Analýza kategorických proměnných - crosstabs

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 než mladí?
- 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**

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:
  - pohlaví X účast ve volbách
  - Sociální třída X vzdělání

Příklad

- Souvislost mezi přežitím (0/1) a cestovní třídou (1-3) mezi účastníky potopení Titanicu

SPSS

- Analyze → Descriptive Statistics → Crosstabs
  - Rows: PClass
  - Columns: Survived

Pozorované četnosti (Observed)

Pclass * Survived Crosstabulation

Count

		Survived		
		0	1	Total
Pclass	1st	80	136	216
	2nd	97	87	184
	3rd	368	119	487
Total		545	342	887

SPSS

- Analyze → Descriptive Statistics → Crosstabs
    - Cells: Counts: observed, Percentages: rows

Pozorované četnosti (Observed) + řádková procenta (Row)

Pclass * Survived Crosstabulation

			Survived		
			0	1	Total
Pclass	1st	Count	80	136	216
		% within Pclass	37,0%	63,0%	100,0%
	2nd	Count	97	87	184
		% within Pclass	52,7%	47,3%	100,0%
	3rd	Count	368	119	487
		% within Pclass	75,6%	24,4%	100,0%
Total		Count	545	342	887
		% within Pclass	61,4%	38,6%	100,0%

řádková procenta dávají součet 100 % na konci řádku

Šance

- Pojem důležitý později pro logistickou regresi
  - Pro kontingenční tabulky jen terminologický problém
  - Šance ≠ pravděpodobnost!!!
- Šance = poměr pravděpodobností mezi tím, že jev nastane a nenastane
  - Hod korunou: šance, že padne orel je 1:1 (pravděpodobnost je 0,5)
  - Hod kostkou: šance, že padne šestka je 1:5 (pravděpodobnost je 0,16)

- Šance na přežití v 1. třidě je 63:37 (cca 1,7)
- Šance na přežití ve 3. třidě je 25:75 (cca 0,3)
- Šance na přežití v 1. třidě je zhruba 6x větší než ve 3.třídě
- Pravděpodobnost přežití v 1. třídě vyšší 2,5x (je vyšší o 40 procentních bodů)

SPSS

- Analyze → Descriptive Statistics → Crosstabs
    - Cells: Counts: observed, Percentages: columns

Pozorované četnosti (Observed) + sloupcová procenta (Column)

Pclass * Survived Crosstabulation

			Survived		
			0	1	Total
Pclass	1st	Count	80	136	216
		% within Survived	14,7%	39,8%	24,4%
	2nd	Count	97	87	184
		% within Survived	17,8%	25,4%	20,7%
	3rd	Count	368	119	487
		% within Survived	67,5%	34,8%	54,9%
Total		Count	545	342	887
		% within Survived	100,0%	100,0%	100,0%

sloupcová procenta dávají součet 100 % vespod sloupce

SPSS

- Analyze → Descriptive Statistics → Crosstabs
    - Cells: Counts: observed, Percentages: columns, rows, total

Pclass * Survived Crosstabulation

			Survived		
			0	1	Total
Pclass	1st	Count	80	136	216
		% within Pclass	37,0%	63,0%	100,0%
		% within Survived	14,7%	39,8%	24,4%
		% of Total	9,0%	15,3%	24,4%
	2nd	Count	97	87	184
		% within Pclass	52,7%	47,3%	100,0%
		% within Survived	17,8%	25,4%	20,7%
		% of Total	10,9%	9,8%	20,7%
	3rd	Count	368	119	487
		% within Pclass	75,6%	24,4%	100,0%
		% within Survived	67,5%	34,8%	54,9%
		% of Total	41,5%	13,4%	54,9%
Total		Count	545	342	887
		% within Pclass	61,4%	38,6%	100,0%
		% within Survived	100,0%	100,0%	100,0%
		% of Total	61,4%	38,6%	100,0%

Crosstab věk x přežití?

