n元度数分布の数量化・補遺

京　極　純　一

(1970年5月　受付)

On a Method of Quantifying N-Way Frequency Tables (2)

Jun-ichi Kyogoku

I. We may show the original data for a n-way frequency table with a $N\times S$ matrix $\boldsymbol{B}$, all elements of which are either zero or unity. N is the number of sample individuals and S is the number of all categories summed up through n items. Then $S\times S$ matrix $\boldsymbol{B}'\boldsymbol{B}$ has n sets of marginal frequencies as its (block) diagonal elements and also $\binom{n}{2}$ two way contingency tables as its off diagonal block elements. We may take matrix $\boldsymbol{B}'\boldsymbol{B}$ as a sum of two matrices. If we write the (block) diagonal matrix made up by n sets of marginal frequencies as $\boldsymbol{F}$ and the residual matrix made up by $\binom{n}{2}$ contingency tables as $\boldsymbol{A}$, then $\boldsymbol{B}'\boldsymbol{B}=\boldsymbol{F}+\boldsymbol{A}$.

The quantification problem is to choose an "optimum" scores $\boldsymbol{x}$ ($S\times 1$ column vector) to be allotted to S categories in order to maximize a specified criterion number. The quantification theory Type III, developed by Dr. C. Hayashi originally for two way contingency tables, specifies the following operation;

$$\rho^2=\frac{\boldsymbol{x}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}}=\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}}\longrightarrow \max.$$

The method of quantifying n-way frequency tables, introduced by J. Kyogoku [1], takes up the following operation;

$$\gamma=\frac{\boldsymbol{x}'\boldsymbol{A}\,\boldsymbol{x}}{(n-1)\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}}\longrightarrow \max.$$

So, clearly,

$$\rho^2=\frac{(n-1)\lambda+1}{n},$$

and the maxima of both ρ^2 and λ are given by the same solution vector $\boldsymbol{x}$ simultaneously. Thus, both operations of quantification are equivalent.

II. We may represent n items by n vectors of length unity,and let all cosines of $\binom{n}{2}$ angles between any two vectors be equal to λ. Let the content of simplex P joining the origin and n end points be V_P, the content of simplex Q joining n end points be V_Q, and the height of P from the origin to the base Q be h, then

$$V_P{}^2=\frac{\{(n-1)\lambda+1\}(1-\lambda)^{n-1}}{(n!)^2}=\frac{n^{n-2}\rho^2(1-\rho^2)^{n-1}}{(n-1)^{n-1}\{(n-1)!\}^2},$$

$$V_Q{}^2=\frac{n(1-\lambda)^{n-1}}{\{(n-1)!\}^2}=\frac{n^n(1-\rho^2)^{n-1}}{(n-1)^{n-1}\{(n-1)!\}^2},$$

$$h^2=\frac{n^2\,V_P{}^2}{V_Q{}^2}\doteqdot\rho^2=\frac{(n-1)\lambda+1}{n}.$$

Clearly V_Q is the monotonic decreasing function of ρ^2, and of λ. But, V_P behaves differently. V_P has its maximal value at $\rho=0$, and at $\rho^2=1/n$. Since the basic idea of quantification approach has the minimization of V_P as its goal, we had better set the practical lower limit of λ at zero, the implication of which is that the practical lower limit of ρ^2 is to be $1/n$. (Incidentally, $1/n$ is also the practical lower limit in the component analysis.)

Tokyo University

[1] J. Kyogoku, On a Method of Quantifying N-way Frequency Tables. The Proceedings of the Institute of Statistical Mathematics, Vol. 15, No. 2, 140–160.

これは前稿「n 元度数分布の数量化について」[1] に対する補遺である．まず，I において，(イ) 前稿で紹介した，n 元度数分布を数量化する方法が，林知己夫氏による数量化理論第 III 類——のある場合——と同等であることを説明する．つぎに，II において，上の結果，(ロ) 数量化理論第 III 類における相関係数ないし相関比に実用上の下限を考えるべきであることを説明する．

I

1·1) 数量化の対象となるすべての個体が，n 元の各元ごとに，その元を構成するいくつかの項目のどれか一つに必ず該当し，n 元全部について通計 n 回該当している；——この場合に関する数量化理論第 III 類と，この場合を前提する，前稿で紹介した，n 元度数分布を数量化する方法が同等であることを．以下において，説明する．

1·2) この場合，資料としての原表のうち，全部で n 元あるなかの，たとえば，第 j 元に関する部分は表1のような形である．表のなかの b_{ijl} などは，対象個体ごとに，該当または非該当のどちらかを示す記号である．これらの記号を，該当を示す場合に 1，非該当を示す場合に 0，という数値でおきかえるのが，データ処理の慣例であり，それによって右端縦列と下端横行の計が与えられる。本稿の場合，右端縦列の計は，すべて，1 である．

