COMPLEX SURVEY VARIANCE ESTIMATION AND CONTINGENCY TABLE ANALYSIS USING REPLICATION

Paul Flyer, Keith Rust, and David Morganstein, Westat, Inc.
David Morganstein, 1650 Research Blvd., Rockville, MD 20850

KEY WORDS:

1. Review - Practical Issues in the Use of Complex Survey Data

The purpose of this paper is to discuss Westat's computer software for the estimation of variance when a complex survey design has been used and our general approach to sampling error estimation and survey inference. We define a complex sample design as any probability design other than a simple random sample. These designs typically involve both stratification and multiple stages of sample selection. Additionally, different sample units generally have different probabilities of selection and often there are correlations between the observations. These facets of the sample design almost always lead to the use of sample weights. Although the word complex is often thought of as referring to the sample design, the sample design is not the only element to which the term should be applied. Often the population parameters for which estimates are desired and the estimates themselves are also complicated. Estimators are often nonlinear functions of sample variables and weights, where the weights themselves are random variables which are adjusted for nonresponse and poststratification.

1.1 Sample Design

It is a relatively common misconception that variance estimation software completely "corrects" for the use of complex sampling. In fact, for the sample designs commonly used in practice, unbiased estimates do not exist. For example, in many applications of multi-stage sampling, stratification is used to the maximum extent possible (i.e., one unit is selected from each stratum). Another popular sample selection technique for which unbiased estimates of variance do not generally exist is systematic sampling from an ordered list. This technique is also frequently used in the selection of second-stage units for one PSU per stratum sample designs. For PSU's selected with certainty, this means that the segments serve as first-stage units. None of these techniques have unbiased variance estimation procedures, even for linear statistics. Frequently, variance is estimated using formulas appropriate for 'with-replacement' sampling even though a more efficient 'without-replacement' method was used. This leads to variance estimates with a positive bias. Additionally, it must be kept in mind that designs allowing for the "unbiased" estimation of variance only produce unbiased estimates for linear statistics. For nonlinear statistics, no unbiased estimate of variance will typically exist. Usually, for nonlinear statistics the true sampling variance cannot even be explicitly expressed in terms of parameters of the joint population distribution of the variables involved. Approximations for use with nonlinear estimates will be discussed later.

1.2 Estimation of Basic Parameters

In most large scale surveys, the sampling units have different probabilities of selection. For each of these sampled units (selected without replacement), a weight is calculated reflecting the probability of selection. If p_i is the probability of selection for the i-th unit selected for the sample, the base weight, w_i, is simply calculated as the reciprocal of p_i for each sampled unit.

This base weight can be used to produce unbiased estimates of population totals. If X is a population total for a particular random variable and x_i is the value of x for the i-th selected sample unit, the Horvitz-Thompson (unbiased) estimate for X is

$$\hat{X} = \sum_{i=1}^{n} w_i x_i .$$

Most parameters estimated from large scale surveys are either population totals, like the above, or functions of population totals. By far the most common function is the ratio of two population totals, say Y/X, which is typically estimated using the following nonlinear function of estimated population totals:

$$\hat{R}_1 = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i x_i} .$$

If y and x are suitably defined, the ratio estimator can be used to provide estimates of means for population subgroups. For example, if z is the variable for which the mean is desired, define x and y using the following:

y_i	=	z_i,	if i is from subpopulation of interest
		0,	otherwise .
x_i	=	1,	if i is from subpopulation of interest
		0,	otherwise .

When defined in this manner, R is the mean for the subpopulation of interest. More complex functions of population totals can also be calculated. For example the odds ratio is frequently of interest in health surveys. Suppose there are two responses: "success" and "failure". Define the following variables:

y_{1i}	=	1,	if i is in subpopulation 1 and is a "success"
		0,	otherwise;
x_{1i}	=	$1-y_{1i}$,	if i is in subpopulation 1
		0,	otherwise;
y_{2i}	=	1,	if i is in subpopulation 2 and is a "success"
		0,	otherwise;
x_{2i}	=	$1-y_{2i}$,	if i is in subpopulation 2
		0,	otherwise .

The ratio of the odds of a success in population 1 to the odds of a success in population 2 is calculated as:

$$\hat{R}_2 = \frac{\sum_{i=1}^{n} w_i y_{1i} \sum_{i=1}^{n} w_i x_{2i}}{\sum_{i=1}^{n} w_i x_{1i} \sum_{i=1}^{n} w_i y_{2i}} .$$

1.3 Weight Adjustments

It was shown above that the base weight, w_i, based upon the probability of selection, is an integral part of

population estimates and it is found in most of the formulas presented in sampling textbooks. In practice, this simple weight is seldom used directly for estimation. Instead, the simple base weight is adjusted for nonresponse and then poststratified to either known population totals or to precise survey estimates derived from larger surveys (e.g., CPS). Both of these modifications to the base weight take on the form of ratio adjustments. As we will discuss later, the effect of these adjustments, which generally is to reduce the mean square error, should be incorporated in the variance estimation system.

One common way of compensating for survey nonresponse is through a nonresponse adjustment to the weights used for estimation. Typically, it is assumed that the nonresponding units, within a particular nonresponse category, are a random sample of the initially selected cases. This means that the completed interviews can be "weighted up" to compensate for the nonresponding units. Generally, this estimation procedure results in the use of nonlinear functions of population totals. Therefore, there is no explicit unbiased estimator of variance.

As an example of a nonlinear nonresponse adjustment, consider the situation where sampled units are assigned to one of c nonresponse adjustment cells. These cells are formed based upon characteristics known for all selected persons, regardless of whether or not the unit completed the survey. For example, in household surveys the region of the country and the sex of the sampled respondent are known, even if the respondent refuses to cooperate with the survey. If n_j units (note, n_j is generally a random variable) were initially selected for the j-th nonresponse cell, the nonresponse adjustment factor for the j-th cell is calculated as:

$$a_j = \frac{\sum_{i=1}^{n_j} w_{ji}}{\sum_{i=1}^{n_j} r_{ji} w_{ji}} ,$$

where the ji subscript refers to the j-th nonresponse adjustment cell and the i-th sample unit, and r_{ji} is 1 if the sampled unit responded and 0, otherwise. Using the nonresponse adjustments, the estimate of a total takes on the following form:

$$\hat{Y} = \sum_{j=1}^{c} \sum_{i=1}^{n_j} a_j \, r_{ji} \, w_{ji} \, y_{ji} \ .$$

Because the response indicator variable, r_{ji}, is a random variable, the adjustment factor, a_j, is also a random variable, which makes the estimated total a nonlinear combination (product) of random variables.

