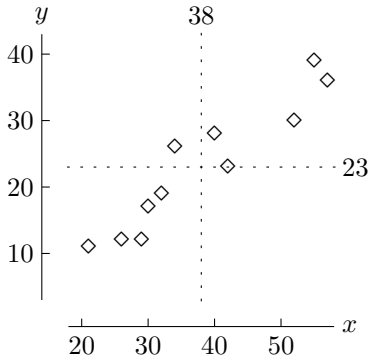
5　相関

5.1　散布図

次のデータは種目 X の記録 x（kgw）と種目 Y の記録 y（m）である。

x	21	26	29	30	32	34	40	42	52	55	57
y	11	12	12	17	19	26	28	23	30	39	36

点 (x_i, y_i) を座標平面上にとったものを**散布図**または相関図という。



x が増加（減少）すると y も増加（減少）する傾向がみられるとき，**正の相関**をもつといい，反対に x が増加（減少）すると y は減少（増加）する傾向がみられるとき，**負の相関**をもつという。上のデータは正の相関をもつ。

5.2　共分散

共分散（covariance）　x の偏差と y の偏差の積の平均値のことを共分散といい，s_{xy} と表す。$s_{xy} > 0$ のとき正の相関，$s_{xy} < 0$ のとき負の相関となる。

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

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

上のデータの共分散を定義式 (1) で求める。平均値は $\bar{x} = 38$，$\bar{y} = 23$ である。

$$s_{xy} = \frac{(21 - 38)(11 - 23) + \cdots + (57 - 38)(36 - 23)}{11} = 102$$

共分散（2）　共分散については公式 $s_{xy} = \overline{xy} - \bar{x}\bar{y}$ が知られている。

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum (x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y}) \\ &= \frac{1}{n}\sum x_i y_i - \bar{x}\cdot\frac{1}{n}\sum y_i - \frac{1}{n}\sum x_i \cdot \bar{y} + \frac{1}{n}\sum \bar{x}\bar{y} \\ &= \overline{xy} - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\ &= \overline{xy} - \bar{x}\bar{y} \end{aligned}$$

定義式 (1) のかわりに，(2) を用いてもよい。

$$s_{xy} = \frac{x_1y_1 + \cdots + x_ny_n}{n} - \bar{x}\bar{y}, \qquad s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \bar{x}\bar{y} \tag{2}$$

上のデータの共分散を公式 (2) で求める。

$$s_{xy} = \frac{21 \cdot 11 + \cdots + 57 \cdot 36}{11} - 38 \cdot 23 = 102$$

共分散（3）　公式 (2) をさらに変形した (3) を用いることもできる。

$$s_{xy} = \frac{n\left(x_1y_1 + \cdots + x_ny_n\right) - \left(x_1 + \cdots + x_n\right)\left(y_1 + \cdots + y_n\right)}{n^2}$$

$$s_{xy} = \frac{n\sum x_iy_i - \sum x_i \sum y_i}{n^2} \tag{3}$$

上のデータの共分散を公式 (3) で求める。

$$\begin{aligned} s_{xy} &= \frac{11\left(21 \cdot 11 + \cdots + 57 \cdot 36\right) - \left(21 + \cdots + 57\right)\left(11 + \cdots + 36\right)}{11^2} \\ &= 102 \end{aligned}$$

分散と共分散　分散は $s_x^2 = \overline{x^2} - \bar{x}^2$，共分散は $s_{xy} = \overline{xy} - \bar{x}\bar{y}$ なので，x と x の共分散は x の分散に等しい。

$$s_{xx} = \overline{xx} - \bar{x}\bar{x} = s_x^2$$

5.3　相関係数

相関係数（correlation coefficient）　x，y の共分散を各変数の標準偏差で割った値を相関係数といい，r_{xy} と表す。

$$\begin{aligned} r_{xy} = \frac{s_{xy}}{s_xs_y} &= \frac{\dfrac{1}{n}\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\dfrac{1}{n}\sum\left(x_i - \bar{x}\right)^2}\sqrt{\dfrac{1}{n}\sum\left(y_i - \bar{y}\right)^2}} \\ &= \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}\sqrt{\sum\left(y_i - \bar{y}\right)^2}} \end{aligned}$$

Cauchy-Schwarz の不等式 $\left|\sum f_ig_i\right| \leqq \sqrt{\sum f_i^2}\sqrt{\sum g_i^2}$ において，$f_i = x_i - \bar{x}$，$g_i = y_i - \bar{y}$ とすると，$|r_{xy}| \leqq 1$ が得られる。相関係数は常に次の範囲内にある。

$$-1 \leqq r_{xy} \leqq 1$$

相関係数が 1 または -1 に近いときは強い相関，0 のときは無相関となる。次のように判定することが多い。境界値（0.2，0.4，0.7）は目安である。

相関係数	強弱・正負	相関係数	強弱・正負
$-1.0 \sim -0.7$	強い負の相関	$0.7 \sim 1.0$	強い正の相関
$-0.7 \sim -0.4$	中程度の負の相関	$0.4 \sim 0.7$	中程度の正の相関
$-0.4 \sim -0.2$	弱い負の相関	$0.2 \sim 0.4$	弱い正の相関
$-0.2 \sim 0.2$	無相関		

