EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Statistické zpracování dat v energetice

Petra Schreiberová

VŠB TECHNICKÁ UNIVERZITA OSTRAVA

Poděkování

Studijní materiál byl vytvořen v rámci realizace projektu „Vzdělávání pro praxi - inovace studijních programů VŠB-TU Ostrava“, registrační číslo CZ.02.2.69/0.0/0.0/18_056/0013302.

ISBN 978-80-248-4651-4

Obsah

1 Základy programu R

R je prostředí pro statistické výpočty, jedná se o volně šiřitelnou implementaci jazyka S, které je dostupné zdarma na adrese `http://www.r-project.org`.

Po spuštění programu se otevře okno s úvodním textem a v posledním řádku se objeví symbol >. Pro nastavení vhodného pracovní adresáře použijeme posloupnost příkazů z nabídky **Menu - File - Change dir**. Text za znaménkem # je komentář. Další knihovny nainstalujeme příkazem `install.packages()` a nápovědu vyvoláme příkazem `help(xxx)` nebo zadáním `?xxx`.

1.1 Aritmetické operace, funkce

Pro sčítání používáme plus, pro odečítání minus (případně pomlčka), pro násobení hvězdička a pro dělení lomítko.

```
> 1+1
[1] 2
> 3-1
[1] 2
> 3*1
[1] 3
> 3/2
[1] 1.5
```

R používá desetinnou tečku, počet desetinných míst změníme funkcí `options(digits=n)`, kde za `n` zadáme o jedno menší číslo. Odmocninu zadáváme klasicky, zbytek po dělení jako %% a dělení beze zbytku %/%.

Na jediné číslo se pohlíží jako na jednoprvkový vektor, tzn. že pracujeme s posloupnostmi čísel (vektory). Výhodou je, že lze provést více výpočtů najednou. Vektor zadáme pomocí funkce `c()`. V případě nestejně dlouhých vektorů je dána délka výsledku délkou delšího vektoru, přičemž kratší vektor se opakuje.

Řešený příklad

Určete hodnoty BMI pro 3 pacienty jako váha v kilogramech vydělená druhou mocninou výšky v metrech.

```
> c(56,85,74)/(c(167,172,192)/100)^2
[1] 20.07960 28.73175 20.07378
```

R obsahuje základní matematické funkce a konstanty.

`cos()`	funkce kosinus
`sin()`	funkce sinus
`log()`	přirozený logaritmus
`log10()`	dekadický logaritmus
`exp(1)`	Eulerova konstanta
`pi`	konstanta π

1.2 Proměnné, vektory, objekty

Je vhodné vytvořit proměnnou, ke které přiřadíme číslo (vektor). Jako přiřazovací znaménko lze použít buď $=$ nebo kombinace znaků $<-$, která je vhodnější a v některých případech se jedná o jedinou možnost. Název proměnné může tvořit jakákoli kombinace číslic a písmen (musí začínat písmenem) a dalších symbolů. Je třeba rozlišovat velká a malá písmena. Není vhodné používat název "data". Další možností je využití funkce `assign()`. Počet prvků vektoru vypíšeme pomocí funkce `length()`.

```
> vyska <- c(164,186,192)
> vyska
[1] 164 186 192
> length(vyska)
[1] 3
> x <- FALSE
> x
[1] FALSE
> y <- "ahoj"
> y
[1] ahoj
> assign("vyska m",158)
> `vyska m`
[1] 158
```

Typ proměnné vypíšeme pomocí příkazu `class()`.

```
> x=2
> x
[1] 2
> class(x)
[1] "numeric"
```

Proměnné, které již nebudou potřeba můžeme vymazat pomocí příkazu `rm`.

Vektor lze zadat i aritmetickou posloupností, případně využít funkcí `seq(from=a,to=b)`, `seq(from=a,by=k,length.out=n)`, `rep((a:b),times=k)`, `rep((a:b),each=k)`. K jednotlivým pozicím přistupujeme pomocí [].

```
> v <- 1:5
> v
```

```
[1] 1 2 3 4 5
> v[3]
[1] 3
> v[2:4]
[1] 2 3 4
> seq(from=4,by=-0.1,length.out=5)
[1] 4.0 3.9 3.8 3.7 3.6
> rep((1:4),each=2)
[1] 1 1 2 2 3 3 4 4
```

Matici vytvoříme příkazem `matrix()`. V případě vyplňování po řádcích, musíme nastavit argument "byrow"na TRUE.

```
> A<-matrix(1:9,nrow=3,ncol=3)
> A
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> B<-matrix(1:9,nrow=3,ncol=3,byrow=TRUE)
> B
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

Vektory lze do matice spojovat horizontálně či vertikálně - příkazy `cbind, rbind`.

```
> C <- cbind(A,c(5,5,5))
> C
     [,1] [,2] [,3] [,4]
[1,]    1    4    7    5
[2,]    2    5    8    5
[3,]    3    6    9    5
```

Matice může být tvořena i náhodnými čísly vygenerovanými funkcí `runif()` nebo posloupností čísel vytvořenou pomocí příkazu `seq()`. Transponování matice provádíme příkazem `t()`, vlastní čísla vypíše funkce `eigen()`. V případě lineárních rovnic lze využít funkci `det()` k nalezení determinantu.

K jednotlivým prvkům vektoru, matice přistupujeme různými způsoby.

```
> C[2,]
[1] 2 5 8 5
> C[1,3]
[1] 7
> C[2,c(3,4)]
[1] 8 5
```

Datové tabulky - datové tabulky lze zadat pomocí funkce `data.frame()`.

```
> mesic<-c("leden","duben","cervenec","rijen")
> t<-c(25,23,11,15)
> v<-c(33,25,21,37)
> tabulka<-data.frame(mesic,t,v)
> tabulka
     mesic  t  v
1    leden 25 33
2    duben 23 25
3 cervenec 11 21
4    rijen 15 37
```

Standardní **logické operace** jsou & - AND, | - OR, ! - negace.

1.3 Vstup a výstup souborů

Pomocí příkazu `read.table()` lze načíst data uložené v textovém souboru. Argument `header=FALSE` určuje, že v prvním řádku jsou rovnou data a ne názvy. Dalšími užitečnými argumenty jsou `sep=""` pro změnu oddělovače hodnot a `dec=","`, který umožňuje načíst data oddělené desetinnou čárkou.

Tabulku v EXCEL můžeme zkopírovat do schránky a pak příkazem `read.delim("clipboard")` importovat do R. V případě, že jsou hodnoty oddělené desetinnou čárkou, použijeme funkci `read.delim2("clipboard")`.

Pro výstup do souboru použijeme funkci `write.table()`.

2 Popisná statistika

V případě statistického zpracování dat je používaným objektem "data.frame", který vytvoříme funkcí `data.frame(col1,col2,...)`. Editovat hodnoty lze příkazem `edit()`.

Řešený příklad

Vytvořte data frame pro rozměry součástky. Vypište délku pro všechny součástky.

```
> soucID <- c(1,2,3)
> delka <- c(25,28,27)
> sirka <- c(1.2, 1.5, 1.3)
> soucdata <- data.frame(soucID,delka,sirka)
> soucdata
  soucID delka sirka
1      1    25   1.2
2      2    28   1.5
3      3    27   1.3
> soucdata$delka
[1] 25 28 27
```

Data ze souborů lze načíst příkazem `read.table()`.

```
> pokus <- read.table(file="pokus.txt", header=FALSE)
> pokus
  V1 V2 V3
1  1  2  3
2  4  5  6
> pokus[,1]
[1] 1 4
> pokus[,1:2]
  V1 V2
1  1  2
2  4  5
> names(pokus)
[1] "V1" "V2" "V3"
```

Příkaz `dim()` zobrazí počet řádků a sloupců v souboru.

Popisná statistika bývá prvním krokem k odhalení informací skrytých ve velkém množství proměnných a jejich variant. Využíváme základních číselných charakteristik.

2.1 Četnosti, číselné charakteristiky

absolutní četnost f_i - počet prvků souboru spadající do i-té třídy

relativní četnost - poměr četnosti třídy k celkovému počtu dat

$$\varphi_i = \frac{f_i}{n}$$

kumulativní absolutní četnost - počet prvků souboru v i-té třídě a třídách předcházejících

$$F_i = \sum_{l=1}^{i} f_l$$

kumulativní relativní četnost - poměr kumulativní četnosti třídy k celkovému počtu dat

$$\phi_i = \frac{F_i}{n}$$

Řešený příklad

Náhodně vygenerujeme 55 celých čísel z intervalu $[25, 60]$ a určíme četnosti pro 7 tříd.

```
> x<-floor(runif(55, min=25, max=60))
> range(x)
[1] 26 59
> length(x)
[1] 55
> tridy = seq(25, 60, by=5)
> tridy
[1] 25 30 35 40 45 50 55 60
> x.tr = cut(x, tridy, right=FALSE)   # FALSE znamena uzavreny
   interval zleva
> x.fr = table(x.tr)
> x.fr
x.tr
[25,30) [30,35) [35,40) [40,45) [45,50) [50,55) [55,60)
      6       8       5       8      12       7       9
> cbind(x.fr)
        x.fr
[25,30)    6
[30,35)    8
[35,40)    5
[40,45)    8
[45,50)   12
[50,55)    7
[55,60)    9
> x.relfr = x.fr / length(x)   #vypocet relativnich cetnosti
> x.relfr
x.tr
   [25,30)    [30,35)    [35,40)    [40,45)    [45,50)    [50,55)
      [55,60)
```

```
0.10909091 0.14545455 0.09090909 0.14545455 0.21818182 0.12727273
   0.16363636
> options(digits=1)
> x.relfr
x.tr
[25,30) [30,35) [35,40) [40,45) [45,50) [50,55) [55,60)
   0.11    0.15    0.09    0.15    0.22    0.13    0.16
> cbind(x.fr, x.relfr)
        x.fr x.relfr
[25,30)    6    0.11
[30,35)    8    0.15
[35,40)    5    0.09
[40,45)    8    0.15
[45,50)   12    0.22
[50,55)    7    0.13
[55,60)    9    0.16
> x.cumfr = cumsum(x.fr)  #vypocet kumulativnich cetnosti
> x.cumfr
[25,30) [30,35) [35,40) [40,45) [45,50) [50,55) [55,60)
      6      14      19      27      39      46      55
> x.cumrelfr = x.cumfr / length(x)   #vypocet kumulativnich
   relativnich cetnosti
> x.cumrelfr
[25,30) [30,35) [35,40) [40,45) [45,50) [50,55) [55,60)
   0.11    0.25    0.35    0.49    0.71    0.84    1.00
> cbind(x.cumfr, x.cumrelfr)
        x.cumfr x.cumrelfr
[25,30)       6       0.11
[30,35)      14       0.25
[35,40)      19       0.35
[40,45)      27       0.49
[45,50)      39       0.71
[50,55)      46       0.84
[55,60)      55       1.00
```

aritmetický průměr - je citlivý na extrémní (odlehlé) hodnoty

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

Další využívané průměry - vážený aritmetický průměr (výpočet aritmetického průměru hodnot uspořádaných do tabulky četností), useknutý průměr, geometrický průměr (pro analýzu vývoje ukazatele v čase), harmonický průměr (v indexní teorii) a kvadratický průměr.

modus $\hat{x}$ - hodnota, která má největší absolutní četnost (z dat uspořádaných v tabulce je modus možno odhadnout jako střed třídy s nejvyšší absolutní četností); modů může být více, nebo i žádný

kvantily x_p - hodnota kvantilu říká, že $100p$ procent hodnot souboru nabývá hodnoty stejné nebo menší, než je hodnota kvantilu x_p; nejčastější kvantily jsou medián, kvartily, decily a percentily

medián $x_{0,5}; \tilde{x}$ - hodnota, která rozděluje seřazený soubor na dvě části o stejném počtu prvků
dolní kvartil $x_{0,25}$ - čtvrtina hodnot je menší nebo rovna této hodnotě

horní kvartil $x_{0,75}$ - tři čtvrtiny hodnot jsou menší nebo rovny této hodnotě

variační rozpětí - stejně jako průměr je citlivé na extrémní (odlehlé) hodnoty

$$R = x_{\max} - x_{\min}$$

interkvartilové rozpětí - rozdíl mezi horním a dolním kvartilem, není citlivý na extrémní hodnoty

$$\mathrm{IQR} = x_{0,75} - x_{0,25}$$

výběrový rozptyl - nejrozšířenější míra variability, vystihuje rozptýlení jednotlivých hodnot kolem aritmetického průměru

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2$$

výběrová směrodatná odchylka - kladná odmocnina výběrového rozptylu, je uvedena ve stejných jednotkách jako aritmetický průměr

$$s = \sqrt{s^2}$$

Nevýhodou rozptylu i směrodatné odchylky je skutečnost, že neumožňují porovnávat variabilitu proměnných vyjádřených v různých jednotkách. K tomuto slouží další charakteristika - variační koeficient.

variační koeficient - udává relativní variabilitu vztaženou k průměru, uvádí se v procentech, pomáhá odhalit odlehlé hodnoty; použití zejména při srovnání variability dvou různorodých proměnných, které jsou vyjádřeny v různých měrných jednotkách

$$V_x = \frac{s}{\bar{x}}$$

Je-li $V_x > 50$ %, může to ukazovat na nesourodost souboru a není např. vhodné používat aritmetický průměr jako charakteristiku polohy.

šikmost - vyjadřuje, jsou-li hodnoty kolem průměru rozloženy symetricky, nebo zdali převažují spíše hodnoty podprůměrné či nadprůměrné

$$A = \frac{n}{(n-1)(n-2)s^3} \sum_{j=1}^{n} (x_j - \bar{x})^3$$

- $A > 0$ - kladné zešikmení (převládají nízké hodnoty)
- $A = 0$ - symetrické (hodnoty rozloženy rovnoměrně)

- $A < 0$ - záporné zešikmení (převládají vysoké hodnoty)

špičatost - vyjadřuje, jaký průběh má rozdělení hodnot kolem zvoleného středu, čím více je rozdělení špičatější, tím více jsou hodnoty soustředěny kolem daného středu

$$\tilde{e} = \frac{n(n+1)}{(n-1)(n-2)(n-3)s^4}\sum_{j=1}^{n}(x_j - \bar{x})^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

- $\tilde{e} > 0$ - špičaté rozdělení (hodnoty koncentrovány kolem středu)
- $\tilde{e} = 0$ - normální rozdělení (hodnoty rozloženy normálně)
- $\tilde{e} < 0$ - ploché rozdělení (hodnoty nejsou koncentrovány kolem středu)

Základní polohové popisné statistiky získáme příkazem `summary()`. Pro výpočet šikmosti a špičatosti je potřeba doinstalovat balíček `install.packages("moments")`, případně i pro určení centrálních momentů balíček `install.packages("e1071")`.

Řešený příklad

Mějme datový soubor o spotřebě energie (kWh) v domácnostech. Určete absolutní, relativní četnosti a základní číselné charakteristiky pro spotřebu.

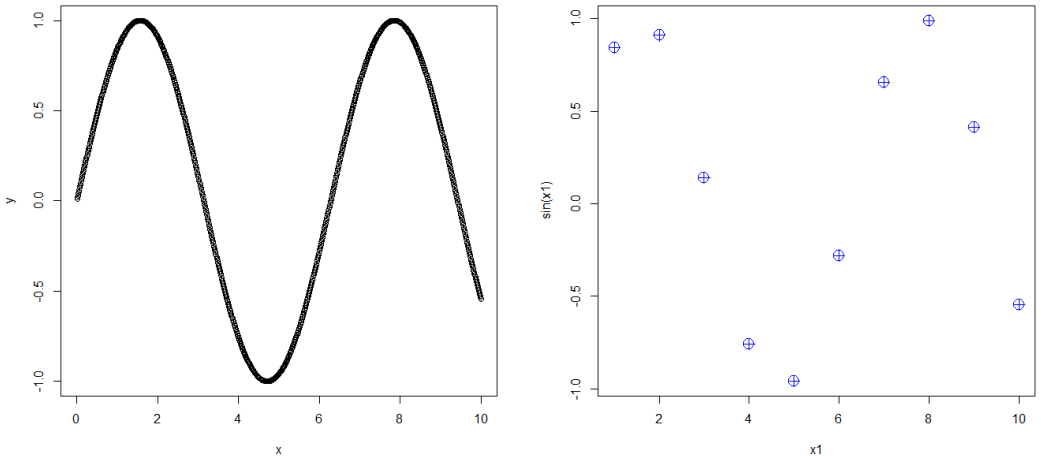
```
> spotreba <- c(3100, 2975, 3095, 2854, 3256, 3124, 2864)
> sort(spotreba) %data seřadíme
[1] 2854 2864 2975 3095 3100 3124 3256
> range(spotreba) %variační rozpětí
[1] 2854 3256
> table(spotreba) %absolutní četnosti
> prop.table(table(spotreba)) %relativní četnosti
> x <- summary(spotreba)  %základní statistiky
>x
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2854    2920    3095    3038    3112    3256
> quantile(spotreba,probs=c(0.1,0.4,0.7)) %kvantily
   10%    40%    70%
  2860.0 3023.0 3104.8
> sd(spotreba) %směrodatná odchylka
[1] 147.2602
> var(spotreba) %rozptyl
[1] 21685.57
> library(e1071)
> skewness(spotreba) %šikmost
[1] -0.01965164
> m3=moment(spotreba, order=3, center=TRUE)
> m3
[1]  -62755.99
> a3=m3/s^3
> a3                  %šikmost
[1] -0.01965164
```

```
> kurtosis(spotreba) %špičatost
[1] -1.658012
>IQR <- x[5]-x[2]  %interkvartilové rozpětí
> IQR(spotreba)
[1] 192.5
```

2.2 Grafy

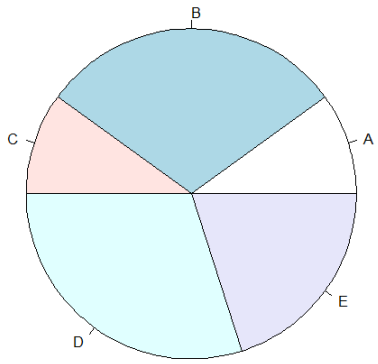
Základní funkcí je `plot()`.

```
> x<-1:1000/100
> y<-sin(x)
> plot(x,y)
> x1<-1:10
> plot(x1,sin(x1), pch=10, col="blue", cex=2)
```



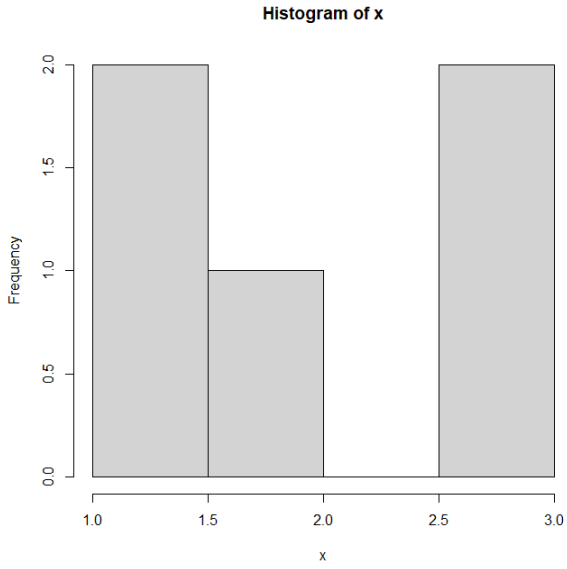
Koláčový graf je znázorněn pomocí výsečí kruhu, kde každé kategorii odpovídá jedna výseč; velikosti obsahů výsečí odpovídají četnostem kategorie.

```
> x<-c(1,3,1,3,2)
> nazev<-c("A","B","C","D","E")
> pie(x, labels=nazev)
```



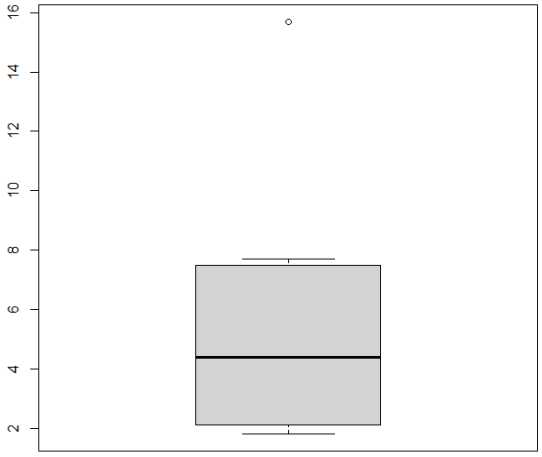
Histogram je sloupcový graf, kdy na vodorovnou osu znázorníme třídy a na svislou osu jejich četnosti; jednotlivé hodnoty četností jsou zobrazeny jako výšky sloupců.

```
> hist(x)
```



Box plot znázorňuje významné a extrémní hodnoty v souboru.

```
> y<-c(3.3,2.3,1.8,5.5,7.7,7.3,1.9,15.7)
> boxplot(y)
```


16
14
12
10
8
6
4
2

3 Náhodné veličiny

Náhodnou veličinou X rozumíme funkci $X : \Omega \to \mathbb{R}$, která zobrazí každý výsledek náhodného experimentu ω na právě jedné reálné číslo $X(\omega) = x$.

Pokud je oborem hodnot náhodné veličiny spočetná množina, jde o diskrétní náhodnou veličinu. V případě, že oborem hodnot je interval, jde o spojitou náhodnou veličinu.

3.1 Diskrétní rozdělení

Každá diskrétní náhodná veličina je popsána pravděpodobnostní funkcí, definovanou jako

$$p(x) = P(X = x).$$

Jelikož hodnoty pravděpodobnostní funkce reprezentují pravděpodobnost, musí funkce $p(x)$ splňovat následující vlastnosti:

- $p(x) > 0,\ \forall x$
- $\sum_{\forall x} p(x) = 1.$

Další důležitou funkcí definující diskrétní náhodnou veličinu je distribuční funkce

$$F(t) = P(X \leq t),\ t \in \mathbb{R}.$$

Vlastnosti:

- $F(x)$ je neklesající
- $F(x)$ je zprava spojitá
- $\lim_{t \to -\infty} = 0,\ \lim_{t \to +\infty} = 1.$

Střední (očekávaná) hodnota náhodné veličiny je dána vzorcem

$$\mu = E(X) = \sum_i x \cdot p(x).$$

Hodnotu rozptylu vypočteme jako

$$\sigma^2 = D(X) = \sum_i (x - \mu)^2 \cdot p(x),$$

případně lze použít i formuli

$$\sigma^2 = E(X^2) - E(X).$$

Přímo z rozptylu je definovaná směrodatná odchylka $\sigma = \sqrt{\sigma^2}$.

Řešený příklad

Náhodná veličina značí počet hlav ve třech hodech mincí. Popište pravděpodobnostní rozdělení této náhodné veličiny a určete střední hodnotu, rozptyl a směrodatnou odchylku.

$X \ldots$ počet hlav ve 3 hodech mincí

Pro $x = \{0, 1, 2, 3\}$ spočteme pravděpodobnosti a zapíšeme do tabulky:

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Nainstalujeme doplněk pro pravděpodobnost a knihovnu spustíme `library(prob)`.

