9.3.5 Korelace

Předpoklady: 9304

Zatím jsme se zabývali vždy pouze jedním znakem, ve statistickém výzkumu jsme však u každého jednotlivce (statistické jednotky) sledovali znaků více. Určitě spolu některé znaky souvisí (například výška a hmotnost) $\Rightarrow$ jde souvislost zachytit matematicky (výpočtem)?

Korelační koeficient znaků x a y: $r_{x,y} = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y}$.

Jak vzorec pozná, že spolu dva znaky souvisí?
Vyzkoušíme jeho funkci na konkrétním případě několika studentů uvedených v tabulce:

Výška	205	150	180	155
Hmotnost	95	51	55	85

Protože se ve vzorci vyskytují ještě průměry, musíme předpokládat, že známe průměrné hodnoty výšky (například 175 cm) a hmotnosti (například 75 kg).

Př. 1: Projdi hodnoty uvedené v tabulce a najdi sloupce, které podporují hypotézu, že větší lidé jsou v průměru těžší. Které sloupce této hypotéze odporují?

Hypotézu podporují sloupce, ve kterých je jak výška, tak hmotnost větší než průměr, nebo sloupce, ve kterých jsou obě hodnoty menší než průměr. Naopak hypotéze odporují sloupce, ve kterých je jedna z hodnot větší než průměr a druhá je menší $\Rightarrow$

- hypotézu podporují sloupce 1 (obě hodnoty větší než průměr) a 2 (obě hodnoty menší než průměr)
- hypotéze odporují sloupce 3 a 4 (jedna hodnota větší než průměr, druhá menší).

Př. 2: Dosaď jednotlivé sloupce tabulky do výrazu $(x_i - \bar{x})(y_i - \bar{y})$ a zhodnoť, jak přispívají k celkovému součtu $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

- 1. sloupec: $(x_i - \bar{x})(y_i - \bar{y}) = (205-175)(95-75) = 30 \cdot 20 = 600$ $\Rightarrow$ získali jsme **kladné** číslo, které je tím větší, čím větší jsou obě hodnoty v porovnání s průměry.
- 2. sloupec: $(x_i - \bar{x})(y_i - \bar{y}) = (150-175)(51-75) = (-25)\cdot(-24) = 600$ $\Rightarrow$ získali jsme **kladné** číslo, které je tím větší, čím menší jsou obě hodnoty v porovnání s průměry.
- 3. sloupec: $(x_i - \bar{x})(y_i - \bar{y}) = (180-175)(55-75) = 5\cdot(-20) = -100$ $\Rightarrow$ získali jsme **záporné** číslo, které je tím větší, čím více se obě hodnoty liší od svých průměrů.
- 4. sloupec: $(x_i - \bar{x})(y_i - \bar{y}) = (155-175)(85-75) = (-20)\cdot 10 = -200$ $\Rightarrow$ získali jsme **záporné** číslo, které je tím větší, čím více se obě hodnoty liší od svých průměrů.

V příkladu jsme si ukázali, že statistické jednotky, které potvrzují hypotézu „větší je těžší“, přispívají do sumy kladnými čísly, statistické jednotky, které hypotézu popírají, přispívají zápornými čísly.

Zkusíme rozvažovat obecně a sledovat hodnotu součinu v sumě:

- vysoká a těžká statistická jednotka (v souladu s představou, že oba znaky spolu souvisí) $\Rightarrow$ $x_i > \overline{x}$, $y_i > \overline{y}$ $\Rightarrow$ součin $(x_i - \overline{x})(y_i - \overline{y})$ je součinem dvou kladných čísel $\Rightarrow$ do sumy přidáváme kladné číslo (zvětšujeme její hodnotu),
- malá a lehká statistická jednotka (v souladu s představou, že oba znaky spolu souvisí) $\Rightarrow$ $x_i < \overline{x}$, $y_i < \overline{y}$ $\Rightarrow$ součin $(x_i - \overline{x})(y_i - \overline{y})$ je součinem dvou záporných čísel $\Rightarrow$ do sumy přidáváme kladné číslo (zvětšujeme její hodnotu),
- vysoká a lehká statistická jednotka (odporuje představě, že oba znaky spolu souvisí) $\Rightarrow$ $x_i > \overline{x}$, $y_i > \overline{y}$ $\Rightarrow$ součin $(x_i - \overline{x})(y_i - \overline{y})$ je součinem kladného čísla $(x_i - \overline{x})$ a záporného čísla $(y_i - \overline{y})$ $\Rightarrow$ do sumy přidáváme záporné číslo (zmenšujeme její hodnotu),
- malá a těžká statistická jednotka (odporuje představě, že oba znaky spolu souvisí) $\Rightarrow$ $x_i > \overline{x}$, $y_i > \overline{y}$ $\Rightarrow$ součin $(x_i - \overline{x})(y_i - \overline{y})$ je součinem záporného čísla $(x_i - \overline{x})$ a kladného čísla $(y_i - \overline{y})$ $\Rightarrow$ do sumy přidáváme záporné číslo (zmenšujeme její hodnotu).

