第8回　相関係数

１．復習――統計的分析の基礎

（1）平均値（算術平均）各個体の観測値の総和を個体数で割ったもの。

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

（2）分散　各観測値の平均からの偏差の２乗の平均

$$S_x^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2$$

分散は、各データの観測値の 2 乗の平均から、平均の 2 乗を引いたものに等しい（証明済み)。

（3）標準偏差　分散の平方根

$$S_x = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

（4）標準測度　各個体の観測値から平均値を引いたものを平均値からの偏差という。標準測度は平均値からの偏差を標準偏差で割ったもの。

平均からの偏差を d_i とすると、 $d_i = x_i - \bar{x}$

平均からの偏差の平均は 0　　$\bar{d} = 0$

標準測度を z_i とすると、

$$z_i = \frac{d_i}{S_x} = \frac{x_i - \bar{x}}{S_x}$$

標準測度の平均は 0、分散は１。

$$\bar{z} = \frac{1}{n} \sum \frac{x_i - \bar{x}}{S_x} = \frac{\bar{d}}{S_x} = 0$$

$$S_z^2=\frac{1}{n}\sum(z_i-\bar{z})^2=\frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{S_x}\right)^2=\frac{1}{n}\sum\frac{(x_i-\bar{x})^2}{S_x^2}=\frac{S_x^2}{S_x^2}=1$$

２．共分散と相関係数

2 組の対応する変数、x, y をもつ n 個のデータがあるとする。

$$\begin{cases}x_1\\y_1\end{cases}\quad\begin{cases}x_2\\y_2\end{cases}\quad\cdots\quad\begin{cases}x_i\\y_i\end{cases}\quad\cdots\quad\begin{cases}x_n\\y_n\end{cases}$$

各変数の平均値からの偏差を考える。

$$\begin{cases}x_1-\bar{x}\\y_1-\bar{y}\end{cases}\quad\begin{cases}x_2-\bar{x}\\y_2-\bar{y}\end{cases}\quad\cdots\quad\begin{cases}x_i-\bar{x}\\y_i-\bar{y}\end{cases}\quad\cdots\quad\begin{cases}x_n-\bar{x}\\y_n-\bar{y}\end{cases}$$

両 者 の 積 $(x_i-\bar{x})(y_i-\bar{y})$ の平均を共分散という。

共 分 散 $S_{xy}=\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\sum x_i\, y_i-\bar{x}\,\bar{y}=\frac{1}{n^2}\left\{n\sum x_i\, y_i-\left(\sum x_i\sum y_i\right)\right\}$

【証明】

$$\begin{aligned}S_{xy}&=\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})\\&=\frac{1}{n}\sum(x_iy_i-\bar{x}y_i-x_i\bar{y}+\bar{x}\,\bar{y})\\&=\frac{1}{n}\sum x_iy_i-\frac{1}{n}\sum\bar{x}y_i-\frac{1}{n}\sum x_i\bar{y}+\frac{1}{n}\sum\bar{x}\,\bar{y}\\&=\frac{1}{n}\sum x_iy_i-\bar{x}\,\bar{y}-\bar{x}\,\bar{y}+\bar{x}\,\bar{y}\\&=\frac{1}{n}\sum x_iy_i-\bar{x}\,\bar{y}\end{aligned}$$

相関係数（ピアソンの積率相関係数）は、対応する変数を標準測度になおしたときの共分散。(標準化された共分散)。

対応する変数を標準測度に直して考えると、

$$\begin{cases} u_1 = \dfrac{x_1 - \overline{x}}{S_x} \\ v_1 = \dfrac{y_1 - \overline{y}}{S_y} \end{cases} \quad \cdots \quad \begin{cases} u_i = \dfrac{x_i - \overline{x}}{S_x} \\ v_i = \dfrac{y_i - \overline{y}}{Sy} \end{cases}$$

このとき両者の積の平均を x と y の相関係数（ピアソンの積率相関係数）という。

$$r_{xy} = \frac{1}{n}\sum u_i \ v_i \ = \overline{uv} \qquad ①$$

相関係数は、共分散をそれぞれの標準偏差の積で割ったものに等しい。

$$r_{xy} = \frac{1}{n}\sum u_i \ v_i = \frac{S_{xy}}{S_x \cdot S_y} \qquad ②$$

通常、相関係数の定義式としては②式を用いる。

【②式の証明】

$$\frac{1}{n}\sum u_i \ v_i = \frac{1}{n}\sum \frac{x_i - \overline{x}}{S_x} \ \frac{y_i - \overline{y}}{Sy}$$

$$= \frac{1}{n}\sum \frac{(x_i - \overline{x})(y_i - \overline{y})}{S_x \cdot S_y}$$

$$= \frac{S_{xy}}{S_x \cdot S_y}$$

相関係数は $-1 \leqq r_{xy} \leqq 1$ である。

【証明】

$$\frac{1}{n}\sum (u_i \pm \ vi)^2 \geqq 0$$

$$\Leftrightarrow \frac{1}{n}\sum u_i^2 \pm \ 2\frac{1}{n}\sum u_i v_i + \frac{1}{n}\sum v_i^2 \geqq 0$$

u_i, v_i は標準測度だから、分散は１、平均は０ 。

$\Leftrightarrow\ 1\pm\ 2r_{xy}+1\geqq 0$

$\Leftrightarrow\ 2\pm\ 2r_{xy}\geqq 0$

$\Leftrightarrow\ 1\pm\ r_{xy}\geqq 0 \qquad \therefore -1\leqq r_{xy}\leqq 1$

２つの量的変数間の関連を見るには、相関係数がよく使われる。相関係数は、0 のとき無相関、1 のとき完全（正）相関、−1 のとき完全逆相関となる。