Chapter XXI
Sampling error estimation for survey data*

Donna Brogan
Emory University
Atlanta, Georgia
United States of America

Abstract

Complex sample survey designs deviate from simple random sampling, including aspects such as unequal probability sampling, multistage sampling and stratification. Weighted analyses are necessary for unbiased (or nearly unbiased) estimates of population parameters. Variance estimation for estimators depends upon the sampling plan specifics and requires approximate methods, generally Taylor series linearization or replication techniques.

Standard statistical software packages generally cannot be used to analyse sample survey data since they typically assume simple random sampling of elements. These packages yield biased point estimates of population parameters (in an unweighted analysis) and/or underestimation of standard errors for point estimates. Using the sampling weight variable with standard packages yields appropriate point estimates of population parameters. However, estimated standard errors usually are still incorrect because the variance estimation procedure typically does not take into account the clustering and/or stratification of the sampling plan.

The present chapter gives an overview of eight software packages with capability for sample survey data analysis, including approximate cost, variance estimation methods, analysis options, user interface, and advantages/disadvantages. Four of the packages are free, hence possibly of interest to developing countries that have a limited budget for software acquisition.

A complex sample survey data set from Burundi illustrates that incorrect analyses are obtained from standard statistical software. Annotated descriptive analyses with the Burundi survey for five of the eight reviewed packages (STATA, SAS, SUDAAN, WesVar and Epi-Info) show how to use these packages. Finally, numerical results from the five software packages are compared for common analytical objectives with the Burundi survey data. All five packages give equivalent variance estimation results whether Taylor series linearization or balanced repeated replication (BRR) is used.

Key terms: Taylor series linearization, replication methods, ultimate cluster, variance estimation, complex sample surveys, software packages.

---

* This chapter includes an Annex (English only) containing illustrative and comparative analyses of data from the Burundi Immunization Survey using five statistical software packages. The contents of the CD-ROM, including program codes and output for each of the software packages, may be downloaded directly from the UN Statistics Division website (http://unstats.un.org/unsd/hhsurveys/) or the CD-ROM may be made available upon request from the UN Statistics Division (statistics@un.org).

A. Survey sample designs

1. As illustrated in many chapters in the present publication, the sample designs for household surveys are complex ones, typically involving stratified multistage sampling. A consequence of the use of a complex sample design is that standard statistical methods and software cannot be applied uncritically for the analysis of household survey data. In particular, the responding units in a survey are assigned weights that compensate for unequal selection probabilities, unit non-response, and non-coverage and that may be used to make weighted survey distributions for certain variables conform to known distributions for those variables. These weights need to be employed in the survey analysis. Also, the computation of sampling errors for survey estimates needs to take into account the fact that the survey sample was selected using a complex sample design. Fortunately, there are now a number of specialist software packages for survey analysis that compute sampling errors correctly for weighted survey estimates from complex sample designs. The present chapter describes and reviews some of these packages.

2. As preparation for the discussion of the survey analysis packages, the next two sections review the issue of weighted analyses and methods of variance estimation with complex sample designs. The following sections compare eight software packages for variance estimation for estimates derived from complex sample survey data and illustrate the use of five of them with data from a sample survey in Burundi. The chapter ends with some conclusions and recommendations. The annex contained in the CD-ROM associated with this publication provides annotated data analyses for three analyses conducted with the selected five software packages.

B. Data analysis issues for complex sample survey data

1. Weighted analyses

3. In many household surveys the units of analysis - households or persons - are selected with unequal probabilities, and weights are needed to compensate for these unequal selection probabilities in the analyses. Further, even when the units are selected with equal probability, weights are often needed to compensate for unit non-response and also for benchmarking, such as post-stratification (see chap. XIX). These weights should be used in the analyses to estimate population parameters. Unweighted estimators (not recommended) may be badly biased for population parameters, depending upon the specific survey. The value of the sample weight variable, denoted by *WTVAR*, for a given respondent sample element R in the data set can be interpreted as the number of elements in the population represented by that R. The sum of the value of *WTVAR* over all Rs in the data set estimates the number of elements in the population.

4. Sometimes, the sampling weight variable *WTVAR* is "normed" by multiplying it by (number of Rs) / (sum of value of *WTVAR* over all Rs). The sum of the value of the "normed weight variable" *WTNORM* over all Rs is the sample size for analysis (number of Rs). It does not matter whether the sample weight variable *WTVAR* or the normed weight variable WTNORM is used to obtain a point estimate of an "average" population parameter such as a

mean or proportion: both yield the same calculation. However, the normed weight variable WTNORM cannot be used to directly estimate population parameter totals, for example, the total number of malnourished children in the population.

2. Variance estimation overview

5. Variance estimation is important because it indicates precision of estimators, leading to confidence intervals for and testing hypotheses about population parameters. Variance estimation for estimators based on complex sample survey data must recognize the following factors: (a) most estimators are non-linear (a ratio of linear estimators is common); (b) estimators are weighted; (c) the sampling plan will generally have used stratification prior to first-stage sampling (and perhaps also at subsequent sampling stages); and (d) elements in the sample will generally not be statistically independent owing to multistage cluster sampling. In almost all cases, it is not possible to obtain a closed-form algebraic expression for the estimated variance. Thus, the research literature on variance estimation for complex sample survey data contains several approximate methods from which sample survey data analysts can choose.

6. The two most commonly used approaches to approximating the estimated variance are Taylor series linearization (TSL) (Wolter, 1985; Shah, 1998) and replication techniques (Wolter, 1985; Rust and Rao, 1996). These approaches are discussed more fully in section C. Most software packages that analyse sample survey data implement only one of these two methods. For estimators that are smooth functions of the sample data (for example, totals, means, proportions, differences between means/proportions, etc.), both methods give comparable variance estimates and neither is clearly preferred. For estimators that are non-smooth functions of the sample data (for example, medians), a particular replication procedure, balanced repeated replication, seems preferred over Taylor series linearization and jackknife, another replication method (Korn and Graubard, 1999). There is a substantial literature on comparison of variance estimation techniques, including particular instances where one method may be preferred over another [for example, see Korn and Graubard (1999) and their many references and, also, Kish and Frankel (1974)].

3. Finite population correction (FPC) factor(s) for without replacement sampling

7. For simplicity, consider initially the estimate of a population mean from a sample of size n selected with equal probability from a population of size N, and compare two sample designs. In one design, the elements are selected by simple random sampling, that is to say, they are selected without replacement. In the other design, they are selected by unrestricted sampling, that is to say, with replacement (also termed simple random sampling with replacement). The difference in the variance for the sample means with these two designs is that a finite population correction (*fpc*) term is included in the variance with the simple random sample but not in that with the unrestricted sample (see chap. VI). The *fpc* term is $(1-f)$ where $f = n/N$ is the sampling fraction. The *fpc* is bounded above by 1.0 and reflects the reduction in variance resulting from sampling without replacement. If the sampling fraction f is small, the *fpc* term is close to 1.0 and has minimal impact on the variance. It can then be safely ignored in variance estimation. In other words, the without replacement sample may be treated as if it had been sampled with replacement. A small sampling fraction generally is considered to be up to 5 or 10

per cent. On the other hand, if *f* is large, ignoring the *fpc* term when the sample is selected without replacement will lead to an overestimate of the variance. In a stratified random sample design with different sampling fractions in different strata, the *fpc* term may be small enough to be ignored in some strata but not others.

8. Most household surveys are based on complex sample designs applied to very large populations. The PSUs are generally selected using probability proportional to size (PPS) without replacement sampling, making the concept of "sampling fraction" more complex. However, the number of PSUs is often large and the PSU sampling fraction in each stratum is fairly small, giving a value close to 1.0 for all first-stage *fpc* terms. Thus, a common approximation in the analysis of complex sample survey data is one where the PSUs have been sampled with replacement in each stratum. If this approximation is made in the presence of some strata with large first-stage sampling fractions, the variance will be overestimated to some extent. Such overestimation is often accepted in view of the complexity of variance estimation without the approximation. Note that if sampling is done with replacement at the first stage of sampling in any stratum, there is no approximation involved for that stratum.

4. Pseudo-strata and pseudo-PSUs

9. For the purpose of variance estimation, sometimes the strata and PSUs are not identified as they actually were used in the sampling plan. Modifications in defining strata and PSUs for variance estimation may be made to make the sampling plan actually used fit into one of the sampling plan options available in a software package. When such modifications are made, the newly defined strata and PSU variables for variance estimation sometimes are called pseudo-strata and pseudo-PSUs.

10. A common example arises when a very large number of strata are defined prior to first-stage sampling, with only one PSU selected (sampled) within each stratum. Variance estimation is impossible with only one PSU per stratum, since between PSU variability within the stratum cannot be estimated. In this situation, two strata are collapsed or combined into one pseudo-stratum, thus giving two sample PSUs within that pseudo-stratum. Collapsing strata is carried out strategically, not arbitrarily, and is based on knowledge of the PSU stratification variable(s) and method of PSU sampling (Kish, 1965).

11. Another example arises with implicit stratification. A country may, for example, be stratified by north and south, with PSUs defined by villages. Within each stratum, the PSUs are ordered by geographical proximity, followed by selection of a probability sample of many (say, 30) PSUs within each stratum using systematic PPES (probability proportional to estimated size) sampling (Kish, 1965). The geographical ordering of the population PSUs within stratum, combined with systematic sampling, results in implicit geographical stratification of the villages (PSUs) within each of the north and south strata. In order to recognize the implicit stratification in variance estimation, the sampling plan typically would be described as encompassing 15 northern pseudo-strata and 15 southern pseudo-strata, each with two sampled PSUs or pseudo-PSUs. The first two PSUs sampled from the north sampling frame would go into the first pseudo-stratum, the next two PSUs sampled into the second pseudo-stratum, etc.

12. Korn and Graubard (1999) give several additional examples where pseudo-strata and pseudo-PSUs are formed for variance estimation purposes, for example, to reduce the number of replicates and computational load. Also, appendix D of the WesVar User's Guide (2002) gives guidance and examples in describing various sampling plans to variance estimation software based on replication methods.

5. A common approximation (*WR*) to describe many complex sampling plans

13. Complex sample surveys typically use multistage cluster sampling. In addition, stratification of population PSUs prior to first-stage sampling is usual. Further, stratification of second and subsequent stage units (within a sample PSU) may occur before sampling at these stages. However, the approximate methods of variance estimation commonly used for these complex designs do not need to take into account all stages of sampling and stratification. Complex sampling at later stages is automatically covered appropriately under the "with replacement" approximation for the first stage of sampling discussed above. In fact, few sample survey software packages have the capability to include all stages of sampling separately in variance estimation in cases where the first stage with replacement approximation is not made.

14. It is very common to use the ultimate cluster variance estimate (UCVE**)** for complex designs, first proposed by Hansen, Hurwitz and Madow (1953) and discussed also in Wolter (1985). The ultimate cluster variance estimate may be implemented with either Taylor series linearization or a replication technique. The UCVE approach treats the PSUs as if they were sampled with replacement within first-stage strata. Then, each R (sample respondent element in the data set) needs to be identified only by the first stage stratum and PSU (within stratum) from which it was selected. Information on sampling stages below the PSU level but before the element stage is not needed for the purpose of variance estimation. Thus, the description of the actual sampling plan is simplified so that it looks like stratified one-stage cluster sampling, that is to say, a stratified sample of completely enumerated ultimate clusters. This ultimate cluster approach yields a good approximation for estimating the variance provided that the first stage with replacement assumption is reasonable. This common approximation (UCVE) sometimes is denoted as *WR* (with replacement) in the sample survey literature, and *WR* is used with that meaning hereinafter.

15. Thus, when the sampling plan is described as *WR*, only three survey design variables are needed for variance estimation:

(a) The sample weight variable *WTVAR* (which is needed as well for point estimates);

(b) The stratification variable (or pseudo-stratification variable) *STRATVAR* used prior to first stage (PSU) sampling;

(c) The PSU (or pseudo-PSU) variable, denoted by *PSUVAR*.

16. Each sample respondent R must have a value for each one of these three variables in the basic data file. For example, a particular R may represent 8,714 elements in the population (*WTVAR* has the value 8,714) and may have been selected from stratum or pseudo-stratum #6 (*STRATVAR* has the value 6) and from PSU or pseudo-PSU #3 within stratum 6 (value of 3 for *PSUVAR*, within STRATAVR = 6).

