Commun. Fac. Sci. Univ. Ank. Ser. $A_1$
V. 36, Number 2, pp 43–54 (1987)

# MULTIPLICITY ESTIMATORS FOR MULTISTAGE SAMPLE SURVEYS: A REVIEW

H. ÖZTAŞ AYHAN
*Department of Statistics, METU Ankara, Turkey*

## ABSTRACT

Recent developments in multiplicity surveys have been examined. A comparison of the conventional survey estimator and the multiplicity estimator has been presented. Multiplicity estimators for stratified multistage cluster samples have been discussed.

New methods to reduce the amount of multiplicity information is presented. Among these "completely nested inverse multiplicity estimator", "superstage inverse multiplicity estimator", and the "use of multiple counting rules" are considered. Computational procedures for the sampling variances of various estimators have been mentioned.

KEY WORDS: Incomplete frames, Counting rules, Stratified cluster sampling, Network sampling, Multiplicity estimators.

## INTRODUCTION

Sample surveys require a frame for accessing the target population. In many cases, frames in which each target population element is linked to either do not exist, or they are too expensive to use because the target population is very rare. In order to increase the feasibility of the survey, frames in which target elements have multiple associations or linkages with the sampling units are used. The linkages are determined by means of counting rules that define which target elements are linked to which frame units. The number of linkages between a target element and a frame are called the element's multiplicity.

Survey which employ such frames are called surveys with multiplicity or network surveys. All multiplicity surveys require that the multiplicity of each sample target element be determined in order to produce unbiased estimates. Often multiplicity information is difficult to obtain either because of recall problems in interview surveys or because the information is lacking in record surveys.

ISSN 0251-0871, A. Üniv. Basımevi

## MULTIPLICITY ESTIMATION

A series of articles explaining the statistical properties of multiplicity (network) estimator for various types of sample designs have indicated that network sampling is a feasible approach to identify the desired elements (Sirken 1970 a,b; Sirken 1972a, b; Sirken 1975; Sirken and Levy 1974; Nathan 1976; Levy 1977).

In a multiplicity survey each event may be linked to more than one household by a specified counting rule. When the multiplicity of each event is correctly reported, an unbiased estimate of the frequency of events of a given type can be obtained from a probability sample of households by weighting each report by the reciprocal of its multiplicity (Nathan, 1976).

Sirken (1970b) has shown that multiplicity estimates with wide counting rules will, in general, reduce the sampling variance as compared with the conventional estimate.

In general, higher response errors might be expected from reporting by more distant relatives, with differential biases according to the counting rule. In order to compaie different counting rules, it is necessary to consider and to evaluate the total mean square error and its components (Nathan, 1976).

A dual system network estimator is proposed by Sirken (1979). Two other dual system network estimators are proposed by Casady, Nathan and Sirken (1985) as potential improvements.

Recently Czaja, Snowden and Casady (1986) used multiplicity counting rules for improving the efficiency of surveys to locate and estimate characteristics of rare populations.

## MULTIPLICITY ESTIMATORS

### *General Approach*

In the household survey with multiplicity, sample households report information about their own residents as well as about other persons who live elsewhere, such as relatives or neighbours, as specified by a multiplicity rule adopted in the survey.

Suppose that a household survey is undertaken to estimate N, the number of individuals in population Ω with a particular attribute. Households are taken as the enumeration units.

Let

$$M = \text{number of households in population } \Omega$$

$$H_i = \text{the i}^{\text{th}} \text{ household } (i = 1, 2, \ldots, M)$$

The different individuals in the population having a specified characteristic are denoted by $I_\alpha$ ($\alpha = 1, 2, \ldots, N$) where N is the number to be estimated.

In the typical household surveys, each individual with the attribute would be reported by one and only one household in population Ω, the household of which he is a resident. This is known as the conventional survey.

