ON THE ESTIMATION OF THE RELATIVE EFFICIENCY OF SAMPLING PROCEDURES

BY J. N. K. RAO

(Received March 19, 1962)

1. Introduction

There are several sampling methods available to increase the precision of the estimator of the population total or mean. When a supplementary variable, x, which is correlated with the characteristic of interest, y, is available for all the N units in the population, one can use any of the following methods: (1) Draw a simple random sample of n units and use a ratio estimator utilizing x. (2) Stratify the population according to x and draw a stratified random sample. (3) Draw a sample of n units with probabilities proportional to sizes x (p.p.s.) with replacement (p.p.s. sampling without replacement is not considered here). Cochran [1] and Des Raj [2] have investigated the conditions on the relative accuracy of these procedures, assuming certain models. However, in practice these results are only of limited use in deciding the sampling procedure, since it may not be possible to verify these conditions.

Now, suppose the researcher has chosen a sampling procedure from whatever knowledge he has and draws the sample accordingly. Then it would be of interest to the researcher, for future guidance, to estimate the amount of gain or loss in efficiency he would obtain if he had used a different sampling procedure. Cochran ([1], p. 97) estimates the variance of the estimator in simple random sampling from a stratified random sample and thus estimates the efficiency of stratification. Sukhatme ([4], p. 135) extends this to p.p.s. sampling with replacement. Mokashi [3] estimates the variance of the ratio estimator in simple random sampling from a stratified random sample. Therefore, his result can be used to estimate the variance in method (1) from a sample drawn by method (2).

The purpose of the present paper is to consider the above three methods in all possible ways and estimate the variance of the estimator in one method from a sample drawn by another method. The criterion of *unbiasedness* is used in finding these estimated variance (except, when estimating the variance of the ratio estimator, we use the usual procedure of substituting sample estimates for each term in the variance formula). The precision of these estimated variances is not investigated.

2. Method (1) vs method (2)

First we consider a simple method of deriving Cochran's result. Let there be L strata with N_t units in the tth stratum ($\sum_1^L N_t=N$) and n_t is the number of units in the sample from the tth stratum ($\sum_1^L n_t=n$). Then it is well known that

$$\hat{Y}_{st}=\sum_1^L N_t\bar{y}_t\,, \tag{1}$$

where $\bar{y}_t$ is the sample mean for the tth stratum, is an unbiased estimate of the population total $Y=\sum_1^L N_t\bar{Y}_t$ and that the estimator of the variance of $\hat{Y}_{st}$ is

$$v(\hat{Y}_{st})=\sum_1^L N_t^2\left(\frac{1}{n_t}-\frac{1}{N_t}\right)s_t^2 \tag{2}$$

where s_t^2 is the sample mean square of the y's for the tth stratum. We are interested in estimating, from the stratified random sample, the variance of the estimate in simple random sampling, namely

$$V(N\bar{y})=\frac{N(N-n)}{n(N-1)}\left(\sum_1^L\sum_1^{N_t}y_{tj}^2-Y^2/N\right) \tag{3}$$

where $\bar{y}$ is the sample mean. Since

$$E\,\frac{N_t}{n_t}\sum_1^{n_t}y_{tj}^2=\sum_1^{N_t}y_{tj}^2 \tag{4}$$

and

$$\text{est } Y^2=\hat{Y}_{st}^2-v(\hat{Y}_{st})\,, \tag{5}$$

it follows that

$$\text{est } V(N\bar{y})=\frac{N(N-n)}{n(N-1)}\left[\sum_1^L\frac{N_t}{n_t}\sum_1^{n_t}y_{tj}^2-\frac{1}{N}\{\hat{Y}_{st}^2-v(\hat{Y}_{st})\}\right]. \tag{6}$$

It is easily verified that (6) is equivalent to the expression given in Cochran [1], namely

$$\text{est } V(N\bar{y})=\frac{N^2(N-n)}{n(N-1)}\left[\sum_1^L W_ts_t^2-\sum_1^L W_ts_t^2/n_t+\sum_1^L W_ts_t^2/n_t\right.$$
$$\left.-\sum_1^L W_ts_t^2/N+\sum_1^L W_t\bar{y}_t^2-(\sum_1^L W_t\bar{y}_t)^2\right] \tag{7}$$

where $W_t=N_t/N$. For computational purposes the form (6) seems to be

more convenient, since $\hat{Y}_{st}$ and $v(\hat{Y}_{st})$ are computed beforehand.

