2014 年 11 月 18 日

各種平均が満たすべき条件

新潟工科大学 情報電子工学科 竹野茂治

1 はじめに

確率・統計では、2 次元データ (x_k, y_k) で相関係数や回帰直線などを計算するのに、次のような平均値を使用することがある。

$$\begin{aligned} &\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k\,,\ \ \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k\,,\ \ \overline{x^2} = \frac{1}{n}\sum_{k=1}^{n} x_k^2\,,\ \ \overline{y^2} = \frac{1}{n}\sum_{k=1}^{n} y_k^2\,, \\ &\overline{xy} = \frac{1}{n}\sum_{k=1}^{n} x_k y_k \end{aligned} \tag{1}$$

分散 s_x^2, s_y^2、共分散 s_{xy}、相関係数 r、回帰直線などは、元のデータの個々の値やデータの個数 n を知らなくても、上記 (1) の値だけでいずれも求めることができる。

$$s_x^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \overline{x^2} - (\bar{x})^2 \tag{2}$$

$$s_y^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2 = \overline{y^2} - (\bar{y})^2 \tag{3}$$

$$s_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}) = \overline{xy} - \bar{x}\,\bar{y} \tag{4}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}} \tag{5}$$

なお、(2) から (5) の式は、いずれも最初の式が定義であり、次の式がそれを展開して (1) の値で計算できる形に変形したものである。

これにより、例えば、「$\bar{x} = 3$, $\bar{y} = 4$, $\overline{x^2} = 21$, $\overline{y^2} = 19$, $\overline{xy} = 16$ のとき、相関係数 r を求めよ」のような問題を作ることができる。

ただ、その場合例えば「$\bar{x} = 3$, $\overline{x^2} = 5$」のような値でもいいかというと、それはだめで、実際のデータではこのような組は起こりえない。

そこで本稿では、(1) の 5 種類の平均に対して成り立つ関係を調べ、どのような値であればそれらが実際のデータの平均になりうるか、という条件を求めることを目標とする。そういう条件が得られれば、それを満たすような平均は心配せずに問題として出せることになる。

なお本稿は、学生にはあまり意味はなく、試験問題、演習問題を作成する側に関係する内容だが、私自身の備忘録もかねてまとめておくことにする。

2 シュワルツの不等式

今回の評価では、次のシュワルツの不等式を利用する。

定理 1. 実数 $x_k,\, y_k\ (k = 1, 2, \ldots n)$ に対して、次の不等式が成り立つ。

$$\left(\sum_{k=1}^{n} x_k y_k\right)^2 \leq \left(\sum_{k=1}^{n} x_k^2\right)\left(\sum_{k=1}^{n} y_k^2\right) \tag{6}$$

等号が成立するのは、$x_1 : y_1 = x_2 : y_2 = \cdots = x_n : y_n$ の場合である。

n 次元ベクトル $\overrightarrow{x} = (x_1, x_2, \ldots, x_n),\ \overrightarrow{y} = (y_1, y_2, \ldots, y_n)$ の内積 $\overrightarrow{x}$・$\overrightarrow{y}$、および長さ $|\overrightarrow{x}|$ は

$$\begin{aligned} \overrightarrow{x}\cdot\overrightarrow{y} &= \sum_{k=1}^{n} x_k y_k = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n, \\ |\overrightarrow{x}| &= \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \end{aligned}$$

と定義されるので、これを用いれば不等式 (6) は簡単に

$$(\overrightarrow{x}\cdot\overrightarrow{y})^2 \leq |\overrightarrow{x}|^2 |\overrightarrow{y}|^2 \quad (\text{または } |\overrightarrow{x}\cdot\overrightarrow{y}| \leq |\overrightarrow{x}||\overrightarrow{y}|)$$

と書くことができるし、等号が成立する条件も、$\overrightarrow{x} \,/\!/\, \overrightarrow{y}$ と書き表すことができる。

3 平均が満たすべき条件

本節では、各平均値に以下のように名前をつけ、それらが満たすべき条件を見ていくことにする:

$$\bar{x} = X_1,\ \bar{y} = Y_1,\ \overline{x^2} = X_2,\ \overline{y^2} = Y_2,\ \overline{xy} = Z \tag{7}$$

