# ESTIMATION OF RARE EVENTS BY SIMPLE CLUSTER SAMPLING WITH MULTIPLICITY

PAUL S. LEVY
University of Illinois at the Medical Center
School of Public Health

## 1. Introduction

The use of sample surveys with multiplicity has been advocated by Sirken (1970) for estimating the number of demographic events (e.g. births, deaths) occurring in a particular time period. Since Sirken's original article, the theory of multiplicity estimation (also called network sampling) has been extended to stratified random sampling (Sirken, 1972; Levy, 1977), estimation of proportions (Sirken and Levy, 1974) and estimation of response errors (Nathan, 1976). In addition, sample surveys with multiplicity have been used in a wide variety of applications (Sirken, 1972; Sirken and Levy, 1974; Sirken et al., 1975; Nathan et al., 1977). In this report, the theory of multiplicity estimation is extended to simple cluster sampling, and an unbiased estimator is proposed for estimating the total number of events under this type of sampling design.

A survey with multiplicity is one in which an element (e.g. birth, death, individual having some attribute, etc.) may be linked to more than one enumeration unit by an algorithm or counting rule. For example, a counting rule in a survey with multiplicity might link a birth to the households of the grandparents as well as to the parents' household whereas a conventional counting rule would link the birth only to the household of the parents.

## 2. Development of the Estimator:

### 2.1 Population Parameters

Let us suppose that a population contains N enumeration units (e.u.'s) grouped into M primary sampling units (PSU's) with PSU i containing $N_i$ e.u.'s; i = 1, ..., M, and that a counting rule links Y events labeled $I_1, \ldots, I_Y$ to enumeration units according to an indicator variable, $\delta_{\alpha ij}$ given by

$$\delta_{\alpha ij} = \begin{cases} 1 & \text{if event } I_\alpha \text{ is linked to e.u. } j \text{ in PSU } i \text{ by the counting rule} \\ 0 & \text{otherwise} \end{cases}$$

where
$\alpha = 1, \ldots, Y$, $i = 1, \ldots, M$ and $j = 1, \ldots, N_i$.

For any counting rule, the following parameters can be defined which characterize the network linking the enumeration units to the elements:

$$s_{\alpha i} = \sum_{j=1}^{N} \delta_{\alpha ij} ;$$

$$s_\alpha = \sum_{i=1}^{M} s_{\alpha i} ;$$

$$t_{\alpha i} = \begin{cases} 1 & \text{if } s_{\alpha i} > 0 \\ 0 & \text{if } s_{\alpha i} = 0 \end{cases}$$

$$t_\alpha = \sum_{i=1}^{M} t_{\alpha i}$$

The parameter, $S_{\alpha i}$, denotes the number of enumeration units in a particular PSU (i) that are linked to a particular element, $I_\alpha$, whereas $s_\alpha$ denotes the total number of enumeration units linked to an element by a counting rule and is referred to as the multiplicity of the element with respect to the counting rule. Clearly, for conventional counting rule, each $s_\alpha$ would be equal to unity. The parameter, $t_\alpha$, denotes the number of PSU's in which a particular element, $I_\alpha$, is linked to one or more enumeration units.

Let $(z_{\alpha i} : \alpha = 1, \ldots, Y; i = 1, \ldots, M; t_{\alpha i} = 1)$ be any set of weights defined for all $(\alpha, i)$ such that $t_{\alpha i} = 1$ with the property:

$$\sum_{i=1}^{M} z_{\alpha i} \, s_{\alpha i} = 1 \qquad \alpha = 1, \ldots, Y$$

We then define the following parameters for each PSU.

$$\lambda'_{ij} = \sum_{\alpha=1}^{Y} z_{\alpha i} \, \delta_{\alpha ij} , \quad j = 1, \ldots, N_i$$

$$Y_i^* = \sum_{j=1}^{N_i} \lambda'_{ij} \; ; \quad E_i = \sum_{\alpha=1}^{Y} z_{\alpha i}^2 \, s_{\alpha i} / Y_i^*$$

