ON THE ESTIMATE OF THE VARIANCE IN UNEQUAL PROBABILITY SAMPLING

By J. N. K. RAO

(Received July 10, 1961)

1. Introduction

The general theory of unequal probability sampling without replacement is given by Horvitz and Thompson (1952). Their estimator of the population total Y is

$$\hat{Y}=\sum^{n}\frac{y_i}{\Pi_i} \tag{1}$$

where y_i is the value of the character for the ith unit and Π_i is the probability of selecting the ith unit in a sample of size n drawn from the population of size N. The variance of $\hat{Y}$ is

$$V(\hat{Y})=\sum^{N}\frac{y_j^2}{\Pi_j}+2\sum_{i<i'}^{N}\frac{\Pi_{ii'}}{\Pi_i\Pi_{i'}}y_iy_{i'}-Y^2 \tag{2}$$

where $\Pi_{ii'}$ is the probability for the ith and i'th units to be both in the sample. Horvitz and Thompson's unbiased estimator of variance is

$$v_{HT}(\hat{Y})=\sum^{n}\frac{1-\Pi_j}{\Pi_j}y_j^2+2\sum_{i<i'}^{n}\frac{\Pi_{ii'}-\Pi_i\Pi_{i'}}{\Pi_i\Pi_{i'}\Pi_{ii'}}y_iy_{i'} \tag{3}$$

Yates and Grundy (1953) suggest an alternative estimator of variance which is believed to be less often negative than Horvitz and Thompson estimator of variance. Their estimator of variance is

$$v_{YG}(\hat{Y})=\sum_{i<i'}^{n}\frac{\Pi_i\Pi_{i'}-\Pi_{ii'}}{\Pi_{ii'}}\left(\frac{y_i}{\Pi_i}-\frac{y_{i'}}{\Pi_{i'}}\right)^2. \tag{4}$$

It is shown by Sen (1953) and Des Raj (1956) that Yates and Grundy estimator of variance is always positive for the following two important sampling systems: (a) The first unit is selected with probabilities p_i (Usually p_i are probabilities proportional to the size of the units denoted by p.p.s.) and the remaining $(n-1)$ units in the sample are selected with equal probabilities and without replacement. This scheme

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

is due to Midzuno (1950). (b) The first unit is selected with p.p.s. and the second unit is selected with p.p.s. of the remaining units, the sample size being two. This scheme is due to Horvitz and Thompson (1952).

It will be of interest to identify more sampling systems which yield simple expressions for the probabilities Π_i and $\Pi_{ii'}$ as in the case of sampling systems (a) and (b), and for which Yates and Grundy estimator of variance (4) is always positive. The purpose of this note is to identify a new sampling system with $n>2$ which yields simple expressions for Π_i and $\Pi_{ii'}$ as in the case of sampling systems (a) and (b), and for which Yates and Grundy estimator of variance is always positive.

2. The new sampling system

The sampling system is as follows: (c) The first unit is selected with p.p.s., the second unit with p.p.s. of the remaining units as in sampling system (b) and the remaining $(n-2)$ units in the sample are selected with equal probabilities and without replacement. Now from the above description of the sampling system (c) it is easily seen that

$$\Pi_i=p_i+p_i\sum_{j\neq i}^{N}\frac{p_j}{(1-p_j)}+\sum_{j\neq k\neq i}^{N}\sum\frac{p_jp_k}{(1-p_j)}\cdot\frac{n-2}{N-2}\,. \tag{5}$$

Noting that $\sum_{i}^{N}p_i=1$, (5) can be simplified as

$$\Pi_i=\frac{(N-n)}{(N-2)}p_i\left[\frac{1}{1-p_{i'}}+A_{ii'}\right]+\frac{n-2}{N-2} \tag{6}$$

where

$$A_{ii'}=\sum_{j\neq(i,i')}^{N}\frac{p_j}{(1-p_j)}\,. \tag{7}$$

Also it is easily seen that

$$\begin{aligned}\Pi_{ii'}&=p_ip_{i'}\left(\frac{1}{1-p_i}+\frac{1}{1-p_{i'}}\right)+\left(\sum_{j\neq(i,i')}^{N}p_j\right)\left(\frac{p_i}{1-p_i}+\frac{p_{i'}}{1-p_{i'}}\right)\cdot\frac{n-2}{N-2}\\&\quad+(p_i+p_{i'})\left[\sum_{j\neq(i,i')}^{N}\frac{p_j}{(1-p_j)}\right]\cdot\frac{n-2}{N-2}+\frac{(n-2)(n-3)}{(N-2)(N-3)}\sum_{j\neq j'\neq(i,i')}^{N}\sum\frac{p_jp_{j'}}{(1-p_j)} \qquad (8)\\&=p_ip_{i'}\left(\frac{1}{1-p_i}+\frac{1}{1-p_{i'}}\right)\cdot\frac{N-n}{N-2}+\frac{(n-2)(N-n)}{(N-2)(N-3)}(p_i+p_{i'})\\&\quad+\frac{(n-2)(N-n)}{(N-2)(N-3)}(p_i+p_{i'})A_{ii'}+\frac{(n-2)(n-3)}{(N-2)(N-3)}\,. \qquad (9)\end{aligned}$$

For the special case of equal probabilities $p_i=1/N$, it is well known that $\Pi_i=\frac{n}{N}$ and $\Pi_{ii'}=\frac{n(n-1)}{N(N-1)}$. Now, if we substitute $p_i=\frac{1}{N}$ in (6) and (9) we obtain $\frac{n}{N}$ and $\frac{n(n-1)}{N(N-1)}$ respectively, thus providing a check.