表 1

項目 / 対象個体	第 j 元					計
	$j1$	··	jl	··	js	
1	b_{1j1}	··	b_{1jl}	··	b_{1js}	1
⋮	⋮		⋮		⋮	⋮
i	b_{ij1}	··	b_{ijl}	··	b_{ijs}	1
⋮	⋮		⋮		⋮	⋮
N	b_{Nj1}	··	b_{Njl}	··	b_{Njs}	1
計	f_{j1}	··	f_{jl}	··	f_{js}	N

1·3) 表1から右端縦列と下端横行の二つの計を除いた残りの，0 と 1 だけを要素とする $N\times S_j$ の行列を $\boldsymbol{B}_j$ とする．(N は対象個体の総数，S_j は第 j 元のなかの項目の数である。) n 元全部についての資料は，$\boldsymbol{B}_j$ のような行列を n 箇横に並列した $N\times S$ の行列 $\boldsymbol{B}$ で与えられる．(S は n 元全部についての項目数の総数である．$S=\sum_j S_j$)

$$\boldsymbol{B}=[\boldsymbol{B}_1\cdot\cdot\boldsymbol{B}_j\cdot\cdot\boldsymbol{B}_k\cdot\cdot\boldsymbol{B}_n] \tag{1·3·1}$$

元と元との間の二元クロス表を示すものは $S\times S$ の行列 $\boldsymbol{B}'\boldsymbol{B}$ である．前稿と同じ表記法を使うために，一般に，

$$\left.\begin{aligned}&\boldsymbol{B}_j'\boldsymbol{B}_k=\boldsymbol{A}_{jk},\quad \boldsymbol{B}_k'\boldsymbol{B}_j=\boldsymbol{A}_{kj}=\boldsymbol{A}'_{jk}\\&\boldsymbol{B}_j'\boldsymbol{B}_j=\boldsymbol{F}_j\end{aligned}\right\} \tag{1·3·2}$$

ときめると，$\boldsymbol{B}'\boldsymbol{B}$ は次のようになる．

$$\boldsymbol{B}'\boldsymbol{B}=\begin{pmatrix}\boldsymbol{F}_1 & \cdot\cdot & \boldsymbol{A}_{1j} & \cdot\cdot & \boldsymbol{A}_{k1} & \cdot\cdot & \boldsymbol{A}_{1n}\\ \vdots & & \vdots & & \vdots & & \vdots\\ \boldsymbol{A}_{j1} & \cdot\cdot & \boldsymbol{F}_j & \cdot\cdot & \boldsymbol{A}_{jk} & \cdot\cdot & \boldsymbol{A}_{jn}\\ \vdots & & \vdots & & \vdots & & \vdots\\ \boldsymbol{A}_{k1} & \cdot\cdot & \boldsymbol{A}_{kj} & \cdot\cdot & \boldsymbol{F}_k & \cdot\cdot & \boldsymbol{A}_{kn}\\ \vdots & & \vdots & & \vdots & & \vdots\\ \boldsymbol{A}_{n1} & \cdot\cdot & \boldsymbol{A}_{nj} & \cdot\cdot & \boldsymbol{A}_{nk} & \cdot\cdot & \boldsymbol{F}_n\end{pmatrix} \tag{1·3·3}$$

なお，一般に，$\boldsymbol{F}_j$ を対角要素行列とする対角行列 $\boldsymbol{F}$，および，$\boldsymbol{B}'\boldsymbol{B}$ の対角要素行列 $\boldsymbol{F}_j$ などをすべてゼロ行列でおきかえた行列 $\boldsymbol{A}$ を定義すると，当然に，次の関係が成立つ．

$$\boldsymbol{B}'\boldsymbol{B}=\boldsymbol{F}+\boldsymbol{A} \tag{1·3·4}$$

さて，数量化によって，第 j 元の S_j 箇の項目に与えられるべき数値を要素とする $S_j\times 1$ の列ベクトルを $\boldsymbol{x}_j$ とする。この $\boldsymbol{x}_j$ のような列ベクトルを縦に n 箇つないだ $S\times 1$ の列ベクトルを $\boldsymbol{x}$ とする．

$$\boldsymbol{x}=\begin{pmatrix}\boldsymbol{x}_1\\ \vdots\\ \boldsymbol{x}_j\\ \vdots\\ \boldsymbol{x}_k\\ \vdots\\ \boldsymbol{x}_n\end{pmatrix} \tag{1·3·5}$$

