Analýza dat v neurologii

LXX. Kovariance

L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková
Institut biostatistiky a analýz,
LF MU, Brno

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

Minulý díl seriálu jsme věnovali úvodu do analýzy kovariance, kterou jsme představili jako jeden ze základních ukazatelů vztahu dvou kvantitativních proměnných. Označíme-li tyto proměnné *X* a *Y*, pak kovarianci značíme cov(*X*, *Y*).

Připomeňme z minulého dílu, že odhad kovariance kalkulujeme podle následujícího vztahu:

$cov(X, Y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \times (y_i - \bar{y})}{N - 1}$, kde

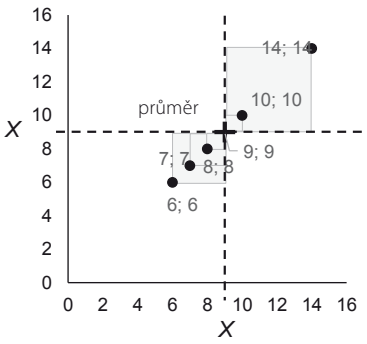
- x_i, y_i jsou jednotlivé hodnoty proměnných *X* a *Y* naměřené párově u $i = 1$ až $i = N$ jedinců v analyzovaném souboru,
- $\bar{x}$, $\bar{y}$ jsou aritmetické průměry proměnných *X* a *Y*.

V tomto díle se dále zaměříme na vybrané vlastnosti kovariance jako statistického ukazatele, představíme postupy pro testování její statistické významnosti a doplníme užitečné informace k jejímu využití. Zamysleme se nejprve v několika následujících poznámkách nad výpočtem hodnoty kovariance dle výše uvedeného vztahu, neboť již z něj lze odvodit interpretační význam kovariance, ale také její limity.

Hodnota kovariance je jednoznačně závislá na rozložení hodnot proměnných *X* a *Y* kolem jejich aritmetického průměru, neboť čitatel je součtem násobků vzdáleností každé jednotlivé hodnoty x_i a y_i od průměru $\bar{x}$, respektive $\bar{y}$. Pokud hodnoty *X* a *Y* vykazují na měřených subjektech stejný trend (vztah), pak rostou stejným směrem od průměru a kovariance nabývá kladných hodnot, tím větších, čím je tento vztah průkaznější. Naopak, pokud hodnoty *X* a *Y* jdou v pozici vůči svým průměrům opačným směrem, je kovariance číselně záporná a vyjadřuje zá-

Příklad číselně dokládá platnost vztahu kovariance *(X, X)* = rozptyl *(X)* neboli *cov(X, X) = var(X)*.

x	$x - \bar{x}$	x	$x - \bar{x}$	$(x - \bar{x}) * (x - \bar{x})$
7	–2	7	–2	4
6	–3	6	–3	9
10	1	10	1	1
14	5	14	5	25
8	–1	8	–1	1
$\bar{x} = 9$		$\bar{x} = 9$		$\sum (x - \bar{x}) * (x - \bar{x}) = 40$



$$cov(X, X) = var(X)$$

$$cov(X, X) = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) * (x_i - \bar{x})}{N - 1}$$

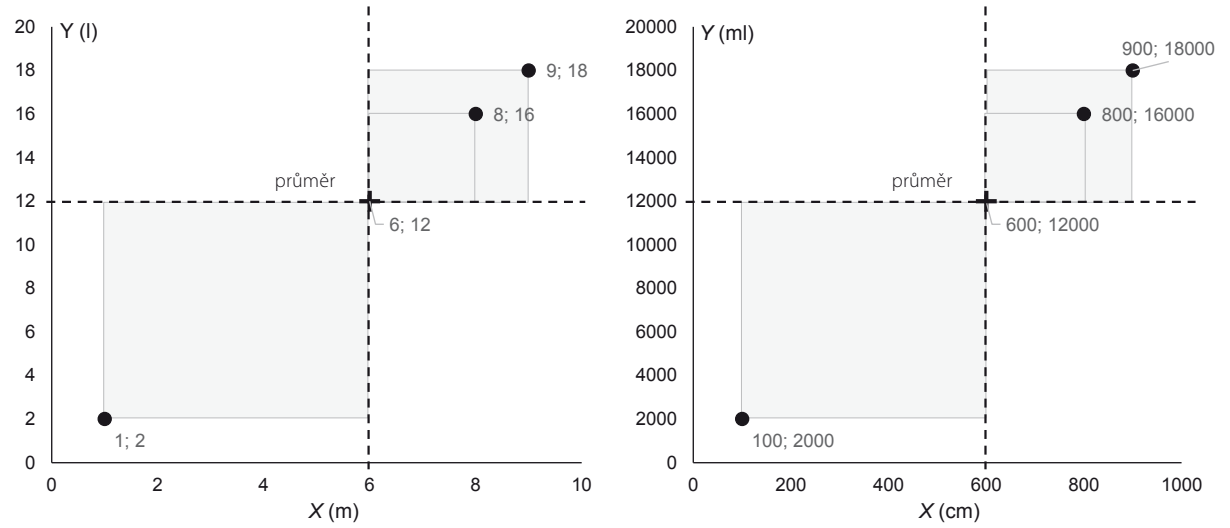
$$= \frac{(7-9)*(7-9)}{4} + \frac{(6-9)*(6-9)}{4} + \frac{(10-9)*(10-9)}{4} + \frac{(14-9)*(14-9)}{4} + \frac{(8-9)*(8-9)}{4} = 10$$