Yates and Grundy estimator of variance (4) is always positive when

$$\Pi_i\Pi_{i'}-\Pi_{ii'}>0 \tag{10}$$

for every pair (i, i'). So, it is sufficient if we prove that (10) holds for our sampling system (c) for every pair (i, i'). After some simplification using (6) and (9) we find that

$$\begin{aligned}\Pi_i\Pi_{i'}-\Pi_{ii'}=\frac{(N-n)}{(N-2)^2}\Big[&\frac{(n-2)}{(N-3)}\{(1-p_i-p_{i'})-A_{ii'}(p_i+p_{i'})\}\\&-\frac{p_ip_{i'}(1-p_i-p_{i'})(N-n)}{(1-p_i)(1-p_{i'})}+(N-n)A_{ii'}\frac{p_ip_{i'}(2-p_i-p_{i'})}{(1-p_i)(1-p_{i'})}\\&+(N-n)p_ip_{i'}A_{ii'}^2\Big].\end{aligned} \tag{11}$$

Consider now the term

$$M=(1-p_i-p_{i'})-A_{ii'}(p_i+p_{i'}) \tag{12}$$

in (11). Since

$$1-p_j>p_i+p_{i'} \qquad \text{for } j\neq(i, i')$$

we have that

$$A_{ii'}(p_i+p_{i'})=\sum_{j\neq(i,i')}^{N}\frac{p_j}{(1-p_j)}(p_i+p_{i'})<\sum_{j\neq(i,i')}^{N}p_j=1-p_i-p_{i'} \tag{13}$$

so that

$$M>(1-p_i-p_{i'})-(1-p_i-p_{i'})=0\,. \tag{14}$$

Therefore

$$\begin{aligned}\Pi_i\Pi_{i'}-\Pi_{ii'}>\frac{(N-n)}{(N-2)^2}\Big[&(N-n)p_ip_{i'}A_{ii'}^2+(N-n)A_{ii'}\frac{p_ip_{i'}(2-p_i-p_{i'})}{(1-p_i)(1-p_{i'})}\\&-\frac{(N-n)p_ip_{i'}(1-p_i-p_{i'})}{(1-p_i)(1-p_{i'})}\Big].\end{aligned} \tag{15}$$

To prove the that r.h.s. of (15) is greater than zero, one can use the proof of Sen (1953) and Des Raj (1956) for the sampling system (b), which consists of finding the minimum of $A_{ii'}$ and substituting it in (15).

However, we give here an elementary alternative proof to show that r.h.s. of (15) is greater than zero. This proof, of course, can be used as an alternative proof to show that Yates and Grundy estimator of variance is always positive for sampling system (b). The proof is as follows: Since

$$A_{ii'}=\sum_{j\neq(i,i')}^{N}\frac{p_j}{(1-p_j)}>\sum_{j\neq(i,i')}^{N}p_j=1-p_i-p_{i'} \tag{16}$$

by substituting for $A_{ii'}$ from (16) in r.h.s. of (15), it immediately follows that

$$\Pi_i\Pi_{i'}-\Pi_{ii'}>\frac{(N-n)}{(N-2)^2}\left[(N-n)p_ip_{i'}A_{ii'}^2+\frac{(N-n)p_ip_{i'}}{(1-p_i)(1-p_{i'})}(1-p_i-p_{i'})^2\right] \tag{17}$$

which is greater than zero. Hence, Yates and Grundy estimator of variance is always positive for sampling system (c).

IOWA STATE UNIVERSITY, U.S.A.

REFERENCES

[1] Des Raj, "Some estimators in sampling with varying probabilities without replacement," *Journal of the American Statistical Association*, Vol. 51 (1956), pp. 269-284.

[2] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from finite universe," *ibid*, Vol. 47 (1952), pp. 663-685.

[3] H. Midzuno, "An outline of the theory of sampling systems," *Annals of the Institute of Statistical Mathematics*, Vol. 1 (1950), pp. 149-156.

[4] A. R. Sen, "On the estimate of the variance in sampling with varying probabilities," *Journal of the Indian Society of Agricultural Statistics*, Vol. 5 (1953), pp. 119-127.

[5] F. Yates and P. M. Grundy, "Selection without replacement from within strata with probability proportional to size," *Journal of the Royal Statistical Society, Series B*, Vol. 15 (1953), pp. 253-261.