# 医学における統計学の応用について

## 第3編　相関・回帰分析を中心として

新潟大学医学部衛生学教室（主任：山本正治教授）
遠　藤　和　男

# An Application of Statistical Methods to Medical Science

## Part III　On the Analysis of Correlation and Regression

Kazuo ENDOH

*Department of Hygiene and Preventive Medicine, Niigata University School of Medicine*
*(Director: Prof. Masaharu YAMAMOTO)*

The author has already reported how to deal with chi-square test for a difference of ratios and how to apply Student's or Welch's t-test for a difference of means elsewhere [1)2)]. Present report is chiefly stressed on statistical methods for the analysis of correlation and regression. Since these methods provide important clues to not only epidemiological but also other medical studies, the author offers some examples and comments for application of these methods.

Key words: analysis of correlation and regression, Pearson's product-moment method, Spearman's rank correlation coefficient, multiple regression analysis
相関・回帰分析，ピアソンの積率法，スピアマンの順位相関係数，重回帰分析

## はじめに

第1編[1)]では，$\chi^2$ 検定を中心とした比率の差の検定について，また第2編[2)]では，2群の平均値の差の検定を中心として，連続変量の評価について述べた．本編で述べる相関・回帰分析は，疫学の分野で因果関係[3)]を裏付ける重要な指標を提供するし，他の医学的分野でも広く応用されているので，医学研究者の参考となるように若干の具体例を示した．特に，最近盛んに応用されている多変量解析[4)]のうち，重回帰分析についても具体例をあげて解説を加えた．

### 1. どちらを x，どちらを y とするか？

1人が（x，y）という1組のデータを持っていたと仮定する．x，yは単位が同じでも，異なっていてもどちらでもよい．ただ単に相関係数を求めるためには，x，yは入れ替わってもかまわない．しかし，yをxで説明

Reprint requests to: Kazuo ENDOH, Department of Hygiene and Preventive Medicine, Niigata University School of Medicine, Niigata City, 951, JAPAN.

別刷請求先：〒951 新潟市旭町通1番町
新潟大学医学部衛生学教室　遠藤和男

表 1　飛び離れた値を含むデータ

| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 平均値 | 標準偏差 |
|---|---|---|---|---|---|---|---|---|---|---|
| x | 1.7 | 3.0 | 1.1 | 4.0 | 2.1 | 4.6 | 3.8 | 10.0 | 3.79 | 2.786 |
| y | 3.9 | 2.8 | 2.0 | 4.1 | 2.1 | 2.6 | 1.6 | 8.0 | 3.39 | 2.063 |

No. 8 の x, y について，Grubbs-Smirnov の棄却検定を実行すると，

$T_0=\frac{10.0-3.79}{2.786}=2.229>T_8(0.05)=2.032$，$T_0=\frac{8.0-3.39}{2.063}=2.235$

したがって，x，y とも飛び離れた値とみなすことができる．

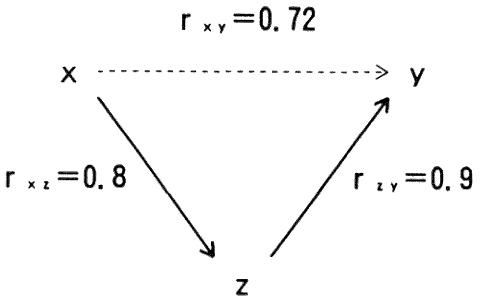


図 1　見かけ上の因果関係

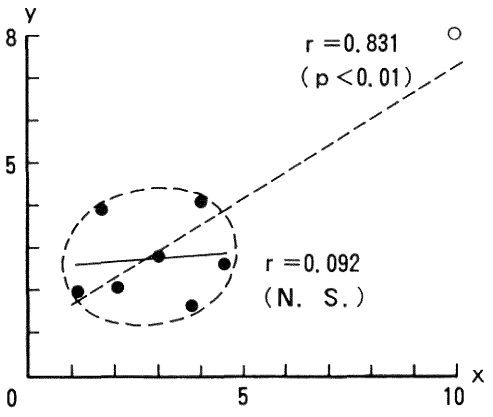


図 2　飛び離れた値を含む場合

して一次関数（回帰方程式）として表すことが多いので，xを**独立変数**（説明変数），yを**従属変数**（目的変数）と呼んでいる．したがって単位が等しい場合は，時間的に古い変数をxとし，また単位が異なる場合でも，より信頼性の高いと考えられる変数をxとした方がよい．また，相関係数が有意であった場合には，「正または負の相関関係が認められる」のであって，疫学的な“因果関係”[4]が立証されたわけではない．

