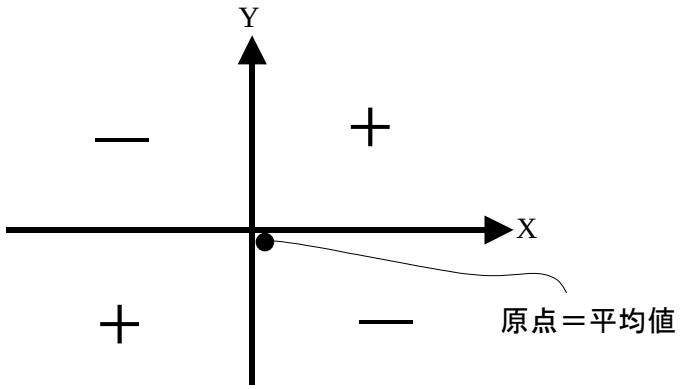
共分散・相関係数・回帰分析

一人の被験者から複数のデータがえられた場合，それぞれのデータの関係を調べたいと思うことはごく自然なことであろう。共分散（covariance），相関係数（correlation coefficient），回帰係数（regression coefficient）とはそのような願望に基づいて発案されたといってもよいであろう。

たとえば，前回配布した資料にあるように，もっとも素朴な分析としては 2 つの変数をそれぞれグラフの軸として，個々の被験者のデータをプロットし散布図を作成する方法が考えられる。このように散布図を作成することは，これから説明する相関係数の分析や回帰分析などでも省略してはならない分析ステップの一つであるが，これだけでは数量的な判断を行うことは難しい。この不具合を解消するために考え出された統計値の一つとして共分散がある。

共分散とは次のような式によって求められる。　$\mathrm{Cov}(x,y)=S_{xy}=\sum\frac{(x_i-\bar{x})\times(y_i-\bar{y})}{N}$



図．$x \times y$ の値。プロットされるデータが特定の象限に集まるほど，共分散の値も極端な値（＋∞～－∞）になることが上記の図から読みとれる。

上記の共分散によっても変数間の対応関係を把握することはできるが，この数値には変数の単位の違いによって影響が出やすいという問題がある。たとえば，ある反応がミリの尺度で計測された場合と，メートルで計測された場合を考えてみるとこの影響の大きさがよく分かるであろう。全く同じデータでも共分散の値が 1000 倍違うことになる。このような不具合を解消するためにはどうしたら良いだろうか？勘の良い読者はきっと「個々のデータを標準化して共分散を求めてやれば良い」ことに気づくであろう。

実際に，$U_i=\frac{x_i-\bar{x}}{S(x)}$，$V_i=\frac{y_i-\bar{y}}{S(y)}$ とおけば，$\mathrm{Cov}(x,y)=\sum\frac{U_i\times V_i}{N}$ となり，個々の変数の観測尺度に依存しない数値を求められることが分かる。このように個々のデータを標準化して求めた共分散というものが，ピアソンの積率相関係数（Pearson’s product-moment correlational coefficient：通称「相関係数」，r と略記される）となっている。r_{xy} は以下のようにも書き換えられる。

$$r_{xy}=\sum\frac{U_i\times V_i}{N}=\sum\frac{\frac{(x-\bar{x})}{Sx}\times\frac{(y-\bar{y})}{Sy}}{N}=\frac{1}{Sx\times Sy}\times\sum\frac{(x-\bar{x})\times(y-\bar{y})}{N}=\frac{S_{xy}}{S_x\times S_y}$$

なお，数学的に眺めると，上記の相関係数はベクトルの内積の角度と密接な関係にある。ベクトルの内積

とは，$\vec{x} \bullet \vec{y} = (x_1 y_1 + x_2 y_2 + ... + x_n y_n) = |\vec{x}| \times |\vec{y}| \times \cos\theta$ （θ はベクトル x とベクトル y とが交わる角度）である。この内積の式から，$\cos\theta = \dfrac{(x_1 y_1 + x_2 y_2 + ... + x_n y_n)}{|\vec{x}| \times |\vec{y}|}$ であることが分かるが，相関係数はデータベクトルが交わる角度にもなっている。

たとえば，平均値からの偏差として各観測データ（N人の被験者からなるデータ）を表現してやると，実験データは以下のようにN次元のベクトルと見なすことができる。

$$\vec{x_i} = \{(x_1 - \bar{x}), (x_2 - \bar{x}), ..., (x_n - \bar{x})\}$$

そして，このベクトルの長さは

$$|\vec{x}| = \sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2} = \sqrt{\sum (x_i - \bar{x})^2} = \sqrt{N} \times \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} = \sqrt{N} \times Sx$$

と表現できる。同様に，

$$\vec{x} \bullet \vec{y} = \{(x_1 - \bar{x})(y_1 - \bar{y}) + ... + (x_n - \bar{x})(y_n - \bar{y})\} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= N \times \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N} = N \times Sxy \text{ となる。}$$

したがって，相関係数 $r_{xy} = \dfrac{N \times Sxy}{(\sqrt{N} \times Sx) \times (\sqrt{N} \times Sy)} = \dfrac{\vec{x} \bullet \vec{y}}{|\vec{x}| \times |\vec{y}|} = \cos\theta$ となることが分かる。

重回帰分析や因子分析では「それぞれの変数どうしが直交している」という表現がよく出てくるが，直交しているとは，「データベクトルが直角に交わっている」ことを意味している。このようなときには，共分散の理屈を図示したように散布図が球状になり，いわゆる無相関な関係にある。

さて，相関係数以外に調べたいものとはどのようなものであろうか？おそらく，ある変数 x から別の変数 y の振る舞いを予測することになるであろう。このとき，x と y が完全に一直線の線上に並ぶような関係にないとしても，そのような直線関係にあることを想定することはそれほど無謀な話ではないであろう。回帰分析とはこのように直線の相関関係を想定した予測式を求める分析である。具体的には次のような 1 次の線形式を想定して，係数 b と定数 a を最小二乗法（method of least square）によって求める。$\hat{y} = a + bx_i$　詳しい説明は教科書に譲るが，$\hat{y} = \bar{y} + r_{xy} \dfrac{Sy}{Sx}(x_i - \bar{x})$ という関係にあることをよく理解しておいて欲しい。