Statistika

(MD360P03Z, MD360P03U)
ak. rok 2007/2008

Karel Zvára

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

(naposledy upraveno 11. prosince 2007)

příklad: je výběr reprezentativní?

- bylo provedeno šetření mezi ženami ve věku 18 až 50 roků
- mezi 498 náhodně oslovenými ženami bylo celkem 180 žen svobodných, 239 žen vdaných, 75 žen rozvedených a 4 ovdovělé
- stejné údaje v procentech: 36,14 % svobodných, 47,99 % vdaných, 15,06 % rozvedených, 0,80 % ovdovělých
- je známo, že v celé populaci žen v ČR uvedeného věkového rozpětí je 34,27 % svobodných, 52,02 % vdaných, 12,50 % rozvedených a 1,20 % ovdovělých
- lze výběr považovat za reprezentativní?

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
- krevní skupiny
  - n=4 (skupiny 0, A, B, AB)
  - $\pi_0, \pi_A, \pi_B, \pi_{AB}$ – psti skupin 0, A, B, AB
  - X_0, X_A, X_B, X_{AB} – počty osob se skupinami 0, A, B, AB

vlastnosti multinomického rozdělení

- každá složka má binomické rozdělení: $X_j \sim \text{bi}(n, \pi_j)$
- 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
- $X_1 = 12,\ X_2 = 21,\ X_3 = 14,\ X_4 = 15,\ X_5 = 21,\ X_6 = 17$
- hypotéza $H_0 : \pi_1 = \ldots = \pi_6 = 1/6$ dá očekávané četnosti
  $n\pi_1 = \ldots = n\pi_6 = 100/6 = 16{,}67$ (vždy více než 5))
- $$\chi^2 = \frac{(12-16{,}67)^2}{16{,}67} + \ldots + \frac{(17-16{,}67)^2}{16{,}67} = 4{,}16$$
- $$\chi^2 < \chi^2_5(0{,}05) = 11{,}07, \qquad p = 52{,}7\ \%$$
- neprokázali jsme, že by kostka nebyla symetrická
- [chisq.test(c(12,21,14,15,21,17),p=rep(1,6)/6)]

příklad: hrací kostka B (1)

- $n = 100$ hodů kostkou
- $X_1 = 15, X_2 = 16, X_3 = 7, X_4 = 6, X_5 = 15, X_6 = 41$
- hypotéza $H_0 : \pi_1 = \ldots = \pi_6 = 1/6$ dá očekávané četnosti
  $n\pi_1 = \ldots = n\pi_6 = 100/6 = 16{,}67$

$$\chi^2 = \frac{(15-16{,}67)^2}{16{,}67} + \ldots + \frac{(41-16{,}67)^2}{16{,}67} = 48{,}32$$

- $$\chi^2 > \chi^2_5(0{,}05) = 11{,}07$$
- zřejmě je nutno zamítnout hypotézu, že kostka je symetrická
- na 5% hladině jsme prokázali, že není symetrická

příklad: hrací kostka B (2), jiná H_0

- $n = 100$ 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:
  $n\pi_1 = \ldots = n\pi_5 = 100/10 = 10, n\pi_6 = 100/2 = 50$

$$\chi^2 = \frac{(15-10)^2}{10} + \ldots + \frac{(15-10)^2}{10} + \frac{(41-50)^2}{50} = 12{,}72$$

- $$\chi^2 > \chi^2_5(0{,}05) = 11{,}07$$
- zřejmě je nutno zamítnout i tuto hypotézu
  [chisq.test(c(15,16,7,6,15,41),p=c(1,1,1,1,1,5)/10)]

příklad: hrací kostka B (3) (použít jen část informace)

- $n = 100$ hodů kostkou
- $X_6 = 41$
- nulová hypotéza: $\pi_6 = 5/10 = 1/2$
- hypotéza o psti jediného z možných výsledků (pst šestky) – binomické rozdělení
- dříve jsme určili přibližný 95% interval spolehlivosti pro pravděpodobnost šestky: (0,31; 0,51)
- $1/2$ je v tomto intervalu, na 5% hladině **nelze** zamítnout

[binom.test(41,100)]

příklad: je výběr reprezentativní?

- provedeme test hypotézy, že pravděpodobnosti čtyř skupin žen jsou rovny procentům v populaci

	svobodné	vdané	rozvedené	ovdovělé	celkem
populace	34,27 %	52,02 %	12,50 %	1,20 %	100 %
výběr	180	239	75	4	498
výběr (rel.)	36,14 %	47,99 %	15,06 %	0,80 %	100 %
oček. čet.	170,69	259,07	62,26	5,99	498
přínos	0,51	1,55	2,61	0,66	5,33

$$\frac{(180-170{,}69)^2}{170{,}69}+\frac{(239-259{,}07)^2}{259{,}07}+\frac{(75-62{,}26)^2}{62{,}26}+\frac{(4-5{,}99)^2}{5{,}99}$$

- výsledná hodnota chí-kvadrát je $\chi^2 = 5{,}33$, ale $\chi^2_3(0{,}05) = 7{,}81$
- neprokázali jsme, že by výběr nebyl reprezentativní, můžeme jej za reprezentativní považovat

test homogenity r výběrů

- například, zda mají kostky A, B stejné šestice psti (ať už je ta šestice jakákoliv)
- $X_{i1}, \ldots, X_{ik}$ 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

test homogenity r výběrů

- neznámé pravděpodobnosti π_j odhadneme pomocí marginálních relativních četností $n_{\bullet j}/n$
- očekávané četnosti tak budou $o_{ij} = n_{i\bullet}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} n_{\bullet j}}{n}$
- 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 výsledná statistika χ^2-rozdělení $\chi^2_{(r-1)(k-1)}$
- hypotézu o shodě pravděpodobností v r populacích zamítáme, je-li $\chi^2 \geq \chi^2_{(r-1)(k-1)}(\alpha)$
- je třeba, aby očekávané četnosti byly dost velké, aspoň 5

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

- $$X^2 = \frac{(12-13{,}5)^2}{13{,}5} + \frac{(21-18{,}5)^2}{18{,}5} + \ldots + \frac{(41-29)^2}{29} = 18{,}13$$

- $$\chi^2 > 11{,}07 = \chi^2_5(0{,}05), \qquad p = 0{,}3\ \%$$

- hypotézu o shodě pstí na kostkách A a B **zamítáme**

příklad – vzdělání matek

	porodnice		
vzdělání	Praha	venkov	celkem
základní	23	11	34
střední	30	17	47
VŠ	17	1	18
celkem	70	29	99

	porodnice		
vzdělání	Praha	venkov	celkem
základní	24,0	10,0	34
střední	33,2	13,8	47
VŠ	12,7	5,3	18
celkem	70	29	99

$$\chi^2 = 6{,}12, \qquad p = 4{,}7\ \%$$

- kdyby rozdělení vzdělání bylo všude stejné, očekáváme tři možnosti v poměru 34:47:18 (marg. četnosti!), celkem 99
- pražských 70 matek by stejný poměr dalo při **očekávaných** četnostech 70·34/99=24,0, resp. 70·47/99=33,2 resp. 70·18/99=12,7
- podobně pro matky z venkova dostaneme 9,96, po zaokrouhlení 10,0, pro další četnosti 13,8 resp. 5,3