§1.4　2次元のデータの共分散と相関係数

例えば19歳の男性において身長と体重とにどれぐらい関連があるのか調べるとします．そのためには，19歳の男性を沢山集めて身長と体重とのデータをとります．このとき，各人の身長と体重とを対にして記録する必要があります．身長だけのデータと体重だけのデータがあっても身長と体重との関連は分かりません．

一般に，変量 X と変量 Y との間の関連について興味があるとき，X の値と Y の値との順序対の集まりのデータが必要です．このように，ある量の値とまたある量の値との順序対の集まりであるデータを2次元のデータといいます．数学では，変量 X の値 x と変量 Y の値 y とを順序対 (x,y) で表し，座標平面の点と考えます．

2次元のデータについて共分散 (covariance) と相関係数 [1] (correlation coefficient) とを定義します．

定義　正の自然数 n に対して，変量 X の値 x と変量 Y の値 y との順序対 (x,y) が次のように n 個あるとする：

$$(x_1,y_1)\,,\ (x_2,y_2)\,,\ (x_3,y_3)\,,\ (x_4,y_4)\,,\ \dots\,,\ (x_n,y_n)\,.$$

X の値の平均を $\overline{x}$ とおき，Y の値の平均を $\overline{y}$ とおく：

$$\overline{x}=\frac{1}{n}\sum_{k=1}^{n}x_k\ ,\qquad \overline{y}=\frac{1}{n}\sum_{k=1}^{n}y_k\ .$$

この2次元のデータにおける X と Y との共分散 c_{XY} を次のように定義する：

$$c_{XY}=\frac{1}{n}\sum_{k=1}^{n}\{(x_k-\overline{x})(y_k-\overline{y})\}\ .$$

更に，X の値の標準偏差を s_X とおき，Y の値の標準偏差を s_Y とおく：

$$s_X=\sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2}\ ,\qquad s_Y=\sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k-\overline{y})^2}\ .$$

X と Y との共分散 c_{XY} を用いて，この2次元のデータにおける X と Y との相関係数 r_{XY} を次のように定義する：$s_X\neq 0$ かつ $s_Y\neq 0$ のとき，

$$r_{XY}=\frac{c_{XY}}{s_Xs_Y}=\frac{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}\{(x_k-\overline{x})(y_k-\overline{y})\}}{\sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})^2}\sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(y_k-\overline{y})^2}}$$

$$=\frac{\displaystyle\sum_{k=1}^{n}\{(x_k-\overline{x})(y_k-\overline{y})\}}{\sqrt{\left\{\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})^2\right\}\left\{\displaystyle\sum_{k=1}^{n}(y_k-\overline{y})^2\right\}}}\ .$$

共分散について次の定理が成り立つ．

定理 1.4.1　正の自然数 n に対して，変量 X の値 x と変量 Y の値 y との順序対 (x,y) が次のように n 個あるとする：

$$(x_1,y_1)\,,\ (x_2,y_2)\,,\ (x_3,y_3)\,,\ (x_4,y_4)\,,\ \dots\,,\ (x_n,y_n)\,.$$

X の値の平均を $\overline{x}$ とおき，Y の値の平均を $\overline{y}$ とおく：

$$\overline{x}=\frac{1}{n}\sum_{k=1}^{n}x_k\ ,\qquad \overline{y}=\frac{1}{n}\sum_{k=1}^{n}y_k\ .$$

この2次元のデータにおける X と Y との共分散 c_{XY} は

$$c_{XY}=\frac{1}{n}\sum_{k=1}^{n}(x_k\,y_k)-\overline{x}\,\overline{y}\ .$$

証明　共分散の定義より，

$$\begin{aligned}c_{XY}&=\frac{1}{n}\sum_{k=1}^{n}\{(x_k-\overline{x})(y_k-\overline{y})\}=\frac{1}{n}\sum_{k=1}^{n}(x_k\,y_k-x_k\,\overline{y}-\overline{x}\,y_k+\overline{x}\,\overline{y})\\&=\frac{1}{n}\sum_{k=1}^{n}(x_k\,y_k)-\frac{1}{n}\sum_{k=1}^{n}(x_k\,\overline{y})-\frac{1}{n}\sum_{k=1}^{n}(\overline{x}\,y_k)+\frac{1}{n}\sum_{k=1}^{n}(\overline{x}\,\overline{y})\\&=\frac{1}{n}\sum_{k=1}^{n}(x_k\,y_k)-\overline{y}\,\frac{1}{n}\sum_{k=1}^{n}x_k-\overline{x}\,\frac{1}{n}\sum_{k=1}^{n}y_k+\frac{1}{n}\cdot n\,\overline{x}\,\overline{y}\\&=\frac{1}{n}\sum_{k=1}^{n}(x_k\,y_k)-\overline{y}\,\overline{x}-\overline{x}\,\overline{y}+\overline{x}\,\overline{y}\\&=\frac{1}{n}\sum_{k=1}^{n}(x_k\,y_k)-\overline{x}\,\overline{y}\ .\end{aligned}$$

