OPTIMUM ALLOCATION IN STRATIFIED RANDOM MULTIPLICITY SAMPLING

Paul S. Levy
University of Illinois at the Medical Center
School of Public Health

This work considers estimation of dichotomous characteristics of rare populations by sample survey. In particular, we investigate optimum allocation when a network sampling rule (also called multiplicity rule) is used to link members of the rare population to enumeration units and when the enumeration units are chosen by stratified random sampling. For this situation, formulae are derived for optimum allocation under the restriction that an enumeration unit is linked to at most one member of the rare population, and the cost efficiency of network sampling rules is compared to that of conventional enumeration rules.

1. Introduction

This report is concerned with the estimation of the prevalence of dichotomous attributes among members of "rare" population groups where a "rare" population group is considered here to be one which consists of less than 3% of the total population. In the field of health statistics, we often are interested in making inferences about rare populations. For example, approximately 1% of the total U.S. population dies during a given year, and hence, decedents represent a rare population group. Another common problem is that of association between two disease conditions. For example, we may be interested in determining whether gout occurs more frequently among diabetics than among non-diabetics. Since recent surveys have estimated the prevalence of diabetes to be in the neighborhood of 3.0 per 100 persons [6], diabetics constitute a rare population group.

Often, characteristics of rare populations are examined not from random samples of the rare population groups but from "collections" of individuals from these groups selected for convenience. For example, association between two diseases is often examined from hospitalized patients and, as Berkson has pointed out [1], the same association between two conditions found among hospitalized patients may not exist in the general population. Thus, in order for valid inferences to be made concerning rare population groups, one is often necessarily confronted with the problem of seeking them out by sample survey.

The problem of locating members of rare population groups by sample survey for purposes of estimating their characteristics has received recent attention among survey statisticians. Sudman [7], suggested the use of stratification, Bayesian optimum allocation and sequential analysis to increase cost efficiency. Sirken, on the other hand, has developed the theory of network sampling (discussed below) as a tool for measurement of characteristics in rare populations [3], and has used it in a wide variety of applications [4], [5], [6].

The concepts, formulation and notation used here are similar to what was used by Sirken and Levy [5]. In particular, this work investigates again the effect which manipulation of enumeration rules (or counting rules) has on the precision of estimators derived from sample surveys. An enumeration (counting) rule is an algorithm for linking enumeration units with elements. For example, in household surveys for estimating deaths, an enumeration rule might specify that a death can be reported only in the household of the decedent. Another enumeration rule might specify that a death can be reported in the decendent's household or in the household of any sibling of the decedent. The former counting rule is an example of a conventional counting rule since it links every element (decedent) to only one enumeration unit (household), while the latter is a network counting rule since an element can be linked to more than one enumeration unit. The rationale behind network rules is that they can increase the "yield" of rare events in a sample survey which results in estimates having increased precision. In a series of papers, Sirken [3], [4], and Sirken and Levy [5], have developed expressions for unbiased estimates of characteristics and for the variances of these estimates when network counting rules are used.

In this report, methods are developed for optimum allocation in stratified random sampling when a network estimator is used to estimate the prevalence of a dichotomous attribute in a rare population group.

2. Method

Let us suppose that the rare population group consists of Y members or elements, X of which have some dichotomous attribute, A, and Y-X do not. Let us suppose further that the total population consists of L enumeration units grouped into H strata with L_h enumeration units appearing in the h^{th} stratum, h = 1,..., H, and that within each stratum, X_h elements having attribute A and $Y_h - X_h$ elements not having attribute A are linked to enumeration units in the stratum by a conventional counting rule. For convenience, we assume that within each stratum (h), elements with the labels I_{hi}, i = 1, ..., X_h have attribute A whereas elements with labels I_{hi}, i = X_h + 1,..., Y_h do not have attribute A.

