Analýza kontingenčních tabulek

4 základních kroků/otázek

0) Příprava
1) Existuje nějaký vztah?
2) Jak silný je vztah?
3) Kde přesně se vztah objevuje?

Krok 0 - Příprava

Před analýzou samotnou je nutné ještě udělat dvě základní věci. Zaprvé je nutné formulovat hypotézu, kterou budeme testovat. Tu už většinou známe dopředu, případně ji formulujeme na základě proměnných v tabulce. Důležité je aby bylo možné naší hypotézu převést na nulovou hypotézu.

Druhou věc, kterou je nutné udělat, je kontrola souhrnné údaje o vzorku. Jde hlavně o kontrolu chybějících proměnných. Pokud chybí větší část respondentů (čtvrtina a více), mělo by to mít nějaké vysvětlení (např. byli odfiltrováni – neptáme se na délku rodičovské lidí, kteří nemají dítě).

Nakonec je dobré překontrolovat samotné proměnné a rozhodnout, jestli jsou nominální nebo ordinální.

Krok 1 – Existuje vztah?

Nejdříve je nutné otestovat, jestli je mezi našimi proměnnými nějaký vztah. K tomu využijeme Pearsonův X^2 testu nezávislosti. Restu samotného se skládá ze dvou částí: ověřování předpokladů a výpočtu testu samotného.

Pro správný výpočet X^2 je nutné, aby byla kontingenční tabulka dostatečně zaplněná, tedy aby nemělo více než 20 % buněk očekávanou hodnotu menší než 5 a zároveň aby v žádná buňka neměla očekávanou hodnotu menší než 1. Pokud tabulka není dostatečně zaplněná, je možné sloučit řádky/sloupce, s příliš malými očekávanými hodnotami (např. můžeme sloučit kategorie „velmi spokojen“ a „spíše spokojen“), sloučené kategorie ale musí dávat smysl (nelze tedy sloučit „spíše spokojen“ a „spíše nespokojen“). Druhou možností je určitý řádek/sloupec vyřadit úplně. Často se např. stává, že odpověď „nevím“ má velmi malou očekávanou četnost a je proto možné ji vyřadit. I tady je ale nutné, aby vyřazení dávalo logicky smysl (nejde např. vyřazovat kategorie, které jsou součástí škály).

Pokud je tabulka dostatečně zaplněná, je možné spočítat X^2 test samotný. Pro výpočet testu je nutné formulovat nulovou hypotézu (negace původní hypotézy, kterou zkoumáme). Interpretace výsledků funguje klasicky, jako u většiny testů, zajímá nás tedy hlavně p hodnota, která říká, jestli máme důkaz pro zamítnutí nulové hypotézy. Pokud je p hodnota větší než kritická hodnota (většinou 0,05), tak nemáme důkaz, že by vztah mezi našimi proměnnými existoval a nemůžeme zamítnout nulovou hypotézu. Pokud je p hodnota menší, máme důkaz o existenci vztahu mezi proměnnými a nulovou hypotézu zamítnout můžeme.

Na základě X^2 testu jsme tedy zjistili, jestli existuje vztah mezi našimi proměnnými nebo ne. Pokud pro existenci vztahu nemáme důkaz, tak analýza kontingenčních tabulek právě skončila. Pokud máme důkaz pro zamítnutí nulové hypotézy (p hodnota je menší než 0,05), tak můžeme dál pokračovat v analýze vztahu mezi našimi proměnnými.

Krok 2 – Jak silný je vztah?

Jakmile víme, že existuje vztah mezi našimi proměnnými, je nutné zjistit, jak je vztah silný, a případně kterým jaká je jeho povaha. K tomu se využívají ukazatel obecně zvané (symetrické) koeficienty kontingence. To, které koeficienty použijeme, závisí na tom, zda jsou naše proměnný nominální nebo ordinální. Všechny koeficienty mají svou hodnotu statistické signifikance, která se interpretuje stejně jako u X^2 testu.

<u>Koeficienty pro nominální proměnné</u>

V případě, že je alespoň jedna z našich proměnných nominální, používáme tyto koeficienty, jejichž hodnoty se pohybují v rozmezí <0;1>[1] (0 = žádný vztah; 1 = dokonalý vztah):

[1] Technicky ne všechny koeficienty nabývají hodnot <0:1>. Fí koeficient může podle některých definic v určitých situacích nabývat hodnot <-1;1>. Záleží i na požitém softwaru. (např. SAS má tendenci u tabulek 2x2 vracet záporné hodnoty Cramerova V).

