Kap. 2 Opisná štatistika viacrozmerného štatistického súboru

Definícia dátového (štatistického) súboru:

Viacrozmerný (dvojrozmerný) dátový súbor je postupnosť vektorov, v ktorej nezáleží na ich poradí.

Definičný obor – objekty.

Obor hodnôt – vektor znakov.

Povaha znaku: kvalitatívny – nominálny

– ordinálny

kvantitatívny

Príklady.

1. Známky študentov gymnázia z angličtiny a matematiky.
2. Investičná analýza.

Kontingenčná tabuľka:

matica $(f_{i,j})_{i\in I, j\in J}$

$f_{i,j}$ = početnosť vektora (u_i, v_j)

marginálne početnosti

$f_{i,*} = \sum_{j \in J} f_{i,j}$ – početnosť hodnoty u_i v prvej zložke

$f_{*,j} = \sum_{i \in I} f_{i,j}$ – početnosť hodnoty v_j v druhej zložke

Číselné charakteristiky 2-rozmerného dátového súboru – stochastická väzba medzi zložkami

Dátová kovariancia pre dátový súbor $x,y = (x_i,y_j)_{i,j=1,2,\ldots,n}$

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

Dátový korelačný koeficient pre dátový súbor $x,y = (x_i,y_j)_{i,j=1,2,\ldots,n}$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Vlastnosti korelačného koeficientu:

1. $|r_{xy}| \leq 1$
2. $|r_{xy}| = 1 \Rightarrow \exists a, b \in \mathbb{R}, \forall i, y_i = a + bx_i$

Regresná priamka

Určíme konštanty a, b, tak, aby sme minimalizovali

$$SSE = \Sigma(a + bx_i - y_i)^2$$

Riešením systému rovníc, získaných z derivácií podľa a, b, položených = 0, dostaneme

$$a = \overline{y} - b\overline{x}, \qquad b = s_{xy}/s_x^2$$

Regresná priamka má tvar

$y - \overline{y} = (s_{xy}/s_x^2)(x - \overline{x})$ alebo $(y - \overline{y})/s_y = r_{xy}(x - \overline{x})/s_x$

Koeficient determinácie

Ak označíme

$SST = \Sigma(y_i - \overline{y})^2 \ (= ns_y^2)$ Sum of Squares Total

$SSE = \Sigma(y_i - a - bx_i)^2$ Sum of Squares due to Error

$SSR = \Sigma(\overline{y} - a - bx_i)^2$, Sum of Squares due to Regression

potom

$SST = SSE + SSR$ and $0 \leq 1 - SSE/SST = SSR/SST \leq 1$

Číslo

$$R^2 = SSR/SST = r_{xy}^2$$

sa nazýva koeficient determinácie. Čím je bližší k 1, tým tesnejšia je lineárna závislosť medzi zložkami x a y dátového súboru.

Dôkaz vlastností korelačného koeficientu r_{xy}

1.

$0 \leq \Sigma((x_i - \overline{x})/s_x \pm (y_i - \overline{y})/s_y)^2 = \Sigma(x_i - \overline{x})^2/s_x^2 + \Sigma(y_i - \overline{y})^2/s_y^2 \pm 2\Sigma((x_i - \overline{x})(y_i - \overline{y}))/s_x\, s_y = 1 + 1 \pm 2r_{xy}$

$0 \leq 2 + 2r_{xy} \Rightarrow r_{xy} \geq -1$

$0 \leq 2 - 2r_{xy} \Rightarrow r_{xy} \leq 1$

2.

$|r_{xy}| = 1 \Rightarrow 0 = \Sigma((x_i - \overline{x})/s_x + (y_i - \overline{y})/s_y)^2$ alebo

$0 = \Sigma((x_i - \overline{x})/s_x - (y_i - \overline{y})/s_y)^2 \Rightarrow (x_i - \overline{x})/s_x + (y_i - \overline{y})/s_y = 0$ alebo $(x_i - \overline{x})/s_x - (y_i - \overline{y})/s_y = 0$ pre $\forall i \Rightarrow y_i = \overline{y} +$

$(s_y/s_x)\overline{x} - (s_y/s_x)x_i$ alebo $y_i = \overline{y} - (s_y/s_x)\overline{x} + (s_y/s_x)x_i$ pre $\forall i$

Dôkaz rovnosti $SST = SSE + SSR$

$SST = \Sigma(y_i - \overline{y})^2 = \Sigma(y_i - a - bx_i - (\overline{y} - a - bx_i))^2 = \Sigma(y_i - a - bx_i - (\overline{y} - a - bx_i))^2 = \Sigma(y_i - a - bx_i)^2 + \Sigma(\overline{y} - a - bx_i))^2 - 2\Sigma((y_i - a - bx_i)(\overline{y} - a - bx_i)) = SSE + SSR - 2\Sigma((y_i - a - bx_i)(\overline{y} - a - bx_i))$

$\Sigma((y_i - a - bx_i)(\overline{y} - a - bx_i)) = \Sigma((y_i - \overline{y} + b\overline{x} - bx_i)(\overline{y} - \overline{y} + b\overline{x} - bx_i)) = \Sigma((y_i - \overline{y} + b(\overline{x} - x_i))b(\overline{x} - x_i)) = b\Sigma((y_i - \overline{y})(\overline{x} - x_i)) + b^2\Sigma(\overline{x} - x_i)^2 = nb(-s_{xy}) + nb^2s_x^2 = n(-s_{xy}^2/s_x^2 + s_{xy}^2s_x^2/(s_x^2)^2) = 0$

Dôkaz rovnosti $r_{xy}^2 = SSR/SST$

$SSR/SST = 1 - SSE/SST = \Sigma(\overline{y} - a - bx_i)^2/(ns_y^2) = \Sigma(\overline{y} - \overline{y} + b\overline{x} - bx_i))^2/(ns_y^2) = b^2\Sigma(\overline{x} - x_i)^2/(ns_y^2) = (s_{xy}/s_x^2)^2n(s_x^2)/(ns_y^2) = s_{xy}^2/s_x^2s_y^2 = r_{xy}^2$