proLékaře.cz | 1.5.2025

Analýza dat v neurologii

LXXVIII. Směsné míry korelace u vícerozměrných souborů kvantitativních a kvalitativních dat

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

Institut biostatistiky a analýz,
LF MU, Brno

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

Typickým výstupem reálných experimentů jsou tzv. mnohorozměrné (vícerozměrné) soubory dat, kdy je *N* jedinců popisováno *K* proměnnými a zápis datového souboru vytváří datovou matici $N \times K$. S rozšiřujícím se arzenálem různých vyšetřovacích metod a zejména s nástupem molekulárně biologických a genetických vyšetření se tento trend týká i klasického klinického výzkumu a výsledné datové matice zahrnují i mnoho desítek proměnných. Logicky vzniká potřeba vyhodnotit vzájemnou korelaci všech těchto proměnných, přičemž zdaleka ne vždy jde o proměnné kvantitativní, tedy spojité. V reálné praxi stojíme i před úkolem vyjádřit korelaci spojitých (metrických) proměnných (např. koncentrace látky v krvi, povrch těla pacienta apod.) s proměnnými ordinálními či binárními (např. dávka léčiva v několika kategoriích či toxicita léčby ve stupních dle grade). Těmto problémům budeme stručně věnovat tento díl seriálu.

Představme si, že máme za úkol popsat korelaci mezi spojitou proměnnou a proměnnou binární (diskrétní). Pro tento účel se používají tzv. **biseriální korelace**, které vedou k odhadu tzv. **biseriálního korelačního koeficientu**. Literatura rozlišuje několik typů těchto korelací podle toho, o jakou diskrétní proměnnou jde. Avšak než se pustíme do dalšího výkladu, musíme zdůraznit, že korelace v těchto případech dává smysl, pouze pokud lze diskrétní proměnnou vzestupně či sestupně jednoznačně uspořádat (tedy musí jít o binární znak nabývající hodnoty 0 či 1 anebo o znak ordinální, kde mají kategorie jasné pořadí). Pokud by diskrétní proměnná byla neuspořádaná, tedy dána v podstatě náhodnými kategoriemi bez pořadí (např. nominální znaky), pak korelace postrádá jakýkoli smysl a nelze ji vyčíslit.

Poměrně často používaným typem biseriálních korelací je tzv. bodově biseriální korelace vyjadřující sílu vztahu mezi spojitou proměnnou a proměnnou binární. Bodový biseriální koeficient korelace proměnných *X* (binární) a *Y* (spojitá) vypočítáme dle relativně jednoduchého vztahu, který dokumentuje příklad 1. Koeficient můžeme značit jako R_{bis} nebo R_{pb} z anglického „point biserial". Postup je jednoduchý, hodnoty *Y* rozdělíme podle toho, zda k nim příslušná hodnota *X* je rovna 1 nebo 0 a následně pracujeme s průměrem hodnot *Y* v rámci každé z těchto skupin. Příklad koreluje s proměnnou *X*, která značí podání léku proti horečce při infekci (ano/ne),

Hodnotíme vztah mezi podáním léku u pacientů (proměnná *X*, kde 0 = ne a 1 = ano) a koncentrací určitého markeru v periferní krvi (proměnná *Y*).

X	*Y*
0	3,6
0	4,1
0	4,8
0	5,1
0	5,3
0	6
1	5,7
1	6,3
1	6,8
1	7,2

Pro výpočet biseriální korelace z dat je potřebné zjistit:

$\overline{y_0}$ a $\overline{y_1}$ – průměrné koncentrace markeru ve skupinách s podáním a bez podání léku;

s – směrodatná odchylka změřených koncentrací markeru pro celý hodnocený soubor;

n, n_0 a n_1 – celkový počet pacientů a počty ve skupinách dle podání léku.

$$R_{bis} = \frac{\overline{y_1} - \overline{y_0}}{s}\sqrt{\frac{n_0 n_1}{n(n-1)}} = \frac{6{,}5 - 4{,}8}{1{,}1}\sqrt{\frac{6 \times 4}{10 \times 9}} = 0{,}76016$$

Po dosazení do rovnice získáme hodnotu biseriální korelace 0,76016.

Stejně jako jiné typy korelačních koeficientů i biseriální korelační koeficient je možné doplnit statistickou významností, v tomto hodnoceném příkladu je $p = 0{,}011$. Závěrem příkladu můžeme konstatovat, že podání léku statisticky významně koreluje s koncentrací markeru v krvi.

Příklad 1. Výpočet bodového biseriálního korelačního koeficientu a test jeho statistické významnosti.

proLékaře.cz | 1.5.2025

Hodnotíme vztah mezi podáním léku u pacientů (proměnná *X*, kde 0 = ne a 1 = ano) a koncentrací určitého markeru v krvi (proměnná *Y*). Pro účely příkladu budeme předpokládat nenormalitu dat a nutnost využití pořadové biseriální korelace.

