母分散 σ^2 の不偏推定量 U^2 とこれに関連する問題について

船　木　勝　也

統計学や関連学科のマーケッティング・リサーチなどの文献において，有限母集団からの任意標本を論じているにもかかわらず，不偏分散が

$$U^2=\frac{1}{n-1}\sum^{n}(X_i-\bar{X})^2$$

と定義されている書が目につく。ここではその事が，どういう事とのかかわりで，どう問題になってくるのか，またそれと関連した事柄のどこが問題になってくるかを議論する上で，問題点が明確に表われてくる点からみて，素材として大屋祐雪氏の論文をとりあげさせて頂く。

大屋氏は経済統計学会の機関誌『統計学』第12号において「現段階における『統計』的標本調査の論理」を構成された。その「論理構成」の一部で「抽出集計の論理構造」をとりあげられ，抽出集計の論理構造の数理的命題として6個の命題を提示された(上書20—21ページ参照せよ)。この6個の命題は，配列順位という点では，誤りのないものと思われるが，〈命題2〉と〈命題5〉とは数理的に厳密にいうと，その内容が合致しない——結合矛盾する——のではないか，というのが私の疑問である。もしそうだとしても，それは氏の「標本調査の論理」構成に影響を及ぼすほどの大きな矛盾ではない。このノートの「結論」で述べるような内容に〈命題〉を書きかえる方が，論理的に完全ではなかろうか，というだけのことである。

この2つの命題はつぎのごときものである。

〈命題2〉母平均 μ，母分散 σ^2，個体の数Nの有限母集団から，n 個の個体を非復元抽出するときの標本平均 $\bar{X}$ の平均値は母平均 μ に等しく，分散は母分散 σ^2の $\frac{N-n}{N-1}\cdot\frac{1}{n}$ 倍に等しい。すなわち，

$$E(\bar{X})=\mu \qquad\qquad \sigma_{\bar{X}}^2=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$$

である。

〈命題5〉母分散 σ^2の母集団からの大きさ n の標本の不偏分散 $U^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ の期待値は母分散に等しい。すなわち，関係式

$$E(U^2)=\sigma^2$$

が成立する。

またこの2つの命題は，周知のことであるが，命題2の $\sigma_{\bar{X}}^2$ という達成精度を計算する過程において通常は σ^2が未知であるから，抽出された n 個の個体からなる標本から計算しえる不偏分散U^2を，n が大である（実験，経験上から少くとも $n\geqq 50$，できれば $n\geqq 100$である）ならば，σ^2の代りに用いてさしつかえない，すなわち $\sigma_{\bar{X}}^2$ を

$$\sigma_{\bar{X}}^2=\frac{N-n}{N-1}\cdot\frac{U^2}{n}$$

として計算しえるという数理的，実験的論理により結合されている。

ところで〈命題5〉の証明はつぎのようになされる。

証明

$$U^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

の分子はつぎのように展開できる。

$$\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}\{(X_i-\mu)-(\bar{X}-\mu)\}^2$$

$$=\sum^{n}(X_i-\mu)^2-2n(\bar{X}-\mu)^2+n(\bar{X}-\mu)^2$$

$$=\sum^{n}(X_i-\mu)^2-n(\bar{X}-\mu)^2$$

$$\therefore E\{\sum^{n}(X_i-\bar{X})^2\}=E\{\sum^{n}(X_i-\mu)^2\}-nE\{(\bar{X}-\mu)^2\}$$

$$=n\sigma^2-n\left(\frac{\sigma^2}{n}\right)=(n-1)\sigma^2$$

$$\therefore E(U^2)=E\left\{\frac{\sum^{n}(X_i-\bar{X})^2}{n-1}\right\}=\frac{1}{n-1}E\{\sum^{n}(X_i-\bar{X})^2\}$$

$$=\frac{1}{n-1}\cdot(n-1)\sigma^2=\sigma^2$$

（証明おわり）

上記の証明の骨子は $E\{\sum^{n}(X_i-\bar{X})^2\}$ の展開式の第2項 $nE\{(\bar{X}-\mu)^2\}=n\left(\frac{\sigma^2}{n}\right)$ である。

$$\sigma_{\bar{X}}^2 \equiv Var\,(\bar{X})=E\{[\bar{X}-E(\bar{X})]^2\}=E\left\{(\bar{X}-\mu)^2\right\}$$

$$=\frac{\sigma^2}{n}$$

ということは，この証明の対象である母集団が無限母集団であるということである。だがこの〈命題5〉を〈命題2〉と結合させると，論理的に矛盾する。〈命題2〉の母集団は有限母集団だからである。

