Základy statistiky a analýzy dat v R

17VSADR – Skriptování a analýza dat v jazyce R

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova v Praze

[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

Obsah

Míry polohy a variability

- založena na funkcích `mean()`, `median()`, `sd()`, `var()`, `summary()`
- s výhodou lze kombinovat s funkcí `apply()`

```
    apply(mtcars, 2, mean)     # průměr
    apply(mtcars, 2, median)   # medián
    apply(mtcars, 2, sd)       # směrodatná odchylka
    apply(mtcars, 2, var)      # rozptyl
    apply(mtcars, 2, summary)  # 6-number statistics

    lapply(                    # vše najednou
      list(
        "mean", "median", "sd", "var", "summary"
      ),
      function(x) apply(mtcars, 2, x)
    )
```

Korelace

- standardně nás zajímá Pearsonův korelační koeficient, Spearmanův korelační koeficient

```
library(MASS)
data(Animals)

cor(
  Animals$body, Animals$brain,
  method = "pearson"
)      # -0.00534

cor(
  Animals$body, Animals$brain,
  method = "spearman"
)      # 0.71630
```

Kontingenční tabulky

- pomocí funkce table()

```
# kontingenční tabulka
table(mtcars$cyl, mtcars$gear,
      dnn = list("cyl", "gear"))

     gear
cyl   3  4  5
  4   1  8  2
  6   2  4  1
  8  12  0  2

# chí-kvadrát test
chisq.test(table(mtcars$cyl, mtcars$gear))

# data: table(mtcars$cyl, mtcars$gear)
# X-squared = 18.036, df = 4,
# p-value = 0.001214
```

Explorativní analýza dat

- navzdory očekávání relativně nová disciplína
- anglicky Exploratory Data Analysis (EDA)
- založena na vzevrubných grafických náhledech, porovnáních

Anscombeův kvartet

- čtyři dvojice proměnných s podobnými popisnými statistikami

```
    summary(anscombe)   # popisné statistiky

    cor(                # korelace
      anscombe[, c(1:4)],
      anscombe[, c(5:8)]
    )
            y1         y2         y3         y4
x1   0.8164205  0.8162365  0.8162867 -0.3140467
x2   0.8164205  0.8162365  0.8162867 -0.3140467
x3   0.8164205  0.8162365  0.8162867 -0.3140467
x4  -0.5290927 -0.7184365 -0.3446610  0.8165214
                        # příslušná x_i a y_i
                        # mají podobné korelace
```

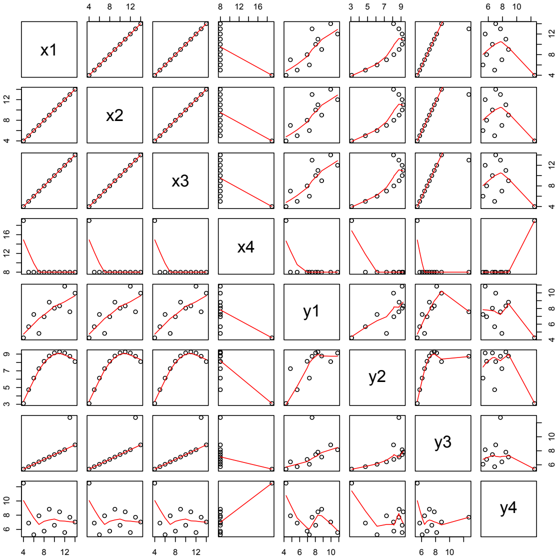
Anscombeův kvartet

- dvojice proměnných se tedy zdají podobné, ale ...

```
pairs(anscombe)

# eventuálně
pairs(anscombe, panel = panel.smooth)
```

Anscombeův kvartet



Alternativní rozdělení

- náhodná veličina X sleduje alternativní (Bernoulliho) rozdělení, tedy $X \sim \mathcal{A}(p)$ s parametrem $0 \leq p \leq 1$, pokud nabývá jen dvou hodnot 0 a 1 s pravděpodobnostmi $P(X = 0) = 1 - p$ a $P(X = 1) = p$
- pravděpodobnostní funkce je tedy

$$P_X(x) = P(X = x) = \begin{cases} p^x(1-p)^{1-x}, & x \in \{0, 1\} \\ 0, & \text{jinak} \end{cases}$$

- snadno nahlédneme, že

$$\mathbb{E}(X) = p$$
$$\text{var}(X) = p(1-p)$$

Alternativní rozdělení

- ať náhodná veličina $X \sim \text{alt}(\texttt{prob})$
- v R získáme pravděpodobnostní hustotu $f_X(x)$, distribuční funkci $F_X(q)$, kvantilovou funkci $Q_X(p)$ a náhodný generátor výběru z X pomocí

```
dbinom(x, size = 1, prob)
pbinom(q, size = 1, prob)
qbinom(p, size = 1, prob)

set.seed(1)
rbinom(n = 100, size = 1, prob = 0.78)
        # výběr o 100 pozorování
        # z alternativního rozdělení s p = 0,78
```

Příklad

- Student odpoví na otázku v testu bodovanou jedním bodem správně s pravděpodobností $p = 0{,}7$. Jaká je očekávaná střední hodnota a rozptyl počtu bodů, které student za takovou otázku získá?

