ある Biased Estimator を使用する時の注意

多　賀　保　志

(1953年7月　受付)

On the control method of a biased estimator

Yasusi TAGA

In estimating a parameter, often occur the cases when it is inevitable to use a biased estimetor. (For example, in estimation of the ratio of cheaters in the whole testers used for a survey-cheater problem.) In such cases, we must try to minimize the mean square error of the estimate under the condition that the total cost is fixed, namely the total sample size is constant. We could succeed in solving the problem above mentioned by numerical calculation, but it is difficult to get an exact solution analytically.

In conclusion, as a tester does cheating about a few samples when he is annoyed to call back them, we had better take less testers (namely more samples per tester) in order to diminish the bias of our estimate.

Institute of St·tistical Mathematics

我々が使用する推定量としては，通常不偏性 (unbiasedness) を持つことが必須條件となつているが，ある種の問題に於ては偏りのある推定量 (biased estimate) の使用を避けることが出来ない場合がある．例えばある調査終了後，サンプリングによつて嘘僞の報告をした調査員の比率を推定する問題 (cheater problem) を考えると，抽出された調査員の受持つたサンプルを全部調査すれば bias は生じないが，一部調査する限り実際に cheating を行つていても見逃されてしまう場合が出て来，然も予めその bias の大きさを知ることは出来ないから，どうしても過小な biased estimator を使わざるを得ないことになる．ではこの様な場合如何なる処置を講じたら良いであろうか．費用さえ許すならば全数調査を行つて bias の解消に努めるべきであろうが，多くの場合本調査に費用を使い果して，この様な吟味調査を行う余裕に乏しいのが普通であるから，費用一定という條件下に平均自乗誤差 (分散と偏差の自乗の和) を最小ならしめる様なサンプリング企画を立てるべきであろう．以下この問題について述べて行くが，我々の経験によると，cheater (cheating を行つた調査員) の全調査員に対する比率は 0.15～0.20，サンプルの数にして 0.05～0.10 の程度であるから，cheater は受持ちサンプル中の $1/3$ 位について cheating を行うことになる．従つて cheater を発見するには，調査員を多く抽出するよりも，むしろ一調査員当りのサンプル数を多くとつた方が bias を小さく出来る (その代り variance は大きくなる)．だからこの bias と variance を睨み合せて最良の sampling design を考えるべきであろう．これは結局平均自乗誤差を最小に押えれば良いことになり，我々の設定した簡単なモデルについて計算してみると，抽出すべき調査員の数が少い程 (従つて一調査員当りのサンプル数が多い程)，bias は小さく且つ平均自乗誤差は抽出調査員数が 35 人 (90 人中) の時最小となる．

勿論1人の cheater が多くのサンプルについて cheating を行つている様な場合 (例えば10人中5人以上) ならば，抽出調査員数を多くとつた方が良くなることはいうまでもない．要は bias と variance の兼合いが大切なのである．

調査員番号	1	2	……	t
サンプル数	n_1	n_2	……	n_t
インチキ数	r_1	r_2	……	r_t
郵便用サンプル数	m_1	m_2	……	m_t

或る調査に於て，n人のサンプルを調べるのにT人の調査員を使用したとする．第i調査員の受持ちサンプル数をn_i人，その中で cheating を行つたサンプル数をr_i人とする．($0 \leqq r_i \leqq n_i$) 今この cheating の有無を check する為，調査終了後直ちに被調査者を再訪問するか又は郵便調査を行つて，調査員が来訪して面接調査をしたか否かを確めるとしよう．その際訪問するにせよ，郵便によるにせよ，全サンプルについて再調査を行うことは困難である場合が多いから（主として費用と労力の点で），調査員なりサンプルなり，その一部を抽出して調査する方法を考える．

今T人の調査員の中からt人を equally probablly に抽出し，その各々の受持ちサンプルn_{ij}人よりm_{ij}人を抽出したとする．推定すべき parameter は，T人の調査員母集団中 cheating を行つたものの比率Pであるとする．即ちi調査員の受持サンプルn_{ij}人中，1人でも cheanting が行われた場合に 1，1人も cheanting が行われない場合に 0 となる様に標識X_iを定めておけば，

$$\overline{X} = \frac{1}{T}\sum_{i=1}^{T} X_i = P$$

となる訳である．我々はこのPを estimate する統計量として

$$\bar{x} = \frac{1}{t}\sum_{i=1}^{t} x_i$$

を考える．但しx_iはn_i人より抽出されたm_i人中1人でも cheating が行われていれば 1，さうでない時0なる値をとる確率変数とする．すると

$$P(x_i = 0) = \frac{n_i - r_i}{n_i}\,\frac{n_i - r_i - 1}{n_i - 1}\cdots\frac{n_i - r_i - m_i}{n_i - m_i} \equiv q_i$$

$$P(x_i = 1) = 1 - q_i \equiv p_i$$

となる．$\bar{x}$の平均を考えると

$$E(\bar{x}) = \frac{1}{T}\sum_{i=1}^{T} p_i = \frac{1}{T}\sum_{j=1}^{L} p_{ij} = \frac{L}{T} - \frac{1}{T}\sum_{j=1}^{L} q_{ij} = P - \frac{1}{T}\sum_{j=1}^{L} q_{ij}$$

但し$r_i=0$（cheating なし）ならば$q_i=1$即ち$p_i=0$となり，$r_i=0$ならばm_iを十分大（$m_i=n_i-r_i$）とすると$q_i=0$，即ち$p_i=1$となる。従つてT人中L人が cheating を行つたとすれば，$(T-L)$人については$p_i=0$となる．上記の$\sum_{j=1}^{L} p_{ij}$は$p_i=0$を除いた時の和を表わすものとする．

