ファイ係数について[1)]

On the phi coefficient

行動科学科　岡 本 安 晴

2変量 X と Y の関係を示す指標の1つに相関係数 r があるが、X 及び Y が0と1の2値しかとらない場合には相関係数はファイ係数 ϕ となる。この ϕ の特徴として、値が概して小さいといわれている(cf.　柏木、1974)。X と Y が、2次元正規分布に従う確率変数 x と y から導かれたものであると考えられる場合に、x と y の真の相関係数 ρ とファイ係数 ϕ とのずれはどの様であるのかという問題は、ファイ係数 ϕ を用いて種々の解釈を行う場合、重要な意味をもちうる。又、ファイ係数 ϕ を真の相関係数 ρ の推定値の1つとしてみた場合、ρ と ϕ 及び他の推定法との関係を知っておくことも、ファイ係数 ϕ を使う際の1つの指針となりうる。本稿では、前者の問題については、ρ の色々な値について確率分布関数から ϕ を計算することにより ρ と ϕ の関係を調べてみた。後者の問題については、種々の ρ の値に対して決まるそれぞれの2次元正規分布に従う乱数により生成されたデータに対して、相関係数 r、ファイ係数 ϕ、及び0－1型データより推定する相関係数（r_{0-1} と表わしておく）を計算することにより、r と ϕ 及び r_{0-1} の関係を調べた。

1　相関係数とファイ係数の関係（標本誤差を含まない場合）

0－1型の確率変数 X と Y が2次元正規分布に従う確率変数 x と y から導かれる、即ち

$$
\mathrm{X}=\begin{cases} 0 & x<\mathrm{c}_x\text{のとき} \\ 1 & x\geq \mathrm{c}_x\text{のとき} \end{cases}
$$

$$
\mathrm{Y}=\begin{cases} 0 & y<\mathrm{c}_y\text{のとき} \\ 1 & y\geq \mathrm{c}_y\text{のとき} \end{cases} \tag{1}
$$

なる関係にあるとする。相関係数及びファイ係数について考えるときには、x 及び y の分散は1であると仮定してよい。相関係数を ρ とおけば、x と y は次の確率密度関数 $f(x, y)$ に従うことになる。

1）プログラムはすべて FACOM OS IV PASCAL で書かれ、実行に際しては金沢大学情報処理センターを利用した。

$$f\ (x,\ y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2\,(1-\rho^2)}\,(x^2-2\rho xy+y^2)\right\} \tag{2}$$

$\alpha,\ \beta,\ \gamma,\ \delta$ などを次のようにおく。

$$\alpha = \int_{c_y}^{\infty}\int_{c_x}^{\infty} f\ (x,\ y)\ dxdy \tag{3}$$

$$\beta = \int_{c_y}^{\infty}\int_{-\infty}^{c_x}\ (x,\ y)\ dxdy$$

$$\gamma = \int_{-\infty}^{c_y}\int_{c_x}^{\infty} f\ (x,\ y)\ dxdy$$

$$\delta = \int_{-\infty}^{c_y}\int_{-\infty}^{c_x} f\ (x,\ y)\ dxdy$$

$$p = \alpha + \beta = \int_{c_y}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy \tag{4}$$

$$q = 1 - p$$

$$p' = \alpha + \gamma = \int_{c_x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \tag{5}$$

$$q' = 1 - p'$$

このとき、X の Y のファイ係数 ϕ は

$$\phi = \frac{\alpha\delta - \beta\gamma}{\sqrt{pqp'q'}}$$

で与えられる。

表一1　相関係数 ρ とファイ係数 ϕ

表1－1　$\rho=0.1$のときの ϕ の値

p' \ p	0.1	0.3	0.5	0.7	0.9
0.1	.037	.046	.047	.043	.031
0.3		.058	.061	.057	.043
0.5			.064	.061	.047
0.7				.058	.046
0.9					.037

表－1－2　$\rho=0.3$のときの ϕ の値

p' \ p	0.1	0.3	0.5	0.7	0.9
0.1	.129	.146	.139	.118	.078
0.3		.182	.184	.167	.118
0.5			.194	.184	.139
0.7				.182	.146
0.9					.129