Příklad

- Student odpoví na otázku v testu bodovanou jedním bodem správně s pravděpodobností $p = 0{,}7$. Jaká je očekávaná střední hodnota a rozptyl počtu bodů, které student za takovou otázku získá?
- *Řešení.*

$$\mathbb{E}(\text{počet bodů za otázku}) = p = 0{,}70$$
$$\text{var}(\text{počet bodů za otázku}) = p(1-p) = 0{,}21$$

$\square$

Binomické rozdělení

- předpokládejme, že v každém z $n \in \mathbb{N}$ nezávislých náhodných pokusů[1] může nastat úspěch s pravděpodobností p a neúspěch s pravděpodobností $1 - p$, kde $0 \leq p \leq 1$
- náhodná veličina X vrací počet úspěchů, které během n pokusů nastanou
- pak náhodná veličina X sleduje binomické rozdělení, tedy $X \sim \text{binom}(n, p)$ s parametry $n \in \mathbb{N}$ a $0 \leq p \leq 1$
- pravděpodobnostní funkce, tedy pravděpodobnost, že během n nezávislých náhodných pokusů nastane úspěch právě x-krát, je

$$P_X(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x \in \{0, 1, 2, \ldots\} \\ 0, & \text{jinak} \end{cases}$$

[1] všimněme si, že jednotlivé pokusy samy o sobě sledují alternativní rozdělení

Binomické rozdělení

- pokud náhodná veličina X sleduje binomické rozdělení, tedy $X \sim \text{binom}(n, p)$, můžeme odvodit, že

$$\begin{aligned} \mathbb{E}(X) &= np \\ \text{var}(X) &= np(1-p) \end{aligned}$$

Binomické rozdělení

- ať náhodná veličina $X \sim \text{binom}(\texttt{size}, \texttt{prob})$
- v R získáme pravděpodobnostní hustotu $f_X(x)$, distribuční funkci $F_X(q)$, kvantilovou funkci $Q_X(p)$ a náhodný generátor výběru z X pomocí

```
dbinom(x, size, prob)
pbinom(q, size, prob)
qbinom(p, size, prob)

set.seed(1)
rbinom(n = 10, size = 1, prob = 0.5)
            # výběr výsledků (0, 1) deseti opakování
            # hodu spravedlivou kostkou
```

Příklad

- Házíme desetkrát klasickou šestistěnnou kostkou. S jakou pravděpodobností padne právě čtyřikrát šestka? S jakou pravděpodobností padne nejvýše dvakrát?

Příklad

- Házíme desetkrát klasickou šestistěnnou kostkou. S jakou pravděpodobností padne právě čtyřikrát šestka? S jakou pravděpodobností padne nejvýše dvakrát?
- *Řešení.*

```
dbinom(
  x = 0:10, size = 10, prob = 1 / 6
)[5]   # 0.054

sum(
  dbinom(
    x = 0:10, size = 10, prob = 1 / 6
  )[1:3]
)      # 0.775
```

□

Poissonovo rozdělení

- ať náhodná veličina X sleduje původně binomické rozdělení, tedy $X \sim \text{binom}(n, p)$ s parametry $n \in \mathbb{N}$ a $0 \leq p \leq 1$
- pak je-li n dostatečně velké $(n > 30)$ a p dostatečně malé $(p < 0{,}1)$, lze binomické rozdělení aproximovat jednoparametrickým, Poissionovým rozdělením s $\lambda = np$, tedy $X \sim \text{Po}(\lambda)$
- náhodná veličina X sledující Poissonovo rozdělení vrací počet úspěchů, které nastanou během „velkého" počtu n pokusů, přičemž dílčí pravděpodobnost p nastání jednoho každého pokusu je „malá"
- pravděpodobnostní funkce, tedy pravděpodobnost, že během n nezávislých náhodných pokusů nastane úspěch právě x-krát, je

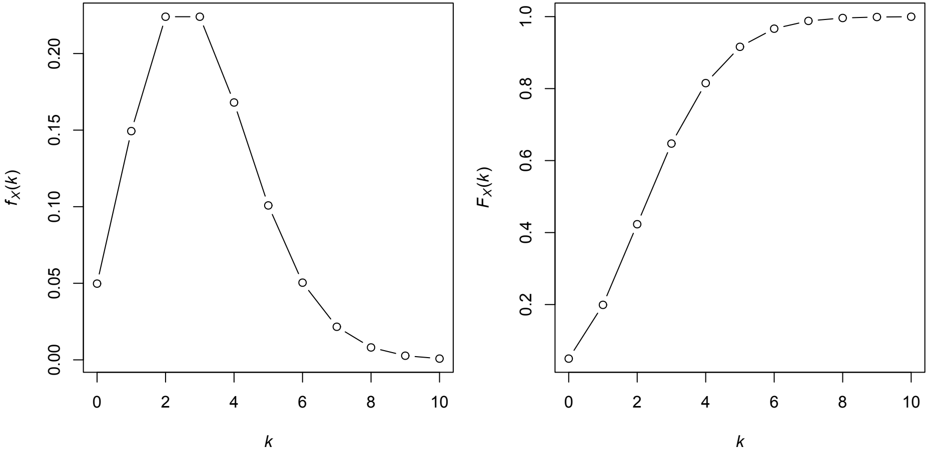
$$P_X(x) = P(X = x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x \in \{0, 1, 2, \ldots\} \\ 0, & \text{jinak} \end{cases}$$

Poissonovo rozdělení

- pokud náhodná veličina X sleduje Poissonovo rozdělení, tedy $X \sim \text{Po}(\lambda)$, můžeme odvodit, že

$$\mathbb{E}(X) = \lambda$$
$$\mathrm{var}(X) = \lambda$$

Poissonovo rozdělení pro $\lambda = 3$



Poissonovo rozdělení

- ať náhodná veličina $X \sim \text{Po}(\texttt{lambda})$
- v R získáme pravděpodobnostní hustotu $f_X(x)$, distribuční funkci $F_X(q)$, kvantilovou funkci $Q_X(p)$ a náhodný generátor výběru z X pomocí

```
        dpois(x, lambda)
        ppois(q, lambda)
        qpois(p, lambda)

        set.seed(1)
        rpois(n = 100, lambda = 3)
            # výběr o 100 pozorování
            # z Poissonova rozdělení o lambda = 3
```

Příklad

- Ve vybraném periodiku se objevují průměrně čtyři překlepy na každých jeho deset stránek. S jakou pravděpodobností bude na náhodně vybrané stránce periodika
  - (i) žádný překlep?
  - (ii) jeden překlep?
  - (iii) dva překlepy?
  - (iv) více než dva překlepy?

