数量化に於ける標本誤差

青山博次郎

(1954 年 10 月受付)

On Sampling Errors in Certain Problems of Quantification.

By Hirojiro AOYAMA

In this paper we treat the sampling errors in certain problems of quantification. The exact formulas of these errors are so complicated that we treat only the upper bounds of the absolute sampling errors with certain reliability. We have found that the order of these upper bounds are $1/\sqrt{n}$ as that of the ordinary sampling error of the estimated mean.

Institute of Statistical Mathematics

§1 緒　　言

数量化の種々の方法がこれまで取扱われているが，それは点推定としての数量化であり，得られた標本集団を恰も母集団の如く考えて最適数量の導出を行つているものとも考えられるのである．従つてサンプルの大きさが小さい時に，このような数量化を行うと多大の標本誤差を有するため，予測性は低いものといわねばならぬ．本稿では屢々用いられている 2，3 の数量化の方法に就て，標本誤差を考慮したときの影響を考えてみることにする．

§2 外部基準のある場合の数量化 I

外部基準 y が存しており，m 個の item $A, B, \cdots C$ (各々は $s_A, s_B, \cdots s_C$ 個のカテゴリーをもつ) を数量化するのに，各 item 毎に相関係数を最大にする如く数量化し，その上で回帰平面で y の推定を行う場合を考えよう．

例えば〔1〕に於て取扱つた学校の推定点を，地域，特性，規模を数量化して求める場合がこれである．

まず次の定理を証明しておこう．

定理 1．k 次の行列式 $F=|f_{ij}|$ の各要素の絶対値の最大値を f_m とするとき，f_{ij} の微小な変化(絶対値は ε より小) に対する F の変化 δF は

$$|\delta F| \leqq k^2(k-1)^{\frac{k-1}{2}} f_m^{k-1}\varepsilon$$

を満足する．

証明．f_{ij} の余因数を F_{ij} とおくとき

$$\delta F = \sum_i \sum_j \frac{\partial F}{\partial f_{ij}}\delta f_{ij} = \sum_i \sum_j F_{ij}\delta f_{ij}$$

Hadamard の定理により

$$|\delta F| \leqq k^2(k-1)^{\frac{k-1}{2}} f_m^{k-1}\varepsilon \qquad \text{(証明終)}$$

いま item A について次表の如き分布が得られているものとしよう．

y \ A	$A_1 \cdots A_i \cdots A_{s_A}$	計
y_1	$f_{A11} \cdots f_{Ai1} \cdots f_{As_A1}$	f_1
⋮	⋮ ⋮ ⋮	⋮
y_k	$f_{A1k} \cdots f_{Aik} \cdots n_{As_Ak}$	f_k
計	$n_{A1} \cdots n_{Ai} \cdots n_{As_A}$	n

A_i には $\hat{y}_i = \frac{1}{n_{Ai}} \sum_{l=1}^{k} f_{Ail} y_l = x_{Ai}$

なる数量を与えるのである．従つて x_{Ai} の分散は近似的に σ_{Ai}^2/n_{Ai} となる．一般に item $A, B, \cdots C$ (j で表わす) に対し

$$\sigma_{Ji}^2 = \frac{1}{N_{Ji}} \sum_{l=1}^{k} F_{Jil}(y_l - \bar{y}_i)^2$$

$$\sigma_j^2 = \frac{1}{N} \sum_{i=1}^{s_j} N_{Ji}(\bar{y}_i - \bar{y})^2 \qquad (j = A, B \cdots, C \text{ 或は } 1, 2, \cdots, m)$$

$$\sigma_0^2 = \frac{1}{N} \sum_{l=1}^{k} F_l(y_l - \bar{y})^2$$

(但し N, F 等は母集団に於て n, f に対応するもの.)

$$\max_{i,j}\left(\frac{\sigma_{Ji}^2}{n_{Ji}}\right) = \frac{\sigma^{*2}}{n^*} \tag{2·1}$$

