2024 年度秋学期　統計学　第６回

データの関係を知る (1) ー相関関係

多変量データと多変量解析

第４回の講義で,「データの分布」について説明しました。「(測定対象や現象が) 分布する」とは,「ある測定対象や現象から得られる数量が大小ばらばらである」という意味です。例えば,「日本人男性の身長」は分布する，ということができます。この例での「身長」のように，大小いろいろな値になる数量のことを**変量**といいます。統計データ解析とは，一言でいえば，分布している変量から情報を引きだす手法ということができます。

世の中には２つ以上の変量で表現されるデータもたくさんあります。例えば試験の点数の場合でも，一人の人の成績は数学，英語，…といった複数の科目の点数（変量）の組み合わせで評価されます。このように，ひとつの個体（人など）が複数の変量の組み合わせで表されているデータを**多変量データ**といい，多変量データの分布を取り扱う統計手法を**多変量解析**といいます。この講義では，多変量解析の中で一番基本的な「相関分析」「回帰分析」について説明します。相関とは，２つの変量の間の関連のしかたをとらえる考え方です。

今日は，まず相関の考え方について説明し，さらに２つの項目についてのの分布の関係をグラフで表現する「散布図」，２つの量のばらつきの関係を表す「共分散」「相関係数」を説明します。さらに，表題のように，一見関係のないことがらに相関関係があるように見えるとき，その構造を分析するための,「層別」の考え方と「偏相関係数」を概説します。

相関関係と散布図

相関関係とその正負

「各県について，人口と店の数」「日本の各都市について，緯度と年平均気温」などのように，２つの変量からなるデータを考えてみましょう。

例えば,「人口と店の数」では，人口が多い町では店の数も多い傾向があるでしょうし,「緯度と気温」では，緯度が高くなると気温が低くなる傾向があるでしょう。これらはあくまで「傾向」であって，店の数が人口だけで決まったり，気温が緯度だけで決まるわけではありません。しかし，そのような傾向があるのは確かです。

このような,「変量どうしの，互いの増減の傾向の関係」を**相関関係**といいます。「人口と店の数」のように,「人口が多いと店の数も多い」という関係を**正の相関関係**といい,「緯度と気温」のように「緯度が高いと気温は低い」という関係を**負の相関関係**といいます。一般的に説明するために，上の例で出てきた「人」「県」「都市」など，数値の組が対応づけられている実体を，以後**個体**とよぶことにします。

散布図

ひとつの変量の分布を目に見えるように表現するために，ヒストグラムを用いることを第４回の講義で説明しました。これに対して，多変量データの分布を目に見えるように表現するのに用いられるのが**散布図**です。

表１は，日本のいくつかの都市の緯度と年平均気温[1]を表しています。このデータは，各都市が緯度と

[1]日本列島大地図館（小学館）［理科年表より転載］より

表 1: 日本の都市の緯度と気温

地名	緯度（度）	気温（℃）
札幌	43.05	8.0
青森	40.82	9.6
秋田	39.72	11.0
仙台	38.27	11.9
福島	37.75	12.5
宇都宮	36.55	12.9
水戸	36.38	13.2
東京	35.68	15.3
新潟	37.92	13.1
長野	36.67	11.4
静岡	34.97	16.0
名古屋	35.17	14.9
大阪	34.68	16.2
鳥取	35.48	14.4
広島	34.40	15.0
高知	33.55	16.3
福岡	33.92	16.0
鹿児島	31.57	17.3
那覇	26.20	22.0

