proLékaře.cz | 3.5.2024



# Analýza dat v neurologii

## XXII. Rozbor složitějších kontingenčních tabulek je účinným nástrojem pro studium vztahů kategoriálních znaků

L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková
Institut biostatistiky a analýz
Masarykova univerzita, Brno

✉
doc. RNDr. Ladislav Dušek, Dr.
Institut biostatistiky a analýz
Masarykova univerzita, Brno
e-mail: dusek@cba.muni.cz

Minulý díl seriálu jsme věnovali analýze kontingenčních tabulek a hodnocení nezávislosti dvou kategoriálních znaků. Připomeňme si, že jde o tabulky četností, kde proti sobě stavíme výskyt kategorií dvou náhodně sledovaných znaků; v nejjednodušším případě jde o tabulky typu 2 × 2. Standardním testem pro hodnocení hypotézy vzájemně nezávislého výskytu znaků je test využívající statistiku $\chi^2$. Ačkoli díl XXI vyčerpal základy této problematiky, nechceme dané téma ještě opustit. Připravili bychom čtenáře o další zajímavé přístupy, které při aplikaci na klinická data přinášejí cenné interpretace.

Tento díl věnujeme hodnocení nezávislosti dvou znaků u složitějších tabulek typu $R \times C$. Tyto kontingenční tabulky jsou tvořeny více kategoriemi dvou znaků, a mají tedy více než dva řádky ($R$) nebo sloupce ($C$) anebo obojí. Hodnocení nezávislosti znamená, že zkoumáme společný výskyt dvou náhodných znaků, ke kterým přistupujeme zcela rovnocenně. Oba kategoriální znaky sledujeme u každého prvku (subjektu, pacienta…) náhodného výběru o velikosti $n$. Příkladem může být klinické stadium nemoci v kategoriích I–IV a výskyt potíží určitého typu (např. A, B, C). Náhodný výběr $n$ pacientů bychom zde přepsali do kontingenční tabulky o rozměru 4 × 3. Pro posouzení hypotézy o nezávislém výskytu stadia a potíží použijeme test $\chi^2$ (viz díl XXI), u kterého lze teoreticky očekávat tři možnosti výsledku:

**1.** Nezamítneme nulovou hypotézu a prokážeme, že oba znaky se u pacientů vyskytují vzájemně nezávisle. To znamená, že vypočtená hodnota testové statistiky $\chi^2$ je pro tabulku s primárními četnostmi numericky malá a nepřekračuje kritickou mez pro danou hladinu chyby I. druhu ($\alpha$) a pro daný počet stupňů volnosti: $(R - 1) \times (C - 1) = 6$. Tímto naše testování nutně končí, závěr je konečný.

**2.** Zamítneme nulovou hypotézu, což znamená, že prokazujeme existující vztah (vazbu) mezi kategoriemi obou znaků. Nabízí se ale otázka, zda by dílčí rozbor poměrně velké tabulky 4 × 3 neupřesnil, mezi kterými kategoriemi tato vazba statisticky průkazně existuje a mezi kterými ne. Tedy zda je možné sloučit některé dílčí kategorie v rámci řádků (stadium) nebo sloupců (typ potíží) na základě dílčích testů nezávislosti. Pokud například dokážeme z původní tabulky separovat určitou část (např. 4 kategorie řádků × 2 kategorie sloupců), mezi kterými potvrdíme nezávislost výskytu, můžeme tyto dvě kategorie sloupců sloučit. To znamená sečíst jejich pozorované četnosti a původní tabulku tak zjednodušit. Logicky tak přesněji určujeme prokázaný vztah a posilujeme klinickou interpretaci výsledku.

**3.** Zamítneme nulovou hypotézu, a tedy prokážeme, že existuje vztah ve výskytu klinického stadia a typu potíží. Dalším detailním rozborem původní tabulky ale nenajdeme žádné kategorie řádků a sloupců, které by bylo možné sloučit na základě dílčích testů hypotézy nezávislosti. Původní tabulku nemůžeme nijak zjednodušit a v podstatě konstatujeme velmi silnou závislost kategorií obou znaků, jejichž vzájemná vazba ovlivňuje četnost v každém poli původní tabulky.

