2. 相関関係

1. 相関係数の定義

2 つの変量 x と y から 2 次元データ

$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_N, y_N)$$

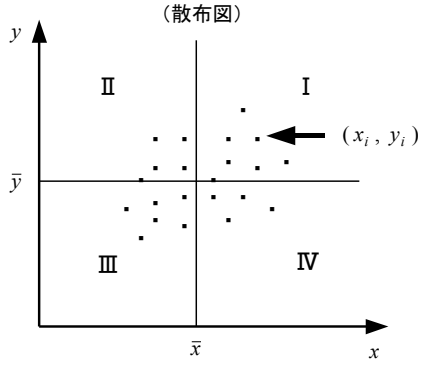
が得られたとき，x と y の相関関係を表す尺度として「相関係数」がある。

いま，$x_1, x_2, \cdots, x_N$ の平均を $\bar{x}$，標準偏差を σ_x とし，$y_1, y_2, \cdots, y_N$ の平均を $\bar{y}$，標準偏差を σ_y とする。

散布図において，点 (x_i, y_i) の位置を考える。x_i は $\bar{x}$ の近辺に多く，y_i は $\bar{y}$ の近辺に多いので，多くの点は，平均を座標とする点 $(\bar{x}, \bar{y})$ の近くに集まっていると考えられる。

そこで，点 $(\bar{x}, \bar{y})$ を通り両軸に平行な 2 直線を引いて，平面を下図のように 4 つの領域 I ～Ⅳに分割してみる。

（散布図）



このとき，点 (x_i, y_i) が領域 I にあれば，$x_i - \bar{x} > 0$，$y_i - \bar{y} > 0$ より，$(x_i - \bar{x})(y_i - \bar{y}) > 0$ であり，領域 II にあれば，$x_i - \bar{x} < 0$，$y_i - \bar{y} > 0$ より，$(x_i - \bar{x})(y_i - \bar{y}) < 0$ である。

よって，点 (x_i, y_i) に対して，x 座標の偏差 $x_i - \bar{x}$ と y 座標の偏差 $y_i - \bar{y}$ との積 $(x_i - \bar{x})(y_i - \bar{y})$ を考えれば，次のことがわかる。

点 (x_i, y_i) が I または Ⅲ にある $\iff$ $(x_i - \bar{x})(y_i - \bar{y}) > 0$

点 (x_i, y_i) が II または Ⅳ にある $\iff$ $(x_i - \bar{x})(y_i - \bar{y}) < 0$

一方，x と y の間に正の相関があれば，I または Ⅲ に位置する点が多いため，

$$\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_N - \bar{x})(y_N - \bar{y})$$

の値は正になると考えられる。逆に，負の相関があれば，II または Ⅳ に点が多く集まるため，この式の値は負になるだろう。また，無相関であれば I ～Ⅳに点が均等に散在するため，上式の値は 0 に近づくであろう。

従って，上式は相関を表す尺度と考えられるが，その値は N に関係するため，N で割った式を考え，それを x と y の共分散（covariance）と呼び，σ_{xy} で表す。

$$\text{共分散}\quad \sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

つまり，<u>共分散とは「偏差の積の平均」のことである</u>。

ただし，共分散は名数のため，x と y の単位の取り方により，その値が変わるという欠点がある。例えば，x が身長，y が体重の場合，それぞれの単位をどのように取るかにより，同じデータであっても共分散の値は異なってくる。

そこで，共分散 σ_{xy} をそれぞれの標準偏差 σ_x，σ_y で割った値を考え，それを x と y の相関係数（correlation coefficient）と呼び r_{xy} で表す。（相関係数を単に r で表すこともある）。

$$\text{相関係数}\quad r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{1}{N}\sum_{i=1}^{N} \frac{x_i - \bar{x}}{\sigma_x} \cdot \frac{y_i - \bar{y}}{\sigma_y}$$