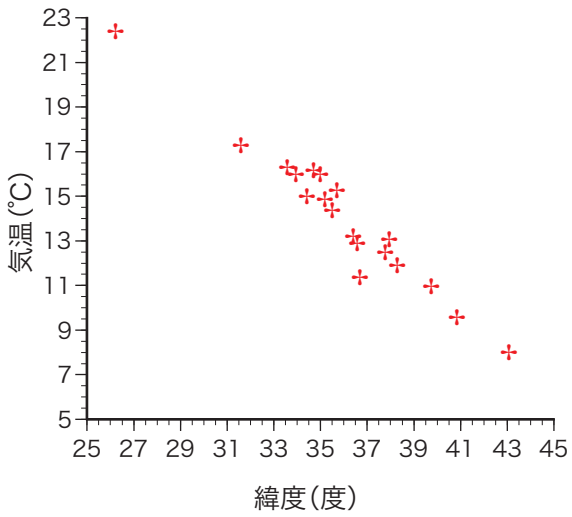


図 1: 散布図：緯度と気温の関係

気温の２つの変量で表されている多変量データです。このデータの分布を目に見えるように，緯度と気温の２つの変量をそれぞれ横軸・縦軸とし，各都市を対応する緯度・気温の位置に配置します。例えば，札幌市は北緯 43.05 度，年平均気温 8.0 ℃ですから，横軸 43.05，縦軸 8.0 の位置に印をつけます。このようにして個体（ここでは都市）を配置した，図 1 のような図を散布図といいます。この場合は変量が２つなので，散布図は横軸･縦軸でできる平面になります。変量が３つ以上になると軸も３つ以上になりますが，この場合も紙の上に描けないだけで，理屈には違いはありません。

相関関係の強弱

図 1 の散布図を見ると，一見して各都市がほぼ直線に沿って並んでおり，「緯度が高（低）いと気温が低（高）い」という負の相関関係が見てとれます。このように，負の相関関係は，散布図上では右下がりの直線上にデータが分布するように表現されます。

図 2 に，さまざまな散布図を示します。(b) は，「緯度と気温」の例と同じ，負の相関関係です。(a) と (c) は，散布図上の点がどちらも右上がりに並んでおり，どちらも正の相関関係を表します。(a) と (c) を比べると，(a) は，ほぼ右上がりの一直線上に点がならんでおり，２つの変量の関係がはっきりしています。これを「強い相関関係」といいます。一方 (c) は，右上がりにならんではいるものの，数値のばらつきが大きくなっていて，２つの変量の関係がそれほどはっきりしていません。これを「弱い相関関係」といいます。また，(d) は，右上がりでも右下がりでもなく，２つの変量の間に増減の傾向がありません。これは「相関関係がない」あるいは「無相関」といいます。

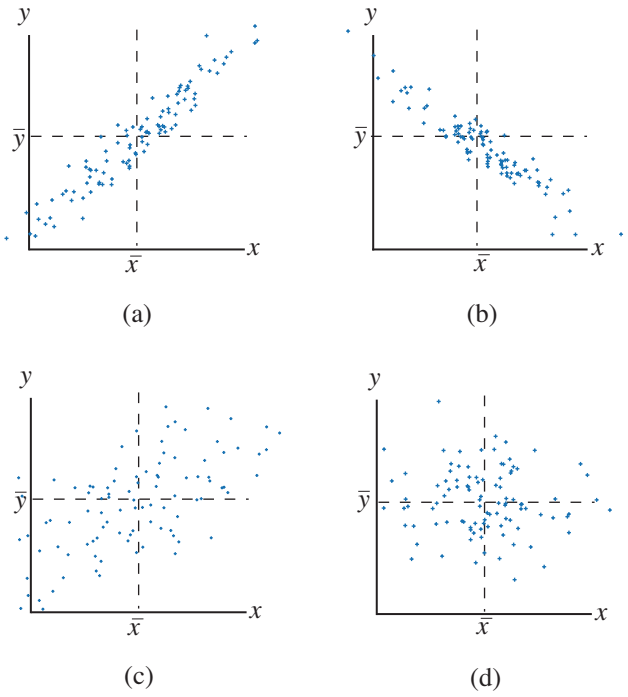


図 2: 相関関係の正負と強弱

共分散と相関係数

相関関係の正／負，強い／弱いを，数値で表すにはどうしたらよいでしょうか？ これを表すのが相関係数です。データが $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ の n 組であるとき，x と y との相関係数 r_{xy} は

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

で表されます。上の式の中央の部分で，分母は，x, y それぞれの標準偏差の積です。分子は，x, y のそれぞれの偏差を同時に平均したもので，**共分散**といいます。

共分散の意味

共分散の意味を，図 3 で考えてみましょう。散布図の平面を，x の平均および y の平均を境にして四分割します。各領域で，$(x_i - \bar{x})(y_i - \bar{y})$ の値を考えてみます。

（イ）では，$x_i - \bar{x} > 0, y_i - \bar{y} > 0$ で，$(x_i - \bar{x})(y_i - \bar{y}) > 0$ であり，(x_i, y_i) が右上に行くほどこの積の値は大きくなります。また，（ハ）では $x_i - \bar{x} < 0, y_i - \bar{y} < 0$ でやはり $(x_i - \bar{x})(y_i - \bar{y}) > 0$ であり，(x_i, y_i) が左下に行くほどこの積の値が大きくなります。これに対して，（ロ）や（ニ）では $(x_i - \bar{x})(y_i - \bar{y}) < 0$ となります。