X	*Y*	Pořadí
0	3,6	1
0	4,1	2
0	4,8	3
0	5,1	4
0	5,3	5
0	6	7
1	5,7	6
1	6,3	8
1	6,8	9
1	7,2	10

Stejně jako u jiných neparametrických statistických metod i zde výpočet vychází z pořadí změřených hodnot. Pro dosazení do vztahu pro výpočet korelace je potřebné:

R_0 a R_1 – průměrné pořadí koncentrace markeru ve skupině s podáním a bez podání léku;

n – celkový počet pacientů.

$$R_{bis} = \frac{2(R_1 - R_0)}{n} = \frac{2(8{,}3 - 3{,}7)}{10} = 0{,}91667$$

Po dosazení do rovnice získáme hodnotu biseriální korelace 0,91667.

Stejně jako jiné typy korelačních koeficientů i pořadový biseriální korelační koeficient je možné doplnit statistickou významností, v hodnoceném příkladu je $p < 0{,}001$. Závěrem příkladu můžeme konstatovat, že podání léku koreluje statisticky významně s koncentrací markeru v krvi.

Příklad 2. Výpočet pořadové biseriální korelace.

Pro hodnocení vztahu dvou kategoriálních proměnných můžeme využít koeficient kontingence. Jeho výpočet vychází ze standardního testu dobré shody pro kontingenční tabulky.

	Skupina	
Výsledek	experimentální	kontrolní
dobrý	33	47
uspokojivý	76	153
špatný	6	25

V našem příkladu hodnotíme vztah mezi skupinami s experimentální a kontrolní léčbou a výsledkem léčby popsaným 3 kategoriemi.

Celková velikost vzorku je $n = 340$, tabulka má $r = 3$ řádky a $c = 2$ sloupce a chi-kvadrát statistika má hodnotu 4,912.

Vztah pro základní formu koeficientu kontingence je: $$KK = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{4{,}912}{340 + 4{,}912}} = 0{,}1193$$

Nevýhodou takto kalkulované statistiky avšak je, že i při úplné závislosti je hodnota *KK* menší než 1. Tento problém řešíme vztažením hodnoty *KK* na její maximální možnou hodnotu, tedy tzv. standardizací *KK*.

Vztah pro maximální možnou hodnotu koeficientu kontingence je: $$KK_{max} = \left(\frac{r-1}{r} \times \frac{c-1}{c}\right)^{\frac{1}{4}} = \left(\frac{2}{3} \times \frac{1}{2}\right)^{\frac{1}{4}} = 0{,}7598$$

Vztah pro standardizovanou hodnotu koeficientu kontingence je: $$KK_{stand} = \frac{KK}{KK_{max}} = \frac{0{,}1193}{0{,}7598} = 0{,}157$$

Statistická významnost koeficientu kontingence odpovídá hodnotě chi-kvadrát testu pro příslušnou kontingenční tabulku (v tomto příkladu jde o hodnotu p = 0,086). Můžeme tedy konstatovat, že vztah mezi skupinou pacientů a výsledky léčby je relativně slabý ($KK_{stand} = 0{,}157$) a statisticky nevýznamný ($p = 0{,}086$).

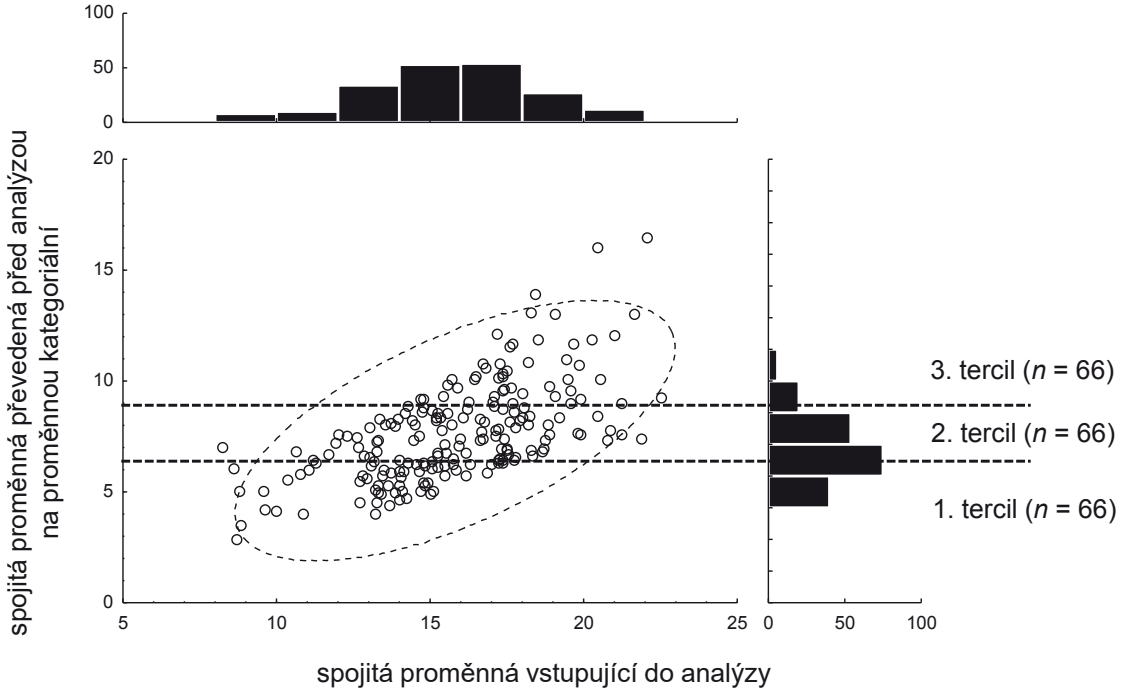
Příklad 3. Koeficient kontingence.