したがって〈命題2〉と結合できる〈命題5〉の母集団は有限母集団でなければならず，有限母集団を前提とした不偏分散が求められねばならない。

以上，2つの命題結合上の論理矛盾を指摘した。矛盾を解消するには形式上は〈命題2〉の方を無限母集団からの抽出，すなわち $\sigma_{\bar{X}}^2=\frac{\sigma^2}{n}$ と考えれば解決するのであるが，全数調査の抽出集計——標本論理が純粋な形であらわれる——の場合をも考えて，社会標本調査の調査対象（研究対象）としての母集団は，N個の個体から構成される「有限母集団」であるという事実は動かすことができない。そこで，この事実の線にそった〈命題5〉の不偏分散を考えるべきである。

まず $\sigma^2=\sum_{i=1}^{N}(X_i-\mu)^2/N=[\sum^{N}X_i^2-N\mu^2]/N$，

$$S^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/n=[\sum^{n}X_i^2-n\bar{X}^2]/n \quad \cdots\cdots(1)$$

と定義すると，不偏分散 U^2 は

$$U^2=\left(\frac{N-1}{N}\frac{n}{n-1}S^2\right)=\frac{N-1}{N}\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$=\left(1-\frac{1}{N}\right)\frac{1}{n-1}\sum^{n}(X_i-\bar{X})^2$$

である。

証明1

〈命題2〉により

$$\sigma_{\bar{X}}^2=E\{[\bar{X}-E(\bar{X})]^2\}=E\{(\bar{X}-\mu)^2\}=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$$

であるから

$$E\{\sum^{n}(X_i-\bar{X})^2\}=E\{\sum^{n}(X_i-\mu)^2\}-nE\{(\bar{X}-\mu)^2\}$$

$$=n\sigma^2-n\cdot\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}=\frac{N}{N-1}(n-1)\sigma^2$$

となる。

$$\therefore E(U^2)=E\left\{\frac{N-1}{N}\frac{1}{n-1}\sum^{n}(X_i-\bar{X})^2\right\}$$

$$=\frac{N-1}{N}\frac{1}{n-1}\cdot\frac{N}{N-1}(n-1)\sigma^2=\sigma^2$$

証明 2

$\sigma_{\bar{X}}^2=\frac{N-n}{N-1}\frac{\sigma^2}{n}$ を用いないで $E(U^2)=\sigma^2$ を証明すれば，つぎのとおりである。

$$E(S^2)=\frac{1}{n}E(\sum X_i^2-n\bar{X}^2)=\frac{1}{n}\{E\sum^{n}X_i^2-nE(\bar{X}^2)\} \quad \cdots\cdots(2)$$

$$E(\bar{X}^2)=E\left\{\frac{(X_1+X_2+\cdots+X_n)^2}{n^2}\right\}=E\left\{\frac{X_1^2+X_2^2+\cdots+X_n^2}{n^2}\right.$$

$$\left.+\frac{2}{n^2}\underbrace{(X_1X_2+\cdots\cdots+X_{n-1}X_n)}_{\binom{n}{2}\text{ヶ}}\right\}$$

$$E\sum_{i=1}^{n}X_i^2=\sum_{i=1}^{n}EX_i^2=n\cdot\frac{1}{N}\sum_{i=1}^{N}X_i^2 \quad \cdots\cdots(3)$$

$$E\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}X_iX_j=\binom{n}{2}EX_iX_j=\frac{n(n-1)}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}X_iX_j$$

$$\therefore E(\bar{X}^2)=n\left\{\frac{1}{n^2N}\sum_{i=1}^{N}X_i^2+\frac{2n(n-1)}{n^2N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}X_iX_j\right\}\cdots\cdots(4)$$

$$2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}X_iX_j=X_1(X_2+X_3+\cdots+X_N)+X_2(X_1+X_3+\cdots+X_N)+\cdots\cdots$$

$$+X_N(X_1+X_2+\cdots+X_{N-1})$$

$$=X_1(N\mu-X_1)+X_2(N\mu-X_2)+\cdots\cdots+X_N(N\mu-X_N)$$

$$=N\mu\sum^{N}X_i-\sum^{N}X_i^2=N^2\mu^2-\sum^{N}X_i^2 \quad \cdots\cdots(5)$$

(5)を(4)へ代入して

$$E(\bar{X}^2)=\frac{1}{nN}\sum^{N}X_i^2+\frac{n-1}{nN(N-1)}(N^2\mu^2-\sum^{N}X_i^2)$$

$$=\frac{1}{nN}\left\{1-\frac{n-1}{N-1}\right\}\sum^{N}X_i^2+\frac{N(n-1)}{n(N-1)}\mu^2$$

$$=\frac{1}{nN}\left(\frac{N-n}{N-1}\right)\sum^{N}X_i^2+\frac{N(n-1)}{n(N-1)}\mu^2 \cdots\cdots(6)$$