Z hlediska analýz a interpretací je nejzajímavější varianta číslo 2. Detailním průzkumem původní tabulky můžeme „objevit" kategorie znaku, které znamenají určitý zlom ve vztahu ke znaku druhému. Můžeme tak značně zjednodušit i poměrně objemnou tabulku s mnoha řádky a sloupci a získat prakticky využitelnou a statisticky potvrzenou informaci. Při detailním průzkumu velkých tabulek doporučujeme provést přepočet jednotlivých polí na dílčí procenta v rámci sloupců nebo řádků, v tabulce se pak lépe orientujeme a naše kroky při analýze jsou cílenější. Popis takové analýzy je uveden v příkladu 1, kde sloučením kategorií sloupců došlo ke zjednodušení původní tabulky 2 × 4 na konečnou 2 × 2. Kategorie sloupců zde vyjadřovaly gradient spotřeby kofeinu a analýza prokázala závislost spotřeby kofeinu na rodinném stavu člověka. Detailní rozbor dále odhalil, že nemá význam rozlišovat čtyři kategorie denní dávky kofeinu a daný vztah k rodinnému stavu určují dávky nižší nebo vyšší než 150 mg/den. Dvě kategorie z oblasti nízkého příjmu kofeinu a dvě kategorie v oblasti vysokého příjmu kofeinu byly sloučeny na základě dílčích statistických testů. Analýza tak identifikuje jistý bod zlomu („cut-off") ve spotřebě kofeinu pro další hodnocení nebo uvažování.

Příklad 2 je další ukázkou rozboru složitější kontingenční tabulky, tentokrát pro dva nominální znaky (zkoumá vztah výskytu krevních skupin a různých onemocnění). Opět zde dílčím slučováním vzájemně nezávislých kategorií dospíváme k pojmenování skutečného vztahu (statisticky významně vyšší výskyt krevní skupiny 0 u pacientů s žaludečními vředy), který byl v původní tabulce v podstatě maskován velkým množstvím vstupních kategorií. Příklad dále dokládá možnou kontrolu výpočtu, kterou u průzkumu složité kontingenční tabulky máme. Všechny dílčí tabulky, které vedly ke sloučení kategorií řádků nebo sloupců, plus konečná

proLékaře.cz | 3.5.2024



1. Zkoumáme vztah mezi rodinným stavem a spotřebou kofeinu u *n* náhodně vybraných jedinců
   Data uspořádáme do 2 × 4 kontingenční tabulky:

| | Spotřeba kofeinu (mg/den) | | | | |
|---|---|---|---|---|---|
| Rodinný stav | 0 | 1–150 | 151–300 | > 300 | Celkem |
| vdaná/ženatý | 802 | 1 537 | 38 | 21 | 2 398 |
| rozvedená/ý, žijící odděleně, ovdovělá/ý | 36 | 46 | 598 | 242 | 922 |
| Celkem | 838 | 1 583 | 636 | 263 | 3 320 |

2. Pomocí testu dobré shody testujeme hypotézu nezávislosti výskytu obou jevů ($H_0$: výskyt kategorií rodinného stavu a spotřeby kofeinu jsou vzájemně nezávislé)

3. Test dobré shody je dán vztahem

$$\chi^2_{(r-1)*(c-1)} = \sum_{x=1}^{r}\sum_{y=1}^{c}\frac{(f_{x,y}-F_{x,y})^2}{F_{x,y}}$$

   kde $f_{x,y}$ jsou pozorované četnosti v buňkách tabulky, $F_{x,y}$ jsou očekávané četnosti, *r* je počet řádků, *c* je počet sloupců.

