層別副次抽出法の最良抽出比について

多　賀　保　志

(1954 年 4 月受付)

On the Optimum Sampling Ratios of Stratified Sub-samplings.

Yasushi TAGA

It is often needed to decide the optimum Sampling ratios of sub-samplings in designing various surveys. Applying our result below to the Book Survey conducted October last year, it was cleared that the so the called between-variances were very large in comparison with the within-variances. Therefore, the sampling ratio of book stores (primary samplingunits) must be taken comparatively large. And our stratification by their subscribed capitals or amounts sold is very effective—effects are about 30% ~60%.

We are now studying the relation between orders of cost, variance and bias, and will soon publish it.

副次抽出法によつて，サンプリングを行う場合，母集団に於ける外分散が内分散に比べて非常に大きい時は，どうしても第1次抽出単位の層別が必要となつてくる．(ここでは簡単の為2回抜きの場合について考えることにする)．

各層から第1次抽出単位を1個づつ抽出する場合の費用と精度の関係については既に述べたことがあるから，ここでは各層から任意有限個の第1次抽出単位を抽出することにして，調査費用を一定とした時に推定の精度を最大ならしめる第1次及び第2次抽出単位の「最良抽出比」を求めて見よう．(勿論母集団は有限とし，費用函数は抽出される第1次抽出単位数及び平均サンプル数について線型であるとする．)——昨年 10 月東京都で行われた読書調査の例について見ると，次の様な興味ある結果が得られた：

1) 外分散は内分散に比べて遙かに大きい．従つて層別の効果は可成り大である．(約30~60%)
2) 単純集計が出来る様に，各層に於ける第1次及び第2次の抽出比の積を一定にした場合と，さういう条件をつけない場合とを比較すると，調査費用が同じならば両者の精度は殆ど差がない．従つて集計の時間と費用を考える時，問題なく前者の方法が優れている．
3) 費用とサンプル数の関係を見ると，案外得られるサンプル数が少いのに気付く．これは外分数が大きいからに外ならない．更に精密な層別を要すると考えられる．

さて初めから順を追つて述べてゆくことにする．先づ母集団は R 個の層より成るものとし，第 i 層には M_i 個の第1次抽出単位が含まれるとする．それら第1次抽出単位の大きさを夫々

$$N_{i1},\ N_{i2}\cdots,\ N_{iM_i}\quad (i=1,2,\cdots,R)$$

とし，この中から m_i 個を等確率で抽出する：

$$N_{i(1)},\ N_{i(2)}\cdots,\ N_{i(m_i)},\ (i=1,2,\cdots,R)$$

更にこれらの中から，次の様に第2次抽出単位を抽出することにする：

$$N_{i(j)} \longrightarrow n_{ij} \quad \begin{pmatrix} i=1,2,\cdots\cdots R \\ j=1,2,\cdots\cdots m_i \end{pmatrix}.$$

ここで $\frac{m_i}{M_i}=\frac{1}{r_{i1}}$，$\frac{n_{ij}}{N_{i(j)}}=\frac{1}{r_{i2}}$ を夫々第1次及び第2次の抽出比と呼ぶことにし，ある標識 X についての母集団総和 $X=\sum_{i}^{R}\sum_{j}^{M_i}\sum_{k}^{N_{ij}}X_{ijk}$ を推定する為の推定量として，

$$x=\sum_{i=1}^{R}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{N_{i(j)}}{n_{ij}}x_{ij}, \quad \left(x_{ij}=\sum_{k=1}^{n_{ij}}x_{ijk}\right)$$

$$=\sum_{i=1}^{R}r_{i1}r_{i2}x_i, \qquad \left(x_i=\sum_{j=1}^{m_i}x_{ij}\right)$$

を用いると，

$$E(x)=X$$

$$\sigma_x^2=\sum_{i=1}^{R}a_i r_{i1}(r_{i2}-1)+\sum_{i=1}^{R}(r_{i1}-1)b_i \qquad (1)$$