Now consider a household survey with multiplicity. In this type of survey, each individual with the attribute would be reported by at least one household, the household of which he is a resident. In addition, he would be reported by other households in population Ω of which he is a nonresident as specified by the multiplicity rule adopted in the survey.

The total number of households in population Ω reporting the individual is referred to as his multiplicity. Every multiplicity rule is based on a system of linking individuals living in different households.

The difference between the conventional survey and the survey with multiplicity may be described as follows, which is based on Sirken (1970a).

Consider the indicator variables,

$$\nu_{\alpha,i} = \begin{cases} 1 & \text{if } I_\alpha \text{ is a resident of } H_i \\ 0 & \text{otherwise.} \end{cases}$$

and

$$\mu_{\alpha,i} = \begin{cases} 1 & \text{if } I_\alpha \text{ is not a resident of } H_i \\ & \text{and is reported by } H_i \\ 0 & \text{otherwise.} \end{cases}$$

Define two variates,

(1) Number of individuals reported by $H_i$ in the conventional survey is,

$$r_i = \sum_{\alpha=1}^{N} \nu_{\alpha,i}$$

(2) Weighted number of individuals reported by $H_i$ in the survey with multiplicity is,

$$t_i = \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} (\mu_{\alpha,\, i} + \nu_{\alpha,\, i})$$

where

$$s_\alpha = \sum_{i=1}^{M} (\mu_{\alpha,\, i} + \nu_{\alpha,\, i})$$

$s_\alpha$ is the number of households reporting $I_\alpha$, that is the multiplicity of $I_\alpha$.

Here $t_i$, which is the variate based on the multiplicity survey requires the multiplicity $s_\alpha$ of every individual reported by $H_i$. The survey procedure used to collect this information would probably depend on the type of multiplicity rule adopted.

Assume that a sample of m households is selected without replacement, then

$$N_r = \frac{M}{m} \sum_{i=1}^{m} r_i$$

is the estimate of N derived from the conventional survey, and

$$N_t = \frac{M}{m} \sum_{i=1}^{m} t_i$$

is the estimate of N derived from the multiplicity survey.

It is important to mention that $N_t$, unlike other estimators for multiplicity surveys that have been investigated (Birnbaum and Sirken, 1965), does not require matching individuals reported by different households to eliminate duplicate reports. Both estimates, $N_r$ and $N_t$ are unbiased because

$$E(r_i) = E(t_i) = \frac{N}{M}$$

and their variances are respectively

$$Var(N_r) = \frac{M-m}{M-1}.\frac{M^2}{m} Var(r)$$

and

$$Var(N_t) = \frac{M-m}{M-1}.\frac{M^2}{m} Var(t)$$

It follows that

$$Var(N_t) = Var(N_r)(1-\delta)$$

where the parameter delta,

$$\delta = \frac{Var(N_r) - Var(N_t)}{Var(N_r)} = \frac{Var(r) - Var(t)}{Var(r)}$$

$\delta$ is a measure of the relative gain in sampling efficiency resulting from the survey with multiplicity.

*Stratified and Multi Stage Approach*

Let us now examine the multiplicity estimation in stratified multi stage sample designs. The methods are illustrated foi a stratified, two stage probability sample. Let

N = The total number of elements in a population of interest.

$(i = 1, 2, \ldots, N)$

Suppose we wish to estimate a total for this population.

$$Y = \sum_{i=1}^{N} Y_i$$

where

$Y_i$ = the value of element i.

Now suppose that the elements of the taiget population are linked to a sampling frame with some elements having multiple linkages.

Let

$H$ = Total number of strata indexed by

$(h = 1, 2, \ldots, H)$

$M_h$ = Total number of first stage units in stratum h.

$(k = 1, 2, \ldots, M_h)$

$A_{hk}$ = Total number of second stage units in stratum h.