17. *WR* is the default or only sampling plan description for most sample survey software packages or procedures. For example, *WR* is default, with Taylor series linearization, in SUDAAN, SAS, STATA, Epi-Info, PC-CARP and CENVAR. *WR* is default, with BRR and jackknife, in WesVar and SUDAAN. Note that single-stage sampling of elements, such as simple random sampling or stratified random sampling, is a special case of multistage sampling where the population PSUs on the sampling frame are the population elements and each sample PSU contains only one element (in other words, no clustering of sample elements). Software packages that have only the *WR* sampling plan description available may provide the option of incorporating *fpc* terms in variance estimation when single stage without replacement sampling of elements is used (for example, SAS, STATA, WesVar).

18. Using *WR* to approximate the actual complex sampling plan may overestimate variances slightly. However, survey data analysts generally are willing to accept some degree of overestimation for the relative simplicity of the *WR* approximation. Note, though, that the overestimation may be appreciable if there are several strata where first-stage sampling is without replacement and with large sampling fractions. In this situation, it may be desirable to use a software option that can incorporate the first stage *fpc* factors.

6. Variance estimation techniques and survey design variables

19. Public release sample survey data sets typically are already set up for variance estimation using one of the two major approaches, Taylor series linearization or replication techniques. Occasionally a public release data set will be set up to use both variance estimation approaches. The relevant sample design variables for variance estimation should be included in the public release data set, with corresponding documentation on how these variables are defined and how to use them.

20. If Taylor series linearization is used for the data set, look for three survey design variables in the documentation: the sample weight variable *WTVAR*, the first stage stratification variable *STRATVAR*, and the PSU variable *PSUVAR*. (Of course, the variables will not have the names used here.) If a replication method is used for the data set, look for the sample weight variable *WTVAR* and several replicate weight variables, often named something like REPL01--REPL52 (for 52 replicate weight variables). It is not necessary to know the *STRATVAR* or *PSUVAR* variables if replicate weight variables are available in the data set.

21. Surveyors who field their survey and prepare their own data set for analysis need to include relevant survey design variables and assign a value to these variables for each sample respondent element (R) in the data set. The minimum set of variables needed is: sample weight variable *WTVAR*, first-stage stratification (or pseudo-stratification) variable *STRATVAR*, and PSU (or pseudo-PSU) variable *PSUVAR* within stratum. These three survey design variables

approximate the actual sampling plan as *WR* and allow direct use of Taylor series linearization or allow personal or software calculation of replicate weights for replication techniques for variance estimation. If one wishes to incorporate *fpc* terms and/or additional stages of sampling or stratification into variance estimation, one needs additional survey design variables in the data set as well as sample survey software with these capabilities (for example, SUDAAN).

22. An unfortunately common situation is the acquisition of a sample survey data set that does not include any survey design variables or any replicate weight variables. Assuming that probability sampling was used, it is necessary to construct the survey design variables *WTVAR* for estimation, and *STRATVAR* and *PSUVAR* for variance estimation. Hopefully, enough details of the sampling plan can be obtained from written documentation or personal contact with the sampling personnel so that survey design variables can be constructed. If limited information is available, some crude approximations can be made. For example, if no selection probabilities can be reconstructed, it might be reasonable to assume an equal probability sample of elements and just use a post-stratification adjustment to obtain values for *WTVAR*. If PSUs cannot be exactly identified, proxy PSUs might be developed if certain geographical identifiers are known. Be aware in such cases of limitations of the data analysis if sample design variables are imprecise.

7. Analysis of complex sample survey data

23. There are many theoretical and practical issues involved in the analysis of complex sample survey data beyond conducting a weighted analysis and correctly estimating variances of estimators. These issues are well addressed and illustrated in the recent comprehensive book by Korn and Graubard (1999), including topics such as fitting models (for example, logistic regression) to sample survey data, goodness-of-fit for models, variance estimation for subpopulations, combining multiple surveys and forming pseudo-strata and pseudo-PSUs. See also other chapters in the present section of the present publication.

C. Variance estimation methods

1. Taylor series linearization for variance estimation

24. Assume a complex sampling plan with stratification of PSUs, multistage sampling, and unequal probability sampling of elements. The linear estimator $\Sigma\, w_i y_i$, a weighted sum, estimates the population total for the y variable, where w_i is the value of the sample weight variable *WTVAR* for sample element i, y_i is the value of the y variable for sample element i, and the summation Σ is over all elements in the sample, i=1, 2, …, m. If y is a dichotomous variable coded 1 for male diabetic and 0 otherwise, then the population total being estimated is the total number of male diabetics. The estimated variance of $\Sigma\, w_i y_i$ can be obtained directly under the *WR* assumption discussed above.

25. Now let x_i be a dichotomous variable coded 1 for male and 0 for female. Then the estimated prevalence of diabetes among males is given by $[\Sigma\ w_i y_i\ ] / [\Sigma\ w_i x_i\ ]$, a ratio of two linear estimators (or two weighted sums). Under the *WR* assumption, the estimated variance of this ratio estimator cannot be obtained directly. Even if simple random sampling has been used as opposed to complex sampling methods, estimating the variance of this non-linear function, a ratio, is not direct and requires some approximate method.

26. The algebraic expression for the non-linear estimator above can be expanded in an infinite Taylor series centred at the (estimated) expected value of the numerator and the (estimated) expected value of the denominator. The non-linear estimator then is approximated algebraically by retaining only the leading terms in the infinite Taylor series, resulting in an algebraic expression that now is a linear (no longer non-linear) function of sample data; that is to say, the non-linear ratio estimator has been "linearized". Now the estimated variance of the linearized function (including relevant covariance terms) can be obtained directly under the *WR* assumption, just as the estimated variance of $\Sigma\ w_i y_i$ was obtained. In this process, the variance of the linearized function is estimated within each stratum separately (since sampling is independent across strata) and then the stratum specific estimated variances are summed to obtain the variance of the estimator.

27. When the Taylor series linearization approach is used, a unique approximate variance estimation formula needs to be derived and programmed not only for every different non-linear estimator, but also for each possible sampling plan where that estimator might be used (*WR* being one such sampling plan). This characteristic is viewed as a disadvantage of the Taylor series linearization approach to variance estimation. In fact, a given software package that analyses sample survey data with Taylor series linearization may not include the combination of the specific estimator that one wishes to use with the actual or approximate sampling plan that one has used.

28. All software programs using Taylor series linearization require the specification of the design variables *WTVAR*, *STRATVAR* and *PSUVAR*, as needed for the *WR* sampling plan approximation. Additional sampling plans may be available with Taylor series linearization, depending upon the software package; their use may require additional design variables.

2. Replication method for variance estimation

29. The replication method for variance estimation of sample survey estimators, although known theoretically for quite some time, has experienced increased utilization with the advent of high-speed computing capability. The replication method is computer-intensive but more flexible than the Taylor series linearization method in terms of the number of different estimators for which estimated variances can be computed.

30. The general idea of replication methods is as follows. First, the entire or full sample is used, as in the Taylor series method, to obtain a point estimate of the population parameter of interest; that is to say, the estimator formula for the population parameter is applied to the full sample. Only the sampling weight variable *WTVAR* is needed for this calculation.

31. Second, in order to estimate the variance of this estimator, many different subsamples or "replicates" are formed from the full sample in such a manner that each replicate reflects the sampling plan and weighting procedures and adjustments of the full sample. Each replicate is defined by the value of a replicate weight variable. For example, *REPWTj* is the replicate weight variable for replicate #*j*, where $j = 1, 2, 3, \ldots, G$ (total number of replicates). An observation in the full sample has a value of zero for *REPWTj* if that observation is not included in replicate #*j* and a positive value if it is included in replicate #*j*. The sum of the values of *REPWTj* over the observations in the full sample is an estimate of the number of elements in the population.

32. Third, the estimator formula is applied to each replicate to obtain a point estimate of the population parameter of interest (the replicate estimate), yielding *G* replicate estimates of the same population parameter.

33. Fourth, based on the variability of the *G* replicate estimates, an estimated variance of the full sample estimator is computed.

34. Replicates can be formed in different ways, resulting in various replication techniques. Two major approaches to forming replicates, each with variations, are balanced repeated replication (BRR) and jackknife (both discussed below). Public release sample survey data sets that are set up for variance estimation with a particular replication method typically include the replicate weight variables with the data set. In this case, the secondary data analyst must use variance estimation software that includes the specific replication technique for which the replicate weights in the data set were generated.

35. However, one may wish to use a replication technique for variance estimation when the replicate weights are not already in the data set. Some software packages that implement replication variance estimation approaches also compute the replicate weights. The minimum survey design variables needed for a software package to form replicate weights are: sample weight variable *WTVAR*, stratification variable *STRATVAR*, and PSU variable *PSUVAR* within stratum. If the full sample has been adjusted for non-response and/or has been post-stratified, then this information may also be accepted as input by the software package in the calculation of replicate weights (for example, WesVar). One can always calculate replicate weights oneself (without a software package), but this strategy is recommended only for those who are knowledgeable about the details of replication techniques.

3. Balanced repeated replication (BRR)

36. Balanced repeated replication (BRR) is a specific replication technique that can be used for very general designs, namely, stratified multistage sampling. However, it was developed for the specific situation with exactly two PSUs selected (sampled) per stratum, generally sampled with unequal probability with or without replacement. It also is generally used with the *WR* approximation to the complex sampling plan (the UCVE approach).

37. With BRR, each replicate contains exactly half of the sample PSUs, one PSU from each stratum; frequently each replicate is called a "half-sample". The total number of possible different replicates is 2^L, where L is the number of strata. However, it is not necessary to use all 2^L replicates, which might require inordinate computing time. Rather, a smaller and "balanced" set of replicates can yield the same variance estimate that would be obtained from all possible replicates. G "balanced" replicates are formed, using a Hadamard matrix (Wolter, 1985), so that each sample PSU appears in the same number of replicates and each pair of sample PSUs from two different strata appears in the same number of replicates. The minimum number G of replicates required is the smallest integer that is greater than or equal to L but divisible by 4. For example, 49 strata, each with two sampled PSUs, would require 52 BRR replicates. Observations in sample PSUs that are not included in replicate j have a value of zero for the replicate weight variable *REPWTj*, and observations in sample PSUs that are included in replicate j have a value that is twice their sampling weight in the full sample, although this may be adjusted for non-response and/or post-stratification.

38. A common variation on the BRR technique defined above was developed by Fay (Judkins, 1990) because standard BRR can be problematic if estimation is desired for a small domain or for a population ratio when the denominator has few cases in the full sample. In Fay's method, observations in the sample PSUs that are not chosen for replicate j are not zeroed out, as they are in standard BRR. Rather, their sampling weight is diminished by a multiplicative factor K ($0 \leq K < 1$), whereas the observations in the sample PSUs chosen for the replicate have their sampling weight enhanced by the multiplicative factor $(2 - K)$. Setting $K = 0$ yields the standard BRR technique. A commonly recommended value is $K = 0.3$ for Fay's method.

4. Jackknife replication techniques (JK)

39. The general idea of jackknife techniques is to delete one sample PSU at a time to form replicates and then reweight each replicate as necessary so that it makes inference to the population represented by the full sample. A sample PSU could comprise a single element, as in the case of simple random sampling or stratified random sampling, or a sample PSU could contain several elements as in the approximate sampling plan *WR*.

40. Consider first the case where no stratification is used prior to PSU sampling and each of G sample PSUs (with approximately the same number of elements) resembles the full sample. A total of G replicates are formed by deleting one sample PSU at a time. For replicate j with the replicate weight variable *REPWTj*, observations in the deleted sample PSU #j have a value of zero for *REPWTj*. Each observation in the remaining (non-deleted) sample PSUs have a value for *REPWTj* that equals the sampling weight for that observation multiplied by the factor $[\, G / (\, G - 1) \,]$.

41. A second example is L strata with exactly two PSUs selected per stratum; this is to say, the design discussed above for BRR. Deleting one sample PSU at a time would result in $2L$ replicates. For each of the $2L$ replicates the remaining sample PSU in the stratum with the deleted sample PSU would have the sampling weight for each observation multiplied by 2 (and the deleted sample PSU would have the sampling weight for its observations multiplied by zero). However, this technique usually is implemented with only L replicates rather than $2L$ replicates,

where only one sample PSU, chosen at random, is deleted within each of the L strata. For linear estimators, the variance estimator using only the L replicates is algebraically equivalent to the variance estimator using the $2L$ replicates.

