統計講座（第4回）

相関関係に関する統計学的方法

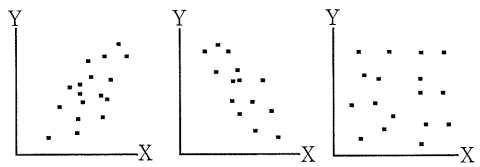
高木　廣文

相関関係と相関図について

実際の研究では，いくつかの調査項目間の関係を調べることに興味がある場合が多い．2つの変数が身長や体重などのように，量的データとして測定されている場合，2変数間の相互関係は「相関correlation」と呼ばれている．

2変数間の相関関係の有無を視覚的に調べ，直観的に把握するために，「相関図correlation diagram」もしくは「散布図scatter plots」と呼ばれる図を作成することが多い．

相関図とは，2つの変数を縦軸と横軸にとり（**図1**ではXとY），各ケースのデータに対応する座標に＊やxで印を付けたものである．通常は，**図1** (a)～(c)のような形状の相関図が得られるだろう．



(a) 正の相関　(b) 負の相関　(c) 相関なし

図1　相関関係の分類

相関図を作成すると，変数間の関係が視覚的に把握できるが，大きく分けるパターンが3通りあることがわかる．

図1－(a)のように，一方の変数が大きい場合，他方の変数も大きく，逆に一方が小さい場合，他方も小さいというように，2変数の大小関係が対応している場合，2変数間に「正の相関（順相関）positive correlation」があるという．

逆に，**図1**－(b)のように，一方の変数が大きい場合，他方の変数は小さく，また一方が小さい場合，他方は大きいというように，2変数の大小関係が逆に対応している場合，2変数間に「負の相関（逆相関）negative correlation」があるという．

また，**図1**－(c)のように，2変数間の大小関係に一定の傾向が認められない場合，「相関がない（無相関）null correlation」という．

（例1）**表1**に心エコー図法を用いて，Simpson法，Teichhols法，およびDoppler法の3方法で1回心拍出量を求めた199人のデータの一部を示した（心臓血管研究所での計測結果による）．

表1　1回心拍出量のデータ

Simpson法	Teichhols法	Doppler法
60	78	82
64	74	53
60	58	70
68	78	73
46	49	51
44	70	50
64	68	69
63	75	57
70	70	68
⋮	⋮	⋮
⋮	⋮	⋮
62	83	89
83	78	96
61	54	58
94	83	80
74	83	69
69	83	93
43	65	83
110	95	119
54	54	66
55	54	63

Hirofumi TAKAGI：新潟大学医療技術短期大学部
〒951-8518　新潟県新潟市旭町通2番町746
e-mail : takagi @ clg.niigata-u.ac.jp
http://www.clg.niigata-u.ac.jp/~takagi/

3方法による1回心拍出量にどのような関係があるかを見るために，2変数ごとの組み合わせで相関図を作成し，**図2**に示した．

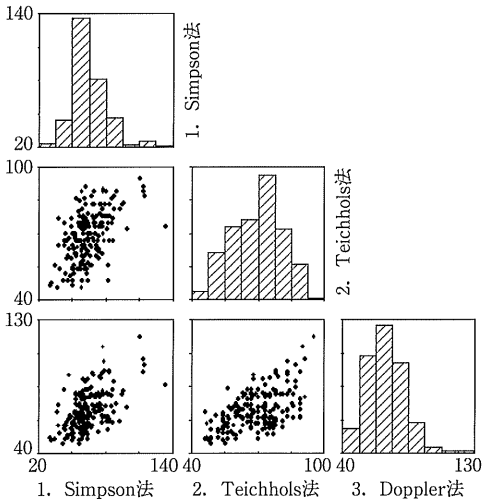


図2　3方法による心拍出量の相関図

図2で同一の測定法の組み合わせについては，その分布を示すためにヒストグラムを表示している．ヒストグラムからは，Simpson法による1回心拍出量のデータは他と比べるとやや小さな値のところにピークがあり，Teichhols法ではやや大きいところにピークがある．そして，Doppler法は両者の中間的な分布の形状を示している．