Příklad

- Ve vybraném periodiku se objevují průměrně čtyři překlepy na každých jeho deset stránek. S jakou pravděpodobností bude na náhodně vybrané stránce periodika
  (i) žádný překlep?
  (ii) jeden překlep?
  (iii) dva překlepy?
  (iv) více než dva překlepy?
- *Řešení.*

```
dpois(x = 0, lambda = 0.4)   # (i) 0.670

dpois(x = 1, lambda = 0.4)   # (ii) 0.268

dpois(x = 2, lambda = 0.4)   # (iii) 0.054

1 - sum(dpois(x = 0:2, lambda = 0.4))
                             # (iv) 0.008
```

Hypergeometrické rozdělení

- předpokládejme, že uvažujeme sestavu N prvků, kde M prvků z nich má určitou vlastnost a $N - M$ prvků tuto vlastnost nemá
- ze sestavy těchto prvků postupně vybereme n prvků bez vracení (!)
- náhodná veličina X, která vrací počet prvků s určitou vlastností mezi n prvky, sleduje hypergeometrické rozdělení, $X \sim \text{hy}(N, M, n)$
- pravděpodobnostní funkce, tedy pravděpodobnost, že mezi n náhodně bez vracení vybranými prvky bude x-krát prvek s určitou vlastností, je

$$P(X = x) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, & \max\{n - N + M, 0\} \leq x \leq \min\{M, n\} \\ 0, & \text{jinak} \end{cases}$$

Hypergeometrické rozdělení

- pokud náhodná veličina X sleduje hypergeometrické rozdělení, tedy $X \sim \text{hy}(N, M, n)$, můžeme odvodit, že

$$\mathbb{E}(X) = n\frac{M}{N}$$

$$\text{var}(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}$$

Hypergeometrické rozdělení

- ať náhodná veličina $X \sim \mathrm{hy}(\mathtt{m} + \mathtt{n}, \mathtt{m}, \mathtt{k})$
- v R získáme pravděpodobnostní hustotu $f_X(x)$, distribuční funkci $F_X(q)$, kvantilovou funkci $Q_X(p)$ a náhodný generátor výběru z X pomocí

```
        dhyper(x, m, n, k)
        phyper(q, m, n, k)
        qhyper(p, m, n, k)

        set.seed(1)
        rhyper(nn = 10, m = 3, n = 7, k = 5)
            # výsledek deseti pokusů:
            # v každém pokusu je na výstupu počet
            # vytažených bílých koulí z urny,
            # ve které jsou původně 3 bílé a 7 černých
            # koulí
```

Příklad

- V sérii po dvě stě kusech je deset zmetků. Při přejímce náhodně vybereme pět kusů a podrobíme je destrukční zkoušce. Pokud není mezi pěti vybranými ani jeden zmetek, sérii přijmeme. S jakou pravděpodobností to nastane?

Příklad

- V sérii po dvě stě kusech je deset zmetků. Při přejímce náhodně vybereme pět kusů a podrobíme je destrukční zkoušce. Pokud není mezi pěti vybranými ani jeden zmetek, sérii přijmeme. S jakou pravděpodobností to nastane?
- *Řešení.*

```
dhyper(x = 0, m = 10, n = 200 - 10, k = 5)
    # 0.772

# anebo též
choose(10, 0) * choose(190, 5) / choose(200, 5)
    # 0.772
```

□

Normální rozdělení

- náhodná veličina X sleduje normální rozdělení, tedy $X \sim \mathcal{N}(\mu, \sigma^2)$ o parametrech $\mu \in \mathbb{R}$ a $\sigma \in \mathbb{R} : \sigma \geq 0$, pokud pro její pravděpodobnostní hustotu $f_X(x)$ a pro libovolné $x \in \mathbb{R}$ platí

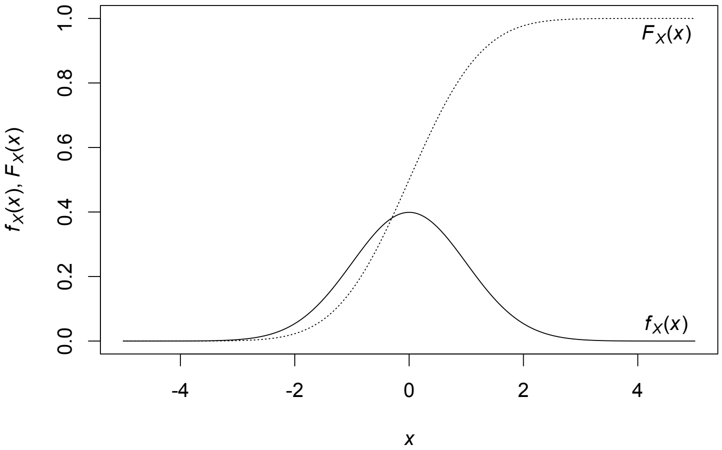
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- platí, že

$$\mathbb{E}(X) = \mu$$
$$\mathrm{var}(X) = \sigma^2$$

- náhodná veličina X sledující normální rozdělení obvykle vhodně popisuje fenomény založené na vzájemném působení mnoha drobných nezávislých, eventuálně navzájem se rušících vlivů
  - např. veličina X může popsat distribuci náhodných chyb, tělesné výšky dané populace, inteligenčního kvocientu, sytost barvy vlasů (šedotónové pixely) dané populace a mnoho dalšího

