Závislost kategoriálních dat

Kontingenční tabulky

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

doc. Ing. Petr Seďa, Ph.D.

Úvod

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Co se dnes dozvíte?

- dvoustupňové třídění,
- kontingenční tabulka,
- čtyřpolní tabulka,
- míry kontingence, test nezávislosti.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Závislost statistických znaků:

Statistické znaky, které při analýze vícerozměrného statistického souboru zkoumáme, mohou být a také většinou jsou **na sobě závislé**.

Např. názor studentů na placení školného je ovlivněn jejich politickou orientací, známka u zkoušky ze statistiky je již částečně předznamenána výsledkem studenta z matematiky a nájem z bytu je ovlivněn jeho plochou.

Tuto závislost statistických znaků na sobě navzájem by bylo dobré umět **měřit nebo dokonce modelovat**.

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Závislost statistických znaků:

Jednostranná závislost - znak X působí na znak Y, avšak znak Y již nepůsobí zpětně na znak X.

		Typ znaku Y (důsledek)	
		kategoriální	kvantitativní
Typ znaku X (příčina)	kategoriální	analýza závislosti v kontingenčních, resp. v asociačních tabulkách	ANOVA
	kvantitativní	Diskriminační analýza, logistická regrese…	regresní a korelační analýza

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Kontingenční tabulka:

Kontingenční tabulka je vhodná k zobrazení **nominálních, popř. ordinálních znaků** dvourozměrného souboru.

Při zobrazení **metrických (číselných) znaků** je bez úprav vhodná k zobrazení znaků diskrétních (nespojitých) s **malým počtem různých obměn** (hodnot).

Chceme-li pomocí kontingenční tabulky zobrazit metrické spojité znaky, musíme nejprve provést jejich **kategorizaci** pomocí intervalového rozdělení četností stejně jako u jednoduchého třídění.

Kontingenční tabulka

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Dvoustupňové třídění:

kontingenční tabulka četností pro 2 statistické znaky

sdružené četnosti rozdělení znaků A a B

	b_1	b_2	…	Σ
a_1	n_{11}	n_{12}	…	n_{10}
a_2	n_{21}	n_{22}	…	n_{20}
…	…	…	…	…
Σ	n_{01}	n_{02}	…	n_{00}

počet prvků s vlastnostmi a_1 a b_1

marginální četnosti rozdělení znaku A

počet prvků s vlastností a_2

počet prvků souboru

marginální četnosti rozdělení znaku B

počet prvků s vlastností b_2

index 0 představuje „všechny prvky“

Příklad 1 – kontingenční tabulka

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Tabulka vyjadřuje závislost mezi vzděláním a politickou orientací u vzorku $n = 50$ osob:

n_{ij}		orientace			$\sum$
		levice	**střed**	**pravice**	
vzdělání	**ZŠ**	5	5	2	**12**
	SŠ	3	13	8	**24**
	VŠ	1	10	3	**14**
$\sum$		**9**	**28**	**13**	**50**

Měření asociace kvalitativních znaků v kontingenční tabulce

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Asociace:

- míra závislosti mezi četnostmi výskytu hodnot dvou kvalitativních znaků X a Y v kontingenční tabulce

hypotetické (očekávané) sdružené četnosti e_{ij} ← expected

$$e_{ij} = \frac{n_{i0} \cdot n_{0j}}{n}$$

pro nezávislé hodnoty x_i a y_j platí:

$$e_{ij} = n_{ij}$$

míra (ne)závislosti kvalitativních znaků G

$$G = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

r – počet řádků (obměn znaku X)
s – počet sloupců (obměn znaku Y)

Měření asociace kvalitativních znaků v kontingenční tabulce

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Míry asociace:

G jako Pearsonova χ^2 míra asociace znaků X a Y

$G = 0$ znaky X a Y jsou nezávislé

$G = n \cdot h$ znaky X a Y jsou maximálně závislé

$h = \min(r - 1;\ s - 1)$

Cramerův kontingenční koeficient V

$$V = \sqrt{\frac{G}{n \cdot h}}$$

$V = 0$ znaky jsou nezávislé

$V = 1$ znaky jsou maximálně závislé

Měření asociace kvalitativních znaků v kontingenční tabulce

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Cramerův kontingenční koeficient → slovní vyjádření

