SOME ALTERNATIVES TO PROBABILITY PROPORTIONAL TO SIZE AND "QUOTA" OF ELEMENTARY UNITS FOR TIGHT BUDGETS IN TWO- OR MULTI-STAGE CLUSTER SAMPLING DESIGN

Francis C. Madigan, S.J.
Research Institute for Mindanao Culture, Xavier University
P.O. Box 24, 9000 Cagayan de Oro City

ABSTRACT

In two- and multi-stage cluster sampling designs, researchers often resort to a selection of clusters with probability proportional to size, a non-random selection of ultimate units (households) and a "quota" of elementary units (respondents) due to large objectives and small funds. Concerned with the potential problems with this approach, the paper proposes four "resort steps" when clusters are selected with equal probability as alternatives. The first two steps utilize a ratio-mean approach with selection of unequal clusters in the first stage by equal probability. The last two steps change the sampling design and present an innovative approach to handling probability estimation.

INTRODUCTION

Research directors often find themselves in the situation of small funding to accomplish large research objectives. In such cases a truly probability sampling is likely to be sacrificed to cover other budgetary needs like travel and wages of field personnel. The first of usual solutions is cluster sampling consisting of: (1) selection of clusters with probability proportional to size (PPS); (2) non-random selection of sampling units (households) in the last stage of a multi-stage sampling design; and (3) "quota" of elementary units (respondents). One specific example is the World Health Organization (WHO)-Expanded Programme on Immunization (EPI) cluster sampling or "30 x 7" EPI sampling strategy (Lemeshow and Stroh, 1988). It is characterized as a two-stage PPS cluster sampling methodology without random selection at the second stage. It consists of: (1) randomly selecting 30 clusters from within each geographical area for which immunized children are desired; (2) randomly selecting a starting-point household within each sample cluster; and (3) selecting seven children from each sample cluster. Selection begins in the starting household and then

continues to the next nearest household until seven children are selected in each sample cluster.

This solution probably occurs most frequently when information on the occurrence of the main variable of interest, especially on its magnitude per cluster or site, is not available from other sources (e.g. the census, Department of Health (DOH) data, or recent surveys). Such variables might be, for instance, immunization coverage, abortion prevalence or incidence of common diseases. Regrettably, while this sampling methodology saves funds, it also has potential problems particularly in the technical aspects of estimation. With the WHO-EPI "30 x 7" sampling design for instance, the following are of important concerns (Lemeshow and Stroh, 1988: 11-12):

(1) The risk that surveys of adjacent households could either over- or under-estimate the true population coverage depending upon where the starting household happens to fall;

(2) Estimates are not self-weighting;

(3) No actual full control over the interviewer; he or she may not follow strictly the prescribed procedure of randomly selecting the initial household and selecting the successive households thus resulting into either over- or under-representation of the true population coverage;

(4) Households unoccupied at first visit being not revisited may pose inadequate representation of the true population coverage;

(5) Households being selected on grounds of convenience; and

(6) Use of nondocumented evidence of immunization status.

The second usual solution to financial constraints is still cluster sampling but clusters are selected with equal probability, and fixed percentages of sample elements in each cluster (e.g. persons recuperating from coronary thrombosis) are maintained, so as to retain equal overall selection probabilities for each household in the cluster. Some approximations of these percentages are nonetheless required either from previous surveys or other independent data sources (e.g. census or DOH data). This solution, like the first, maintains the advantage that stratum sampling fractions in stages of sampling before the last stage are the same, so that one has only differences in the last stage to contend with. However, the researcher may encounter complicated mathematics of calculating the sampling variance and therefore of testing for statistical significance especially if available cases of the desired variable differ greatly between clusters of the same strata. Formulas like those of Set 1 in the Appendix may be used for computing means and variances for this approach.

Some Other Alternatives

This paper presents other alternatives to handling limited budgets in a

two- or multi-stage cluster sampling design, with clusters selected at random with equal probability from all available clusters. The first alternative ("resort step" 1) elaborates the second usual solution noted earlier, i.e., drawing unequal-size clusters with equal probability, but in situations where available percentages of the elementary units differ among clusters.[1] It suggests post-stratification for sorting similar clusters into percentage strata in an effort to try to solve the problem of differing percentages by stratum weighting, provided that too many substrata can be avoided.[2] In effect, it is workable if the varying percentages of the elementary units are not too many and diverse.