```
> vysl <- tosscoin(3)
> vysl
  toss1 toss2 toss3
1     H     H     H
2     T     H     H
3     H     T     H
4     T     T     H
5     H     H     T
6     T     H     T
7     H     T     T
8     T     T     T
> x<-c(0,1,2,3)
> p<-c(1/8,3/8,3/8,1/8)
> A<-probspace(x,p)
> A
  x probs
1 0 0.125
2 1 0.375
3 2 0.375
4 3 0.125
> mu <- sum(x * p)
> mu
[1] 1.5
> sigma2 <- sum((x-mu)^2 * p)
> sigma2
[1] 0.75
> sigma <- sqrt(sigma2)
> sigma
[1] 0.8660254
> F = cumsum(p)
> F
[1] 0.125 0.500 0.875 1.000
```

Charakteristiky lze rovnou určit s využitím balíku `library(distrEx)`.

```
> X <- DiscreteDistribution(supp=0:3, prob = c(1,3,3,1)/8)
> E(X); var(X); sd(X)
[1] 1.5
[1] 0.75
[1] 0.8660254
```

3.2 Základní pravděpodobnostní modely

3.2.1 Alternativní rozdělení $A(p)$

Rozdělení nula-jedničkové veličiny. Popisuje výsledek náhodného pokusu, kdy v případě, že nastane sledovaný jev A, je $x = 1$ s pravděpodobností p a v případě, že jev A nenastane, je $x = 0$ s pravděpodobností $1 - p$.

Pravděpodobnostní funkce je dána

$$p(x) = p^x(1-p)^{1-x},\ x = 0, 1.$$

Vlastnosti:

$$E(X) = p,\ D(X) = p(1-p)$$

3.2.2 Binomické rozdělení $Bi(n, p)$

Jedná se o jedno z nejčastěji používaných rozdělení. Popisuje experimenty sestávající z n Bernoulliho pokusů (Pokusů, které mají dva možné výsledky ("úspěch", "neúspěch"). Pravděpodobnost "úspěchu"je p a pravděpodobnost "neúspěchu"je $1 - p$.), kde pravděpodobnost "úspěchu"je stejná pro všechny pokusy. Pokusy jsou tedy vzájemně nezávislé.

Pravděpodobnostní funkce je dána

$$p(x) = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0, 1, \ldots, n.$$

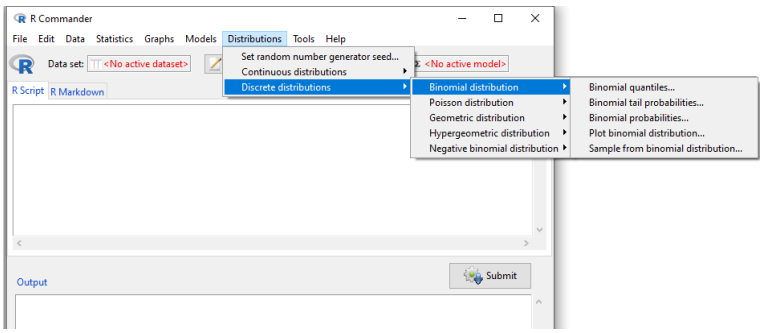
Vlastnosti:

$$E(X) = np,\ D(X) = np(1-p)$$

Pravděpodobnostní a distribuční funkce s jejich grafy lze určit v R COMMANDERU.

```
> install.packages("Rcmdr")
> library(Rcmdr)
```

Diskrétní rozdělení najdeme v menu `Distributions-Discrete distribution-Binomial distribution`.



3.2.3 Hypergeometrické rozdělení $H(N, M, n)$

Používáme ho při výběru bez vracení, tzn. závislé výběry. Je-li v populaci N sledovaný znak M-krát, pak pravděpodobnost, že ve výběru n jednotek bude nacházet právě x jednotek se sledovaným znakem je

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},\ x = \max(0, n-N+M), \ldots, \min(M, n).$$

Vlastnosti:

$$E(X) = n\frac{M}{N}, \quad D(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}.$$

Menu `Distributions-Discrete distribution-Hipergeometric distribution`, parametry zadáváme v pořadí $x, M, N-M, n$.

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

Je nejdůležitější používané rozdělení. Za určitých podmínek k němu podle centrální limitní věty konvergují jiná rozdělení. Popisuje pravděpodobnostní modely chování jevů v technice, ekonomii i přírodních vědách.

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

Vlastnosti:

$$E(X) = \mu, \quad D(X) = \sigma^2.$$

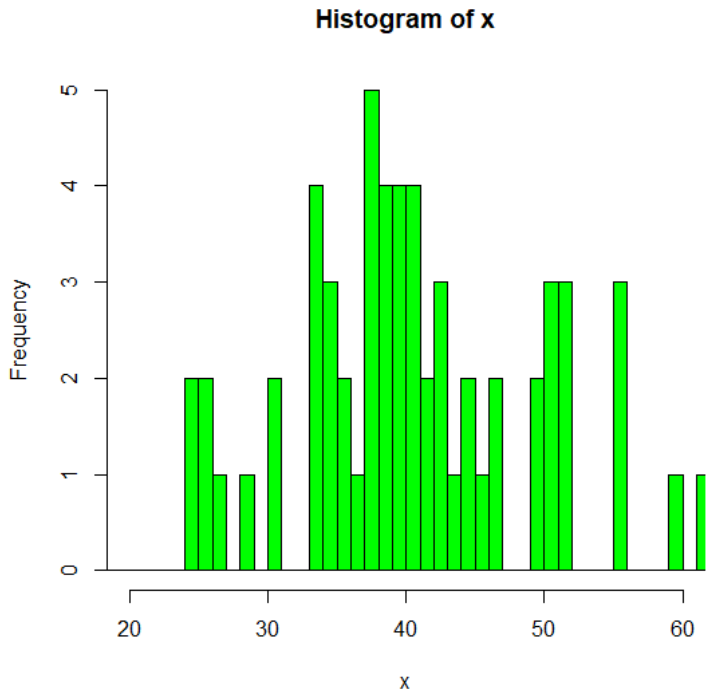
Pravidlo 3σ:

$$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9976$$

Menu `Distributions-Continuous distribution-Normal distribution`.

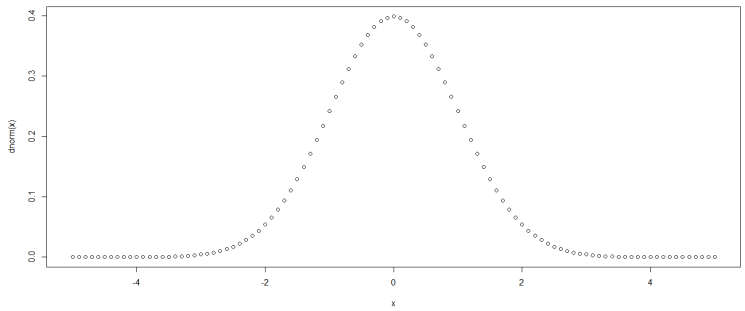
Sérii náhodných čísel si lze vygenerovat pomocí funkce `rnorm`. Vyzkoušejte, zda při větším počtu generovaných čísel se průběh funkce přibližuje ke Gaussově křivce.

```
> x <- rnorm(60, mean = 40, sd = 10)
> hist(x, br=40, xlim=c(20,60), col="green")
```



Funkce `dnorm(b)` vrací hustotu rozdělení, tj. pravděpodobnost, že naměříme hodnotu mezi b a $b + \delta x$.

```
> dnorm(0.5)
[1] 0.3520653
> x <- -50:50/10
> plot(x, dnorm(x))
```

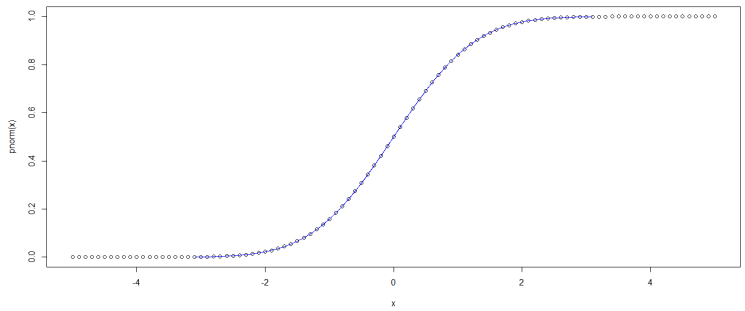


Funkce `pnorm(b)` vrací hodnotu distribuční funkce, tzn. pravděpodobnost, že pro veličinu naměříme hodnoty do b. Odchylka od postupu s využitím lichoběžníkové metody je způsobena nepřesností numerické metody.

```
> x <- -500:50/100
> 0.01*sum(dnorm(x))
[1] 0.693221
> pnorm(0.5)
[1] 0.6914625
```

Funkce `qnorm(p)` vrací hodnotu p-kvantilu. Jedná se o inverzní funkci k distribuční.

```
> qnorm(0.5, mean=20, sd=5)
[1] 20
> x <- -50:50/10
> p <- 1:999/1000
> plot(x, pnorm(x))
> lines(qnorm(p),p, col="blue")
```



Z diskrétních rozdělení je často používáno např. **Poissonovo rozdělení -** $Po(\lambda)$ - počet jevů v prostorové/časové jednotce, **geometrické rozdělení -** $Ge(p)$**, záporně binomické rozdělení -** $NBi(n,p)$ - počet neúspěchů do 1. úspěchu, resp. do n-tého úspěchu.

Ze spojitých rozdělení je v praxi používané **rovnoměrné rozdělení -** $R(a,b)$ - v simulačních metodách, **logaritmicko-normální rozdělení -** $LN(\mu,\sigma^2)$ - v teorii spolehlivosti, **exponenciánlní rozdělení -** $Exp(\lambda)$ - v teorii splehlivosti, teorii hromadné obsluhy.

Ve statistických analýzách se využívá rozdělení odvozených z normálního rozdělení - **Pearsonovo** (χ^2) **rozdělení** - funkce `dchisq, pchisq, qchisq, rchisq`, **Studentovo** $t-$ **rozdělení** - funkce `dt, pt, qt, rt`, **Fischerovo-Snedecorovo** (F) **rozdělení** - funkce `df, pf, qf, rf`.

3.2.5 Centrální limitní věty

a) **Moivreova-Laplaceova věta**

Vyjadřuje konvergenci binomického rozdělení k normálnímu. Pokud $X \sim Bi(n,p)$, pak pro velké n platí

$$U = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1).$$

Aproximace se zlepšuje s rostoucím rozptylem a dává dobré výsledky při

$$np(1-p) > 9 \;\Leftrightarrow\; \min\{np, np(1-p)\} > 5.$$

b) **Lindebergova-Lévyho věta**

Pro dosti velké n má součet i průměr nezávislých stejně rozdělených náhodných veličin se stejným rozptylem a stejnou střední hodnotou asymptoticky normální rozdělení $N(n\mu, n\sigma^2)$, resp. $N(\mu, \frac{\sigma^2}{n})$. Platí

$$U = \frac{X - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma^2}\sqrt{n} \sim N(0,1).$$

Řešený příklad

Životnost určitého výrobku se řídí exponenciálním rozdělením se střední hodnotou 3 roky. Určete pravděpodobnost, že průměrná životnost 200 prodaných kusů daného výrobku bude alespoň 42 měsíců.

X_i = životnost i-tého výrobku

$X_i \sim Exp(\frac{1}{3}) \Rightarrow E(X_i) = 3,\ D(X_i) = 9$

$\bar{X}$ = průměrná životnost 200 výrobků

Z L-L věty víme, že

$$\bar{X} = \frac{1}{200}\sum_{i=1}^{200} X_i \sim N\left(3, \frac{9}{200}\right).$$

Tedy

$$P(\bar{X} > 3.5) = 1 - F(3.5) = 0.009$$

```
> 1- pnorm(c(3.5), mean=3, sd=sqrt(9/200))
[1] 0.009211063
> pnorm(c(3.5), mean=3, sd=sqrt(9/200), lower.tail=FALSE)
[1] 0.009211063
```

Příklad

a) Nechť X a Y jsou náhodné veličiny s χ^2 rozdělením s 8 a 5 stupni volnosti. Která z následujících pravděpodobností je větší $P(X < 3)$, $P(Y > 3)$? [$P(X < 3) = 0.07$, $P(Y > 3) = 0.70$]

b) Z chovného rybníka bylo vyloveno 25 línů. Vypočítaná průměrná hmotnost byla 2.75 kg a směrodatná odchylka byla odhadnuta na 0.5 kg. V rybníku bylo nasazeno 2000 plůdků a počítá se s 10 % úmrtností. Jaká je pravděpodobnost, že výlov celého rybníka, tj. hmotnost vylovených línů přesáhne 4900 kg? [CLV, $\bar{X} \sim N(1800 \cdot 2.75, 1800 \cdot 0.5$, $P(\bar{X} > 4900) = 0.952$]

4 Odhady parametrů

V praxi je pro rozhodování důležité získat informace a využít je na odhady parametrů. Jedním ze základních úkolů statistické indukce je odhad neznámých parametrů základního souboru pomocí náhodného výběru. Intervalové odhady umožňují odhadnout nejistotu v odhadu parametru náhodné veličiny.

Používáme dva typy odhadů:

- **bodový odhad** - odhadujeme jedním číslem
- **intervalový odhad** - hledáme interval, ve kterém hledaný parametr leží

Intervalový odhad (IO) spočívá v nalezení **intervalu spolehlivosti** (T_D, T_H) pro parametr θ s pravděpodobností $1 - \alpha$, kterou označujeme jako **úroveň spolehlivosti**. Vyžadujeme co největši spolehlivost odhadu, přitom ale co nejmenší šířku intervalu. S rostoucí spolehlivostí se zvětšuje šířka IO a tím ale klesá významnost získané informace. Nejčastěji se volí spolehlivost 0.95, resp. 0.99, 0.9. Je zřejmé, že šířku IO snižuje rostoucí rozsah výběru.

Zapisujeme:

$$P(T_D < \theta < T_H) = 1 - \alpha.$$

V případě, že jsou obě meze intervalu konečné, nazýváme tento interval **oboustranný**. Je-li jedna z mezí nekonečno, hovoříme o **jednostranném intervalu**.

Meze intervalu jsou určeny na základě odhadovaného parametru, použitém náhodném výběru a jeho výběrovém rozdělení.

4.1 Intervaly spolehlivosti pro normální výběr

Předpokládejme, že náhodný výběr pochází z normálního rozdělení $N(\mu, \sigma^2)$, kde μ je odhadovaný parametr a hodnota rozptylu σ^2 je známá. Pak statistika

$$U = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$$

má rozdělení $N(0,1)$ a hledaný IO je ve tvaru

$$\mu = (\bar{x} - d,\ \bar{x} + d),\ d = \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}},$$

kde $u_{1-\frac{\alpha}{2}}$ jsou kvantily rozdělení $N(0,1)$. Hodnotu d lze interpretovat jako statistickou chybu průměru a s rostoucím rozsahem souboru její hodnota klesá.

Potřebný rozsah souboru s požadovanou chybou odhadu Δ určíme podle vzorce

$$n \geq \left(\frac{\sigma}{\Delta} u_{1-\frac{\alpha}{2}}\right)^2.$$

Řešený příklad

Napětí na jističi se měří přístrojem, jehož systematická chyba je rovna 0 a náhodné chyby se řídí normálním rozdělením se směrodatnou odchylkou 6 V. Kolik měření je potřeba provést, aby s pravděpodobností 95 % bylo stanoveno napětí s chybou menší než 3 V?

```
> sigma = 6
> alpha = 0.05
> delta = 3
> (qnorm(1-alpha/2)*sigma/delta)^2   %odhad rozsahu
[1] 15.36584
```

Ve většině reálných situací ale parametr rozptylu σ^2 není znám, proto ho musíme nahradit bodovým odhadem a použijeme statistiku

$$T = \frac{\bar{X} - \mu}{S}\sqrt{n},$$

která má Studentovo rozdělení $t(n-1)$. Interval spolehlivosti je ve tvaru

$$\mu = (\bar{x} - d,\ \bar{x} + d),\ d = \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}},$$

kde $t_{1-\frac{\alpha}{2}}$ jsou kvantily rozdělení $t(n-1)$.

Řešený příklad

Opakovaná měření stejné náhodné veličiny dala následující výsledky

17	37	70	34	64	52	72	21	47	32	94	62	71	34	43

Za předpokladu, že chyby jednotlivých měření mají normální rozdělení a jsou nezávislé, odhadněte, jakou hodnotu překročí střední hodnota s pravděpodobností 5 %?

Určíme potřebné výběrové charakteristiky a hodnotu kvantilu rozdělení $t(n-1)$ pro jednostranný test:

$$n = 15,\ \bar{x} = 50,\ S = 21.7,\ t_{1-\alpha}(n-1) = 1.76$$

Odhad hodnoty je

$$h \geq 50 + \frac{21.7}{\sqrt{15}} \cdot 1.76 \doteq 59.87$$

```
> mereni <- c(17,37,70,34,64,52,72,21,47,32,94,62,71,34,43)
> m <- mean(mereni)
> s <- sd(mereni)
> n <- length(mereni)
> alpha = 0.05
> q <- qt(1-alpha, df=n-1)          %kvantil jednostranný odhad
```

```
> h <- m + s/sqrt(n)*q                    %výpočet horní hranice odhadu
> h
[1] 59.87262
```

K určení intervalu spolehlivosti pro parametr σ^2 se využívá statistiky

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2},$$

které má Pearsonovo rozdělení $\chi^2(n-1)$. Pak platí

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}\right) = 1-\alpha.$$

Řešený příklad

Opakovaná měření koncentrace stejné látky dala následující hodnoty

0.2	0.23	0.21	0.16	0.18	0.19	0.14	0.18	0.21

Najděte 90% odhad pro rozptyl a směrodatnou odchylku.

Určíme potřebné výběrové charakteristiky a hodnoty kvantilů rozdělení $\chi^2(n-1)$ pro oboustranný test:

$$n = 9,\ \bar{x} = 0.189,\ S^2 = 7.6\cdot 10^{-4},\ \chi^2_{1-\frac{\alpha}{2}} = 15.51\ \chi^2_{\frac{\alpha}{2}} = 2.73$$

Intervalový odhad pro rozptyl je

$$\sigma^2 \in \left(\frac{8\cdot 7.6\cdot 10^{-4}}{15.51}, \frac{8\cdot 7.6\cdot 10^{-4}}{2.73}\right) \doteq (3.9\cdot 10^{-4}, 2.2\cdot 10^{-3}).$$

Intervalový odhad pro směrodatnou odchylku je

$$\sigma \in \left(\sqrt{3.9\cdot 10^{-4}}, \sqrt{2.2\cdot 10^{-3}}\right) \doteq (1.98\cdot 10^{-2}, 4.7\cdot 10^{-2}).$$

```
> k <- c(0.20,0.23,0.21,0.16,0.18,0.19,0.14,0.18,0.21)
> m <- mean(k)
> r <- var(k)
> n <- length(k)
> alpha = 0.1
> c1 <- qchisq(1-alpha/2, df=n-1)    %kvantil pro dolní mez
> c2 <- qchisq(alpha/2, df=n-1)      %kvantil pro horní mez
> h1 <- (n-1)*r/c1    %výpočet dolní hranice odhadu pro rozptyl
> h1
[1] 0.0003926463
> h2 <- (n-1)*r/c2    %výpočet horní hranice odhadu pro rozptyl
> h2
[1] 0.00222821
```

```
> hs1 <- sqrt(h1)     %výpočet dolní hranice odhadu pro směrodatnou
    odchylku
> hs1
[1] 0.0198153
> hs2 <- sqrt(h2)     %výpočet horní hranice odhadu pro směrodatnou
    odchylku
> hs2
[1] 0.04720392
```

4.2 Asymptotické intervaly spolehlivosti

Mějme náhodný výběr $X_1, \ldots, X_n$ z libovolného rozdělení s neznámými parametry střední hodnoty a rozptyl. Nechť rozsah souboru je velký - $n > 30$. Pak lze pro odhad intervalu spolehlivosti pro parametr střední hodnoty použít statistiku

$$U = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n},$$

která má podle centrální limitní věty rozdělení $N(0,1)$. Tedy asymptotický intervalový odhad je ve tvaru

$$P\left(\bar{x} - u_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \bar{x} + u_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Potřebný rozsah souboru s požadovanou chybou odhadu Δ určíme podle vzorce

$$n \geq \left(\frac{S}{\Delta} u_{1-\frac{\alpha}{2}}\right)^2.$$

Intervalový odhad parametru binomického rozdělení π lze s využitím výběrového poměru P vyjádřit ve tvaru

$$P\left(P - u_{1-\frac{\alpha}{2}} \frac{\sqrt{P(1-P)}}{\sqrt{n}} < \pi < P + u_{1-\frac{\alpha}{2}} \frac{\sqrt{P(1-P)}}{\sqrt{n}}\right) = 1 - \alpha.$$

Potřebný rozsah souboru s požadovanou chybou odhadu Δ určíme podle vzorce

$$n \geq u^2_{1-\frac{\alpha}{2}} \frac{P(1-P)}{\Delta^2}.$$

Řešený příklad

Předpokládejme, že v náhodném výběru 500 žen ve věku 20-45 let má 80 z nich vyšší hladinu cukru. Určete 95% interval spolehlivosti pro procento žen ve věku 20-45 let s vyšší hladinou cukru.

Určíme výběrový poměr a hodnotu kvantilu rozdělení $N(0,1)$:

$$P = 0.16,\ u_{1-\frac{\alpha}{2}} = 1.96$$

Odhad hodnoty π je

$$\pi \in (0.128,\ 0.192).$$

Poměr žen ve společnosti ve věku 20-45 let s vyšší hladinou cukru je s 95% spolehlivostí mezi 12.8 % a 19.2 %.