0	nezávislé proměnné
0,0 – 0,2	velmi slabá závislost
0,2 – 0,4	slabá závislost
0,4 – 0,7	střední závislost
0,7 – 0,9	vysoká závislost
0,9 – 1,0	velmi vysoká závislost
1	absolutní závislost

Příklad 2 – kontingenční tabulka

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Je politická orientace závislá na vzdělání? Spočtěte Cramerův koeficient kontingence.

n_{ij}		orientace			∑
		levice	střed	pravice	
vzdělání	ZŠ	5	5	2	**12**
	SŠ	3	13	8	**24**
	VŠ	1	10	3	**14**
∑		**9**	**28**	**13**	**50**

e_{ij}		orientace			∑
		levice	střed	pravice	
vzdělání	ZŠ	2,16	6,72	3,12	**12**
	SŠ	4,32	13,4	6,24	**24**
	VŠ	2,52	7,84	3,64	**14**
∑		**9**	**28**	**13**	**50**

$$G = \frac{(5-2,16)^2}{2,16} + \frac{(5-6,72)^2}{6,72} + \ldots + \frac{(3-3,64)^2}{3,64} = 7,11$$

Cramerův koeficient: $V = \sqrt{\frac{G}{n \cdot h}} = \sqrt{\frac{7,11}{50 \cdot 2}} = 0,27$ ← slabá závislost

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Vyhodnocení výsledků:

Pokud vyjde hodnota míry asociace G nebo Cramerův koeficientu kontingence nenulová, znamená to, že mezi oběma sledovanými znaky **existuje určitá asociace** (závislost).

To však nelze takto s jistotou tvrdit v případě, že zkoumaný soubor je výběrový, tedy představuje pouze vzorek ze základního souboru.

Již z předchozího studia statistiky víte, že vzorek není věrným obrazem základního souboru, každý výběr je **zatížen určitou chybou**, kterou nemůžeme zcela ovlivnit, a tak lze závěry našich statistických analýz stanovit pouze s určitou spolehlivostí.

V praxi to znamená, že pokud není hodnota míry asociace G dostatečně vysoká, **nemusí to být přesvědčivý důkaz** o závislosti obou znaků.

Měření asociace kvalitativních znaků v kontingenční tabulce

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Test nezávislosti:

H_0: sledované znaky jsou nezávislé ($G = 0$)

H_1: sledované znaky jsou závislé ($G > 0$)

testové kritérium: Pearsonova míra asociace G

kritická hodnota: $\chi^2_{1-\alpha}(v)$, kde $v = (r-1)\cdot(s-1)$

$G > \chi^2_{1-\alpha}(v)$ zamítáme nulovou hypotézu, že sledované znaky jsou nezávislé

Příklad 2 – kontingenční tabulka

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Je politická orientace závislá na vzdělání? Použijte test nezávislosti v kontingenční tabulce.

- testujeme závislost politické orientace na vzdělání na hladině významnosti $\alpha = 0{,}05$

testové kritérium: $G = 7{,}11$

stupeň volnosti $v = 2\ x\ 2 = 4$

kritická hodnota: $\chi^2_{0{,}95}\ (4) = 9{,}49$

Závěr: závislost politické orientace na vzdělání **není** prokázána, protože testové kritérium je menší než kritická hodnota testu.

Míry asociace ve čtyřpolní tabulce

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Speciální typ kontingenčních tabulek, které používáme k sledování závislosti dvou **dichotomických znaků**, tj. kategoriálních znaků nabývajících pouze dvou variant. (asociace = vztah dvou dichotomických znaků).

Na asociační (čtyřpolní) tabulku lze sice nahlížet jako na **speciální případ** kontingenčních tabulek a při analýze používat jejich aparát, nicméně vhodnější je využít **specifické metody a charakteristiky asociace**.

U závislosti dvou dichotomických znaků můžeme měřit nejen **sílu, ale také její směr**.

$X\ (okolnosti)\backslash Y\ (výskyt\ události)$	$y_{[1]}$ (úspěch)	$y_{[2]}$ (neúspěch)	Celkem
$x_{[1]}$ (I.)	a	b	$a+b$
$x_{[2]}$ (II.)	c	d	$c+d$
Celkem	$a+c$	$b+d$	n

Pozitivní závislost znamená, že odpověď x_1 na jednu otázku znamená také převážně odpověď y_1 na otázku druhou.