This will be even more practical if not too many strata already existed prior to substratification, as if, e.g., only a few broad strata like rural and urban, etc. pre-existed. These substrata are then treated as strata in the analysis. Post-stratification is not very expensive in most cases. Basically, formulas for means and variances would be used in these cases, as indicated in Set 2 in the Appendix.

The second broad alternative pertains to situations wherein percentages of the variables of interest are not available or "resort step" 1 is daunting and ineffective due to the many and diverse percentages at the cluster level. The percentage of nonchristian households, or the percentage of couples willing to participate in a trial of a new method of periodic abstinence ("rhythm") are clear examples. It basically suggests a "networking" or "snow-balling" approach to be done initially. In each sample cluster, the assigned interviewer makes rapid first-phase round of every tenth household, asking at most four short questions bearing on the main variables of interest. Using the above examples of studies concerned with religion and periodic abstinence we would thus have:

What is the religion of the household head? Are there other households of the same group living here in this area? (**IF YES**:) Can you tell me their names and how to find them in this area?

Have you learned about periodic abstinence or rhythm? (**IF YES**) Have you tried to use it? (**IF YES**:) Are you currently using some form of natural family planning? (**WRITE ANSWER**.) Do you know of others in this area who have used or are using natural family planning or rhythm? (**LIST NAME, "ADDRESS" AND EASIEST WAY TO REACH THESE PERSONS**.)

Then the interviewer estimates the percentage of potential sample elements in his or her assigned cluster. This information is then cabled, phoned or brought to the central office, depending upon the distance involved. Depending on the degree of similarity or disparity between resulting cluster percentages, one of the following two different approaches may

then be applied: (1) "resort step" 1 as explained in great detail earlier using the same Set 2 mean and variance formulations which is now termed "resort step" 2; and (2) a revision of the sample design if resulting cluster percentages are too disparate for the "resort step" 1 approach.

The revision of the sample design relates to the third broad alternative offered in this paper. The first stage (or stages) of cluster sampling are treated as an equal probability method by which a subsample (the sample clusters) of the entire survey area of interest has been drawn. A sampling list is prepared in the central office for each cluster in the subsample. The percentage estimated as above by the interviewer in the field is applied to the total number of households (or other ultimate sampling units) reported by the interviewer in each cluster. This produces an estimated complete listing of sample elements believed to be residing in this particular cluster. This is done for all sample clusters and for each cluster, each elementary unit is identified as Sample Element A of Cluster 1, Sample Element B of Cluster 1, Sample Element C of Cluster 1, etc. From the complete list of sample elements in all sample clusters, the desired sample size is selected through simple random sampling across these sample clusters. A unique number is assigned to each randomly selected sample element per sample cluster and sent back to each of the sample clusters where the interviewer is waiting before actual interview.

Meanwhile, while the interviewer is waiting for the list of sample elements to be interviewed to be provided by the Central Office, he or she is instructed to prepare his or her own list of sample elements he or she can locate for interview in his or her assigned cluster based on the "networking" or "snowballing" he or she would have done initially and to number these potential respondents chronologically.

If the number of sample cases in a given sample cluster selected by the Central Office turns out to be less than or equal to the number of persons located and listed by the interviewer, then the interviewer terminates his work in the cluster after accomplishing the total number of sample respondents determined by the Central Office in his or her assigned area. Otherwise, he or she attempts to go back to every tenth of the interviewed sampled elements to inquire about other cluster cases with the same rare characteristic sought. If additional cases are identified, the interviewer assigns to them consecutive sampling numbers in order of identification. If this informant system proves insufficient, the interviewer then makes a random start in the cluster and interviews every tenth household not yet interviewed so as to find out if it contains one of the rare cases in question. If it does not, inquiry is then made whether the respondent knows of any such cases in the cluster

other than those already identified. These further searches for the desired cases are undertaken until the number of cases in that sample cluster sufficiently and reasonably allows the drawing out of the remaining undiscovered number of sample cases prepared at the Central Office. He or she then puts in a hat the numbers of such cases and draws out as many numbers as needed to cover the number selected by the Central Office from the cluster.

In the event that despite such searches, the interviewer has not yet completed the required sample size but finds conclusive evidence that no more undiscovered cases of the desired characteristic exist in the cluster, he or she writes down the reasons for this conclusion, submits this report to his/her supervisor, and terminates the search in that cluster for further rare cases. Note, however, that all decisions must be explained in writing to the supervisor and subsequently handed in to the Central Office.

