第 1 章 相関係数の描像

浅川 伸一

2006 年 10 月 12 日

1　相関関係、共変関係、因果関係

広辞苑第四版電子ブック版によれば、相関、共変、および因果関係の意味は以下のとおり。

相関関係 一方が他方との関係を離れては意味をなさないようなものの間の関係。父と子、右と左など。相関関係にある概念を相関概念という。

共変関係 項目がない。(;_;)

因果関係 原因とそれによって生ずる結果との関係。

x：蛙が鳴く, y:雨が降る, とすると x → y は蛙が鳴くと雨が降る。これは因果関係ではない。蛙の口を縛っても音を出せないようにしても雨は降る (以上吉田 (1998) より)。この場合は x, y とも気圧の変化などの別の変数 z の影響によって受ける共変関係。その他にも
`http://info.pref.fukui.jp/nougyou/murano_kurashi/otenki/yuki.html`
などがある。秋に鳥の餌が低い所に挿してある時は小雪、高きは大雪
鳥の餌を低いところから高いところに挿し替えたら大雪になるか？

南風原 (2002) より

1. x が大きい人ほど y も大きい。
2. x が大きくなると y も大きくなる。
3. x が大きくすると y も大きくなる。
4. x が大きいから y も大きい。

一日の最高気温と花粉症の発生件数。直接因果関係を意味しない間接的な相関、もしくは偽相関。杉花粉の量。

風が吹くと桶屋が儲かるは？

ボーイ (ガール) スカウトに入ると外向的な性格になる。ボーイスカウト経験の有無が無作為に決定されない限り、経験が性格を決定すると言うのは難

しい。元もとそういう性格傾向を持っていたために経験を受け入れたのかも知れない。親の養育態度、親の性格、親の好みが影響したのかも知れない。

データから因果関係を導くのはそう簡単な話ではない。例えば、ある課題中に脳内のある部位に活性化が認められたからと言って、その部位がその課題を遂行するための中枢であるとは限らない。おばあちゃん細胞仮説、否定的な文脈で用いられることが多い。統計学を学んでいる間にある部位が活性化したとして、その部位を統計学中枢と呼んで良いのだろうか。

2 相関図

二つの関連する変量 x, y がに対して、x を横軸、y を縦軸として個々のデータを点で表したグラフを散布図 scattergram という。x と y との数字は、それぞれ対応した意味を持つデータでなければならない。例えば、ある実験中におけるある被験者の頭頂葉における血流量の変化と後頭葉に置ける血流量の変化、ある人の中間テストの成績と期末テストの成績。ところがあるクラスの中間テストの成績と別のクラスの中間テストの成績とは散布図として描けない。個々のデータが対応しないから。

下のグラフは、n 行 m 列目のグラフを (n, m) と表すとすると、それぞれの相関係数は (1,1)=0.0, (1,2)=0.2, (1,3)=0.4, (2,1)=0.6, (2,2)=0.8, (2,3)=0.9, (3,1)=1.0,

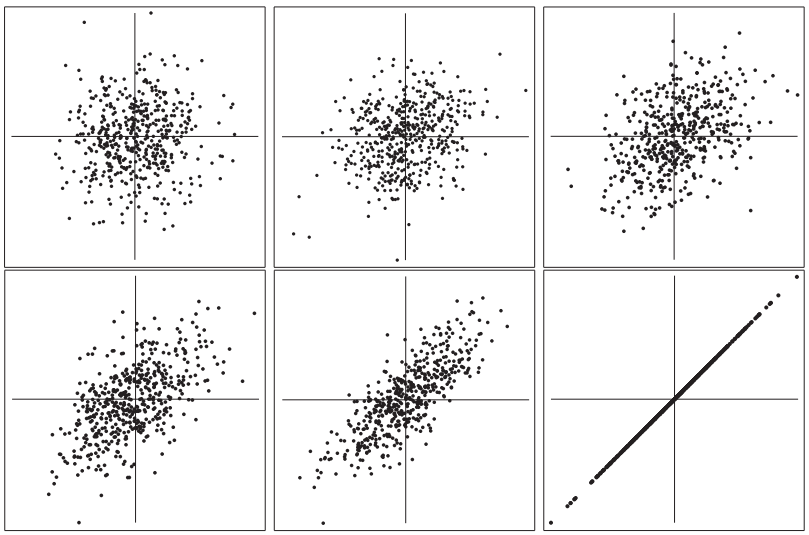

図 1: 散布図 scattergram. n 行 m 列目のグラフを (n, m) と表すとすると、それぞれの相関係数は (1,1)=0.0, (1,2)=0.2, (1,3)=0.4, (2,1)=0.6, (2,2)=0.8, (2,3)=1.0,

(1,1)=-0.2, (1,2)=-0.4, (1,3)=-0.6, (2,1)=-0.8, (2,2)=-0.9, (2,3)=-1.0.

