二重標本抽出에서의 危險率에 의한 두가지 要因의 比較調査

金 益 贊*

Two-Factor Comparative Surveys with Risks in Double Sampling

Kim Ik-chan

요 약

二重 標本抽出에서의 最適 設計로서 2 × 2 table로 주어지는 두가지 要因을 比較分析함에 있어서, 비용을 一般化한 危險率이 주어질때 이들 要因들의 同一한 精度를 最大化하기 爲한 標本의 크기 및 最小 分散을 算出하였다.

1. Double sampling and required precision

In many sample surveys the principle objective is to compare several sectors of a finite population. Specially, there may be several factors of interest and each of these factors may have been divided into several categories.

If the elements (N_{ij}) represented by the cells in a 2×2 table are not identifiable in advance, one cannot sample independently in each of them. However, one may select a large preliminary sample (n') and identify the subpopulation to which each sampled element (n'_{ij}) belongs. Then, for each sub-population, a sub-sample (n_{ij}) is selected for analytical surveys. Such a double sampling procedure is useful if the risk of identifying an element is small relative to the risk of securing the necessary information in the main

* 사범대학 수학교육과

survey. Now cousider the optimal design for two-factor comparative surveys. The two factor α and τ are represented by a 2×2 table with (i, j)th cell denoting ith category of α and jth category of τ.

The two categories for each factor are compared by considering

$$D\alpha = W_{.1}(\overline{Y}_{11} - \overline{Y}_{21}) + W_{.2}(\overline{Y}_{12} - \overline{Y}_{22})$$
$$D\tau = W_{1.}(\overline{Y}_{11} - \overline{Y}_{12}) + W_{2.}(\overline{Y}_{21} - \overline{Y}_{22}) \qquad (1)$$

where N_{ij}= total number of units in the (i, j)th cell, $W_{ij}=N_{ij}/N$, $W_{i.}=\Sigma_j W_{ij}$, $W_{.j}=\Sigma_i W_{ij}$, $\overline{Y}_{ij}$= true mean for (i, j)th cell and $N = \Sigma_i \Sigma_j N_{ij}$, the size of population.

Using the double sampling method, the unbiased estimators are given by

$$\widehat{D}\alpha = \frac{n_{.1}'}{n'}(\bar{y}_{11} - \bar{y}_{21}) + \frac{n_{.2}'}{n'}(\bar{y}_{12} - \bar{y}_{22})$$
$$\widehat{D}\tau = \frac{n_{1.}'}{n'}(\bar{y}_{11} - \bar{y}_{12}) + \frac{n_{2.}'}{n'}(\bar{y}_{21} - \bar{y}_{22}) \qquad (2)$$

note that $n'_{i.} = \Sigma_j n'_{ij}$, $n'_{.j} = \Sigma_i n'_{ij}$ are obatin from the preliminary sample n' and sample mean $\bar{y}_{ij}$ from n_{ij}.

Let $n_{ij} = n_{ij}' \nu_{ij}$, $0 < \nu_{ij} \leq 1$ and $w_{ij} = \frac{n_{ij}'}{n'}$ then

n_{ij}, w_{ij}, $\bar{y}_{ij}$ are vandom variables and

Lemma 1. $E(n_{ij}) = E(n_{ij}'\nu_{ij}) = E(w_{ij} n' \nu_{ij}) = n' \nu_{ij} E(w_{ij}) = n' \nu_{ij} W_{ij}$

Lemma 2. $E\left(\frac{1}{n_{ij}}\right) \simeq \frac{1}{E(n_{ij})}$

Proof : See [ ]

