

相関 (correlation)

2種類のデータのあいだになんらかの関係がある場合、統計学的な関係がみられるときに、「相関がある」という。

- データの大小に関して、一方の値が変わるにつれて、もう一方の値も変わる
 - 身長と体重
 - 収縮期血圧と拡張期血圧

相関係数 (correlation coefficient)

相関の種類

- 線形相関: 相関(関係)を示すグラフが1本の直線で近似できる
 - 順相関: 相関が正の場合 (グラフが右肩あがりの直線)
 - 逆相関: 相関が負の場合 (グラフが右肩さがりの直線)
- 非線形相関: 相関を示すグラフが指数関数や2次・3次関数のように直線状をしていない

相関係数 (ピアソンの積率相関係数)

- 相関係数 (ピアソン(Pearson)の積率相関係数) r は、相関の程度をあらわし、次の値をとる。

$$-1 \leq r \leq +1$$

- 完全相関: 相関係数が ± 1 の場合
- 無相関: 相関係数が 0 の場合
- 相関係数 r は、次の式で求められる

$$\begin{aligned} r &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \\ &= \frac{\frac{1}{n} S_{xy}}{s_x s_y} \end{aligned}$$

- 標本数: n
- 標準偏差: s_x, s_y
- 偏差積和 (偏差の積の合計):

$$S_{xy} = \sum_{i=1}^n d_x \cdot d_y = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- または、次の式でも求められる (統計量だけから計算できる)

$$\begin{aligned} r &= \frac{1}{s_x s_y} \left(\frac{\sum x_i y_i}{n} - \bar{x} \bar{y} \right) \\ &= \frac{1}{s_x s_y} \left(\frac{T_{xy}}{n} - \bar{x} \bar{y} \right) \end{aligned}$$

- 積和 (2変数の積の合計):

$$T_{xy} = \sum_{i=1}^n x_i y_i$$

偏差積和

- 偏差積和(偏差の積和) S_{xy} とは、偏差(各データと平均の差)の積の総和である。

$$\begin{aligned} S_{xy} &= \sum_{i=1}^n d_x d_y \\ &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

- 標本数: n
- 偏差: d_x, d_y

共分散

- 共分散 s_{xy} は、偏差積和を標本数で割ったもの。

$$\begin{aligned} s_{xy} &= \frac{1}{n} S_{xy} \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^n d_x \cdot d_y \end{aligned}$$

- 標本数: n
 - 偏差: d_x, d_y
- 共分散を使うと、相関係数は次のようになる。

$$r = \frac{s_{xy}}{s_x s_y}$$

偏差平方和

- 偏差平方和: S_{xx} は、偏差の二乗の合計を計算したもの

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n d_x^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

- 偏差平方和を使うと、相関係数は次のようになる。

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \\ &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \end{aligned}$$

相関関係と因果関係

相関関係から因果関係を確定するときの注意点

- 関連の時間性
 - 原因は結果の前にあるか

2. 関連の密接性

- 原因が結果に密接に関連するか

3. 関連の特異性

- 原因が結果のは性にどの程度かわっているか

4. 関連の普遍性

- 対象や時期、方法などが異なっても、類似した結果が得られるか

5. 関連の合理性

- 従来の理論や経験から考えて矛盾がないか

疑似相関（見かけの相関）

- 直接の相関はないが何かある要因が2つの事象と相関しているために、2つの事象に相関がみられるケースがある。
- 相関関係があるからといって、それが必ずしも因果関係であるとは限らない場合がある。
 - 「ビアホールでの1日当たりの生ビールの売り上げ数」と「アイスクリーム店のお客の数」
 - 「進行性の疾患をもつ患者の疾患についての知識」と「その疾患の進行度」

相関の程度

相関係数の値から、相関の程度を次のように記述できる。

-1.0	相関係数 $r < -0.7$	強い負の相関
-0.7	相関係数 $r < -0.4$	かなりな負の相関
-0.4	相関係数 $r < -0.2$	やや負の相関
-0.2	相関係数 r 0.2	ほとんど相関がない
0.2	相関係数 r 0.4	やや正の相関
0.4	相関係数 r 0.7	かなりな正の相関
0.7	相関係数 r 1	強い正の相関

なお、標本数が少ない場合は、母相関係数の推定や検定（後日説明）が必要となる。

順位相関係数

- 相関がない場合や順位に意味がある・順位だけしかわからない場合には、順位データ（データを小さいほうから並べた順位）をもとに相関を求めてみる。
 - 英語のテストの順位と数学のテストの順位の相関
 - 薬と奇形児発生の相関
 - 2つの銘柄の株価の相関
- スピアマン (Spearman) の順位相関係数 r_s は、相関係数と同様、次の値をとる。

$$-1 \leq r_s \leq +1$$

- 同一順位の場合は、次のように扱う（平均順位）
 - 2位が2つある場合：2位と3位の中間 $(2+3)/2=2.5$ 位を順位とする
 - 4位が3つある場合：4位と5位と6位の中間 $(4+5+6)/3=5$ 位を順位とする
- 順位相関係数は、次のようにして求められる。

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

- 標本数: n
- 番目の順位差: d_i