これで，相関係数は無名数になる。

上記の式を見てわかるように，<u>相関係数は「標準測度の積の平均」である</u>。すなわち，x_i の標準測度を z_i，y_i の標準測度を w_i とすると

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}\ （x_i\ \text{の標準測度}），\quad w_i = \frac{y_i - \bar{y}}{\sigma_y}\ （y_i\ \text{の標準測度}）$$

$$r_{xy} = \frac{1}{N}\sum_{i=1}^{N} z_i w_i\ （\text{標準測度の積の平均}）$$

なお，z_i の標準偏差を σ_z，w_i の標準偏差を σ_w，z_i と w_i の共分散を σ_{zw}，z_i と w_i の相関係数を r_{zw} とすれば，

$$\bar{z} = 0,\quad \sigma_z = 1,\quad \bar{w} = 0,\quad \sigma_w = 1$$

であるから，

$$r_{zw} = \frac{\sigma_{zw}}{\sigma_z \sigma_w} = \sigma_{zw} = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(w_i - \bar{w}) = \frac{1}{N}\sum_{i=1}^{N} z_i w_i = r_{xy}$$

すなわち，相関係数 r_{xy} は，標準測度の相関係数 r_{zw} と一致し，データを標準化しても相関係数は変わらないことが分かる。

以上をまとめると次のようになるが，相関係数を求めるときには，(3)の共分散の公式もよく使用される。

● 相関係数の定義と性質

2 つの変量 x と y から得られた 2 次元データ

$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_N, y_N)$$

に対して，(1)と(2)のように定義する。

(1) x と y の共分散 σ_{xy}

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \text{ ⇐(偏差の積の平均)}$$

(2) x と y の相関係数 r_{xy}

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{x\text{と}y\text{の共分散}}{(x\text{の標準偏差})\,(y\text{の標準偏差})} = \frac{1}{N}\sum_{i=1}^{N}\frac{x_i - \bar{x}}{\sigma_x}\cdot\frac{y_i - \bar{y}}{\sigma_y}$$

(3) データを標準化しても，相関係数の値は変わらない。

(4) 共分散に関する公式

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\cdot\bar{y} = (\text{積の平均}) - (\text{平均の積})$$

(4)を証明すると，次のようになる。$\sum_{i=1}^{N}$ は $\sum_{i}$ で表す。

$N\bar{x} = \sum_i x_i$，$N\bar{y} = \sum_i y_i$ であるから，

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$

$$= \sum_i x_i y_i - \sum_i x_i \bar{y} - \sum_i \bar{x} y_i + \sum_i \bar{x}\bar{y} = \sum_i x_i y_i - \bar{y}\sum_i x_i - \bar{x}\sum_i y_i + N\bar{x}\bar{y}$$

$$= \sum_i x_i y_i - N\bar{x}\bar{y} - N\bar{x}\bar{y} + N\bar{x}\bar{y} = \sum_i x_i y_i - N\bar{x}\bar{y}$$

$$\therefore\ \sigma_{xy} = \frac{1}{N}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_i x_i y_i - \bar{x}\bar{y}$$

2. 相関係数の値の範囲

相関係数 r_{xy} のとり得る値の範囲は，$-1 \le r_{xy} \le 1$ である。この証明方法はいろいろあるが，以下では数 I で証明できる方法を説明する。

● 相関係数 r_{xy} の値の範囲

2 つの変量 x と y に関する 2 次元データ $(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_N, y_N)$ について，以下が成り立つ。

(1) $-1 \le r_{xy} \le 1$　すなわち　$|r_{xy}| \le 1$

(2) $r_{xy} = 1 \iff$ すべての点 (x_i, y_i) が直線 $y = ax + b\ (a > 0)$ 上にある。

(3) $r_{xy} = -1 \iff$ すべての点 (x_i, y_i) が直線 $y = ax + b\ (a < 0)$ 上にある。

（証明）データを標準化すると，計算が楽になる。

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}, \quad w_i = \frac{y_i - \bar{y}}{\sigma_y}$$

