

# Popisná statistika v R

**Matematická statistika** se zabývá získáváním informací z empirických dat.

Součástími matematické statistiky jsou:

- teorie odhadu: určuje odhady neznámých parametrů základního souboru na základě empirických dat získaných z výběrového souboru,
- testování statistických hypotéz: ověřuje hypotézy o základním souboru,
- statistická predikce: snaží se o kvalifikovaný odhad budoucího vývoje sledovaných veličin na základě jejich dosavadního vývoje.

V tomto článku se pokusíme ukázat základní nástroje popisné statistiky v jazyce a prostředí R.

Pro správnou funkci kódu v tomto článku bude potřeba nainstalovat balíček *psych*:

```
# balíčky pro tuto kapitolu
library(psych)
```

## Základní pojmy

*Statistický soubor* je množina statistických jednotek stejného typu a shodného vymezení. Rozlišujeme:

- základní soubor (populace),
- výběrový soubor (výběr, vzorek).

*Statistická jednotka* je prvek *statistického souboru*, který při statistickém zkoumání sledujeme. Statistickou jednotku vymezujeme:

- věcně (co to je?),
- časově (kdy to sledujeme?) a
- prostorově (kde to sledujeme?).

*Statistický znak* je vlastnost, kterou u statistických jednotek sledujeme. Mohou být:

- číselné (měřitelné), které se ještě dělí na:
  - intervalové a
  - poměrové (kardinální),
- slovní (kategorické), které rozdělujeme na:
  - pořadové (ordinální) a
  - jmenovité (nominální) - alternativní (jen dvě možnosti) či množné.

U statistického znaku se můžeme zabývat četnostmi.

*Absolutní četnost* je absolutní počet výskytů daného znaku v dané hodnotě (či intervalu).

```
# u diskretních hodnot je to jednoduché
table(mtcars$cyl)

# u spojitých hodnot musíme kouzlit
table(cut(mtcars$qsec,
          breaks = 5))
```

*Relativní četnost* určuje podíl dané hodnoty (nebo intervalu) vůči celé sadě naměřených hodnot.

```
# u diskretních hodnot je to jednoduché
prop.table(table(mtcars$cyl))

# u spojitých hodnot musíme zase trochu kouzlit
prop.table(table(cut(mtcars$qsec,
                     breaks = 5)))
```

*Kumulativní četnosti* (absolutní i relativní) udávají součet všech pozorování, která nepřekračují určitou hodnotu.

```
# ukázka absolutních kumulativních četností
cumsum(table(mtcars$cyl))
```

## Základní zobrazení dat

### Diagram rozptýlení

Zejména spojité hodnoty můžeme zobrazit na jednorozměrném *diagramu rozptýlení*.

```
# použijeme uhrn srážek v amerických městech za rok
stripchart(x = precip,
           dlab = "Diagram rozptýlení (množství srážek v palcích za rok)")
```

## Polygon četností

Absolutní četnosti názorně ukáže *polygon četností*.

```
# použijeme počet válců aut v seznamu
plot(x = table(mtcars$cyl),
     type = "b",
     xlab = "počet válců",
     ylab = "absolutní četnost")
```

## Součtová křivka

Kumulativní četnosti zobrazí *součtová křivka*.

```
# použijeme počet velkých objevů za rok podle World Almanac
plot(x = cumsum(table(discoveries)),
     type = "b",
     xlab = "počet objevů za rok",
     ylab = "kumulativní četnost")
```

## Krabicový graf

*Krabicový graf* znázorňuje nejmenší a největší hodnotu, dolní a horní kvartil a medián.

```
boxplot(x = precip,
        ylab = "srážky v palcích za rok")
```

## Empirická distribuční funkce

Pomocí *empirické distribuční funkce* lze také dobře znázornit rozdělení četností.

```
plot(x = ecdf(precip),
     main = "",
     xlab = "srážky v palcích za rok",
     ylab = "distribuční funkce (četnost)")
```

## Histogram

*Histogram* zobrazí distribuci spojitých dat pomocí sloupců stejné šířky, které znázorňují intervaly.

```
hist(x = precip,
     main = "",
     xlab = "srážky v palcích za rok",
     ylab = "četnost (počet měst)")
```

Podobně znázorní *kumulativní (součtový) histogram* kumulativní četnosti.

```
# nejprve vyrobím obyčejný histogram
qsec.hist <- hist(discoveries)

# pak zamením četnosti kumulativními četnostmi
qsec.hist$counts <- cumsum(qsec.hist$counts)

# a nakonec vygeneruji kumulativní histogram pomocí plot()
plot(x = qsec.hist,
     main = "",
     xlab = "počet objevů",
     ylab = "kumulativní četnost")
```

