データの分析【相関係数】　p.187～189

1 学習内容の説明 ⇒ 2 問題演習 ⇒ 3 振り返り（確認テスト・相互採点・リフレクションの記入）

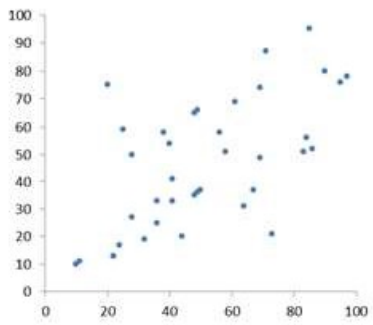
【内容目標】散らばりの度合いを表す値を求められるようになろう

□相関係数

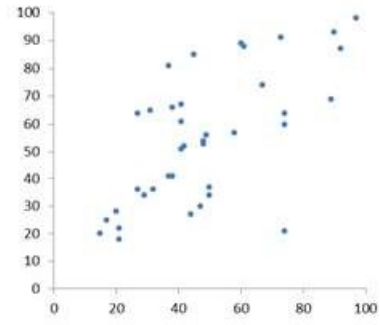
下の図は、ある２クラスの国語と英語のテストの散布図である。どの程度強い相関関係であるかを他の人に伝える際にどうすれば的確に伝えられるだろうか？

また微妙な散らばり具合の散布図の相関の強さを比較するとき、どうすれば判断することができるだろうか。

１組

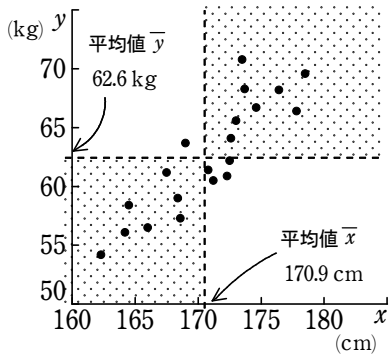


２組



２つの変量からなるデータが与えられたとき、データの値から相関関係を調べる方法として、どの程度直線的であるかを数値で表すことで比較することができる。

185 ページの身長と体重のデータでは，x と y の間に正の相関があると考えられる。このデータの散布図に，身長の平均値 $\overline{x}$，体重の平均値 $\overline{y}$ を記すと右の図のようになり，点の多くが色をつけた部分にある様子がわかる。



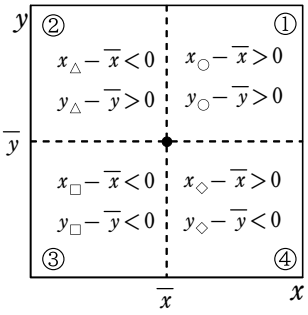
２つの変量 x，y からなるデータが与えられたとき，データの値から，x と y の間の相関を調べる方法を考えてみよう。

２つの変量 x，y からなるデータとして n 個の値の組

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

が得られているとする。

$x_1,\ x_2,\ \cdots\cdots,\ x_n$ の平均値を $\overline{x}$，$y_1,\ y_2,\ \cdots\cdots,\ y_n$ の平均値を $\overline{y}$ とし，$\overline{x}$，$\overline{y}$ を境界として，データの散布図を右の図のように ①，②，③，④ の領域に分ける。



このとき，データの散布図について，次のことが考えられる。

点の多くが ① と ③ にあるとき，x と y の間に正の相関がある。

点の多くが ② と ④ にあるとき，x と y の間に負の相関がある。

ここで，たとえば，点 $(x_1,\ y_1)$ について考えてみると，

$(x_1,\ y_1)$ が ① または ③ にあるとき　$(x_1-\overline{x})(y_1-\overline{y})>0$，

$(x_1,\ y_1)$ が ② または ④ にあるとき　$(x_1-\overline{x})(y_1-\overline{y})<0$

である。

⟹　x と y の間の相関を調べるのに，

x の偏差と y の偏差の積 （**共分散**） を用いると相関関係を調べられる

データの分析【相関係数】　p.187～189

x の偏差と y の偏差の積について，その平均値

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

を x と y の **共分散** といい，s_{xy} で表す。x と y の間に，正の相関があるとき共分散は正となり，負の相関があるとき共分散は負となる。

相関の強弱をみるために，共分散 s_{xy} を，x の標準偏差 s_x と y の標準偏差 s_y の積 s_xs_y で割った値を考える。この値を x と y の **相関係数** といい，r で表す。