とおくと，$r_{xy} = \frac{1}{N}\sum_i z_i w_i$ であった。ここで，

$$A = \frac{1}{N}\sum_i (z_i \pm w_i)^2$$

とおくと，$\sum_i z_i^2 = N$，$\sum_i w_i^2 = N$ であるので，

$$A = \frac{1}{N}\sum_i (z_i^2 \pm 2 z_i w_i + w_i^2) = \frac{1}{N}\sum_i z_i^2 \pm 2 \cdot \frac{1}{N}\sum_i z_i w_i + \frac{1}{N}\sum_i w_i^2$$
$$= 1 \pm 2 r_{xy} + 1 = 2(1 \pm r_{xy})$$

ここで，$A \geq 0$ であるから，$1 \pm r_{xy} \geq 0$ となり，ゆえに $-1 \leq r_{xy} \leq 1$ となる。

$r_{xy} = 1$ となるのは，$1 - r_{xy} = 0$ すなわち

$$\frac{1}{N}\sum_i (z_i - w_i)^2 = 0$$

の場合であり，これはすべての $(z_i - w_i)^2$ が 0，すなわち

$$z_i = w_i \quad (i = 1, 2, \cdots, N)$$

が成立することと同値である。これを書き直すと

$$\frac{x_i - \bar{x}}{\sigma_x} = \frac{y_i - \bar{y}}{\sigma_y}$$

であり，これを変形すると

$$y_i = \frac{\sigma_y}{\sigma_x}(x_i - \bar{x}) + \bar{y} \quad (i = 1, 2, \cdots, N)$$

この等式は，すべての点 (x_i, y_i) が直線

$$y = \frac{\sigma_y}{\sigma_x}(x - \bar{x}) + \bar{y}$$

の上にあることを示す。

同様にして，$\sigma_{xy} = -1$ となるのは，すべての点 (x_i, y_i) が直線

$$y = -\frac{\sigma_y}{\sigma_x}(x - \bar{x}) + \bar{y}$$

の上にある場合であることが分かる。

ここで，$\frac{\sigma_y}{\sigma_x} > 0$ であるから，(2)と(3)は証明された。

さて，分かったことは，相関係数の r_{xy} の値は必ず－1 以上 1 以下であり，計算によって $r_{xy} = 1.2$ などとなった場合は，計算の誤りである。

また，$r_{xy} = \pm 1$ となるのは，すべての点 (x_i, y_i) が一定の直線 $y = ax + b$ $(a \neq 0)$ 上にあるとき，すなわち

$$y_i = a x_i + b \quad (i = 1, 2, \cdots, n)$$

が成立する場合であり，この場合は最も強い相関関係になる。$r_{xy} = \pm 1$ のとき，x と y の間に完全

相関があるという。

では，$-1 < r_{xy} < 1$ の場合はどうであろうか。容易に推測できるように，$r_{xy} > 0$ の場合には正の相関があり，その値が 1 に近いほど相関が強くなると推測できる。同様に，$r_{xy} < 0$ の場合には負の相関があり，その値が－1 に近いほど相関が強くなる。（これらは回帰直線で明確になる。）

相関係数の値から相関関係をどのように判断すべきかは一概にはいえず，また，難しい問題でもある。つまり，データの種類，分析の目標などによって異なり，明確な判断基準はない。