Survey estimates are often improved by adjusting sample totals to equal known population totals. This is called poststratification. In this type of adjustment, sample elements are classified into a number of poststratification cells that are possibly different from both the original stratification cells and the nonresponse adjustment cells. This procedure is very similar to the ratio adjustment calculated for nonresponse, except that the cell total used for adjustment is a known constant. The poststratification adjustment, in the absence of nonresponse adjustment, for the k-th poststratification cell can be calculated using the following:

$$b_k = \frac{N_k}{\sum_{i=1}^{n_k} w_{ki}} ,$$

where N_k is the population total for the k-th, n_k is the number of sample units for the k-th poststratification cell. This adjustment factor is then used in an analogous manner to the nonresponse adjustment factor.

Nonresponse and poststratification adjustments are typically used concurrently. When the nonresponse cells cut across poststratification cells, the nonresponse adjustment is generally calculated first. The poststratification adjustment is then calculated using the nonresponse adjusted weights. This process is performed using the following:

$$a_j = \frac{\sum_{k=1}^{d} \sum_{i=1}^{n_{kj}} w_{kji}}{\sum_{k=1}^{d} \sum_{i=1}^{n_{kj}} r_{kji} w_{kji}} ; \text{ and}$$

$$b_k = \frac{N_k}{\sum_{j=1}^{c} \sum_{i=1}^{n_{kj}} a_j \, r_{kji} w_{kji}} ,$$

where n_{kj} is then number of sample units in the k-th poststratification cell and the j-th nonresponse cell, d is the number of poststrata cells, and the subscript i refers to the i-th sample unit within the kj-th combined adjustment cell.

An estimated total for a variable y, adjusting for both nonresponse and poststratification, takes on the following form:

$$\hat{Y} = \sum_{k=1}^{d} \sum_{j=1}^{c} \sum_{i=1}^{n_{kj}} b_k \, a_j \, r_{kji} \, w_{kji} \, y_{kji} \ .$$

It can be seen that the estimated total is a function of both the randomly chosen observations, and random weight adjustments, which are a function of the particular sample drawn. Note that in calculating the weight adjustments, the original stratification variables and clusters (if any were used) have not necessarily been incorporated into the formation of adjustment cells. This will have important implications in the estimation of variances.

In addition to calculating totals for the complete population, most large surveys produce estimates of means. In the presence of weights, means are typically calculated using ratio estimates. General ratio estimates are calculated using:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{i=1}^{n} b_i \, a_i \, r_i \, w_i \, y_i}{\sum_{i=1}^{n} b_i \, a_i \, r_i \, w_i \, x_i} \ .$$

The above estimates can also be calculated for subgroups of the population. For example, means can be calculated separately by sex. This is accomplished by introducing an indicator variable into both the numerator variable and the denominator variable. The indicator variable is 1 when the sampled unit is a member of the subgroup of interest, and 0 otherwise.

To a certain extent the weight adjustments and estimators described above represent only the most basic procedures

followed in large scale surveys. Often there is more than one stage of nonresponse adjustment and poststratification adjustment. The weights might also be trimmed to reduce the effect on variance of extreme weights. Additionally, more complex weighted estimates such as odds-ratios, weighted averages of subpopulation means, correlation coefficients, etc., are frequently calculated.

2. Variance Estimation and Inference

Complex sample designs usually involve selection schemes that lead to statistical dependence among selected sample units. Often, this dependance takes the form of a positive correlation between members of the sample. This positive correlation leads to a negative bias in estimates of variance based upon the assumption of simple random sampling (i.e., as is common to most statistical packages). To produce at least approximately unbiased estimates of variance, aspects of the sample design must be taken into account in estimating variance. This has led to the creation of a number of computer programs that utilize various approximations to adjust for the complex nature of the sample design. The following design elements have an impact on the estimation of variance and will be discussed:

- With- or without-replacement sampling;
- Stratification;
- Multiple stages of selection;
- Unequal probability of selection; and
- Certainty selection of clusters.

In spite of the fact that most complex sample designs have a net increase in variance over simpler sample designs, certain aspects of the sample design can actually result in a decrease in variance. The use of without-replacement sampling is one of these; under simple random sampling, the variance is reduced by a factor equal to one minus the sampling rate. In more complex sample designs, an analogous situation exists. For simplicity, estimates of variance are calculated under the assumption that with-replacement sampling has been used. This is especially true when unequal probability sampling has been used because in without-replacement sampling, unbiased estimates of variance require known joint probabilities of selection of all sampled units. Formulas based upon with-replacement sampling require knowing only the probability for selection of each sampled unit. Additionally, sampling past the first stage is not explicitly taken into account in with-replacement formulas, since this is not necessary for obtaining unbiased variance estimates when with-replacement sampling of first-stage units is used.

The simplicity of variance estimation under the assumption of with-replacement sampling has a price. When with-replacement formulae are used for without-replacement designs, the variance estimates are biased. This bias is actually equal to two times the reduction in variance brought about by using without-replacement sampling. This results in an interesting paradox. Without-replacement sampling is used to lower variance, but the easiest approach to variance estimation results in treating estimates as if they were less precisely measured than would have been the case had a with-replacement design been used. Often, the first-stage sampling rate will be low enough, or the first stage component of variance small enough, that it is reasonable to assume this effect will not be too great.

Often, prior to selecting the sample, the population is grouped into strata based upon characteristics thought to be related to the variable or variables of interest. This stratification usually reduces the variance of the population estimates. The stratification of primary sampling units can be taken to the point where only two are selected from each stratum. This allows the maximum stratification to be used while still allowing for the estimation of variance. If the stratification is ignored when variance is estimated, a positive bias will be introduced because the resulting estimate of variance will contain a "between strata" variance component. Where stratification has been carried out to the point where only one unit is selected per stratum, this "between strata" component is not easily avoided. In this situation, similar strata are combined so that a "within strata" variance component can be estimated with what is hoped to be a little, albeit positive, bias.

