Statistika

(MD360P03Z, MD360P03U)
ak. rok 2007/2008

Karel Zvára

karel.zvara@mff.cuni.cz
http://www.karlin.mff.cuni.cz/∼zvara

(naposledy upraveno 26. listopadu 2007)

Mannův-Whitneyův (Wilcoxonův) test

pořadová obdoba dvouvýběrového t-testu

- porovnáváme stejný kvantitativní znak ve dvou populacích
- máme dva **nezávislé** výběry z těchto populací
- co když nelze předpokládat normální rozdělení?
- nechť $X_1, \ldots, X_{n_1}$ a $Y_1, \ldots, Y_{n_2}$ jsou **nezávislé** výběry ze spojitého rozdělení (například věk matek, střední délka života mužů při narození ve dvou skupinách zemí, potratovost ...)
- H_0 tvrdí, že obě rozdělení jsou stejná (mezi populacemi není rozdíl, zpravidla nás zajímá, že není rozdíl v mírách polohy)
- specielně to znamená, že **populační mediány** jsou shodné
- postup založen na pořadí bez ohledu na výběr
- idea: kdyby nebyl mezi populacemi rozdíl, byla by takto zjištěná průměrná pořadí v obou výběrech podobná

příklad: potraty na 1000 obyv. (Čechy vers. Morava)

v roce 2003

kraj	Pha	Stč	Jč	Pl	KV	Ús	Lb
potratovost	4,03	4,02	4,11	4,70	5,65	5,80	4,98
pořadí	7	6	8	10	12	13	11

kraj	HK	Par	Vys	JM	Ol	Zl	MS
potratovost	4,33	3,38	3,57	3,70	3,65	3,42	3,87
pořadí	9	1		4	3	2	5

- H_0 : shoda populací (zejm. mediánů), H_1 : neshoda
- nejasné, kam patří kraj Vysočina; vynecháme jej
- průměrné pořadí českých krajů: 77/9=8,56
  W_1=7+6+8+10+12+13+11+9+1=77
- průměrné pořadí moravských krajů: 14/4=3,5
  W_2=4+3+2+5=14

přibližné rozhodování (n_1, n_2 desítky)

- W_1, W_2 součty pořadí, W_1 standardizujeme

$$Z = \frac{W_1 - n_1(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

- za hypotézy (není rozdíl mezi populacemi) je použitím centrální limitní věty $Z \sim \mathrm{N}(0,1)$
- hypotézu zamítáme, je-li $|Z| \geq z(\alpha/2)$
- náš příklad: [wilcox.test(potr∼Cechy)]

$$Z = \left| \frac{77 - 9 \cdot 14/2}{\sqrt{9 \cdot 4 \cdot 14/12}} \right| = 2{,}16 > 1{,}96 = z(0{,}05/2) \qquad p = 3{,}1\ \%$$

- na 5% hladině jsme prokázali rozdíl

přesný výpočet p-hodnoty Wilcoxonova testu

- zajímá nás, nakolik je náš výsledek ($W_1 = 77, W_2 = 14$) výjimečný
- máme celkem $n_1 + n_2 = 13$ pozorování, čtyři z nich (tolik jich je v menší skupině, z Moravy) lze vybrat celkem $\binom{13}{4} = 715$ způsoby
- kolik z těchto způsobů vede k tak extrémně nestejným průměrným pořadím?
- budeme hledat, kolik čtveřic označených za moravské by dalo v součtu nejvýš 14, jak nám doopravdy vyšlo
- vždy platí $W_1 + W_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2 = 91$ (součet čísel $1 + 2 + \ldots + n_1 + n_2$)
- stačí zabývat se jedinou ze statistik W_1, W_2, zpravidla tou pro menší výběr

přehled možných čtveřic v nichž je součet pořadí nejvýš 14

(čtveřice vybíráme z čísel 1, 2, . . . ,13)