42. The most general sampling plan is stratified multistage sampling with L strata (prior to PSU sampling) and two or more PSUs sampled per stratum. Each sample PSU is deleted to form a replicate; the number of replicates G is equal to the total number of sample PSUs in the full sample (n). Within stratum h, the value for the replicate weight variable *REPWTj* for each observation in the deleted sample PSU is the sample weight variable *WTVAR* multiplied by zero. The value of the variable *REPWTj* for each observation remaining in stratum h from which the sample PSU was deleted is the sample weight variable *WTVAR* multiplied up by the factor [n_h / ($n_h - 1$)], where n_h is the number of sample PSUs within stratum h in the full sample.

5. Some common errors made by users of variance estimation software

43. Several software packages require the user to sort the input data set by some of the survey design variables, for example, by *STRATVAR* and by *PSUVAR* within *STRATVAR* (as explained in para. 35). Forgetting to sort may yield incorrectly estimated variances, although most software programs will emit an error message if the data set is not sorted correctly.

44. Users of public release data sets may specify incorrect survey design variables because of an inadequate review of the sample survey documentation. An incorrectly specified sample weight variable will result in biased estimators and incorrectly estimated variances; that is to say, all analyses will be wrong. If the sample weight variable is correct but the stratification and/or PSU variable is incorrect, point estimates will be correct but estimated variances will be incorrect.

45. Some public release data sets have multiple data files with different survey design variables for different files. Different data files may have varying units of analysis, for example, person, household or family, so careful attention is needed to interpretation of output. Some survey variables may be measured on only a probability subsample of the full sample, requiring a different sample weight variable than variables measured on the entire sample. Careful and thorough reading of the documentation is essential for all sample surveys, whether the sampling plan is simple or inordinately complex.

D. Comparison of software packages for variance estimation

46. Web links to a full array of software packages for sample survey data packages, including the eight reviewed in this article, can be found at the informative web site www.fas.harvard.edu/~stats/survey-soft/survey-soft.html. See also Carlson (1998) for a review of software packages for complex sample survey data. Note that SPSS is not included among the software packages reviewed. As of early 2003, SPSS had had no capability for complex sample survey variance estimation but it did release an add-on module in late 2003 when this chapter was in press.

47. The remainder of this chapter reviews and compares eight software packages for variance estimation with complex sample survey data: SAS, SUDAAN, STATA, Epi-Info, WesVar, PC-CARP, CENVAR and IVEware. The first five of the eight packages are illustrated with descriptive analyses using data from a sample survey conducted in Burundi in 1989; population proportions, means and totals are estimated and domains are compared on these parameters. Results from the Burundi analyses are summarized in the chapter in table XXI.1, and detailed tables and annotated example programs and output for each package are given in the annex on the CD-ROM. The annotated examples in the annex can help users learn how to use the first five variance-estimation software packages.

Table XXI.1. Comparison of PROCS in five software packages: estimated percentage and number of women who are seropositive, with estimated standard error, women with recent birth, Burundi, 1988-1989

Software package and PROC	% Seropos	s.e. of % Seropos	95% CI % Seropos	Number Seropos	s.e. # Seropos	95% CI # Seropos
SAS 8.2 MEANS [a/] No weight	74.88% wrong	2.12% wrong	N-APP	N-APP	N-APP	N-APP
SAS 8.2 MEANS [b/] With weight	67.20%	2.30% wrong	N-APP	N-APP	N-APP	N-APP
SAS 8.2 SURVEYMEANS	67.20%	3.83%	59.38%, 75.02%	142,485	8848.10	124415, 160556
SUDAAN 8.0 CROSSTAB and DESCRIPT Taylor and BRR	67.20%	3.83%	N-AV	142,485	8848.10	N-AV
STATA 7.0 Svymean	67.20%	3.83%	58.38%, 75.02%	N-AV	N-AV	N-AV
STATA 7.0 Svytotal	N-AV	N-AV	N-AV	142,485	8848.10	124415, 160556
Epi-Info 6.04d CSAMPLE [c/]	67.20%	3.83%	59.70%, 74.71% [c/]	N-AV	N-AV	N-AV
WesVar 4.2	67.20%	3.83%	59.38%, 75.02%	142,485	8848.10	124415, 160556

Note: Abbreviations used: CI = Confidence interval, N-APP = not applicable, N-AV = not available, s.e. = standard error.

[a/] Incorrectly specified analysis; ignores sampling weight, clustering and stratification.

[b/] Incorrectly specified analysis; sampling weight incorporated but not clustering and stratification.

[c/] Confidence interval given by Epi-Info 6.04d is narrower than that of other software packages. Epi-Info 6.04d used z=t=1.96 to construct the 95 per cent confidence interval, whereas the other software packages used t = 2.042 from the Student t-distribution with 30 ddf (denominator degrees of freedom for the sample survey, calculated as number of PSUs minus number of pseudo-strata). Using the actual survey ddf is preferred.

48. Among the five packages illustrated with the Burundi survey data, three (STATA, SAS and Epi-Info) include sample survey procedures within a general statistical software package. All three use Taylor series linearization for variance estimation. The remaining two illustrated packages (WesVar and SUDAAN) were developed especially for sample survey variance estimation. WesVar uses replication methods and SUDAAN offers both Taylor series linearization and replication methods.

49. Three additional software packages (PC-CARP, CENVAR and IVEware) are reviewed but not illustrated with the Burundi survey data. PC-CARP and CENVAR both use Taylor series linearization for variance estimation. IVEware uses both Taylor series linearization and replication methods.

50. The eight packages reviewed here include many, but not all, of the possible options for sample survey variance estimation. Three (Epi-Info, CENVAR and WesVar 2) were chosen because they offer basic descriptive analyses and can be downloaded from the Web at no cost, an appealing feature for analysts with a limited or no budget for software purchases. Two (PC-CARP and WesVar 4) were chosen because, although not free, they are low in cost compared with to other options and offer descriptive analyses as well as design-based linear and logistic regression. Two moderately priced packages (SUDAAN and STATA) were chosen because they offer, along with descriptive analyses, comprehensive choices for design-based regression models. Although expensive, SAS was chosen because of its dominance in the data management and analysis arena and its relatively new PROCS for sample survey data analysis. Finally, the recently released IVEware (beta version) was chosen because it offers comprehensive descriptive analyses and design-based regression models, along with multiple imputation procedures. IVEware is free (downloadable from the Web) but runs as a SAS callable software application (thus requiring SAS).

51. Table XXI.2 summarizes all eight software packages on a wide variety of characteristics, including sampling plans covered, methods of variance estimation, and types of analyses.

Table XXI.2. Attributes of eight software packages with variance estimation capability for complex sample survey data

ATTRIBUTE	SAS 8.2	SUDAAN 8.0	STATA 8.0	Epi-Info 6.04d	WesVar 4.2	PC-CARP	CENVAR	IVEware
Taylor series	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes Desc
Replication methods BRR and JK	No	BRR JK	No	No	BRR JK	No	No	JK Models
Replicate weights formed	No	No-BRR Yes-JK	No	No	Yes BRR/JK	No	No	Yes JK
Input data set	SAS	SAS, SPSS, ASCII	STATA	Epi-Info	SAS, SPSS, STATA, ASCII, ODBC	ASCII	ASCII	SAS
Estimate total	Yes	Yes	Yes	No	Yes	Yes	Yes	No
CI on total	Yes	No	Yes	No	Yes	Yes	Yes	No
LC on totals	No	Yes	Yes	No	Yes	Yes	Yes	No
Estimate mean	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CI on mean	Yes	No	Yes	Yes-narrow	Yes	Yes	Yes	Yes
LC on means	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Estimate proportions	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CI on proportion	Yes	No	Yes	Yes-narrow	Yes	Yes	Yes	Yes
LC on proportions	No	Yes	Yes	Yes-error	Yes	Yes	Yes	Yes
Estimate ratio	Yes	Yes	Yes	No	Yes	Yes	Yes	No
CI on ratio	Yes	No	Yes	No	Yes	Yes	Yes	No
LC on ratios	No	Yes	Yes	No	Yes	Yes	Yes	No
Domain analyses	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Compare domains	No-8.2 Yes-9.0	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subpopulation analyses	No	Yes	Yes	No	Yes	Yes	Yes	Yes
Standardized rates/means	No-8.2 Yes-9.0	Yes	Yes	No	Yes	No	No	No
Chi-square tests	No-8.2 Yes-9.0	Yes	Yes	No	Yes	Yes	No	No
Logistic regression	No	Yes	Yes	No	Yes	Yes	No	Yes
Odds ratio	No	Yes	Yes	Yes	Yes	Yes	No	Yes
Risk ratio	No	Yes	Yes	Yes	Yes	No	No	No
Linear regression	Yes	Yes	Yes	No	Yes	Yes	No	Yes

ATTRIBUTE	SAS 8.2	SUDAAN 8.0	STATA 8.0	Epi-Info 6.04d	WesVar 4.2	PC-CARP	CENVAR	IVEware
Additional regression models	No	Yes	Yes	No	No	No	No	Yes
Describes many sample stages	No	Yes	No	No	No	No	No	No
Design effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Free trial software	No	No	No	NA Free	Yes	No	NA Free	NA Free
General statistical package	Yes	No	Yes	Yes	No	No	No	No
Manage data capability	Yes	No	Yes	Yes	Yes	No	No	No
Run via input programs	Yes	Yes	Yes	No-6.04d Yes-2002	No	No	No	Yes
Run via short commands	No	No	Yes	No	No	No	No	No
Run via menu selection	No	No	No	Yes	Yes	Yes	Yes	No
Sort data set by stratum and PSU	No	Yes	No	Yes	No	Yes	Yes	No
Training offered by developer	Yes	Yes	Yes	No	Yes	No	No	No
Written/online manual	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tutorials for survey procedures	No	No	No	No	Yes	No	Yes	No
Cost	High	Medium	Medium	Free	Low V4 Free V2	Low	Free	Free
Annual renewal fee	High	Medium	None	None	None	None	None	None
Impute data	No	No	No	No	No	Yes	No	Yes

Abbreviations used: ASCII = American Standard Code for Information Interchange, BRR = balanced repeated replication, CI = confidence interval, JK = jackknife, LC = linear contrast, NA = not available, ODBC = Open DataBase Connectivity, V = version.

E. The Burundi sample survey data set

52. All numerical examples in this chapter use a data set from a tetanus toxoid (TT) immunization coverage sample survey conducted in Burundi in 1989. A brief summary of the Burundi sample survey design follows; more detail is provided in section I of the annex on the CD-ROM. For additional information on this survey's methodology and its published results, see the report by the Expanded Programme on Immunization (EPI) (1996) of the World Health Organization (WHO).

1. Inference population and population parameters

53. The population of inference for this survey is women of Burundi who gave birth between Easter of 1988 and February/March of 1989. The population parameter of interest is percentage (or proportion) of women who were seropositive for tetanus antitoxin, thus protecting their newborn against neonatal tetanus.

2. Sampling plan and data collection

54. The sampling plan was a modification suggested by Brogan and others (1994) of the cluster sample survey methodology developed at the WHO for its Expanded Programme on Immunization. The modification yields a probability sample of dwellings or housing units and hence a probability sample of women, which the standard WHO EPI cluster sampling methodology may not do (ibid.).

55. Burundi was stratified into two geographical areas, the capital Bujumbura (urban stratum) and the rest of the country (rural stratum). Primary sampling units (PSUs) were geographical areas, *collines* within the rural stratum and *quartiers* or *avenues* within the urban stratum. The PSU sampling frame for each stratum was ordered by geographical proximity. Systematic ppes (probability proportional to estimated size) sampling was used to select 30 sample PSUs per stratum. Since 96 per cent of the inference population resides in rural Burundi, and since the same number of sample PSUs was allocated to each stratum, urban women were substantially oversampled. The specific ordering of the PSUs on the sampling frame, combined with systematic ppes sampling of PSUs, yields implicit geographical stratification within each stratum.

56. Further stages of probability sampling within sample PSUs were conducted to obtain a sample of occupied dwellings. All survey-eligible women within a sampled dwelling were selected for the sample. Seropositivity of tetanus antitoxin titre was determined from a finger prick blood sample. The survey response rate was essentially 100 per cent, an unusually high response rate. A total of 206 urban and 212 rural women were interviewed.

