# Analýza dat v neurologii

## LX. Analýza trendu ve vztahu kategoriálních znaků

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 minulém díle seriálu věnovaném kauzalitě vztahu znaků jsme zdůraznili, že jedním ze silných důkazů ve prospěch kauzality je statisticky významný vztah „dávka–odpověď". Skutečně, pokud s dávkou (úrovní) expozice nějakého faktoru narůstají i následky tohoto působení na zkoumaný biologický systém, jde s vysokou pravděpodobností o příčinný vztah. Proto je průkaz závislosti následků expozice na její dávce jedním z klíčových přístupů při zkoumání příčinnosti nejrůznějších vztahů v biologii i medicíně.

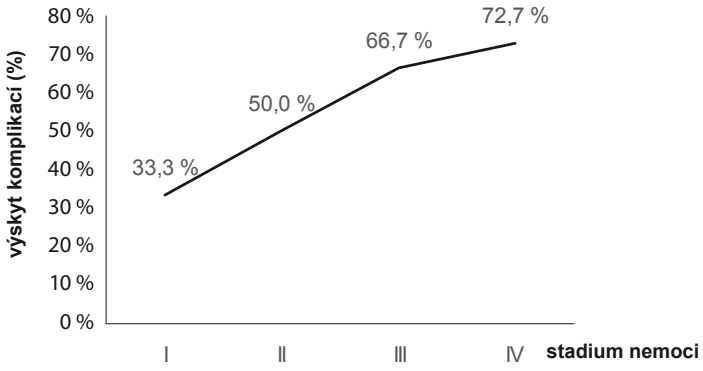
Vstupují-li takto do vzájemné interakce dva spojité (kvantitativní) faktory (např. expozice jako dávka léku přímo v koncentračních jednotkách a jako následek např. pokles tělesné teploty ve stupních Celsia), pak lze jednoduše zakreslit bodový či čárový graf a trend vztahu zviditelnit. Následně lze kvantifikovat změnu teploty na jednotku změny koncentrace léku a vztah hodnotit korelační či regresní analýzou. Těmto kvantitativním analýzám budeme věnovat další díly seriálu. U kategoriálních znaků je hodnocení vztahu „dávka–účinek" bohužel méně graficky atraktivní, nicméně i zde využitelné.

V tomto díle se zaměříme na testy trendů, které lze hodnotit z větších tabulek četností $R \times C$ ($R$ – řádky (*rows*); $C$ – sloupce (*columns*)), u kterých kategorie minimálně jednoho nebo 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 dů-

Pro vizualizaci trendu v kontingenčních tabulkách lze využít standardních čárových a skládaných sloupcových grafů. Příklady níže vizualizují frekvenční tabulky popisující vztah ordinální a binární proměnné (trend) a vztah dvou ordinálních (trendových) proměnných.

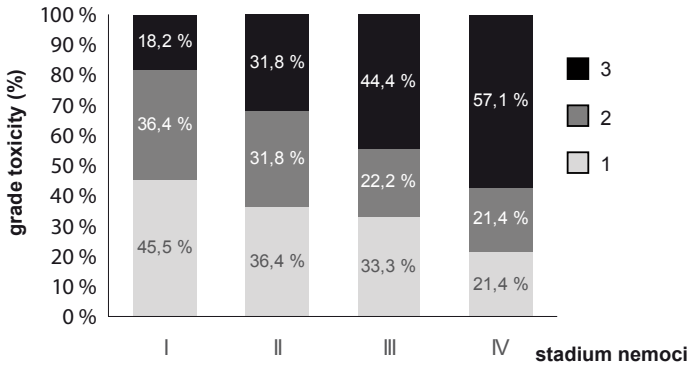
| Komplikace | Stadium nemoci I | II | III | IV |
|---|---|---|---|---|
| **N** | | | | |
| **ano** | 20 | 35 | 20 | 40 |
| **ne** | 40 | 35 | 10 | 15 |
| **%** | | | | |
| **ano** | 33,3 % | 50,0 % | 66,7 % | 72,7 % |
| **ne** | 66,7 % | 50,0 % | 33,3 % | 27,3 % |