図2の3方法の組み合わせによる相関図は，いずれも右上がりの形状となっており，正の相関があることがわかる．それでは，どの2方法の組み合わせで相関関係が一番強いといえるのだろうか．この問題に答えるには，より客観的に相関関係を示す指標が必要である．

相関係数

2つの変数の関係を検討するためには，すでに説明したように相関図を描き，視覚的に眺めることは極めて重要なことである．しかし，図を見て相関がある，相関が強いといっても，それは単なる主観的な表現に過ぎない．実際に，どの程度の相関があるのかがわからない．したがって，より客観的に相関の程度を示すような数値があると便利である．

データが量的な2変数間の相関の程度を表す指標は「相関係数correlation coefficient」と呼ばれている．なお，データが質的な2変数間の関係の程度は「関連係数」によって示すことができる．

まず，ここでは通常よく用いられる指標である「ピアソンの積率相関係数Pearson's product moment correlation coefficient」について説明する．この呼称は極めて長いので，「単相関係数simple correlation coefficient」もしくは，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ある．

2変数XとYの単相関係数r_{XY}は，次のようにして求めることができる．

各ケースについて変数XとYのデータが，1番目のデータが（x_1，y_1），2番目のデータが（x_2，y_2），...，i番目のデータが（x_i，y_i），...，そしてN番目のデータが（x_N，y_N）のように観測されたとしよう．また，変数XとYの平均値をM_X，M_Yとし，分散をs_X^2，s_Y^2とする．

相関係数の計算では，まず「共分散covariance」（s_{XY}とする）と呼ばれる次の量を計算する必要がある．

$$s_{XY}=\frac{1}{N}\sum_{i=1}^{N}(x_i-M_X)(y_i-M_Y)$$

である．各変数の分散と共分散を用いて，相関係数は，

$$r_{XY}=\frac{s_{XY}}{s_X s_Y}$$

として求められる．

相関係数r_{XY}は，−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，また0に近いほど相関が弱いことを示す．また，相関係数の符号に対応して，正の相関か負の相関かがわかる．ただし，実際には±1に近い相関係数が観察されることはまれである．一方の変数から他方の変数の変数をどの程度説明できるかは，相関係数の2乗によって表され，「決定係数coefficient of determination」と呼ばれている．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%ということになる．これはかなり強い相関関係の存在を示すことになる．

（例２）１回心拍出量についての３方法のデータから，各２変数の組み合わせについて単相関係数を求めてみよう．

表２． 方法別心拍出量の平均値と標準偏差

方法名	平均値	標準偏差
1）Simpson法	64.683	15.285
2）Teichhols法	69.291	11.128
3）Doppler法	69.588	13.026

表３． ピアソンの単相関係数行列

変数名	1）	2）	3）
1）Simpson法	1.000	0.542	0.549
2）Teichhols法	0.542	1.000	0.574
3）Doppler法	0.549	0.574	1.000

表３に示したように，１回心拍出量推定のための３方法はすべて正の相関係数であった．相関係数はSimpson法とTeichholos法は0.542と最も小さく，Simpson法とDoppler法は0.549，Teichhols法とDoppler法は0.574とやや大きかった．

相関係数に関する推定と検定

相関係数により量的な２変数間の関係を示すことはできるが，それはあくまでも収集されたデータに関してのものである．実際に収集されたデータは，その背後にある母集団のほんの一部分に過ぎない．仮に，母集団での母相関係数が０であっても，データ収集時の偶然の変動によって，標本のデータから計算された相関係数は０でない値になるかもしれない．このような問題に答えるのが，母相関係数に関する検定と推定の方法といえる．