となる．

ここに

$$a_i=\sum_{j=1}^{M_i}\frac{N_{ij}\sigma_{ij}^2}{1-N_{ij}^{-1}},\quad b_i=\frac{M_i\sigma_{bi}^2}{1-M_i^{-1}}$$

$$\sigma_{ij}^2=\frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}}(X_{ijk}-\bar{X}_{ij})^2$$

$$\sigma_{bi}^2=\frac{1}{M_i}\sum_{j=1}^{M_i}(X_{ij}-\bar{X}_i)^2$$

$$\bar{X}_{ij}=\frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}}X_{ijk},\quad X_{ij}=\sum_{k=1}^{N_{ij}}X_{ijk}.$$

$$\bar{X}_i=\frac{1}{M_i}\sum_{j=1}^{M_i}X_{ij}.$$

この場合サンプル数 $n=\sum_{i=1}^{R}\sum_{j=1}^{m_i}\frac{N_{i(j)}}{r_{i2}}$ は確率変数であるが，その平均を求めると，

$$\bar{n}=E(n)=\sum_{i=1}^{R}\sum_{j=1}^{M_i}\frac{N_{ij}}{r_{i1}r_{i2}} \qquad (2)$$

となる．調査費用が近似的に $\bar{n}$ 及び $m\left(=\sum_{i=1}^{R}m_i\right)$ の一次式として表わされる場合に最良抽出比を求めて見よう．それには

$$C=C_0+c_1m+c_2\bar{n}=C_0+c_1\sum_{i=1}^{R}\frac{M_i}{r_{i1}}+c_2\sum_{i=1}^{R}\sum_{j=1}^{M_i}\frac{N_{ij}}{r_{i1}r_{i2}} \qquad (3)$$

が一定なる条件の下に，σ_x^2 を最小ならしめる様に r_{i1} 及び $r_{i2}(i=1,2,\cdots,R)$ を定めたらよい．

容易に

$$r_{i1}'=\lambda'^{\frac{1}{2}}\sqrt{\frac{c_1M_i}{b_i-a_i}},\quad r_{i2}'=\sqrt{\frac{c_2N_i}{c_1M_i}\cdot\frac{b_i-a_i}{a_i}} \qquad (4')$$

但し

$$\lambda'^{\frac{1}{2}}=\frac{1}{C-C_0}\sum_{i=1}^{R}\left\{\sqrt{c_1M_i(b_i-a_i)}+\sqrt{c_2N_ia_i}\right\} \qquad (5')$$

となることが判る．

この時（1）式は

$$\sigma_x'^2 = \lambda'(C - C_0) - b', \qquad \left(b' = \sum_{i=1}^{R} b_i\right) \tag{6'}$$

となる.

次に単純集計が出来る様に $r_{i1}r_{i2} = r$（一定）なる条件をつけると，上の各式は次の様になる：

$$r_{i1}'' = \lambda''^{\frac{1}{2}}\sqrt{\frac{c_1 M_i}{b_i - a_i}}, \quad r_{i2}'' = \sqrt{\frac{c_2 N}{c_1 M_i} \cdot \frac{b_i - a_i}{a}} \tag{4''}$$

但し

$$\lambda''^{\frac{1}{2}} = \frac{1}{C - C_0}\left\{\sum_{i=1}^{R}\sqrt{c_1 M_i(b_i - a_i)} + \sqrt{c_2 N a}\right\} \tag{5''}$$

$$N = \sum_{i=1}^{R} N_i, \qquad a = \sum_{i=1}^{R} a_i$$

$$\sigma_x''^2 = \lambda''(c - c_0) - b' \tag{6''}$$

となる．$\left(\text{この時 } \bar{n}'' = N/r'', \text{ 但し } r'' = r_{i1}''r_{i2}'' = \lambda''^{\frac{1}{2}}\sqrt{\frac{c_2 N}{a}}\right)$

最後に層別を行わない場合を考えると，

$$r_1 = \lambda^{\frac{1}{2}}\sqrt{\frac{c_1 M}{b - a}}, \qquad r_2 = \sqrt{\frac{c_2 N}{c_1 M} \cdot \frac{b - a}{a}} \tag{4}$$

但し

$$\lambda^{\frac{1}{2}} = \sqrt{c_1 M(b-a)} + \sqrt{c_2 N a} \tag{5}$$

$$\sigma_x^2 = \lambda(C - C_0) - b, \quad \left(b = \frac{M\sigma_b^2}{1 - M^{-1}}\right) \tag{6}$$

となる．$\left(\text{この時 } \bar{n} = N/r,\ r = r_1 r_2 = \lambda^{\frac{1}{2}}\sqrt{\frac{c_2 N}{a}}\right)$.