表－1－3　$\rho=0.5$のときの ϕ の値

p' \ p	0.1	0.3	0.5	0.7	0.9
0.1	.249	.257	.225	.177	.103
0.3		.318	.313	.271	.177
0.5			.333	.313	.225
0.7				.318	.257
0.9					.249

表－1－4　$\rho=0.7$のときの ϕ の値

p' \ p	0.1	0.3	0.5	0.7	0.9
0.1	.409	.378	.298	.211	.111
0.3		.479	.453	.365	.211
0.5			.494	.453	.298
0.7				.479	.378
0.9					.409

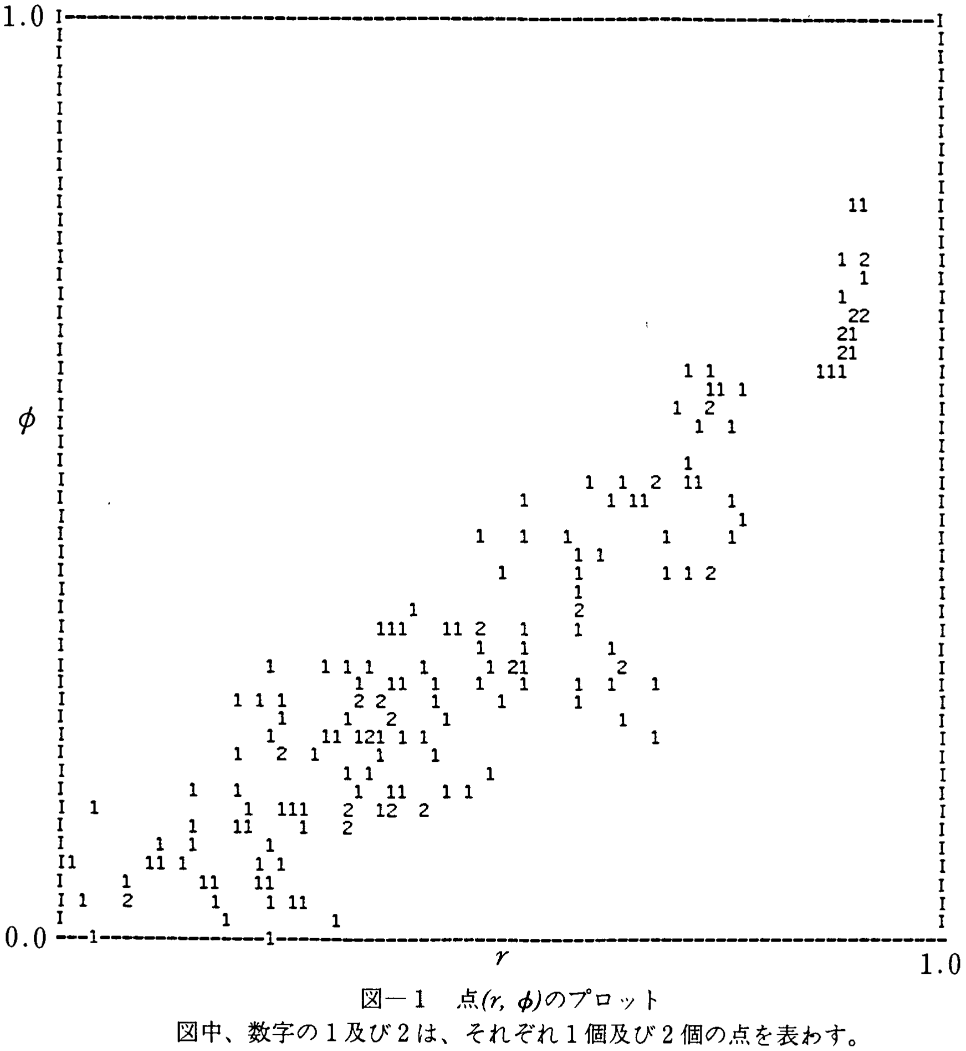
表－1－5　$\rho=0.9$のときの ϕ の値

p' \ p	0.1	0.3	0.5	0.7	0.9
0.1	.654	.493	.333	.218	.111
0.3		.703	.609	.426	.218
0.5			.713	.609	.333
0.7				.703	.493
0.9					.654