Φ (fí) – nejednoduší na výpočet, jedná se o X^2 statistiku vydělenou počtem pozorování. Vhodná pro tabulky 2x2, u větších nepřesný.

Koeficient kontingence (C) – Lepší verze fí koeficientu, je možné použít i na tabulky větší než 2x2. Problémem je, že maximální hodnota, kterou může koeficient nabýt, není vždy 1, ale závisí na velikosti tabulky (tzn. počtu řádků/sloupců). Platí, že čím menší tabulka, tím menší je maximální možná hodnota koeficientu. U menších tabulek je tedy třeba dát si pozor při interpretaci, protože koeficient opticky podhodnocuje těsnost/sílu vztahu mezi proměnnými. Koeficient je možné standardizovat tak, aby měl maximální hodnotu vždy rovnou 1, ale není to standardem (např. SPSS verze 24 koeficient nestandardizuje automaticky, je třeba přepočítávat ručně).

Cramerovo V – V podstatě se jedná o korekci fí koeficientu pro větší tabulky. Je možné využít pro tabulky jakékoliv velikosti a maximální hodnota je vždy 1. Pravděpodobně nejpoužívanější ze zde popsaných koeficientů.

Intepretace probíhá tak, že si spočítáte koeficienty a buď a) vezmete ten, který považujete za nejpřesnější (Cramerovo V, případně normalizovaný koeficient kontingence) a orientujete se podle něj nebo b) zprůměrujete ty nejpřesnější (např. Cramerovo V a koeficient kontingence) a orientujete se podle průměru. Interpretace je jednoduše čím více, tím lépe, tzn.:

Hodnota 0 – žádný vztah

Hodnoty kolem 0,2 slabý vztah

Hodnoty kolem 0,5 - střední vztah

Hodnoty kolem 0,7 a výše – silný vztah

Hranice jsou ale pouze orientační a záleží na kontextu (povaze dat).

<u>Koeficienty pro ordinální proměnné</u>

Pokud jsou obě naše proměnné ordinální, používáme koeficienty, které nám kromě síly vztahu dokáží říct také směr vztahu (jedná se vlastně korelační koeficienty pro ordinální data). Ordinální koeficienty se pohybují v intervalu <-1;1> a interpretují se stejně jako klasické korelační koeficienty. Nejznámější jsou:

Gamma – základní míra založená na výpočtu konkordantních a diskordantních párů. Nevýhodou je, že ignoruje remízy (páry, které nejsou ani konkordantní ani diskordantní), takže její míra přesnosti klesá s rostoucí velikostí tabulky (tzn. se zvětšujícím se množstvím řádků a sloupců).

Kendallovo tau-b – korekce gamma, zohledňující remízy. Používá se na čtvercové tabulky (např. 4x4, 6x6,...)

Kendallovo tau-c – Jako Kendallovo tau-b, ale na obdélníkové tabulky.

Interpretace je stejná jako u korelačních koeficientů, tzn. čím dále od 0, tím silnější vztah. Pozitivní koeficient značí, že s rostoucí mírou jedné proměnné roste i druhá (např. s rostoucí úrovní vzdělání roste zájem o politiku). Negativní koeficient značí, že s rostoucí mírou jedné proměnné druhá klesá (např. s rostoucím věkem klesá spokojenost se životem).

Teď tedy víme nejen, že vztah mezi proměnnými existuje, ale i jak je silný. Pearsonův X^2 test a koeficienty spolu často sedí, takže pokud vyjde test nesignifikantní, hodnoty koeficientů budou kolem 0.

Krok 3 – Kde přesně vztah je?

Poslední částí analýzy je zjisti, kde přesně se vztah proměnných projevuje. Je totiž možné, že se vztah neobjevuje napříč celou tabulkou, ale v pouze některé její části (např. spokojenost se životem souvisí se vzděláním ale pouze od základního vzdělání po střední vzdělání s maturitou, u vysokoškoláků jejich vzdělání na spokojenost mít vliv nemusí). Přesnou podobu vztahu zjistíme pomocí Pearsonových adjustovaných standardizovaných reziduí. Podíváme se na jednotlivé buňky a vyznačíme všechny rezidua, jejichž hodnota je vyšší než 1,96 nebo menší než -1,96. Pokud je hodnota rezidua vyšší než 1,96, znamená to, že v dané buňce je statisticky signifikantně více pozorování, než jich bylo očekáváno (pokud by mezi proměnnými vztah nebyl). Naopak, pokud je hodnota rezidua nižší než -1,96, znamená to, že v dané buňce je statisticky významně méně pozorování, než bylo očekáváno. Cílem je nalézt v rozmístění signifikantních reziduí interpretovatelnou strukturu.