V případě vztahu ordinální proměnné (trend) a binární proměnné je nejjednodušší vizualizací čárový graf popisující procentuální výskyt jedné z kategorií.




| Grade toxicity léčby | Stadium nemoci I | II | III | IV |
|---|---|---|---|---|
| **N** | | | | |
| **1** | 50 | 40 | 15 | 15 |
| **2** | 40 | 35 | 10 | 15 |
| **3** | 20 | 35 | 20 | 40 |
| **%** | | | | |
| **1** | 45,5 % | 36,4 % | 33,3 % | 21,4 % |
| **2** | 36,4 % | 31,8 % | 22,2 % | 21,4 % |
| **3** | 18,2 % | 31,8 % | 44,4 % | 57,1 % |

V případě vztahu dvou ordinálních (trendových) proměnných je nejjednodušší vizualizací skládaný sloupcový graf procentuálního zastoupení jednotlivých ordinálních kategorií.




Příklad 1. Vizualizace trendů v kontingenčních tabulkách.

Standardní test dobré shody (Chí-kvadrát test) je určen pro analýzu vztahu dvou kategoriálnách proměnných bez ohledu na pořadí kategorií a je tedy nepoužitelný pro hodnocení trendu ve vztahu ordinálních dat. Jako test vztahu ordinálních dat (trendu) se používá tzv. Goodmanovo a Kruskalovo Gama (často nazývané pouze gama). Pro výpočet Gama statistiky potřebujeme zjistit:

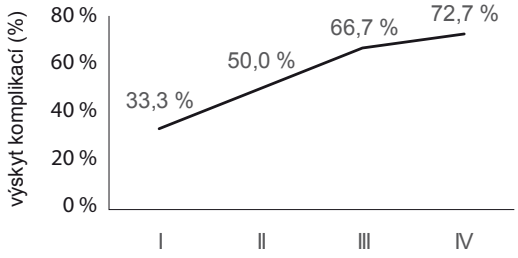
- $N_s$ – počet souhlasných párů (konkordantních) a $N_d$ – počet nesouhlasných párů (diskordantních).
- Souhlasné a nesouhlasné páry jsou definovány takto: pokud seřadíme dvojice pozorování *X*/*Y* podle *X* ve vzestupném pořadí a podíváme se na konkrétní dvojici, pak tato dvojice je souhlasná, pokud je hodnota *Y* vyšší než hodnota *Y* předchozí dvojice. Pro nesouhlasné dvojice platí, že hodnota *Y* je menší než předchozí hodnota *Y*.

Gama je potom vypočteno následovně: $G = \frac{N_s - N_d}{N_s + N_d}$

Testová statistika pro Gama je vypočtena aproximací Studentova rozdělení *t*, kde *n* odpovídá počtu pozorování: $t \approx G\sqrt{\frac{N_s + N_d}{n(1-G^2)}}$

**A) Kontingenční tabulka s pozorovatelným trendem**

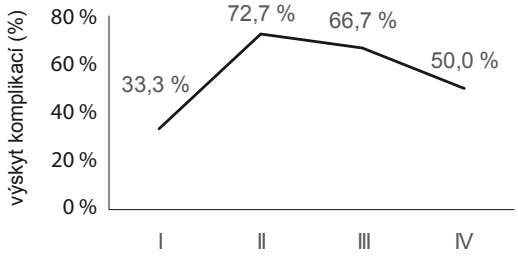
| | Stadium | | | |
|---|---|---|---|---|
| **Komplikace** | **I** | **II** | **III** | **IV** |
| **N** | | | | |
| **ano** | 20 | 35 | 20 | 40 |
| **ne** | 40 | 35 | 10 | 15 |
| **%** | | | | |
| **ano** | 33,3 % | 50,0 % | 66,7 % | 72,7 % |
| **ne** | 66,7 % | 50,0 % | 33,3 % | 27,3 % |




Chí-kvadrát test $p < 0,001$; Gama = 0,452 ($p < 0,001$): Chí-kvadrát test identifikoval statisticky významný vztah mezi proměnnými, Gama statistika identifikovala statisticky významný trend ve vztahu obou proměnných.

