ a $Y \sim Bin(n_2, p_2)$. Označme $\hat{p}_1 = x/n_1, \hat{p}_2 = y/n_2$. Predpokladajme, že $p_1(1-p_1) + p_2(1-p_2) \neq 0$.

Z CLV vyplýva, že pri dostatočne veľkých hodnotách n_1 a n_2 môžeme použiť aproximáciu

$$p_1 \sim N\left[p_1, \frac{p_1(1-p_1)}{n_1}\right], p_2 \sim N\left[p_2, \frac{p_2(1-p_2)}{n_2}\right].$$

Keďže x a y sú nezávislé veličiny, dostávame potom

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1).$$

Ak plati H_0, dostaneme v čitateli $p_1 = p_2$. V menovateli však neznáme hodnoty p_1 a p_2 stále ostávajú. Preto sa za ne dosadzujú ich odhady (Waldov prístup). Dá sa dokázať, že sa limitne rozdelenie ani potom nezmení a ostane $N(0,1)$.

Ako odhad p_1 použijeme $\hat{p}_1$, ako odhad p_2 zasa $\hat{p}_2$. Silný zákon veľkých čísel zaručuje, že $\hat{p}_1 \to p_1$ a $\hat{p}_2 \to p_2$ skoro všade. Potom

$$U_a = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}},$$

kde ak $|U_a| \geq u_{\alpha/2}$, zamietame H_0. Podobne testujeme jednostranné hypotézy.

9.2.1 Príklady v S-PLUS a R

Príklad 75 *Experiment - z* 50 *zubov vystaveným tepelnému šoku sa* 21 *zlomilo, vypočítajte IS pre populačnú pravdepodobnosť zlomenia zuba po tepelnom šoku*

```
# funkcia na výpočet dvojstranného IS pre rozdiel pravdepodobností
# DH - dolná hranica IS
# HH - horná hranica IS
ProbIS <- function(k, n, alpha = 0.05)
{
# k - pocet priaznivych vysledkov
# n - pocet pozorovani
p <- k/n
stderr <- sqrt(p * (1 - p)/n1)
DH <- rozd - qnorm(1 - alpha/2) * stderr
HH <- rozd - qnorm(1 - alpha/2) * stderr
return(cbind(DH,p, HH))
}
```

Príklad 76 *Experiment - z* 50 *zubov vystaveným tepelnému šoku sa* 21 *zlomilo. Z kontrolných* 50 *zubov sa zlomilo len* 11. *Otestujte, či rozdiel populačných pravdepodobností zlomenia zuba po tepelnom šoku a u kontroly je rovnaká.*

Riešenie:
$n_1 = n_2 = 50, X = 21, Y = 11, p_1 = 0.42, p_2 = 0.22, p = 0.32$
$U_a = 2.195, U_b = 2.144, u_{0.05} = 1.645$
funkcia na výpočet dvojstranného IS pre rozdiel pravdepodobností
DH - dolná hranica IS

```
# HH - horná hranica IS
ProbDiffIS <- function(k1, k2, n1, n2, alpha = 0.05)
{
# k1, k2 - pocet priaznivych vysledkov
# n1, n2 - pocty pozorovani
p1 <- k1/n1
p2 <- k2/n2
rozd <- p1 - p2
stderr <- sqrt(p1 * (1 - p1)/n1 + p2 * (1 - p2)/n2)
DH <- rozd - qnorm(1 - alpha/2) * stderr
HH <- rozd - qnorm(1 - alpha/2) * stderr
return(cbind(DH,rozd, HH))
}
```

9.3 Neparametrické dvojvýberové testy

Predpokladajme, že $X_1, ..., X_{n_1}$ je náhodný výber z nejakého spojitého rozdelenia, $Y_1, ..., Y_{n_2}$ je náhodný výber z rovnakého spojitého rozdelenia a je oproti prvému rozdeleniu posunuté o nejakú konštantu δ. Znamená to, že veličiny $X_1, ..., X_{n_1}$ a $Y_1 - \delta, ..., Y_{n_2} - \delta$ majú rovnaké rozdelenie. Predpokladá sa, že oba výbery sú nezávislé. Potom

- $H_0 : \delta = 0$,
- $H_1 : \delta \neq 0$.

9.3.1 Wilcoxonov test

Nech $n_1 + n_2 = n$, kde n_1 je počet pozorovaní v prvom výbere, n_2 je počet pozorovaní v druhom výbere. Nech $R_1, R_2, ..., R_{n_1}$ sú poradia prvého výberu v rámci usporiadaného združeného výberu. Potom pre *Wilcoxonovu štatistiku* S_W platí

$$W_X = S_W = \sum_{i=1}^{n_1} R_i.$$

Pre strednú hodnotu a rozptyl S_W za platnosti H_0 platí

$$E_0 [S_W] = \frac{n_1 (n_1 + n_2 + 1)}{2},$$

$$Var_0 [S_W] = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12},$$

kde $n_1, n_2 \geq 10$. Potom platí

$$Z = \frac{S_W - E_0 [S_W]}{\sqrt{Var_0 [S_W]}} \sim N (0, 1) .$$

Test prislúchajúci S_W sa nazýva *Wilcoxonov test.*

Pozn.: Nulovú hypotézu zamietame, ak priemery takto získaných poradí spočítané pre každý pôvodný výber sa od seba príliž líšia (dáta nie sú dosť premiešané). Prakticky je však výhodnejšie počítať súčet poradí len v jednom z výberov. Platí

$$W_X + W_Y = (n_1 + n_2) (n_1 + n_2 + 1)/2,$$

čo platí vždy. Pri dostatočne veľkom výbere použijeme štatistiku z ($n_1 + n_2 \geq 12$). Nulovú hypotézu zamietame, ak $|z| \geq u_{\alpha/2}$.

S ohľadom na zhody sa používa modifikovaná z-štatistika, bližšie pozri Hollander, Wolfe (1999). V prípade zhôd sa urobí priemer poradí zhôd (*strednoporadie*). V tomto prípade musíme výpočet smerodajnej odchýlky čitateľa upraviť a dostaneme

$$Z_{kor} = \frac{S_W - E_0\left[S_W\right]}{\sqrt{Var_0\left[S_W\right] - \frac{n_1 n_2 \sum_j \left(t_j^3 - t_j\right)}{12(n_1+n_2)(n_1+n_2-1)}}} \sim N\left(0,1\right),$$

kde t_j su počty zhodných pozorovaní.

9.3.2 Mann-Whitney test

Tento test ekvivalentný s vyššie uvedeným Wilkoxonovým testom.

Nech $\{x_i\}_{i=1}^{n_1}$ a $\{y_j\}_{j=1}^{n_2}$ sú množiny pozorovaní v prvom výbere (napr. pôvodný typ liečby), resp. v druhom výbere (napr. nový typ liečby). Nech (x_i, y_j) sú možné páry pozorovaní, medzi ktorými môžu nastať nasledovné dve situácie $x_i < y_j$ a $x_i > y_j$. Potom platí

$$S_{MW} = \#\left(x_i, y_j\right), \text{ kde } x_i > y_j,$$

$$S_{MW} = S_W - \frac{1}{2} n_1 \left(n_1 + 1\right),$$

kde S_{MW} nazývame *Mann-Whitney štatistika (S_{MW}).* Pre strednú hodnotu a rozptyl S_{MW} za platnosti nulovej hypotézy platí

$$E_0\left[S_{MW}\right] = \frac{n_1 n_2}{2},$$

$$Var_0\left[S_{MW}\right] = Var_0\left[S_W\right] = \frac{n_1 n_2 \left(n_1 + n_2 + 1\right)}{12},$$

Potom

$$Z = \frac{S_{MW} - E_0\left[S_{MW}\right]}{\sqrt{Var_0\left[S_{MW}\right]}} \sim N\left(0,1\right).$$

Pozn.: Test je založený na veličinách

$$U_X = n_1 n_2 + \frac{n_1\left(n_1+1\right)}{2} - W_X, U_Y = n_1 n_2 + \frac{n_2\left(n_2+1\right)}{2} - W_Y.$$

Veličina U_X vyjadruje počet dvojíc X_i, Y_j, kde platí $X_i < Y_j$, podobne U_Y vyjadruje počet dvojíc X_i, Y_j, kde platí $X_i > Y_j$. Pri zhode $X_i = Y_j$ sa pripočíta $\frac{1}{2}$ k U_X aj ku U_Y. Musí platiť $U_X + U_Y = n_1 n_2$ (počet všetkých dvojíc X_i, Y_j).

Pre malé rozsahy výberov by sme pri testovaní nulovej hypotézy o zhode oboch rozdelení mali namiesto normovanej veličiny z použiť špeciálne tabuľky (napr. Likeš a Laga, 1978). Kritický obor je v nich popísaný pomocou štatistiky $U = \min(U_X, U_Y)$. Nulovú hypotézu na hladine významnosti najviac α zamietame, ak $U \leq v_{n_1,n_2}\left(\alpha\right)$, kde $v_{n_1,n_2}\left(\alpha\right)$ je kritická hodnota. Pretože U má diskrétne rozdelenie, skutočná pravdepodobnosť chyby I. druhu môže byť výrazne nižšia ako požadovaná α.

9.4 Neparametrické párové testy

Do skupiny neparametrických párových testov patria znamienkový test a Wilcoxonov znamienkovaný test.

9.4.1 Znamienkový test

Požiadavka na normálne rozdelenie rozdielu $X_i - Y_i$ je niekdy príliž prísna. Hlavne pre malý rozsah výberu jeho nesplnenie môže znehodnotiť rozhodovanie pomocou párového t-testu. Oproti párovému t-testu má *znamienkový test* oveľa slabšie predpoklady, kde postačuje, aby náhodná veličina $D = X - Y$ mala nejaké spojité rozdelenie, kde stačí vedieť, ktorá z nasledovných udalostí nastala

$$D_i : X_i < Y_i, X_i = Y_i, X_i > Y_i.$$

Testujeme

- H_0 : *medián rozdielu* $D_i = X_i - Y_i$ *sa rovná nule* oproti
- obojstrannej alternatíve H_1 : *medián je nenulový.*

Pozorovania, kde nastalo $X_i = Y_i$ neberieme do úvahy a podľa toho upravíme rozsah súboru n. Označme počet kladných D_i ako D^+. H_0 zamietame približne na α, ak pre

$$Z = \frac{|D^+ - n/2| - 1/2}{\sqrt{n/4}}$$

platí, že $|Z| \geq u_{\alpha/2}$. Existujú aj tabuľky na presný výpočet (napr. Likeš a Laga, 1978).

Spomenutý test má jednoduchú motiváciu. Ak platí H_0, má náhodný jav $X_i > Y_i$ pravdepodobnosť výskytu 1/2, kde oba rôzne výsledky majú rovnakú pravdepodobnosť a zhoda $X_i = Y_i$ nastáva pri spojitom rozdelení s nulovou pravdepodobnosťou. Preto pre D^+, za platnosti H_0, platí $D^+ \sim Bin\,(n, 0.5)$ s $E\,[D^+] = n/2$ a $Var\,[D^+] = n/4$. Štatistitiku Z dostaneme modifikáciou - odpočítaním 1/2 v čitateli - normovaného protajšku veličiny D^+.

9.4.2 Wilcoxonov znamienkovaný test

Wilcoxonov znamienkovaný test (Wilcoxon signed rank test, neparametrická alternativa pároveho t-testu).

