数理解析研究所講究録
第749巻 1991年 162-173

Basuの推定量とその周辺

関東学院大経済　布能英一郎(Eiichiro Funo)

§0. はじめに

本稿は有限母集団サンプリングにおいて、Basu(1971)によって提唱された推定量、および、これに関連した種々の推定量の導出を紹介し、これらが、ある種の条件の下で自乗損失関数に関して許容的であることを示す。また、本稿で扱う種々の推定量のうちのベイズ的でない推定量に対し、Polya pseudo posterior法を用いると、ベイズ的な扱いも出来る。この事についても述べる。

§1. 有限母集団サンプリングにおける Basu の推定量

有限母集団Uの総数を Nとし、ユニット i($1\leqq i\leqq N$) における観測値をy_iとする。ここでベクトル $y=(y_1,y_2,\ldots,y_N)$を未知母数とし、$y\in\mathbb{R}^N=\Theta$であると考える。$\{1,2,\ldots,N\}$ の部分集合 sをサンプルと呼び，n(s)をサンプル sの標本数とする。

与えられた$y\in\Theta$と $s=\{i_1, i_2,\ldots,i_{n(s)}\}$ 但し $1\leqq i_1<\ldots<i_{n(s)}$に対し、$y(s)=(y_{i_1},y_{i_2},\ldots,y_{i_{n(s)}})$とおく。

ここで推定したい量は$\sum_{i=1}^{N} y_i$ であるとする。Basu(1971)は次のような推定量を提案した。

$$e(y,s)=\sum_{i\in s} y_i + \sum_{j\notin s}[(1/n(s))\sum_{i\in s}\{(y_i)/(m_i)\}]m_j$$

ここで、$(m_1,m_2,\ldots,m_N)$ は $y=(y_1,y_2,\ldots,y_N)$ の guess。これを Basu の推定量という。

以後、一般性を失うことなく、i_1を1, i_2を2, $\ldots$, i_nをnとする indexのつけ替えを行って、 $s=\{1,2,\ldots,n\}$ としてよい。また、記号の簡略化のため、

$$\mu=\frac{\sum_{i=1}^{N} y_i}{N},\quad \sigma^2=\frac{\sum_{i=1}^{N}(y_i-\mu)^2}{N},\quad \bar{y}=\frac{\sum_{i=1}^{n} y_i}{n},\quad s^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}$$

とおく。

§2. 他の推定量：

2.1.: $\quad t_0=\sum_{i\in s} y_i+(N-n)\bar{y},\qquad t_\infty=\sum_{i\in s} y_i+\sum_{j\notin s} m_j$

ここで、t_∞を求めるのに、$(y_{n+1},y_{n+2},\ldots,y_N)$の guessである$(m_{n+1},m_{n+2},\ldots,m_N)$すべてが与えられる必要はなく、

$\sum_{i=n+1}^{N} y_i/(N-n)$のguess μ_*が与えられていれば十分である。

さて、t_0, t_∞ および $e(y,s)$ の具体的意味を次の例で与えてみよう。

某国には N人の国民がいて、各国民の所得をy_iとする。y_iは税務署から見れば未知の値である。さて、各国民の申告所得額m_iをもってy_iを推定してもよいかも知れないが、申告所得額は必ずしも信用できない。そこで税務署は「所得調査」を行なうのだが、国民全員の所得調査をすることは、膨大な

時間・労力・費用がかかるので不可能である。そこで税務署は n人に対して所得調査を実施した。所得調査された人々の index を i=1,2,...,n とすると、税務署は所得調査によって $y_1, y_2, ..., y_n$を知ることができたのである。

このような状況で、所得調査できなかった国民の index を j=n+1,n+2,...,N とすると、未知の所得y_jに関して、推定量 $t_\infty, t_0, e(y,s)$は、それぞれ次の様に行っていると言える。

ⓐ t_∞では、申告所得m_jを信用してy_jをm_jで推定した。

ⓑ t_0では、申告所得m_jを全く信用せず、y_jを$\bar{y}$,すなわち所得調査で知り得た$y_1, y_2, ..., y_n$の平均で推定した。

ⓒ Basuの推定量では、所得調査を行ったi=1,2,,,nに対し真の所得y_iと申告所得m_iとの「ごまかし比」y_i/m_i を得ているので、所得調査を受けなかった人には、申告所得に「ごまかし比の平均」を乗じて推定した。