さて調査員の行つた cheating の比率は

$$P = \frac{L}{T}$$

で表わされるから，$\bar{x}$をPの estimate とすると平均的に言つて under estimate となり，bias の大きさは$d = \frac{1}{T}\sum_{j=1}^{L} q_{ij}$となつている．この$d$は$m_i$の大きさが小さいと比較的大きくなり得るから（普通$r_i$は$n_i$に比して小さいから，$m_i=1$の程度ならば$q_i \fallingdotseq 1$，即ち$d \fallingdotseq P$となる），なるべく$d$を小さくするには$m_i$を大きくしなければならぬ。然し$m = \sum_{i=1}^{t} m_i$を一定とすれば，$m_i$を大きくすると$t$が小さくなるから，bias dは小さくなつても，$\bar{x}$の variance が大きくなつてくる．そこでσ_x^2を求めると

$$\sigma_x^2 = \frac{1}{tT}\sum_{i=1}^{T} p_i q_i + \frac{T-t}{T-1}\,\frac{\sigma_p^2}{t}$$

但し　$\sigma_p^2=\frac{1}{T}\sum_{i=1}^{T}(p_i-\bar{p})^2$,　$\bar{p}=\frac{1}{T}\sum_{i=1}^{T}p_i$

となる．第一項は，いわば内分散，第二項は外分散（調査員間）に相当する項である．

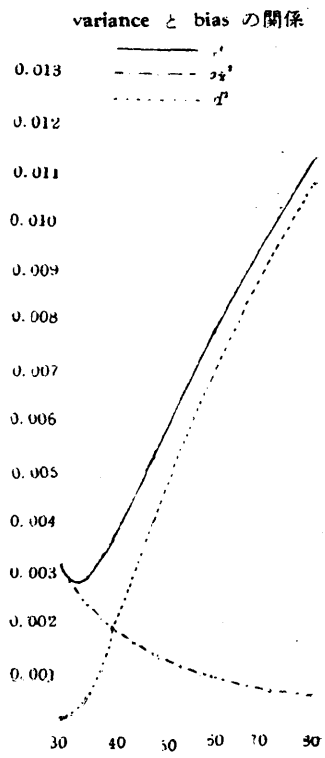
こゝで optimum allocation を求めるには $\sum_{i=1}^{t}m_i=m$（一定）なる條件下に平均自乗誤差 τ^2 を min ならしめれば良い．何となれば，

$$P\{|\bar{x}-P|\leqq h\tau\}\geqq w(h-1)$$

なる故，信賴度 $w(h-1)$ を fix すれば h が fix され，從つて τ の大きさによつて信賴区間 $(\bar{x}-h\tau,\ \bar{x}+h\tau)$ の巾 $2h\tau$ が決ることになるからである．さて τ^2 の大きさを求めると，

$$\begin{aligned}\tau^2&=E(\bar{x}-P)^2\\&=E(\bar{x}-E(\bar{x}))^2+(E(\bar{x})-P)^2\\&=\sigma_{\bar{x}}^2+d^2\\&=\left(\frac{1}{tT}\sum_{i=1}^{T}p_iq_i+\frac{T-t}{T-1}\frac{\sigma_p^2}{t}\right)+\left(\frac{1}{T}\sum_{j=1}^{L}q_{ij}\right)^2\end{aligned}$$

となる．所が初めの括弧内は $\frac{1}{t}$ の order であるのに，第二の括弧内は t を大きくするにつれて大きくなるのであるから，余り多くの調査員をとつても第二項の大きさがものを言つて，かえつて τ^2 は大となる．寧ろ抽出する調査員数をへらして各調査員当りのサンプル数 m_i を大きくとり，第二項 $\left(\frac{1}{T}\sum_{j=1}^{L}q_{ij}\right)^2$ を小さくする様にした方が精度の良い結論が得られることになる．但し m_i の optimum allocation，從つて抽出調査員数 t の大きさを実際に求めることは困難である．然し当研究所で行つた「国民性の調査」の Pretest について，大体の見当をつけて見よう．



先づ $T=90$, $n=720$, $n_i=\frac{720}{90}=8$, $m=240$, $m_i=\frac{240}{t}$, $r_i\leqq 2$, $L=15$ とすると

$$q_i=\frac{7}{8}\cdot\frac{6}{7}\cdot\cdot\frac{7-\frac{240}{t}}{8-\frac{240}{t}}\qquad(r_i=1\text{ の時})$$

更に，$r_{ij}=2$ なる場合は存在しないと仮定すれば，下表のような結果を得る．

m_i / t	8	7	6	5	4	3
	30	35	40	48	60	80
$\frac{1}{90t}\Sigma p_iq_i$	0	0.00052	0.00078	0.00081	0.00069	0.00049
$\frac{90-t}{89}\frac{\sigma_x^2}{t}$	0.00312	0.00188	0.00110	0.00053	0.00020	0.00003
$\sigma_{\bar{x}}^2$	0.00312	0.00240	0.00188	0.00135	0.00089	0.00052
$\left(\frac{1}{90}\Sigma q_i\right)$	.0	0.00043	0.00174	0.00891	0.00694	0.01085
τ^2	0.00312	0.00283	0.00362	0.00526	0.00783	0.01137

この結果より見ると，$t=35$（$m_i=7$）の場合に τ^2 は最も小さくなつている．結論として調査員の cheating 率を推定する場合は通常 r_i が小さいから，m_i を或る程度大きく（従つて t を小さく）した方が精度の良い（bias の小さい）推定が出来ることになる．

（統計数理研究所）