Rovnako ako pri znamienkovom teste, musia mať rozdiely $D_i = X_i - Y_i$ spojité rozdelenie, kde sa naviac predpokladá, že rozdelenie je symetrické okolo mediánu. Teda, krivka hustoty rozdielov D_i musí byť symetrická a ako stred symetrie nepripadá do úvahy nič iné ako medián.

- H_0 : *medián ako stred symetrie je rovný nule*,
- H_1 : *medián je rôzny od nuly.*

Majme pozorovanie (x_i, y_i) pre i-tu štatistickú jednotku (napr. pred a po liečbe), $i = 1, ...n$. Najprv vypočítame rozdiel $d_i - y_i$ medzi každým párom pozorovaní. Ak sú medzi rozdielmi d_i nuly, príslušné pozorovania sa vylúčia a n sa podľa toho zmenší. *Wilcoxonova znamienkovaná štatistika* T_W^+ sa vypočíta z absolútnych hodnôt z_i zoradených podľa veľkosti. Označme znamienkom plus (+) kladné D_i a znamienkom mínus (−) záporné rozdiely D_i. Potom $T_W^+ = \sum (+poradia)$ a podľa napr. Hollander, Wolf (1999) a CLV

$$E_0\left[T_W^+\right] = \frac{n\,(n+1)}{4}$$

$$Var_0\left[T_W^+\right] = \frac{n\,(n+1)\,(2n+1)}{24}$$

$$Z = \frac{T_W^+ - E_0\left[T_W^+\right]}{\sqrt{Var_0\left[T_W^+\right]}} \sim N(0,1).$$

Pozn.: Môže sa stať, že s outliermi nie je rozdiel medzi porovnávanými skupinami pri použití t-testu. Po vylúčení outlierov sa už rozdiely prejavia. Pre malé počty pozorovaní treba použiť tabuľky, napr. Likeš a Laga (1978). Ak sú medzi rozdielmi nulové hodnoty (teda zhody), vypočítame aritmetický priemer z pozorovaní asociovaných so zhodnými rozdielmi (*strednoporadia*) a urobíme znova rozdiel priemerov a o prislušný počet sa upraví n.

Teda podobne ako pri dvojvýberovom Wilcoxonovom teste (ak $t_1, t_2, ...$ sú počty zhodných pozorovaní) platí

$$Z_{kor} = \frac{T_W^+ - E_0\left[T_W^+\right]}{\sqrt{Var_0\left[T_W^+\right] - \frac{1}{2}\frac{1}{24}\sum_j\left(t_j^3 - t_j\right)}} \sim N(0,1).$$

9.4.3 Príklady v S-PLUS a R

Príklad 77 *neparametrické testy, dáta z kapitoly "t.testy", funkcia "`wilcox.test`"*

```
wilcox.test(mich,mu=990)
wilcox.test(gain.high,gain.low)
wilcox.test(wear.A,wear.B,paired=T)
# vstupy
# "mu" - parameter posunu polohy pre rozdelenie"x"
# formulácia alternatívy alternative=''two.sided'' je default, ďalšie voľby sú ''greater'', ''less''
# jednovýberový test = zadáme "x" ako vektor dát,
# dvojvýberový test = zadáme "x" a "y" ako vektory dát
# párový test = zadáme "x" a "y" ako vektory dát, stanovíme argument paired=T, defult je paired=F
# exaktné rozdelenie testovacej štatistiky (tabuľky) - correct=T je default
# korekcia spojitosti pre normálnu aproximáciu p-hodnoty - correct=T je default
# výstupy
# názov použitého testu "method"
# testovacia štatistika "rank-sum statistic"
# počet pozorovaní "n"
# p-hodnota "p-value"
# alternatívna hypotéza "alternative hypothesis"
```

9.5 Korelačná analýza

Definícia 78 *Nech* $E\left(X^2\right)$, $E\left(Y^2\right) < \infty$. *Potom je*

1. *kovariancia* $cov(X,Y)$ *náhodných veličín* X *a* Y *definovaná ako*

$$cov(X,Y) = E\left[(X - E(X))\ \ (Y - E(Y))\right] = E(XY) - E(X)E(Y).$$

   Ak $X = Y$, *potom* $cov(X,X) = D(X) = E\left(X^2\right) - \left[E(X)\right]^2$.

2. *koeficient korelácie* $\rho_{X,Y}$ *náhodných veličín* X *a* Y *definovaný ako*

   $\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$, *pre* $Var(X)\,Var(Y) > 1$.

Definícia 79 *Nech* $\mathbf{X} = (X_1, X_2, ..., X_n)'$ *je náhodný vektor, ktorého zložky majú konečný druhý moment. Potom*

1. *kovariančná (variančná) matica* $Var(\mathbf{X})$*s rozmermi* $n \times n$ *tohoto náh. výberu má prvky definované ako*

   $cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))]$,

   $cov(X_i, X_i) = D(X_i)$.

   Táto matica je symetrická a pozitívne definitná.

2. *korelačná matica* $|Corr(\mathbf{X})| \leq 1$ *je matica s prvkami*

   $\rho_{ij} = corr(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}}, i, j \in \langle 1, n \rangle$,

   $\rho_{ii} = corr(X_i, X_i) = 1, i \in \langle 1, n \rangle$.

Definícia 80 *Matica sa nazýva pozitívne definitná ak*

1. $Var(\mathbf{X}) = Var(\mathbf{X})^T$,
2. *pre* $\forall a = (a_1, ..., a_n)^T \in R^n$ *platí, že* $a^T Var(X)\, a > 0$.

9.5.1 Korelačný koeficient

Prvky **výberovej kovariančnej matice** $\mathbf{S}_{p\times p}$ (vieme, že $s_{ij} = s_{ji}, \forall i, j$)

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j), i, j = 1, 2..., p$$

Korelačný koeficient (product moment correlation coeficient, sample correlation coeficient).
Majme nezávislé náhodné vektory (X_1, Y_1), $(X_2, Y_2), ..., (X_n, Y_n)$, nezávislé medzi dvojicami, všeobecne nie vnútri dvojice, potom môžeme ako odhad *korelačného koeficientu* (KK) použiť

$$\begin{aligned} R_{X,Y} &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \\ &= \frac{S_{X,Y}}{\sqrt{S_X}\sqrt{S_Y}} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sqrt{\sum_{i=1}^{n} X_i - n\overline{X}}\sqrt{\sum_{i=1}^{n} Y_i - n\overline{Y}}}, \end{aligned}$$

- zápis pre (i, j)-ty člen výberovej korelačnej matice $\mathbf{R}$ $(r \in \langle -1, 1 \rangle)$

$$r_{i,j} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} = \frac{\sum_{k=1}^{n} x_{ki}x_{kj} - n\overline{x}_i\overline{x}_j}{\sqrt{\sum_{k=1}^{n} x_{ki} - n\overline{x}_i}\sqrt{\sum_{k=1}^{n} x_{kj} - n\overline{x}_j}}, \text{ kde } s_{ii} = s_i^2, s_{jj} = s_j^2,$$

čo nie je nevychýlený odhad. Jeho nevychýlená podoba má tvar

$$R_{ij}^* \approx R_{ij}\left[1 + \frac{1 - R_{ij}^2}{2(n-4)}\right].$$

Pozn.: Ak majú náhodné vektory (X_i, Y_i) aspoň "približne" normálne rozdelenie, môžeme výberový korelačný koeficient r použiť na testovanie nulovosti populačného korelačného koeficientu ρ. Ak zamietame hypotézu o nulovosti ρ, zamietame súčasne ak hypotézu o nezávislosti náhodných veličín.

Testovanie hypotéz a IS pre *KK* $\rho_{X,Y}$

Majme náhodný výber $(X,Y) \sim N_2(\mu,\Sigma)$, $corr(X,Y)=0, n \geq 3$, potom

$$T(R) = \frac{\sqrt{n-2}R}{\sqrt{1-R^2}} \sim t_{n-2}$$

Teda $H_0 : \rho_{X,Y} = 0, H_1 : \rho_{X,Y} \neq 0$. H_0 zamietame, ak $|T(r)| > t_{n-2}(\alpha/2)$ (podobne jednostranné alternatívy)

Ak $\rho_{X,Y}=0$, potom za platnosti $(X,Y) \sim N_2(\mu,\Sigma), corr(X,Y)=0, n \geq 3$ platí

$$f(r) = \frac{\Gamma[1/2(n-1)]}{\Gamma[1/2(n-2)]\sqrt{\pi}} \left(1-r^2\right)^{1/2(n-4)},$$

kde $\Gamma[p] = \int_0^\infty x^{p-1}e^{-x}dx$, ak $p \in Z$ potom $\Gamma[p] = (n-1)!$. Teda (pre $n \to \infty$, často $n > 100$)

$$R \sim N\left(\rho_{X,Y}, \frac{\left(1-\rho_{X,Y}^2\right)^2}{n-1}\right).$$

Fisherova Z premenná - transformácia stabilizujúca rozptyl ($n \geq 3$, jej rozdelenie sa rýchlo blíži k normálnemu*)*

$$Z = \frac{1}{2}\ln\frac{1+R}{1-R}, Z \sim N\left(\frac{1}{2}\ln\frac{1+\rho_{X,Y}}{1-\rho_{X,Y}}, \frac{1}{n-3}\right)$$

Odporúča sa uvažovať o Z ako o normálnej premennej už pre $n \geq 10$, ak $\rho_{X,Y}$ nie je blízko ± 1.

$H_0 : \rho_{X,Y} = \rho_0, H_1 : \rho_{X,Y} \neq \rho_0$. H_0 zamietam, ak $U = \sqrt{n-3}\,|z-\xi_0| > u(\alpha/2)$, kde $z = \frac{1}{2}\ln\frac{1+r_0}{1-r_0}, \xi_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}$

1. $100 \times (1-\alpha)\%IS$ pre ρ odvodíme nasledovné (Potocký, 1998)

   $Pr\left(\sqrt{n-3}\,|z-\xi| \leq u(\alpha/2)\right) = 1-\alpha$, $\xi = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$

   $$\left(Z - \frac{u(\alpha/2)}{\sqrt{n-3}} \leq \xi \leq Z + \frac{u(\alpha/2)}{\sqrt{n-3}}\right)$$

   Ak $\rho = th(\xi)$ (monotónna fcia ξ), potom

   $$\left(th\left[Z - \frac{u(\alpha/2)}{\sqrt{n-3}}\right] \leq \rho \leq th\left[Z + \frac{u(\alpha/2)}{\sqrt{n-3}}\right]\right)$$

2. $100 \times (1-\alpha)\%IS$ pre ρ : (Andel, 1998)

   $$\left(\frac{D-1}{D+1}, \frac{H-1}{H+1}\right),$$

   kde $D = \exp\left[2Z - 2\frac{u(\alpha/2)}{\sqrt{n-3}}\right], H = \exp\left[2Z + 2\frac{u(\alpha/2)}{\sqrt{n-3}}\right]$

   Pozn.: je zrejmé, že (1) je identické s (2), kde $D = th(Z - \frac{u(\alpha/2)}{\sqrt{n-3}})$ a $H = th(Z + \frac{u(\alpha/2)}{\sqrt{n-3}})$

Testovanie hypotéz o ρ_1, ρ_2

Nech $(X_1,Y_1) \sim N_2(\mu_1,\Sigma_1)$ s ρ_{X_1,Y_1} a R_{X_1,Y_1}, rozsahom n_1. Nech $(X_2,Y_2) \sim N_2(\mu_1,\Sigma_1)$ s ρ_{X_2,Y_2} a R_{X_2,Y_2}, rozsahom n_2. Nech $H_0 : \rho_1 = \rho_2, H_1 : \rho_1 \neq \rho_2$. Potom

$\mathrm{Z}_1 - Z_2 \sim N\left(0, \frac{1}{n_1-3} + \frac{1}{n_2-3}\right), Z_i = \frac{1}{2}\ln\frac{1+R_i}{1-R_i}, i = 1,2,$

$$Z_{1,2} = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \sim N(0,1)$$

H_0 zamietame, ak $|z_{1,2}| > u(\alpha/2)$

9.5.2 Spearmanov korelačný koeficient

Spearmanov korelačný koeficient (r_S)

Predpokladajme, že $(X_1, Y_1)^T, ..., (X_n, Y_n)^T$ je výber z dvojrozmerného rozdelenia. Nech $R_1, ..., R_n$ sú poradia veličín $X_1, ..., X_n$ a $Q_1, ..., Q_n$ sú poradia veličín $Y_1, ..., Y_n$ v spoločnom (združenom) výbere.