The parameter, $\lambda'_{ij}$, represents the basic summary information obtained from enumeration units concerning elements, while the weights $\{z_{\alpha i}\}$ are functions of the particular network linking enumeration to elements and are chosen to make estimates of Y unbiased. Some possible choices of $z_{\alpha i}$ might be $1/s_\alpha$ or $1/(s_{\alpha i} t_\alpha)$ for those counting rules which link elements to enumeration units in more than one PSU. For counting rules which link elements to enumeration units in only one PSU, the $z_{\alpha i}$ might be set equal to $1/s_{\alpha i}$, and for conventional counting rules, the $z_{\alpha i}$ would be equal to 1. The $E_i$ are generalizations of parameters found by Sirken and Levy (1974) to be involved in the variances of

estimates obtained from multiplicity surveys, while the parameters, $Y_i^*$, although not necessarily integer valued, could be interpreted as being the "effective number" of elements linked to PSU i by the enumeration rule. It can be shown that $\sum_{i=1}^{M} Y_i^* \equiv Y$

Using nomenclature similar to that in Hansen, Hurwitz, and Madow (1953), we define for the variable, $\lambda'_{ij}$, the between PSU variance

$$S_{1Y}^{2*} = \sum_{i=1}^{M} (Y_i^* - \bar{Y})^2 \; / \; (M - 1)$$

the within PSU variance,

$$S_{2Y}^{2*} = \frac{1}{\bar{N}} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (\lambda'_{ij} - \bar{Y}_i^*)^2$$

and the intra-class correlation coefficient

$$\delta = \left( \frac{M-1}{M} S_{1Y}^{2*} - \bar{N} S_{2Y}^{2*} \right) / \left( \frac{M-1}{M} S_{1Y}^{2*} + \bar{N} (\bar{N}-1) S_{2Y}^{2*} \right)$$

where

$$\bar{Y} = \sum_{i=1}^{M} Y_i^* / M = Y/M \text{ (since } \sum_{i=1}^{M} Y_i^* = Y\text{)} ,$$

and

$$\bar{N} = \sum_{i=1}^{M} N_i / M .$$

With these definitions, the following theorems can be proved.

Theorem 1.

The variance, $S_{1Y}^{2*}$ among PSU's with respect to $\lambda'_{ij}$ is equal to the expression given by:

$$S_{1Y}^{2*} = \left[ M \bar{Y} (E_k - \bar{Y}) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{\alpha\alpha'} \right] / (M-1).$$

where

$$E_k = \sum_{\alpha} \sum_{i} z_{\alpha i}^2 \; s_{\alpha i}^2 \; / \; Y$$

and

$$v_{\alpha\alpha'} = \sum_{i} z_{\alpha i} \; z_{\alpha' i} \; s_{\alpha i} \; s_{\alpha' i} \; ; \; \alpha' \neq \alpha.$$

PROOF

$$\sum_{i=1}^{M} (Y_i^* - \bar{Y})^2 = \sum_{i=1}^{M} (Y_i^*)^2 - M \bar{Y}^2$$

$$= \sum_{i=1}^{M} \left( \sum_{j=1}^{N_i} \sum_{\alpha=1}^{Y} z_{\alpha i} \; \delta_{\alpha ij} \right)^2 - M \bar{Y}^2$$

$$= \sum_{i=1}^{M} \left( \sum_{\alpha=1}^{Y} z_{\alpha i} \; s_{\alpha i} \right)^2 - M \bar{Y}^2$$

$$= \sum_{i=1}^{M} \sum_{\alpha=1}^{Y} z_{\alpha i}^2 \; s_{\alpha i}^2 + \sum_{i} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} z_{\alpha i} z_{\alpha' i} s_{\alpha i} s_{\alpha' i} - M \bar{Y}^2$$

$$= M \bar{Y} \; (E_k - \bar{Y}) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{\alpha\alpha'}$$

q.e.d.

Theorem 2.

The within PSU variance with respect to $\lambda'_{ij}$ is given by

$$S_{2Y*}^2 = \frac{1}{N} \sum_{i=1}^{M} \frac{N_i^2}{N_i - 1} \bar{Y}_i^* \; (E_i - \bar{Y}_i^*) + \frac{1}{N} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{i\alpha\alpha'}$$

where $v_{i\alpha\alpha'} = \sum_{i=1}^{N_i} z_{\alpha i} \; z_{\alpha' i} \; \delta_{\alpha ij} \; \delta_{\alpha' ij}$ for $\alpha' \neq \alpha$

