Statistika

(MD360P03Z, MD360P03U)
ak. rok 2007/2008

Karel Zvára

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

(naposledy upraveno 17. prosince 2007)

příklad: předvolební průzkum

zprávy TV XY	strana A	strana B	celkem
sledoval	**11**	**4**	15
nesledoval	**6**	**9**	15
celkem	17	13	30

zprávy TV XY	strana A	strana B	celkem
sledoval	**73 %**	**27 %**	100 %
nesledoval	40 %	60 %	100 %
celkem	57 %	43 %	100 %

zprávy TV XY	strana A	strana B	celkem
sledoval	**65 %**	**31 %**	50 %
nesledoval	35 %	69 %	50 %
celkem	100 %	100 %	100 %

- 30 voličů bylo dotázáno, které ze dvou stran dají přednost
- souvisí odpovědi se sledováním večerních zpráv na dané TV stanici?
- znamená něco nestejné zastoupení příznivců stran u těch, kteří sledovali?
- znamenají něco nestejné podíly těch, kteří sledovali mezi příznivci dvou stran?

test nezávislosti kvalitativních znaků

- 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 jsou $o_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ (podmíněné psti stejné)
- výpočet χ^2 a jeho hodnocení stejné jako u homogenity

příklad: souvisí plánované těhotenství se vzděláním?

vzdělání	plánované ne	plánované ano	celkem
základní	20	14	34
střední	16	31	47
VŠ	5	13	18
celkem	41	58	99

vzdělání	plánované ne	plánované ano	celkem
základní	58,8 %	42,1 %	100 %
střední	34,0 %	66,0 %	100 %
VŠ	27,8 %	72,2 %	100 %
celkem	41,4 %	58,6 %	100 %

- je souvislost mezi odpověďmi o plánovaném těhotenství a vzděláním matek?
- kdyby byly znaky nezávislé, byly by podmíněné pravděpodobnosti pro jednotlivá vzdělání stejné, tedy jejich odhady by byly podobné
- test vlastně porovnává procenta u jednotlivých vzdělání
- chí-kvadrát test porovnává skutečně zjištěné četnosti s tím, jaké četnosti bychom v průměru očekávali, kdyby platila nulová hypotéza

příklad: plánovaná těhotenství

skutečné četnosti (očekávané četnosti)

vzdělání	plánované ne	plánované ano	celkem
základní	20 (14,08)	14 (19,92)	34
střední	16 (19,46)	31 (27,54)	47
VŠ	5 (7,46)	13 (10,54)	18
celkem	41	58	99

- odhad pravděpodobnosti, že má matka základní vzdělání: $\hat{P}(vzdel = zakladni) = 34/99$
- odhad pravděpodobnosti, že jde o plánované těhotenství: $\hat{P}(tehot = plan) = 58/99$
- **jsou-li** vzdělání a plánovanost **nezávislé**, pak $P((vzdel = zakladni) \cap (tehot = plan)) = P(vzdel = zakladni) \cdot P(tehot = plan) \doteq (34/99) \cdot (58/99)$
- očekávaný počet matek se základním vzděláním a plánovaným těhotenstvím (**za platnosti nulové hypotézy**) odhadneme: $99 \cdot (34/99) \cdot (58/99) = 34 \cdot 58/99 \doteq 19{,}92$

příklad: plánovaná těhotenství

skutečné četnosti (očekávané četnosti)

vzdělání	plánované ne	plánované ano	celkem
základní	20 (14,08)	14 (19,92)	34
střední	16 (19,46)	31 (27,54)	47
VŠ	5 (7,46)	13 (10,54)	18
celkem	41	58	99

$$\chi^2 = \frac{(20-14{,}08)^2}{14{,}08} + \frac{(14-19{,}92)^2}{19{,}92} + \frac{(16-19{,}46)^2}{19{,}46} + \frac{(31-27{,}54)^2}{27{,}54} + \frac{(5-7{,}46)^2}{7{,}46} + \frac{(13-10{,}54)^2}{10{,}54} = 6{,}68$$

příklad: souvisí plánované těhotenství se vzděláním?

