相関についての検定

Tests for Correlation

髙木廣文

看　護　研　究

第47巻 第6号 別刷

2014年10月15日発行

KANGO-KENKYU(*The Japanese Journal of Nursing Research*)

Vol.47 No.6/Sep.－Oct.2014

医学書院

第5回

相関についての検定

Tests for Correlation

髙木廣文　東邦大学看護学部長

相関関係とはどんな関係か

一般的に言って，身長が高ければ体重は重いし，手足も長くなるでしょう。人間は高齢になればなるほどさまざまな病気にかかる人も増えてきます。このように，世の中に存在する物事の間には，相互になんらかの関係があるのが普通です。このような「関係」を，統計学ではどのように扱っているのかということについて，今回は扱ってみたいと思います。

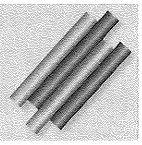
英国人フランシス・ゴルトン Francis Galton は，チャールズ・ダーウィン Charles Darwin の従弟であり，ダーウィンの『種の起源』(1859年)の出版に触発され，進化と遺伝の問題，それを科学的に扱う生物統計学を研究するようになったといわれています。1865年には『遺伝的才能と性格(Hereditary Talent and Character)』という本の中で，天才が遺伝することと結婚により人類を向上させる可能性を論じています。1869年には『遺伝的天才』を発表し，天才が遺伝することを証明しようと試み，知性の分布をガウス分布(正規分布)に従うものと仮定しました。また，ゴルトンは，スイートピーの親種と子種の観察を行ない，大きな親種の子種の平均値は親種より小さくなり，小さな親種の子種の平均値は親種の平均値より大きくなるような傾向があることを見いだしました。当初，このような傾向は「reversion(先祖がえり)」と呼ばれていましたが，その後，「**回帰** regression」と呼ばれるようになりました。

さらにゴルトンは，人間の両親と子どもの身長にも類似性があることを初めて見いだしました。大きな身長の両親のグループの子どもの身長の平均値は大きいけれども，そのグループの両親の平均値より小さく，また小さな身長の両親のグループの子どもの平均値はその両親の平均値よりも大きく，いずれのグループの子どもの身長の平均値は，両親の身長平均から全体の平均値の方向に向かうという「**回帰傾向**」があることを見いだしています。

両親の身長と子どもの身長の類似性のように，2変数間の量的な関係は「**相関** correlation」と呼ばれるようになり，相関の程度を数量的に表わす「**相関係数**」も，ゴルトンにより初めて定義されました。しかし，現在使用されている相関係数は，後述するようにゴルトンの弟子のカール・ピアソン Karl Pearson によって考え出されたものです。

0022-8370/14/￥600/論文/JCOPY

相関係数と無相関の検定



1. 相関係数の定義

ある2変数間に大きな正の相関または負の相関がある場合，一方の変数の増加・減少が他方の変数の増加・減少と関係があることになります。そのような場合，一方の変数の値を知ることで，他方の変数の値を予測することが，ある程度できるはずです。

いま2変数をXとYとし，Xの値(観測値x)からYの値を予測するための式を考えてみます。最も簡単な式は，Xの値に適当な数値をかけ，これに一定の数を加えるという一次式を使う方法です。すなわち，

$$Y\text{の予測値}\ \hat{y}=[\text{係数}\ b]\times[X\text{の値}\ x]+[\text{定数}\ a]$$

という式を仮定します。記号のみで書けば，

$$\hat{y}=bx+a$$

と書くことができます。

式を使って予測するためには，係数bの値と定数aの値を求める必要があります。ここで，上の式は「**回帰式** regression equation」と呼ばれ，係数bは「**回帰係数** regression coefficient」と呼ばれています。そして，この数式が表わす一次直線は「**回帰直線** regression line」と呼ばれています。

相関図の上に直線を引く方法は無数にあるので，どれか最適な1本に決める必要があります。回帰係数bと定数aを求めるには，回帰式で予測されたYの予測値$\hat{y}$と観測値yとに差がなければ，回帰式による予測はうまくいっていると考えてよいでしょう。実際の計算では，予測値$\hat{y}$と観測値yの差の2乗について，全ケースの合計が最小になるように，回帰係数bと定数aを定めています。すなわち，

$$L=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-bx_i-a)^2$$

を最小にするように，回帰係数bと定数aを求めればよいのです。この方法は「**最小2乗法** least squares method」と呼ばれています。実際に計算すると，

$$\text{回帰係数}\ b=\frac{X\text{と}Y\text{の共分散}}{X\text{の分散}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

