

Testy nezávislosti

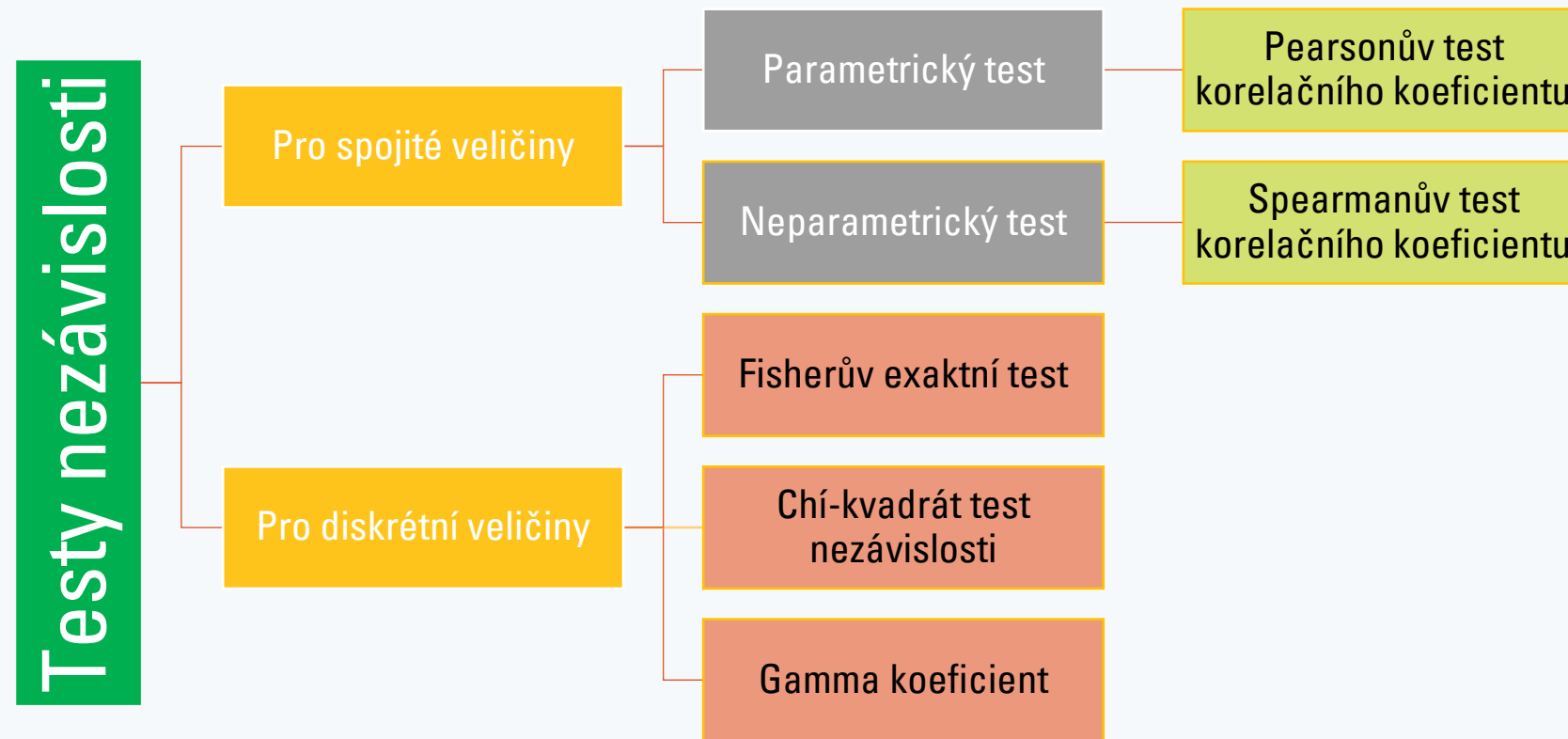
Pavla Pecherková

15. 4. 2021

Co jsou a k čemu slouží testy nezávislosti

- Testy, které nám určí, zda spolu souvisí nějak dvojice diskrétních nebo dvojice spojitých náhodných veličiny (nelze kombinovat spojitou náhodnou veličinu s diskrétní).
- Opakování:
 - Diskrétní data: data, která jsou spočetná a konečná
 - Spojitá data: data, která jsou nekonečná nebo nespočetná
- Proč to potřebujeme?
 - Potřebujete zda zjistit, zda jedna veličina ovlivňuje druhou
 - jsou data vhodná k regresi? K čemu by nám bylo počítat regresi na dvou nezávislých datech.
 - ovlivňuje jedna druhou?
 - Mají mezi sebou nějaký vztah?

Testy nezávislosti - dělení



Testy korelačního koeficientu (spojitá data)

- Kdy jsme slyšeli o korelačním koeficientu?

