Přednáška 11 – Testy pro diskrétní data
Poslední teoretická přednáška

Testy pro jeden výběr | Testy pro dva výběry

- dotazníky – nehody, okolnosti nehody, aj.

http://staff.utia.cas.cz/uglickich/pdfka/volbaTH.pdf

Test podílu pro jeden výběr

- alternativní rozdělení, $x \in \{0,1\}$, 0 – neúspěch, 1 – úspěch
- výběrový podíl $p = \frac{\text{počet úspěchů}}{\text{počet dat}}$ Příklad: podíl studentů s modrýma očima

Předpoklady:

- $n > 30$
- $np \geq 5, \quad n(1-p) \geq 5$

- np=100·0.2=20> 5 – můžeme použít test
- np=100·0.8=80> 5 – můžeme použít test
- np=1000·0.0001=0.1< 5 – nemůžeme

Nulová hypotéza:

$H_0 : p = p_0$ podíl se rovná předpokládané hodnotě

Alternativní hypotéza:

$H_A : p \neq (>, <) p_0$

určuje směr testu

$T = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$ aproximativně pro $n > 30$

Příklad: Na úseku silnice s maximální povolenou rychlostí 80 km/h kontrolujeme rychlost vozidel. Testujeme hypotézu, že podíl řidičů, kteří překračují rychlost o více než 3km, je menší než 20%.

78	86	65	93	92	85	76	79	...$\rightarrow n > 30$

- diskrétní N.V. – překročení povolené rychlosti o více než 3 km/h
- úspěch – rychlost > 83, neúspěch – rychlost ≤ 83
- jeden výběr, testujeme podíl

Nulová hypotéza – podle tvrzení

H_0 : p=0.2 (nebo p<0.2) – podíl řidičů překračujících rychlost o více než 3 km/h se rovná nebo je menší než 0.2

Alternativní hypotéza – opačné tvrzení

H_A : p>0.2 – podíl je větší než 0.2 – pravostranný test

$$p = \frac{\text{počet úspěchů}}{\text{počet dat}} = \frac{\text{počet rychlostí}_{>83}}{\text{počet řidičů}} = \frac{1(86)+1(93)+1(92)+1(85)+\dots}{n}$$

χ^2 test dobré shody

- test rozdělení výběru
- diskrétní rozdělení

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- O_i – pozorované četnosti
- E_i – očekávané četnosti
- pouze pravostranný

Nulová a alternativní hypotézy:

H_0 : data mají teoretické rozdělení
H_A : data nemají teoretické rozdělení

Předpoklady:

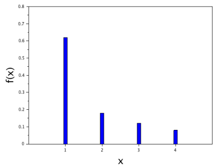
- všechny četnosti > 2
- aspoň 80% četností > 5

Příklad: Síť autorizovaných prodejců vozů nabízí vozy značky ABC s následujícími typy karoserie: hatchback, sedan, kombi a coupé. Doposud vozy se prodávaly takto:

kombi 62%, hatchback 18%, sedan 12%, coupé 8%.

V rámci rozšíření sítě byla otevřena nová pobočka s autorizovaným prodejcem, kde se zatím prodalo 120 vozů s karoserií kombi, 40 vozů hatchback, 18 sedanů a 22 coupé. Zajímá nás, zda se prodej na nové pobočce neliší od ostatních prodejců.

- diskrétní N.V. prodej $\in$ {kombi, hatchback, sedan, coupé}



- je dáno teoretické rozdělení – pf f(prodej)

	kombi	hatchback	sedan	coupé
f(prodej)	0.62	0.18	0.12	0.08

Nulová hypotéza:

H_0 : prodej na nové pobočce má stejné rozdělení

Alternativní hypotéza:

H_A : nemá stejné rozdělení

- pozorované četnosti $O = [120 \quad 40 \quad 18 \quad 22]$
- očekávané četnosti (v případě shody rozdělení):

$$E = [0.62 \cdot 200 \;\; 0.18 \cdot 200 \;\; 0.12 \cdot 200 \;\; 0.08 \cdot 200] = [124 \;\; 36 \;\; 24 \;\; 16]$$