The above discussion is for the estimation of variance for linear statistics. Most statistics of interest are not linear. What is more, as discussed previously the weights often are modified for nonresponse and/or poststratification. The approximations typically employed are discussed in the next section. In summary, approximate variance estimators are required for the following reasons:

- No explicit variance estimator is available because the design does not allow for one in the case of systematic sampling or one PSU per stratum designs;
- Adjustments have been made to the sample weights;
- The variance of a nonlinear estimator is desired; and
- It is too much trouble to use one of the exact formulas.

2.1 Alternative Approximate Estimators

Two differing approaches are currently in widespread use for the estimation of survey sampling errors for complex parameter estimators: linearization and replication. The method of linearization provides a general approach through the use of linear approximation to the nonlinear estimation of interest. Explicit formulas for the estimate of variance for these linear approximations can then be derived. Variance estimation is achieved by estimating the variance of a linear combination of simple estimators whose variance is close to that of the complex estimator of interest. Linearized variance estimators have been described in a number of textbooks for estimating variance with ratio estimators (Cochran, 1977; Kish, 1965). Binder (1983) has developed a general method for obtaining an appropriate linearization estimator for a high proportion of the cases likely to be met in practice. It is important to keep in mind that the linearization approach does not actually yield an estimate of variance. Instead, the linearization approach provides a linear approximation to the quantity for which variance is to be estimated, after which the usual textbook formulas for the variance of a linear statistic are applied.

Replication (sample re-use) methods repeat the estimation process on a sequence of subsets of the full survey data set, and then compute the variance from the variation among these subsample estimates. The available replication methods differ as to their specification of the sample subsets or replicates and subsequent variance estimation formulae. Three general approaches in use are known as balanced repeated replication (BRR) or balanced half-sampling, jackknifing or jackknife repeated replication, and bootstrapping. Each method has variations of application which affect the number of replicate estimates derived in a given case. Wolter (1985), Rust (1985), and

Kalton (1977) give details of the implementation of BRR and jackknifing. Rao and Wu (1984, 1988) discuss the application of bootstrapping for a variety of sample designs.

Even for relatively simple quantities like means and totals, typical survey estimators involve the use of nonresponse and ratio adjustments to the weighting, resulting in weights that are random quantities, dependent upon the sample actually selected. A question to be addressed when comparing linearization with replication is the relative contribution of these adjustments to the variances of parameter estimates. While linearization can be undertaken in a manner accounting for this weight variability, such variance estimation does become cumbersome, whereas it remains relatively straightforward with replication. On the other hand, if the variability in weights can be safely ignored, for many parameters estimated from surveys, linearization can be undertaken straightforwardly in a much less computationally intensive manner than replication. Kish and Frankel (1974), and Bean (1975) have suggested that the contribution of such variation in weights to variance can be reasonably ignored, whereas Lemeshow (1979) cautions against this. Lemeshow's findings from simulation studies suggested that a substantial increase in the bias and variance of variance estimates could result from ignoring variability in the weights. Lago *et al.* (1987) describe an example of the analysis of data from the Hispanic Health and Nutrition Examination Survey in which the effects of ignoring the randomness of the poststratification adjustments led to a large upward bias in variance estimates.

2.2 Comparisons Appearing in the Statistical Literature

A number of investigations have been conducted into the properties, both theoretical and empirical, of linearization and replication. Though these studies did not investigate the effect of nonresponse and poststratification adjustments, the results are still of some interest. The results of many of these are reviewed in Rust (1985). Important among the empirical studies has been the work of Kish and Frankel (1974), and Frankel (1971), who undertook a large scale empirical study comparing the properties of linearization, BRR, and the jackknife. Their major finding was the similarity in performance of all three methods across a range of parameter estimators of varying complexity, from means to multiple correlation coefficients. They concluded that there was evidence that linearization gave somewhat greater accuracy (as measured by the mean square error) in variance estimation, but that replication methods, and in particular BRR, gave confidence interval coverage which was slightly closer to the nominal coverage rate.

Subsequent investigations have in the main concentrated on the aspects of bias and precision of variance estimation. Rao and Wu (1985) examined the asymptotic biases of linearization, BRR and jackknifing, considering a number of alternative forms of the replication methods. They showed that jackknifing in a number of forms and linearization were almost equivalent, while BRR was not nearly as equivalent to the other two procedures. These results were for multistage designs in which two primary sampling units (PSUs) are selected independently per stratum. Considering the combined ratio estimator specifically, Rao and Wu showed that the biases of the jackknife and BRR exceeded that of linearization under a particular population model.

Empirical studies, although few in number, are generally consistent with the theoretical results of Rao and Wu. Hansen and Tepping (1985) and Kovar, Rao and Wu (1988) used simulated data from a design with two PSUs selected independently from each of 32 strata. For the ratio estimator, both studies concluded that all methods performed well when the coefficient of variation of the denominator was below 10 percent. With a larger coefficient of variation for the denominator, BRR and the bootstrap became substantially positively biased, while the linearization and jackknife variance estimators showed slight negative bias. Kovar *et al.* concluded that for regression and correlation coefficients the substantial positive biases of BRR and the bootstrap, and the slight negative bias of linearization and the jackknife were evident regardless of the coefficients of variation of the component variables. Thus, these empirical results accorded well with the asymptotic results of Rao and Wu, demonstrating the similarity in performance of linearization and the jackknife, and the divergence of BRR from these two.

2.3 Inference

One might regard the results of such investigations as indications that the less biased methods of linearization and jackknifing are superior to BRR in terms of the resulting quality of variance estimation. Since the practical advantages and disadvantages of BRR are similar to those of the jackknife, if this conclusion is well-founded then it would seem that BRR should begin to lose favor. However, it must be remembered that the primary purpose of variance estimation in surveys is for making inference about parameters of the population, rather than about sampling errors. Thus, as suggested by Kish and Frankel (1974), the coverage of confidence intervals formed from variance estimates would seem to be of primary importance in assessing the relative merits of variance estimation techniques. Such assessment involves consideration of the joint properties of the parameter estimate and its variance estimate, making investigation of this issue complex. Frankel examined intervals of the form $\hat{\theta} \pm z_{\alpha/2} \sqrt{v(\hat{\theta})}$, where $\hat{\theta}$ is the estimated parameter, $v(\hat{\theta})$ is the estimated variance and $z_{\alpha/2}$ is the $\alpha/2$ critical value from the standard normal distribution. Empirical studies by Bean (1975), Campbell and Meyer (1978), Kovar *et al.* (1988), are notable for the similarities of their conclusions to those reached by Kish and Frankel (1974). In these investigations there was evidence that, in considering confidence intervals, BRR was somewhat superior to linearization and jackknifing (Bean did not consider jackknifing). These studies also indicate that the use of the confidence interval coefficients derived from an appropriate t-distribution may improve confidence interval coverage, but that the use of the number of strata as the degrees of freedom may not always be appropriate.