$(\alpha = 1, 2, \ldots, A_{hk})$

Suppose that some sample design which allows unbiased estimates of the total for a characteristic $T_{hk\alpha}$ has been used, for example.

$$E(\hat{t}) = \sum_{h=1}^{H} \sum_{k=1}^{M_h} \sum_{\alpha=1}^{A_{hk}} T_{hk\alpha}$$

For this estimator several different selection techniques (srs, unequal prob. sampl.) can be used as long as the expected value over all possible samples is given by the above formula.

Each of the target elements (i) is linked to one or more of the ultimate sampling units hk$\alpha$. For this situation any estimator, $\hat{X}$, for a total of the target population that is linked to this sampling frame will be unbiased if,

$$E(\hat{X}) = \sum_{h=1}^{H} \sum_{k=1}^{M_h} \sum_{\alpha=1}^{A_{hk}} \sum_{i=1}^{N} X(i, hk\alpha)$$

Here $X(i, hk\alpha)$ is the value for the $i^{th}$ element that is linked to sampling unit (hk$\alpha$).

More than one element can be linked to an ultimate sampling unit, and $X_{hk\alpha}$ is the sum over (i) of the $X(i, hk\alpha)$.

For $\hat{X}$ to be unbiased for Y it is only necessary that the sum of $X(i, hk\alpha)$ over the entire sampling frame be equal to $Y_i$

$$Y_i = \sum_{h=1}^{H} \sum_{k=1}^{M_h} \sum_{\alpha=1}^{A_{hk}} X(i, hk\alpha)$$

For reducing the amount of multiplicity information some methods were needed. Let us examine these methods which are based on Lessler (1981).

*Completely Nested Inverse Multiplicity Estimator*

This technique involves weighting the information obtained for each member of the target population by the inverse of the number of associations he / she has with the various stages of the nested sampling frame structure. Lets see how this provides unbiased estimates.

We make use of a series of indicator functions.

Let,

$$\begin{aligned} \theta(i, h) &= 1 \quad \text{if element (i) is linked to stratum (h).} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Then,

$$\gamma(l, i) = \sum_{h=1}^{H} \theta(i, h)$$

$$= \text{the total number of strata element (i) is linked to.}$$

In some studies, $\gamma$ (l, i) would be greater than 1 if the selected person had moved from one region of the country to another during the survey period and reselected in this region.

Likewise, let

$$\begin{aligned} \theta(i, hk) &= 1 \quad \text{if element (i) is linked to first stage unit (k) in stratum (h).} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

and

$$\begin{aligned} \theta(i, hk\alpha) &= 1 \quad \text{if element (i) is linked to second stage unit } (\alpha) \text{ in first stage unit (hk).} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Then the multiplicity at each of these states of sampling is given by,

$$\gamma(2, h, i) = \sum_{k=1}^{M_h} \theta(i, hk)$$

and

$$\gamma(3, hk, i) = \sum_{\alpha=1}^{A_{hk}} \theta(i, hk\alpha)$$

The sampling frame is a nested structure with three levels of nesting, a stratum level and two levels (or stages) of sampling within each stratum.

We will define a multiplicity measure for each level of the sampling frame as the number of associations or linkages that element (i) has with the units at a particular level.

Using this we define,

$$X\,(i, hk\alpha) = \frac{\theta\,(i,\ h)}{\gamma\,(1,\, i)} \cdot \frac{\theta\,(i,\, hk)}{\gamma\,(2,\, h,\, i)} \cdot \frac{\theta\,(i,\ hk\alpha)}{\gamma\,(3,\, hk,\, i)} Y_i$$

Here the expected value of X̂ is the sample estimate of Y.

$$E\ (\hat{X}) = \sum_{h=1}^{H} \sum_{k=1}^{M_h} \sum_{\alpha=1}^{A_{hk}} \sum_{i=1}^{N} X\,(i,\ hk\alpha)$$

$$= \sum_{i=1}^{N} Y_i \,.\, \sum_{h=1}^{H} \frac{\theta\,(i,\, h)}{\gamma\,(1,\, i)} \,.\, \sum_{k=1}^{M_h} \frac{\theta\,(i,\, hk)}{\gamma\,(2,\, h,\, i)} \,.$$

$$\sum_{\alpha=1}^{A_{hk}} \frac{\theta\,(i,\, hk\alpha)}{\gamma\,(3,\, hk,\, i)} = \sum_{i=1}^{N} Y_i$$

