USE OF TRANSFORMATION IN PPS SAMPLING

By

S. Mohanty

College of Agriculture, Bhubaneswar

1. Introduction

A general theory of unequal probabilities sampling without replacement was first given by Horvitz and Thompson (1952). Since then a large number of papers have been published on this topic. Durbin (1953) has pointed out that the H.T. estimate is less efficient than the *pps* sampling with replacement for some set of inclusion probabilities. Besides this it suffers from a major drawback that the estimate of $\hat{V}(\hat{Y})_{HT}$ assumes negative values for some samples. But when the Midzuno scheme is used the estimate becomes more efficient than the sampling with replacement. Under this system the estimate of the variance is never negative. The main advantage of this method of sampling besides its simplicity and non-negative estimate is that, it is possible to compute a set of revised probabilities of selection P_i''s such that the inclusion probabilities π_i resulting from the revised probabilities are proportional to the initial probabilities of selection. It is desirable to do so since $P_i's$ can be chosen proportional to, some known size measure. This is possible only when the initial probabilities of selection P_i satisfy the condition

$$P_i > \frac{n-1}{n(N-1)}, \text{ for all } i\text{'s}, \qquad \ldots(1.1)$$

This restriction on the initial probabilities of selection naturally limits the uses of this simple and efficient system in practice.

The object of this paper is to transform the auxiliary variate X_i, on which the initial probabilities of selection depends such that the restriction 1.1 is satisfied for all i's.

2. The Suggested Method

All the initial probabilities of a pps sampling scheme can not be less than $\frac{n-1}{n(N-1)}$. To prove this let us assume that all P_i's are less than $\frac{n-1}{n(N-1)}$. Then, $n(N-1)P_i < (n-1)$, for all i's. Summing this over i's and simplifying, we have $n > N$. This being impossible,

all P_i's cannot be less than $\frac{n-1}{n(N-1)}$. All or some of the P_i's must be greater than $\frac{n-1}{n(N-1)}$. Hence we have the following Lemma:

Lemma 2.1. For a *pps* sampling scheme, all or some of the probabilities of selection must be greater than $\frac{n-1}{n(N-1)}$.

If all P_i's are greater than $\frac{n-1}{n(N-1)}$, the Midzuno scheme with revised probabilities can be applied directly. But difficulty arises, when some of the probabilities are less than $(n-1)/n(N-1)$. Let P_1 be the smallest of these probabilities. Let us transform X to Z through a linear transformation

$$Z_i = \frac{X_i}{\bar{X}} + c, \qquad \ldots(2.1)$$

where $\bar{X}$ is the mean of X and c is an unknown constant. Under the transformation $Z\ (=\sum_i^N Z_i) = N(1+c)$.

The new set of probabilities are given by

$$P_i' = \frac{NP_i + c}{N(1+c)}, \qquad \ldots(2.2)$$
$$i = 1, 2, \ldots\ldots\ldots, N.$$

Under this linear transformation, the correlation between Y and Z remains same as that of Y and X, since X's are usually positive. To make the restriction (1.1) valid for new set of probabilities, we must find c such that the lowest new probability

$$P_i' > \frac{n-1}{n(N-1)} \ (=k, \text{ say}) \qquad \ldots(2.3)$$

Substituting the values of P_1' from (2.2) and simplifying, we have

$$c > \frac{N(k-P_1)}{1-Nk}. \qquad \ldots(2.4)$$

c will be always positive as k is greater than P_1, the lowest initial probability and $(1-Nk) = \frac{N-n}{n(N-1)}$. The R.H.S. of (2.4) being known, we can select c such that the new set of probabilities P_i' 's become greater than $(n-1)/n(N-1)$. Using these new set of probabilities P_i', the Midzuno system of sampling can be used. Hence, we have the following theorem :

THEOREM 2.1

Under the *pps* sampling scheme, if some of the selection probabilities are less than $\frac{n-1}{n(N-1)}$ $(=k)$, a linear transformation of the probabilities of the form $P_i'=\frac{NP_i+c}{N(1+c)}$ may be used, where c is a random constant satisfying the condition

$$c>\frac{N(k-P_1)}{1-Nk}, \text{ where}$$

P_1 being the smallest initial probability, such that the P_i' 's are greater than $(n-1)/n(N-1)$.

By the linear transformation with positive c, the smaller probabilities are increased and larger probabilities are decreased since

$$\sum_{i=1}^{N} P_i'=1.$$

It can be very easily shown that for any $P_i>P_j$ $(i\neq j)$, $P_i'>P_j'$, which suggest that through the above transformation probabilities maintain the same order of magnitude as the initial probabilities.

REMARK 2.1

The transformation considered in 2.1 does not affect the correlation between the study variable and the auxiliary variable. However the length of the intercept cut on the study variable (Y) axis is changed in the situation when the regression is linear. So naturally one may think that the efficieny may be reduced by the transformation which increases the intercept while satisfying the condition (1.1). But how much efficieny is reduced can only be calculated from the actual observations.

