Katedra pravděpodobnosti a matematické statistiky

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

**Zdeněk Pustějovský**

# Permutační testy

9. listopadu 2021

# Obsah


1. Permutační testy - dvouvýběrové
   - Princip, výpočet p-hodnoty, pro a proti...
2. Porovnání se známými dvouvýb. testy
   - t-test, Wilcoxonův test
3. Permutační testy nezávislosti
4. K-výběrové permutační testy
   - ANOVA, Kruskalův-Wallisův test

# Permutační testy - dvouvýběrové

- Dva vzájemně nez. náhodné výběry
  $\mathbf{X} = (X_1, ..., X_n) \sim F_X, \mathbf{Y} = (Y_1, ..., Y_m) \sim F_Y$
  - Existují hustoty $f_X$ a $f_Y$
- Testujeme hypotézu $H_0 : F_X(x) = F_Y(x)\ \forall x \in \mathbb{R}$
- Máme testovou statistiku $U$
- Označme $\mathbf{Z} = (X_1, ..., X_n, Y_1, ..., Y_m)$

# Princip

- Pro jednoduchost budeme předpokládat, že v datech nenastávají shody, tj.

$$z_{(1)} < z_{(2)} < \ldots < z_{(n+m)}$$

- Za $H_0$ platí

$$\mathrm{P}(\mathbf{Z} = (z_1, \ldots, z_{n+m}) | \mathbf{Z}_{(.)} = (z_{(1)}, \ldots, z_{(n+m)})) = \frac{1}{(n+m)!}$$

pro každou permutaci složek $(z_{(1)}, \ldots, z_{(n+m)})$

# Výpočet p-hodnoty

- Označme $u$ hodnotu $U$ pro naše data $(x_1, ..., x_n, y_1, ..., y_m)$
- Spočteme hodnoty $u^*$ pro všechny permutace $\mathbf{z}^* = ({z_1}^*, ..., {z_{n+m}}^*)$
- p-hodnotu dostaneme jako (za $H_0$)

$$p = \mathrm{P}(|u^*| \geq |u| \mid \mathbf{Z}_{(.)} = (z_{(1)}, ..., z_{(n+m)}) = \frac{\#\{|u^*| \geq |u|\}}{n!}$$

# Pro a proti

- Pro:
  - Není třeba znát rozdělení $U$ za $H_0$
  - Někdy můžeme zobecnit model (K-S test)
  - I pro malé rozsahy výběrů
- Proti:
  - Výpočetně náročné
  - Špatně se určují intervaly spolehlivosti

# Co když testuji "slabší" $H_0$?

- Např. rovnost středních hodnot, rozptylů...

# Co když testuji "slabší" $H_0$?

- Pak už je test jen přibližný
- Obvykle dodržuje hladinu asymptoticky
  - Dokonce (obvykle) lépe než klasické asymptotické testy

# Dvouvýběrový t-test

- Model $\mathcal{F} = \{F_X \sim N(\mu_X, \sigma^2), F_Y \sim N(\mu_Y, \sigma^2)\}$
- $H_0 : \mu_X - \mu_Y = 0$ proti $H_1 : \mu_X - \mu_Y \neq 0$
- Testová statistika

$$T_{n,m} = \frac{\overline{X_n} - \overline{Y_m}}{\sqrt{S^2_{n,m}(\frac{1}{n} + \frac{1}{m})}},$$

  kde $S^2_{n,m} = \frac{n-1}{n+m-2} S^2_X + \frac{m-1}{n+m-2} S^2_Y$
- Zamítáme $H_0 \iff |T_{n,m}| \geq t_{n+m-2}(1 - \frac{\alpha}{2})$
- p-hodnota je $2(1 - F(|t|))$