$\Rightarrow$ Pokud většina jednotek odpovídá představě, že oba znaky spolu souvisí, získáme sumací kladné číslo, pokud je počet členů, které představě odpovídají, přibližně stejný jako počet členů, které ji vyvrací, získáme sumací číslo blízké nule.

Jaký význam mají zbývající části vzorce?

- $\frac{1}{n}$ - známe z výpočtu průměru i rozptylu, zabraňuje tomu, aby při větším počtu členů vyšel větší výsledek.
- $s_x \cdot s_y$ - sumou sčítáme násobky odchylek od průměrů $\Rightarrow$ pro soubory s větším rozptylem bychom získali větší hodnotu i při menší míře závislosti $\Rightarrow$ po vydělení součinem $s_x \cdot s_y$ odstraníme závislost na rozptylu hodnot a získáme výsledek v intervalu $\langle -1;1 \rangle$.

Př. 3: Co vypovídá o vztahu veličin *x* a *y* hodnota korelace blízká:
a) 1 b) –1 c) 0?

a) $r(x, y)$ se blíží 1

1 je nejvyšší možná hodnota koeficientu r $\Rightarrow$ součiny $(x_i - \overline{x})(y_i - \overline{y})$ musely do sumy přispívat kladnými čísly $\Rightarrow$ veličiny *x*, *y* jsou svázány úzkým vztahem „větší *x* znamená větší *y*“.

b) $r(x, y)$ se blíží -1

-1 je nejnižší možná hodnota koeficientu r $\Rightarrow$ součiny $(x_i - \overline{x})(y_i - \overline{y})$ musely do sumy přispívat téměř pořád zápornými čísly (popíraly hypotézu „větší znamená těžší“) $\Rightarrow$ veličiny *x*, *y* jsou svázány úzkým vztahem „větší *x* znamená menší *y*“.

c) $r(x, y)$ se blíží 0

součiny $(x_i-\overline{x})(y_i-\overline{y})$ musely do sumy přispívat stejně kladnými i zápornými čísly $\Rightarrow$ veličiny *x*, *y* nejsou svázány vztahem „větší *x* znamená menší *y*“ (ani vztahem opačným).

Pedagogická poznámka: Následující odvození opět pouze ukáži pomocí projektoru.

Tvar $r_{x,y}=\dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{s_x\cdot s_y}$ umožňuje interpretovat vnitřní logiku vzorce, ale pro praktické výpočty je příliš složitý.
Čitatel zlomku je možné upravit takto:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i-\overline{x}y_i-x_i\overline{y}+\overline{xy})=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i+\overline{xy}-\overline{x}y_i+\overline{xy}-x_i\overline{y}-\overline{xy})=$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_iy_i+\overline{x}(\overline{y}-y_i)+\overline{y}(\overline{x}-x_i)-\overline{xy})=\frac{1}{n}\sum_{i=1}^{n}x_iy_i+\frac{1}{n}\sum_{i=1}^{n}\overline{x}(\overline{y}-y_i)+\frac{1}{n}\sum_{i=1}^{n}\overline{y}(\overline{x}-x_i)-\frac{1}{n}\sum_{i=1}^{n}\overline{xy}$$

Upravíme jednotlivé sumy:

- $\frac{1}{n}\sum_{i=1}^{n}\overline{x}(\overline{y}-y_i)=\frac{\overline{x}}{n}\sum_{i=1}^{n}(\overline{y}-y_i)=\frac{\overline{x}}{n}\cdot 0$ (z minulé hodiny průměr je taková hodnota, aby se odchylky na obě strany navzájem odečetly),
- $\frac{1}{n}\sum_{i=1}^{n}\overline{y}(\overline{x}-x_i)=\frac{\overline{y}}{n}\sum_{i=1}^{n}(\overline{x}-x_i)=\frac{\overline{y}}{n}\cdot 0$ (z minulé hodiny průměr je taková hodnota, aby se odchylky na obě strany navzájem odečetly),
- $\frac{1}{n}\sum_{i=1}^{n}\overline{xy}=\frac{n\overline{xy}}{n}=\overline{xy}$ (*n*-krát sčítáme stále stejnou hodnotu součinu průměrů $\overline{xy}$).

