On the number of alternatives and correlation.

選択肢数と相関係数について[1)]

岡　本　安　晴

1　はじめに

質問紙などにおいては、一般にデータは離散的である。これは、解答者に要求される反応型が言語的であるという利点がある。例えば、お菓子の甘さの評定のときに、100点満点で評定する場合と、甘い—甘くないの2件法で評定する場合とを比べてみればよい。2件法の場合、データの分析において2項目間の相関を求めるのにファイ係数が用いられることがある。背後に想定される要因が連続型のときに、それらの連続型の要因間の相間係数の推定値として用いられたファイ係数には大きな偏りの生じることがある（cf. 岡本（1987））。質問紙等における選択肢数については、2件法の他、「甘い—どちらでもない—辛い」の3件法など種々の多件法がある。選択肢数が多い場合には、等間隔に数値を割り当てて連続型のデータとして処理されることがあるが、本稿では、この場合における算出された相関係数の偏りについての検討を試みる。

2　モデル

2変数 x と y が相関係数 ρ の2次元正規分布

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

に従い、かつ、$(k+1)$ 個の値 $C_{x,i}(i=0,1,\cdots, k)$ により x の値域が

$$C_{x,i-1}<x\leqq C_{x,i}$$

のようにカテゴリーに分けられているとする。但し、$C_{x,0}=-\infty$、$C_{x,k}=+\infty$。

確率変数 X を次のようにとる。

$$X=i \quad if \ C_{x,i-1}<x\leqq C_{x,i}$$

y についても同様に

1）計算はすべてパーソナルコンピュータ PC9801vm2（NEC）で行い、XY—プロッタ DXY—980（Roland DG）により作図した。プログラムは Turbo Pascal（Borland）で書かれている。

$$-\infty = C_{y,0} < C_{y,1} < \cdots < C_{y,k} = +\infty$$

なる $(k+1)$ 個の値 $C_{y,i}$ により k 個のカテゴリーに分け、そのカテゴリーの値をとる確率変数を Y とおく。

このとき、N 個のデータ (X,Y) に対して相関係数 r_{XY} は次式により算出される。

$$r_{XY} = \frac{\frac{1}{N}\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{N}\Sigma(X-\bar{X})}\sqrt{\frac{1}{N}\Sigma(Y-\bar{Y})}}$$

但し、

$$\bar{X}=\frac{1}{N}\Sigma X, \quad \bar{Y}=\frac{1}{N}\Sigma Y$$

$N\to\infty$ すると、

$$\lim_{N\to\infty} r_{XY} = \frac{(\sum_{i,j} ijP_{xy,ij}) - (\sum_i iP_{x,i})(\sum_j jP_{y,j})}{\sqrt{(\sum_i i^2P_{x,i}) - (\sum_i iP_{x,i})^2}\sqrt{(\sum_j j^2P_{y,i}) - (\sum_j jP_{y,j})^2}}$$