Now consider what advantage this estimator has over the traditional multiplicity estimator. This has been written it looks as if we need more information than before, namely all the nested multiplicities. However, we need only to know them for the sampling unit that was selected in the sample not for all units. There is however, an estimator that will require even less information.

*Superstage Inverse Multiplicity Estimator*

Any two or more adjacent stages in a sampling structure may be thought of as a single "superstage" for measuring an element's multiplicity with those levels of the sampling frame. For example, the first and second stages of sampling in our example could be considered a superstage with

$$S_h = \sum_{k=1}^{M_h} A_{hk}$$

$=$ total number of superstage units in stratum h, each indexed by $(j = 1, 2, \ldots, S_h)$

Then, as before we let

$$\theta\ (S, i, hj) = 1 \quad \text{if element (i) is linked to superstage unit (j) in stratum (h).}$$

$$= 0 \quad \text{otherwise}$$

Then multiplicity of element (i) with the units in the superstage is

$$\gamma\ (S, i, h) = \sum_{j=1}^{S_h} \theta\ (S, i, hj)$$

Using this we define

$$X\ (i, hk\alpha) = \frac{\theta\ (i, h)}{\gamma\ (1, i)} \ \frac{\theta\ (i, hk)\ \theta\ (i, hk\alpha)}{\gamma\ (S, i, h)} Y_i$$

Under the earlier condition of the E $(\hat{t})$, the E $(\hat{X})$ for the same design used is given by

$$E\ (\hat{X}) = \sum_{h=1}^{H} \sum_{k=1}^{M_h} \sum_{\alpha=1}^{A_{hk}} \sum_{i=1}^{N} X\ (i, hk\alpha)$$

This equals Y, because summing over the first stage units and second stage units is equivalent to summing over the superstage, i.e.,

$$\sum_{k=1}^{M_h} \sum_{\alpha=1}^{A_{hk}} \frac{\theta\ (i, hk)\ \theta\ (i, hk\alpha)}{\gamma\ (S, i, h)} = 1$$

The traditional estimator which measures an element's total multiplicity is equivalent to considering all parts of the sampling structure a single superstage.

### *Use of Multiple Counting Rules*

A counting rule is a mechanizm for defining a linkage between the frame or sampling units and the elements of the population of interest. In the previous sections, we have been assuming a single co-counting rule.

In general, suppose that there are C counting rules indexed by $c = 1, 2, \ldots, C$. We can then define a counting rule multiplicity.

Let,

$$w(c, i) = 1 \quad \text{if element (i) is linked to the frame by means of counting rule (c).}$$

$$= 0 \quad \text{otherwise.}$$

Then

$$C_i = \sum_{c=1}^{C} w(c, i)$$

= total number of counting rules by which element (i) is linked to the frame.

Now, $C_i$ can be defined within any level of the sampling structure or it can be defined overall.

The expected value of $\hat{Y}$ has the form,

$$E(\hat{Y}) = \sum_{h=1}^{H} \sum_{k=1}^{M_h} \sum_{\alpha=1}^{A_{hk}} \sum_{c=1}^{C} \sum_{i=1}^{N} \frac{w(c, i)}{C(i)} \cdot$$

$$\frac{\theta(h, i, c)}{\gamma(1, i, c)} \cdot \frac{\theta(hk, i, c)}{\gamma(2, i, c)} \cdot \frac{\theta(hk\alpha, i, c)}{\gamma(3, i, c)} Y_i$$

$$= \Sigma\, \Sigma\, \Sigma\, \Sigma\, \Sigma\; X(hk\alpha, i, c) = \sum_{i=1}^{N} Y_i$$

The unbiased estimator of Y has the same form as before, except that now each of the indicator functions and multiplicity measures refers to a specific event.

