相関係数の定義とコーシー・シュワルツの不等式から導かれる不等式

にしもと のりよし
西元 教善

§1. はじめに

数学Ⅰの「データの分析」で、相関係数を扱う。2つの変数 $x: x_1, x_2, \dots, x_n$ と $y: y_1, y_2, \dots, y_n$ の標準偏差をそれぞれ s_x, s_y とし、 x と y の共分散を s_{xy} としたとき、 x と y の相関係数 r は

$r = \frac{s_{xy}}{s_x s_y}$ で定義されるが、 r の値については

「 $-1 \leq r \leq 1$ であることが知られている」という、証明抜きの扱いで、主眼は、その値から2つの変数 x, y の間の相関関係を判断することである。

不等式の証明は、数学Ⅱで扱うが、 $-1 \leq r \leq 1$ 、すなわち、 $s_{xy}^2 \leq s_x^2 s_y^2$ 、つまり、次の不等式

$$\left\{ \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \right\}^2 \leq \sum_{k=1}^n (x_k - \bar{x})^2 \sum_{k=1}^n (y_k - \bar{y})^2 \quad \dots\dots (*)$$

を、証明する機会はない。

コーシー・シュワルツの不等式は、教科書の本文中で扱われることもなく、問題集・参考書等で不等式 $(a^2 + b^2)(x^2 + y^2) \geq (ax + by)^2$ として扱われる程度であるため、無理からぬことではあるが、探究的な学習の時間等で

$$\left(\sum_{k=1}^n a_k^2 \right) \left(\sum_{k=1}^n b_k^2 \right) \geq \left(\sum_{k=1}^n a_k b_k \right)^2$$

(等号が成り立つのは、 $b_k = ca_k$ ($k=1, 2, \dots, n$) を満たす実数 c が存在するとき) $\dots\dots (**)$ を扱うこともできるだろう。

すると、(**)から(*)が示され、 $-1 \leq r \leq 1$ であることや、等号成立条件が、正の定数 c に対して

$$r = \pm 1 \iff y_k - \bar{y} = \pm c(x_k - \bar{x})$$

($k=1, 2, \dots, n$) (複号同順)

を満たすときであること、つまり、点 (x_k, y_k) ($k=1, 2, \dots, n$) が、直線 $y - \bar{y} = \pm c(x - \bar{x})$ (複号同順) 上にあることがわかるので、相関係数の値と散布図の関係が、よりわかるようになる。

本稿では、 $s_{xy}^2 \leq s_x^2 s_y^2$ とコーシー・シュワルツの不等式から得られる不等式について、考察する。

§2. $s_{xy}^2, s_x^2 s_y^2$ を $\bar{x}, \bar{x}^2, \bar{y}, \bar{y}^2, \bar{xy}$ で表す $s_x^2 = \overline{x^2} - (\bar{x})^2$ であることは、教科書でも扱っている(ことが多い)が、 $s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$ は、教科書ではあまり見かけない。問題集では、当然知っているものとして、使っている場合もある。証明は、次の通りである。

証明

$$\begin{aligned} s_{xy} &= \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \\ &= \frac{1}{n} \sum_{k=1}^n (x_k y_k - \bar{y} x_k - \bar{x} y_k + \bar{x} \cdot \bar{y}) \\ &= \frac{1}{n} \sum_{k=1}^n x_k y_k - \bar{y} \cdot \frac{1}{n} \sum_{k=1}^n x_k \\ &\quad - \bar{x} \cdot \frac{1}{n} \sum_{k=1}^n y_k + \frac{1}{n} \sum_{k=1}^n \bar{x} \cdot \bar{y} \\ &= \overline{xy} - \bar{y} \cdot \bar{x} - \bar{x} \cdot \bar{y} + \frac{1}{n} \cdot n \bar{x} \cdot \bar{y} \\ &= \overline{xy} - \bar{x} \cdot \bar{y} \end{aligned}$$

証明終

よって、 $s_x^2 = \overline{x^2} - (\bar{x})^2, s_y^2 = \overline{y^2} - (\bar{y})^2,$