および，

$$\text{定数}\ a=\bar{y}-b\times\bar{x}$$

により，回帰係数bと定数aが求められます。ここで，$\bar{x}$，$\bar{y}$は，変数XとYの平均値です。

なお，上式の**共分散** covariance とは，

$$\textbf{共分散}\ S_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

によって定義される統計量です。

共分散は，2つの変数のそれぞれの偏差の積和となっています。2つの変数の偏差が両方とも正の場合，その積は正になるし，両方とも負の場合も，やはりその積は正になります。逆に，一方の変数の偏差が正で他方が負の場合には，その積は負になります。したがって，2つの変数の偏差の正負の傾向が一致していれば，その積は正になるし，逆の傾向があれば負になります。2つの変数の偏差の積をすべてのケースについて合計した場合，その値が正ならば2変数の大小について同一の傾向があることを示しているし，その値が負ならば逆の傾向があることを示していることになります。

この共分散の性質を用いれば，2変数間の相関関係を示すことができます。しかし，身長と体重の共分散の単位が，cm・Kgになってしまうよう

に，変数の測定単位とその大きさの影響を受けてしまい，共分散の数値のみでは相関関係の有無や強弱の判断が困難です。そこで，変数の単位の相違をなくすために，各変数の標準偏差で共分散を割り算することで単位をなくし，大きさを標準化することができます。

2つの変数 X，Y の標準偏差を S_x，S_y とすると，相関係数 r_{xy} は，

$$r_{xy}=\frac{S_{xy}}{S_x\times S_y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

によって定義されます。この相関係数は，正式には「**ピアソンの積率相関係数** Pearson's product moment correlation coefficient」と呼ばれています。しかし，この呼称は極めて長いので「**単相関係数** simple correlation coefficient」とも呼ばれており，より簡単に「**相関係数**」といえば，このピアソンの積率相関係数を示すのが普通です。

相関係数 r_{xy} は正の場合には正の相関を，負の場合には負の相関を示し，−1から1までの範囲の値をとり，その絶対値が1に近いほど相関関係が強いことを，数値が0に近いほど相関が弱いことを示します。そして，相関係数の絶対値が大きいほど，直線的な関係が強いことになり，前述した回帰直線により，一方の変数の値からもう一方の変数の値を予測できることになります。一方の変数から他方の変数をどの程度説明できるかは，相関係数の2乗によって表わされ，「**決定係数** coefficient of determination」と呼ばれています。例えば，相関係数が0.5の場合，決定係数は $0.5^2=0.25$ となり，一方から他方への説明力は25%ということになり，かなり強い相関関係の存在を示します。

さて，ある2つの変数について，標本から得られた相関係数が大きいからといって，母集団でも相関が大きいとは必ずしも明らかではありません。標本抽出に伴う偶然の変動が常に存在するため，その影響を受けるからです。そのため，母相関係数に関する検定や推定が必要になってきます。

ここではまず，「**無相関の検定**」について説明し，その後，母相関係数の信頼区間の構成方法について説明します。

2. 無相関の検定の方法

2つの変数 X と Y について，無作為に N 人の標本からデータを収集したところ，標本相関係数 r_{xy} が得られたとします。検定のための帰無仮説は，「母相関係数 $\rho=0$」です。この帰無仮説が正しいと仮定した場合に，偶然に標本相関係数が r_{xy} となる確率を求めて，その確率が極めて小さければ，帰無仮説はあり得ないことになります。

さて，検定のための統計量 t_0 は，

$$t_0=|r_{xy}|\times\sqrt{\frac{N-2}{1-r_{xy}^2}}$$

によって求められます。この統計量 t_0 は，自由度 $\nu=N-2$，のt分布に従うことから検定が行なえます。

コンピュータを用いて，この統計量 t_0 の有意確率(p値)を求め，$p\leq 0.05$(有意確率が5%以下)ならば，5%水準で有意な相関があるとすればよいでしょう。

もしも，p値が5%より大きい場合，帰無仮説を棄却できません。そのような場合には，「母相関係数が0でないとはいえない」という結論になります。すでに述べているように，帰無仮説が棄却できない場合には，帰無仮説「母相関係数が0である($\rho=0$)」ことを証明したことにはならない点に注意が必要です。

【例題1】

ある町の40歳以上の住民から無作為に選んだ50人の空腹時血糖値と肥満度BMIの単相関係数は0.483であったとします。この町の40歳以上の住民の空腹時血糖値と肥満度について相関があると考えてよいでしょうか。

まず，検定統計量 t_0 を求めてみます。

$$t_0 = |0.483| \times \sqrt{\frac{50-2}{1-0.483^2}}$$
$$= 0.483 \times \sqrt{62.605} = 3.822$$