4. Vypočítáme očekávané četnosti, např.:

$$F_{1,1} = \frac{2398*838}{3320} = 605{,}3$$

5. Po dosazení získáváme hodnotu testové statistiky:

$$F_{x,y} = \frac{\sum_{z=1}^{c} f_{x,z} * \sum_{z=1}^{r} f_{z,y}}{N}$$

$$\chi^2_3 = \frac{(802-605{,}3)^2}{605{,}3} + \ldots + \frac{(242-73)^2}{73} = 2651{,}12$$

6. Hodnotu testové statistiky porovnáme s kritickou hodnotou pro α = 0,05 a stupně volnosti ν = 3:

$$\chi^2_{(1-\alpha)}{}^{(\nu=3)} = 7{,}82.$$

7. Výsledek: získaná hodnota testové statistiky významně převyšuje kritickou hodnotu:
   $\chi^2_{(3)} > \chi^2_{(1-\alpha)}{}^{(\nu=3)}$, a na hladině významnosti 0,05 tedy zamítáme $H_0$ ($p < 0{,}05$).

8. Datovou tabulku lze vyjádřit pomocí procentuálních zastoupení kategorií v rámci sloupců následovně:

| | Spotřeba kofeinu (mg/den) | | | |
|---|---|---|---|---|
| Rodinný stav | 0 | 1–150 | 151–300 | > 300 |
| vdaná/ženatý | 95,7 % | 97,1 % | 6,0 % | 8,0 % |
| rozvedená/ý, žijící odděleně, ovdovělá/ý | 4,3 % | 2,9 % | 94,0 % | 92,0 % |

9. Z procentuálního zastoupení je patrné, že složitou kontingenční tabulku 2 × 4 můžeme smysluplně rozdělit na dvě menší (obě 2 × 2): první vznikne z 1. a 2. sloupce (tabulka A), druhá z 3. a 4. sloupce (tabulka B).

10. Pro obě dílčí kontingenční tabulky znovu samostatně otestujeme hypotézu nezávislosti výskytu kategorií:

    **A:** $\chi^2_{(1)} = 3{,}24 < \chi^2_{(0{,}95)}{}^{(\nu=1)} = 3{,}84$
    **–> nezamítáme $H_0$ ($p > 0{,}05$)**

    **B:** $\chi^2_{(1)} = 1{,}23 < \chi^2_{(0{,}95)}{}^{(\nu=1)} = 3{,}84$
    **–> nezamítáme $H_0$ ($p > 0{,}05$)**

    Podle provedených testů není v rámci obou tabulek statisticky významný rozdíl ve výskytu kategorií znaků.

11. Na základě výsledku v bodě 10 můžeme sloučit pozorované četnosti v rámci dílčích tabulek a získáme tak výslednou kontingenční tabulku 2 × 2:

| | Spotřeba kofeinu (mg/den) | | |
|---|---|---|---|
| Rodinný stav | 0–150 | ≥ 151 | Celkem |
| vdaná/ženatý | 2 339 | 59 | 2 398 |
| rozvedená/ý, žijící odděleně, ovdovělá/ý | 82 | 840 | 922 |
| Celkem | 2 421 | 899 | 3 320 |

12. Pro sloučenou kontingenční tabulku opět vypočteme hodnotu testové statistiky pro test dobré shody a porovnáme s kritickou hodnotou:

    $\chi^2_{(1)} = 2650{,}21 > \chi^2_{(0{,}95)}{}^{(\nu=1)} = 3{,}84$
    **–> zamítáme $H_0$ ($p < 0{,}05$)**

**Závěr:** Zamítáme nulovou hypotézu, že rodinný stav a spotřeba kofeinu jsou vzájemně nezávislé. Detailní rozbor původní kontingenční tabulky umožnil nalézt kategorie znaků, které jsou za prokázaný vztah odpovědné. Ostatní kategorie mohly být sloučeny. Vidíme tedy, že lidé žijící v manželství spotřebují méně kofeinu než lidé žijící samostatně; hranicí oddělující obě skupiny je spotřeba 150 mg kofeinu/den.