(3)と(6)とを(2)へ代入して

$$E(S^2)=\frac{1}{n}\left[\frac{n}{N}\sum^{N}X_i^2-\frac{N-n}{N(N-1)}\sum^{N}X_i^2-\frac{N(n-1)}{N-1}\mu^2\right]$$

$$=\frac{1}{n}\left[\frac{N(n-1)}{N(N-1)}\sum^{N}X_i^2-\frac{N(n-1)}{N-1}\mu^2\right]$$

$$=\frac{n-1}{n}\frac{1}{N-1}[\sum^{N}X_i^2-N\mu^2]$$

$$=\frac{n-1}{n}\frac{N}{N-1}\sigma^2$$

$$\therefore E(U^2)=E\left\{\frac{N-1}{N}\frac{1}{n-1}\sum^{n}(X_i-\bar{X}^2)\right\}=\sigma^2$$

(証明おわり)

次に　$\sigma_p{}^2=\sum_{i=1}^{N}(X_i-\mu)^2/(N-1)$，$\hat{\sigma}^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ ……(7)

と定義すると，全く同じ証明法により

$$E(\hat{\sigma}^2)=\sigma_p{}^2$$

をえる。

結論

母分散および標本分散を(1)式のように定義すれば，有限母集団を前提とした不偏分散 U^2 は $\frac{N-1}{N}\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ であり，母分散および標本分散を(7)式のように定義しなおせば，不偏分散は $\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ である。ただしこの場合は，〈命題2〉の標準誤差 $\sigma_{\bar{x}}{}^2$ は $\left(1-\frac{n}{N}\right)\frac{\sigma_p{}^2}{n}$ となる。

なお，標本の大きさnが小さく（実験，経験上よりn≦100）で正規分布の信頼区間（大屋氏論文〈命題4〉）が適用できない場合——ただし属性調査の場合は構成比に応じた別個の標本数が必要；W. G. Cochran, *Sampling Techniques,*[注] P. 41参照せよ——は，正規母集団からの抽出であるかぎり t 分布の信頼区間が採用できるが（大屋氏論文〈命題6〉），t 分布の統計量 $t=\frac{\bar{X}-\mu}{U/\sqrt{n}}$ のUは $U^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ としてのみ定義可能であるから，標本調査の数理命題1～6の母分散および標本分散は(7)式で定義した方がその限りでは論理一貫する。

ただし，こうすると母分散の本来の定義である分母の1/Nをゆがめてしまうことになる。

もともと分散は，各変量の基準値からの偏差自乗 $(X_i-a)^2$ の集合を考え，その集合の均等分配値として，算術平均 $\sum^{N}(X_i-a)^2/N$ を考える時に，偏差自乗和が最小になる基準値として平均 $\mu=\sum^{N}X_i/N$ をとった値

$$\sigma^2=\frac{1}{N}\sum^{N}(X_i-\mu)^2$$

である。Nで割って均等分配値を考えたわけであるのに，分子をN-1で割ったのでは均等分配値にならないし，またN-1で割らなければならない必然性がない。

もし，N-1で割ることの必然性をもちだそうとするならば，アメリカ統計協会機関誌J. A. S. A.誌上のW. E. Deming and F. F. Stephan論文「標本としてのセンサスの解釈について」（1941年）を端緒とする増山元三郎氏，北川敏男氏達「推計理論派」の主張される母集団を含む時系列祖母集団（又は時系列超母集団）を想定し，そこでの各個体要素が確率変数であることを論証する以外にない。だが，かかる想定は，われわれ「技術論的

標本理論派」は容認できない。社会的母集団，ましてや祖母集団の各要素が確率変数に従うとの証明は未だなされていない。

したがって，社会的有限母集団以降のところで議論する限り，母分散，標本分散の分母は，それぞれ$1/N$，$1/n$で定義し，不偏分散が問題になるときのみ

$$U^2=\left(1-\frac{1}{N}\right)\frac{1}{n-1}\sum^{n}(X_i-\bar{X})^2$$

と定義しなおす。そして〈命題6〉

正規母集団からの大きさnの標本の統計量$t=(\bar{X}-\mu)\Big/U/\sqrt{n}$は自由度n-1の$t$分布に従って分布する。(ただし，ここでの$U^2$は$U^2=\sum^{n}(X_i-\bar{X})^2/n-1$)

の次に，関連〈命題7〉をおこす。

〈命題7〉正規母集団の前提をはずした場合でも，log-normal扱いできる程度の軽度の正の非対称分布($\mu-m_o<3(\mu-m_e)$)ならば，正規分布の頑健性の理論により，t分布理論の近似的適用が可能である。