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i-\overline{x}y_i-x_i\overline{y}+\overline{xy})=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{xy}$$

Praktičtější vztah pro výpočet korelace: $r_{x,y}=\dfrac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{xy}}{s_x\cdot s_y}$.

Samotný korelační koeficient se často označuje jako Pearsonův korelační koeficient, výraz v čitateli se nazývá kovariance (termín dobře popisuje o co jde: KO(společné) VARIANCE(odchylky).

Př. 4: V tabulce je uvedeno prvních šest dvojic znaků známka z matematiky a známka z fyziky. Urči jejich korelační koeficient.

x (známka z matematiky	3	2	2	4	3	2
y (známka z fyziky)	3	2	1	3	2	2

Pomocné výpočty: $\overline{x}=\dfrac{3+2+2+4+3+2}{6}=2{,}67$, $\overline{y}=\dfrac{3+2+1+3+2+2}{6}=2{,}17$.

$$s_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2}=\sqrt{\frac{1}{6}\left(3^2+2^2+2^2+4^2+3^2+2^2\right)-2{,}67^2}=0{,}733$$

$$s_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2-\overline{y}^2}=\sqrt{\frac{1}{6}\left(3^2+2^2+1^2+3^2+2^2+2^2\right)-2{,}17^2}=0{,}677$$

$$r_{x,y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}}{s_x \cdot s_y} = \frac{\frac{1}{6}\left(3\cdot 3+2\cdot 2+2\cdot 1+4\cdot 3+3\cdot 2+2\cdot 2\right) - 2{,}67\cdot 2{,}17}{0{,}733\cdot 0{,}677} = 0{,}751$$

Hodnota $r_{x,y} = 0{,}751$ znamená již značnou míru závislosti.

Ruční výpočet korelačního koeficientu je značně zdlouhavý i pro pouhých šest dvojic hodnot. Výpočet je možné (za příznivých okolností) urychlit tím, že sestavíme a využijeme tabulku četností, tentokrát četností dvojic hodnot znaků *x* a *y* $\Rightarrow$ tabulka nemůže mít pouze jeden řádek na zápis četností, sledujeme dvojici znaků a každá možná dvojice hodnot potřebuje své políčko.

Pedagogická poznámka: Následující příklad není veden jako příklad, abych ho mohl jednak postupně vysvětlovat u tabule (hlavně začátek je těžký) a jednak libovolně urychlovat tak, aby na příklad 5 zbylo alespoň deset minut.

Například pro dvojice znaků „známka z matematiky“ (pět hodnot) a „doba strávená studiem“ (pět hodnot), potřebujeme $5\times 5 = 25$ políček.

		Známka z matematiky				
		1	**2**	**3**	**4**	**5**
Doba strávená studiem	**1**					
	2			2	1	
	3	1	2	8	2	
	4		3			
	5					

Trojka v druhém sloupci a čtvrté řádce znamená, že tři žáci mají z matematiky dvojku (druhý sloupec) a zároveň tráví studiem trochu větší než průměrné množství času.

Z tabulky můžeme snadno získat i četnosti pro jednotlivé znaky, například 2 z matematiky má pět žáků, které získáme součtem hodnot ve druhém sloupci tabulky.

x – známka z matematiky, *y* – doba strávená studiem

Pomocné výpočty: $\overline{x} = \frac{1\cdot 1+2\cdot 5+3\cdot 10+4\cdot 3}{19} \doteq 2{,}79$, $\overline{y} = \frac{2\cdot 3+3\cdot 13+4\cdot 3}{19} = 3$.

