UIT Report Vol. 12, No. 1, pp. 53~57, 1981. 연구논문집 제12권 제 1 호 pp. 53~57, 1981

Note on a Comparison of Stratified with Simple Random Sampling in the Estimation of Proportions

Yeo, Sung-Chil
Dept. of Industrial Engineering
(Received November 10, 1980)

〈**Abstract**〉

With intelligent use, stratified random sampling nearly always gives a more precise estimate than simple random sampling. But it is not true that any stratification gives a more efficiency than a simple random sample.

On the other hand, the estimation of population proportions are often required in the field of survey sampling. In this paper a comparison is made between simple random sampling and stratified random sampling with proportional and optimum allocation in the estimation of proportions.

比率의 推定에서 層化와 單純確率抽出의 比較에 관한 硏究

余 成 七
産業工學科
(1980. 11. 10 접수)

〈**要 約**〉

單純確率抽出과 比例 및 最適割當에 의한 層化確率抽出에서 比率의 推定이 요구될때 구해진 推定量들 사이에 有効性을 비교하였다.

Ⅰ. Introduction

Consider how much stratified random sampling results in a more precise estimate than simple random sampling.(1) If adequate stratification and appropriate allocation are used for the sample, stratified random sampling nearly always results in a more efficiency than is given by a comparable simple random sampling. It is not true, however, that any stratification gives a smaller variance than a simple random sampling. If the number of units in sample within strata are far from optimum, stratified sampling may have a larger variance. In fact, even stratification with optimum allocation for fixed total sample size may give a larger variance, though this result is an academic curriosity rather than something likely to happen in practice.

On the other hand, we sometimes wish to estimate the proportions of units in the population that possess some characteristic or attribute or fall into some defined class. Sampling for proportions is widely used in market research, public opinion surveys, and quality control. In this paper we give a comparison between simple random sampling and stratified random sampling with proportional and optimum allocation in

(1) Interesting discussions of this question were provided by Armitage [1] and Evans [3].

the estimation of proportions. This comparison shows how the gain due to stratification is achieved. For simple random, proportional and optimum stratified random sampling, the variances of the estimated proportions are denoted by V_{ran}, V_{prop}, and V_{opt}, respectively.

Ⅱ. Notation

We introduce the notations that are used in this paper. Here the suffix h the stratum and i the unit within the stratum.

N is the total number of units in the population.

L is the number of strata.

n is the fixed total size of sample.

$f=\frac{n}{N}$ is the sampling fraction in the population.

$A=\sum^{N} X_i$ is the number of units in some defined class C in the population.

$a=\sum^{n} x_i$ is the number of units in C in the sample.

$P=\frac{A}{N}$ is the proportion of units in C in the population.

$p=\frac{a}{n}$ is the proportion of units in C in the sample.

N_h is the total number of nuits in the hth stratum

n_h is the number of units in the sample of the hth stratum.

$W_h=\frac{N_h}{N}$ is the stratum weight in the hth stratum.

$f_h=\frac{n_h}{N_h}$ is the sampling fraction in the hth stratum.

$A_h=\sum_{i}^{N_h} X_{hi}$ is the number of units in C in the hth stratum.

$a_h=\sum_{i}^{n_h} x_{hi}$ is the number of units in C in the sample of the hth stratum.

$P_h=\frac{A_h}{N_h}$ is the proportion of units in C in the hth stratum.

$p_h=\frac{a_h}{n_h}$ is the proportion of units in C in the sample of the hth stratum.

Ⅲ. Comparison between simple and stratified random sampling for proportions.

In this section we have ignored $1/N_h$ and hence $1/N$ since in nearly all applications, even if the fpc is not negligible, terms in $1/N_h$ are negligible.

1. Comparison between simple random and proportional stratified random sampling.

To compare the efficiencies of these procedures, we need to compare the variances of the estimators obtained by the two different methods.

