
Analysis of Complex Sample Data Using Replication

J. Michael Brick
David Morganstein
Richard Valliant

July 29, 2000

WESTAT

1. Introduction

The analysis of data from surveys and observational samples may be complicated by the necessity of applying appropriate estimation weights and variance estimation techniques. These techniques generally require specialized software that is difficult to learn and use and is based on concepts that are not familiar to many analysts. In some cases, the providers of the data do not even supply analysts with the information necessary to implement the techniques. Recent advances in software help to reduce these difficulties and bring practice more into line with theory. In this paper, we describe how replication methods provide a solution to some of these problems.

In many studies, data are collected from individuals or units sampled using complex sample designs that include varying probabilities and non-independent selections. For example, in household surveys persons may be sampled from geographically clustered households. In this scheme, persons within the same geographical cluster have a higher probability of being sampled together than do persons in different clusters. Similarly, the samples of patients in clinical trials are not independent because several patients are sampled from the same clinics and each clinic has its own practices and procedures. Another type of dependence occurs when patient data are collected at several points in time.

In addition to the complexities due to clustering, the probability of selecting a particular unit may vary depending on factors such as the size or location of the unit. For example, in a sample of hospitals the probability of selecting a hospital may be proportional to the number of admissions. These types of design features make analysis of the data more difficult.

Typically, the goal of a study is to estimate population quantities that can be written $\theta = g(y_1, \dots, y_N)$ by using a sample statistic, $\hat{\theta} = h(y_1, \dots, y_n)$, where the y_i are observations on unit i and may be vector-valued. For example, the pupil-teacher ratio may be the population quantity θ that is being estimated from a survey of schools. The statistic from the survey $\hat{\theta}$ is the estimated total number of pupils divided by the total number of teachers.

Many investigations have shown that ignoring the sample design and using simple random sampling methods leads to biased estimates (e.g., Landis et al. 1982; Kish 1992; Korn and Graubard 1995; Brogan 1998). In the pupil-teacher ratio example, the estimates in both the numerator and denominator of the ratio must be appropriately weighted to minimize the bias of the estimated ratio. These studies also show that ignoring the sample design or the population structure as reflected in a sample leads to biased and misleading estimates of the standard error.

Confidence intervals and statistical tests will be incorrect if the complexities of the design and estimation methods are not taken into account in the analysis. Analysts who naively ignore the sample design or data structure may substantially underestimate the standard error, especially when clustering and unequal probabilities of selection have been used.

Analysts may find it difficult to deal with these complexities because standard statistical software, such as SPSS and SAS, is not designed to address these issues. Even when the standard software can be manipulated to produce unbiased point estimates by using weights, it generally cannot be easily used to estimate appropriate standard errors. This is true even for simple estimates of totals or means. In fact, for many complex sample designs and estimation methods, there are no direct analytic methods that can be used to produce unbiased estimates of the standard errors of the estimates. The only alternative is to approximate the quantities.

One approach to approximating the standard error of an estimator is to use replication methods such as the jackknife and balanced repeated replication (BRR) methods. An alternative approach is to linearize the estimator using a Taylor series expansion and then use standard sample survey variance estimation methods to estimate the precision of the linearized statistic. Both methods are described in Wolter (1985). In this paper, we concentrate on replication methods and how they can be used simply and accurately to produce approximately unbiased estimates of standard errors for a variety of statistics. Theoretical and empirical studies have found that these methods perform well (e.g., Kish and Frankel 1974; Krewski and Rao 1981; Kovar, Rao, and Wu 1988; Rao, Wu, and Yue 1992; Shao 1996). Rust and Rao (1996) provide a good review of the replication literature as applied to complex sample surveys.

In Section 2, we discuss how replication methods can be easily applied to estimate standard errors in complex samples and we present some advantages of replication. Replication methods can be implemented using WesVarTM, which extends the capabilities of earlier software, WesVarPC (Morganstein and Brick 1996), by including many new features and a powerful, user-friendly interface. Before describing the benefits of replication methods, we briefly review these methods and how they can be used in complex samples.

2. Replication Methods

Replication is appealing because the basic idea is easy to understand and explain. To estimate sampling errors, one repeatedly selects subsamples from the realized full sample; in other words, one subsamples from the sample. The desired statistics are computed from each

subsample, and the variability among these subsample or replicate estimates is used to compute the standard error of the full-sample estimate. Replication mimics one of the standard approaches to deriving theoretical statistical properties, which is to measure the average performance of estimators across many samples selected in a prescribed way.

Different replication methods use different approaches to subsample from the full sample. The subsamples are called replicate samples, and the statistics calculated from these replicates are called replicate estimates.