- U regrese – počítali jsme sílu závislosti mezi výběrem X a Y
- Nabývají hodnot mezi $< -1, 1 >$



Co to znamená, kdy jsou závislé a nezávislé?



Testy nezávislosti:

H₀: náhodné veličiny jsou nezávislé (korelace je 0)

H_A: náhodné veličiny nejsou nezávislé (korelace není 0)

Korelační koeficient

Lineární
(normalita dat)

Nelineární
(není normalita dat)

Pearsonův k. k.

Spearmanův k. k.

$$r_p = \frac{\text{cov}[x, y]}{\sigma_x \sigma_y}$$

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

Testy korelačního koeficientu

H₀: náhodné veličiny jsou (lineárně) nezávislé

H_A: náhodné jsou (lineárně) závislé



Pro regresi **potřebujeme zamítnout**

PROČ?

H₀ zamítáme

- $r \neq 0$
- Data spolu korelují (souvisí)
- Mezi daty je souvislost, můžeme hledat jaká

H_A nezamítáme

- $r = 0$
- Data spolu nekoreluji (nesouvisí)
- K čemu hledat souvislost tam kde není?

Pearsonův test korelačního koeficientu

Předpoklad:

1. předpokládáme $N(\mu, \sigma^2)$
2. párové výběry.

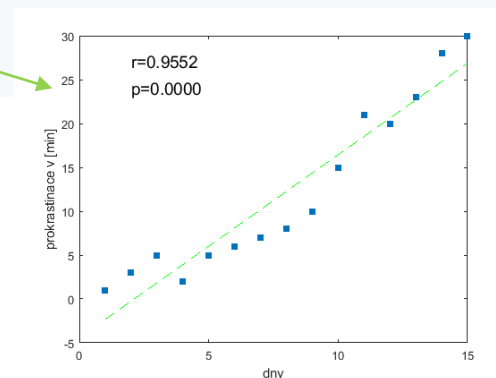
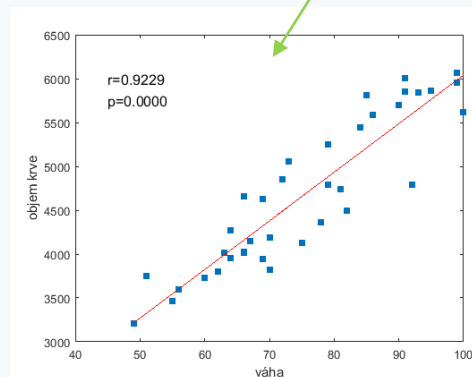
H_0 : náhodné veličiny jsou lineárně nezávislé

H_A : náhodné veličiny jsou lineárně závislé

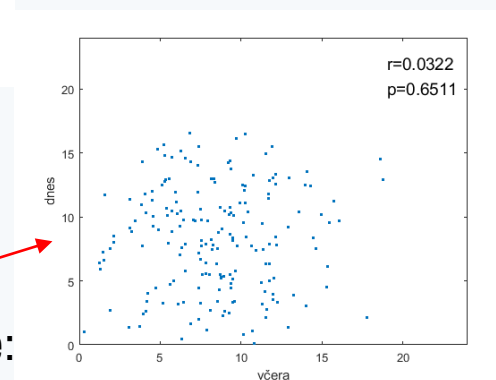
Statistika:

$$T = \frac{r_p}{\sqrt{\frac{1 - r_p^2}{n - 2}}} \sim t(n - 1)$$

H_0 zamítáme:



H_0 nezamítáme:



Spearmanův test korelačního koeficientu

Předpoklad:

- párové výběry.

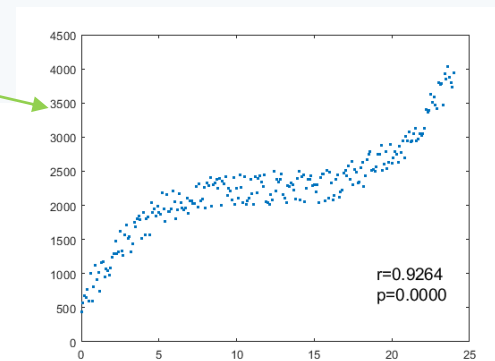
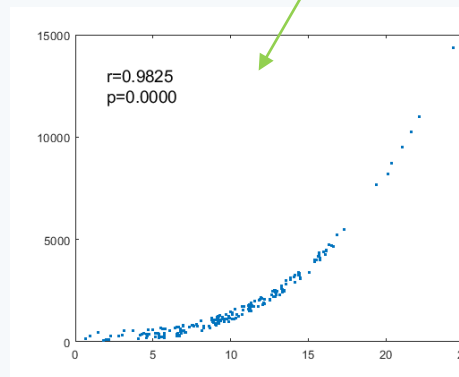
H₀: náhodné veličiny jsou nezávislé