Pravděpodobnostní hustota a distribuční funkce normálního rozdělení



Distribuční funkce normálního rozdělení

- ať náhodná veličina X sleduje normální rozdělení, tedy $X \sim \mathcal{N}(\mu, \sigma^2)$ o parametrech $\mu \in \mathbb{R}$ a $\sigma \in \mathbb{R} : \sigma \geq 0$
- její distribuční funkce $F_X(x)$ má tvar

$$F_X(x) = \int_{-\infty}^{x} f_X(\tau) \mathrm{d}\tau = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\tau-\mu)^2}{2\sigma^2}} \mathrm{d}\tau$$

- takový výraz ale nelze běžně integrovat, navíc numerické řešení integrálu by bylo nutné tabelizovat pro všechny kombinace (x, μ, σ), což není možné, proto zavádíme odvozenou náhodnou veličinu $U \equiv \frac{X-\mu}{\sigma}$, která sleduje standardní normální rozdělení $U \sim \mathcal{N}(0, 1^2)$

Standardní normální rozdělení

- ať náhodná veličina X sleduje normální rozdělení, tedy $X \sim \mathcal{N}(\mu, \sigma^2)$ o parametrech $\mu \in \mathbb{R}$ a $\sigma \in \mathbb{R} : \sigma \geq 0$
- náhodná veličinu $U \equiv \frac{X-\mu}{\sigma}$, sleduje standardní normální rozdělení $U \sim \mathcal{N}(0, 1^2)$ s pravděpodobnostní hustotou

$$f_U(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

distribuční funkcí

$$F_U(x) = \int_{-\infty}^{x} f_U(\tau) \mathrm{d}\tau = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau^2}{2}} \mathrm{d}\tau$$

a charakteristikami

$$\mathbb{E}(U) = 0$$

$$\mathrm{var}(U) = 1$$

Distribuční funkce standardního normálního rozdělení

- ať náhodná veličinu $U \equiv \frac{X-\mu}{\sigma}$, sleduje standardní normální rozdělení $U \sim \mathcal{N}(0, 1^2)$ a má distribuční funkci

$$F_U(x) = \int_{-\infty}^{x} f_U(\tau)\mathrm{d}\tau = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau^2}{2}} \mathrm{d}\tau$$

- takový výraz je stále nepraktický pro rutinní výpočty, ale díky standardizaci již lze smysluplně tabelizovat
- běžně píšeme

$$F_U(x) = P(U \leq x) = \Phi(x)$$

  a pro malé hodnoty $x \geq 0$ hodnoty $\Phi(x)$ tabelizujeme
- díky sudosti pravděpodobnostní hustoty $f_U(x)$ je $P(U \leq x) = 1 - P(U \leq -x)$, lze pro hodnoty $x < 0$ využít vztah

$$\Phi(-x) = P(X \leq -x) = 1 - P(X \leq x) = 1 - \Phi(x)$$

Normální rozdělení

- ať náhodná veličina $X \sim \mathcal{N}(\mathtt{mean}, \mathtt{sd}^2)$
- v R získáme pravděpodobnostní hustotu $f_X(x)$, distribuční funkci $F_X(q)$, kvantilovou funkci $Q_X(p)$ a náhodný generátor výběru z X pomocí

```
dnorm(x, mean, sd)
pnorm(q, mean, sd)
qnorm(p, mean, sd)

set.seed(1)
rnorm(n = 100, mean = 0, sd = 1)
    # výběr o 100 pozorování
    # ze standardního normálního
    # rozdělení
```

Příklad

- Pomocí úvahy nebo tabulek distribuční funkce standardního normálního rozdělení
  (i) najděte $\Phi(0)$.
  (ii) najděte $x \in \mathbb{R}$ takové, aby $\Phi(x) = 0{,}5$.
  (iii) najděte $x \in \mathbb{R}$ takové, aby $\Phi(x) = \Phi(-x)$.
  (iv) najděte $\Phi(1{,}96)$.
  (v) najděte $\Phi(-1{,}96)$.
  (vi) najděte $x \in \mathbb{R}$ takové, aby $\Phi(x) = 0{,}025$.
  (vii) najděte $x \in \mathbb{R}$ takové, aby $\Phi(x) = 1$.
  (viii) najděte $x \in \mathbb{R}$ takové, aby $\Phi(x) \leq 2$.
  (ix) najděte $x \in \mathbb{R}$ takové, aby $\Phi(x) = 0$.
  (x) najděte $x \in \mathbb{R}$ takové, aby $\Phi(x) = +\infty$.
  (xi) najděte $\Phi(+\infty)$.
  (xii) najděte $\Phi(-\infty)$.
  (xiii) najděte $x \in \mathbb{R}$ takové, aby $\Phi(x) - \Phi(-x) = 0{,}5$.
  (xiv) najděte $x \in \mathbb{R}$ takové, aby $P(|U| \leq x) = 0{,}95$.

Příklad

- Náhodná veličina X sleduje normální rozdělení $\mathcal{N}(20, 16)$. Jaká je pravděpodobnost, že nabude hodnoty
  (i) menší než 16?
  (ii) větší než 20?
  (iii) v rozmezí mezi 12 a 28?
  (iv) menší než 12 nebo větší než 28?

