Statistika

(D360P03Z, D360P03U)

Karel Zvára

29. listopadu 2004

porovnání středních hodnot nezávislých výběrů

- zřejmě $\mathsf{H}_0 : \mu_1 = \mu_2$ (není **žádný** rozdíl mezi populacemi, **nulová** hypotéza)
- alternativy
  - $\mathsf{H}_1 : \mu_1 \neq \mu_2$ (není-li důvod k jednostranné alternativě)
  - $\mathsf{H}_1 : \mu_1 > \mu_2$ (bylo cílem dokázat, že hoši X větší dívek Y)
  - $\mathsf{H}_1 : \mu_1 < \mu_2$ (bylo cílem dokázat, že hoši X menší dívek Y)
- rozhodování založeno na porovnání průměrů $\bar{X}$ a $\bar{Y}$; čím více se liší, tím spíše zamítnout hypotézu
- je třeba porovnat s mírou přesnosti, s jakou rozdíl průměrů $\bar{X} - \bar{Y}$ odhadne skutečný rozdíl populačních průměrů $\mu_1 - \mu_2$

kritické obory dvouvýběrového t-testu

- o hypotéze $\mathsf{H}_0 : \mu_1 = \mu_2$ se rozhoduje pomocí

$$T = \frac{\bar{X} - \bar{Y}}{\widehat{\mathsf{S.E.}}(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

- $\mathsf{H}_1 : \mu_1 \neq \mu_2$ zamítáme pokud $|T| \geq t_{n_1+n_2-2}(\alpha)$
- $\mathsf{H}_1 : \mu_1 > \mu_2$ zamítáme pokud $T \geq t_{n_1+n_2-2}(2\alpha)$
- $\mathsf{H}_1 : \mu_1 < \mu_2$ zamítáme pokud $T \leq -t_{n_1+n_2-2}(2\alpha)$
- s^2 je odhad σ^2 založený na obou výběrech

$$s^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} s_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_Y^2$$

provedení v MS Excelu (stejné rozptyly)

	Excel	Soubor 1	Soubor 2
průměr	Stř. hodnota	139.133	140.833
rozptyl	Rozptyl	42.981	33.788
rozsah	Pozorování	15	12
spol. odhad rozpt.	Společný rozptyl	38.936	
$H_0 : \mu_1 - \mu_2 =$	Hyp. rozdíl stř. hodnot	0	
stupně vol.	Rozdíl	25	
T	t stat	-0.733	
p jednostr. testu	P(T<=t) (1)	0.244	
$t_{n_1+n_2-2}(2\alpha)$	t krit (1)	1.708	
p oboustr. testu	P(T<=t) (2)	0.488	
$t_{n_1+n_2-2}(\alpha)$	t krit (2)	2.060	

při oboustranné alternativě nelze nulovou hypotézu zamítnout

problém nestejných rozptylů

- předpoklad o stejném rozptylu v obou souborech nemusí být ve skutečnosti splněn
- lze jej ověřit porovnáním odhadů rozptylu F-testem $F = \frac{s_X^2}{s_Y^2}$
- hypotéza $\mathsf{H}_0 : \sigma_1^2 = \sigma_1^2$ se proti $\mathsf{H}_1 : \sigma_1^2 \neq \sigma_2^2$ zamítá, když je

  buď $F = \frac{s_X^2}{s_Y^2} \geq F_{n_1-1,n_2-1}(\alpha/2)$ nebo $\frac{1}{F} = \frac{s_Y^2}{s_X^2} \geq F_{n_2-1,n_1-1}(\alpha/2)$
- vlastně se větší odhad rozptylu dělí menším odhadem, k tomu se musí zvolit správné pořadí stupňů volnosti a hladina
- Excel: uvádí kritickou hodnotu a p-hodnotu pro jednostrannou alternativu; p-hodnotu je třeba vynásobit dvěma

MS Excel: Dvouvýběrový F-test pro rozptyl

		Soubor 1	Soubor 2
průměr	Stř. hodnota	139.13	140.83
rozptyl	Rozptyl	42.98	33.79
rozsah	Pozorování	15	12
stupně vol.	Rozdíl	14	11
F	F	1.27	
p	P$(F <= f)$ (1)	0.349	
	F krit (1)	2.739	

ve skutečnosti je P$(F > 1{,}27) = 0{,}349$, takže $p = 2 \cdot 0{,}349 = 0{,}698$
pro oboustrannou alternativu bylo použíto $F_{14,11}(0{,}025) = 3{,}359$

dvouvýběrový t-test při nestejných rozptylech

- není-li udržitelný předpoklad o stejných rozptylech, lze použít přibližný t-test (Welchův, jiný odhad S.E.$(\bar{X} - \bar{Y})$)

$$T = \frac{\bar{X} - \bar{Y}}{\widehat{\text{S.E.}}(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}}$$