このように，〈命題6〉(引用書20頁)でうちきるよりも，〈命題7〉をおこして接続させた方がよいと考えられる。

F分布理論ついても，〈命題7〉の前提条件を満たしている限りでの近似的適用ということになろう。

この分野でいえば，母集団がnormalの前提を設けていないのは，x^2分布だけである。

パラメトリックな推定，検定論では，たとえば標本相関係数rからの母相関係数ρの推定，検定は2変量正規母集団の前提に立っている。しかしながら社会的母集団の分布型で一番多いのはlog-normal扱いできない程度に歪んだ正の非対称分布であり，ついで多いのは逆J型分布であって，

正規分布はきわめて少ない。社会的母集団では，正規分布は normal ではなく abnormal (例外) である。社会的経済量は原始共産社会でも想定しない限り，もともと偏った分布が一般的である。(誤解がないように断っておくが，私は自由平等の理想を否定しているわけではない。不平等の弊害がでてくればこそ，それを是正しようとする動きがでてくるわけであって，それは非常に大事な事であるが，ここでの議論とは別次元の話しである。)

distribution-free のノンパラメトリックな推定・検定論，たとえば順位相関係数 r_s による推定・検定あたりが，もっと活用さるべきであって，検出力は落ちるがやむをえない。正規又は準正規の前提を満しえない場合の r による検定・推定，また R^2 の計算は意味がない。

以上は，論理構造を展開する上での問題点をみてきた。

次にその諸命題〈1～7〉が現実の統計表に記載されている標本統計ではどうなるかについて。そこでは全母集団の推定値もあれば，階層別，地域別，要因別などの構造統計表に従った部分母集団の推定値もある。

全母集団にかんするかぎり，問題はない。有限母集団が歪んだ分布であっても，中心極限定理という理論上の裏づけがあり，さらに実験的，経験的結果から n が大きい ($n \geqq 50$，できれば $n \geqq 100$) ならば，標本平均の分布はほぼ正規分布とみていいことが立証されている。

上記の標本数がとれない小標本 ($n < 50$) の場合はどうか。log-normal 扱いできる程度の歪んだ正の非対称分布の母集団からの標本であり，小標本の randomness さえ保証されていれば，t 分布の理論が近似的に適用できる。

問題となる場合は，クロス・セクション・データで部分母集団についての標本統計についてである。

大屋祐雪氏は，ここでとりあげた論文で，次の 3 つをあげておられる。

(1)　目標精度の措定ある層別集計＝層別集計のマス目

(2)　目標精度の措定なき層別集計＝層別集計のマス目

(3)　集計時階層別のマス目

上記マス目のうち(1)については全母集団の推定値と同じように考えてよい。

(2), (3)についてはどうか。「内包規定のすくない全母集団から，漸時，内包規定の豊かな α 部分母集団になってきている事情を考える（引用書25頁）」と，$n_{ij} \geqq 50$ による正規分布理論の近似的適用および $n_{ij} < 50$ のときの t 分布理論の近似的適用は，みたされ難くなるのではなかろうか。また $N_{ij} = 50$ のときの $(1 - 1/N_{ij}) = 0.98$ であるから，$\bar{X}_{ij}$ の達成精度が問題となるとき，

$$\mu_{ij} = \bar{X}_{ij} \pm t\sqrt{\frac{N_{ij} - n_{ij}}{N_{ij} - 1}} \frac{U_{ij}}{\sqrt{n_{ij}}}$$

の区間推定で $(1 - 1/N_{ij})$ 修正項は無視しえない。

また，多標識有限母集団において，そのすべての標識にかんして，部分母集団がそれぞれ正規分布になるように層別することは困難である。たとえば総務庁統計局『家計調査年報』の全国勤労者世帯について，年間収入18階層，両端2階層を除外したデータについて，横軸に可処分所得，縦軸に消費支出総額，10大費目別消費支出額をとった11箇のヒストグラムを眺めると，どの時点をとっても log-normal 扱いできない程度の偏った正の非対称分布である。同局『就業構造基本調査（全国編）』の年間所得階層別雇用者数の分布も，上記の非正規である。このようにしてみていくと，次のような一般的表現が妥当である。すなわち，標識 $a(=x_i)$ にかんして部分母集団 α, β 等が正規になっていても，標識 $b(=y_i)$ にかんしてはひどい非正規になっていないともかぎらない。

従って，以上の検討を欠いた実証分析での計算は，信頼性に欠ける。

注

SMALLEST VALUES OF *np* FOR USE OF THE NORMAL APPROXIMATION

p	*np*＝number observed in the *smaller* class	*n*＝sample size
0.5	15	30
0.4	20	50
0.3	24	80
0.2	40	200
0.1	60	600
0.05	70	1400
～0*	80	∞

*This means that *p* is extremely small, so that *np* follows the Poisson distribution.

Source : W.G.Cochran, *Sampling Techniques,* John Wiley & Sons, 1953 (Modern Asia First Edition, 1959), P.41