Příklad

- Náhodná veličina X sleduje normální rozdělení $\mathcal{N}(20, 16)$. Jaká je pravděpodobnost, že nabude hodnoty
  (i) menší než 16?
  (ii) větší než 20?
  (iii) v rozmezí mezi 12 a 28?
  (iv) menší než 12 nebo větší než 28?
- *Řešení.*
  (i) $0{,}158$.
  (ii) $0{,}500$.
  (iii) $0{,}955$.
  (iv) $0{,}046$. □

Příklad

- Náhodná veličina X vrací hodnotu chyby měření určitým přístrojem a sleduje normální rozdělení se střední hodnotu $\mu = 0{,}20$ a rozptylem $\sigma^2 = 0{,}64$.
  - (i) Jaká je pravděpodobnost, že absolutní hodnota chyby měření bude menší než $1{,}0$?
  - (ii) Jaká je horní hranice chyby měření, které se může přístroj dopustit s pravděpodobností $0{,}95$?

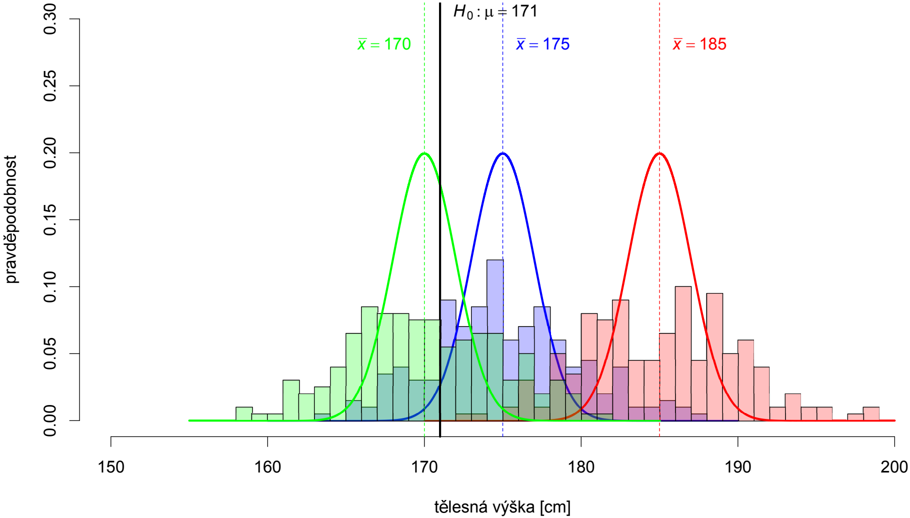
Příklad

- Náhodná veličina X vrací hodnotu chyby měření určitým přístrojem a sleduje normální rozdělení se střední hodnotu $\mu = 0{,}20$ a rozptylem $\sigma^2 = 0{,}64$.
  - (i) Jaká je pravděpodobnost, že absolutní hodnota chyby měření bude menší než $1{,}0$?
  - (ii) Jaká je horní hranice chyby měření, které se může přístroj dopustit s pravděpodobností $0{,}95$?
- *Řešení.*
  - (i) $0{,}775$.
  - (ii) $1{,}516$. □

Princip testování hypotéz

- je založen na definování tzv. nulové hypotézy, kterou lze eventuálně vyvrátit nalezením významného protipříkladu
- nulovou hypotézou může být např. tvrzení, že průměrná výška v populaci je 171 cm
- protipříkladem je ve statistice myšlen dostatečně velký soubor hodnot, které jsou dostatečně „v rozporu“ s nulovou hypotézou
- protipříkladem může být např. výběr sto lidí, kde je průměrná výška 175 cm a směrodatná odchylka 10 cm

Testy hypotéz



Hladina významnosti

hladina významnosti

předpokládejme, že nulová hypotéza platí; pak pravděpodobnost toho, že za její platnosti dostanu data, která jsem nasbíral, je nazývaná *hladina významnosti* či p-hodnota

- pokud platí nulová hypotéza, měla by být p-hodnota co největší

chyba prvního typu

předpokládejme, že nulová hypotéza platí; pokud ji zamítnu, dělám chybné rozhodnutí, a takové chybné rozhodnutí se nazývá *chyba prvního typu*

- p-hodnota udává pravděpodobnost chyby prvního typu, tedy pravděpodobnost chybného závěru; proto by měla být p-hodnota co nejmenší, pokud hodláme zamítnout nulovou hypotézu

Hladina významnosti

- je-li obvykle

$$\text{hladina významnosti} \equiv p\text{-hodnota} \leq 0{,}05$$

lze již nulovou hypotézu zamítnout (riziko chyby prvního typu je malé)

Testy normality

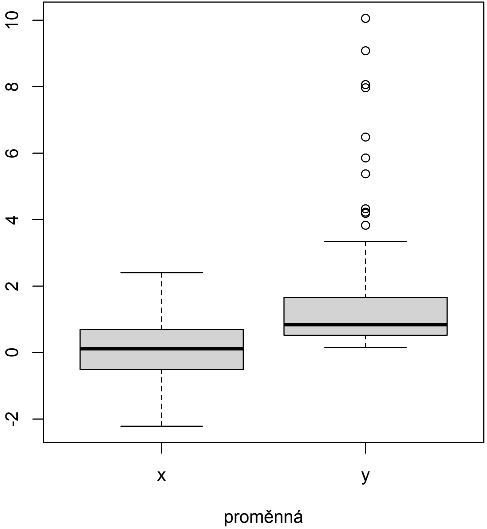
- existuje jich celá řada
- vždy testují nulovou hypotézu H_0 o normálním rozdělení zkoumaného souboru
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α
- jeden z nejpoužívanějších je Shapirův-Wilkův test

```r
set.seed(1)
x <- rnorm(100); y <- exp(rnorm(100))

shapiro.test(x)
# Shapiro-Wilk normality test
# W = 0.9956, p-value = 0.9876

shapiro.test(y)
# Shapiro-Wilk normality test
# W = 0.66135, p-value = 7.401e-14
```