表一1に、$\rho=0.1, 0.3, 0.5, 0.7, 0.9$ の場合に、p 及び p' がそれぞれ 0.1, 0.3, 0.5, 0.7, 0.9 の値をとるときの ϕ を計算したものをまとめた。ϕ はかなり ρ より小さくなっていることが分かるが、この傾向は、p 又は p' のいずれかのみが小さいときに著しい。p と p' がともに 0.5 のときには、ϕ は比較的 ρ に近くなっている。しかし、例えば、$\phi \fallingdotseq 0.1$ ぐらいの値は、$\rho=0.3, 0.5, 0.7, 0.9$ のいずれの表にも現われており、ϕ のみの値から ρ を推定することは危険である。

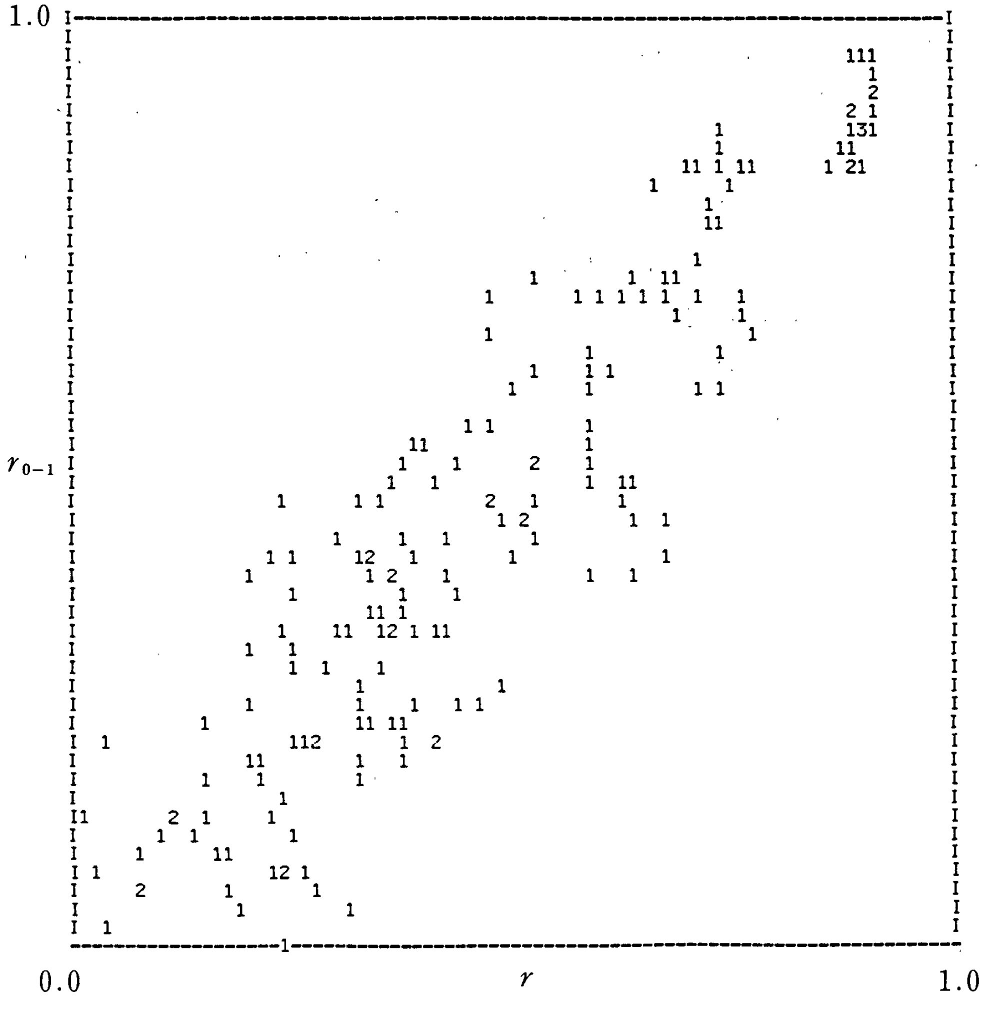
2　相関係数とファイ係数の関係（標本値の場合）

連続変量と考えられるものが何らかの基準なりデータ収集上の都合により 0－1 型のデータとして与えられている場合について、元の連続変量のデータから算出されるはずであった相関係数の標本値 r と、0－1 型のデータから計算されたファイ係数 ϕ との関係をシミュレーションにより調べた。確率密度関数(2)式に従う乱数により 1 系列100対のデータ (x, y) からなる177系列のデータを作成した。この (x, y) の各系列に対して、式(1)により 0－1 型のデータ対 (X, Y) の系列177系列を作成した。これらの系列に対して、(x, y) の系列には相関係数 r を、(X, Y) の系列にはファイ係数 ϕ を算出し、点 (r, ϕ) をプロットしたものが図一1である。ϕ が r に比べて小さいことがよく表わされている。



図一1　点 (r, ϕ) のプロット
図中、数字の1及び2は、それぞれ1個及び2個の点を表わす。

0－1型のデータ(X, Y)の系列よりρを推定する方法としては、4分相関係数r_{tet}(*cf.* M. Kendall & A. Stuart, 1979)によるもの、(3)式をもとにしたものからNewton-Raphsonの繰り返し法により求めるもの(Kirk, 1973)などがあるが、ここでは、(3)式、(4)式及び(5)式が、それぞれρ、c_y及びc_xに関して単調関数であるので、これらの式から直接、根を求めることにした。(4)式と(5)式によりc_yとc_xを求め、これらの値をもとにして(3)式からρを求めたが、こうして求められた推定値をここではr_{0-1}と表わすことにする。点(r, r_{0-1})をプロットしたものが図－2である。図－1の対角線$\phi=r$の下側に集まっている点(r, ϕ)の分布に比べて、図－2の点(r, r_{0-1})の分布は対角線$r_{0-1}=r$を中心としていることが分かる。