Příklad 1. Analýza složitější kontingenční tabulky pro nominální a ordinální znak.

proLékaře.cz | 3.5.2024



1. Zkoumáme vztah mezi výskytem onemocnění žaludku a krevní skupinou pacientů

| Krevní skupina | Žaludeční vředy | Rakovina žaludku | Kontrola | Celkem |
|---|---|---|---|---|
| 0 | 983 | 383 | 2 892 | 4 258 |
| A | 679 | 416 | 2 625 | 3 720 |
| B | 134 | 84 | 570 | 788 |
| Celkem | **1 796** | **883** | **6 087** | **8 766** |

2. Pomocí testu dobré shody testujeme nezávislost obou jevů ($H_0$: onemocnění žaludku a krevní skupina pacientů jsou nezávisle se vyskytující veličiny)

3. Test dobré shody je dán vztahem

$$\chi^2_{(r-1)*(c-1)} = \sum_{x=1}^{r}\sum_{y=1}^{c}\frac{(f_{x,y} - F_{x,y})^2}{F_{x,y}}$$

kde $f_{x,y}$ jsou pozorované četnosti v buňkách tabulky, $F_{x,y}$ jsou očekávané četnosti, $r$ je počet řádků, $c$ je počet sloupců.

4. Očekávané četnosti vypočítáme dle vztahu

$$F_{x,y} = \frac{\sum_{z=1}^{c} f_{x,z} * \sum_{z=1}^{r} f_{z,y}}{N}$$

5. Následně získáme hodnotu testové statistiky:

$$\chi^2_4 = \frac{(983-872)^2}{872} + \ldots + \frac{(570-547)^2}{547} = 40{,}61$$

6. Hodnotu testové statistiky porovnáme s kritickou hodnotou testu pro $\alpha = 0{,}05$ $\chi^2_{(1-\alpha)}{}^{(v=4)} = 9{,}49$.

7. Výsledek testu je následující:

$\chi^2_{(4)} > \chi^2_{(1-\alpha)}{}^{(v=4)}$, **zamítáme $H_0$ ($p < 0{,}05$)**

8. Datovou tabulku lze vyjádřit pomocí dílčích procent řádkových kategorií v rámci sloupců:

| Krevní skupina | Žaludeční vředy | Rakovina žaludku | Kontrola |
|---|---|---|---|
| 0 | 54,7 % | 43,4 % | 47,5 % |
| A | 37,8 % | 47,1 % | 43,1 % |
| B | 7,5 % | 9,5 % | 9,4 % |

9. Z procentuálního zastoupení je patrné, že ze složité kontingenční tabulky můžeme vytvořit jednodušší tabulku typu 3 × 2, která otestuje hypotézu o nezávislosti výskytu typu krevní skupiny pro nemocné rakovinou a pro zdravé lidi.

10. Pro tuto dílčí kontingenční tabulku znovu vypočteme hodnotu testové statistiky pro test dobré shody a porovnáme s kritickou hodnotou:

$\chi^2_{(2)} = 5{,}64 < \chi^2_{(0{,}95)}{}^{(v=2)} = 5{,}99$,
**nezamítáme $H_0$ ($p > 0{,}05$)**

Dle dílčího testu není ve výskytu krevních skupin rozdíl mezi nemocnými rakovinou a zdravými lidmi a tyto dva sloupce původní tabulky lze tedy sloučit.

11. Z procentuálního zastoupení kategorií je patrné, že se můžeme pokusit sestavit ještě jednodušší tabulku 2 × 2, která otestuje hypotézu o shodné distribuci krevních skupin A a B mezi kombinovaným vzorkem (sloučená skupina s rakovinou + kontrola) vs nemocní s žaludečními vředy (dílčí test neuvažuje krevní skupinu 0). Upravená tabulka je zde:

| Krevní skupina | Žaludeční vředy | Rakovina + kontrola | Celkem |
|---|---|---|---|
| A | 679 | 3 041 | 3 720 |
| B | 134 | 654 | 788 |
| Celkem | **813** | **3 695** | **4 508** |

12. I pro tuto dílčí kontingenční tabulku vypočteme hodnotu testové statistiky pro test dobré shody a porovnáme s kritickou hodnotou:

$\chi^2_{(1)} = 0{,}68 < \chi^2_{(0{,}95)}{}^{(v=1)} = 3{,}84$,
**nezamítáme $H_0$ ($p > 0{,}05$)**

Na základě tohoto výsledku můžeme tedy přistoupit ke sloučení kategorií v původní tabulce, tentokrát řádků s krevními skupinami A a B.