１）無相関の検定

母集団での相関係数（母相関係数）が０という仮説のもとで，まずデータから計算された数値の相関係数が偶然得られる確率（p値，有意確率）を求める．求めたp値がある基準より小さければ，母相関係数が０の場合に，偶然によってデータから計算されたような相関係数が観察されるとは考えにくいことになる．したがって，母相関係数が０であるという仮説を棄却し，母相関係数が０ではない，ある値をもつと考えることにする．これが，「無相関の検定」と呼ばれている検定方法の考え方である．

実際には，直接p値を計算することはできないので，検定のための統計量をまず計算する．

標本数をN，２変数X，Yの単相関係数をr_{XY}としたとき，

$$t_0 = |r_{XY}| \cdot \sqrt{\frac{N-2}{1-r_{XY}^2}}$$

を求める．

上記のt_0は自由度（N－2）のt分布に従うので，その分布の両側確率からp値を求めることができる．

検定の方法は，一般にp値が0.05（５%）より小さい場合に，帰無仮説を棄却し，母相関係数は０ではない（有意な）相関があるものとする．逆にp値が0.05（５%）より大きい場合，偶然の結果として観測したような相関係数が得られる可能性を否定できない．このため，帰無仮説を棄却できないので，母相関係数が０なのかそうでないのかは判断できないことになる（有意な相関があるとはいえない）．p値が５%より大きい場合には，このように無相関の仮説が証明された訳ではないので，母相関係数は０であるとはいえない点に注意が必要である．

（例３）１回心拍出量推定のための３方法の各２方法の母相関係数について，無相関の検定を行ってみよう．

Simpson法とTeichholos法の単相関係数は0.542であり，標本数は199人だったので，

$$t_0 = |0.542| \cdot \sqrt{\frac{199-2}{1-0.542^2}} = 9.052$$

と計算できる．また，自由度は197となる．この位の大きさの自由度のt分布では，t値が２程度が両側５%を与える目安になるので，とくに数値表を引かなくても，t値が9.052であることから有意な相関があることがわかる．実際にt分布から，p値を求めると，0.000000未満という極めて小さな

値になっている．

同様に，Simpson法とDoppler法の単相関係数は0.549であったので，$t_0=9.219$となる．また，Teichhols法とDoppler法の単相関係数は0.574であったので，$t_0=9.838$となり，いずれもp値は0.000000未満であり，高度に有意な相関があることを示している．

2）母相関係数の信頼区間

無相関の検定では，母相関が0か否かという情報は得られるが，実際にどの程度の大きさといってよいのかという情報は得られない．この目的のためには，母相関係数の信頼区間を推定する必要がある．

母相関係数の信頼区間は，次のように求めることができる．まず，単相関係数r_{XY}の分布はそのままでは正規分布で近似できないため，

$$Z_r=\frac{1}{2}\cdot\ln\left[\frac{1+r_{XY}}{1-r_{XY}}\right]$$

という変換を行う．これは「フィッシャーのZ変換」と呼ばれている．このような変換を行うと，Z_rは分散が1／（N－3）の正規分布として扱えるので，このZ_rの95%信頼区間を求めることができる．すなわち，