（なお，慣例により，対角行列 $\boldsymbol{F}$ の要素をウェイトとして，$\boldsymbol{x}$ 全体の平均がゼロであるのみならず，さらに，各元に対応する $\boldsymbol{x}_j$ などの部分ベクトルについても，平均がゼロであるものとする．あとの条件を追加しても，数量化理論第 III 類の相関係数ないし相関比に影響しない．

1·4)　数量化理論第 III 類においては，対象個体に与えられるべき数値を要素とする $N\times 1$ の列ベクトルを $\boldsymbol{y}$ として，上の $\boldsymbol{x}$ とこの $\boldsymbol{y}$ との間に，行列 $\boldsymbol{B}$ をウエイトに，成立つ相関係数 ρ を指標として採用し，次の式を解くことによって，数量化する．

$$n\rho^2=\frac{\boldsymbol{x}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}}=\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}}\longrightarrow \max. \tag{1·4·1}$$

これに対して，前稿で紹介した，n 元度数分布を数量化する方法は，n 元を示す n 本のベクトルの n 個の終点の集中度を手掛りに，λ という指標を採用し，次の式を解くことによって，数量化する．

$$\lambda=\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{(n-1)\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x}}\longrightarrow \max. \tag{1·4·2}$$

(1·4·1) と (1·4·2) から，明らかに，

$$\rho^2=\frac{(n-1)\lambda+1}{n} \tag{1·4·3}$$

であり，ρ^2 が最大のときと λ が最大のときが一致する．しかも，同一の列ベクトル $\boldsymbol{x}$ が，同時に，ρ^2 の最大値と λ の最大値を与える．そして，この関係は，次々に固有値と固有ベクトルを算出する，いわゆる多次元数量化においても，同様に，成立つ．すなわち，本稿の場合における数量化理論第 III 類と n 元度数分布を数量化する方法とは同等である．

なお，n 本のベクトルの n 箇の終点が，その重心のまわりでもつ距離の平方和 D は，ρ^2 を使って書きかえると，次のようになる．

$$D=(1-\rho^2)\cdot\boldsymbol{x}'\boldsymbol{F}\boldsymbol{x} \tag{1·4·4}$$

ρ を大きくすることがもつ，幾何学的なひとつの意味が示されている．

II

2·1)　(1·4·2) を解くと，たとえば

$$\sum_{\substack{\alpha\\(\alpha\neq j)}}\boldsymbol{A}_{j\alpha}\boldsymbol{x}_\alpha=(n-1)\lambda\boldsymbol{F}_j\boldsymbol{x}_j \tag{2·1·1}$$

のような形の式が n 箇連立する．各元ごとに $\boldsymbol{x}_j$ の分散が 1 となるように寸法をきめても，λ は影響をうけないので，解の $\boldsymbol{x}_j$ の寸法をすべて，1 にそろえる．(2·1·1) の両辺に左から $\boldsymbol{x}'_j$ をかけ，両辺を N で割ると，一般に，第 j 元と第 r 元との間の相関係数を $r_{j\alpha}$ として，

$$\sum_{\alpha} r_{j\alpha} = (n-1)\lambda \qquad (2\cdot1\cdot2)$$
$$(j \neq \alpha)$$

となる. (2·1·2) を j について辺々加え, $n(n-1)$ で両辺を割ると, (1·4·2) ないし (2·1·1) を解いてえられる $\boldsymbol{x}$ ベクトルによって元と元との間に成立つすべての相関係数の平均を $\bar{r}$ として,

$$\bar{r} = \lambda \qquad (2\cdot1\cdot3)$$

となる.

ここで, 2元の場合には相関係数そのもの, 3元の場合には3通りの相関係数すべてが λ に等しい. しかし, 4元以上の場合には, 相関係数の箇数の方が, (2·1·2) のような式の箇数より多くなり, λ から直接に相関係数を知ることはできない.

2·2) さて, 一般に,

$$1 \geqq \rho^2 \geqq 0 \qquad (2\cdot2\cdot1)$$

であるから, (1·4·3) によって,

$$1 \geqq \lambda \geqq -\frac{1}{n-1} \qquad (2\cdot2\cdot2)$$

となり, λ の上限と下限が与えられる. ρ と λ のこれらの形式上の下限が実用上いかなる意味をもつか, 以下において, とりあげる.

