# Základy statistiky v R

## Úvod do jazyka R

Lubomír Štěpánek[1, 2]

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

ČVUT FBMI

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

# Obsah

# Normální rozdělení

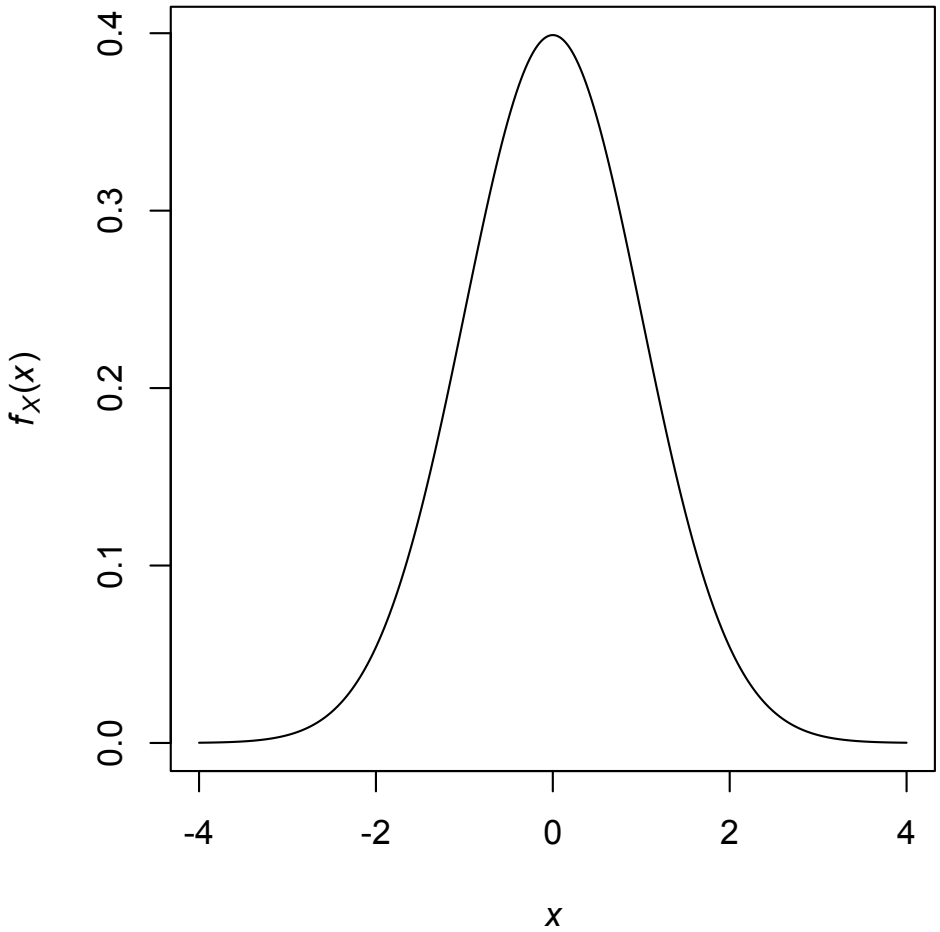
- náhodná veličina $X$ sleduje normální rozdělení $\mathcal{N}(\mu, \sigma^2)$
- tedy $X \sim \mathcal{N}(\mu, \sigma^2)$, pak

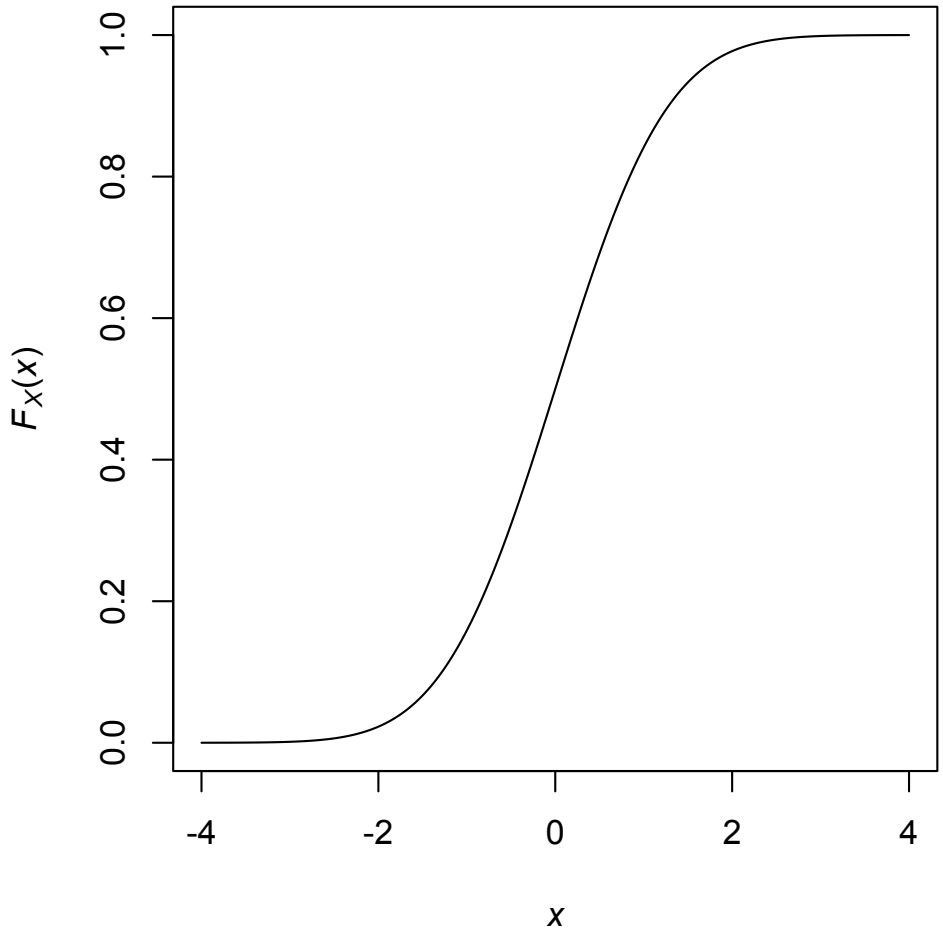
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- dále

