Marie Turčičová

Korelace

Korelace měří sílu (těsnost) vzájemné závislosti dvou náhodných veličin (znaků) X a Y. Znakem X může být třeba délka a znakem Y hmotnost dítěte. Skutečnou (teoretickou) korelaci těchto dvou znaků v populaci (např. v populaci dětí) udává populační korelační koeficient, který se značí $\rho_{X,Y}$ a je dán vzorcem

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \cdot \operatorname{var} Y}}, \tag{1}$$

kde $\operatorname{cov}(X,Y) = \mathrm{E}(X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)$ je kovariance veličin X a Y. S tou už jsme se také kdysi setkali. Kovariance (stejně jako korelace) vyjadřuje vzájemnou závislost veličin, ale její nevýhoda je, že závisí na měřítku[1]. Díky vydělení odmocninou z rozptylů se ale tato závislost odstraní a korelace je již na měřítku veličin X a Y nezávislá. Korelace má řadu zajímavých vlastností:

- $-1 \le \rho_{X,Y} \le 1$ (tj. nabývá hodnot od -1 do 1)
- měří sílu lineární závislosti (tj. $\rho_{X,Y}$ nabývá hodnoty 1 nebo -1 pokud X je lineární funkcí Y, to jest $X = a + bX$ pro nějaké a, b)
- Jsou-li X, Y nezávislé, pak $\rho_{X,Y} = 0$. (Pozor! Opačná implikace platí jen pro normální rozdělení! Viz další bod.)
- Pochází-li vektor (X,Y) z dvojrozměrného normálního rozdělení a $\rho_{X,Y} = 0$, pak jsou X a Y nezávislé.
- Je-li $\rho_{X,Y} \neq 0$, pak jsou X a Y určitě závislé. (To platí bez ohledu na rozdělení).

Pearsonův korelační koeficient

Odhadem populačního korelačního koeficientu je Pearsonův korelační koeficient, se kterým jsme se setkali už dříve. Máme-li k dispozici realize veličin X a Y naměřených na n objektech, tj. naše data jsou tvaru

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

pak Pearsonův korelační koeficient vypočteme jako

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{2}$$

Chceme-li otestovat hypotézu, že znaky X a Y jsou v populaci nezávislé (tj. jejich populační korelace $\rho_{X,Y} = 0$), můžeme k tomu využít následující postup:

Máme-li výběr $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$, který pochází z dvojrozměrného normálního rozdělení, pak lze testovat hypotézy

$$H_0 : \rho_{X,Y} = 0 \quad (\text{tj. } X, Y \text{ jsou nezávislé})$$
$$H_1 : \rho_{X,Y} \neq 0 \quad (\text{tj. } X, Y \text{ nejsou nezávislé})$$

pomocí statistiky $T = \frac{r_{X,Y}}{\sqrt{1-r_{X,Y}^2}}\sqrt{n-2}$, která má v případě, že by H_0 platila, rozdělení t_{n-2}. Nulovou hypotézu zamítám, pokud $r_{X,Y}$ je dost daleko od nuly, tj. pokud $|T| \ge qt_{n-2}(1 - \frac{\alpha}{2})$.

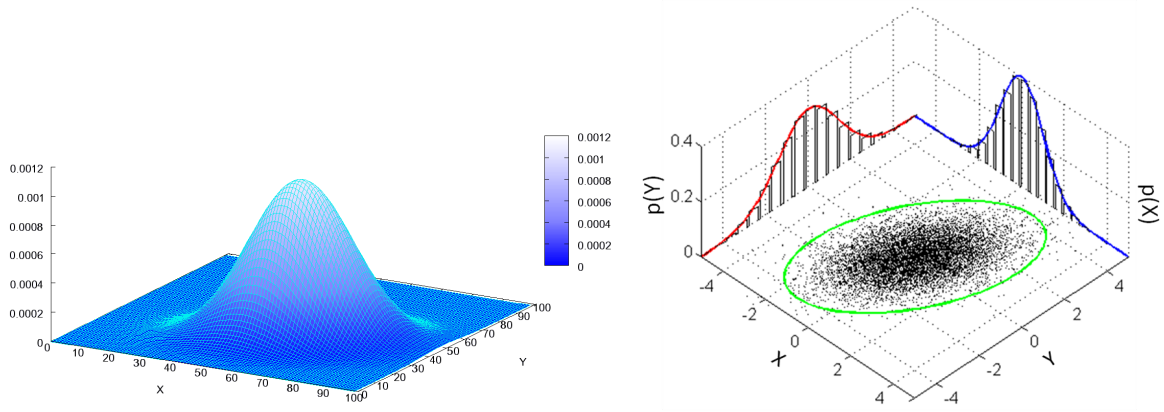
Kde samozřejmě $qt_{n-2}(1 - \frac{\alpha}{2})$ značí $(1 - \frac{\alpha}{2}) \cdot 100\%$ kvantil t-rozdělení s $n-2$ stupni volnosti.