$$Z_U=Z_r+1.96/\sqrt{(N-3)}$$
$$Z_L=Z_r-1.96/\sqrt{(N-3)}$$

を計算する．

上記のZ_UとZ_Lは変換したZ_rの95%信頼区間の上限と下限を与えるが，このままでは母相関係数の信頼区間とならない．

母相関係数の95%信頼区間を求めるためには，これらの値を逆変換する．すなわち，

$$r_U=[\exp(2Z_U)-1]/[\exp(2Z_U)+1]$$
$$r_L=[\exp(2Z_L)-1]/[\exp(2Z_L)+1]$$

を求める．ここで，exp（x）は自然対数の底eのx乗を計算することを表している．

結局，母相関係数の95%信頼区間は，

$$r_L\leqq\text{母相関係数}\leqq r_U$$

と求められる．なお，99%信頼区間を求めるには，上の式の1.96を2.58として計算すればよい．

（例4）1回心拍出量推定のための3方法の各2方法の母相関係数について，95%信頼区間を求めてみよう．

Simpson法とTeichholos法の単相関係数0.542をZ変換すると，

$$Z_r=\frac{1}{2}\cdot\ln\left[\frac{1+0.542}{1-0.542}\right]=0.60698$$

となる．

標本数は199人であったので，

$$Z_U=0.607+1.96/\sqrt{(199-3)}=0.747$$
$$Z_L=0.607-1.96/\sqrt{(199-3)}=0.467$$

とZ値の上限と下限を計算できる．

さらに，上記の値を逆変換すると，

$$r_U=[\exp(2\times0.747)-1]/[\exp(2\times0.747)+1]=0.633$$
$$r_L=[\exp(2\times0.467)-1]/[\exp(2\times0.467)+1]=0.436$$

となる．したがって，母相関係数の95%信頼区間は（0.436，0.633）となる．

同様に，Simpson法とDoppler法の単相関係数は0.549であったので，95%信頼区間は（0.444，0.639）となる．また，Teichholos法とDoppler法の単相関係数は0.574であったので，その95%信頼区間は（0.473，0.660）と求められる．

順位相関係数について

データが量的であっても，その数値があまり信用できない場合や，データの大小関係のみが与えられている順序尺度による場合，測定データの数値自体を用いるのではなく，その順位をもとにした相関係数を用いることがある．そのような相関係数は「順位相関係数rank correlation coefficient」と呼ばれている．

2変数のデータが1からN（標本数）までの順位で与えられている場合，すでに説明したピアソ

ンの積率相関係数は計算が簡単になり，「スピアマンの順位相関係数Spearman's rank correlation coefficient」と呼ばれているものに一致する．すでに説明した単相関係数の計算と同様に行えばよいが，より簡単に以下のように求めることもできる．

いま，2つの変数X，Yのi番目のデータが（x_i, y_i）（i=1，2,...，n）のように，順位で与えられているとすると，順位相関係数r_Sは，

$$r_S=1-\frac{6\Sigma_i(x_i-y_i)^2}{N(N^2-1)}$$

によって求められる．なお，データに同順位がある場合には，上の式は若干の修正が必要となるがここでは，その説明は省略する．

順位相関係数の求め方には他にも方法がある．すなわち，データの大小関係のみから相関係数を求める方法として「ケンドールの順位相関係数」がある．

2つの変数のデータが，1からNまでの順位で与えられている場合，N組のデータから任意に2組取り出すことを考えてみよう．そのような取り出し方は，${}_NC_2=N(N-1)/2$通りある．また，変数Xのデータの順位の大小関係と，Yのデータの順位の大小関係が一致する組の個数をPとし，一致しない組の個数をQとしよう．

ケンドールの順位相関係数τは

$$\tau=\frac{2(P-Q)}{N(N-1)}$$

により求められる．同順位がある場合，この式は若干修正が必要であるが，説明は省略する．

（例5）1回心拍出量についての3方法のデータから，各2変数の組み合わせについて，順位相関係数を求めてみよう．

表4に1回心拍出量についての3方法の各組み合わせによるスピアマンの順位相関係数（**表4**の左下側）とケンドールの順位相関係数（**表4**の右上側）を示した．

表4　順位相関係数行列

変数名	1）	2）	3）
1）Simpson法	1.000	0.390	0.372
2）Teichhols法	0.536	1.000	0.382
3）Doppler法	0.514	0.528	1.000

対角線より下側がスピアマンの順位相関係数
上側がケンドールの順位相関係数

先に求めたピアソンの単相関係数に比べると，スピアマンの順位相関係数は大差ないがやや小さく，ケンドールの順位相関係数はかなり小さくなっていることがわかる．通常はこのような傾向が認められるが，場合によってスピアマンの順位相関係数の方が，ピアソンの単相関係数よりも大きくなることもある．