If equal precision is desired for $\widehat{D}\alpha$ and $\widehat{D}\tau$, we use the objective function ;

$$\overline{V} = \frac{1}{2}\{V(\widehat{D}\alpha) + V(\widehat{D}\tau)\}$$
$$= \frac{1}{2}\left[E\left\{\Sigma\Sigma\left(\frac{(n_{.j}')^2 + (n_{i.}')^2}{(n')^2}\right) \cdot \frac{S_{ij}^2}{n_{ij}}\right\} + V\left\{\frac{n_{.1}'}{n'}(\overline{Y}_{11} - \overline{Y}_{21}) + \frac{n_{.2}'}{n}(\overline{Y}_{12} - \overline{Y}_{22})\right\} + V\left\{\frac{n_{1.}'}{n'}(\overline{Y}_{11} - \overline{Y}_{12}) + \frac{n_{2.}'}{n'}(\overline{Y}_{21} - \overline{Y}_{22})\right\}\right]$$

where S_{ij}^2 is the true variance in (i, j)th cell. Using lemma 1, 2 and the approximation

$$E\left\{\frac{(n_{.j}')^2+(n_{i.}')^2}{n'^2\, n_{ij}}\right\} \simeq \frac{\{E(n_{.j}')\}^2+\{E(n_{i.}')\}^2}{n'^2 E(n_{ij})} = \frac{W_{.j}^2+W_{i.}^2}{n'\, W_{ij}\, \nu_{ij}}$$

$$= \frac{\tilde{g}_{ij}}{n'\nu_{ij}}$$

then the objective function reduces to

$$\bar{V} = E\left(\Sigma\Sigma\frac{g_{ij}^2}{n_{ij}}\right) \simeq \Sigma\Sigma\frac{g_{ij}^2}{n' W_{ij}\, \nu_{ij}} \tag{3}$$

where $2g_{ij}^2 = \tilde{g}_{ij} W_{ij} S_{ij}^2$

2. Determine the risks in double sampling

Let Ω be a parameter space on random variable X and U be a numerical function defined on Ω whose value we wish to estimate on the basis of the outcome of an experiment $x \in X$.

Let A be the space of actions on real ling R^1 and a non-randomized decision function δ^* defined on X be a numerical function specifying for each x the number $a \in A$ which will be chosen to estimate U when that x is observed. Then the loss function $L(U, \delta^*)$ defined on $\Omega \times A$ is the loss incurred when U is estimated by δ^*.

If we define the loss ;

$$L(U.\delta^*) = |\delta^* - U| + \Sigma_i\Sigma_j C_{ij} n_{ij} + C' n'$$

and replace δ^* with strata mean $\bar{y}_{st} = \Sigma\,\Sigma\, w_{ij}\bar{y}_{ij}$, U with population mean $\bar{Y}$, then the risk function R is defined by

$$R(U.\delta^*) = E\,|\,\bar{y}_{st} - \bar{Y}| + \Sigma\Sigma\, C_{ij}\chi_{ij} + C'n' \tag{4}$$

where C' is the cost of classification per unit and C_{ij} the cost of measuring a unit in (i, j)th cell.

Lemma 3. An estimator $\bar{y}_{st}$ is unbiased estimator of $\bar{Y} = \Sigma\Sigma W_{ij} \bar{Y}_{ij}$

Proof: $E(\bar{y}_{st}) = E[E(\Sigma\Sigma w_{ij}\bar{y}_{ij} \mid w_{ij})] = E(\Sigma\Sigma w_{ij})E_2(\bar{y}_{ij})$

$$= E(\Sigma\Sigma w_{ij}\bar{Y}_{ij}) = \bar{Y}$$

where the subscript 2 refer to an average over all random sample of n_{ij} units that can be drawn from a given n'_{ij} units.

Now the specified cost of taking the sample is generalized by the risk R and

Theorem 1.

$$R(U, \delta^*) \leq n'MD_p + \Sigma\Sigma W_{ij}MD_{ij} \mid n_{ij} - n'_{ij} \mid + \Sigma\Sigma C_{ij}n_{ij}\nu + C'n' \qquad (5)$$

where MD_p is the true man deviation and MD_{ij} the (i, j)th cell mean deviation.