Now consider the converse problem: To estimate the variance of the estimator in method 2, namely

$$V(\hat{Y}_{st}) = \sum_1^L N_t^2\left(\frac{1}{n_t}-\frac{1}{N_t}\right)S_t^2 \tag{8}$$

where S_t^2 is the population mean square of the y's in the tth stratum, from a simple random sample of n units. Let n_t' denote the number of units in the simple random sample that belong to the tth stratum. Note that n_t' is a random variable and $\sum_1^L n_t'=n$. Define the new variables ${}_ty_i$ by

$${}_ty_i=y_i \text{ if the } i\text{th unit in the population belongs to the } t\text{th stratum}$$
$$=0 \text{ otherwise.} \tag{9}$$

Then

$$E\frac{N}{n}\sum_1^n {}_ty_i^2=\sum_1^N {}_ty_i^2=\sum_1^{N_t} y_{tj}^2 . \tag{10}$$

Also, since

$$V\left(\frac{1}{n}\sum_1^n {}_ty_i\right)=E\left(\frac{1}{n}\sum_1^{n_t'} y_{tj}\right)^2-\frac{N_t^2\bar{Y}_t^2}{N^2}, \tag{11}$$

$$\text{est}\,(N_t\bar{Y}_t^2)=\frac{N^2}{N_t}\left[\left(\frac{1}{n}\sum_1^{n_t'}y_{tj}\right)^2-v\left(\frac{1}{n}\sum_1^n {}_ty_i\right)\right] \tag{12}$$

where

$$v\left(\frac{1}{n}\sum_1^n {}_ty_i\right)=\frac{(N-n)}{Nn}\frac{1}{n-1}\left[\sum_1^{n_t'} y_{tj}^2-\frac{\left(\sum_1^{n_t'} y_{tj}\right)^2}{n}\right]. \tag{13}$$

Therefore, from (10) and (12), we have

$$\text{est}\,S_t^2=\text{est}\,(\sum_1^{N_t} y_{tj}^2-N_t\bar{Y}_t^2)/(N_t-1)$$

$$=\frac{N}{(N_t-1)}\left[\sum_1^{n_t'}\frac{y_{tj}^2}{n}\left\{1+\frac{(N-n)}{N_t(n-1)}\right\}-\frac{\left(\sum_1^{n_t'} y_{tj}\right)^2}{nN_t}\frac{(N-1)}{(n-1)}\right], \tag{14}$$

and hence, from (8),

$$\text{est } V(\hat{Y}_{st})=N\sum_{1}^{L}\frac{N_t(N_t-n_t)}{n_t(N_t-1)}\left[\left\{1+\frac{(N-n)}{N_t(n-1)}\right\}\frac{\sum_{1}^{n_t'}y_{tj}^2}{n}-\frac{(N-1)}{(n-1)}\frac{\left(\sum_{1}^{n_t'}y_{tj}\right)^2}{nN_t}\right]. \tag{15}$$