```
> n = 500
> P = 80/500
> P
[1] 0.16
> alpha = 0.05
> q = qnorm(1-alpha/2)
> d = q*sqrt(P*(1-P))/sqrt(n)
> d
[1] 0.03213385
> m1 = P-d
> m1
[1] 0.1278662
> m2 = P+d
> m2
[1] 0.1921338
```

Příklad

a) Zajímá nás, zda změnou dodavatele nedošlo i ke změně kvality našich výrobků. Zatímco dříve bylo mezi našimi výrobky v průměru 5 % zmetků, zjistila výstupní kontrola mezi 250 nově vyrobenými výrobky 16 nevyhovujících. Na základě 95% intervalu spolehlivosti rozhodněte, zda došlo ke změně kvality výrobků.
[$0.034 < \pi < 0.094$, změna neměla vliv]

b) Soubor (70,84,89,70,74,70) je náhodným výběrem z normálního rozdělení $N(\mu, \sigma^2)$. Určete 95% interval spolehlivosti pro rozptyl σ^2.
[$\sigma^2 \in (26.64,\ 411.296)$]

5 Testování hypotéz

Testování statistických hypotéz patří mezi základní metody statistické indukce a mezi metody kvantitativní teorie rozhodování. Umožňuje posoudit, zda data získané experimentem nepopírají předpoklad, který jsme učinili před provedením experimentu.

Statistická hypotéza je tvrzení o parametrech (rozdělení) základního souboru. V případě tvrzení týkající se parametru se jedná o parametrickou statistickou hypotézu (srovnávací testy, ANOVA, test o střední hodnotě, ...). Pokud se předpoklad týká jiné vlastnosti rozdělení, hovoříme o neparametrické statistické hypotéze (o typu rozdělení, o závislosti, ...).

Test statistické hypotézy je proces, kterým ověřujeme, zda lze statistickou hypotézu pokládat za správnou. Vstupem do testu jsou dvě hypotézy - nulová H_0 a alternativní H_1 (tvrzení popírající nulovou hypotézu). Pravdivost nulové hypotézy nelze na základě dat dokázat, lze ji na základě dat vyvrátit.

Postup - klasický přístup:

- Formulace hypotéz
- Výběr testového kritéria + ověření předpokladů testu
- Volba hladiny významnosti
- Určení hodnotu testového kritéria x_{obs} a kritické hodnoty (kritický obor)
- Rozhodnutí testu na základě vztahu hodnoty testového kritéria a kritické hodnoty

Nevýhodnou klasického přístupu je, že není na první pohled zřejmé, jak rozhodnutí závisí na změně hladiny významnosti. Pro rozhodování o výsledku testu se používá p**-hodnota**, což je nejvyšší hladina významnosti, na které již nelze nulovou hypotézu zamítnout. Jedná se o postup, který nazýváme čistý test významnosti.

Výpočet p-hodnoty:

tvar H_1	$p-$hodnota
$\theta < \theta_0$	$p-$hodnota = $F(x_{obs})$
$\theta > \theta_0$	$p-$hodnota = $1 - F(x_{obs})$
$\theta \neq \theta_0$	$p-$hodnota = $2 \cdot min\{F(x_{obs}, 1 - F(x_{obs})\}$

Rozhodnutí na základě p-hodnoty:

$p-$hodnota	rozhodnutí
$p-$hodnota $< \alpha$	zamítáme H_0 ve prospěch H_1
$p-$hodnota $\geq \alpha$	nezamítáme H_0

Rozhodování se řídí podle následujících pravidel:

rozhodnutí/skutečnost	H_0 platí	H_0 neplatí
nezamítáme H_0	správné rozhodnutí pravděpodobnost $1 - \alpha$	chyba II. druhu pravděpodobnost β
zamítáme H_0	chyba I. druhu pravděpodobnost α	správné rozhodnutí ravděpodobnost $1 - \beta$

Chyba I. druhu - nesprávné zamítnutí nulové hypotézy (pravděpodobnost chyby - hladina významnosti).

Chyba II. druhu - nesprávné nezamítnutí nulové hypotézy (pravděpodobnost, že se nedopustíme chyby nazýváme síla testu).

S klesající hladinou významnosti roste pravděpodobnost chyby II. druhu. Zvýšením rozsahu souboru snížíme obě pravděpodobnosti chyb, tzn. sílu testu ovlivníme volbou testové statistiky a dostatečným počtem pozorování.

5.1 Jednovýběrové testy

5.1.1 Testy o parametrech normálního rozdělení

Ekvivalentem intervalů spolehlivosti je jednovýběrový test, a proto stejně jako u intervalových odhadů máme tři testy o parametrech normálního rozdělení.

- **Jednovýběrový z-test**

  $$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0 \quad (\text{případně } \mu < \mu_0,\ \mu > \mu_0)$$

  Předpoklad: $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$, kde hodnota σ^2 je známa

  Testové kritérium: $U = \frac{\bar{x}-\mu_0}{\sigma}\sqrt{n} \sim N(0,1)$

  Kritická hodnota: $u_{1-\frac{\alpha}{2}}$ u oboustranného testu a $u_{1-\alpha}$ u jednostranného testu

- **Jednovýběrový t-test**

  $$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0 \quad (\text{případně } \mu < \mu_0,\ \mu > \mu_0)$$

  Předpoklad: $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$, kde hodnota σ^2 není známa

  Testové kritérium: $T = \frac{\bar{x}-\mu_0}{s}\sqrt{n} \sim t_{n-1}$

  Kritická hodnota: $t_{1-\frac{\alpha}{2}}(n-1)$ u oboustranného testu a $t_{1-\alpha}(n-1)$ u jednostranného testu

  V případě velkého rozsahu souboru $(n > 30)$ nemusí být splněn předpoklad o normalitě výběru a lze použít testové kritérium

  $$U = \frac{\bar{x}-\mu_0}{s}\sqrt{n},$$

  jehož normalita je zaručena z centrální limitní věty a pro rozhodnutí používáme kvantily rozdělení $N(0,1)$.

  R COMMANDER - menu `Statistics-Means-Single-sample t-test` nebo funkce `t.test()`.

```
> x <- rnorm(15, mean=35)
> x
 [1] 35.94176 35.36492 35.37699 35.32756 33.78589 34.61224
   33.86090 34.74667
 [9] 36.33886 33.54628 35.53323 34.33238 35.69775 35.54201
   36.45806
> m <- mean(x)
> m
[1] 35.0977
> n
[1] 15
> s <- sd(x)
> s
[1] 0.9124104
> T <-(m - 35)*sqrt(n)/s % testujeme H0: střední hodnota = 35
> T                      % výpočet testové statistiky
[1] 0.4147159
> qt(p=0.975,df=(n-1))           % výpočet kritické hodnoty
[1] 2.144787                   % |T| < krit => H0 nelze zamítnout
>t.test(x, mu=35, conf.level=0.95)    % funkce t.test

      One Sample t-test

data:  x                       % vyhodnocení pomocí p-hodnoty
t = 0.41472, df = 14, p-value = 0.6846
alternative hypothesis: true mean is not equal to 35
95 percent confidence interval:
 34.59242 35.60298
sample estimates:
mean of x
  35.0977
```

- **Jednovýběrový test o rozptylu**

$$H_0 : \sigma^2 = \sigma_0^2, \qquad H_1 : \sigma^2 \neq \sigma_0^2 \quad (\text{případně } \sigma^2 < \sigma_0^2,\ \sigma^2 > \sigma_0^2)$$

Předpoklad: $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$, kde oba parametry jsou neznámé

Testové kritérium: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1)$

Kritické hodnoty: $\chi^2_{1-\frac{\alpha}{2}}(n-1)$, $\chi^2_{\frac{\alpha}{2}}(n-1)$ u oboustranného testu a $\chi^2_{1-\alpha}(n-1)$, $\chi^2_{\alpha}(n-1)$ u jednostranného testu

Využití funkce `varTest()` v balíčku "EnvStats"

```
> install.packages("EnvStats")
> library(Envstats)
```

Řešený příklad

Automat vyrábí pístové kroužky o daném průměru. Výrobce udává, že směrodatná odchylka průměru kroužku je 0.1 mm. K ověření této informace bylo náhodně vybráno 60 kroužků a vypočtena směrodatná odchylka jejich průměru 0.08 mm. Lze tento rozdíl považovat za statisticky významný ve smyslu zlepšení kvality produkce? Předpokládejte, že průměr pístových kroužků má normální rozdělení.

$H_0 : \sigma = 0.1,\ H_1 : \sigma < 0.1$

```
> n = 60
> s = 0.08
> sigma = 0.1
> alpha = 0.05       %hladina významnosti
> C <- (n-1)*s^2/sigma^2    %vypocet testovaci hodnoty
> C
[1] 37.76
> q <- qchisq(alpha,n-1)    %vypocet kriticke hodnoty
> q
[1] 42.33931
> p.hodnota = pchisq(C,n-1)
> p.hodnota
[1] 0.01416051
```

p-hodnota $\doteq 0.01 \Rightarrow H_0$ lze zamítnout ve prospěch alternativní hypotézy

2. způsob - klasický přístup: $C = 37.76 < 42.34 = q \Rightarrow H_0$ lze zamítnout na hladině významnosti 0.05

Závěr: Směrodatná odchylka průměru kroužku je statisticky významně menší než 0.1 mm.

5.1.2 Test o parametru binomického rozdělení

$$H_0 : \pi = \pi_0, \qquad H_1 : \pi \neq \pi_0 \quad (\text{případně } \pi < \pi_0,\ \pi > \pi_0)$$

Předpoklad: $x_1, \ldots, x_n$ je výběr z alternativního rozdělení, kde rozsah souboru musí splňovat, že $n > 30,\ n > \frac{9}{P(1-P)}$, kde P je relativní četnost.

Testové kritérium: $U = \frac{P-\pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n}$, které má podle Moivreovy-Laplaceovy věty přibližně rozdělení $N(0,1)$.

Kritická hodnota: $u_{1-\frac{\alpha}{2}}$ u oboustranného testu a $u_{1-\alpha}$ u jednostranného testu

R COMMANDER - menu `Statistics-Proportions-Single-sample proportion test` nebo funkce `binom.test()`.

Řešený příklad

Firma udává, že 1 % jejich výrobů nesplňuje požadovaná kritéria. V testované dodávce 2 500 ks bylo nalezeno 28 nevyhovujících výrobků. Potvrzuje tento výsledek tvrzení firmy?

$H_0 : \pi = 0.01,\ H_1 : \pi \neq 0.01$

```
> n = 2500
> x = 28
> P = x/n          %relativni cetnost
> P
[1] 0.0112
> alpha = 0.05
> binom.test(x,n,0.01,conf.level=0.95)

    Exact binomial test

data:  x and n
number of successes = 28, number of trials = 2500, p-value =
   0.5452
alternative hypothesis: true probability of success is not equal
   to 0.01
95 percent confidence interval:
 0.007454864 0.016146715
sample estimates:
probability of success
                0.0112
> U <- (P-0.01)/sqrt(0.01*(1-0.01))*sqrt(n)    %vypocet testovaci
   hodnoty
> U
[1] 0.6030227
> q<-qnorm(1-alpha/2)
> q                          %vypocet kriticke hodnoty
[1] 1.959964
```

p-hodnota $\doteq 0.55 \Rightarrow H_0$ nelze zamítnout

2. způsob - klasický přístup: $U = 0.6 < 1.96 = q \Rightarrow H_0$ nelze zamítnout

Závěr: Výsledek potvrzuje tvrzení firmy.

5.1.3 Shapiro-Wilkův test, grafické ověření normality

Pro použití jednovýběrových testů o parametrech normálného rozdělení je potřeba ověřit předpoklad testu, a to normalitu dat (při malém počtu měření). Jedním z testů ověřující normalitu je Shapiro-Wilkův test `shapiro.test(x)`, který je jeden z nejsilnějších testů normality:

H_0 : X podléhá normálnímu rozdělení

Testové kritérium: $W = \frac{s^2_{norm}}{s^2}$

Kritické hodnoty jsou tabelovány.

Posouzení, zda data splňují normalitu, lze udělat na základě grafické analýzy a hodnot číselných charakteristik. Vyhodnocení krabicového grafu (symetrie, malé množství odlehlých hodnot), velmi nízká hustota bodů daleko od střední hodnoty a vysoká v její blízkosti (`plot(sort(x))`),k porovnání teoretického a skutečného profilu slouží Q-Q graf (`qqnorm(x), qqline(x)`), porovnání histogramu s křivkou normálního rozdělení.

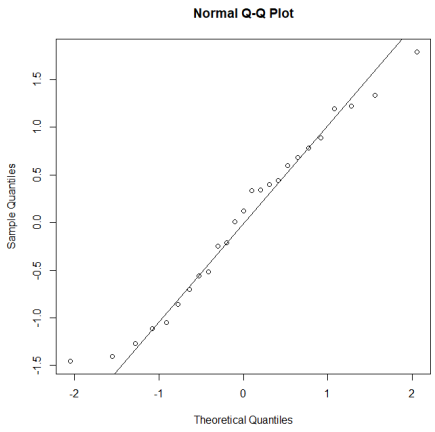
Grafické ověření i testy lze určit v R COMMANDERU
- menu `Statistics-Summaries-Test of normality: Shapiro-Wilks`,
- menu `Graphs-Boxplot`, `Graphs-Quantile-comparison plot`.

V případě, že data nepodléhají normalitě a tedy nelze použít t-test, využijeme neparametrický test střední hodnoty - mediánový test, Wilcoxonův test.

Řešený příklad

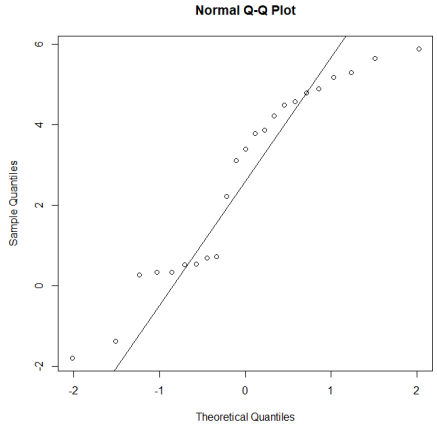
Vygenerujte náhodná data s normálním rozdělením a data, která nemají normální rozdělení a na základě Q-Q grafu a S-W testu ověřte normalitu dat.

```
> x <- rnorm(25)
> qqnorm(x)
> qqline(x)
> shapiro.test(x)

    Shapiro-Wilk normality test

data:  x
W = 0.96658, p-value = 0.5601
```



Vidíme, že body leží kolem přímky a p-hodnota je větší než 0.05, tzn. že nezamítáme nulovou hypotézu a data pochází z normálního rozdělení.

```
> x <- c(rnorm(10),rnorm(13, mean = 4))
> qqnorm(x)
> qqline(x)
> shapiro.test(x)

    Shapiro-Wilk normality test

data:  x
W = 0.90736, p-value = 0.03595
```



Body jsou více vzdálené od přímky a p-hodnota je menší než 0.05, tzn. že na hladině významnosti 0.05 lze nulovou hypotézu o normalitě zamítnout.

5.1.4 Wilcoxonův znaménkový test

$$H_0 : Me = m_0, \qquad H_1 : Me \neq m_0 \quad (\text{případně } Me < m_0,\ Me > m_0)$$

Počítáme pořadí od nejmenších po největší číslům $|x_i - m_0|$. R^+ a R^- označuje součet těchto pořadí pro kladné a záporné $x_i - m_0$, nulové hodnoty vynecháme a ke stejným hodnotám určíme průměrné pořadí.

Testové kritérium: $T = min(R^+, R^-)$.

Kritická oblast: $T \leq T_{\frac{\alpha}{2}}$ - kvantil jednovýběrové Wilcoxonovy statistiky T (funkce `qsignrank()`) u oboustranného testu, $R^- \leq T_\alpha$ u pravostranného testu a $R^+ \leq T_\alpha$ u levostranného testu

R COMMANDER - menu `Statistics-Nonparametric tests-Single-sample Wilcoxon test` nebo funkce `wilcox.test()`.

Řešený příklad

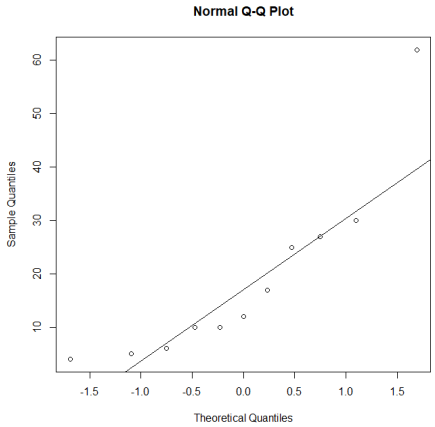
Opakovaná měření stejné náhodné veličiny dala následující výsledky

17	25	6	62	10	10	30	27	4	5	12

Z předchozích měření víme, že průměrný výsledek měření je 20. Lze na základě dat toto potvrdit?

Ověříme normalitu dat.

```
> x<- c(17,25,6,62,10,10,30,27,4,5,12)
> qqnorm(x)
> qqline(x)
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.80653, p-value = 0.01147  %na hl. významnosti 0.05 zamítáme
   předpoklad normality
```




```
> wilcox.test(x,mu=20,alternative="greater",conf.level=0.95,conf.
   int=TRUE) %použijeme neparametrický test

        Wilcoxon signed rank test with continuity correction

data:  x
V = 22, p-value = 0.8472      %Medián výsledku měření není
   statisticky významně větší než 20.
alternative hypothesis: true location is greater than 20
95 percent confidence interval:
 8.999968      Inf
```

5.1.5 Testy dobré shody

V praxi je velmi často potřeba ověřit, zda náš odhad ohledně rozdělení studovaného výběru je správný. Shodu mezi teoretickým a odhadovaným rozdělením ověřujeme testy dobré shody. Jedná se o neparametrické testy.

- **Chí-kvadrát test**
  Jedná se o nejznámější test a ověřuje, zda se pozorované četnosti jednotlivých variant shodují s očekávanými.

  Náhodý výběr rozdělíme do k tříd a určíme pozorované e_i a teoretické (očekávané) četnosti o_i.

  Testové kritérium: $\chi^2 = \sum_{i=1}^{k} \frac{(e_i - o_i)^2}{o_i} \sim \chi^2(k-r-1)$, kde r je počet odhadovaných parametrů.

  Důležité je, aby byl splněn předpoklad testu, že všechny očekávané četnosti byly větší než 5. V případě, že není předpoklad splněn, buď rozšíříme rozdah výběru, nebo dodatečně sloučíme varianty, které spolu souvisí.

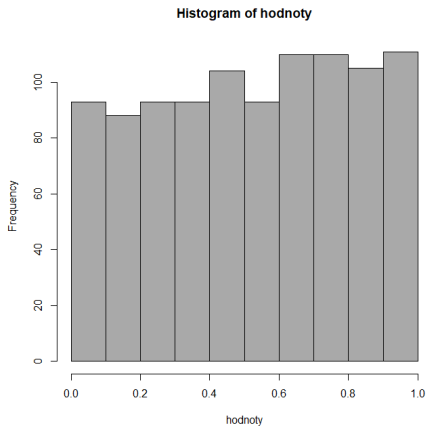
  Kritická hodnoty: $\chi^2_{1-\alpha}(k-r-1)$ - kvantily Chí-kvadrát rozdělení (funkce `qchisq()`)

  R COMMANDER - menu `Statistics-Summaries-Frequency Distributions` nebo funkce `chisq.test()`.

Řešený příklad

Ověřte, zda generátor náhodných čísel z rovnoměrného rozdělení na interval $\langle 0, 1 \rangle$ opravdu generuje výběr z daného rozdělení.

```
> hodnoty = runif(1000,0,1)
> hist(hodnoty, col="darkgray")
```




```
> breaks=seq(0,1,by=0.1)              % hranice trid
> breaks
 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> hodnoty.cut=cut(hodnoty,breaks,right=FALSE)
> hodnoty.freq=table(hodnoty.cut)
> hodnoty.freq
```

```
hodnoty.cut
  [0,0.1) [0.1,0.2) [0.2,0.3) [0.3,0.4) [0.4,0.5) [0.5,0.6)
   [0.6,0.7) [0.7,0.8)
       93         88         93         93        104         93
      110        110
[0.8,0.9)   [0.9,1)
      105        111
> cbind(hodnoty.freq)          %tabulka cetnosti
          hodnoty.freq
[0,0.1)             93
[0.1,0.2)           88
[0.2,0.3)           93
[0.3,0.4)           93
[0.4,0.5)          104
[0.5,0.6)           93
[0.6,0.7)          110
[0.7,0.8)          110
[0.8,0.9)          105
[0.9,1)            111
> pozcet=c(cbind(hodnoty.freq))
> pozcet
 [1]  93  88  93  93 104  93 110 110 105 111
 > p=rep(c(0.1), times=10)      % teoreticke cetnosti
> chisq.test(pozcet,p=p)

        Chi-squared test for given probabilities

data:  pozcet
X-squared = 7.02, df = 9, p-value = 0.635

> e=pozcet                     % vypocet dle vzorce
> o=p*sum(pozcet)
> sum((e-o)^2/o)               % testove kriterium
[1] 7.02
> qchisq(0.95,df=9)            % kriticka hodnota
[1] 16.91898
```

Kritická hodnota je větší než testová statistika (p-hodnota=0.63), nelze zamítnout H0, že generátor generuje čísla z rozdělení R(0;1).

- **Kolmogorovův-Smirnovův jednovýběrový test**
  Při ověřování dobré shody mezi empirickým a teoretickým rozdělením se spojitou distribuční funkcí dáváme tomuto testu přednost před chi-kvadrát testem v případech výběru malého rozsahu. V případě, že jsou data setříděna do tabulky rozdělení četností, představuje empirická distribuční funkce kumulované relativní četnosti.

  Testové kritérium je maximální odchylka teoretické a pozorované distribuční funkce: $D = \sup_x |F_n(x) - F_0(x)| = \max\{D_i\}$, kde $D_i = \max\left\{\left|F_0(x) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - F_0(x)\right|\right\}$

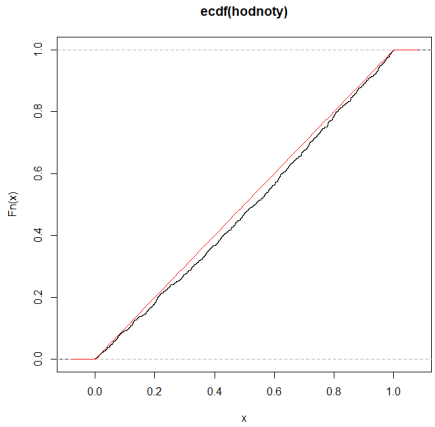
Kritická hodnoty: $d_{n;\alpha}$, při malém rozsahu jsou kvantily tabelovány, v případě velkého rozsahu lze hodnoty aproximovat podle vztahu: $d_{n;\alpha} \doteq \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}$

Funkce `ks.test()`.

Řešený příklad

Ověřte, zda generátor náhodných čísel z rovnoměrného rozdělení na interval $\langle 0, 1 \rangle$ opravdu generuje výběr z daného rozdělení.

