相関とは

森島（もりしま）　充（みつる）

「相関とは要するに何だろう？」と思いながら定義式を見ていたら「cos だ」と気がつきました。これを足掛かりに，相関の図形的意味について考えます。

§1. 定義

2 つのデータ，

$x_1,\ x_2,\ x_3,\ \cdots,\ x_n \qquad y_1,\ y_2,\ y_3,\ \cdots,\ y_n$

に対して，n 次元空間の点，

$\mathrm{X}(x_1,\ x_2,\ x_3,\ \cdots,\ x_n)\quad \mathrm{Y}(y_1,\ y_2,\ y_3,\ \cdots,\ y_n)$

$\mathrm{P}(\overline{x},\ \overline{x},\ \overline{x},\ \cdots,\ \overline{x})\quad \mathrm{Q}(\overline{y},\ \overline{y},\ \overline{y},\ \cdots,\ \overline{y})$

を考えます。

このとき，$\overrightarrow{\mathrm{PX}}\neq\vec{0}$，$\overrightarrow{\mathrm{QY}}\neq\vec{0}$ であれば，相関係数 r は，

$$r=\frac{\overrightarrow{\mathrm{PX}}\cdot\overrightarrow{\mathrm{QY}}}{|\overrightarrow{\mathrm{PX}}||\overrightarrow{\mathrm{QY}}|}$$

と書けます。

すなわち，$\overrightarrow{\mathrm{PX}}$ と $\overrightarrow{\mathrm{QY}}$ の n 次元空間における「なす角」を θ とすると，

$$r=\cos\theta$$

です。これより，$-1\leqq r\leqq 1$ であり，

$r=1 \iff \overrightarrow{\mathrm{PX}}/\!/\overrightarrow{\mathrm{QY}}$（同じ向き）

$r=-1 \iff \overrightarrow{\mathrm{PX}}/\!/\overrightarrow{\mathrm{QY}}$（逆向き）

$r=0 \iff \overrightarrow{\mathrm{PX}}\perp\overrightarrow{\mathrm{QY}}$

となります。

§2. n 次元空間における図形的意味

原点を O，点(1，1，1，…，1)を E とすると，$\overrightarrow{\mathrm{PX}}=(x_1-\overline{x},\ x_2-\overline{x},\ x_3-\overline{x},\ \cdots,\ x_n-\overline{x})$ より，

$$\begin{aligned}\overrightarrow{\mathrm{OP}}\cdot\overrightarrow{\mathrm{PX}}&=\overline{x}\,\overrightarrow{\mathrm{OE}}\cdot\overrightarrow{\mathrm{PX}}\\&=\overline{x}\left\{\sum_{j=1}^{n}(x_j-\overline{x})\right\}\\&=0\end{aligned}$$

です。したがって $\overrightarrow{\mathrm{OP}}\neq\vec{0}$，$\overrightarrow{\mathrm{PX}}\neq\vec{0}$ であれば，$\overrightarrow{\mathrm{OP}}\perp\overrightarrow{\mathrm{PX}}$ です。

このことは標準偏差の公式

$$\sigma=\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j-\overline{x})^2}=\sqrt{\frac{1}{n}\sum_{j=1}^{n}x_j{}^2-(\overline{x})^2}$$

を変形して，

$$\sum_{j=1}^{n}(x_j-\overline{x})^2=\sum_{j=1}^{n}x_j{}^2-n(\overline{x})^2$$

$$\sum_{j=1}^{n}(x_j-\overline{x})^2=\sum_{j=1}^{n}x_j{}^2-\sum_{j=1}^{n}(\overline{x})^2$$

$$\sum_{j=1}^{n}(x_j-\overline{x})^2+\sum_{j=1}^{n}(\overline{x})^2=\sum_{j=1}^{n}x_j{}^2$$

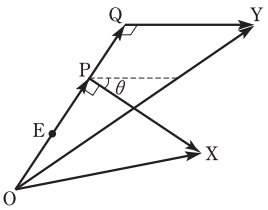
すなわち，

$$\mathrm{PX}^2+\mathrm{OP}^2=\mathrm{OX}^2$$

であることからも導かれます。

また，$\overrightarrow{\mathrm{OP}}\perp\overrightarrow{\mathrm{PX}}$ は，$\overrightarrow{\mathrm{OP}}$ が $\overrightarrow{\mathrm{OX}}$ の直線 OE への正射影であることも示しています。

2 点 P，Q は直線 OE 上にあるので，これを図にすると右のようになります。



この図を見ると，改めて，相関とは点 X，Y の，点 P，Q からの「ずれの方向」の違いを表す指標であることが分かります。

しかし，相関が強いときはともかく，相関が弱いときは「ずれの方向」の自由度はかなり高くなります。

上の図は 3 次元的な図なので，例えば相関係数が 0，すなわち $\overrightarrow{\mathrm{PX}}\perp\overrightarrow{\mathrm{QY}}$ のとき，$\overrightarrow{\mathrm{PX}}$ が決まれば $\overrightarrow{\mathrm{QY}}$ は大きさを除いて固定されるような印象がありますが，$\overrightarrow{\mathrm{OQ}}\cdot\overrightarrow{\mathrm{QY}}=0$ と $\overrightarrow{\mathrm{PX}}\cdot\overrightarrow{\mathrm{QY}}=0$ の 2 つの条件しかないので，n 次元空間であれば，$\overrightarrow{\mathrm{QY}}$ は $n-2$ 次元の自由度をもちます。

