

Závislost dvou kvantitavních proměnných

Lineární korelace

”Existuje pozitivní korelace mezi výdaji za reklamu a prodejem výrobků”.

”IQ a spotřeba alkoholu nejsou korelované”.

Korelační koeficient r :

popisná míra síly lineárního (přímkového) vztahu mezi dvěma proměnnými.

Korelační koeficient dvou proměnných x a y :
vztahem

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{yx}}{s_y s_x} = r_{yx} \in \langle 0, 1 \rangle$$

s_x a s_y – směrodatné odchylky veličin x resp. y

$$s_x = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad s_y = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

s_{xy} – kovariance

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Interpretace korelačního koeficientu

$r_{xy} = 1$ – *přímá* lineární závislost x a y

$r_{xy} = -1$ – *nepřímá* lineární závislost x a y

$r_{xy} = 0$ – závislost x a y *není lineární*

r_{xy} blízké -1 nebo 1 – *silná lineární závislost* x a y

r_{xy} blízké nule – *slabá lineární závislost* x a y

$r_{xy} > 0$ – *kladně lineárně korelované* veličiny

$r_{xy} < 0$ – *záporně lineárně korelované* veličiny

- Některá úskalí při používání lineárního k. k.

r_{xy} – popisuje sílu lineární závislosti mezi x a y

Používat pouze tehdy, když bodový diagram naznačuje, že data jsou *soustředěna kolem přímky*.

- Korelace není příčinnost

Veličiny mohou být silně korelované, to však neznamená, že je mezi nimi vztah *příčinný*.

Příklad:

V tabulce jsou uvedena data týkající se počtu hodin, které každý z osmi náhodně vybraných studentů (veličina x) věnoval přípravě na test z matematiky, který se měl uskutečnit za 14 dní a počet bodů získaných za test (veličina y).

x	10	15	12	20	8	16	14	22
y	92	81	84	74	85	80	84	80

Silně záporně korelované ($r = -0.779$), neznamená to, že větší počet hodin věnovaný přípravě na test je příčinou horšího výsledku testu.

Dvě veličiny mohou být silně korelované z toho důvodu, že obě jsou vázány s jinými veličinami, tzv. *skryté veličiny*, které jsou příčinou variability veličin, které zkoumáme.

Testy hypotéz o koeficientu korelace

Předpokládejme, X a Y jsou dvě náhodné veličiny

Jestliže $\rho_{xy} = 0 \implies$

$$T = \frac{r_{xy}}{\sqrt{\frac{1-r_{xy}^2}{n-2}}} \sim t(n-2)$$

$$\left(T = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \sqrt{n-2} \right)$$

Test hypotézy pro korelační koeficient s $H_0 : \rho = 0$

- *Testová statistika:* $T = \frac{r_{xy}}{\sqrt{\frac{1-r_{xy}^2}{n-2}}} \sim t(n-2)$

- *Kritické hodnoty H_0 :*

pro oboustranný test: $\pm t_{1-\alpha/2}(n-2)$

pro levostranný test: $-t_{1-\alpha}(n-2)$

pro pravostranný test: $t_{1-\alpha}(n-2)$

Příklad:

V padesátých letech došlo k úniku radioaktivního odpadu ze skládky v Hanfordu do řeky Columbia River. V devíti okrscích níže po proudu řeky bylo zjišťováno vystavení radioaktivitě X . Současně se sledovala úmrtnost na rakovinu Y (úmrtí na 100tisíc obyvatel za rok v letech 1959-1964).

Zjištěné údaje jsou v následující tabulce:

okrsek	1	2	3	4	5	6	7	8	9
X	8.3	6.4	3.4	3.8	2.6	11.6	1.2	2.5	1.6
Y	210	180	130	170	130	210	120	150	140

Poskytují nám údaje dostatek argumentů pro to, abychom udělali na 1% hladině významnosti závěr, že vystavení radioaktivitě a úmrtnost na rakovinu jsou kladně lineárně korelované.