H_A: náhodné veličiny jsou závislé

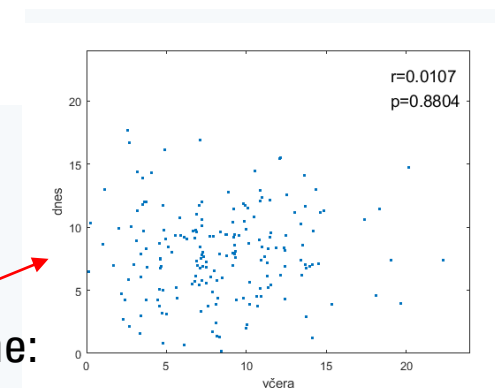
Statistika:

$$T = \frac{r_S}{\sqrt{\frac{1 - r_S^2}{n - 2}}} \sim t(n - 1)$$

H₀ zamítáme:



H₀ nezamítáme:



Testy nezávislosti pro diskrétní data

Diskrétní data:

- Nominální data
 - diskrétní data, kdy seřadit data nemá smysl. Hodnota nemá další vlastnost než označení
 - Příklad: krevní skupina 0, A, B, AB; barva vozidla: červená=1, zelená=2,..., černá=10
- Ordinální data
 - diskrétní data, které je možné seřadit.
 - Zpravidla se řadí vzestupně, lze tedy určit, která hodnota je menší a která větší. Vždy musí být seřazena jedním směrem
 - Příklad: dosažené vzdělání: ZŠ=1, SŠ=2, VŠ=3, intenzita světla v době nehody: tma=1, šero=2, světlo=3)

Kontingenční tabulka

- důležitá tabulka, která jednoduše shrnuje základní údaje
- zobrazuje frekvenční součet (kombinací) dvou diskrétních, a to jak nominálních tak ordinálních, dat

	D1=1	D1=2	D1=3
D2=1	N_{11}	N_{12}	N_{13}
D2=2	N_{21}	N_{22}	N_{23}
D2=3	N_{31}	N_{32}	N_{33}

Kde N_{xy} je počet kombinací, kde první veličina nabývá hodnoty x a druhá veličina y .

Chí-kvadrát test nezávislosti

- Podmínky:
 - Alespoň 80% buněk >5
 - Všechny buňky >2
 - Nominální diskrétní data
- Test podobný testu dobré shody pro rozdělení, kdy porovnáváme pozorované (O) a očekávané (E) četnosti.
- Statistika:

H0: data jsou nezávislá

HA: data nejsou nezávislá

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(r-1)(c-1)$$

kde r je počet řádků a c je počet sloupců.

Zjišťovali jsme mezi kuřáky, kolik je zde mužů/žen a do jaké kategorie spadají. Získali jsme následující zápis:

Žena	Žena	Muž	Muž	Muž	Žena	Muž	Muž		Žena
0-35	35-65	35-65	65+	0-35	65+	65+	0-35		35-65

Zjistěte, zda jsou výběry nezávislé (tzn. jestli žen v daných věkových skupinách je přiměřeně stejně jako mužů – tedy nezáleží na pohlaví).

Věková skupina	Ženy	Muži
0-35	1075	1074
35-65	1797	1628
65+	796	657

Frekvence/
četnost

Kontingenční tabulka = pozorovaná četnost

Věková skupina	Ženy	Muži	Celkem
0-35	1075	1074	2149
35-65	1797	1628	3425
65+	796	657	1453
Celkem	3668	3359	7027

Sdružená a marginální p.

f(věk,pohlaví)	Ženy	Muži	f(věk)
0-35	0,1530	0,1528	0,3058
35-65	0,2527	0,2317	0,4874
65+	0,1133	0,0935	0,2068
f(pohlaví)	0,5220	0,4780	

$\frac{1075}{7027}$
 $\frac{657}{7027}$

$f(\text{věk}) * f(\text{pohlaví})$

Teoretický f(v,p)	Ženy	Muži
0-35	0,1596	0,1462
35-65	0,2544	0,2330
65+	0,1079	0,0988

Teor. $f(v,p) * 7027$

Teoretická četnost	Ženy	Muži
0-35	1121,7	1027,3
35-65	1787,8	1637,2
65+	758,4	694,6

[tabulka,chi,p]=crosstab(xo,yo)

tabulka = 3x2

	1	2
1	1075	1074
2	1797	1628
3	796	657

chi = 8.0646

p = 0.0177

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(1075 - 1121,7)^2}{1121,7} + \dots + \frac{(657 - 694,6)^2}{694,6} = 8,0646$$

Fisherův exaktní test

■ Předpoklady:

- tabulka 2x2
- celkový počet dat <1000
- nominální diskretní data

H₀: náhodné veličiny jsou nezávislé (podíly jsou shodné)

H_A: náhodné veličiny nejsou nezávislé (podíly nejsou shodné)

Náhodné veličiny X \ Y	Y=0	Y=1
X=0	a	b
X=1	c	d

Referenční pravděpodobnost (pozor, to není p-hodnota) pro danou kombinaci četností je dána hypergomeotrickým rozdělením jako:

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

kde n je celková četnost a a, b, c, d jsou četnosti v jednotlivých políčkách tabulky.