これらの結果を前記の読書調査に適用して見ると，次の様になる：

［i］ 抽出単位 {第1次抽出単位：書店 / 第2次抽出単位：書籍（各書店で一日の間に売れるもの）}

［ii］ 層別の方法——書店の規模によつて，大きい方より順に，A・B・C・Dの4層に分けた．各層に於ける書店数 M_i 及び売上書籍冊数 N_i は第1表の通り（但し表中の N_i の値は調査の結果よりの推定値である）.

第1表　母集団の大きさ

i	1	2	3	4	計
M_i	6	36	68	349	459
N_i	6982	7839	5956	10336	31113

［iii］ 調査費用——調査員の謝金と交通費が大部分を占めるが，原則として1つのカウンターに1人の調査員を配置したから，調査費用は大部分抽出店数 m によつてきまると考えてよい．

そこで　　$c_1 = 400,\ c_2 = 100,\ C_0 = 0$

とすれば，

$$C = 400\,m + 100\,\bar{n}$$

と表わされる.

［iv］ 書籍売上高についての推定 ——— $a_i,\ b_i,\ X_i$ の推定値を求めると第2表の様になる.

第 2 表　内分散と外分散の大きさ及び売上高

i	1	2	3	4	計
a_i	112807	68267	26647	50412	258133
b_i	15708019	26907597	2681359	1625934	46922909
X_i(万円)	107.1	124.7	84.3	158.4	474.5(万円)

これより C の種々な値に対して，上の3つの場合につき相対精度を比較して見ると次の様になる：

第 3 表　相対精度の比較及び最良抽出比

C(万円)	1	2	3	5	8	10
σ_x'/X	0.244	0.165	0.128	0.088	0.054	0.036
σ_x''/X	0.248	0.168	0.130	0.090	0.056	0.038
σ_x/X	0.545	0.376	0.299	0.220	0.158	0.131
r''	850.2	425.1	283.4	170.1	106.3	85.0

これによつて見るに，σ_x'/X と σ_x''/X との間には殆ど差がない．つまり単純集計可能の為の条件，$r_{i1}r_{i2}=r''$（一定）なる条件をつけてもつけなくても，推定の精度には殆ど変りがないことが判る．従つて集計の手数及び費用を考えた時，後者の方法が遙かに優れていることが判るであらう．この意味でこの r'' を最良抽出比と呼ぶことにする．又層別しないでサンプリングした場合の精度 σ_x/X は，この両者の2倍以上になることが判る．（つまり層別の効果は 50% を上廻ることになる．）サンプル数について見ると，何れの場合も，調査費用に対して殆ど直線的に増加する．（[第1図] 参照）

[v] 文庫率についての推定——全売上冊数を 100 とした時，その中でいわゆる文庫本の占める割合 (%) を「文庫率」と呼ぶことにし，母集団における文庫率を推定して見よう．この場合は二項分布によつて計算を行うことになるが，大体 [iv] と同じ様な結果が得られる．ただ層別の効果は前程よくないが，それでも 30% 位となつている．（[第2図] 参照）

[註] a_i 及び b_i を推定するのに用いた式は次の通りである．

$$\hat{a}_i=\frac{M_i}{m_i}\sum_{j=1}^{m_i}N_{i(j)}S_{ij}^2=r\sum_{j=1}^{m_i}n_{ij}S_{ij}^2$$