とおく．平均的には $n^* = nP^*$ (P^* はある item j の i カテゴリーに属するものの比率) と考えてよい．

一方 y の推定値 y' は原点を適当に変換して

$$y' = -\left(\frac{R_{01}\sigma_0}{R_{00}\sigma_1}x_1 + \frac{R_{02}\sigma_0}{R_{00}\sigma_2}x_2 + \cdots + \frac{R_{0m}\sigma_0}{R_{00}\sigma_m}x_m\right) \tag{2·2}$$

とおけるものとする．実際の計算では R_{00}, R_{0j}, σ_0, σ_j, x_j はすべて標本値を用いて推定する．このときの y を y'' としておこう*．上式に於て y と x_j の母集団相関係数を ρ_{0j}, x_j と $x_{j'}$ との母集団相関係数を $\rho_{jj'}$ とし，

$$R = \begin{vmatrix} \rho_{00} & \rho_{01} & \rho_{02} \cdots & \rho_{0m} \\ \rho_{10} & \rho_{11} & \rho_{12} \cdots & \rho_{1m} \\ \vdots & & & \vdots \\ \rho_{m0} & \rho_{m1} & \rho_{m2} \cdots & \rho_{mm} \end{vmatrix}$$

に関する ρ_{0j} の余因数を R_{0j} とおいた．

R_{00}, R_{0j} は共に m 次の行列式であるから，定理1により

$$|\delta R_{00}| \leqq (m-1)^{\frac{m-1}{2}} \cdot m^2 \cdot \max|\delta\rho_{jj'}|$$

item j と j' の相関曲面がガウス型であると仮定すれば，$\rho_{jj'}$ の標本推定値 $r_{jj'}$ の分散は $(1-\rho_{jj'}^2)/n$ であるから（この仮定が成立しないときも $O(1/n)$ であることは同様である）3シグマの誤差を考え，R_{00} の標本絶対誤差は高々

$$|\delta R_{00}| \leqq (m-1)^{\frac{m-1}{2}} \cdot m^2 \cdot \frac{3}{\sqrt{n}} \tag{2·3}$$

なることが「相当高い信頼度」をもつていえる．(以下誤差の不等式を標本誤差の不等式におきかえるとき「」内の言葉を附加すべきであるが，簡単のために省略する)

$|\delta R_{0j}|$ についても (2·3) と同様の式が成立つ．従つて標本絶対誤差は高々次のように評価される．

---

* このとき (2·2) における誤差項 $\pm k s_{0.12\cdots m}\sqrt{1/n + \sum\sum a_{ij}x_i x_j}$ は数量化の性質上考慮しないこととする．それは y' よりむしろ x_j の数量化に主眼をおいているからである．

$$\left|\delta\left(\frac{R_{0j}}{R_{00}}\right)\right|=\left|\frac{R_{0j}}{R_{00}}\left(\frac{\delta R_{00}}{R_{00}}+\frac{\delta R_{0j}}{R_{0j}}\right)\right|\leqq\frac{(m-1)^{\frac{m-1}{2}}|\delta R_{00}|}{{R_{00}}^2}+\frac{|\delta R_{0j}|}{|R_{00}|}$$

$$\leqq\frac{3}{\sqrt{n}}m^2(m-1)^{\frac{m-1}{2}}\left(\frac{(m-1)^{\frac{m-1}{2}}}{{R_{00}}^2}+\frac{1}{|R_{00}|}\right)\qquad(2.4)^{**}$$

また β_2, β_{2j} をそれぞれ y 及び item j の分布の尖度とすると（ガウス型なら 3 とおけばよい）

$$\left|\delta\left(\frac{\sigma_0}{\sigma_j}\right)\right|\leqq\frac{3\sigma_0}{2\sqrt{n}\,\sigma_j}(\sqrt{\beta_2-1}+\sqrt{\beta_{2j}-1})\qquad(2\cdot5)$$