Spearmanova štatistika má tvar

$$S_N = \sum_{i=1}^{n} R_i Q_i$$

Spearmanov korelačný koeficient

$$R_S = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(Q_i - \frac{n+1}{2}\right)}{\sigma_R \sigma_Q},$$

kde

$$\sigma_R = \sigma_Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2}.$$

Potom

$$\begin{aligned} R_S &= \frac{12}{n(n^2-1)}\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(Q_i - \frac{n+1}{2}\right) \\ &= \frac{12}{n(n^2-1)}\left(S_N - \frac{n\left(n+1\right)^2}{4}\right) \\ &= 1 - \frac{6}{n\left(n^2-1\right)}\sum_{i=1}^{n}\left(R_i - Q_i\right)^2. \end{aligned}$$

Vlastnosti:

1) $|R_S| \leq 1$, pričom $R_S = 1$ práve vtedy, keď $R_i = Q_i$ pre všetky i, a rovnosť $R_S = -1$ nastáva práve vtedy, keď $R_i = n + 1 - Q_i$ pre všetky i.

2) Ak $R = (R_1, ..., R_n)$, $Q = (Q_1, ..., Q_n)$ sú nezávislé rovnomerne rozdelené náhodné vektory, tak $E[R_S] = 0, Var[R_S] = \frac{1}{n-1}$.

3) R_S je symetricky rozdelený okolo nuly a za platnosti asymptotickej normality koeficientu R_S ($n \to \infty$, pre $n > 30$) platí pre kritickú hodnotu nasledovný vzťah $r_S(\alpha) = u(\alpha/2)/\sqrt{(n-1)}$.

4) Platí $E[S_N] = n\left(\frac{n+1}{2}\right)^2, Var[S_N] = \frac{1}{n-1}\left(\frac{n(n^2-1)}{12}\right)^2$, potom

$$R_S = \frac{1}{\sqrt{(n-1)\,Var[S_N]}}\left(S_N - E[S_N]\right).$$

5) Spearmanova štatistika je symetricky rozdelená okolo svojej strednej hodnoty a túto vlastnosť má aj R_S.

6) Ak $(X_1, Y_1)^T, ..., (X_n, Y_n)^T$ je výber z dvojrozmerného normálneho rozdelenia s korelačným koef. ρ, potom plati

$$R = 2\sin\left(\frac{\pi}{6}\right) R_S$$

Testovanie hypotéz

H_0 : premenné X, Y sú nezávislé

H_1 :

1) obojstranná alternatíva

zamietame hypotézu nezávislosti	$\lvert r_S\rvert > r(\alpha/2, n)$
nezamietame hypotézu nezávislosti	$\lvert r_S\rvert \leq r(\alpha/2, n)$

2) alternatíva kladnej závislosti

"príjímame" alternatívu kladnej závislosti	$r_S > r(\alpha, n)$
nezamietame hypotézu nezávislosti	$r_S \leq r(\alpha, n)$

3) alternatíva zápornej závislosti

"príjímame" alternatívu zápornej závislosti	$r_S < r(\alpha, n)$
nezamietame hypotézu nezávislosti	$r_S \geq r(\alpha, n)$

Kritické hodnoty $r(\alpha/2, n)$ su tabelované a používajú sa približne do $n = 30$.

Ak sa medzi $X_1, ..., X_n$ alebo $Y_1, ..., Y_n$ vyskytujú zhody, tak R_i, Q_i sa vytvárajú pomocou procesu strednoporadí (ako pri iných neparametrických testoch, pozri vyššie) a Spearmanov korelačný koeficient založený na týchto strednoporadiach je určený vzťahom

$$R_S = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(Q_i - \frac{n+1}{2}\right)}{\sigma_R \sigma_Q},$$

kde

$$\sigma_R = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)^2}, \sigma_Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_i - \frac{n+1}{2}\right)^2},$$

teda

$$\begin{aligned} R_S &= \frac{1}{n\sigma_R\sigma_Q}\left(S_N - \frac{n(n+1)^2}{4}\right) \\ &= \frac{12}{n(n^2-1)}\left(S_N - \frac{n(n+1)^2}{4}\right)\frac{1}{d_R d_Q}, \end{aligned}$$

kde

$$d_R = \sqrt{1 - \frac{1}{n(n^2-1)}\sum_j t_j^{(X)}\left[\left(t_j^{(X)}\right)^2 - 1\right]},$$

$$d_Q = \sqrt{1 - \frac{1}{n(n^2-1)}\sum_j t_j^{(Y)}\left[\left(t_j^{(Y)}\right)^2 - 1\right]},$$

kde t_j predstavujú počty zhôd pri príslušnej hodnote indexu j. Asymptoticky (pre $n \to \infty$) platí, že

$$\sqrt{n-1}R_S \sim N(0,1)$$

a postupujeme rovnako ako predtým, teda ($n \to \infty$, pre $n > 30$) pre kritickú hodnotu platí nasledovný vzťah $r_S(\alpha) = u(\alpha/2)/\sqrt{(n-1)}$.

9.5.3 Príklady v S-PLUS a R

Príklad 81 *funkcia "`cor`"*

```
# vstupy
# zadáme "x" a "y" ako vektory dát
# chýbajúce pozorovania, na.method=''fail'' je default, pokiaľ ich chceme vyradiť na.method=
''omit''
# výstupy
# Pearsonov korelačný koeficient alebo korelačnú maticu (ak zadáme ako vstup maticu premenných)

# korelačnú matica pre dáta "longley.x"
cor(longley.x)
```

Príklad 82 *funkcia "cor.test", dáta "state.x77"*

```
# vstupy
# zadáme "x" a "y" ako vektory dát
# metóda - method=''pearson'', je default, alternatíva je napr. method=''spearman''
# formulácia alternatívy alternative=''two.sided'' je default, ďalšie voľby sú ''greater'' (väčší ako nula), ''less'' (menší ako nula)
# výstupy
# názov použitého testu "method"
# testovacia štatistika "normal-z"
# p-hodnota "p-value"
# alternatívna hypotéza "alternative hypothesis"
# odhad korelačného koeficientu "sample estimates"
murder <- state.x77[,''Murder'']
illit <- state.x77[,''Illiteracy'']
cor.test(murder,illit)
cor.test(murder,illit,method=''s'')
```

Príklad 83 *IS pre korelačný koeficient, dáta "state.x77"*

```
IScor _ function(x, y, conf.level = 0.95)
{
 z <- atanh(cor(x, y))
 a <- qnorm(1 - (1 - conf.level)/2)/(length(x) - 3)^0.5
 CI.Z <- c(z - a, z + a)
 conf.int <- tanh(CI.Z)
}

IScor(murder,illit)
```

9.6 Testovanie normality

9.6.1 Chi-kvadrát test dobrej zhody

Nech $\xi_1, ..., \xi_n$ je náhodný výber. Chceme testovať nulovú hypotézu H_0, že ide o výber z $N\left(\mu, \sigma^2\right)$, kde parametre μ, σ^2 nie sú známe. Najprv vytvoríme triedy (podobne ako pri histograme)

$$(-\infty, b_1), \langle b_1, b_2), ..., \langle b_{k-2}, b_{k-1}), \langle b_{k-1}, \infty),$$

kde $k \geq 4$. Označme i-tu triedu ako J_i. Pravdepodobnosť p_i, že daná veličina ξ_j $(j = 1, ..., n)$ padne do J_i, je rovná

$$p_i = p_i(\mu, \sigma) = \int_{J_i} f(x)\, dx, \text{ kde } f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Iteračne riešime sústavu

$$\frac{1}{n}\sum_{i=1}^{k} \frac{X_i}{p_i} \int_{J_i} x f(x)\, dx, \sigma^2 = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i}{p_i} \int_{J_i} (x-\mu)^2 f(x)\, dx.$$

Pozor: μ závisia na p_i a $f(x)$. Riešenia sústavy sú $\widehat{\mu}$ a $\widehat{\sigma}$. Potom

$$\chi^2 = \sum_{i=1}^{k} \frac{[X_i - np_i(\widehat{\mu}, \widehat{\sigma})]^2}{np_i(\widehat{\mu}, \widehat{\sigma})} \sim \chi^2_{k-3}.$$

9.6.2 Kolmogorov - Smirnovov test dobrej zhody

Zhodu rozdelení veličiny v dvoch populáciách by sme mohli porovnávať, keby sme poznali ich distribučné funkcie (CDF). Rozhodnutie o zhode distribučných funkcií môžeme založiť na porovnávaní ich odhadov, teda empirických distribučných funkcií. Empirickú distribučnú funkciu môžeme písať ako kumulatívnu relatívnu početnosť (pozri vyššie)

$$\widehat{F}^X(x) = \frac{\#x_i < x}{n},$$

kde n je počet pozorovaní v náhodnom výbere.

Tento test hodnotí maximálny rozdiel empirických distribučných funkcií

$$D = \max_{\forall x} \left| \widehat{F}^X(x) - \widehat{F}^Y(x) \right|.$$

Podobný vzťah platí aj pre porovnanie empirickej a teoretickej distribučnej funkcie

$$D^* = \max_{\forall x} \left| \widehat{F}^X(x) - F^X(x) \right|.$$

Na rozhodovanie slúžia tabuľky (e.g. Likeš a Laga, 1978) alebo približná kritická hodnota

$$D_{n_1,n_2}(\alpha) = \sqrt{\frac{n_1+n_2}{2n_1n_2} \log \frac{2}{\alpha}},$$

kde nulovu hypotézu zamietame, ak $D \geq D_{n_1,n_2}(\alpha)$.

Nech F_1 a F_2 sú dve *CDF*. V jednovýberovej situácii je F_1 empirická *CDF* a F_2 je *CDF* nulovej hypotézy. V dvojvýberovej situácii sú obe F_1 a F_2 empirické *CDF*.