$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$ であるから

$$\begin{aligned} s_{xy}^2 &= (\overline{xy} - \bar{x} \cdot \bar{y})^2 = (\overline{xy})^2 - 2\overline{xy} \cdot \bar{x} \cdot \bar{y} + (\bar{x})^2 (\bar{y})^2 \\ s_x^2 s_y^2 &= \{ \overline{x^2} - (\bar{x})^2 \} \{ \overline{y^2} - (\bar{y})^2 \} \\ &= \overline{x^2} \cdot \overline{y^2} - \overline{x^2} (\bar{y})^2 - (\bar{x})^2 \overline{y^2} + (\bar{x})^2 (\bar{y})^2 \end{aligned}$$

である。

§3. $\bar{x}$, $\bar{x}^2$, $\bar{y}$, $\bar{y}^2$, $\overline{xy}$ に関する不等式

§2で考察したように, s_{xy}^2 は $\bar{x}$, $\bar{y}$, $\overline{xy}$ で表され, $s_x^2 s_y^2$ は $\bar{x}$, $\bar{y}$, $\bar{x}^2$, $\bar{y}^2$ で表されるので, $s_{xy}^2 \leq s_x^2 s_y^2$, すなわち, $s_x^2 s_y^2 \geq s_{xy}^2$ に, §2の結果を代入すると

$$\begin{aligned} & \bar{x}^2 \cdot \bar{y}^2 - \bar{x}^2 (\bar{y})^2 - (\bar{x})^2 \bar{y}^2 + (\bar{x})^2 (\bar{y})^2 \\ & \geq (\overline{xy})^2 - 2\overline{xy} \cdot \bar{x} \cdot \bar{y} + (\bar{x})^2 (\bar{y})^2 \end{aligned}$$

よって, $\bar{x}$, $\bar{x}^2$, $\bar{y}$, $\bar{y}^2$, $\overline{xy}$ に関する不等式

$$\bar{x}^2 \cdot \bar{y}^2 - (\overline{xy})^2 \geq \bar{x}^2 (\bar{y})^2 - 2\overline{xy} \cdot \bar{x} \cdot \bar{y} + (\bar{x})^2 \bar{y}^2 \quad \dots\dots ①$$

が成り立つ。

§4. (2乗の和の積) $\geq$ (積の和の2乗) + (0以上の式) を考える

(※)より, $\left(\sum_{k=1}^n x_k^2\right) \left(\sum_{k=1}^n y_k^2\right) \geq \left(\sum_{k=1}^n x_k y_k\right)^2$ であるから,

$$\frac{1}{n} \sum_{k=1}^n x_k^2 \cdot \frac{1}{n} \sum_{k=1}^n y_k^2 \geq \left(\frac{1}{n} \sum_{k=1}^n x_k y_k\right)^2, \text{つまり,}$$

$$\bar{x}^2 \cdot \bar{y}^2 \geq (\overline{xy})^2 \text{ が成り立つ。}$$

よって, $\sqrt{x^2} \sqrt{y^2} \geq |xy|$ であるから

$$-\sqrt{x^2} \cdot \sqrt{y^2} \leq -|xy| \leq -xy$$

が成り立つ。

コーシー・シュワルツの不等式

$$\left(\sum_{k=1}^n x_k^2\right) \left(\sum_{k=1}^n y_k^2\right) \geq \left(\sum_{k=1}^n x_k y_k\right)^2 \text{ では, } x_k, y_k$$

($k=1, 2, \dots, n$) は任意の実数で, 特に条件はないが, あえて, $x_k > 0, y_k > 0$ ($k=1, 2, \dots, n$) とすると, $\bar{x}, \bar{y}, \overline{xy}$ はすべて正の数になるから

$$-\sqrt{x^2} \sqrt{y^2} \leq -|xy| = -\overline{xy} < 0$$

$$\text{より } -2\sqrt{x^2} \sqrt{y^2} \bar{x} \cdot \bar{y} \leq -2\overline{xy} \cdot \bar{x} \cdot \bar{y} < 0$$