3. Weighting procedures and set-up for variance estimation

57. The sample weight variable *W* provided in the Burundi data set was revised to *W2* so that the value of *W2* for a sample respondent *R* is an estimate of the number of women in the inference population represented by that *R*. The value of *W2* is approximate and used only to

illustrate the estimation of population totals with the various software packages. Substantive conclusions regarding population totals for survey-eligible women in Burundi in 1989 should not be drawn from the analyses in this chapter**.** It is important to note that estimated proportions and means reported in this chapter agree with previously published results with this data set (Expanded Programme on Immunization, 1996) since the revised *W2* is a scalar multiple of *W* that was used for previous analyses. The value of *W2* was 959.3 for rural sample women and 42.0 for urban sample women, reflecting the substantial oversampling of urban women. The Burundi sampling plan was approximated by the common description *WR* for the purpose of variance estimation, that is to say, the UCVE approach with low first-stage sampling fractions.

58. Since PSUs were implicitly stratified, the sampling plan within each of the urban and rural strata was regarded as two PSUs sampled from each of 15 pseudo-strata. Describing the sampling plan as a total of 30 pseudo-strata, each with two sample PSUs, is preferred over describing it as 2 strata, each with 30 sample PSUs, because the former yields less biased variance estimation, since it takes the implicit stratification into account. The pseudo-stratification variable *PSTRA* was coded 1 through 30 and the pseudo-PSU variable *PPSU* was coded 1 or 2 within each pseudo-stratum.

59. When Taylor series linearization is used for variance estimation, only the variables *W2*, *PSTRA* and *PPSU* are needed. When replication techniques are used, however, replicate weights are required. WesVar was used to calculate BRR replicate weights from the variables *W2*, *PSTRA* and *PPSU*. These replicate weights calculated by WesVar were used both in WesVar and in SUDAAN for variance estimation using BRR.

4. Three examples for survey data analyses

60. The annex contains annotated data analyses for the three examples below, using five software packages for sample survey data (sects. II-VI). The examples below illustrate common descriptive and analytical analyses performed on sample survey data, namely, (a) estimation of proportions, totals and means for the entire population and for domains or strata; and (b) comparison of domains or strata on means or proportions. The inference population is survey-eligible women in Burundi in early 1989.

> Example 1: Estimate number of women (population total) and percentage of women (population proportion/percentage) who were seropositive (*IMMUNE* variable, 1= seropositive, 2 = seronegative). The variable *BLOOD* (1 = seropositive, 0 = seronegative) is a recode of *IMMUNE*.
>
> Example 2: Estimate the population parameters of example 1 among urban and rural women (*RUR_URB* variable, coded 1 = rural, 2 = urban). Determine whether rural/urban residence is statistically independent of seropositivity (*IMMUNE*).
>
> Example 3: Estimate mean international units of antitoxin per millitre (ml) (*IUML*), for the inference population of women and by rural/urban residence. Determine whether rural/urban residence is related to mean *IUML*.

Note: estimation of mean international units of antitoxin per millitre (ml) (*IUML*) may give misleading results because of the right skewed distribution of this variable. It might be better to use the median or to transform *IUML* before analysis, for example, to the natural logarithim of *IUML*. In this chapter, mean *IUML* (without transformation) is estimated to show the capabilities of the five software packages, not to illustrate substantive results concerning *IUML*.

F. Using non-sample survey procedures to analyse sample survey data

61. The present section illustrates that incorrect use of simple random sample formulae for analysis of complex sample survey data can result in biased point estimates and biased (usually too small) estimated standard errors. See Brogan (1998 and in press) for another illustration. Any statistical software package could have been used for this illustration, and answers comparable with those obtained with SAS (used in this section) would have been obtained.

62. The population parameter to be estimated is the proportion of women in the inference population who are seropositive. The indicator variable *BLOOD* is calculated and coded as 1 = seropositive and 0 = seronegative. Thus, the mean of *BLOOD* is the proportion of women who are seropositive. PROC MEANS in SAS estimates the mean of *BLOOD* as 0.74880, with estimated standard error of 0.02124 (row 1 of table XXI.1). These two calculations are biased because the incorrectly applied PROC MEANS ignores the sampling weight variable for estimating the population proportion and, in addition, ignores the sampling weight, PSU and stratification variables for calculating the estimated standard error of the point estimate.

63. PROC MEANS in SAS, then, is used with the sampling weight variable W2 on a WEIGHT statement. The mean of *BLOOD* is estimated as 0.67203, with estimated standard error (weighted) of 0.02299 (row 2 of table XXI.1). In this analysis PROC MEANS obtains an appropriate point estimate for the population proportion. However, the incorrectly applied PROC MEANS yields a biased estimated standard error because it ignores the PSU and stratification variables.

64. Finally PROC SURVEYMEANS in SAS is used to provide an appropriate analysis for complex sample survey data. (Details on how to use SURVEYMEANS appear in the next section.) The point estimate of the population proportion is 0.67203, with estimated standard error of 0.03830 (row 3 of table XXI.1). PROC SURVEYMEANS takes into account the sampling weight variable for calculating the point estimate and the sampling weight, PSU and stratification variables for calculating the estimated standard error.

65. A comparison of these three analyses in table XXI.1 shows that the unweighted (incorrect) point estimate of 0.7488 (74.88 per cent) differs quite a bit from the weighted (correct) point estimate of 0.6720 (67.20 per cent). The unweighted point estimate is too high because a higher proportion of urban women, compared with rural women, are seropositive (illustrated later in the chapter), and urban women are overrepresented in the sample since they were oversampled: they constitute about half of the sample but only 4 per cent of the inference population. Thus, in an unweighted analysis for making inference to the country, urban women

are given much more influence than they should and bias upward the estimated population proportion.

66. A comparison of the two analyses that yield the correct weighted point estimate illustrates that, even with a WEIGHT statement, the incorrectly applied PROC MEANS in SAS seriously underestimates the standard error, an incorrect calculation of 0.02299 (2.30 per cent) compared with the correct PROC SURVEYMEANS calculation of 0.03830 (3.83 per cent). This occurs primarily because PROC MEANS, with or without a WEIGHT statement, ignores the clustering of women within sample PSUs, whereas SURVEYMEANS recognizes the clustering for variance estimation. Since the intra-class correlation coefficient is positive for most measured variables in complex sample surveys, correct variance estimation procedures that take into account the clustering usually yield larger estimated standard errors.

67. In general, biased point estimates of population parameters are obtained if sample survey data are not analysed with the appropriate sample weight variable. Further, even if the sample weight variable is incorporated into the analysis, yielding appropriate point estimates of population parameters, the standard errors typically are underestimated when sample elements are clustered in survey data and the clustering is not recognized in variance estimation. Underestimation of standard errors results in confidence intervals that are too narrow and statistical tests of significance with p-values that are too small, in other words, the level of statistical significance is overstated.

68. The magnitude of underestimation of variance by ignoring clustering of sample survey data is approximated by the expression $[1 + \rho (b - 1)]$ where ρ is the intra-class correlation coefficient between population elements and b is the average number of sample elements per sample cluster (PSU) (see chap. VI). For example, if the value of the expression is 2, then taking the clustering into account approximately doubles the estimated variance that one would obtain by ignoring the clustering. Note that the Burundi PSU variable named *PPSU* identifies for the software what sample elements are clustered together within the same sample PSU (for a given stratum).

69. In addition to the impact of clustering on estimated variance, substantial variation in the value of the sampling weight variable across respondents increases estimated variance. Thus, if the sampling weight variable is ignored in the analysis, the estimated standard error is underestimated (and the estimator of the population parameter is biased).

G. Sample survey procedures in SAS 8.2

1. Overview of SURVEYMEANS and SURVEYREG

70. Version 8.2 in SAS contains two recently developed procedures (they first appeared in V 8.0) for analysis of sample survey data: SURVEYMEANS and SURVEYREG. SAS includes the common sampling plan description *WR* for which the basic three survey design variables are required. Finite population correction terms can be applied for single-stage sampling designs such as stratified random sampling and simple random sampling. Taylor series linearization is used for variance estimation. SAS V9 contains two new PROCS for complex sample survey data, SURVEYFREQ for analysis of categorical variables and SURVEYLOGISTIC for logistic regression. Additional SAS procedures for sample survey data are under development.

71. The syntax for specifying the relevant survey design variables for *WR* is the same for both SURVEYMEANS and SURVEYREG. The keyword STRATA is used to designate the stratification variable, the keyword CLUSTER is used to designate the PSU variable, and the keyword WEIGHT is used to specify the sampling weight variable (as in other SAS procedures such as MEANS). These statements, appropriate for a given survey, must be in each SAS sample survey procedure and generally will not change as long as the same sample survey data set is being analysed. For the Burundi data set, the SAS statements below describe the sample survey design for SAS PROC SURVEYMEANS or PROC SURVEYREG:

STRATA *PSTRA*
CLUSTER *PPSU*
WEIGHT *W2*

72. If the STRATA statement is missing, SAS assumes the sampling plan had no stratification of PSUs prior to first-stage sampling. If the CLUSTER statement is missing, SAS assumes that the sample elements are not clustered, i.e., that each sample cluster contains exactly one element, i.e., that elements were sampled at the first (and only) stage of sampling, i.e., that simple random or stratified random sampling was used. If the WEIGHT statement is missing, SAS assumes that each *R* has the same value for the weighting variable and SAS assigns the value 1.0 to the weighting variable. If all three survey design statements (STRATA, CLUSTER, WEIGHT) are missing, this is equivalent to specifying simple random sampling from an infinite population, the assumption for most of the non-survey PROCS in SAS.

2. SURVEYMEANS

73. This procedure estimates population means and totals for continuous variables and population proportions and totals for categorical variables, using sample survey data. Estimated standard errors and coefficients of variation are provided for all point estimates, as well as confidence intervals for population parameters. Specific statistics can be requested on the PROC statement, or one can take the default printout for statistics, or one can use ALL on the PROC statement to obtain all statistics that can be calculated by SURVEYMEANS.

74. Variables to be analysed (both continuous and categorical) appear on the VAR statement. The CLASS statement lists the variables on the VAR statement that are categorical; SAS then assumes that all other variables on the VAR statement are continuous.

75. The DOMAIN statement with one or more categorical variables is used to specify domains for analysis of all variables on the VAR statement. SAS automatically provides analyses for the marginal, in other words, the entire population, in addition to the domain analyses. A program without a DOMAIN statement provides estimates for the entire population only. Although the BY statement in SURVEYMEANS can be used to obtain estimates for domains, this is not recommended for sample survey data because the appropriate formulae for variance estimation are not used when the BY statement is used. Use the DOMAIN statement for analysis of domains.

76. SAS V8.2 does not have a statement that allows a subpopulation to be analysed, for example, only older women. However, subpopulation analyses can be conducted by first defining an indicator variable, for example, *OLDERFEM*, which indicates whether the sample element belongs to the subpopulation. Then, the statement DOMAIN *OLDERFEM* can be used to obtain the desired analyses; ignore the SAS output for the sample elements who are not older women. Do not use the SAS IF statement to subset the data set to women who are older before going into PROC SURVEYMEANS, since the standard errors may be calculated incorrectly inasmuch as SURVEYMEANS may not know the full number of strata and sample PSUs in the sample survey.

3. SURVEYREG

77. This procedure performs linear regression for sample survey data according to the design-based approach (Korn and Graubard, 1999), that is to say, the analysis takes into account the survey design variables. As with linear regression for non-survey data, the dependent variable is continuous (or assumed to be so), and the independent variables can be a mixture of continuous and categorical variables. The MODEL statement includes the dependent variable and all independent variables. Any categorical variable on the MODEL statement must also appear on the CLASS statement, and the CLASS statement must precede the MODEL statement in the SAS program. SURVEYREG forms dummy indicator variables (coded 1 or 0) for categorical independent variables, with the highest coded value of the variable defined as the reference group. Other options in SURVEYREG, as well as its output, are similar to the (non-survey) linear regression in SAS.

78. SAS Version 8.2 has no sample survey procedures to compare domains on means or proportions, although these capabilities are under development. An example question for this situation is, Do rural and urban women in the Burundi inference population differ on mean *IUML* units or on proportion who are seropositive? SURVEYFREQ in V9.0 can be used to conduct a chi-square test on the two variables residence (rural/urban) and seropositivity (yes/no). Until domain comparison procedures are full developed in SAS for sample survey data, SURVEYREG can be used as follows to compare domains.

