Analýza dat v neurologii

LXIX. Kovariance

V minulém díle seriálu jsme otevřeli problematiku korelační analýzy, která je velmi široce používanou metodikou. V nejširším slova smyslu označujeme pojmem korelace kvantifikaci vzájemného vztahu proměnných, které jsou kvantitativní. Kvantifikace síly a vyhodnocení statistické významnosti takových vztahů je základním úkolem statistiky, která pro tento účel vyvinula několik velmi dobře interpretovatelných ukazatelů. Jedním ze základních ukazatelů vztahu dvou kvantitativních proměnných je tzv. kovariance (covariance). Příklady výpočtu kovariance tedy v tomto díle zahájíme výklad nástrojů korelační analýzy. V následujících dílech se posuneme k výkladu korelace a různé ukazatele budeme mezi sebou srovnávat zejména z hlediska jejich interpretace.

Kovariance je kvantitativním ukazatelem vzájemné souvislosti dvou náhodných veličin. Značíme ji *cov(X, Y)* a v přesné definici jde o střední hodnotu součinu rozdílu náhodných veličin a jejich středních hodnot. V zjednodušeném výkladu můžeme kovarianci představit jako hodnotu společného rozptylu proměnných *X* a *Y*, jejichž závislost studujeme. Výpočet hodnoty kovariance také skutečně vychází z rozptylu *X* a *Y*:

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

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 průměry proměnných X a Y.

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

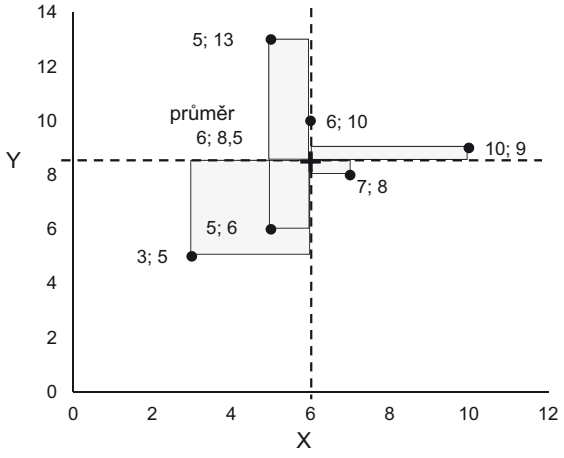
✉
doc. RNDr. Ladislav Dušek, Ph.D.
Institut biostatistiky a analýz,
LF MU, Brno
e-mail: dusek@iba.muni.cz

Jinou formou zápisu vztahu pro výpočet kovariance může být:
cov(X, Y) = cov(Y, X) = E(*X* – E[*X*]) (*Y* – E[*Y*]), kde E(*X*), resp. E(*Y*) značí střední hodnoty veličiny *X*, resp. *Y*.

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

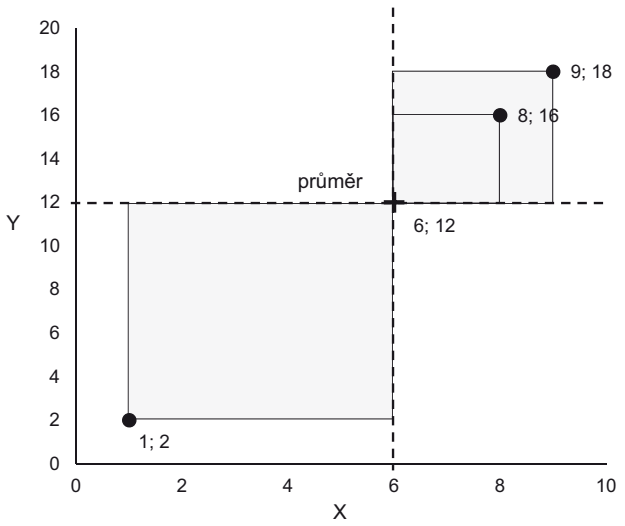
x	$x - \bar{x}$	y	$y - \bar{y}$	$(x - \bar{x}) * (y - \bar{y})$
3	–3	5	–3,5	10,5
10	4	9	0,5	2
5	–1	13	4,5	–4,5
5	–1	6	–2,5	2,5
7	1	8	–0,5	–0,5
6	0	10	1,5	0
$\bar{x} = 6$		$\bar{y} = 8,5$		$\sum (x - \bar{x}) * (y - \bar{y}) = 10$

$$cov(X,\ Y) = \frac{(3-6)*(5-8{,}5)}{5} + \frac{(10-6)*(9-8{,}5)}{5} + \frac{(5-6)*(13-8{,}5)}{5} + \frac{(5-6)*(6-8{,}5)}{5} + \frac{(7-6)*(8-8{,}5)}{5} + \frac{(6-6)*(10-8{,}5)}{5} = 2$$



Příklad 1. Výpočet kovariance v jednoduchém číselném příkladu.