Testy normality



Dvouvýběrový t-test

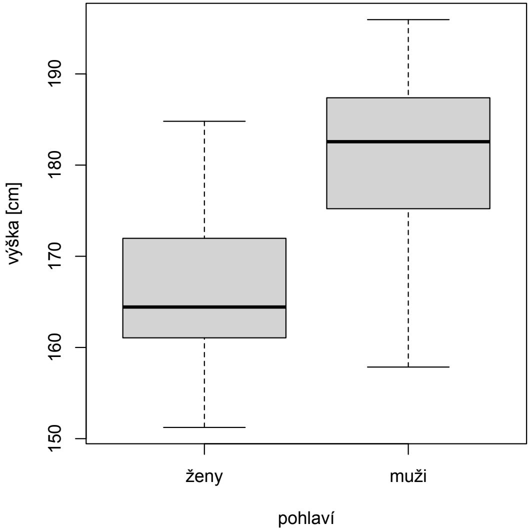
- testuje nulovou hypotézu $H_0 : \mu_1 = \mu_2$ o statisticky nevýznamném rozdílu ve středních hodnotách dvou výběrů
- předpokládá normalitu obou výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
set.seed(1)
muzi <- rnorm(30, 180, 10)
zeny <- rnorm(30, 165, 10)

t.test(muzi, zeny)
# Welch Two Sample t-test
# t = 6.5125, df = 56.741, p-value = 2.093e-08
# ...

t.test(muzi, zeny)$p.value   # 2.093108e-08
```

Dvouvýběrový t-test



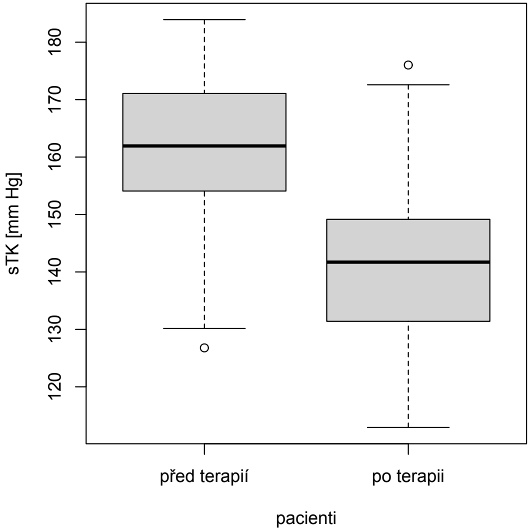
Párový t-test

- testuje nulovou hypotézu $H_0 : \mu_1 = \mu_2$ o statisticky nevýznamném rozdílu ve středních hodnotách jednoho výběru ve dvou situacích
- předpokládá normalitu obou výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
set.seed(1)
pacienti_pred <- rnorm(50, 160, 15)
pacienti_po <- rnorm(50, 140, 15)

t.test(
  pacienti_pred, pacienti_po, paired = TRUE
)
# Paired t-test
# t = 7.1546, df = 49, p-value = 3.823e-09
# ...
```

Párový t-test



F-test

- testuje nulovou hypotézu $H_0 : \sigma_1^2 = \sigma_2^2$ o statisticky nevýznamném rozdílu v rozptylech dvou výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
set.seed(1)
muzi <- rnorm(30, 180, 10)
zeny <- rnorm(30, 165, 10)

var.test(muzi, zeny)
# F test to compare two variances
# F = 1.3501, num df = 29, denom df = 29,
# p-value = 0.4238

var.test(muzi, zeny)$p.value   # 0.4237845
```

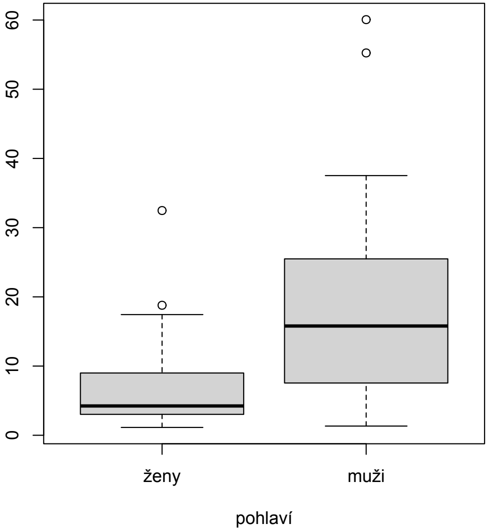
Wilcoxonův dvouvýběrový test

- testuje nulovou hypotézu $H_0 : \mu_1 = \mu_2$ o statisticky nevýznamném rozdílu ve středních hodnotách dvou výběrů
- **nepředpokládá** normalitu obou výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
set.seed(1)
muzi <- exp(rnorm(30, 2.5, 1))
zeny <- exp(rnorm(30, 1.5, 1))

wilcox.test(muzi, zeny)
# Wilcoxon rank sum test
# W = 710, p-value = 7.215e-05
```