Thus, in considering the relative qualities of these different methods of variance estimation, further theoretical developments and empirical investigations of confidence interval coverage properties appear warranted. Such studies should also include consideration of bootstrap methods to assist in determining situations in which these present a better practical alternative than the established methods.

In addition to making inference using simple confidence intervals, it is frequently of interest to conduct tests of independence in two-way cross-tabulations using chi-squared statistics. Consider the situation where two variables, A with r levels and B with c levels, are cross-tabulated, producing weighted cell proportions p_{ij}, row

totals expressed as proportions, $p_{.j}$ and column totals expressed as proportions, $p_{i.}$. These estimates of proportion have estimated variances v_{ij}, for p_{ij}, $v_{.j}$ for $p_{.j}$, and $v_{i.}$ for $p_{i.}$.

The usual Pearson's chi-squared statistic, with (r-1)(c-1) degrees of freedom, can be calculated for the weighted sample estimates using the following formula:

$$X^2 = \frac{n \sum_{i=1}^{r} \sum_{j=1}^{c} (p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}}, \qquad (2.3.1)$$

where n is the unweighted sample size. It is widely accepted that the use of this statistic will result in too frequent rejection of the null hypothesis of independence. Various approaches have been suggested for modifying the Pearson chi-squared statistic to take into account the sample design. The simplest approaches are based upon correcting the Pearson chi-squared statistic by some form of average design effect. The adjusted chi-squared statistic is simply calculated using the following formula:

$$X^2_a = X^2 / \bar{b}, \qquad (2.3.2)$$

where $\bar{b}$ is the "average" design effect. Fellegi (1980) suggested that the average cell design effect might be used as an adjustment factor. One suggested average design effect is calculated using the following:

$$\bar{b} = \frac{n}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{v_{ij}}{p_{ij}(1-p_{ij})}. \qquad (2.3.3)$$

Rao and Scott (1979) proposed calculating the adjustment based upon the average eigenvalue of the following matrix:

$$\mathbf{D} = \mathbf{P}_0^{-1} \mathbf{V},$$

where $\mathbf{P}_0$ is the variance covariance matrix of the cell proportions under the null hypothesis of independence and simple random sampling and $\mathbf{V}$ is the variance covariance of the cell proportions taking into account the sample design (not under the null hypothesis). The matrix $\mathbf{D}$ has been called the generalized design effect matrix. The formula for the Rao and Scott average design effect reduces to the following relatively simple expression:

$$\bar{b} = \frac{n}{rc-1} \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{v_{ij}}{p_{ij}}. \qquad (2.3.4)$$

Note that the calculation of these adjustment factors only involves the use of the cell proportions and variances. This means that they can often be used with access to published crosstabulations of proportions and their associated variances.

Rao and Scott (1981, 1984) have also developed methods which are based upon the asymptotic behavior of Pearson's chi-squared statistic. Under the null hypothesis of independence the chi-squared statistic can be written as a weighted function of (r-1)(c-1) asymptotically independent χ^2_1 random variables, W_i,

$$X^2 \approx \sum_{i=1}^{(r-1)(c-1)} \delta_i W_i.$$

where "$\approx$" means asymptotically 'distributed as' and the δ_i's are the eigenvalues of

$$(\mathbf{V}_0(\mathbf{h}))^{-1}\mathbf{V}(\mathbf{h})$$

where

h_{ij} = $p_{ij} - p_{i.}p_{.j}$;

$\mathbf{h}$ = $(h_{11}, ..., h_{rc})$;

$\mathbf{V}(\mathbf{h})$ is the variance-covariance matrix for $\mathbf{h}$; and

$\mathbf{V}_0(\mathbf{h})$ is the variance-covariance matrix for $\mathbf{h}$ under the null hypothesis of independence and multinomial sampling.

It can be seen that X^2 has asymptotic mean and variance,

$$E(X^2) = \sum_{t=1}^{(r-1)(c-1)} \delta_t = (r-1)(c-1)\bar{\delta}$$

$$V(X^2) = 2 \sum_{t=1}^{(r-1)(c-1)} \delta_t^2,$$

respectively. Rao and Scott show that the expectation and variance can be written in terms of the variance of the h_{ij}'s:

$$(r-1)(c-1)\bar{\delta} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{v(h_{ij})}{p_{i.}p_{.j}} \text{ and}$$

$$\sum_{t=1}^{(r-1)(c-1)} \delta_t^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{i'=1}^{r} \sum_{j'=1}^{c} \frac{[cov(h_{ij},h_{i'j'})]^2}{p_{i.}(1-p_{.j})p_{i'.}(1-p_{.j'})}$$

One approach suggested by Rao and Scott is to standardize X^2 to have asymptotic mean equal to (r-1)(c-1), which is what would be expected if X^2 actually asymptotically follows a χ^2-distribution with (r-1)(c-1) degrees of freedom. This adjusted chi-squared variate takes on the form of (2.3.2), where

$$\bar{b} = \frac{n}{(r-1)(c-1)} \left[\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{v_{ij}}{p_{i.}p_{.j}} - \sum_{i=1}^{r} \frac{v_{i.}}{p_{i.}} - \sum_{j=1}^{c} \frac{v_{.j}}{p_{.j}} \right]. \qquad (2.3.5)$$

This expression is based upon a linear approximation to the h_{ij} in terms of the p_{ij}'s. Note that this approximation only requires estimates of the variance of the cell proportions and row proportions.

Rao and Scott have also proposed a more complicated approximation that standardizes for both the asymptotic expectation and variance following the approach of Satterthwaite (1946). In this approach, Pearson's chi-squared statistic is standardized to have asymptotic expectation ν and variance 2ν, where

$$\nu = \frac{[\Sigma\delta_t]^2}{\Sigma\delta_t^2}. \qquad (2.3.6)$$

This is done by modifying Pearson's chi-squared statistic to the following:

$$X^2_a = X^2 \frac{\Sigma\delta_t}{\Sigma\delta_t^2}, \qquad (2.3.7)$$

which is then treated as a chi-squared variate with ν degrees of freedom. As with the chi-squared statistic that has been standardized to have the correct expectation, the correction factor can be based upon either $v(h_{ij}, h_{i'j'})$, which can be calculated using replication, or upon the variance-covariance matrix of a linear approximation to the h_{ij}'s using the p_{ij}'s.