$$\hat{b}_i=M_i r_{i1}^2 S_{bi}^2-r(r_{i1}-1)\sum_{j=1}^{m_i}n_{ij}S_{ij}^2$$

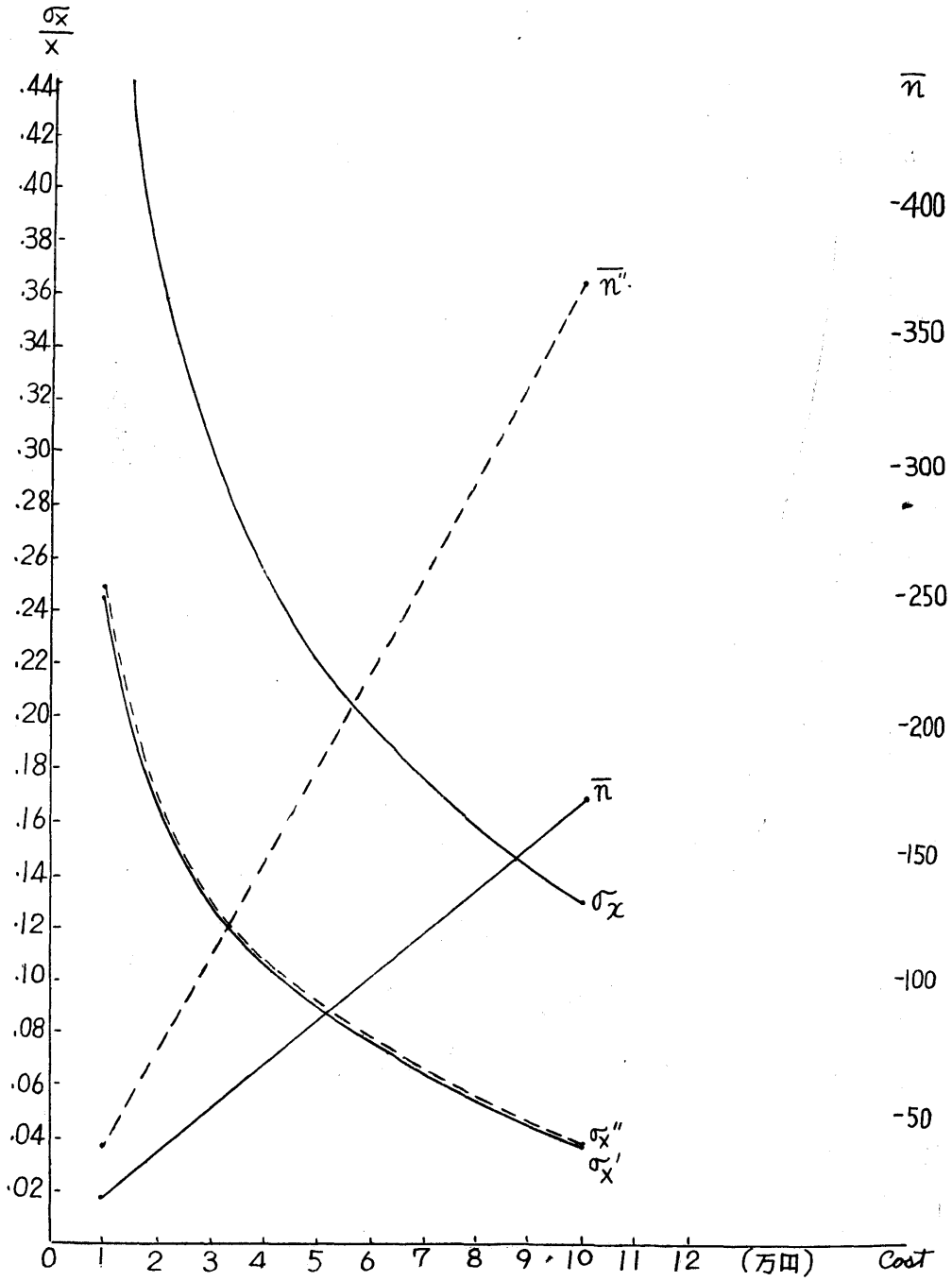
但し

$$S_{ij}^2=\frac{1}{n_{ij}-1}\sum_{k=1}^{n_{ij}}(x_{ijk}-\bar{x}_{ij})^2,\quad \bar{x}_{ij}=\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}x_{ijk}$$