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

# Normální rozdělení



# Normální rozdělení

- v R získáme pravděpodobnostní hustotu $f_X(x)$, distribuční funkci $F_X(x)$, kvantilovou funkci $Q_X(p)$ a náhodný generátor výběru z $X$ pomocí

```
dnorm(x = 1, mean = 0, sd = 1)     # 0.2419707
pnorm(q = 0, mean = 0, sd = 1)     # 0.5
qnorm(p = 0.5, mean = 0, sd = 1)  # 0

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

# Poissonovo rozdělení

- náhodná veličina $X$ sleduje normální rozdělení $Poiss(\lambda)$
- tedy $X \sim Poiss(\lambda)$, pak

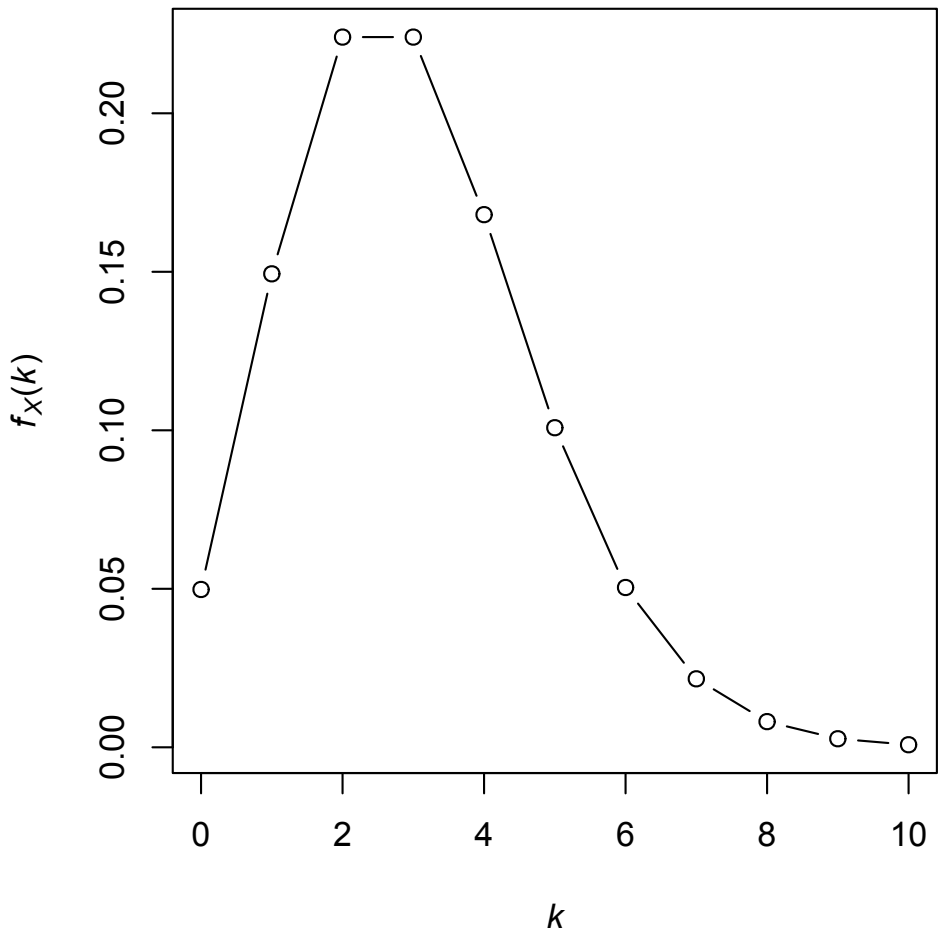
$$f_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

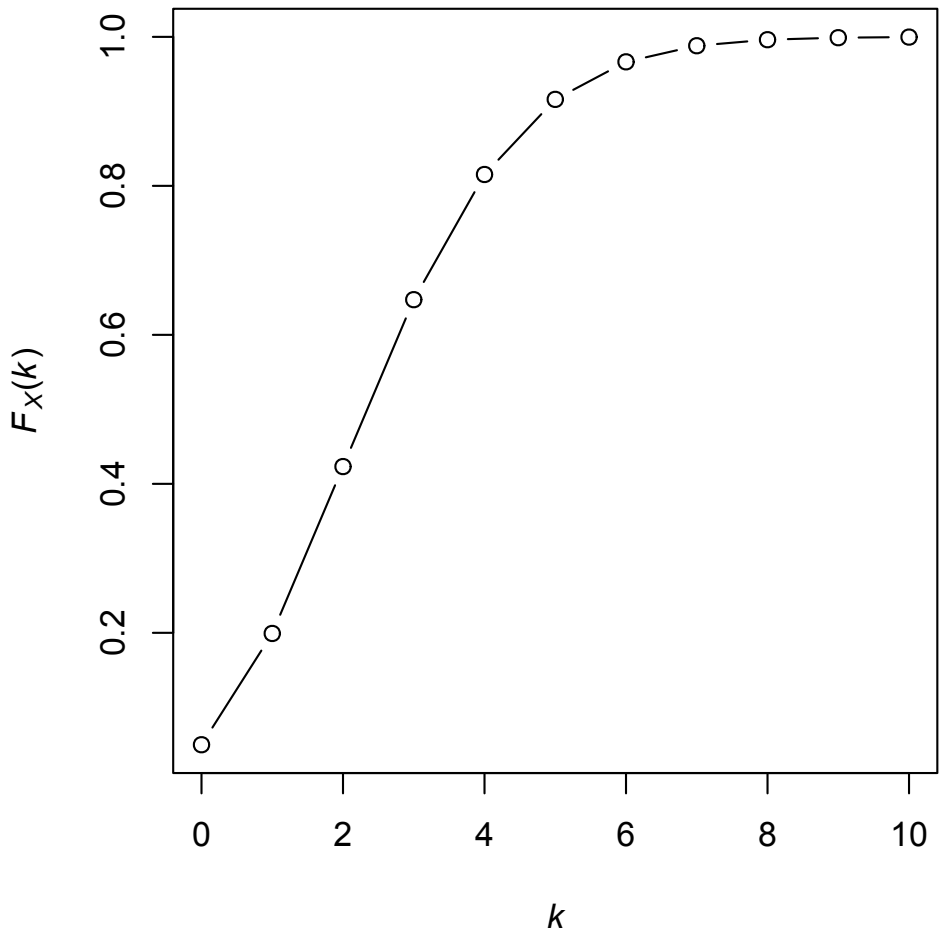
- dále

$$\mathbf{E}(X) = \lambda,$$

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

# Poissonovo rozdělení



# Poissonovo rozdělení

- v R získáme pravděpodobnostní hustotu $f_X(k)$, distribuční funkci $F_X(k)$, kvantilovou funkci $Q_X(p)$ a náhodný generátor výběru z $X$ pomocí

```
1  dpois(x = 3, lambda = 3)    # 0.2240418
2  ppois(q = 3, lambda = 3)    # 0.6472319
3  qpois(p = 0.5, lambda = 3)  # 3
4
5  set.seed(1)
6  rpois(n = 100, lambda = 3)
7       # výběr o 100 pozorování
8       # z Poissonova rozdělení o lambda = 3
```

# Míry polohy a variability

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

```
1  apply(mtcars, 2, mean)      # průměr
2  apply(mtcars, 2, median)    # medián
3  apply(mtcars, 2, sd)        # směrodatná odchylka
4  apply(mtcars, 2, var)       # rozptyl
5  apply(mtcars, 2, summary)   # 6-number statistics
6
7  lapply(                     # vše najednou
8    list(
9      "mean", "median", "sd", "var", "summary"
10   ),
11   function(x) apply(mtcars, 2, x)
12 )
```

# Korelace

- standardně nás zajímá Pearsonův korelační koeficient, Spearmanův korelační koeficient

```
1   library(MASS)
2   data(Animals)
3
4   cor(
5     Animals$body, Animals$brain,
6     method = "pearson"
7   )       # -0.00534
8
9   cor(
10    Animals$body, Animals$brain,
11    method = "spearman"
12  )       # 0.71630
```

# Kontingenční tabulky

- pomocí funkce `table()`