Proof : $R(U.\delta^*) = E \mid \bar{y}_{st} - \bar{Y} \mid + \Sigma\Sigma C_{ij} n_{ij} + C'n'$

$$\mid \bar{y}_{st} - \bar{Y} \mid = \mid \Sigma\Sigma w_{ij}\bar{y}_{ij} + \bar{Y} \mid$$

$$= \mid \Sigma\Sigma w_{ij}\bar{y}_{ij}' + \Sigma\Sigma w_{ij}(\bar{y}_{ij} - \bar{y}_{ij}') - \bar{Y} \mid$$

$$\leq \mid \Sigma\Sigma w_{ij}\bar{y}_{ij}' - \bar{Y} \mid + \mid \Sigma\Sigma w_{ij}(\bar{y}_{ij} - \bar{y}_{ij}') \mid$$

And Since

$$MD_p = \frac{1}{N}\Sigma_i\Sigma_j \mid y_{ij} - \bar{Y} \mid \quad \text{and} \quad E \mid y_{ij} - \bar{Y} \mid \leq \frac{n'}{N}\Sigma\Sigma \mid y_{ij} - \bar{Y} \mid$$

$$= n' \cdot MD_p$$

So $E_2 \mid \Sigma\Sigma w_{ij}(\bar{y}_{ij} - \bar{y}_{ij}') \mid = \Sigma\Sigma W_{ij} E_2 \mid (\bar{y}_{ij} - \bar{Y}_{ij}) - (\bar{y}_{ij}' - \bar{Y}_{ij}) \mid$

$$= \Sigma\Sigma W_{ij} \mid E_2(\bar{y}_{ij} - \bar{Y}_{ij}) - E_2(\bar{y}_{ij}' - \bar{Y}_{ij}) \mid$$

$$\leq \Sigma\Sigma W_{ij} \mid n_{ij} MD_{ij} - n_{ij}' MD_{ij} \mid$$

Therefore $E(\bar{y}_{st} - \bar{Y}) \leq n' MD_p + \Sigma\Sigma W_{ij} MD_{ij} \mid n_{ij} - n_{ij} \mid$

This completes the proof.

Theorem 2.

Let the expected risk be $R^* = \{R(U.\delta^*)\}$, then

$$R^* \leq n'B + n'\Sigma\Sigma v_{ij} W_{ij} D_{ij} \qquad (6)$$

where $B = C' + MD_p + \Sigma\Sigma W_{ij}^2 MD_{ij}$ and $D_{ij} = C_{ij} - W_{ij} MD_{ij}$

Proof : since $E(\Sigma\Sigma W_{ij} MD_{ij} \mid n_{ij} - n_{ij}' \mid) = \Sigma\Sigma W_{ij} MD_{ij}\ E(n_{ij} \mid 1 - \frac{1}{\nu_{ij}} \mid)$

$$= \Sigma\Sigma W_{ij}\ MD_{ij}\ n' \nu_{ij} W_{ij} \mid \frac{1}{\nu_{ij}} - 1 \mid$$

$$= \Sigma\Sigma n'\ MD_{ij}\ W_{ij}^2\ (1 - \nu_{ij})$$

Hence

$$R^* = E(R) \le E(n'MD_p) + E(\Sigma\Sigma W_{ij} MD_{ij} \mid n_{ij} - n'_{ij} \mid) + n'\Sigma\Sigma C_{ij} \nu_{ij} W_{ij} + C'n'$$

$$= n'MD_p + n'\Sigma\Sigma MD_{ij} W_{ij}^2 (1 - \nu_{ij}) + n'\Sigma\Sigma C_{ij} \nu_{ij} W_{ij} + C'n'$$

$$= n'(C' + MD_p + \Sigma\Sigma W_{ij}^2 MD_{ij}) + n'\Sigma\Sigma \nu_{ij} W_{ij}(C_{ij} - W_{ij} MD_{ij})$$

$$= n'B + n'\Sigma\Sigma \nu_{ij} W_{ij} D_{ij}$$

3. Optimum design for two factor comparative surveys with specfied risk

We consider the optimum design to find those values of the preliminary sample size n′ and main sample size n which maximize, for a given risk, the equal precision of comparisons of two-factor with categories.