故に (2·2) は (2·1), (2·4), (2·5) を用いて

$$|\delta y''|\leqq\frac{3\sigma_0}{\sqrt{n}}m^2(m-1)^{\frac{m-1}{2}}\left(\frac{(m-1)^{\frac{m-1}{2}}}{{R_{00}}^2}+\frac{1}{|R_{00}|}\right)\sum_j\frac{|X_j|}{\sigma_j}$$

$$+\frac{3\sigma_0\sqrt{\beta_2-1}}{2\sqrt{n}}\sum_j\left|\frac{R_{0j}}{R_{00}}\right|\frac{|X_j|}{\sigma_j}+\frac{3\sigma_0}{2\sqrt{n}}\sum_j\left|\frac{R_{0j}}{R_{00}}\right|\frac{|X_j|}{\sigma_j}\sqrt{\beta_{2j}-1}$$

$$+\frac{3\sigma_0\sigma^*}{\sqrt{n^*}}\sum_j\left|\frac{R_{0j}}{R_{00}}\right|\frac{1}{\sigma_j}\qquad(2\cdot6)$$

これより標本絶対誤差は大体サンプルの大きさ n の平方根に逆比例する一定値をこえないことが分る．

§3 外部基準のある場合の数量化 II

多くの item を同時に外部基準によつて数量化するときは，例えば［2］にある

$$f_\rho^{(i)}z_\rho^{(i)}+\sum_{j\neq i}\sum_u f_{\rho u}^{(ij)}z_u^{(j)}=\sum_s f_{\rho s}^{(i)}y_s\qquad(3\cdot1)$$

を用いる．この式の両辺を $f_\rho^{(i)}$ で割つた式を行列表示して

$$AZ=B\qquad(3\cdot2)$$

としよう．ここで行列 $A=(a_{i\rho})$，ベクトル $Z=(z_1,\cdots\cdots z_k)$，$B=(b_1,\cdots\cdots b_k)$，k はカテゴリーの総数とする．

(3·2) を満足する Z を求めると

$$Z=A^{-1}B\qquad(3\cdot3)$$

先ず次の定理を証明しておく．

定理 2. 定理 1 と同一の条件が成立つとき逆行列 F^{-1} の i 行 k 列の要素についての微小変化は

$$\varepsilon\left\{\frac{(k-1)^2(k-2)^{\frac{k-2}{2}}{f_m}^{k-2}}{|F|}+\frac{|F_{ij}|k^2(k-1)^{k-1}{f_m}^{2k-2}}{F^2}\right\}$$

を越えない．

証明. 定理 1 と同様にして

$$|\delta F_{jk}|\leqq(k-1)^2(k-2)^{\frac{k-2}{2}}{f_m}^{k-2}\varepsilon$$

従つて逆行列の i 行 k 列の要素 F_{jk}/F について

** 分散を考えて $D^2\left(\frac{R_{0j}}{R_{00}}\right)\fallingdotseq\left(\frac{R_{0j}}{R_{00}}\right)^2\left(\frac{\sigma_{R0j}{}^2}{\bar{R}_{0j}{}^2}+\frac{\sigma_{R00}{}^2}{\bar{R}_{00}{}^2}-\frac{2\rho\cdot\sigma_{R0j}\sigma_{R00}}{\bar{R}_{0j}\bar{R}_{00}}\right)$ において $\rho=-1$ とおき，$D\left(\frac{R_{0j}}{R_{00}}\right)$ の最大値を考えたことに当る．本文中では R_{0j} の平均値 $\bar{R}_{0j}$ などを簡単のため R_{0j} と記してある．以下も同様．

$$\left|\delta\left(\frac{F_{jt}}{F}\right)\right|=\left|\frac{F_{jt}}{F}\left(\frac{\delta F_{jt}}{F_{jt}}+\frac{\delta F}{F}\right)\right|\leqq\frac{(k-1)^2(k-2)^{\frac{k-2}{2}}f_m{}^{k-2}\varepsilon}{|F|}$$

$$+\frac{k^2(k-1)^{\frac{k-1}{2}}f_m{}^{k-1}\varepsilon}{F^2}\cdot(k-1)^{\frac{k-1}{2}}f_m{}^{k-1}$$

$$=\varepsilon\left\{\frac{(k-1)^2(k-2)^{\frac{k-2}{2}}f_m{}^{k-2}}{|F|}+\frac{k^2(k-1)^{k-1}f_m{}^{2k-2}}{F^2}\right\}\qquad\text{(証明終)}$$