```
1   # kontingenční tabulka
2   table(mtcars$cyl, mtcars$gear)
3
4      3  4  5
5   4  1  8  2
6   6  2  4  1
7   8 12  0  2
8
9   # chí-kvadrát test
10  chisq.test(table(mtcars$cyl, mtcars$gear))
11
12  # data: table(mtcars$cyl, mtcars$gear)
13  # X-squared = 18.036, df = 4,
14  # p-value = 0.001214
```

# Exploratívní 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

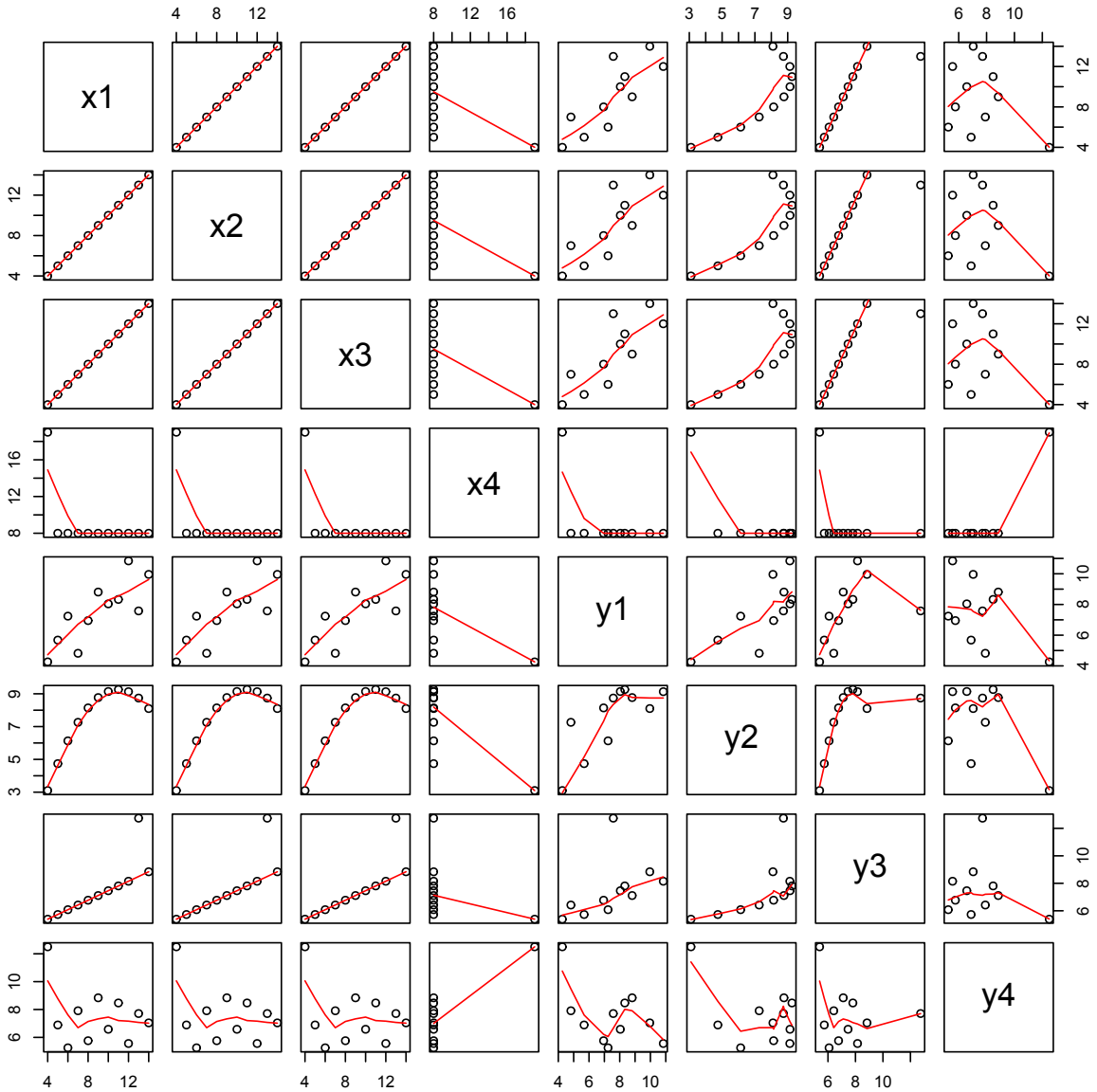
```
1   summary(anscombe)   # popisné statistiky
2
3   cor(                # korelace
4     anscombe[, c(1:4)],
5     anscombe[, c(5:8)]
6   )
7            y1          y2          y3          y4
8   x1  0.8164205  0.8162365  0.8162867 -0.3140467
9   x2  0.8164205  0.8162365  0.8162867 -0.3140467
10  x3  0.8164205  0.8162365  0.8162867 -0.3140467
11  x4 -0.5290927 -0.7184365 -0.3446610  0.8165214
12                 # příslušná x_i a y_i
13                 # mají podobné korelace
```

# Anscombeův kvartet

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