```
> ks.test(hodnoty,"punif",min(hodnoty),max(hodnoty))

        One-sample Kolmogorov-Smirnov test

data:  hodnoty
D = 0.040931, p-value = 0.07012
alternative hypothesis: two-sided

> plot(ecdf(hodnoty))
> curve(punif(x,min(hodnoty),max(hodnoty)), add=TRUE,col="red
   ")
```



Na hladině významnosti 0.05 nezamítáme nulovou hypotézu, tj. nelze tvrdit, že generátor negeneruje čísla z rozdělení R(0;1).

5.2 Dvouvýběrové testy

5.2.1 Testy o shodě parametrů dvou normálních souborů

Předpokládejme, že máme dva nezávislé náhodné výběry, které pocházejí z normálních rozdělení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$.

- **Dvouvýběrový F-test o shodě rozptylů**

$$H_0 : \sigma_1^2 = \sigma_2^2, \qquad H_1 : \sigma_1^2 \neq \sigma_2^2 \quad (\text{případně } \sigma_1^2 < \sigma_2^2,\ \sigma_1^2 > \sigma_2^2)$$

Testové kritérium: $F = \frac{s_1^2}{s_2^2} \sim F(m-1, n-1)$

Kritické hodnoty: $F < F_{\frac{\alpha}{2}}$ nebo $F > F_{1-\frac{\alpha}{2}}$ - kvantily Fisherova rozdělení (funkce `qf()`) u oboustranného testu

R COMMANDER - menu `Statistics-Variances-Two-variances F-test` nebo funkce `var.test()`.

Řešený příklad

Vygenerujte dva soubory s normálním rozdělením a na základě F-testu rozhodněte, zda lze tvrdit, že mají shodné rozptyly.

```
> x <-rnorm(10, mean=12.3, sd=3.3)
> x
 [1] 10.909418 12.370540 15.681135 16.531512  8.989301
   11.837525 10.621985
 [8]  6.994720 14.395227 13.321544
> y <-rnorm(10, mean=8.5x, sd=3.3)
> y
 [1]  6.132603  9.664394  9.485209 12.287006  9.361378
   9.804891  7.964152
 [8] 13.410421 15.054612 12.622732
> r1=var(x)
> r1
[1] 8.74054
> r2=var(y)
> r2
[1] 7.318308
> F = r2/r1        %vypocet testove hodnoty
> F
[1] 0.8372833
> qf(0.025,9,9)          %vypocet kritickych hodnot
[1] 0.2483859
> qf(0.975,9,9)
[1] 4.025994         %F nelezi v kritickem oboru, nezamitame H0
> var.test(y,x)

        F test to compare two variances

data:  y and x
F = 0.83728, num df = 9, denom df = 9, p-value = 0.7957  %na
   zaklade p-hodnoty nelze H0 zamitnout, rozptyly lze
   povazovat za shodne
```

```
alternative hypothesis: true ratio of variances is not equal
   to 1
95 percent confidence interval:
 0.2079693 3.3708977
sample estimates:
ratio of variances
         0.8372833
```

- **Dvouvýběrový t-test o shodě středních hodnot, pokud** $\sigma_1^2 = \sigma_2^2$

$$H_0 : \mu_1 = \mu_2, \qquad H_1 : \mu_1 \neq \mu_2 \quad (\text{případně } \mu_1 < \mu_2,\ \mu_1 > \mu_2)$$

Testové kritérium: $T = \frac{(\bar{x}-\bar{y})}{S}\sqrt{\frac{mn}{m+n}}$, kde $S = \sqrt{\frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}} \sim t(m+n-2)$

Kritické hodnoty: kvantily t-rozdělení

R COMMANDER - menu `Statistics-Means-Independent samples t-test` nebo funkce `t.test(x,y,var.equal=TRUE)`.

Řešený příklad

Na datech z předchozího příkladu ověříme, zda lze tvrdit, že soubory mají shodné střední hodnoty.

```
>  t.test(y,x,var.equal=TRUE)

        Two Sample t-test

data:  y and x
t = -1.252, df = 18, p-value = 0.2266   %na zaklade p-hodnoty
   nelze H0 zamitnout, soubory lze povazovat za shodne
alternative hypothesis: true difference in means is not equal
   to 0
95 percent confidence interval:
 -4.248913  1.075811
sample estimates:
mean of x mean of y
 10.57874  12.16529
```

- **Dvouvýběrový (Aspinové-Welchův) test o shodě středních hodnot, pokud** $\sigma_1^2 \neq \sigma_2^2$

$$H_0 : \mu_1 = \mu_2, \qquad H_1 : \mu_1 \neq \mu_2 \quad (\text{případně } \mu_1 < \mu_2,\ \mu_1 > \mu_2)$$

Testové kritérium: $T = \frac{(\bar{x}-\bar{y})}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}} \sim t(\nu)$

Kritické hodnoty: kvantily t-rozdělení s ν stupni volnosti, kde $\nu \cong \frac{\left(\frac{s_1^2}{m}+\frac{s_2^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_1^2}{m}\right)^2+\frac{1}{n-1}\left(\frac{s_2^2}{n}\right)^2}$

R COMMANDER - menu `Statistics-Means-Independent samples t-test` nebo funkce `t.test(x,y)`.

Řešený příklad

Oštěpařky Bára a Anežka provedly po řadě 9 a 7 hodů. Výsledky v metrech jsou zazna- menány v tabulce. Na hladině významnosti 5 % rozhodněte, zda Bářin výkon je srovnatelný jako Aneščin.

Bára	45	52	48	60	55	51	59	47	50
Anežka	60	69	75	49	53	50	37		

Ověříme normalitu dat.

```
> B <- c(45,52,48,60,55,51,59,47,50)
> A <- c(60,69,75,49,53,50,37)
> shapiro.test(A)

        Shapiro-Wilk normality test

data:  A
W = 0.96947, p-value = 0.8946
> shapiro.test(B)

        Shapiro-Wilk normality test

data:  B
W = 0.94266, p-value = 0.6102
```

U obou testů vyšla p-hodnota větší jak 0.05, tudíž oba soubory jsou normální.

Provedeme F-test.

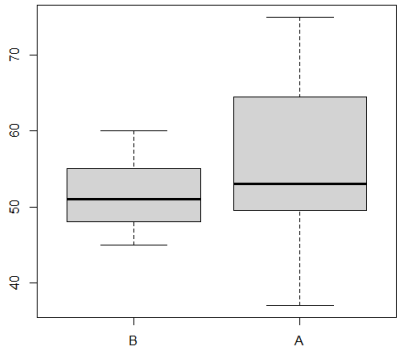
```
> var.test(B,A)

        F test to compare two variances

data:  B and A
F = 0.16253, num df = 8, denom df = 6, p-value = 0.02222
alternative hypothesis: true ratio of variances is not equal
   to 1
95 percent confidence interval:
 0.0290247 0.7560278
sample estimates:
```

```
ratio of variances
         0.1625274
```

Nulovou hypotézou je, že poměr rozptylů je rovný nule, tedy že oba rozptyly jsou stejné. Na základě p-hodnoty 0.022 zamítáme nulovou hypotézu a použijeme Welchův test.

```
> t.test(B,A)

        Welch Two Sample t-test

data:  B and A
t = -0.82108, df = 7.5226, p-value = 0.4368
alternative hypothesis: true difference in means is not equal
   to 0
95 percent confidence interval:
 -16.334526   7.826589
sample estimates:
mean of x mean of y
 51.88889  56.14286
```

Hodnota p-value je 0.4368, tedy nezamítáme nulovou hypotézu a lze tvrdit, že obě oštěpařky mají srovnatelné průměrné výsledky.



Párový test

Jsou-li oba výběry normální závislé s $m = n$ (párová měření), počítáme $D_i = x_i - y_i$. Test o shodě dvou středních hodnot prováděný na základě dvou závislých výběrů můžeme převést na jednovýběrový test o střední hodnotě aplikovaném na tyto rozdíly – jde o párový dvouvýběrový t-test.

R COMMANDER - menu `Statistics-Means-Paired t-test` nebo funkce `t.test(x,y,paired=TRUE)`.

5.2.2 Neparametrické testy

Wilcoxonův párový test

Jedná se o test o shodě úrovní, kde místo středních hodnot porovnáváme mediány a výběry jsou závislé - párová měření.

$$H_0 : Me(X) = Me(Y), \qquad H_1 : Me(X) \neq Me(Y)$$

Počítáme pořadí od nejmenších po největší číslům $|x_i - y_i|$. T^+ a T^- označuje součet těchto pořadí pro kladné a záporné $x_i - y_i$, nulové hodnoty vynecháme a ke stejným hodnotám určíme průměrné pořadí.

Testové kritérium: $T = min(T^+, T^-)$.

Kritická oblast: $T \leq T_{\frac{\alpha}{2}}$ - kvantil jednovýběrové Wilcoxonovy statistiky T (`qsignrank()`) u oboustranného testu, $T^- \leq T_\alpha$ u pravostranného testu a $T^+ \leq T_\alpha$ u levostranného testu

R COMMANDER - menu `Statistics-Nonparametric tests-Paired-samples Wilcoxon test` nebo funkce `wilcox.test(x,y,paired=TRUE, alternative="two.sided")`.

Řešený příklad

Změřili jsme váhu u 10 myší před a po změně stravování. Bude nás zajímat, zda měla strava vliv na váhu.

Před	200.1	190.9	172.7	195.5	241.4	176.9	172.2	175.5	205.2	183.7
Po	392.9	393.2	345.1	393	434	427.9	422	383.9	392.3	352.2

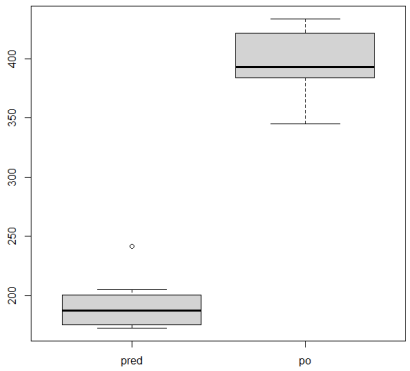
Ověříme normalitu dat.

```
> pred<-  c(200.1, 190.9, 172.7, 195.5, 241.4, 176.9, 172.2,
   175.5, 205.2, 183.7)
> po <-c(392.9, 393.2, 345.1, 393, 434, 427.9, 422, 383.9, 392.3,
   352.2)
> data <- data.frame(stav=rep(c("pred", "po"),each = 10),vaha=c(
   pred,po))
> print(my_data)
   stav  vaha
1  pred 200.1
2  pred 190.9
3  pred 172.7
4  pred 195.5
5  pred 241.4
6  pred 176.9
7  pred 172.2
8  pred 175.5
9  pred 205.2
10 pred 183.7
11   po 392.9
```

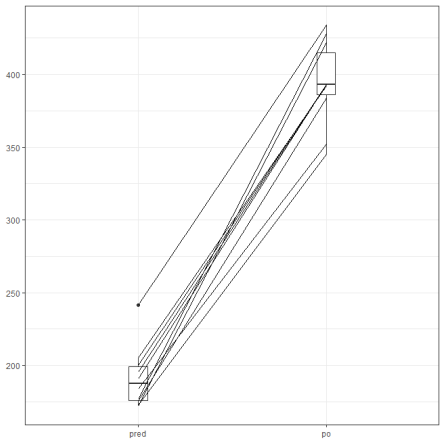
```
12   po 393.2
13   po 345.1
14   po 393.0
15   po 434.0
16   po 427.9
17   po 422.0
18   po 383.9
19   po 392.3
20   po 352.2
>  shapiro.test(pred)

        Shapiro-Wilk normality test

data:  pred
W = 0.84289, p-value = 0.04778
```

U prvního souboru zamítáme normalitu a použijeme tedy Wilcoxonův párový test. Znázorníme data i jejich páry.

```
> boxplot(pred,po,names = c("pred", "po"))
```




```
> install.packages("PairedData")
> pred <- subset(my_data,stav == "pred", vaha, drop=TRUE)
> po <- subset(my_data,stav == "po", vaha, drop=TRUE)
> library(PairedData)
> pd <- paired(pred, po)
> plot(pd, type = "profile") + theme_bw()
```




```
> res <- wilcox.test(pred,po, paired = TRUE)
> res

        Wilcoxon signed rank exact test

data:  pred and po
V = 0, p-value = 0.001953
alternative hypothesis: true location shift is not equal to 0
```

Hodnota p-value je 0.001953, což je menší než hladina významnosti , proto zamítáme nulovou hypotézu a lze tvrdit, že medián váhy myši před testem je významně odlišná od mediánu váhy po testu.

Přesněji lze použít jednostranný test, tzn. alternativní hypotéza bude ve tvaru: median(pred) < media(po).

```
> res <- wilcox.test(pred,po, paired = TRUE,alternative = "less")
> res

        Wilcoxon signed rank exact test

data:  pred and po
V = 0, p-value = 0.0009766
alternative hypothesis: true location shift is less than 0
```

Mannův-Whitneyův test

Jedná se o test o shodě úrovní, kde místo středních hodnot porovnáváme mediány a výběry jsou nezávislé ze spojitých rozdělení se stejným roztylem a tvarem. značení výběrů se volí tak, aby platilo $n_1 \geq n_2$.

$$H_0 : Me(X) = Me(Y), \qquad H_1 : Me(X) \neq Me(Y)$$

Smícháme výběry a k jednotlivým měřením určíme pořadí. Určíme součet pořadí u obou výběrů R_1, R_2.

Výpočet testových statistik: $T_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$ a $T_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$, platí $T_1 + T_2 = n_1 n_2$.
Testové kritérium: $T = min(T_1, T_2)$.

Kritické hodnoty Mannova-Whitneyova testu jsou tabelovány. Pokud je pozorovaná hodnota testového kritéria menší nebo rovna příslušné kritické hodnotě, nulová hypotéza se zamítá.

Lze použít modifikaci testu - **dvouvýběrový Wilcoxonův test.**
Jeho testová statistika je

$$W = R_1 - \frac{n(n+1)}{2}$$

a má W rozdělení, jehož kvantily jsou tabelované (funkce `qwilcox(p,n1,n2)`).

R COMMANDER - menu `Statistics-Nonparametric tests-Two-sample Wilcoxon test` nebo funkce `wilcox.test()`.

Řešený příklad

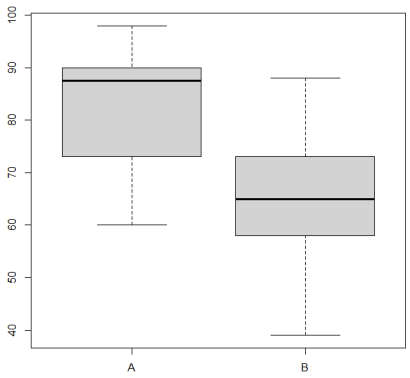
Zajímá nás, který z výukových programů je efektivnější. Náhodným výběrem jsme vybrali výsledky závěrečných testů 14 studentů, kteří absolvovali program společnosti A a 15 studentů, kteří absolvovali program společnosti B. Závěrečný test byl v obou skupinách stejný a jeho výsledky jsou následující:

A	85	87	92	98	90	88	75	72	60	93	88	89	62	73	
B	65	57	74	43	39	88	62	69	70	72	59	60	80	83	50

```
> A <- c(85,87,92,98,90,88,75,72,60,93,88,89,62,73)
> B <- c(65,57,74,43,39,88,62,69,70,72,59,60,80,83,50)
```

Jedná se o nezávislé náhodné výběry, proto použijeme MWT.

```
> boxplot(A,B,names = c("A", "B"))
```



```
> wilcox.test(A,B,alternative="greater", exact = FALSE)

  Wilcoxon rank sum test with continuity correction

data:  A and B
W = 177.5, p-value = 0.0008303
alternative hypothesis: true location shift is greater than 0
> qwilcox(0.95,15,14)
[1] 143
```

Hodnota testového kritéria překročila kritickou hodnotu, proto H0 zamítáme a lze tvrdit, že výukové programy nejsou srovantelné a program od společnosti A dává lepší výsledky (p-hodnota = 0.0008).

Dvouvýběrový test o populačních poměrech

$$H_0 : \pi_1 = \pi_2, \qquad H_1 : \pi_1 \neq \pi_2 \quad (\text{případně } \pi_1 < \pi_2,\ \pi_1 > \pi_2)$$

Předpoklad: X, Y jsou náhodné výběry, které pocházejí z alternativních rozdělení. Výběry musí být dostatečného rozsahu, tzn. $n_1 > \frac{9}{p_1(1-p_1)}$, $n_2 > \frac{9}{p_2(1-p_2)}$, kde p_1, p_2 jsou odpovídající výběrové poměry.

Testové kritérium: $U = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$, které má rozdělení $N(0,1)$.

Kritická hodnota: $u_{1-\frac{\alpha}{2}}$ u oboustranného testu a $u_{1-\alpha}$ u jednostranného testu

V praxi se častěji využívá Pearsonův chí-kvadrát test.

R COMMANDER - menu `Statistics-Proportions-Two-sample proportions test` nebo funkce `prop.test()`.

Řešený příklad

Testovali jsme výrobky od dvou výrobců (A,B). Firma A tvrdí, že její výrobky jsou spolehlivější než výrobky od firmy B. Pro ověření tohoto tvrzení jsme provedli průzkum a zjistili, že z 300 prodaných výrobků od firmy A bylo reklamováno 10 výroků a z 440 prodaných výrobků firmy B bylo reklamováno výrobků 18. Má firma A pravdu?

Hypotézy: $H0 : \pi_A = \pi_B$, $H1 : \pi_A < \pi_B$,

```
> prop.test(x=c(10,18), n=c(300,440), alternative = "less")

        2-sample test for equality of proportions with continuity
   correction

data:  c(10, 18) out of c(300, 440)
X-squared = 0.11161, df = 1, p-value = 0.3692
alternative hypothesis: less
95 percent confidence interval:
```

```
 -1.00000000  0.01828922
sample estimates:
    prop 1     prop 2      %vyberove pomery
0.03333333 0.04090909
```

Na hladině významnosti 0.05 nezamítáme nulovou hypotézu (p-hodnota > 0.05), tvrzení firmy A nelze považovat za pravdivé.

Kolmogorovův Smirnovův test pro dva výběry

Jedná se o test shody rozdělení u dvou výběrů.

Princip je obdobný jako u jednovýběrového K-S testu. Seřadíme všechny hodnoty do neklesající posloupnosti a určíme kumulativní relativní četnosti jednotlivých výběrů $F(x), G(x)$.

Testové kritérium: $D_{m,n} = \max|F(x) - G(x)|$

Kritická hodnota: $d_{1-\alpha}$, kvantily Kolmogorova rozdělení a jsou tabelovány

Řešený příklad

Vybrali jsme 13 polí stejné kvality a na 5 z nich jsme použili nový způsob hnojení. Výnosy v tunách na hektar jsou uvedeny v tabulce:

Nový	5.0	4.5	4.2	5.4	4.4			
Běžný	5.7	5.5	4.3	5.9	5.2	5.6	5.8	5.1

Pochází oba výběry ze stejného rozdělení?