Without loss the generality, we can assume that inequality in (6) change to equality, therefore

$$R^* = n'B + n'\Sigma\Sigma v_{ij} W_{ij} D_{ij} \tag{7}$$

Let find the values of n′ and ν_{ij} which maximize (3)

$$\bar{V} = \Sigma\Sigma \frac{a_{ij}}{n' W_{ij} \nu_{ij}}$$

subject to (7) and $O < \nu_{ij} \le 1$, where $a_{ij}^2 = 2g_{ij}^2$ are known constants. We determine first the optimal ν_{ij} for a given n′ and then the optimal n′.

By Cauchy inequality ;

$$\Sigma\alpha_h^2\ \Sigma\beta_h^2 - (\Sigma\alpha_h\beta_h)^2 = \sum_{ij>i}\sum (\alpha_i\beta_j - \alpha_j\beta_i)^2$$

then $(\Sigma\alpha_h)^2(\Sigma\beta_h)^2 \geq (\Sigma\alpha_h\beta_h)^2$

And if $\frac{\beta_1}{\alpha_1} = \frac{\beta_2}{\alpha_2} = \cdots = \frac{\beta_2}{\alpha_2} =$ constant the

equality holds

Now let $R' = R^* - n'B = \Sigma\Sigma n'\nu_{ij} W_{ij} D_{ij}$, (8)

then the preduct $R'\overline{V} = \left(\Sigma\Sigma \frac{a_{ij}^2}{n'W_{ij}\nu_{ij}'}\right)\left(\Sigma\Sigma n'\ \nu_{ij} W_{ij} D_{ij}\right)$

Using the Cauchy inequalits,

Put $\alpha_h = \frac{a_{ij}}{\sqrt{n'W_{ij}\nu_{ij}}}$, $\beta_h = \sqrt{n'\nu_{ij}W_{ij}D_{ij}}$

then $\frac{\beta_h}{\alpha_h} = \frac{n'W_{ij}\nu_{ij}\sqrt{D_{ij}}}{a_{ij}}$ and

$$\frac{n'W_{11}\nu_{11}\sqrt{D_{11}}}{a_{11}} = \frac{n'W_{12}\nu_{12}\sqrt{D_{12}}}{a_{12}} = \cdots = \frac{\Sigma\Sigma n'W_{ij}\nu_{ij}D_{ij}}{\Sigma\Sigma a_{ij}\sqrt{D_{ij}}}$$

$$= \frac{R^* - B}{\Sigma\Sigma a_{ij}\sqrt{D_{ij}}}$$

Hence $\frac{n'W_{ij}\nu_{ij}\sqrt{D_{ij}}}{a_{ij}} = \frac{R^* - B}{\Sigma\Sigma a_{ij}\sqrt{D_{ij}}}$

So the optimal ν_{ij} for fixed n′ is given by

$$n'W_{ij}\nu_{ij} = \frac{a_{ij}(R^* - n'B)}{\sqrt{D_{ij}}\ \Sigma\Sigma a_{ij}\sqrt{D_{ij}}} \quad (9)$$

Provided $n'W_{ij}\nu_{ij} \leq n'W_{ij}$ for all i, j; that is

$$\frac{a_{ij}(R^* - nB')}{\sqrt{D_{ij}}\ \Sigma\Sigma a_{ij}\sqrt{D_{ij}}} \leq n'W_{ij}$$

Hence $n' \geq \dfrac{R^*}{B + W_{ij}\cdot\dfrac{\sqrt{D_{ij}}}{a_{ij}}\Sigma\Sigma a_{ij}\sqrt{D_{ij}}}$; i, j = 1, 2

$$= \left[B + W_{(1,1)}\cdot\frac{\sqrt{D_{(1,1)}}}{a_{(1,1)}}\cdot\Sigma\Sigma a_{ij}\sqrt{D_{ij}}\right]^{-1}\cdot R^* \equiv m_{11}' \quad (10)$$

where (1, 1) denotes the group with the smallest value of $W_{ij}\sqrt{D_{ij}}/a_{ij}$