まず、(2), (3) の定義よりこれらは非負の値なので、

$$\overline{x^2} \geq (\bar{x})^2, \quad \overline{y^2} \geq (\bar{y})^2,$$

すなわち

$$X_2 \geq X_1^2, \quad Y_2 \geq Y_1^2 \tag{8}$$

の関係が成り立つ。これらは、以下のようにシュワルツの不等式から導くこともできる:

$$\begin{aligned}
(\bar{x})^2 &= \left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2 = \frac{1}{n^2}(x_1 + x_2 + \cdots + x_n)^2 \\
&= \frac{1}{n^2}(\overrightarrow{x} \cdot \overrightarrow{n})^2 \quad (\overrightarrow{n} = (1, 1, \ldots, 1)) \\
&\leq \frac{1}{n^2}|\overrightarrow{x}|^2|\overrightarrow{n}|^2 = \frac{1}{n^2}|\overrightarrow{x}|^2 n = \frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2) = \overline{x^2}
\end{aligned}$$

同じように、シュワルツの不等式より、

$$\begin{aligned}
(\overline{xy})^2 &= \frac{1}{n^2}(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^2 = \frac{1}{n^2}(\overrightarrow{x} \cdot \overrightarrow{y})^2 \\
&\leq \frac{1}{n^2}|\overrightarrow{x}|^2|\overrightarrow{y}|^2 = \frac{1}{n^2}(x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2) = \overline{x^2}\,\overline{y^2}
\end{aligned}$$

となるので、

$$Z^2 \leq X_2Y_2 \tag{9}$$

の関係が成り立つことがわかる。

さらに、(4) の s_{xy} の定義式にシュワルツの不等式を用いると、

$$\begin{aligned}(s_{xy})^2 &= \frac{1}{n^2}\{(x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})\}^2 \\ &= \frac{1}{n^2}\{(\overrightarrow{x}-\bar{x}\,\overrightarrow{n})\cdot(\overrightarrow{y}-\bar{y}\,\overrightarrow{n})\}^2 \le \frac{1}{n^2}|\overrightarrow{x}-\bar{x}\,\overrightarrow{n}|^2|\overrightarrow{y}-\bar{y}\,\overrightarrow{n}|^2 \\ &= \frac{1}{n^2}\left(\sum_{k=1}^{n}(x_k-\bar{x})^2\right)\left(\sum_{k=1}^{n}(y_k-\bar{y})^2\right) = s_x^2 s_y^2 \end{aligned} \tag{10}$$

となるので、これに (2), (3), (4) の右辺の式を代入して平均の式に置き換えると

$$(\overline{xy}-\bar{x}\,\bar{y})^2 \le \{\overline{x^2}-(\bar{x})^2\}\{\overline{y^2}-(\bar{y})^2\}$$

となり、よって

$$(Z-X_1Y_1)^2 \le (X_2-X_1^2)(Y_2-Y_1^2) \tag{11}$$

の関係が成り立つことになる。

なお、この (10) の関係は、相関係数 r が $-1 \le r \le 1$ を満たすことを意味している。

4 条件を満たすデータの存在

前節で、実際のデータに対しては X_1, X_2, Y_1, Y_2, Z が不等式 (8), (9), (11) を満たすことを見たが、逆に X_1, X_2, Y_1, Y_2, Z がこれらの不等式を満たす実数であるとき、それらが (7) のような平均値になるようなデータ (x_k, y_k) が存在するかどうかを考えてみることにする。

まず、$X_2 = 0$ の場合は、(8) より $X_1 = 0$、(9) より $Z = 0$ となるので、(11) の両辺も 0 であり、よって、この場合は $n = 2$ として、$x_1 = x_2 = 0$ とすれば X_1, X_2 は OK で、y_1, y_2 は

$$y_1 + y_2 = 2Y_1, \quad y_1^2 + y_2^2 = 2Y_2$$

を満たさなければならないが、

$$y_1 y_2 = \frac{(y_1+y_2)^2-(y_1^2+y_2^2)}{2} = \frac{4Y_1^2-2Y_2}{2} = 2Y_1^2 - Y_2$$

となるから、y_1, y_2 は 2 次方程式

$$\lambda^2 - 2Y_1\lambda + 2Y_1^2 - Y_2 = 0$$

の 2 解として求まる。実際、この 2 次方程式の判別式 D は、(8) より

$$\frac{D}{4} = Y_1^2 - (2Y_1^2 - Y_2) = Y_2 - Y_1^2 \geq 0$$

となるので実数解を持つ。

$Y_2 = 0$ の場合も同様なので、以後、$X_2 > 0, Y_2 > 0$ の場合を考える。

今、新たに

$$\hat{X}_1 = \frac{X_1}{\sqrt{X_2}}, \quad \hat{Y}_1 = \frac{Y_1}{\sqrt{Y_2}}, \quad \hat{Z} = \frac{Z}{\sqrt{X_2 Y_2}}$$

と書くこととすると、条件 (8), (9), (11) は、

$$|\hat{X}_1| \leq 1, \quad |\hat{Y}_1| \leq 1, \quad |\hat{Z}| \leq 1 \tag{12}$$
$$(\hat{Z} - \hat{X}_1\hat{Y}_1)^2 \leq \{1 - (\hat{X}_1)^2\}\{1 - (\hat{Y}_1)^2\} \tag{13}$$

