5 データの相関と散布図

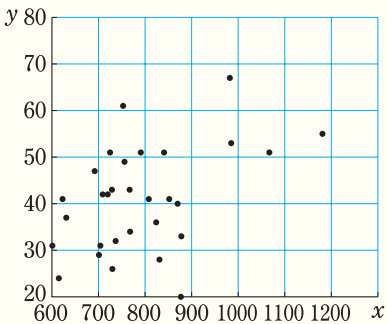
これまでは，1つの変量からなるデータの分析を行ってきたが，ここでは，2つの変量がある場合に，それらの間にどのような関係があるか調べる方法について考えていこう。

次のデータは，あるシーズンにおけるアメリカのバスケットボールリーグ30チームの3点シュートの成功本数 x（本）と勝利数 y（勝）である。

x	982	1181	870	1067	756	841	985	852	615	725	692	877	601	753	709
y	67	55	40	51	49	51	53	41	24	51	47	20	31	61	42

824	730	768	704	720	808	729	623	737	831	631	701	791	767	878
36	26	34	31	42	41	43	41	32	28	37	29	51	43	33

このデータを見やすくするために，各チームの2つの変量 x，y の値の組 (982, 67), (1181, 55), ……, (878, 33) を平面上の点として図に表すと，右のようになる。このような図を **散布図** という。



右の散布図では，x，y の一方が増加すると，他方も増加する傾向が見てとれる。このとき，2つの変量 x と y の間に **正の相関がある** という。また，一方が増加すると，他方が減少する傾向があるときは，2つの変量 x と y の間に **負の相関がある** という。

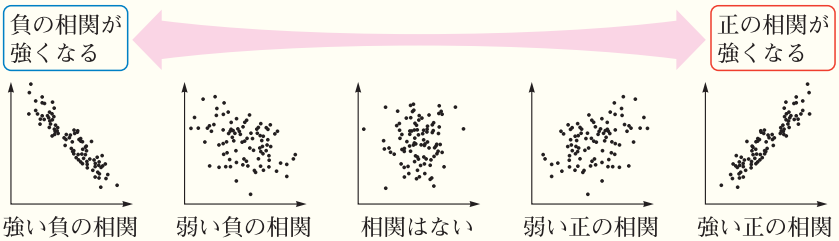
どちらの傾向も見られないときは，2つの変量 x と y の間に **相関はない** という。

▶ p.181 1

6 相関係数

次の5つの散布図のように，相関といっても，その傾向が明確にみられる場合とそうでない場合がある。

その傾向がはっきりみられるものほど，相関は強いという。



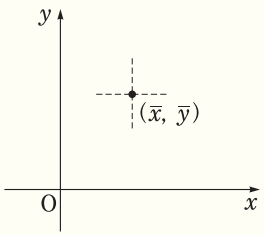
散布図で2つの変量のおよその相関はわかるが，さらに，その相関の程度を数値で表すことを考えていこう。

相関係数

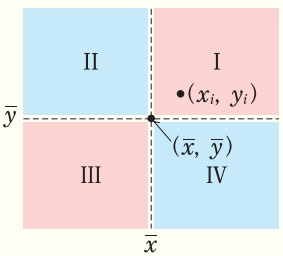
2つの変量を x，y として，それら n 個の値を

$(x_1,\ y_1),\ (x_2,\ y_2),\ \cdots\cdots,\ (x_n,\ y_n)$

で表し，x，y の平均値を，それぞれ $\overline{x}$，$\overline{y}$，標準偏差を，それぞれ s_x，s_y で表す。



右の図のように，2つの変量の平均値を座標とする点 $(\overline{x},\ \overline{y})$ を中心にして平面を4つの区域に分ける。



このとき，境界線上にない点 $(x_i,\ y_i)$ の位置と，変量 x，y のそれぞれの偏差の積

$$(x_i-\overline{x})(y_i-\overline{y})$$

の符号の関係は，次のようになる。

(1) 点 $(x_i,\ y_i)$ が区域Ⅰか Ⅲ に入るとき，$(x_i-\overline{x})(y_i-\overline{y})>0$ となる。

(2) 点 $(x_i,\ y_i)$ が区域Ⅱか Ⅳ に入るとき，$(x_i-\overline{x})(y_i-\overline{y})<0$ となる。

したがって，変量 x と y の間に正の相関があれば，区域Ⅰか Ⅲに点が多く集まるから，x と y の偏差の積の平均値

$$\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots\cdots+(x_n-\overline{x})(y_n-\overline{y})\} \quad \cdots\cdots ①$$

は正となる。

逆に，変量 x と y の間に負の相関があれば，区域Ⅱか Ⅳに点が多く集まるから，x と y の偏差の積の平均値は負となる。

また，変量 x と y の間に相関がほとんどない場合は，点が区域Ⅰ，Ⅱ，Ⅲ，Ⅳにばらつくから，x と y の偏差の積の平均値は 0 に近い値となる。

①を変量 x と y の **共分散** といい，s_{xy} で表す。

$$s_{xy}=\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots\cdots+(x_n-\overline{x})(y_n-\overline{y})\}$$

そして，共分散 s_{xy} を変量 x の標準偏差 s_x と変量 y の標準偏差 s_y の積で割った値を考える。この値を変量 x と y の **相関係数** といい，r で表す。