- 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

- kdyby nebyla závislost, u každého vzdělání by bylo stejné procento plánovaných těhotenství, totiž 58/99=58,6 %
  - u zákl. vzdělání $x/34 = 58/99$ tedy $x = 34 \cdot 58/99 = 19{,}9$
  - u středního vzdělání $x/47 = 58/99$ tedy $x = 47 \cdot 58/99 = 27{,}5$
  - u vysokoškoláček $x/18 = 58/99$ tedy $x = 18 \cdot 58/99 = 10{,}5$
- všechny očekávané četnosti jsou dostatečně velké

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

příklad: vzdělání snoubenců

ženich	nevěsta základní	nevěsta střední	nevěsta VŠ	celkem
základní	24	12	3	39
střední	7	24	3	34
VŠ	3	9	15	27
celkem	34	45	21	100

- u 100 náhodně vybraných snoubenců bylo zjištěno vzdělání (základní = základní nebo neúplné střední)
- lze považovat vzdělání snoubenců za nezávislá?
- jsou četnosti dost velké?
- nejmenší očekávané četnost (při nezávislosti): $27 \cdot 21/100 = 5{,}67$

příklad: vzdělání snoubenců

ženich	nevěsta základní	nevěsta střední	nevěsta VŠ	celkem
základní	24 (13,2)	12 (17,6)	3 (8,2)	39
střední	7 (11,6)	24 (15,3)	3 (7,1)	34
VŠ	3 (9,2)	9 (12,2)	15 (5,7)	27
celkem	34	45	21	100

- $\chi^2 = 43{,}2 > \chi^2_4(0{,}05) = 9{,}5$, $p < 0{,}1$ %
- na 5 % hladině jsme prokázali závislost
- vzdělání snoubenců nelze považovat za nezávislá
- četnosti na diagonále jsou větší, než očekáváme za nezávislosti
- četnosti daleko od diagonály (velký rozdíl ve vzdělání) jsou menší, než očekáváme za nezávislosti

čtyřpolní tabulka

speciální případ kontingenční tabulky

-

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

- **sílu závislosti** lze měřit ϕ-koeficientem [phi coefficient] (čtyřpolní korelační koeficient)

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

- ϕ je (jako každý korelační koeficient) mezi –1 a 1
- pro

11	4	15
6	9	15
17	13	30

vyjde

$$\phi = \frac{11 \cdot 9 - 4 \cdot 6}{\sqrt{15 \cdot 15 \cdot 17 \cdot 13}} = 0{,}34$$

příklad: předvolební průzkum

- $\phi > 0$ znamená, že četnosti na hlavní diagonále (indexy 1,1 a 2,2) převládají nad četnostmi na vedlejší diagonále (indexy 1,2 a 2,1)
- v našem příkladu

TV XY	strana A	strana B	celkem
sledoval	**11**	**4**	15
nesledoval	**6**	**9**	15
celkem	17	13	30

vychází $\phi = 0{,}34 > 0$
(tedy kladné), protože je $11{\cdot}9 > 6{\cdot}4$

čtyřpolní tabulka – prokazování závislosti

- chí-kvadrát porovnávající teoretické a očekávané četnosti lze upravit na tvar

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

- nezávislost se na hladině α zamítá, je-li $\chi^2 \geq \chi^2_1(\alpha)$
- příklad (předvolební průzkum)

$$\chi^2 = \frac{30 \cdot (11 \cdot 9 - 4 \cdot 6)^2}{15 \cdot 15 \cdot 17 \cdot 13} = 3{,}39 = 30 \cdot 0{,}34^2$$

- závislost jsme na 5% hladině neprokázali, neboť

$$3{,}39 < 3{,}84 = \chi^2_1(0{,}05), \qquad p = 6{,}5\ \%$$

malé očekávané četnosti ve čtyřpolní tabulce

- stále je třeba, aby byly očekávané četnosti dost velké (≥ 5)
- **Yatesova korekce** umožní rozhodnutí i při menších četnostech tím, že zmenší čitatele

$$\chi^2_{\text{Yates}} = \frac{n(|ad-bc|-n/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

- nezávislost se zamítá, je-li opět $\chi^2_{\text{Yates}} \geq \chi^2_1(\alpha)$
- **Fisherův exaktní test** počítá přímo p-hodnotu

příklad: souvislost délky kojení a plánování těhotenství

	Praha a venkov			venkov		
těhot.	neplán.	plán.	celkem	neplán.	plán.	celkem
ve 24. t. nekojí	35	36	71	13	9	22
ve 24. t. kojí	6	22	28	1	6	7
celkem	41	58	99	14	15	29

