

TESTOVÁNÍ NEPARAMETRICKÝCH HYPOTÉZ



11. cvičení

Neparametrické hypotézy

- ❖ Hypotézy o vlastnostech populace (typ rozdělení, závislost proměnných...)

Testy dobré shody

Testují shodu mezi výběrovým (empirickým) a teoretickým rozdělením.

- ❖ **χ^2 – test dobré shody** - ověřuje, jestli se pozorované (angl. "observed") absolutní četnosti O_i jednotlivých variant náhodné veličiny shodují s očekávanými (angl. „expected“) absolutními četnostmi E_i , tj. četnostmi, které bychom očekávali v případě platnosti nulové hypotézy.
- ❖ **Kolmogorovův – Smirnovův test pro jeden výběr** - ověřuje hypotézu, zda pořázený výběr pochází z rozdělení se spojitou distribuční funkcí $F(x)$.

χ^2 – test dobré shody

0. Předpoklady testu:

$$n \cdot \pi_{0,i} > 5, \quad i = 1, 2, \dots, k$$

Očekávané četnosti musí být větší než 5 (alespoň 80% očekávaných četností musí být větších než 5, zbylých 20% nesmí být menší než 2)

χ^2 – test dobré shody

1. Volba nulové hypotézy:

- a) H_0 : Výběr pochází z populace, v níž jsou **relativní četnosti jednotlivých variant rovny číslům**, populace musí být roztržitelná podle nějakého znaku do k skupin.
- b) H_0 : Výběr pochází z rozdělení určitého typu, jehož parametry jsou dány (**úplně specifikovaný model**).
- c) H_0 : Výběr pochází z rozdělení určitého typu, přičemž neověřujeme informace o parametrech rozdělení, parametry modelu odhadujeme (**neúplně specifikovaný model**).

Alternativní hypotézu volíme ve tvaru:

H_A : neplatí H_0 (H_0)

χ^2 – test dobré shody

2. Výpočet pozorované hodnoty testové statistiky:

$$T(\underline{X}) = G = \sum_{i=1}^k \frac{(n_i - n \cdot \pi_{0,i})^2}{n \cdot \pi_{0,i}} \rightarrow \chi_{k-h-1}^2$$

n ... rozsah výběru

k ... počet variant

h ... počet odhadovaných parametrů modelu

n_i ... pozorované četnosti jednotlivých variant

$\pi_{0,i}$... očekávané relativní četnosti jednotlivých variant

$n \cdot \pi_{0,i}$... očekávané absolutní četnosti jednotlivých variant

χ^2 – test dobré shody

3. Výpočet p -value:

$$p - value = 1 - F_0(x_{OBS})$$

4. Formulace závěru vztaženého ke konkrétnímu příkladu.

Podívejte se na flash animaci „Testy dobré shody – řešený příklad“ na stránkách MI21 za a), b) i c)

<http://mi21.vsb.cz/flash-animace/testy-dobre-shody-reseny-priklad>

Kolmogorovův-Smirnovův test

Používá se pro ověření hypotézy, zda pořázený výběr pochází z rozdělení se **spojitou distribuční funkcí** $F(x)$.

$F(x)$ musí být úplně specifikovaná.

Výhody oproti X^2 – testu dobré shody:

- ❖ větší síla testu
- ❖ nemá omezující podmínky (lze použít při výběrech malého rozsahu)
- ❖ vychází z jednotlivých pozorování a nikoliv z údajů seřazených do skupin

Kolmogorovův-Smirnovův test

1. Formulace nulové a alternativní hypotézy:

H_0 : Náhodný výběr pochází z rozdělení se spojitou distribuční funkcí $F_0(x)$.

H_A : Náhodný výběr nepochází z rozdělení se spojitou distribuční funkcí $F_0(x)$.