The researcher can choose between either of at least two approaches when it comes to statistical generalization. He can now define the population studied as limited to the clusters selected into the sample ("resort step" 3). He can then generalize to these only, using formulas for means and variances for simple random sampling.

The second approach under the third alternative in this paper attempts to generalize back to the original population and is termed "resort step" 4. It argues that since the first stage of sampling selected a set of areas by equal probability, this may therefore be considered as a legitimate area probability delimitation for the study of the original population. It next argues that selection of a random sample across clusters of the whole probability sample gave each person listed in the sampled area an equal chance of being drawn into the final subsample. It further argues that since the probabilities of the first and second stages are both known, it is possible to compute approximate means and variances (Set 3 in the Appendix).[3]

CONCLUSION

This paper has presented four "resort steps" which may be taken when financial or other constraints make it difficult to obtain a truly representative probability sample in two- or multi-stage cluster sampling design. The point of the paper has been to inform social scientists that several alternatives do exist in modifying classical cluster sampling techniques owing to financial and time constraints but still yield estimates of variables of interest having a reasonable level of precision.

NOTES

[1] *The four usual problems of this approach are: (1) the sample size is not fixed but variable; (2) the ratio mean is not an unbiased estimator of the population mean; (3) practical variance formulas are not unbiased esti-*

mators of the true variance; and (4) the variance formulas are complicated. If the sample size is handled by crude post-stratification by size, and the usual caution of keeping the coefficient of variation of sample size at .2 or less, the bias can hopefully be kept acceptably small, and the results will still be very good although slightly approximate.

[2] Weights (necessary because of the size of strata) will be kept quite simple and easy to use — if possible in a range from 1 to 10 — even if the weighing is rather crude and approximate. Possibly, too, actual variances can be calculated for a few more important variables and then the design effect used with SRS formulas where variables seem to behave like those already actually calculated with the ratio-mean formulas.

[3] The first idea of the approach of "resort step" 4 was suggested to me by Dr. Michael A. Costello of Xavier University in a discussion concerned with elaborating a sample design for a study of opinions regarding the proposed Muslim Autonomous Region in Mindanao.

REFERENCE

Lemeshow S. and G. Stroh. 1988 *Sampling Techniques for Evaluating Health Parameters in Developing Countries.* A Working Paper prepared for Board on Science and Technology for International Development, Office of International Affairs, National Research Council, Washington, DC,: National Academy Press.

Appendix: Formulas

Set 1: $r = \dfrac{\sum_h \sum_a Y_{ha}}{n_{ha}}$

Where r = ratio mean; y = summation yi; a = cluster; n = no. of cases; h = stratum; f = sampling fraction $\frac{a}{A}$

$$\text{var}(r) = 1/N^2 \left[\sum \frac{(1 - f_h)}{(a_h - 1)} (a_h y^2_{ha} - y^2_h) \right]$$

$$+ r^2 \left[\sum \frac{(1 - f_h)}{(a_h - 1)} (a_h n^2_{ha} - n^2_h) \right]$$

$$- 2r \left[\sum \frac{(1 - f_h)}{(a_h - 1)} (a_h Y_{ha} n_{ha} - Y_h n_h) \right]$$

Note: The coefficient of variation, Cn, should be kept below .1, or at most .2, to avoid more than negligible bias in the ratio mean.

Set 2: $r = \dfrac{\sum_h \sum_a W_{ha} Y_{ha}}{\sum_h \sum_a W_{ha} n_{ha}}$

Where w=weight of stratum and/or cluster

$$\text{var}(r) = 1/N^2 \left[\sum \frac{(1 - f_h)}{(a_h - 1)} \left(a_h \sum_a^A w^2_{ha} y^2_{ha} - w^2_h y^2_h\right) \right]$$

$$+ r^2 \left[\sum \frac{(1 - f_h)}{(a_h - 1)} \left(a_h \sum_a^A w^2_{ha} n^2_{ha} - w^2_h n^2_h\right) \right]$$

$$- 2r \left[\sum \frac{(1 - f_h)}{(a_h - 1)} \left(a_h \sum_a^A w_{ha} n_{ha} - w_h y_h w_h n_h\right) \right]$$

Set 3: $\bar{y} = \sum_j^n y_N$

$$\text{var}(\bar{y}) = 1/N^2 \left[\sum \frac{(1 - f_h)}{(a_h - 1)} \frac{(1)}{(n - 1)} \left(\sum_j^n y^2_j - n y^2 \right) \right]$$