相関関係と因果関係について

医学的な研究では，病気の原因や治療などに興味があることが多く，かつ応用を目指すため，ある変数間で統計学的な有意な相関が認められると，即「AはBの原因である」といった表現をしがちである．しかし，これは全くの間違いである．

「2変数間に因果関係のある場合，相関関係は必ず存在する」ということは常に正しい．このような関係を図に描いてみよう．いま，2つの変数をAとBとし，Aが原因でBがその結果であるとする．これを図に示すと，**図3**のように描くことができる．

ところが，相関関係は**図4**の(a)や(b)のような場合でも存在する．**図4**は変数として，AとBの他に未知の変数であるXという「第3の変数」が存在する場合を示している．このような場合，

図3　因果関係がある場合

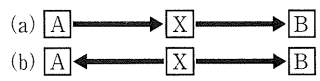


図4　第3の変数が存在する場合

AとBには相関関係は観察されても，それが直接，2変数間の因果関係を示すものでないことは明ら

かであろう

実際には，第3の変数は研究者にとっては知られているとは限らないので，有意な相関が認められても，それが因果関係を示すのかそうでないのかは判断できないことになる．

ちなみに，第3の変数が存在することにより，2変数間に相関がみられる場合，それは「疑似相関spurious correlation」と呼ばれるが，現実には何が疑似で，何が真なのかわからないのが普通であろう．なお，相関関係や因果関係を**図3**や**図4**のように，矢印を用いて示したものは「パス・ダイアグラムpath diagram」などと呼ばれ，仮説的なモデルを図示するのに有効である．場合によっては，変数間の相関関係から，変数間の影響の程度を算出し，矢印（パスpath）の上側にその係数（パス係数）をかいて示すこともあり，「パス解析path analysis」と呼ばれている．

相関関係の使用上の注意

実際に相関係数を用いる場合のいくつかの問題点について，以下に簡単に説明しよう．

1）曲線相関の問題

通常よく用いられるピアソンの積率相関係数は，2変数間の直接的な関係を示すものである．したがって，**図5**のように，2変数の間に曲線的な関係がある場合，単相関係数はその関係を適切に示している数値を与えることはできない．**図5**のような場合には，単相関係数はほとんどゼロになるだろう．このような場合，2変数の関係を最も適切に表示する方法は相関図を描くことである．数式を用いて，その関係を示したいのならば，2次曲線を当てはめたり，必要によってはさらに高次の曲線を当てはめる必要があるかもしれない．また，順位相関係数が適している場合もあるだろう．単に相関係数を数値として求めるだけでなく，まず相関図を描くことは重要である．

2）外れ値の問題

相関図を描いた場合，多くのデータが集まっている部分から離れた位置にある点が「外れ値outlier」である．**図6**で，もし外れ値が(1)の位置にあり，(2)になければ，相関係数は正の方向に大きくなる．逆に，(1)になく，(2)に外れ値があれば，負の方向に相関を偏らせ，大多数の部分では正の相関があるにもかかわらず，全体では0に近い相関係数となるだろう．

このように，相関図を描いたとき，他の多くのデータから極めて離れている点については，外れ値として特別の配慮が必要になる．

まず第一に，そのデータは誤りではないかという問題がある．これを調べるには，原データ（調査票や原票など）に戻って厳重にチェックする必要がある．

データに誤りがない場合には，この外れ値を含めて分析するか，除外するかという問題が生じる．一つの方法としては，外れ値を含めた場合と除いた場合の2通りの分析を行い，あまり差がないようであれば含めてしまうという考え方がある．しかし，一般に分析結果の相違は大きいことが多いので，異質なものとして取り除いた方が無難である．しかし，データチェックや，何故そのような