Let us suppose that within each stratum, a simple random sample of ℓ_h enumeration units is

drawn for purposes of estimating the prevalence, X/y of attribute A among members of the rare population. For each stratum, enumeration unit specific prevalence rates are given by:

$$R_{hx} = X_h / L_h$$

and

$$R_{hy} = Y_h / L_h$$

Likewise, for each stratum we define

$$\pi_h = L_h / L$$

$$\gamma_h = X_h / Y_h = R_{hx} / R_{hy}$$

and over all strata we define

$$R_y = Y/L$$

$$R_x = X/L$$

and

$$\gamma = X/Y = R_x / R_y$$

In this formulation, we will assume that within each stratum, the counting rule used specifies that each of the Y_h elements of the rare population is linked to at least one of the L_h enumeration units in the stratum but to no enumeration unit in another stratum. If the counting rule used is a conventional one, then each element would be linked to one and only one enumeration unit in the same stratum. On the other hand if the counting rule used is a network sampling rule, then an element may be linked to more than one enumeration unit. In addition, whether the counting rule used is a conventional or network rule, we place the important restriction that no enumeration unit is linked to more than one element of the rare population.

From the simple random sample of ℓ_h enumeration units in each stratum, we obtain the estimate γ' of γ given by:

$$\gamma' = \left(\sum_{h=1}^{H} L_h \, x_h/\ell_h \right) / \left(\sum_{h=1}^{H} L_h \, y_h/\ell_h \right) \qquad (1)$$

where

$$y_h = \sum_{j=1}^{\ell_h} (\lambda'_{hi_j} + \lambda''_{hi_j})$$

$$x_h = \sum_{j=1}^{\ell_h} \lambda'_{hi_j}$$

$$\lambda'_{hi} = \sum_{\alpha=1}^{X_h} \delta_{h\alpha i} / W_{h\alpha},$$

$$\lambda''_{hi} = \sum_{\alpha=X_h+1}^{Y_h} \delta_{h\alpha i} / W_{h\alpha},$$

$\delta_{h\alpha i}=1$ if element $I_{h\alpha}$ is linked to enumeration unit i by the counting rule, $\delta_{h\alpha i} = 0$ otherwise, $i_1,\ldots,i_{\ell_h}$ represent the indices of enumeration units chosen in the sample,

and $W_{h\alpha} = \sum_{i=1}^{L_h} \delta_{h\alpha i}$ = the multiplicity of element $I_{h\alpha}$ (i.e., the number of enumeration units linked to element $I_{h\alpha}$ by the counting rule).

Since γ' is a ratio estimate based on stratified random sampling, its approximate variance is given by:

$$\sigma_{\gamma'}^2 = \left(\sum_{h=1}^{H} \pi_h^2 \, S_{hz'}^2 \, (L_h-\ell_h)/(\ell_h \, L_h) \right) / R_y^2 \qquad (2)$$

where

$$S_{hz'}^2 = S_{hx}^2 + \gamma^2 \, S_{hy}^2 - 2\gamma \, S_{hxy}$$

and where S_{hx}^2, S_{hy}^2 and S_{hy} are within stratum variance and covariances with respect to the distribution of elements having attribute A and not having attribute A among enumeration units in the stratum.

For enumeration rules based on network sampling in which an enumeration unit is linked to at most one element, Sirken and Levy [5] have shown that:

$$S_{hx}^2 = R_{hx} \, (E_{hx} - R_{hx})$$

$$S_{hy}^2 = R_{hy} \, (E_{hy} - R_{hy}) \qquad (3)$$

and

$$S_{hxy} = R_{hx} \, (E_{hx} - R_{hy})$$

where

$$E_{hx} = \left(\sum_{\alpha=1}^{X_h} 1 / W_{h\alpha} \right) / X_h$$

and

$$E_{hy} = \left(\sum_{\alpha=1}^{Y_h} 1 / W_{h\alpha} \right) \Big| \, Y_h \, .$$

Thus, $S_{hz'}^2$ is given by:

$$s^2_{hz'} = R_{hx}(E_{hx} - R_{hx}) + \gamma^2 R_{hy}(E_{hy} - R_{hy}) - 2\gamma R_{hx}(E_{hx} - R_{hy}) \quad (4)$$