79. If it is desired to compare rural and urban women in the inference population on mean *IUML*, use the MODEL statement in SURVEYREG with the continuous variable *IUML* as the dependent variable and the domain variable designating rural/urban as the independent categorical variable. Part of the standard output from SURVEYREG is a test of the null hypothesis that the population regression coefficient for rural/urban (with one degree of freedom) is equal to zero. This null hypothesis regarding the regression coefficient is equivalent to the null hypothesis that rural and urban women in the inference population have the same mean *IUML*.

80. If it is desired to compare urban and rural women in the inference population on proportion who are seropositive (a dichotomous variable), use the indicator variable *BLOOD* (1= seropositive, 0=seronegative) as the dependent variable. (Note that *BLOOD* is simply a recode of the *IMMUNE* variable where 1=seropositive and 2=seronegative.) On the MODEL statement in SURVEYREG, define *BLOOD* as the dependent variable and the domain variable designating rural/urban as the independent categorical variable. The null hypothesis that the regression coefficient is zero is equivalent to the null hypothesis that the proportion seropositive is the same for rural and urban women in the population of inference.

4. Numerical examples

81. Section II of the annex on the CD-ROM illustrates the use of SURVEYMEANS and SURVEYREG to work the three examples listed in paragraph 60. Review of the annotated SAS programs (user-written) and annotated SAS output should prepare readers to write their own SAS programs for SURVEYMEANS and SURVEYREG and interpret the output.

82. Table XXI.1, row 3, summarizes the SURVEYMEANS output in section II of the annex for estimating the percentage and number of women in the Burundi inference population who are seropositive, with estimated standard error and confidence interval; most of these results were discussed in section F of this chapter. Table XXI.3, row 1 (in annex, sect. VII, on the CD-ROM), summarizes the SURVEYMEANS output for estimating the percentage seropositive for each of the two domains of rural and urban women, 66.51 per cent and 83.50 per cent, respectively. Table XXI.4, row 1 (in annex, sect. VII, on the CD-ROM), summarizes the SURVEYREG output that compares rural and urban women, yielding a t-value of -3.52, with a p-value of 0.0014 for testing the null hypothesis that rural and urban women do not differ on the percentage who are seropositive. Thus, rural and urban women in the inference population differ on percentage who are seropositive: urban women have a higher seropositivity prevalence rate.

5. Advantages/disadvantages/cost

83. If one already is a SAS/STAT user, then the sample survey procedures in SAS are available at no additional cost and use familiar syntax. Further, the full capabilities of SAS for data management and new variable formation are also available. Technical support and documentation for the sample survey procedures are subsumed under the regular system of SAS support. Compared with that of other sample survey packages reviewed, the cost of SAS is high.

84. SAS 8.2 has no capability to compare domains to each other, although SURVEYREG can be used as a temporary solution for this type of analysis. The addition of SURVEYFREQ in V9.0 provides domain comparisons on categorical variables.

85. SAS uses only Taylor series linearization for variance estimation. For stratified multistage cluster sampling, it handles only the common sampling plan description *WR*. However, it can incorporate *fpc* terms into single-stage stratified random sampling or simple random sampling.

86. The capability of SAS 8.2 for sample survey data analysis is basic and descriptive and may fit the analysis needs of many users. The addition of SURVEYFREQ in V9.0 provides descriptive and analytical capability for categorical variables. Sample survey procedures still under development, for example, logistic regression, should make SAS more comparable in the future with other software packages that offer comprehensive sample survey analyses.

H. SUDAAN 8.0

1. Overview of SUDAAN

87. SUDAAN (Research Triangle Institute, 2001) is a specialty software package originally developed for the analysis of complex sample survey data, but now generalized for the analysis of correlated data using techniques such as longitudinal data analysis and generalized estimating equations (GEE). SUDAAN is an acronym for SUrvey DAta ANalysis. The procedures for descriptive and analytical statistics are DESCRIPT, CROSSTAB, and RATIO. Design-based modelling procedures include linear regression, logistic regression (including multinomial), log-linear regression and survival analysis.

88. SUDAAN 8.0 is programmed in C language, with user-provided command statements similar to those of SAS. Input data sets can be either SAS, SPSS or ASCII files. SUDAAN is available to run by itself (standalone SUDAAN) or in conjunction with SAS (SAS-callable SUDAAN). SAS users generally would prefer SAS-callable SUDAAN.

89. SUDAAN is the only sample survey package to include both of the two most common approaches to variance estimation: Taylor series linearization and replication methods. The latter approach in SUDAAN includes balanced repeated replication (BRR), with or without the Fay adjustment factor, and jackknife methods. All replication methods in SUDAAN assume the common sampling plan description referred to previously as *WR*. If BRR is used for variance estimation, the BRR replicate weights must be provided with the input data set; SUDAAN does not generate BRR replicate weights. SUDAAN will generate replicate weights for the jackknife delete one (PSU) method or will accept jackknife replicate weights provided with the input data set for the jackknife delete one method and variations on this method.

90. The sample survey design is described to SUDAAN in three statements: (a) by choosing an option for the DESIGN keyword on the PROC statement; (b) by specifying the stratification and clustering variables on the NEST statement; and (c) by specifying the sample weight variable on the WEIGHT statement. The input data set to SUDAAN must be sorted by all of the

variables that appear on the NEST statement, generally the first-stage stratification variable and then the PSU variable within each stratum.

91. Unlike most other software packages with sample survey capability, second and subsequent stages of sampling and stratification in multistage sampling can be described to SUDAAN for variance estimation, alleviating the necessity of always using the common sampling plan description *WR*. In addition, SUDAAN has extensive capability for incorporating into variance estimation the finite population correction (*fpc*) terms at multiple stages of without replacement sampling. The SUDAAN manual, available in print or a pdf file, gives several examples of how to describe sampling plans to SUDAAN (see chap. III).

92. The default sampling plan for SUDAAN is *WR* as defined above, whether for Taylor series linearization, BRR or jackknife. Using the SUDAAN syntax DESIGN = *WR* on the PROC statement invokes not only the UCVE approach and first-stage sampling with replacement or without replacement but with small first-stage sampling fractions, but also the use of Taylor series linearization. With DESIGN = *WR*, the NEST statement contains one or more justification variables (usually just one) and one PSU variable. If the option DESIGN = is missing from the PROC statement, SUDAAN assumes DESIGN = *WR*.

93. The SUDAAN syntax DESIGN = BRR invokes the common sampling plan description *WR* (as discussed previously) with balanced repeated replication for variance estimation. The BRR replicate weight variables must be in the input data set, and the REPWGT statement in the SUDAAN program gives the variable names for the replicate weight variables.

94. The SUDAAN syntax DESIGN = JACKKNIFE, in the absence of JACKWGTS and JACKMULT statements, invokes the common sampling plan description *WR* with variance estimation by the delete one jackknife technique where SUDAAN generates the jackknife replicate weights. The SUDAAN syntax DESIGN = JACKKNIFE, with the JACKWGTS statement, invokes the common sampling plan description *WR* with the jackknife weights provided to SUDAAN as variables in the input data set.

95. The sample survey design for the Burundi survey and specification of Taylor series linearization for variance estimation are described to SUDAAN as follows:

```
PROC …….    DESIGN = WR ……
NEST    PSTRA    PPSU
WEIGHT    W2
```

96. The sample survey design for the Burundi survey and specification of BRR (balanced repeated replication) for variance estimation are described to SUDAAN as follows:

```
PROC …….    DESIGN = BRR ……
WEIGHT    W2
REPWGT  REPLWT01-REPLWT32
```

Note above that the REPWGT statement identifies the replicate weight variables included in the input data set. These 32 replicate weight variables are based on the 30 pseudo-strata, with 2 PSUs per pseudo-stratum, and were obtained by using WesVar. Note also that the NEST statement is absent when BRR is used; SUDAAN does not need to know the stratification and PSU variables, since it uses only the replicate weight variables for variance estimation.

2. DESCRIPT

97. The DESCRIPT procedure estimates population totals and means for continuous variables as well as population totals and percentages for categorical variables. The VAR statement lists the variables (dependent) to be analysed. For a given DESCRIPT program, all variables on the VAR statement must be continuous or all variables must be categorical. If categorical variables are on the VAR statement, then the CATLEVEL statement must also be used to indicate for which levels of each categorical variable estimates are desired. For example, the two statements below estimate the percentage of the inference population in Burundi who are seropositive and not seropositive [assuming *IMMUNE* is coded 1, 2 or . (dot) for missing].

VAR	*IMMUNE*	*IMMUNE*
CATLEVEL	1	2

98. Estimates are provided for domains by using a TABLES statement that contains one or more categorical variables. Domains can be compared with each other via linear contrasts using the CONTRAST, PAIRWISE or DIFFVAR statements. Standardized rates and means can be estimated, for example, an age-adjusted prevalence for disease, by using the STDVAR and STDWGT statements. Linear and higher-level (quadratic, etc.) trends on means or percentages can be assessed across levels of some categorical variable by using the POLY (POLYNOMIAL) statement; SUDAAN uses orthogonal polynomial linear contrasts for these analyses.

99. All variables on a TABLES, CONTRAST, PAIRWISE, DIFFVAR, STDVAR or POLY statement must also appear on a SUBGROUP statement, and a required LEVELS statement indicates the highest coded value in the analysis for each categorical variable on the SUBGROUP statement.

100. The SUBPOPN statement in SUDAAN, which can be used in all PROCS, restricts analyses to a subpopulation, for example, only older women. Use the SUBPOPN statement with the full sample survey data set input into SUDAAN instead of subsetting the input data set to the subpopulation of interest before using SUDAAN, since the latter procedure may result in incorrectly estimated standard errors inasmuch as some sample PSUs may be missing from the subsetted data set.

3. CROSSTAB

101. The CROSSTAB procedure is for categorical variables only. The TABLES statement in CROSSTAB indicates the one-way, two-way or multi-way tables for which population percentages and totals are estimated. Corresponding SUBGROUP and LEVELS statements are required for all variables on the TABLES statement.

102. The TEST statement in CROSSTAB requests chi-square tests for testing the null hypothesis that two categorical variables are statistically independent. One chi-square test is based on a Pearson type test (CHISQ), using "observed minus expected" calculations on estimated population totals. The other chi-square test is based on estimated population odds (LLCHISQ). Odds ratios and relative risks (prevalence ratios, really), with confidence intervals, are estimated for 2 x 2 tables by using RISK = ALL on the PRINT statement. Finally, a Cochran-Mantel-Haenszel test (use CMH on the TEST statement) is available to assess statistical independence of two variables while controlling on ("stratifying" on) a third variable.

4. Numerical examples

103. Section III of the annex on the CD-ROM illustrates the use of CROSSTAB and DESCRIPT to work the three examples listed in paragraph 60, using SAS-CALLABLE SUDAAN (SAS Version 8.2 and SUDAAN Version 8.0). Both Taylor series linearization and BRR (balanced repeated replication) are used for variance estimation. Review of the annotated SUDAAN programs (user-written) and annotated SUDAAN output should aid readers in writing their own SUDAAN programs and interpreting the output. Only selected SUDAAN analyses discussed in TABLES 1, 3, 4, 5, and 6 are included and annotated in the annex, section III.

104. Table XXI.1, row 4 summarizes the CROSSTAB and DESCRIPT output in section III (annex) for estimating the percentage and number of women in the Burundi inference population who are seropositive, with estimated standard error. The CROSSTAB and DESCRIPT results from SUDAAN are identical for a given method of variance estimation (as expected), and the Taylor Series and BRR results are identical (not always true). The SUDAAN results agree with results from SAS SURVEYMEANS. Note that CROSSTAB and DESCRIPT do not calculate confidence intervals for estimated population percentages or totals.

105. Table XXI.3, row 2, in the annex, section VII (CD-ROM), shows that identical output is obtained from CROSSTAB and DESCRIPT (whether with Taylor series or BRR) for estimating the percentage who are seropositive, but for each of the two domains of rural and urban women. The SUDAAN CROSSTAB and DESCRIPT results agree with SAS SURVEYMEANS.

106. Table XXI.4, row 2 in the annex, section VII (CD-ROM), summarizes the DESCRIPT output (with Taylor series and BRR) that uses a linear contrast to compare rural with urban women on percentage who are seropositive. There is a negligible difference in the estimated standard error with Taylor series and BRR. The conclusion is: urban and rural women in the Burundi inference population differ on seropositivity prevalence; urban women have a higher prevalence. Note that the DESCRIPT linear contrast results agree with using SAS SURVEYREG to compare two domains.