（証明終り）

相関係数について次の定理が成り立つ．

定理 1.4.2　変量 X と Y との相関係数 r_{XY} について $-1\leq r_{XY}\leq 1$ ．

証明　正の自然数 n に対して，変量 X の値 x と変量 Y の値 y との順序対 (x,y) が次のように n 個あるとする：

$$(x_1,y_1)\,,\ (x_2,y_2)\,,\ (x_3,y_3)\,,\ (x_4,y_4)\,,\ \dots\,,\ (x_n,y_n)\,.$$

X の値の平均を $\overline{x}$ とおき，Y の値の平均を $\overline{y}$ とおく：

$$\overline{x}=\frac{1}{n}\sum_{k=1}^{n}x_k\ ,\qquad \overline{y}=\frac{1}{n}\sum_{k=1}^{n}y_k\ .$$

X の標準偏差および Y の標準偏差は 0 でないとする:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}\neq 0\ ,\qquad \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j-\overline{y})^2}\neq 0\ .$$

自然数 $k=1,2,3,\dots,n$ に対して，$\tilde{x}_k=x_k-\overline{x}$ ，$\tilde{y}_k=y_k-\overline{y}$ とおく．相関係数 r_{XY} の定義より，

$$r_{XY}=\frac{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}\{(x_k-\overline{x})(y_k-\overline{y})\}}{\sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\dfrac{1}{n}\displaystyle\sum_{j=1}^{n}(y_j-\overline{y})^2}}=\frac{\displaystyle\sum_{k=1}^{n}(\tilde{x}_k\,\tilde{y}_k)}{\sqrt{\displaystyle\sum_{i=1}^{n}\tilde{x}_i{}^2}\sqrt{\displaystyle\sum_{j=1}^{n}\tilde{y}_j{}^2}}\ .$$

これより，複号同順で，

$$\begin{aligned}\sum_{k=1}^{n}\left(\frac{\tilde{x}_k}{\sqrt{\displaystyle\sum_{i=1}^{n}\tilde{x}_i{}^2}}\pm\frac{\tilde{y}_k}{\sqrt{\displaystyle\sum_{j=1}^{n}\tilde{y}_j{}^2}}\right)^2&=\sum_{k=1}^{n}\left(\frac{\tilde{x}_k{}^2}{\sqrt{\displaystyle\sum_{i=1}^{n}\tilde{x}_i{}^2}^2}\pm 2\frac{\tilde{x}_k}{\sqrt{\displaystyle\sum_{i=1}^{n}\tilde{x}_i{}^2}}\frac{\tilde{y}_k}{\sqrt{\displaystyle\sum_{j=1}^{n}\tilde{y}_j{}^2}}+\frac{\tilde{y}_k{}^2}{\sqrt{\displaystyle\sum_{j=1}^{n}\tilde{y}_j{}^2}^2}\right)\\&=\frac{\displaystyle\sum_{k=1}^{n}\tilde{x}_k{}^2}{\displaystyle\sum_{i=1}^{n}\tilde{x}_i{}^2}\pm 2\sum_{k=1}^{n}\frac{\tilde{x}_k\,\tilde{y}_k}{\sqrt{\displaystyle\sum_{i=1}^{n}\tilde{x}_i{}^2}\sqrt{\displaystyle\sum_{j=1}^{n}\tilde{y}_j{}^2}}+\frac{\displaystyle\sum_{k=1}^{n}\tilde{y}_k{}^2}{\displaystyle\sum_{j=1}^{n}\tilde{y}_j{}^2}\\&=1\pm 2\frac{\displaystyle\sum_{k=1}^{n}(\tilde{x}_i\,\tilde{y}_k)}{\sqrt{\displaystyle\sum_{i=1}^{n}\tilde{x}_i{}^2}\sqrt{\displaystyle\sum_{j=1}^{n}\tilde{y}_j{}^2}}+1\\&=2\pm 2r_{XY}\ .\end{aligned}$$