2. Výběrový soubor $X_1, X_2, \dots, X_n$ uspořádáme vzestupně podle velikosti $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$

Empirická distribuční funkce je pak dána vztahem:

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ i / n, & X_{(i)} < x < X_{(i+1)}, i = 1, \dots, n-1 \\ 1, & x \geq X_{(n)} \end{cases}$$

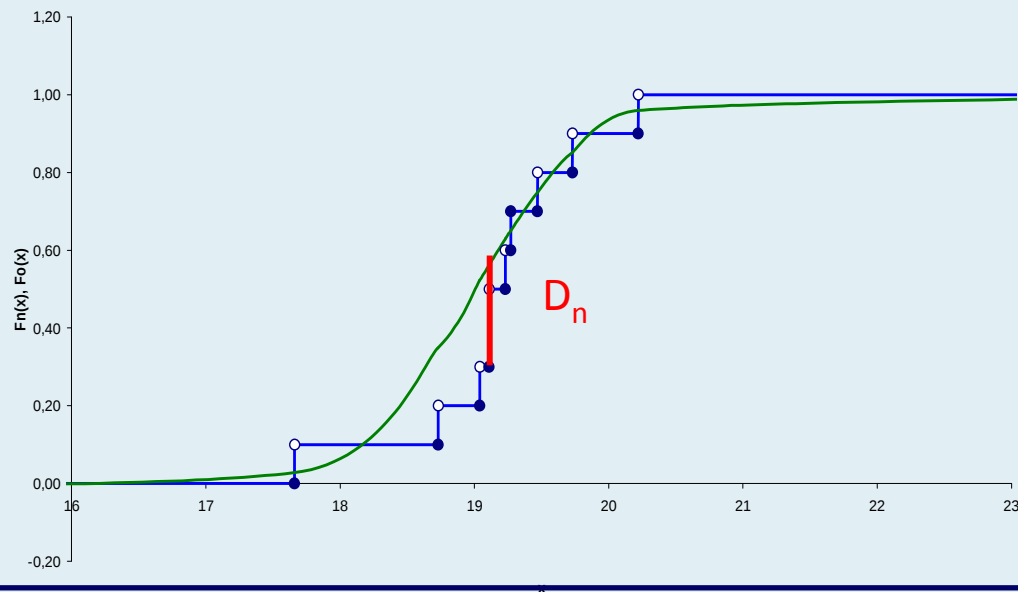
Kolmogorovův-Smirnovův test

3. Testová statistika:

$$T(\underline{X}) = D_n = \sup_x |F_n(x) - F_0(x)| = \max(D_1^*, D_2^*, \dots, D_n^*)$$

$$D_i^* = \max \left\{ \left| F_0(x_i) - \frac{i-1}{n} \right|, \left| \frac{i}{n} - F_0(x_i) \right| \right\} \quad \text{pro } i = 1, 2, \dots, n$$

Stanovení D_n



Kolmogorovův-Smirnovův test

4. Rozhodnutí:

Nulovou hypotézu zamítáme, pokud pozorovaná hodnota testové statistiky D_n překročí kritickou hodnotu $D_{n(a)}$.

Malé $n \Rightarrow$ speciální tabulky kritických hodnot $D_{n(a)}$.

Velké $n \Rightarrow$ aproximace kritických hodnot $D_{n(a)}$ podle vztahu:

$$D_{n(a)} \cong \sqrt{\frac{1}{2n} \ln\left(\frac{2}{a}\right)}$$

Podívejte se na flash animaci „Kolmogorovův-Smirnovův test – řešený příklad“ na stránkách MI21

<http://mi21.vsb.cz/flash-animace/kolmogorovuv-smirnovuv-test-reseny-priklad>

Vazby mezi znaky

U statistických jednotek zkoumáme většinou více znaků najednou:

- ❖ spotřeba, objem motoru, hmotnost a zrychlení automobilů,
- ❖ školní prospěch a pocit deprese u dětí, apod.

Mezi jednotlivými znaky hledáme závislosti:

- ❖ spotřebou automobilu a jeho hmotností
- ❖ pocitem deprese u dětí a školním prospěchem.