となる。

(7) を満たすデータに対しても新たに記号を導入し、

$$\overrightarrow{\hat{x}} = \frac{\overrightarrow{x}}{\sqrt{nX_2}}, \quad \overrightarrow{\hat{y}} = \frac{\overrightarrow{y}}{\sqrt{nY_2}}, \quad \overrightarrow{\hat{n}} = \frac{\overrightarrow{n}}{\sqrt{n}}$$

とすると、

$$\begin{aligned} \overrightarrow{\hat{x}} \cdot \overrightarrow{\hat{n}} &= \frac{\overrightarrow{x} \cdot \overrightarrow{n}}{n\sqrt{X_2}} = \frac{x_1 + x_2 + \cdots + x_n}{n\sqrt{X_2}} = \frac{\bar{x}}{\sqrt{X_2}} \\ \overrightarrow{\hat{y}} \cdot \overrightarrow{\hat{n}} &= \frac{\overrightarrow{y} \cdot \overrightarrow{n}}{n\sqrt{Y_2}} = \frac{y_1 + y_2 + \cdots + y_n}{n\sqrt{Y_2}} = \frac{\bar{y}}{\sqrt{Y_2}} \end{aligned}$$

より $\bar{x} = X_1, \bar{y} = Y_1$ は

$$\overrightarrow{\hat{x}} \cdot \overrightarrow{\hat{n}} = \hat{X}_1, \quad \overrightarrow{\hat{y}} \cdot \overrightarrow{\hat{n}} = \hat{Y}_1 \tag{14}$$

と書き換えられる。また、

$$|\overrightarrow{\hat{x}}|^2 = \frac{|\overrightarrow{x}|^2}{nX_2} = \frac{\overline{x^2}}{X_2}, \quad |\overrightarrow{\hat{y}}|^2 = \frac{|\overrightarrow{y}|^2}{nY_2} = \frac{\overline{y^2}}{Y_2}$$

より、$\overline{x^2} = X_2$, $\overline{y^2} = Y_2$ は、

$$|\overrightarrow{\hat{x}}| = 1, \quad |\overrightarrow{\hat{y}}| = 1 \tag{15}$$

に置き換わる。そして、

$$\overrightarrow{\hat{x}} \cdot \overrightarrow{\hat{y}} = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{n\sqrt{X_2 Y_2}} = \frac{\overline{xy}}{\sqrt{X_2 Y_2}}$$

より、$\overline{xy} = Z$ は、

$$\overrightarrow{\hat{x}} \cdot \overrightarrow{\hat{y}} = \hat{Z} \tag{16}$$

に置き換わる。

つまり、(12), (13) を満たす $\hat{X}_1$, $\hat{Y}_1$, $\hat{Z}$ に対して、(14), (15), (16) を満たすベクトル $\overrightarrow{\hat{x}}$, $\overrightarrow{\hat{y}}$ が存在することを示せばよいことになる。

実際これは、$n = 3$ の 3 次元で解を求めることができる。(15) より $\overrightarrow{\hat{x}}$, $\overrightarrow{\hat{y}}$ は単位ベクトルで、$\overrightarrow{\hat{n}}$ も単位ベクトルであることに注意する。

今、$\overrightarrow{\hat{x}}$ と $\overrightarrow{\hat{n}}$ のなす角を α、$\overrightarrow{\hat{y}}$ と $\overrightarrow{\hat{n}}$ のなす角を β、$\overrightarrow{\hat{x}}$ と $\overrightarrow{\hat{y}}$ のなす角を γ とすると、(14), (16) は

$$\cos\alpha = \hat{X}_1, \quad \cos\beta = \hat{Y}_1, \quad \cos\gamma = Z_1' \tag{17}$$

を意味する。

(12) より、(17) を満たす α, β, γ は一意に存在する。(17) を (13) に代入すると、

$$|\cos\gamma - \cos\alpha\cos\beta| \leq \sqrt{1-\cos^2\alpha}\sqrt{1-\cos^2\beta} = \sin\alpha\sin\beta$$

となり、これは加法定理により

$$\cos(\alpha+\beta) \leq \cos\gamma \leq \cos(\alpha-\beta) \tag{18}$$

と書き換えられる。

すなわち、(17) により一意に決定する $[0,\pi]$ の範囲の角 α, β, γ が (18) を満たすとき、それらがなす角となるような単位ベクトル $\overrightarrow{\hat{x}}, \overrightarrow{\hat{y}}$ が存在することを確認することが目標となる。