- dvojvýberová situácia
  $H_0 : F_1(x) = F_2(x)$, pre $\forall x$
  $H_1 : F_1(x) \neq F_2(x)$, pre aspoň jedno x
  alternatíva $T = \sup_x |F_1(x) - F_2(x)|$
- jednovýberová alternatíva ("less")
  $H_0 : F_1(x) \geq F_2(x)$, pre $\forall x$
  $H_1 : F_1(x) < F_2(x)$, pre aspoň jedno x
  alternativa $T^- = \sup_x |F_2(x) - F_1(x)|$
- jednovýberová alternatíva ("greather")
  $H_0 : F_1(x) \leq F_2(x)$, pre $\forall x$
  $H_1 : F_1(x) > F_2(x)$, pre aspoň jedno x
  alternativa $T^+ = \sup_x |F_2(x) - F_1(x)|$

Pozn.:
- v porovnaní s chí-kvadrát testom dobrej zhody môžeme pracovať aj s výbermi s malým rozsahom,
- základom testovacej štatistiky sú rozdiely v napr. teoretických a empirických kumulatívnych početnostiach (ako pri chí-kvadrát teste dobrej zhody),
- najpr utvoríme usporiadaný výber,
- zistíme empirické početnosti n_{e_i},
- očakávané (teoretické) početnosti n_{o_i},
- empirické kumulované početnosti N_{e_i},
- očakávané kumulované početnosti N_{o_i},

- vypočítame

$$d = \frac{1}{n} \max |N_{e_i} - N_{o_i}|$$

a v tabuľkách, alebo ak $n > 100$ asymtoticky, na základe asymetrického rozdelenia veličiny d_n, ktoré odvodil Kolmogorov (1941)

$$d_n(0.95) \approx \frac{1.358}{\sqrt{n}}, d_n(0.99) \approx \frac{1.628}{\sqrt{n}},$$

ak $d > d_n(\alpha)$, zamietame hypotézu H_0.

9.6.3 Príklady v S-PLUS a R

Príklad 84 *funkcia "`chisq.gof`"*

POZN.: `chisq.gof` testuje len dvojstrannú alternatívu

```
# vstupy
# zadáme "x" a zadefinujeme typ rozdelenia, napr. distribution=''normal'' (default) alebo ''chisquare'', ''t'', ''f'', ''exponential'' a pod.
# počet tried a zlomové body sú explicitne nastavené a nebudeme ich meniť
# výstupy
# názov použitého rozdelenia "method"
# testovacia štatistika "Chi-square"
# df súvisiace s testovacou štatistikou "parameters"
# p-hodnota "p-value"
# alternatívna hypotéza "alternative hypothesis"
# počty pozorovaní v jednotlivých intervaloch "counts"
# očakávané počty pozorovaní v jednotlivých intervaloch "expected"

# nagenerujte pseudonáhodné čísla z exponenciálneho rozelenia (n=50) a porovnajte s normálnym a exponenciálnym rozdelením
x <- rexp(50, rate=1.0)
chisq.gof(x)
chisq.gof(x,dist=''exponential'',rate=1.0)
# nagenerujte pseudonáhodné čísla z poisonovho rozdelenia (n=50), rozdeľte tento vektor v kvartiloch ako zlomových bodoch na 4 skupiny
x <- rpois(50,lambda=3)
breaks <- quantile(x)
# vložte minimálnu hodnotu do delenia
# zlomové body - argument cut.points
breaks[1] <- cut.points[1] - 1
z <- chisq.gof(x,cut.points=breaks,dist=''poisson'',lambda=3)
z$count
z$expected
```

Príklad 85 *funkcia "`ks.gof`"*

```
# vstupy
# zadáme "x" a zadefinujeme typ rozdelenia, napr. distribution=''normal'' (default) alebo ''chisquare'', ''t'', ''f'', ''exponential'' a pod.
# zadáme "x" a "y" ako vektory dát
# formulácia alternatívy alternative=''two.sided'' je default (pre dvojvýberový test je len táto možnosť), ďalšie voľby sú ''greater'', ''less''
# počet tried a zlomové body sú explicitne nastavené a nebudeme ich meniť
# výstupy
```

```
# názov použitého rozdelenia "method"
# testovacia štatistika "ks"
# df súvisiace s testovacou štatistikou "parameters"
# p-hodnota "p-value"
# alternatívna hypotéza "alternative hypothesis"

# jednovýberový prípad
z <- rnorm(100)
ks.gof(z, distribution = ''normal'')
ks.gof(z, distribution = ''chisquare'', df = 2)
# dvojvýberový prípad
x <- rnorm(90)
y <- rnorm(8, mean = 2.0, sd = 1)
ks.gof(x, y)

# dáta "mich" a zobrazenie ich kumulatívnej distribučnej funkcie
cdf.compare(mich,dist=''normal'',mean=mean(mich),sd=sqrt(var(mich)))
ks.gof(mich,dist=''normal'')
chisq.gof(mich, dist = ''normal'', n.param.est = 2, mean = mean(mich), sd =
sqrt(var(mich)))
```

10 Lineárne regresné modely

Všeobecná formulácia štatistického modelu:

1.) Základný model pre dáta $\mathbf{x}$: je **model náhodného výberu** - $\mathbf{x}$ je realizácia náhodného výberu $\mathbf{X}$ z nejakého rozdelenia $F_\theta, \theta \in \Theta$,

$\mathbf{X} = [X_1, X_2, ..., X_n]$ je iid, $E[\mathbf{X}] = \mu(1, 1, ..., 1)^T, D[\mathbf{X}] = \sigma^2 \mathbf{I}_n$

2.) **Regresný model na** $Y_i, i =, ..., n$

$y_i = g(\mathbf{x}_i) + \varepsilon_i$, kde $g(x_i)$ je ľubovoľná spojitá funkcia na R.
A) $g(\cdot)$ je neparametrická funkcia - neparametrický model,
B) $g(\cdot)$ je parametrická funkcia - parametrický model, kde $y_i = g_\theta(\mathbf{x}_i) + \varepsilon_i, \theta \in \Theta, \theta = (\theta_1, \theta_2, ..., \theta_n)^T \in E^k$, do tejto množiny patria klasické lineárne a nelineárne modely (LM a NLM).

10.1 Jednoduchý lineárny regresný model

Definujem nasledovné pojmy:

- Nech y_i reprezentuje *hodnotu odpovede* (*response variable)* na i-ej štatistickej jednotke (na grafe v dvojdimenzionálnej pravouhlej súradnicovej sústave ju znázorňuje- me na vertikálnej y-ovej osi), a x_i je hodnota *nezávislej premennej* (*explanatory variable,* na grafe znázornenej na horizontálnej x-ovej osi).

- *Jednoduchý lineárny regresný model* môžeme definovať ako

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

kde β_0 je *intercept* a β_1 je *sklon* (*slope)* priamky lineárneho vzťahu ε_i sú *náhodné chyby*.

- Pre náhodné chyby predpokladáme ich nezávislosť (nezávislé náhodné premenné) a ich normálne rozdelene, $\varepsilon_i \sim N(\mu, \sigma^2)$. Ďalej, $Y \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$. Naviac

$$\widehat{\beta} \sim N\left(\beta, \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}\right).$$

- *Regresné koeficienty*, β_0 a β_1, sú odhadnuté metódou najmenších štvorcov (MNŠ) ako $\widehat{\beta}_0$ a $\widehat{\beta}_1$. Potom

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

kde sme $\widehat{\beta}_0$ a $\widehat{\beta}_1$ dostali nasledovne pomocou normálnych rovníc

$$S^2_{\mathbf{y}}(\theta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = (-2)\sum (y_i - \beta_0 - \beta_1 x_i) = 0, \frac{\partial S}{\partial \beta_1} = (-2)\sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\widehat{\beta}_0 = \frac{\sum y_i - \beta_1 \sum x_i}{n} = \overline{y} - \widehat{\beta}_1 \overline{x}$$

$$\widehat{\beta}_1 = \frac{\sum (y_i - \overline{y})(x_i - \overline{x})}{\sum (x_i - \overline{x})^2} = r_{x,y} \frac{\sqrt{\sum (y_i - \overline{y})^2}}{\sqrt{\sum (x_i - \overline{x})^2}}$$

V maticovej podobe

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \mathbf{X}^T \mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

Ak $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\mathbf{A}$, potom

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{1}{\sum x_i^2 - n\overline{x}^2}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}.$$

Regresia prechádzajúca počiatkom (model $y_i = \beta_1 x_i + \varepsilon_i$, teda $\beta_0 = 0$, $\beta_1 = \beta$) bude mať normálne rovnice a odhad $\widehat{\beta}$

$$S_{\mathbf{y}}^2(\theta) = \sum_{i=1}^{n}(y_i - \beta_1 x_i)^2 \Longrightarrow \widehat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Testovanie hypotéz
$H_0 : \beta_1 = 0$.
ANOVA tabuľka regresného modelu $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ bude

Zdroj rozptylu	suma štvorcov	df	priemerné štvorce (MS)
SS_R	$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$	$df_R = 1$	$MS_R = SS_R/df_R$
SS_e	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	$df_e = n-2$	$MS_e = SS_e/df_e$
SS_T	$\sum_{i=1}^{n}(y_i - \overline{y})^2$	$df_T = n-1$	

Ďalej

$$\widehat{\sigma}^2 = s^2 = \frac{SS_e}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Za platnosti H_0 bude $MSR = MS_R/MS_e$ mať rozdelenie $F_{1,n-2}$.

Nakoniec, testovacia štatistika a ISs pre β_0 a β_1 bude mať tvar

$$T_i = \frac{\left|\beta_i - \widehat{\beta}_i\right|}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t_{n-2}, i = 1, 2$$

Najčastejšie $\beta_i^{(0)} = 0$. Potom $(1-\alpha)\times 100\% IS$ pre β_i, $i = 1, 2$, bude mať tvar

$$\widehat{\beta}_0 \pm t_{n-2}(\alpha)\,\widehat{\sigma}\sqrt{\frac{\sum x_i^2}{n\sum(x_i - \overline{x})^2}},$$

$$\widehat{\beta}_1 \pm t_{n-2}(\alpha)\,\widehat{\sigma}\sqrt{\frac{1}{\sum(x_i - \overline{x})^2}}$$

a na základe kontrastového vektora $\mathbf{c} = (1, x)^T$, kde x je nejaké dané číslo (rozšírenú definíciu $\mathbf{c}$ pozri nižšie), môžeme písať $(1-\alpha)\,\%IS$ pre regresnú priamku

$$\beta_0 + \beta_1 x_i \pm t_{n-2}(\alpha)\,\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum(x_i - \overline{x})^2}}.$$

Ak počítame IS pre všetky $x \in \langle \min_{\forall x_i} x_i, \max_{\forall x_i} x_i\rangle$, dostaneme dve vetvy hyperboly, ktoré tvoria *pás spoľahlivosti okolo regresnej priamky* (Scheffeho metóda). Ten ale zaručuje prekrytie jednej hodnoty x so spoľahlivosťou $1-\alpha$, ale nie celej priamky. Je možné odvodiť aj taký pás, ktorý pokryje celú regresnú priamku so spoľahlivosťou $1-\alpha$. Hovoríme mu *pás spoľahlivosti pre regresnú priamku* a bude mať tvar

$$\beta_0 + \beta_1 x_i \pm \sqrt{2F_{2,n-2}}\,\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum(x_i - \overline{x})^2}}.$$

Pozn.1: Pre regresný model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ nás môže zaujímať nasledovná nulová hypotéza $H_0^* : \beta = (\beta_0, \beta_1)^T = \beta^{(0)} = (0,1)^T$ oproti $H_1^* : \beta = (\beta_0, \beta_1)^T \neq (0,1)^T$, kedy by za platnosti H_0^* bola odpoveď y_i rovná x_i až na náhodnú chybu ε_i, teda $y_i = x_i + \varepsilon_i$, kedy použijeme testovaciu štatistiku

$$Z = \frac{1}{2s^2}\left(\beta - \beta^{(0)}\right)\mathbf{X}^T\mathbf{X}\left(\beta - \beta^{(0)}\right)^T \sim F_{2,n-2}.$$

Pozn.2: Pri *kvadratickej závisloti* máme lineárny model v tvare

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

potom nás bude zaujímať nulová hypotéza $H_0^{**} : \beta_2 = 0$ oproti $H_1^{**} : \beta_2 \neq 0$, kedy ide o *test linearity regresie*. Tento test je založený na štatistike

$$T_3 = \frac{|\beta_2|}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{22}}} \sim t_{n-3}$$

Inokedy nás môže zaujímať $H_0^{***} : (\beta_1, \beta_2)^T = \mathbf{0}$ oproti $H_1^{***} : (\beta_1, \beta_2)^T \neq \mathbf{0}$, čo je v tomto prípade test závislosti y_i na x_i. Na to potrebujeme testovaciu štatistiku

$$Z = \frac{1}{2s^2}(\beta_1, \beta_2)\left(\mathbf{X}_{sub}^T\mathbf{X}_{sub}\right)(\beta_1, \beta_2)^T \sim F_{2,n-3},$$

kde $\mathbf{X}_{sub}$ je 2×2 subblok matice $\mathbf{X}$ prislúchajúci testovaným koeficientom. Vo všeobecnosti má $Z \sim F_{q,n-k}$, kde k je celkový počet koeficientov a q je počet porovnávaných koeficientov.

