Dvouvýběrový Wilcoxonův test

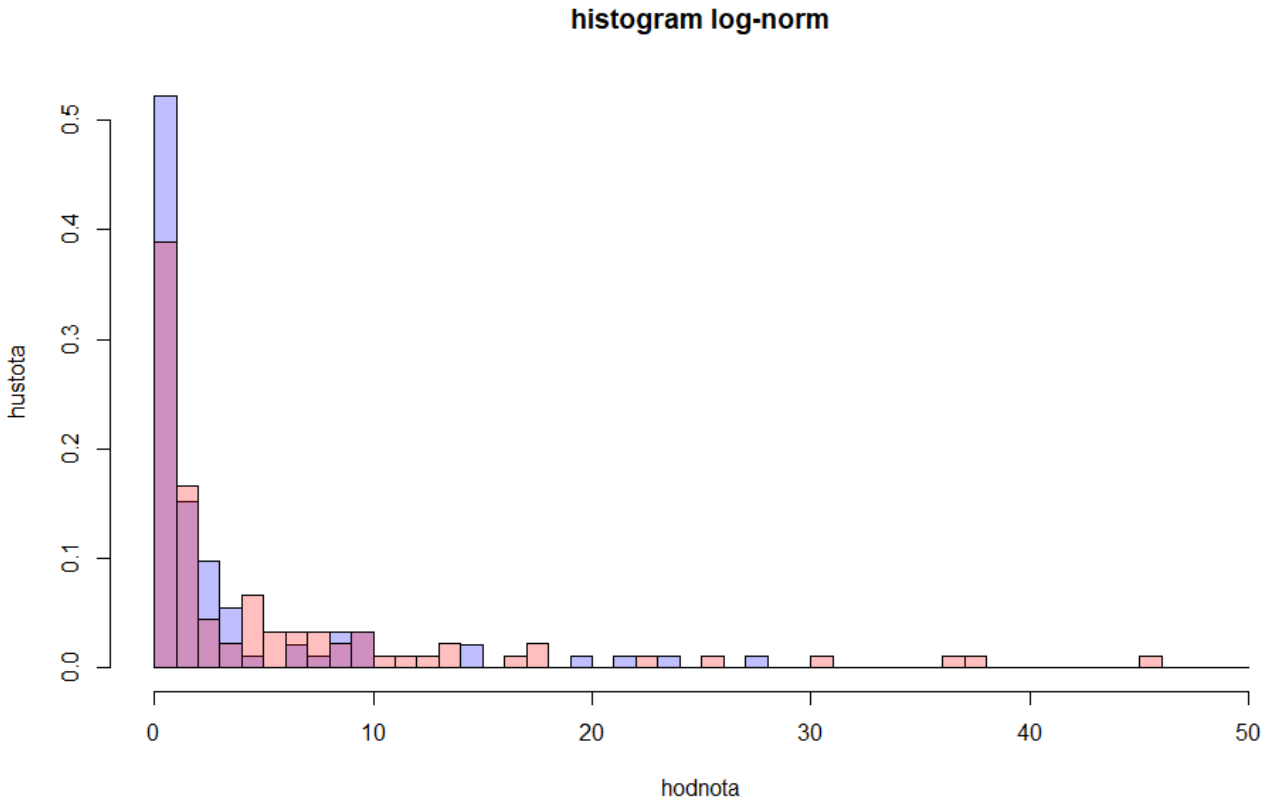
Kateřina Lipavská

26.10.2021

Čím se budeme zabývat

Příklad dat

log-normální rozdělení s jinými středními hodnotami a stejnými rozptyly, t-test je silný v normálním modelu



Uspořádaný náhodný výběr

- náhodný výběr $X_1, \ldots, X_n$
- předpoklady: $f << \lambda$, spojité rozdělení
- $X_{(R_i)} = X_i$