```
> x = c(5,4.5,4.2,5.4,4.4)
> y = c(5.7,5.5,4.3,5.9,5.2,5.6,5.8,5.1)
> ks.test(x,y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.675, p-value = 0.07925
alternative hypothesis: two-sided
```

Na hladině významnosti 0.05 nezamítáme nulovou hypotézu (p-hodnota > 0.05), oba výběry pocházejí ze základních souborů se stejným rozdělením.

6 Analýza rozptylu, ANOVA

Pro porovnání dvou průměrů používáme t-test, ale v případě, že chceme porovnávat průměry více souborů, nelze použít t-test pro porovnání "každý s každým". Tento postup je ale z důvodu mnohonásobného porovnávání nesprávý. Je to proto, že jednotlivé testy jsou nezávislé a proto je obtížné pravděpodobnost chyby prvního druhu alespoň v jednm testu odhadnout. Správným postupem je provést analýzu rozptylu.

Prvním krokem je provedení explorační analýzy - vizualizace (boxplot, bodový graf), základní charakteristiky. Krabicový graf využijeme i k identifikaci odlehlých hodnot, které mohou způsobit selhání analýzy rozptylu. V případě, že se odlehlé hodnoty vyskytly z důvodu hrubých chyb, překlepů, poruch, atd. lze je z dalšího zpracování vyloučit. Jinak musíme použít neparametrický Kruskalův-Wallisův test.

Testujeme nulovou hypotézu $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ a v případě zamítnutí nulové hypotézy nás zajímá, které dvojice zamítnutí způsobily.

Předpokladem jednofaktorové ANOVy je rovnost rozptylů (homoskedasticita), nezávislé výběry a normalita rozdělení (není tak důležitý přepodklad, hlavně u větších rozsahů). Pro ověření shody rozptylů lze použít následující testy.

- **Barlettův test** - je citlivý na porušení normality
- **Leveneův test** - je méně citlivý na porušení předpokladu normality, používáme ho, když nelze použít Barlettův test
- **Hartleyův test** - lze použít v případě shodných rozsahů
- **Cochranův test** - lze použít v případě shodných rozsahů

R COMMANDER - menu `Statistics-Variances-Bartlett's/Levene's test` nebo funkce `bartlett.test(), levene.test()`.
Myšlenkou analýzy rozptylu je, že celkovou variabilitu závisle proměnné rozdělíme do dvou částí, na variabilitu mezi skupinami a variabilitu uvnitř skupin. Variabilitu jednotlivých pozorování kolem celkového průměru charakterizuje celkový součet čtverců

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

a celkový rozptyl

$$MST = \frac{SST}{n-1}.$$

Variabilitu mezi skupinami charakterizují následující veličiny - meziskupinový součet čtverců

$$SSB = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$$

a rozptyl mezi skupinami

$$MSB = \frac{SSB}{k-1}.$$

Variabilitu uvnitř skupin popisuje tzv. reziduální součet čtverců

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

a reziduální rozptyl

$$MSB = \frac{SSE}{n-k}.$$

Testová hodnota se nazývá F-poměr a určí se podle vztahu

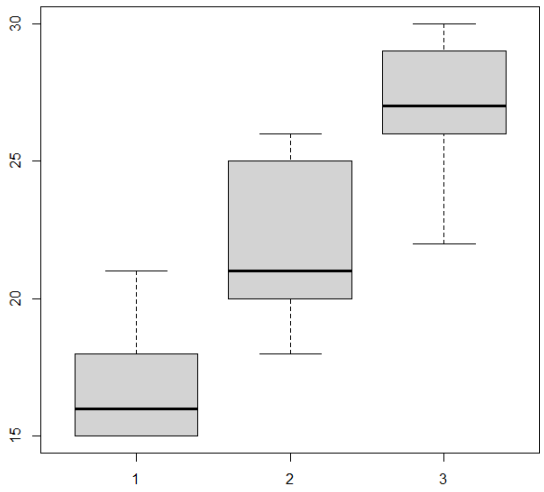
$$\text{F-hodnota} = \frac{MSB}{MSE}.$$

V případě nezamítnutí nulové hypotézy je závěr jasný a testování končí. Pokud však zamítneme H_0 ve prospěch H_1 , byla by naše analýza nekompletní, pokud bychom neidentifikovali, mezi kterými dvěma soubory existují statisticky významné rozdíly, kolik takových dvojic je a jaký je mezi nimi vztah. Tento další proces se nazývá post hoc analýza a spočívá v porovnávání středních hodnot všech dvojic populací, tzv. mnohonásobném porovnávání. Pro řešení problému mnohonásobného porovnávání existuje několik metod, jako například Fisherovo LSD, Scheffého a Tukeyova metoda.
R COMMANDER - menu `Statistics-Means-One-way ANOVA` nebo funkce `aov()`.

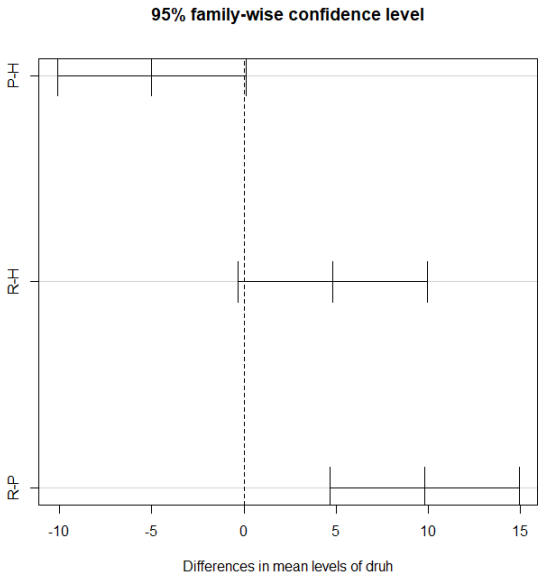
Řešený příklad

5 rostlin bylo rozděleno náhodně do tří skupin po pěti. Rostliny první skupiny byly pěstovány (každá ve zvláštním květináči) v písčité půdě, rostliny druhé skupiny v hlinité půdě a třetí skupiny v rašelině. Výšky rostlin na vrcholu sezóny:
písčitá: 15, 16, 18, 15, 21
hlinitá: 21, 20, 18, 25, 26
rašelina: 22, 26, 27, 30, 29
Má typ půdy vliv na výšku rostlin? Které skupiny se navzájem liší? Zkontrolujte homogenitu variancí.

```
> rostlina <- data.frame(druh=factor(rep(c("P","H","R"),c(5,5,5)))
   ,vyska=c(15,16,18,15,21,21,20,18,25,26,22,26,27,30,29))
> vysP <- rostlina$vyska[rostlina$druh=="P"]
> vysH <- rostlina$vyska[rostlina$druh=="H"]
> vysR <- rostlina$vyska[rostlina$druh=="R"]
> boxplot(vysP, vysH,vysR)
```




```
> bartlett.test(vyska~druh, data=rostlina)

        Bartlett test of homogeneity of variances

data:  vyska by druh
Bartlett's K-squared = 0.29587, df = 2, p-value = 0.8625

> aov <-aov(vyska~druh,data=rostlina)
> summary(aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
druh         2  240.1  120.07      13 0.000991 ***
Residuals   12  110.8    9.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = vyska ~ druh, data = rostlina)

druh
    diff         lwr        upr     p adj
P-H -5.0 -10.1271129  0.1271129 0.0561479
R-H  4.8  -0.3271129  9.9271129 0.0672929
R-P  9.8   4.6728871 14.9271129 0.0007081
> plot(TukeyHSD(aov))
```



Kruskalův-Wallisův test
Jedná se o neparametrickou obdobu jednocestné ANOVy. Používá se při nejistotě splnění předpokladů normality a homoskedesticity dat. Jedná se o vícevýběrový test shody mediánů. Nechť máme k nezávislých výběrů z rozdělení se spojitou distribuční funkcí o rozsazích $n_1, \ldots, n_k$, pak testujeme hypotézu ve tvaru

$$H_0 : \tilde{x}_1 = \cdots = \tilde{x}_n.$$

Pro výpočet testové statistiky použijeme obdobný postup jako v případě Mannova-Whitneyova testu. Pozorované hodnoty seřadíme, určíme jejich pořadí a spočteme součet pořadí prvků T_i, které patří do i-tého výběru $(i = 1, \cdots, k)$.

Testová statistika: $Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1) \sim \chi^2(k-1)$.

V případě, že $Q \geq \chi^2_{1-\alpha}(k-1)$, zamítáme nulovou hypotézu.

R COMMANDER - menu `Statistics-Nonparametric tests-Kruskal-Wallis test` nebo funkce `kruskal.test()`.

V případě zamítnutí provedeme post hoc analýzu. V případě stejného rozsahu souborů využijeme Neményho metodu - pokud je číslo $|T_i - T_j|$ větší nebo rovno kritické hodnotě (jsou tabelovány), zamítneme hypotézu, že i-tý a j-tý výběr pochází ze stejného rozdělení (mediány se statisticky významně liší).

Pokud jsou rozsahy výběrů různé, využijeme Dunnové metodu. Distribuční funkce i-tého a j-tého výběru se významně liší, pokud

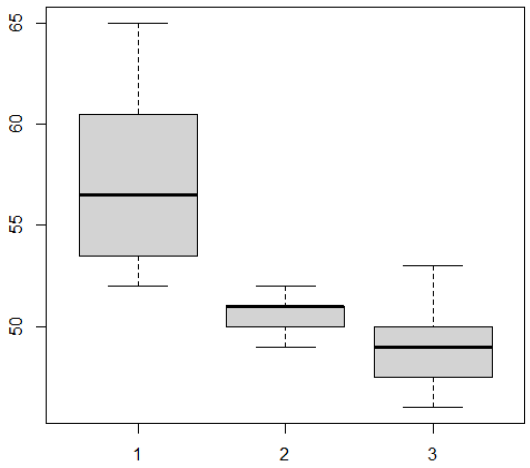
$$\left| \frac{T_i}{n_i} - \frac{T_j}{n_j} \right| > \sqrt{\frac{1}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) n(n+1) \chi^2_{1-\alpha}(k-1)}$$

Pro využití funkcí `posthoc.kruskal.dunn.test(), posthoc.kruskal.dunn.test()` je potřeba balíčku `PMCMR`.

Řešený příklad

Byla sledována doba bezporuchového chodu přístrojů tří různých značek A, B a C. Výsledky jsou
A: 55,54,58,61,52,60,53,65
B: 52,50,51,51,49
C: 47,53,49,50,16,48,50
Předpokládáme, že výběry pocházejí z exponenciálního rozdělení. Je mezi značkami rozdíl v kvalitě?

```
> pristroj<-data.frame(typ=factor(rep(c("A","B","C"),c(8,5,7))),
   doba=c
   (55,54,58,61,52,60,53,65,52,50,51,51,49,47,53,49,50,46,48,50))
> dobaA <- pristroj$doba[pristroj$typ=="A"]
> dobaB <- pristroj$doba[pristroj$typ=="B"]
> dobaC <- pristroj$doba[pristroj$typ=="C"]
> boxplot(dobaA,dobaB,dobaC)
```




```
> kruskal.test(doba~typ, data=pristroj)

        Kruskal-Wallis rank sum test

data:  doba by typ
Kruskal-Wallis chi-squared = 13.412, df = 2, p-value = 0.001224
```

Z výsledků vidíme, že zamítáme nulovou hypotézu (p-hodnota<0.05) o shodě mediánu. Provedeme post hoc analýzu, konkrétně Dunnové test (z důvodu různých rozsahů).

```
> require(PMCMR)
> posthoc.kruskal.dunn.test(x=pristroj$doba, g=pristroj$typ, p.
   adjust.method="none")
```

```
Warning in posthoc.kruskal.dunn.test.default(x = pristroj$doba, g
   = pristroj$typ,  :
  Ties are present. z-quantiles were corrected for ties.

        Pairwise comparisons using Dunn's-test for multiple
                        comparisons of independent samples

data:  pristroj$doba and pristroj$typ

  A       B
B 0.01957 -
C 0.00039 0.38959

P value adjustment method: none
```

Na základě analýzy lze tvrdit, že přístroje od značky A se kvalitou statisticky liší od přístrojů značek B, C, které jsou srovnatelné. Doba bezporuchového chodu u značky A je významně větší než u zbývajících značek.

7 Analýza závislostí, kontingenční tabulky

Pro sledování závislosti dvou nebo více kategoriálních proměnných využíváme **kontingenční tabulky**. Uvažujme náhodný vektor $Z = (X, Y)$, který má diskrétní rozdělení. Po uskutečnění výběru o rozsahu n zjistíme počet případů, kdy se ve výběru vyskytla dvojice (i, j) a označíme ho n_{ij}, tzn. jde o absolutní četnost. Kontingenční tabulka je definována jako matice (n_{ij}).

$X \setminus Y$	y_1	y_2	$\cdots$	y_s
x_1	n_{11}	n_{12}	$\cdots$	n_{1s}
x_2	n_{21}	n_{22}	$\cdots$	n_{2s}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}

Můžeme určit další číselné charakteristiky jako marginální četnosti, součty, relativní četnosti, případně znázornit graficky - mozaikový graf, shlukový graf. Čím je mozaikový graf členitější, tím silnější závislost lze mezi znaky předpokládat.

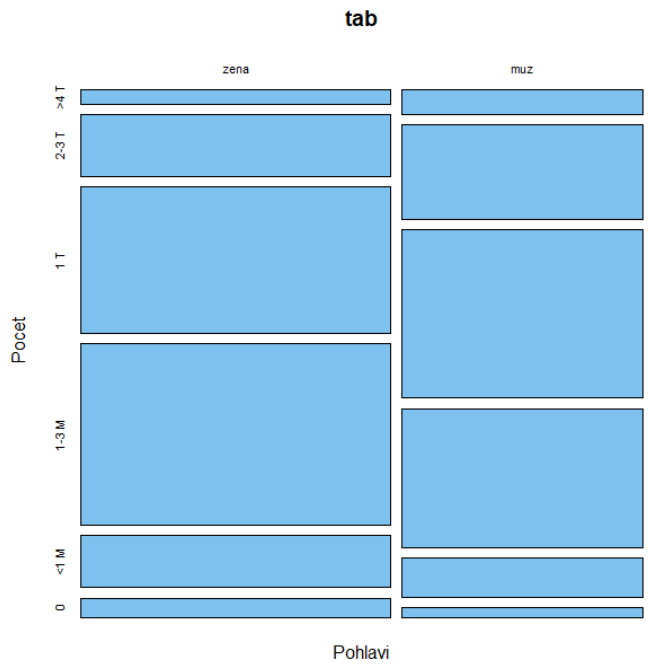
Řešený příklad

Na základě průzkumu jsme zjistili údaje o využití auta pro cestu do práce u žen a mužů.

pohlaví\ počet jízd	$> 4T$	$2-3T$	$1T$	$1-3M$	$< 1M$	0
žena	15	65	155	191	55	20
muž	20	78	138	114	132	8

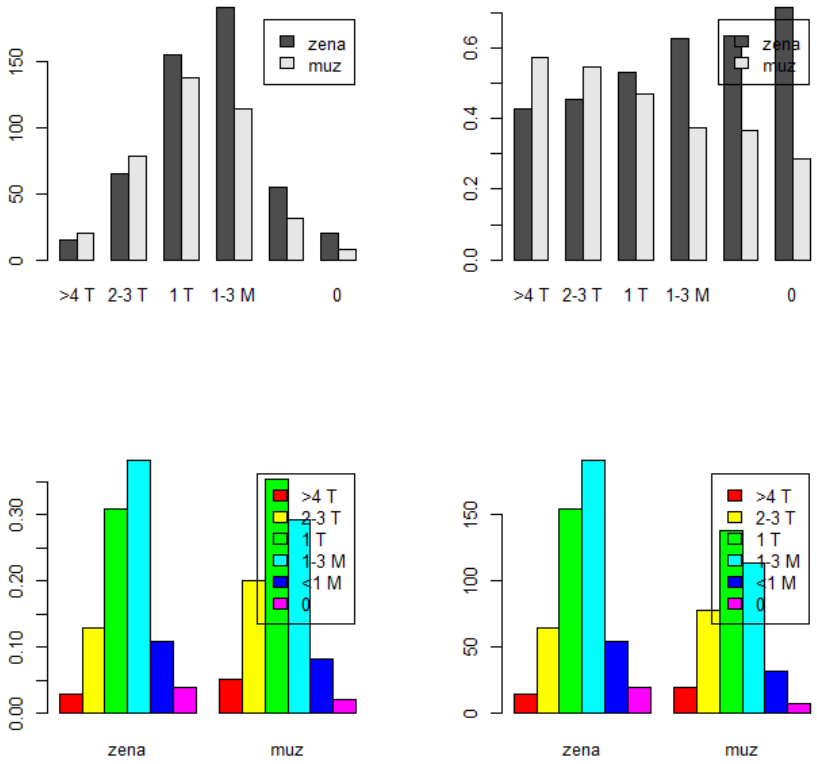
```
> tab=matrix(c(15,20,65,78,155,138,191,114,55,32,20,8),nrow=2)
> dimnames(tab)=list(Pohlavi=c("zena","muz"),Pocet=c(">4 T","2-3 T
   ","1 T","1-3 M","<1 M","0"))
> tab
       Pocet
Pohlavi >4 T 2-3 T 1 T 1-3 M <1 M  0
   zena   15    65 155   191   55 20
   muz    20    78 138   114   32  8
> rowSums(tab)  % součty v řádcích, lze využít i funkci apply(tab
   ,1,sum)
zena  muz
 501  390
> colSums(tab)  % součty ve sloupcích, lze využít i funkci apply(
   tab,2,sum)
 >4 T 2-3 T   1 T 1-3 M  <1 M     0
   35   143   293   305    87    28
> prop.table(tab) % tabulka relativních četností
       Pocet
Pohlavi       >4 T      2-3 T       1 T     1-3 M      <1 M
       0
```

```
   zena 0.01683502 0.07295174 0.1739618 0.2143659 0.0617284
   0.022446689
   muz  0.02244669 0.08754209 0.1548822 0.1279461 0.0359147
   0.008978676
> prop.table(tab,marg=1)  % tabulka marginálních četností
       Pocet
Pohlavi       >4 T     2-3 T       1 T     1-3 M       <1 M
      0
   zena 0.02994012 0.1297405 0.3093812 0.3812375 0.10978044
   0.03992016
   muz  0.05128205 0.2000000 0.3538462 0.2923077 0.08205128
   0.02051282
> prop.table(tab,marg=2)
       Pocet
Pohlavi      >4 T     2-3 T       1 T     1-3 M      <1 M
   0
   zena 0.4285714 0.4545455 0.5290102 0.6262295 0.6321839
   0.7142857
   muz  0.5714286 0.5454545 0.4709898 0.3737705 0.3678161
   0.2857143
> mosaicplot(tab,color = "skyblue2")
```




```
> par(mfrow=c(2,2))
> barplot(tab,beside=T,legend=T)  % shlukové a sloupcové grafy
> barplot(prop.table(tab,mar=2),beside=T,legend=T)
> barplot(prop.table(t(tab),mar=2),beside=T,legend=T,col=rainbow
   (6))
```

```
> barplot(t(tab),beside=T,legend=T,col=rainbow(6))
```



Po grafické analýze a určení číselných charakteristik můžeme testovat některé z následujících hypotéz: hypotéza nezávislosti, hypotéza symetrie,...

7.1 χ^2 test nezávislosti

Testujeme nulovou hypotézu H_0, že sledované znaky jsou statisticky nezávislé, je založen na porovnávání teoretických četností n'_{ij} s pozorovanými n_{ij}, funkce `chisq.test()`.

Teoretické četnosti:

$$n'_{ij} = \frac{n_{i.} \cdot n_{.j}}{n},$$

kde $n_{i.}, n_{.j}$ jsou příslušné marginální četnosti

Testová statistika:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}},$$

která má v případě platnosti přibližně χ^2 rozdělení s $(s-1)(r-1)$ stupni volnosti

$p-$hodnota:

$$p - hodnota = 1 - F(\chi^2)$$

- podmínky použití testu:

- žádná z očekávaných četností nesmí být menší než 2
- alespoň 80% z očekávaných četností musí být větší než 5

Pro předcházející příklad:

```
> n<-sum(tab)
> n
[1] 891
> Pohlavi<-c(sum(tab[1,]),sum(tab[2,]))
> Pohlavi
[1] 501 390
> Pocet<-c(sum(tab[,1]),sum(tab[,2]),sum(tab[,3]),sum(tab[,4]),sum
   (tab[,5]),sum(tab[,6]))
> Pocet
[1]  35 143 293 305  87  28
> teor_cetnosti<-matrix(ncol=6,nrow=2)
> for (i in 1:2)
+ for (j in 1:6)
+ teor_cetnosti[i,j]<-Pohlavi[i]*Pocet[j]/n
> teor_cetnosti
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 19.68013 80.40741 164.7508 171.4983 48.91919 15.74411
[2,] 15.31987 62.59259 128.2492 133.5017 38.08081 12.25589
> chisq.test(tab,correct=FALSE)

