Přednáška 9 – Testy nezávislosti

Budeme testovat, zda jsou dva výběry nezávislé

Testy pro spojitá data

Testy pro diskrétní data

Parametrické testy s předpokladem normality

Neparametrické testy bez předpokladu normality

Nulová a alternativní hypotézy pro všechny testy nezávislosti:

H_0 : veličiny jsou nezávislé

H_A : veličiny nejsou nezávislé

- Směr testu nevolíme
- Tabulka testů hypotéz:

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

Použijeme za předpokladu:

- 2 párové spojité výběry
- normalita všech výběrů

- Pearsonův korelační koeficient – míra korelace

  - $0 < r_p \leq 1$ – přímá závislost
  - $-1 \leq r_p < 0$ – nepřímá závislost
  - pro nezávislé veličiny $r_p = 0$

Nulová hypotéza Pearsonova testu:

$H_0 : r_p = 0$ veličiny jsou lineárně nezávislé

Alternativní hypotéza – opačné tvrzení:

$H_A : r_p \neq 0$ veličiny nejsou lineárně nezávislé

$$T = \frac{r_p}{\sqrt{\frac{1-r_p^2}{n-2}}} \sim \text{Studentovo rozdělení}$$

Příklad: lineární závislost mezi plochou bytu a jeho cenou

- čím větší je byt, tím je dražší?
- závislost nemusí být nutně lineární
- oba párové výběry – normální

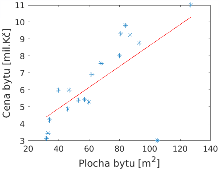
Nulová hypotéza:

H_0 : plocha a cena jsou lineárně nezávislé

Alternativní hypotéza:

H_A : nejsou lineárně nezávislé

- p-hodnota$=0.0018 < 0.05$
- $r_p = 0.68$ – čím větší je byt, tím je dražší
- data jsou vhodná k lineární regresi



plocha, m^2	cena, Kč
84	9 811 200
105	3 000 000
40	5 990 000
57	5 420 000
127	11 000 000
80	8 000 000
32	3 150 000
46	4 876 000
33	3 450 000
93	8 754 000
34	4 249 929
53	5 400 000
87	9 240 250
47	5 988 050
81	9 296 600
60	5 280 000
62	6 894 250
68	7 566 600

Spearmanův test korelačního koeficientu neparametrická alternativa

Použijeme za předpokladu:

- 2 párové spojité výběry
- alespoň 1 výběr nemá normalitu

Spearmanův koeficient pořadové korelace

- na základě Pearsonova koeficientu
- s využitím pořadí hodnot výběrů místo dat:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad r_s \in (-1, 1)$$

- d_i – rozdíly pořadí z výběrů

- $0 < r_s \leq 1$ – přímá závislost
- $-1 \leq r_s < 0$ – nepřímá závislost
- pro nezávislé veličiny $r_s = 0$

Nulová hypotéza Spearmanova testu:

$H_0 : r_s = 0$ veličiny jsou nezávislé

Alternativní hypotéza – opačné tvrzení:

$H_A : r_s \neq 0$ veličiny nejsou nezávislé

Příklad: závislost mezi dobou strávenou u obrazovky mobilu a denním počtem kroků

- oba párové výběry nejsou normální

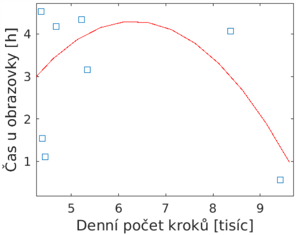
Nulová hypotéza:

H_0 : počet kroků a doba jsou nezávislé

Alternativní hypotéza:

H_A : nejsou nezávislé

- p-hodnota je $0.03 < 0.05$
- data jsou vhodná k regresi (polynomiální, exponenciální, atd.)



denní počet kroků	čas, [h]
8374.6163	4.07
9419.2094	0.57
4364.4742	4.53
4682.0057	4.18
9885.3437	0.54
5214.719	4.34
4119.8086	4.57
4446.0034	1.11
5337.3764	3.16
4384.4456	1.54

Pro zajímavost:

Pearsonův test:

- p-hodnota je 0.1263
- data nejsou vhodná k lineární regresi

Testy nezávislosti pro diskrétní veličiny

χ^2 test nezávislosti

Předpoklady:

- 2 diskrétní veličiny
- všechny četnosti > 2, aspoň 80% četností > 5
- kontingenční tabulka

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

- O_i – pozorované četnosti (data)
- E_i – očekávané četnosti (v případě nezávislosti)

Nulová hypotéza H_0 :

veličiny jsou nezávislé

Alternativní hypotéza H_A :

veličiny nejsou nezávislé

Příklad: Ptali jsme se zákazníků různých věkových kategorií, v jakém e-shopu nejraději objednávají potraviny. Mezi zákazníky ve věku od 20 do 30 let preferovalo 56 e-shop A, 42 B a 19 C. Ve věku 31-45 let 46 zákazníků volilo e-shop A, 17 B a 37 využívalo služeb rozvozu potravin C. Z věkové kategorie od 46 do 70 let 25 zákazníků by si objednalo potraviny v e-shopu A, 36 v B a 44 v C. Ha hladině významnosti 0.05 testujte tvrzení, že věk a volba e-shopu jsou nezávislé.