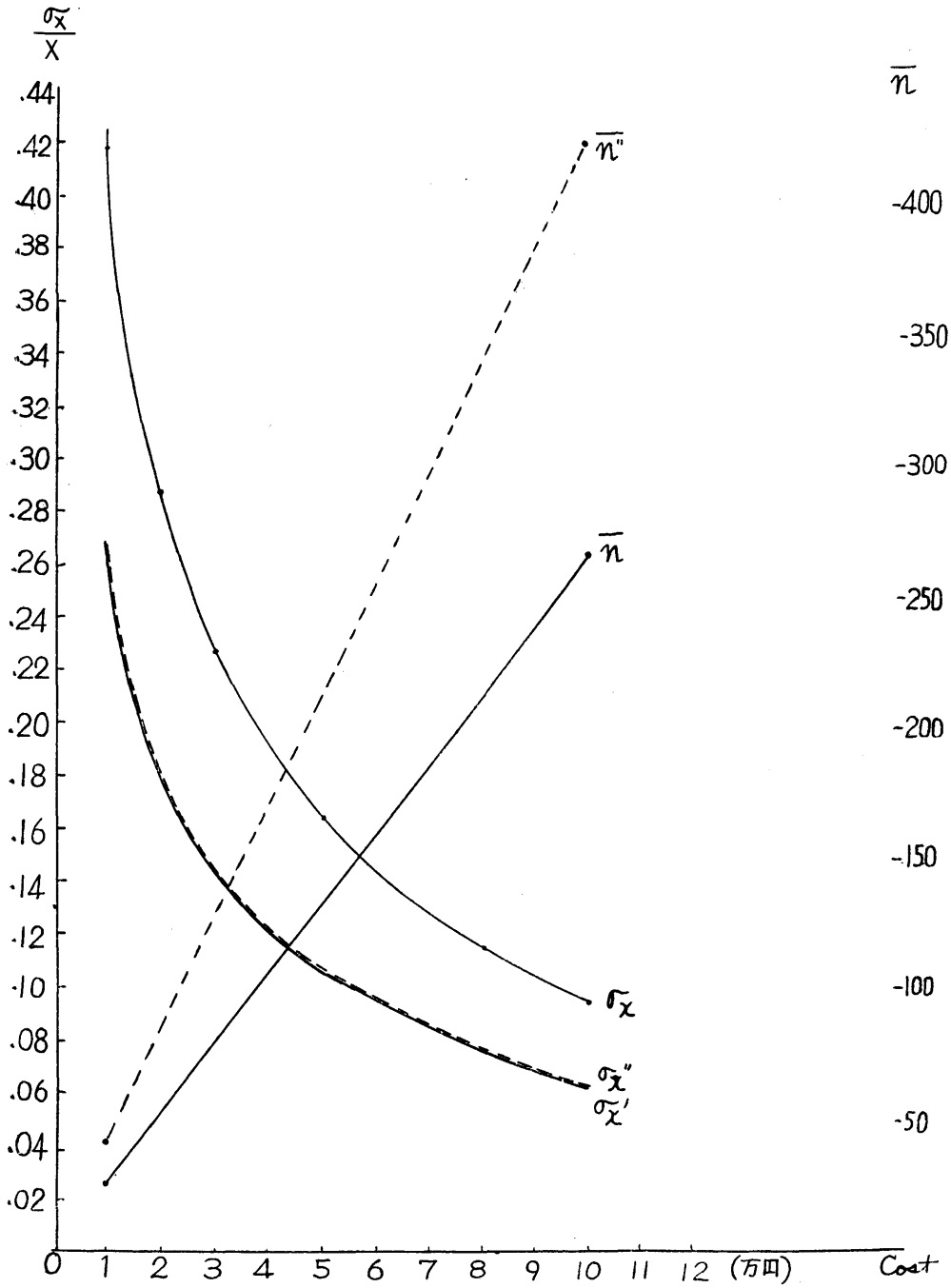
$$S_{bi}^2=\frac{1}{m_i}\sum_{j=1}^{m_i}(x_{ij}-\bar{x}_i)^2,\quad x_i=\frac{1}{m_i}\sum_{j=1}^{m_i}x_{ij}$$

参考文献

1) 林知己夫：サンプリング調査はどう行うか，東京大学出版部，1951.
2) 林知己夫，丸山文行：ある層化法について，講究録 4-10，1949.
3) C. Hayashi, F. Maruyama, M.D. Ishida: On Some Criteria for Stratification, A. I. S. M., vol. II, No. 2.
4) Y. Taga: On Optimum Balancing between Sample Size and Number of Strata in Sub-sampling, A. I. S. M. vol IV, No. 2, 1953.



［第 1 図］　副次抽出法に於ける調査費用と相対精度（及びサンプル数）の関係 I ── 書籍売上高について



［第 2 図］ 副次抽出法に於ける調査費用と相対精度（及びサンプル数）の関係II——文庫率について