Wilcoxonův dvouvýběrový test



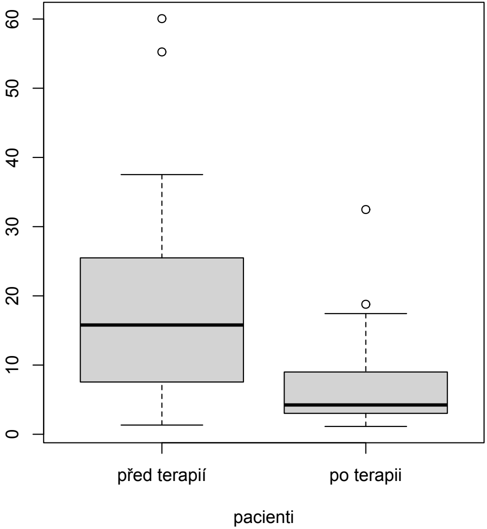
Wilcoxonův párový test

- testuje nulovou hypotézu $H_0 : \mu_1 = \mu_2$ o statisticky nevýznamném rozdílu ve středních hodnotách jednoho výběru ve dvou situacích
- **nepředpokládá** normalitu obou výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
set.seed(1)
pacienti_pred<- exp(rnorm(30, 2.5, 1))
pacienti_po <- exp(rnorm(30, 1.5, 1))

wilcox.test(
  pacienti_pred, pacienti_po, paired = TRUE
)
# Wilcoxon signed rank test
# V = 400, p-value = 0.0002833
# ...
```

Wilcoxonův párový test



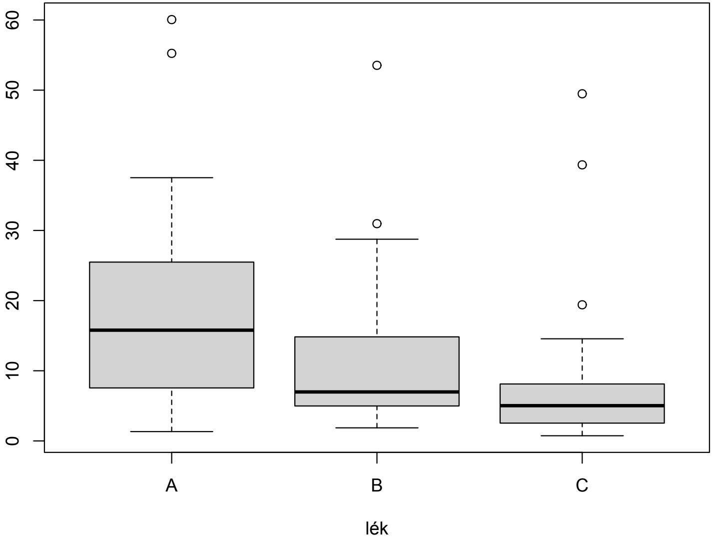
Kruskal-Wallisův test

- testuje nulovou hypotézu $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ o statisticky nevýznamném rozdílu ve středních hodnotách k výběrů
- **nepředpokládá** normalitu výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
set.seed(1)
lek_A <- exp(rnorm(30, 2.5, 1))
lek_B <- exp(rnorm(30, 2.0, 1))
lek_C <- exp(rnorm(30, 1.5, 1))

my_data <- data.frame(
  "mira" = c(lek_A, lek_B, lek_C),
  "lek" = c(rep("A", 30), rep("B", 30),
            rep("C", 30))
)
```

Kruskal-Wallisův test



Kruskal-Wallisův test

- výsledek testu

```
kruskal.test(mira ~ lek, my_data)
# Kruskal-Wallis rank sum test
# Kruskal-Wallis chi-squared = 16.945,
# df = 2, p-value = 0.0002092
```

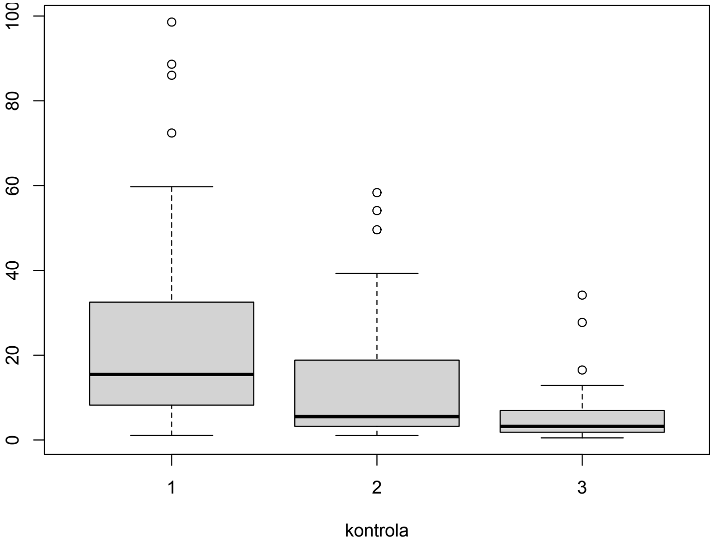
Friedmanův test

- testuje nulovou hypotézu $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ o statisticky nevýznamném rozdílu ve středních hodnotách jednoho výběru v k situacích
- **nepředpokládá** normalitu výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
set.seed(2)
cas_1 <- exp(rnorm(30, 2.5, 1))
cas_2 <- exp(rnorm(30, 2.0, 1))
cas_3 <- exp(rnorm(30, 1.5, 1))

friedman.test(cbind(cas_1, cas_2, cas_3))
# Friedman rank sum test
# Friedman chi-squared = 18.6, df = 2,
# p-value = 9.142e-05
```

Friedmanův test



χ^2 test nezávislosti

- χ^2-test nezávislosti testuje nulovou hypotézu H_0 o nezávislosti mezi řádky a sloupci kontingenční tabulky
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```r
chisq.test(
  matrix(c(
    12, 20, 30,
    18, 14, 10
  ), nrow = 2, byrow = T)
)
# Pearson's Chi-squared test
# X-squared = 8.7357, df = 2, p-value = 0.01268
```