Kovariance je jedním z parametrických ukazatelů vztahu dvou spojitých proměnných. Jak naznačuje její název, jde o hodnocení rozptylu sdíleného dvěma spojitými proměnnými, a také její výpočet odpovídá vzorci pro rozptyl, pouze modifikovanému pro dvě proměnné (pokud bychom počítali kovarianci proměnné na sebe samotnou, dostaneme její rozptyl).



Vzorec pro výpočet kovariance je:

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

kde *N* je počet objektů v souboru (každý popsaný hodnotami obou proměnných x_i a y_i) a $\bar{x}$ a $\bar{y}$ průměrné hodnoty jednotlivých proměnných.

$$cov(X,\ Y) = \frac{(9-6) * (18-12)}{2} + \frac{(8-6) * (16-12) + (1-6) * (2-12)}{2} = 38$$

Kovariance (sdílený rozptyl) hodnocených proměnných je 38.

Příklad 2. Výpočet kovariance s grafickým znázorněním.

Z tohoto vztahu je zřejmé, že velikost rozptylu hodnot *X* a *Y* kolem průměru těchto proměnných určuje číselnou hodnotu kovariance. Obecně kovariance vyjadřuje, jak se hodnoty obou proměnných pohybují vůči sobě. Vyjadřuje, zda se tyto proměnné pohybují ve stejném směru (kladná kovariance, větší než 0), nebo ve směru opačném (záporná kovariance, menší než 0). Nulová hodnota kovariance znamená, že proměnné *X* a *Y* nemají žádný vztah a různé hodnoty *X* se vyskytují zcela náhodně pro různé hodnoty *Y*, resp. různé hodnoty *Y* se vyskytují náhodně (mohou nabývat libovolných hodnot) pro různé hodnoty *X*. V takovém případě se v čitateli vzorce pro výpočet kovariance náhodně potkávají kladné i záporné vzdálenosti konkrétních hodnoty x_i a y_i od průměrů proměnných a v součtu se vzájemně vynulují. Dále platí:

- pokud je *cov(X, Y)* větší než 0, pak je souvislost mezi veličinami *X* a *Y* pozitivní, tzn., že čím je větší *X*, tím je větší *Y* a naopak;
- pokud je *cov(X, Y)* menší než 0, pak je souvislost mezi veličinami *X* a *Y* negativní, tzn., že čím je větší *X* tím je menší *Y* a naopak;
- platí, že nezávislé veličiny mají *cov(X, Y)* rovnu nule, ale bohužel neplatí, že by *cov(X, Y)* rovnou nule znamenalo, že *X* a *Y* jsou nezávislé; mezi proměnnými může existovat jiný než lineární vztah;
- sama hodnota kovariance nevypovídá nic o relativní síle vazby *X* a *Y*, neboť je vyjádřena přímo v jednotkách *X* a *Y*; např. hodnota kovariance hmotnosti a výšky postavy bude numericky větší, pokud výšku vyjádříme v cm, než když ji vyjádříme v metrech.

Zejména poslední bod ve výše uvedeném výčtu vlastností kovariance je velmi podstatný. Říká totiž, že hodnota kovariace není nijak ohraničena a je odvislá od jednotek proměnných *X* a *Y*. Z tohoto důvodu nelze mezi sebou přímo srovnávat absolutní hodnoty kovariance odhadnuté na různých souborech dat a je tedy nutné tento ukazatel nějakou formou standardizovat, např. pomocí výpočtu tzv. korelačního koeficientu. Této problematice se bude podrobně věnovat příští díl seriálu.

Výpočet hodnoty kovariance zde dokládá číselný příklad 1, doplněný grafickým znázorněním na příkladu 2.