        Pearson's Chi-squared test

data:  tab
X-squared = 20.028, df = 5, p-value = 0.001235
> chisq.test(tab,correct=FALSE)$residuals  %Pearsonovy residuály,
   tj. hodnoty $(O-E)/\sqrt{E}$
       Pocet
Pohlavi      >4 T     2-3 T        1 T     1-3 M       <1 M
     0
   zena -1.054980 -1.718231 -0.7596758  1.489163  0.8694041
   1.072585
   muz   1.195724  1.947458  0.8610232 -1.687830 -0.9853902
   -1.215677
```

p-hodnota<0.05, proto zamítáme nulovou hypotézu ve prospěch alternativy, tzn. že využití auta souvisí s pohlavím.

V případě nesplnění podmínek testu lze použít Yatesovu korekci, nevýhodou je, že test má menší sílu, v R je korekce prováděna automaticky, parametr correct=FALSE korekci vypne.

Pro posouzení síly vztahu závislosti lze použít pro čtvercové kontingence **Pearsonův koeficient kontingence**, který může nabývat hodnoty z $(0;1)$, přičemž 0 znamená nezávislost

a 1 silnou závislost, maximální hodnota závisí na velikosti tabulky - není vždy 1:

$$P = \frac{\chi^2}{\chi^2 + n}$$

- další používané koeficienty: Cramerův a Čuprovův koeficient kontingence

8 Korelační a regresní analýza

Regrese a korelace slouží k popisu vztahu dvou kontinuálních proměnných (i když obecně může být nezávislých proměnných několik.). V případě regresní analýzy jsme shopni určit, která z proměnných je závislá a která nezávislá. Předpokládáme, že nezávislá proměnná je určena (změřena) přesně, i když v praxi stačí, že zatížení chybou měření je u nezávislé proměnné mnohem menší než u závislé.

Cílem regresní analýzy je nalezení modelu (regresní rovnice), pomocí které lze predikovat hodnotu závislé proměnné při dané hodnotě proměnné nezávislé a určení koeficientu determinace, který definuje míru vysvětlené variability.

Jednoduchá regresní analýza je nejjednodušším typem regrese a určuje, že máme pouze jednu nezávislou (vysvětlující) proměnnou. V případě více nezávislých proměnných hovoříme o **mnohonásobné regresní analýze**.

Regrese může být lineární, tzn. že pro popis vztahu mezi proměnnými využíváme funkce lineární v parametrech, případně funkce, které lze na lineární vhodnou transformací převést, jinak se jedná o regressi nelineární.

8.1 Jednoduchá lineární regrese

Regresní model

$$y = f(x)$$

vysvětluje vztah mezi veličinou y a hodnotách x prostřednictvím regresní funkce f. Známe-li n pozorovaných dvojic (x_i, y_i) a předpokládáme, že hodnoty y_i jsou naměřeny s určitou chybou e_i, dostáváme n rovnic

$$y_i = f(x_i) + e_i, \quad i = 1, 2, \ldots, n.$$

Příklady regresních funkcí:

- regresní přímka: $y = b_0 + b_1 x$
- regresní parabola: $y = b_0 + b_1 x + b_2 x^2$
- regresní hyperbola: $y = b_0 + \frac{b_1}{x}$
- logaritmická funkce: $y = b_0 + b_1 \ln x$
- exponenciální funkce: $y = b_1^{b_0 x}$

Lineární regresní model je takový, kde je odhadovaná závislost (v populaci) popsána lineární funkcí

$$Y = \beta_0 + \beta_1 x.$$

Cílem regrese je nalézt parametry β_0, β_1. Pro nalezení parametrů se využívá **metoda nejmenších čtverců**, která spočívá v tom, že hledáme parametry, pro které je součet čtverců chyb modelu minimální. Hledáme minimum funkce

$$G(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2$$

a tedy řešíme soustavu rovnic

$$\frac{\partial G(\beta_0, \beta_1)}{\partial \beta_0} = 0,$$
$$\frac{\partial G(\beta_0, \beta_1)}{\partial \beta_1} = 0.$$

Po úpravách dostáváme nejlepší nestranné bodové odhady parametrů regresní přímky, mající nejmenší rozptyl ze všech nestranných odhadů

$$b_0 = \bar{Y} - b_1 \bar{x},$$
$$b_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2},$$

kde $\bar{x}, \bar{Y}$ značí průměry zadaných hodnot.

8.2 Verifikace modelu

Po nalezení konkrétního odhadu regresní funkce na základě výběru si musíme položit otázku, zda je nalezený odhad kvalitní, zda byl zvolen vhodný typ regresní funkce, atd. Je tedy potřeba provést verifikaci modelu a hodnocení kvality modelu.

Minimum funkce

$$g(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n} Y_i^2 - b_0 \sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} x_i Y_i$$

se nazývá **reziduální součet čtverců** a označujeme ho SSE.

Celkový součet čtverců

$$S_T^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

značí celkovou kvadratickou chybu modelu.

Koeficient determinace R^2
Udává kvalitu modelu a určuje, jaké procento rozptylu vysvětlované proměnné je vysvětleno modelem; nabývá hodnot od 0 do 1; čím je vyšší hodnota, tím je model vhodnější a lépe vystihuje naměřená data.

$$R^2 = 1 - \frac{SSE}{S_T}.$$

Pokud není významný rozdíl mezi koeficienty determinace u různých modelů, vždy je

vhodné zvolit model jednodušší. Hodnota R^2 roste s počtem koeficientů k, proto je nutné modely s více koeficienty porovnávat pomocí upraveného koeficientu determinace

$$R^2_{ADJ} = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

Reziduální rozptyl je nevychýlený odhad rozptylu σ^2

$$\widehat{\sigma^2} = s_e^2 = \frac{SSE}{n-k-1}.$$

Celkový F-test při kterém testujeme, zda hodnota vysvětlované proměnné závisí na lineární kombinaci vysvětlujících proměnných.

H_0 : zvolený model není statisticky významný $(b_1 = \cdots = b_k = 0)$
$H_1 : \neg H_0$

Testovací kritérium: $F = \frac{S_R^2}{k} : \frac{SSE}{n-k-1}$

Testovací statistika má F-rozdělení pravděpodobnosti s k a $n-k-1$ stupni volnosti. Kritickou hodnotou je kvantil $F(k, n-k-1)$ a $p - hodnota = 1 - F(x_{OBS})$.

Často nás zajímá, zda je možné model zjednodušit tak, že hodnoty Y_i nezávisí na x_i, tudíž potřebuje testovat hypotézu

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0.$$

Dílčí t-test testuje oprávněnost setrvání koeficientů v regresním modelu, každý koeficient se testuje zvlášť a je zkoumá, zda směrodatná chyba s_{b_j} odhadů koeficientů není natolik významná, že je možné příslušné koeficienty považovat za nulové.
Testová statistika

$$T = \frac{b_1}{s}\sqrt{\sum x_i^2 - n\bar{x}}$$

má Studentovo rozdělení o $n-2$ stupních volnosti. Pokud $|T| \geq t_{n-2}(1-\alpha/2)$ zamítáme H_0 a potvrdíme lineární závislost.

8.3 Intervaly spolehlivosti

Pro parametr β_1 lze zkontruovat intervalový odhad o spolehlivosti $1-\alpha$

$$\left\langle b_1 - \frac{t_{n-2}(1-\alpha/2)s}{\sqrt{\sum x_i^2 - n\bar{x}^2}}, b_1 + \frac{t_{n-2}(1-\alpha/2)s}{\sqrt{\sum x_i^2 - n\bar{x}^2}} \right\rangle.$$

Odhad pro $\beta_0 + \beta_1 x$:

$$\left\langle b_0 + b_1 x - t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}, b_0 + b_1 x + t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x_i^2 - n\bar{x}^2}} \right\rangle.$$

S využitím intervalů pro všechna x_i sestrojíme pás spolehlivosti kolem regresní přímky.

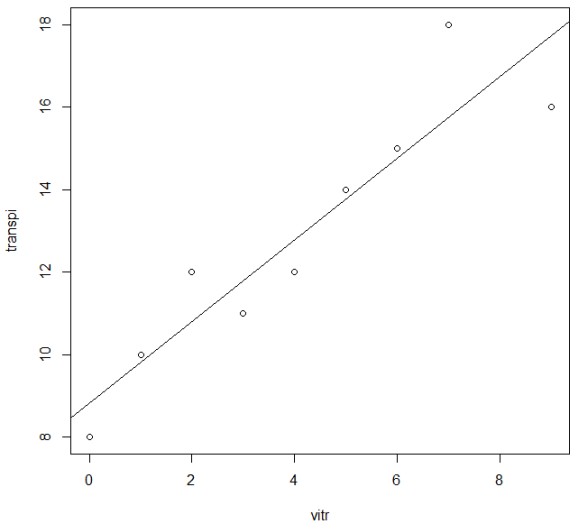
Řešený příklad

Byla studována závislost transpirace na rychlosti větru. Byla získána následující data

Vitr	2	9	5	6	7	3	4	1	0
Transpi	12	16	14	15	18	11	12	10	8

Pro lineární regresi používáme funkce `lm()`, která provede odhad modelu. Pro další analýzy se využívá funkcí `summary(), anova(), plot()` a další.

```
> vetry<-data.frame(vitr=c(2,9,5,6,7,3,4,1,0),transpi=c
   (12,16,14,15,18,11,12,10,8))
> lm.vetry<-lm(transpi~vitr,data=vetry)
> lm.vetry

Call:
lm(formula = transpi ~ vitr, data = vetry)

Coefficients:                %nalezene koeficienty
(Intercept)          vitr
     8.8242        0.9887
> coef(lm.vetry)
(Intercept)         vitr
  8.8241935   0.9887097
> plot(transpi~vitr,data=vetry)
> abline(lm.vetry)
```



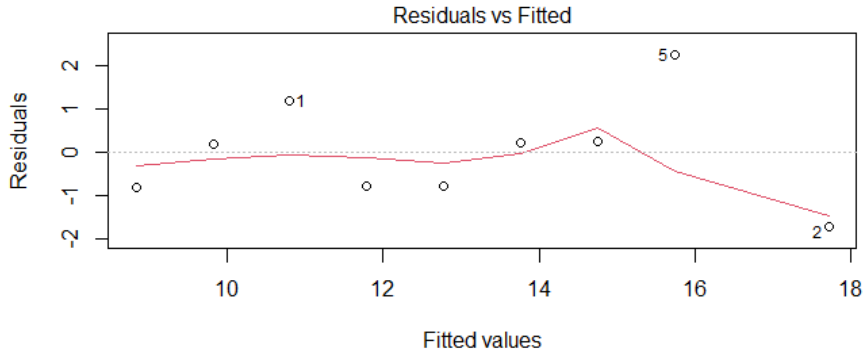
Odhad hodnoty transpirace při hodnotě větru 10 spolu s 95% intervalovým odhadem.

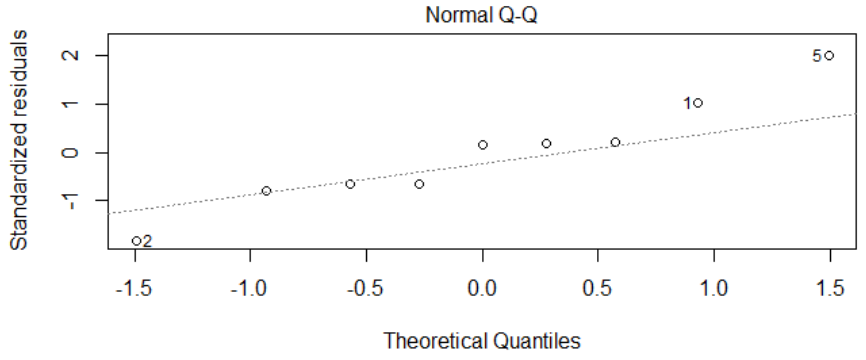
```
> a <- data.frame(vitr = 10)
> result <-  predict(lm.vetry,a)
```

```
> result
       1
18.71129
> predict(lm.vetry, a, interval = "confidence")
       fit      lwr      upr
1 18.71129 16.33056 21.09202
```

Vykreslíme rezidua - měly by být rovnoměrně rozloženy kolem osy x a porovnáme je s normálním rozdělením (Q-Q graf).

```
> par(mfrow=c(2,1))
> plot(lm.vetry,which=1)
> plot(lm.vetry,which=2)
```





Vypíšeme základní výsledky

```
> summary(lm.vetry)

Call:
lm(formula = transpi ~ vitr, data = vetry)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7226 -0.7903  0.1871  0.2435  2.2548
```

```
Coefficients:       %tabulka koeficientu
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.8242     0.7668  11.508 8.42e-06 ***      %
   konstantni clen
vitr          0.9887     0.1547   6.389 0.000371 ***      %sklon
   primky
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.284 on 7 degrees of freedom
Multiple R-squared:  0.8536,    Adjusted R-squared:  0.8327
F-statistic: 40.82 on 1 and 7 DF,  p-value: 0.000371
```

Celkový F-test:

```
> anova(lm.vetry)
Analysis of Variance Table

Response: transpi
          Df Sum Sq Mean Sq F value   Pr(>F)
vitr       1 67.342  67.342  40.825 0.000371 ***
Residuals  7 11.547   1.650
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

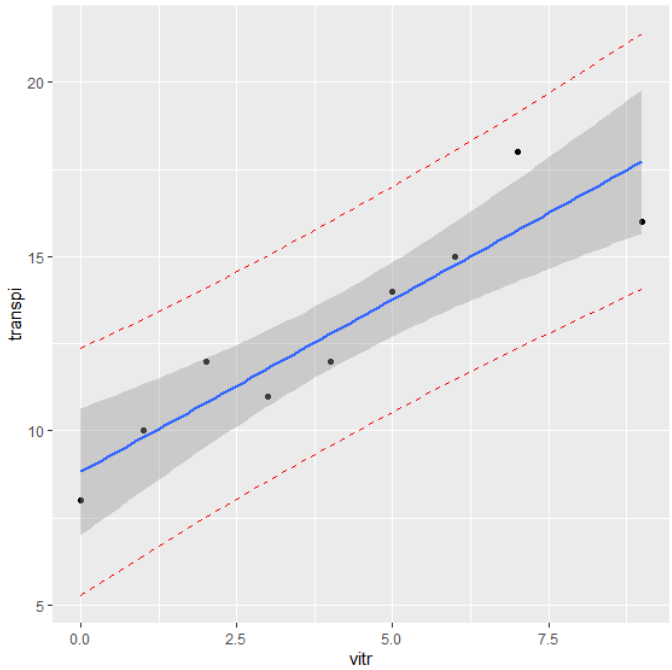
Celková suma čtverců je rozdělena do regresní sumy čtverců (hodnota 67.342) a residuální sumy čtverců (hodnota 11.547).

Intervaly spolehlivosti pro parametry:

```
> confint(lm.vetry,level=0.99)
                0.5 %    99.5 %
(Intercept) 6.1407928 11.507594
vitr        0.4471945  1.530225
```

Znázorníme konfidenční a predikční pás.

```
> pred <- predict(lm.vetry, interval ="prediction") %pridame
   predikce
> mydata<-cbind(vetry,pred)
> library("ggplot2")
> p<-ggplot(mydata,aes(vitr,transpi))+geom_point()+stat_smooth(
   method=lm)
> p+geom_line(aes(y=lwr),color="red", linetype = "dashed")+
   geom_line(aes(y=upr),color="red",linetype="dashed")
```



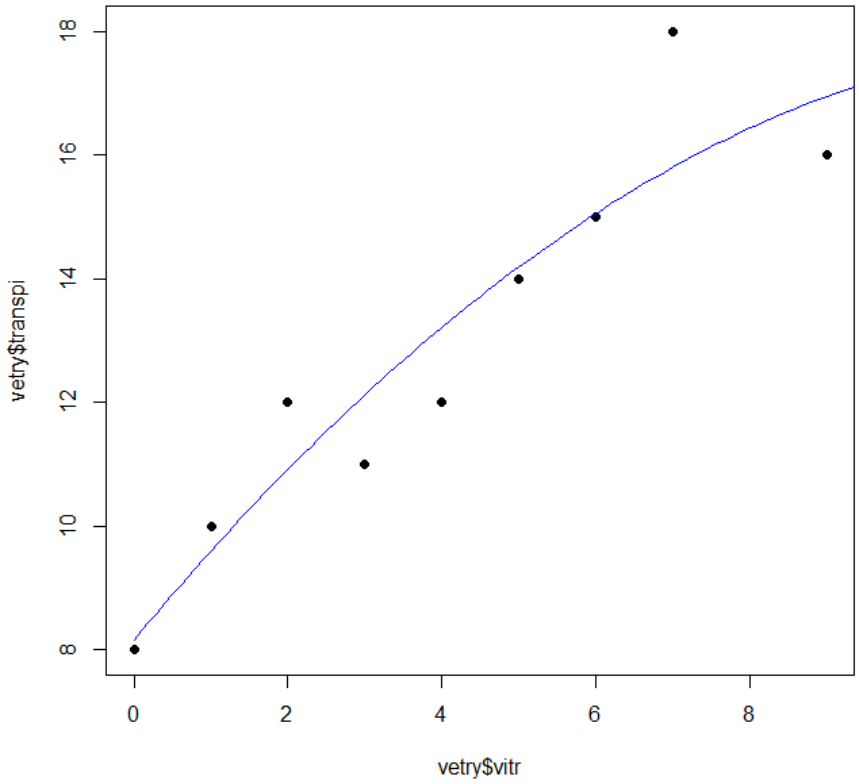
V případě volby kvadratického modelu musíme první vytvořit novou proměnnou $vitr^2$.

```
> vetry$vitr2 <- vetry$vitr^2
> kvadrModel <- lm(transpi ~ vitr + vitr2, data=vetry)
> summary(kvadrModel)

Call:
lm(formula = transpi ~ vitr + vitr2, data = vetry)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2139 -0.9455 -0.1636  0.4017  2.2030

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.16364    1.00891   8.092 0.000191 ***
vitr         1.49199    0.52352   2.850 0.029186 *
vitr2       -0.05736    0.05701  -1.006 0.353155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.283 on 6 degrees of freedom
Multiple R-squared:  0.8748,    Adjusted R-squared:  0.833
F-statistic: 20.95 on 2 and 6 DF,  p-value: 0.001964
```

A vykreslíme:

```
> vitrValues <- seq(0, 10, 0.1)
> transpiPredict <- predict(kvadrModel,list(vitr=vitrValues, vitr2
   =vitrValues^2))
> plot(vetry$vitr, vetry$transpi, pch=16)
```

```
> lines(vitrValues, transpiPredict, col='blue')
```



Případná transformace se provede zadáním funkce přímo ve funkci `lm()`. Například:

```
> lm.vetry.loglog <- lm(log(transpi)~log(vitr+1),data=vetry)
> coef(lm.vetry.loglog)
  (Intercept) log(vitr + 1)
    2.0627007     0.3250512
```

8.4 Korelační analýza

Při regresi jsme vycházeli z předpokladu, že mezi proměnnými existuje funkční závislost a že nezávislá proměnná není zatížená chybou. V korelační analýze předpokládáme, že dvě proměnné jsou pouze korelovány, jsou zatížey chybou a mezi proměnnými není nutně funkční závislost.

Mírou těsnosti lineárního vztahu je **korelační koeficient**. Využíváme funkci `cor()`.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}.$$

Pro předchozí příklad dostáváme:

```
> cor(vetry$transpi, vetry$vitr)
[1] 0.9239223
```

V případě přímkové regrese výběrový korelační koeficient roven odmocnině z indexu determinace.

Čím více se hodnota korelačního koeficientu blíží k hodnotě jedna, tím je závislost silnější, obě hodnoty společně rostou.

Čím blíže je hodnota korelačního koeficientu k hodnotě -1, tím je závislost silnější, rostou-li hodnoty jedné proměnné, hodnoty druhé klesají.

Je-li hodnota blízká nule, nejsou proměnné závislé.

O síle závislosti vypovídá nejen korelační koeficient, ale i rozsah souboru n.

Slovní interpretace:

$\lvert r\rvert = 0$	korelační nezávislost
$0 < \lvert r\rvert < 0.3$	nízký stupeň korelační závislosti
$0.3 \leq \lvert r\rvert < 0.5$	mírný stupeň korelační závislosti
$0.5 \leq \lvert r\rvert < 0.7$	střední stupeň korelační závislosti
$0.7 \leq \lvert r\rvert < 0.9$	vysoký stupeň korelační závislosti
$0.9 \leq \lvert r\rvert < 1$	velmi vysoký stupeň korelační závislosti
$\lvert r\rvert = 1$	funkční závislost

Příklad

a) Za prvních sedm měsíců roku má firma záznamy o počtu hodin provozu výrobní linky a o nákladech na údržbu v tis. Kč.

hodiny	275	350	250	325	375	400	300
náklady	149	170	140	164	192	200	165

Proveďte regresní a korelační analýzu. Zvolte regresní přímku.

b) Byla zjišťována závislost délky trvání vegetační sezóny na nadmořské výšce plochy. Byly naměřeny hodnoty:

výška (m)	600	650	665	750	850	880	950	1000	1005
délka sezóny (dny)	150	144	145	140	110	105	110	99	103

Proveďte regresní a korelační analýzu.

8.5 Mnohonásobná lineární regrese

Model mnohonásobné regrese vyjadřuje vztah mezi jednou vysvětlovanou a mnoha vysvětlujícími proměnnými. Nezávislé proměnné označíme $x_1, x_2, \ldots$. i-té pozorování první nezávislé proměnné bude označeno x_{1i}. V případě jednoduché regrese byla pro lineární model regresní rovnice ve tvaru

$$Y = \beta_0 + \beta_1 x$$

a obrazem byla přímka. Pokud budeme mít dvě nezávislé proměnné, bude regresní rovnice ve tvaru

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

a obrazem bude rovina v prostoru.

Model lze zobecnit pro více proměnných:

$$Y = \beta_0 + \sum_j \beta_j x_j.$$

Koeficienty β_j označujeme jako parciální regresní koeficienty, jejichž hodnoty odhadujeme na základě výběru. Obdobně jako u jednoduché regrese lze pomocí t-testu (test o významnosti vlivu proměnné) ověřit platnost hypotézy, že hodnota koeficientu je rovna nule: testová statistika $T_j = \frac{b_j}{SE(b_j)}$, kde $SE(b_j)$ je střední chyba parametru, má t-rozdělení s $n - m - 1$ stupni volnosti, kde m je počet vysvětlujících proměnných. Pro koeficient b_0 nemá test reálný smysl. Test příslušného parciálního koeficientu testuje, zda testované proměnná přináší v daném souboru proměnných významné množství informace k vysvětlení variability závislé proměnné.

Pro ověření, že regresní model nevysvětluje žádnou část variability závislé proměnné, opět využijeme F-test (viz. kapitola 8.2).

Stejně jako u jednoduché regrese se vypočítá hodnota koeficientu determinace, ale jelikož se jedná o vychýlený odhad koeficientu determinace základního souboru (čím méně dat a více nezávislých proměnných, tím je hodnota vyšší), musíme raději použít korigovaný koeficient determinace.

V případě, že celková regrese je průkazná (F-test, koeficient determinace), ale žádný z parciálních regresních koeficientů průkazně odlišný od nuly není, tak to většinou znamená, že vysvětlující proměnné jsou vzájemně korelované. Což by ale být nemělo, měly by být nezávislé.

Řešený příklad

Byla studována závislost transpirace nejen na rychlosti větru, ale i teplotě a vlhkosti vzduchu. Byla získána následující data

Vitr	2	9	5	6	7	3	4	1	0
Transpi	12	16	14	15	18	11	12	10	8
Teplota	10	12	8	16	22	7	11	15	5
Vlhkost	50	80	62	95	45	32	92	46	58

```
> hodnoty<-data.frame(vitr=c(2,9,5,6,7,3,4,1,0),transpi=c
   (12,16,14,15,18,11,12,10,8), teplota=c(10,12,8,16,22,7,11,15,5)
   ,vlhkost=c(50,80,62,95,45,32,92,46,58))
> hodnoty
  vitr transpi teplota vlhkost
1    2      12      10      50
2    9      16      12      80
3    5      14       8      62
4    6      15      16      95
5    7      18      22      45
6    3      11       7      32
7    4      12      11      92
8    1      10      15      46
9    0       8       5      58
> model<-lm(transpi~vitr+teplota+vlhkost,data=hodnoty)
> summary(model, corr=T)

Call:
lm(formula = transpi ~ vitr + teplota + vlhkost, data = hodnoty)

Residuals:
      1       2       3       4       5       6       7       8
      9
 1.0834 -0.9368  1.0012  0.2464  0.4368 -0.5668 -0.1442 -0.9999
  -0.1199

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.21085    1.26298   6.501  0.00129 **
vitr         0.88176    0.15055   5.857  0.00206 **
teplota      0.17938    0.07652   2.344  0.06602 .
vlhkost     -0.01703    0.01730  -0.985  0.36999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9656 on 5 degrees of freedom
Multiple R-squared:  0.9409,    Adjusted R-squared:  0.9055
F-statistic: 26.54 on 3 and 5 DF,  p-value: 0.001692

Correlation of Coefficients:
        (Intercept) vitr  teplota
vitr     0.26
teplota -0.60       -0.52
vlhkost -0.75       -0.44  0.16
```

Lze vidět, že p-hodnota F-testu je <0.05. To znamená, že alespoň jedna z vysvětlujících proměnných významně ovlivňuje variabilitu vysvětlované proměnné. Z tabulky koeficientů vidíme, že změna vlhkost nemá významný vliv na hodnotu transpirace, lze ji z modelu

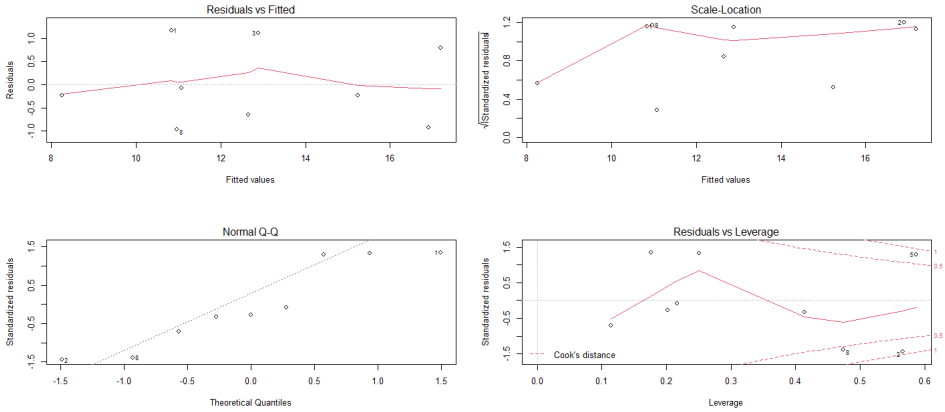
odstranit a vytvoříme model pro dvě vysvětlující proměnné - teplota, vítr.

```
> model2<-lm(transpi~vitr+teplota,data=hodnoty)
> print(model2)

Call:
lm(formula = transpi ~ vitr + teplota, data = hodnoty)

Coefficients:
(Intercept)          vitr      teplota
     7.2810       0.8159        0.1914
> summary(model2)

Call:
lm(formula = transpi ~ vitr + teplota, data = hodnoty)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9672 -0.6494 -0.2378  0.7981  1.1737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.28104    0.83662   8.703 0.000127 ***
vitr         0.81588    0.13453   6.065 0.000912 ***
teplota      0.19135    0.07535   2.539 0.044121 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9631 on 6 degrees of freedom
Multiple R-squared:  0.9295,    Adjusted R-squared:  0.9059
F-statistic: 39.52 on 2 and 6 DF,  p-value: 0.0003511
> cor(hodnoty$vitr,hodnoty$teplota)
[1] 0.5059396
```

Nalezený model je ve tvaru: transpirace $= 7.28 + 0.82 \cdot \text{vitr} + 0.19 \cdot \text{teplota}$.