【例1】xとyとの相関係数 $r_{xy}=0.72$ と有意であったとする．しかし，**図 1** のように自分が考慮に入れな

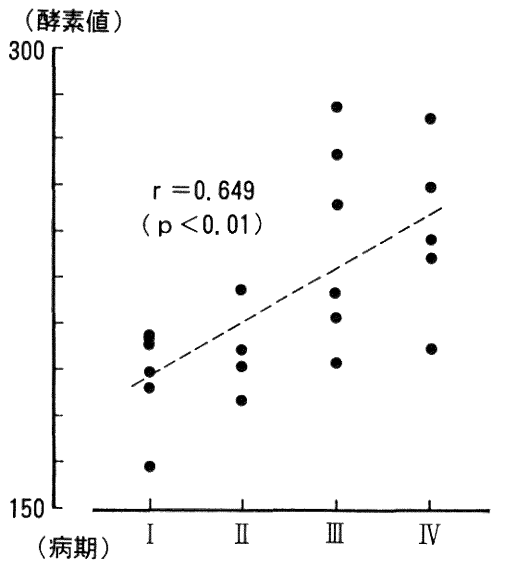


図 3-1　順序尺度による散点図

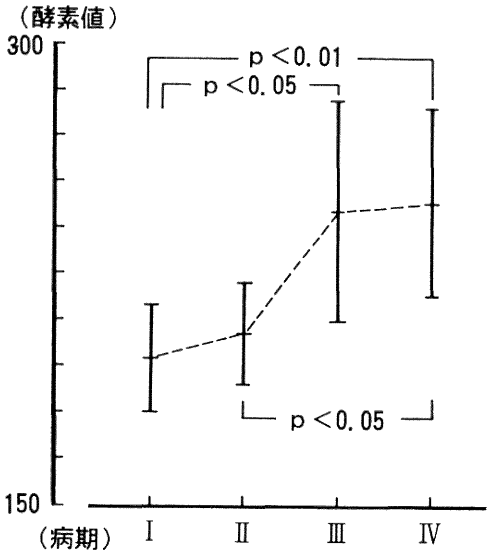


図 3-2　病期別の平均値±標準偏差

かった真の因子zが存在する場合も多い．したがって相関関係は，因果関係が認められるための必要条件の一つ（関連の強固性）ではあるが，十分条件とはなり得ない．

### 2. 散点図を描くことの重要性

相関・回帰分析を行う場合，**散点図**（scatter diagram，散布図，相関図）は重要な情報を提供するが，手で描くと案外時間がかかるため，以下のような誤りを犯すこともある．

【例2】**表 1** に示すようなデータがあったとする．No. 8のデータは，x，yとも飛び離れているようにみえる．**飛び離れた値の棄却検定**（Grubbs-Smirnov の検定）[5)6)]を実施しない場合，**図 2** に示すように $r=0.831$（$p<0.01$）と，有意な相関があると誤って判断してしまう．棄却検定によって No. 8 のデータはx，yとも棄却できる．$n=7$として改めて相関係数を算出すると，$r=0.092$ と有意性は認められなくなってしまう．

【例3】病期別にある酵素の値を測定したところ，**表 2** のような結果を得た．ここで病期Ⅰ～Ⅳは，連続変数[7)]ではなく順序尺度[7)]となっており，また必ずしも等間隔性が保証されているわけではない．ところが，この病期をx軸にとると散点図は，**図 3-1** のようになる．相関分析では通常，xとyとは1：1の対応をしているが，**図 3-1** では，病期xと酵素の値yとは1：複数であり，しかも一定していない．動物実験等で1：}と，繰り返し数}が一定している場合には，共分散分析[8)]が適用できる．しかし患者データでは，各患者を繰り返しと考えられないので，以下のように考えるとよい．