Because these approximate chi-square tests have not been used extensively in the past, a consensus has not developed as to which approximation tends to be best. The Fellegi and first Rao and Scott statistics are the simplest to calculate, but are thought to be excessively conservative. The second Rao and Scott statistic attempts to compensate for the suspected conservatism shown by the simpler statistics, but it is not clear at the present time if this is generally achieved in practice. In general, the Satterthwaite adjusted statistic which does require more computational effort is likely to be the best approximation. Westat's limited experience has been that the simpler statistics are not too different from each other, but that the more complex statistics, as a group, can be either higher or lower. It is recommended that preliminary tests always be carried out using all four approximations. If one or more of the alternatives to the Satterthwaite adjusted statistic is comparable, then it may not be necessary to calculate the more complicated statistic for each table analyzed.

2.4 Westat's Preference for Replication

The above discussion implies that the linearization approach is best suited to situations where both the the sample design and estimators are used repeatedly. The steps required are to linearize a specific estimator and determine its variance under a specific sample design. While it is always possible to develop a linearization for a new estimator and to use the formula appropriate for estimating the variance of the linear components given the sample design (or an approximation to it), this will not always be practical for a one time survey. In practice, the user of computer software attempts to find a preprogrammed estimator which approximates the estimator of interest and also assumes a sample design for which the variance of the linear statistics will not be too dissimilar. This generally means assuming that a with-replacement sample design has been used and that weights adjusted for nonresponse and poststratification are in fact not subject to sampling variation. For example, when estimating the variance of a ratio estimator, it is assumed that the variance of

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{i=1}^{n} b_i a_i r_i w_i y_i}{\sum_{i=1}^{n} b_i a_i r_i w_i x_i}$$

can be approximated by using the variance estimator appropriate for

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{i=1}^{n} w_i^* y_i}{\sum_{i=1}^{n} w_i^* x_i} ,$$

where $w_i^* = b_i a_i r_i w_i$ is taken as being a function only of the i-th unit itself, rather than the sample as a whole. As discussed above, the adequacy of the approximation resulting from the use of this assumption may be questionable, especially if the nonresponse and poststratification adjustments have had a substantial impact on the level of sampling variance.

Conversely, replication procedures are well-suited to situations where a variety of complex estimators are needed, including secondary analyses of survey data which were not specifically envisioned and planned for at the time of initial analysis of the survey data. Replication procedures can be adapted to incorporate more readily into the variance estimates the effects of nonresponse and poststratification adjustment. Additionally, replication procedures are easily adaptable to a wide range of sample designs without the need for reprogramming. The combination of these factors lies behind Westat's preference for the use of replicated variance estimation procedures in the analysis of survey data.

The typical circumstances under which Westat conducts a survey include essentially a one-time survey design, rather than say a monthly or quarterly series in which the same quantities are estimated each time. Following the completion of data collection, Westat derives a data file with survey weights adjusted for nonresponse and frequently poststratified to agree with "known" totals. A report of the major survey findings is produced, including estimates of population summary characteristics, frequently in the form of means and proportions for the total population and for subgroups. For these major findings estimates of sampling error are provided. Finally, a data file is assembled, which includes survey weights and an appropriate set of replicate weights, with instructions for the estimation of sampling errors for estimates derived using the survey weights. Frequently the data file is released in the form of a public use data file, with accompanying documentation.

By using replicate weights and Westat's own replication based variance estimation software, the various aspects of variance estimation involving the survey data can be handled via a single approach. The use of replication can account for nonresponse and poststratification adjustment in a standard way, and the replicate weights derived can then be used for Westat's own analyses of the data, as well as providing a very simply implemented means of variance estimation for secondary users. With replicated procedures specialized variance estimation software is not required, and appropriate variance estimates can be readily derived, albeit somewhat laboriously, using standard statistical packages and/or one-time programs. When numerous sampling error estimates are required, it becomes much more convenient to employ prewritten software suitable for a wide range of applications, and it is this need that the WESVAR program is intended to fill (along with the WESREG and WESLOG programs designed for specific modelling applications).

2.5 Procedures for Creating Replicates

The use of replication via a set of replicate weights is described in Dippo, Fay and Morganstein (1986). The general procedure is to form, implicitly, a succession of replicate data sets, using a version of BRR or jackknife, each comprised of a subsample of the full data set. Each replicate in turn is weighted appropriately to represent the same population as represented by the full set of data. This weighting procedure includes the nonresponse adjustment and poststratification, as implemented with the full sample. It is this latter component which permits this procedure to reflect appropriately the effect on variance of these aspects of estimation. The weight associated with one repetition of this procedure constitutes a single set of replicate weights, which are attached on the data file to their respective units. Those dropped from the data set for a given replicate receive a zero weight for that replicate.

There are some practical difficulties associated with the task of forming replicate weights, and these fall into two main classes. The first involves the method of forming the replicate data sets in such a way that the sample design actually used is reflected appropriately in variance estimation, at least approximately, while at the same time keeping the number of replicates formed to a manageable

level. The second involves the methods of reflecting appropriately the procedures for nonresponse and poststratification adjustment.

The formation of replicate groups involves two distinct processes. The first of these is to set up replicates and their complementary drop-out sets in such a way that the features of the sample design are appropriately reflected, leading to little bias. This must be achieved by approximating the sample design as a stratified, (possibly) multi-stage design, in which at least two PSU's are selected from within each stratum using with-replacement sampling -- that is, selections within strata are independent. Often this involves the approximation of a two PSU per stratum design. For example, when a one PSU per stratum design is used, a standard approach is to pair strata and treat the data as having been drawn as two PSU's per stratum with replacement. This collapsing of strata in general leads to a small positive bias in variance estimation, provided that strata similar with respect to survey characteristics are paired, and that this pairing is performed on the basis of frame stratum information, not sample data. Pairing on the basis of sample characteristics (or frame characteristics) of the selected PSU's likely results in a substantial negative bias in variance estimation.