## VARIANCE OF VARIOUS ESTIMATORS

No special procedures are needed for deriving the variance formulas for these estimators or for estimating their variance. This is because if one takes the sum of X (hk$\alpha$, i, c) over (i) and (c), one is left with $X_{hk\alpha}$ for each sampling unit as in ordinary sampling without multiplicity.

## CONCLUSION

The cumulated knowledge in sampling theory of the recent past had considerable contribution to the sampling frame problems. The survey statisticians have realized that it was not practical and was costly to create a perfect frame. Instead it was more efficient to take the available units with their known selection probabilities and also with their multiplicities. This was very important for designing sample surveys in terms of time and cost.

We have also seen that, the multiplicity estimators can also be unbiased like the conventional estimators. We hope to see the application of this type of work to the "sample designs for large scale surveys" as well as "sampling of rare items."

The sampling errors for the multiplicity survey are not necessarily smaller than those for the conventional survey in which sample households report for their own residents only, in most instances it should be feasible to assure a suitable reduction in sampling error by selecting appropriate multiplicity rules. Using alternative statistical models, it can be shown that under specified conditions, sampling errors for the multiplicity survey are necessarily smaller than those for the conventional survey and the results of recent literature give insight regarding the factors contributing to the efficiency of the multiplicity survey.

## REFERENCES

BIRNBAUM, Z.W. and SIRKEN, M.G. 1965. Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics*, PHS Publ. No. 1000 Series 2 (11), National Center for Health Statistics, Washington, D.C.

CASADY, R.J., NATHAN, G. and SIRKEN, M.G. 1985. Alternative Dual System Network Estimators. *International Statistical Review*, 53 (2), 183–197.

CZAJA, R.F., SNOWDEN, C.B. and CASADY, R.J. 1986. Reporting Bias and Sampling Errors in a Survey of a Rare Population Using Multiplicity Counting Rules. *Jour. Amer. Statist. Assoc.*, 81, 411–419.

LESSLER, J.T. 1981. Multiplicity Estimators with Multiple Counting Rules for Multistage Sample Survey. *Proc. Social Statist Sect., A.S.A.*, 12–16.

LEVY, P.S. 1977. Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence of Attributes in Rare Populations. *Jour. Amer. Statist. Assoc.*, 72, 758–763.

NATHAN, G. 1976. An Empirical Study of Response and Sampling Errors for Multiplicity Estimates with Different Counting Rules. *Jour. Amer. Statist. Assoc.*, 71, 808–815.

SIRKEN, M.G. 1970a. Household Surveys with Multiplicity. *Jour. Amer. Statist. Assoc.*, 65, 257–266.

SIRKEN, M.G. 1970b. Survey Strategies for Estimating Rare Health Attributes. *Proc. 6th Berkeley Symp. Statist. and Prob.* Univ. of Calif. Press, pp. 135–144.

SIRKEN, M.G 1972a. Stratified Sample Surveys with Multiplicity. *Jour. Amer. Statist. Assoc*, 67, 224–227.

SIRKEN, M.G. 1972b. Variance Components of Multiplicity Estimators. *Biometrics* 28, 869–873.

SIRKEN, M.G. 1975. Network Surveys. *in Proceedings, Bull. Intern. Statist. Inst.* (40 th Session) Tokyo, pp. 332–342.

SIRKEN, M.G. 1979. A Dual System Network Estimator. *Proc. Survey Rese. Meth. Sect., A.S.A.*, 340–342.

SIRKEN, M.G. and LEVY, P.S. 1974. Multiplicity Estimation of Proportions Based on Ratios of Random Variables. *Jour. Amer. Statist. Assoc.*, 69, 68–73.