Příklad

Popis dat

Proměnné v tabulce:	Vzdělání (4 kategorie)
	Politická orientace (6. kategorií)
Celkový počet pozorování:	1039
Počet platných pozorování:	1025
Počet chybějících pozorování:	14

Pearsonův X^2 test

X^2	*DF*	*p hodnota*
129,096	15	<0,001

0 buněk (0,0%) má očekávanou četnost menší než 5. Minimální očekávaná četnost je 15,08.

		Levice	Levý střed	Střed	Pravý střed	Pravice	Neví
(Neúplné) základní	*četnost*	27	31	44	22	16	38
	řádková %	15%	17%	25%	12%	9%	21%
	adj. stand. rezidua	1	0,2	0,8	-2,8	-3,5	6
Střední bez maturity a vyučení	*četnost*	62	70	85	60	50	31
	řádková %	17%	20%	24%	17%	14%	9%
	adj. stand. rezidua	3,1	1,6	0,7	-1,8	-2,7	-0,6
Střední s maturitou	*četnost*	35	52	78	80	58	25
	řádková %	11%	16%	24%	24%	18%	8%
	adj. stand. rezidua	-1,4	-0,7	0,7	2,5	-0,4	-1,3
VOŠ a Vysokoškolské	*četnost*	8	21	24	42	64	2
	řádková %	5%	13%	15%	26%	40%	1%
	adj. stand. rezidua	-3,3	-1,4	-2,5	2,1	7,6	-3,9

Koeficienty kontingence/asociace

	koeficient	*p hodnota*
Fí	0,355	<0,001
Cramerovo V	0,205	<0,001
Koeficient kon. (C)	0,334	<0,001
Kendallovo tau-b	0,095	<0,001
Kendallovo tau-c	0,098	<0,001
Gamma	0,123	<0,001

Interpretace

Pomocí tabulky budeme testovat hypotézu, že politická orientace respondentů závisí na jejich dosaženém vzdělání. Obě proměnné jsou teoreticky ordinální, nicméně proměnná politická orientace obsahuje kategorii „nevím“ a je tedy v analýze považovaná za nominální.

Existenci vztahu ověříme Pearsonovým X^2 testem. Žádná buňka nemá očekávanou hodnotu menší než pět a nejmenší očekávaná je 15,08, předpoklady testu jsou tedy splněné a můžeme pokračovat s analýzou. Nulová hypotéza bude znít „Neexistuje vztah mezi dosaženým vzděláním a politickou orientací“. P hodnota testu je menší než 0,05 (dokonce je menší než 0,001), máme tedy důkaz pro zamítnutí nulové hypotézy a můžeme říct, že existuje vztah mezi vzděláním a politickou orientací.

Co se týče koeficientů, je nutné použít koeficienty pro nominální proměnné. To proto, že v proměnné politická orientace se nachází odpověď „nevím“. Pokud bychom chtěli použít ordinální koeficienty, museli bychom odpověď „nevím“ vyřadit, nicméně za ceno možné ztráty informací. Jak Cramerovo V, tak koeficient kontingence jsou statisticky signifikantní a dohromady ukazují slabý, až středně silný vztah.

Rezidua ukazují, že lidé se základním vzděláním mají méně často pravicovou nebo pravostředovou politickou orientaci. Naopak častěji odpovídají, že svou politickou orientaci neznají (nebo nejsou schopní se zařadit). U střední školy bez maturity a vyučených je výrazně více zastoupena kategorie „levice“ a naopak výrazně méně zastoupená kategorie „pravice“. U respondentů se středoškolským vzděláním s maturitou je výrazně více zastoupena kategorie „pravý střed“. U vysokoškoláků se výrazně častěji vyskytují kategorie „pravý střed“ a „pravice“, naopak výrazně méně zastoupeny jsou kategorie „levice“, „levý střed“ a také kategorie „nevím“.

Analýza přinesla dva poznatky. Zaprvé, se vzrůstajícím vzděláním se lidé posouvají od levice k pravici. Za druhé, samotná znalost výrazů pravice a levice je spojená s úrovní dosaženého vzdělání. Lidé se základním vzděláním volí výrazně často kategorii „nevím“ (zde ji zvolilo 21%). Možným vysvětlením je, že s termíny nejsou obeznámeni natolik, aby se podle nich dokázali zařadit. Naopak vysokoškoláci voli kategorii „nevím“ výrazně méně (pouze 1%), pravděpodobně proto, že jejich chápání politické teorie je dostatečně rozvinuté, aby pro ně pojmy pravice a levice měli konkrétní význam.