n 元に対応する n 本のベクトルの長さが, すべて, 1, n 本のベクトルの間に成立つ $\binom{n}{2}$ 通りの角度の余弦が, すべて, λ であるとする. 原点と n 箇の終点からなる単体 P の体積を V_P, 頂点である原点からみて単体 P の底をなす, n 箇の終点だけからなる単体 Q の体積を V_Q, 原点から底 Q にいたる高さを h であらわすと,

$$\left.\begin{aligned} V_P^2 &= \frac{\{(n-1)\lambda+1\}(1-\lambda)^{n-1}}{(n!)^2} = \frac{n^{n-2}\rho^2(1-\rho^2)^{n-1}}{(n-1)^{n-1}\{(n-1)!\}^2} \\ V_Q^2 &= \frac{n(1-\lambda)^{n-1}}{\{(n-1)!\}^2} = \frac{n^n(1-\rho^2)^{n-1}}{(n-1)^{n-1}\{(n-1)!\}^2} \\ h^2 &= \frac{n^2 V_P^2}{V_Q^2} = \rho^2 = \frac{(n-1)\lambda+1}{n} \end{aligned}\right\} \qquad (2\cdot2\cdot3)$$

となる. [2] すなわち, 本稿の場合の数量化理論第 III 類における相関係数ないし相関比は, 単体 P において, 頂点である原点から底をなす単体 Q にいたる高さを示している. また, 本稿の場合の数量化理論第 III 類も, 前稿を紹介した, n 元度数分布を数量化する方法も, ともに, 単体 Q の体積 V_Q を最小にすることと同等であることが, (2·2·3) から明らかである.

2·3) 数量化理論の出発点には, 前稿でふれたように, 単体 P の体積 V_P を最小にする, という基本的着想がある. そこで, V_P の変化と V_Q の変化との関係をみるために,

$$\left.\begin{aligned} \alpha &= 1 - \frac{V_P}{\max.(V_P)} \\ \beta &= 1 - \frac{V_Q}{\max.(V_Q)} \end{aligned}\right\} \qquad (2\cdot3\cdot1)$$

という二つの指標を作ってみる．当然に，$\beta, h(=\rho)$，$\lambda(=\bar{r})$ の三つのうち，二つは，つねに，残りの一つの単調増加函数となり，区間の上限で，それぞれ，最大限を，区間の下限で，それぞれ，最小値をとる．

これに対して，α の変化は β の変化と一致しない．すなわち，表2のように，$\lambda=0$，ないし，$\rho^2=1/n$，のところで，V_P は，つねに，最大値をとる．（幾何学的には，ほとんど自明のことである。）なお，図1は，$n=5$ の場合について，$\alpha, \beta, h(=\rho)$ および $\lambda(=\bar{r})$ の変化を例示したものである．

表 2

	α	V_p
$\begin{cases}\lambda=1\\ \rho^2=1\end{cases}$	1 (max.)	0 (min.)
$\begin{cases}\lambda=0\\ \rho^2=1/n\end{cases}$	0 (min.)	$1/n!$ (max.)
$\begin{cases}\lambda=-1/(n-1)\\ \rho^2=0\end{cases}$	1 (max.)	0 (min.)

従って，単体 P の体積 V_p を最小にする，という数量化理論の基本的着想に立ちかえるならば，n 元度数分布を数量化する方法においては，

$$\lambda \geqq 0 \qquad (2\cdot3\cdot2)$$

という下限を λ に与え，実用上は λ を正の場合に限ることが適当となる。同様にして，本稿の場合の数量化理論第 III 類においては，

$$\rho^2 \geqq \frac{1}{n} \qquad (2\cdot3\cdot3)$$

という下限を ρ^2 に与え，実用上は，ρ^2 を $1/n$ より大きい場合に限ることが適当となる．（これは，よく知られているように，成分分析における実用上の下限でもある．）ρ も λ もいわゆる多次元数量化を許すものであるから，この形で，実用上の下限を，あらかじめ，定めうることは，数値計算の手順の上で，便利なことである．

（$n=5$ の例）

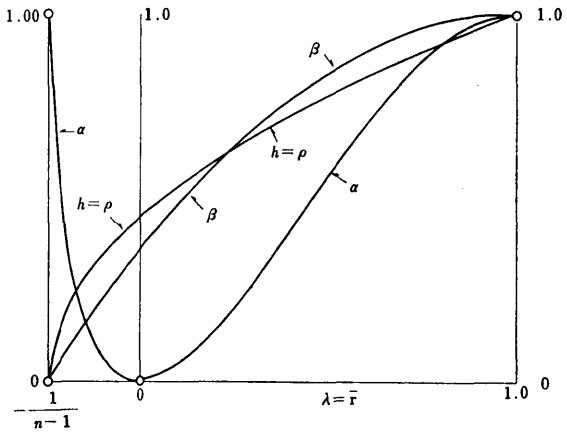

図 1

東京大学教養学部

参　考　文　献

[1] 京極純一：n 元度数分布の数量化について，統計数理研究所彙報，第15巻，第2号．
[2] M. G. Kendall: A Course in the Geometry of n Dimensions. C. Griffin & Co., London, 1961.