相関係数

1　因果関係と相関関係

原因となる変数 x があり，その結果生じる変数 y があれば，それらの変数間には因果関係があるという．このとき，x を**独立変数**とか**説明変数**と呼び，y を**目的変数**や**従属変数**などと呼ぶ．このような関係にある変数であれば，前述の最小 2 乗法や後述する回帰分析など数式で表せる関数系の依存関係を議論することができる．

しかし，変数 x と y の両方が何らかの原因の結果であり，独立変数ではない場合には，その間に関係があることがわかっても，数式で因果関係を表すことはできない．このような 2 つの変数においては，その間に「相関」があるという．そのような 2 つの変数の間に潜む線形な関係の強弱を表すものが相関係数である．例えば，夏の気温とアイスクリームやビールの売り上げには相関があるとされており，製造現場や販売店ではその関係に基づいて製品の製造・出荷また調達などを行っている．JIS Z 8101-1 によると，相関とは「二つの確率変数の分布法則の関係。多くの場合，線形関係の程度を指す。」と定義されている．

ただ単に相関係数というと，それは「ピアソンの積率相関係数」を指す．以下では，その相関係数について説明する．

2　相関係数の導出

図 1 (a) に示すように散布図としてあらわされる変量 x と y があるとする．このとき，以下のように変数変換を行って改めて図 1 (b) に示すような関係に描き直してみる．

$$X = x - \bar{x}, \quad Y = y - \bar{y}$$

そうすると，以下のように第 I から第 IV 象限にあるデータの関連が明らかとなる．

- 第 I 象限：　$X > 0,\ Y > 0,\ XY > 0$
- 第 II 象限：　$X < 0,\ Y > 0,\ XY < 0$
- 第 III 象限：　$X < 0,\ Y < 0,\ XY > 0$
- 第 IV 象限：　$X > 0,\ Y < 0,\ XY < 0$

象限 I〜IV の各領域において積 XY の和を求め，それを $\sum_I XY$ のように標記すると，

$$\sum XY = \sum_{I} XY + \sum_{II} XY + \sum_{III} XY + \sum_{IV} XY$$

のように表せせる．図 1 (b) に示すような分布の場合，データの個数的には第 I 象限と第 III 象限に多く含まれているため，$\sum_I XY$ と $\sum_{III} XY$ が支配的となり，$\sum XY > 0$ と考えられる．第 II 象限と第 IV 象限にデータの個数

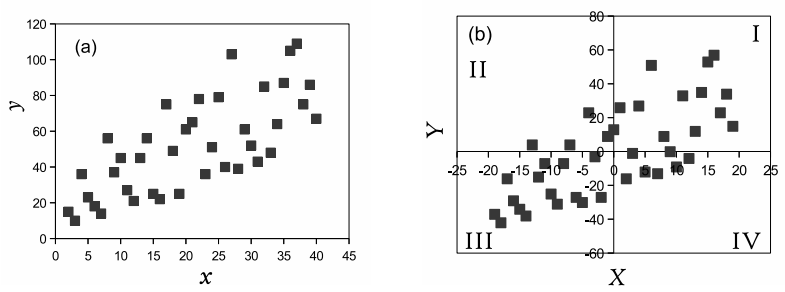

図 1: 散布図の例．(a) のグラフを平均値を基準にプロットしなおしたものが (b)．

が多い場合には $\sum XY < 0$，均等に 4 つの象限に分布していると $\sum XY \simeq 0$ となるであろう．そすると，$\sum XY$ の値は 2 つの変数 x と y の間の関係を表す尺度として使えることがわかる．すなわち，

$$\sum XY = \sum (x - \bar{x})(y - \bar{y})$$

を見れば，傾向がわかることになる．しかし，変数 x と y はともに次元を持つ量であるのが普通であるため，この積は変数の大きさの影響を受けるとともに，変数の個数の多少によっても変化する．そこで，残差を用いて以下のように規格化する．

$$\begin{aligned} r &= \sum \left(\frac{x_i - \bar{x}}{\sqrt{\sum (x_i - \bar{x})^2}} \right) \left(\frac{y_i - \bar{y}}{\sqrt{\sum (y_i - \bar{y})^2}} \right) \\ &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \end{aligned} \tag{1}$$

これまでも見てきたように残差は通常 S の記号を用いて表すので，以下のように残差を用いて表すと，

$$\begin{aligned} S_x &= \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 \\ S_y &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 \\ S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} \end{aligned}$$

式 (1) は以下のようにあらわせる．

$$r = \frac{S_{xy}}{\sqrt{S_x S_y}} \tag{2}$$

この r を相関係数という．ここで，Caushy-Schwarz の不等式，

$$\left(\sum x_i^2 \right) \left(\sum y_i^2 \right) \geq \left(\sum x_i y_i \right)^2 \tag{3}$$

より，$S_{xy}^2 \leq S_x S_y$ となるので，相関係数 r は，

$$-1 \leq r \leq 1$$

でなければならない．

$r > 0$ のとき，x もしくは y のどちらかが増加すると，もう一方も増加する傾向があるので，**正の相関**といい，反対にどちらかが増加するともう一方が減少する関係にある時は $r < 0$ となり，**負の相関**という．$|r| \ll 1$ のときには，$(x,\ y)$ の組は互いに無関係に散らばる．この状態を無相関という．

