ON TWO SYSTEMS OF UNEQUAL PROBABILITY SAMPLING WITHOUT REPLACEMENT*

BY J. N. K. RAO

(Received Feb. 4, 1963; revised Aug. 13, 1963)

1. Introduction

A general theory of sampling with unequal probabilities and without replacement is given by Horvitz and Thompson [5]. Their estimator of the population total Y is

$$\hat{Y}=\sum_{1}^{n} y_i/\pi_i \tag{1}$$

where y_j is the value of the characteristic for the jth unit and π_j is the probability for selecting the jth unit in a sample of size n drawn from a population of N units. The variance of $\hat{Y}$ is

$$V(\hat{Y})=\sum_{1}^{N}\frac{y_j^2}{\pi_j}+\sum_{i\neq i'}^{N}\frac{P_{ii'}}{\pi_i\pi_{i'}}y_iy_{i'}-Y^2 \tag{2}$$

where $P_{ii'}$ is the probability for the units i and i' to be both in the sample. If a supplementary variate x_i, which is approximately proportional to y_i, is available for all the units in the population, considerable reduction in the variance of $\hat{Y}$ can be achieved by making π_i proportional to x_i. Therefore, we confine here to sampling procedures for which $\pi_i\alpha x_i$. Also, we consider the case $n=2$ only.

Now, all procedures of unequal probability sampling without replacement need not necessarily lead to estimators with variance (2) always smaller than the variance, $V'(\hat{Y})$, in sampling with probabilities proportional to x_i with replacement (see Durbin [3]), where

$$V'(\hat{Y})=\sum_{1}^{N}\pi_j\left(\frac{y_j}{\pi_j}-\frac{Y}{n}\right)^2. \tag{3}$$

To show this, consider the following example taken from Des Raj [2] (it was used to show that the Yates and Grundy estimator of variance of $\hat{Y}$ need not always be positive):

* Research sponsored by Army Research Office, Durham, N.C., under Grant No. DA-ARO (D)-31-124-G93.

$$N=4,\ n=2,\ y_1=1,\ y_2=2,\ y_3=3,\ y_4=4 \quad \text{and}$$

$$P_{12}=P_{34}=\frac{3}{8},\ P_{13}=P_{14}=P_{23}=P_{24}=\frac{1}{16}.$$

Then $\pi_1=\pi_2=\pi_3=\pi_4=\frac{1}{2}$ and, from (2) and (3),

$$V(\hat{Y})=12.5 \text{ and } V'(\hat{Y})=10.0 \text{ so that } V(\hat{Y})>V'(\hat{Y}).$$

It is therefore useful to identify systems of unequal probability sampling without replacement for which $V(\hat{Y})$ is always smaller than $V'(\hat{Y})$ in sampling with replacement. Recently, Brewer and Undy [1] have proved this property for the following sampling system due to Narain [6]: Select the first unit with probabilities proportional to the revised sizes x_i', and the second unit with probabilities proportional to the revised sizes of the remaining units, where

$$\pi_i=p_i'+p_i'\sum_{j\neq i}^{N}\frac{p_j'}{1-p_j'}=2p_i \tag{4}$$

where $p_i'=x_i'/X$, $p_i=x_i/X$, $X=\sum_N x_i$ and $\sum_N p_i'=1$. The sampling system considered by Hartley and Rao [4] has this property asymptotically.

It is interesting to note that this desirable property leads to another useful property, namely that the Yates and Grundy estimator of variance of $\hat{Y}$,

$$v(\hat{Y})=\frac{\pi_i\pi_{i'}-P_{ii'}}{P_{ii'}}\left(\frac{y_i}{\pi_i}-\frac{y_{i'}}{\pi_{i'}}\right)^2, \tag{5}$$

is always positive. This follows from Narain [6] where it is shown that a necessary condition for $V(\hat{Y})$ with $\pi_i=2p_i$ to be always smaller than $V'(\hat{Y})$ is $\pi_i\pi_{i'}>P_{ii'}$ for all pairs i and $i'(i\neq i')$, the sample size being two. Sen [8] and Des Raj [2] have proved that $v(\hat{Y})$ is always positive for Narain's system by showing that $\pi_i\pi_{i'}>P_{ii'}$ directly.

The purpose of the present note is to show that two other well-known systems of unequal probability sampling without replacement lead to an estimator $\hat{Y}$ with $V(\hat{Y})$ always smaller than $V'(\hat{Y})$. The first system, due to Horvitz and Thompson [5], is based on Midzuno's scheme and is as follows: select the first unit with probabilities proportional to the revised sizes x_i'' and the second unit with equal probabilities without replacement, where the x_i'' are determined from

$$\pi_i=\frac{N-2}{N-1}p_i''+\frac{1}{N-1}=2p_i \tag{6}$$

where $p_i''=x_i''/X$ and $\sum_N p_i''=1$. From (6) we have that

$$(7) \qquad p_i''=2\frac{N-1}{N-2}p_i-\frac{1}{N-2}.$$

The advantage of this system is that the revised sizes can be easily computed from (7), whereas Narain's system involves a iterative solution. However, since p_i'' must be greater than zero, the system is valid only for $p_i>\frac{1}{2(N-1)}$ so that it is only of limited use. Denote the system as sampling system (a).