Before proceeding further, we must define some terminology. Strata and primary sampling units (PSUs) are the variables from the full sample needed to specify replicate subsamples. These are standard terms in finite population sampling theory; see, for example, texts by Cochran (1977) and Särndal et al. (1992). Generally speaking, strata are groups of units that are sampled as if they were separate populations, while PSUs are clusters sampled within a stratum. For example, if three clinics are designated for a study and patients are sampled within each clinic, clinics are the strata and patients are the PSUs.

In this paper, we consider BRR and a variation known as the Fay method (Fay 1989), as well as three methods of jackknife replication; however, other methods, such as bootstrapping, are also available. BRR is a method of half-sample replication applicable to designs in which exactly two PSUs have been sampled from each stratum. The first jackknife method we consider, called JK1, corresponds to the situation in which there is only one stratum and PSUs have been randomly sampled from it. The second jackknife method, JK2, handles the same design as BRR (i.e., where exactly two PSUs have been selected per stratum). The third method of jackknife replication, JK_n, is for general stratified samples where two or more PSUs have been selected in each stratum. JK1 is a special case of the more general JK_n method. The types of samples that can be handled by each of these methods are discussed below.

The estimated variance, $v(\hat{\theta})$, of an estimate, $\hat{\theta}$, based on the replicate estimates is

$$v(\hat{\theta}) = c \sum_{g=1}^G h_g (\hat{\theta}_g - \hat{\theta})^2, \quad (1)$$

where

$\hat{\theta}$ is the estimate of θ based on the full sample,

$\hat{\theta}_g$ is the g -th estimate of θ based on the observations included in the g -th replicate,

G is the number of replicates,

c is a constant that depends on the replication method, and

h_g is a factor associated with replicate g and the replication method.

The values of c are given below:

Method	c
BRR, Fay	$1/G$
JK1	$(G-1)/G$
JK2	1
JKn	1

The other factor, h_g , is equal to unity for all methods except JKn, where h_g is the ratio of the number of PSUs in the stratum minus 1 to the number of PSUs. The bootstrap variance estimator also has the form (1) with $c = 1/G$ and $h_g = 1$.

Rust and Rao (1996) describe replication methods more completely and suggest how the methods might be appropriate for specific sample designs. We present a few examples of the application of these methods to suggest the generality of the approach, especially when more advanced methods, such as partial balancing or combining strata (Rust 1986), are used in conjunction with these methods. Many national samples are multi-stage samples in which the first-stage units are highly stratified and two PSUs have been sampled per stratum. These designs are well suited for BRR, Fay, or JK2 with little or no modification. Simple random samples and unequal probability systematic samples are frequently easy to handle with JK1, even when the units selected are clusters rather than individual units. Random digit dialing telephone surveys using a list-assisted method (Brick et al. 1995) fit into this method. The JKn method may be appropriate for clinical trials and for establishment surveys where establishments are stratified and a different number are sampled from each stratum. Because these designs often sample very heavily in some strata, it is possible to modify h_g in the JKn method to include a finite population correction factor and account for this “without replacement” sampling feature.

Replication methods can be applied to nearly all types of sample survey and experimental designs. The manual for WesVar includes a more detailed prescription of how various sample designs can be handled with one or more of the replication methods. The only requirement is that

the user be able to define strata and PSUs for the sample design. Even this requirement can be relaxed if the provider of the survey data attaches replicate weights to the data file, as discussed in Section 3.

3. Some Advantages of Replication

Easy to Explain

As mentioned in Section 2, an important feature of replication is the ease with which users can be informed about the method, even if they do not have advanced training in variance estimation for complex samples. Replication is a well-known method in statistics, and users readily accept and understand the basic premise. Often, it is only necessary to state that the method involves dividing the full sample into subsamples and estimating the variability from the subsamples.

Theoretically Sound

Many authors have studied replication variance estimators using either design-based or model-based analysis. By “design-based” we mean that properties are computed with respect to the method of randomization used in a specified sample design. “Model-based” analyses calculate statistical properties with respect to a mathematical model. For a broad array of sample designs, estimators, and models, the BRR, Fay, and jackknife methods provide variance estimators that are supported by either theoretical approach.