(1) 各病期の平均値及び標準偏差を算出する．スチューデント法等，2群の平均値の差の検定[2)]を実施して，**図 3-2** のように有意差を表示する．

**表 2**　順序尺度によるデータ

| 病期 / 人数 | Ⅰ | Ⅱ | Ⅲ | Ⅳ |
|---|---|---|---|---|
| 1 | 205 | 201 | 248 | 202 |
| 2 | 206 | 221 | 265 | 276 |
| 3 | 161 | 197 | 197 | 237 |
| 4 | 190 | 185 | 220 | 254 |
| 5 | 194 | — | 212 | 230 |
| 6 | — | — | 281 | — |
| 平均値 | 191.2 | 201.0 | 237.2 | 239.8 |
| 標準偏差 | 18.24 | 14.97 | 32.73 | 27.59 |

病期Ⅰ～Ⅳを $x=1\sim4$ とした場合，
$r=0.649$（$p<0.01$）と判定してしまう．

(2) 順序尺度が等間隔である場合には，順序尺度 $x_i$ と各平均値 $\bar{y}_i$ は1：1に対応している．直線性を確かめたい場合には，改めて相関係数を算出して評価する．

### 3. 異質なデータを一緒にした場合

医学データは，男女で異なる傾向を示す場合が多い．男女別々に分析すると，それぞれ有意な相関が認められるにもかかわらず，男女を一緒にしたために，みかけ上相関が認められなくなる場合がある．

【例4】学部1年生に疫学及び医統計の試験を実施したところ，**表 3** のような成績であった（仮想データ）．男女を別々にして分析すると，**図 4-1** に示したように $r=0.661$（$p<0.01$）と有意である．しかし男女を一緒にして相関分析を行うと，**図 4-2** のように $r=0.100$ と有意性は認められなくなってしまう．例数が少ないからといって，異質なデータを一緒にすることは危険である．

この例は，男女を説明変数の1つと考えて，多変量解析の1つである．判別分析（Discriminant analysis）[4)]

**表 3**　異質なデータを一緒にした場合

| No. | 性別 | 医統計 | 疫　学 | 合　計 |
|---|---|---|---|---|
| 1 | 男 | 50 | 34 | 84 |
| 2 | 男 | 46 | 38 | 84 |
| 3 | 男 | 44 | 35 | 79 |
| 4 | 男 | 42 | 32 | 74 |
| 5 | 男 | 40 | 28 | 68 |
| 6 | 男 | 36 | 38 | 74 |
| 7 | 男 | 36 | 24 | 60 |
| 8 | 男 | 34 | 31 | 65 |
| 9 | 男 | 32 | 26 | 58 |
| 10 | 男 | 30 | 22 | 52 |
| 11 | 女 | 34 | 50 | 84 |
| 12 | 女 | 38 | 46 | 84 |
| 13 | 女 | 35 | 44 | 79 |
| 14 | 女 | 32 | 42 | 74 |
| 15 | 女 | 28 | 40 | 68 |
| 16 | 女 | 38 | 36 | 74 |
| 17 | 女 | 24 | 36 | 60 |
| 18 | 女 | 31 | 34 | 65 |
| 19 | 女 | 26 | 32 | 58 |
| 20 | 女 | 22 | 30 | 52 |

男女を一緒にして相関係数を求めると，$r=0.100$ と有意差を求めないが，男女を別々にした場合には，それぞれ，$r=0.661$（$p<0.05$）と有意である．