107. Table XXI.5, rows 1 and 2, in the annex, section VII (CD-ROM), shows results from the two different chi-square tests available in CROSSTAB: Pearson (CHISQ) and log-linear (LLCHISQ). Results using Taylor Series and BRR are identical. The estimated seropositivity prevalence is significantly higher for urban women than for rural women (using CHISQ), and the estimated odds of seropositivity is significantly higher for urban women than for rural women (using LLCHISQ).

108. Table XXI.6, row 1 in the annex, section VII (CD-ROM), shows the estimated odds ratio (0.393) and prevalence ratio (0.797) for seropositivity (rural to urban), each with a 95 per cent confidence interval. Taylor Series and BRR have negligible differences in the upper limit for the 95 per cent confidence interval on odds ratio. The estimated odds ratio and prevalence ratio differ in magnitude because the prevalence of seropositivity is not low.

5. Advantages/disadvantages/cost

109. SUDAAN is a comprehensive sample survey (and correlated data) software package with analytical strengths for both descriptive and modelling analyses. It has extensive capability to estimate and test user-specified contrast matrices on population parameters, including regression coefficients. It runs in both mainframe and PC environments. SAS users likely have an advantage in learning SUDAAN, since its syntax is similar to SAS. However, some of the syntax of SUDAAN is esoteric, perhaps requiring more learning time than other packages.

110. Compared with that of other software packages reviewed in this chapter, the cost of SUDAAN is high, especially if used as SAS-Callable SUDAAN because, then, SAS also is required. Technical support is provided for licensed users. The SUDAAN Users Manual for Version 8.0, primarily for reference as opposed to learning SUDDAN, has several detailed annotated examples of analyses with NHANES-III (National Health and Nutrition Examination Survey-III) data which can be useful for learning how to use SUDAAN.

111. SUDAAN is the only software package illustrated here that includes both major approaches to variance estimation, Taylor series linearization and replication methods. However, SUDAAN does not construct replicate weights for balanced repeated replication (BRR), requiring the user to provide these weights. SUDAAN constructs replicate weights for the jackknife delete one procedure and will also accept jackknife replicate weights if they are included in the input data set.

112. SUDAAN also is the only software package reviewed here that has extensive capability for describing several stages of sampling, stratification and *fpc* terms for incorporation into variance estimation. Further, it has several different definitions for design effect calculations to allow one to exclude from the design effect the effects of oversampling and/or of unequal weighting.

113. ASCII data input into SUDAAN is cumbersome, making the other two data input options preferable, namely, a SAS or SPSS data set. A SAS data set input into standalone SUDAAN must be SAS Version 6.04 or a SAS transport file. SAS-Callable SUDAAN can read any data set that SAS can read. SUDAAN output can be saved electronically to a SAS data file format for further use in SAS or spreadsheet software such as EXCEL. SUDAAN has very limited capability for recoding variables and no capability for data management. Thus, it is prudent to undertake any necessary recoding and formation of new variables in either SAS or SPSS (depending upon type of input data set) before using SUDAAN.

I. Sample survey procedures in STATA 7.0

1. Overview of STATA

114. STATA is a general statistical software package that added extensive capability for sample survey data analysis in 1995. STATA 7.0 is illustrated here; Version 8.0 was released in 2003. Only Taylor series linearization is used for variance estimation. The common sampling plan description *WR* is default. STATA can incorporate *fpc* terms into variance estimation for single stage without replacement sampling plans (simple random sampling and stratified random sampling) and for one stage without replacement cluster sampling (stratified or not) where equal probability sampling is used for clusters (PSUs) within a stratum and all elements in a sampled PSU are included in the sample.

115. The breadth of sample survey analyses of STATA compares favorably with that of SUDAAN, with mathematical statistical capability for user-specified contrast matrices on population parameters, including regression coefficients. STATA runs interactively with short and simple commands, making it relatively easy to learn. However, user-written programs can be submitted in batch mode if desired. STATA is case-sensitive, and commands to STATA are typed in lower case. STATA allocates a default amount of memory into which it loads a copy of the input data set. If this memory is insufficient for large data sets, the memory can be increased with the set memory command.

116. The sample survey commands in STATA begin with the name svy (for survey). Descriptive commands are available for estimating a population mean (svymean), a population total (svytotal), a population proportion (svyprop), and percentages and totals in two-way tables (svytab). Confidence intervals on population proportions from svytab use a logit transform so that estimated lower and upper limits are constrained within (0,1). Eight different chi-square tests for sample survey data in two-way tables are available in svytab. Available modelling procedures include linear regression, logistic regression (including multinomial with a nominal or ordered variable), Poisson regression, and probit models.

117. The svyset command is used to specify the sampling plan to STATA. To describe the common sampling plan *WR* (default), three keywords for the command svyset are typed into STATA interactively. The keyword strata precedes the stratification variable name, the keyword **psu** precedes the PSU variable name, and the keyword pweight precedes the sampling weight variable name. Thus, the sampling plan for the Burundi survey is described to STATA V7 as:

svyset strata *pstra*
svyset psu *ppsu*
svyset pweight *w2*

118. As indicated earlier for the sample survey procedures in SAS, omission of the strata keyword in STATA implies no stratification of PSUs prior to first-stage sampling. Omission of the psu keyword implies one-stage sampling of elements and no clustering of sampled elements. Omission of the pweight keyword implies equally weighted sample elements, with a default

value of 1.0 for the weighting variable. The syntax for the svyset command is revised in STATA V8.

119. The command svydes instructs STATA to output the survey design variables it has attached to the data set (from the svyset commands) and to summarize the number of strata, the number of PSUs per stratum, and the average number of observations per PSU within each stratum. This is a very useful summary of characteristics of the sample survey design.

2. SVYMEAN, SVYPROP, SVYTOTAL, SVYLC

120. The svymean command estimates a population mean, either for a continuous variable or for an indicator variable coded 1 or 0 (that is to say, an estimated population proportion). Output options include estimated standard error, estimated coefficient of variation, design effect and confidence interval on the population parameter.

121. The svyprop command is for categorical data: it estimates the proportion of the population that is at each level of the categorical variable, along with estimated standard error. Fewer output options are available with svyprop, compared with svymean.

122. The svytotal command estimates a population total for either a continuous or an indicator (0, 1) variable, with estimated standard error, estimated coefficient of variation, design effect and confidence interval.

123. Each of the three commands above can be used to estimate population parameters for domains by using the option by on the command line, for example**,** by (stra**)** or by (urb_rur) to analyse the two domains of rural and urban women in Burundi. STATA uses correct variance estimation formulae for domains with the by statement in its svy commands.

124. In addition, each of the three commands above can be used with a subpop option on the command line to perform estimation of population parameters for a subpopulation, for example, only older women. Do not use the STATA "if" statement for subpopulation analyses because estimated variances may be incorrect; use the subpop option.

125. The svylc command estimates user specified linear combinations of domain means, proportions or totals, along with estimated standard error, t-test, p-value, and confidence interval. This command can be used to compare domains with each other. In V8.0, the svylc command is replaced by lincom**.** The command svylc continues to work in V8.0 but is no longer documented.

3. SVYTAB

126. The svytab command in STATA is for two-way tables. It estimates population percentages (row, column or total) with estimated standard errors, population totals for table cells with estimated standard errors, and confidence intervals. A logit transform is used to obtain confidence intervals on population proportions so that estimated lower and upper limits are constrained to be in the interval (0, 1). Eight different chi-square tests are available to test the

null hypothesis of statistical independence of the two categorical variables in the table. The command subpop is available for use with svytab.

4. Numerical examples

127. Section IV of the annex (CD-ROM) illustrates the use of STATA commands to work the three examples listed in paragraph 60. Each worked example is a log file of the interactive session with STATA. Review of the annotated STATA log (user commands and STATA output) should aid readers in using the sample survey commands in STATA and interpreting the output.

128. The commands svymean and svytotal were used with the indicator variable *BLOOD* (1=seropositive, 0=seronegative). Table XXI.1 (rows 5 and 6) shows the estimated number and percentage of women who are seropositive, with confidence intervals. The STATA calculations agree with SAS SURVEYMEANS and with SUDAAN DESCRIPT and CROSSTAB.

129. Table XXI.3 (row 3) in the annex, section VII (CD-ROM) shows the estimated percentage of women who are seropositive, by rural/urban residence. The STATA svytab point estimates and estimated standard errors agree with SAS SURVEYMEANS and with SUDAAN DESCRIPT and CROSSTAB. However, the confidence intervals for domains differ slightly between STATA svytab and SAS SURVEYMEANS because STATA svytab uses a logit transform to obtain confidence intervals.

130. Table XXI.4 (row 3) in the annex, section VII (CD-ROM), presents the STATA **svylc** results for the linear contrast that compares rural and urban women on percentage who are seropositive, indicating a significant difference between the two domains. The STATA results agree with SUDAAN DESCRIPT and with using SAS SURVEYREG for domain comparisons.

131. Table XXI.5 (rows 3 through 5) in the annex, section VII (CD-ROM), presents the STATA svytab results for three chi-square tests of the null hypothesis that seropositivity is statistically independent of rural/urban residence. All three svytab chi-square tests have similar (and small) p-values. The default chi-square test for STATA svytab (row 3) is a Pearson type chi-square test proposed by Rao and Scott (1981; 1984) with a second-order correction. The other two chi-square tests in svytab (rows 4 and 5) are the same chi-square tests as in SUDAAN CROSSTAB, and STATA and SUDAAN yield the same calculations for these two tests.

132. Since the svytab command in STATA does not produce odds ratios or prevalence ratios, the command svylogit was used to estimate odds ratio (urban to rural) for seropositivity. The STATA odds ratio, with confidence interval, is in table XXI.6 (row 2) in the annex, section VII (CD-ROM). The STATA svylogit command gives the same calculations as SUDAAN CROSSTAB for point estimate and confidence interval.

5. Advantages/disadvantages/cost

133. STATA is a comprehensive general statistical analysis package and also has extensive analytical capability for sample survey data, including descriptive and design-based modelling

procedures. It provides many modelling procedures for sample survey data. STATA has received very good reviews as a statistical package, is relatively easy to learn, and has an active users group. Compared with other software packages reviewed in this chapter, its cost is moderate.

134. STATA accepts user-defined contrast matrices of estimated population parameters, including regression coefficients, for those who wish to test their own specific hypotheses or estimate combinations of population parameters. In general, it allows great flexibility in conducting statistical analyses for those with the requisite mathematical statistical background.

135. STATA uses only Taylor series linearization and is limited to the common sampling plan description *WR*. However, it can include in variance estimation for without replacement sampling the *fpc* terms for one-stage sampling of elements and for one-stage cluster sampling. It is somewhat difficult, but possible, to extract STATA analytical results (for example, unweighted sample sizes, point estimates, standard errors) for export to other data formats.

J. Sample survey procedures in Epi-Info 6.04d and Epi-Info 2002

1. Overview of Epi-Info

136. Epi-Info has been developed over many years by the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO). This software is available at no cost as a download from the CDC web site: http://www.cdc.gov/epiinfo/ .

137. Two versions of Epi-Info are available: the last DOS-based Epi-Info Version 6.04d and the most recent Windows-based Epi-Info 2002.

138. The capabilities of Epi-Info include development of a questionnaire or research data-collection form, customized data entry, data analysis and word processing. Its analytical and statistical capabilities are oriented towards epidemiologists worldwide. Output (analytical results) from Epi-Info analyses can be sent to the screen, to a printer, or to an electronic file.

139. Both versions of Epi-Info (DOS or Windows) have capability for basic descriptive analyses of complex sample survey data. Only the common sampling plan description *WR* is available. The input data set must be sorted by two of the three survey design variables: the stratification variable *STRATVAR* and by the PSU variable *PSUVAR* within stratum. Epi-Info does not incorporate any *fpc* terms into variance estimation. Also, it does not estimate population totals. Taylor series linearization is used for variance estimation.

140. The analytical capability of Epi-Info for complex sample survey data originally was developed for the Behavioral Risk Factor Surveillance System (BRFSS), a CDC-sponsored annual health sample survey programme for States in the United States of America (Brogan, 1998 and in press) and for the WHO cluster sample methodology used worldwide by the Expanded Programme on Immunization (EPI) to estimate vaccination coverage among children (Brogan and others, 1994). However, the sample survey procedures in Epi-Info may be used for any complex sample survey that can be described by the common sampling plan description *WR*.

