A TECHNIQUE FOR REDUCING SURVEY SAMPLING SIZE*

R. Hoffman**, E. Gomolka***, and H. Phillips****

> What can we do when the survey must produce several, perhaps many, variates? This calls for caution before resorting to disproportionate sampling... If the different answers point toward the same design, there is no great problem. But if the answers are contradictory, the choice involves good judgment and theory that is beyond existing theoretical developments.[1]

Introduction

This paper briefly describes the increasing need for the use of survey type sampling in the development and evaluation of action research projects. The standard methods appear to have some limitations when one is faced with stratum or cells within the social survey which have a finite (often very small) number of possible observations and/or surveys where there are a large number of attributes which one may wish to sample.

A method and accompanying computer based procedure is presented. The procedure specifies the sample size to be taken for each cell without regard to whether the particular cell is sampled.

It is shown that the method presented results in a lower total sample than fixed sample size methods and that the method converges to fixed sample size when the number of possible observations in each cell becomes very large. Finally, an extension to the method using an economical pilot sample, and the basis for the selection of acceptable sample size for the cells which are used, is discussed as a corollary result.

The Problem

The objective of the social survey is primarily to describe a population in order to make inferences relating to phenomena which may or may not exist within the population. Typically, social survey techniques have made use of base data, such as census data, which are regularly collected by Federal agencies. Often, however, in action-oriented research programs conducted by the government and universities, the data required to evaluate these programs is very rarely available from the standard sources. Thus, the researcher must survey the target population for the data needed to evaluate the program. Similar problems often occur when attempting to obtain information with which to develop new programs. Hence, in evaluating or developing new government programs, the researcher must often collect his own data. But with the increased need for social survey sampling, greater attention will be required for the development of sampling techniques which deal with

*The research of the first author was supported in part by a Public Health Service grant to the Center for Planning and Development Research, and a Department of Transportation grant (Cal. MTD-11) to the School of Criminology, both of the University of California at Berkeley. The authors received additional support from the School of Management, State University of New York at Buffalo.

**Department of Management Science, State University of New York at Buffalo and Visiting, School of Criminology, University of California at Berkeley.

***School of Management, State University of New York at Buffalo.

**** Department of Management Science, State University of New York at Buffalo.

1. Kish, Leslie, Survey Sampling, New York (Wiley), 1967, pp.96-97.

2. See, P.G. Gray and T. Corlett, "Sampling for the Social Survey," Journal of the Royal Statistical Society, Part II, pp.150-199 (1950), for a basic discussion of multi-stage procedures for social survey sampling when the survey is to be used as an alternative to, or to update, a census. Similar procedures are appropriate for the problems we discuss.

cases where the usual assumptions are not appropriate.[3]

Two cases where significant problems may arise with the use of the usual techniques are: 1)a finite number of possible observations within a cell or stratum to be sampled, and 2)sampling of populations to obtain information on relatively rare events. We find many of the cells within the population may have a small number of possible observations. For example, if we are evaluating the effects of an existing poverty program it may be useful to stratify on the basis of head of household yearly income. But it is likely that in very affluent suburbs we may find in these strata there are very few individuals with incomes below $3000. Alternatively, a converse situation will hold in ghetto neighborhoods. However, it is often very important to obtain information on these unusual situations. A second source of finiteness is that the events of interest may be relatively rare, e.g., murder and rape in a given precinct in a given city. Hence, the cost of sampling (including the questionnaire cost and the cost of transmitting the interviewer to the area (load cost) becomes a major problem, when one considers the number of interviews which are often necessary to obtain a relevant sample.

The methods are, of course, quite standard if one wishes to obtain data from the population on but one attribute. However, this is not the typical case. If we survey a large population for one attribute, the value of this information will most often not be sufficient to justify the costs of sampling. Additionally, one may not know a priori those attributes which will be of importance.[4] Hence, the "optimal method" of Neyman as described in Mills[5] results in difficulty since, with its use, sample size will be functionally related to the variance of but one attribute.

However, a prevalent case upon which we focus our attention has a finite (and many times very small) number of possible observations in many of the strata or cells. When this is the case, proportional sampling implies that one may sample say 40 of a population of 400 houses in one subdivision but 1 of a population of 8 houses in another subdivision. Clearly, the sample of 40 may give us very acceptable results (if we wish to obtain inferences relating to the subdivision containing 400 homes), but the sample of 1 out of 8 very likely will not.[6]

The obvious alternative is to use a fixed sample size. When this is done, we may find that the appropriate sample size, say 30, may leave us with a sample size greater than population size.[7] Clearly, the information obtained from a complete enumeration of the attributes of the population will will exceed that obtained from any sample of the population.[8] Complete enumeration of some cells results in oversampling of such cells relative to

---

3. See, D. Raj, Sampling Theory, New York: McGraw-Hill, (1968), p.36, for a standard discussion. However, when there are but a finite possible number of observations the sampling fraction n_i/N_i is not small, hence the assumptions relating to the dependence of the variance of mean upon n(sample size) and σ^2(variance) are not appropriate.

4. As a collection of 300 attributes or more is not unusual, it will be necessary to obtain information on many attributes to determine the crucial variables.

