Analýza kategorických proměnných

POLb 1139

Peter Spáč

Vícerozměrná analýza

- Jednorozměrná analýza přináší informace o jednotlivých proměnných
- Cílem (nejen) statistiky je identifikovat vztahy mezi proměnnými za účelem lepšího poznání reality
- Praktickým vyjádřením této snahy je vícerozměrná analýza – souhrn postupů, které zahrnují vícero proměnných

Vícerozměrná analýza

- Jaký je vztah mezi vzděláním a výškou příjmu?
- Souvisí čas odevzdání seminární práce s jejím hodnocením?
- Mají starší lidé vyšší pravděpodobnost účasti ve volbách?
- Liší se známky studentů v závislosti na tom, zda výuka probíhá osobně anebo online?

Co je důležité vědět?

- Jaké postupy jsou vhodné pro jaká data
- Jaké jsou silné stránky a limity daných postupů
- Jak chápat a interpretovat zjištění daných postupů
- V čem je rozdíl mezi statistickou a věcnou významností
- A nakonec, že je omyl domnívat se, že software udělá 99 % práce

Vztahy dvou proměnných

- Podoba analýzy závisí na typu proměnných

- Kontingenční tabulky (crosstabs):
  - Dvě kategorické proměnné – nominální, ordinální
  - Nižší počet kategorií v proměnných (podmínka jsou minimálně dvě)

- Korelace (correlation):
  - Dvě kardinální proměnné, kardinální a ordinální, dvě ordinální
  - Specifický případ – kardinální a dichotomická proměnná

Kontingenční tabulky

- Cross-tabulation, crosstabs

- Vztah mezi dvěma kategorickými proměnnými
  - Nominální, ordinální

- Příklady:
  - Věkové skupiny v. účast ve volbách
  - Pohlaví v. příjmové skupiny

Pozorované četnosti (observed)

Pclass	Survived		Total
	0	1	
1st	129	193	322
2nd	160	119	279
3rd	573	138	711
Total	862	450	1312

Pozorované četnosti (observed) + řádková procenta (row)

Pclass		Survived		Total
		0	1	
1st	Count	129	193	322
	% within row	40.1 %	59.9 %	100.0 %
2nd	Count	160	119	279
	% within row	57.3 %	42.7 %	100.0 %
3rd	Count	573	138	711
	% within row	80.6 %	19.4 %	100.0 %
Total	Count	862	450	1312
	% within row	65.7 %	34.3 %	100.0 %

Pozorované četnosti (observed) + sloupcová procenta (column)

Pclass		Survived		Total
		0	1	
1st	Count	129	193	322
	% within column	14.9 %	42.9 %	24.5 %
2nd	Count	160	119	279
	% within column	18.6 %	26.4 %	21.3 %
3rd	Count	573	138	711
	% within column	66.5 %	30.7 %	54.2 %
Total	Count	862	450	1312
	% within column	100.0 %	100.0 %	100.0 %

Pclass		Survived		Total
		0	1	
1st	Count	129	193	322
	% within row	40.1 %	59.9 %	100.0 %
	% within column	14.9 %	42.9 %	24.5 %
	% of total	9.8 %	14.7 %	24.5 %
2nd	Count	160	119	279
	% within row	57.3 %	42.7 %	100.0 %
	% within column	18.6 %	26.4 %	21.3 %
	% of total	12.2 %	9.1 %	21.3 %
3rd	Count	573	138	711
	% within row	57.3 %	42.7 %	100.0 %
	% within column	66.5 %	30.7 %	54.2 %
	% of total	43.7 %	10.5 %	54.2 %
Total	Count	862	450	1312
	% within row	57.3 %	42.7 %	100.0 %
	% within column	100.0 %	100.0 %	100.0 %
	% of total	65.7 %	34.3 %	100.0 %

Existuje vztah mezi vzděláním a volební účastí?

Vzdělání	Účast		Total
	Ne	Ano	
ZŠ	67	65	132
SŠ bez M	336	441	777
SŠ s M	346	655	1001
VŠ	98	293	391
Total	847	1454	2301

Existuje vztah mezi vzděláním a volební účastí?

Vzdělání		Účast		Total
		Ne	Ano	
ZŠ	Count	67	65	132
	% within row	50.8 %	49.2 %	100.0 %
SŠ bez M	Count	336	441	777
	% within row	43.2 %	56.8 %	100.0 %
SŠ s M	Count	346	655	1001
	% within row	34.6 %	65.4 %	100.0 %
VŠ	Count	98	293	391
	% within row	25.1 %	74.9 %	100.0 %
Total	Count	847	1454	2301
	% within row	36.8 %	63.2 %	100.0 %

Existuje vztah mezi vzděláním a volební účastí?