```
1  pairs(anscombe)
2
3  # eventuálně
4  pairs(anscombe, panel = panel.smooth)
```

# Anscombeův kvartet



# Implementace většiny metod pomocí Shiny

**Odkaz**

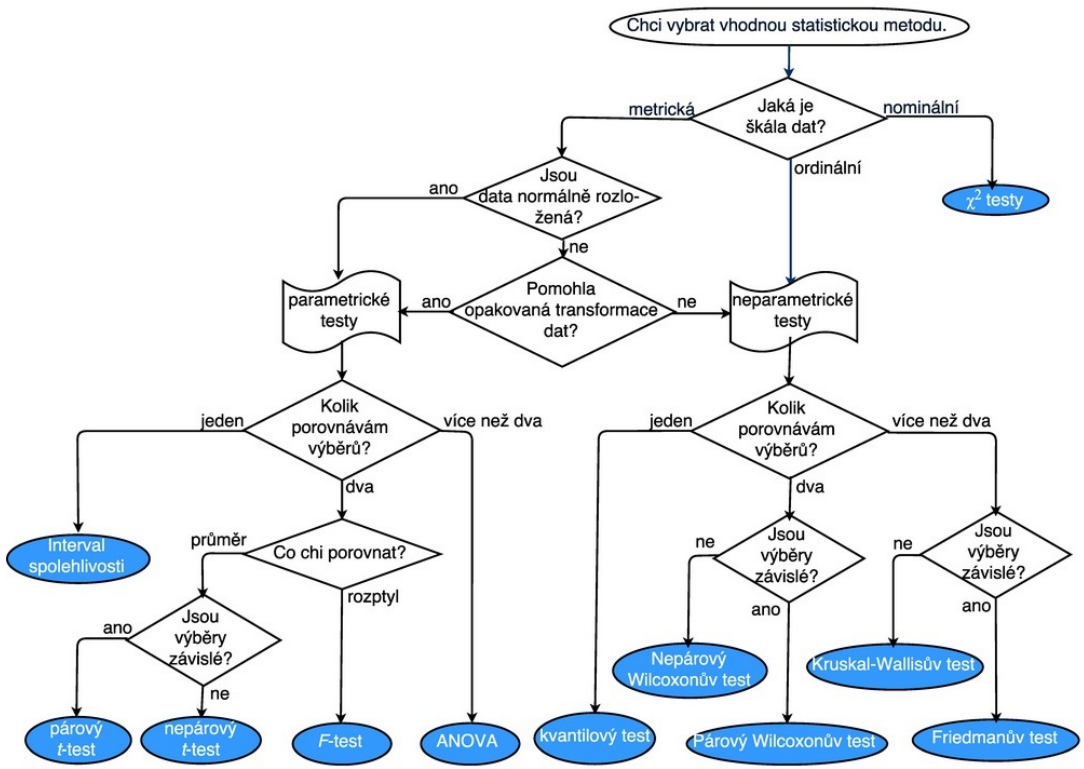
http://shiny.statest.cz:3838/statisticke_nastroje/

Statistické nástroje

Úvod | Nahrání dat | Výběr metody | Testování normality | *t*-testy | *F*-test | Wilcoxonovy testy | Friedmanův test | Kruskal-Wallisův test | $\chi^2$ 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 $\chi^2$ **testům**

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

Statistické nástroje verze 1.0.0



Počet návštěv: 391

Statistické nástroje

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

## Parametry analýzy

☑ Zobrazit originální výstup z R?

## Výsledky Friedmanova testu

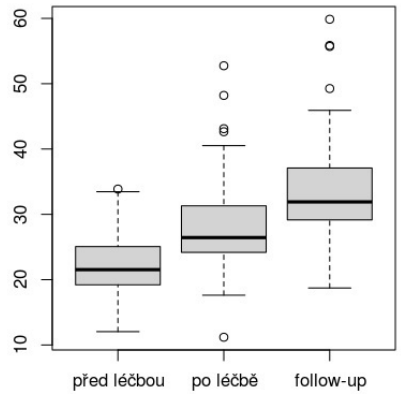
| parametr | hodnota |
|---|---|
| Friedmanova statistika | 95.280 |
| počet stupňů volnosti | 2 |
| p-hodnota | < 0.00001 |

## 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

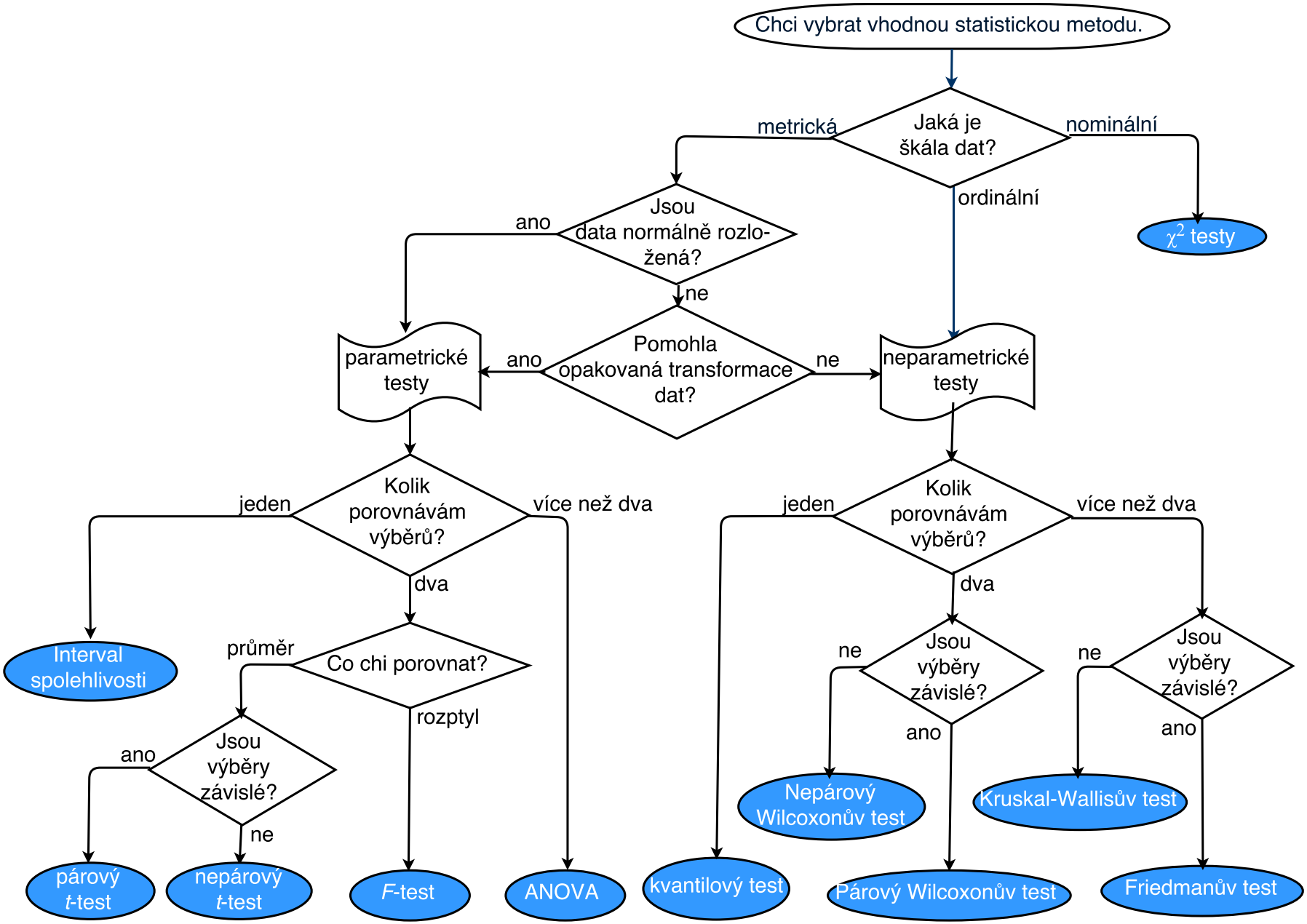



Stáhni diagram!

Statistické nástroje verze 1.0.0



Počet návštěv: 391

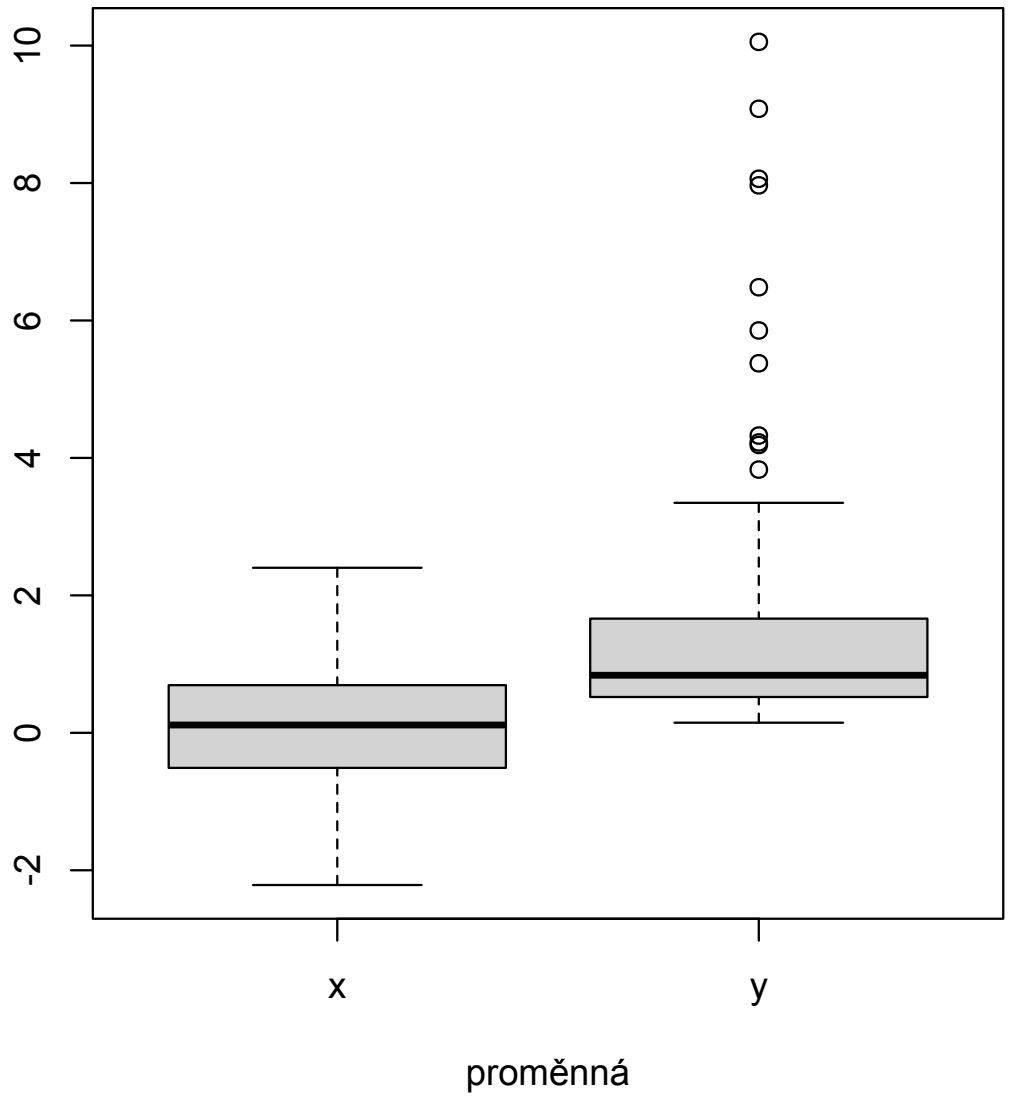

Chci vybrat vhodnou statistickou metodu.
Jaká je škála dat?
metrická
nominální
ordinální
Jsou data normálně rozložená?
ano
ne
$\chi^2$ testy
parametrické testy
Pomohla opakovaná transformace dat?
ano
ne
neparametrické testy
Kolik porovnávám výběrů?
jeden
dva
více než dva
Interval spolehlivosti
Co chi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

# 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 $\alpha$
- jeden z nejpoužívanějších je Shapirův-Wilkův test