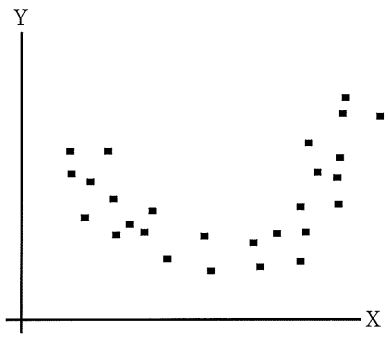


図5　曲線相関

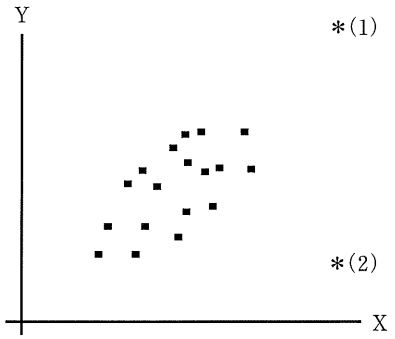


図6　外れ値の影響

データが観察されたのかの考察なしに，何でも外れ値ならば除けばよいというのも問題である．データを除くのだから，それなりのきちっとした理由付けが必要となる．

外れ値のある場合には，観測されたデータをそのまま用いるのではなく，順位などに基づく分析を行うのもよいかもしれない．例えば相関係数の計算には，順位相関を求めるなどの方法がある．いずれにしろ，データに外れ値を含む場合，その取り扱いによっては結果がまったく異なったものになる可能性が高い．データをよく吟味するためにも，相関図を作成しておくことが必要であろう．

3）打ち切りデータ

大学の入学試験の成績と入学後の成績にはあまり相関がないらしい．これが事実だとすると入試などはやる必要がないことになる．しかし，この論理は少し考えてみるとおかしいことがすぐにわかるはずである．大学入学後の成績のデータが手に入るのは，入試の成績が合格点以上である一部のものについてだけである．もしも入試の成績にかかわらず全員を入学させ，その後の成績と比べてみたらどうなるだろう．入試の成績が悪いものは，やはり入学後も成績がよくないかもしれないし，本当に何の関係も両者にはみられないかもしれない．常識的には，両者に正の相関があるものと考えらるのが普通であるが，なぜ入試の成績と入学後の成績に相関があまり認められないのであろうか．

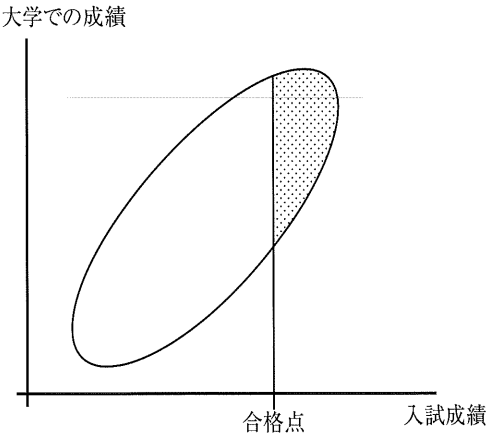


図7　打ち切りデータ

図7にそのような関係を考えるための相関図を示した．仮に正の相関関係が両者にある場合でも，図に示したように，入試の成績により，ある点（合格点）でデータが打ち切られてしまう．すなわち，合格点に達しない者のデータは入手することができず，合格点以上のわずかなデータしか観測することができない．この結果，合格点のところで集団が2分され，一方のデータのみが相関係数の計算に使用されることになる．このようなデータは「打ち切りデータtruncated data」と呼ばれる．両者に正の相関があるとしても，**図7**に示したように，実際に使用される灰色の部分のデータでは，それほど相関は認められなくなるだろう．このように，データが何らかの基準（例えば，血圧値やコレステロール値など）によって打ち切られたものでないかを確かめることは，分析の前によく考慮しておく問題である．

4）層別化について

いま数学の成績と英語の成績の相関を調べているとしよう．相関係数を求めたところ，弱い正の相関が認められるだけであったと仮定しよう．この結果からすると，英語ができることと数学ができることには関係がたいしてないことになる．ところで，**図8**を見てみよう．これは，男女別に2科目の相関図を描いた場合である．