Typy závislosti:

- ❖ Jednostranná závislost
- ❖ Oboustranná závislost

Jednostranná závislost

V případě, kdy znak X působí na znak Y , ale znak Y nepůsobí zpětně na znak X .

$$X \longrightarrow Y$$

- ❖ vztah mezi typem absolvované střední školy a (bodovým) výsledkem přijímací zkoušky z matematiky
- ❖ vztah mezi výškou a váhou

		Typ znaku Y (důsledek)	
		kategoriální	kvantitativní
Typ znaku X (příčina)	kategoriální		ANOVA
	kvantitativní		regresní a korelační analýza

Metody analýzy jednostranné závislosti

Oboustranná závislost

- ❖ Ve vztahu nelze jednoznačně určit příčinu a důsledek
- ❖ Znak X působí na znak Y a znak Y působí zpětně na znak X

$$X \longleftrightarrow Y$$

- ❖ vztah mezi výdaji domácností na oblečení a na potraviny
- ❖ Výběr vhodné metody při zjištění závislostí závisí na typu analyzovaných veličin

		Typ znaku Y		
		kategoriální	ordinální	kvantitativní
Typ znaku X	kategoriální	analýza záv. v kontingenčních tabulkách, analýza záv. v asociačních tabulkách		
	ordinální		analýza závislosti ordinálních znaků	
	kvantitativní			analýza závislosti v normálním rozdělení

Metody analýzy oboustranné závislosti

Testy v kontingenční tabulce

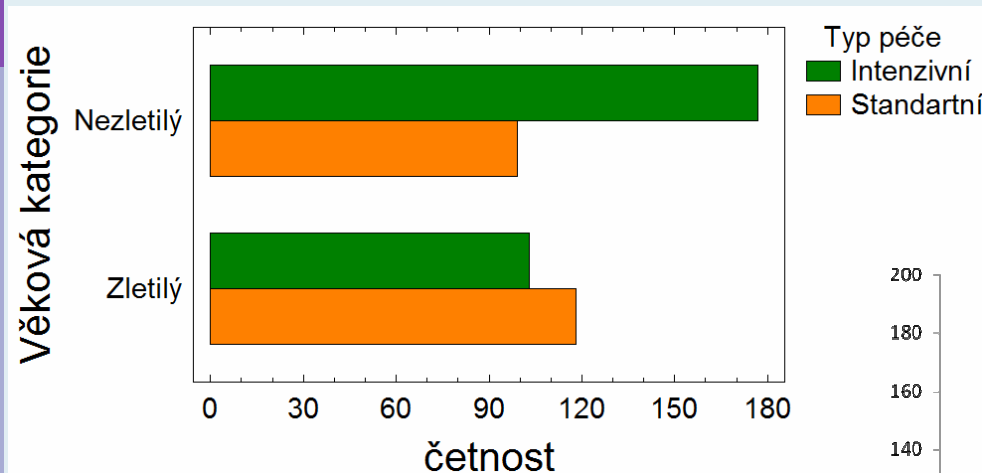
- ❖ Testuje závislost dvou kategoriálních proměnných
- ❖ Kontingenční tabulka:

X/Y	Y_1	Y_2	$\dots$	Y_n	$\sum_j$
X_1	n_{11}	n_{12}	$\vdots$	n_{1n}	$n_{1.}$
X_2	n_{21}	n_{22}	$\vdots$	n_{2n}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_m	n_{m1}	n_{m2}	$\dots$	n_{mn}	$n_{m.}$
$\sum_i$	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.n}$	n