$$var(X) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$= \frac{(7-9)*(7-9)}{4} + \frac{(6-9)*(6-9)}{4} + \frac{(10-9)*(10-9)}{4} + \frac{(14-9)*(14-9)}{4} + \frac{(8-9)*(8-9)}{4} = 10$$

Příklad 1. Výpočet kovariance vychází z hodnot rozptylu proměnných, jejichž vztah studujeme.

Kovariance, stejně jako rozptyl, je svojí hodnotou závislá na číselných jednotkách hodnocených proměnných. Pouhá změna jednotek, např. za účelem zpřesnění měření, může vést k zásadní změně hodnoty kovariance. Příklady níže ukazují ovlivnění výsledné hodnoty kovariance změnou jednotek proměnné *X* (z m na cm) a *Y* (z l na ml).



$cov(X, Y) = 38$

$cov(X, Y) = 3\,800\,000$

Přestože se směr závislosti *X* a *Y* nijak kvalitativně nezměnil a vizuálně jsou oba grafy v podstatě totožné, hodnoty kovariance se výrazně liší. Problémem využití kovariance jako míry závislosti proměnných tedy je neexistující maximální možná hodnota, tj. hodnota vyjadřující maximální sílu vztahu *X* a *Y* (všechny body *X* a *Y* by v takové situaci ležely na přímce).

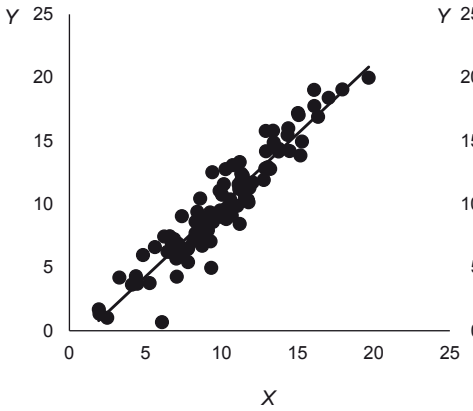
Příklad 2. Hodnoty kovariance jsou určovány rozptylem, a tedy jednotkami proměnných, jejichž vztah zkoumáme.

Hodnota kovariance je stochastickým odhadem a může být statisticky testována. Typickou nulovou hypotézou je cov*(X, Y)* = 0, alternativní hypotézou pak cov*(X, Y)* ≠ 0. Testová statistika *t* má Studentovo rozdělení s n – 2 stupni volnosti.

$$t = \frac{\frac{cov(X,Y)}{s_x s_y}\sqrt{N-2}}{\sqrt{1-\left(\frac{cov(X,Y)}{s_x s_y}\right)^2}}$$

Příklady 3A–C ukazují výsledky testu při různých hodnotách kovariance.

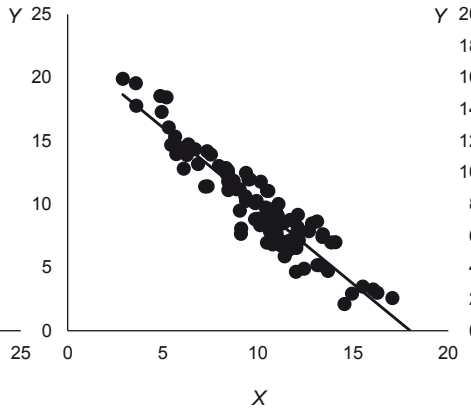
3A) významná kladná kovariance



cov(X, Y) = 13,896
t = 28,462
p < 0,001

Statisticky významná kladná kovariance, mezi proměnnými existuje kladný, statisticky významný, vztah.