$$\sum_{k=1}^{n}\left(\frac{\tilde{x}_k}{\sqrt{\displaystyle\sum_{i=1}^{n}\tilde{x}_i{}^2}}\pm\frac{\tilde{y}_k}{\sqrt{\displaystyle\sum_{j=1}^{n}\tilde{y}_j{}^2}}\right)^2\geq 0\quad\text{なので，}\quad 2+2r_{XY}\geq 0\quad\text{かつ}\quad 2-2r_{XY}\geq 0\ ,$$

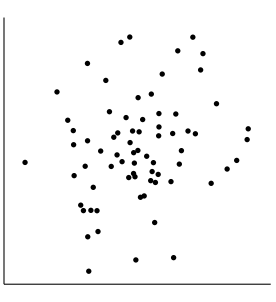
$1+r_{XY}\geq 0$ かつ $1-r_{XY}\geq 0$ ，$r_{XY}\geq -1$ かつ $r_{XY}\leq 1$ ，つまり $-1\leq r_{XY}\leq 1$ ．

（証明終り）

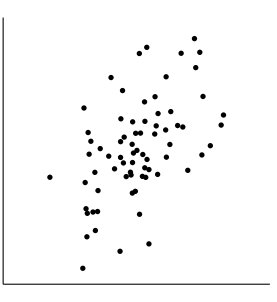
正の自然数 n に対して，変量 X の値と変量 Y の値との n 個の順序対

$$(x_1,y_1)\,,\ (x_2,y_2)\,,\ (x_3,y_3)\,,\ (x_4,y_4)\,,\ \dots\,,\ (x_n,y_n)$$

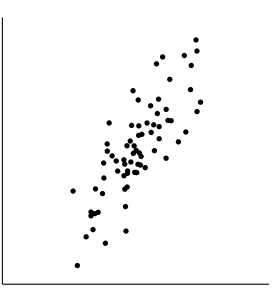
を要素とする2次元のデータがあるとします．各要素を座標平面の座標と考えて，座標平面において座標が表す点の集まりを散布図といいます．散布図と相関係数は例えば次のようになります．