- kde $s_{\bar{X}-\bar{Y}}$ je střední chyba $\bar{X} - \bar{Y}$

$$s_{\bar{X}-\bar{Y}} = \sqrt{v_1 + v_2} \qquad v_1 = \frac{s_X^2}{n_1} \qquad v_2 = \frac{s_Y^2}{n_2}$$

- H_0 se zamítá, je-li $|T| \geq t_f(\alpha)$, kde $f = \dfrac{s^4_{\bar{X}-\bar{Y}}}{\frac{v_1^2}{n_1-1} + \frac{v_2^2}{n_2-1}}$

- náš příklad $T = -0{,}713,\ f = 24{,}69,\ t_f(0{,}05) = 2{,}061,\ p = 0{,}482$

provedení v MS Excelu (nestejné rozptyly)

		Soubor 1	Soubor 2
průměr	Stř. hodnota	139.133	140.833
rozptyl	Rozptyl	42.981	33.788
rozsah	Pozorování	15	12
$H_0 : \mu_1 - \mu_2 =$	Hyp. rozdíl stř. hodnot	0	
stupně vol. f	Rozdíl	25	
T	t stat	-0.713	
p jednostr. testu	P(T<=t) (1)	0.241	
$t_f(2\alpha)$	t krit (1)	1.708	
p oboustr. testu	P(T<=t) (2)	0.482	
$t_f(\alpha)$	t krit (2)	2.060	

při oboustranné alternativě nelze nulovou hypotézu zamítnout

souvislost s bodově biseriálním korel. koef.

- bodově biseriální korelační koeficient vypovídá o síle závislosti mezi spojitou a nula-jedničkovou veličinou (v současném označení)

$$r_{\text{bis}} = \frac{\bar{X} - \bar{Y}}{s_{\text{all}}} \sqrt{\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)}}$$

- pomocí r_{bis} lze vyjádřit testovou statistiku dvouvýběrového t-testu

$$T = \frac{r_{\text{bis}}}{\sqrt{1 - r_{\text{bis}}^2}} \sqrt{n - 2}$$

- stejný vztah platí i pro (Pearsonův) korelační koeficient r
- náš příklad $r_{\text{bis}} = -0{,}139$

porovnání podílů (příklad)

- podíl matek, které označily těhotenství za plánované:
  matky jen se základním vzděláním: 14 z 34 (41,2 %)
  matky aspoň s maturitou: 44 z 65 (67,7 %)
  matky bez ohledu na vzdělání: 58 z 99 (58,6 %)
- $n_1 = 34, f_1 = 0{,}412, n_2 = 65, f_2 = 0{,}677, f = 0{,}586$
- nutno odhadnout rozptyl $f_1 - f_2$:

$$\widehat{\mathrm{var}}\,(f_1 - f_2) = \widehat{\mathrm{var}}\,f_1 + \widehat{\mathrm{var}}\,f_2 = f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$Z = \frac{0{,}412 - 0{,}677}{\sqrt{0{,}586 \cdot 0{,}414\left(\frac{1}{34} + \frac{1}{65}\right)}} = -2{,}54 \qquad p = 2(1 - \Phi(|-2{,}54|)) = 1{,}1\ \%$$

porovnání podílů

- dva nezávislé výběry
  Y_1 absolutní četnost jevu v prvním výběru rozsahu n_1
  Y_2 absolutní četnost jevu ve druhém výběru rozsahu n_2
- hypotéza $\mathsf{H}_0 : \pi_1 = \pi_2$, tj. podíly jevu v obou populacích stejné
- statistika Z porovnává relativní četnosti $f_1 = Y_1/n_1, f_2 = Y_2/n_2$

$$Z = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad f = \frac{Y_1 + Y_2}{n_1 + n_2}$$

$\mathsf{H}_1 : \pi_1 \neq \pi_2$ zamítat pro $|Z| \geq z(\alpha/2)$
$\mathsf{H}_1 : \pi_1 < \pi_2$ zamítat pro $Z \leq -z(\alpha)$
$\mathsf{H}_1 : \pi_1 > \pi_2$ zamítat pro $Z \geq z(\alpha)$

souvislost s čtyřpolním korel. koef.

výběr	výskyt jevu		celkem
	ano	ne	
1	Y_1	$n_1 - Y_1$	n_1
2	Y_2	$n_2 - Y_2$	n_2
celkem	$Y_1 + Y_2$	$n_1 + n_2 - Y_1 - Y_2$	$n_1 + n_2$

- pomocí čtyřpolního korelačního koeficientu $r_{2,2}$ lze Z zapsat jako $Z = \sqrt{n_1 + n_2}\ r_{2,2}$
- příklad

14	20	34
44	21	65
58	41	99

$r_{2,2} = \frac{14 \cdot 21 - 44 \cdot 20}{\sqrt{34 \cdot 65 \cdot 58 \cdot 41}} = -0{,}256,$

$Z = \sqrt{99} \cdot (-0{,}256) = -2{,}54,\ \chi^2 = 99 \cdot (-0{,}256)^2 = 6{,}47,\ p = 1{,}1\ \%$