2. Epi-Info Version 6.04d (DOS), CSAMPLE module

141. Epi-Info for DOS was a joint development effort of CDC and WHO. Data input for Epi-Info 6.04d is a dBase file or an ASCII file which Epi-Info then converts into an Epi-Info data file *.rec. Software packages exist to convert SAS or SPSS or other types of data files into an Epi-Info *.rec file, for example, DBMS-COPY (http://www.dataflux.com/conceptual/). Epi-Info 6.04d runs as an interactive program and cannot be run in batch mode. The DOS version may be preferred over the Windows version by those who have older computers, older operating systems and/or limited hard-drive storage space.

142. The CSAMPLE module in Epi-Info Version 6.04d conducts analyses for complex sample survey data. CSAMPLE estimates a population mean (for a continuous variable or for an indicator variable coded 1/0) or a population percentage (for a categorical variable), along with estimated standard error, confidence interval(s) and design effect. These estimates also are provided for domains formed by levels of a categorical variable. In addition, CSAMPLE estimates the difference between domain means or domain percentages, with corresponding estimated standard error of the estimated difference and a confidence interval on the population difference. CSAMPLE estimates odds ratio and risk ratio for 2 x 2 tables. Note that CSAMPLE does not estimate population totals.

143. When the CSAMPLE module is opened in Epi-Info 6.04d, a data input screen appears where the user specifies variables to be used in the analysis. The user selects a variable for each of the three survey design boxes: STRATA (the stratification variable), PSU (the PSU or cluster variable) and WEIGHT (the sampling weight variable). For the Burundi survey, the specification to Epi-Info was as follows:

STRATA	*PSTRA*
PSU	*PPSU*
WEIGHT	*W2*

144. The user specifies the analysis variable (or dependent variable) in the box called MAIN. This variable can be continuous, such as *IUML*, or categorical, such as *IMMUNE*. If the estimated population mean is desired for the continuous (or assumed to be continuous) variable specified in MAIN, then the user clicks on the option MEANS. If the estimated population percentages are desired for the categorical variable specified in MAIN, then the user clicks on the option TABLE.

145. If estimated means or percentages are desired for domains, then the variable that defines the domains is specified in the box called CROSSTAB and the analysis variable is specified in the MAIN box.

146. In addition, CSAMPLE can estimate the difference between two domains on the mean of an analysis variable. The user can specify the two levels of the CROSSTAB variable that define the two domains to be compared with each other.

3. Epi-Info 2002 (Windows)

147. Epi-Info 2002, a Windows application, has been developed by CDC. Data input for Epi-Info 2002 data analysis is via a MicroSoft Access 1997 file (*.mdb) or a dBase file. Epi-Info 2002 can also read the *.rec files prepared for the DOS versions of Epi-Info. The software runs interactively but has an option to run in batch mode.

148. Epi-Info 2002 has three complex sample procedures located in the Analyze Data section under Advanced Statistics. Complex Sample Frequencies estimates a one-way percentage distribution for a categorical variable, with estimated standard error and confidence intervals. Complex Sample Tables estimates row and column percentages for a two-way table of categorical variables [labelled exposure (row) and outcome (column)], with estimated standard errors and confidence intervals for row percentages. If the table is 2 x 2, the procedure also estimates odds ratio and risk ratio, with confidence intervals. Complex Sample Means estimates the mean for a continuous variable, with estimated standard error and confidence interval, including estimation of mean for domains formed by a categorical variable. If the domain variable is at two levels, the difference between domain means also is estimated, with estimated standard error and confidence interval.

149. In all three complex sample procedures, the survey design variables are identified in three boxes labelled Weight, PSU and Stratify By (the sample survey stratification variable). In order to obtain estimated standard errors and confidence intervals as output, double click on OPTIONS:SET and then choose Statistics = Advanced.

4. Numerical examples

150. Section V of the annex (CD-ROM) illustrates the use of CSAMPLE in Epi-Info 6.04d to work the three examples in paragraph 60. Each worked example contains the output from Epi-Info, annotated with comments. Review of the annotated output should aid readers in interpreting the CSAMPLE output.

151. Table XXI.1 (row 7) gives the Epi-Info 6.04d estimate for percentage of women who are seropositive. The Epi-Info point estimate and estimated standard error agree with SAS SURVEYMEANS, STATA svymean and SUDAAN DESCRIPT and CROSSTAB. The 95 per cent confidence interval on seropositivity prevalence is narrower than the confidence intervals given by SAS SURVEYMEANS and STATA svymean**.** This occurs because Epi-Info uses z = 1.96 in its 95 per cent confidence interval calculation rather than the Student-t value of 2.042 with 30 df, the denominator degrees of freedom for the Burundi survey [number of PSUs (60) less number of pseudo-strata (30)].

152. Table XXI.3 (row 4) in section VII of the annex (CD-ROM) gives the Epi-Info estimates of seropositivity prevalence by rural/urban residence. The Epi-Info point estimates and estimated standard errors agree with SAS SURVEYMEANS, STATA svytab, and SUDAAN DESCRIPT and CROSSTAB. The Epi-Info domain confidence intervals are narrower compared with those from SAS SURVEYMEANS and STATA svytab because Epi-Info uses z = 1.96.

153. Table XXI.4 (row 4) in section VII of the annex (CD-ROM) gives the result of the Epi-Info linear contrast that compares rural and urban women on seropositivity prevalence. The estimated contrast value (-16.99 per cent) agrees with SAS SURVEYREG, with SUDAAN DESCRIPT and with STATA svylc. Epi-Info does not give the estimated standard error of the estimated difference, and the 95 per cent confidence interval that Epi-Info gives on the contrast value is in error.

154. Table XXI.6 (row 3) in section VII of the annex (CD-ROM) gives the Epi-Info estimated odds ratio (urban to rural) and estimated prevalence ratio of seropositivity, with 95 per cent confidence interval. The Epi-Info point estimates agree exactly with SUDAAN CROSSTAB and with STATA svylogit, and the Epi-Info confidence intervals are in close agreement with SUDAAN and STATA.

5. Advantages/disadvantages/cost

155. A major advantage of Epi-Info is its cost: it can be downloaded free from the CDC web site. Further, it is available for both DOS and WINDOWS operating systems, permitting wide flexibility on hardware and software required to run Epi-Info. The sample survey capability of Epi-Info certainly would appeal to those who already are Epi-Info users for other types of epidemiological or statistical analyses.

156. Epi-Info uses only Taylor series linearization and handles only the common sampling plan description *WR*. The CSAMPLE module in the DOS release and its counterpart in the Windows release (three procedures under Advanced Statistics) are adequate for basic descriptive statistics for complex sample survey data. This includes estimation of population means or percentages for the entire population and for domains, as well as comparison of domains. Epi-Info has no sample survey capability for estimating population totals, for conducting chi-square tests, for incorporating the *fpc* (finite population correction) terms into variance estimation, or for design-based modelling analyses (for example, logistic regression or linear regression).

K. WesVar 4.2

1. Overview of WesVar

157. WesVar is a software package dedicated to the analysis of sample survey data. Replication methods (Rust and Rao, 1996) are used for variance estimation: BRR, including the optional Fay factor, and three jackknife variations. WesVar does not have capability for Taylor series linearization. Sample survey designs that lend themselves well to BRR have several strata and exactly two sample PSUs per stratum. Jackknife methods, like Taylor series linearization, can be applied to a design with any number (> = 2) of sample PSUs per stratum.

158. The default sampling plan for WesVar is the common sampling plan *WR* referred to earlier. WesVar has capability to include *fpc* factors in variance estimation, but only for jackknife techniques and only for one-stage sampling of elements.

159. WesVar 4.2 can read the following types of input data sets: PC-SAS for DOS, SAS transport, SAS (versions 6-8), SPSS, STATA, ASCII, and ODBC-compliant files such as Microsoft Excel or Access. Consistent with the assumed common sampling plan *WR*, if replicate weights are to be constructed, WesVar requires the stratification, PSU and weight variables for each observation. Once the replicate weights are on the file, however, PSU and strata identifiers are not needed: this is a confidentiality advantage of replication methods for public use files. WesVar is the only package among those reviewed that can adjust basic survey weights for non-response, post-stratification and raking. After preparation of the input data set is completed, it is saved as a WesVar (*.var) file for data analysis and any future data management.

160. A full range of descriptive statistics is available: estimated population means, percentages, percentiles and totals, along with estimated standard error, coefficient of variation, confidence interval and design effect. A particular strength of WesVar, and replication methods in general, is the ability to obtain point estimates (with estimated standard error) of user-specified functions of population parameters, for example, prevalence ratios. Design-based regression analyses are available in WesVar: linear, logistic and multinomial logistic.

161. A download of WesVar Version 4 is available from the WESTAT web page for a thirty-day trial period. WesVar Version 2 is available for download from the web page and can be used for an unlimited time at no cost (see http://www.westat.com/wesvar). WesVar Version 4, compared with Version 2, accepts a wider variety of input data sets, has better capability for file handling and data management, adjusts replicated weights for non-response, and includes many more analytical options. A user could begin with WesVar Version 2 and then upgrade to Version 4, if needed.

2. Using WesVar Version 4.2

162. The user interacts with WesVar via pop-up menus in a Windows environment. When the WesVar software is opened, the first menu contains four options. The first option, new WesVar data file, (1) reads in an input data set that is not a WesVar data set; (2) creates replicate weights or accepts replicate weights already in the input data set; (3) recodes, transforms, labels and formats variables; (4) performs post-stratification, raking and non-response adjustments; (5) defines subpopulations for analysis; and (6) modifies the default ddf if requested, and then saves the data set as a WesVar file. The second option, open WesVar data file**,** reads in a WesVar data file and allows all of the six operations just listed above.

163. The third option, New WesVar Notebook**,** accepts analysis requests for a WesVar data file, runs the requests, displays the output, and saves the requests and output in a "notebook", WesVar's system for organizing requested analyses and resulting output. One of two types of analysis is requested: tables or regression (linear, logistic or multinomial). After the tables or regression choice is made, many options are available to specify the analysis. Navigating the menu screens for analysis and reading the output are not straightforward, but the WesVar User's Guide has several useful examples to illustrate menu navigation and output organization.

164. If the requests and output from a previous WesVar session were saved in a notebook, then the fourth option on the first menu could be chosen: open WesVar Notebook**.** New

analysis requests can be added to an existing notebook and then saved. All analyses related to a specific WesVar data file or to a specific project can be organized into one or more notebooks.

165. One of five replication methods in WesVar must be specified in order to construct replicate weights or to recognize replicate weights that already exist in the input data file. These replication methods are:

(a) Balanced repeated replication (BRR)-exactly two sample PSUs per stratum;
(b) Fay's perturbation method (FAY) with BRR;
(c) Jackknife delete one with no explicit stratification (JK1);
(d) Jackknife with exactly two sample PSUs per stratum (JK2);
(e) Jackknife with two or more sample PSUs per stratum (JKn).

166. Appendices A and D in the WesVar User's Guide contain an excellent overview of these five replication methods and illustrate via examples how to translate different sampling plans into one of these five methods.

3. Numerical examples

167. Since the Burundi input data set did not contain replicate weights, it was necessary to choose one of the five available replication techniques and then request WesVar to calculate the replicate weights. The Burundi survey design variables needed by WesVar were: *PSTRA*, *PPSU* and *W2*. Since the Burundi sampling plan is *WR,* with 30 pseudo-strata and exactly two sample PSUs per stratum, BRR or JK2 are the best choices. BRR was chosen with no Fay perturbation factor. Further, no non-response adjustments or post-stratification or raking was carried out for the replicates since these adjustments were not carried out on the full data set when Taylor series linearization was used.

168. Section VI (CD-ROM) illustrates the use of WesVar to work the three examples listed in paragraph 60. Each worked example contains the output from WesVar 4.1 or 4.2, although the input menu screens for the requested analyses are not shown. Review of the annotated WESVAR output should aid readers in interpreting the WesVar output.

169. Table XXI.1, row 8, shows that WesVar agrees with all other sample survey software packages on the estimated percentage and estimated number of women who are seropositive (with standard errors). The WesVar confidence intervals agree with SAS and STATA but not Epi-Info, which are too narrow.