$\overrightarrow{\hat{x}}$ と $\overrightarrow{\hat{n}}$ はなす角が α であるが、そのようなベクトル全体の終点は単位球上で一つの円 C_1 を描く (図 1)。同様に $\overrightarrow{\hat{n}}$ とのなす角が β であるようなベクトルの終点も単位球上の円 C_2 となる。よって、それらの円 C_1, C_2 上にそれ

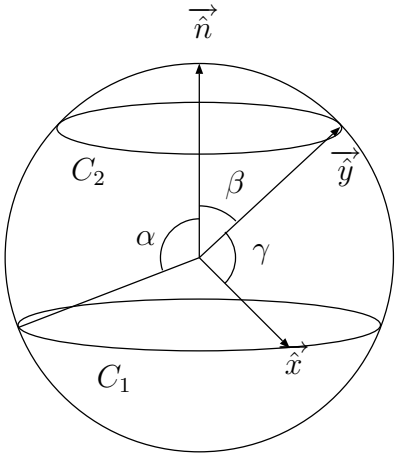

図 1: 円 C_1, C_2

ぞれ終点があるベクトル $\overrightarrow{\hat{x}}, \overrightarrow{\hat{y}}$ のなす角が、(18) を満たす γ となりうるかを考えればよいことになる。

今、$\alpha \geq \beta$ と仮定する ($\alpha < \beta$ の場合は、単に X_i と Y_i を入れ換えて考えればよい)。不等式 (18) は、$\alpha+\beta \leq \pi$ の場合は

$$\alpha+\beta \geq \gamma \geq \alpha-\beta \tag{19}$$

を意味し、$\alpha+\beta > \pi$ の場合は

$$2\pi-(\alpha+\beta) \geq \gamma \geq \alpha-\beta \tag{20}$$

を意味することに注意する。

まず、$\alpha+\beta \leq \pi$ の場合、図 1 からもわかるが、$\overrightarrow{\hat{x}}$ と $\overrightarrow{\hat{y}}$ とのなす角が最小となるのは $\overrightarrow{\hat{x}}$, $\overrightarrow{\hat{y}}$, $\overrightarrow{\hat{n}}$ が同一平面にあるときで、その角は $(\alpha-\beta)$、なす角が最大となるのもやはりこれらのベクトルが同一平面にあるときで、その角は $(\alpha+\beta)$ となる。それ以外の場合は、その 2 つのベクトルのなす角はこの最小角と最大角の間を連続的に変化するので、結局 (19) を満たすすべての γ に対し、それがなす角となるような $\overrightarrow{\hat{x}}$, $\overrightarrow{\hat{y}}$ の配置が存在することがわかる。

$\alpha+\beta > \pi$ の場合も、図 2 より、この場合のなす角の最小値は $(\alpha-\beta)$、最大値は $2\pi-(\alpha+\beta)$ であることがわかり、$\overrightarrow{\hat{x}}$, $\overrightarrow{\hat{y}}$ のなす角はその間を連続的に変化するので、(20) を満たすすべての γ に対し、それがなす角となるような $\overrightarrow{\hat{x}}$, $\overrightarrow{\hat{y}}$ の配置が必ず存在する。

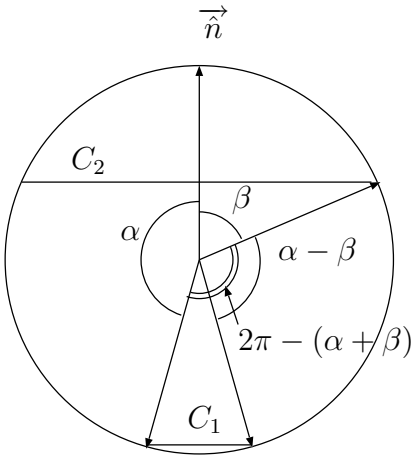

図 2: $\alpha+\beta > \pi$ の場合

これで、(12), (13) を満たす $\hat{X}_1, \hat{Y}_1, \hat{Z}$ に対して、(14), (15), (16) を満たすベクトル $\overrightarrow{\hat{x}}$, $\overrightarrow{\hat{y}}$ が存在することを示すことができたことになり、よって、(8), (9), (11) を満たす X_1, X_2, Y_1, Y_2, Z に対し、それらが (7) のような平均値となるようなデータ (x_k, y_k) が実際に存在することが示されたことになる。

5 最後に

本稿では、2 次元データの統計的平均の値が、実際のデータとして起こりうるための条件を考察した。最後は、計算というよりも幾何学的な存在証明に

なったので、むしろわかりにくいかもしれないが、存在することの保証は正しく得られていると思う。

これで、不等式 (8), (9), (11) を満たしているかだけ確認すれば良いことがわかったので、今後は安心して「適当な」値を出せることになる。