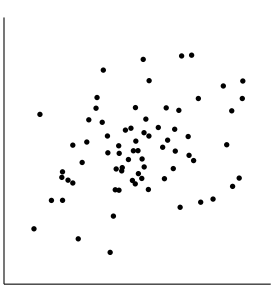
相関係数 0.2



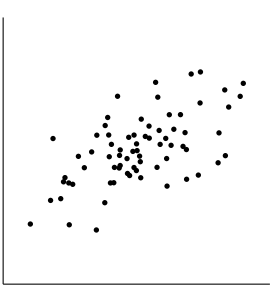
相関係数 0.5



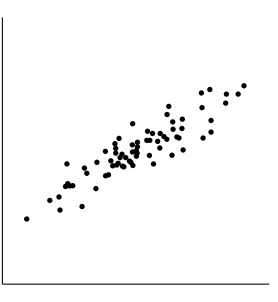
相関係数 0.8



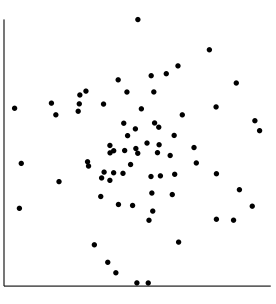
相関係数 0.3



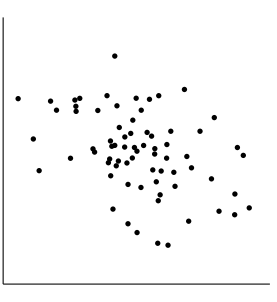
相関係数 0.6



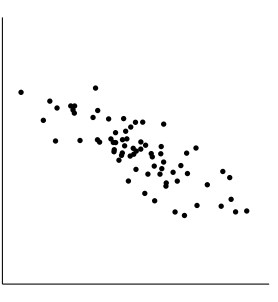
相関係数 0.9



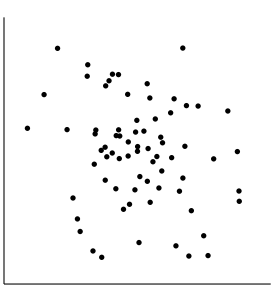
相関係数 0



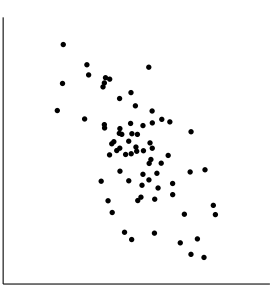
相関係数 −0.4



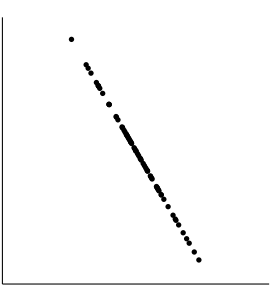
相関係数 −0.8



相関係数 −0.2



相関係数 −0.6



相関係数 −1

これらの図から分かるように，変量 X と Y との相関係数 r_{XY} について以下のことが成り立ちます．

(1)　$r_{XY}>0$ のとき，X の値が増加すると Y の値も増加する傾向がある．$r_{XY}<0$ のとき，X の値が増加すると Y の値は減少する傾向がある．

(2)　XY 座標平面において，r_{XY} の絶対値 $|r_{XY}|$ が 1 に近いほどデータを表す点はある直線の近くに分布する．

変量 X と Y との相関係数 r_{XY} について，$r_{XY}>0$ のとき X と Y とは正の相関があるといい，$r_{XY}<0$ のとき X と Y とは負の相関があるといいます．また，相関係数 r_{XY} の絶対値 $|r_{XY}|$ が 1 に近いとき X と Y との相関は強いといい，相関係数 r_{XY} の絶対値 $|r_{XY}|$ が 0 に近いとき X と Y との相関は弱いといいます．

相関係数は，座標平面において2次元のデータの散布図が直線に近いかどうかを示します．相関が弱いこと，つまり相関係数が 0 に近いことは，データの散布図が直線に近くないということを意味するだけで，必ずしも関連が薄い訳ではありません．例えば変量 X の値 x と変量 Y の値 y との対 (x,y) のデータが次のようにあるとします：

$$(-2,4)\,,\quad (-1,1)\,,\quad (0,0)\,,\quad (1,1)\,,\quad (2,4)\,.$$

このデータにおける X と Y とのデータの共分散は 0 ですから，相関係数も 0 です．つまり，このデータにおいて X と Y とは正の相関も負の相関もありません．しかし，このデータにおいて，X の値 x と Y の値 y とについて $y=x^2$ となる関係があります．

例題 1.4　4人の学生A君, B君, C君, D君のある科目の中間試験の点数 X と期末試験の点数 Y とを調べると次のようになった．

学生	A	B	C	D
中間試験の点数 X	75	67	81	73
期末試験の点数 Y	67	71	83	71

X と Y との共分散と相関係数とを求める．

X の平均は

$$\frac{1}{4}(75+67+81+73)=74\ .$$

X の分散は

$$\frac{1}{4}\{(75-74)^2+(67-74)^2+(81-74)^2+(73-74)^2\}=\frac{1}{4}(1+49+49+1)$$
$$=25\ .$$

Y の平均は

$$\frac{1}{4}(67+71+83+71)=73\ .$$

Y の分散は

$$\frac{1}{4}\{(67-73)^2+(71-73)^2+(83-73)^2+(71-73)^2\}=\frac{1}{4}(36+4+100+4)$$
$$=36\ .$$

X と Y との共分散は

$$\frac{1}{4}\{(75-74)(67-73)+(67-74)(71-73)+(81-74)(83-73)+(73-74)(71-73)\}$$
$$=\frac{1}{4}(-6+14+70+2)=20\ .$$

X と Y との相関係数は

$$\frac{20}{\sqrt{25}\sqrt{36}}=\frac{20}{30}=\frac{2}{3}\fallingdotseq 0.667\ .$$

終

問題 1.4　4人の学生A君, B君, C君, D君の身長（単位は cm） X と体重（単位は kg） Y とについて調べると次のようになりました．

学生	A	B	C	D
身長（単位は cm） X	167	159	177	161
体重（単位は kg） Y	57	61	73	61

X と Y との共分散と相関係数とを求めなさい．

---

[1] 相関係数といわれるものにはいくつか異なるものがあります．ここで述べるのはピアソンの相関係数といわれます．