1	1	1	1	1	1	1	1	1	1	1	2	1	1
2	2	2	2	2	2	3	2	2	2	3	3	2	2
3	3	3	4	3	4	4	3	4	5	4	4	3	4
4	5	6	5	7	6	5	8	7	6	6	5	9	8
10	11	12	12	13	13	13	14	14	14	14	14	15	15

- nejvýš 14 mohl být součet pořadí za platnosti hypotézy s pravděpodobností $p_1 = 12/715 = 0{,}01678$
- protože máme oboustrannou alternativu, musíme vzít v úvahu také situaci, kdy by byla na Moravě velká pořadí, p-hodnotu nutno zdvojnásobit: $p = 24/715 = 3{,}4$ %

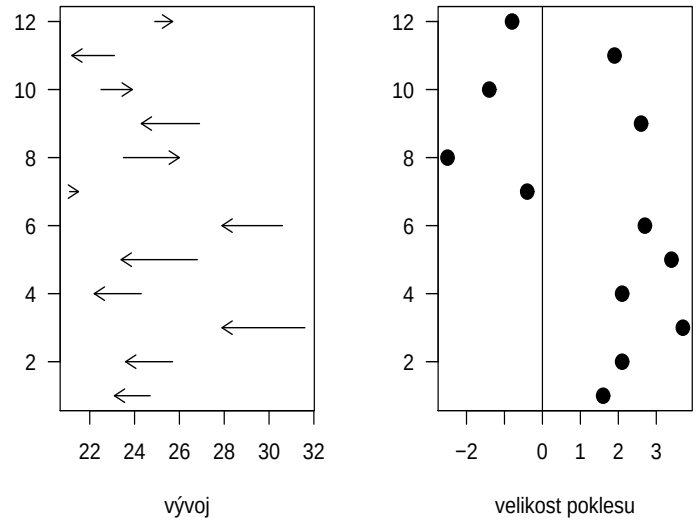
příklad: klesá potratovost? (párový t-test zde nevhodný)

potratů na 100 těhotenství

Y_i	Z_i	X_i	R_i^+
24,7	23,1	1,6	4
25,7	23,6	2,1	6
31,6	27,9	3,7	12
24,3	22,2	2,1	7
26,8	23,4	3,4	11
30,6	27,9	2,7	10
21,1	21,5	-0,4	1
23,5	26,0	-2,5	8
26,9	24,3	2,6	9
22,5	23,9	-1,4	3
23,1	21,2	1,9	5
24,9	25,7	-0,8	2

- použijeme údaje z 12 okresů v letech 2000 (Y_i) a 2001 (Z_i)
- hypotéza H_0 : v obou letech potratovost stejná, rozdíly dány náhodným kolísáním; H_1 : potratovost klesá (jednostranná alt.)
- za H_0 by rozdíly měly kolísat **symetricky kolem nuly**
- za H_1 by měly převládat kladné rozdíly, spíše velké
- průměrné pořadí z 8 kladných rozdílů: 8 (součet $W = 64$), průměrné pořadí ze 4 záporných rozdílů 3,5 (součet 14)

příklad: klesá potratovost?



párový Wilcoxonův (Wilcoxon signed rank) test

- nechť $(Y_1, Z_1) \ldots, (Y_n, Z_n)$ **nezávislé** dvojice, rozdíly $X_i = Y_i - Z_i$ mají **spojité** rozdělení
- H_0 : Y_i, Z_i mají stejné rozdělení (populace jsou stejné)
- mají-li Y_i, Z_i stejné rozdělení, pak rozdíly $X_i = Y_i - Z_i$ jsou symetricky rozděleny kolem nuly
- postup
  - vyloučit nulové hodnoty X_i (tedy shodné hodnoty Y_i, Z_i), podle toho případně zmenšit n
  - určit pořadí R_i^+ **absolutních hodnot** $|X_i| = |Y_i - Z_i|$
  - určit W, tj. součet pořadí původně kladných hodnot X_i
  - podle W rozhodnout

rozhodování

- na základě centrální limitní věty lze použít

$$Z = \frac{W - \mathrm{E}\, W}{\mathrm{S.E.}(W)} = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