We compare (15) with the large sample estimator of variance of the ratio estimator $\hat{Y}_R=(\bar{y}/\bar{x})X$, namely

$$v(\hat{Y}_R)\doteq\frac{N(N-n)}{n(n-1)}\sum_{1}^{n}\left(y_i-\frac{\bar{y}}{\bar{x}}x_i\right)^2, \tag{16}$$

to estimate the efficiency of $\hat{Y}_R$ over $\hat{Y}_{st}$. Also, it may be of interest to compare (15) with the estimator of variance of $N\bar{y}$, namely

$$v(N\bar{y})=\frac{N(N-n)}{n(n-1)}\sum_{1}^{n}(y_i-\bar{y})^2. \tag{17}$$

Example 1: Cochran ([1], p. 113) gives the data of a simple random sample of 49 large cities from the population of 196 large cities in the United States. Here y_i and x_i denote the size of the ith city (in 1000's) in 1930 and 1920, respectively. The population total, X, is known from the previous census and is equal to 22,919. Consider now the following stratification: $L=2$ and stratum 1 is composed of all cities of size less than 100 and stratum 2 consists of all cities of size 100 and larger. From the information on x_i, we have $N_1=132$ and $N_2=64$. Let $n_1=25$ and $n_2=26$. The stratification and the allocation of the sample we employ here are, obviously, not the best and are meant only for illustration. From Cochran's data it is seen that

$$n_1'=35,\ n_2'=14,\ \sum_{1}^{n_1'}y_{1j}=2496,\ \sum_{1}^{n_1'}y_{1j}^2=194{,}258,$$

$$\sum_{1}^{n_2'}y_{2j}=3768 \text{ and } \sum_{1}^{n_2'}y_{2j}^2=1{,}333{,}624.$$

Using formula (15) we find

$$\text{est } V(\hat{Y}_{st})=196\times 17788.8.$$

And from (16) and (17)

$$v(\hat{Y}_R)=196\times 1861.5 \text{ and } v(N\bar{y})=196\times 45476.6.$$

Therefore, the estimate of the per cent gain in efficiency of $\hat{Y}_{st}$ over $N\bar{y}$ is $((45476.6/17788.8)-1)\times 100=155.6\%$. On the other hand, the estimate of the per cent gain in efficiency of $\hat{Y}_R$ over $\hat{Y}_{st}$ is $((17788.8/1861.5)-1)=855.6\%$.

3. Method (1) vs method (3)

Let $p_i=x_i/X$ denote the probability for drawing ith unit in the first draw. The problem is to estimate the variance of $N\bar{y}$ (equation 3) and

$$V(\hat{Y}_R)\doteq\frac{N(N-n)}{n(N-1)}\left[\sum_1^N y_i^2-\frac{2Y\left(\sum_1^N y_ix_i\right)}{X}+\frac{Y^2}{X^2}\sum_1^N x_i^2\right] \tag{18}$$

from a *p.p.s.* sample of n units drawn with replacement. The *p.p.s.* estimator of Y is

$$\hat{Y}_{pps}=\sum^n y_i/np_i \tag{19}$$

and the estimator of variance is

$$v(\hat{Y}_{pps})=\frac{X^2}{n(n-1)}\sum^n\left(\frac{y_i}{x_i}-\frac{1}{n}\sum^n\frac{y_i}{x_i}\right)^2. \tag{20}$$

Now

$$E\sum^n\frac{y_i^2}{np_i}=\sum_1^N y_i^2\,, \tag{21}$$

and, since

$$V(\hat{Y}_{pps})=E(\hat{Y}_{pps}^2)-Y^2, \tag{22}$$

$$\text{est } Y^2=\hat{Y}_{pps}^2-v(\hat{Y}_{pps}). \tag{23}$$

Therefore,

$$\text{est } V(N\bar{y})=\frac{N(N-n)}{n(N-1)}\left[\sum^n\frac{y_i^2}{np_i}-\frac{1}{N}\{\hat{Y}_{pps}^2-v(\hat{Y}_{pps})\}\right]. \tag{24}$$

To estimate $V(\hat{Y}_R)$ we use the usual procedure of replacing each term in (18) by its unbiased estimator. Therefore,

$$\text{est } V(\hat{Y}_R)\doteq\frac{N(N-n)}{n(N-1)}\left[\sum^n\frac{y_i^2}{np_i}-\frac{2}{X}\hat{Y}_{pps}\left(\sum^n\frac{y_ix_i}{np_i}\right)\right.$$

$$\left.+\frac{(\sum^n x_i^2/np_i)}{X^2}\{\hat{Y}_{pps}^2-v(\hat{Y}_{pps})\}\right]. \tag{25}$$