- bez ohledu na místo: $\chi^2 = 6{,}43$, $p = 1{,}1\ \%$,
  $\chi^2_{\text{Yates}} = 5{,}33$, $p = 2{,}1\ \%$ (nejm. četnost $41 \cdot 28/99 = 11{,}6$)
  Fisherův exaktní test: $p = 1{,}3\ \%$
- venkov: $\chi^2 = 4{,}27$, $p = 3{,}9\ \%$,
  $\chi^2_{\text{Yates}} = 2{,}66$, $p = 10{,}3\ \%$ (nejm. četnost $14 \cdot 7/29 = 3{,}4$)
  Fisherův exaktní test: $p = 8{,}0\ \%$

Simpsonův paradox

dílčí tabulky mohou ukazovat na závislost jiného směru, než jejich součet

venkov	A	B	celkem
sledoval	34	5	39
nesledoval	28	2	30
celkem	62	7	69

město	A	B	celkem
sledoval	4	29	33
nesledoval	6	35	42
celkem	10	64	74

$\phi_{\text{venkov}} = -0{,}10$ $\phi_{\text{město}} = -0{,}04$

celkem	A	B	celkem
sledoval	38	34	72
nesledoval	34	37	71
celkem	72	71	143

$\phi_{\text{celkem}} = 0{,}05$

- po spojení dvou tabulek se záporným ϕ-koeficientem vyšla tabulka s kladným ϕ-koeficientem

závislost mezi nula-jedničkovým a kvantitativním znakem

- dva nezávislé výběry, např. hoši $X_1, \ldots, X_{n_0}$ a dívky $X_{n_0+1}, \ldots, X_{n_0+n_1}$, vždy normální rozdělení jako pro dvouvýběrový t-test
- otázka: jak silně souvisí sledovaná vlastnost a pohlaví?
- označme pohlaví formálně $Y_i = 0$ pro chlapce a $Y_i = 1$ pro děvčata
- korelační koeficient $r_{X,Y}$ mezi těmito veličinami se dá zapsat také jako

$$r_{\text{bis}} = \frac{\bar{X}_1 - \bar{X}_0}{S}\sqrt{\frac{n_0 n_1}{n(n-1)}}$$

- S je směrodatná odchylka spočítaná bez ohledu na pohlaví, $n = n_0 + n_1$ je celkový počet měření v obou výběrech
- r_{bis} **bodově-biseriální korelační koeficient**

příklad: výška desetiletých

- stejná data jako dvouvýběrový test (data ze str. 170)
- $\bar{X}_0 = 139{,}13, \qquad n_0 = 15$
- $\bar{X}_1 = 140{,}83, \qquad n_1 = 12$
- $S^2 = 38{,}18, \qquad S = 6{,}18$
- $$r_{\text{bis}} = \frac{140{,}83 - 139{,}13}{6{,}18}\sqrt{\frac{15 \cdot 12}{15 + 12}} = 0{,}493$$
- H_0 : nezávislost
- má-li X normální rozdělení, lze použít stejný test, jako u korelačního koeficientu; je to ekvivalentní dvouvýběrovému t-testu (při stejných populačních rozptylech

přehled korelačních koeficientů

- základním je (momentový) Pearsonův
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
- když místo hodnot x_i, y_i dosadíme jejich pořadí R_i, Q_i, dostaneme (pořadový) Spearmanův korelační koeficient
$$r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$
- je-li jedna z veličin nula-jedničková, vyjde bodově-biseriální korelační koeficient r_{bis}
- jsou-li obě veličiny nula-jedničkové, dostaneme ϕ-koeficient (čtyřpolní korelační koeficient)

přehled testů o populačních mírách polohy

rozdělení	normální	spojité
populační parametr (o čem je hypotéza)	populační průměr	populační medián
jeden výběr	jednovýběrový t-test	znaménkový Wilcoxon
výběr dvojic	párový t-test	znaménkový Wilcoxon
dva nezávislé výběry	dvouvýběrový t-test	Mann-Whitney