10.2 Mnohorozmerný lineárny regresný model

LM: $y_i = \mathbf{x}_i^T\theta + \varepsilon_i, (\mathbf{x}_i)_{1\times k}, \theta_{k\times 1} \in E^k, i = 1, ..., n$, $\sum_{i=1}^n \mathbf{x}_i\theta_i = \mathbf{x}^T\theta$ (skalárny súčin), $E_\theta[Y_i] = \mathbf{x}_i^T\theta$ závisí na i

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times k}\theta_{k\times 1} + \varepsilon_{n\times 1},$$

kde $E[\varepsilon] = \mathbf{0}$, $Cov[\varepsilon] = \sigma^2\mathbf{I}$ a $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ je matica plánu regresného experimentu, ďalej $E_\theta[Y] = \mathbf{X}\theta$, $Cov_\theta[Y] = Cov[\varepsilon] = \sigma^2\mathbf{I}$.

Pozn.:

- $\mathbf{y}_{n\times 1} = (y_1, y_2, ..., y_n)^T$,
- $\mathbf{X}_{n\times k} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,k-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,k-1} \end{bmatrix}$,
- $\theta = (\theta_0, ..., \theta_{k-1})^T$,
- $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$.

Inak môžeme písať regresný model aj nasledovne

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{k-1}x_{i,k-1} + \varepsilon_i.$$

Linearita: je v $\theta \Rightarrow$ závislosť $g_\theta(\cdot)$ na θ je lineárna
$E_{c_1\theta_1 + c_2\theta_2}[Y] = c_1E_{\theta_1}[Y] + c_2E_{\theta_2}[Y], \theta_1, \theta_2 \in \Theta = E^k$
napr. $y_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \varepsilon_i, i = 1, ..., n; \theta = (\theta_1, \theta_2, \theta_3)^T \in E^3$
Odhady MNŠ

$$\widehat{\theta}(\mathbf{y}) = \arg\min_{\theta\in\Theta} S_{\mathbf{y}}^2(\theta),$$

kde

$$S_{\mathbf{y}}^2(\theta) = \sum_{i=1}^{n} \left(y_i - g_\theta(\mathbf{x}_i)\right)^2.$$

normálne rovnice - k-rovníc o k-neznámych, $\widehat{\theta}(\mathbf{y}) : \frac{\partial S_{\mathbf{y}}^2(\theta)}{\partial \theta}|_{\widehat{\theta}(\mathbf{y})} = 0$
NR pre LRM, kde $\theta = \beta$
$\mathbf{y} = \mathbf{X}\theta + \varepsilon, E_\theta[\mathbf{Y}] = \mathbf{X}\theta, Cov[\mathbf{Y}] = \sigma^2\mathbf{I}_n$

- MNŠ odhad β je definovaný ako
$$\widehat{\beta}(\mathbf{y}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y},$$
- Platí
$$E_\beta\left[\widehat{\beta}\right] = \beta, \forall\beta \in E^k,$$
$$Cov_{\sigma^2}\left[\widehat{\beta}\right] = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1},$$
kde pre *reziduálny rozptyl* platí
$$\begin{aligned} s^2 &= \widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\left(y_i - \left(\mathbf{X}\widehat{\beta}\right)_i\right)^2 = \frac{1}{n-k}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 = \frac{1}{n-k}\left\|\widehat{\varepsilon}\right\|^2 \\ &= \frac{1}{n-k}\left\|\mathbf{y} - \mathbf{X}\widehat{\beta}\right\|^2 = \frac{1}{n-k}\left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)^T\left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right), \end{aligned}$$
kde $\|\mathbf{A}\|^2 = \mathbf{A}^T\mathbf{A}$, $\widehat{\sigma}^2$ je nevychýlený konzistentný odhad pre σ^2, čo vyplýva z $E\left[\widehat{\varepsilon}^T\widehat{\varepsilon}\right] = \sigma^2(n-k)$, kde $(n-k)$ je počet stupňov voľnosti modelu (df), $\widehat{\sigma}^2$ má najmenší rozptyl medzi všetkými kvadratickými nevychýlenými odhadmi,
$$\begin{aligned} SS_e &= \left\|\mathbf{y} - \mathbf{X}\widehat{\beta}\right\|^2 = \mathbf{y}^T\mathbf{y} - \widehat{\beta}^T\mathbf{X}^T\mathbf{X}\widehat{\beta} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{y}^T\mathbf{P}\mathbf{y}, \end{aligned}$$
kde $\mathbf{P} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$, nazývame *reziduálny súčet štvorcov.*
- diagonálne elementy matice $Cov\left[\widehat{\beta}\right]$, $Cov[\beta_i, \beta_j] = Var[\beta_i] = \widehat{\sigma}^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{ij}$, kde $i = j$, sú rozptyly parametrov $\widehat{\beta}_i$, mimodiagonálne elementy sú kovariancie medzi pármi $\widehat{\beta}_i, \widehat{\beta}_k$, kde $i \neq j$. Odmocniny diagonálnych elementov tejto matice sú teda štandardné chyby parametrov $\widehat{\beta}_i$;
- potom je štandardná chyba β_i daná vzťahom
$$se(\beta_i) = \widehat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}},$$
- platí $\varepsilon \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$, z čoho vyplýva, že $\mathbf{y} \sim N\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$. Naviac
$$\widehat{\beta} \sim N\left(\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right);$$
- platí
$$T_i = \frac{\beta_i - \widehat{\beta}_i}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t_{n-k},$$

pretože

$$\frac{\beta_i - \widehat{\beta}_i}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim N(0,1)$$

a naviac

$$\frac{SS_e}{\sigma^2} \sim \chi^2_{n-k}$$

a β a s^2 sú nezávislé. Preto

$$\frac{\left(\beta_i - \widehat{\beta}_i\right)/\sigma\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}}{\sqrt{SS_e/\sigma^2}}\sqrt{n-k} = T_i \sim t_{n-k}.$$

Práve pomocou vyššie spomenutého vzorca môžeme testovať nulovú hypotézu

$$H_0 : \beta_i = \beta_i^{(0)},$$

voči alternatíve $H_1 : \beta_i \neq \beta_i^{(0)}$, kedy H_0 zamietame na hladine významnosti α, ak platí

$$T_i = \frac{\left|\beta_i - \widehat{\beta}_i\right|}{s_{\mathbf{y}}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t_{n-k}.$$

Najčastejšie $\beta_i^{(0)} = 0$. Potom $(1-\alpha)\times 100\%IS$ pre β_i bude mať tvar

$$\widehat{\beta}_i \pm t_{n-k}(\alpha/2)\,\widehat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}.$$

- $(1-\alpha)\times 100\%CS$ (elipsa spoľahlivosti, konfidenčná elipsa) má tvar

$$\frac{\left(\widehat{\beta} - \beta^{(0)}\right)^T \mathbf{X}^T\mathbf{X}\left(\widehat{\beta} - \beta^{(0)}\right)}{k\widehat{\sigma}^2} \sim F_{k-1,n-k},$$

kde

$$\frac{\left(\widehat{\beta} - \beta^{(0)}\right)^T \mathbf{X}^T\mathbf{X}\left(\widehat{\beta} - \beta^{(0)}\right)}{\sigma^2} \sim \chi^2_k$$

a

$$\frac{(n-k)\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}$$

sú nezávislé.

- **Kontrasty.** Nech $\mathbf{c} = (c_1, c_2, ..., c_k)^T$ je *vektor kontrastov*. Potom $E\left[\mathbf{c}^T\widehat{\beta}\right] = \mathbf{c}^T\beta$ a

$$T = \frac{\left|\mathbf{c}^T\beta - \mathbf{c}^T\widehat{\beta}\right|}{s\sqrt{\mathbf{c}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{c}}} \sim t_{n-k},$$

čo vedie ku $(1-\alpha)\times 100\%CS$ pre $\mathbf{c}^T\beta$ v podobe

$$\mathbf{c}^T\widehat{\beta} \pm t_{n-k}(\alpha)\,s\sqrt{\mathbf{c}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{c}}.$$

Ak však chceme nájsť $(1-\alpha)\times 100\%ISs$ pre p lineárnych kombinácií $\mathbf{c}_j^T\beta$ súčasne, spoločná pravdepodobnosť ISs už nebude $1-\alpha$. Uvedieme jednu z viacerých možností riešenia tohoto problému.

- *Bonferroniho metóda.* Nech $E_1, ..., E_p$ sú náhodné udalosti. Potom

$$\Pr\left(\cap_{j=1}^{p} E_j\right) \geq 1 - \sum_{i=1}^{p} \Pr\left(E_j\right),$$

teda

$$\Pr\left(\cap_{j=1}^{p}\left[\mathbf{c}_j^T\beta \in \mathbf{c}_j^T\widehat{\beta} \pm t_{n-k}\left(\frac{\alpha}{2k}\right) s\sqrt{\mathbf{c}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{c}_i}\right]\right) \geq 1 - k\frac{\alpha}{k} = 1 - \alpha.$$

- **Regresná diagnostika**. Vyššie sme definovali

$$SS_e = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2.$$

Teraz nám ostáva dodefinovať *celkovú sumu štvorcov*

$$SS_T = \|\mathbf{y} - \overline{\mathbf{y}}\|^2$$

a *regresnú sumu štvorcov*

$$SS_R = \|\widehat{\mathbf{y}} - \overline{\mathbf{y}}\|^2.$$

Potom môžeme vytvoriť ANOVA tabuľku regresného modelu $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ nasledovne

Zdroj rozptylu	df	Priemerné štvorce (Mean Squares (MS))
SS_R	$df_R = k - 1$	$MS_R = SS_R/df_R$
SS_e	$df_e = n - k$	$MS_e = SS_e/df_e$
SS_T	$df_T = n - 1$	

- Priemerná suma štvorcov

$$MSR = MS_R/MS_e \sim F_{k-1,n-k}$$

slúži v F-teste na testovanie všeobecnej nulovej hypotézy

$$H_0 : \beta_1 = ... = \beta_{k-1} = 0.$$

Korelácia medzi pozorovanými hodnotami y_i a modelom predikovanými hodnotami $\widehat{y}_i$ je známa ako *mnohorozmerný koeficient korelácie (multiple correlation coeficient, R)*. Hodnota R^2 poukazuje na *proporciu rozptylu odpovede* (*podiel vysvetlenej variability, koeficient determinacie, proportion of variance of the response variable)*

$$R^2 = 1 - \frac{SS_e}{SS_T}$$

korigovaný (*adjustovaný) koeficient determinácie* (má tendenciu nechávať príliž veľa premenných v modeli)

$$R^2_{adj} = 1 - \frac{n-1}{n-k}\frac{SS_e}{SS_T} = 1 - \frac{n-1}{n-k}\left(1 - R^2\right)$$

Pozn.: Pri interpretácii R^2 treba byť opatrný. Tento koeficient síce hovorí o tom, *koľko variability je vysvetlené modelom pri daných hodnotách nezávislej premennej*, ale pokiaľ β_1 pri jednoduchej lineárnej regresii ako $\beta_{y|x}$, a naopak $\beta_{x|y}$ pri opačnej závislosti, dostali by sme

$$\beta_{y|x}\beta_{x|y} = R^2_{xy}.$$

Ďalej, pokiaľ závislosť y_i na x_{1i} adjustujeme modelovanú závislosť aj voči x_{2i}. Potom má zmysel hovoriť aj o *parciálnom korelačnom koeficiente*

$$\rho_{xy,z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{\left(1 - \rho^2_{xz}\right)\left(1 - \rho^2_{yz}\right)}},$$

kde je z nová premenná, ktorej závislosť na premenných x, y meriame. V skutočnosti ide o klasický korelačný koeficient medzi dvoma novými veličinami, ktoré dostaneme ako reziduály predpovede veličiny x, resp. y pomocou z. Pri odhade použijeme empirický proťajšok ρ v podobe r.