Depending upon the sample design actually employed, procedures other than pairing of PSU's are available. For example, if systematic selection is used, replicates can be formed by dropping every r-th selected PSU from the data, where the selected PSU's are sorted in their original order of sample selection. The technique of pairing is frequently employed with multistage designs, and gives rise to the greatest need for methods to reduce the extent of replication. Thus, we concentrate on this approach in the remainder of this section.

If 2r PSU's are selected, resulting in r pairs, then a total of r replicates can be formed by dropping one member of each pair from the data in turn. In many instances and particularly when some PSU's are included with certainty, resulting in the second-stage units becoming the true PSU's within each of these certainties, the resulting number of replicates r may be several hundred in magnitude. In such a case the derivation of the full set of replicate weights and their use in variance estimation becomes burdensome. An approach is needed which reduces the amount of replication without introducing further bias into the variance estimates.

The approach of "partial balancing" or combining of pseudostrata handles this situation. Applied correctly this approach reduces the number of replicates without introducing bias. Inevitably there will be some loss in the precision of variance estimation (as reflected in the actual degrees of freedom of the variance estimator) but often this is of little consequence in comparison with the great reductions in the extent of replication required. The procedure consists of first randomly designating which member of each pair is to be the one dropped in replication, with probability one half and independently from pair to pair. Then two or more pairs are combined, in that a single replicate is formed by dropping simultaneously all of those PSU's belonging to the combined group designated to be dropped via the above random procedure. Thus, if the r pairs are partitioned into r'< r groups, only r' replicates result, and hence r' replicate weights are formed. Methods of achieving this combining without undue loss of degrees of freedom are discussed by Lee (1972, 1973), who refers specifically to BRR, and Rust (1986), who makes particular reference to jackknifing.

2.6 Procedures for Defining Adjustment Cells

The second area of practical consideration in using replication procedures is the procedure for defining nonresponse classes and poststrata. In developing survey weights it is good practice to ensure that the number of respondents falling within a given nonresponse class or poststratum is sufficiently large in expectation as to ensure that undue variability in the resulting adjustment factor does not result. When replicates are to be formed, it is important to ensure further that these small and unstable sample sizes do not occur in any substantial number of replicates, or else a bias will be introduced into the variance estimates. In particular, there is no unbiased procedure available for handling the case where a replicate estimate is undefined because a nonresponse cell or poststratum is empty. This problem is seldom of great concern when using the jackknife procedure, since most of the sample units are retained in any given replicate. With BRR, however, more careful attention is required since generally only about one-half of the units appear in a given replicate.

As a result of these considerations, it may be that on occasion fewer and somewhat larger nonresponse classes and poststrata are utilized than would be the case if a linearization approach to variance estimation were used. Such differences in procedure will almost always have a negligible effect on the precision and bias of the overall estimator, since most of the gains from using nonresponse classes and poststratification can generally be obtained using only a few cells, each having a large sample size within each cell. Thus, the major concern is the practical one of actually developing classes of sufficient size, rather than concern about precision of the parameter estimator.

3. WESVAR

3.1 Historical Development of WESVAR

The WESVAR procedure has been previously described by Mohadjer *et al.* (1986). This program was based upon an earlier version called NASSVAR (Binzer and Morganstein, 1983). The latest version of WESVAR uses the same basic approach as these earlier versions, but adds features to make the program both easier to use and provide additional analytic capabilities. The main change in the program is the inclusion of a TABLES statement. This statement allows the user to specify one variable or two variables together to create "cells" within which estimation takes place. For example, estimates may be desired for cells formed by the combination of respondent's age and sex. Previously, users had to create indicator variables to represent the cells of the table. Several analytic improvements have also been added in conjunction with the TABLE statement. Chi-squared tests of independence are now available based upon the distribution of the sample weights. Additionally, complex functions of cell estimators can now be formed using the FUNCTION statement. For example, log-odds ratios can be contrasted using the FUNCTION statement. Additionally, substantial computational efficiencies were incorporated via a GROUP option. The following sections describe the various analysis options available in WESVAR.

3.2 Weighting

The procedure WESVAR assumes the presence of the survey weights necessary for the estimation of sampling errors. This computer file must contain a SAS variable

containing the full sample weight, as well as one SAS variable for each of the replicate weights. The user of WESVAR lists, on a WEIGHT statement, the SAS variable containing the full sample weight, followed by the SAS variables containing the replicate weights. WESVAR makes very simple checks on the weights to insure that none of the weights are missing and that the full sample weight is positive and that the replicate weights are positive. It is the user's responsibility to insure that the weights specified are appropriate for the replication technique WESVAR is directed to use, and the actual design of the sample.

3.3 Simple Statistics - Sampling Errors

The parameters that will typically be of interest to users of WESVAR are totals, ratio means, proportions, general ratios or other functions of totals. Frequently, it also is of interest to analyze these variables for subgroups of the population. This type of analysis will frequently include the use of crosstabulations. WESVAR can be used to estimate such sample statistics and their sampling errors.

WESVAR operates by calculating totals for the variables of interest. These variables are listed on the VAR statement. Additional variables can be created by manipulating these totals using the COMPUTE statement. If a variable used in a compute statement has not been listed on a VAR statement, the total for that variable is calculated, but statistics are not printed.

If there are n records in the file and the variable of interest is represented by 'y', the population total is estimated by the following formula:

$$\hat{Y} = \sum_{i=1}^{n} w_i y_i ,$$

where w_i is the full sample weight and y_i is the observed value of y for the i-th unit in the sample (ignoring any stratification). WESVAR calculates this quantity for each variable listed on the VAR statement. A similar quantity is also estimated by WESVAR for each of the replicates using the replicate weights instead of the full sample weight. The replicate estimates are not printed by WESVAR, but are used for variance estimation and are available in one of the optional output data sets. A similar process is followed for each COMPUTE variable. If x is another variable in the data set, specifying

```
COMPUTE   R = Y/X;
```

leads WESVAR to calculate

$$\hat{R} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i x_i} .$$

The important thing to note is that the equation given calculates the weighted total for each variable used on the right-hand side of the COMPUTE statement, and then calculates the quantity of interest using these weighted sums.

The variance of the estimates produced using the VAR and COMPUTE statements are calculated in an identical fashion. If the population parameter of interest is represented by the symbol θ, then $\hat{\theta}$ is used to represent the full sample estimate and $\hat{\theta}_k$ is used to represent the estimate for the k-th replicate. The parameter θ can be any of the parameters discussed in the previous section: total, ratio mean, proportion, general ratio, etc.

