2026 年度　データサイエンス数学基礎/統計学　（担当：松澤　寛）　レジメ

No.4　2 次元のデータ

散布図と相関

例 10 人の学生の数学と物理の得点

	1	2	3	4	5	6	7	8	9	10
数学	20	70	30	45	30	55	85	50	70	60
物理	15	50	25	50	20	50	80	40	85	75

- 平面上に縦軸と横軸をとり，横軸と縦軸にそれぞれ別の量をとり、データが当てはまるところに点を打って（プロットするという）示したもの.

- ひとつのデータが高いと, もう一つのデータが高い傾向があるとき**正の相関**があるという.
- ひとつのデータが増える, もう一つのデータが小さくなる傾向があるとき**負の相関**があるという.

相関を表す量

- 相関を数値的に評価するとき，**相関係数**が用いられる.
- データを $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ とする.
- 次の量を**共分散**という $s_{xy} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

  $s_{xy} > 0$ のとき, x が大きければ y も大きい
  $s_{xy} < 0$ のとき, x が大きければ y は小さい

  と考えられる.

- 共分散の計算方法

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{x}\cdot\overline{y}) \\ &= \frac{1}{n}\sum_{i=1}^{b} x_i y_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n} x_i - \overline{x}\frac{1}{n}\sum_{i=1}^{n} y_i + \overline{x}\cdot\overline{y} \\ &= \overline{xy} - \overline{y}\cdot\overline{x} - \overline{x}\cdot\overline{y} + \overline{x}\cdot\overline{y} = \overline{xy} - \overline{x}\cdot\overline{y} \end{aligned}$$

- $\sigma_x = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$: $x_1, \cdots, x_n$ の標準偏差
- $\sigma_y = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}$: $y_1, \cdots, y_n$ の標準偏差
- $r = \dfrac{s_{xy}}{\sigma_x \sigma_y}$ を**相関係数**という.

相関係数の性質

- 2 つのデータ $\{x_i\}_i, \{y_i\}_i$ に対して, $z_i = tx_i - y_i$ の分散について考える：

$$\begin{aligned}
0 \leq v_z &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{z})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(tx_i - y_i) - (t\overline{x} - \overline{y})\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \overline{x})t - (y_i - \overline{y})\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \overline{x})^2 t^2 - 2(x_i - \overline{x})(y_i - \overline{y})t + (y_i - \overline{y})^2\} \\
&= t^2 \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 - 2t\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= t^2 v_x - 2ts_{xy} + v_y
\end{aligned}$$

- すべての t に対して, 0 以上より, D を判別式とすると,

$$\begin{aligned}
&D/4 = (s_{xy})^2 - v_x v_y \leq 0 \\
&(s_{xy})^2 \leq v_x v_y \\
&\frac{(s_{xy})^2}{v_x v_y} \leq 1
\end{aligned}$$

- よって, 相関係数 $r = \dfrac{s_{xy}}{\sigma_x \sigma_y}$ は $-1 \leq r \leq 1$ を満たす.
- 特に, 判別式 D が 0 ($|\dfrac{s_{xy}}{\sigma_x \sigma_y}| = 1$) のとき,

$$0 = v_z = t_0^2 v_x - 2t_0 s_{xy} + v_y$$

を満たす t_0 が存在する. 分散の定義から, すべての i について,

$$\begin{aligned}
0 &= z_i - \overline{z} \\
&= (t_0 x_i - y_i) - (t_0 \overline{x} - \overline{y}) \\
y_i - \overline{y} &= t_0(x_i - \overline{x})
\end{aligned}$$

2 つのデータ $\{x_i\}_i, \{y_i\}_i$ は 1 次式の関係にある.

例えば, $\dfrac{s_{xy}}{\sigma_x\sigma_y}=1$, 即ち, $s_{xy}=\sigma_x\sigma_y$ のとき,

$$
\begin{gathered}
t_0^2 v_x - 2t_0 s_{xy} + v_y = 0 \\
t_0^2 v_x - 2t_0 \sigma_x\sigma_y + v_y = 0 \\
(t_0\sigma_x - \sigma_y)^2 = 0 \\
t_0 = \frac{\sigma_y}{\sigma_x}
\end{gathered}
$$

よって,

$$
\begin{aligned}
y_i - \overline{y} &= t_0(x_i - \overline{x}) = \frac{\sigma_y}{\sigma_x}(x_i - \overline{x}) = \frac{\sigma_x\sigma_y}{\sigma_x^2}(x_i - \overline{x}) = \frac{s_{xy}}{v_x}(x_i - \overline{x}) \\
y_i &= \frac{s_{xy}}{v_x}x_i + \overline{y} - \frac{s_{xy}}{v_x}\overline{x}
\end{aligned}
$$

これを, y の x による **回帰直線** という.