Wilcoxonův test

- $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ nezávislé n. v.
- Model posunutí:
  $F = \{F_X, F_Y \text{ spojité}, \exists \delta \in \mathbb{R} : F_X(x) = F_Y(x - \delta) \forall x \in \mathbb{R}\}$
- H_0: $\delta_{XY} = 0$, H_1: $\delta_{XY} \neq 0$
- Sdružený výběr $\mathbf{Z} = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$
- R_i pořadí X_i v $\mathbf{Z}$
- Testová statistika $W_{n,m} = \sum_{i=1}^{n} R_i$
- $U = \frac{W_{n,m} - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \xrightarrow[\mathrm{H}_0]{d} N(0,1)$
- Kritický obor: zamítáme H_0 právě, když $|U| \geq u_{1-\frac{\alpha}{2}}$

Mannova-Whitneyho statistika

- $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ ze spojitých rozdělení
- Model $F^* = \{X \sim F_X, Y \sim F_Y,$ spojité d. f.$\}$
- Testovaný parametr $\theta = P(X \leq Y)$
- H_0^*: $\theta = \frac{1}{2}$, H_1^*: $\theta \neq \frac{1}{2}$
- Testová statistika $W_{n,m}^* = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}(X_i \leq Y_j)$
- $nmW_{n,m}^* + W_{n,m} = nm + \frac{n(n+1)}{2}$

U-statistika

- 2 nezávislé náhodné výběry
- $\{X_1, \ldots, X_n\} \sim F_X$, $\{Y_1, \ldots, Y_m\} \sim F_Y$
- funkce h,

$$U_{n,m} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} h\left(X_i, Y_j\right)$$

Vyjádření testové statistiky

Vyjádření pomocí součtu nezávislých n. v.

$$W^*_{n,m} = \frac{1}{n}\sum_{i=1}^{n} g_1(X_i) + \frac{1}{m}\sum_{j=1}^{m} g_2(Y_j) + r_{n,m},$$

kde $g_1(x) = 1 - F_Y(x)$ a $g_2(y) = F_X(y)$.

Testová statistika bez zbytkového členu:

$$\widetilde{W}_{n,m} = \frac{1}{n}\sum_{i=1}^{n} g_1(X_i) + \frac{1}{m}\sum_{j=1}^{m} g_2(Y_j)$$

Momenty a asymptotické rozdělení

$$\mathbf{var}\,\widetilde{W}_{n,m} = \frac{1}{n}\delta_{0,1} + \frac{1}{m}\delta_{1,0}$$

Z předchozího a z Cramér-Sluckého věty dostáváme, že

$$\frac{W^*_{n,m} - \theta}{\sqrt{\mathbf{var}\,\widetilde{W}_{n,m}}} = \sqrt{\frac{nm}{m\delta_{0,1} + n\delta_{1,0}}}(W^*_{n,m} - \theta) \xrightarrow{d} N(0,1),$$

kde $\delta_{0,1} = \mathbf{var}(g(X_i))$ a $\delta_{1,0} = \mathbf{var}(h(Y_j))$.

Mannův-Whitneyův test

- Model posunutí
  $F = \{F_X, F_Y \text{ spojité}, \exists \delta \in \mathbb{R} : F_X(x) = F_Y(x - \delta) \forall x \in \mathbb{R}\}$
- H_0: $\delta_{XY} = 0$, H_1: $\delta_{XY} \neq 0$
- za H_0 je $\theta = \frac{1}{2}$, $\delta_{0,1} = \delta_{1,0} = \frac{1}{12}$
- Kritický obor $\sqrt{\frac{12mn}{m+n}} |W^*_{n,m} - \frac{1}{2}| \geq u_{1-\frac{\alpha}{2}}$

Zobecněný Mannův-Whitneyův test

- Model $F^* = \{X \sim F_X, Y \sim F_Y, \text{ spojité d. f.}\}$
- H_0: $\theta = \frac{1}{2}$, H_1: $\theta \neq \frac{1}{2}$
- Testová statistika $\widetilde{U}_{n,m} = \frac{W^*_{n,m} - \frac{1}{2}}{\sqrt{\frac{1}{n}\hat{\delta}_{0,1} + \frac{1}{m}\hat{\delta}_{1,0}}}$
- Kritický obor $|\widetilde{U}_{n,m}| \geq u_{1-\frac{\alpha}{2}}$