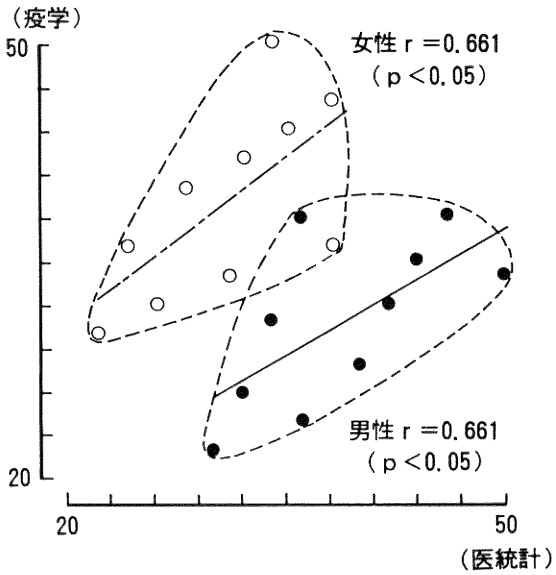


図 4-1　男女を別々にした場合

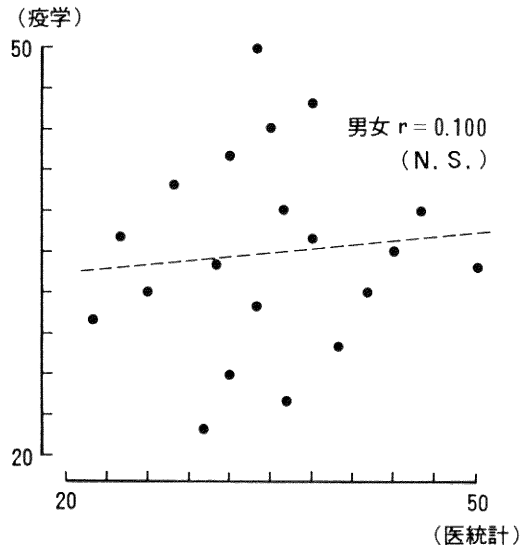


図 4-2　男女を一緒にした場合

の例題として用いることもできるが，詳細は省略する．

### 4．相関係数の有意性の検定について

図 5 に示すように相関係数 r の分布は，母相関係数 $\rho$ が 0 に近い場合には正規分布に近似できる．また $\rho$ が大きくて 1 に近い場合には独特な分布を示す．

したがって相関係数の有意性の検定は，以下のように

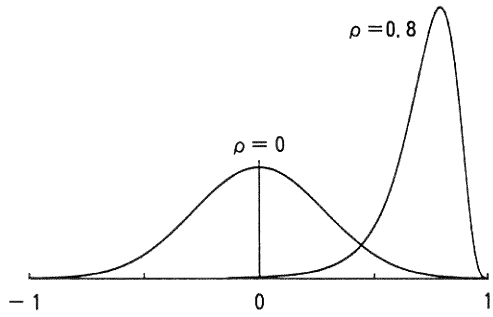


図 5　標本相関係数 r の分布

大きく 2 とおりに分けられる．

(1) $|r| \geqq 0.75$ の場合，**Fisher の Z 変換**[9] を用いる．

$Z = \frac{1}{2} \ln \frac{1+r}{1-r}$（ln は自然対数）として，

$Z_0 = Z\sqrt{n-3} \sim N(0, 1)$　　…… 式①

なお，Z 変換の値は統計数値表[10] に掲載されているので，わざわざ計算する必要はない．

(2) $|r| < 0.75$ の場合，t 分布を用いる（$n < 50$）．

$t_0 = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} \sim t_{n-2}$　　…… 式②

$t_0 > t_{n-2}\left(\frac{\alpha}{2}\right)$ なら，有意水準 $\alpha$ で有意である．

ただし，n が大きくなると $t_{n-2}(\frac{\alpha}{2})$ の値が正確には求められないし，図 5 のように正規分布に近づくため，$n \geqq 50$ の場合には，次式による方が便利である．

$Z_0 = \frac{r}{1-r^2} \sqrt{n-1} \sim N(0, 1)$　　…… 式③

式②と式③との区別は厳密ではない．しかし，式①の代わりに式②を用いると，有意性の判断を誤る場合があるので，十分に注意する必要がある．

【例 5】$n = 5$，$r = 0.880$ の時，$|r| \geqq 0.75$ の条件を無視して式②を用いると，

$$t_0 = \frac{0.88}{\sqrt{1-0.88^2}} \sqrt{3} = 3.209 > t_3\left(\frac{0.05}{2}\right) = 3.182$$

となり $p < 0.05$ で有意である．一方，式①によると，

$Z_0=1.375768\sqrt{2}=1.946<Z\left(\frac{0.05}{2}\right)=1.96$

であるから，有意性は認められなくなってしまう．

### 5. 正規分布に従わないデータ

これまでx，yはともに正規分布に従うと仮定して，ピアソンの積率法（Pearson's product-moment method）によって相関係数を求めてきた．すなわち，相関係数rは次式によって定義される．

$$r=\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\,\Sigma(y_i-\bar{y})^2}} \quad \cdots\cdots \text{式④}$$