- 2 diskrétní veličiny:

$\text{věk} \in \{\text{20-30 let},\ \text{31-45 let},\ \text{46-70 let}\}$, $\text{e-shop} \in \{A, B, C\}$

- kontingenční tabulka – O

věk \ eshop	A	B	C
20-30 let	56	42	19
31-45 let	46	17	37
46-70 let	25	36	44

Princip výpočtu E:

- $\frac{O}{n} = f(\text{věk}, \text{e-shop})$
- Pokud platí

$f(\text{věk}, \text{e-shop}) = f(\text{věk}) f(\text{e-shop})$

veličiny jsou nezávislé

- $f(\text{věk}) \cdot f(\text{e-shop}) \cdot n = E$

Postup:

1. $\frac{O}{332} = f(\text{věk}, \text{e-shop})$ – sdružené rozdělení

věk \ e-shop	A	B	C
20-30 let	0.17	0.13	0.06
31-45 let	0.14	0.05	0.11
46-70 let	0.08	0.11	0.15

2. marginální rozdělení $f(\text{věk})$ a $f(\text{e-shop})$

věk \ e-shop	A	B	C
20-30 let	0.17	0.13	0.06
31-45 let	0.14	0.05	0.11
46-70 let	0.08	0.11	0.15
$f(\text{e-shop})$	0.39	0.29	0.32

$f(\text{věk})$
0.17+0.13+0.06=0.36
0.14+0.05+0.11=0.3
0.08+0.11+0.15= 0.34

3.

$f(\text{věk}) \cdot f(\text{e-shop})$ = nové sdružené rozdělení pro nezávislé veličiny:

$f(\text{věk}) \cdot f(\text{e-shop})$=

věk \ e-shop	A	B	C
20-30 let	0.14	0.1	0.12
31-45 let	0.12	0.09	0.09
46-70 let	0.13	0.1	0.11

4. $f(\text{věk}) \cdot f(\text{e-shop}) \cdot 332 = E$

věk \ e-shop	A	B	C
20-30 let	47	33	40
31-45 let	40	30	30
46-70 let	43	33	37

- p-hodnota$= 0.0000033$
- zamítáme H_0, že věk zákazníků a výběr e-shopu jsou nezávislé

Fisherův exaktní test

Předpoklady:

- nominální binární data
- kontingenční tabulka 2×2

X \ Y	$Y = 1$	$Y = 2$
$X = 1$	a	b
$X = 2$	c	d

$$p^* = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!\,(a+c)!\,(c+d)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

- p^* – referenční pravděpodobnost

Nulová hypotéza H_0 :

veličiny jsou nezávislé

Alternativní hypotéza H_A :

veličiny nejsou nezávislé

Příklad:

pohlaví \ operační systém	A	B
muži	19	26
ženy	31	22

- operační systém $\in$ {A, B}, pohlaví $\in$ {muži,ženy}
- p-hodnota$= 0.1555$

Gamma koeficient (Goodmanovo-Kruskalovo gamma)

Předpoklady:

- ordinální kategorická data

Nulová hypotéza H_0 :

$\gamma = 0$ veličiny jsou nezávislé

Alternativní hypotéza H_A :

$\gamma \neq 0$ veličiny nejsou nezávislé

$$T = \gamma \sqrt{\frac{N_c + N_d}{n(1-\gamma^2)}}$$

- míra asociace mezi veličinami

$$\gamma = \frac{N_c - N_d}{N_c + N_d}, \quad \gamma \in \langle -1, 1 \rangle$$

- N_c – počet dat s přímou závislostí
- N_d – počet dat s nepřímou závislostí

- $0 < \gamma \leq 1$ – přímá závislost
- $-1 \leq \gamma < 0$ – nepřímá závislost
- pro nezávislé veličiny $\gamma = 0$

Příklad: vazba mezi vzděláním a dobou po získání řidičského průkazu

doba \ vzdělání	bez maturity	s maturitou	VŠ
do 5 let	25	38	11
6-10 let	56	47	37
více než 10 let	53	46	54

- doba $\in$ {do 5 let, 6-10 let, víc než 10 let }
- vzdělání $\in$ {bez maturity, s maturitou, VŠ}
- uspořádaná data v souladu pro obě veličiny
- hlavní diagonála – přímá vazba (“vyšší vzdělání – déle mám řidičák”)
- vedlejší diagonála – nepřímá vazba

$$N_c = 25*(47+37+46+54)+38*(37+54)+11*0+56*(46+54)+47*54+37*0+53*0+46*0+54*0 = 16196$$

$$N_d = 11*(56+47+53+46)+38*(56+53)+25*0+37*(53+46)+47*53+56*0+54*0+46*0+53*0 = 12518$$

- $\gamma = \frac{16196-12518}{16196+12518} = 0.128$
- p-hodnota=0.0592

Yule's Q koeficient – speciální případ Gamma koeficientu

Předpoklady:

- ordinální binární data
- kontingenční tabulka 2×2

X \ Y	$Y = 1$	$Y = 2$
$X = 1$	a	b
$X = 2$	c	d

$$Q = \frac{N_c - N_d}{N_c + N_d} = \frac{ad - bc}{ad + bc}$$

- H_0, H_A, T – stejné

Příklad: závislost mezi dobou učení a výsledkem u zkoušky

doba učení \ prospěch	neprospěl	prospěl
málo se učil/a	13	5
hodně se učil/a	7	19

- $Q = \frac{13*19-7*5}{13*19+7*5} = 0.752$
- p-hodnota=4.8866e-07