2023年11月吉日

データの分析で学んだ公式からの考察
― 回帰直線と相関係数の関係 ―

統計学　数理研究班

研究背景

数学Iで学んだデータの分析の単元で，相関係数rは

$$-1 \leqq r \leqq 1$$

と解説してあるが，証明が与えられていない．

分布図の傾向を最もよく表す1次直線として，回帰直線を導入してるが，理論的な解説がない．

研究目的

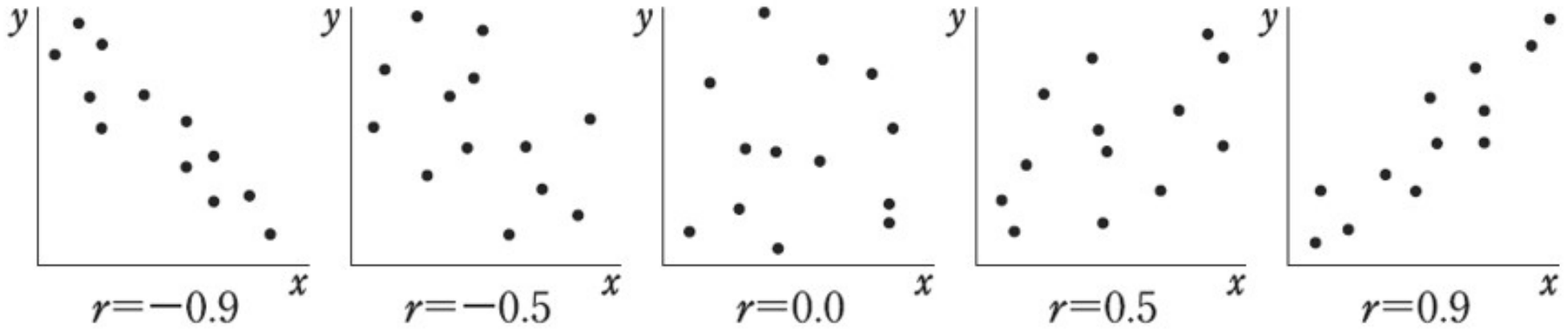
これらの理論的な根拠を学習するとともに，相関係数と回帰直線との関係性について研究を行った．

主に次の3点について研究を行った．

1 $-1 \leqq r \leqq 1$の証明．
2 回帰直線の方程式を求める．
3 相関係数と回帰直線の傾きとの関係を研究．

正の相関，負の相関

2つの変量からなるデータにおいて，一方が増えると他方も増える傾向がみられるとき，2つの変量の間には正の相関があるという．また，一方が増えると他方が減る傾向がみられるとき，2つの変量の間には負の相関があるという．どちらの傾向もみられないときは，2つの変量の間には相関がないという．



相関係数

2つ変量x，yからなるデータとしてn個の値の組

$$(x_1,\ y_1),\ (x_2,\ y_2),\ \cdots,\ (x_n,\ y_n)$$

が得られているとするとき，x_1，x_2，$\cdots$，x_nの平均値を$\overline{x}$，y_1，y_2，$\cdots$，y_nの平均値を$\overline{y}$とおく．

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$
$$\overline{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$$

相関係数rは共分散s_{xy}を，xの標準偏差s_xとyの標準偏差s_yの積$s_x s_y$で割った値として定義される．

$$
\begin{aligned}
r &= \frac{s_{xy}}{s_x s_y} \\
&= \frac{\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}}{\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})\}}\sqrt{\frac{1}{n}\{(y_1-\overline{y})^2+\cdots+(y_n-\overline{y})\}}} \\
&= \frac{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})^2}\sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(y_k-\overline{y})^2}}
\end{aligned}
$$

$-1 \leqq r \leqq 1$の証明

次のtの2次関数を考える$(f(t) \geqq 0)$.

$$
\begin{aligned}
f(t) &= \frac{1}{n}\sum_{k=1}^{n}\{t(x_k-\overline{x})-(y_k-\overline{y})\}^2 \\
&= \frac{t^2}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2 - \frac{2t}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y}) \\
&\qquad\qquad + \frac{1}{n}\sum_{k=1}^{n}(y_k-\overline{y})^2 \\
&= {s_x}^2t^2 - 2s_{xy}t + {s_y}^2
\end{aligned}
$$

