Statistika

(D360P03Z, D360P03U)

akademický rok 2004/2005

Karel Zvára

3. ledna 2005

multinomické rozdělení

- zobecnění binomického rozdělení na k-tici náhodných veličin $X_1, \ldots, X_k$
- parametry $n, \pi_1, \ldots, \pi_k$ $(0 < \pi_j < 1, \quad \pi_1 + \ldots + \pi_k = 1)$
- n **nezávislých** pokusů
- v každém pokusu **právě jeden** z k možných výsledků
- j-tý výsledek s pravděpodobností π_j
- X_j – počet pokusů, v nichž nastal j-tý možný výsledek, tedy nutně

$$X_1 + \ldots + X_k = n$$

příklady multinomického rozdělení

- předvolební průzkum
  - n – počet tázaných
  - π_j – skutečný podíl voličů j-té strany v populaci
  - X_j – počet (četnost) voličů j-té strany ve výběru
- hody hrací kostkou
  - n – počet hodů
  - $\pi_1, \ldots, \pi_6$ – pravděpodobnosti jednotlivých stran kostky
  - $X_1, \ldots, X_6$ – absolutní četnosti jednotlivých stran kostky

vlastnosti multinomického rozdělení

- každá složka má binomické rozdělení: $X_j \sim \text{bi}\big(n, \pi_j\big)$
- střední hodnota: $\mu_{X_j} = n\pi_j$, rozptyl: $\sigma^2_{X_j} = n\pi_j(1 - \pi_j)$
- (pro zajímavost) kovariance: $\text{cov}(X_j, X_t) = -n\pi_j\pi_t \quad j \neq t$
- asymptotická vlastnost **chí-kvadrát** (velká n, $n\pi_j \geq 5$)

$$\chi^2 = \sum_{j=1}^{k} \frac{(X_j - n\pi_j)^2}{n\pi_j} \sim \chi^2_{k-1}$$

- X_j – **empirické četnosti**,
  $n\pi_j$ – **očekávané (teoretické) četnosti**

příklad: hrací kostka A

- test **jednoduché** hypotézy
- $n = 100$ hodů kostkou
- $X_1 = 12, X_2 = 21, X_3 = 14, X_4 = 15, X_5 = 21, X_6 = 17$
- očekávané četnosti za hypotézy $\pi_1 = \ldots = \pi_6 = 1/6$:
  $n\pi_1 = \ldots = n\pi_6 = 100/6 = 16{,}67$

$$\chi^2 = \frac{(12-16{,}67)^2}{16{,}67} + \ldots + \frac{(17-16{,}67)^2}{16{,}67} = 4{,}16 < \chi^2_5(0{,}05) = 11{,}07$$

$$p = 52{,}7\ \%$$

příklad: hrací kostka B (1)

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- očekávané četnosti za hypotézy $\pi_1 = \ldots = \pi_6 = 1/6$:
  $n\pi_1 = \ldots = n\pi_6 = 100/6 = 16{,}67$

$$\chi^2 = \frac{(15-16{,}67)^2}{16{,}67} + \ldots + \frac{(41-16{,}67)^2}{16{,}67} = 48{,}32 > \chi^2_5(0{,}05) = 11{,}07$$

- zřejmě je nutno zamítnout hypotézu, že kostka je symetrická

příklad: hrací kostka B (2)

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- nulová hypotéza: $\pi_1 = \ldots = \pi_5 = 1/10, \pi_6 = 5/10 = 1/2$
- očekávané četnosti za hypotézy:
  $n\pi_1 = \ldots = n\pi_5 = 100/10 = 10, n\pi_6 = 100/2 = 50$

$$\chi^2 = \frac{(15-10)^2}{10} + \ldots + \frac{(15-10)^2}{10} + \frac{(41-50)^2}{50} = 12{,}72 > \chi^2_5(0{,}05) = 11{,}07$$

- zřejmě je nutno zamítnout i tuto hypotézu

příklad: hrací kostka B (3)

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- nulová hypotéza: $\pi_6 = 5/10 = 1/2$
- hypotéza o jediné četnosti – binomické rozdělení
- na 6. přednášce jsme určili přibližný 95% interval spolehlivosti pro pravděpodobnost šestky: (0,31; 0,51)
- 1/2 je v tomto intervalu, na 5% hladine **nelze** zamítnout