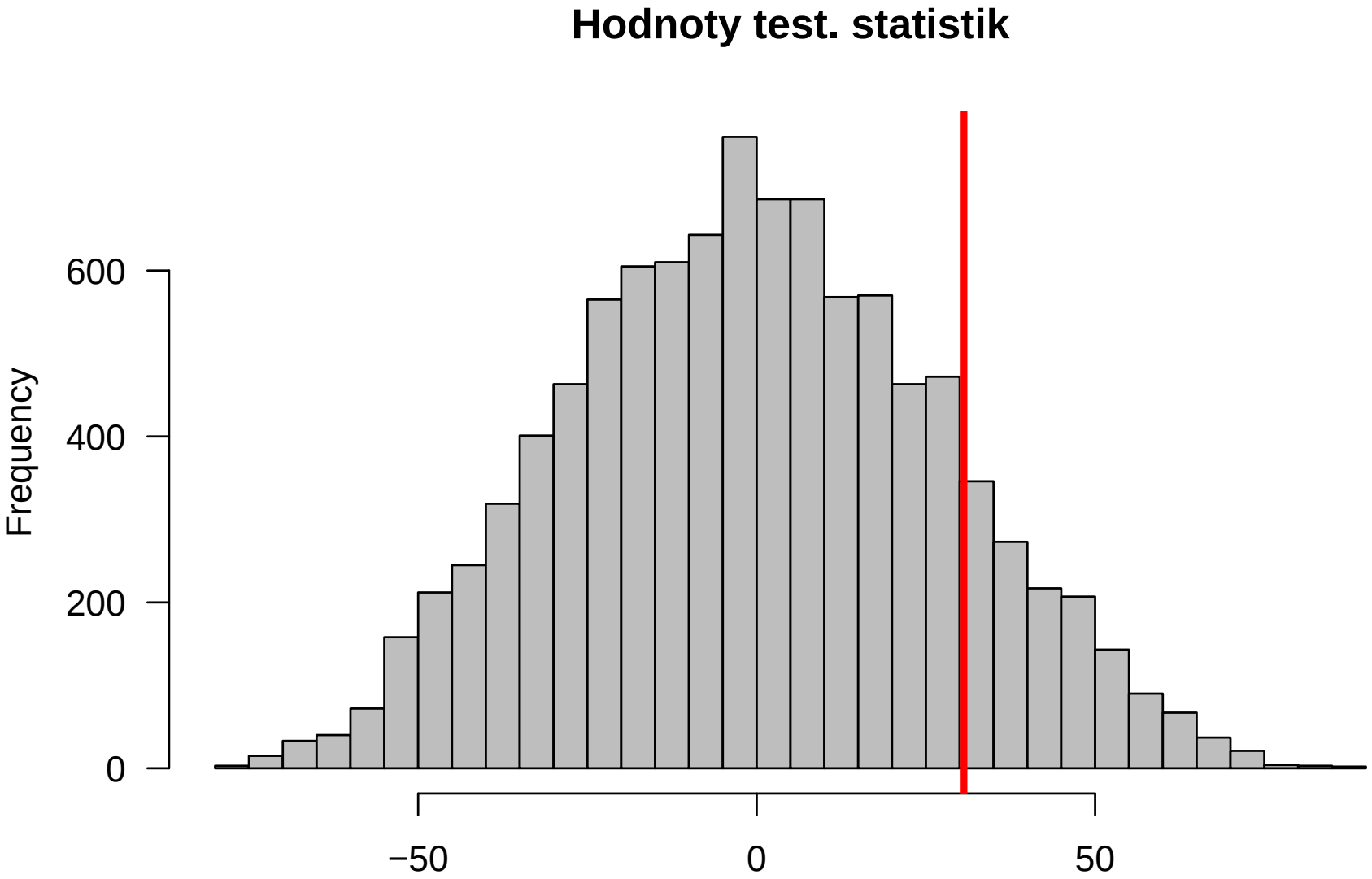
# Příklad

Tabulka: Čas přežití

| Skupina | Počet | Průměr | Data | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Léčba | 7 | 86.86 | 94 | 197 | 16 | 38 | 99 | 141 | 23 |
| Neléčba | 9 | 56.22 | 52 | 104 | 146 | 10 | 51 | 30 | 40 |
| | | | 27 | 46 | | | | | |