では，図 2 に示されているそれぞれの分布で，$\sum_i (x_i - \bar{x})(y_i - \bar{y})$ の値はどうなるでしょうか？ (a) の場合は，個体は先の図 3 の（イ）（ハ）の部分に多く分布していますから，正の大きな値になります。一方，(b) の場合は，個体は（ロ）（ニ）の部分に多く分布していますから，負で絶対値の大きな値になります。(c) の場合は，個体は（イ）（ハ）の部分に多く分布してはいますが，(a) と比べると（ロ）（ニ）

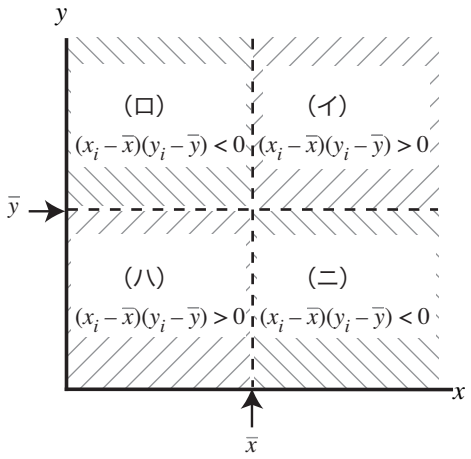


図 3: 共分散の概念

の部分にもかなり分布しています。（イ）（ハ）の部分と（ロ）（ニ）の部分とで，$(x_i-\bar{x})(y_i-\bar{y})$ の値が打ち消し合うので，$\sum_i(x_i-\bar{x})(y_i-\bar{y})$ の値は，正の数ではありますが (a) の場合よりも小さくなります。さらに，(d) の場合は，個体は（イ）（ロ）（ハ）（ニ）のすべての部分にほぼ等しく分布しているので，打ち消しあって 0 に近い値になります。

この $\sum_i(x_i-\bar{x})(y_i-\bar{y})$ を，分布しているデータのサイズ n に影響されないように，n で割って「合計」でなく「平均」にしたものが共分散です。つまり正の相関があるとき正の値，負の相関のとき負の値，どちらでもないときは 0 に近い値になります。

共分散と相関係数の関係

相関係数は共分散を x, y それぞれの標準偏差の積で割ったものとなっていますが，これは図 4 の左右の分布で相関係数が同じになるようにするためです。図 4 の左右は，ばらつきは異なっていますが，相関の強さは同じです。なお，相関係数は -1 から 1 の範囲の値をとり，1 がもっとも強い正の相関，-1 がもっとも強い負の相関，0 は相関がないことをあらわします。

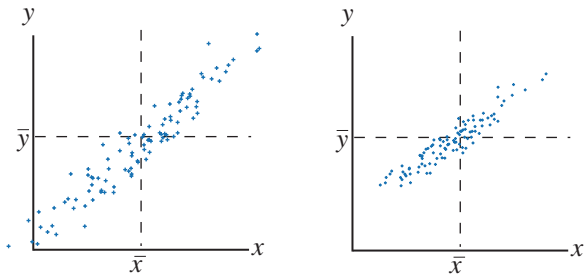
中くらいの強さの相関関係とは

ところで，図 5 は，(a) は相関係数 0.5 の散布図の例，(b) は相関係数 0.7 の散布図の例です。これを見てもわかるように，相関係数 0.5 は中くらいの強さの相関ではなく，0.7 くらいで中くらいの強さの相関になります。このようになる理由と「中くらい」の意味は，次回の回帰分析についての講義で説明します。

相関係数を計算する

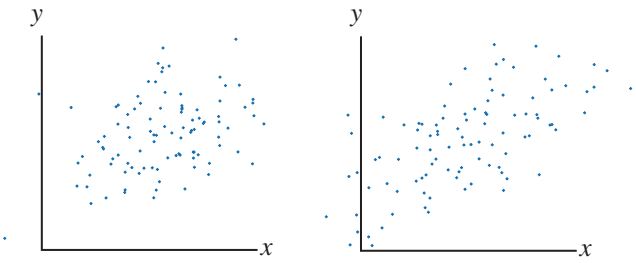
表 2 は，表 1 のデータのうち，長野〜鹿児島の数値を使って，相関係数を計算したものです。「緯度（気温）の分散 = ［緯度（気温）の偏差］^{2}の和/データサイズ」，「共分散 = ［緯度の偏差 × 気温の偏差］の和/データサイズ」，「相関係数 = 共分散/［$\sqrt{\{緯度の分散\}} \times \sqrt{\{気温の分散\}}$］」です。

