2009 年度前期　データ解析序説　第４回

データの関係を知る (1) ー共分散と相関

多変量データと多変量解析

第２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. 今回から, この講義は「多変量解析編」に入ります. その１回目と２回目では, 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つの項目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

「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.

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２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

[1] 日本列島大地図館 (小学館)［理科年表より転載］より

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

表 1: 日本の年の緯度と気温

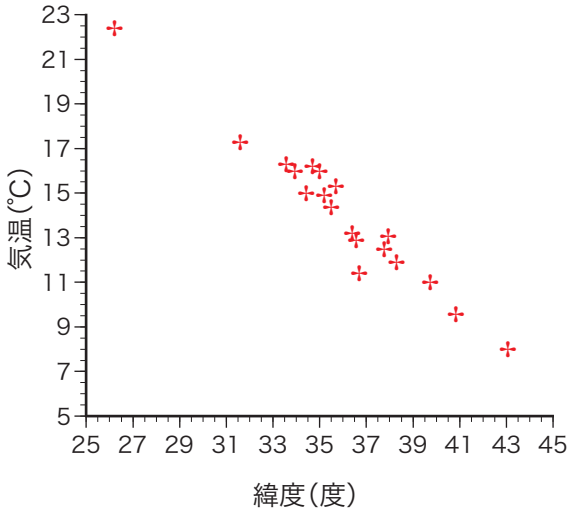

図 1: 散布図：緯度と気温の関係

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．

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．また，正の相関関係では右上がりの直線上に並ぶことになります．別紙[2]［資料 1 ］に，いろいろな散布図を示します．これを見ると，「人口と小売商店数」の関係では，各データがほぼ右上がりの一直線上に乗っており，「強い正の相関がある」ことがわかります．これに対し，「平均不快日数とルームエアコンの保有率」では各データのばらつきが大きくなっています．これを，「弱い正の相関がある」といいます．

相関係数

相関関係の強い／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

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

[2]「統計学入門」（東京大学出版会）44 ページ（受講者にのみ配付）

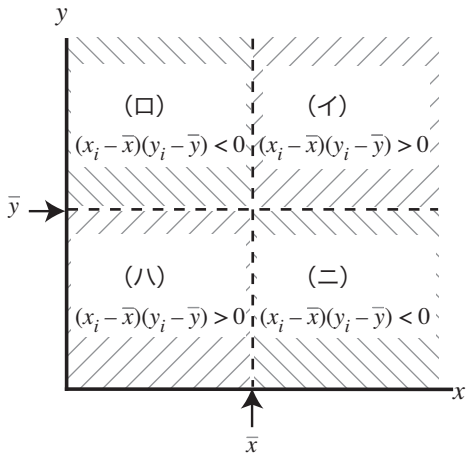

図 2: 共分散の概念

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．

共分散の意味を，図 2 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 3 の 3 つの分布で，$\sum_i (x_i - \bar{x})(y_i - \bar{y})$ の値はどうなるでしょうか？（グレーの部分にデータがおもに分布しているとします．）(a) の場合は先の図 2 の（イ）（ハ）の部分に多く分布していますから正の大きな値，(b) の場合は（ロ）（ニ）の部分に多く分布していますから負の大きな値，(c) の場合は（イ）（ロ）（ハ）（ニ）のすべての部分に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，グレーの部分に分布しているデータの個数 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．

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．なお，［資料２］[3]に示すように，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．このことについては，次回の回帰分析についての講義で説明します．

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．

[3]「統計学入門」（東京大学出版会）44 ページ（受講者にのみ配付）

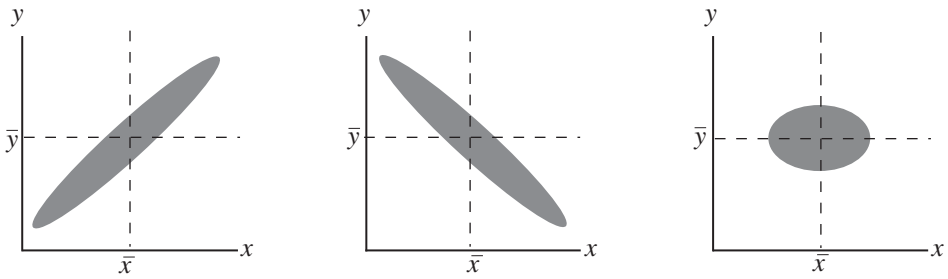

図 3: 正負の相関

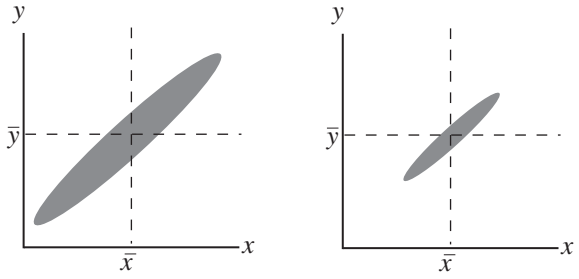

図 4: 同じ相関係数をもつ分布

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．

層別と相関，「みかけ上の相関」と偏相関係数

「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」という量の間にあるために，本来相関はないはずの「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**といいます．

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

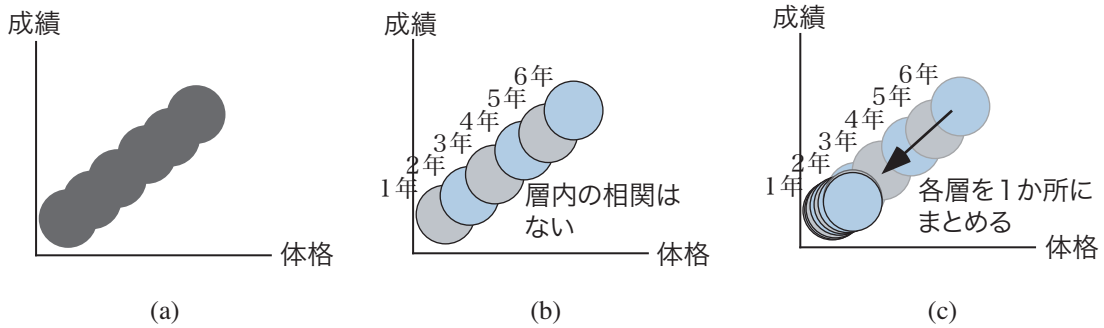

図 5: 層別の相関

図 5(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 5(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 5(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．

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 5(b) の 6 つの分布を図 5(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．

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

偏相関係数については，第 6 回の「重相関」のところでもう一度説明しますが，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．簡単に言えば，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．

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[4]」という仮定から導かれるものだからです．しかし，その仮定が正しいかどうかは，相関係数や偏相関係数をからはわからず，別の観点からの考察が必要です．

[4]「学年が成績，体格それぞれを**説明している**」といいます．