3. Optimum Allocation

Let us suppose that the average cost per enumeration unit in stratum h is given by the equation:

$$c_h = c_1 + c_2 R_{hy} \bar{W}_{hy} \quad (5)$$

where c_1 is the cost component associated with screening a sample enumeration unit and determining whether it is linked to a member of the rare population, c_2 is the cost component assocaited with interviewing the member or members of the rare population linked to the sample enumeration unit, determining whether the individual has attribute A, and determining the total number of enumeration units linked to the individual (i.e., determining the multiplicity, $W_{h\alpha}$, of individual $I_{h\alpha}$), and

$\bar{W}_{hy} = \sum_{\alpha=1}^{Y_h} W_{h\alpha} / Y_h$ is the average multiplicity of elements in stratum h. Then, the total cost, c, for a survey of ℓ_h enumeration units in stratum h, h = 1, ..., H, is given by:

$$c = c_1 \ell_1 + c_2 \ell_2 + \ldots + c_H \ell_H \quad (6)$$

Where c_h is given by equation (5) for h = 1, ..., H.

With cost function of the form given by equation (6), it is a well known result [2], that optimum allocation $\tilde{\ell}_h$, of sample to strata at fixed total cost, c, is given by:

$$\tilde{\ell}_h = c\pi_h \tilde{S}_h / \sqrt{c_h} \Big/ \sum_{h=1}^{H} \pi_h \tilde{S}_h \sqrt{c_h} \quad (7)$$

where

$$\tilde{S}_h = S_{hz'} / R_y \quad (8)$$

By substitution of (4), (5), and (8) into equation (7), we can obtain an explicit formula for $\tilde{\ell}_h$, and for a conventional rule, if we assume that the cost component, c_2, is the same as that for multiplicity rule, this explicit formula is given by:

$$\tilde{\ell}_h = \frac{c\pi_h \sqrt{R_{hy}} \sqrt{a_h} \Big/ \sqrt{c_1 + c_2 R_{hy}}}{\sum_h \pi_h \sqrt{R_{hy}} \sqrt{a_h} \Big/ \sqrt{c_1 + c_2 R_{hy}}} \quad (9)$$

where

$$a_h = \gamma_h (1 - \gamma_h R_{hy}) + \gamma (1 - R_{hy})(\gamma - 2\gamma_h) .$$

If $\gamma_h = \gamma$ for all h, the explicit formula is given by:

$$\tilde{\ell}_h = \frac{c\pi_h \sqrt{R_{hy}} \sqrt{E_{hx} - \gamma E_{hy}} \Big/ \sqrt{c_1 + c_2 \bar{W}_{hy} R_{hy}}}{\sum_h \pi_h \sqrt{R_{hy}} \sqrt{E_{hx} - \gamma E_{hy}} \Big/ \sqrt{c_1 + c_2 \bar{W}_{hy} R_{hy}}} \quad (10)$$

4. Comparison of Variances Under Different Counting Rules

If we make the simplifying assumptions that $E_{hx} = E_{hy} = E$, that $\bar{W}_{hy} = \bar{W}$ and that $\gamma_h = \gamma$ for h = 1, ..., H, it can be shown that the formula for the approximate variance of γ' (equation (3)) for the optimum allocation, $\hat{\ell}_h$ at cost c reduces to the form given by:

$$\sigma^2_{\gamma'} = \frac{\gamma(1-\gamma) E}{cR_y^2} \left[\sum_h \pi_h \sqrt{R_{hy}} \sqrt{c_1 + c_2 \bar{W} R_{hy}}\right]^2 \quad (11).$$

If we assume that the cost component, c_2, remains the same whether a conventional or network sampling rule is used, then the variance of γ' at total cost, c, for a conventional counting rule is given by equation (3) with E and $\bar{W}$ set equal to unity. Thus, the ratio of the variance γ' under network sampling to that under a conventional rule at the same total cost is given by:

$$\frac{E\left[\sum_{h=1}^{H} \pi_h (1 + c_2 \bar{W} R_{hy} / c_1)^{\frac{1}{2}}\right]^2}{\left[\sum_{h=1}^{H} \pi_h (1 + c_2 R_{hy} / c_1)^{\frac{1}{2}}\right]^2} \quad (12)$$

For most values of c_2, c_1, and $\bar{w}$ likely to encountered in practice, expression (12) will be approximately equal to E, and hence, for most applications, network sampling will result in estimates having lower variance at the same total cost than estimates based on a conventional counting rule.

