Analýza dat v neurologii

LII. Odhad poměru šancí u složitějších tabulek četností

L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková

Institut biostatistiky a analýz
Masarykova univerzita, Brno

✉
doc. RNDr. Ladislav Dušek, Ph.D.
Institut biostatistiky a analýz
MU, Brno
e-mail: dusek@iba.muni.cz

V tomto díle seriálu reagujeme na podnět čtenáře, který po prostudování předchozích dílů 42–49 vznesl tento dotaz: **Většina učebnicových příkladů pracuje jen s tabulkou četností 2 × 2. Jak ale postupovat při asociační analýze, kde je zdrojová tabulka s více řádky a sloupci?**

Je logické, že reálné situace vedoucí ke studiu vztahu „expozice-účinek" zdaleka nemusí vycházet pouze z nejjednodušší tabulky četností. Složitější tabulky označujeme obecně *R* × *C*, což ukazuje, že se skládají z více než dvou řádků a dvou sloupců (*R*: „rows", *C*: „columns"). Taková tabulka dává do vzájemné souvislosti dva znaky, které nabývají hodnot v rámci více než dvou kategorií. Ačkoli jsou pro tyto složitější situace dostupné komplexnější techniky modelování (např. logistická regrese, log-lineární modely, asociační modely), lze je zvládnout i relativně jednodušími a standardními postupy. Některým z nich jsme se věnovali již v díle 22 našeho seriálu. Výhodou postupů, které zde budeme dokumentovat na konkrétních příkladech, je fakt, že experimentátor kontroluje logiku postupu a určuje, jaké kategorie bude dávat do vzájemného vztahu a jeho platnost testovat.

I zde můžeme testovat jednak platnost obecné hypotézy nezávislosti obou znaků (např. klasickým chí-kvadrát testem) a následně využít výpočet poměru šancí (*OR*) pro kvantifikaci míry vztahu „expozice-účinek". Problém ale je, že ve složitější tabulce se nabízí více kategorií a tedy více odhadů *OR*.

Zjišťujeme, zda existuje vztah mezi krevními skupinami pacientů (A, B, 0) a výskytem určitých diagnóz (dg. I, II, III). Prvním krokem analýzy je výpočet řádkových a sloupcových procent v tabulce a stanovení hypotéz o možných dílčích tabulkách, u kterých by bylo možné provést sloučení řádků nebo sloupců tabulky. Dílčí hypotézy vztahu mezi řádky a sloupci dílčích tabulek jsou testovány pomocí testu dobré shody.

1) Vstupní tabulka 3 × 3, dle výsledků testu dobré shody je mezi řádky a sloupci tabulky statisticky významný vztah s $p < 0{,}001$.

Krevní skupina	Diagnóza		
	I	II	III
A	28	20	41
B	45	30	59
0	240	25	51
Řádková %	**I**	**II**	**III**
A	31,5 %	22,5 %	46,1 %
B	33,6 %	22,4 %	44,0 %
0	75,9 %	7,9 %	16,1 %
Sloupcová %	**I**	**II**	**III**
A	8,9 %	26,7 %	27,2 %
B	14,4 %	40,0 %	39,1 %
0	76,7 %	33,3 %	33,8 %

2) Vyhodnocením procentuálního zastoupení kategorií se nabízí možnost testovat dílčí tabulky:
- krevní skupina A vs. B,
- diagnóza II vs. III.

Po otestování pomocí testu dobré shody nemůžeme zamítnout nulovou hypotézu o náhodném vztahu řádků a sloupců ani v jedné z dílčích tabulek ($p = 0{,}940$, resp. $p = 0{,}991$); vzhledem k vyššímu p u dílčí tabulky „diagnóza II vs. III" budeme v dalším kroku analýzy počítat se sloučenou skupinou II + III.

3) Vyhodnocením procentuálního zastoupení kategorií ve sloučené tabulce se nabízí možnost testovat dílčí tabulku krevní skupina A vs. B.

Krevní skupina	Diagnóza	
	I	II + III
A	28	61
B	45	89
0	240	76
Řádková %	**I**	**II + III**
A	31,5 %	68,5 %
B	33,6 %	66,4 %
0	75,9 %	24,1 %
Sloupcová %	**I**	**II + III**
A	8,9 %	27,0 %
B	14,4 %	39,4 %
0	76,7 %	33,6 %

Po otestování pomocí testu dobré shody nemůžeme zamítnout nulovou hypotézu o náhodném vztahu řádků a sloupců v dílčí tabulce ($p = 0{,}741$); v dalším kroku analýzy budeme počítat se sloučenou skupinou A + B.

4) Výsledná tabulka 2 × 2.

Krevní skupina	Diagnóza	
	I	II + III
A + B	73	150
0	240	76
Řádková %	**I**	**II + III**
A + B	32,7 %	67,3 %
0	75,9 %	24,1 %
Sloupcová %	**I**	**II + III**
A + B	23,3 %	66,4 %
0	76,7 %	33,6 %

Dle výsledku testu dobré shody je mezi řádky a sloupci tabulky statisticky významný vztah s $p < 0{,}001$. Krevní skupina 0 zvyšuje riziko výskytu diagnózy I s poměrem šancí *OR* (95% IS): 6,489 (4,435; 9,493).