For simple random sampling, the variance of sample mean $\bar{x}$ is

$$V(\bar{x})=(1-f)\frac{S^2}{n}, \tag{3.1}$$

where $S^2=\frac{1}{N-1}\sum^{N}(X_i-\bar{X})^2$ is the population variance and where $\bar{X}$ is the population mean.

To find the result for proportions, we define X_i as 1 if the unit is in some defined class C and as 0 if it is in C'. Then for the population, we have

$$\bar{X}=\frac{A}{N}=P \quad . \tag{3.2}$$

Also, for the sample,

$$\bar{x}=\frac{a}{n}=p \quad . \tag{3.3}$$

Note that $\sum^{N} X_i^2=A=NP$, and hence

$$S^2=\frac{1}{N-1}(\sum^{N} X_i^2-N\bar{X}^2)=\frac{N}{N-1}PQ, \tag{3.4}$$

where $Q=1-P$.

With the term $1/N$ negligible, the slightly simpler formula for S^2 is $S^2=PQ$.

Hence for simple random sampling, the variance of sample proportion p is

$$V_{ran}=(1-f)\frac{PQ}{n}. \tag{3.5}$$

On the other hand, for stratified random sampling, the variance of sample mean $\bar{x}_{st}$ is

$$V(\bar{x}_{st})=\Sigma W_h^2\frac{(1-f_h)}{n_h}S_h^2, \tag{3.6}$$

or equivalently

$$V(\bar{x}_{st}) = \Sigma \frac{W_h^2 S_h^2}{n_h} - \Sigma \frac{W_h^2 S_h^2}{N_h}, \qquad (3.7)$$

where $S_h^2 = \frac{1}{N_h - 1} \sum_i^{N_h} (X_{hi} - \bar{X}_h)^2$ is the stratum variance, and where $\bar{X}_h = \frac{1}{N_h} \sum_i^{N_h} X_{hi}$ is the stratum mean.

To find the result for proportions, let X_{hi} be a variate which has the value 1 when the unit is in C, and zero otherwise.

Then for the subpopulations in the hth stratum, we have

$$\bar{X}_h = \frac{A_h}{N_h} = P_h. \qquad (3.8)$$

Also, for the sample in the hth stratum.

$$\bar{x}_h = \frac{a_h}{n_h} = p_h. \qquad (3.9)$$

Thus for the proportions in the whole population, the estimate appropriate to stratified random sampling is

$$p_{st} = \Sigma \frac{N_h p_h}{N}. \qquad (3.10)$$

Note that $\sum_i^{N_h} X_{hi}^2 = A_h = N_h P_h$, and hence

$$S_h^2 = \frac{1}{N_h - 1} \left(\sum_i^{N_h} X_{hi}^2 - N_h \bar{X}_h^2 \right)$$

$$= \frac{N_h}{N_h - 1} P_h Q_h, \qquad (3.11)$$

where $Q_h = 1 - P_h$.

If terms in $1/N_h$ are negligible, the slightly simpler formula for S_h^2 is $S_h^2 = P_h Q_h$.

Hence for stratified random sampling, the variance of sample proportion p_{st} is

$$V(p_{st}) = \Sigma W_h^2 \frac{(1 - f_h)}{n_h} P_h Q_h. \qquad (3.12)$$

To obtain the result for proportional allocation, we substitute $n_h = W_h \cdot n$ into the equation (3.12). (2)

Therefore we have

$$V_{\text{prop}} = \frac{(1 - f)}{n} \Sigma W_h P_h Q_h. \qquad (3.13)$$

Note that the relation between the stratum variances and the whole population variance is as follows:

$$(N-1)S^2 = \Sigma (N_h - 1) S_h^2 + \Sigma N_h (\bar{X}_h - \bar{X})^2 \text{ (3)} \qquad (3.14)$$

The first term on the right hand side shows the stratum variances and the second term shows the disperson due to variation among stratum means.