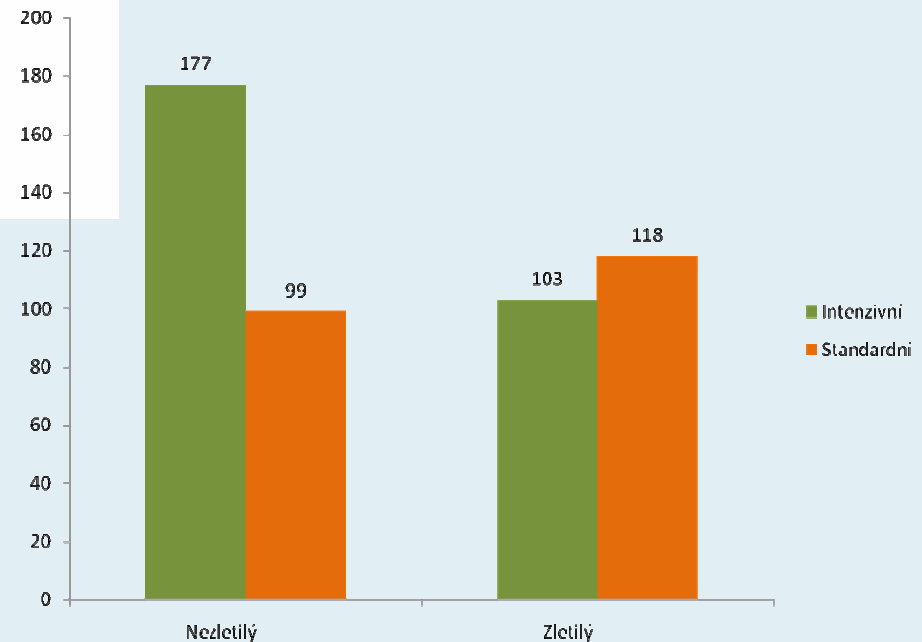
- ❖ Dvourozměrná tabulka četností, z jejichž hodnot můžeme usoudit na závislost či nezávislost mezi dvěma kategoriálními proměnnými