The user can direct WESVAR to use one of three replication approaches for estimating variance: balanced repeated replication (BRR), jackknife #1 (JK1), and jackknife #2 (JK2). The two jackknife estimates are designed to handle different sampling situations.

JK1 is usually applied when no stratification has been used to select the sample. It can also be adapted to handle the situation where only a few strata have been used. To form the replicates for use with JK1, sampled PSU's are grouped into G random subsets of equal or nearly equal size (G≤#PSU's) with each subset resembling the full sample. Replicates are formed by deleting a single group.

The basic sample design assumed for JK2 is the same as that used for BRR, two first-stage selections (PSU's) made with replacement in each of L strata. The primary difference between BRR and JK2 is in the formation of replicates once the PSU's have been grouped into pairs. In JK2, one PSU is deleted from a single stratum to form the replicate. This process is repeated in turn for each stratum. This means that if there are L strata, or pseudostrata, then L replicates will be created for use with JK2.

The three replication techniques calculate the variance estimate for $\hat{\theta}$ using a slightly different formula, for a given number of replicates G:

$$\text{BRR:}\quad v(\hat{\theta}) = \frac{1}{G} \sum_{k=1}^{G} (\hat{\theta}_k - \hat{\theta})^2 ,$$

$$\text{JK1:}\quad v(\hat{\theta}) = \frac{G-1}{G} \sum_{k=1}^{G} (\hat{\theta}_k - \hat{\theta})^2 \text{ , and}$$

$$\text{JK2:}\quad v(\hat{\theta}) = \sum_{k=1}^{G} (\hat{\theta}_k - \hat{\theta})^2 .$$

In some situations (Wolter, 1985; Judkins, 1987) replicates are formed in such a way that each term $(\hat{\theta}_k - \hat{\theta})^2$ must be multiplied by an adjustment factor, F_k, to produce unbiased estimates of variance. The FACTOR statement allows these factors to be specified, one factor for each replicate, and is described in Section 3.7.

3.4 Subgroup Analyses -- TABLE Statement

The estimation of variance for variables in WESVAR is controlled primarily through three statements: COMPUTE, TABLE and FUNCTION. As mentioned previously, the COMPUTE statement is used to manipulate estimated totals calculated for variables in the data set using simple arithmetic operations. The COMPUTE statement is based upon the manipulation of weighted totals calculated for the entire sample. The TABLE statement allows these totals to be estimated and manipulated for subgroups formed by one or two categorical variables.

TABLE requests/options;

As an example of a subgroup analysis, consider estimating the ratio mean of the average number of years of schooling for two different regions of the country. The following statements can be used to estimate these quantities:

```
COMPUTE MEAN = Y/C;
```

TABLE REGION: ,

where C is a variable which is equal to "1" for all records and REGION is a SAS variable equal to either "1" or "2". WESVAR calculates the following for the new SAS variable MEAN:

$$\hat{\bar{Y}}_1 = \frac{\sum_{i=1}^{n_1} w_i y_i}{\sum_{i=1}^{n_1} w_i} \quad \text{and} \quad \hat{\bar{Y}}_2 = \frac{\sum_{j=1}^{n_2} w_j y_j}{\sum_{j=1}^{n_2} w_j},$$

where n_1 is the number of sample elements in the first region, n_2 is the number in the second and where the mean for region 1 is represented by $\hat{\bar{Y}}_1$ and the mean for region 2 is represented by $\hat{\bar{Y}}_2$. Two-way tables can be formed by simply using the following syntax: TABLE variable 1*variable 2;.

A number of options are available for use with the TABLE statement. These options are primarily designed to control the display of statistics calculated for the cells formed by the variables making up the table request. For example, the options can be used to determine if the statistics should be printed in the form of a table or simply listed one cell at a time. Other options determine if the results should be percentaged by rows, columns or overall. Additionally, chi-square statistics are available for testing independence.

3.5 Complex Comparisons of Table Values - FUNCTION Statement

Since it is often of interest to make comparison among subgroup estimates, a FUNCTION statement has also been included in WESVAR to allow comparisons among subgroups to be made using functions of the crosstabulation cells.

Continuing with the previous example, the estimate of the difference between the two regions, $D = [\bar{Y}_1 - \bar{Y}_2]$, can be written as

$$\hat{D} = \hat{\bar{Y}}_1 - \hat{\bar{Y}}_2 .$$

The following sequence of statements will calculate this estimator

```
COMPUTE MEAN = Y/C;
TABLE REGION;
FUNCTION REGION [1]-[2] FOR MEAN;.
```

The FUNCTION statement can also be used to calculate functions of two-way table cells. For example, if A takes on 3 values (1,3,4) and B takes on 2 values (0,1), the log-odds ratio of A=1 to A=3 (B=1 signifies that the event of interest has occurred), the following FUNCTION statement would be used:

```
FUNCTION  'LOG-ODDS A1/A3'  A*B

LOG( ([1,1]*[3,0]) / ([3,1]*[1,0]) ) FOR WEIGHT_;.
```

The function of table cells may include any of the five standard arithmetic operators: multiplication (*), division (/), addition (+), subtraction (-) and exponentiation (**). The "-" can be used as a prefix operator to indicate a negative number. Additionally, the following functions are available: EXP (raises e, the base of natural logarithms, to the specified power), SQRT (square root), LOG (natural logarithm), LOG2 (logarithm to the base 2) and LOG10 (logarithm to the base 10). Normal SAS rules of precedence govern the order of evaluation; parentheses may be used to change this order if desired.

3.6 Inference - Confidence Intervals and Chi-square Tests of Independence

As mentioned previously, one of the primary purposes of the estimation of variance is to make inferences about population quantities. WESVAR approaches this through either the use of confidence intervals or through the use of chi-square tests of independence. Two-sided confidence intervals can be created for any basic statistic estimated by WESVAR using the VAR, COMPUTE and FUNCTION statements. It is possible to specify the level of the confidence interval in the PROC statement. For the level specified, WESVAR finds the critical value appropriate for that level of the confidence using the standard normal distribution. Confidence intervals take on the following form:

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{v(\hat{\theta})} .$$

Simple tests of significance between two groups can be created by using the FUNCTION statement to compute the appropriate difference. If the interval does not contain 0, the two groups can be concluded to be significantly different. More complex tests can be created by considering other weighted combinations of cell estimates. At a later date, WESVAR will be able to utilize the t-distribution with user-specified degrees of freedom.