```
1  set.seed(1)
2  x <- rnorm(100); y <- exp(rnorm(100))
3
4  shapiro.test(x)
5  # Shapiro-Wilk normality test
6  # W = 0.9956, p-value = 0.9876
7
8  shapiro.test(y)
9  # Shapiro-Wilk normality test
10 # W = 0.66135, p-value = 7.401e-14
```

# Dvouvýběrový t-test

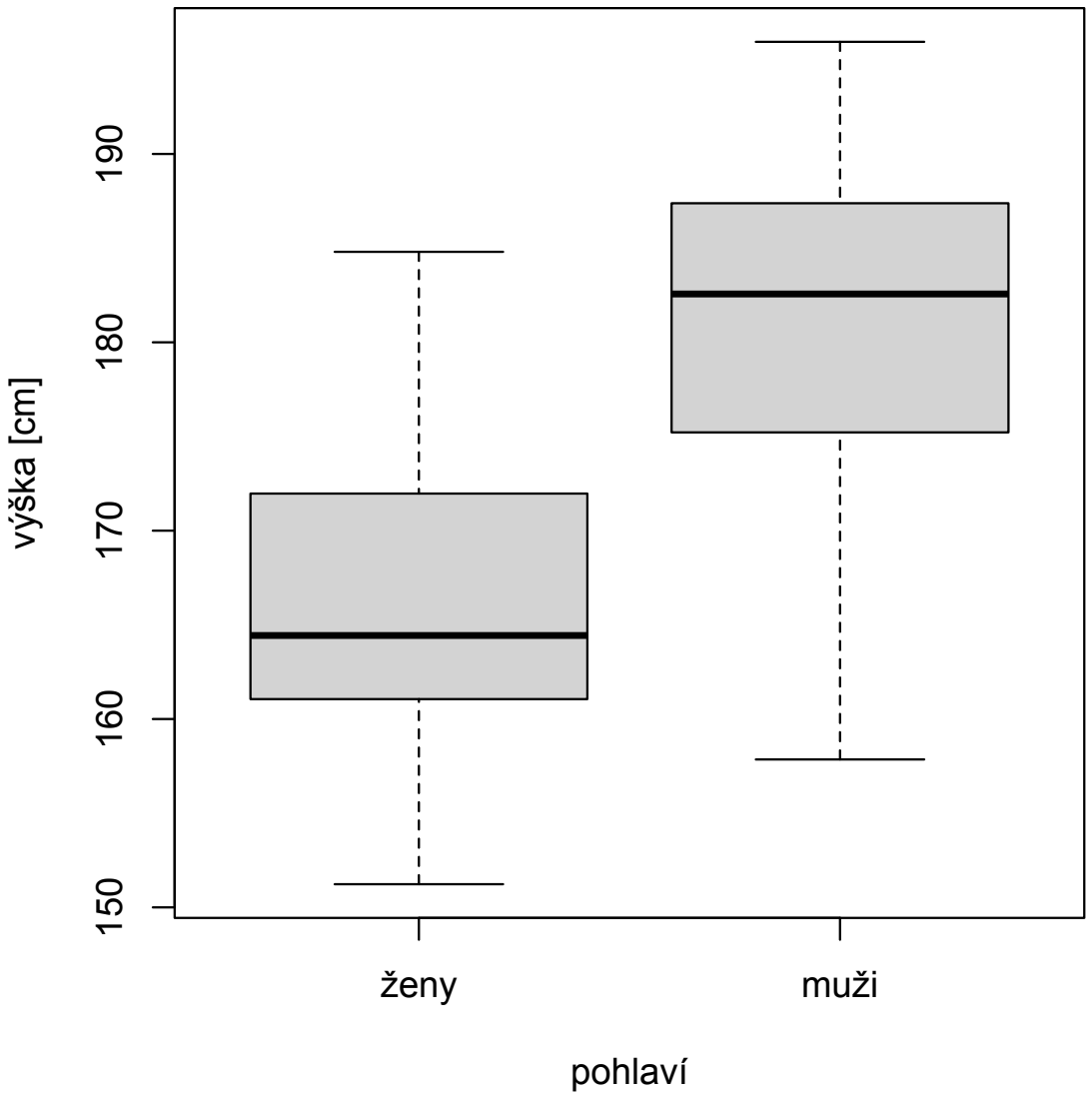


# Dvouvýběrový $t$-test

- 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 $\alpha$