Age * Survived Crosstabulation

			Survived		
			0	1	Total
Age	1	Count	2	12	14
		% within Age	14,3%	85,7%	100,0%
	2	Count	7	4	11
		% within Age	63,6%	36,4%	100,0%
	3	Count	2	5	7
		% within Age	28,6%	71,4%	100,0%
	4	Count	3	8	11
		% within Age	27,3%	72,7%	100,0%
	5	Count	2	4	6
		% within Age	33,3%	66,7%	100,0%
	6	Count	1	2	3
		% within Age	33,3%	66,7%	100,0%
	7	Count	3	2	5
		% within Age	60,0%	40,0%	100,0%
	8	Count	4	2	6
		% within Age	66,7%	33,3%	100,0%
	9	Count	6	2	8
		% within Age	75,0%	25,0%	100,0%
	10	Count	2	0	2
		% within Age	100,0%	0,0%	100,0%
	11	Count	3	1	4
		% within Age	75,0%	25,0%	100,0%
	12	Count	1	1	2
		% within Age	50,0%	50,0%	100,0%
	13	Count	0	2	2
		% within Age	0,0%	100,0%	100,0%
	14	Count	4	3	7
		% within Age	57,1%	42,9%	100,0%
	15	Count	2	4	6
		% within Age	33,3%	66,7%	100,0%
	16	Count	13	7	20
		% within Age	65,0%	35,0%	100,0%
	17	Count	10	6	16
		% within Age	62,5%	37,5%	100,0%
	18	Count	23	13	36
		% within Age	63,9%	36,1%	100,0%
	19	Count	22	11	33
		% within Age	66,7%	33,3%	100,0%

řešení

- Rekódování věku do kategorií
  - Věcný smysl hranic intervalů (dospělost, důchodový věk,….)
  - Kvantily (různé možnosti počtu kategorií)
- Transform -> recode into different variable
  - Input: age
  - Output name: vek_kat, potom change
  - Old and new values
  - Range, lowest through values: 20 -> new value 1 -> Add
  - Range 21 through 30 -> new value 2 -> Add
  - Range 31 through 40 -> new value 3 -> Add
  - Range, value through Highest: 41 -> new value 4 -> Add
  - Continue, ok

Příklad č. 2:

- Jak spolu souvisí vzdělání s volební účastí?

- Data: reprezentativní vzorek

SPSS

- Soubor ESS9CZ – data z European Social Survey 2018
- Analyze → Descriptive Statistics → Crosstabs
  - Rows: Vzdelani_4kat
  - Columns: ucast
  - Cells: Counts: observed

Existuje vztah mezi vzděláním a volební účastí?

Vzdelani_4kat * Ucast Crosstabulation

Count

		Ucast		
		Ne	Ano	Total
Vzdelani_4kat	ZŠ	67	65	132
	SŠbezM	336	441	777
	SŠsM	346	655	1001
	VŠ	98	293	391
Total		847	1454	2301

SPSS

- Soubor ESS9CZ
- Analyze → Descriptive Statistics → Crosstabs
  - Rows: Vzdelani_4kat
  - Columns: ucast
  - Cells: Counts: observed, percentages: Rows

Existuje vztah mezi vzděláním a volební účastí?

Vzdelani_4kat * Ucast Crosstabulation

			Ucast		
			Ne	Ano	Total
Vzdelani_4kat	ZŠ	Count	67	65	132
		% within Vzdelani_4kat	50,8%	49,2%	100,0%
	VŠ	Count	98	293	391
		% within Vzdelani_4kat	25,1%	74,9%	100,0%
Total		Count	847	1454	2301
		% within Vzdelani_4kat	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) – četnost, která by se v konkrétní kategorii měla pozorovat za předpokladu nezávislosti obou proměnných
- Základní prvky pro výpočet chí-kvadrátu

	muži	ženy	celkem
Volební účast ano	100	100	200
Volební účast ne	100	100	200
celkem	200	200	**400**

	muži	ženy	celkem
Volební účast ano	240	60	300
Volební účast ne	80	20	100
celkem	320	80	**400**

	muži	ženy	celkem
Volební účast ano	50	150	300
Volební účast ne	150	50	100
celkem	200	200	**400**

SPSS

- Analyze → Descriptive Statistics → Crosstabs
    - Cells: Counts: observed, expected

Pozorované četnosti (Observed) + očekávané četnosti (Expected)

Vzdelani_4kat * Ucast Crosstabulation

			Ucast		
			Ne	Ano	Total
Vzdelani_4kat	ZŠ	Count	67	65	132
		Expected Count	48,6	83,4	132,0
	SŠbezM	Count	336	441	777
		Expected Count	286,0	491,0	777,0
	SŠsM	Count	346	655	1001
		Expected Count	368,5	632,5	1001,0
	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
  - **Tj. statisticky významné**

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

- Analyze → Descriptive Statistics → Crosstabs
  - Statistics: Chi-square, Phi and Cramers V

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	50,225[a]	3	,000
Likelihood Ratio	50,934	3	,000
Linear-by-Linear Association	50,061	1	,000
N of Valid Cases	2301		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 48,59.