[1] Počítáte-li kovarianci váhy dítěte a jeho délky v centimetrech, dostanete jiné číslo, než když je délka udaná v metrech. Tento „problém“ u korelace není.

Poznámka 1 *Pro použití tohoto testu je samozřejmě nutné ověřit předpoklad, že data pocházejí z dvojrozměrného normálního rozdělení. Sice existuje mnohorozměrná verze Shapiro-Wilkova testu, my se ale spokojíme s okometrickou metodou založenou na bodovém diagramu (scatter plotu).*

Hustota jednorozměrného normálního (Gaussova) rozdělení je křivka zvonovitého tvaru. Hustota dvojrozměrného normálního rozdělení má podobný tvar, akorát ve dvou rozměrech. Připomíná trochu horu Říp, nebo homoli cukru. Prohlédnout si ji můžete na Obrázku 1(a). Když se podíváte shora na libovolný horizontální řez touto hustotou, uvidíte elipsu nebo kružnici. Proto by i data pocházející z dvojrozměrného normálního rozdělení měla po vykreslení do bodového diagramu tvořit elipsu/kružnici (viz Obrázek 1(b)), kdy směrem od středu bodů postupně ubývá.

Normalitu budeme tedy ověřovat tak, že data vykreslíme do bodového diagramu a podíváme se, zda body přibližně tvoří elipsu.



(a) Hustota

(b) Bodový diagram hodnot pocházejících z dvojrozměrného normálního rozdělení

Obrázek 1: Dvojrozměrné normální rozdělení (obrázky jsou převzaté z wikipedie)

Spearmanův korelační koeficient

Tento korelační koeficient meří sílu monotónní závislosti[2], ne pouze lineární. Hodí se například pokud:

- chceme testovat nezávislost pomocí výběrového korelačního koeficientu, ale je porušen předpoklad normality
- nebo když hodnoty znaků X a Y na n jedincích nelze přímo změřit, ale je k dispozici pouze jejich pořadí (např. máme dva someliéry, kteří hodnotí n lahví vína - na základě ochutnávky jsou schopní vínům přiřadit pouze jejich pořadí).

Mějme opět náš náhodný výběr

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

a pro i od 1 do n si označme jako R_i pořadí X_i v rámci $X_1, \ldots, X_n$ a Q_i jako pořadí Y_i v rámci $Y_1, \ldots, Y_n$. Pak Spearmanův korelační koeficient lze spočítat jako

$$r^s_{X,Y} = \frac{\sum_{i=1}^n (R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \cdot \sum_{i=1}^n (Q_i - \bar{Q})^2}} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n (R_i - Q_i)^2. \tag{3}$$

Je to vlastně Pearsonův korelační koeficient spočítaný z hodnot pořadí

$$\begin{pmatrix} R_1 \\ Q_1 \end{pmatrix}, \begin{pmatrix} R_2 \\ Q_2 \end{pmatrix}, \ldots, \begin{pmatrix} R_n \\ Q_n \end{pmatrix}.$$

[2] to jest nabývá hodnoty 1 nebo -1 když X je čistě rostoucí, nebo čistě klesající funkcí Y

Na Spearmanově korelačním koeficientu lze založit test nezávislosti, který již nepotřebuje předpoklad normálního rozdělení.

Testujeme-li hypotézy

$$H_0 : X, Y \text{ jsou nezávislé}$$
$$H_1 : X, Y \text{ nejsou nezávislé}$$

pak nulovou hypotézu zamítneme, pokud $|\sqrt{n-1} \cdot r^s_{X,Y}| \geq qnorm(1 - \frac{\alpha}{2})$.

Kde $qnorm(1 - \frac{\alpha}{2})$ značí kvantil rozdělení $N(0, 1)$.

Poznámka 2 *Korelace neznamená kauzalitu (příčinnost). To, že mají dvě veličiny mezi sebou vysokou korelaci, neznamená, že jedna z nich je příčinou druhé! Vtipnou ilustraci tohoto faktu naleznete na:* `https://www.tylervigen.com/spurious-correlations`*, kde jsou vykresleny reálné hodnoty veličin, které spolu zjevně nejsou v příčinném vztahu, ačkoli mají vysokou hodnotu Pearsonova korelačního koeficientu.*