To find the result for proportions, if we substitute the equations (3.4) and (3.11) into the both sides of the equation (3.12), respectively, then the equation (3.14) reduces to

$$PQ = \Sigma W_h P_h Q_h + \Sigma W_h (P_h - P)^2. \qquad (3.15)$$

From the equation (3.15), we obtain that

$$V_{\text{ran}} = V_{\text{prop}} + \frac{(1 - f)}{n} \Sigma W_h (P_h - P)^2. \qquad (3.16)$$

Since $\Sigma W_h (P_h - P)^2$ is the variance between the stratum proportions and the population proportion, the precision of proportional stratified random sampling is greater when the variation between strata is large.

2. Comparison between proportional and optimum stratified random sampling

From the equation (3.13), equivalently we have

$$V_{\text{prop}} = \frac{1}{n} \Sigma W_h P_h Q_h - \frac{1}{N} \Sigma W_h P_h Q_h. \qquad (3.17)$$

On the other hand, in the optimum allocation for a fixed total size of sample n, the variance of sample mean $\bar{x}_{st}$ is minimized when

$$n_h = \frac{N_h S_h}{\Sigma N_h S_h} n. \qquad (3.18)$$

To find $V_{\min}(\bar{x}_{st})$ for stratified random sampling with optimum allocation for a fixed n (4), we substitute the equation (3.18) into the equation (3.7). Hence we have

$$V_{\min}(\bar{x}_{st}) = \frac{1}{n} (\Sigma W_h S_h)^2 - \frac{1}{N} \Sigma W_h S_h^2. \qquad (3.19)$$

Since $S_h^2 = P_h Q_h$, in terms of proportions, the

(2) For the proportional allocation, we have $n/N = n_h/N_h$, that is, $f = f_h$.
(3) See Cochran [2] p. 100. Where the first term on the right hand side is sometimes called the within variance because it shows the variation within each stratum, and the second term is called the between variance because it shows the variation between strata.
(4) This allocation is sometimes called Neyman allocation.

equation (3.19) reduces to

$$V_{opt}=\frac{1}{n}(\Sigma W_h\sqrt{P_hQ_h})^2-\frac{1}{N}\Sigma W_hP_hQ_h^{(5)}. \tag{3.20}$$

By the equations (3.17) and (3.20), their difference is

$$V_{prop}-V_{opt}=\frac{1}{n}\left[\Sigma W_hP_hQ_h-(\Sigma W_h\sqrt{P_hQ_h})^2\right]. \tag{3.21}$$

Now let $\sqrt{\overline{P_hQ_h}}=\Sigma W_h\sqrt{P_hQ_h}$.

Then we have

$$\begin{aligned}&\Sigma W_hP_hQ_h-(\Sigma W_h\sqrt{P_hQ_h})^2\\&=\Sigma W_hP_hQ_h-2\Sigma W_h\sqrt{P_hQ_h}\sqrt{\overline{P_hQ_h}}\\&\quad+\Sigma W_h\overline{P_hQ_h}\\&=\Sigma W_h(\sqrt{P_hQ_h}-\sqrt{\overline{P_hQ_h}})^2.\end{aligned}$$

Hence the equation (3.21) becomes

$$V_{prop}-V_{opt}=\frac{1}{n}\left[\Sigma W_h(\sqrt{P_hQ_h}-\sqrt{\overline{P_hQ_h}})^2\right]. \tag{3.22}$$

or equivalently

$$V_{prop}=V_{opt}+\frac{1}{n}\left[\Sigma W_h(\sqrt{P_hQ_h}-\sqrt{\overline{P_hQ_h}})^2\right]. \tag{3.23}$$

The term $\sqrt{\overline{P_hQ_h}}$ may be considered a weighted average of $\sqrt{P_hQ_h}$ (the stratum standard deviation for proportions). Hence the term $\Sigma W_h\times(\sqrt{P_hQ_h}-\sqrt{\overline{P_hQ_h}})^2$ may be considered as showing the dispersion of $\sqrt{P_hQ_h}$ around its mean $\sqrt{\overline{P_hQ_h}}$. The greater the scatter of $\sqrt{P_hQ_h}$ around $\sqrt{\overline{P_hQ_h}}$, the larger the term $(\sqrt{P_hQ_h}-\sqrt{\overline{P_hQ_h}})^2$, and hence the greater the difference between V_{prop} and V_{opt}.