```
1  set.seed(1)
2  muzi <- rnorm(30, 180, 10)
3  zeny <- rnorm(30, 165, 10)
4
5  t.test(muzi, zeny)
6  # Welch Two Sample t-test
7  # t = 6.5125, df = 56.741, p-value = 2.093e-08
8  # ...
9
10 t.test(muzi, zeny)$p.value   # 2.093108e-08
```

# Dvouvýběrový $t$-test

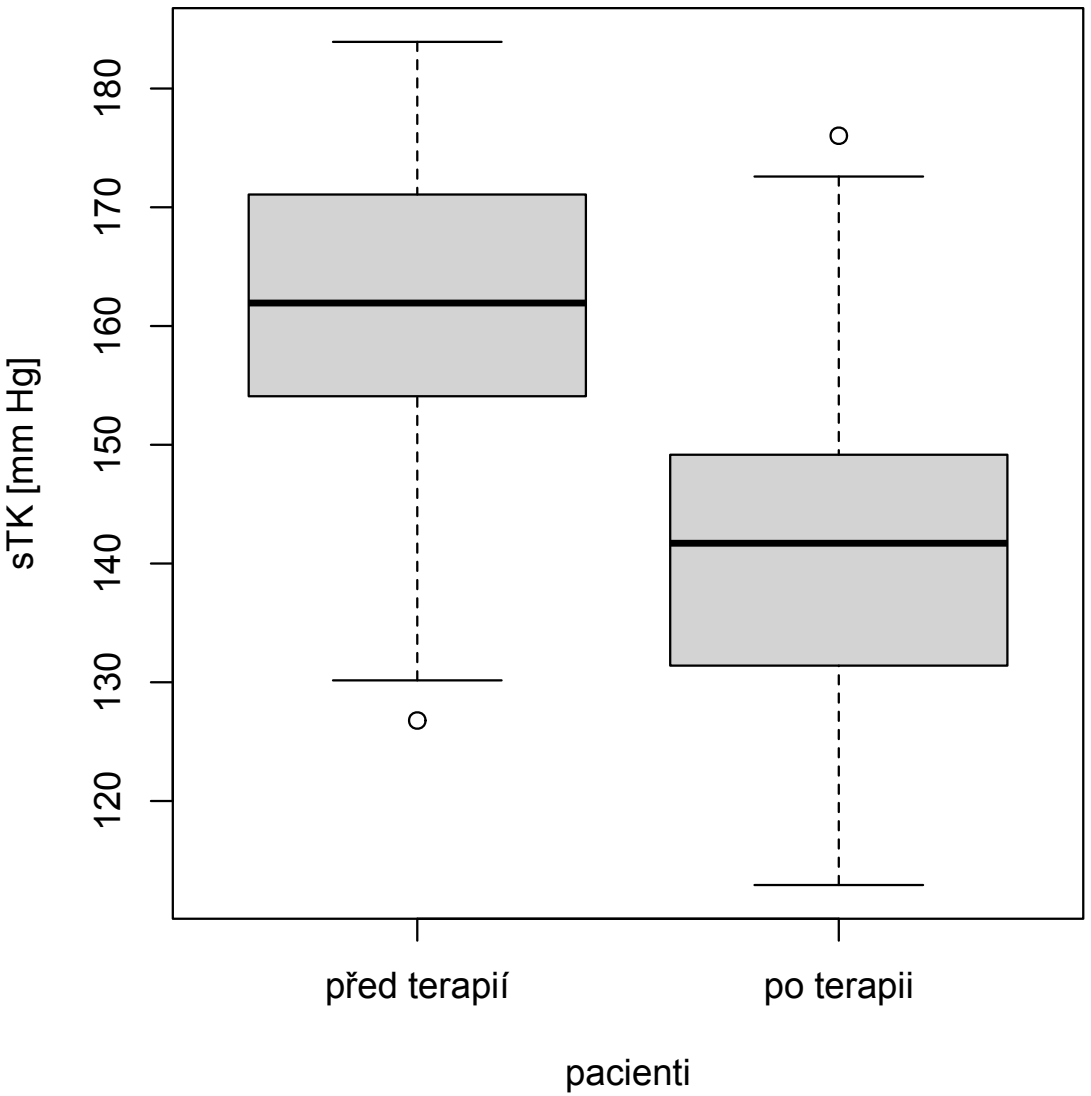


# Párový $t$-test

- 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 $\alpha$

```
1  set.seed(1)
2  pacienti_pred <- rnorm(50, 160, 15)
3  pacienti_po <- rnorm(50, 140, 15)
4
5  t.test(
6    pacienti_pred, pacienti_po, paired = TRUE
7  )
8  # Paired t-test
9  # t = 7.1546, df = 49, p-value = 3.823e-09
10 # ...
```

# 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 $\alpha$

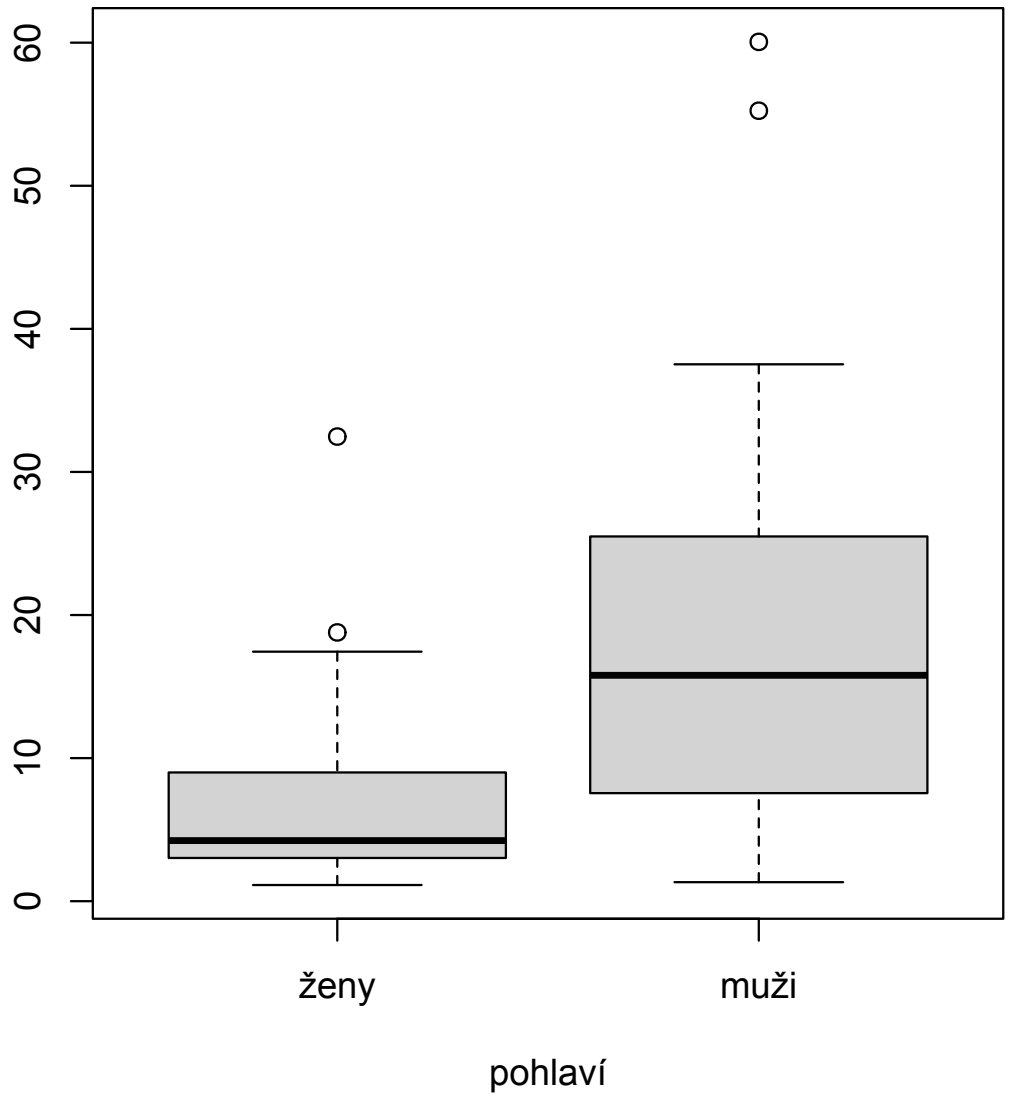
```
1  set.seed(1)
2  muzi <- rnorm(30, 180, 10)
3  zeny <- rnorm(30, 165, 10)
4
5  var.test(muzi, zeny)
6  # F test to compare two variances
7  # F = 1.3501, num df = 29, denom df = 29,
8  # p-value = 0.4238
9
10 var.test(muzi, zeny)$p.value   # 0.4237845
```

# 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 $\alpha$