Mannův-Whitneyův (Wilcoxonův) test

- 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 . . .)
- postup založen na pořadí bez ohledu na výběr
- idea: kdyby nebyl mezi populacemi rozdíl, byla by průměrná pořadí v obou výběrech podobná

příklad: potratovost (Čechy vers. Morava)

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
- 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, použitím centrální limitní věty

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

- za hypotézy (není rozdíl mezi populacemi) je $Z \sim \mathrm{N}(0, 1)$
- hypotézu zamítáme, je-li $|Z| \geq z(\alpha/2)$
- náš příklad:

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

- na 5% hladině jsme prokázali rozdíl

přesný výpočet p-hodnoty

- zajímá nás, nakolik je náš výsledek ($W_1 = 77, W_2 = 14$) výjimečný
- máme celkem $n_1 + n_2 = 13$ pozorování, čtyři z nich (Morava) lze vybrat celkem $\binom{13}{4} = 715$ způsoby
- kolik z nich 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

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

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

párové testy

- předpoklad **nezávislosti** porovnávaných výběrů musí opravdu být splněn, jinak dostaneme nesmysl
- typické porušení předpokladu nezávislosti je u párových dat
  - měření na stejných objektech ve dvou různých časech
  - měření na stejných objektech před zásahem a po něm (ošetření)
  - měření na rodičích
- postup
  - spočítají se a hodnotí rozdíly (změny)
  - přejde se k úloze s jediným výběrem
  - mají-li rozdíly normální rozdělení, pak párový t-test

příklad: výška rodičů

- rozhodnout o tvrzení, že populační průměr výšek otců je o 10 cm větší než populační průměr výšek matek
- otcové: $\bar{Y} = 179{,}26, s_Y = 6{,}78, n_1 = 99$
  matky: $\bar{Z} = 166{,}97, s_Z = 6{,}11, n_2 = 99$
- otcové jsou (ve výběru) v průměru o $\bar{Y} - \bar{Z} =$ 12,29 cm vyšší
  směrodatná odchylka **rozdílů** je 8,14 (méně, než kdyby byly výšky rodičů nezávislé . . . $6{,}78^2 + 6{,}78^2 = 9{,}13^2$)
  střední chyba rozdílu průměrů je $8{,}14/\sqrt{99} = 0{,}819$
- rozhodneme podle statistiky

$$T = \left|\frac{12{,}29 - 10}{0{,}819}\right| = 2{,}801 > 1{,}984 = t_{98}(0{,}05/2) \qquad p = 0{,}6\ \%$$

párový t-test:

- nechť $(Y_1, Z_1) \ldots, (Y_n, Z_n)$ nezávislé dvojice, $X_i = Y_i - Z_i$
- nechť $X_i \sim \mathsf{N}\left(\mu, \sigma^2\right)$
- neznámé $\sigma > 0$ odhadneme pomocí $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- $\mathsf{H}_0 : \mu = \mu_0$ (μ_0 známá konstanta, zpravidla 0)

$$T = \frac{\bar{X} - \mu_0}{\widehat{\mathsf{S.E.}}(\bar{X})} = \frac{\bar{X} - \mu_0}{s}\sqrt{n}$$

- hypotézu H_0 zamítáme (kritický obor):
  - $\mathsf{H}_1 : \mu \neq \mu_0$ (oboustranná alternativa) $\quad |T| \geq t_{n-1}(\alpha)$
  - $\mathsf{H}_1 : \mu > \mu_0$ (jednostranná alternativa) $\quad T \geq t_{n-1}(2\alpha)$
  - $\mathsf{H}_1 : \mu < \mu_0$ (jednostranná alternativa) $\quad T \leq -t_{n-1}(2\alpha)$

příklad: klesá potratovost?

Y_i	24.7	25.7	31.6	24.3	26.8	30.6	21.1	23.5	26.9	22.5	23.1	24.9
Z_i	23.1	23.6	27.9	22.2	23.4	27.9	21.5	26.0	24.3	23.9	21.2	25.7
X_i	1.6	2.1	3.7	2.1	3.4	2.7	-0.4	-2.5	2.6	-1.4	1.9	-0.8
R_i^+	4	6	12	7	11	10	1	8	9	3	5	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 64)
  průměrné pořadí ze 4 záporných rozdílů 3,5 (součet 14)

párový ilcoxonův test

- nechť $(Y_1, Z_1) \ldots, (Y_n, Z_n)$ nezávislé dvojice, $X_i = Y_i - Z_i$
- $\mathrm{H}_0 : Y_i, Z_i$ mají stejné rozdělení (populace jsou stejné)
- mají-li Y_i, Z_i stejné rozdělení, pak $X_i = Y_i - Z_i$ jsou symetricky rozdělena 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 součet pořadí původně kladných hodnot X_i
  - podle W rozhodnout

rozhodování

- na základě centrální limitní věty lze použít

$$Z = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

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

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