2.2.: 更に別の推定量

0<M<+∞ なる定数 Mを1つ選んでおいて

$$t_M = \sum_{i=1}^{n} y_i + (N-n)\left(\frac{M}{M+n}\mu_* + \frac{n}{M+n}\bar{y}\right)$$

なる推定量を考える。推定量t_Mはμ_*と $\bar{y}$の convex combination であるが、Ericson(1969) は次のような意味づけを与えている。

Pを平均がμ_*であるような任意の分布とする。この時、

$y_1 \sim P$

と仮定し、更に帰納的に i=2,3,...に対し、

$$y_{i+1} \mid y_1, y_2, \ldots, y_i \sim \frac{M}{M+i}P + \frac{i}{M+i}E_i$$

と仮定する。但し、E_i は $y_1, y_2, \ldots, y_i$ の empirical distributionである。そうすると、

$$E\{y_{i+1} \mid y_1, y_2, \ldots, y_i\} = \frac{M}{M+i}\int y\,dP + \frac{i}{M+i}\int y\,dE_i$$

$$= \frac{M}{M+i}\mu_* + \frac{i}{M+i}\cdot\frac{1}{i}\sum_{j=1}^{i} y_j$$

を得る。よって、特に j>nに対し、

$$E\{y_j \mid y_1, y_2, \ldots, y_n\} = \frac{M}{M+n}\mu_* + \frac{n}{M+n}\bar{y}$$

であるから、

$$E\left(\sum_{i=1}^{N} y_i \;\middle|\; y_1, y_2, \ldots, y_n\right)$$

$$= \sum_{i=1}^{n} y_i + (N-n)E\{y_j \mid y_1, y_2, \ldots, y_n\}$$

$$= \sum_{i=1}^{n} y_i + (N-n)\left(\frac{M}{M+n}\mu_* + \frac{n}{M+n}\bar{y}\right) = t_M$$

を得る。

2.3.: これと同様な考え方で、

$$E(\sigma^2 \mid y_1, y_2, \ldots, y_n) = E\left(\frac{\sum_{i=1}^{N}(y_i-\mu)^2}{N} \;\middle|\; y_1, y_2, \ldots, y_n\right)$$

を求めてみよう。分布Pの分散をσ_*^2と置く。j,k>n, j≠k にて

$$E(y_j^2 \mid y_1, y_2, \ldots, y_n) = \frac{M}{M+n}(\sigma_*^2+\mu_*^2) + \frac{1}{M+n}\sum_{i=1}^{n} y_i^2$$

$$E(y_j y_k | y_1, y_2, \ldots, y_n) = \frac{(n\bar{y}+M\mu_*)^2 + \sum_{i=1}^{n} y_i^2 + M(\sigma_*^2+\mu_*^2)}{(M+n)(M+n+1)}$$

が計算によって示されるので、後は単なる代数計算で

$$E(\sigma^2 | y_1, y_2, \ldots, y_n) = \frac{n(N+M)(NM+Nn+n)}{N^2(M+n)(M+n+1)} s^2$$

$$+ \frac{(N-n)M(NM+Nn-M)}{N^2(M+n)(M+n+1)} \sigma_*^2 + \frac{(N-n)nM(N+M)}{N^2(M+n)(M+n+1)} (\bar{y}-\mu_*)^2$$

が導ける。$v_M = E(\sigma^2 | y_1, y_2, \ldots, y_n)$ と置くと、

$$\lim_{M \to 0} v_M = \frac{n(N+1)}{(n+1)N} s^2 \equiv v_0,$$

$$\lim_{M \to \infty} v_M = \frac{n}{N} s^2 + \frac{n(N-n)}{N^2} (\bar{y}-\mu_*)^2 + \frac{(N-n)(N-1)}{N^2} \sigma_*^2 \equiv v_\infty$$

が得られる。

2.4.: Basuの推定量 $e(y,s)$ を B_0 と書き直す。すなわち、

$$B_0 = \sum_{i=1}^{n} y_i + \left(\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{y_i}{m_i} \right) \sum_{j=n+1}^{N} m_j$$