PROOF

$$\sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (\lambda'_{ij} - \bar{Y}_i^*)^2$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left( \sum_{j=1}^{N_i} (\lambda'_{ij})^2 - N_i (\bar{Y}_i^*)^2 \right)$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left( \sum_{j=1}^{N_i} \left( \sum_{\alpha=1}^{Y} z_{\alpha i} \; \delta_{\alpha ij} \right)^2 - N_i (\bar{Y}_i^*)^2 \right)$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left( \sum_{j=1}^{N_i} \sum_{\alpha=1}^{Y} z_{\alpha i}^2 \delta_{\alpha ij}^2 + \sum_{j=1}^{N_i} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} z_{\alpha i} z_{\alpha' i} \delta_{\alpha ij} \delta_{\alpha' ij} - N_i (\bar{Y}_i^*)^2 \right).$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left( \sum_{\alpha=1}^{Y} z_{\alpha i}^2 s_{\alpha i} + \sum_{i=1}^{N_i} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} z_{\alpha i} z_{\alpha' i} \delta_{\alpha i j} \delta_{\alpha' i j} - N_i (\bar{Y}_i^*)^2 \right)$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left( N_i \bar{Y}_i^* (E_i - \bar{Y}_i^*) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{i\alpha\alpha'} \right)$$

q.e.d.

Corollary 1.

The intra class correlation coefficient, $\delta$, is given by

$$\delta = \frac{A - B}{A + (N-1)B}$$

where

$$A = M \bar{Y} (E_k - \bar{Y}) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{\alpha\alpha'}$$

and

$$B = \sum_{i=1}^{M} \frac{N_L^2}{N_i - 1} \bar{Y}_i^* (E_i - \bar{Y}_i^*) + \sum_{i=1}^{M} \frac{N_L^2}{N_i - 1} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{i\alpha\alpha'}$$

Proof follows from Theorems 1 and 2 and the definition of $\delta$.

q.e.d.

Corollary 2.

If the assumption is made that an enumeration unit is linked to no more than one element then $v_{i\alpha\alpha'} = 0$ for all i, $\alpha$, and $\alpha'$; $S_{2Y}^{2*}$ is given by

$$S_{2Y}^{2*} = \frac{1}{N} \sum_{i=1}^{M} \frac{N_i^2}{N_i - 1} (E_i - \bar{Y}_i^*)$$

and

$$\delta = \frac{A - B}{A + (N-1)B}$$

where

$$A = M \bar{Y} (E_k - \bar{Y}) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{\alpha\alpha'}$$

and

$$B = \sum_{i=1}^{M} \frac{N_i^2}{N_i - 1} \bar{Y}_i^* (E_i - \bar{Y}_i^*)$$

## 2.2 Estimation of Y, the Total Number of Events from the Sample

Let us assume that the sample design used to estimate the number of events, Y, in the population is a simple two stage cluster sample as defined by Hansen, Hurwitz, and Madow (1953). In other words, a simple random sample of m PSU's is taken from the M PSU's in the population, and within each sample PSU (i), a simple random sample of $n_i$ enumeration units is taken from the $N_i$ enumeration units in the PSU, with the second stage sampling fraction, $n_i/N_i$, the same for each PSU.

If (for convenience) the sample PSU's are labelled 1, ...., m and the sample enumeration units within each sample PSU are denoted $i_1$, ..., $i_{n_i}$ where i = 1, ..., m, then the estimator Y' of Y as given by

$$Y' = \frac{M}{m} \sum_{i=1}^{m} \frac{N_i}{n_i} \sum_{j=1}^{n_i} \lambda'_{ij} \quad (1)$$

is an unbiased estimator of Y as shown below in Theorem 3.

Theorem 3.

The estimator, Y' of Y as defined in equation ( 1 ) is an unbiased estimator of Y.

PROOF

For a given sample PSU, i, the expected value over all possible second stage samples of $\sum_{j=1}^{n_i} \lambda'_{ij}$ is given by

$$E\left(\sum_{j=1}^{n_i} \lambda'_{ij}\right) = \frac{n_i}{N_i} \sum_{j=1}^{N_i} \lambda'_{ij} = \frac{n_i}{N_i} Y_i^*$$

Thus, the expected value of Y' over all possible samples is given by

$$E(Y') = \frac{M}{m} \underset{\text{all } i}{E} \left( \sum_{i=1}^{M} \frac{N_i}{n_i} \underset{\text{all } j}{E} \left( \sum_{j=1}^{n_i} \lambda_{ij} \right) \right)$$

$$= \frac{M}{m} \frac{m}{M} \sum_{i=1}^{M} Y_i^*.$$

$$= \sum_{i=1}^{M} Y_i^*.$$

$$= Y$$

q.e.d.