```
1  set.seed(1)
2  muzi <- exp(rnorm(30, 2.5, 1))
3  zeny <- exp(rnorm(30, 1.5, 1))
4
5  wilcox.test(muzi, zeny)
6  # Wilcoxon rank sum test
7  # W = 710, p-value = 7.215e-05
```

# Wilcoxonův dvouvýběrový test

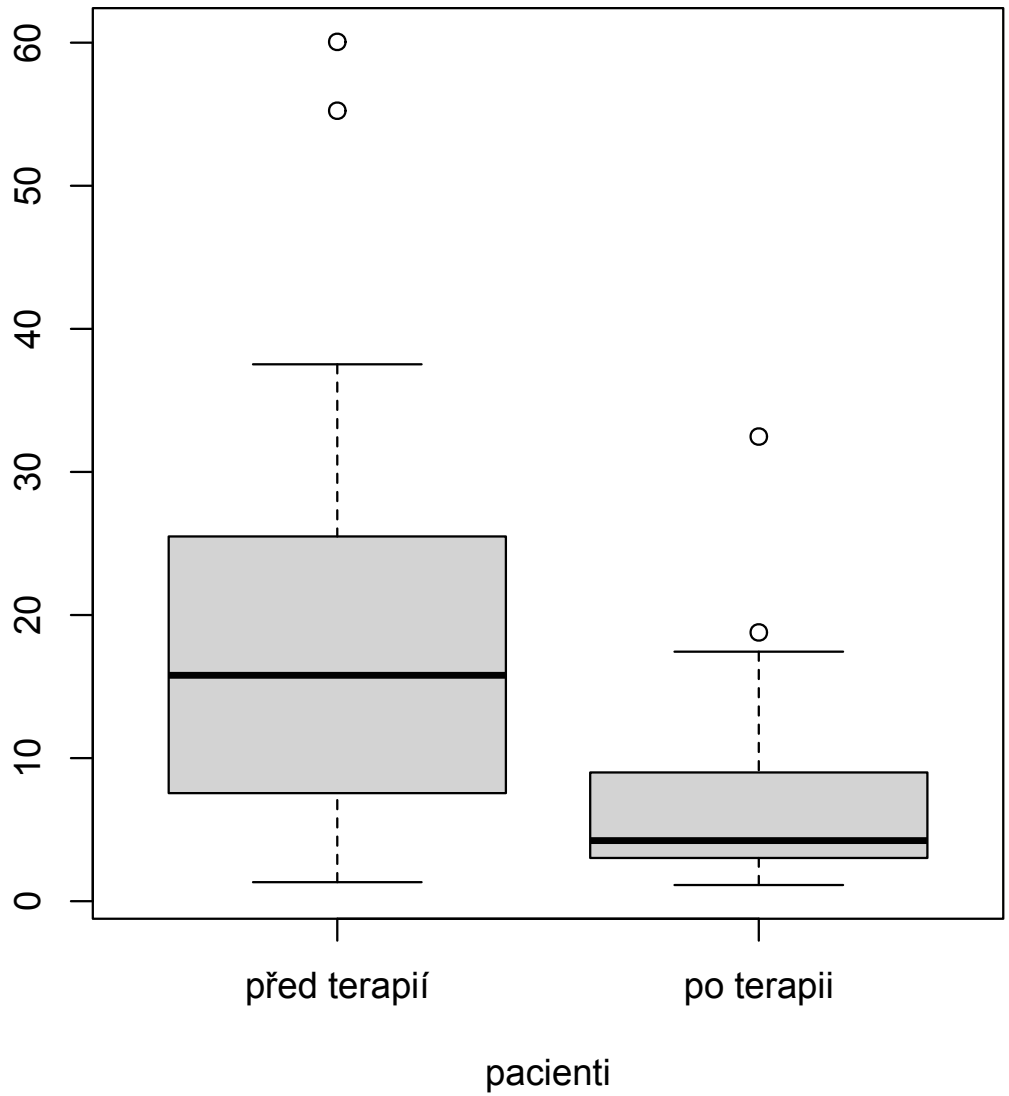


# Wilcoxonův párový test

- 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 $\alpha$

```
1  set.seed(1)
2  pacienti_pred<- exp(rnorm(30, 2.5, 1))
3  pacienti_po <- exp(rnorm(30, 1.5, 1))
4
5  wilcox.test(
6    pacienti_pred, pacienti_po, paired = TRUE
7  )
8  # Wilcoxon signed rank test
9  # V = 400, p-value = 0.0002833
10 # ...
```

# Wilcoxonův párový test

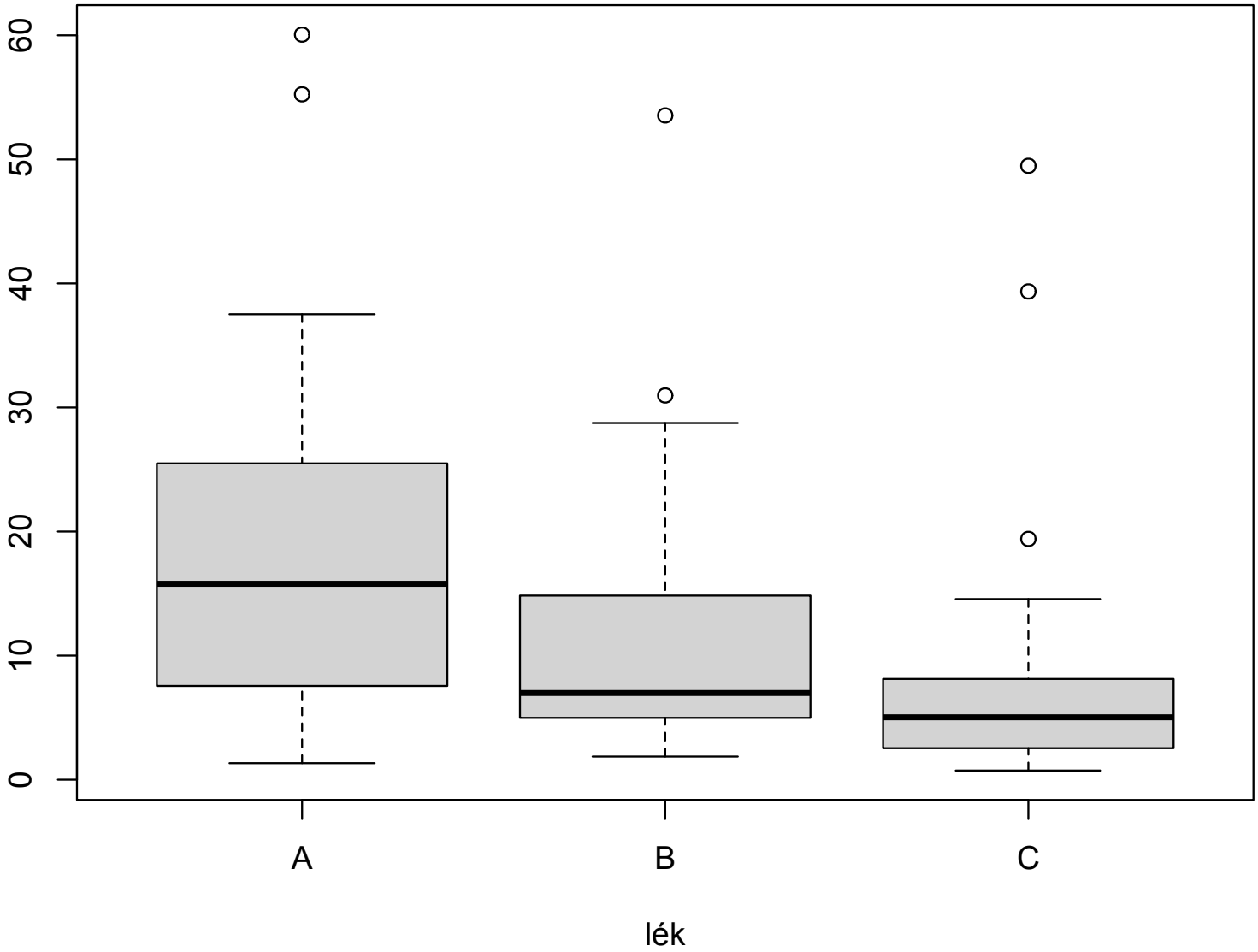


# Kruskal-Wallisův test

- 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 $\alpha$