相関係数の求め方（1）　分散や共分散を定義式 (1) によって求める。

変数	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
a.	**21**	**11**	-17	-12	289	144	204
b.	**26**	**12**	-12	-11	144	121	132
c.	**29**	**12**	-9	-11	81	121	99
d.	**30**	**17**	-8	-6	64	36	48
e.	**32**	**19**	-6	-4	36	16	24
f.	**34**	**26**	-4	3	16	9	-12
g.	**40**	**28**	2	5	4	25	10
h.	**42**	**23**	4	0	16	0	0
i.	**52**	**30**	14	7	196	49	98
j.	**55**	**39**	17	16	289	256	272
k.	**57**	**36**	19	13	361	169	247
合計	418	253	0	0	1496	946	1122
平均	38	23	0	0	136	86	102

平均は $\bar{x} = 38$，$\bar{y} = 23$，分散は $s_x^2 = 1496/11 = 136$，$s_y^2 = 946/11 = 86$，共分散は $s_{xy} = 1122/11 = 102$ だから，相関係数は

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{102}{\sqrt{136}\sqrt{86}} = 0.943$$

または

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{1122}{\sqrt{1496}\sqrt{946}} = 0.943$$

である。

相関係数の求め方（2）　分散や共分散を公式 (2) によって求める。

変数	x	y	x^2	y^2	xy
a.	**21**	**11**	441	121	231
b.	**26**	**12**	676	144	312
c.	**29**	**12**	841	144	348
d.	**30**	**17**	900	289	510
e.	**32**	**19**	1024	361	608
f.	**34**	**26**	1156	676	884
g.	**40**	**28**	1600	784	1120
h.	**42**	**23**	1764	529	966
i.	**52**	**30**	2704	900	1560
j.	**55**	**39**	3025	1521	2145
k.	**57**	**36**	3249	1296	2052
合計	418	253	17380	6765	10736
平均	38	23	1580	615	976

平均は $\bar{x} = 38$，$\bar{y} = 23$，分散は $s_x^2 = 1580 - 38^2 = 136$，$s_y^2 = 615 - 23^2 = 86$，共分散は $s_{xy} = 976 - 38 \cdot 23 = 102$ だから，相関係数は

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{102}{\sqrt{136}\sqrt{86}} = 0.943$$

である。

相関係数の求め方（3）　分散や共分散を公式 (3) によって求める。「相関係数の求め方（2）」の表を利用すると，平均は $\bar{x} = 418/11 = 38$，$\bar{y} = 253/11 = 23$，分散は $s_x^2 = (11 \cdot 17380 - 418^2)/11^2 = 16456/11^2 = 136$，$s_y^2 = (11 \cdot 6765 - 253^2)/11^2 = 10406/11^2 = 86$，共分散は $s_{xy} = (11 \cdot 10736 - 418 \cdot 253)/11^2 = 12342/11^2 = 102$ だから，相関係数は

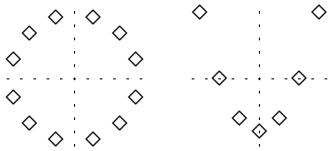
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{102}{\sqrt{136}\sqrt{86}} = 0.943$$

または

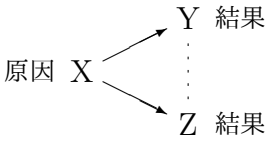
$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{12342}{\sqrt{16456}\sqrt{10406}} = 0.943$$

である。

相関関係と因果関係　相関とは，増加や減少という直線的な関係を表すものである。下の散布図が表す 2 つの変数には何らかの強い関係があるが，相関はない。



また，相関関係と因果関係は同じではない。因果関係を示すためには，その現象の全体像を解明する必要がある。X を共通の原因として Y と Z が起こる場合，X と Y，そして X と Z が互いに相関をもつことから，Y と Z にも間接的な相関が生じることがある。しかし，Y と Z の関係は因果関係とはいえない。このように因果関係とはいえない相関関係のことを**見かけの相関**または擬似相関という。[*i]　　*i



[*i]共通の原因となりやすいものに，天候，年齢，集団の規模等がある。

参考文献

統計学入門（基礎統計学）
東京大学教養学部統計学教室（編）東京大学出版会 978-4-13-042065-5

統計学
久保川 達也 （著） 東京大学出版会 978-4-13-062921-8

はじめての統計学
道家 暎幸（共著） コロナ社 978-4-339-06113-0

確率統計 新版 （新版数学シリーズ）
岡本 和夫 （ほか著） 実教出版 978-4-407-32171-5

統計学序論 改訂版
山本 義郎 （著） 東海大学出版部 978-4-486-02133-9

確率統計（高専テキストシリーズ）
上野 健爾（監修） 森北出版 978-4-627-05561-2

基本統計学 第４版
宮川 公男 （著） 有斐閣 978-4-641-16455-0

新統計入門
小寺 平治 （著） 裳華房 978-4-7853-1099-8

Schaum's Outline of Introduction to Probability and Statistics
Seymour Lipschutz （著） McGraw-Hill Education 978-0-07-176249-6

A Dictionary of Statistics
Graham Upton （著） Oxford Univ Pr 978-0-19-967918-8

www5e.biglobe.ne.jp/~emm386/statistics/