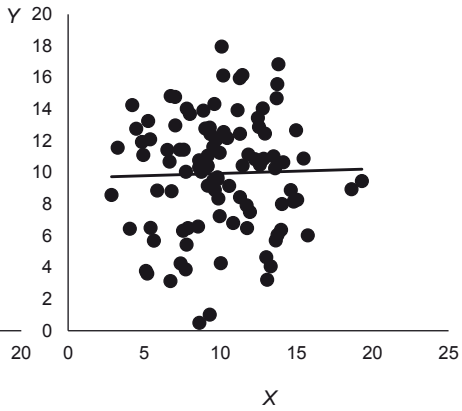
3B) významná záporná kovariance



cov(X, Y) = –10,886
t = –27,447
p < 0,001

Statisticky významná záporná kovariance, mezi proměnnými existuje záporný, statisticky významný, vztah.

3C) nevýznamná kovariance



cov(X, Y) = 0,326
t = 0,265
p = 0,791

Statisticky nevýznamná kovariance, nelze zamítnout nulovou hypotézu *cov(X, Y)* = 0. Mezi proměnnými neexistuje statisticky významný vztah.

Příklad 3. Testování statistické významnosti kovariance.

V případech, kdy potřebujeme posoudit vzájemný vztah mezi více než dvěma proměnnými, lze samozřejmě odhadovat hodnotu kovariance mezi všemi dvojicemi testovaných proměnných. Tab. 1 ukazuje příklad vstupního datového souboru se čtyřmi proměnnými, jejichž vzájemné vztahy je možno zapsat v podobě tzv. kovarianční matice (tab. 2) a znázornit formou maticového grafu (graf 1). Na hlavní diagonále kovarianční matice jsou hodnoty rozptylu jednotlivých proměnných X1–X4, neboť platí vztah *cov(X,X) = var(X)*.

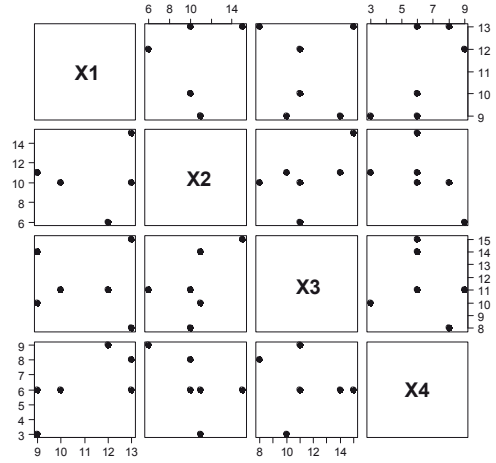
Tab. 1. Vstupní data.

X1	X2	X3	X4
13	10	8	8
13	15	15	6
12	6	11	9
10	10	11	6
9	11	10	3
9	11	14	6
$\bar{x} = 11$	$\bar{x} = 10{,}5$	$\bar{x} = 11{,}5$	$\bar{x} = 6{,}33$

Tab. 2. Kovarianční matice.

	X1	X2	X3	X4
X1	3,6	0,4	–0,4	2,6
X2		8,3	4,1	–3,2
X3			6,7	–0,8
X4				4,3

Graf 1. Maticový graf X1, X2, X3, X4.



Příklad 4. Kovarianční matice a její interpretace.

porný vztah obou proměnných. Nulová či nule blízká hodnota kovariance potom dokládá neexistenci vztahu *X* a *Y*, jejichž hodnoty na sobě nijak nezávisí a vyskytují se v pozici vůči svým průměrným hodnotám zcela náhodně.

Čím jsou tedy hodnoty proměnných *X* a *Y* více „rozptýleny" kolem jejich průměru, tím je hodnota kovariance numericky vyšší, ať již v záporných nebo kladných číslech. Proto se o kovarianci v odborné literatuře někdy píše jako o společném rozptylu proměnných *X* a *Y*, jejichž závislost studujeme. Její výpočet totiž skutečně vychází z výpočtu pro rozptyl, který jednoduše definujeme jako průměrný čtverec vzdálenosti od průměru. Pokud tedy dosadíme do vztahu pro výpočet kovariance místo hodnoty *Y* hodnotu *X*, dostaneme vztah pro výpočet rozptylu proměnné *X*, který označme *var(X)*:

$$cov(X, X) = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \times (x_i - \bar{x})}{N - 1} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1} = var(X)$$

A obdobně by samozřejmě platilo, že *cov(Y, Y) = var(Y)*. Příklad 1 dokládá na konkrétním souboru dat platnost tohoto vztahu mezi kovariancí a rozptylem. Mnohé čtenáře nyní jistě napadá legitimní otázka, zda jsme zde ve výkladu již nepřešli k příliš detailním matematickým podrobnostem a zda tyto informace mají prakticky využitelný výstup. Odpověď zní jednoznačně ano, neboť z výše uvedeného vyplývají zásadní interpretační omezení odhadu kovariance. Absolutní hodnoty kovariance totiž nejsou určovány pouze silou vztahu proměnných *X* a *Y*, ale zejména jejich jednotkami a tedy i velikostí jejich rozptylu, který je číselně rovněž určen jednotkami *X* a *Y*. Budeme-li např. zkoumat vztah mezi výškou a hmotností lidské postavy, vyjde kovariance v absolutních hodnotách zcela jinak při měření výšky v metrech nebo v centimetrech. V tomto smyslu je kovariance číselně nestandardizovaný ukazatel a velikost kovariance není nijak omezena. Tento fakt dokládá příklad 2 tohoto dílu seriálu.