13. Dílčí analýzy směřují k tabulce 2 × 2, kde testujeme hypotézu o nezávislosti výskytu krevní skupiny 0 a sloučených skupin A + B mezi vzorkem lidí s žaludečním vředy oproti kombinovanému vzorku rakovina žaludku + kontrola. V této kontingenční tabulce již uvažujeme všechny pacienty z původního datového souboru.

$\chi^2_{(1)} = 34{,}29 > \chi^2_{(0{,}95)}{}^{(v=1)} = 3{,}84$,
**zamítáme $H_0$ ($p < 0{,}05$)**

**Závěr:** Zamítáme hypotézu o nezávislém výskytu krevních skupin a různých kategorií pacientů. Byl prokázán statisticky významný vztah mezi vyšším výskytem krevní skupiny 0 u pacientů s žaludečními vředy.

Průběh výpočtu lze shrnout takto:

| Provedené analýzy | Stupně volnosti | $\chi^2$ |
|---|---|---|
| Skupiny 0, A, B u pacientů s rakovinou (r) vs kontrola (k) | 2 | 5,64 |
| Skupiny A, B u pacientů s vředy vs kombinovaný vzorek (r + k) | 1 | 0,68 |
| Skupiny 0, A + B u pacientů s vředy vs kombinovaný vzorek (r + k) | 1 | 34,29 |
| Celkem | **4** | **40,61** |

Celkový součet dílčích testových statistik $\chi^2$ odpovídá původní hodnotě $\chi^2 = 40{,}61$. Tento kontrolní součet potvrzuje aditivitu testu dobré shody. Stejný závěr platí pro součty stupňů volnosti.

Příklad 2. Analýza složitější kontingenční tabulky pro dva nominální znaky.

proLékaře.cz | 3.5.2024



zjednodušená tabulka sčítají „své" hodnoty testové statistiky $\chi^2$ přesně do hodnoty $\chi^2$ vypočítané pro původní tabulku. Postupným slučováním vzájemně nezávislých kategorií tak vlastně naše zorné pole zbavujeme „nadbytečných" třídění a „koncentrujeme" hodnotu statistiky $\chi^2$ v konečné tabulce prokazující závislost.

Pevně věříme, že samostatný díl věnovaný složitějším kontingenčním tabulkám má smysl především pro čtenáře z řad lékařů. Tento typ asociačních analýz dvou i více znaků je v klinickém výzkumu velmi častý a zasahuje mnoho typů hodnocení diagnostiky a léčby, od zkoumání významu rizikových a prognostických faktorů až po analýzy vysvětlující léčebnou odpověď nebo výskyt komplikací vazbou k různým znakům. Tabulky typu 2 × 2 samozřejmě problém nepředstavují, neboť jejich výstupy jsou jednoznačné. Pokud ale u objemnější tabulky s více řádky a sloupci dosáhneme vysoce průkazného potvrzení vztahu sledovaných znaků, zkoumání tím nekončí, naopak spíše začíná. Málokdy je každá kategorie jednoho znaku nějak provázána s každou kategorií znaku druhého, a měli bychom tedy plně využít prostor pro detailní určení směru a významu zjištěné asociace. To znamená nejen obecné konstatování existující závislosti, ale také pojmenování kategorií znaků, mezi kterými vztah průkazně existuje.