また，この図から，$n=2$ のときは，相関係数は $r=\pm 1$ に限られ，$n=1$ のときは，$\overrightarrow{\mathrm{PX}}=\vec{0}$，$\overrightarrow{\mathrm{QY}}=\vec{0}$ となるので相関係数が存在しないことが分かります。

関連して，

$$\sigma=\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j-\overline{x})^2}=\frac{|\overrightarrow{\mathrm{PX}}|}{\sqrt{n}}$$

ですから，標準偏差 σ は，点Xの点Pからの「ずれの大きさ」を表していることが分かります。$\sqrt{n}$ で割っているのは，データの値の個数が変化しても指標として比較できるように調整するためだと考えることができます。

次に散布図における相関の図形的意味を考えます。

§3. 散布図における図形的意味

散布図とは，2つのデータ，

x_1，x_2，x_3，…，x_n　　y_1，y_2，y_3，…，y_n

に対して，xy 平面の点，

$\mathrm{A}_j(x_j,\ y_j)$　$(j=1,\ 2,\ 3,\ \cdots,\ n)$

を図示したものでした。

点$(\overline{x},\ \overline{y})$をGとすると，点Gは質量が等しい n 個の質点 A_j の重心になります。

ここで，簡単のためにデータを「平行移動」して，以下は点Gが原点Oに重なるようにします。

すなわち，改めて $\overline{x}=0$，$\overline{y}=0$ であるデータ

x_1，x_2，x_3，…，x_n　　y_1，y_2，y_3，…，y_n

について，

$\mathrm{A}_j(x_j,\ y_j)$　$(j=1,\ 2,\ 3,\ \cdots,\ n)$

を定めることにします。

このとき，§1の n 次元空間の点P，Qも原点Oに重なるので，

$r=\pm1$ のとき，$\overrightarrow{\mathrm{OX}}/\!/\overrightarrow{\mathrm{OY}}$

となります。したがって，k を実数として，

$\overrightarrow{\mathrm{OY}}=k\overrightarrow{\mathrm{OX}}$

と書けますから，xy 平面上の点 A_j の座標は，

$(x_j,\ y_j)=(x_j,\ kx_j)$

となり，点 A_j は直線 $y=kx$ 上に並びます。相関係数を考える際は，$\overrightarrow{\mathrm{OX}}\neq\vec{0}$，$\overrightarrow{\mathrm{OY}}\neq\vec{0}$ ですから，$k\neq0$ です。

データのスケールを適当に変えれば，k の値も変えられるので，$k=1$ か $k=-1$ に限定することもできます。ですから相関図の「傾き」にはあまり意味がなさそうですが，経年変化などを考える場合には意味があるといえます。

次に，$r=0$ のときを考えます。

このとき，n 次元空間において，

$$\overrightarrow{\mathrm{OX}}\cdot\overrightarrow{\mathrm{OY}}=\sum_{j=1}^{n}x_jy_j=0$$

となりますが，n 次元空間で $\overrightarrow{\mathrm{OX}}\perp\overrightarrow{\mathrm{OY}}$ であることを，2次元平面の散布図で考えようと思っても簡単ではなさそうです。しかし，内積のもう1つの意味が「面積の和」であることを利用すると簡単にイメージがもてます。

点 A_j を頂点とする右の図のような長方形の面積を S_j とすると，x_jy_j の値は

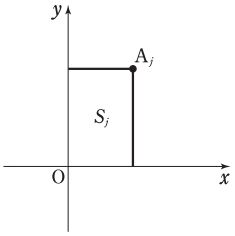
第1・3象限で

$x_jy_j=S_j$

第2・4象限で

$x_jy_j=-S_j$

です。



したがって，$r=0$ とは，

「点 A_j $(j=1,\ 2,\ 3,\ \cdots,\ n)$ の重心を原点として，第1・3象限にできる長方形の面積 S_j の和と第2・4象限にできる長方形の面積 S_j の和が等しいことである」…(＊)

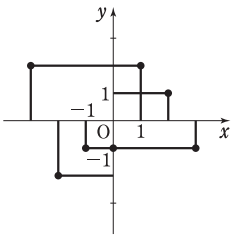
といえます。

これによって，$r=0$ のとき，点 A_j は原点の周りにバランスよく散らばることになります。

しかし，(＊)はそれ以上の意味をもたないので，同じ $r=0$ でも様々な例を考えることができます。

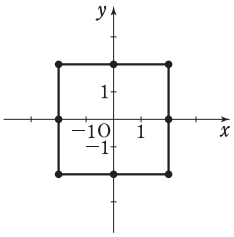
例1

x	-3	-2	-1	0	1	2	3
y	2	-2	-1	-1	2	1	-1



例 2

x	-2	-2	0	2	2	2	0	-2
y	0	2	2	2	0	-2	-2	-2



例 3

$$(x_j,\ y_j)=\left(\cos\frac{2j}{n}\pi,\ \sin\frac{2j}{n}\pi\right)$$
$$(n\geqq 3,\ j=0,\ 1,\ 2,\ \cdots,\ n-1)$$

例 2，例 3 は，とても「相関係数 0 」とは思えませんが，$\overrightarrow{\mathrm{OX}}\perp\overrightarrow{\mathrm{OY}}$ であっても n 次元空間では $n-2$ 次元の自由度をもつことや，散布図では(＊)の意味しかないことを考えると理解できます。

逆に，相関が弱いからといって「法則性がない」とは判断できない，という点に注意する必要がありそうです。

(東京都立調布南高等学校)