$f(t) = {s_x}^2t^2 - 2s_{xy}t + {s_y}^2$ は，$f(t) \geqq 0$ であるから，
係数について

$$D/4 = {s_{xy}}^2 - {s_x}^2{s_y}^2 \leqq 0$$

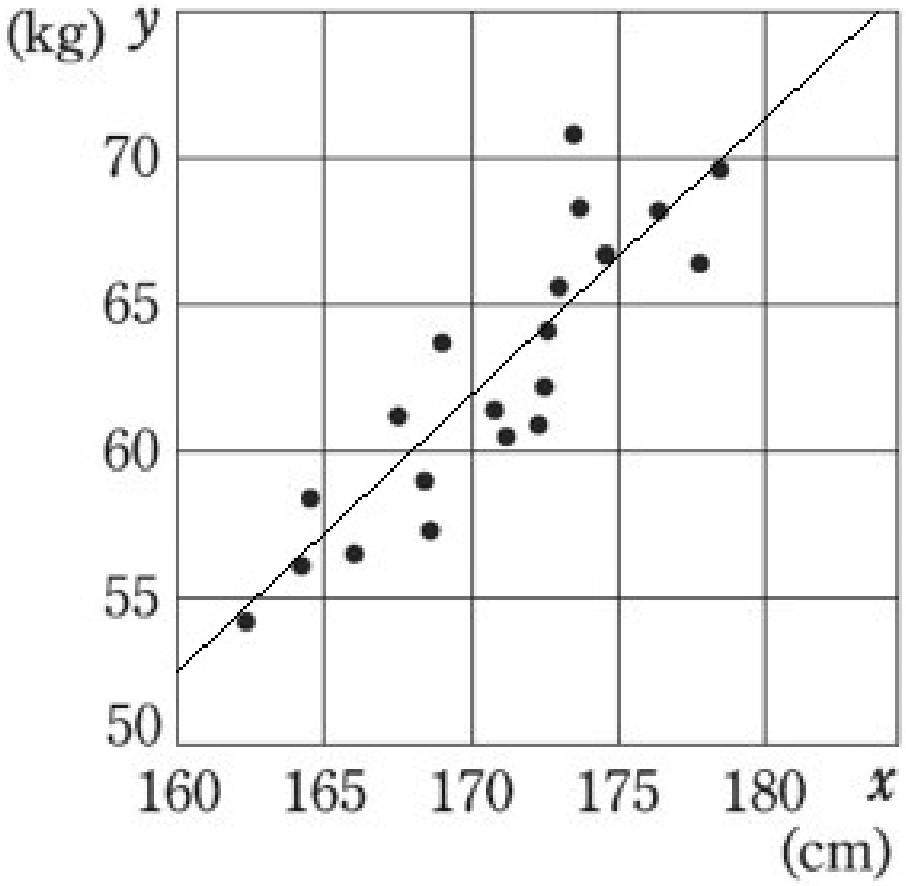
したがって

$$\frac{{s_{xy}}^2}{{s_x}^2{s_y}^2} - 1 \leqq 0 \quad \text{ゆえに} \quad r^2 - 1 \leqq 0$$

これから　$(r+1)(r-1) \leqq 1$　よって　$-1 \leqq r \leqq 1$

回帰直線

回帰直線とは、散布図上のデータに最もよく当てはまるよう引いた直線のことである．右の図はある高校の1年生男子の身長xと体重yの散布図について，引いた直線が回帰直線の例として示したものである．



2つ変量x, yからなるデータとしてn個の値の組

$$(x_1,\ y_1),\ (x_2,\ y_2),\ \cdots,\ (x_n,\ y_n)$$

について，$\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(y_k-ax_k-b)^2$ を最小にする定数a, bを求める．

$$F(a,\ b)=\frac{1}{n}\sum_{k=1}^{n}(y_k-ax_k-b)^2$$

とおく．

Fをa，bについて偏微分すると

$$\begin{aligned}\frac{\partial F}{\partial a} &= -\frac{2}{n}\sum_{k=1}^{n}(y_k - ax_k - b)x_k \\ &= -2\{E(xy) - aE(x^2) - b\overline{x}\} \\ \frac{\partial F}{\partial b} &= -\frac{2}{n}\sum_{k=1}^{n}(y_k - ax_k - b) \\ &= -2(\overline{y} - a\overline{x} - b)\end{aligned}$$

Fが最小となるとき，$\dfrac{\partial F}{\partial a} = \dfrac{\partial F}{\partial b} = 0$ であるから

$$aE(x^2) + b\overline{x} = E(xy), \quad a\overline{x} + b = \overline{y}$$

したがって

$$a = \frac{E(xy) - \overline{x}\,\overline{y}}{E(x^2) - \overline{x}^2} = \frac{\sigma_{xy}}{{\sigma_x}^2}, \quad b = -a\overline{x} + \overline{y}$$

このとき，直線$y = ax + b$，すなわち

$$y - \overline{y} = \frac{\sigma_{xy}}{{\sigma_x}^2}(x - \overline{x})$$

を回帰直線という．

結論

回帰直線

$$y - \overline{y} = \frac{\sigma_{xy}}{{\sigma_x}^2}(x - \overline{x})$$

によって，ある値x_iに対するyの予測値を計算できる．また，相関係数rと回帰直線の傾きaの符号が一致する．