Grafické výstupy pro analýzu závislosti dvou kategoriálních proměnných

❖ Shlukový sloupcový graf

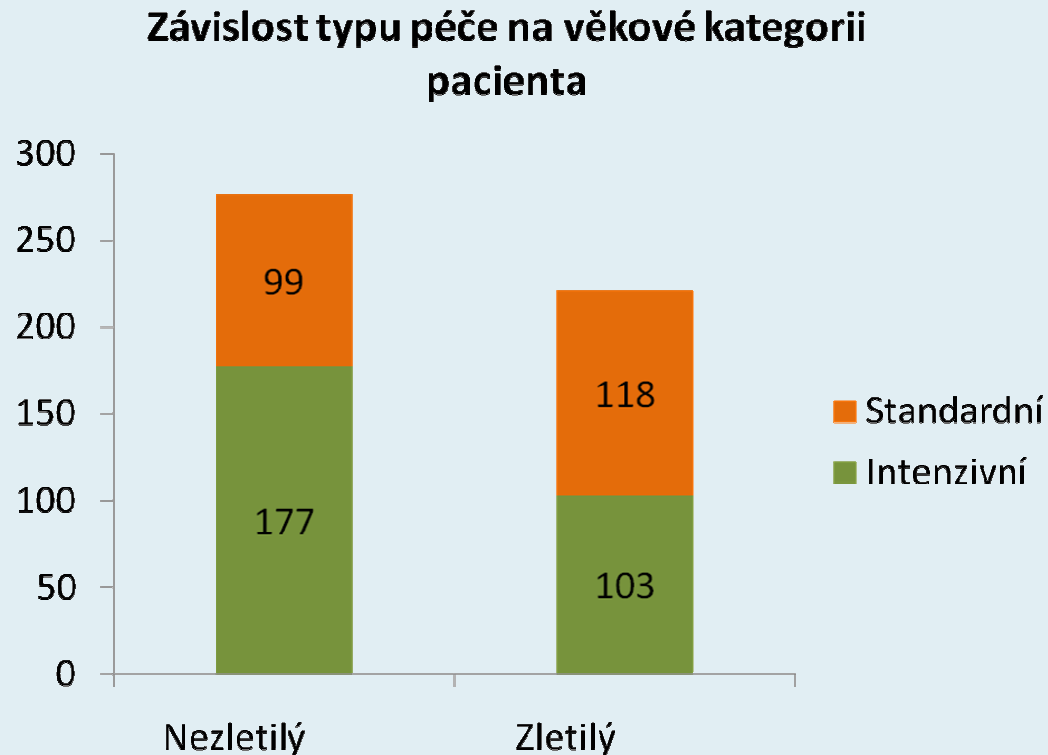


Závislost typu péče na věkové kategorii pacienta



Grafické výstupy pro analýzu závislosti dvou kategoriálních proměnných

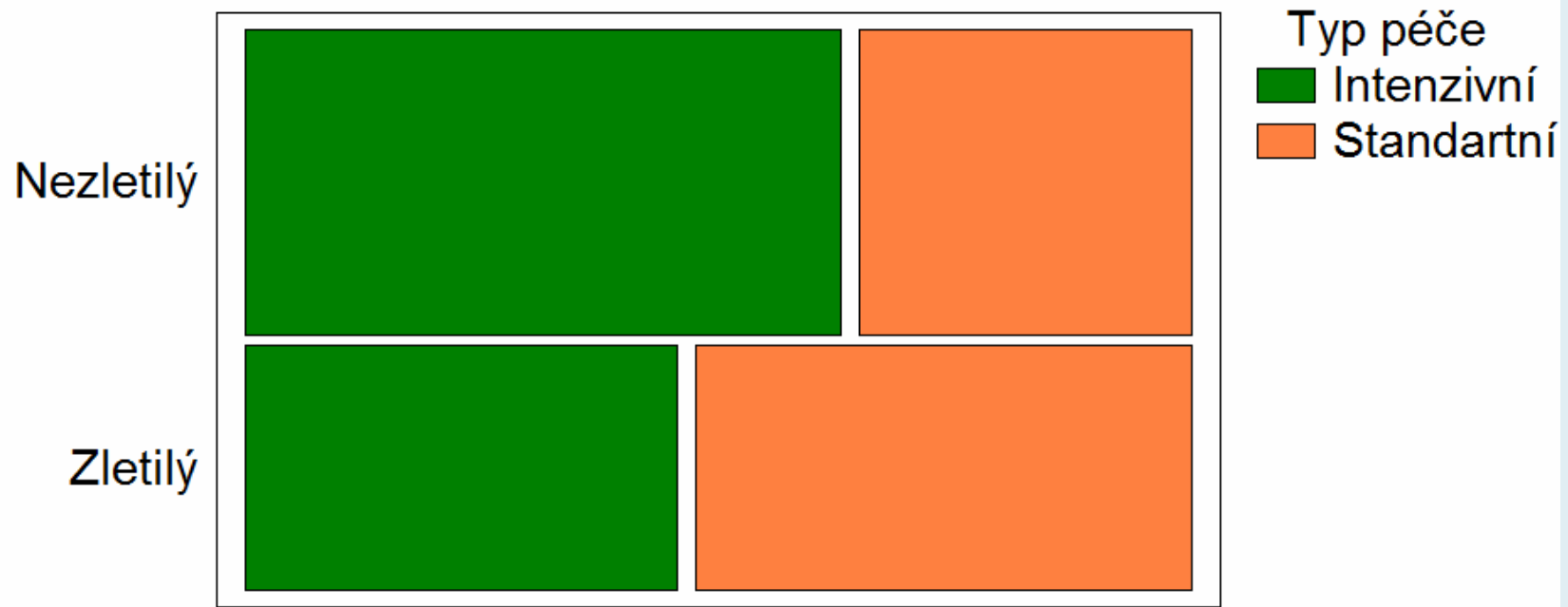
❖ Kumulativní sloupcový graf (excel)



Grafické výstupy pro analýzu závislosti dvou kategoriálních proměnných

❖ Mozaikový graf (Statgraphics)