**表 4**　一次回帰式で説明できない場合

| No. | 術後日数 | 酸素値 | ln（日数） |
|---|---|---|---|
| 1 | 1 | 35 | 0 |
| 2 | 2 | 45 | 0.693 |
| 3 | 3 | 36 | 1.099 |
| 4 | 4 | 47 | 1.386 |
| 5 | 5 | 44 | 1.609 |
| 6 | 6 | 50 | 1.792 |
| 7 | 14 | 46 | 2.639 |
| 8 | 21 | 51 | 3.045 |
| 9 | 28 | 48 | 3.332 |
| 10 | 35 | 49 | 3.555 |

術後日数をx，酸素値をyとすると，r=0.580（p=0.079）と有意でないが，日数の対数値をxにとった場合には，r=0.738（p<0.05）と有意である．

**表 5**　表4のデータの順位

| No. | $R_i$ | $S_i$ | $D_i$ | $D_i^2$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 2 | 4 | −2 | 4 |
| 3 | 3 | 2 | 1 | 1 |
| 4 | 4 | 6 | −2 | 4 |
| 5 | 5 | 3 | 2 | 4 |
| 6 | 6 | 9 | −3 | 9 |
| 7 | 7 | 5 | 2 | 4 |
| 8 | 8 | 10 | −2 | 4 |
| 9 | 9 | 7 | 2 | 4 |
| 10 | 10 | 8 | 2 | 4 |

$\Sigma D_i^2=38<\underline{D}_{0.025}=58$ Spearman の相関係数 $r_s$ は，

$r_s=1-\frac{6\Sigma D_i^2}{n^3-n}=0.770$ と，日数の対数値をxにとった場合の有意性（p<0.05）と同じである．

しかしながら，x，yのうち少なくとも一方が正規分布に従わないか，またはnの数が少なくて正規性が不明な場合，ノンパラメトリック検定技法を用いなくてはならない．代表的な方法として，**スピアマンの順位相関係数**（Spearman's rank correlation coefficient）[11]の求め方の例をあげる．

【例6】手術後のある測定値について，**表 4** のような成績が得られた．手術後の日数をx，測定値をyとすると，ピアソンの式④による相関係数 $r_p=0.580$ と有意差は認められない（**図 6-1**）．しかし，日数の自然対数をとってx軸とすれば，r=0.738（p<0.05，**図 6-2**）

**表 6**　重回帰分析用データ

| No. | 年齢（歳） | 身長（cm） | 体重（Kg） | 肥満度（%） | 最大血圧（mmHg） |
|---|---|---|---|---|---|
| 1 | 25 | 170.0 | 60.0 | −4.76 | 120 |
| 2 | 28 | 172.0 | 63.0 | −2.78 | 126 |
| 3 | 30 | 160.0 | 56.0 | 3.70 | 118 |
| 4 | 35 | 164.5 | 53.5 | −7.84 | 124 |
| 5 | 38 | 164.0 | 55.0 | −4.51 | 126 |
| 6 | 40 | 161.5 | 51.5 | −6.96 | 122 |
| 7 | 45 | 160.0 | 60.0 | 11.11 | 128 |
| 8 | 48 | 168.0 | 63.0 | 2.94 | 138 |
| 9 | 50 | 161.0 | 58.3 | 6.19 | 130 |
| 10 | 55 | 159.4 | 60.0 | 12.23 | 132 |
| 11 | 58 | 158.4 | 56.6 | 7.69 | 142 |
| 12 | 60 | 150.7 | 53.0 | 16.15 | 134 |
| 13 | 65 | 158.5 | 71.8 | 36.37 | 136 |
| 14 | 68 | 159.6 | 59.0 | 9.99 | 144 |
| 15 | 70 | 152.0 | 59.0 | 26.07 | 140 |

Broca-桂変法により，標準体重＝［身長（cm）−100］×0.9 として，

$$\text{肥満度}=\frac{\text{実測体重}-\text{標準体重}}{\text{標準体重}}\times 100\ (\%)$$

で求めた．

**表 7**　肥満度を説明変数に加えた場合

| 説明変数 | 偏相関係数 | F 値 | 有意差 |
|---|---|---|---|
| 年　齢 | 0.924 | 58.399 | p<0.001 |
| 身　長 | 0.263 | 0.745 | N.S. |
| 体　重 | −0.150 | 0.229 | N.S. |
| 肥満度 | 0.108 | 0.117 | N.S. |