相関係数を母集団で計算したときのギリシャ文字 ρ(ロー)の英語表記 rho の頭文字

相関係数 r（correlation coefficient）

$$r=\frac{s_{xy}}{s_xs_y}$$

$\frac{(x\text{ と }y\text{ の共分散})}{(x\text{ の標準偏差})\times(y\text{ の標準偏差})}$ 【計算方法１】

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

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

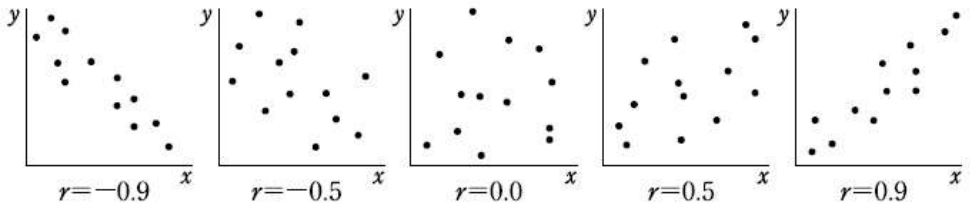
$\frac{(x\text{ と }y\text{ の総和})}{\sqrt{(x\text{ の総和})\times(y\text{ の総和})}}$ 【計算方法２】

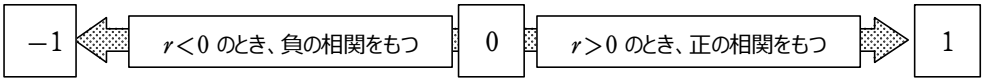
相関係数 r については，

$-1\leqq r\leqq 1$ であることが知られている。

r の値は，正の相関が強いほど 1 に近づき，負の相関が強いほど -1 に近づく。

また，相関がないとき，r の値は 0 に近い値をとる。





【参考】

相関係数	相関関係
0	相関がない
0.0～±0.2	ほとんど相関がない
±0.2～±0.4	やや相関がある(低い相関)
±0.4～±0.7	相関がある
±0.7～±0.9	強い相関がある(高い相関)
±0.9～±1.0	きわめて強い相関がある
±1.0	完全な相関

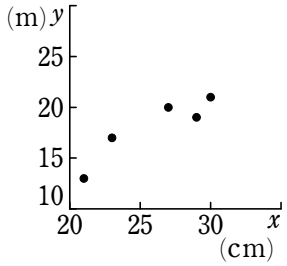
ちなみに前述の１組、２組の散布図の相関係数は 0.65 と 0.62 となるので、**微妙ながら１組の方が正の相関が強い**と言うことがわかる。

データの分析【相関係数】　p.187～189

例１１）　次の表は，同じ種類の５本の木について，根もとの太さ x(cm) と高さ y(m) を測定した結果である。

x と y の相関係数 r を求めよう。

	①	②	③	④	⑤
x	21	27	29	23	30
y	13	20	19	17	21



散布図は右の図のようになる。

x，y のデータの平均値は

$\overline{x}=\frac{1}{5}\times 130=26$，$\overline{y}=\frac{1}{5}\times 90=18$

	x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})(y-\overline{y})$	$(x-\overline{x})^2$	$(y-\overline{y})^2$
①	21	13	−5	−5	25	25	25
②	27	20	1	2	2	1	4
③	29	19	3	1	3	9	1
④	23	17	−3	−1	3	9	1
⑤	30	21	4	3	12	16	9
計	130	90			○ 45	△ 60	□ 40
平均	26	18			9	12	8

$$\frac{(x\text{と}y\text{の総和})}{\sqrt{(x\text{の総和})\times(y\text{の総和})}}=\frac{○}{\sqrt{△\times□}}$$ 【計算方法２】を利用

上の表から，相関係数 r は　$r=\frac{45}{\sqrt{60\times 40}}\fallingdotseq 0.92$　　終

相関係数が正で１に近いから，x と y には強い正の相関があると考えられる。

【参考】相関表

散布図にすると重なる点がある場合、正しく表すことができない。

またデータの数が多すぎると点だらけになってしまう。

そのようなときは相関表を用いると良い。

	0～20	20～40	40～60	60～80	80～100	計
80以上100未満			2	3	6	11
60～80		1	3	14	4	22
40～60	2		6	12	2	22
20～40		4	2			6
0～20	5	1		1		7
計	7	6	13	30	12	68