現在では，こんなふうに手計算で相関係数を求めることは，まずありません。パソコンで表計算ソフトウェアや統計ソフトウェアを用いれば，数値を並べるだけで相関係数が求められます。しかし，計算の途中を見ないで結果だけが得られるため，数値を入力し間違えるなどの単純なミスに気づかず，大失



(a) (b)

図 4: 同じ相関係数をもつ分布



(a) (b)

図 5: 相関係数と散布図

敗をする可能性があります。

それを防ぐには，まず，結果をよく見ることです。今回の計算の場合，相関係数は -0.844 と求められました。これは「強い負の相関」を表していて，散布図で見た結果と一致しています。これがもし，相関係数を求めた結果が正の数であれば，明らかに計算が間違っています。

また，時々はこのような表を作って，計算の過程が明らかになるように計算してみて，同じ結果が得られるかを確かめることをお勧めします。私は，研究で用いる計算でも，必ず一部はこのようにして計算しています。

ちょっと問題

次の記述について，何がどうおかしいか説明してください。

1. 国民所得と酒の消費量の間には正の相関がある。だから，国民が酒をたくさん飲めば所得が増える。
2. ある電器製品の普及台数は，発売以来毎年倍に増えている。発売後の年数と普及台数の相関係数は，非常に強い相関であるから，ほぼ 1 である。

解説

1. 相関関係は，「因果関係」については何も述べていません。所得が増えたから酒をたくさん飲むのか，酒を

表 2: 相関係数を求める

地名	緯度（度）	気温（℃）	緯度の偏差	左の 2 乗	気温の偏差	左の 2 乗	両偏差の積
長野	36.67	11.4	2.18	4.752	−3.878	15.037	−8.454
静岡	34.97	16.0	0.48	0.230	0.722	0.522	0.347
名古屋	35.17	14.9	0.68	0.462	−0.378	0.143	−0.257
大阪	34.68	16.2	0.19	0.036	0.922	0.850	0.175
鳥取	35.48	14.4	0.99	0.980	−0.878	0.770	−0.869
広島	34.40	15.0	−0.09	0.008	−0.278	0.077	0.025
高知	33.55	16.3	−0.94	0.884	1.022	1.045	−0.961
福岡	33.92	16.0	−0.57	0.325	0.722	0.522	−0.412
鹿児島	31.57	17.3	−2.92	8.526	2.022	4.089	−5.905
	緯度の平均 = 34.49	気温の平均 = 15.278		緯度の分散 = 1.800		気温の分散 = 2.562	共分散 = −1.812 相関係数 = −0.844

たくさん飲むから所得が増えるのか，この問題の場合は常識的に前者とわかりますが，相関関係自体は前者であるとも後者であるとも言っていませんし，そもそも因果関係があるかないかも言っていません。因果関係を推定するには，また別の方法が必要です。

2. 「年数と台数の相関関係が 1 である」とは，年数と台数の増減に直線的な関係があることを意味しています。この問題の場合は，年数と台数の間に指数関数の関係があるので，相関係数はほぼ 1 にはなりません。このような場合，散布図で台数のほうを「対数目盛り」で表すと，年数と「台数の対数」の増減に直線的な関係ができ，相関関係として取り扱うことができます。対数目盛りについては，この講義では今後は扱いませんが，スライドで簡単に説明していますので，見てください。

層別と相関，「みかけ上の相関」と偏相関係数

見かけ上の相関

「小学生については，身体が大きいと試験の成績が良い」という説があります。明らかにおかしな話ですが，これは事実です。

種明かしをすると，これは，小学校の全学年の児童を対象に同じ問題で試験をした場合の話でした。こういう場合ならば，「体の大きさ」と「試験の成績」には正の相関関係が見られるはずです。

これは，「原因→結果」という因果関係が，「学年」→「体格」および「学年」→「成績」のそれぞれにあるために，本来無関係なはずの「体格」と「成績」にも相関が現れるという現象です。これを**みかけ上の相関（疑似相関）**といいます。