test homogenity r výběrů

- $X_{i1}, \ldots, X_{ik}$ i-tý výběr z multinomického rozdělení s parametry $n_{i\bullet}, \pi_{i1}, \ldots, \pi_{ik} \qquad (i = 1, \ldots, r)$
- H_0 : pravděpodobnosti jsou ve všech srovnávaných populacích stejné: $\pi_{i1} = \pi_1, \ldots, \pi_{ik} = \pi_k$ (nezávisí na populaci)
- četnosti uspořádáme do kontingenční tabulky
  - n_{ij} – počet j-tých výsledků v i-tém výběru
  - $n_{i\bullet} = \sum_j n_{ij}$ jsou řádkové marginální četnosti (rozsahy výběrů)
  - $n_{\bullet j} = \sum_i n_{ij}$ jsou sloupcové marginální četnosti (četnosti možných výsledků bez ohledu na výběr)
  - $n = \sum_i n_{i\bullet} = \sum_j n_{\bullet j} = \sum_i \sum_j n_{ij}$ je celkový počet pozorování
- očekávané četnosti tak budou $o_{ij} = n_{i\bullet}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} n_{\bullet j}}{n}$

test homogenity r výběrů

- empirické četnosti porovnáme s četnostmi očekávanými

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - o_{ij})^2}{o_{ij}}$$

- platí-li hypotéza, má výsledné chí-kvadrát rozdělení chí-kvadrát s $(r-1)(k-1)$ stupni volnosti
- hypotézu o shodě pravděpodobností v r populacích zamítáme, je-li $\chi^2 \geq \chi^2_{(r-1)(k-1)}(\alpha)$

mají obě kostky stejné šestice pravděpodobností?

- empirické četnosti (kontingenční tabulka)

A	12	21	14	15	21	17	100
B	15	16	7	6	15	41	100
	27	37	21	21	36	58	200

- očekávané četnosti (za hypotézy): 27·100/200=13,5, ...

A	13,5	18,5	10,5	10,5	18	29	100
B	13,5	18,5	10,5	10,5	18	29	100
	27	37	21	21	36	58	200

$$X^2 = \frac{(12-13{,}5)^2}{13{,}5} + \frac{(21-18{,}5)^2}{18{,}5} + \ldots + \frac{(41-29)^2}{29} = 18{,}13 > 11{,}07 = \chi^2_5(0{,}05)$$

hypotézu o shodě pstí na obou kostkách **zamítáme** ($p = 0{,}3\ \%$)

příklad: vzdělání matek

	vzdělání			
porodnice	základní	střední	VŠ	celkem
Praha	23 (24,0)	30 (33,2)	17 (12,7)	70
venkov	11 (10,0)	17 (13,8)	1 (5,3)	29
celkem	34	47	18	99

- v závorce jsou očekávané četnosti za hypotézy, že podíly tří vzdělanostních skupin jsou v obou populacích shodné

$$\frac{(23-24)^2}{24} + \frac{(30-33{,}2)^2}{33{,}2} + \frac{(17-12{,}7)^2}{12{,}7} + \frac{(11-10)^2}{10} + \frac{(17-13{,}8)^2}{13{,}8} + \frac{(1-5{,}3)^2}{5{,}3}$$

$$\chi^2 = 6{,}12 > \chi^2_2(0{,}05) = 5{,}99, \qquad p = 4{,}7\ \%$$

- bylo třeba $o_{ij} \geq 5$ pro všechna i, j

test nezávislosti

- vyšetřujeme **současně** dva znaky v nominálním měřítku u n nezávislých statistických jednotek
- n_{ij} je počet jednotek, kde je současně i-tá hodnota prvního znaku a j-tá hodnota druhého znaku
- celkem je i-tá hodnota prvního znaku u $n_{i\bullet} = \sum_j n_{ij}$ jednotek, j-tá hodnota druhého znaku u $n_{\bullet j} = \sum_i n_{ij}$ jednotek
- kdyby byly znaky nezávislé, byl by pro každou hodnotu jednoho znaku poměr mezi četnostmi hodnot druhého znaku podobný, proto očekávané četnosti $o_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ (podmíněné psti stejné)
- výpočet χ^2 a jeho hodnocení stejné jako u homogenity

příklad: plánovaná těhotenství

- u každé matky zjišťovány dva znaky: dosažené vzdělání, zda těhotenství plánováno

vzdělání	základní	střední	VŠ	celkem
neplánováno	20 (14,1)	16 (19,5)	5 (7,5)	41
plánováno	14 (19,9)	31 (27,5)	13 (10,5)	58
celkem	34	47	18	99

- všechny očekávané četnosti dostatečně velké

$$\chi^2 = 6{,}68 > 5{,}99 = \chi^2_2(0{,}05), \qquad p = 3{,}5\ \%$$

čtyřpolní tabulka

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

$$\chi^2 = \frac{n(a \cdot d - b \cdot c)^2}{(a+b)(c+d)(a+c)(b+d)}$$

úmrtí na vnější příčiny v roce 2003

příčina	muži	ženy	celkem
dopr. nehody	1097 (1130,3)	362 (328,7)	1459
sebepoškození	1365 (1331,7)	354 (387,3)	1719
celkem	2462	716	3178

$$\chi^2 = \frac{3178(1097 \cdot 354 - 362 \cdot 1365)^2}{1459 \cdot 1719 \cdot 2462 \cdot 716} = 8{,}05 > \chi^2_1(0{,}05) = 3{,}84 \qquad p = 0{,}5\ \%$$

prokázali jsme neshodu mezi muži a ženami, souvislost s pohlavím vlastně prokázali neshodu podílů při nehodách (odhady 44,6 %, 50,6 %)

zobecnění dvouvýběrového t-testu

(požadoval normální rozdělení v aspoň dvou nezávislých výběrech)