Pro odhad kovariance tedy není definována maximální hodnota, která by vyjadřovala nejsilnější možný vztah zkoumaných proměnných (jejich hodnoty by v takovém případě ležely přesně na přímce). Naopak, situaci ještě komplikuje fakt, že kovariance je statistika tzv. parametrická, což znamená, že předpokladem pro její výpočet je smysluplná výpovědní hodnota aritmetického průměru jako středu normálního (Gaussova) rozdělení hodnot. Předpokládáme tedy, že proměnné *X* a *Y* naplňují definici normálního rozdělení, které známe jako rozdělení symetrické, bez odlehlých hodnot a s hodnotou aritmetického průměru rovnou mediánu. Významně odlehlé hodnoty jedné nebo obou zkoumaných proměnných silně ovlivňují číselnou hodnotu kovariance, neboť v čitateli pro její výpočet se objeví velká čí-

selná hodnota rozdílů $x_i - \bar{x}$ nebo $y_i - \bar{y}$. V extrémním případě tak může jedna jediná hodnota vést k vysoké hodnotě kovariance, která by po jejím vyloučení z výpočtu byla nulová nebo blízká nule. Z tohoto důvodu nesmí být kontrola rozdělení hodnot zkoumaných proměnných podceněna.

Z výše uvedeného vyplývá, že z absolutních hodnot kovariance nelze prvoplánově usuzovat sílu vztahu zkoumaných proměnných a dále že hodnoty kovariance odhadnuté v různých studiích jsou jen obtížně srovnatelné. O to větší význam má testování statistické významnosti kovariance, které by mělo být téměř povinným doplňkem publikovaných hodnot. Kovariance je stochastický ukazatel a o jejích hodnotách lze tedy formulovat různé hypotézy a jejich platnost ověřovat statistickými testy. Standardní hypotézou je nulová hypotéza, že kovariance je rovna nule a mezi proměnnými *X* a *Y* tedy není žádný prokazatelný vztah. Zamítnutím této hypotézy statistickým testem na dané hladině významnosti potvrzujeme statisticky významný vztah mezi zkoumanými proměnnými.

Připomeňme, že statistické testy pracují s tzv. testovou statistikou, kterou počítáme dle definovaného vztahu a výsledek vyhodnocujeme pomocí pravděpodobnosti. Testová statistika odhadu kovariance má Studentovo rozdělení pro *N – 2* stupňů volnosti a její výpočet zde dokládá příklad 3. Ze vztahu pro testovou statistiku je zřejmé, že čím větší je hodnota kovariance, ať již kladná nebo záporná, tím větší absolutní hodnoty dosahuje hodnota statistiky *t* a tím větší je pravděpodobnost zamítnutí nulové hypotézy $cov(X, Y) = 0$.

Fakt, že pro kovarianci není definována maximální možná hodnota a číselná hodnota odhadu kovariance závisí na jednotkách a rozptylu zkoumaných proměnných, nemusí být vždy nevýhodou. Například zkoumáme-li vztah dvou proměnných, které mají finanční význam (*X*: investice v Kč; *Y*: výnosy v Kč), je absolutní číselná hodnota kovariance přímo využitelná pro posouzení síly vztahu. Obecně při posuzování vztahu dvou proměnných, které si vzájemně odpovídají jednotkami i číselným rozsahem, může mít absolutní hodnota kovariance přímou interpretaci.

Tento díl seriálu uzavřeme příkladem 4, který znázorňuje situaci, kdy potřebujeme posoudit hodnotu kovariance pro více než 2 proměnné. Potřeba vyjádřit se současně o větším počtu proměnných je v praxi velmi častá a vede k vícerozměrnému přístupu v korelační analýze. Při současném zpracování *K* proměnných hodnotíme kovarianci pro $K * (K - 1)/2$ dvojic proměnných, které sestavujeme do tzv. kovarianční matice, jejíž řádky i sloupce jsou věnovány postupně první až *K*-té proměnné. Na průsečíku *i*-tého řádku a *j*-tého sloupce je uvedena kovariance *i*-té a *j*-té proměnné. Kovarianční matice je čtvercová (symetrická podle hlavní diagonály) a na diagonále obsahuje rozptyly zkoumaných proměnných, neboť platí výše zdůvodněný vztah $cov(X, X) = var(X)$.