s dobou, do kdy dojde k poklesu tělesné teploty pacientů. Příklad také dokládá statistický test významnosti tohoto koeficientu, tedy ověření platnosti hypotézy $R_{bis} = 0$.

Velmi užitečnou modifikací výše uvedeného výpočtu je tzv. pořadový biseriální korelační koeficient, který je využitelný za situací, kdy spojitá proměnná *Y* nesplňuje předpoklad normálního rozdělení hodnot. Výpočet je srovnatelný s výše uvedeným bodovým korelačním koeficientem, jen pracuje s průměrným pořadím hodnot *Y* pro kategorii *X* = 1 a pro kategorii *X* = 0. Příklad 2 dokumentuje odhad tohoto koeficientu na datech hodnotících vliv podpůrné předoperační terapie (proměnná *X* nabývající hodnoty 1/0) a doby rekonvalescence pacienta po operaci (spojitá proměnná *Y*).

proLékaře.cz | 1.5.2025

Biseriální a polyseriální korelace mohou být využívány i za situace, kdy primárně sice pracujeme se spojitou proměnnou, ale přesné určení kvantitativní hodnoty u jednotlivých měření není možné nebo není smysluplné. Může např. jít o situaci, kdy měření není dostatečně přesné a primární spojitá hodnota je jen odhadem. Jiným příkladem je situace, kdy sice máme primárně spojitou proměnnou (např. koncentrace nějakého markeru v krvi), ale pro klinickou praxi má smysl pracovat pouze s jejími kategoriemi odlišujícími normální a patologické hladiny. V těchto případech lze spojitou proměnnou nahradit vzestupně uspořádanými intervaly hodnot a vzniká tak skrytě spojitá proměnná, která do následného výpočtu vstupuje jako proměnná binární či kategoriální. Častým způsobem rozdělení spojité proměnné na kategorie je členění dle hodnoty percentilů. V příkladu níže jsou takto využity tercily, které dělí soubor na 3 stejně početné kategorie.



Příklad 4. Schematické znázornění kategorizace spojité proměnné před korelační analýzou s jinou spojitou proměnnou.

Zobecněním biseriálních korelací jsou tzv. korelace polyseriální, které analyzují vztah spojité proměnné s proměnnou kategoriální (ordinální). Proměnná *X* zde tedy nenabývá pouze hodnot ano/ne, ale je uspořádanou škálou hodnot, které např. vyjadřují odstupňovanou a rostoucí dávku podaného léčiva apod. Tyto korelace předpokládají, že za kategoriemi proměnné *X* existuje skrytá spojitá proměnná, jejíž hodnoty kategorie *X* reprezentují. Obdobným předpokladem jsou vybaveny také tzv. korelace polychorické, které odhadují sílu vztahu dvou diskrétních proměnných. Tyto metody již svou složitostí překračují rámec této kapitoly a je také nutno poznamenat, že metodou první volby při studiu vzájemného vztahu (asociace) dvou diskrétních znaků jsou jednoznačně kontingenční tabulky (např. díl 21 a 22 našeho seriálu). Pro ověření závislosti kategoriálních znaků uspořádaných v kontingenční tabulce standardně používáme chí-kvadrát test nezávislosti dvou znaků. Jako nadstavba analýzy kontingenčních tabulek se využívá tzv. koeficient kontingence, jehož výpočet přibližuje příklad 3.

Na závěr je nutné zdůraznit, že výše uvedené typy biseriálních a polyseriálních korelací mohou být využívány i za situace, kdy primárně pracujeme se spojitou proměnnou, ale přesné určení kvantitativní hodnoty u jednotlivých měření není možné, např. při odečítání počtu kolonií při bakteriologickém výsevu na živné půdě nebo při hodnocení stupně vyrážky na kůži. V těchto případech lze spojitou proměnnou nahradit pouze vzestupně uspořádanými intervaly hodnot a vzniká tak skrytě spojitá proměnná, kterou pro následný výpočet zastupuje proměnná binární či kategoriální, rozdělená do skupin hodnot. Tímto způsobem vlastně elegantně řešíme nepřesnost primárních měření, aniž bychom museli nějak modifikovat experimentální plán. Daný postup schematicky znázorňuje graf uvedený v příkladu 4.