Common Procedure

Another benefit of replication is that the same procedure is used to compute the standard error for nearly all statistics. The variance of means, totals, ratios, and more complex functions of the sample data are all computed using equation (1). For example, in the pupil-teacher ratio example the statistic can be written

$$\hat{\theta} = \frac{\sum w_i x_i}{\sum w_i y_i},$$

where w_i is the weight, x_i is the number of pupils, and y_i is the number of teachers in sample school i . The difference between the pupil-teacher ratios in different types of schools can be

specified easily in WesVar by writing it as a difference of ratios. The variance of the difference is then computed using (1). Another example is the log-odds ratio for a two-way table, which can be written as

$$\hat{\theta} = \log \left(\frac{\hat{N}_{11}\hat{N}_{22}}{\hat{N}_{12}\hat{N}_{21}} \right),$$

where \hat{N}_{ij} is the full-sample estimate of the total in row i and column j ($i = 1, 2; j = 1, 2$) of a table. This definition of $\hat{\theta}$ is then used in (1) to estimate the variance. These statistics and many other more complicated statistics can be computed simply and quickly using WesVar. It requires no programming skill; the simple Windows interface in WesVar is used to define the statistic, and the program computes the standard error and confidence interval.

General Adjustment or Estimation Methods

An important benefit of replication methods is that they can be used to reflect adjustments made in the estimation process. In most sample surveys, the observed data are first weighted by the reciprocal of the probability of selection and then adjusted. For example, almost all surveys are subject to unit nonresponse, and the full-sample weights are adjusted to compensate for it (Brick and Kalton 1996). If the replicate weights are adjusted similarly, then the variance estimates reflect this adjustment. Estimation strategies such as poststratification or raking can also be considered as adjustments of the base weights. These methods use auxiliary data to reduce the bias and the variances of the estimates. Replication offers a simple and elegant method of incorporating these adjustments in the estimates of variance (Valliant 1993; Yung 1996). WesVar has a mechanism that enables the user to make nonresponse, poststratification, and raking adjustments, if they are not already included in replicate weights with the data file. (Below, we discuss some of the benefits that users enjoy if the data provider does include replicate weights with the data file.)

Weights Contain Design Information

Another benefit of replication is that the replicate weights can be created once for a study and stored on the data file, along with the full-sample weight. Many government agencies are

beginning to include replicate weights on their public use releases. This procedure has three important advantages:

1. The user does not have to know any of the features of the design to estimate standard errors properly (the information is contained in the replicate weights);
2. The agency can reflect all stages of estimation appropriately in the replicate weights (including nonresponse); and
3. The software for replicate variance estimation does not have to be modified to compute the estimates properly; equation (1) is still used.

Domain Estimates

Because the replicate weights carry all of the information needed for estimating variances, replication is extremely well suited for the analysis of subsets or domains. For example, a sample of adults might have been selected, but some analysts may wish to estimate characteristics for only a subset of the adults (e.g., females between the ages of 30 and 44). This type of analysis is simple to do with replication methods without concerns about the implications for variance estimation. The analyst can identify the subset of the data and extract a subset file that contains only the records for the domain of interest, with the full sample and replicate weights. The detailed analysis for the domain can be conducted with the subset file because the replicate weights contain all of the information needed to properly estimate the standard error of the estimates. Thus, the concern about domain analysis discussed by Graubard and Korn (1996) for linearization methods does not apply with replication methods.

Missing Data

Incomplete or missing data are a common but difficult problem in sample surveys. We have already discussed a replication approach for handling unit nonresponse adjustments of the weights. For item-level missing data, the problem is more difficult and does not have a general solution, irrespective of the method of variance estimation. In these circumstances, a reasonable goal is to handle the missing items consistently when computing the estimate and its variance.

With replication, the same procedure used for handling missing data for the full-sample estimate is applied to the replicate estimates. For example, in estimating a total with missing item responses, the missing data can be excluded in computing the full-sample estimate and the

replicate estimates. This may result in a biased estimate because of the item nonresponse (the user should evaluate this or any other strategy for dealing with missing data), but the standard error of the estimate can be computed appropriately for the procedure because the full and replicate estimates are handled consistently. This method is used in WesVar.

Longitudinal Designs

Estimates from panel or longitudinal studies are more precise for measuring change over time than are estimates from independent cross-sectional studies. An increase in precision results from the correlation between the units sampled over time. Replication methods provide a simple and elegant means of incorporating this correlation. The replicates for the baseline sample are used for the followup data collections so that the replicate estimates contain the appropriate correlation. Hinkins et al. (1996) discuss this issue in more detail for a particularly complex situation. The WesVar manual describes how to construct replicates in several longitudinal designs for the purpose of estimating change over time.