5. Discussion

Although methodology for the use of network enumeration rules in stratified random sampling has been developed previously [4], this is the first work in which optimum allocation has been considered for network sampling. The scope of this work, however, is limited to the estimation of dichotomous attributes (proportions), and the methodology developed here depends on several restrictions which will be reiterated below.

Perhaps the most important restriction on the results developed here is to counting rules in which each enumeration unit is linked to at most one element. If this restriction were not

made, algebraic expressions for the variances of estimates under network sampling rules become extremely complex and involved [3], [5]. Whether or not this restriction makes sense depends, of course, on the particular enumeration rule and the particular application. For example, in surveys for the prevalence of diseases which have a genetic component, one might expect a clustering of elements among particular enumeration units in which case, the expressions for optimum allocation developed here would not be valid. In other cases, the results developed here seem safe to use.

The other restriction on enumeration rules imposed here is that an element can be linked to enumeration units only in the same stratum. While this restriction is easy to impose, it could also decrease the yield of elements obtained in a sample and hence increase the sampling variance of the estimator.

For each stratum, the average cost per sample enumeration unit $c_1 + c_2 R_{hy} \bar{W}_{hy}$ depends on three factors, namely c_1, the cost of screening an enumeration unit for purposes of determining whether any elements of the rare population unit are linked to it, c_2, the cost of obtaining the desired information concerning a sample element including the multiplicity of the element and $R_{hy} \bar{W}_{hy}$, the average number of elements per enumeration unit for a particular counting rule in a stratum. Although it was assumed here that the cost components, c_1 and c_2 are independent of the particular enumeration rule used, it should be recognized that this might not be a safe assumption to make when the process of determining the multiplicity of an element is complex or difficult and adds considerably to the cost.

Perhaps the most interesting result is that obtained for the ratio of the sampling variance of γ' under a network sampling rule to that under a conventional counting rule in the situation where γ_h is the same for all strata and the multiplicity parameters are the same over all strata (expression (12)). This ratio is equal to E, the inverse harmonic mean of the multiplicities multiplied by a factor which in most situations will be only slightly greater than unity. Thus, the parameter E, would give a rough indication of the extent to which the sampling variance can be reduced by use of a network counting rule over a conventional rule at equivalent field costs.

Finally, it should be mentioned that the important problem of measurement error in determining the multiplicities of elements, which was not treated here should be taken into consideration in choosing a particular counting rule, since errors in the multiplicities could introduce serious biases.

References

1. Berkson, J., "Limitation of the Application of Fourfold Table Analysis to Hospital Data," Biometrics Bulletin, 2, 47-53, 1946.

2. Hansen, Hurwitz, and Madow, "Sample Survey Method and Theory," Volume 1, John Wiley and Sons, New York, 1953.

3. Sirken, M.G., "Household Surveys with Multiplicity," Journal of the American Statistical Association, 65, 257-266, 1970.

4. Sirken, M.G., "Stratified Sample Surveys with Multiplicity," Journal of the American Statistical Association, 67, 224-227, 1972.

5. Sirken, M.G., and Levy, P.S., "Multiplicity Estimation of Proportions Based on Ratios of Random Variables," Journal of the American Statistical Association, 69, 68-73, 1974.

6. Sirken, M.G., Inderfurth, G.P., Burham, C.E., and Danchik, K.M., "Household Sample Survey of Diabetes: Design Effects of Counting Rules," American Statistical Association, Proceedings of the Social Section, 659-663, 1975.

7. Sudman, S., "On Sampling Very Rare Human Populations, Journal of the American Statistical Association, 67, 335-339, 1972.