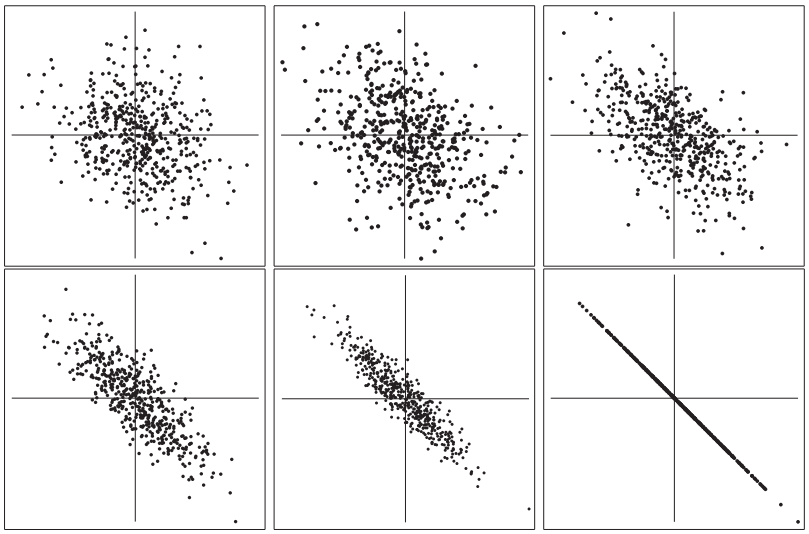

図 2: 負の相関を表す散布図。相関係数はそれぞれ (1,1)=-0.2, (1,2)=-0.4, (1,3)=-0.6, (2,1)=-0.8, (2,2)=-0.9, (2,3)=-1.0.

となる。

`http://case.f7.ems.okayama-u.ac.jp/statedu/flash/scattergram.html` なども参考にすること。相関係数の数値と対応する散布図とをそれぞれ確認すること。

図では縦軸横軸とも平均が 0 になるように描かれているが描画する上では大切ではない。相関係数を計算するときには重要。

3 相関係数の定義

対応のある 2 変数 x, y が n 個あったとき、標本相関係数を r は以下のように定義される。

$$r = \frac{s_{xy}}{s_x s_y} \tag{1}$$

ここで s_x, s_y は x, y それぞれの標準偏差である。s_{xy} は共分散 covariance と呼ばれる量である。言葉で表現すれば、相関係数とは二つの変量の共分散をそれぞれの変量の標準偏差で除したものである。共分散は以下で定義される。

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{2}$$

ここで $\bar{x}, \bar{y}$ はそれぞれの平均である[1]。x の標準偏差は次のように表すことができた。(復習：前期の対応する式はどれだろうか)

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

これらを用いて相関係数の定義式を書き下すと次のようになる。

$$\begin{aligned} r &= \frac{s_{xy}}{s_x s_y} && (4)\\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} && (5)\\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} && (6) \end{aligned}$$

分散 s_x^2, s_y^2 およびその開平である標準偏差 s_x, s_y は正であるから、相関係数が負となるためには共分散が負 $s_{xy}0$ でなければならない。

4 相関に関するいくつかの注意点

重要な注意点として幾つか列挙する。

1. 分断によって相関係数が変化してしまう場合がある。入学試験の成績と入学後の成績など。被験者を選んでしまうような調査ではしばしば問題になる。ネット上での調査など。

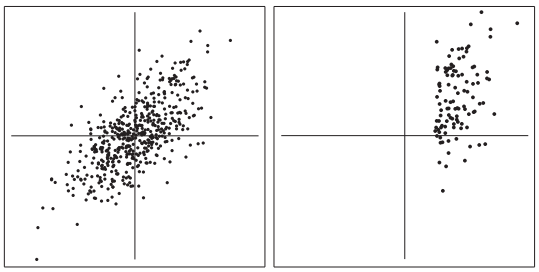

図 3: 分断前と分断後

2. 曲線相関がある場合には相関係数を求めることが不適切である場合がある。計算はできてしまうが意味を成さない。U shape, log curve, and so on.