図－2　点(r, r_{0-1})のプロット
図中、数字の1、2、3はそれぞれ1個、2個、3個を表わす。

以上、表－1から図－1及び図－2までをみれば、変数の相関を推定するものとしてファイ係数を用いるときには注意が必要であることが改めて強調されねばならない。ファイ係数ϕには、標本数Nが極端に小さくないときには、$x^2=N\phi^2$が近似的に自由度1のカイ2乗分布に従う(*cf.*芝祐順、*1975*)という性質があるため、統計的検定を行いたいときには都合がよいようである。しかし、相関係数は、単に0でないというだけではその情報量は少

ないと言わなければならない。一方の変数で他方の変数の変動がどの程度説明できるのかということをおさえておかないと、それらの関係が問題とすべき程のものなのかどうかが判断できないが、節1節で指摘したように、ϕ の値だけでは ρ が一意に決定できないという事情を考慮すれば、この無相関の仮説検定だけでは無意味な場合が多いと思われる。パーソナルコンピュータの高性能化と低価格化の進んでいる現状では、できるだけ計算機の利用により背後に働いていると考えられる連続型の変数 x と y の相関係数 ρ を求めるようにするべきであると思われる。

引　用　文　献

柏木繁男　*1974*　0－1型データの処理　続　有恒・八木　冕（監修）　東　洋（編）　心理学研究法　15　データ解析　II　東京大学出版会　pp. 193-231.

Kendall, M. & Stuart, A. 1979 The advanced theory of statistics. Voll. 2. 4th ed. Charles Griffin & Company Limited.

Kirk, D. B. 1973 On the numerical approximation of the bivariate narmal (tetrachoric) correlation coefficient. Psychometrika, 1973, 38, 259-268.

芝　祐順　１９７５　相関関係の分析　続　有恒・八木　冕（監修）　東　洋（編）　心理学研究法　14　データ解析 I　東京大学出版会　pp.81-105.