```
1  set.seed(1)
2  lek_A <- exp(rnorm(30, 2.5, 1))
3  lek_B <- exp(rnorm(30, 2.0, 1))
4  lek_C <- exp(rnorm(30, 1.5, 1))
5
6  my_data <- data.frame(
7    "mira" = c(lek_A, lek_B, lek_C),
8    "lek" = c(rep("A", 30), rep("B", 30),
9              rep("C", 30))
10 )
```

# Kruskal-Wallisův test



# Kruskal-Wallisův test

- výsledek testu

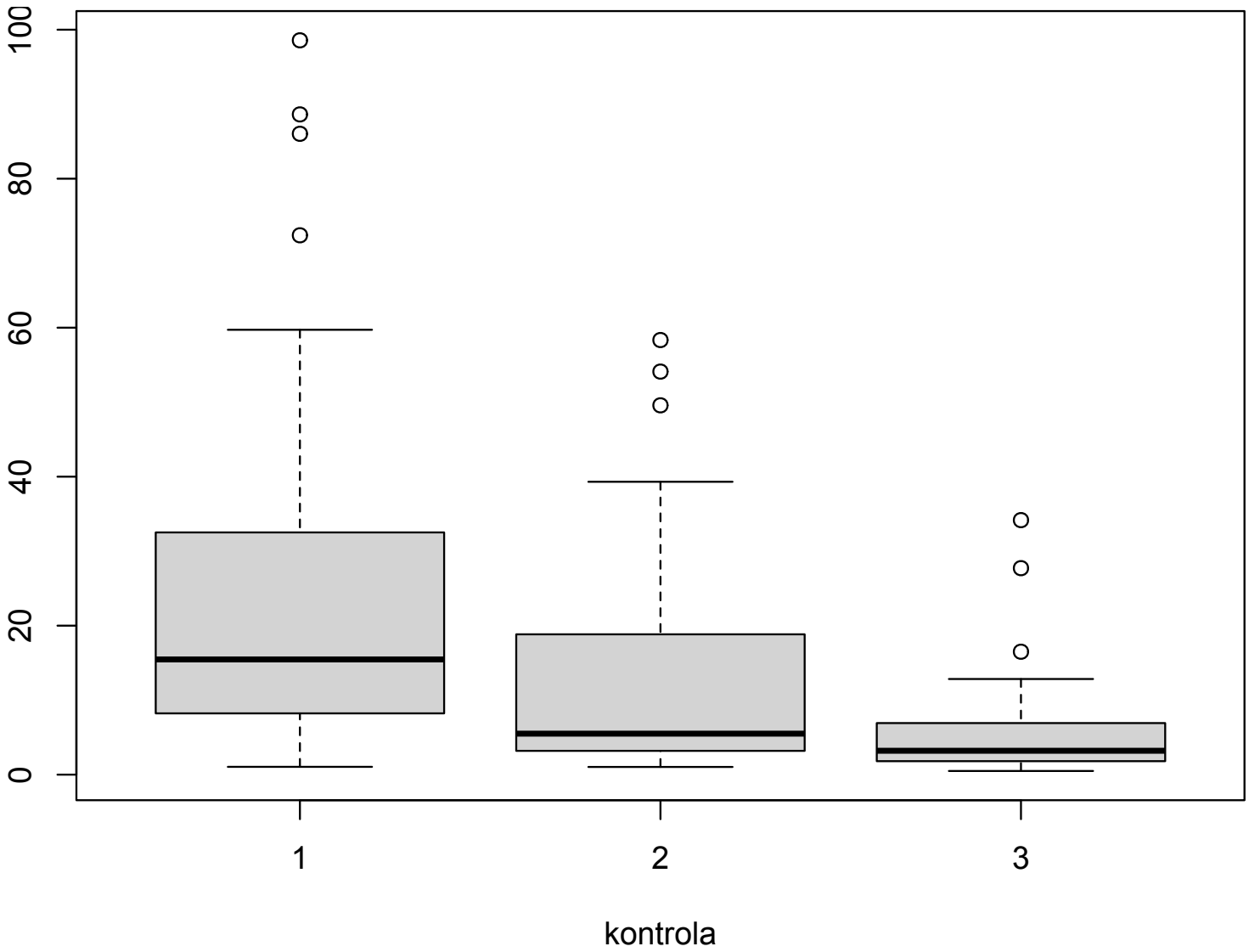
```
kruskal.test(mira ~ lek, my_data)
# Kruskal-Wallis rank sum test
# Kruskal-Wallis chi-squared = 16.945,
# df = 2, p-value = 0.0002092
```

# 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 $\alpha$

```
1  set.seed(2)
2  cas_1 <- exp(rnorm(30, 2.5, 1))
3  cas_2 <- exp(rnorm(30, 2.0, 1))
4  cas_3 <- exp(rnorm(30, 1.5, 1))
5
6  friedman.test(cbind(cas_1, cas_2, cas_3))
7  # Friedman rank sum test
8  # Friedman chi-squared = 18.6, df = 2,
9  # p-value = 9.142e-05
```

# Friedmanův test



# $\chi^2$ test nezávislosti

- $\chi^2$-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 $\alpha$

```
1  chisq.test(
2    matrix(c(
3      12, 20, 30,
4      18, 14, 10
5    ), nrow = 2, byrow = T)
6  )
7  # Pearson's Chi-squared test
8  # X-squared = 8.7357, df = 2, p-value = 0.01268
```

# $\chi^2$ testy dobré shody

- $\chi^2$-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 $\alpha$

```
1  chisq.test(
2    c(10, 15, 14),
3    p = c(1/3, 1/3, 1/3)
4  )
5  # Chi-squared test for given probabilities
6  # X-squared = 1.0769, df = 2, p-value = 0.5836
```

# Literatura


ZVÁRA, Karel. *Základy statistiky v prostředí R*.
Praha, Česká republika: Karolinum, 2013. ISBN 978-80-246-2245-3.


WICKHAM, Hadley. *Advanced R*.
Boca Raton, FL: CRC Press, 2015. ISBN 978-1466586963.

## Děkuji za pozornost!

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