The second system, denoted as sampling system (b), is as follows: Select two units with probabilities proportional to the revised sizes x_i^* with replacement. If the two units are identical, reject both selections and make two further selections using the same method, the process being continued until two different units are selected in the sample. The revised sizes x_i^* are obtained by iteration from

$$(8) \qquad \pi_i=2\frac{p_i^*(1-p_i^*)}{1-\sum_1^N p_i^{*2}}=2p_i$$

where $p_i^*=x_i^*/X$ and $\sum_N p_i^*=1$. It is shown in Rao [7] that the sampling systems of Hartley and Rao, Narain and system (b) have exactly the same asymptotic variance of $\hat{Y}$, and, hence, $V(\hat{Y})$ is always smaller than $V'(\hat{Y})$ asymptotically.

2. Sampling system (a)

We follow Brewer and Undy's method to derive $V(\hat{Y})$. Now the variance of $\hat{Y}$ is

$$(9) \qquad V(\hat{Y})=E\left\{\frac{1}{2}\left(\frac{y_1}{p_1}+\frac{y_2}{p_2}\right)-Y\right\}^2$$
$$=\frac{1}{4}E\left\{\left(\frac{y_1}{p_1}-Y\right)^2+2\left(\frac{y_1}{p_1}-Y\right)\left(\frac{y_2}{p_2}-Y\right)+\left(\frac{y_2}{p_2}-Y\right)^2\right\}$$

where the subscripts 1 and 2 refer to the order of selection. Since the probability of selecting the ith unit in the first draw is p_i'' and the conditional probability of selecting the jth unit given that the ith unit is selected in the first draw is $1/(N-1)$,

$$(10) \qquad V(\hat{Y})=\frac{1}{4}\left[\sum_1^N p_i''\left(\frac{y_i}{p_i}-Y\right)^2+\sum_1^N p_i''\sum_{j\neq i}^N\frac{1}{N-1}\left(\frac{y_j}{p_j}-Y\right)^2\right.$$

$$+2\sum_1^N p_i''\left(\frac{y_i}{p_i}-Y\right)\sum_{j\neq i}^N\frac{1}{N-1}\left(\frac{y_j}{p_j}-Y\right)\Big]$$

$$=\frac{1}{4}\Big[\frac{N-4}{N-1}\sum_1^N p_i''\left(\frac{y_i}{p_i}-Y\right)^2+\frac{1}{N-1}\sum_1^N\left(\frac{y_i}{p_i}-Y\right)^2$$

$$+\frac{2}{N-1}\Big\{\sum_1^N p_i''\left(\frac{y_i}{p_i}-Y\right)\Big\}\Big\{\sum_1^N\left(\frac{y_i}{p_i}-Y\right)\Big\}\Big].$$

Substituting now for p_i'' from (7) in (10) and noting that $\sum_N p_i\left(\frac{y_i}{p_i}-Y\right)=0$, we find that

$$(11)\quad V(\hat{Y})=\frac{N-4}{2(N-2)}\sum_1^N p_i\left(\frac{y_i}{p_i}-Y\right)^2-\frac{1}{2(N-1)(N-2)}\Big[\Big\{\sum_1^N\left(\frac{y_i}{p_i}-Y\right)\Big\}^2$$

$$-\sum_1^N\left(\frac{y_i}{p_i}-Y\right)^2\Big]$$

$$=V'(\hat{Y})-\frac{1}{N-2}\sum_1^N\left(p_i-\frac{1}{2(N-1)}\right)\left(\frac{y_i}{p_i}-Y\right)^2$$

$$-\frac{1}{2(N-1)(N-2)}\Big\{\sum_1^N\left(\frac{y_i}{p_i}-Y\right)\Big\}^2$$

since

$$(12)\qquad V'(\hat{Y})=\frac{1}{2}\sum_1^N p_i\left(\frac{y_i}{p_i}-Y\right)^2.$$

Now from (7) we know that system (a) is valid for $p_i>\frac{1}{2(N-1)}$ only. Therefore, it follows from (11) that $V(\hat{Y})$ for system (a) is always smaller than $V'(\hat{Y})$. Since a necessary condition for $V(\hat{Y})$ to be always smaller than $V'(\hat{Y})$ is that $\pi_i\pi_{i'}>P_{ii'}$ for all pairs i, $i'(i\neq i)$, it follows that the Yates and Grundy estimator of variance is always positive for system (a). Sen [8] and Des Raj [2] have proved that $v(\hat{Y})$ is always positive by showing that $\pi_i\pi_{i'}>P_{ii'}$ directly.