χ^2 testy dobré shody

- χ^2-test dobré shody testuje nulovou hypotézu H_0 o statisticky nevýznamné odlišnosti mezi předpokládaným a testovaným rozdělením
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
chisq.test(
  c(10, 15, 14),
  p = c(1/3, 1/3, 1/3)
)
# Chi-squared test for given probabilities
# X-squared = 1.0769, df = 2, p-value = 0.5836
```

Implementace většiny metod pomocí Shiny

Odkaz

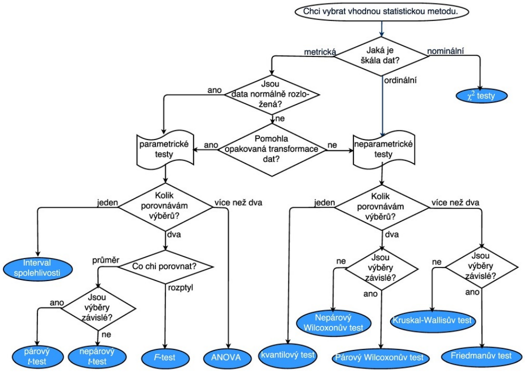
http://shiny.statest.cz:3838/statisticke_nastroje/

Statistické nástroje

Úvod | Nahrání dat | Výběr metody | Testování normality | *t*-testy | *F*-test | Wilcoxonovy testy | Friedmanův test | Kruskal-Wallisův test | χ^2 testy | ANOVA | O aplikaci

Vývojový diagram pro výběr statistické metody

Pomocí vývojového diagramu je na základě vložených dat a výzkumných hypotéz možné odhadnout, která statistická metoda nejlépe odpovídá výzkumnému záměru. Poté je možné přejít přímo k záložce, která nabízí aparát pro realizaci analýzy, a to pomocí tlačítek pod diagramem.



K **testování normality** | Ke **Kruskal-Wallis testu** | K **χ^2 testům**

K ***t*-testům** | K ***F*-testu** | K **ANOVA** | K **Wilcoxonovým testům** | K **Friedmanovu testu**

Statistické nástroje verze 1.0.0

Počet návštěv: 391

ČVUT FBMI

Statistické nástroje

Úvod | Nahrání dat | Výběr metody | Testování normality | *t*-testy | *F*-test | Wilcoxonovy testy | Friedmanův test | Kruskal-Wallisův test | χ^2 testy | ANOVA | O aplikaci

Parametry analýzy

☑ Zobrazit originální výstup z R?

Výsledky Friedmanova testu

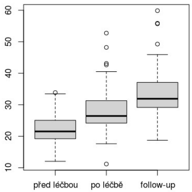
parametr	hodnota
Friedmanova statistika	95.280
počet stupňů volnosti	2
p-hodnota	< 0.00001

Originální výstup z R

```
	Friedman rank sum test

data:  as.matrix(my_data())
Friedman chi-squared = 95.28, df = 2, p-value < 2.2e-16
```

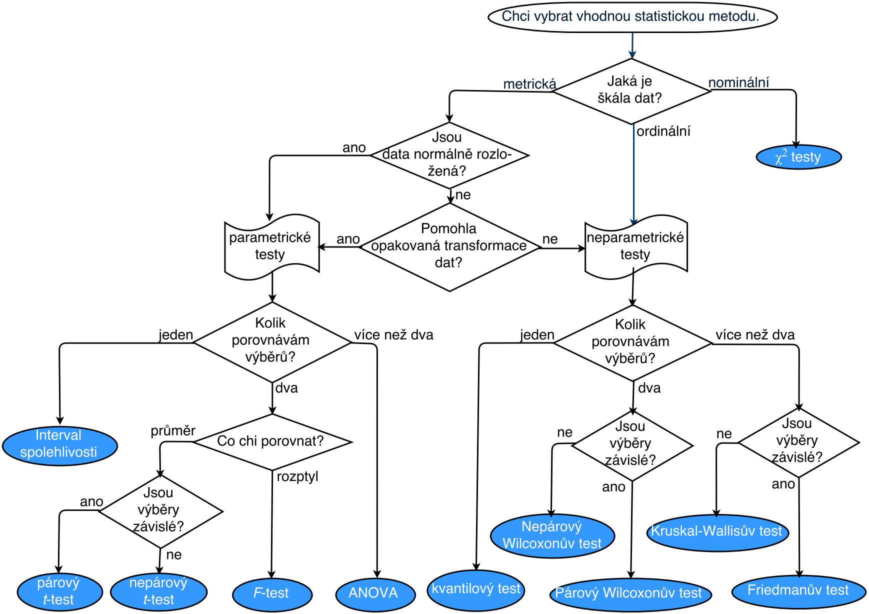
Diagram



Stáhni diagram!

Statistické nástroje verze 1.0.0

Počet návštěv: 391


Chci vybrat vhodnou statistickou metodu.
Jaká je škála dat?
metrická
nominální
ordinální
χ^2 testy
Jsou data normálně rozložená?
ano
ne
Pomohla opakovaná transformace dat?
ano
ne
parametrické testy
neparametrické testy
Kolik porovnávám výběrů?
jeden
dva
více než dva
Interval spolehlivosti
Co chi porovnat?
průměr
rozptyl
Jsou výběry závislé?
ano
ne
párový t-test
nepárový t-test
F-test
ANOVA
Kolik porovnávám výběrů?
jeden
dva
více než dva
Jsou výběry závislé?
ne
ano
Jsou výběry závislé?
ne
ano
kvantilový test
Nepárový Wilcoxonův test
Párový Wilcoxonův test
Kruskal-Wallisův test
Friedmanův test

Literatura

- Karel Zvára. *Základy statistiky v prostředí R*. Praha, Česká republika: Karolinum, 2013. ISBN: 978-80-246-2245-3.
- Hadley Wickham. *Advanced R*. Boca Raton, FL: CRC Press, 2015. ISBN: 978-1466586963.

Děkuji za pozornost!

lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz

GitHub

github.com/LStepanek/17VSADR_Skriptovani_a_analyza_dat_v_jazyce_R