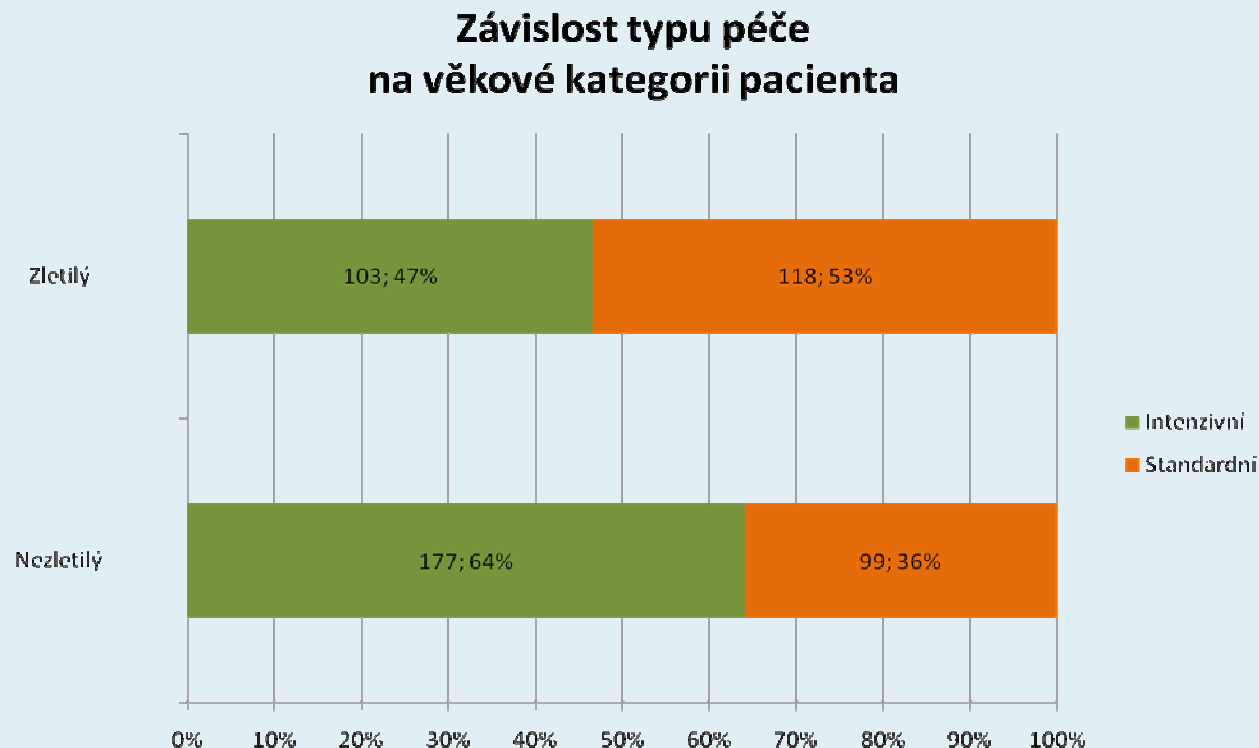
Závislost typu péče na věkové kategorii pacienta



Grafické výstupy pro analýzu závislosti dvou kategoriálních proměnných

❖ 100% skládaný pruhový graf

Všechny grafy se používají k explorační analýze závislosti



χ^2 -test nezávislosti v kontingenční tabulce

Základní pojmy:

- ❖ **Pozorované** (empirické) **četnosti** O_{ij} (Observed frequency) – zjištěné sdružené četnosti
- ❖ **Očekávané** (teoretické) **četnosti** E_{ij} (Expected frequency) – sdružené četnosti očekávané za předpokladu nezávislosti proměnných (aneb platí-li H_0)

$$E_{ij} = n_{ij}^* = \left(\frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} \right) \cdot n = \frac{n_{i.} \cdot n_{.j}}{n}$$

- ❖ **Odchyly, rezidua** (Deviation) – rozdíly mezi očekávanými a pozorovanými četnostmi = $E_{ij} - O_{ij}$

χ^2 -test nezávislosti v kontingenční tabulce

0. Předpoklady testu:

- ❖ Žádná očekávaná četnost nesmí klesnout pod 2
- ❖ Alespoň 80% četností musí být větších než 5

1. Nulová a alternativní hypotéza:

- ❖ H_0 : Proměnné v kontingenční tabulce jsou nezávislé.
- ❖ H_A : Proměnné v kontingenční tabulce jsou závislé.

