2　2 次元データ

クラス 40 人の数学と英語の点になんらかの関係があるかどうかなど、2 つの変量をもつ場合を考える。
ここでは、2 つ変量 x, y のデータが n 個の x, y の値の組として、次のように与えられているとする。

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

2.1　相関

- **散布図** 上記の x, y の値の組を座標とする点を平面上にとったもの ($cf.$ 教科書 pp.41, 42)。

- **共分散**, **相関係数**

$x_1, x_2, \cdots, x_n$ と $y_1, y_2, \cdots, y_n$ の平均値をそれぞれ $\overline{x}, \overline{y}$ で標準偏差を s_x, s_y で表す。
このとき、x と y の共分散 s_{xy} を

$$s_{xy} = \frac{1}{n}\{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

と定め、x と y の相関係数 r を $r = \dfrac{s_{xy}}{s_x s_y}$ と定める。ただし、$s_x > 0$ かつ $s_y > 0$ のときのみ相関係数は考えるものとする。

問 2.1 $s_{xy} = \overline{xy} - \overline{x}\,\overline{y}$ を示せ。ただし、$\overline{xy} = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i$. ヒント: $s_x^2 (= v_x) = \overline{x^2} - \overline{x}^2$ の証明と同様。

定理 2.1 (1)　相関係数 r について、$-1 \leqq r \leqq 1$ となる。
(2)　$r = 1$ となるのは、n 個のデータが正の傾きをもつ直線上に集中しているとき、
(3)　$r = -1$ となるのは、n 個のデータが負の傾きをもつ直線上に集中しているときに限る。

証明: コーシー・シュワルツの不等式: $(a_1b_1 + a_2b_2 + \cdots + a_nb_n)^2 \leqq (a_1{}^2 + a_2{}^2 + \cdots + a_n{}^2)(b_1{}^2 + b_2{}^2 + \cdots + b_n{}^2)$ で $a_k = x_k - \overline{x}$, $b_k = y_k - \overline{y}$ を代入することで (1) はすぐにわかる。また、この不等式で等号が成立するための条件は、ある定数 c があってすべての k に対して $b_k = ca_k$ となることであるから、[*1]

$c > 0$ のとき $r = 1$ であり $y_k - \overline{y} = c(x_k - \overline{x})$ となること、即ち、直線 $y = c(x - \overline{x}) + \overline{y}$ に集中、

$c < 0$ のとき $r = -1$ であり $y_k - \overline{y} = c(x_k - \overline{x})$ となること、即ち、直線 $y = c(x - \overline{x}) + \overline{y}$ に集中、

から (2), (3) は従う。　　□

- **正の相関**, **負の相関**　　変量 x と y の間に、

一方の値が増加すると他方も増加する傾向があるとき、2 つの変量 x, y の間に正の相関があるという。
一方の値が増加すると他方は減少する傾向があるとき、2 つの変量 x, y の間に負の相関があるという。
正の相関も負の相関もみられないとき、相関がないという。

おおよその目安となる基準は以下のようである ($cf.$ [2], p.60)。

(i)　相関係数 $= 0.7 \sim 1.0$ (または $= -0.7 \sim -1.0$): かなり強い正の相関 (負の相関) がある。
(ii)　相関係数 $= 0.4 \sim 0.7$ (または $= -0.4 \sim -0.7$): 中程度の正の相関 (負の相関) がある。
(iii)　相関係数 $= 0.2 \sim 0.4$ (または $= -0.2 \sim -0.4$): 弱い正の相関 (負の相関) がある。
(iv)　相関係数 $= -0.2 \sim 0.2$: ほとんど相関がない。

x, y のそれぞれのデータの平均値からのずれ (偏差) からなる n 次元ベクトルを考えると、r はこの 2 つのベクトルの内積を長さの積で割ったものだから「なす角」とみなせる。つまり、次のように考えられる。

・x, y のデータの平均値からの偏差が比較的同じ方向を向いている $\longleftrightarrow$ 正の相関関係がある。

・x, y のデータの平均値からの偏差が比較的反対の方向を向いている $\longleftrightarrow$ 負の相関関係がある。

[*1] コーシー・シュワルツの不等式とその等号成立のための条件は、$\sum_{k=1}^{n}(a_k t + b_k)^2$ を t について平方完成することで証明できる。

注意 2.1 (1) 二つの変量 x, y に強い正の相関があっても、実際にその二つの間に因果関係があるとは限らない。例えば、「サラリーマンの年収と血圧を調べると正の相関がある」について (実際に調べるとかなり強い正の相関があるらしい)、これは年収と血圧がともに年齢とともに上昇する傾向があることによっている。このように実際に因果関係があるかは相関係数だけではなく他の要因も調べなければならない。