Neplatí model posunutí

- Záleží na rozsazích
- Zobecněný test dodržuje hladinu asymptoticky
- $\theta = \frac{1}{2}$, např. $N(0,1)$ a $R(-1,1)$
- $\theta \neq \frac{1}{2}$, např. $N(0,1)$ a $R(-\frac{3}{4}, \frac{5}{4})$

$\theta = \frac{1}{2}$, zamítnutí v %

Procento zamítnutí pro $N(0,1)$ a $R(-1,1)$

Rozsah X	Rozsah Y	Klasický test	Zobecněný test
10	30	8,92	7,71
30	10	2,74	6,24
30	90	8,40	5,65
90	30	2,75	5,68
50	50	5,51	5,39

z BP T. Šlampiaka

$\theta \neq \frac{1}{2}$, zamítnutí v %

Procento zamítnutí pro $N(0,1)$ a $R(-\frac{3}{4}, \frac{5}{4})$

Rozsah X	Rozsah Y	Klasický test	Zobecněný test
10	30	16,71	14,59
30	10	9,53	17,08
30	90	30,89	24,78
90	30	26,76	36,97
50	50	32,10	31,82

z BP T. Šlampiaka

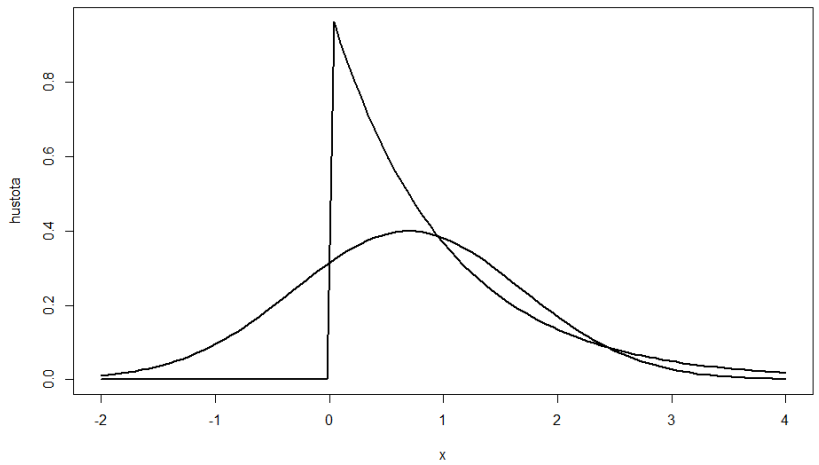
Varování

Wilcoxonův ani Mannův-Whitneyův test neporovnává mediány ani střední hodnoty.

Např. $X \sim N(\ln(2), 1)$ a $Y \sim \mathrm{Exp}(1)$, $\theta = P(X \leq Y) \doteq 0,44$

Zobecněný Mann-Whitney

Rozsahy	p-hodnota
50	$0,62$
100	$0,15$
250	$0,04$
500	$2,7 \cdot 10^{-5}$
1000	$5,6 \cdot 10^{-8}$



Shodná pozorování

- H_0: $\theta = P(X < Y) + \frac{1}{2}P(X = Y)$
- Wilcoxon: průměrná pořadí a úprava testové statistiky

$$\frac{W_{n,m} - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1-kor.)}{12}}}$$

- Mann-Whitney (i zobecněný):
  $W^*_{n,m} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\mathbb{1}(X_i < Y_j) + \frac{1}{2}\cdot\mathbb{1}(X_i = Y_j)\right]$
  $+$ úprava odhadu rozptylu

Závěr

- zopakování Wilcoxonova testu
- Mannův-Whitneyův test a jeho zobecnění
- je potřeba myslet na to, co testujeme a za jakých předpokladů

Zdroje

Chung, E., Romano, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample U-statistics. Journal of Statistical Planning and Inference, 168, 97-105.

Šlampiak, T. (2016). Wilcoxonův dvouvýběrový test. Bakalářská práce. MFF UK.

Omelka, M. (2021). Poznámky k přednášce Matematická statistika 1. KPMS MFF UK.

Poděkování

Děkuji za pozornost.

Děkuji doc. Omelkovi za rady k úpravám prezentace a poskytnutí zdrojů.