【演習】
相関係数を求める演習を行う．

3　相関の強さ

相関係数とは，2 つの変数の間のなんらかの関係を示す以上のものではなく，因果関係とは直接結びつかない．しかし，相関係数の大きさによって，それら 2 つの変数間の関係の強さについては議論できる．一般的には，表 1 のように強さは判断されている．しかし，実際に相関があるかどうかについては，厳密には後述の相関分析を行う必要がある．

表 1: 相関係数と相関の強さ

相関係数	相関の強さ
$\|r\| < 0.2$	ほとんど相関はない
$0.2 < \|r\| < 0.4$	弱い相関がある
$0.4 < \|r\| < 0.7$	相関がある
$0.7 < \|r\| < 0.9$	強い相関がある
$\|r\| > 0.9$	ほぼ完全な相関がある

4 偏相関係数

繰り返しになるが，相関があるといって 2 つの変数間に何らかの因果関係があるということを直接説明しているわけではない．$(x,\ y)$ の組に相関があるということは，x が原因で y が変化する，y が原因で x が変化する，もしくは，何か別の要因で x と y が対応する変化をする，のどれかであることを示しているに過ぎない．そこで，相関においては，何か別の要因が働いて見かけ上相関が表れていることも考えられる．そこで，2 つの変数の相関が第 3 の変数によって変化するような場合に，第 3 の変数の影響を取り除いて求めた相関係数を**偏相関係数**といい，以下の式で表される．

$$r_{xy\cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}} \tag{4}$$

ここで，式 (4) において，r の添え字 $xy\cdot z$ とは，変数 z の影響を取り除いた x と y の間の相関（偏相関係数）を表し，r_{xy} などは 2 つの変数間の相関係数を表している．

【実際の事例】
ネット上には様々なデータが統計の学習用に公開されている．今回は，その中の店舗の売り上げに関する情報を利用して相関について考える．

表 2 のようなデータがある．これはあるチェーン店の 15 の店舗における各種の情報と売り上げをまとめたもの[1]である．ここから売り上げに寄与する要因を分析することができるが，例えば，店舗内の品数と売り上げについて見てみよう．

表 2: あるチェーン店の各種情報

支店名	通行人	最寄駅からの時間	店舗面積	駐車台数	従業員数	商品数	売り上げ
三条	716	25	44	16	7	125	78
京都南	2208	30	25	8	3	132	34
長岡京	1880	3	68	18	10	110	145
生駒	1416	20	30	10	5	70	51
高槻	904	10	67	32	10	82	98
枚方	1850	3	66	10	10	82	115
池田	1039	15	52	15	7	82	75
東大阪	2394	1	113	50	20	125	258
堺	711	12	30	12	7	102	70
八尾	738	10	39	10	7	70	65
和歌山	1322	11	60	23	8	72	82
宝塚	793	18	34	10	3	97	32
西宮	1733	3	96	40	10	145	190
西神	1569	4	55	28	10	92	168
加古川	1770	6	80	32	8	80	195

[1] http://mo161.soci.ous.ac.jp/@d/DoDStat/DataList/indexj.html

店舗内の商品数と売り上げの関係をグラフにすると図 2 のようになり，商品数と売り上げの相関係数を求めると 0.31 となる．弱い相関があることがその値からは認められるが，表を見ただけでもある程度わかるように，通行人や店舗面積，駐車台数などと売り上げには正の相関があると思われるし，また，最寄駅からの時間（分）とも負の相関があるように見える．そこで，例えば，商品数に大きく関係すると思われる店舗面積（任意単位）の影響を除いた偏相関係数を求めると，0.12 と小さな偏相関係数となり，商品数は売り上げとほとんど関係が無いことになってしまう．

このように，見かけの相関があるために実際の影響を読み間違うこともあるので，分析は慎重に行う必要がある．なお，今回の事例のように複数の要因が関与していると思われる場合には後述の重回帰分析を行い，どの要因がどの程度寄与しているのかを詳細に調べる手法もある．

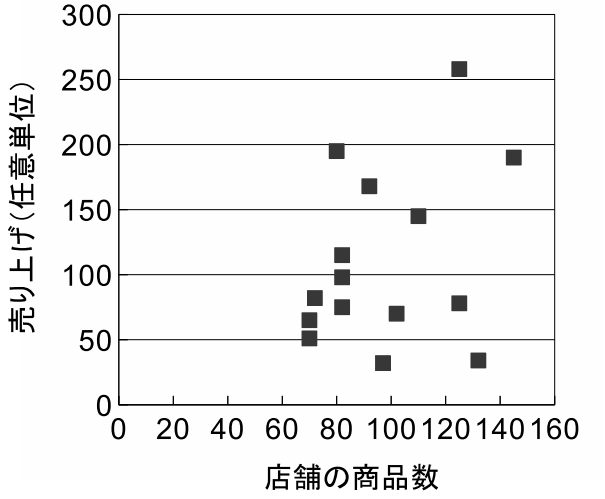

図 2: 商品数と売り上げの関係