5. See, F.C. Mills, Statistical Methods, 3rd Ed., New York: Holt, Rinehart, and Winston, (1955), p.180. The other alternatives proportional and fixed sample size, can be used.

6. The arguments in favor of the proportional sampling method have often been based upon ease of computation; e.g., Kish, op.cit., Chapter 3. However, as an appropriate weighting scheme can always be found and with the availability of high-speed computing equipment, the extra costs of weighting and special handling are minimal.

7. See previous example for such a case: e.g., total subdivision population of 8 houses.

8. Raj, op.cit., p.27, implies that complete enumeration is appropriate if each cell is to be sampled (i.e., no savings in load costs) but good sampling procedure will still result in lower costs with little if any loss of precision. Also, using standard techniques such as those described in Raj, op.cit., p.81 and p.206, it may not be necessary to sample all cells.

those cells in which but a fraction are taken. Hence, when one has a finite number of possible observations in a cell, if a method can be found which results in equal precision and reliability for each cell, it is possible to reduce the sample size in the cells with a finite possible number of observations. This, of course, would result in lowered sampling costs.

The Method and Computer Based Procedure

The method for the selection of the sample size for each cell can be developed using the basic homoscedasticity assumption.

This method is derived from the basic equations for the standard error.

Here S_x^2 is the unbiased estimator of σ_x^2. Then we obtain n(sample size) from the standard relationship between precision and reliability.

$$e^2 = z^2 \sigma_x^2 \quad [9]$$

Substituting s_x^2 for σ_x^2

$$e^2 = z^2 s_x^2 = z^2 \frac{s^2}{n}$$

Solving for N and applying the exact form of the finite population correction factor[10] (fpc), as it is appropriate in the cases of interest.

$$n = \frac{z^2 x^2}{e^2} \cdot \left(1 - \frac{n-1}{N-1}\right)$$

We define $C_1 = \frac{s^2}{e^2}$, where we accept the standard homoscedasticity assumption.

Then, for populations with a finite number of observations, the "t" distribution is utilized. This is necessary, as quite often the sample size will be small.

---

9. T. Yamane, Elementary Sampling Theory, Englewood Cliffs, N.J.: PrenticeHall, Inc., 1967, p.81.

10. Raj, op.cit., p.36, 1- = fpc

$$n = C_1 t^2 \left(1 - \frac{(n-1)}{(N-1)}\right)$$

$$n(N-1) = C_1 t^2 (N-1) - C_1 t^2 (n-1)$$

$$n(N-1+C_1 t^2) = C_1 t^2 (N-1) + C_1 t^2$$

$$n = \frac{C_1 t^2 + N}{C_1 t^2 + (N-1)}$$

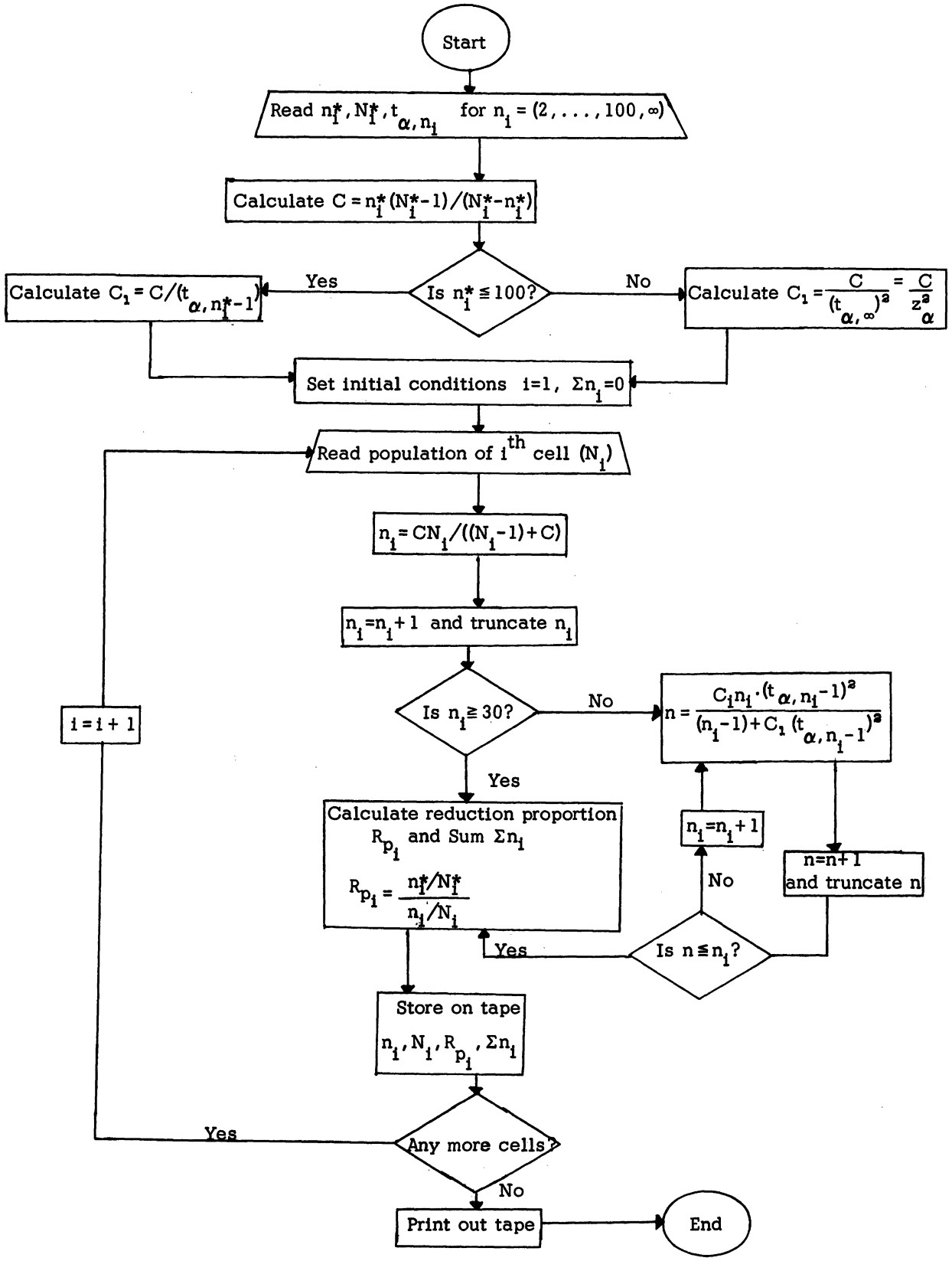
if n sufficiently large then $t^2 \cong z^2$