として、先程 t_M を構成したのと同様の操作を行ってみる。つまり、

$$B_M = \sum_{i=1}^{n} y_i + \left(\frac{M}{M+n} \xi_* + \frac{n}{M+n} \bar{\xi} \right) \sum_{j=n+1}^{N} m_j$$

なる推定量を作る。なお、ここで $\xi_i = y_i / m_i$ $(i=1,2,\ldots,N)$,

$\bar{\xi} = \sum_{i=1}^{n} \xi_i / n$ であり、ξ_* は $\sum_{i=n+1}^{N} \xi_i / (N-n)$ の guessである。

そうすると、

$$\lim_{M\to\infty} B_M = B_0, \quad \lim_{M\to\infty} B_M = \sum_{i=1}^{n} y_i + \xi_* \sum_{j=n+1}^{N} m_j \equiv B_\infty$$

を得る。

2.5.:

Basu の推定量では、ratio $\xi_i = y_i/m_i$ を用いたが、difference $d_i = y_i - m_i$ $(i=1,2,\ldots,N)$ を用いることもできる。$\bar{d}$ を $d_1,\ldots,d_n$ の算術平均, i.e., $\bar{d} = \sum_{i=1}^{n} d_i/n$, d_* を $\sum_{i=n+1}^{N} d_i/(N-n)$ の guess とすると、

$$B_0' = \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} m_j + (N-n)\bar{d}$$

$$B_M' = \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} m_j + (N-n)\left(\frac{M}{M+n} d_* + \frac{n}{M+n}\bar{d}\right)$$

$$B_\infty' = \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} m_j + (N-n) d_*$$

が得られる。

§3. 推定量の自乗損失下での許容性

この節では、§1, §2 で導出した推定量のうち、t_M, t_∞, t_0, $B_M, B_\infty, B_0, B_M', B_\infty', B_0'$ の自乗損失下での許容性を調べる。

定理

条件1: $y_1, y_2, \ldots, y_N$ の値は r個の有限な値 $\alpha_1 < \alpha_2 < \ldots < \alpha_r$ しか取らない。

条件2: $\alpha_1 \leqq \mu_* \leqq \alpha_r$, $\alpha_1 - \alpha_r \leqq d_* \leqq \alpha_r - \alpha_1$,

$$\min_{i,j} \{\alpha_i/m_j\} \leqq \xi_* \leqq \max_{i,j} \{\alpha_i/m_j\}$$

を仮定する。また、推定量 B_0, B_Mを考える際には y_i/m_i $(i=1,2,\ldots,n)$が well-defined であるために$m_i \neq 0$ $(i=1,2,\ldots,n)$ を当然のこととして仮定するが、B_0の場合には、更に強く、すべての $i=1,2,\ldots,N$に対し、$m_i \neq 0$ を仮定する。

そうすると、推定量 $t_M, t_\infty, t_0, B_M, B_\infty, B_0, B_M', B_\infty', B_0'$ はすべて自乗損失下で許容的である。

Remark 1. 条件1 を Hartley and Rao(1968,1969) は scale load situationと呼んでいる。

Remark 2. 条件2 は guess μ_* が$\alpha_1, \alpha_2, \ldots \alpha_r$ の最大値を越えたり、最小値を下回ったりする事は無いという、ごく自然な仮定である。guess d_*, ξ_* についての条件も同様である。

Basuの推定量B_0の許容性を、まず証明する。

m_j $(j=1,2,\ldots,N)$ は既知ゆえ、未知ベクトル $(y_{n+1}, y_{n+2}, \ldots, y_N)$の推定問題は $(\xi_{n+1}, \xi_{n+2}, \ldots, \xi_N)$の推定問題となる。

\$2.2. で行ったように、$\xi_1, \xi_2, \ldots, \xi_n$ のempirical distribution を E_n とし、各 $j=n+1, n+2, \ldots, N$に対し

$$\xi_j \mid \xi_1, \xi_2, \ldots, \xi_n \sim \text{ i.i.d. } E_n \qquad (*)$$

を仮定すると、$E[\xi_j \mid \xi_1, \xi_2, \ldots, \xi_n] = \bar{\xi}$ ゆえ、

$$E\left(\sum_{i=1}^{N} y_i \;\middle|\; \xi_1, \xi_2, \ldots, \xi_n \right) = \sum_{i=1}^{N} E\left(\xi_i \;\middle|\; \xi_1, \ldots, \xi_n \right) m_i$$

$$= \sum_{i=1}^{n} \xi_i m_i + \sum_{j=n+1}^{N} E[\xi_j \mid \xi_1, \xi_2, \ldots, \xi_n]\, m_j$$