Příklad 1. Zjednodušení *R* × *C* tabulky četností s pomocí dílčích testů dobré shody.

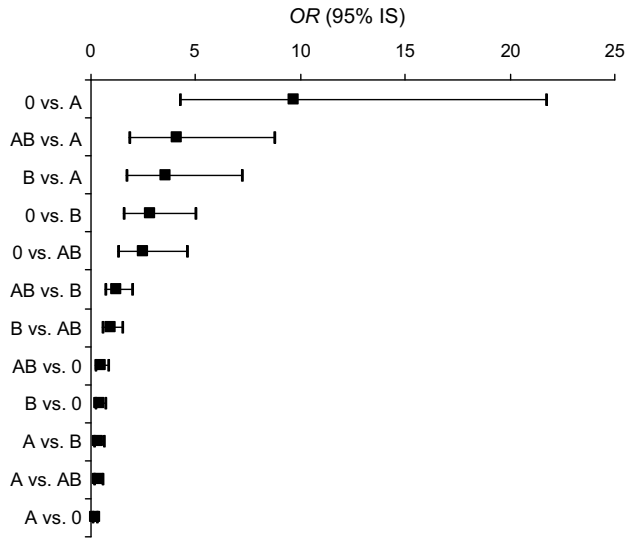
Zjišťujeme, zda existuje vztah mezi krevními skupinami pacientů (A, B, AB, 0) a výskytem určité diagnózy (dg. ne/ano). Tabulku rozdělíme na všechny možné dílčí tabulky 2 × 2 a vypočteme pro každou z nich poměr šancí (*OR*) pro výskyt diagnózy v závislosti na krevní skupině. Dílčí tabulky seřadíme podle hodnoty *OR* a identifikujeme tabulky s nejvýznamnějšími vztahy mezi krevními skupinami a výskytem hodnocené diagnózy.

Diagnóza	Krevní skupina			
	A	**B**	**AB**	**0**
ne	80	120	66	30
ano	10	52	33	36
Sloupcová %	**A**	**B**	**AB**	**0**
ne	88,9 %	69,8 %	66,7 %	45,5 %
ano	11,1 %	30,2 %	33,3 %	54,5 %

Z tabulky je možné vytvořit šest tabulek 2 × 2. Pro každou z nich je spočten poměr šancí (*OR*) pro vliv jedné krevní skupiny oproti druhé na výskyt sledované diagnózy. Pro každou 2 × 2 tabulku jsou spočítány dva odhady *OR*, v pozici rizikové kategorie se vystřídají postupně obě krevní skupiny v tabulce, celkem je tedy spočítáno 12 *OR*.

Testovaná kombinace	OR (95% IS)
0 vs. A	9,600 (4,242; 21,724)*
AB vs. A	4,000 (1,836; 8,717)*
B vs. A	3,467 (1,665; 7,219)*
0 vs. B	2,769 (1,545; 4,964)*
0 vs. AB	2,400 (1,266; 4,551)*
AB vs. B	1,154 (0,679; 1,960)
B vs. AB	0,867 (0,510; 1,472)
AB vs. 0	0,417 (0,220; 0,790)*
B vs. 0	0,361 (0,201; 0,647)*
A vs. B	0,288 (0,139; 0,601)*
A vs. AB	0,250 (0,115; 0,545)*
A vs. 0	0,104 (0,046; 0,236)*

* Statisticky významná hodnota OR.



Příklad 2. Zjednodušení *R* × *C* tabulky četností pomocí testování všech možných dílčích tabulek („*cross-classification*").

Bude tudíž záležet na nosných hypotézách experimentu a prioritách, které si při hodnocení stanovíme. V několika bodech stručně shrneme možné přístupy k problému, přičemž každý z nich je doplněn číselným příkladem.

Je možné postupovat následovně:

1. Pokusíme se tabulku zjednodušit, například postupným prováděním dílčích testů dobré shody (chí-kvadrát test). Dílčí testy povedou ke slučování řádků a sloupců, mezi kterými nebude prokázána závislost (četnosti takových kategorií lze sečíst, aniž přijdeme o významnou informaci). Po zjednodušení na výsledné tabulce vyhodnotíme vztah „expozice-účinek" a kvantifikujeme jej pomocí poměru šancí. Tento postup dokumentuje příklad 1.
2. Rozdělíme tabulku *R* × *C* na dílčí tabulky četností 2 × 2 a ty separátně vyhodnotíme a odhadneme z nich plynoucí dílčí poměry šancí. Tomuto postupu se v mezinárodní literatuře někdy říká „*cross-classification*" a v podstatě vede k vzájemnému testování vztahu všech jednotlivých kategorií, každé s každou. Výpočet je dokumentován v příkladu 2.
3. V posledním postupu nebudeme testovat vztah všech kategorií obou znaků, ale zvolíme adekvátní podmnožinu dílčích tabulek a vztahů a ty otestujeme. Standardně tento přístup vede ke zvolení jedné kategorie jako referenční a k ní jsou potom vztahovány odhady poměru šancí všech dalších kategorií. V epidemiologické literatuře se setkáváme s pojmem „*case-referent study*", který je používán místo standardního „*case-control study*". Volba referenční kategorie má tu výhodu, že všechny dílčí odhady *OR* jsou vzájemně srovnatelné, neboť se vztahují ke stejnému referenčnímu základu. Postup výpočtu dokumentují příklady 3a a 3b.