Negativní závislost znamená, že respondenti, kteří odpověděli na jednu otázku x_1, na druhou odpovídali spíše y_2 a naopak.

Míry asociace ve čtyřpolní tabulce

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

$X\ (okolnosti) \backslash Y\ (výskyt\ události)$	$y_{[1]}$ (úspěch)	$y_{[2]}$ (neúspěch)	Celkem
$x_{[1]}$ (I.)	a	b	$a+b$
$x_{[2]}$ (II.)	c	d	$c+d$
Celkem	$a+c$	$b+d$	n

Koeficient asociace ve čtyřpolní tabulce lze vypočítat jako: $r = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$.

Kladná **hodnota koeficientu r znamená pozitivní závislost**, záporná pak závislost negativní. Absolutní hodnota $|r|$ vyjadřuje intenzitu této závislosti a **má stejný význam jako Cramerův koeficient kontingence**.

Všimněte si, že ve čtyřpolní tabulce mezi mírou asociace G a koeficientem asociace r platí vztah: $G = N \cdot r^2$.

Test nezávislosti ve čtyřpolní tabulce

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

X (okolnosti)\Y(výskyt události)	$y_{[1]}$ (úspěch)	$y_{[2]}$ (neúspěch)	Celkem
$x_{[1]}$ (I.)	a	b	$a+b$
$x_{[2]}$ (II.)	c	d	$c+d$
Celkem	$a+c$	$b+d$	n

H_0: sledované znaky jsou nezávislé $(G = 0)$

H_1: sledované znaky jsou závislé $(G > 0)$

Pro čtyřpolní tabulku lze souhrnnou **χ^2-míru asociace** vypočítat dle vztahu:

$$G = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

Statistika G má v případě nezávislosti obou znaků rozdělení χ^2 o 1 stupni volnosti.

Kritický obor je tedy vymezen jako množina hodnot vyšších než kvantil $\chi^2_{1-\alpha}(1)$.

Příklad 3 – čtyřpolní tabulka

Při průzkumu vlivu TV reklamy na prodej nového výrobku odpovědělo 200 respondentů na dotaz, zda viděli reklamu na nový výrobek a zda si výrobek také koupili. Na obě otázky odpovídali na alternativní škále ANO - NE.

a) Určete koeficient asociace a zhodnoťte závislost mezi znalostí reklamy a nákupem u testovaného vzorku osob.

b) Proveďte na hladině významnosti 5% test nezávislosti a zhodnoťte závislost mezi znalostí reklamy a nákupem, budeme-li považovat výběr respondentů za reprezentativní.

Znalost \ Nákup	NE	ANO	CELKEM
NE	58	15	73
ANO	89	38	127
CELKEM	147	53	200

Ad a) koeficient asociace: $r = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = 0{,}102.$

Velmi nízká hodnota koeficientu asociace znamená **velmi slabou pozitivní závislost** mezi znalostí reklamy a nákupem u testovaného vzorku osob.

Příklad 3 – čtyřpolní tabulka

SLEZSKÁ UNIVERZITA
FAKULTA VEŘEJNÝCH POLITIK V OPAVĚ

Znalost \ Nákup	NE	ANO	CELKEM
NE	58	15	73
ANO	89	38	127
CELKEM	147	53	200

Ad b) H_0: Účinnost TV reklamy na prodej nového výrobku není prokázána.
H_1: Účinnost TV reklamy na prodej nového výrobku je prokázána.

Testovou statistiku G spočítáme dle vztahu: $G = N \cdot r^2 = 200 \cdot 0{,}102^2 = 2{,}08$.

Protože hodnota testového kritéria $G = 2{,}08$ je nižší než kvantil $\chi^2_{0{,}95}(1) = 3{,}84$, **nemůžeme** na hladině významnosti 5% **zamítnout nulovou hypotézu**. U zákaznické populace nebyla prokázána účinnost TV reklamy na prodej nového výrobku.

Literatura

1. Ramík J. a Čemerková Š. *Statistika A.* Opava, Karviná: SLU, 2000. **(kapitola 9).**

Děkuji za pozornost.