$= \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} \xi\, m_j = B_0$ を得る。このことから、$\sum_{i=1}^{N} y_i$ の許容的推定量を考察することは (*)におけるE_nの mean ξ の許容的推定量を考察することに帰着する。j=n+1,...,N に対し、$Z_j = \xi_j \mid \xi_1,\ldots,\xi_n$ なる記号を導入する。y_jは $\alpha_1,\ldots,\alpha_r$の値しか取らないとの仮定より、Z_jは有限個の値$\beta_1,..,\beta_k$しか取らない。β_1を取る確率をθ_1, β_2をとる確率を$\theta_2,..,\beta_k$を取る確率をθ_kとすると、 $\xi = \sum_{\ell=1}^{k} \beta_\ell \theta_\ell$ である。

さて、θ_iを次のように推定できる：$\xi_1, \xi_2,\ldots,\xi_n$は既知ゆえこの n個の標本の中で$\beta_1$の値を取るものが$x_1$個, β_2の値を取るものがx_2個,....,β_kの値を取るものがx_k個とすると、

$$(X_1, X_2, \ldots, X_k)' \sim \text{Multinomial}(n, (\theta_1, \theta_2, \ldots, \theta_k)')$$

である。ここでθ_ℓの推定量x_ℓ/nは自乗損失下で許容的。

そして、$\sum_{\ell=1}^{k} x_\ell = n$, $\sum_{\ell=1}^{k} \beta_\ell x_\ell = \sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} (y_i/m_i)$ であるから $\xi = \sum_{\ell=1}^{k} \beta_\ell \theta_\ell$ の推定量 $\sum_{\ell=1}^{k} \beta_\ell x_\ell / n = (1/n) \sum_{i=1}^{n} (y_i/m_i)$は自乗損失下で許容的。以上でBasuの推定量が許容的であることが示せた。

推定量B_ϖは、事前分布 $T=(T_1, T_2, \ldots, T_k)$ such that

$$E[(\beta_1, \beta_2, \ldots, \beta_k)(T_1, T_2, \ldots, T_k)'] = \xi_*$$

に関する Bayes解であるから、許容性は直観的には明らかである。ただ、T～Multivariate Beta$(a_1, a_2, .. a_k)$ such that

$$\sum_{\ell=1}^{k} \beta_\ell a_\ell / \{a_1+a_2+\ldots+a_k\} = \xi_*$$

に選べる保証は無い。たとえば、$\xi_*=\beta_1$ ならば、a_1以外はすべて 0にしなくてはならない。このようなときには、事前分布をθ_1に確率 1 を入れるようなものに選ぶ必要がある。よって、数学的に厳密に証明するには、Multivariate Beta に事前分布を選べる場合と、特定のθ_iにはzero probabilityを入れるような事前分布の場合にわかれるが、いづれの場合もθ_iの推定量 $a_i/(a_1+a_2+\ldots+a_k)$は自乗損失下で許容的。以後の議論はB_0の場合に同じ。

推定量B_MについてもB_0とB_∞のconvex combinationゆえ、適当な事前分布を選んで、この Bayes解としてB_Mが求まる。但し、B_Mの事前分布をMultivariate Beta に選べない事もあるが、その際の処理はB_∞の場合に同じ。

推定量t_0, t_M, t_∞はB_0, B_M, B_∞の特殊な場合、すなわち、

$$m_{n+2}=1, \ldots, m_N=1, \quad \xi_{n+1}=y_{n+1}, \xi_{n+2}=y_{n+2}, \ldots, \xi_N=y_N$$

とした場合と見做せば、自明である。

推定量B_0', B_M', B_∞'の場合も同様にして出来る。

<u>$4. Polya posterior 法による pseudo posterior distributionの構成</u>

4.1.: n個の標本 $y_1, y_2, \ldots, y_n$ の値が、r個の異なった値 $\alpha_1, \alpha_2, \ldots, \alpha_r$ $(r \leqq n)$ のいずれかであって、更に n個の標本のうち α_j の値を取るものが k_j 個あったとする。