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_j^{*2} n_j - \overline{x}^2} = \sqrt{\frac{1}{19}\left(1\cdot 1^2+5\cdot 2^2+10\cdot 3^2+3\cdot 4^2\right) - 2{,}79^2} = 0{,}764$$

$$s_y = \sqrt{\frac{1}{n}\sum_{j=1}^{r} y_j^{*2} n_j - \overline{y}^2} = \sqrt{\frac{1}{19}\left(2^2\cdot 3+3^2\cdot 13+4^2\cdot 3\right) - 3^2} = 0{,}562$$

$$r_{x,y} = \frac{\frac{1}{19}\left(2\cdot 3\cdot 2+1\cdot 4\cdot 2+1\cdot 1\cdot 3+2\cdot 2\cdot 3+8\cdot 3\cdot 3+2\cdot 4\cdot 3+3\cdot 2\cdot 4\right) - 2{,}79\cdot 3}{0{,}764\cdot 0{,}562} = -0{,}494$$

Co znamená záporná hodnota korelačního koeficientu?

Žáci, kteří se více snaží (více hodin), mají lepší známku z matematiky $\Rightarrow$ více hodin studia znamená menší známku z matematiky $\Rightarrow$ oba znaky jsou na sobě závislé, nadprůměrným hodnotám času, odpovídají podprůměrné hodnoty známky (většina členů v sumě by byla záporná).

Př. 5: Sestav tabulku relativních četností a urči korelaci znaků Známka z matematiky a Maturita z matematiky. Studentům, kteří maturovat nebudou, přiřaď hodnotu 0, studentům, kteří maturovat budou, hodnotu 1. Ještě před výpočtem odhadni hodnotu korelačního koeficientu.

Hodnota korelačního koeficientu bude zřejmě záporná, protože žáci nižší (lepší) známkou z matematiky budou s větší pravděpodobností z matematiky maturovat (a budou mít u znaku Maturita z matematiky hodnotu 1 místo hodnoty 0).

		Známka z matematiky (x)				
		1	**2**	**3**	**4**	**5**
Maturita z matematiky (y)	**0**		2	8	3	
	1	1	3	2		

x – známka z matematiky, y – maturita z matematiky

Pomocné výpočty: $\overline{x} = \frac{1\cdot 1+2\cdot 5+3\cdot 10+4\cdot 3}{19} \doteq 2,79$, $\overline{y} = \frac{6\cdot 1+13\cdot 0}{19} = 0,316$.

$$s_x = \sqrt{\frac{1}{n}\sum_{j=1}^{r} x_j^{*2} n_j - \overline{x}^2} = \sqrt{\frac{1}{19}\left(1\cdot 1^2 + 5\cdot 2^2 + 10\cdot 3^2 + 3\cdot 4^2\right) - 2,79^2} = 0,764$$

$$s_y = \sqrt{\frac{1}{n}\sum_{j=1}^{r} y_j^{*2} n_j - \overline{y}^2} = \sqrt{\frac{1^2\cdot 6 + 0^2\cdot 13}{19} - 0,316^2} = 0,465$$

$$r_{x,y} = \frac{\frac{1}{n}\sum_{j=1}^{r} x_j y_j - \overline{xy}}{s_x \cdot s_y} = \frac{\frac{1}{19}\left(1\cdot 1\cdot 1 + 2\cdot 2\cdot 0 + 3\cdot 2\cdot 1 + 8\cdot 3\cdot 0 + 2\cdot 3\cdot 1 + 3\cdot 4\cdot 0\right) - 2,79\cdot 0,316}{0,764\cdot 0,465} =$$

$$= -0,556$$

Na závěr je nutné upozornit, že pomocí korelace zjišťujeme vzájemnou souvislost dvou znaků. Ze vzájemné souvislosti však nijak nevyplývá příčinný vztah dokonce ani skutečná vzájemná souvislost vztah. I velmi vysoké hodnoty korelačního koeficientu mohou být dílem náhody, zejména dnes, kdy je možné automaticky prohledávat obrovská množství informací. Velmi zajímavé příklady náhodných korelací jsou uvedeny na stránkách Spurious correlations.

Navíc i v případě, že vysoká hodnota korelačního koeficientu odhalí skutečný vztah dvou veličin, nic nám neříká o směru příčinné souvislosti. Například je zřejmé, že výška platu koreluje s cenou soukromého automobilu. Tvrdit však, že si musíme koupit drahé auto, aby nám zvýšili mzdu, by bylo velmi odvážné. Každý cítí, že příčinná souvislost je zřejmě opačná.

Shrnutí: Korelace umožňuje zachytit vzájemnou souvislost dvou veličin.