ここに、

$$P_{x,i} = \lim_{N\to\infty} \quad (X=i \text{ であるデータの比率})$$

$$= \int_{C_{x,i-1}}^{C_{x,i}} \int_{-\infty}^{\infty} N(x,y)\mathrm{dydx}$$

$$P_{y,j} = \lim_{N\to\infty} \quad (Y=j \text{ であるデータの比率})$$

$$= \int_{C_{y,j-1}}^{C_{y,j}} \int_{-\infty}^{\infty} N(x,y)dxdy$$

$$P_{xy,ij} = \lim_{N\to\infty} \quad (X=i \text{ かつ } Y=j \text{ であるデータの比率})$$

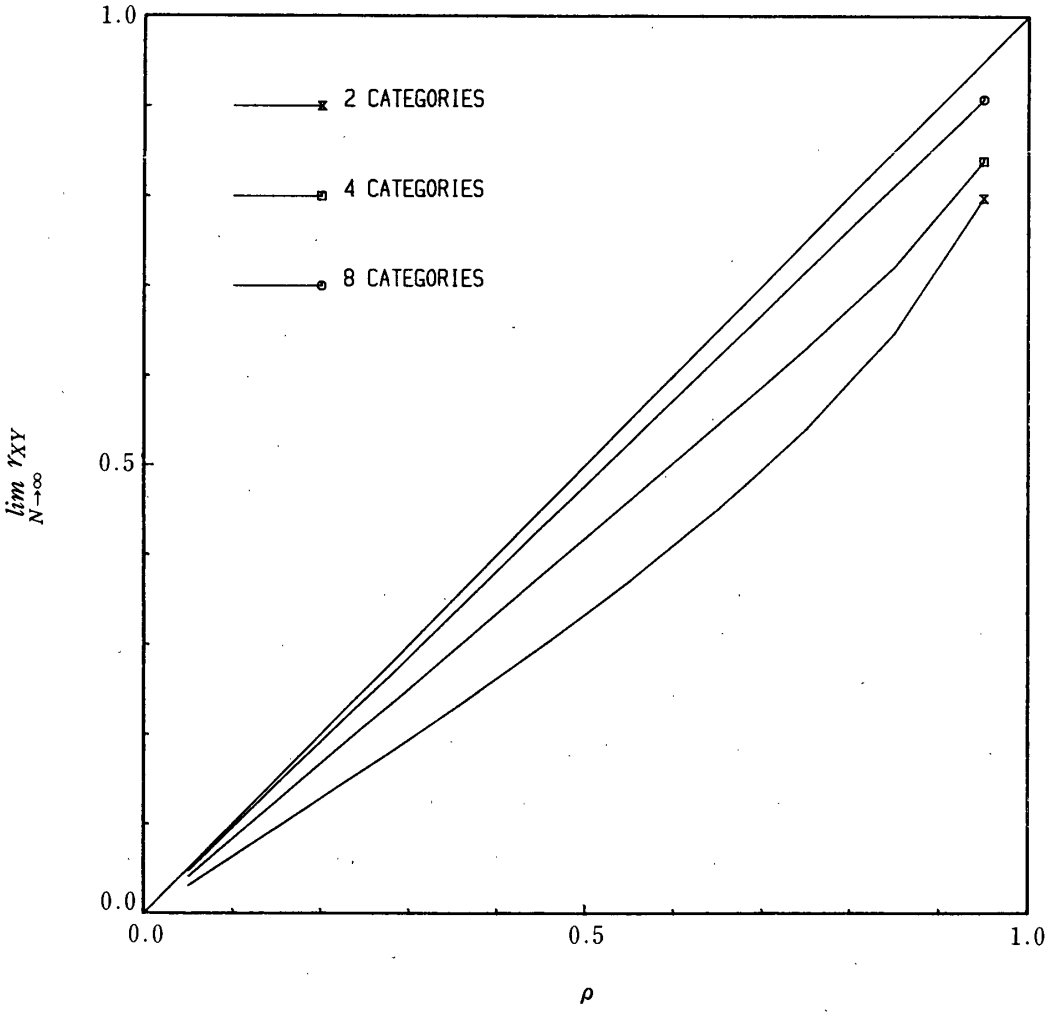
$$= \int_{C_{x,i-1}}^{C_{x,i}} \int_{C_{y,j-1}}^{C_{y,j}} N(x,y)dxdy$$

3 結　　果

$k=2$、4、8の場合の $\lim_{N\to\infty} r_{XY}$ をρの関数として描いたものが図－1及び2である。$C_{x,i}$, $C_{y,j}$ として表－1に示す値を用いたものが図－1であり、表－2の値に対応するものが図－2である。標準正規分布における $C_{x,i}$, $C_{y,j}$ の位置については図－3を参考にして頂きたい。

図－1（表－1）は、データが中央のカテゴリーを中心にして対称に分布している場合であるが、選択肢数が8の場合は $\lim_{N\to\infty} r_{XY}$ はρに十分近いと思われる。選択肢数が2であっても（ファイ係数の場合）ρとの関係はほぼ線型であるとみなせるが、因子分析等における相関係数の推定値として用いた場合にはcommunalityの低下が予想される。図－2は表－2に示されるようなデータのカテゴリー上の分布に偏りがある場合である。選択肢数が2の場合にはρが1に近づくと $\lim_{N\to\infty} r_{XY}$ はほぼ一定となっている。又、ρの値が0から1の範囲の全域において、$\lim_{N\to\infty} r_{XY}$ の値はρに比べて相対的に随分と小さい。即ち、岡本(1987)においても示されたように、このような場合にファイ係数を用いることは極めて危険である。選択肢数が4以上のときは、ρとの関係はほぼ線型でありρとともに増加している。これは、表－2のような偏りの場合は、選択肢数が4以上のときデータの分布が2つ以上のカテゴリーに散らばっているため図－1の選択肢数が2の場合程度の性質の良さがみられたものと思われる。