男女別にみると，英語と数学の成績には正の相関が認められるが，全体でみるとこの関係が弱まってしまうことになる．一般に，男は数学が得意

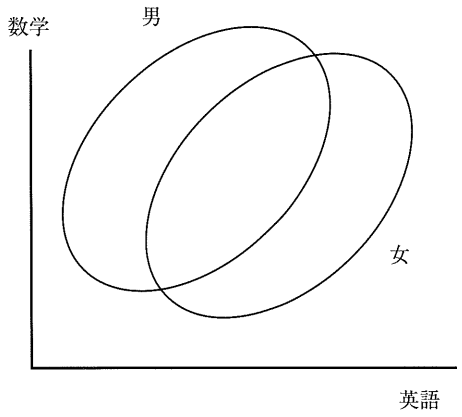


図8　層別化と相関

で，女は英語が得意であるという傾向をもつようである．**図8**はやや大げさに表示してあるが，このような状況もあり得るわけである．このような場合，男女別に相関係数を求める必要がある．これは，人間の性という変数によって集団を分割する方法であり，一般にある変数によって集団を分割することを「層別化stratification」と呼び，分割された各小集団を層と呼ぶ．

層別化のための変数としてよく用いられるものとして，性，年齢（5歳ごと，10歳ごと等），職業，疾患の重症度などがある．健常者のみを調べているつもりであっても，その中に特定の病気を持った人達が標本として混在していないかなどを，相関係数を計算する以前に吟味することは重要である．これは単に相関に関してのみならず，他の分析の場合でも必要な考え方である．とくに標本数が多いときには性別，年齢別などの分析をすることで，まったく異なる結果を得ることもある．

5）有意な相関と意味のある相関

2変数の相関係数を求めた場合，通常は前述した「無相関の検定」を行うことが多い．この検定の結果，相関は有意（5％水準や1％水準で）となった場合，多くの人はそれが2変数間の相関に極めて意味のある関係が証明されたと考えてしまうのではないだろうか．ところが，実際の検定の意味は，単に「母集団で2つの変数の相関係数は0ではない」と考えてよいことを示しているだけなのである．例えば，相関係数が0.1であるとしよう．これは，ほとんど2変数間に相関のない状態と考えられるが，仮に標本数が500あれば，検定のためのt値を計算すると，$t_0 = 2.243$となり，5％水準で有意となる．確かに0.1は0ではないのであるから，相関が有意となっても一向に構わないことである．しかし，この結果を過大に評価すべきではない．もし，この結果から一方の値を知ることで他方も予測できるなどとして，回帰直線などを求めても意味のないことである．

相関係数の有意性の検定は重要な方法ではあるが，これによってのみ結果を評価することは，正しいデータ解析の方法とは考えられない．有意でかつ大きな値をもつような相関が価値のあるものと考えられよう．仮に，相関係数が大きくても，標本数が少なすぎれば有意にはなりにくいものである．そのような場合，検定結果が有意でなくとも，すぐに2変数間の相関はないものと結論づけるのではなく，次に調べるときには，より多くの標本数を得られるような計画を立て，その結果と比較できるようにすべきであろう．

検定が有意だからといって，実質的には無意味な場合もあり，逆に有意でなくとも意味のある相関を示していることもあり得るのである．

おわりに

今回は，相関関係に関する話題をかなり広範に盛り込んだ．これは，相関関係というのは重要であるが，間違った解釈や誤用もその分だけ多くあると考えられるからである．

今回，説明した各種の計算については，これまで同様に私のホームページで計算することができる．

参考文献

1）高木廣文：ナースのための統計学，医学書院（東京），1984
2）高木廣文，三宅由子：よくわかる医療・看護のための統計学入門，メディカ出版（大阪），1991
3）柳井晴夫，高木廣文　編著：統計学－基礎と実践，メヂカルフレンド社（東京），1995