WESVAR can also be used to generate chi-square tests of independence for two variables. These test statistics are available as option in the TABLE statement. Option CHISQ1 will produce the three simpler average design effect corrected chi-squared statistics, described in Section 2.3 and specified by expression (2.3.1) through (2.3.5). The choice of the CHISQ2 option results in the Satterthwaite adjusted chi-squared statistic given by (2.3.7), being calculated.

3.7 General Replicated Variance Estimation Formulae -- FACTOR Statement

The previously discussed estimators of variance can be called "traditional" replicate estimators. In some applications, the use of the variance estimation formulae presented earlier may not be appropriate for the particular replicates being used. WESVAR contains a FACTOR statement to allow other replicate estimators to be calculated. The FACTOR statement allows the squared deviation produced from each replicate to be multiplied by a specified value. The list of values in the FACTOR statement contain these factors, with the i-th factor used to adjust the i-th replicate. The values are listed after the word FACTOR with a space separating each value. Currently, there must be as many factors as there are replicates if the FACTOR statement is used. The resulting variance estimate is used for all computations, including the chi-square statistics. If the i-th factor is represented by F_i and the corresponding parameter estimate is $\hat{\theta}_i$, the replicate term $(\hat{\theta}_i - \hat{\theta})^2$, is multiplied by the adjustment factor, to give $F_i(\hat{\theta}_i - \hat{\theta})^2$. As an example of the syntax, consider the situation with four replicates and an adjustment of .667 for each replicate. The following FACTOR statement would be used:

```
FACTOR    .667 .667 .667 .667;  .
```

One application of the FACTOR statement is the implementation of Fay's modified BRR approach to variance estimation (Dippo, Fay and Morganstein, 1984; Judkins, 1987). In this approach, the usual BRR replicate weights within a stratum, i.e. either 0 or 2, are replaced by weights of k and 2-k for $0 \le k < 1$. The F_i's are set to $1/(1-k)^2$ for all replicates.

The FACTOR statement can also be used in conjunction with a jackknife estimator when more than two PSU's are selected in one or more strata. For example, if in each stratum three PSUs have been sampled, a jackknife estimator of variance can be based upon replicates formed by dropping each PSU in turn. This leads to forming 3*L replicate estimates. Setting F_i to 2/3 with WESVAR option JK2 will produce unbiased estimates of variance for linear statistics. Wolter (1985) discusses the general situation of weighting jackknife replicate estimators.

3.8 Availability

The WESVAR procedure is a Westat-written SAS procedure that can be accessed from SAS after installing a load module on the computer where SAS is running. Currently, WESVAR is operating on DEC/VAX and IBM mainframe computers. WESVAR can be installed on other computers where SAS supports user-written procedures. Since SAS does not support user-written procedures for IBM compatible PCs, no short term plans exist for making WESVAR available for these computers. A manual for the current β version is available upon request. Arrangements for program installation can be made by contacting David Morganstein at Westat.

References

Bean, J.A. (1975). *Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples.* Vital and Health Statistics, Series 2, No. 65. National Center for Health Statistics, Public Health Service. Washington, D.C.: U.S. Government Printing Office.

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review* 51, 279-292.

Binzer, G., and Morganstein, D. (1983). Automation of Estimation and Sampling Error for Computations using PROCs NASSTIM and NASSVAR. *SAS Users Group International* 8, 838-843.

Campbell, C., and Meyer, M. (1978). Some properties of T Confidence Intervals for Survey Data. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 437-442.

Cochran, W.G. (1977). *Sampling Techniques.* Third Edition. New York: Wiley.

Dippo, C.S., Fay, R.E., and Morganstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 113-121.

Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. *Journal of the American Statistical Association* 75, 665-670.

Frankel, M.R. (1971). *Inference from Survey Samples.* Ann Arbor: Institute for Social Research, The University of Michigan.

Hansen, M.H. and Tepping, B.J. (1985). *Estimation of Variance in NAEP.* Unpublished Westat Manuscript.

Judkins, D. (1987). Modified Balanced Repeated Replications. *Proceedings of the American Statistical Association Section Survey Research Section*, 492-495.

Kalton, G. (1977). Practical Methods for Estimating Survey Sampling Errors. *Bulletin of the International Statistical Institute* 47, 495-514.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Kish, L. and Frankel, M.R. (1974). Inference from Complex Surveys. *Journal of the Royal Statistical Society Series B* 36, 1-37.

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, Supplement, 25-45.

Lago, J., Massey, J.T., Ezzati, T., Johnson, C. and Fulwood, R. (1987). Evaluation of the Design Effects for the Hispanic Health and Nutrition Examination Survey. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 595-600.

Lee, K-H. (1972) Partially balanced designs for half sample replication method of variance estimation. *Journal of the American Statistical Association* 67, 324-334.

Lee, K-H. (1973) Using Partially balanced designs for the half sample replication method of variance estimation. *Journal of the American Statistical Association* 68, 312-14.

Lemeshow, S. (1979). The Use of Unique Statistical Weights for Estimating Variances with the Balanced Half-Sample Technique. *Journal of Statistical Planning and Inference* 3, 315-323.

Mohadjer, L., Morganstein, D., Chu, A. and Rhoads, M. (1986). Estimation and Analysis of Survey Data using SAS Procedures WESVAR, NASSREG, and NASSLOG. *Proceedings of the American Statistical Association Social Statistics Section*, 258-263.

Rao, J.N.K. and Scott, A. J. (1979). Chi-squared Test for Analysis of Categorical Data from Complex Surveys. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 58-66.

Rao, J.N.K. and Scott, A. J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association* 76, 221-230.

Rao, J.N.K. and Scott, A. J. (1984). On Chi-squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Sample Surveys. *Annals of Statistics* 76, 46-60.

Rao, J.N.K. and Wu, C.F.J. (1984). Bootstrap Inference for Sample Surveys. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 106-112.

Rao, J.N.K. and Wu, C.F.J. (1985). Inference from Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics. *Journal of the American Statistical Association* 80, 620-630.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association* 83, 231-241.

Rust, K.F. (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics* 1, 381-397.

Rust, K.F. (1986). Efficient Replicated Variance Estimation. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 81-87.

Satterthwaite, F.E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics* 2, 110-114.

Wolter, K.M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.