したがって

$$\begin{aligned} & \bar{x}^2 (\bar{y})^2 - 2\overline{xy} \cdot \bar{x} \cdot \bar{y} + (\bar{x})^2 \bar{y}^2 \\ & \geq (\sqrt{x^2})^2 (\bar{y})^2 - 2\sqrt{x^2} \cdot \sqrt{y^2} \bar{x} \cdot \bar{y} + (\sqrt{y^2})^2 (\bar{x})^2 \\ & = (\sqrt{x^2} \bar{y} - \sqrt{y^2} \bar{x})^2 \quad \dots\dots ② \end{aligned}$$

よって, ①, ②より

$$\bar{x}^2 \cdot \bar{y}^2 - (\overline{xy})^2 \geq (\sqrt{x^2} \bar{y} - \sqrt{y^2} \bar{x})^2 \quad \dots\dots ③$$

(③の左辺) $= \bar{x}^2 \cdot \bar{y}^2 - (\overline{xy})^2$

$$\begin{aligned} & = \frac{1}{n} \sum_{k=1}^n x_k^2 \cdot \frac{1}{n} \sum_{k=1}^n y_k^2 - \left(\frac{1}{n} \sum_{k=1}^n x_k y_k\right)^2 \\ & = \frac{1}{n^2} \left\{ \left(\sum_{k=1}^n x_k^2\right) \left(\sum_{k=1}^n y_k^2\right) - \left(\sum_{k=1}^n x_k y_k\right)^2 \right\} \end{aligned}$$

(③の右辺) $= (\sqrt{x^2} \bar{y} - \sqrt{y^2} \bar{x})^2$

$$\begin{aligned} & = \left(\sqrt{\frac{1}{n} \sum_{k=1}^n x_k^2} \cdot \frac{1}{n} \sum_{k=1}^n y_k \right. \\ & \quad \left. - \sqrt{\frac{1}{n} \sum_{k=1}^n y_k^2} \cdot \frac{1}{n} \sum_{k=1}^n x_k \right)^2 \\ & = \frac{1}{n^3} \left(\sqrt{\sum_{k=1}^n x_k^2} \sum_{k=1}^n y_k - \sqrt{\sum_{k=1}^n y_k^2} \sum_{k=1}^n x_k \right)^2 \end{aligned}$$

であるから, ③より

$$\begin{aligned} & \frac{1}{n^2} \left\{ \left(\sum_{k=1}^n x_k^2\right) \left(\sum_{k=1}^n y_k^2\right) - \left(\sum_{k=1}^n x_k y_k\right)^2 \right\} \\ & \geq \frac{1}{n^3} \left(\sqrt{\sum_{k=1}^n x_k^2} \sum_{k=1}^n y_k - \sqrt{\sum_{k=1}^n y_k^2} \sum_{k=1}^n x_k \right)^2 \end{aligned}$$

すなわち

$$\begin{aligned} & \left(\sum_{k=1}^n x_k^2\right) \left(\sum_{k=1}^n y_k^2\right) - \left(\sum_{k=1}^n x_k y_k\right)^2 \\ & \geq \frac{1}{n} \left(\sqrt{\sum_{k=1}^n x_k^2} \sum_{k=1}^n y_k - \sqrt{\sum_{k=1}^n y_k^2} \sum_{k=1}^n x_k \right)^2 \end{aligned}$$

よって

$$\begin{aligned} & \left(\sum_{k=1}^n x_k^2\right) \left(\sum_{k=1}^n y_k^2\right) \\ & \geq \left(\sum_{k=1}^n x_k y_k\right)^2 + \frac{1}{n} \left(\sqrt{\sum_{k=1}^n x_k^2} \sum_{k=1}^n y_k - \sqrt{\sum_{k=1}^n y_k^2} \sum_{k=1}^n x_k \right)^2 \end{aligned}$$

等号が成り立つのは, 定数 c に対して, $y_k = cx_k$ ($k=1, 2, \dots, n$) を満たすときである。

§5. コーシー・シュワルツの不等式の系

以上から, コーシー・シュワルツの不等式の【系】として, 次が成り立つ。

【系】

$x_k > 0, y_k > 0$ ($k=1, 2, \dots, n$) のとき

$$\begin{aligned} & \left(\sum_{k=1}^n x_k^2\right) \left(\sum_{k=1}^n y_k^2\right) \\ & \geq \left(\sum_{k=1}^n x_k y_k\right)^2 + \frac{1}{n} \left(\sqrt{\sum_{k=1}^n x_k^2} \sum_{k=1}^n y_k - \sqrt{\sum_{k=1}^n y_k^2} \sum_{k=1}^n x_k \right)^2 \end{aligned}$$