Výše uvedené rozbory vypadají na první pohled velmi pracně – vždyť komu by se chtělo propočítávat různé dílčí testy dobré shody například u tabulky o rozměru 10 × 10? Ve skutečnosti tomu tak ale není. Jakmile již máme k dispozici vstupní tabulku četností, je provádění dílčích analýz velmi snadné. A není k tomu nutný ani specializovaný software. Stačí do internetového vyhledávače zadat heslo „chi square calculator" nebo „chi square analysis" a internet nabídne až desítky volně přístupných nástrojů pro hodnocení i velmi objemných tabulek. Sofistikovanější kalkulátory nabízejí i různé možnosti testu, například pro tabulku 2 × 2 alternativní aplikaci Fisherova exaktního testu nebo Yatesovu korekci pro testovou statistiku $\chi^2$ (viz díl XXI seriálu). I zde se tedy uživatel musí orientovat v terminologii a možnostech výpočtu, nudnou práci však lehce obstará stroj. I u kontingenčních tabulek tedy naštěstí stále platí, že lidé ovládají stroje, nikoli naopak :-)



# Neurovaskulární kongres 2010

Ve dnech 10.–12. 6. 2010 proběhl v Olomouci Neurovaskulární kongres 2010, zahrnující 38. české a slovenské cerebrovaskulární sympozium a 10. neurosonologické dny. Odborná i společenská část kongresu se konala v akademických prostorách Univerzity Palackého v Olomouci, instituce s téměř pětsetletou tradicí. Program kongresu byl multidisciplinární s účastí nejen lékařů různých odborností, ale také zdravotních sester, radiologických asistentů, fyzioterapeutů a dalších zájemců o širokou problematiku cévních mozkových příhod.

V rámci odborné části kongresu, pořádaného poprvé pod patronátem European Stroke Organisation a organizačně zajištěného Konferenčním servisem UP, zaznělo celkem 114 přednášek a dále bylo prezentováno 13 posterů. Mezi pozvanými zahraničními přednášejícími byli například Carlos A. Molina ze Španělska, Eva Bartels a Peter Schellinger z Německa, Alejandro Rabinstein z USA a László Oláh z Maďarska. Vědecká sdělení pokrývala široké spektrum problematiky cévních mozkových příhod od diagnostiky přes konzervativní, radiointervenční, neurochirurgickou a cévní chirurgickou léčbu akutního stadia ischemického i hemoragického iktu, problematiku neurointenzivní péče, primární a sekundární prevence, organizaci přednemocniční a nemocniční péče o pacienty s ikty až po léčbu následků iktů (včetně rehabilitace, léčby spasticity, bolestí hlavy, vaskulární demence a deprese). V rámci 10. neurosonologických dnů byl uspořádán neurosonologický kurz, ve kterém se mohli účastníci seznámit se základy vyšetřování vertebrálních tepen, možností využití peroperačního ultrazvukového monitoringu v neurochirurgii nebo využití transkraniální sonografie u neurodegenativních onemocnění, a to včetně praktických ukázek vyšetření.

Na kongresu byl předán diplom cerebrovaskulární sekce ČNS ČLS JEP prof. MUDr. Pavlu Kalvachovi, CSc., za jeho celoživotní přínos na poli cerebrovaskulárních onemocnění. Více než 400 účastníků kongresu se kromě odborného programu mohlo zúčastnit také programu společenského, a to první večer na nádvoří Zbrojnice za doprovodu Peter Lipa Jazz Bandu a následující večer na terase Přírodovědecké fakulty UP za doprovodu Blue Canisters a s vyhlídkou na slunce zapadající za panoramatem historického centra Olomouce. Účastníci kongresu měli možnost navštívit také Arcidiecézní muzeum. Počasí po celou dobu kongresu bylo příznivé rovněž pro individuální prohlídky historického centra a věříme, že umocňovalo příjemný dojem z času stráveného v Olomouci.

*doc. MUDr. Roman Herzig, Ph.D., FESO*
*předseda kongresu*
*doc. MUDr. David Školoudík, Ph.D.*
*vědecký sekretář kongresu*