# A COMPARISON OF BALANCED REPEATED REPLICATION TO DIRECT REPLICATION FOR VARIANCE ESTIMATION

Virginia Richards and Daniel H. Freeman, Jr., Yale University

Survey sampling has been faced with an important paradox since its inception. On one hand there is a need to produce estimates with minimum mean square error at minimum cost. This has led to the development of complex sampling designs which optimally utilize available resources. These designs tend to minimize the expected mean square error of the corresponding estimators. Unfortunately, estimating the standard errors of these estimators is a complex and tedious task. On the other hand analysts have increased in their statistical sophistication and are demanding that survey sample data be utilized in fitting statistical models of population phenomena. This process of model fitting can be misleading when the underlying sampling structure is ignored.

This is paralleled by the widespread awareness that sampling variance is not the only source of variation for an estimator. There is, in addition, response variance arising from the respondents themselves. The way a respondent answers a question is subject to an element of chance and this variability can be accentuated by an interviewer's interpretation of the question or of the answer. In some circumstances this response variance may dominate sampling variance.

Research in the development of techniques of indirect estimation suggest ways of dealing with both the problem of estimating response variance and the problem of estimating standard errors of estimators from complex sampling designs. It has been shown that for appropriately designed samples all sources of variation, other than those due to under-coverage and other forms of systematic bias, may be estimated through estimation techniques related to replication. The parallel development in techniques of variance estimation for estimators from complex surveys has likewise led to the technique of replication, and further from direct replication to pseudo-replication and balanced repeated replication.

It is by linking direct estimates to replication and pseudo-replication that it may be possible to construct variance estimators which incorporate both sampling and response variance. In order to establish this linkage it is necessary to review the literature on response error and that on indirect estimation. To date the literature in these two areas has been parallel, but separate.

Hansen, Hurwitz, and Bershad (1961) and Hansen, Hurwitz, and Pritzker (1964) have described a model for the total variance of a survey estimator from a simple random sample. This model partitions the total survey variance into two components, the sampling variance and the response variance. The sampling variance of an estimator based on n observations reflects the variability of the estimator over the possible samples of size n from a population of size N. If the survey is a complete census then the sampling variance component equals zero. The response variance reflects the variability of an estimator from a particular sample over a conceptual (or actual) series of repeated trials. The response variance is further partitioned into the simple response variance and the correlated response variance. The simple response variance reflects the variability of the individual response deviations, that is the deviation of an observed value on a given individual on a given trial from the expected value for that individual over repeated trials. The correlated response variance reflects the correlations among response deviations of different units in a given sample and a given trial.

Koch (1973) extends this response error model to multivariate response error situations and complex survey designs. The formulation is based on indicator functions as described by Cornfield (1944) and a Horvitz and Thompson (1952) type estimator. The statistic is $\underset{\sim}{x}_t = \sum_{i=1}^{N} w_i U_i \underset{\sim}{Y}_{it}$

where $U_i = \begin{cases} 1 & \text{if population element } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$

$\underset{\sim}{Y}_{it}$ is defined to be a vector random variable with p components with t indexing conceptual repeated trials

and $w_i$ are known coefficients.

The $\{U_i\}$ specify the survey design and the $E(U_i)$ and $E(U_iU_{i'})$ specify the probability of selection. The random variables $\{U_i\}$ reflect sampling errors and the $\{Y_{it}\}$ reflect response errors. The sampling variance and the response variance for this statistic are derived as well as an interaction variance term. Koch shows the relationship of the components of the multivariate response variance to be

$$RV = \frac{1}{n}[(SRV) + (n-1)(CRV)]$$

where SRV = simple response variance and CRV = correlated response variance which is completely analogous to what Hansen, et al. found for the univariate case. Koch further partitions the CRV into a simple correlated component (SCRV) and the interaction response variance (IRV). Under certain simplifying assumptions, the $\text{var}(\underset{\sim}{x}_t) = \frac{1}{n}[(SRV) + (n-1)(CRV)] + [SV]$ where SV = sampling variance.

Direct methods for the estimation of upper and lower bounds for $\text{var}(\underset{\sim}{x}_t)$ are discussed by Koch, Freeman, and Freeman (1975). They note that if the sampling variance is the most important source of error, then the lower bound is the most appropriate estimator. If simple response variance is the most important source of error, then the upper bound is most appropriate.

Hansen, Hurwitz, and Bershad (1961) show that if the response deviations are uncorrelated and the sample is selected by simple random sampling with replacement, then the usual sample estimate of the sampling variance will reflect not only the sampling variance but also the response variance.

It is also possible to include both sampling variance and response variance in the variance estimate when the method of replicated sampling is used in the sample design (Deming 1960). Replicated sampling also has the distinct advantage that it allows one to bypass complicated formulations when obtaining estimates of the variance of estimators. It is these two properties of replicated sampling which make it the natural link between these two areas of research.

Replicated samples, interpenetrating samples, and random groups are three names describing essentially the same technique. The idea behind replicated sampling is that a sample of size n can be designed to include k independent sub-samples (replicates) of size n/k, each of which reflect the design of the entire sample in all respects except size. With interpenetrating samples and random groups a sample of size n is selected and then randomly divided into k groups of size n/k. With these methods of replicated samples, the desired statistic, $x_\gamma$, $\gamma = 1,\ldots,k$ is computed for each replicate, each $x_\gamma$ being an estimate of the population value. The $x_\gamma$ may represent any statistic such as a mean, total, regression coefficient, etc. The estimate for the entire sample is the average of the replicate estimates,

$\bar{x} = \frac{1}{k} \sum_{\gamma}^{k} x_\gamma$. The variance of this estimate is given by $\text{var}(\bar{x}) = \frac{1-f}{k(k-1)} \sum_{\gamma}^{k} (x_\gamma - \bar{x})^2$ where 1-f is the finite pop'n correction.