自由度調整済決定係数 $R^2=0.874$（P<0.001）

$F_{10}^4\left(\frac{0.01}{2}\right)=7.343$，$F_{10}^4\left(\frac{0.001}{2}\right)=13.406$

年齢だけが有意であると誤って判定する．

表 8　肥満度を説明変数から除外した場合

| 説明変数 | 偏相関係数 | F　値 | 有意差 |
|---|---|---|---|
| 年　齢 | 0.924 | 63.806 | p<0.001 |
| 身　長 | 0.738 | 13.178 | P<0.01 |
| 体　重 | −0.356 | 1.598 | N.S. |

自由度調整済決定係数 $R^2=0.884$ ($P<0.001$)

$F_{11}^{3}\left(\frac{0.01}{2}\right)=7.600$, $F_{11}^{3}\left(\frac{0.001}{2}\right)=13.655$

年齢のほか身長にも有意差が認められる.

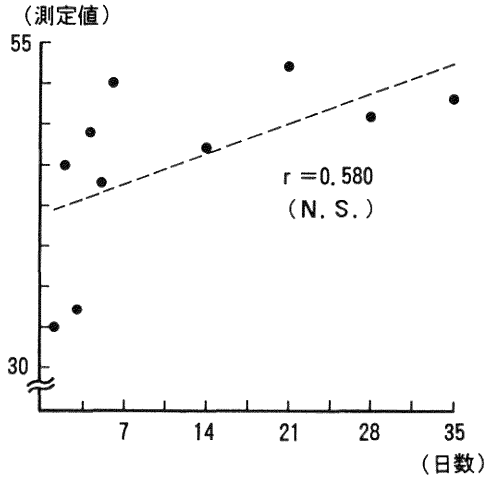


図 6-1　一次回帰式によらない場合

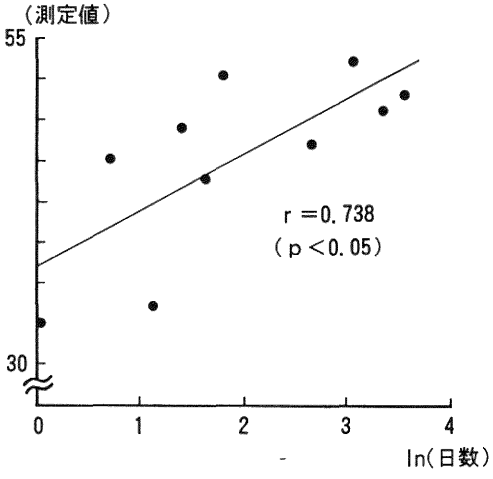


図 6-2　日数の自然対数をとった場合

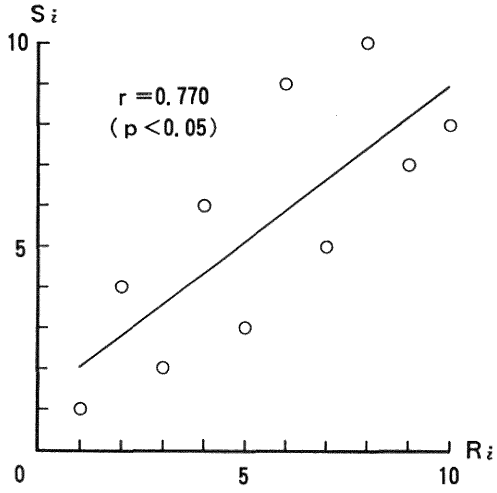


図 6-3　順序を x, y 軸とした場合

となる. ただし, n=10と小さく正規性が疑わしいので, スピアマンの順位相関係数 $r_s$ を求めてみる. 表 5のように, $x_i$, $y_i$ について順位 $R_i$, $S_i$ とその差 $D_i$ を求めて, 次式に代入する.