さて A の各要素は内部比率であるから，その分散は近似的に

$$\frac{F^{(ij)}_{\rho u}}{F^{(j)}_{\rho}}\left(1-\frac{F^{(ij)}_{\rho u}}{F^{(i)}_{\rho}}\right)\Big/n\frac{N^{(i)}_{\rho}}{N}\qquad(3\cdot4)$$

となるから，この最大値を c^2/n とおく．

また B の要素についても，その分散は σ^{*2}/f^* を越えない．但し item i のカテゴリー ρ に属するもののyの分散を $\sigma^2_{(i)\rho}$ とおくとき

$$\max_{i,\rho}\left(\frac{\sigma^2_{(i)\rho}}{f^{(i)}_{\rho}}\right)=\frac{\sigma^{*2}}{f^*}\qquad(3\cdot5)$$

従つて $z^{(i)}_{\rho}$ の標本誤差 E は，カテゴリーの総数を k，各カテゴリー毎の $|\bar{y}|$ の最大値を M とおくと

$$|E|=|\sum_{i=1}^{k}(b_i\delta a_{ii}^{-1}+a_{ii}^{-1}\delta b_i)|$$

$$\leqq k\left\{\frac{3cM}{\sqrt{n}}\left(\frac{(k-1)^2(k-2)^{\frac{k-2}{2}}}{|\det A|}+\frac{k^2(k-1)}{|\det A|^2}\right)+\frac{3\sigma^*(k-1)^{\frac{k-2}{2}}}{\sqrt{f^*}\,|\det A|}\right\}\qquad(3\cdot6)$$

故に終局的に $z=\sum_{i=1}^{m}z^{(i)}_{\rho}$ によつて数量化するときに生ずる標本絶対誤差は高々 (3·6) の右辺の m 倍になる．この値は f^* が平均的にはサンプルの大きさ $\sqrt{n}$ に比例するから，n を大きくすれば概ね標本絶対誤差は $\sqrt{n}$ に逆比例していることが分る．

§4 $HX=\lambda X$ 型の数量化

Guttman の paired comparison による数量化 [3] はこの型のものである．このときの標本絶対誤差は摂動論を用いて評価することができる．

先ず定理を証明なしに掲げておこう．

定理 3. 対称行列 K の固有値 $\alpha_1,\cdots,\alpha_k$ が凡て相異なり，固有ベクトルを $X^{(1)},\cdots,X^{(k)}$ とする．K と微小な行列 J だけ異なる対称行列 H の固有値を $\lambda_1,\cdots,\lambda_k$，固有ベクトルを $Y^{(1)},\cdots,Y^{(k)}$ とすると

$$\lambda_i=\alpha_i+(X^{(i)},\ JX^{(i)})\qquad(4\cdot1)$$

$$Y^{(i)}=X^{(i)}+\sum_{j\neq i}c_jX^{(j)}\qquad(4\cdot2)$$

但し

$$c_j\fallingdotseq\frac{1}{\alpha_i-\alpha_j}(X^{(j)},JX^{(i)})\qquad(4\cdot3)$$

ここに $(X^{(j)},JX^{(i)})$ は内積を示す．

この定理は良く知られており，K が等しい固有値をもつ場合にもほぼ同様な結果が得られている．(省略)

Guttman の例［3］によれば H の要素について

$$H_{jk}=\frac{1}{n(N-1)\binom{N}{2}}\sum_{i=1}^{n}(f_{ij}f_{ik}+g_{ij}g_{ik})$$

n： 判定者数
N： 対象者数

上式は

$$H_{jk}=\frac{2}{Nn}\sum_{i}\left(\frac{f_{ij}f_{ik}}{(N-1)^2}+\frac{g_{ij}g_{ik}}{(N-1)^2}\right)$$