- příklad: souvisí výška otce se vzděláním matky?
  (součet čtverců = součet čtverců odchylek od průměru)

vzdělání	rozsah	průměr	součet čtverců	směr. odch.
základní	34	177,1	1188,7	6,0
střední	47	179,5	1909,8	6,4
VŠ	19	182,8	1027,1	7,8
celkem	99	179,3	4511,2	6,8

- **součet** součtů čtverců 1188,7+1909,8+1027,1=4125,6, kdežto když nehledíme na vzdělání, tak 4511,2, rozdíl je 385,6
- je to dost k tomu, aby bylo prokázáno, že se tři skupiny liší populačním průměrem?

model analýzy rozptylu (r **nezávislých** výběrů)

$$X_{11}, X_{12}, \ldots, X_{1n_1} \sim \mathrm{N}\left(\mu_1, \sigma^2\right) \quad \text{první výběr (skupina)}$$
$$X_{21}, X_{22}, \ldots, X_{2n_2} \sim \mathrm{N}\left(\mu_2, \sigma^2\right) \quad \text{druhý výběr (skupina)}$$
$$\ldots$$
$$X_{r1}, X_{r2}, \ldots, X_{rn_r} \sim \mathrm{N}\left(\mu_r, \sigma^2\right) \quad r\text{-tý výběr (skupina)}$$

- zobecnění hypotézy dvouvýběrového t-testu $\mathrm{H}_0 : \mu_1 = \ldots = \mu_r$
- rozhoduje se na základě rozkladu součtu čtverců

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{i=1}^{r} n_i(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet})^2$$

variabilita: celková = mezi výběry + uvnitř výběrů

označení: $\bar{X}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, $\bar{X}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} X_{ij}$

$$\begin{aligned}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\bullet\bullet})^2 &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left((X_{ij} - \bar{X}_{i\bullet}) + (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})\right)^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet})^2 + \sum_{i=1}^{r} n_i(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2\end{aligned}$$

protože

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet})(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}) = \sum_{i=1}^{r}(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet}) = 0$$

tabulka analýzy rozptylu

variabilita	součet čtv.	st. vol.	prům. čtv.	F	p
mezi výběry	385,6	2	192,8	4,49	0,014
uvnitř výběrů	4125,6	96	43,0		
celková	4511,2	98			

- součet čtverců uvnitř skupin = součet součtů čtverců, stupně volnosti v řádku uvnitř skupin: $n - r$
- průměrné čtverce: vždy součet čtverců/stupně volnosti; platí-li H_0, pak obojí průměrné čtverce podobné; čím je prům. čtverec mezi skupinami větší, tím spíš H_0 zamítnout
- H_0 zamítnout, je-li $F \geq F_{r-1,n-r}(\alpha)$ (tabelováno)
- příklad: $F_{2,96}(0{,}05) = 3{,}09$, na 5% hladině jsme **prokázali**, že výška otců souvisí se vzděláním matek (ale hladina testu)

analýza rozptylu – předpoklady

- r **nezávislých** výběrů (nelze obejít)
- stejné rozptyly ve všech skupinách (lze testovat)
- normální rozdělení (lze ověřit stejně jako u regrese, pomocí reziduí)
- někdy pomůže transformace, např. logaritmy hodnocené veličiny
- Kruskalův-Wallisův test – zobecnění dvouvýběrového Wilcoxonova testu: hodnotí se pořadí místo původních hodnot, zda jsou průměrná pořadí podobná

příklad: věk matky $\sim$ vzdělání matky

vzdělání	rozsah n_i	průměr $\bar{X}_{i\bullet}$	prům. pořadí	součet pořadí T_i
základní	34	23,4	30,1	1025
střední	47	26,3	55,7	2618
VŠ	18	28,5	72,6	1307

- rozhoduje se pomocí upravené F-statistiky z analýzy rozptylu:

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{T_i^2}{n_i} - 3(n+1)$$

- kde T_i je součet pořadí i-té skupiny
- kritický obor: $Q \geq \chi^2_{r-1}(\alpha)$ (dostatečně velká n_i)
- $Q = 29{,}48 > 5{,}99, \qquad p < 0{,}0001$