相関係数

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

$$\text{相関係数}=\frac{(x\text{ と }y\text{ の共分散})}{(x\text{ の標準偏差})\times(y\text{ の標準偏差})}$$

$$=\frac{\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}}{\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}}\sqrt{\frac{1}{n}\{(y_1-\overline{y})^2+(y_2-\overline{y})^2+\cdots+(y_n-\overline{y})^2\}}}$$

相関係数 r は相関の強さを測る指標であり，$-1\leqq r\leqq 1$ である。

注目 例えば，m（メートル）で測った値を cm（センチメートル）で表すなど，変量 x，y の単位のとり方を変えると，相関の強さは変わらないのに，共分散の値は変わってしまう。そこで，単位のとり方によって相関の程度を表す値が変わらないように，共分散を x，y それぞれの標準偏差の積で割った値を考えているのである。それが相関係数である。

一般に，相関係数 r の値と相関について，次のようなことがいえる。

(I)　r の値が 1 に近いほど，2 つの変量 x と y の正の相関が強い。

(II)　r の値が -1 に近いほど，2 つの変量 x と y の負の相関が強い。

(III)　r の値が 0 に近いほど，2 つの変量 x と y の相関は弱い。

例 5　次の表は，ある 5 人の生徒に行った小テスト (各 10 点満点) の英語と国語の得点である。

	A	B	C	D	E
英語	6	8	5	3	8
国語	6	7	8	5	9

この英語と国語の得点をそれぞれ x，y として，その相関係数を求める。

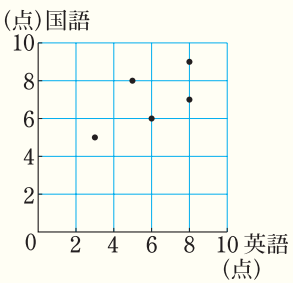
	x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
A	6	6	0	-1	0	1	0
B	8	7	2	0	4	0	0
C	5	8	-1	1	1	1	-1
D	3	5	-3	-2	9	4	6
E	8	9	2	2	4	4	4
計	30	35	0	0	18	10	9
平均	6	7	0	0	$\frac{18}{5}$	2	$\frac{9}{5}$
	$\overline{x}$	$\overline{y}$			s_x^2	s_y^2	s_{xy}

$$r=\frac{s_{xy}}{s_x s_y}=\frac{9}{5}\div\left(\sqrt{\frac{18}{5}}\times\sqrt{2}\right)$$

$$=\frac{3\sqrt{5}}{10}\ (\fallingdotseq 0.67)$$

したがって，この 5 人の英語と国語の得点の間には正の相関があるといえる。

また，散布図は右のようになる。



問 5 次の表は，ある5人の生徒に行った小テスト（各10点満点）の国語の得点 x と数学の得点 y である。x と y の相関係数を求めよ。

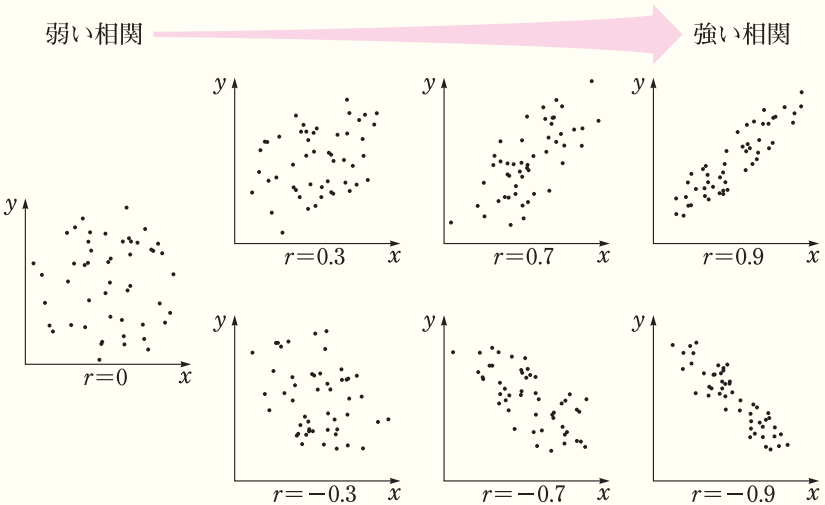
	A	B	C	D	E
x	7	6	3	9	5
y	10	7	6	4	8

▶ p.181 3

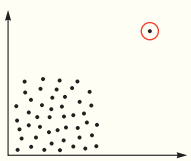
171ページのアメリカのバスケットボールリーグのチームの3点シュートの成功本数と勝利数の相関係数を実際に計算すると，約0.45になり，3点シュートの成功本数と勝利数の間に正の相関があることがわかる。

散布図と相関係数

相関係数 r は，2つの変量の相関の強さを測る指標であり，散布図と相関係数の関係はおよそ次のようになる。



注_ 右の図のように，全体の傾向から外れたデータの値があるとき，相関係数はその値に影響され，全体の傾向を示さないことがある。この外れた値のことも **外れ値** という。
このため，相関の強さを調べるときには，相関係数だけで判断するのではなく，散布図をかいてみることも大切である。



▶ p.181 2