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

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,148	,000
	Cramer's V	,148	,000
N of Valid Cases		2301	

Cramer’s V (statology.org)

Degrees of freedom	Small	Medium	Large
1	0.10	0.30	0.50
2	0.07	0.21	0.35
3	0.06	0.17	0.29
4	0.05	0.15	0.25
5	0.04	0.13	0.22

- df = minimum(řádky -1, sloupce -1)
- Tabulky 2x2, 4x2 nebo 10x2 mají stále jen 1 df
- Odlišné od df pro Chí kvadrát !!!

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 adj. standardizované rezidua:
  - Vyjadřují standardizovaný rozdíl mezi pozorovanými a očekávanými četnostmi

SPSS

- Analyze → Descriptive Statistics → Crosstabs
  - Cells: Counts: observed, expectedl
  - Statistics: Chi-square et al.

Pozorované četnosti (Observed) + očekávané četnosti (Expected) + nestandard. rezidua (Unstandardized)

Vzdelani_4kat * Ucast Crosstabulation

			Ucast		
			Ne	Ano	Total
Vzdelani_4kat	ZŠ	Count	67	65	132
		Expected Count	48,6	83,4	132,0
		Residual	18,4	-18,4	
	SŠbezM	Count	336	441	777
		Expected Count	286,0	491,0	777,0
		Residual	50,0	-50,0	
	SŠsM	Count	346	655	1001
		Expected Count	368,5	632,5	1001,0
		Residual	-22,5	22,5	
	VŠ	Count	98	293	391
		Expected Count	143,9	247,1	391,0
		Residual	-45,9	45,9	
Total		Count	847	1454	2301
		Expected Count	847,0	1454,0	2301,0

Standard. 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 → 95 %
  - $\pm$ 2,58 → 99 %
  - $\pm$ 3,29 → 99,9 %
- Adj. standard. rezidua mají výpočet částečně odlišný (obsah pod odmocninou)

SPSS

- Analyze → Descriptive Statistics → Crosstabs
    - Cells: Counts: observed, expected
        - Residuals: unstandardized

SPSS

- Analyze → Descriptive Statistics → Crosstabs
    - Cells: Counts: observed, expected
        - Residuals: adjusted standardized

Pozorované četnosti (Observed) + očekávané četnosti (Expected) + adj. standard. rezidua (Adj. St. Res.)

Vzdelani_4kat * Ucast Crosstabulation

			Ucast		
			Ne	Ano	Total
Vzdelani_4kat	ZŠ	Count	67	65	132
		Expected Count	48,6	83,4	132,0
		Adjusted Residual	3,4	-3,4	
	SŠbezM	Count	336	441	777
		Expected Count	286,0	491,0	777,0
		Adjusted Residual	4,6	-4,6	
	SŠsM	Count	346	655	1001
		Expected Count	368,5	632,5	1001,0
		Adjusted Residual	-2,0	2,0	
	VŠ	Count	98	293	391
		Expected Count	143,9	247,1	391,0
		Adjusted Residual	-5,3	5,3	
Total		Count	847	1454	2301
		Expected Count	847,0	1454,0	2301,0

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

Vzdelani_4kat * Ucast Crosstabulation

			Ucast		
			Ne	Ano	Total
Vzdelani_4kat	ZŠ	Count	67	65	132
		Adjusted Residual	3,4	-3,4	
	SŠbezM	Count	336	441	777
		Adjusted Residual	4,6	-4,6	
	SŠsM	Count	346	655	1001
		Adjusted Residual	-2,0	2,0	
	VŠ	Count	98	293	391
		Adjusted Residual	-5,3	5,3	
Total		Count	847	1454	2301

Ne každý statisticky významný výsledek má smysl

group * mayor Crosstabulation

			mayor				
			coalition	idependent	opposition	SMER	Total
group	1	Count	85	365	252	297	999
		Expected Count	89,0	380,0	240,8	289,2	999,0
		Adjusted Residual	-,5	-1,2	1,0	,7	
	2	Count	82	395	222	278	977
		Expected Count	87,1	371,6	235,5	282,9	977,0
		Adjusted Residual	-,7	1,9	-1,2	-,4	
	3	Count	94	354	232	273	953
		Expected Count	84,9	362,5	229,7	275,9	953,0
		Adjusted Residual	1,3	-,7	,2	-,3	
Total		Count	261	1114	706	848	2929
		Expected Count	261,0	1114,0	706,0	848,0	2929,0

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
- Jde jen o souvislost, vliv 3. proměnných dokáže odfiltrovat logistická regrese