- p-hodnota $= 0.2286$

Test o shodě dvou podílů

- dva výběry
- dvě diskrétní N.V. $x, y \in \{0,1\}$, 0 – neúspěch, 1 – úspěch
- výběry nemusí být párové
- stejné předpoklady jako pro test podílu

Nulová hypotéza:

$H_0 : p_1 = p_2$ podíly jsou stejné

Alternativní hypotéza:

$H_A : p_1 \neq (>,<) p_2$ určuje směr testu

$$T = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \sim N(0,1), \text{ kde } p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$

Příklad: Policie ČR tvrdí, že na silnici s maximální povolenou rychlostí 80 km/h směrem z Prahy je menší podíl řidičů překračujících povolenou rychlost než směrem do Prahy. Testujeme tvrzení policie na hladině významnosti 0.05.

z Prahy	93	86	85	93	92	85	86	79	$\ldots \to n_1 > 30$
do Prahy	78	96	75	83	97	105	81	79	$\ldots \to n_2 > 30$

- 2 diskrétní N.V. : překročení rychlosti na silnici z Prahy
  překročení rychlosti na silnici do Prahy
- úspěch – rychlost > 80, neúspěch – rychlost ≤ 80
- dva výběry, testujeme shodu podílů

Nulová hypotéza – podle tvrzení

H_0 : $p_z = p_{do}$ (nebo $p_z < p_{do}$) – podíl řidičů překračujících povolenou rychlost na silnici směrem z Prahy je menší než směrem do Prahy

Alternativní hypotéza – opačné tvrzení

H_A : $p_z > p_{do}$ – pravostranný test

Dva výběrové podíly:

$$p_z = \frac{\text{rychlosti z Prahy}_{>80}}{n_1} = \frac{1(93)+1(86)+1(85)+1(93)+1(92)+1(85)+\ldots}{n_1}$$

$$p_{do} = \frac{\text{rychlosti do Prahy}_{>80}}{n_2} = \frac{1(96)+1(83)+1(97)+1(105)+1(81)+\ldots}{n_2}$$

McNemarův test

Předpoklady:

- dva párové diskrétní výběry
- binární data
- kontingenční tabulka 2×2

- pouze pravostranný

- testování efektu nějaké události
- shoda četností před a po – změna?

před \ po	NE	ANO
NE	a	b
ANO	c	d

Nulová hypotéza:

$H_0 : b = c$ četnosti jsou stejné, tj., není žádná změna

Alternativní hypotéza:

$H_A : b \neq c$ četnosti nejsou stejné, tj., je změna

$T = \frac{(b-c)^2}{b+c} \sim \chi^2$-rozdělení

Příklad: Na přechodu pro chodce namontovali nový semafor. Zeptali jsme se 15 lidí na jejich spokojenost s přechodem před instalací semaforu a po instalaci. Testujeme tvrzení, že instalace semaforu nepřispěla ke spokojenosti občanů.

před instalací	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1
po instalaci	1	2	1	2	2	2	2	2	2	1	2	2	2	1	2

- diskrétní N.V.
  spokojenost $\in \{1=\text{NE}, 2=\text{ANO}\}$
- párové výběry
- změna stavu: četnosti b, c

- kontingenční tabulka:

před \ po	NE	ANO
NE	a=4	b=9
ANO	c=0	d=2

Nulová hypotéza:

$H_0 : b = c$ spokojenost respondentů je stejná před instalací semaforu a po ní, tj., semafor nemá vliv

Alternativní hypotéza:

$H_A : b \neq c$ semafor má vliv

- p-hodnota $= 0.0026998 < 0.05$

Poznámka 1:

- Testy pro dva výběry – testy nezávislosti pro diskrétní data
  - χ^2 test nezávislosti
  - Fisherův exaktní test
  - Gamma koeficient
  - Yule's Q koeficient

Poznámka 2:

- poslední přednáška
- příští tyden – konzultace podle domluvy (konkrétní dotazy)
- zkouška, termíny