Rezidua:

```
> residuals(model2)
          1           2           3           4           5
          6
 1.17369668 -0.92014218  1.10876777 -0.23791469  0.79810427
  -0.06812796
          7           8           9
-0.64940758 -0.96718009 -0.23779621
> layout(matrix(c(1,2,3,4),2,2))
> plot(model2)
```

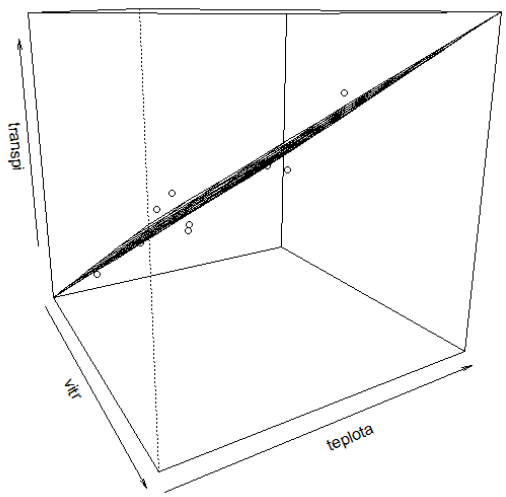


Konfidenční intervaly pro koeficientů jsou:

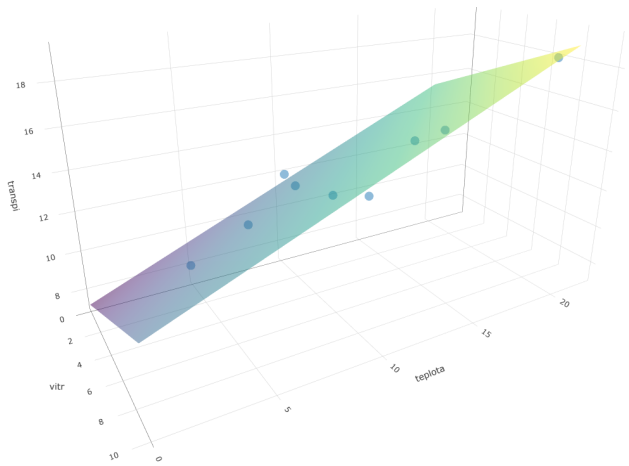
```
> confint(model2)
                  2.5 %     97.5 %
(Intercept) 5.233897391 9.3281879
vitr        0.486701471 1.1450521
teplota     0.006963998 0.3757374
```

Vykreslení dat a regresního modelu - 2 způsoby vykreslení:

```
> hodnoty.marg<-list(vitr=seq(0,10,by=1),teplota=seq(0,30,by=2.5))
> hodnoty.fit<-predict(model2,expand.grid(hodnoty.marg))
> res<-persp(hodnoty.marg$vitr,hodnoty.marg$teplota,matrix(hodnoty
   .fit,11),xlab="vitr",ylab="teplota", zlab="transpi",theta=60)
> points(trans3d(hodnoty$vitr, hodnoty$teplota,hodnoty$transpi,res
   ))
```



```
> install.packages("plotly")
> library(plotly)
 plot_ly(data = hodnoty, z = ~transpi, x = ~vitr, y = ~teplota,
   opacity = 0.5)%>% add_markers()
> x <- seq(0, 10, by = 0.1)
> y <- seq(0, 20, by = 0.2)
> plane <- outer(x, y, function(a, b){model2$coef[1] +model2$coef
   [2]*a + model2$coef[3]*b})
> plot_ly(data = hodnoty, z = ~transpi, x = ~vitr, y =~teplota,
   opacity = 0.5) %>%add_markers() %>%a dd_surface(x = ~x, y = ~y,
    z = ~plane, showscale = FALSE)
```



Příklad

Byla zjišťována závislost výšky rostliny na hladině podzemní vody a na množství dusíku.

hladina vody	5	5	5	10	10	10	15	15	15	20	20	20
množství dusíku	1	2	3	1	2	3	1	2	3	1	2	3
výška rostliny	15	17	20	13	16	17	10	12	15	10	12	13

Vyhodnoťte s využitím mnohonásobné regrese, zda má hladina podzemní vody a množství dusíku vliv na výšku rostliny.

9 Časové řady

Ke zkoumání dynamiky jevů v čase slouží **časové řady.** Mají základní význam pro analýzu příčin, které na tyto jevy působily a ovlivňovaly jejich chování v minulosti, tak pro předvídání jejich budoucího vývoje. Časová řada je uspořádaná (od minulosti do přítomnosti) posloupnost pozorování jistého věcného a prostorově vymezeného ukazatele.

Časové řady lze dělit podle časového omezení, podle délky intervalu, případně podle sledovaných ukazatelů.

1. podle časového omezení
   - **intervalové** - hodnota ukazatele závisí na celém sledovaném intervalu (spotřeba energie, náklady na mzdy)
   - **okamžikové** - hodnoty se vztahují ke konkrétnímu časovému okamžiku, na který nemá vliv předchozí hodnoty (počet uživatelů v konkrétním měsíci)
2. podle délky intervalu
   - **krátkodobé** - interval kratší jak jeden rok
   - **dlouhodobé**
3. podle druhu sledovaných dat
   - **absolutní/primární** - přímo s daty (počet výpadků za měsíc)
   - **odvozené** - pracuje se například s průmerem nebo kumulací ukazatelů (průměrná spotřeba energie)

9.1 Popisné statistiky

U intervalové řady se k interpretaci využívá součet nebo aritmetický průměr

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t.$$

V případě okamžikové řady se provádí **chronologický průměr.** Je důležité rozlišit, zda se výpočet provádí pro stejně vzdálené úseky mezi jednotlivými okamžiky či ne.

- **prostý chronologický průměr** - stejné vzdálenosti

$$\bar{y} = \frac{1}{n-1}\left(\frac{y_1+y_2}{2} + \frac{y_2+y_3}{2} + \cdots + \frac{y_{n-1}+y_n}{2}\right) = \frac{1}{n-1}\left(\frac{y_1}{2} + \sum_{t=2}^{n-1} y_t + \frac{y_n}{2}\right)$$

- **vážený chronologický průměr** - nestejné vzdálenosti, kde d_t, $t = 1, \ldots, n-1$ jsou délky jednotlivých intervalů

$$\bar{y} = \frac{1}{d_1 + d_2 + \cdots + d_{n-1}}\left(\frac{y_1+y_2}{2}d_1 + \frac{y_2+y_3}{2}d_2 + \cdots + \frac{y_{n-1}+y_n}{2}d_{n-1}\right)$$

Řešený příklad

Vždy k prvnímu lednu evidujeme počet zaměstnanců. Počínaje rokem 2010 máme následující údaje: 384, 425, 420, 480, 505, 486, 535, 528. Jaký byl průměrný počet zaměstnanců za sledované roky?

Jedná se o okamžikovou řadu se stejně vzdálenými okamžiky (jeden rok). Použijeme tedy prostý chronologický průměr.

$$\bar{y} = \frac{\frac{384}{2} + 425 + 420 + 480 + 505 + 486 + 535 + \frac{528}{2}}{8-1} = \frac{3307}{7} \doteq 472.43$$

Průměrný počet zaměstnanců byl 472.

Druhý způsob spočívá v určení aritmetického průměru z přepočtené úsekové řady (viz. tabulka).

$$\bar{y} = \frac{3307}{7} \doteq 472.43$$

Rok	Počet zaměstnanců	Přepočítaná úseková řada
2010	384	
2011	425	$\frac{384+425}{2} = 404.5$
2012	420	$\frac{425+420}{2} = 422.5$
2013	480	$\frac{420+480}{2} = 450.0$
2014	505	$\frac{480+505}{2} = 492.5$
2015	486	$\frac{505+486}{2} = 495.5$
2016	535	$\frac{486+535}{2} = 510.5$
2017	528	$\frac{535+528}{2} = 531.5$

9.2 Míry dynamiky

Před vlastní analýzou časové řady je užitečné charakterizovat chování řady pomocí měr dynamiky, které slouží k charakteristice základních rysů chování časových řad a následnému přizpůsobení kritérií pro modelaci.

- **absolutní přírustek** (první diference) - nejjednodušší míra dynamiky, charakterizuje přírustek hodnoty časové řady v čase t oproti času $t-1$

  $$\Delta y_t = y_t - y_{t-1},\ t = 2, 3, \ldots, n$$

  - diferencováním první diference lze určit druhou diferenci

  $$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1},\ t = 3, 4, \ldots, n$$

- **průměrný absolutní přírustek** - za celou časovou řadu

  $$\bar{\Delta} = \frac{\sum_{t=2}^{n} \Delta y_t}{n-1} = \frac{y_n - y_1}{n-1}$$

- **relativní přírustek** - "o kolik procent"se změnila časová řada mezi jednotlivými okamžiky, uvádí se v procentech $\delta_t \cdot 100\%$

$$\delta_t = \frac{\Delta y_t}{y_{t-1}} = \frac{y_t - y_{t-1}}{y_{t-1}} = \frac{y_t}{y_{t-1}} - 1,\ t = 2,3,\ldots,n$$

- **koeficient růstu** - vyjádřený v procentech udává, o kolik procent vzrostla hodnota časové řady v okamžiku t oproti $t-1$

$$k_t = \frac{y_t}{y_{t-1}},\ t = 2,3,\ldots,n$$

- **průměrný koeficient růstu** - udává průměrnou rychlost růstu, či poklesu hodnot

$$\bar{k} = \sqrt[n-1]{k_1 k_2 \cdots k_n} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

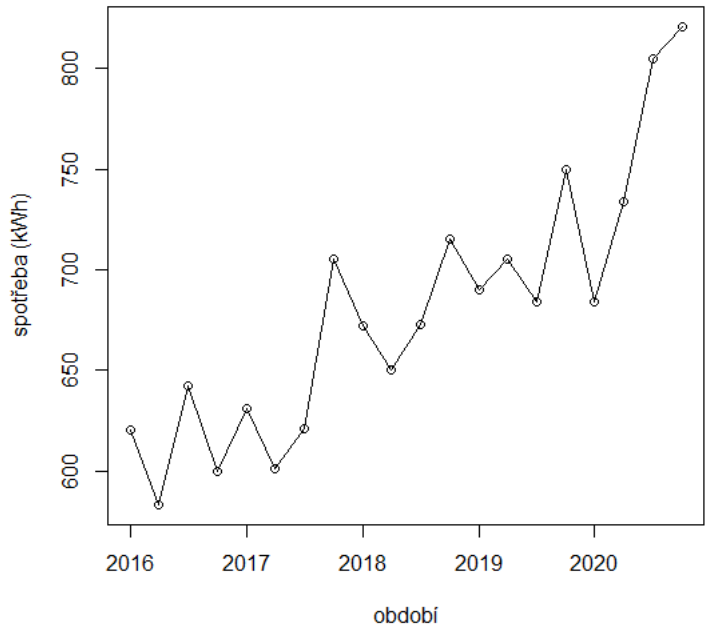
Řešený příklad

Časová řada (viz. tabulka) udává roční spotřebu elektriky (kWh) domácnosti v jednotlivých čtvrletích roku 2016-2020.

Rok	1.čtvrtletí	2.čtvrtletí	3.čtvrtletí	4.čtvrtletí
2016	620	583	642	600
2017	631	601	621	705
2018	672	650	673	715
2019	690	705	684	750
2020	684	734	805	821

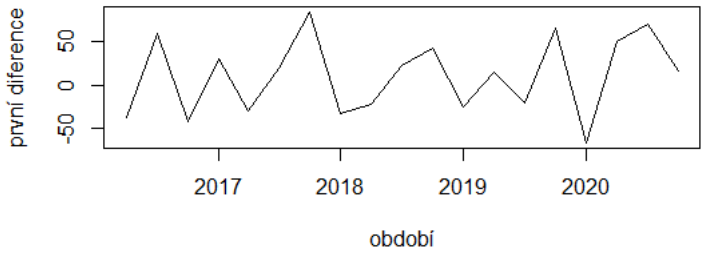
Určete základní míry dynamiky pro tuto časovou řadu.

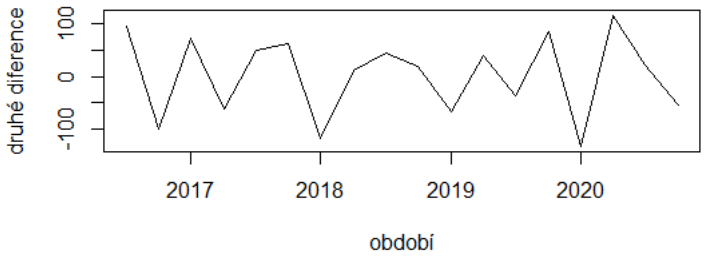
```
> y<-c
   (620,583,642,600,631,601,621,705,672,650,673,715,690,705,684,750,

684,734,805,821)
> spotreba<-ts(y,frequency=4,start=c(2016,1))
> spotreba
     Qtr1 Qtr2 Qtr3 Qtr4
2016  620  583  642  600
2017  631  601  621  705
2018  672  650  673  715
2019  690  705  684  750
2020  684  734  805  821
> plot(spotreba, type="o", xlab="období",ylab="spotřeba (kWh)")
```



Absolutní přírustky a druhé diference:

```
> ap<-diff(spotreba)
> ap
     Qtr1 Qtr2 Qtr3 Qtr4
2016       -37   59  -42
2017   31  -30   20   84
2018  -33  -22   23   42
2019  -25   15  -21   66
2020  -66   50   71   16
> plot(ap, xlab="období",ylab="první defierence")
> dif2<-diff(ap)
> dif2
     Qtr1 Qtr2 Qtr3 Qtr4
2016              96 -101
2017   73  -61   50   64
2018 -117   11   45   19
2019  -67   40  -36   87
2020 -132  116   21  -55
> par(mfrow=c(2,1))
> plot(ap, xlab="období",ylab="první diference")
> plot(dif2, xlab="období",ylab="druhé diference")
```





Průměrný absolutní přírustek

$$\bar{\Delta} = \frac{821 - 620}{20 - 1} = 10.58$$

```
> mean(ap)
[1] 10.57895
```

Lze tedy říci, že spotřeba elektřiny rostla čtvrtletně v průměru o 10.6 kWh.

Průměrný koeficient růstu

$$\bar{k} = \sqrt[19]{\frac{821}{620}} \doteq 1.01$$

Vidíme, že spotřeba elektřiny rostla v průměru o 1.01 %.

9.3 Dekompozice časových řad

Hodnoty časové řady lze rozložit do čtyř základních složek.

- T_t - **trendová složka**
- S_t - **sezónní složka**
- C_t - **cyklická složka**
- e_t - **reziduální složka**

Existují dva modely chování časové řady: aditivní a multiplikativní. V případě, že se hodnota rozptylu nemění v čase, jedná se o **model aditivní** a dekompozice lze zapsat ve tvaru:

$$y_t = T_t + S_t + C_t + e_t.$$

V opačném případě, tzn. variabilita hodnot řady se mění v čase výrazně, využíváme **model multiplikativní**:

$$y_t = T_t \cdot S_t \cdot C_t \cdot e_t.$$

Po vypuštění některých složek z modelu, vznikají různé modifikace.

9.3.1 Trendová složka

Trend vyjadřuje nějaký dlouhodobý směr vývoje sledovaného ukazatele v čase. Analýza trendové složky je jednou z nejdůležitějších částí analýzy.

Pro sezónní očištění dat se využívá metody **klouzavých průměrů** - prosté, centrované. V případě potřeby odstranění sezónního vlivu využijeme klouzavé průměry s délkou odpovídající délce sezónního období. Hlavní nevýhodou použití klouzavých průměrů je nevyrovnání koncové části časové řady.

Prosté klouzavé průměry

Je-li m liché číslo, získáme jejich hodnoty jako aritmetické průměry dané délky.

$$\bar{y_t} = \frac{1}{m} \sum_{i=-p}^{p} y_{t+i} = \frac{y_{t-p} + \cdots + y_{t+p}}{m},\ t = p+1, \ldots, n-p$$

p hodnot na začátku a na konci zůstane nevyrovnáno. Střední body klouzavých částí jsou celá čísla t.

Centrované klouzavé průměry

Používá se v případě, že délka klouzavé části je sudé číslo $m = 2p$. Cetrované klouzavé průměry jsou ve tvaru:

$$y_t = \frac{1}{4p}(y_{t-p} + 2y_{t-p+1} + \cdots + 2y_{t+p-1} + y_{t+p}).$$

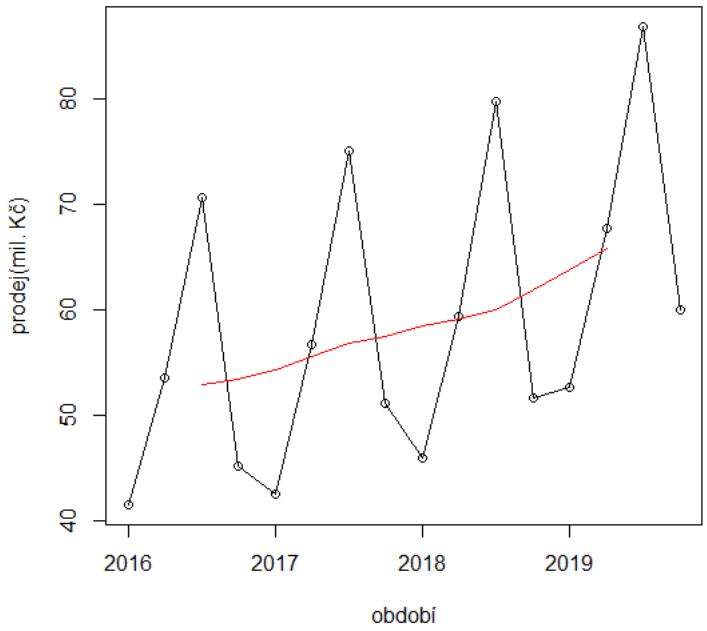
Řešený příklad

Časová řada (viz. tabulka) udává výsledky prodeje za čtvrtletí v období 2016-2019.

Rok	Čtvrtletí	Prodej (mil. Kč)
2016	I.	41.5
	II.	53.6
	III.	70.7
	IV.	45.2
2017	I.	42.6
	II.	56.7
	III.	75.1
	IV.	51.1
2018	I.	45.9
	II.	59.4
	III.	79.8
	IV.	51.6
2019	I.	52.7
	II.	67.7
	III.	86.9
	IV.	60

Vyrovnejte časovou řadu klouzavými průměry.

```
> install.packages("forecast")
> library(forecast)
> y<-c
   (41.5,53.6,70.7,45.2,42.6,56.7,75.1,51.1,45.9,59.4,79.8,51.6,
     52.7,67.7,86.9,60)
> prodej<-ts(y,frequency=4,start=c(2016,1))
> trend_prodej=ma(prodej,order=4,centre = T)
> trend_prodej
        Qtr1    Qtr2    Qtr3    Qtr4
2016      NA      NA 52.8875 53.4125
2017 54.3500 55.6375 56.7875 57.5375
2018 58.4625 59.1125 60.0250 61.9125
2019 63.8375 65.7750      NA      NA
> plot(prodej, type="o", xlab="období",ylab="prodej(mil. Kč)")
> lines(trend_prodej,col="red")
```



Trendové funkce

Analytické vyrovnávání trendu matematickou křivkou patří mezi neadaptivní metody. Na problém analýzy trendu se díváme jako na speciální případ regresní závislosti (nezávislá proměnná je čas).

Nejčastěji používané trendové funkce:

- lineární trend
- polynomický trend
- exponenciální trend
- modifikovaný exponenciální trend
- logistický trend
- Gomperzova křivka

Řešený příklad

Časovou řadu pro prodej z minulého příkladu vyrovnejte postupně lineární a kvadra-

tickou trendovou funkcí.