**B) Kontingenční tabulka bez pozorovatelného trendu**

| | Stadium | | | |
|---|---|---|---|---|
| **Komplikace** | **I** | **II** | **III** | **IV** |
| **N** | | | | |
| **ano** | 20 | 40 | 20 | 35 |
| **ne** | 40 | 15 | 10 | 35 |
| **%** | | | | |
| **ano** | 33,3 % | 72,7 % | 66,7 % | 50,0 % |
| **ne** | 66,7 % | 27,3 % | 33,3 % | 50,0 % |




Chí-kvadrát test $p < 0,001$; Gama = 0,147 ($p = 0,156$): Chí-kvadrát test identifikoval statisticky významný vztah mezi proměnnými, Gama statistika nicméně neprokázala statisticky významný trend ve vztahu obou proměnných.

Příklad 2. Testování trendu v kontingenční tabulce 2 × C.

ležitá. Úvodem připomeňme zcela elementární skutečnosti:

- Minimálně jeden ze zkoumaných znaků musí být kategoriální s více než dvěma kategoriemi. Pokud z experimentu získáme nejjednodušší tabulku četností 2 × 2, pak sice můžeme hodnotit sílu vztahu dvou binárních parametrů, ale hovořit o trendu zde postrádá smysl. Trend lze analyzovat ve vztahu jednoho binárního znaku a jednoho znaku s více než dvěma ordinálními kategoriemi a ovšem také ve vztahu dvou a více znaků s více než dvěma ordinálními kategoriemi.
- Velmi nutnou podmínkou analýzy trendu je, aby zkoumané kategoriální proměnné byly ordinální. Tedy aby jejich hodnoty vytvářely jasně řazenou ordinální škálu od nejmenší hodnoty po největší. Pokud by totiž jednotlivé hodnoty kategoriálního znaku mohly být v tabulce seřazeny jakkoli, nemá hodnocení trendu smysl.
- Připomeňme ještě, že ordinální znaky jsou definovány tak, že jejich hodnoty je možno navzájem uspořádat, ale není známa míra (kvantita) toho, jak jsou od sebe jednotlivé kategorie vzdáleny. Tento fakt má pro hodnocení trendu velmi vážné důsledky. V takovém prostoru totiž můžeme prokázat trendovou asociaci znaků (pozitivní, kdy obě ordinální škály spolu klesají či stoupají, nebo negativní, kdy se jejich hodnoty vyvíjejí v opačném směru), ale nemůžeme kvantifikovat diference. Nelze tedy určit, o kolik se změní hodnota jednoho znaku při jednotkové změně hodnoty jiného znaku, neboť vzdálenosti mezi body na ordinální stupnici nejsou kvantifikovatelné. Známe jen pořadí (*rank*) bodů.

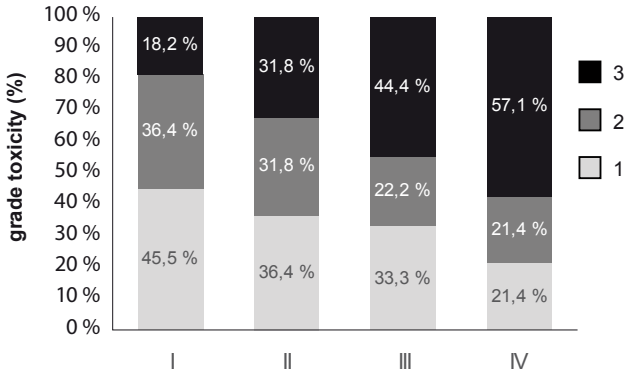
Je nepochybné, že trend v tabulce četností nepůsobí jako nějaký neutrální prvek, ale silně ovlivňuje rozložení četností v jednotlivých polích tabulky. Potvrzení existence statisticky významného trendu tedy také nutně znamená zamítnutí nulové hypotézy o neexistenci vztahu obou znaků; je-li mezi hodnotami znaků trend, musí mezi nimi být i vztah (asociace). Čím významnější trend je, tím více se četnosti v tabulkách mění v závislosti na hodnotách obou znaků a tím průkaznější bude i sama existence obecného vztahu obou znaků.