[1]混乱するのでここでは触れないが、上記の量は全て標本統計量である。標本平均、標本標準偏差、標本共分散、標本相関係数の定義を示した。これに対応する母集団統計量があって、それぞれ母平均、母標準偏差がそんざいする。同じようにして母共分散、母相関係数も存在する。後述。

3. 異なる集団を同一視してしまうと相関係数の意味をなさないことがある。

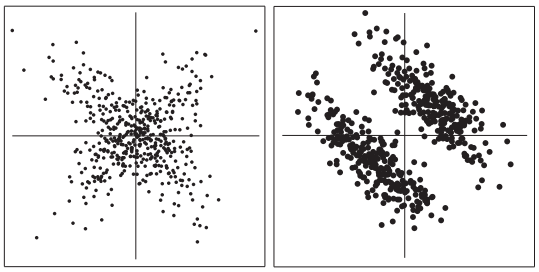

図 4: 集団の混交

5 相関係数の別解釈

2つの n 次元ベクトルを $\boldsymbol{x}, \boldsymbol{y}$ と太字で表記する。ここで $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$, $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ である[2]。このとき $\boldsymbol{x}$ と $\boldsymbol{y}$ との内積 inner product は次のように定義される。

$$(\boldsymbol{x}, \boldsymbol{y}) = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \tag{7}$$

自分自身との内積をベクトルの長さ (を 2 乗したもの) と言う。

$$|\boldsymbol{x}|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \tag{8}$$

これを用いてベクトルの別の定義

$$(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x}||\boldsymbol{y}| \cos\theta \tag{9}$$

が与えられる。ここで θ は二つのベクトルのなす角の余弦 cosine である。ベクトルの内積の定義は
`http://www.nikonet.or.jp/spring/in_pro.htm`
などを参照。上の式から $\cos\theta$ は

$$\cos\theta = \frac{(\boldsymbol{x}, \boldsymbol{y})}{|\boldsymbol{x}||\boldsymbol{y}|} \tag{10}$$

と書くことができる。$\boldsymbol{x}$ と $\boldsymbol{y}$ の各値からそれぞれの平均 $\bar{x}$, $\bar{y}$ を引いたベクトルを改めて $\boldsymbol{x}$, $\boldsymbol{y}$ とする。

$$\boldsymbol{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x})' \tag{11}$$

$$\boldsymbol{y} = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_n - \bar{y})' \tag{12}$$

[2]ベクトルの肩に乗っている ' は転置を表す。すなわち列 (縦) ベクトルを行 (横) ベクトルに、行ベクトルを列ベクトルにする操作を意味する。

するとこのベクトルの長さをデータ数 n で割ったものが分散に一致する。

$$s_x^2 = \frac{1}{n}|\boldsymbol{x}|^2 = \frac{1}{n}(\boldsymbol{x},\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{13}$$

$$s_y^2 = \frac{1}{n}|\boldsymbol{y}|^2 = \frac{1}{n}(\boldsymbol{y},\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{14}$$

さらに $\boldsymbol{x}$ と $\boldsymbol{y}$ との共分散 s_{xy} は以下のように表せる。

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \tag{15}$$

$$= \frac{1}{n}(\boldsymbol{x},\boldsymbol{y}) \tag{16}$$

変数 x と y との相関係数 r_{xy} は x と y との共分散をそれぞれの標準偏差 (分散の開平) で除したものであるから次の重要な関係を得る。

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{17}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{18}$$

$$= \frac{(\boldsymbol{x},\boldsymbol{y})}{|\boldsymbol{x}||\boldsymbol{y}|} \tag{19}$$

$$= \cos\theta \tag{20}$$

上式は散布図の上でイメージされた相関係数を 2 本の n 次元ベクトルのなす角とみなすこともできることを意味している。2 次元上の n 個の点と n 次元上の 2 点とは数学的なとり扱いとしては等価である。このような相関係数の描像を持つ必要がある。相関係数となす角の対応表を作ってみると以下のようになる。

相関係数	角度
1.0	0 °(重なる)
$\frac{\sqrt{3}}{2}$	30 °
$\frac{\sqrt{2}}{2}$	45 °
$\frac{\sqrt{1}}{2}$	60 °
0.0	90 °(直交する)
$-\frac{1}{2}$	120 °
-1.0	180 °(反対向き)