Cvičení 13. Neparametrické testy, testy dobré shody

Testy dobré shody

Název testu	Předpoklady testu	Testová statistika
χ^2 test dobré shody	Očekávané četnosti ≥ 2, alespoň 80% očekávaných četností > 5	$G = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

Analýza závislosti v kontingenční tabulce

Název testu	Předpoklady testu	Testová statistika
Analýza závislosti v kontingenční tabulce	Očekávané četnosti ≥ 2, alespoň 80% očekávaných četností > 5	$K = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

- Koeficient korelace $CC = \sqrt{\frac{K}{K+n}}$ (pro čtvercové kontingenční tabulky)
- Korigovaný koeficient kontingence $CC_{cor} = \frac{CC}{CC_{max}}$, kde $CC_{max} = \sqrt{\frac{min(r;s)-1}{min(r;s)}}$ (pro obdélníkové kontingenční tabulky)
- Cramerův koeficient $V = \sqrt{\frac{K}{n(min(r;s)-1)}}$.

Tyto koeficienty se mohou vyskytovat v intervalu (0;1). Čím blíže jsou 1, tím je závislost mezi X a Y těsnější.

Analýza závislosti v asociační tabulce

- Odhad poměru šancí: $\widehat{OR} = \frac{ad}{bc}$
- Intervalový odhad OR: $\left\langle \widehat{OR} \cdot e^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\cdot z_{1-\frac{\alpha}{2}}} ; \widehat{OR} \cdot e^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\cdot z_{1-\frac{\alpha}{2}}} \right\rangle$
- Odhad relativního rizika: $\widehat{RR} = \frac{a(c+d)}{c(a+b)}$
- Intervalový odhad RR: $\left\langle \widehat{RR} \cdot e^{-\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}}\cdot z_{1-\frac{\alpha}{2}}} ; \widehat{RR} \cdot e^{\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}}\cdot z_{1-\frac{\alpha}{2}}} \right\rangle$

	A_1	A_2	P_B
B_1			
B_2			
P_A			

Příklad č. 1

x_i	1	2	3	4	5	6
O_i	979	1002	1015	980	1040	984
$\tilde{\pi}_i$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
E_i	1000	1000	1000	1000	1000	1000

H_0: kostka je férová $\Rightarrow \pi = \tilde{\pi}$

H_A: $\neg H_0$

χ^2-test

$$G = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 2{,}926$$

$df = 6 - 1 = 5$

p-hodnota $= 1 - F_{\chi^2_5}(6) \doteq 0{,}711$

⇒ na $\alpha = 0{,}05$ nezamítám H_0

⇒ není stat. významně odlišné od férové kostky

Příklad 2.

X_i	0	1	2	3	4 a více
O_i	52	48	36	10	4
$\tilde{\pi}_i$	0,301	0,361	0,217	0,087	0,034
E_i	45,2	54,2	32,5	13,0	5,1

$\mathrm{dpois}(x;\ \lambda t = 1{,}2)$

χ^2 test

předpoklady OK $\cdot E_i \geq 5$

H_0: data pochází z $Poi(\lambda t = 1{,}2)$

H_A: $\neg H_0$

$G = 3{,}03$ $\quad$ $df = 5-1 = 4$

p-hodnota $= 0{,}552$

$\Rightarrow$ nezamítám H_0 $(\alpha = 0{,}05)$

Příklad č. 3

data jsou z $Poi(\lambda t = ?)$

$\underline{E(X) = \underline{\lambda t}}$ $\quad$ $\overline{X} = \dfrac{\sum_{i=1}^{5} x_i \cdot O_i}{\sum_{i=1} O_i} \doteq 1{,}07$

x_i	0	1	2	3	4 a více
O_i	52	48	36	10	4
$\tilde{\pi}_i$	0,331	0,366	0,202	0,075	0,026
E_i	49,6	54,9	30,4	11,2	3,4

H_0: data pocházejí z Poi

H_A: $\neg H_0$

$G = 2{,}15$ $\qquad$ $df = 5-1-1 = 3$

p-hodnota $= 0{,}541$

$\Rightarrow$ nezamítám H_0

Příklad 4.

- OP ? ✓

H_0: data pochází z norm. roz.