Příklad 1 dokumentuje schematicky ukázky možných tabulek četností vstupujících do trendové analýzy a jejich kvalitativní grafické znázornění. V případě, že proti sobě v tabulce vystupují binární znak a znak ordinální, pak v podstatě studujeme, zda se ordinální škála liší ve dvou kategoriích daného binárního znaku. Příklad 2 dokumentuje výpočet testu pro trend v kontingenční tabulce 2 × *C* a příklad 3 ten samý výpočet pro kontingenční tabulku *R* × *C*.

Při testování síly vztahu a trendu ve složitějších tabulkách četností musíme při interpretaci vždy dbát na možný vliv různých

**A) Kontingenční tabulka s pozorovatelným trendem**

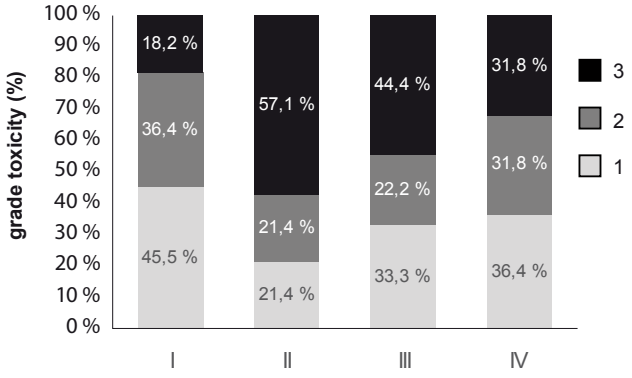
| Grade toxicity léčby | Stadium nemoci I | II | III | IV |
|---|---|---|---|---|
| **N** | | | | |
| **1** | 50 | 40 | 15 | 15 |
| **2** | 40 | 35 | 10 | 15 |
| **3** | 20 | 35 | 20 | 40 |
| **%** | | | | |
| **1** | 45,5 % | 36,4 % | 33,3 % | 21,4 % |
| **2** | 36,4 % | 31,8 % | 22,2 % | 21,4 % |
| **3** | 18,2 % | 31,8 % | 44,4 % | 57,1 % |




Chí-kvadrát test $p < 0{,}001$; Gama = 0,334 ($p < 0{,}001$): Chí-kvadrát test identifikoval statisticky významný vztah mezi proměnnými, Gama statistika identifikovala statisticky významný trend ve vztahu obou proměnných.

**B) Kontingenční tabulka bez pozorovatelného trendu**

| Grade toxicity léčby | Stadium nemoci I | II | III | IV |
|---|---|---|---|---|
| **N** | | | | |
| **1** | 50 | 15 | 15 | 40 |
| **2** | 40 | 15 | 10 | 35 |
| **3** | 20 | 40 | 20 | 35 |
| **%** | | | | |
| **1** | 45,5 % | 21,4 % | 33,3 % | 36,4 % |
| **2** | 36,4 % | 21,4 % | 22,2 % | 31,8 % |
| **3** | 18,2 % | 57,1 % | 44,4 % | 31,8 % |




Chí-kvadrát test $p < 0{,}001$; Gama = 0,115 ($p = 0{,}075$): Chí-kvadrát test identifikoval statisticky významný vztah mezi proměnnými, Gama statistika nicméně neprokázala statisticky významný trend ve vztahu obou proměnných.

Příklad 3. Testování trendu v kontingenční tabulce $R \times C$.

faktorů zkreslení. Interpretace takových analýz směřuje k diskuzi o kauzalitě vztahu dvou nebo více proměnných a nelze ji opřít pouze o výsledek jednoho statistického testu (viz též předchozí díl seriálu). Vedle možnosti zkreslení již samotným designem studie a náběrem probandů vstupují do hry i potenciální zavádějící faktory. Jako příklad uveďme studium výskytu vrozených vývojových vad u dětí, kde z frekvenčních dat může vyplynout silný vztah mezi pořadím narození dítěte a rizikem takové vady. Avšak skutečným rizikovým faktorem tu není pořadí novorozence, ale věk matky.