Pro výpočet p-hodnoty je nutné dopočítat i ostatní pravděpodobnosti pro možné kombinace četností při zachování hodnot marginál. P-hodnota je poté součet všech pravděpodobností menších nebo rovno referenční p^* .

- Vědci zkoumali chování čipmanků východních (Burke da Silva et al. (2002)). Ze vzdálenosti 10 a 100 m simulovali útok dravce a sledovali jejich chování. Z 24 jich 16 varovalo ostatní, v případě 10 m vzdálenosti, u 100 m vzdálenosti to byli 3 z 24. Na hladině významnosti 0,05 testujte tvrzení, že vzdálenost a počty varování jsou nezávislé.

vzdálenost\varování	ano	ne
10 m	16	8
100 m	3	21

```
[h,p]=fishertest([16 8; 3 21])
```

```
h = logical
```

```
1
```

```
p = 2.7507e-04
```



Zamítáme hypotézu o nezávislosti vzdálenosti a počtu varování, protože $p < \alpha$.

Gamma koeficient (Goodmanovo-Kruskalovo gamma)

- Testujeme, jak silný má vliv nezávisle proměnná x na závislé proměnnou y za předpokladu, že obě náhodné veličiny jsou diskrétní a ordinální.
 - Pozor, nelze použít opačně, tedy jak y ovlivňuje x
- Teoreticky lze diskretizací udělat ze spojitě náhodné veličiny ordinální diskrétní → lze použít Gamma koeficient
- Obdobné chování jako regresní koeficient pro spojitě náhodné veličiny → $\gamma \in (-1, 1)$
 - $\gamma = 1$ resp. $\gamma = -1$ je absolutní přímá resp. nepřímá závislost
 - $\gamma = 0$ není žádná závislost

▪ Předpoklady:

- Uspořádaná ordinální diskrétní data

H₀: náhodné veličiny jsou nezávislé ($\gamma = 0$)

H_A: náhodné veličiny nejsou nezávislé ($\gamma \neq 0$)

Statistika:

$$T = \gamma \sqrt{\frac{N_c + N_d}{n(1 - \gamma^2)}} \quad T \sim N(0, 1)$$

kde $\gamma = \frac{N_c - N_d}{N_c + N_d}$, n je počet dat, N_c je počet hodnot vyjadřující přímou vazbu a N_d je počet hodnot vyjadřující nepřímou vazbu.

- Jak spočteme N_c a N_d si ukážeme na příkladu:
- *Chceme zjistit, zda je nějaká závislost mezi stářím a počtem zdravých zubů. Na hladině významnosti 0,05 testujte tvrzení, zda je počet zdravých zubů ovlivněn věkem.*



Zuby\stáří	Junior	Dospělý	senior
0-10	2	12	5
11-21	5	16	6
22-32	12	3	7

$$N_c = 2(16 + 3 + 6 + 7) + 12(6 + 7) + 5(3 + 7) + 16 * 7 = 382$$

$$N_d = 5(5 + 16 + 12 + 3) + 12(5 + 12) + 6(3 + 12) + 16 * 12 = 666$$

$$\gamma = \frac{N_c - N_d}{N_c + N_d} = \frac{382 - 666}{382 + 666} = -0,2710$$

$$p = 0,1057$$

Nezamítáme hypotézu o nezávislosti mezi stářím a počtem zdravých zubů, protože $p > \alpha$.

Yule's Q koeficient

Náhodné veličiny X \ Y	Y=0	Y=1
X=0	a	b
X=1	c	d

- Speciální případ Gamma koeficientu, pro rozměr 2x2.

- Předpoklady:

- Uspořádaná ordinální diskrétní data
- Binární proměnné obě veličiny

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

- kde $N_c = ad, N_d = bc$

- Statistika:

H0: náhodné veličiny jsou nezávislé ($Q = 0$)

HA: náhodné veličiny nejsou nezávislé ($Q \neq 0$)

$$T = Q \sqrt{\frac{N_c + N_d}{n(1 - Q^2)}} \quad T \sim N(0,1)$$

Ovlivňuje síla větru dobu pouštění draka?

Síla větru \ drak	neletěl	letěl
Bezvětrí, mírný	60	15
Velký, orkán	6	42

$$Q = \frac{N_c - N_d}{N_c + N_d} = \frac{60 \cdot 42 - 15 \cdot 6}{60 \cdot 42 + 15 \cdot 6} = 0,9310$$