$$0\leqq\frac{f_{ij}f_{ik}}{(N-1)^2}\leqq 1$$

であるから，$f_{ij}f_{ik}/(N-1)^2$ を確率変数と考えるとこの分散は 1/4 を越えない．故に

$$\left|D^2(H_{jk})\right|\leqq\frac{1}{N^2n}$$

従つて H_{jk} の標本絶対誤差は相当高い信頼度を以て $3/\sqrt{n}N$ を越えないと考えてよい．

$HX=\lambda X$ の最大固有値はこのとき 1 であるから，その次に大きい固有値を λ_2 とすると，解ベクトル $X^{(2)}$ の標本絶対誤差の成分は，定理 3 を用いて高々

$$\frac{3}{N\sqrt{n}}\sum_{\substack{i=1\\ i\neq 2}}^{N}\frac{|X_k^{(i)}|}{\lambda_2-\lambda_i}(|X_1^{(2)}|+\cdots+|X_N^{(2)}|)(|X_1^{(i)}|+\cdots+|X_N^{(i)}|)\qquad(4.4)$$

となる．但し $|X_k^{(i)}|$ はベクトル $X^{(i)}$ の k 成分の絶対値を示す．

これより判定者数 n を増せば標本絶対誤差は $\sqrt{n}$ に逆比例して減少することが分る．

例．4 人の審判者が 3 人の候補者を審査するとき，O_1 は $A>B>C$, O_2 は $A>B>C$, O_3 は $A>C>B$, O_4 は $B>A>C$ と判定したとき Guttman の方法によると

$$H=\begin{pmatrix}0.5833 & 0.2500 & 0.1667\\ 0.2500 & 0.5000 & 0.2500\\ 0.1667 & 0.2500 & 0.5833\end{pmatrix}$$

この固有値は 1, 0.4166, 0.2500 であり，固有ベクトルは

$$X^{(1)}=\left\{\frac{1}{\sqrt{3}},\ \frac{1}{\sqrt{3}},\ \frac{1}{\sqrt{3}}\right\}$$

$$X^{(2)}=\left\{-\frac{1}{\sqrt{2}},\ 0,\ \frac{1}{\sqrt{2}}\right\}$$

$$X^{(3)}=\left\{\frac{1}{\sqrt{6}},\ -\frac{2}{\sqrt{6}},\ \frac{1}{\sqrt{6}}\right\}$$

となる．相関比を最大にするような数量化に対しては $X^{(2)}$ が所要の固有ベクトルとなる．このとき $X^{(2)}$ に対する標本絶対誤差は (4.4) により，高々 {4.04, 6.86, 4.04} である．このように大きい誤差が生じるのは，審判者が 4 人（審判者の集団からのランダムサンプルと考える．勿論 4 人の審判者のみの集団をとるならばここの議論は不要である）であるからで，もし 16 人ならば上の半分の誤差となる訳である．尚誤解を生じないために附言しておけば，正確なる標本誤差は勿論これよりずつと小さいが，正確に求めるのは困難だから上限を抑えたのである．

§5 $HX=\lambda FX$ 型の数量化

H, F を k 次の対称行列とし，F^{-1} は存在し，$F^{-1}H$ は正数の固有値をもつものとしよう．数量化の場合ではこの条件は満足されていると考えてよい．

F の要素の絶対値の最大値を f_m，H の要素のそれを h_m とすれば

$$(F^{-1}H)X=\lambda X \tag{5·1}$$

なる固有値問題を解けばよいのであるから，$F=(f_{ij})$，$H=(h_{ij})$ とし，f_{ij}，h_{ij} の微小変化の絶対値は ε_f，ε_h を越えないとすれば，定理2により $F^{-1}H$ の各要素の微小変化は

$$\left|\sum_{i=1}^{k}\left\{h_{il}\delta\left(\frac{F_{ji}}{\det F}\right)+\frac{F_{ji}}{\det F}\delta(h_{il})\right\}\right|\leqq k\varepsilon_f h_m\left\{\frac{(k-1)^2(k-2)^{\frac{k-2}{2}}f_m{}^{k-2}}{|\det F|}\right.$$
$$\left.+\frac{k^2(k-1)^{k-1}f_m{}^{2k-2}}{|\det F|^2}\right\}+k\varepsilon_h\frac{(k-1)^{\frac{k-1}{2}}f_m{}^{k-1}}{|\det F|} \tag{5.2}$$