2. Výpočet pozorované hodnoty testové statistiky: Pearsonova statistika chí-kvadrát

$$T(\underline{X}) = K = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \rightarrow \chi^2_{(m-1)(n-1)}$$

m ... počet řádků kont. tabulky, n ... počet sloupců kont. tabulky

χ^2 -test nezávislosti v kontingenční tabulce

V ideálním případě nezávislosti:

$$\forall i, j: \quad O_{ij} = E_{ij}$$

proto

$$O_{ij} - E_{ij} = 0$$

$$i=1, \dots, m; j=1, \dots, n$$

3. Výpočet p -value: $p\text{-value} = 1 - F_0(x_{OBS})$

4. Závěr

Podívejte se na flash animaci „Analýza závislosti dvou kategoriálních veličin“ na stránkách MI21

<http://mi21.vsb.cz/flash-animace/analyza-zavislosti-dvou-kategorialnich-velicin>

Yatesova korekce

- ❖ Lze provést v případě, kdy nejsou splněny předpoklady chí-kvadrát testu nezávislosti (extrémně nízké očekávané četnosti).
- ❖ Snižuje pravděpodobnost chyby I. druhu, tím však **snižuje sílu testu**.

Testová statistika (Pearsonova statistika chí-kvadrát):

$$T(\underline{X}) = K = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij} - 0,5)^2}{E_{ij}} \rightarrow X_{(m-1)(n-1)}^2$$

m, n ... počet řádků (sloupců) kont. tabulky

Výpočet p-value: $p - value = 1 - F_0(x_{OBS})$

Síla vztahu

Chí-kvadrát test nezávislosti nevypovídá nic o síle vztahu, pouze zamítá, resp. nezamítá nulovou hypotézu o nezávislosti znaku X a Y . Pro zjištění síly vztahu používáme koeficienty:

- ❖ koeficient kontingence – pro čtvercové kontingenční tabulky

$$CC = \sqrt{\frac{K}{K + n}}$$

- ❖ korigovaný koeficient kontingence – pro obdélníkové k. tab.

$$CC_{cor} = \frac{CC}{CC_{max}}, \text{ kde } CC_{max} = \sqrt{\frac{\min(r; s) - 1}{\min(r; s)}}$$

- ❖ Cramerův koeficient

$$V = \sqrt{\frac{K}{n(\min(r; s) - 1)}}$$

Čím jsou koeficienty blíže 1, tím je závislost mezi znaky X a Y těsnější.

McNemarův test

- ❖ Pouze pro čtyřpolní tabulky (2 sloupce a 2 řádky).
- ❖ Test shody rozdělení pro závislé alternativní proměnné se stejnými kódy.

Nulová a alternativní hypotéza:

H_0 : Procenta „úspěšností“ jsou u obou veličin stejná.

H_A : Procenta „úspěšností“ nejsou u obou veličin stejná.

Testové kritérium:
$$T(\underline{X}) = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \rightarrow \chi^2_{(1)}$$

Předpoklad testu:
$$\frac{(n_{12} + n_{21})}{2} \geq 4$$

Výpočet p-value:
$$p - value = 1 - F_0(x_{OBS})$$

Test

Vyberte správný výraz:

- a) Kolmogorovův-Smirnovův test ve své základní podobě (*lze, nelze*) použít pro testování normality.
- b) Použijeme-li χ^2 test dobré shody pro ověření toho, zda je klasická šestistěnná hrací kostka „férová“, pak má v případě platnosti nulové hypotézy testová statistika rozdělení s (4; 5; 6) stupni volnosti.