Sampling system (b)

For system (b), it is easily seen that the probability of selecting the ith unit in the first draw is

$$p_i^*(1-p_i^*)/(1-\sum_1^N p_i^{*2}),$$

and the conditional probability of selecting the jth unit in the second draw given that the ith unit is selected in the first draw is $p_j^*/(1-p_i^*)$. Therefore, from (9) it follows that

$$(13)\quad V(\hat{Y})=\frac{1}{4}\Bigg[\sum_1^N \frac{p_i^*(1-p_i^*)}{1-\sum_1^N p_i^{*2}}\left(\frac{y_i}{p_i}-Y\right)^2+\sum_1^N \frac{p_i^*(1-p_i^*)}{1-\sum_1^N p_i^{*2}}\sum_{j\neq i}^N \frac{p_j^*}{1-p_i^*}\left(\frac{y_i}{p_i}-Y\right)^2$$

$$+2\sum_1^N \frac{p_i^*(1-p_i^*)}{1-\sum_1^N p_i^{*2}}\left(\frac{y_i}{p_i}-Y\right)\sum_{j\neq i}^N \frac{p_j^*}{1-p_i^*}\left(\frac{y_j}{p_j}-Y\right)\Bigg]$$

$$=\frac{1}{2(1-\sum_1^N p_i^{*2})}\Bigg[\sum_1^N p_i^*(1-p_i^*)\left(\frac{y_i}{p_i}-Y\right)^2+\left\{\sum_1^N p_i^*\left(\frac{y_i}{p_i}-Y\right)\right\}^2$$

$$-\sum_1^N p_i^{*2}\left(\frac{y_i}{p_i}-Y\right)^2\Bigg]$$

$$=V'(\hat{Y})-\frac{1}{2(1-\sum_1^N p_i^{*2})}\Bigg[\sum_1^N p_i^{*2}\left(\frac{y_i}{p_i}-Y\right)^2-\left\{\sum_1^N p_i^*\left(\frac{y_i}{p_i}-Y\right)\right\}^2\Bigg]$$

using (8). Also, from (8),

$$(14)\qquad p_i^*=p_i(1-\sum_1^N p_i^{*2})+p_i^{*2}.$$

Therefore, noting that $\sum_1^N p_i\left(\frac{y_i}{p_i}-Y\right)=0$, we have

$$(15)\qquad \sum_1^N p_i^{*2}\left(\frac{y_i}{p_i}-Y\right)^2-\left\{\sum_1^N p_i^*\left(\frac{y_i}{p_i}-Y\right)\right\}^2$$

$$=\sum_1^N p_i^{*2}\left(\frac{y_i}{p_i}-Y\right)^2-\left\{\sum_1^N p_i^{*2}\left(\frac{y_i}{p_i}-Y\right)\right\}^2$$

$$>\sum_1^N p_i^{*3}\left(\frac{y_i}{p_i}-Y\right)^2-\left\{\sum_1^N p_i^{*2}\left(\frac{y_i}{p_i}-Y\right)\right\}^2$$

$$=\sum_1^N p_i^*(a_i-\sum_1^N p_i^* a_i)^2\geqq 0$$

where $a_i=p_i^*\left(\frac{y_i}{p_i}-Y\right)$. Hence, from (13) and (15), we have that $V(\hat{Y})$ for system (b) is smaller than $V'(\hat{Y})$. Therefore, it follows that the Yates and Grundy estimator of variance is always positive for system (b).

IOWA STATE UNIVERSITY, U.S.A.

REFERENCES

[1] K. R. W. Brewer and G. C. Undy, "Samples of two units drawn with unequal probabilities without replacement," *Aust. J. Statist.*, Vol. 4 (1962), pp. 89–100.

[2] Des Raj, "Some estimators in sampling with varying probabilities without replacement," *J. Amer. Statist. Ass.*, Vol. 51 (1956), pp. 269–284.

[3] J. Durbin, "Some results in sampling theory when the units are selected with unequal probabilities," *J. R. Statist. Soc.*, (B), Vol. 15 (1953), pp. 262–269.

[4] H. O. Hartley and J. N. K. Rao, "Sampling with unequal probabilities and without replacement," *Ann. Math. Statist.*, Vol. 33 (1962), pp. 350–374.

[5] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from finite universe," *J. Amer. Statist. Ass.*, Vol. 47 (1952), pp. 663–685.

[6] R. D. Narain, "On sampling without replacement with varying probabilities," *J. Indian Soc. Agric. Statist.*, Vol. 3 (1951), pp. 169–175.

[7] J. N. K. Rao, "On three procedures of unequal probability sampling without replacement," *J. Amer. Statist. Ass.*, Vol. 58 (1963), pp. 202–215.

[8] A. R. Sen, "On the estimate of the variance in sampling with varying probabilities," *J. Indian Soc. Agric. Statist.*, Vol. 5 (1953), pp. 119–127.