$$r_s=1-\frac{6\Sigma D_i^{2}}{n^3-n}=0.770 \qquad \cdots\cdots \text{式} ⑤$$

もし, $S_i$ にタイの順位があった場合には, 次式によって補正する. タイの数 t, 大きさ u として,

$$\text{補正値 } U=\sum_{i=1}^{t}\frac{u_i^{3}-u_i}{12} \qquad \cdots\cdots \text{式} ⑥$$

$$r_s=1-\frac{6(\Sigma D_i^{2}+U)}{n^3-n} \qquad \cdots\cdots \text{式} ⑦$$

図 6-3 に示すように, 一次回帰式に従わない場合, スピアマン法の方が有利な場合が多い. またタイの順位がない場合, $R_i$, $S_i$ を $x_i$, $y_i$ とみなして求めたピアソンの相関係数 $r_p$ は, $r_s$ と全く等しい値となる.

### 6. 重回帰分析について

単相関分析では, 目的変数 y に対する説明変数 x は1つだけであるが, 説明変数 $x_i$ が複数ある場合には, 重回帰分析と呼ばれ, 最近, 医学の分野でも盛んに応用されてきている. ただし単相関の場合, 説明変数が独立変数と呼ばれているように, 各々の説明変数 $x_i$ が互いに独立でないと, 結果の解釈を誤る場合がある.

【例7】男性の収縮期血圧を目的変数とし, 年齢, 身長, 体重のほか, さらに肥満度を目的変数に加えた場合

を考えてみる．肥満度は様々な算出方法[12]が提唱されているが，**表 6** に示したように Broca-桂変法を用いた場合，標準体重＝[身長(cm)－100]×0.9 で示される．

肥満度のように他の変数の式で表されるような場合，独立した説明変数とは認められない．にもかかわらず，肥満度を説明変数とした場合には，**表 7** に示すように，年齢だけに有意差が認められ，他は有意差なしと判定してしまう．一方，肥満度を説明変数から除外した場合，年齢に加えて身長にも有意差が認められるようになる（**表 8**）．

そのほか，説明変数に性別，既往の有無といった名義尺度[7]を入れた場合，通常の重回帰分析でなく，林の数量化理論Ⅰ類[13]を用いるべきである．詳しくは，専門書[4)14)15)]を参照されたい．

## 参 考 文 献


1) **遠藤和男**：医学における統計学の応用について（第1編）2×2表を中心として．新潟医誌，**102**：147～154，1988．

2) **遠藤和男**：医学における統計学の応用について（第2編）平均値の差の検定を中心として．新潟医誌，**104**：78～85，1990．

3) **近藤東郎，糸川嘉則，山本正治，監訳**：疫学テキスト，第2版，pp 151～161，西村書店（新潟），1986．

4) **田中　豊，乗水共之，脇本和昌，編**：パソコン統計解析ハンドブック，Ⅱ多変量解析編，共立出版（東京），1984．

5) **Grubbs, F.E.**：Sample criteria for testing outlying observations, Ann. Math. Statist., **21**：27～58, 1950.

6) **Smirnov, N.V.**：On the estimation of the maximum term in a series of observation. Dokl. Akad. Nauk. SSSR, **33**：364～350, 1941.

7) **遠藤和男，山本正治**：医統計テキスト，pp 4～6，西村書店（新潟），1992．

8) **田中　豊，乗水共之，編**：パソコン統計解析ハンドブック，Ⅱ実験計画法編，pp 414～426，共立出版（東京），1986．

9) **Fisher, R.A.**：On the "probable error" of a coefficient deduced from a small sample. Merton, **1(4)**：1～32，1921．

10) **統計数値表編集委員会，編**：簡約統計数値表．pp 104～105，日本規格協会（東京），1985．

11) **Spearman, C.**：The proof and measurement of association between two things. Amer. J. Psycol., **15**：72～101，1904．

12) **片桐邦三**：肥満の判定基準―管理の実際―，日本医事新報，**No. 3009**：3～12，1981．

13) **Hayashi, C.**：On the quantification of qualitative data form the mathematico-statistical point of view. Ann. Inst. Statist. Math., **2**：35～47, 1950.

14) **柳井晴夫，高根芳雄**：多変量解析法．朝倉書店（東京），1977．

15) **柳井晴夫，高木廣文，編著**：多変量解析ハンドブック．現代数学社（京都），1986．


（平成4年2月5日受付）