ここで $\det F$ は行列式 F を示す．

従つて (5·2) の右辺を用いて，定理3により固有ベクトル $X^{(i)}$ の標本誤差を評価することができる．

例．相関比を最大にして組別の効果をあげる場合の数量化［4］では

$$\frac{\partial(\sigma_b^2/\sigma^2)}{\partial x_{uv}}=0$$

を満足する x_{uv} を求めることが問題となる．このとき

$$\sum_{j=1}^{R}\sum_{k=1}^{2}h_{uv}(jk)x_{jk}=\eta^2\sum_{l=1}^{R}\sum_{m=1}^{2}f_{uv}(lm)x_{lm} \tag{5·3}$$
$$u=1,2,\cdots,R;\ v=1,2$$

なる連立方程式の解 x_{jk} を求めるのである．問題の仮定より

$$f_{uv}(lm)=\sum_{i=1}^{n}\delta_i(uv)\delta_i(lm)\leqq n$$

$$h_{uv}(jk)=\sum_{t=1}^{R+1}\frac{g_t(jk)g_t(uv)}{n_t}\leqq\sum_{t=1}^{R+1}n_t=n$$

従つて (5·3) の両辺を n で割つた式を $HX=\eta^2FX$ とおくと，H, F の要素はすべて内部比率であるから前述と同様に一定の正数 c をえらぶと各要素の標本絶対誤差は $3c/\sqrt{n}$ を越えないと考えられる．

$$f_m=h_m=1,\quad \varepsilon_f=\varepsilon_h=3c/\sqrt{n}$$

とおいて (5·2) の右辺の値 (E とおく) を求めると

$$E=\frac{6Rc}{\sqrt{n}}\left\{\frac{(2R-1)^2(2R-2)^{R-1}}{|\det F|}+\frac{4R^2(2R-1)^{2R-1}}{|\det F|^2}+\frac{(2R-1)^{\frac{2R-1}{2}}}{|\det F|}\right\} \tag{5·4}$$

従つて解ベクトル $X^{(2)}$ の標本絶対誤差の成分は高々

$$E\sum_{\substack{i=1\\ i\neq 2}}^{2R}\frac{|X_k^{(i)}|}{\lambda_2-\lambda_i}(|X_1^{(2)}|+\cdots+|X_{2R}^{(2)}|)(|X_1^{(i)}|+\cdots+|X_{2R}^{(i)}|) \tag{5·5}$$

となる．

§6 結　　語

数量化に於ける標本絶対誤差は上述の評価により相当高い信頼度を以て $1/\sqrt{n}$ のオーダーであることを知つた．正確な標本誤差については共分散の項が複雑となり実用にならない．従つて得られた標本集団を母集団とみなして差支えない位サンプルの大きさを大きくしておくことが肝要である．

(統計数理研究所)

参 考 文 献

[1] 青山博次郎: 教育調査に於ける諸問題 I, 統計数理研究所彙報, 第1巻, 第1号, 1953.

[2] 青山博次郎: 数量化の一問題, 統計数理研究所講究録, 第8巻, 第4号, 1952.

[3] L. Guttman: An approach for quantifying paired comparisons and rank order, Annals of Math. Stat. Vol. 17, No. 2, 1946.

[4] C. Hayashi: On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view, Annals of the Institute of Stat. Math. Vol. III, No. 2, 1952.

(附記) 定理1の Hadamard の定理を用いた評価は余りにも大きすぎぎる．もし f_{ij} が確率変数であつて，i,j の如何に関せず同一の分散 σ^2 をもち，かつ互いに独立であると仮定すれば $|\delta F|$ の評価として $\sqrt{max\, D^2(F)}$ を用いることができよう．このとき

$$max D^2(F) = k!\{\sigma^{2k} + k\sigma^{2(k-1)}M^2 + {}_kP_{k-2}\sum_{r=2}^{k-1}\frac{\sigma^{2(k-r)}}{(k-r)!}(M^{2r} - m^{2r})\}$$

但し　M, m は f_{ij} の絶対値の最大値及び最小値，P は順列を示す

となるから，通常の場合，定理1の結果よりずつと良い評価式が得られる．(1955.1.8)