- $(1-\alpha)\times 100\%CS$ **pre predikciu**. Nech $\mathbf{x}_{nové}$ je p-vektor nových hodnôt, na základe ktorých a modelu by sme chceli predikovať budúce $\mathbf{y}_{nové}$, teda

$$\widehat{\mathbf{y}}_{nové}=\mathbf{x}_{nové}^T\widehat{\beta}.$$

Platí

$$Var[\mathbf{x}_{nové}^T\widehat{\beta}]=\mathbf{x}_{nové}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{nové}\widehat{\sigma}^2.$$

My potrebujeme rozlišovať medzi

1) *predikciou budúcich pozorovaní* - teda ak máme $\mathbf{x}_{nové}$, aké bude $\mathbf{y}_{nové}$, $\mathbf{y}_{nové}=\mathbf{x}_{nové}^T\beta+\varepsilon$,$E[\varepsilon]=0$, predikujeme $\widehat{\mathbf{y}}_{nové}=\mathbf{x}_{nové}^T\widehat{\beta}$, ale v odhade rozptylu tohoto odhadu musíme zahrnúť rozptyl chýb $\widehat{\varepsilon}$, $D[\widehat{\varepsilon}]$ (**prediction of future values**)
$95\%CI$ teda bude

$$\widehat{\mathbf{y}}_{nové}\pm t_{n-k}(\alpha/2)\,\widehat{\sigma}\sqrt{1+\mathbf{x}_{nové}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{nové}^T},$$

2) *predikciou budúcej priemernej odpovede* - teda ak máme $\mathbf{x}_{nové}$, aké bude priemerné $\mathbf{y}_{nové}$, $\mathbf{y}_{nové}=\mathbf{x}_{nové}^T\beta+\varepsilon$, kde znovu predikujeme $\widehat{\mathbf{y}}_{nové}=\mathbf{x}_{nové}^T\widehat{\beta}$, ale v odhade rozptylu tohoto odhadu musíme zahrnúť rozptyl $\widehat{\beta}$, $D\left[\widehat{\beta}\right]$ (**prediction of the mean response**)

$$\widehat{\mathbf{y}}_{nové}\pm t_{n-k}(\alpha/2)\,\widehat{\sigma}\sqrt{\mathbf{x}_{nové}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{nové}^T}.$$

- **Testovanie hypotéz o submodeloch.** Majme nadmodel Ω a submodel ω, ktorý pozostáva z podmnožiny prediktorov $\widehat{\beta}$ obsiahnutých v Ω. Teda zoberieme ω, ktorý bude reprezentovať nulovú hypotézu a Ω, ktorý bude predstavovať alternatívnu hypotézu. Ak ${}_\omega SS_e-{}_\Omega SS_e$ je malé číslo, potom je ω adekvátny model vo vzťahu ku Ω. To vedie ku testovacej štatistike typu

$$({}_\omega SS_e-{}_\Omega SS_e)/{}_\Omega SS_e,$$

kde je menovateľ použitý ako škála. Tento typ štatistiky vznikol na základe pomeru vierohodnosti

$$\frac{\max_{\beta,\sigma\in\Omega}L(\beta,\sigma|y)}{\max_{\beta,\sigma\in\omega}L(\beta,\sigma|y)}.$$

Test zamieta nulovú hypotézu, ak je tento pomer dosť veľký. Vieme, že $L\left(\widehat{\beta},\widehat{\sigma}|y\right)\approx\widehat{\sigma}^{-n}$, čo nám dáva test, ktorý zamieta, keď

$$\frac{\widehat{\sigma}_\omega}{\widehat{\sigma}_\Omega}>\text{konšt.}\approx\frac{{}_\omega SS_e}{{}_\Omega SS_e}>\text{konšt.}\approx\frac{{}_\omega SS_e}{{}_\Omega SS_e}-1>\text{konšt.}-1\approx\frac{{}_\omega SS_e-{}_\Omega SS_e}{{}_\Omega SS_e}>\text{konšt.}$$

Ak počet parametrov (dimenzia) modelu Ω je k, a dimenzia modelu ω je q, potom podľa *Cochranovej vety,* za platnosti nulovej hypotézy (ω) platí

$$\frac{{}_\omega SS_e-{}_\Omega SS_e}{k-q}\sim\sigma^2\chi^2_{k-q},\frac{{}_\Omega SS_e}{n-k}\sim\sigma^2\chi^2_{n-k}$$

a tieto podiely sú nezávislé, potom platí

$$F=\frac{({}_\omega SS_e-{}_\Omega SS_e)/(k-q)}{{}_\Omega SS_e/(n-k)}\sim F_{k-q,n-k}.$$

Nulovu hypotézu zamietame, ak $F>F_{k-q,n-k}(\alpha)$. Vždy platí, že

$$df=\#_{pozorovaní}-\#_{parametrov},$$

potom

$$F=\frac{({}_\omega SS_e-{}_\Omega SS_e)/(df_\omega-df_\Omega)}{{}_\Omega SS_e/(df_\Omega)}$$

10.2.1 Regresná diagnostika

Grafy diagnostiky lineárneho modelu: hlavnú úlohu zohrávajú reziduály v absolútnej hodnote

$$ {}_{abs}r_i = |y_i - \widehat{y}_i| \,, $$

nemusia mať konštantný (rovnaký) rozptyl (presnosť $\widehat{y}_i$ zavisi na x_i), občas sú štandardizované pred použitím, pozri Cook and Weisberg (1982).

- *reziduály ${}_{abs}r_i$ vs. fitované hodnoty* - ak nejde približne o horizontálny kompaktný oblak bodov, potom to indikuje, že

  - funkcionálna forma nášho modelu je nesprávna alebo

  - máme nekonštantný rozptyl

- *reziduály ${}_{abs}r_i$ vs. realizácie (x, "exploratory variables")* - systematická štruktúra indikuje nedodržanie konštantnosti rozptylu alebo nevhodnosť typu modelu

- *qq-plot reziduálov*

- *'index plot' Cookových vydialeností D_k pre každé pozorovanie (y)* - hodnoty D_k predstavujú vplyv k-teho pozorovania na odhadované regresné koeficienty. Hodnoty $D_k > 1$ naznačujú, že príslušné pozorovania majú nadmerný vplyv na odhadované regresné koeficienty. *Cookove vzdialenosti* definujeme ako

$$ D_k = \frac{1}{(p+1)\, s^2} \sum_{i=1}^{n} \left[\widehat{y}_{i(k)} - \widehat{y}_i \right]^2 , $$

  kde $\widehat{y}_{i(k)}$ sú fitované hodnoty, keď je k-te pozorovanie odstránené z modelu.

10.3 Príklady v S-PLUS a R

Príklad 86 *ANAEROB - zobrazte premennú "oxygen uptake" oproti premennej "expired ventilation" a dopíšte do grafu aj korelačný koeficient zaokrúhlený na dve desatinné miesta. Vkreslite do grafu aj regresnú priamku, ktorej koeficienty dostaneme MNŠ.*

```
ANAEROB_data.frame(ANAEROB)
attach(ANAEROB)
plot(Oxygen,Ventilation,xlab=''Oxygen uptake'', ylab = '' Expired ventila- tion '')
text(locator(1),paste(''Correlation of oxygen uptake
and expired ventilation= '',round(cor(Oxygen,Ventilation),digits=2)))
MODEL_lm(Ventilation~Oxygen)
abline(MODEL)
```

Príklad 87 *ANAEROB lineárny model. Popíšte výstup z modelu. Použite dáta ANAEROB a premenné Oxygen a Ventilation.*

```
summary(MODEL)
```

Príklad 88 *Ako otestovať nulovosť všetkých regresných koeficientov okrem interceptu? Použite dáta ANAEROB a premenné Oxygen a Ventilation.*

Návod: vypočítajte SS_e a SS_R
`k` ... počet parametrov v modeli
`df.model` ... modelové df

```
df.model_summary(MODEL)$df[2]
SSy_sum((DATA$y-mean(DATA$y))^2)
SSe_sum(MODEL$res^2)
```

```
F_((SSy-SSe)/(k-1))/(SSe/df.model)
```
p-hodnota je potom
```
Pvalue_1-pf(F,k-1,df.model)
```

Príklad 89 *Použitie F - štatistiky v simuláciách. Nagenerujte 1000 náhodných permutácií a vypočítajte pomer F - štatistí, ktoré presuahnu F - štatistik vypočítanú z realizácií. Použite dáta* `ANAEROB` *a premenné* `Oxygen` *a* `Ventilation`*.*

Návod:

- extrakcia F-statistiky je nasledovná

```
MODEL.Fstat_summary(MODEL)$fstat
```
jednotlivé položky výstupu sú ... `[1] value, [2] numdf, [3] dendf`

- náhodné permutácie sú generované funkciou ... `sample(VEKTOR)` ... ide o tzv. neparametrický *permutačný test*

```
attach(DATA)
# DATA majú odpoveď Y a maticu plánu (realizácií) X
MODEL_lm(y~X,data= DATA)
Fstat_summary(MODEL)$fstat[1]
F.stat_numeric(1000)
for (i in 1:1000)
{
MODELsimul_lm(sample(y)~X,data= DATA)
F.stat[i]_summary(MODELsimul)$fstat[1]
}
PROP_length(F.stat[F.stat > Fstat])/1000
# hodnota bude blízka p-hodnote vypočítanej na základe teórie normality, pokiaľ nie sú silne porušené predpoklady modelu
PROP
```

Príklad 90 *Testovanie nulovosti regresných koeficientov samostatne. Použite dáta* `ANAEROB` *a premenné* `Oxygen` *a* `Ventilation`*.*

Návod:
```
# c ... hodnota regresného koeficientu za platnosti nulovej hypotézy.
# MODEL$coef ... extrahovanie odhadov hodnôt jednotlivých regresných koeficientov
odhad.beta.i_MODEL$coef[i]
Ti_(odhad.beta-c)/se(odhad.beta)
# p-hodnota je potom
2*pt(Ti,df.model)
```