## 3. Some Relationships Involving $\delta$ When all $v_{i\alpha\alpha'} = 0$

When $v_{i\alpha\alpha'} = 0$ for all $\alpha, \alpha'$ and for all i, then the intraclass correlation coefficient, $\delta$, is given by

$$\delta = (A - B) \;/\; (A + (\bar{N} - 1)\, B)$$

where (assuming $N_i / (N_i - 1) \doteq 1$)

$$A = M \bar{Y} (E_k - \bar{Y}) + \sum_{\substack{\alpha \\ \alpha' \neq \alpha}} \sum v_{\alpha\alpha'}$$

and

$$B = \sum_{i=1}^{M} N_i \bar{Y}_i^* (E_i - \bar{Y}_i^*) .$$

Clearly $A \geq 0$ and $B \geq 0$ since they are quadratic forms. It can be shown that since $B > 0$

$$\frac{d\,\delta}{d\,A} = \frac{B}{[A + (\bar{N} - 1)\, B]^2} \geq 0 ,$$

and therefore, $\delta$ varies directly with A.

On the other hand, since $A > 0$, then

$$\frac{d\,\delta}{d\,B} = \frac{-\bar{N} A}{[A + (\bar{N} - 1)\, B]^2} \leq 0$$

and hence $\delta$ varies inversely with B.

Let us examine $\delta$ for the set of weights $z_{\alpha i} = 1/(t_\alpha s_{\alpha i})$ ; $\alpha = 1, \ldots, Y$; $i = 1, \ldots, M$. For this set of weights, we have:

$$E_i = \sum_{\alpha=1}^{Y} \frac{t_{\alpha i}}{t_\alpha^2 s_{\alpha i}} \;/\; \sum_{\alpha=1}^{Y} \frac{t_{\alpha i}}{t_\alpha} ,$$

$$\bar{Y}_i^* = \frac{1}{N-1} \sum_{\alpha=1}^{Y} \frac{t_{\alpha i}}{t_\alpha} ,$$

$$E_k = \sum_{\alpha=1}^{Y} \frac{1}{t_\alpha} \;/\; Y$$

and

$$\sum_{\substack{\alpha \\ \alpha' \neq \alpha}} \sum_{\alpha'} v_{\alpha\alpha' i} = \sum_{\alpha} \sum_{\alpha'} \sum_{i} \frac{t_{\alpha i}\, t_{\alpha' i}}{t_\alpha\, t_{\alpha'}}$$

If the multiplicities, $s_{\alpha i}$, are increased without increasing the $t_\alpha$ or $t_{\alpha i}$, then the $E_i$ would decrease which would cause a decrease in B since the $\bar{Y}_i^*$ would be unaffected. Since, also, the $v_{\alpha\alpha'}$, and $E_k$ would not be affected by change in the $s_{\alpha i}$, it follows that A would not be affected. Hence, <u>increase</u> in $s_{\alpha i}$ would result in an <u>increase</u> in the intra-class correlation coefficient, $\delta$.

## REFERENCES


1. Sirken, M.G. (1970): "Household Surveys with Multiplicity" <u>Journal of the American Statistical Association</u> <u>65</u>, 257 - 266.

2. Sirken, M.G. (1972): "Stratified Sample Surveys with Multiplicity" <u>Journal of the American Statistical Association</u> <u>67</u>, 224 - 227.

3. Levy, P. S. (1978): "Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence of Attributes in Rare Populations" To appear in <u>Journal of the American Statistical Association, December, 1977.</u>

4. Sirken, M.G. and Levy, P.S. (1974): "Multiplicity Estimation of Proportions Based on Ratios of Random Variables" <u>Journal of the American Statistical Association</u> <u>69</u>, 68 - 73.

5. Sirken, M. G., Indurfurth, G. P., Burnham, C. E., and Danchik, K. M. (1975): "Household Sample Surveys of Diabetes: Design Effects of Counting Rules" <u>American Statistical Association, Proceedings of the Social Statistics Section</u>, 659 - 663.

6. Nathan, G., Schmelz, U. O. and Kenvin, J. (1977): "Multiplicity Study of Marriages and Births in Israel" <u>Vital and Health Statistics, Series 2</u> No. 70, NCHS, Rockville, Maryland.

7. Nathan, G. (1976): "An Empirical Study of Response and Sampling Errors for Multiplicity Estimates with Different Counting Rules" <u>Journal of the American Statistical Association 71,</u> 808 - 815.