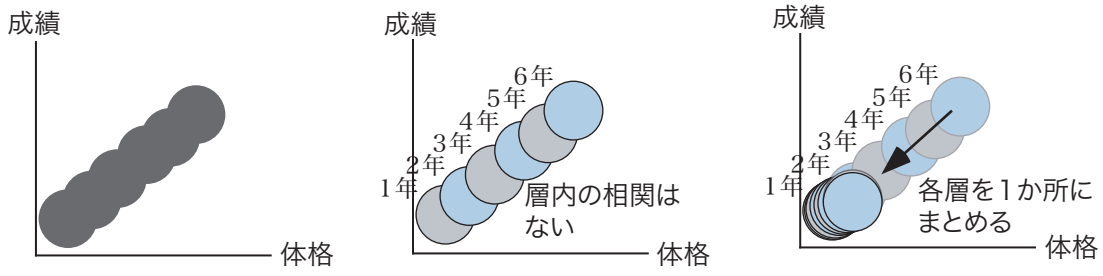


図 6: 層別の相関

層別

小学校 1 年生と 6 年生では体格は大きく違うのは当たり前です。「体格の違い」を問題にするには，各学年を別々に考え，1 つの学年の中での「体格の違い」を問題にする必要があります。このように，ほぼ均質と思われるグループ（ここでは学年）に母集団を分けることを**層別**といいます。

さて，この問題で，「体格」と「成績」の間には正の相関関係があるわけですから，これは次ページの図 6(a) のような分布をしていることになります。しかし，この分布を層別にしてみると，図 6(b) のように，各学年に対応する 6 つの分布が重なっているものと考えられます。各々の分布を別々に見たとき，もし各学年の分布が図 6(b) のようであれば，それぞれの分布では体格と成績には相関がないことがわかります。

偏相関係数

このように学年の影響を除いた相関係数を求めるには，図 6(b) の 6 つの分布を図 6(c) のように 1 か所に重ねてしまい，その重なった分布に対して相関係数を求めればよいことになります。このような操作をして得られる相関係数を**偏相関係数**といいます。

変量 x と y，y と z，z と x の各相関係数を r_{xy}, r_{yz}, r_{zx} とするとき，z の影響を除いた時の x と y の偏相関係数 $r_{xy,z}$ は次式で表されます。

$$r_{xy,z} = \frac{r_{xy} - r_{yz}r_{zx}}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{zx}^2}} \tag{2}$$

この式の詳しい導出は，回帰分析の詳しい知識が必要なので，この講義では説明しません。簡単にいえば，偏相関係数は，x, y, z の 3 つの変量を軸とする 3 次元の散布図を考えて，分布を z 軸のまわりに移動したとするとき，分母は x, y それぞれのばらつき，分子は x, y の共分散にそれぞれ相当する量になっています。

「見かけ上の」とはどういう意味か

さて，ここまでの説明を読んで，「では，『成績の影響を除いた，学年と体格の相関』もほとんどないことにならないのか？」と思った人もいるのではないでしょうか？ これは，偏相関係数を求める (2) 式で x, y, z を y, z, x に入れ替えてもほとんど同じ式が得られるように，数式の上では正しい結論です。

しかし，実際には意味のない結論です。なぜならば，「みかけ上の相関」は，「体格と成績に相関があるように見えるが，実は『学年』という隠れた量があって，学年が成績，体格それぞれの大小に影響し

ている」という，「現象のしくみ」についての仮定から導かれるものだからです。これを，「学年が成績，体格それぞれを**説明している**」と仮定する，といいます。しかし，いま仮定した「現象のしくみ」が正しいかどうかは，相関係数や偏相関係数からはわからず，別の観点からの考察が必要です。

また，「見かけ上の相関」あるいは「疑似相関」とは，「嘘の相関」という意味ではありません。「本当は相関はない」のではなく，「相関はあるが，因果関係など現象のしくみの説明にはなっていない」のです。ですから，現象のしくみを説明するのではなく，単に「体格から成績を推定したい」だけであれば，見かけ上の相関でも，「体格が大きい→成績がいいだろう」という推定は間違いではありません。ただ，「現象のしくみ」という点では，「体格が大きい→学年が高いだろう→成績がいいだろう」というしくみを考える必要がある，というだけです。

演習問題

1. 相関関係と因果関係が異なることを示している例をひとつあげ，なぜそう言えるのか説明してください。
2. 「見かけ上の相関」の例をひとつあげ，なぜそう言えるのか，隠れている変量は何かを説明してください。