Ⅳ. Concluding remarks

From the equations (3.16) and (3.23) with terms in $1/N_h$ negligible, we have

$$V_{ran}=V_{opt}+\frac{1}{n}\left[\Sigma W_h(\sqrt{P_hQ_h}-\sqrt{\overline{P_hQ_h}})^2\right]+\frac{(1-f)}{n}\Sigma W_h(P_h-P)^2. \tag{4.1}$$

In the equation (4.1), there are two components of the decrease in variance as we change from simple random sampling to optimum allocation for a fixed sample of size n. The first component (term on the extreme right) comes from the elimination of differences among the stratum proportions, the second (middle term on the right) comes from elimination of the effect of differences among the stratum standard deviation for proportions.

On the other hand, if terms in $1/N_h$ are not negligible, we substitute the equations (3.4) and (3.11) into the equations (3.1) and (3.6), respectively. Then we have

$$V_{ran}=\frac{N-n}{N-1}\cdot\frac{PQ}{n}, \tag{4.2}$$

and

$$V(p_{st})=\Sigma W_h^2\frac{(1-f_h)}{n_h}\frac{N_h}{N_h-1}P_hQ_h. \tag{4.3}$$

If we substitute $n_h=W_h\cdot n$ into the equation (4.3), then we see that

$$V_{prop}=\frac{(1-f)}{n}\Sigma W_h\frac{N_h}{N_h-1}P_hQ_h. \tag{4.4}$$

By the equation (3.15), we have

$$V_{ran}=\frac{(1-f)}{n(N-1)}\left[\Sigma N_hP_hQ_h+\Sigma N_h(P_h-P)^2\right]. \tag{4.5}$$

Therefore the difference between V_{ran} and V_{prop} is

$$\begin{aligned}&V_{ran}-V_{prop}\\&=\frac{(1-f)}{n(N-1)}\left[\Sigma N_h(P_h-P)^2-\frac{1}{N}\Sigma(N-N_h)\times\frac{N_h}{N_h-1}P_hQ_h\right].\end{aligned} \tag{4.6}$$

From the above equation (4.6), proportional stratification gives a larger variance than simple random sampling if

$$\Sigma N_h(P_h-P)^2<\frac{1}{N}\Sigma(N-N_h)\frac{N_h}{N_h-1}P_hQ_h. \tag{4.7}$$

Mathematically, this can happen. Suppose that

(5) Where the formula (3.19) or (3.20) has a limited range which gives a positive value for V_{opt}, however, covers nearly all applications.

$$\frac{N_h}{N_h-1}P_hQ_h=\Sigma W_hP_hQ_h \text{(6)}$$

for all h, so that proportional allocation is optimum with a fixed sample of size n. Then the inequality (4.7) becomes

$$\Sigma N_h(P_h-P)^2<(L-1)\Sigma W_hP_hQ_h,$$

or

$$\frac{1}{L-1}\Sigma N_h(P_h-P)^2<\Sigma W_hP_hQ_h. \quad (4.8)$$

In the sense of the analysis of variance, the above inequality (4.8) implies that the mean square among strata is smaller than the mean square within strata.

—References—

1. Armitage, P., A comparison of stratified with unrestricted random sampling from a finite population, *Biometrika*, 34 : 273—280 (1947).
2. Cochran, W.G., Sampling Techniques, Third Edition, Wiley: New York, 1977.
3. Evans, W.D., On stratification and optimum allocations, *Jour. Amer. Stat. Assoc.*, 46 : 95—104(1951).

(6) This means that the stratum variances S_h^2 are all equal to the within variance.