It may be pointed out that sometimes it is possible to construct alternative unbiased estimators of variance. For example, since $V(N\bar{y})$ can be written as

$$V(N\bar{y})=\frac{(N-n)}{n(N-1)}\sum_{i<i'}^{N}(y_i-y_{i'})^2, \tag{26}$$

an unbiased estimator of (26) from the *p.p.s.* sample is

$$\text{est } V(N\bar{y})=\frac{(N-n)}{n(N-1)}\sum_{i<i'}^{N}\frac{t_it_{i'}}{n(n-1)p_ip_{i'}}(y_i-y_{i'})^2 \tag{27}$$

where t_i is the number of times ith unit is selected in the *p.p.s.* sample ($\sum_1^N t_i=n$) and $E(t_it_{i'})=n(n-1)p_ip_{i'}$. Note that (24) is, in general, not equal to (27). However, (27) has the advantage that it is always positive (it is not clear whether (24) is always positive).

The converse problem of estimating the variance of $\hat{Y}_{pps}$ from a simple random sample is quite simple. Since

$$V(\hat{Y}_{pps})=\sum_1^N\frac{y_i^2}{np_i}-\frac{Y^2}{n}=\sum_{i<i'}^{N}\frac{p_ip_{i'}}{n}\left(\frac{y_i}{p_i}-\frac{y_{i'}}{p_{i'}}\right)^2, \tag{28}$$

$$\text{est } V(\hat{Y}_{pps})=\frac{N(N-1)}{n(n-1)}\sum_{i<i'}^{n}\frac{p_ip_{i'}}{n}\left(\frac{y_i}{p_i}-\frac{y_{i'}}{p_{i'}}\right)^2. \tag{29}$$

One compares (29) with (16) to estimate the efficiency of $\hat{Y}_R$ over $\hat{Y}_{pps}$.

Example 2: Sukhatme ([4], p. 183) gives the data of a *p.p.s.* sample of $n=34$ villages drawn with replacement from a total of $N=170$ villages. Here y_i is the area under wheat in 1937 and x_i is the total cultivated area in 1931 for the ith village. Using (20), (24) and (25) we find that

$$v(\hat{Y}_{pps})=(170)^2\times 53.38,$$
$$\text{est } V(N\bar{y})=(170)^2\times 306.81,$$

and

$$\text{est } V(\hat{Y}_R)=(170)^2\times 53.88.$$

Therefore, the estimate of the per cent gain in efficiency of $\hat{Y}_{pps}$ over $N\bar{y}$ is $((306.81/53.38)-1)\times 100=475\%$; the estimate of the per cent gain in efficiency of $\hat{Y}_{pps}$ over $\hat{Y}_R$ is $((53.88/53.38)-1)\times 100=1\%$.

4. Method (2) vs method (3)

It is required to estimate the variance of the estimator in method 2 (equation 8) from a sample of n units drawn with *p.p.s.* and with replacement. Let n_t' denote the number of units (not distinct) in the *p.p.s.* sample that belong to the tth stratum so that $\sum_1^L n_t'=n$. Then,

defining the variable ${}_{t}y_i$ as before, we have

$$E\sum^{n}\frac{{}_{t}y_i^2}{np_i}=\sum_{1}^{N}{}_{t}y_i^2=\sum_{1}^{N_t}y_{tj}^2\,, \tag{30}$$

and, since

$$V\left(\sum^{n}\frac{{}_{t}y_i}{np_i}\right)=E\left(\sum^{n_t'}\frac{y_{tj}}{np_{tj}}\right)^2-N_t^2\bar{Y}_t^2 \tag{31}$$

where $p_{tj}=x_{tj}/X$,

$$\text{est}\,N_t\bar{Y}_t^2=\frac{1}{N_t}\left[\left(\sum^{n_t'}\frac{y_{tj}}{np_{tj}}\right)^2-v\left(\sum^{n}\frac{{}_{t}y_i}{np_i}\right)\right] \tag{32}$$

where

$$v\left(\sum^{n}\frac{{}_{t}y_i}{np_i}\right)=\frac{1}{n-1}\left[\sum^{n_t'}\frac{y_{tj}^2}{np_{tj}}-\left(\sum^{n_t'}\frac{y_{tj}}{np_{tj}}\right)^2\right]. \tag{33}$$