and $C_1 z^2 = e$

$$n = \frac{CN}{C + (N-1)}$$

If the researcher is willing to specify an acceptable sample size (n_i^*) for any cell, then given the N_i (population size) for all other cells it is possible to specify the n_i for all other cells. The method uses 1) n_i^*, 2) N_i^* (population size) of cell i^*, and 3) a range of values from the "t" distribution for "t" at a researcher specified level of alpha (α).

The computer based procedure is described in the accompanying flow chart (on the following page).


Start
Read $n_i^*, N_i^*, t_{\alpha, n_i}$ for $n_i = (2, \ldots, 100, \infty)$
Calculate $C = n_i^*(N_i^*-1)/(N_i^*-n_i^*)$
Is $n_i^* \leq 100$?
Yes
No
Calculate $C_1 = C/(t_{\alpha, n_i^*-1})$
Calculate $C_1 = \frac{C}{(t_{\alpha,\infty})^2} = \frac{C}{z_\alpha^2}$
Set initial conditions i=1, $\Sigma n_i = 0$
Read population of i^{th} cell (N_i)
$n_i = CN_i/((N_i-1)+C)$
$n_i = n_i + 1$ and truncate n_i
Is $n_i \geq 30$?
No
$n = \frac{C_1 n_i \cdot (t_{\alpha, n_i-1})^2}{(n_i-1)+C_1 (t_{\alpha, n_i-1})^2}$
i = i + 1
Yes
Calculate reduction proportion R_{p_i} and Sum Σn_i
$R_{p_i} = \frac{n_i^*/N_i^*}{n_i/N_i}$
$n_i = n_i + 1$
No
n=n+1 and truncate n
Is $n \leq n_i$?
Yes
Store on tape $n_i, N_i, R_{p_i}, \Sigma n_i$
Any more cells?
Yes
No
Print out tape
End

Conclusion and Extension

The method can be described simply. If one is willing to accept a sample size (n*) for any cell in a stratified and/or multi-stage sample, then with the standard homoscedasticity assumptions, we can specify the sample size for all cells which will result in a total sample size which is usually less, and never greater, than the total sample size obtained from a fixed sample size procedure of equivalent effectiveness.

It can be shown where:

$N_i \to \infty$ (actually very large N_i)

then

$n_i \to n_f$ (fixed sample size)[11]

since

$$n_i = \frac{C \cdot N_i}{(N_i - 1) + C}$$

the one can show as $N \to \infty$ (or is very large)

$$n_i = \frac{C \cdot N_i}{(N_i - 1) + C} \to \frac{C \cdot N_i}{N_i + C} \to \frac{C \cdot N_i}{N_i} \to C = n_f$$

$n_i \cong C$ and a constant, i.e., fixed sample size.

A very interesting extension relates to the use of a pilot sample, say in a cell with relatively low load costs.[12] If on the basis of the pilot cell sample's variances one is willing to specify a sample size which would be acceptable for the pilot cell, then it is possible to specify the sample size for all cells which could be in the grand sample. This result only requires that one again accept the standard homoscedasticity assumption. It should be noted that it is not necessary that the pilot sample be included in the final grand sample.[13]

---

11. Obviously there exists a fixed sample size which would be acceptable.

12. Perhaps, in a nearby jurisdiction, prison, or housing subdivision.

13. This will occur rather often in practice, as the probability of the pilot cell being in the final strata or cell is a function of the ratio of the pilot population size to the size of the grand population,

$$\Pr(\text{pilot sample in grand sample}) = \frac{n_i}{N} \cdot \frac{N_i\,(\text{pilot})}{N}$$