Disclosure Avoidance

In some surveys, especially establishment surveys, the organizations collecting the data wish to release the information to the public but also want to protect the respondents from being individually identifiable. Approaches to preserve the confidentiality of responses have included suppressing data and adding random noise to some of the responses. The concern for confidentiality sometimes prevents the release of data on sampling strata and PSUs. These variables are needed to compute standard errors with linearization methods and to set up the replicates with replication. However, the replicate weights contain all of the information needed to compute standard errors, and it is unnecessary to include stratum and PSU on the public release file if the replicate weights are included. Because it may be possible to “reverse-engineer” the replicate weights to determine the stratum and PSU, other techniques can be used to make identification very difficult. For example, combining strata or PSUs to form replicates may reduce the chance of disclosure. A more extreme option is to add noise to the replicate weights to further protect the confidentiality of the data. In many circumstances, replication provides an additional mechanism to protect confidentiality and prevent disclosure.

4. Summary

In sample surveys, clinical trials, and other studies with complex designs or estimation methods, standard statistical software cannot produce both unbiased point estimates and appropriate standard errors of the estimates. Special methods and software are needed to avoid the biases that arise when statistical software assumes that the data are independent and identically distributed.

Replication methods are well suited to handling complex designs and estimation procedures. These methods can be applied to nearly all methods of collecting clustered or correlated data in sample surveys and clinical trials. The following are some of the benefits of replication:

- The simplicity of the method,
- A sound theoretical basis,
- The application of a common procedure for computation purposes,
- The feasibility of reflecting a variety of estimation and adjustment methods,
- The encoding of the design information in replicate weights stored on the data file,
- A simple method for analyzing domains,
- The ability to handle item-level missing data in computing standard errors in a manner consistent with the way the estimates are produced,
- A direct method for variance estimation in longitudinal designs, and
- The possibility of protecting the confidentiality of the data.

These benefits, especially when combined with the simple and powerful user interface in WesVar, suggest that replication methods should be considered whenever complex sample data are analyzed.

References

- Brick, J.M., and Kalton, G. (1996). Handling Missing Data in Survey Research. *Statistical Methods in Medical Research*, **5**, 215-238.
- Brick, J.M., Waksberg, J., Kulp, D., and Starer, A. (1995). Bias in List-Assisted Telephone Surveys. *Public Opinion Quarterly*, **59**(2), 218-235.
- Brogan, D.J. (1998). Pitfalls of Using Standard Statistical Software Packages for Sample Survey Data. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics*. New York: John Wiley and Sons.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons.
- Fay, R.E. (1989). Theoretical Application of Weighting for Variance Calculation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 212-217.
- Graubard, B.I., and Korn, E.L. (1996). Survey Inference for Subpopulations. *American Journal of Epidemiology*, **144**, 102-106.
- Hinkins, S., Moriarity, C., and Scheuren, F. (1996). Replicate Variance Estimation in Stratified Sampling with Permanent Random Numbers. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 824-829.
- Kish, L. (1992). Weighting for Unequal Pi. *Journal of Official Statistics*, **8**, 183-200.
- Kish, L., and Frankel, M. (1974). Inference from Complex Samples. *Journal of the Royal Statistical Society B*, **36**, 1-22.
- Korn, E.L., and Graubard, B.I. (1995). Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. *The American Statistician*, **49**, 291-295.
- Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. (1988). Bootstrap and Other Methods to Measure Errors in Survey Estimates. *Canadian Journal of Statistics*, **16**, 25-46.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of Linearization, Jackknife and Balanced Repeated Replication Methods. *Annals of Statistics*, **9**, 1010-1019.

- Landis, R.J., Lepkowski, J.M., Eklund, S.A., and Stehouwer, S.A. (1982). A Statistical Methodology for Analyzing Data from a Complex Survey: The First National Health and Nutrition Examination Survey (DHHS Pub. No. 82-1366). *Vital and Health Statistics, Series 2, No. 92*. Hyattsville, MD: National Center for Health Statistics.
- Morganstein, D., and Brick, J.M. (1996). WesVarPC: Software for Computing Variance Estimates from Complex Designs. *Proceedings of the 1996 Annual Research Conference*, pp. 861-866. Washington, DC: U.S. Bureau of the Census.
- Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some Recent Work in Resampling Methods. *Survey Methodology*, **18**, 209-217.
- Rust, K.F. (1986). Efficient Replicated Variance Estimation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 81-87.
- Rust, K.F., and Rao, J.N.K. (1996). Variance Estimation for Complex Surveys Using Replication Techniques. *Survey Methods in Medical Research*, **5**, 283-310.
- Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Shao, J. (1996). Resampling Methods in Sample Surveys (with Discussion). *Statistics*, **27**, 203-254.
- Valliant, R. (1993). Poststratification and Conditional Variance Estimation. *Journal of the American Statistical Association*, **88**, 89-96.
- Wolter, K. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- Yung, W. (1996). *Contributions to Poststratification in Stratified Multistage Samples*. Doctoral dissertation, Carleton University, Ottawa, Ontario, Canada.