Exercise 91 *Zo stránky* `http://cran.r-project.org/` *stiahnite knižnicu "ellipse" a vložte z nej programy* `ellipse, ellipse.default, ellipse.lm, ellipse.profile, pairs. profile` *do S-PLUS. Použite dáta* `ANAEROB` *a premenné* `Oxygen` *a* `Ventilation`*.*

Návod:
```
plot(ellipse(MODEL,c(1,2)),type=''l'')
# nezabudnite na limitáciu osí !!!
points(0,0)
points(MODEL$coef[1],MODEL$coef[2])
```

Príklad 92 *IS pre modelom odhadnutú odpoveď a pre modelom odhadnutú priemernú odpoveď. Použite dáta* `ANAEROB` *a premenné* `Oxygen` *a* `Ventilation`*.*

Návod:

- `n` ... počet štatistických jednotiek v modeli

ak zadefinujeme nejakú hodnotu `x0`, potom odhad `y0`

```
y0_ MODEL$coef[1]+x0*MODEL$coef[2]
t.cv_qt(0.975,df.model)
# ak X je matica plánu, potom (X^T X)^{-1} bude
X_as.matrix(X)
XtXi_solve(t(X)%*%X)
# (1) ak 95%CI pre modelom odhadnutú priemernú odpoveď je tvaru ŷ0±BM, potom
BM1_sqrt(x0%*%XtXi%*%x0)*t.cv*sqrt(sum(MODEL$res^2)/(n-p))
95CI_c(y0-BM1,y0+BM1)
# (2) ak 95%CI pre modelom odhadnutú odpoveď je tvaru ŷ0±BM, potom
BM2_sqrt(1+x0%*%XtXi%*%x0)*t.cv*sqrt(sum(MODEL$res^2)/(n-p)
95CI_c(y0-BM2,y0+BM2)
```

- pomocou preddefinovaných funkcií:

```
# dostaneme BM1 pre prípad (1)
predict(MODEL,x0,se=T)...[1] fit, [2] se.fit, [3] df
```

Príklad 93 *Nakreslite pás spoľahlivosti pre modelom odhadnuté priemerné odpovede. Použite dáta* ***ANAEROB*** *a premenné* ***Oxygen*** *a* ***Ventilation****.*

Návod:

```
attach(DATA)
# X1 je stĺpec v DATA
# a=min(X1),b=max(X1)
a_min(Weight)
b_max(Weight)
X1.seq_seq(a,b,by=1)
PRED_predict(MODEL,data.frame(X1=X1.seq),se=T)
cv_qt(0.975,summary(MODEL)$df[2]
matplot(X1.seq,cbind(PRED$fit,PRED$fit-cv*PRED$se,PRED$fit+cv*PRED$se),
lty=c(1,2,2), type=''l'', xlab=''X1'',ylab=''Y'')
rug(DATA$X1)
points(X1,y,pch=16)
```

Príklad 94 *Čo tak kvadratická regresia? Použite dáta* ***ANAEROB*** *a premenné* ***Oxygen*** *a* ***Ventilation****. Nakreslite aj graf s pásom spoľahlivosti pre modelom odhadnuté odpovede a pre modelom odhadnuté priemerné odpovede. Urobte taký istý graf aj pre priamku.*

Návod:

```
# lineárna regresia - priamka
plot(X1,Y)
n <- length(Y)
X <- cbind(rep(1,n),X1)
# BETA <- solve(X, Y)
BETA <- MODEL$coef
minX1 <- min(X1)
maxX1 <- max(X1)
minY <- min(Y)
maxY <- max(Y)
SIMUL.X <- seq(minX1, maxX1,by=0.01)
```

```
SIMUL.X <- cbind(rep(1,length(SIMUL.X)), SIMUL.X)
ODHAD.Y <- SIMUL.X%*%BETA
SIGMA.odh <- sqrt(sum(X%*%BETA-Y)^2)/(n-2)
CONF.INT <- qt(0.975,n-2)*SIGMA.odh * sqrt(diag(SIMUL.X%*%solve
(crossprod(X)) %*% t(SIMUL.X)))
PRED.INT <- qt(0.975,n-2)*SIGMA.odh * sqrt(1+diag(SIMUL.X%*%solve
(crossprod(X)) %*% t(SIMUL.X)))
plot(X1,Y,xlim=c(minX1, maxX1),ylim=c(minY, maxY))
lines(SIMUL.X1, ODHAD.Y)
lines(SIMUL.X1, ODHAD.Y-CONF.INT,lty=2)
lines(SIMUL.X1, ODHAD.Y+CONF.INT,lty=2)
lines(SIMUL.X1, ODHAD.Y-PRED.INT,lty=3)
lines(SIMUL.X1, ODHAD.Y+PRED.INT,lty=3)
rug(X1)
# lineárna regresia - parabola
MODELkvadr_lm(Y ~X1 + I(X1^2), data=DATA)
plot(X1,Y)
n <- length(Y)
X <- cbind(rep(1,n),X1,X1^2)
# BETA <- solve(X, Y)
BETA <- MODELkvadr$coef
minX1 <- min(X1)
maxX1 <- max(X1)
minY <- min(Y)
maxY <- max(Y)
SIMUL.X <- seq(minX1, maxX1,by=0.01)
SIMUL.X <- cbind(rep(1,length(SIMUL.X)), SIMUL.X, SIMUL.X^2)
ODHAD.Y <- SIMUL.X%*%BETA
SIGMA.odh <- sqrt(sum((X%*%BETA-Y)^2)/(n-3))
CONF.INT <- qt(0.975,n-3)*SIGMA.odh * sqrt(diag(SIMUL.X%*%solve
(crossprod(X)) %*% t(SIMUL.X)))
PRED.INT <- qt(0.975,n-3)*SIGMA.odh * sqrt(1+diag(SIMUL.X%*%solve
(crossprod(X)) %*% t(SIMUL.X)))
plot(X1,Y,xlim=c(minX1, maxX1),ylim=c(minY, maxY))
lines(SIMUL.X1, ODHAD.Y)
lines(SIMUL.X1, ODHAD.Y-CONF.INT,lty=2)
lines(SIMUL.X1, ODHAD.Y+CONF.INT,lty=2)
lines(SIMUL.X1, ODHAD.Y-PRED.INT,lty=3)
lines(SIMUL.X1, ODHAD.Y+PRED.INT,lty=3)
rug(X1)
```

Príklad 95 *Testovanie submodelov. Použite výsledky modelov* `MODEL` *a* `MODELkvadr` *uvedených vyššie.*

Návod:

- pre *F-štatistiku*

```
F_sum(nested.model$res^2-full.model$res^2)/full.model$res^2 *
df.full/(df.nested-df.full)
```

p-hodnota je potom `1-pf(F,df.nested-df.full,df.full)`

- pre *t-štatistiku*

`T_sqrt(F)`
p-hodnota je potom `2*(1-pt(T,df.full))`

- Test H_0 : $\texttt{model}_\omega$, H_1 : $\texttt{model}_\Omega$

`anova(nested.model,full.model)`

Príklad 96 *Regresná diagnostika*

`plot(MODEL)`

11 Príklad analýzy dát v R

Dáta pochádzajú z autorovej databázy z gynekologického projektu "Sledovanie niektorých parametrov oxidačného stresu počas pôrodu"vo Fakultnej Nemocnici s Poliklinikou Bratislava, pracovisko Ružinov, ktorého sa zučastnil v roku 2002.

Význam stĺpcov je nasledovný:

1. malondialdehyd (MDA) je látka, ktorá sa meria v krvi novorodencov a rodičiek a jej zvýšená hladina indikuje stres, IMDA je meranie prvé a IIMDA je meranie druhé, hypotéza je, že IIMDA by malo byť menšie ako IMDA, teda stres časom klesá
2. lipofuscin detto, ILipo, IILipo
3. vek = vek rodičky
4. t.gr = počet týždňov gravidity
5. poc.deti = počet detí rodičky
6. hmotnost = hmotnosť dieťaťa v gramoch
7. status.AS1min = do7, nad7 (nad7 - dieťa je v dobrej kondícii, pod7 - potrebuje dodatkovú špeciálnu starostlivosť)
8. AS5min

APGAR skóre (AS, z anglického Activity, Pulse, Grimace, Appearance and Respiration) v Apgar teste na hodnotenie kondície novorodenca hneď po pôrode, v 1. a 5.min, tu je nasledovných 5 skórovaných premenných (dávajú sa body od 0b po 2b, spolu teda 10)

- pulzová frekvencia (2b - nad 100 úderov/min, 1b - pod 100 úderov/min, 0b - bez pulzu),
- dýchanie (rýchlosť a sila, 2b - normálne, 1b - pomalé alebo nezvyčajné, 0b absentujúce),
- aktivita a svalový tonus (2b - aktívny, spontánny pohyb, 1b - ruky a nohy skrčené s malým pohybom, 0b - bez pohybu, tzv. "floppy tonus"),
- grimasová odpoveď (známa ako "reflex irritability", 2b - odťahovanie na podnet, kýchanie alebo kašľanie, 1b- iba tvárové grimasy so stimuláciou, 0b - absencia, bez odpovede na stimuláciu),
- farba a vzhľad pokožky (2b - normálna farba všade, aj ruky a nohy sú ružové, 1b - normálna farba, ale ruky a nohy sú namodralé, 0b - modro-sivá alebo bledá farba všade).

Ak bolo AS v 1min po pôrode pod7 a nezlepšilo sa do tejto 5min, sestrička musí pokračovať v "podpore"dieťaťa. AS je teda monitoring momentálneho stavu kondície a zdravia dieťaťa a môže byť použité aj v domácej starostlivosti neskôr.