このとき、次のようにして pseudo posterior distribution を構成する。

壷Uの中には n個のボールがあり、そして各ボールには $\alpha_1, \alpha_2, \ldots, \alpha_r$ $(r \leqq n)$ のいすれかのラベルが貼られているものとする。ラベルが同一なボールは同一のものと見なす。そして各 $i=1,2,\ldots,r$ に対し、ラベル α_i のボールが k_i 個存在するとする。この壷Uより

① randomに1つのボールを壷Uより取り出す。この取り出したボールのラベルを y_{n+1} の値とする。

② ①で取り出したボールと、更にこのボールと同一のラベルを持つ別のボール合計2個を、壷Uの中に入れる。よってこの時点で、壷Uには n+1個のボールがある。

③ ②の状態の壷Uからrandomに1つのボールを取り出す。この取り出したボールのラベルを y_{n+2} の値とする。

④ ③で取り出したボールと、更にこのボールと同一のラベルを持つ別のボール合計2個を、壷Uの中に入れる。よって、この時点で、壷Uには n+2個のボールがある。

この操作を繰り返して、$y_{n+3}, y_{n+4}, \ldots, y_N$ の値を得る。

4.2.: Polya posterior 法を用いると、任意の $j=n+1, n+2, \ldots, N$ に対して、

$$E(y_j|y_1,y_2,\ldots,y_n)=\bar{y}$$

となる。よって、

$$E(\sum_{\ell=1}^{N} y_\ell|y_1,y_2,\ldots,y_n) = \sum_{i=1}^{n} y_i+(N-n)E[y_j|y_1,y_2,\ldots,y_n]$$

$$= n\bar{y}+(N-n)\bar{y} = N\bar{y}$$

を得る。

4.3.: Polya posterior 法をBasuの推定量に用いることもできる。この際は $\xi_i=y_i/m_i$ に対し、4.1.で行った操作をすればよい。

4.4.: Basuの推定量は $m_1,m_2,\ldots,m_N$ が、与えられた定数と思えば non-Bayes推定量である。しかしながら、上記のようにして壷からボールを random に取り出す操作で posterior が求まったと思えば、「Bayes 的」と見ることができる。

References

Basu, D. (1969). Role of sufficiency and likelihood principles in sample survay theory. Sankhya, A 31, 441-454.

Basu, D. (1971). An essay on the logical fundations of survey sampling, part one. Foundations of Statistical Inference, edited by V.P.Godambe and D.A. Sprott. Holt, Rinehart and Winston, Toronto.

Chaudhuri, A. (1978). On estimating the variance of a finite population. Metrika, 25, 65-76.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite population (with discussion). J.Roy.Statist.Soc., Ser.B 31, 195-233.

Ericson, W.A. (1970). On a class of uniformly admissible estimators of a finite population total. Ann.Math.Statist., 41, 1369-1372.

Godambe,V.P. (1966). Bayes and empirical Bayes estimation in sampling finite population(Abstract). Ann.Mat.Statst., 37, 552.

Godambe,V.P. (1969). Admisibility and Bayes estimation in sampling finite populations V. Ann.Math. Statist., 40, 672-676.

Hartley,H.O. and Rao,J.N.K. (1968). A new estimation theory for sample survays. Biometrika, 55, 547-557

Hartley,H.O. and Rao,J.N.K. (1969). A new estimation theory for sample survays II. New Developments in Survay Sampling, edited by N.L.Johnson and H. Smith. Wiley, New York.

Joshi,V.M. (1965). Admisibility and Bayes estimation in sampling finite populationc II and III. Ann. Math. Statist., 36, 1723-1742.

Meeden,G. and Ghosh,M. (1983). Choosing between experiments:Applications to finite population sampling. Ann.Statist., 11, 296-305.

Meeden,G. Ghosh,M.and Vardeman,S. (1985). Some admissible nonparametric and related finite population sampling estimators. Ann.Statist., 13, 811-817.

Scott,A.J. (1975). On admissibility and uniform admissibility in finite population sampling. Ann. Statist., 3, 489-491.

Vardeman,S.and Meeden,G. (1983). Admissible estimators in finite population sampling employing various types of prior information. J. Statist. Planning Inf., 7, 329-341.

Vardeman,S. and Meeden,G. (1984). Admissible estimators for the total of a stratified population that employ prior information. Ann.Statist., 12, 675-684.