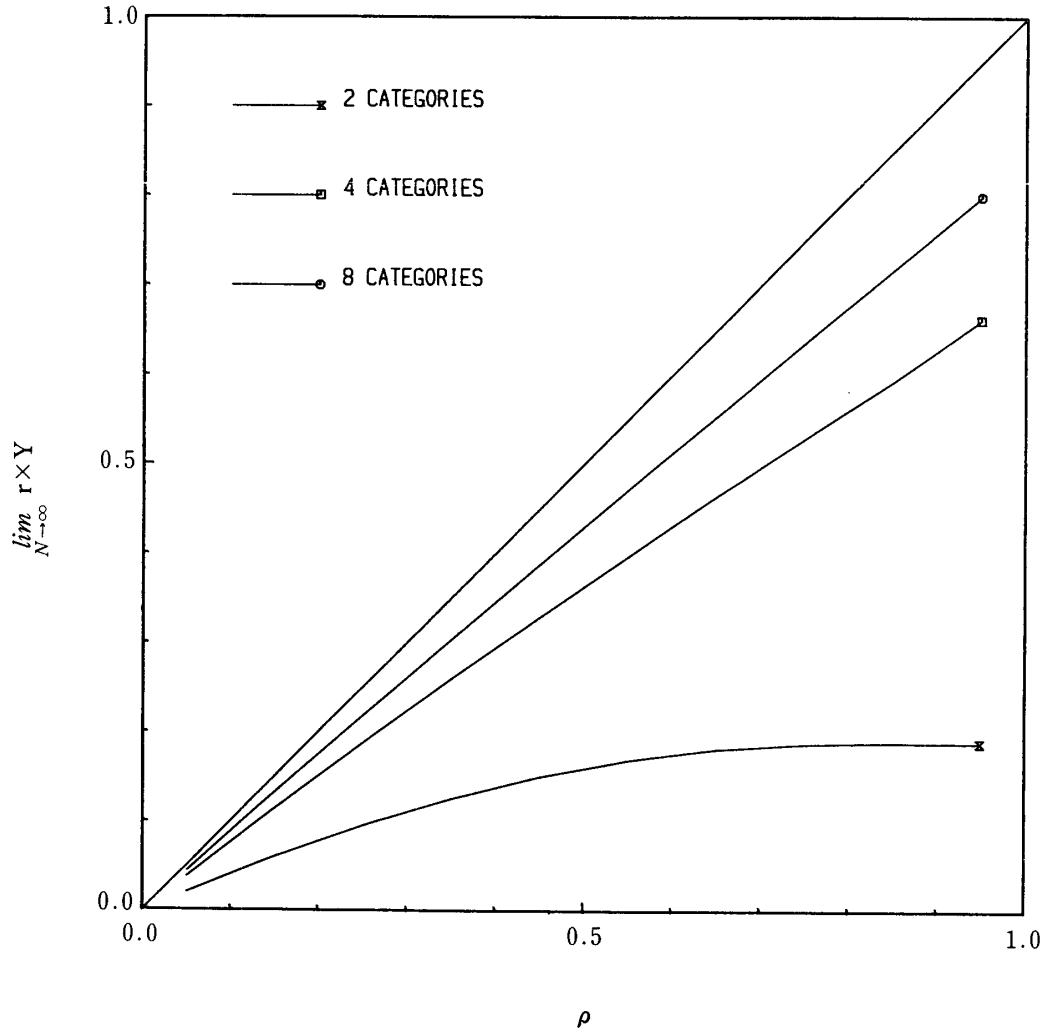
以上、図－1と2の結果だけから判断すれば、選択肢数としては8程度あればカテゴリー値から直接算出された r_{XY} を相関係数の推定値として用いてもよいように思われるが、データのカテゴリー上の分布には注意する必要がある。少なくとも2つ以上のカテゴリーにデータが分布しているかどうか調べて、もしデータの大部分が1つのカテゴリーに集中している場合にはカテゴリー値から直接 r_{XY} を算出して相関係数の推定値とすることは避けねばならない。