Vzdělání		Účast		Total
		Ne	Ano	
ZŠ	Count	67	65	132
	% within row	50.8 %	49.2 %	100.0 %
SŠ b[illegible] M[illegible]	Count	336	441	777
	% within row	43.2 %	56.8 %	100.0 %
SŠ	Count	346	655	1001
	% within row	34.6 %	65.4 %	100.0 %
VŠ	Count	98	293	391
	% within row	25.1 %	74.9 %	100.0 %
Total	Count	847	1454	2301
	% within row	36.8 %	63.2 %	100.0 %

Lidé s vyšším vzděláním se voleb zúčastnili ve vyšší míře.

Dá se ale tento závěr uplatnit i na celou **populaci** ČR?

Pozorované vs. očekávané četnosti

- Klíčové pro pochopení logiky kontingenčních tabulek
- Pozorované četnosti (observed) – reálná pozorování spadající do konkrétní kategorie
- Očekávané četnosti (expected) – početnost, která by se v konkrétní kategorii měla pozorovat za předpokladu nezávislosti obou proměnných
- Základní prvky pro výpočet chí-kvadrátu

Pozorované četnosti (observed) + očekávané četnosti (expected)

Vzdělání		Účast		Total
		Ne	Ano	
ZŠ	Count	67	65	132
	Expected count	48.6	83.4	132.0
SŠ bez M	Count	336	441	777
	Expected count	286.0	491.0	777.0
SŠ s M	Count	346	655	1001
	Expected count	368.5	632.5	1001
VŠ	Count	98	293	391
	Expected count	143.9	247.1	391.0
Total	Count	847	1454	2301
	Expected count	847.0	1454.0	2301.0

Test Chí-kvadrát

- Posuzuje, zda jsou rozdíly mezi pozorovanými a očekávanými četnostmi natolik výrazné, aby nebyly pouze výsledkem náhody

- Je nutné si dát pozor na malé počty pozorování:
  - 5 a méně pozorování v méně než 20 % kategorií
  - Kategorie s nenulovými pozorováními

- Pro malé vzorky (JASP):
  - χ^2 continuity correction
  - Likelihood ratio

	Value	df	p
X^2	50.225	3	< .001
X^2 continuity correction	50.225	3	< .001
Likelihood ratio	50.934	3	< .001
N	2301		

Mezi vzděláním a účastí ve volbách existuje signifikantní vztah → platí pro populaci

	Value
Contingency coefficient	0.146
Cramer's V	0.148

Rezidua

- Testy závislosti mezi proměnnými ukáží, zda mezi proměnnými existuje anebo neexistuje asociace
- Pro věcné pochopení vztahu je důležité poznat více detailů
- Pro tento účel sledujeme standardizovaná rezidua:
  - Vyjadřují standardizovaný rozdíl mezi pozorovanými a očekávanými četnostmi
- JASP automaticky nepočítá – potřebná manuální kalkulace

Standardizovaná rezidua

- **Pro výpočet se využívají z-scores**

- $$z = \frac{pozorovaná\ četnost - očekávaná\ četnost}{\sqrt{očekávaná\ četnost}}$$

- **V následném kroku se naměřená hodnota porovná s používanými hladinami signifikantnosti:**
  - **$\pm 1{,}96 \rightarrow 95\ \%$**
  - **$\pm 2{,}58 \rightarrow 99\ \%$**
  - **$\pm 3{,}29 \rightarrow 99{,}9\ \%$**

Vzdělání		Účast		Total
		Ne	Ano	
ZŠ	Count	67	65	132
	Expected count	48.6	83.4	132.0
SŠ bez M	Count	336	441	777
	Expected count	286.0	491.0	777.0
SŠ s M	Count	346	655	1001
	Expected count	368.5	632.5	1001
VŠ	Count	98	293	391
	Expected count	143.9	247.1	391.0
Total	Count	847	1454	2301
	Expected count	847.0	1454.0	2301.0

Kategorie – ZŠ bez volební účasti

- **Vstupní hodnoty:**
  - **Pozorovaná četnost = 67**
  - **Očekávaná četnost = 48,6**

- $z = \frac{pozor.četn. - očekávaná.četn.}{\sqrt{očekávaná.četn.}}$ = (67 – 48,6) / $\sqrt{48,6}$ = 18,4 / 6,97 = **2,64**