Vyberte správný výraz:

- a) Kolmogorovův-Smirnovův test ve své základní podobě (*lze*, *nelze*) použít pro testování normality.
- b) Použijeme-li χ^2 test dobré shody pro ověření toho, zda je klasická šestistěnná hrací kostka „férová“, pak má v případě platnosti nulové hypotézy testová statistika rozdělení s (*4*; *5*; *6*) stupni volnosti.

Vyberte správný výraz:

- c) Pro úplně specifikovaný test dobré shody se spojitým rozdělením je vhodnější použít (χ^2 test dobré shody, Kolmogorovův-Smirnovův test).
- d) Chceme-li pro ověření shody mezi teoretickým a empirickým rozdělením použít test dobré shody, musí být všechny (*pozorované, očekávané*) četnosti jednotlivých variant, resp. třídících intervalů, větší než 5.

Vyberte správný výraz:

- c) Pro úplně specifikovaný test dobré shody se spojitým rozdělením je vhodnější použít (χ^2 *test dobré shody*, *Kolmogorovův-Smirnovův test*).
- d) Chceme-li pro ověření shody mezi teoretickým a empirickým rozdělením použít test dobré shody, musí být všechny (*pozorované*, *očekávané*) četnosti jednotlivých variant, resp. třídících intervalů, větší než 5.

Vyberte správný výraz:

- e) Čím členitější je mozaikový graf, tím (*slabší, silnější*) závislost mezi veličinami v kontingenční tabulce pozorujeme.
- f) Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak χ^2 test nezávislosti můžeme použít, pokud alespoň (4; 10; 16; 20) očekávaných četností je větších než 5 a ostatní nejsou menší než (0; 1; 2).
- g) Koeficient kontingence (*se vyskytuje v intervalu (0;1); může nabývat hodnot větších než 1*).

Vyberte správný výraz:

- e) Čím členitější je mozaikový graf, tím (*slabší*, *silnější*) závislost mezi veličinami v kontingenční tabulce pozorujeme.
- f) Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak χ^2 test nezávislosti můžeme použít, pokud alespoň (*4*; *10*; *16*; *20*) očekávaných četností je větších než 5 a ostatní nejsou menší než (*0*; *1*; *2*).
- g) Koeficient kontingence (*se vyskytuje v intervalu (0;1)*; *může nabývat hodnot větších než 1*).

Vyberte správný výraz:

- h) (*Kontingenční, Asociační*) tabulka je speciálním případem (*kontingenční, asociační*) tabulky.
- i) Je-li odhad relativního rizika $\widehat{RR}=1,2$, pak (*mezi znaky v asociační tabulce existuje závislost, mezi znaky v asociační tabulce neexistuje závislost, o závislosti znaků v asociační tabulce musí rozhodnout test*).
- j) Kvalita 50 různých výukových materiálů byla dvěma odborníky hodnocena na stupnici od 1 do 5. Vhodnou mírou závislosti mezi hodnocením jednotlivých odborníků je (*Pearsonův, Spearmanův*) korelační koeficient.

Vyberte správný výraz:

- h) (*Kontingenční*, *Asociační*) tabulka je speciálním případem (*kontingenční*, *asociační*) tabulky.
- i) Je-li odhad relativního rizika $\widehat{RR}=1,2$, pak (*mezi znaky v asociační tabulce existuje závislost*, *mezi znaky v asociační tabulce neexistuje závislost*, *o závislosti znaků v asociační tabulce musí rozhodnout test*).
- j) Kvalita 50 různých výukových materiálů byla dvěma odborníky hodnocena na stupnici od 1 do 5. Vhodnou mírou závislosti mezi hodnocením jednotlivých odborníků je (*Pearsonův*, *Spearmanův*) korelační koeficient.