- Rozdíl průměrů je 30.63

# p-hodnota



# Wilcoxonův test

- Model $\mathcal{F} = \{\exists g$ rost. $\exists \delta \in \mathbb{R} : g(X_i) \sim \tilde{F}_X$ spojitá d. f., $g(Y_i) \sim \tilde{F}_Y, \tilde{F}_X(x) = \tilde{F}_Y(x - \delta) \forall x \in \mathbb{R}\}$
- $H_0 : \delta = 0$ proti $H_1 : \delta \neq 0$
- Testová statistika $W_{n,m} = \sum_{i=1}^{n} R_i$, kde $R_i$ jsou pořadí $X_i$ ve sdruženém náhodném výběru.
  - Pro malé rozsahy známe rozdělení za $H_0$ přesně, jinak asymptoticky

# Permutační testy nezávislosti

- Máme náhodný výběr dvojic $\mathbf{Z}_1 = \binom{X_1}{Y_1}, \ldots, \mathbf{Z}_n = \binom{X_n}{Y_n}$
- Testujeme, zda je $X_1$ nezávislé s $Y_1$
- Statistika $U$ např. Pearsonův korelační koef. nebo $\chi^2$-test nezávislosti

$$\hat{\rho}_n = \frac{S_{XY}}{S_X S_Y}$$

# Permutační testy nezávislosti

- Za $H_0$ platí

$$\mathrm{P}\left(\mathbf{Z}_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \mathbf{Z}_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix} \middle| \begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_{(1)} \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix} = \begin{pmatrix} x_n \\ y_{(n)} \end{pmatrix}\right) =$$

$$= \frac{1}{n!}$$

pro každou permutaci složek $(y_{(1)}, \ldots, y_{(n)})$

- p-hodnotu dostaneme jako

$$p = \mathrm{P}(|u^*| \geq |u| \mid \mathbf{Z}_{(.)} = (z_{(1)}, \ldots, z_{(n+m)}) = \frac{\#\{|u^*| \geq |u|\}}{n!}$$

# K-výběrový problém

- Teď máme $K \geq 2$ vz. nezávislých náhodných výběrů
- Obecně testujeme hypotézu nulového rozdílu

$$F_1(x) = F_2(x) = ... = F_K(x) \; \forall x \in \mathbb{R}$$

- Dá se na něj dívat jako na test nezávislosti

# ANOVA

- Model $\mathcal{F} = \{F_k \in \mathcal{L}^2_+, k \in \{1, ..., K\}, F_k$ mají stejné rozptyly$\}$
- $H_0 : \mu_1 = ... = \mu_K$ proti $H_1$ : jinak
- Testová statistika $F_A = \frac{SS_A/(K-1)}{SS_e/(N-K)}$, kde $SS_A = \sum_{k=1}^{K} n_k(\overline{Y}_{k+} - \overline{Y}_{++})^2$ a $SS_e = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (Y_{ki} - \overline{Y}_{k+})$
- Zamítáme $H_0 \iff F_A \geq F_{K-1,N-K}(1-\alpha)$
- p-hodnota je $1 - F^*(s)$

# Kruskalův-Wallisův test

- K-výběrový zobecněný model posunutí
- $H_0 : \delta_1 = ... = \delta_K$
- Za $H_0$ jsou výběry vz. iid
- Testová statistika

$$Q = \frac{12}{N(N+1)} \sum_{k=1}^{K} n_k \left( \overline{R}_{k+} - \frac{N+1}{2} \right)^2$$

- Zamítáme $H_0 \iff Q \geq \chi^2_{K-1}(1-\alpha)$

# Zdroje


- https://www2.karlin.mff.cuni.cz/ omelka/Soubory/nmst434 /nmst434_course-notes.pdf
- Davison, A. C. and Hinkley, D. V. (1997). Bootstrap Methods and their Application. Cambridge University Press, New York. Chapter 4.3.
- Efron, B. and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman & Hall. Chapter 15.
- Lehman E. L. and Romano, J. P. (2005). Testing Statistical Hypotheses. Springer

**Děkuji za pozornost!**