と計算できます。また，自由度 ν は，$\nu = 50-2 = 48$ なので，パソコンで p 値を求めると，$p = 0.0004$ と求められます。したがって，空腹時血糖値と肥満度には，高度に有意な相関があると考えてよいことになります。

母相関係数の信頼区間

無相関の検定で統計学的に有意な結果が得られた場合，母相関係数は 0 でないことになります。それでは，母相関係数が 0 でなければどのくらいの値なのかという問題が出てきます。このような問題に対しては，母相関係数の信頼区間の推定を行なえばよいことになります。

母割合や母平均値の区間推定の場合と同様に，2 つの変数 X と Y の単相関係数 r_{xy} の分布を正規分布や t 分布などで近似できれば，信頼区間の計算が簡単になるでしょう。面倒なことに，標本相関の分布はそのままでは正規分布や t 分布に従いません。母相関係数の信頼区間を求めるためには，まず「**フィッシャーの z 変換**」と呼ばれる変換を行ない，正規分布に従うような統計量を求める必要があります。

単相関係数 r_{xy} を次のように変換して，

$$z_r = \frac{1}{2} \times \ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$$

を求めます。ここで，$\ln(x)$ は x の自然対数を求めることを示しています。

上記の z_r は，帰無仮説（母相関係数 $\rho = 0$）のもとでは，平均が 0，分散 V が，

$$V = \frac{1}{N-3}$$

の正規分布に従う統計量になります。

したがって，上記の統計量 z_r の 95％信頼区間の上限 z_U と下限 z_L は，

$$z_U = z_r + \frac{1.96}{\sqrt{N-3}}$$

および，

$$z_L = z_r - \frac{1.96}{\sqrt{N-3}}$$

によって求められます。99％信頼区間を求めたい場合には，上記の 2 式の 1.96 を 2.58 にかえるだけです。

さて，統計量 z_r の 95％信頼区間の上限 z_U と下限 z_L は求められましたが，このままでは母相関係数の信頼区間ではありません。次の操作として，上限 z_U と下限 z_L の逆変換を行ない，標準得点から相関係数に戻す必要があります。すなわち，

$$r_U = \frac{\exp(2z_U)-1}{\exp(2z_U)+1}$$

および，

$$r_L = \frac{\exp(2z_L)-1}{\exp(2z_L)+1}$$

を求める必要があります。ここで，$\exp(x)$ は，自然対数の底 e の x 乗，e^x を示します。

これらの計算によって，母相関係数の 95％信頼区間の上限 r_U と下限 r_L を求めることができます。

【例題 2】

前述の例題で，50 人の空腹時血糖値と肥満度の単相関係数は 0.483 であり，検定結果は有意でした。それでは，母相関係数がどのくらいの値になるのか，95％信頼区間を求めてみましょう。

まず，単相関係数をフィッシャーのz変換すると，

$$z_r = \frac{1}{2} \times \ln\left(\frac{1+0.483}{1-0.483}\right) = 0.52689$$

となります。

次に，z値の95%信頼区間の上限，下限の計算を行なうと，

$$z_U = 0.52689 + \frac{1.96}{\sqrt{50-3}}$$
$$= 0.52689 + 0.28590 = 0.81279$$

および，

$$z_L = 0.52689 - \frac{1.96}{\sqrt{50-3}}$$
$$= 0.52689 - 0.28590 = 0.24099$$

と求められます。

このままでは，相関係数ではないので，上記の2つのz値をそれぞれ逆変換すると，

$$r_U = \frac{\exp(2\times 0.81279)-1}{\exp(2\times 0.81279)+1} = 0.671$$

および，

$$r_U = \frac{\exp(2\times 0.24099)-1}{\exp(2\times 0.24099)+1} = 0.236$$

となります。したがって，母相関係数ρの95%信頼区間は，

$$0.236 < \rho < 0.671$$

と求められます。思いのほか，母相関係数の推定値の信頼区間が広くなっていることがわかります。

有意となる標本数を求める

今回の例題では有意な結果が得られましたが，一般に無相関の検定が必ず有意になるような標本数を研究計画時に決めるためには，以下のような情報が必要になります。

(1)検定のための第1種の誤り2α(有意水準)
(2)検定のための第2種の誤りβ，または検出力$1-\beta$
(3)2変数の母相関係数ρの値(推定値$\hat{\rho}$でよい)