等号が成り立つのは, 定数 c に対して $y_k = cx_k$ ($k=1, 2, \dots, n$) を満たすときである。

§6. 系を使う証明問題

【問題】

$x_k > 0, y_k > 0$ ($k=1, 2, \dots, n$) とする。

$\sum_{k=1}^n x_k^2 = \sum_{k=1}^n y_k^2 = 1$ のとき

$$\left(\sum_{k=1}^n x_k y_k\right)^2 \leq 1 - \frac{1}{n} \left\{ \sum_{k=1}^n (x_k - y_k) \right\}^2$$

が成り立つことを証明せよ。

【証明】 $x_k > 0, y_k > 0$ ($k=1, 2, \dots, n$) かつ

$\sum_{k=1}^n x_k^2 = \sum_{k=1}^n y_k^2 = 1$ より, §5の【系】を用いて

$$\begin{aligned} 1 \cdot 1 &\geq \left(\sum_{k=1}^n x_k y_k \right)^2 + \frac{1}{n} \left(\sqrt{1} \sum_{k=1}^n y_k - \sqrt{1} \sum_{k=1}^n x_k \right)^2 \\ &= \left(\sum_{k=1}^n x_k y_k \right)^2 + \frac{1}{n} \left(\sum_{k=1}^n y_k - \sum_{k=1}^n x_k \right)^2 \\ &= \left(\sum_{k=1}^n x_k y_k \right)^2 + \frac{1}{n} \left\{ - \sum_{k=1}^n (x_k - y_k) \right\}^2 \end{aligned}$$

よって $\left(\sum_{k=1}^n x_k y_k \right)^2 \leq 1 - \frac{1}{n} \left\{ \sum_{k=1}^n (x_k - y_k) \right\}^2$ 証明終

§7. まとめ

相関係数 r のとりうる値の範囲は, 証明抜きで $-1 \leq r \leq 1$ として扱ってある。『データの分析』履修時は, 「不等式の証明」を扱う前であるから致し方ないにしても, 「不等式の証明」の中で触れてほしいという思いが筆者にはある。また, $-1 \leq r \leq 1$ となる理由を知りたいと思う生徒も多いのではないだろうか。

「 $-1 \leq r \leq 1$ であることが知られている」という現状の扱いについて, 生徒たちに無条件に受け入れさせるだけでよいのであろうか。興味・関心の芽を摘むことがないように, また, 理解の深化を図るためにも, 相関係数 r のとりうる値の範囲について扱うように配慮するべきであると思う。

$-1 \leq r \leq 1$, つまり, $s_{xy}^2 \leq s_x^2 s_y^2$ は, コーシー・シュワルツの不等式 $\left(\sum_{k=1}^n a_k^2 \right) \left(\sum_{k=1}^n b_k^2 \right) \geq \left(\sum_{k=1}^n a_k b_k \right)^2$ において, $a_k = \frac{1}{\sqrt{n}}(x_k - \bar{x})$, $b_k = \frac{1}{\sqrt{n}}(y_k - \bar{y})$ とお

けば導けるが, $x_k > 0, y_k > 0$ ($k=1, 2, \dots, n$) という条件をつけ加え, 『データの分析』でよく使う性質 $s_x^2 = \overline{x^2} - (\bar{x})^2$, $s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$ を活用して

$$\begin{aligned} \left(\sum_{k=1}^n x_k^2 \right) \left(\sum_{k=1}^n y_k^2 \right) &\geq \left(\sum_{k=1}^n x_k y_k \right)^2 \\ &+ \frac{1}{n} \left(\sqrt{\sum_{k=1}^n x_k^2} \sum_{k=1}^n y_k - \sqrt{\sum_{k=1}^n y_k^2} \sum_{k=1}^n x_k \right)^2 \end{aligned}$$

という不等式を導出した次第である。

(山口県立徳山高等学校)