- hypotézu o shodě zamítneme, bude-li $|Z| \geq z(\alpha/2)$
- při jednostranné alternativě porovnat Z a $z(\alpha)$
- pro malý počet dvojic (do deseti) raději použít tabulky
- příklad ($W = 64, n = 12$, jinak přesně je $p = 2{,}6$ %)

$$Z = \frac{64 - 12 \cdot 13/4}{\sqrt{12 \cdot 13 \cdot 25/24}} = 1{,}961 > 1{,}645 = z(0{,}05), p = 2{,}5\ \%$$

poznámky k výpočtu

- nezapomenout vyloučit nulové rozdíly
- shodným absolutním hodnotám rozdílům přiřadíme jejich průměrné pořadí
- Excel nám v takovém případě moc nepomůže, protože řeší problém shod nestandardně, např.:

X_i	4	-2	5	2	-6	-4	2	7
$\lvert X_i \rvert$	4	2	5	2	6	4	2	7
R_i^+	4,5	2	6	2	7	4,5	2	8
Excel	4	1	6	1	7	4	1	8

- v tabulce patrné nestandardní chování Excelu
- [wilcox.test(pokles,alternative="greater")]

párový znaménkový (sign) test

- hodnotí pouze **počet** kladných a záporných rozdílů, nezáleží na tom, jak jsou rozdíly veliké (slabší test než Wilcoxonův)
- H_0 : Y_i, Z_i mají stejné rozdělení; za hypotézy očekáváme, že počty kladných a záporných X_i jsou podobné
- označme Y počet kladných X_i z celkem n nenulových, za hypotézy $Y \sim \mathrm{bi}(n, 1/2)$
- přibližné rozhodování (centrální limitní věta)

$$Z = \frac{Y - n/2}{\sqrt{n/4}} = \frac{2Y - n}{\sqrt{n}}, \text{ zamítat pro } |Z| \geq z(\alpha/2)$$

- při jednostranné alternativě porovnáme Z a $z(\alpha)$

poznámky

- pro znaménkový test není třeba znát hodnoty Y_i, Z_i, stačí vědět, která z možností $Y_i > Z_i$, $Y_i < Z_i$, $Y_i = Z_i$ nastala
- náš příklad o možném poklesu potratovosti ($n = 12, Y = 8$)

$$Z = \frac{2 \cdot 8 - 12}{\sqrt{12}} = 1{,}155, \quad p = \mathrm{P}(Z > 1{,}155) = 0{,}124$$

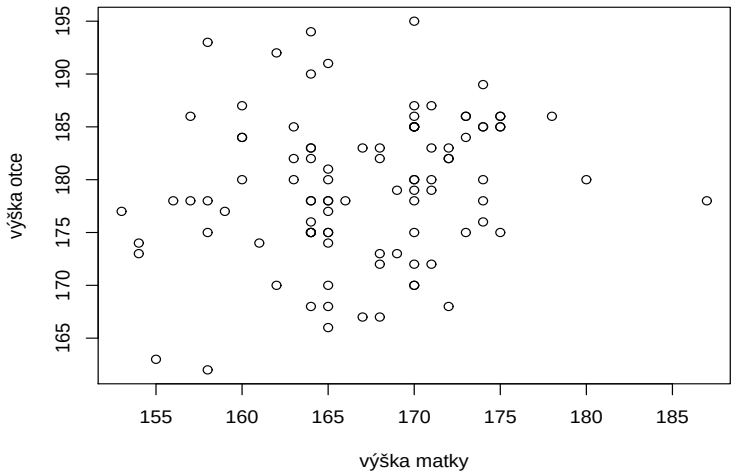
- při malých hodnotách n (do 30) se doporučuje Yatesova korekce

$$Z_{\text{Yates}} = \frac{|2Y - n| - 1}{\sqrt{n}} \operatorname{sign}(2Y - n)$$

- náš příklad (Yatesova korekce, jiným způsobem přesně p =0,194)

$$Z = \frac{|2 \cdot 8 - 12| - 1}{\sqrt{12}} \cdot 1 = 0{,}866, \quad p = 1 - \Phi(0{,}866) = 0{,}193$$

souvisí spolu výšky rodičů?