9. status.pH = pH krvi menšie alebo vačšie ako 7.2, teda pod7.2, nad7.2 10. status.IDP = trvanie prvej doby pôrodnej, t.j. do4hod, nad4hod 11. status.IIDP = trvanie druhej doby pôrodnej, t.j. do10min, nad10min 12. status.porod = spontanny a cisarsky.rez

```
# ---------------------------------------------------------------
# PRIKLAD: R-kod
# ---------------------------------------------------------------

# desatinna bodka
# "<-" priradovanie
# pomocnik ... help(FUNKCIA)

# matematicke a zakladne operacie

# >, <, >=, <=, ==
# odmocnina ... sqrt()
# absolutna hodnota ... abs()
# mocnina ... cislo^exponent
# +,-,*,/ \
# logaritmus ... log(ZAKLAD,BAZA)
# exp(EXPONENT) ... "e"
# zoradenie podla velkosti ... sort(VEKTOR)
```

```
# vektor ... c(CISLO,...,CISLO)
# matica ... matrix(CISLA,pocet.riadkov,pocet.stlpcov)

# ------------------------------------------------------------

# nacitanie dat
# nazov v uvodzovkach + cesta
# header=T ... hlavicka
NOVOROD<-read.table("D:/Dokumenty/SURNE/WEB/asnovorod.txt",header=T)
RODICKY<-read.table("D:/Dokumenty/SURNE/WEB/asdatab.txt",header=T)

# nacitanie nazvov stlpcov

names(NOVOROD)
#  [1] "PC"            "MDA"           "lipofuscin"    "vek"
#  [5] "t.gr"          "poc.deti"      "pH"            "hmotnost"
#  [9] "AS5min"        "status.pH"     "status.IDP"    "status.IIDP"
# [13] "status.AS1min"

names(RODICKY)
#  [1] "pc"           "IMDA"         "IIMDA"        "ILipo"        "IILipo"
#  [6] "vek"          "poc.deti"     "t.gr"         "status.porod" "status.pH"
# [11] "status.IDP"   "status.IIDP"

# priame citanie nazvov stlpcov pri analyze

attach(NOVOROD)
attach(RODICKY)

# zobrazenie dat v tabulke

edit(NOVOROD)

# ------------------------------------------------------------------

# zakladne charakteristiky polohy


summary(NOVOROD)


#        PC             MDA           lipofuscin          vek
#  Min.   : 1.00   Min.   :0.1500   Min.   : 6.130   Min.   :20.00
#  1st Qu.:13.75   1st Qu.:0.4850   1st Qu.: 8.623   1st Qu.:27.00
#  Median :26.50   Median :0.6300   Median : 9.800   Median :29.00
#  Mean   :26.50   Mean   :0.6033   Mean   : 9.785   Mean   :30.25
#  3rd Qu.:39.25   3rd Qu.:0.7000   3rd Qu.:10.903   3rd Qu.:33.00
#  Max.   :52.00   Max.   :1.1700   Max.   :12.620   Max.   :40.00
#       t.gr          poc.deti           pH           hmotnost
#  Min.   :36.00   Min.   :1.000   Min.   :7.030   Min.   :2610
#  1st Qu.:39.00   1st Qu.:1.000   1st Qu.:7.190   1st Qu.:3113
#  Median :40.00   Median :1.500   Median :7.297   Median :3470
#  Mean   :39.83   Mean   :1.635   Mean   :7.275   Mean   :3460
```

```
#  3rd Qu.:41.00   3rd Qu.:2.000   3rd Qu.:7.343   3rd Qu.:3713
#  Max.   :41.00   Max.   :4.000   Max.   :7.457   Max.   :4450
#      AS5min        status.pH    status.IDP   status.IIDP status.AS1min
#  Min.   : 7.000   nad7.2:37   do4hod :24   do10min :26   do7 : 8
#  1st Qu.:10.000   pod7.2:15   nad4hod:28   nad10min:26   nad7:44
#  Median :10.000
#  Mean   : 9.885
#  3rd Qu.:10.000
#  Max.   :10.000

# frekvencna tabulka

table(poc.deti,status.porod)
#         status.porod
# poc.deti cisarsky.rez spontanny
#        1  9           32
#        2  7           13
#        3  2            2
#        7  0            1

#  priemer, rozptyl, smerodajna odchylka, minimum, maximum, kvantily
#  (0,0.25,0.5,0.75,1) - v zatvorke je vektor pozorovani (realizacii)

mean(hmotnost)
#  [1] 3460

var(hmotnost)
#  [1] 178772.5

sqrt(var(hmotnost))
#  [1] 422.815

min(hmotnost)
#  [1] 2610

max(hmotnost)
#  [1] 4450

quantile(hmotnost,c(0,0.25,0.5,0.75,1))
#      0%    25%    50%    75%   100%
#  2610.0 3112.5 3470.0 3712.5 4450.0

# dlzka vektora, pocet riadkov matice - pocet pozorovani

length(hmotnost)
# [1] 52

nrow(NOVOROD)
# [1] 52

# pokial mame v datach chybajuce hodnoty, treba urobit upravu tak, aby boli
# chybajuce hodnoty vyradene - nova funkcia (priklady mean a var)
```

```
Mmean <- function(x) mean(x[!is.na(x)])
Mvar <- function(x) var(x[!is.na(x)])

Mmean(hmotnost)
Mvar(hmotnost)


# vypocet napr. priemeru vnutri kategorialnej premennej, napr. piemer IMDA pre
# oba typy porodov

sapply(split(IMDA,status.porod),mean)

# cisarsky.rez    spontanny
#     0.755000     1.008333


# co tak priradenie nazvu vysledku

AP.IMDA <- sapply(split(IMDA,status.porod),mean)

# co tak zaokruhlenie na 3 desatinne miesta

round(AP.IMDA,3)

# cisarsky.rez    spontanny
#        0.755        1.008

# ------------------------------------------------------------------

# GRAFICKE ZNAZORNENIA

# histogram (absolutna a relativna skala) - nastavenie priblizneho poctu intervalov
# "nclass", vypisanie stredov intervalov, poctu pozorovani v nich

hist(hmotnost)
hist(hmotnost, probability=T)
hist(hmotnost,probability=T, nclass=20)
x <- hist(hmotnost,nclass=20)

# boxplot
# "varwidth=T" ... sirka krabiciek == sqrt(ROZSAH)
# "notch=T" ... IS okolo medianu
# "outpch=16" ... typ bodiek

boxplot(split(IMDA,status.porod),varwidth=T,notch=T,outpch=16)

# vkreslenie aritmetickych priemerov do boxplotu

apIMDA<-sapply(split(IMDA,status.porod),mean)
points(apIMDA,pch=16)

# kvantilovy graf (qq plot)
```

```
qqnorm(IMDA)
qqline(IMDA)

# obrazok typu "scatter plot" + popisy

plot(IMDA,IIMDA,pch=16, main="Vztah IMDA a IIMDA", xlab="IMDA",ylab="IIMDA")

plot(IMDA,IIMDA, main="Vztah IMDA a IIMDA", xlab="IMDA",ylab="IIMDA",type="n")
points(IMDA[status.porod=="cisarsky.rez"],IIMDA[status.porod=="cisarsky.rez"],pch=2)
points(IMDA[status.porod=="cisarsky.rez"],IIMDA[status.porod=="cisarsky.rez"],pch=16)
legend(5,20,c("spontanny","cisarsky.rez"),pch=c(2,16,))

# CDF (kumulativna distribucna fukcia), empiricka a simulovana "rnorm" s aritmetickym
# priemerom a sd ako data

plot.ecdf(hmotnost)
plot.ecdf(rnorm(1000,mean=mean(hmotnost),sd=sqrt(var(hmotnost))))

# graf hustoty

plot(density(hmotnost))

# ------------------------------------------------------------------

# STATISTICKA INFERENCIA

# KS test normality


ks.test(hmotnost,"pnorm",mean(hmotnost),sqrt(var(hmotnost)))

#         One-sample Kolmogorov-Smirnov test

# data:  hmotnost
# D = 0.0647, p-value = 0.9815
# alternative hypothesis: two.sided

ks.test(IMDA,IIMDA)

#         Two-sample Kolmogorov-Smirnov test

# data:  IMDA and IIMDA
# D = 0.1515, p-value = 0.4349
# alternative hypothesis: two.sided

#  F - test na porovnanie rozptylov

var.test(IMDA,IIMDA)

#  F test to compare two variances

# data:  IMDA and IIMDA
# F = 1.5024, num df = 65, denom df = 65, p-value = 0.1033
```

```
# alternative hypothesis: true ratio of variances is not equal to 1
# 95 percent confidence interval:
#  0.9199607 2.4536334
# sample estimates:
# ratio of variances
#           1.502413


# t - test na porovnanie populacnych strednych hodnot

t.test(IMDA,IIMDA)

#         Welch Two Sample t-test

# data:  IMDA and IIMDA
# t = 1.3711, df = 124.963, p-value = 0.1728
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -0.03896776  0.21472534
# sample estimates:
# mean of x mean of y
# 0.9392424 0.8513636

t.test(IMDA,IIMDA,paired=T)

#         Paired t-test

# data:  IMDA and IIMDA
# t = 2.0637, df = 65, p-value = 0.04305
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  0.002833598 0.172923977
# sample estimates:
# mean of the differences
#              0.08787879


# VOLBY v t - teste
# alternative = "two.sided", alternative = "less", alternative = "greater"
# paired = F, paired = T
# var.equal = F, var.equal = T
# conf.level = 0.95
# mu = 0 ... volba "strednej hodnoty za platnosti nulovej hypotezy"

# neparametricky pristup

wilcox.test(IMDA,IIMDA,paired=T)

#         Wilcoxon signed rank test with continuity correction

# data:  IMDA and IIMDA
# V = 1418.5, p-value = 0.04589
# alternative hypothesis: true mu is not equal to 0
```

```
# ------------------------------------------------------------------

# KORELACNA ANALYZA

cor.test(IMDA,IIMDA)

#  Pearson's product-moment correlation

# data:  IMDA and IIMDA
# t = 5.5522, df = 64, p-value = 5.811e-07
# alternative hypothesis: true correlation is not equal to 0
# 95 percent confidence interval:
#  0.3806653 0.7137084
# sample estimates:
#       cor
# 0.5701667

# VOLBY
# alternative = "two.sided", alternative = "less", alternative = "greater"
# method = "pearson", method = "spearman"
# conf.level = 0.95

# ------------------------------------------------------------------

# LINEARNY REGRESNY MODEL

MDAlm01 <- lm(IMDA ~ IIMDA, data = RODICKY)
summary(MDAlm01)

# Call:
# lm(formula = IMDA ~ IIMDA, data = RODICKY)

# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.86898 -0.22835 -0.06587  0.16532  0.93290

# Coefficients:
            Estimate Std. Error t value Pr(>|t|)
# (Intercept)   0.3442     0.1148   2.999  0.00385 **
# IIMDA         0.6989     0.1259   5.552 5.81e-07 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Residual standard error: 0.334 on 64 degrees of freedom
# Multiple R-Squared: 0.3251,     Adjusted R-squared: 0.3145
# F-statistic: 30.83 on 1 and 64 DF,  p-value: 5.811e-07

hist(resid(MDAlm01))
qqnorm(resid(MDAlm01))
qqline(resid(MDAlm01))

# kvantilovy graf, tiez je tu nahlad na normalitu dat, ak su data
```

```
# okolo priamky rozptylene len malo, je to OK, konce casto "utekaju",
# tu to nie je take kriticke - model je OK

plot(fitted(MDAlm01),resid(MDAlm01))

# obrazok fitovanych hodnot voci rezidualom - nesmie byt vidiet nijaku
# zavislost, inak je model ZLY (tu je to OK)

# ISs a predikcne intervaly okolo regresnej priamky

plot(IIMDA,IMDA)
n <- length(IMDA)
X <- cbind(rep(1,n),IMDA)
# BETA <- solve(X, IIMDA)
BETA <- MDAlm01$coef
minIMDA <- min(IMDA)
maxIMDA <- max(IMDA)
minIIMDA <- min(IIMDA)
maxIIMDA <- max(IMDA)
SIMUL.IMDA <- seq(minIMDA, maxIMDA,by=0.01)
SIMUL.X <- cbind(rep(1,length(SIMUL.IMDA)), SIMUL.IMDA)
SIMUL.IIMDA <- SIMUL.X%*%BETA
SIGMA.odh <- sqrt(sum(X%*%BETA-IIMDA)^2)/(n-2)
CONF.INT <- qt(0.975,n-2)*SIGMA.odh * sqrt(diag(SIMUL.X%*%solve(crossprod(X))%*%
t(SIMUL.X)))
PRED.INT <- qt(0.975,n-2)*SIGMA.odh * sqrt(1+diag(SIMUL.X%*%solve(crossprod(X))%*%
t(SIMUL.X)))
plot(IIMDA,IMDA,xlim=c(minIMDA, maxIMDA),ylim=c(minIIMDA, maxIIMDA))
lines(SIMUL.IMDA,SIMUL.IIMDA)
lines(SIMUL.IMDA,SIMUL.IIMDA-CONF.INT,lty=2)
lines(SIMUL.IMDA,SIMUL.IIMDA+CONF.INT,lty=2)
lines(SIMUL.IMDA,SIMUL.IIMDA-PRED.INT,lty=3)
lines(SIMUL.IMDA,SIMUL.IIMDA+PRED.INT,lty=3)
rug(IMDA)
# ------------------------------------------------------------------
```