これまで説明したように，検定の有意水準2αは0.05(5%)とし，βは2αの4倍とします。すなわち，有意水準が5%，検出力は80%に設定するのが普通です。あと必要な情報は，母相関係数ρの推定値$\hat{\rho}$です。通常は，既存の文献を参考にして推定しますが，情報が全くない場合には予備調査が必要となるかもしれません。

標準正規分布の上側α点をz_αとし，必要な標本数をNとすれば，

$$N = 4\times(z_\alpha + z_\beta)^2 \div \left[\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)\right]^2 + 3$$

によって求めることができます。

検定の有意水準を5%，検出力を80%とした場合には，$z_{0.025}=1.96$および$z_{0.2}=0.842$を上式に代入すると，

$$N = 31.4 \div \left[\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)\right]^2 + 3$$

となり，必要な標本数を簡単に計算することができます。

【例題3】

塩分摂取量と最大血圧の母相関係数が0.3と推定されるとき，検定の有意水準を5%，検出力を80%とした場合，必要とされる標本数を求めてみましょう。

$$N = 31.4 \div \left[\ln\left(\frac{1+0.3}{1-0.3}\right)\right]^2 + 3$$
$$= 31.4 \div \left[\ln\left(\frac{1.3}{0.7}\right)\right]^2 + 3 = 31.4 \div 0.3832 + 3$$
$$= 84.9$$

となり，85例の標本数が必要となります。

5. 2つの母相関係数の差の検定が有意となる標本数を求める

実際の研究ではあまりないかもしれませんが，性別などの属性の違いによって，ある2変数間の母相関係数に相違があることを検証したいような場合，その差を検出できるような標本数が研究上必要となります。

2つの母相関係数の推定値を$\hat{\rho}_1$と$\hat{\rho}_0$とし，検定の有意水準を両側2α，検出力を$1-\beta$とします。標準正規分布の上側α点の値をz_αとすると，必要とされる標本数Nは，

$$N = \left(\frac{z_\alpha + z_\beta}{z_1 - z_0}\right)^2 + 3$$

によって求められます。ただし，z_1とz_0はそれぞれの母相関係数の推定値をz変換したものです。すなわち，

$$z_1 = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}_1}{1-\hat{\rho}_1}\right), \quad z_0 = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}_0}{1-\hat{\rho}_0}\right)$$

によって求められた値です。

検定の有意水準を5%，検出力を80%とした場合，$z_{0.025}=1.96$，$z_{0.2}=0.842$の値を上式に代入すると，

$$N = 7.85 \div (z_1 - z_0)^2 + 3$$

のように比較的簡単に計算ができます。

【例題4】

2つの母相関係数の推定値を0.3と0.4とした場合，有意水準5%，検出力80%として，その差を検出するための標本数を求めてみましょう。

$$z_1 - z_0 = 0.5 \times \ln(1.3/0.7) - 0.5 \times \ln(1.4/0.6)$$
$$= 0.30952 - 0.42365 = -0.11413$$

となるので，必要な標本数は，

$$N = 7.85 \div (-0.11413)^2 + 3$$
$$= 7.85 \div 0.01303 + 3 = 605.5$$

と求められます。

実際には，調査などからの対象者の脱落の割合や有効標本の割合などを考慮する必要がありますので，2つの相関係数の僅かな差を検出するためには，これ以上に大きな標本数が必要とされることになります。

文献

- 安藤洋美(1997)．多変量解析の歴史．現代数学社，pp.1-64.
- 髙木廣文，林邦彦(2006)．エビデンスのための看護研究の読み方・進め方．中山書店，pp.63-69.
- 髙木廣文(2009)．ナースのための統計学，第2版．医学書院，pp.97-113.
- 髙木廣文(2014)．統計学のキー・ポイント―「検定」に焦点を当てて　第1回　検定の基本的な考え方．看護研究，47(1)，52-58.
- 髙木廣文(2014)．統計学のキー・ポイント―「検定」に焦点を当てて　第2回　検定と標本数の関係．看護研究，47(2)，150-154.
- 髙木廣文(2014)．統計学のキー・ポイント―「検定」に焦点を当てて　第3回　平均値の差についての検定．看護研究，47(3)，264-269.
- 髙木廣文(2014)．統計学のキー・ポイント―「検定」に焦点を当てて　第4回　割合についての検定．看護研究，47(5)，474-481.
- Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D.G., & Newman, T.B.(2007). *Designing Clinical Research*(3rd ed.). Lippincott Williams & Wilkins.／木原雅子，木原正博訳(2009)．医学的研究のデザイン，第3版．メディカル・サイエンス・インターナショナル，pp.67-97.

たかぎひろふみ◎東邦大学看護学部
〒143-0015　東京都大田区大森西4-16-20