From (30), (32) and (33), we have

$$\begin{aligned}\text{est}\,S_t^2=\frac{1}{(N_t-1)}\Bigg[&\sum^{n_t'}\frac{y_{tj}^2}{np_{tj}}\left\{1+\frac{1}{N_t(n-1)p_{tj}}\right\}\\&-\frac{n}{N_t(n-1)}\left(\sum^{n_t'}\frac{y_{tj}}{np_{tj}}\right)^2\Bigg].\end{aligned} \tag{34}$$

Therefore,

$$\begin{aligned}\text{est}\,V(\hat{Y}_{st})=\sum_{1}^{L}\frac{N_t(N_t-n_t)}{n_t(N_t-1)}\Bigg[&\sum^{n_t'}\frac{y_{tj}^2}{np_{tj}}\left\{1+\frac{1}{N_t(n-1)p_{tj}}\right\}\\&-\frac{n}{N_t(n-1)}\left(\sum^{n_t'}\frac{y_{tj}}{np_{tj}}\right)^2\Bigg].\end{aligned} \tag{35}$$

We compare (35) with (20) to estimate the efficiency of $\hat{Y}_{pps}$ over $\hat{Y}_{st}$.

Consider now the converse problem of estimating $V(\hat{Y}_{pps})$ from a stratified random sample. We have

$$E\sum_{1}^{L}\frac{N_t}{n_t}\sum_{1}^{n_t}\frac{y_{tj}^2}{np_{tj}}=\sum_{1}^{L}\sum_{1}^{N_t}\frac{y_{tj}^2}{np_{tj}}=\sum_{1}^{N}\frac{y_i^2}{np_i} \tag{36}$$

and, since

$$V(\hat{Y}_{st})=E\left(\sum_{1}^{L}N_t\bar{y}_t\right)^2-Y^2, \tag{37}$$

$$\text{est}\,Y^2=\left(\sum_{1}^{L}N_t\bar{y}_t\right)^2-v(\hat{Y}_{st}) \tag{38}$$

where $v(\hat{Y}_{st})$ is given by (2). Therefore, from (28), we have

$$\text{est } V(\hat{Y}_{pps})=\sum_{1}^{L}\frac{N_t}{n_t}\sum_{1}^{n_t}\frac{y_{tj}^2}{np_{tj}}-\frac{1}{n}\left[\left(\sum_{1}^{L}N_t\bar{y}_t\right)^2-v(\hat{Y}_{st})\right]. \tag{39}$$

We compare (39) with (2) to estimate the efficiency of $\hat{Y}_{st}$ over $\hat{Y}_{pps}$.

IOWA STATE UNIVERSITY, AMES, IOWA

REFERENCES

[1] W. G. Cochran, *Sampling Techniques*, John Wiley and Sons, New York, 1953.

[2] Des Raj, "On the relative accuracy of some sampling techniques," *Journal of the American Statistical Association*, Vol. 53 (1958), pp. 98-101.

[3] V. K. Mokashi, "Efficiency of stratification in sub-sampling designs for the ratio method of estimation," *Journal of the Indian Society of Agricultural Statistics*, Vol. 6 (1954), pp. 77-82.

[4] P. V. Sukhatme, *Sampling Theory of Survey with Applications*, Indian Society of Agricultural Statistics, New Delhi and Iowa State College Press, Ames, Iowa, 1954.

Rresearch sponsored by the Office of Ordnance Research, U.S. Army, under Grant No. DA-ARO(D)-31-124-G93.