社会科学の分野では、ポール・ラザースフェルドが 1959 年に、次の 3 つの基準を挙げた。

1. 原因は結果に先行する。
2. 2 つの変量は経験的に相関している。
3. その相関は、別の第三の変数によって説明されない。

自然科学の分野では、米国公衆衛生局長諮問委員会が 1964 年に喫煙と肺がんの因果関係を諮問されたときの判断基準がある。詳しくはいくつかの用語を導入しなければいないので省略する ($cf.$ [1], p.102)。

(2) 一般に、データをまとめ上げてしまうと、部分的に存在する関係等が良く見えなくなってしまう場合が多い。例えば、理系科目が得意の生徒だけが集まったクラスと文系科目が得意の生徒だけが集まったクラスがあったとしよう。それぞれのクラスでは、国語と数学の試験の点数には正の相関があったとしても、二つのクラス全体のデータから国語と数学の試験の点数の間の相関係数を計算すると負になることもあり得る。

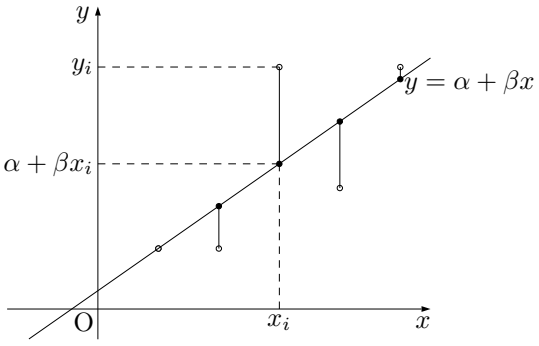
このように、部分的な関係も把握できるように、属性やデータの値などによって、データをいくつかの部分集合に分けて (層別にして) 解析を行うことが重要となる。

一方、一部のデータのみに基づいて計算された相関係数は、実際の相関係数より小さくなりやすいことも注意する必要がある。例えば、大学入試の成績 x と入学後の成績 y の相関関係を考えてみよう。これがある正の相関をもつと想定することは自然である。しかし、このデータを調べることは不可能である。なぜなら、不合格者は大学に入学できないから、入学後の成績のデータが得られない。特に、競争倍率が高く合格者の割合が少ない場合など、合格者のみのデータによって計算される x と y の相関係数は低くなり、場合によっては負の相関となってしまう場合も珍しくない。

このようなある値より小さい (または大きい) 値を持つデータしか存在しない場合は、それは「切断データ」とよばれ、少なくとも一方が切断されている場合には、計算された相関係数の値は一般に低くなる ($cf.$ [2])。

2.2 回帰直線

2 次元データに 強い相関があるとき、$y = \alpha + \beta x$ の関係がある (α, β は定数) と考えられる。x を独立変数、y を従属変数という。



- **最小二乗法**

x_i から予測される値 $\alpha + \beta x_i$ と現実の値 y_i との差の二乗の和 $Q(\alpha, \beta) = \sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2$ が最小となるように係数 α, β の値を定める。

$$\frac{1}{n}Q(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i)$$

$$= \overline{y^2} + \alpha^2 + \beta^2\overline{x^2} - 2\alpha\overline{y} - 2\beta\overline{xy} + 2\alpha\beta\overline{x} = \{\alpha - (\overline{y} - \beta\overline{x})\}^2 + (\overline{x^2} - \overline{x}^2)\beta^2 - 2(\overline{xy} - \overline{x}\,\overline{y})\beta + \overline{y^2} - \overline{y}^2$$

$$= \{\alpha - (\overline{y} - \beta\overline{x})\}^2 + s_x^2\beta^2 - 2s_{xy}\beta + s_y^2 = \{\alpha - (\overline{y} - \beta\overline{x})\}^2 + s_x^2\Big(\beta - \frac{s_{xy}}{s_x^2}\Big)^2 - \frac{s_{xy}^2}{s_x^2} + s_y^2$$

よって、$\beta = \dfrac{s_{xy}}{s_x^2}$, $\alpha = \overline{y} - \beta\overline{x} = \overline{y} - \dfrac{s_{xy}}{s_x^2}\overline{x}$ のとき最小となるため、回帰直線の方程式は $y - \overline{y} = \dfrac{s_{xy}}{s_x^2}(x - \overline{x})$ と表される。(厳密には x から y への回帰直線という。)

参考文献

[1] 青木 繁伸: 統計数字を読み解くセンス 当確はなぜすぐにわかるのか?, 化学同人, 2009.

[2] 田栗 正章, 藤越 康祝, 柳井 晴夫, C.R. ラオ: やさしい統計入門, 講談社ブルーバックス, 2007.