The intercept of the regression line Y on X is given by $\alpha=\bar{Y}-\beta\bar{X}$ and that of Y on Z after transformation by $\alpha'=\alpha-\beta\bar{X}c$, where β is the regression coefficient of Y on X and is usually positive under *pps* strategy. Thus the transformation will reduce the intercept is $|\alpha'|<|\alpha|$ only when α is positive and 2α is greater than $\beta\bar{X}c$. When α is negative this transformation will not reduce the intercept as β. $\bar{X}$ and c are positive.

However, the advantage of this transformation is in the use of the simple Midzuno scheme with revised probabilities in practical cases, when it is not possible to use any other complicated *pps* without replacement strategy.

3. ILLUSTRATION

To illustrate the working of the above transformation, we have considered here the most reported example studied by Yates and Grundy (1953). The three populations with the initial probabilities are given below:

TABLE 3.1
Population

Sl. No.	P_i	*A*	*B*	*C*
1	0.1	0.5	0.8	0.2
2	0.2	1.2	1.4	0.6
3	0.3	2.1	1.8	0.9
4	0.4	3.2	2.0	0.8

The relative performance of the various sampling procedures for sample size 2 have been considered below with and without transformation. For the above populations the Midzuno system of sampling with revised probabilities using H.T. estimate cannot be used as one of the probability is less than $(n-1)/n(N-1)$ i.e. 0.167. Hence, we have transformed the probabilities for $c>0.84$ (obtained from 2.4). In the first population (A), the intercept is negative and hence it is not possible to reduce the intercept by transformation. In the population B, the intercept is positive and 2α is greater than $\beta\bar{X}c$ for c is in the region $1>c>0.84$. In the population C, the intercept is positive and does not satisfy $2\alpha>\beta\bar{X}c$ for $c>0.84$. For illustration we have considered $c=0.9$ for the transformation. The new set of transformed probabilities are $P_1'=0.1710$, $P_2'=0.2237$, $P_3'=0.2764$ and $P_4'=0.3289$. Using these probabilities, the variance of the sample etimate of the population mean is calculated. The variance of the sample estimates of

TABLE 3.2
Variance of the estimate

Sampling Procedure	*Population*					
	A		*B*		*C*	
	For P_i	*For* P'_i	*For* P_i	*For* P'_i	*For* P_i	*For* P'_i
1. *pps* sampling with replacement	0.0313	0.1871	0.0313	0.0121	0.0078	0.0146
2. Midzuno system with revised probabilities & H.T. estimate	Not possible	0.0667	Not possible	0.0053	Not possible	0.0060
3. Horvitz-Thompson estimate	0.0504	0.0803	0.0028	0.0079	0.0036	0.0064

the population mean corresponding to sampling procedures for the three populations are given on previous page for the original and revised probabilities.

It is observed that even though for the population B, the intercept is reduced by the transformation, still the variance of the H.T estiamte increases, whereas the variance of the estimate for the replacement strategy is reduced. Therefore, it cannot be always said that the variance of the estimate will decrease with the decrease of the intercept except in case of *pps* sampling with replacement.

In fact none of the methods for *pps* sampling is always better from the point of view of precision and some are even difficult for the point of estimating the variance for large samples. Therefore, a procedure which is suitable for the practical point of view and efficient than the replacement procedure is usually recommended for the actual survey work. Midzuno strategy being one of them, the linear transformation suggested in the paper greatly increases its scope in practical uses and even some cases may increase the efficiency.

SUMMARY

It is well known that the technique of drawing units with varying probabilities is used in practice in order to incorporate the available supplementary information in sampling procedure so that the resulting variance of the estimate is minimised. Horvitz and Thompson sampling strategy for such situation without replacement is well known. But its major drawbacks are that in some cases (*i*) the estimate of the variance assume negative values and, (*ii*) becomes less efficient than the scheme with replacement. The strategy suggested by Midzuno is free from such defects. Its main advantage is that, it is possible to compute a set of revised probabilities of selection, such that the inclusion probabilities calculated from these revised probabilities are proportional to the initial probabilities of selection. This happens only when the initial probabilities (P_i) satisfy the condition $P_i > \frac{n-1}{n(N-1)}$, for all i's. In this paper it has been shown that through linear transformation of the supplementary variable, it is possible to remove the above restriction when it exists and thus giving Midzuno strategy a wide applicability in practice.

REFERENCES

[1] Durbin, J. (1953) : Some results in sampling theory when the units are selected with unequal probabilities. *J. Roy. Stat. Soc.*, (B), 15, 262-269.

[2] Horvitz, D.G. & Thompson, D.J. (1952) : A generalization of sampling without replacement from a finite universe, *JASA.*, 47, 663-685.

[3] Midzuno, H. (1950) : An outline of the theory of sampling system. *Ann. Inst. Stat. Math.*, 1, 149-156.

[4] Yates, F. & Grundy, P.M. (1953) : Selection without replacement from within strata with probability proportional to size. *J. Roy. Stat. Soc.*, (B), 15, 253-261.