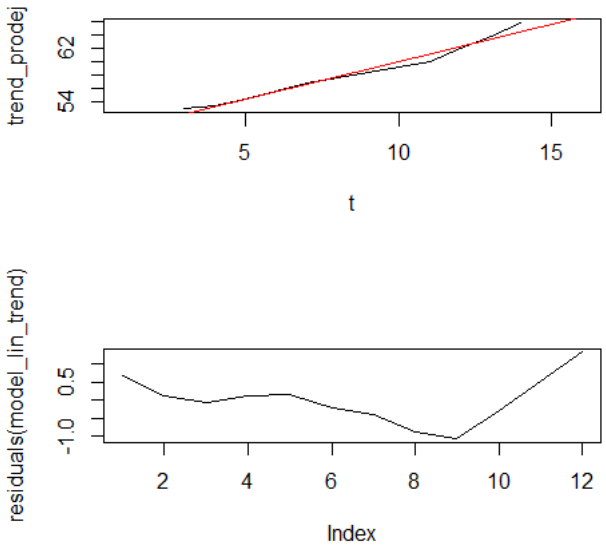
```
> t<-c(1:length(prodej))
> model_lin_trend<-lm(trend_prodej~t)
> summary(model_lin_trend)

Call:
lm(formula = trend_prodej ~ t)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06935 -0.32264  0.02176  0.23854  1.34119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.84964    0.53153   91.90 5.69e-16 ***
t            1.11316    0.05794   19.21 3.18e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6928 on 10 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.9736,    Adjusted R-squared:  0.971
F-statistic: 369.1 on 1 and 10 DF,  p-value: 3.176e-09
```

Z celkových statistik je vidět rovnice lineárního trendu

$$y_t = 48.85 + 1.11 \cdot t,$$

koeficient determinace je roven 97.36 %, což značí uspokojující výsledek modelace. Do grafu očištěných dat od sezonnosti vložíme zobrazíme nalezený lineární model a zobrazíme i graf reziduí.

```
> split.screen(c(2,1))
[1] 1 2
> screen(1)
> plot(t,trend_prodej,type = "l")
> abline(48.84964,1.11316,col="red")
> screen(2)
> plot(residuals(model_lin_trend),type="l")
```



Je vidět, že na začátku a hlavně na konci náš odhad podhodnocuje skutečnost. Zkusíme aplikovat kvadratický trend.

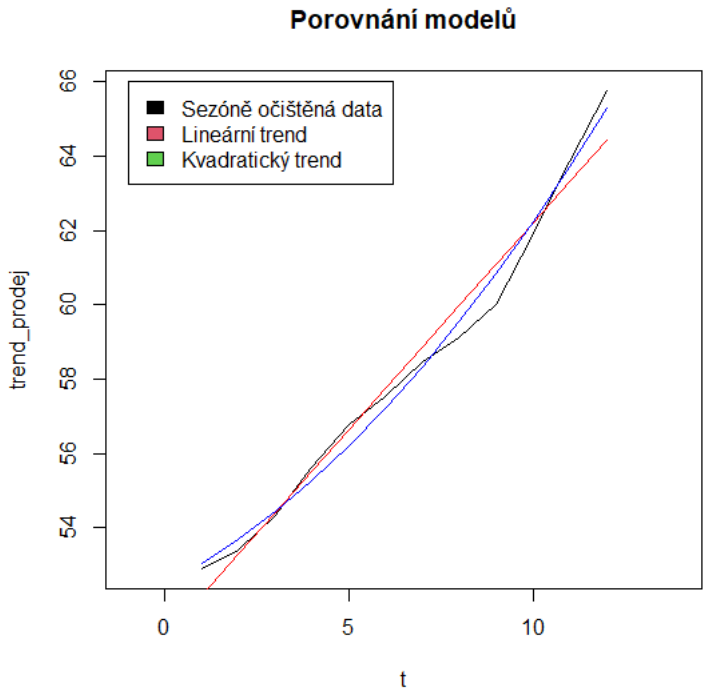
```
> t_2 <- t^2
> model_kvad_trend<-lm(trend_prodej~t+t_2)
> summary(model_kvad_trend)

Call:
lm(formula = trend_prodej ~ t + t_2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80528 -0.28620  0.02378  0.33797  0.59624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.66119    0.83706  61.717 3.88e-13 ***
t            0.32095    0.21733   1.477  0.17384
t_2          0.04660    0.01258   3.704  0.00489 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4597 on 9 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.9896,    Adjusted R-squared:  0.9872
F-statistic: 426.1 on 2 and 9 DF,  p-value: 1.219e-09
```

Kvadratický trend je ve tvaru

$$y_t = 51.66 + 0.32 \cdot t + 0.05 \cdot t^2,$$

je ale vidět, že koeficient u t je statisticky nevýznamný a lze ho z modelu vypustit. Modifikovaný koeficient determinace je u kvadratického modelu nepatrně vyšší než u lineárního. Vytvoříme graf s porovnáním obou modelů.

```
> predict_lin<-predict(model_lin_trend)
> predict_kvad<-predict(model_kvad_trend)
> plot(t-2,trend_prodej,type = "l",main = "Porovnání modelů",xlab
   ="t")
> lines(predict_lin,col="red")
> lines(predict_kvad,col="blue")
> legend(-1,66,legend = c("Sezóně očištěná data","Lineární trend
   ","Kvadratický trend"),col=c("black","red","green"),fill=1:3)
```




```
> plot(residuals(model_lin_trend),type="l",col="red",main = "
   Rezidua")
> lines(residuals(model_kvad_trend),col="green")
> legend(4,1.2,legend = c("Lineární trend","Kvadratický trend"),
   col=c("red","green"),lty=1:1)
```

9.3.2 Sezonní složka

Sezónnost můžeme vyjádřit tak, že od původních dat odečteme trendovou složku a nebudeme uvažovat vliv reziduální složky:

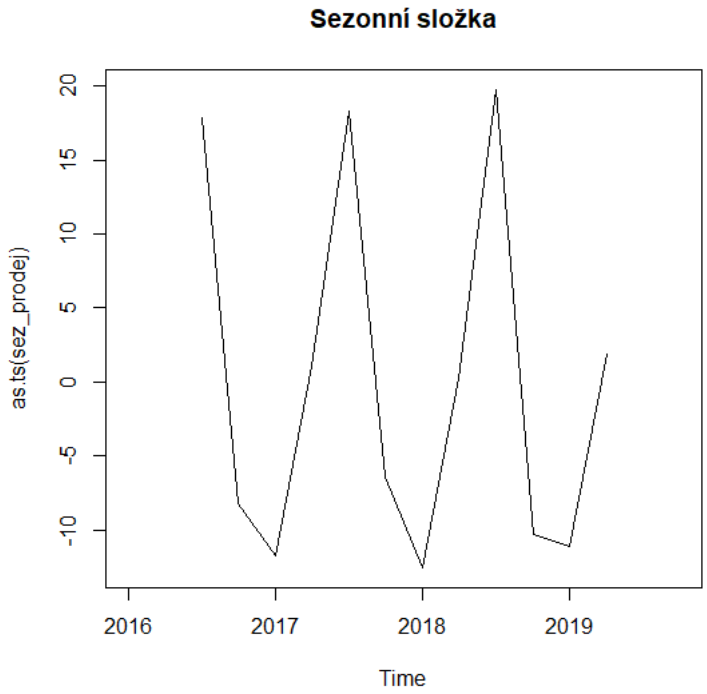
$$S_t = y_t - T_t.$$

Řešený příklad

Pro časovou řadu pro prodej proveďte analýzu sezonního kolísání v jednotlivých čtvrtletích.

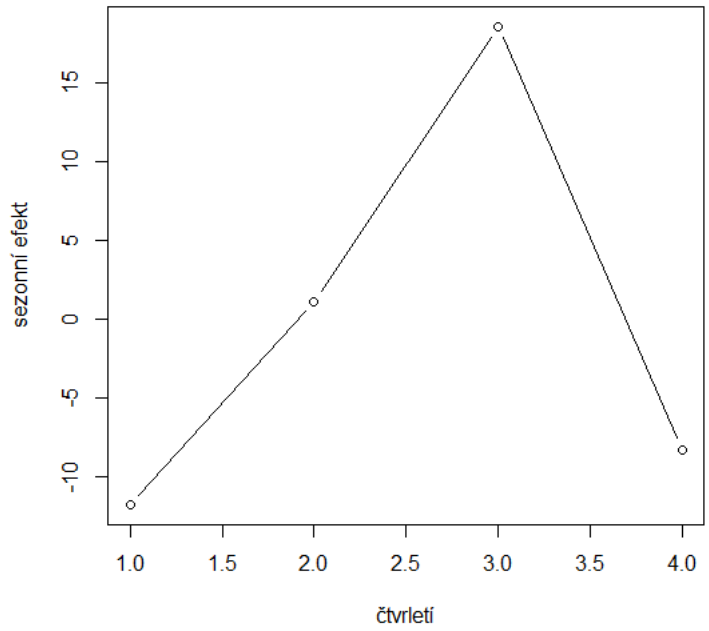
```
> sez_prodej<-prodej-trend_prodej
> sez_prodej
         Qtr1     Qtr2     Qtr3     Qtr4
2016       NA       NA  17.8125  -8.2125
2017 -11.7500   1.0625  18.3125  -6.4375
2018 -12.5625   0.2875  19.7750 -10.3125
2019 -11.1375   1.9250       NA       NA
> plot(as.ts(sez_prodej),main = "Sezonní složka")
```

Zobrazíme graf sezonní složky.



Provedeme sezonní dekompozici časové řady. Určíme průměrnou sezonnost, tj. sezonní hodnoty rozdělíme na období a pro každé období vypočteme průměrnou hodnotu.

```
> index<-seq(1,16,by=4)-1
> mm<-numeric(4)
> for(i in 1:4) {mm[i] <- mean(sez_prodej[index + i],na.rm =TRUE)}
> mm
[1] -11.816667   1.091667  18.633333  -8.320833
> plot.ts(mm,ylab="sezonní efekt", xlab="čtvrletí", cex=1, type="b
   ")
```



Nakonec vytvoříme sezonní složku pro celou časovou řadu.

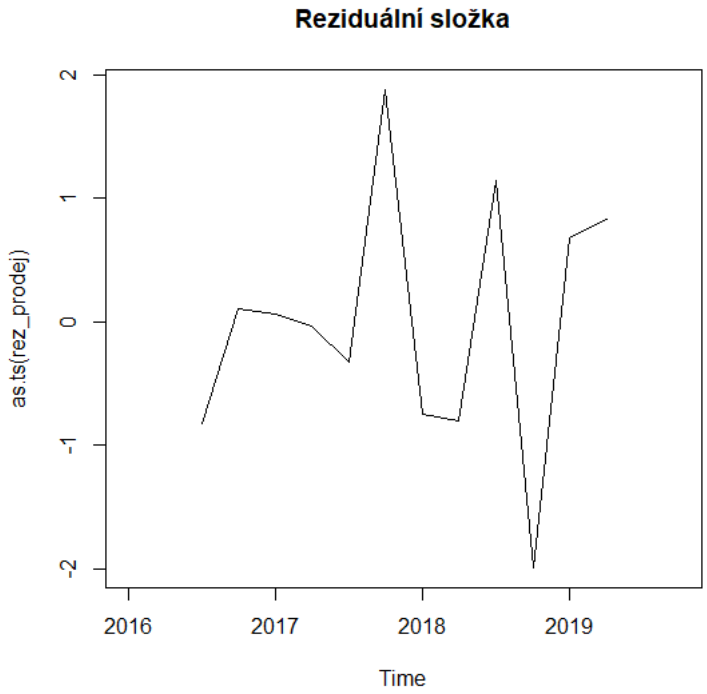
```
> sez_prodej_ts<- ts(rep(mm,4+1)[seq(16)],start=start(sez_prodej),
   frequency=4)
> sez_prodej_ts
           Qtr1       Qtr2       Qtr3       Qtr4
2016 -11.816667   1.091667  18.633333  -8.320833
2017 -11.816667   1.091667  18.633333  -8.320833
2018 -11.816667   1.091667  18.633333  -8.320833
2019 -11.816667   1.091667  18.633333  -8.320833
```

9.3.3 Reziduální složka

Po odstranění trendové a sezonní složky zůstane reziduální složka, která je tvořena náhodnými vlivy:

$$e_t = y_t - T_t - S_t.$$

```
> rez_prodej<-prodej-trend_prodej-sez_prodej_ts
> rez_prodej
            Qtr1         Qtr2         Qtr3         Qtr4
2016          NA           NA -0.82083333  0.10833333
2017  0.06666667 -0.02916667 -0.32083333  1.88333333
2018 -0.74583333 -0.80416667  1.14166667 -1.99166667
2019  0.67916667  0.83333333           NA          NA
```

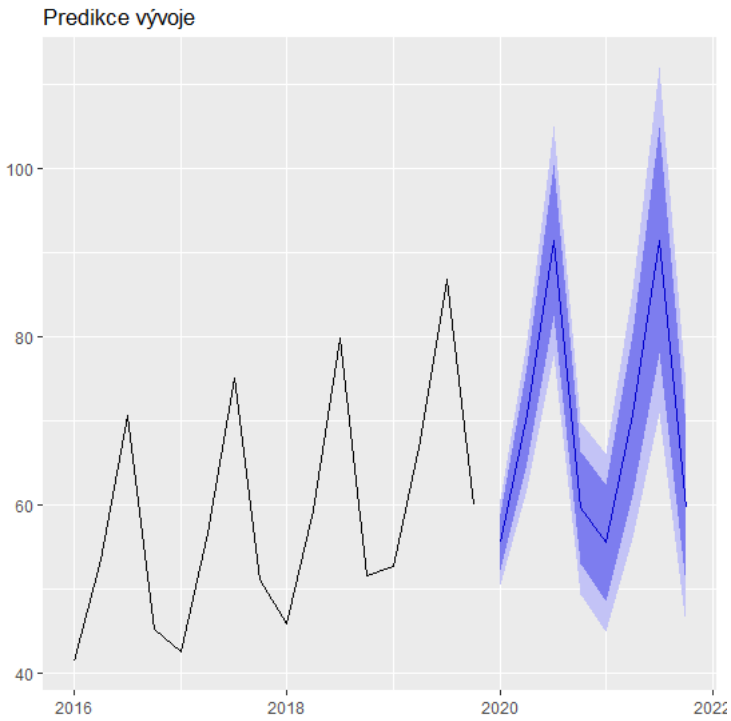


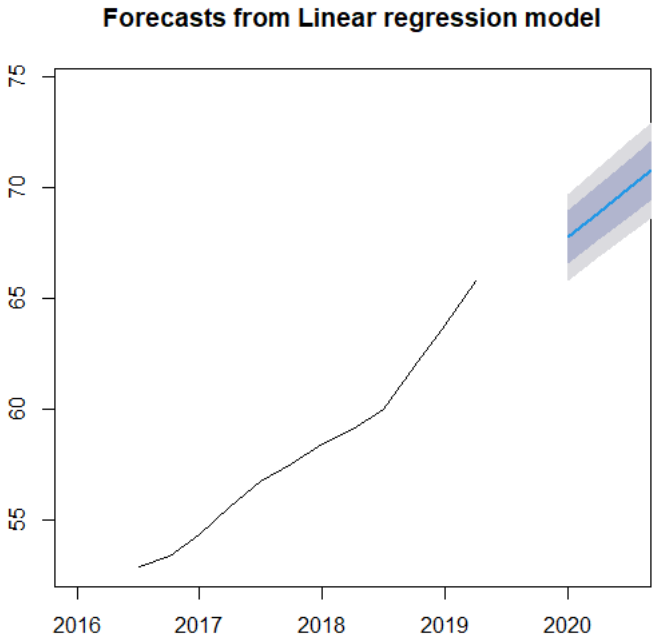
9.3.4 Predikce

Jedním důvodem pro dekompozici je snaha predikovat budoucí vývoj. Odhad budoucích hodnot můžeme provést pomocí dosazením hodnoty do získané rovnice trendu časové řady, kdy známe jednotlivé parametry trendové funkce a průměrné sezonní hodnoty pro jednotlivé čtvrtletí. Lze i využít funkci `predict()`, která odhad provede na základě předchozích dat (odhad je na základě klouzavých průměrů a ne trendové funkce).

```
> predict(prodej)
        Point Forecast    Lo 80     Hi 80    Lo 95     Hi 95
2020 Q1       55.55248 52.21616  58.88880 50.45001  60.65495
2020 Q2       70.79063 65.04767  76.53358 62.00754  79.57372
2020 Q3       91.50692 82.55918 100.45466 77.82253 105.19131
2020 Q4       59.69564 53.00902  66.38226 49.46934  69.92195
2021 Q1       55.55249 48.62778  62.47720 44.96207  66.14292
2021 Q2       70.79064 61.15319  80.42809 56.05144  85.52985
2021 Q3       91.50694 78.07826 104.93561 70.96955 112.04432
2021 Q4       59.69565 50.34368  69.04763 45.39305  73.99826
> autoplot(predict(prodej),xlab="",ylab="",main="Predikce vývoje")
> prodej_pred<-ts(trend_prodej,frequency=4, start=c(2016,1))
> prodej_pred
        Qtr1    Qtr2    Qtr3    Qtr4
2016      NA      NA 52.8875 53.4125
2017 54.3500 55.6375 56.7875 57.5375
2018 58.4625 59.1125 60.0250 61.9125
2019 63.8375 65.7750      NA      NA
> fit<-tslm(prodej_pred~trend)
> nsf<-forecast(fit,h=4)
```

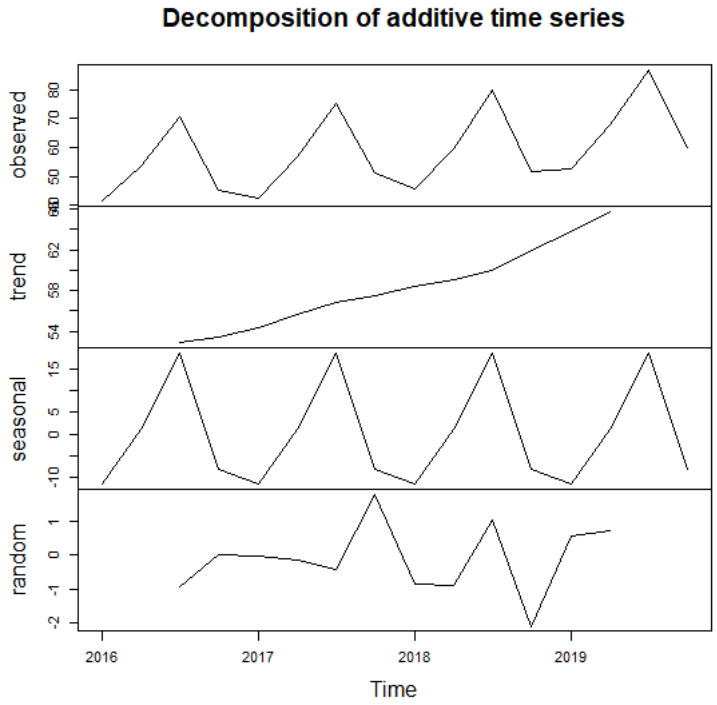
```
> nsf
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2020 Q1         67.77328 66.57505 68.97151 65.82760 69.71896
2020 Q2         68.88644 67.64164 70.13124 66.86514 70.90773
2020 Q3         69.99959 68.70501 71.29417 67.89747 72.10171
2020 Q4         71.11275 69.76554 72.45996 68.92516 73.30034
> plot(forecast(fit,h=5))
```





V programu R lze použít funkci `decompose()`, která provede dekompozici pro všechny složky najednou. Pomocí centrovaného klouzavého průměru sezoně očistí řadu, určí průměr sezonní složky a reziduální hodnotu. Před použitím příkazu je potřeba nadefinovat periodu modelované časové řady. Tato délka je nutná pro sezonní očištění.

```
> cas_rada<-ts(prodej,frequency=4)
> dekom_prodej=decompose(cas_rada,"additive")
> plot(dekom_prodej)
```



Literatura

[1] DVOŘÁKOVÁ, Stanislava. *Statistická analýza a časové řady v příkladech.* 1. vyd. Vysoká škola polytechnická, Jihlava, 2015, 83 s., ISBN 978-80-88064-18-3.

[2] HRNČÍŘ, Tomáš. *Časové řady v jazyce R.* Univerzita Pardubice, 2018, 96 s., diplomová práce, dostupná z https://dk.upce.cz/handle/10195/70676?show=full

[3] LITSCHMANNOVÁ, Martina. *Úvod do statistiky.* VŠB-TU Ostrava, 2011, 379 s., dostupné z https://mi21.vsb.cz/sites/mi21.vsb.cz/files/unit/uvod_do_statistiky.pdf

[4] STUCHLÝ, Jaroslav. *Statistika.* 2. vyd. Vysoká škola technická a ekonomická, České Budějovice, 2017, 258 s., ISBN 978-80-7468-021-2.

[5] SCHREIBEROVÁ, Petra a kol. *Matematika III: Pracovní listy.* 1. vyd. VŠB-TU Ostrava, 2016, 237 s., ISBN 978-80-248-3875-5.

[6] ANDĚL, Jiří. *Statistické metody.* Matfyzpress Praha, 1998, 274 s., ISBN 80-85863-27-8.

[7] CALDA, Emil a Václav DUPAČ. *Matematika pro gymnázia - Kombinatorika, pravděpodobnost a statistika.* Prometheus Praha, 1994, ISBN 80-85849-10-0.

[8] DUPAČ, Václav a Marie HUŠKOVÁ. *Pravděpodobnost a matematická statistika.* Karolinum, Praha, 1999.

[9] HAJKR, Oldřich, Pavel HRADECKÝ, Anna MADRYOVÁ a Matěj TURČAN. *Teorie statistiky.* Ostrava: VŠB - Technická univerzita, 1988, 267 s.

[10] HEBÁK, Petr a Jiří HUSTOPECKÝ. *Průvodce moderními statistickými metodami.* SNTL Praha, 1990, 293 s., ISBN 80-03-00534-5.

[11] HEBÁK, Petr a Jana KAHOUNOVÁ. *Počet pravděpodobnosti v příkladech.* Informatorium, Praha, 2005, ISBN 80-733-040-7.

[12] MAREK, Luboš. *Statistika v příkladech.* 1. vyd. Professional Publishing, Praha, 2013, 404 s., ISBN 978-80-7431-118-5.

Název: Statistické zpracování dat v energetice

Fakulta, katedra: Fakulta strojní, Katedra matematiky a deskriptivní geometrie

Autor: Schreiberová Petra

Místo, rok vydání: Ostrava, 2022

Počet stran: 93

Nakladatel: Vysoká škola báňská – Technická univerzita Ostrava

ISBN 978-80-248-4651-4