- **V následném kroku se naměřená hodnota porovná s používanými hladinami signifikantnosti:**
  - **± 1,96 → 95 %**
  - **± 2,58 → 99 %**
  - **± 3,29 → 99,9 %**

Vzdělání		Účast		Total
		Ne	Ano	
ZŠ	Count	67	65	132
	Expected count	48.6	83.4	132.0
	St. Res.	2.6	-2.0	
SŠ bez M	Count	336	441	777
	Expected count	286.0	491.0	777.0
	St. Res.	3.0	-2.3	
SŠ s M	Count	346	655	1001
	Expected count	368.5	632.5	1001
	St. Res.	-1.2	0.9	
VŠ	Count	98	293	391
	Expected count	143.9	247.1	391.0
	St. Res.	-3.8	2.9	
Total	Count	847	1454	2301
	Expected count	847.0	1454.0	2301.0

Které skupiny podle vzdělání by volily častěji / méně často oproti předpokladu nezávislosti obou proměnných?

Vzdělání		Účast		Total
		Ne	Ano	
ZŠ	Count	67	65	132
	St. Res.	2.6**	-2.0*	
SŠ bez M	Count	336	441	777
	St. Res.	3.0**	-2.3*	
SŠ s M	Count	346	655	1001
	St. Res.	-1.2	0.9	
VŠ	Count	98	293	391
	St. Res.	-3.8***	2.9**	
Total	Count	847	1454	2301

Adjustovaná standardizovaná rezidua

- Stejná logika jako u standardizovaných reziduí + zohledňují počet pozorování
- Pro výpočet se využívají z-scores (o něco komplikovanější rovnice)
- $$z = \frac{pozorovaná\ četnost - očekávaná\ četnost}{\sqrt{n\check{R} * nS * \left(1 - \frac{n\check{R}}{N}\right) * \left(1 - \frac{nS}{N}\right) * \frac{1}{N}}}$$
- nŘ = počet případů v řádku, nS = počet případů v sloupci, N = počet všech případů
- V následném kroku se naměřená hodnota porovná s používanými hladinami signifikantnosti:
  - ± 1,96 → 95 %
  - ± 2,58 → 99 %
  - ± 3,29 → 99,9 %

Kategorie – ZŠ bez volební účasti

- **Vstupní hodnoty:**
  - **Pozorovaná četnost = 67, očekávaná četnost = 48,6**
  - **Případů v řádku = 132, případů v sloupci = 847, počet všech případů = 2301**

- $z = \frac{pozorovaná\ četnost - očekávaná\ četnost}{\sqrt{nŘ * nS * \left(1-\frac{nŘ}{N}\right) * \left(1-\frac{nS}{N}\right) * \frac{1}{N}}} = (67 - 48{,}6) / \sqrt{132 * 847 * \left(1-\frac{132}{2301}\right) * \left(1-\frac{847}{2301}\right) * \frac{1}{2301}} =$

$18{,}4 / \sqrt{28{,}94} = 18{,}4 / 5{,}38 = \underline{3{,}42}$

- **V následném kroku se naměřená hodnota porovná s používanými hladinami signifikantnosti:**
  - **± 1,96 → 95 %**
  - **± 2,58 → 99 %**
  - **± 3,29 →** 99,9 %

Které skupiny podle vzdělání by volily častěji / méně často oproti předpokladu nezávislosti obou proměnných?

Vzdělání		Účast		Total
		Ne	Ano	
ZŠ	Count	67	65	132
	Adj. St. Res.	3.42***	-3.42***	
SŠ bez M	Count	336	441	777
	Adj. St. Res.	4.57***	-4.57***	
SŠ s M	Count	346	655	1001
	Adj. St. Res.	-1.95	1.95	
VŠ	Count	98	293	391
	Adj. St. Res.	-5.29***	5.29***	
Total	Count	847	1454	2301

Shrnutí

- Kontingenční tabulky jako nástroj pro zobrazení vztahu mezi dvěma kategorickými proměnnými
- Pomocí jednotlivých testů je možné identifikovat existenci a sílu vztahu mezi proměnnými
- Důležité je vnímat věcný rozměr zjištění
- Pozor na příliš obsáhlé kontingenční tabulky
  - Náročnější na interpretaci
  - Zbytečné zahlcení publika množstvím údajů (pozorované četnosti, očekávané četnosti, řádková procenta, sloupcová procenta, rezidua)
  - Hrozí, že v části kategorií bude jen malý počet hodnot