H_A: $\neg H_0$

test dobré shody (pearson.test)

p-hodnota << 0,001 (df = 12)

⇒ zamítáme H_0

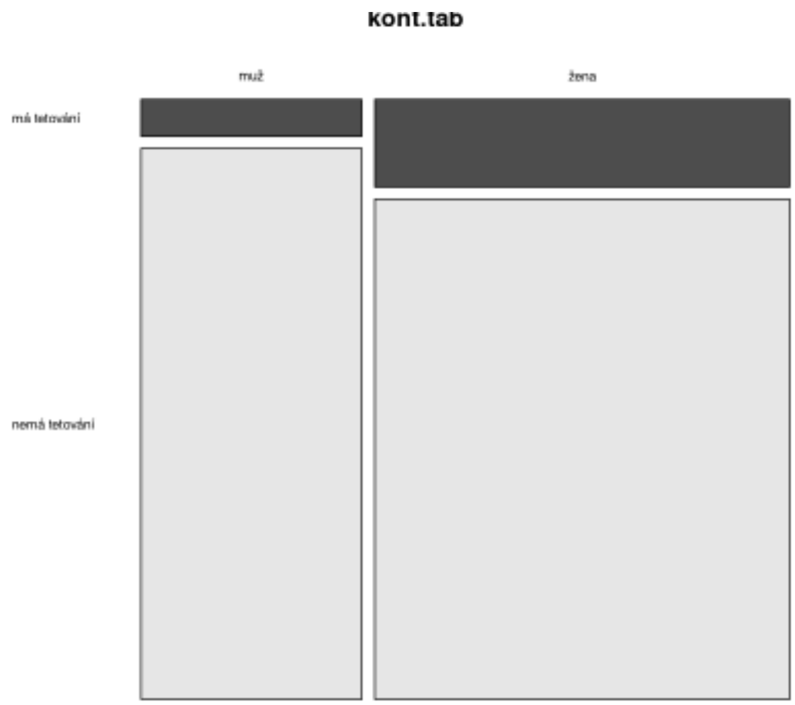
→ data nepochází z norm.

Příklad č. 5:

Explorační posouzení:

Kont. tabulka

		Tetování	
		MÁ	NEMÁ
pohlaví	muž	73	194
	žena	58	331



Cramerovo $V = 0{,}1221$

$\to$ EXPL. posouzení $\Rightarrow$ slabá závislost

H_0: data jsou nezávislá

H_A: $\neg H_0$

(test dobré shody v kont. tab
chisq.test)

očekávané čet.:

	má tetování	nemá tetování
muž	24.6594	182.3406
žena	46.3406	342.6594

E_i

⇒ předpoklady jsou splněny

p-hodnota = 0,003

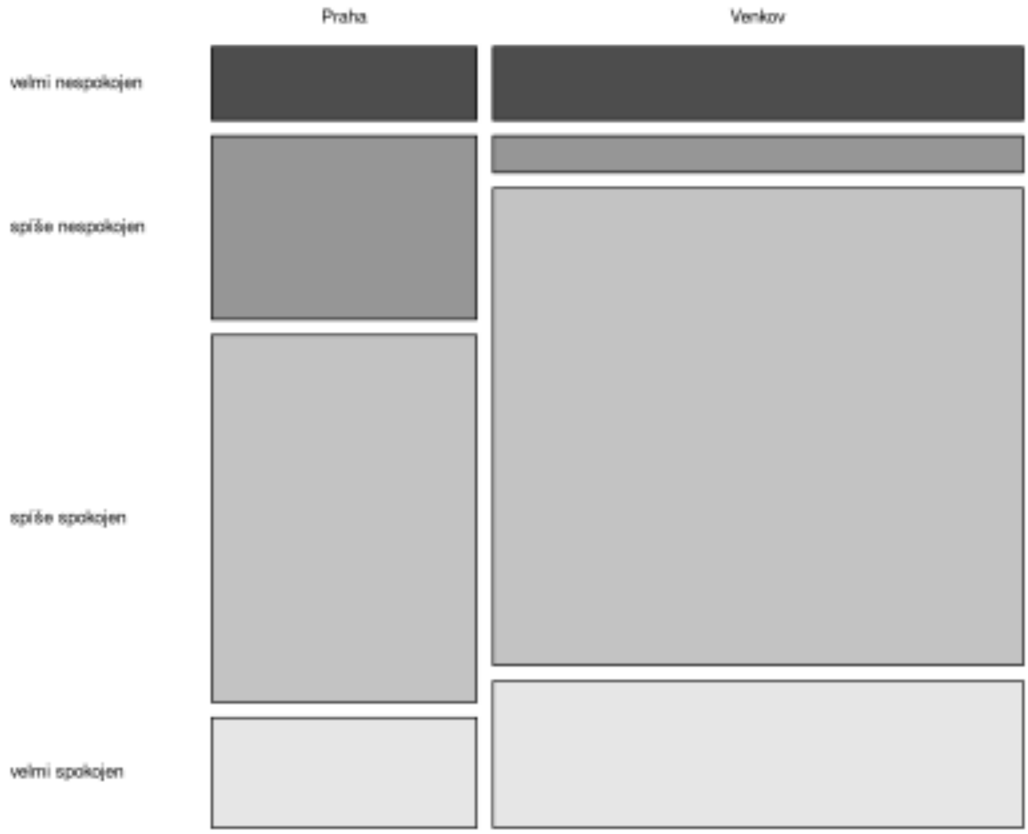
($\alpha = 0,05$) ⇒ zamítnout H_0

⇒ existuje stat. výz. závislost mezi pohlavím a přítomností tet.

(slabá závislost ($V = 0,12$)

Příklad 6.

	velmi nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen
Praha	10	25	50	15
Venkov	20	10	130	40



cramerovo V = 0,296

$\Rightarrow$ středně silná závislost

H_0: data jsou nezávislá

H_A: $\neg H_0$

(CHISQ.test)

PŘEDPOKLADY? ✓

	velmi nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen
Praha	10	11.66667	60	18.33333
Venkov	20	23.33333	120	36.66667

$E_i \geq 5$

$(n-1)\cdot(m-1)$

$df = 3$

p-hodnota << 0,001

⇒ zamítám H_0

→ existuje stat. významná závislost mezi místem práce a spokojeností

Příklad č. 7:

a)

	NEMOC+	ZDRAVÝ	
KUŘÁK	171	3264	3435
NEKUŘÁK	117	4320	4437

b)

Cramerovo V = 0,061

→ velmi slabá / žádná závislost

c) KUŘÁCI

$p_K \doteq 0{,}050$ $\qquad \underline{\pi = ?}$

IO $\qquad$ předp $\qquad n > \dfrac{9}{p\cdot(1-p)}$

$n >> 191$

$$P(0{,}042 < \pi < 0{,}058) \doteq 0{,}95$$

π = absolutní riziko (pravděpod. onemocnění u kuřáků)

NEKUŘÁCI

$p_N = 0{,}026$ $\qquad$ PŘEDPOK. OK

$$P(0{,}021 < \pi < 0{,}032) \doteq 0{,}95$$

absolutní riziko u nekuřáků

d) RR

$$\widehat{RR} = \frac{p_K}{p_N} \doteq 1,89$$

	Outcome +	Outcome -	Total	Inc risk * %	Odds
Exposed +	171	3264	3435	$100 \cdot p_K$ = 4.98	0.0524
Exposed -	117	4320	4437	= 2.64	0.0271
Total	288	7584	7872	3.66	0.0380

Point estimates and 95% CIs:

Inc risk ratio RR	$\widehat{RR}$ = 1.89 (1.50, 2.38)	← 95% IS
Odds ratio OR	$\widehat{OR}$ = 1.93 (1.52, 2.46)	
Attrib risk *	2.34 (1.47, 3.21)	
Attrib risk in population *	1.02 (0.39, 1.65)	
Attrib fraction in exposed (%)	47.03 (33.27, 57.95)	
Attrib fraction in population (%)	27.92 (17.32, 37.17)	

Test that OR = 1: chi2(1) = 30.110 Pr>chi2 = <0.001
Wald confidence limits
CI: confidence interval
* Outcomes per 100 population units

$$P(1,50 < RR < 2,38) \doteq 0,95$$

e) $\check{S}_K = 0,0524$

$\check{S}_N = 0,0271$

f) $\widehat{OR} = 1,93$ $\left(\frac{\check{S}_K}{\check{S}_N}\right)$

$P(1{,}52 < OR < 2{,}46) \geq 0{,}95$

g) H_0: data jsou nezávislá

H_A: $\neg H_0$

(CHISQ.TEST)

PŘEDPOKLADY ✓

	ano	ne
kuřák	125.6707	3309.329
nekuřák	162.3293	4274.671

p-hodnota << 0,001

$\Rightarrow$ zamítám H_0 $\Rightarrow$ existuje stat. významná závislost mezi kouřením a výskytem nemoci