As noted by Hansen, Hurwitz, and Bershad the total response variance can be partitioned into the simple response variance and the correlated response variance. It is the correlated component which is usually responsible for the greater contribution to the total response error. Until recently attempts to estimate the components of response variance have centered around replication in the sense of repeated measurements on the same unit or the method of interpenetrating subsamples. Bailar and Dalenius (1969) present a comprehensive review of these two methods and also designs using a combination of both methods.

Interpenetrating samples, a technique proposed by Mahalanobis (1946), enables one to estimate the correlated component of the response variance. The study is designed so that there is no correlation between the errors of measurement in different subsamples. For example, assuming that interviewers are the only source of correlated measurement error, an interviewer would not be assigned to two different subsamples. The method provides an estimate of the correlated component of the response variance by subtracting the within subsample variance and dividing by the number of interviewers. Deming (1960) proposed a method randomizing the interviewer's assignments over the subsamples in a randomized block design. In this design the variance between interviewers can be computed and the variance between subsamples will not reflect the differences between interviewers. Mahalanobis attempts to measure the total variance and Deming attempts to measure the 'pure' sampling variance, by the variance between subsamples.

In general, with replicated sampling, the stability of the estimate of variance tends to increase with an increasing number of replications. Increasing the number of replicates, however, can be difficult since if you wish to keep the overall sample size constant you are limited in the number of replicates you can have and still maintain a reasonable replicate size. This is one of the considerations which led to the development of the method of pseudoreplication.

The method of pseudoreplication, also called half sample replication, and repeated replication, was described and improved upon by McCarthy (1966). The method is based on a sample being designed with k = 2 replicates, each replicate being a half sample, so the sample is stratified with two independent selections per stratum. Instead of relying on only the two half samples originally chosen, new half samples are created by selecting either the first or second element from each stratum. If there are L strata this yields a possible $2^L$ half samples. The average of all half sample estimates of the mean is equal to the stratified mean, $\bar{y}_{st}$, which is an unbiased estimate of the population mean. Also $E[(\bar{y}_{hs} - \bar{y}_{st})^2] = \text{Var}(\bar{y}_{st})$ for repeated selections of the entire sample. In practice the repeated replication technique is not restricted to the strict design above. The 2 PSU's per stratum design can be obtained by other methods such as by collapsing strata or if there are replicates within a PSU by combining them to obtain two. Using all possible half samples may not be feasible due to extensive calculations.

In developing the method of balanced half sample replication or balanced repeated replication (BRR), McCarthy (1966) showed that by choosing a subset of half samples, such that the between strata contributions to the estimates of variance are eliminated, one can retain all the information available in the total sample. In other words, the same variance estimator is obtained as if all possible half samples were used. Use of a design matrix with the number of columns equal to the number of strata and each row representing the selected half sample constructed so that the columns are orthogonal will result in the desired subset of half samples. The number of half samples required will be at most 3 more than the number of strata.

The Connecticut High Blood Pressure Survey is one which lends itself well to examining the previously discussed concepts. The survey is a complex, multi-staged, probability sample which was designed to include four independent replicated subsamples. The 169 towns of Conn. were divided into 32 strata on the basis of Health Service Area and population size. In each stratum, one PSU (town) was selected. Within each PSU, four independent systematic samples of segments were chosen. The segments were created to be of approximate size of 16 housing units. Within each segment a random sample of housing units was chosen with the sampling fraction equal to 1/4.

The type of estimator analysed by Koch and described above will be used in this survey. A sample estimate will be the average of the four replicate estimates and the variance of the estimate will be estimated as described in the section on replicated sampling. Also, balanced half samples will be created. Sample estimates and variances of the estimates will be calculated by the balanced repeated replication technique. A comparison between replicated sampling with a small number of replicates and balanced repeated replication can be made.

ACKNOWLEDGEMENTS

This research was in part supported by the National Heart Lung and Blood Institute (Grant # 731C-41-57818) and the Connecticut High Blood Pressure Program.

REFERENCES

1. Bailar, B.A., and Dalenius, T. (1969). Estimating the response variance components of the US Bureau of the Census survey model. Sankhya, Ser. B, 31, 341-360.

2. Cornfield, J. (1944). On samples from finite populations. Journal of the American Statistical Association, 39, 236-239.

3. Deming, W.E. (1960). Sample Design in Business Research. New York: John Wiley and Sons.

4. Hansen, M.H. Hurwitz, W.N., Bershad, M.A. (1961). Measurement errors in censuses and surveys. Bulletin of the International Statistical Institute, 38, Part II, 359-374.

5. Hansen, M.H., Hurwitz, W.N., Pritzker, Leon (1964). The estimation and interpretation of gross differences and the simple response variance. in C.R. Rao, ed., Contributions to statistics Presented to Professor P.C. Mahalanobis on the Occasion of His 70th Birthday. Pergamon Press, Ltd.

6. Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.

7. Koch, Gary G. (1973). An alternative approach to multivariate response error models for sample survey data with applications to estimators involving subclass means. Journal of the American Statistical Association 68, 906-913.

8. Koch, G.G., Freeman, D.H., Freeman, J.L.(1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 43, 59-78.

9. Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society, 109, 325-370.

10. McCarthy, P.J. (1966). Replication: An approach to the Analysis of Data from Complex Surveys. National Center for Health Statistics, Series 2, No. 14.

11. McCarthy, P.J. (1969a). Pseudo-replication: Further Evaluation and Application of the Balanced Half-sample Technique. National Center for Health Statistics, Series 2, No.31.

12. McCarthy, P.J. (1969b). Pseudo-replication: half samples. Review of the International Statistical Institute, 37, 239-264.