図－1　(1)式におけるρの関数としての $\lim_{N\to\infty} r_{XY}$ のグラフ。各選択肢数 k ($k=2, 4, 8$) に対して $C_{\alpha,i}$ ($\alpha = x, y;\ i=1,\cdots, k-1$) は表－1に示された値が設定されている。

		$C_{\alpha,1}$	$C_{\alpha,2}$	$C_{\alpha,3}$	$C_{\alpha,4}$	$C_{\alpha,5}$	$C_{\alpha,6}$	$C_{\alpha,7}$
$k=2$	$\alpha=x$	0.0						
	$\alpha=y$	0.0						
$k=4$	$\alpha=x$	−1.5	0.0	1.5				
	$\alpha=y$	−1.5	0.0	1.5				
$k=8$	$\alpha=x$	−2.25	−1.5	−0.75	0.0	0.75	1.5	2.25
	$\alpha=y$	−2.25	−1.5	−0.75	0.0	0.75	1.5	2.25

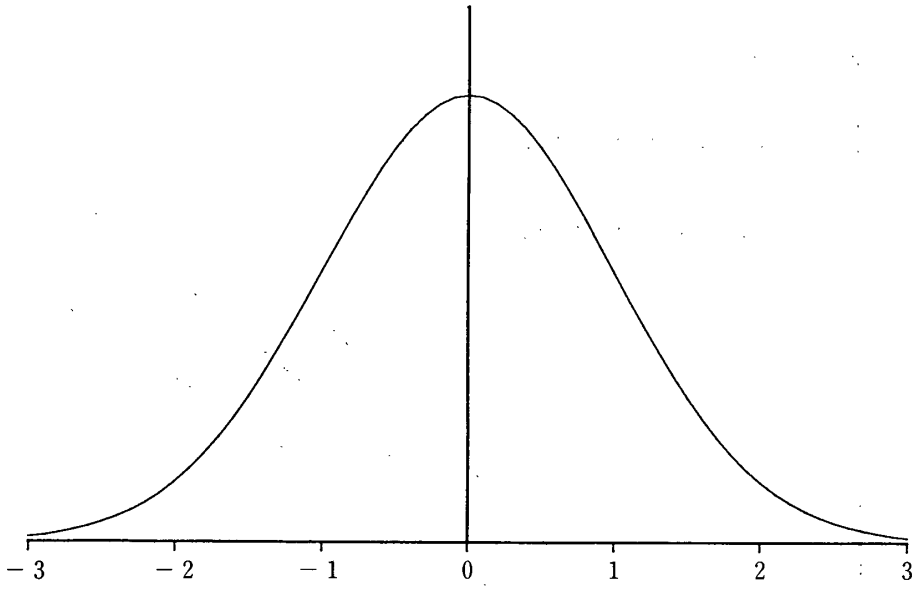
表－1　カテゴリーの分割点 $C_{\alpha,i}(\alpha=x,yji=1,\cdots,k-1)$ の値。分布が中央のカテゴリーを中心として対称になるように値が設定されている。



図－2　(1)式におけるρの関数としての $\lim_{N\to\infty} r_{XY}$ のグラフ。各選択肢数 k ($k=2, 4, 8$) に対して $C_{\alpha,i}$ ($\alpha=x,y;\ i=1,\cdots,k-1$) は表－2に示された値が設定されている。

		$C_{\alpha,1}$	$C_{\alpha,2}$	$C_{\alpha,3}$	$C_{\alpha,4}$	$C_{\alpha,5}$	$C_{\alpha,6}$	$C_{\alpha,7}$
$k=2$	$\alpha=x$	1.0						
	$\alpha=y$	−1.0						
$k=4$	$\alpha=x$	0.0	1.0	2.0				
	$\alpha=y$	−2.0	−1.0	0.0				
$k=8$	$\alpha=x$	−0.5	0.0	0.5	1.0	1.5	2.0	2.5
	$\alpha=y$	−2.5	−2.0	−1.5	−1.0	−0.5	0.0	0.5

表－2　カテゴリーの分割点 $C_{\alpha,i}(\alpha=x,y;\ i=1,\cdots,k-1)$ の値。カテゴリ上の分布に偏りが生じるように値が設定されている。



図－3　標準正規分布のグラフ。

引　用　文　献

岡本安晴　1987　ファイ係数について　金沢大学文学部論集行動科学篇　第7号