The minimum value of $\bar{V}$ for $n' \geq m_{11}'$ after substituting the optimul ν_{ij} in (3), is given by

$$\bar{V}_{11}(n') = \frac{(\Sigma\Sigma\, a_{ij}\sqrt{D_{ij}})^2}{R^* - n'B} \tag{11}$$

so that the minimum occurs at the value $m_{11} = m_{11}'$.

Note that $\nu_{(1,1)} = 1$ when $n' = m_{11}$

To examine values of n′ smaller than m_{11}, set $\nu_{11} = 1$ and use the Cauchy inequality to obtain the remaining ν_{ij}.

This gives

$$n' W_{ij}\nu_{ij} = \frac{a_{ij}}{\sqrt{D_{ij}}}\left\{\left(\frac{R^* - n'B - n'W_{(1,1)}\cdot D_{(1,1)}}{\Sigma\Sigma\, a_{ij}\ \sqrt{D_{ij}}}\right)\Big/ \sum\sum_{(1)(1)} a_{ij}\ \sqrt{D_{ij}}\right\} \quad (i,j) \neq (1,1) \tag{12}$$

Provided

$$n' \geq \left\{B + D_{(1,1)}W_{(1,1)} + \left(W_{(1,2)}\cdot\frac{\sqrt{D_{(1,2)}}}{a_{(1,2)}}\right)\sum\sum_{(1)(1)} a_{ij}\sqrt{D_{ij}}\right\}^{-1}\cdot R^* \equiv m_{12}'$$

where Σ denotes the summation over *i, j* ≠ (1), and (1, 2) denotes the group with the second smallest values of $W_{ij}\sqrt{D_{ij}}\,/\,a_{ij}$

Therefore, the minimum value of $\bar{V}$, for n′ in the range $m_{12}' \leq n' \leq m_{11}'$ is given by

$$\bar{V}_{12}(n') = \frac{a^2_{(1,1)}}{n'W_{(1,1)}} + \frac{(\sum\sum_{(1)(1)} a_{ij}\sqrt{D_{ij}})^2}{R^* - n'(B+D_{(1,1)}\cdot W_{(1,1)})} \tag{13}$$

From this $\bar{V}_{12}$, to find the optimal n′ ovev the range

$m_{12}' \leq n' \leq m_{11}'$, we Put $\dfrac{d\bar{V}_{12}(n')}{dn'} = 0$, then

$$n' = \left[B+D_{(1,1)}\cdot W_{(1,1)} + \frac{\{W_{(1,2)}(B+D_{(1,1)}\cdot W_{(1,1)})\}^{\frac{1}{2}}}{a_{(1,1)}}\cdot\sum\sum_{(1)(1)} a_{ij}\cdot\sqrt{D_{ij}}\right]^{-1}\cdot R^* \tag{14}$$

If $d\bar{V}_{12}(n')/dn'$ does not vanish for $n' \geqq m_{12}'$, we need to see ν'_{11}, ν_{12}, and so forth, = 1 in turn until the turning point of $\bar{V}$ is found, and note that $\bar{V}(n')$ has a unique minimum.

Literature Cited

(1) Aggarwal, O. M. 1960. Bayes and minimax procedures in sampling from finite and infinite population- I , American Statistical Association Journal 206~218.

(2) Booth, G and Sedransk, J. 1969. Planing some two-factor comparative surveys, American Statistical Association Journal 64, 560~573.

(3) Cochran, W. G. 1977. Sampling Techniques 3rd ed, John Wiley and Sons, New-York.

(4) Rao, S. N. K. 1973. On double sampling for stratification and analytical surveys, Biometrica 60, 125~133.

(5) Sedransk, J. 1965. A double sampling scheme for analytical surveys, American Statistical Association 60, 985~1004.