ただし，通常は，次のように判断してよい。

$|r_{xy}| \geq 0.7$ のときは，強い相関

$|r_{xy}| < 0.2$ のときは，無相関

以下は，テキストの判断基準であるが，あくまでも参考である。

また，正の相関を「順相関」，負の相関を「逆相関」ともいう。相関の強弱について，強いことを「高い」，弱いことを「低い」という人もいる。

● 相関係数 r_{xy} の意味

(1) $r_{xy} = \pm 1$ のときは，最も強い相関である（完全相関）。

(2) $r_{xy} = 0$ のときは，相関は全くない。

(3) $r_{xy} > 0$ のときは正の相関（順相関），$r_{xy} < 0$ のときは負の相関（逆相関）があると考えられる。

(4) r_{xy} の値が 1 に近づくほど，正の相関が強くなる。

(5) r_{xy} の値が－1 に近づくほど，負の相関が強くなる。

(6) r_{xy} の値が 0 に近づくほど，相関が弱くなる。

(7) 一般には，次のように考えてよい。

① $|r_{xy}| \geq 0.7$ のとき，強い相関がある。

② $|r_{xy}| < 0.2$ のとき，相関関係はほとんどない（無相関）。

(8) 次は，テキストの判断基準：

① $r_{xy} \geq 0.7$ のとき，強い正の相関

② $0.4 \leq r_{xy} < 0.7$ のとき，正の相関

③ $0.2 \leq r_{xy} < 0.4$ のとき，弱い正の相関

④ $r_{xy} \leq -0.7$ のとき，強い負の相関

⑤ $-0.7 < r_{xy} \leq -0.4$ のとき，負の相関

⑥ $-0.4 < r_{xy} \leq -0.2$ のとき，弱い相関

⑦ $-0.2 < r_{xy} < 0.2$ のとき，無相関

3. 相関係数の計算練習

（例）2 つの変量 x と y を測定して，次の表を得た。以下，相関係数の定義の式から，x と y の相関係数 r_{xy} を求めてみる。

x	7	8	5	6	9
y	4	5	3	3	5

分散の公式　$\sigma_x{}^2 = \dfrac{1}{N}\sum_i x_i{}^2 - \bar{x}^2$

共分散の公式　$\sigma_{xy} = \dfrac{1}{N}\sum_i x_i y_i - \bar{x}\cdot\bar{y}$

x	y	x^2	y^2	xy
7	4	49	16	28
8	5	64	25	40
5	3	25	9	15
6	3	36	9	18
9	5	81	25	45
35	20	255	84	146

変量 x　：平均　$\bar{x} = \dfrac{35}{5} = 7$

分散　$\sigma_x{}^2 = \dfrac{255}{5} - 7^2 = 51 - 49 = 2$

標準偏差　$\sigma_x = \sqrt{2}$

変量 y　：平均　$\bar{y} = \dfrac{20}{5} = 4$

分散　$\sigma_y{}^2 = \dfrac{84}{5} - 4^2 = 16.8 - 16 = 0.8$

標準偏差　$\sigma_y = \sqrt{0.8}$

共分散　$\sigma_{xy} = \dfrac{146}{5} - 7\times 4 = 29.2 - 28 = 1.2$

相関係数　$r_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = \dfrac{1.2}{\sqrt{2}\times\sqrt{0.8}} \fallingdotseq 0.95$

4. 相関係数における外れ値の影響

データには，外れ値（異常値）がつきものである。外れ値とは，他の値から大きく外れた値のことであり，異常値ともいう。

平均，分散，標準偏差など，ほとんどの統計量は，すべての値を計算に入れるため，外れ値の影響を受けるが，相関係数は特に大きな影響を受ける。次の例を見てみよう。

■ 例　学生 5 人に対して，英語の学力と数学の学力の相関を調べる目的で，英語と数学の試験を実施した。次の表がその結果である。

	学生 1	学生 2	学生 3	学生 4	学生 5
英語の成績（x）	40 点	50 点	60 点	70 点	100 点
数学の成績（y）	37 点	39 点	56 点	80 点	40 点

学生 1～学生 4 の間では，x と y の間に正の相関が見られるが，学生 5 の成績（100，40）は外れ値であるといえる。x と y の相関係数 r_{xy} を計算すると，次のようになる。