## Charakteristiky polohy

### Průměr

*Aritmetický průměr* vyjadřuje průměrnou hodnotu daného znaku v souboru, počítá se jako podíl součtu hodnot a počtu pozorování.

```
mean(mtcars$cyl)
```

*Harmonický průměr* je podíl počtu pozorování ku součtu převrácených hodnot. Má smysl tam, kde počítáme s převrácenými hodnotami (hustoty, rychlosti apod.).

```
psych::harmonic.mean(1:10)
```

*Geometrický průměr* se počítá jako n-tá odmocnina součinu hodnot. Má smysl u analýz časových řad, např. při počítání průměrného tempa růstu nebo poklesu.

```
psych::geometric.mean(1:10)
```

*Kvadratický průměr* je odmocnina z průměru kvadratických hodnot.

```
sqrt(mean((1:10)^2))
```

## Modus

*Modus* je hodnota znaku s největší četností.

```
# s modem je to slozite, ale napr. takto
# - vytvorim tabulku cetnosti
# - seradim ji od nejvetsiho
# - a vezmu "nazev" nejcetnejsi hodnoty na prvni miste
# !!! CAVE: idealni je vytvorit reseni na miru situaci,
#         tohle reseni napr. ignoruje vicemodalni rozdeleni
names(sort(-table(c(1, 2, 1, 3))))[1]
```

## Kvantil

*Kvantil* je  $x(p)$  hodnota znaku, pro kterou platí, že podíl  $p$  hodnot uspořádaných v řadě má hodnotu menší nebo rovnou  $x(p)$  a podíl  $(1-p)$  hodnot má hodnotu větší nebo rovnou  $x(p)$ . Takových hodnot stejně jako u mediánu může být víc, záleží na konkrétním vzorci.

```
# 0,75 kvantil z rady 1-10
quantile(1:10, 0.75)
```

*Medián* je 50% kvantil, dělí řadu vzestupně seřazených výsledků na dvě stejně početné poloviny.

```
median(1:10)
```

## Charakteristiky variability

### Variační šíře

*Variační šíře (rozpětí)* je rozdíl mezi nejmenší a největší hodnotou.

```
# variacni rozpeti
sada <- 1:10
max(sada) - min(sada)

# nebo muzeme najit nejmensi a nejvetsi hodnotu jinak
range(sada)

# a spocist rozdil
diff(range(sada))
```

### Kvantilová rozpětí

Podobně fungují *kvantilová rozpětí*. Např.:

- *kvartilové rozpětí*  $x(0,75) - x(0,25)$ ,
- *decilové rozpětí*  $x(0,90) - x(0,10)$ ,
- *percentilové rozpětí*  $x(0,99) - x(0,01)$ .

```
# kvartilove rozpeti na ukazku
diff(quantile(1:10, c(0.75, 0.25)))

# pomoci funkce IQR
IQR(1:10)
```

Poloviční hodnoty uvedených rozpětí se nazývají *odchylky*:

- *kvartilová odchylka*,
- *decilová odchylka*,
- *percentilová odchylka*.

### Průměrná odchylka

*Průměrná odchylka* je aritmetický průměr absolutních odchylek hodnot znaku od jejich aritmetického průměru.

```
x <- 1:10
mean(abs(x - mean(x)))
```

## Rozptyl

*Rozptyl* je aritmetický průměr druhých mocnin odchylek hodnot znaku od jejich aritmetického průměru.

```
var(rnorm(100))
```

## Směrodatná odchylka

Odmocnina z rozptylu je *směrodatná odchylka*.

```
sd(rnorm(100))
```

## Variační koeficient

*Variační koeficient* je podíl směrodatné odchylky a aritmetického průměru.

```
sd(rivers)/mean(rivers)
```

## Charakteristiky koncentrace

### Koeficient šikmosti

*Koeficient šikmosti* udává tvar rozdělení četností. Pokud je kladný, je křivka zešikmená doleva (tj. delší je ocas vpravo). Pokud je záporný, je rozdělení četností zešikmené doprava (a má delší ocas vlevo).

```
# histogram
hist(x = rivers,
     main = "",
     xlab = "delka reky v milich",
     ylab = "cetnost")

# sikmost
psych::skew(rivers)
```

### Koeficient špičatosti

*Koeficient špičatosti* udává špičatost křivky rozdělení četností. Pokud je kladný, je křivka špičatější.

```
x <- rnorm(100)

# histogram
hist(x = x,
     main = "",
     xlab = "hodnoty",
     ylab = "cetnost")

# spicatost
psych::kurtosi(x)
```

## Odkazy

### Použitá literatura

- OLDŘICH, Neubauer. *Základy statistiky*. - vydání. Grada Publishing a.s., 2012. 236 s. ISBN 9788024742731.

### Použité balíčky R

- REVELLE, W. *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, Illinois, USA, 2018. Verze 1.8.10. <https://CRAN.R-project.org/package=psych>