Je jistě patrné, že výše shrnuté postupy analýz jsou velmi rozdílné a vyžadují různý stupeň znalosti řešeného problému a experimentální situace. Rozdíly lze dokumentovat následovně:

1. Postup 1 v podstatě usiluje o snížení počtu kategorií v analýze, ideálně až na úroveň nejjednodušší možné tabulky 2 × 2, jak dokumentuje příklad 1. Pokud získaná experimentální data toto umožňují, jde jistě o efektivní řešení problému, které nadto doloží, že pro daný problém nemá dělení sledovaných znaků na více kategorií smysl.
2. Postup 2 je z navržených metod asi nejméně intuitivní a v podstatě analyzuje všechny možné dílčí kombinace tabulek, což může být vhodné, pokud nejsme schopni prioritizovat prováděná srovnání. Jde o jistou formu pilotní exploratorní analýzy, ze které teprve vzejdou podložené otázky a hypotézy.
3. Postup 3 je v odborné literatuře velice hojně využíván a lze jej označit za jistý druh standardu. Volba jedné z kategorií znaku za referenční (tzn. tato kategorie je označena jako reference ve všech testovaných dílčích tabulkách) umožní vztáh-

Příklad 3a) Zjišťujeme, zda existuje vztah mezi léčebnou odpovědí (vynikající, dobrá, stabilizace, progrese) a výskytem dlouhodobých komplikací (ne/ano) po ukončení léčby. V analýze je třeba zohlednit ordinální charakter léčebné odpovědi, proto je třeba provádět srovnání v 2 × 2 tabulkách jen vůči referenční kategorii (namísto všech kombinací 2 × 2 tabulek). Jako referenční jsme zvolili kategorii „vynikající".

Komplikace	Léčebná odpověď			
	vynikající	**dobrá**	**stabilizace**	**progrese**
ne	76	320	58	29
ano	10	85	23	16
Sloupcová %	**vynikající**	**dobrá**	**stabilizace**	**progrese**
ne	88,4 %	79,0 %	71,6 %	64,4 %
ano	11,6 %	21,0 %	28,4 %	35,6 %

Testovaná kombinace	*OR* (95% IS)
dobrá vs. vynikající	2,019 (1,001; 4,071)*
stabilizace vs. vynikající	3,014 (1,331; 6,824)*
progrese vs. vynikající	4,193 (1,707; 10,298)*

*Statisticky významné *OR*.

S horší léčebnou odpovědí stoupá statisticky významně šance na výskyt dlouhodobých komplikací.

Příklad 3b) Zjišťujeme, zda existuje vztah mezi hladinou určitého proteinu (I < II < III < IV) a výskytem komplikací (ne/ano) po ukončení léčby. V analýze je třeba zohlednit ordinální charakter kategorií hladiny proteinu, proto je třeba provádět srovnání v 2 × 2 tabulkách jen vůči referenční kategorii (namísto všech kombinací 2 × 2 tabulek). Jako referenční jsme zvolili kategorii „I".

Komplikace	Hladina proteinu			
	I	**II**	**III**	**IV**
ne	240	260	320	226
ano	48	50	60	25
Sloupcová %	**I**	**II**	**III**	**IV**
ne	83,3 %	83,9 %	84,2 %	90,0 %
ano	16,7 %	16,1 %	15,8 %	10,0 %

Testovaná kombinace	*OR* (95% IS)
II vs. I	0,962 (0,624; 1,483)
III vs. I	0,938 (0,619; 1,419)
IV vs. I	0,553 (0,330; 0,927)*

*Statisticky významné OR.

Nejvyšší hladina proteinu působí v kontrastu s nejnižší hladinou jako statisticky významný ochranný faktor pro výskyt komplikací.

Příklad 3. Zjednodušení $R \times C$ tabulky četností pomocí testování jednotlivých kategorií proti referenční kategorii.

nout výsledky dílčích analýz ke stejnému základu. Výsledek je velmi dobře čitelný a srozumitelný. Nadto znalý experimentátor umí dobře zvolit referenční kategorii tak, aby dávala i věcný klinický či biologický smysl.

Snad jsme zvolenými příklady dostatečně odpověděli na otázku našeho čtenáře. Na závěr ještě uvádíme zmínku o zvláštní formě tabulek $R \times C$, u kterých kategorie jednoho i obou asociovaných znaků nejsou nominálními položkami, ale vytvářejí ordinální škálu. V takovém systému jde vedle vlastní asociace znaků testovat i její trendovou složku, která může být informačně velmi důležitá. Tomuto problému se budeme věnovat v některém z blízkých dílů našeho seriálu.