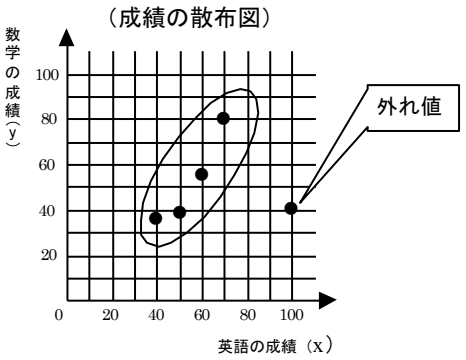
（イ）学生 5 を含めない場合

学生 1～学生 4 に限定すれば，$r_{xy} = 0.95$ であり，x と y の間に強い正の相関がある。

（ロ）学生 5 を含めた場合

学生 1～学生 5 に対して計算すると，$r_{xy} = 0.16$ となり，相関がほとんどない。

このように，異常な点（外れ値）を 1 個含めるか除外するかで，結論が全く違ってくる。従って，実際の分析では，相関係数のみで結論を出すのではなく，散布図をつくりグラフを眺めることも重要になる。



さて，学生 5 人から計算した相関係数は $r_{xy} = 0.16$ となったが，これで「相関がほとんどない」と判断するのは誤りであろう。1 個の点を除けば，強い正の相関があるからである。この場合には，正の相関があると判断するのが妥当であるが，この判断を明確にするために，順位の相関係数を考えることもある。

5. 順位に関する相関係数

相関係数が外れ値の影響を大きく受ける理由は，実際のデータの値で計算しているためであり，データの大小の順位で計算すれば，外れ値の影響を少なくすることができる。前述の例題で説明しよう。

■ 例　前述の例題において，成績の点数ではなく，成績の順位を考える。学生 5 人の英語の成績は，40 点，50 点，60 点，70 点，100 点であるが，最高の 100 点を 1 位とし，最低の 40 点を 5 位とする（順位は逆にしてもよい）。数学も同様とし，英語の成績の順位を x，数学の成績の順位を y で表す。

	学生 1	学生 2	学生 3	学生 4	学生 5
英語の成績の順位（x）	5	4	3	2	1
数学の成績の順位（y）	5	4	2	1	3

ここで，x の値は 1, 2, 3, 4, 5，y の値も 1, 2, 3, 4, 5 なので，$\bar{x} = \bar{y}$，$\sigma_x{}^2 = \sigma_y{}^2$ である。

x	y	x^2	y^2	xy
5	5	25	25	25
4	4	16	16	16
3	2	9	4	6
2	1	4	1	2
1	3	1	9	3
15	15	55	55	52

変量 x　：平均 $\bar{x} = \dfrac{15}{5} = 3$

分散 $\sigma_x{}^2 = \dfrac{55}{5} - 3^2 = 11 - 9 = 2$

変量 y　：平均 $\bar{y} = \bar{x} = 3$，分散 $\sigma_y{}^2 = \sigma_x{}^2 = 2$

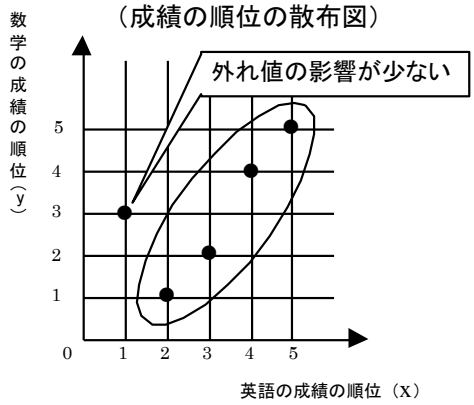
共分散 $\sigma_{xy} = \dfrac{52}{5} - 3 \times 3 = 1.4$

相関係数 $r_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = \dfrac{\sigma_{xy}}{\sigma_x{}^2} = \dfrac{1.4}{2} = 0.7$

相関係数は 0.7 であり，順位の散布図も次のようになるので，成績の順位の間には正の相関がはっきりあると判断できる。

このように，順位で考えれば，外れ値の影響が少なくなり，相関係数がより妥当な値になる。

学生 5 人の学力については，英語の学力と数学の学力の間に，はっきりとした正の相関があると判断してよい。


（成績の順位の散布図）
外れ値の影響が少ない
数学の成績の順位（y）
5
4
3
2
1
0
1
2
3
4
5
英語の成績の順位（x）