prokazování závislosti spojitých veličin

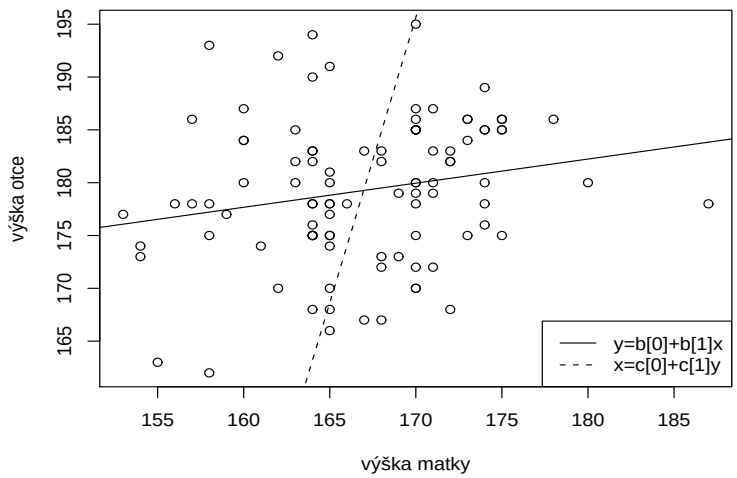
- víme, že pro nezávislé X, Y je $\rho_{X,Y} = 0$
- r_{xy} je odhadem $\rho_{X,Y}$; jak daleko od nuly musí být r_{xy}, abychom na hladině α prokázali zaávislost X, Y?
- za předpokladu, že X, Y mají normální rozdělení (nebo počet pozorovaných dvojic X_i, Y_i je velký), hypotézu nezávislosti zamítáme pokud je $|T| \geq t_{n-2}(\alpha)$, kde

$$T = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

příklad: výšky rodičů

- pro $n = 99$ dvojic byl spočítán korelační koeficient $r = 0{,}205$;
- $$T = \frac{0{,}205}{\sqrt{1 - 0{,}205^2}} \sqrt{97} = 2{,}07 > t_{97}(0{,}05) = 1{,}98$$
- na 5% hladině jsme závislost prokázali
- $t_{97}(0{,}01) = 2{,}63$, tudíž na 1% hladině jsme závislost neprokázali
- výška zpravidla splňuje předpoklad o normálním rozdělení
- [cor.test(vyska.m+vyska.o,data=Kojeni)] [CORREL(x;y)] (pouze výpočet korelačního koeficientu)
- není-li normální rozdělení a nemnoho pozorování, raději použít Spearmanův korelační koeficient

příklad: výšky rodičů



Spearmanův korelační koeficient

- místo původních hodnot x_i, y_i používá jejich pořadí R_i, Q_i
- je to vlastně Pearsonův korelační koeficient použitý na pořadí
- výpočet lze upravit, zjednodušit na

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$

- vhodný pro nelineární monotonní **závislost**, nevadí odlehlé hodnoty
- při testování nemusí být normální rozdělení

příklad: alkohol a úmrtnost na cirhózu

země	spotřeba	úmrtnost	R_i	Q_i	$R_i - Q_i$
Finsko	3,9	3,6	1	3	-2
Norsko	4,2	4,3	2	5	-3
Irsko	5,6	3,4	3	2	1
Holandsko	5,7	3,7	4	4	0
Švédsko	6,0	7,2	5	7	-2
Anglie	7,2	3,0	6	1	5
Belgie	10,8	12,3	7	8	-1
Rakousko	10,9	7,0	8	6	2
SRN	12,3	23,7	9	10	-1
Itálie	15,7	23,6	10	9	1
Francie	24,7	46,1	11	11	0

$$r_S = 1 - \frac{6}{11 \cdot 120} \left(2^2 + 3^2 + \ldots\right) = 0{,}773$$

$r = 0{,}956$ zdánlivě mnohem těsnější závislost!

cirhóza jater a spotřeba alkoholu