170. Table XXI.3, row 5 in the annex (CD-ROM), shows that WesVar agrees with all other software packages on domain point estimates and estimated standard errors. The WesVar confidence intervals are very close to those of SAS SURVEYMEAS but differ slightly from STATA svytab (uses logit transform) and Epi-Info (uses z=1.96 rather than Student t-value).

171. Table XXI.4, row 5, in section VII of the annex (CD-ROM), shows the WesVar linear contrast result to compare rural and urban women on seropositivity prevalence. WesVar agrees with SAS SURVEYREG, SUDAAN DESCRIPT and STATA svylc on estimated standard error

for the linear contrast and Student t-statistic. Confidence intervals on the linear contrast have negligible differences among SAS, STATA and WesVar.

172. Table XXI.5, rows 6 and 7, in section VII of the annex (CD-ROM), show the two Rao/Scott chi-square tests for complex sample survey data as implemented in WesVar. These calculations do not agree exactly with any of the other chi-square tests in other packages.

173. Table XXI.6, row 4 in section VII of the annex (CD-ROM), shows that the WesVar logistic regression procedure produces the same estimated odds ratio and essentially the same confidence interval as do SUDAAN CROSSTAB and STATA svylogit. Table XXI.6, row 5, shows that the WesVar estimated prevalence ratio (by using cell functions in TABLES) agrees with SUDAAN CROSSTAB and Epi-Info, with negligible differences in the confidence intervals between SUDAAN and WesVar.

4. Advantages/disadvantages/cost

174. WesVar uses only replication techniques for variance estimation. Secondary data analysts of public release data sets with replicate weights provided do not have to know details of the sample design (for example, the survey design variables *STRATVAR* and *PSUVAR*), although they do need to specify to WesVar the method that was used to obtain the replicate weights (information obtained from the survey documentation). If the user needs to use WesVar to construct replicate weights for the sample survey data set, some knowledge about replication methods is required and, in addition, the three survey design variables associated with the common sampling plan *WR* must be available (stratification variable *STRATVAR*, PSU variable *PSUVAR* within stratum, sample weight variable *WTVAR*).

175. WesVar has extensive capability for constructing replicate weights for a sample survey data set. Five different replication techniques are available, including the opportunity to adjust for non-response and to conduct post-stratification or raking. In addition, WesVar has options for incorporating a finite population correction term for single-stage sampling using jackknife techniques for variance estimation.

176. For those new to replication techniques for variance estimation, appendix A of the WesVar User's Guide has an excellent overview of the theory and practice of replication techniques, although reading this material requires some background in mathematical statistics. Further, appendix D of the User's Guide gives very useful guidance and several examples for choosing a replication method for a given sampling plan.

177. WesVar is capable of estimating user-defined functions of population parameters, something that is more difficult to do with the Taylor series linearization approach to variance estimation. Thus, it is inherently more flexible than the other software packages reviewed in this chapter in terms of the population parameters it is able to estimate. Although SUDAAN has BRR and jackknife replication methods available for variance estimation, SUDAAN does not allow the user to specify functions of population parameters to be estimated, as does WesVar.

178. Direct output from WesVaris somewhat difficult to work with, compared with most other sample survey software. WesVar output contains one row for each cell of a requested table (as illustrated in section VI of the annex). However, a Table Viewer utility is available as a free download from the WesVar web site. This adjunct program converts the WesVar 4 output into a grid or tabular form to display on the screen or to print or produces an electronic file in this form for pasting into applications such as Microsoft Word or Excel.

179. Compared with that of other software packages reviewed in this chapter, the cost of WESVAR is low. Version 4 is available as a free download for a thirty-day trial period, and version 2 is available as a free download for unlimited use.

L. PC-CARP

180. PC-CARP is a standalone MS-DOS program developed at and available from Iowa State University (Statistics Department). It handles the common sampling plan *WR* discussed above and, for simpler designs, can incorporate *fpc* terms up to two stages of sampling. Taylor series linearization is used for variance estimation.

181. Point estimates, estimated standard errors and confidence intervals are constructed for population and subpopulation totals, means, proportions, quantiles, empirical distribution functions, ratios, and differences of ratios (and hence differences of means, proportions and totals). Also included are design-based linear regression and a two-way contingency table analysis, including a chi-square test. Design effect and coefficient of variation for point estimates are calculated. Three add-on modules are available: PC-CARPL for design-based logistic regression, POSTCARP for post-stratification of sample survey data, and EV CARP for regression analysis with measurement error in the explanatory variables.

182. The user interface is via keyboard-navigated text-based menu screens; mouse use is not supported. Only ASCII files are accepted as input where the input records may be space-delimited or fixed-length with a supporting format statement in FORTRAN syntax. There are no restrictions on number of observations in the data set, and most analyses can accept up to 50 variables. PC-CARP can run on older computer systems with DOS 5.0 or later and Windows 3.1x or Windows 95 or later. It takes only 3 metabytes (Mb) of hard-disk space and only 450 kilobytes (Kb) of random access memory (RAM). Any newer system must support DOS programs in order to run PC-CARP.

183. The one-time purchase price for PC-CARP, compared with that of other software packages reviewed, is low. No annual renewal fee is required. There is a small fee for each of the three add-on modules.

184. No example analyses of the Burundi survey with PC-CARP are reported in this chapter.

M. CENVAR

185. CENVAR is one component of a comprehensive statistical software system called Integrated Microcomputer Processing System (IMPS) that was designed by the United States Bureau of the Census for processing, management and analysis of complex sample survey data. IMPS, including CENVAR, is available at no cost and can be downloaded from http://www.census.gov/ipc/www/imps/download.htm. As of early 2003, part of IMPS is Windows-based and part is still DOS-based. No discussion of IMPS is included in this chapter.

186. CENVAR is adapted from PC-CARP and thus has many of its characteristics. CENVAR supports the same sample designs as PC-CARP, that is to say, the common sampling plan *WR* as well as incorporation of *fpc* terms into variance estimation for simpler one- and two-stage designs using without replacement sampling. Taylor series linearization is used for variance estimation. The software is menu-driven and has no mouse support.

187. Point estimates, estimated standard errors, confidence intervals, coefficients of variation and design effects are constructed for population and subpopulation totals, means, proportions, ratios, and differences of ratios (and hence differences of means, proportions and totals). The remaining options in PC-CARP are not included, namely, design-based linear regression, a two-way contingency table analysis, and quantile estimation. The add-on modules in PC-CARP are not included in CENVAR.

188. The CENVAR User's Guide (1995), about 100 pages long, can be downloaded from the web. It contains useful examples and training exercises from three sample surveys conducted by the Bureau of the Census. CENVAR accepts only ASCII data input and it requires the IMPS Data Dictionary software. The Data Dictionary must be created prior to running CENVAR. Thus, some familiarity with IMPS must be obtained in order to use CENVAR. CENVAR runs in a DOS 3.2 or higher environment on a PC. It requires 10 Mb of disk storage and 640K byes of available memory. No example analyses of the Burundi survey with CENVAR are reported in this chapter.

N. IVEware (Beta version)

189. IVEware (Imputation and Variance Estimation Software) is a SAS callable software application for sample survey data recently developed by the Survey Methodology Program at the University of Michigan. It handles the common sampling plan *WR* and uses either Taylor series linearization or replication methods, depending upon the procedure.

190. The IMPUTE module uses a multivariate sequential regression approach to impute item missing values, including multiple imputed data sets. The DESCRIBE module estimates population and subpopulation means and proportions, subgroup differences and linear contrasts of means and proportions; Taylor series linearization is used. The REGRESS module fits several design-based regression models (linear, logistic, etc.); the jackknife replication technique is used. The SASMOD module allows users to take into account complex sample design features when using several SAS PROCS for data analysis, for example, CATMOD, GENMOD, and MIXED.

A multiple imputation analysis can be performed for the three data analysis modules (DESCRIBE, REGRESS, SASMOD).

191. IVEware runs with SAS V 6.12 or higher and is available for personal computers using Microsoft Windows or Linux operating systems; other platforms are available. Although users do not need to be familiar with the IVEware building blocks of SAS Macro Language, C and FORTRAN, they do need to have a moderate amount of SAS experience and, of course, SAS software. The IVEware software and documentation are available for free download from http://www.isr.umich.edu/src/smp/ive/. No example analyses of the Burundi survey with IVEware are reported in this chapter.

O. Conclusions and recommendations

192. Some data analysts may be surprised that specialized software is needed for variance estimation with complex sample survey data. Although some analysts may want to use software developed for simple random samples for variance estimation with complex sample survey data, we do not recommend this. There are several software options now for variance estimation, including some that are free. Reasons for choosing among these options are likely to be familiarity with the software, cost, ease of use, and whether one is interested in only basic descriptive analyses or more comprehensive analyses

193. If you already use a general statistical package that has sample survey variance estimation capability, then that package is an obvious choice, since the acquisition cost is already paid and the syntax is familiar. STATA users have comprehensive sample survey variance estimation capability in that package and should not need to look elsewhere unless the data set being analysed must use replication methods. SAS users, with the recently released Version 9.0, have increased capability for sample survey variance estimation compared with Version 8.2 and can expect additional capability in the future. However, if SAS V9.0 is not sufficient for your sample survey variance estimation purposes, using the free IVEware package with SAS may meet your needs. Epi-Info users have only basic sample survey data variance estimation capability in that package, but if that is all you need, it will suffice. SPSS, a widely used statistical analysis package, released a complex sample survey add-on module in late 2003, so that this is now a viable choice.

194. If your general statistical software package does not have the necessary sample survey variance estimation capability, then consider a specialized sample survey software package (for example, WesVar, SUDAAN, PC-CARP or CENVAR) or a different general statistical package (for example, STATA or SAS with/without IVEware or SPSS or perhaps Epi-Info). SUDAAN often appeals to SAS users because of its SAS-like syntax and the option to run it as SAS-callable SUDAAN, although in a standalone environment it also accepts SPSS input data sets. WesVar, PC-CARP and CENVAR are all stand-alone programs with their own unique organization, so familiarity with some other statistical package likely is not going to influence choice among these three. PC-CARP and CENVAR may appeal to those who must or prefer to operate in a DOS environment and may not appeal to those who prefer a Windows environment.

195. If cost is a major factor in software selection, then some packages are definitely more preferable. Epi-Info, although free, is limited in the analytical options for sample survey variance estimation but may be fine for basic analyses. CENVAR, also free, has more analytical options than Epi-Info but no design-based regression procedures. WesVar Version 2 is also free. IVEware is free but must run in conjunction with SAS. Low-cost but comprehensive sample survey software includes WesVar Version 4 and PC-CARP. STATA and standalone SUDAAN are moderate in cost, and SAS is expensive.

196. Another factor in choosing software may be the variance estimation method that is used. For example, if you are analysing a public release data set that includes BRR or jackknife replicate weights and no stratum/PSU identifier variables, then a software package that uses only Taylor series linearization will not be useful for you. Among the packages reviewed here, SUDAAN and IVEware offer both Taylor series linearization and replication methods, WesVar offers only replication procedures, and STATA, SAS, PC-CARP, Epi-Info and CENVAR offer only Taylor series linearization.

197. Finally, the choice of software depends upon the analyses you wish to conduct. All of the eight packages reviewed here perform basic and descriptive analyses. Among these eight packages, the ones that go beyond basic analyses include STATA, SUDAAN, WesVar, PC-CARP and SAS (with or without IVEware). Table XXI.2 summarizes and compares many attributes of these eight software packages.

198. The five software packages compared empirically in this chapter (SAS, SUDAAN, STATA, Epi-Info and WesVar) provide the same point estimates for all descriptive and analytical examples, an expected finding. All five software packages produce essentially the same estimated standard errors, whether BRR or Taylor series linearization was used. There are slight variations among the five packages on some of the confidence interval calculations; reasons for this were discussed earlier. Thus, there is no compelling reason to choose among these five packages based on the benchmarking analyses reported in this chapter.

199. The market for specialized sample survey software packages (with focus on variance estimation) may disappear in the future. The trend seems to be to include these capabilities in the standard statistical packages (for example, STATA, SAS and SPSS). Thus, in the future it may be easier for data analysts to obtain and use appropriate software for variance estimation with complex survey data.

Acknowledgements

Appreciation is extended to:

Michael S. Deming, MD, MPH, for providing the Burundi data set and its documentation, for careful reading of multiple manuscript drafts, and for valuable editing suggestions.

Kevin Sullivan, PhD, for instruction and valuable hints in navigating Epi-Info, for careful reading of multiple manuscript drafts, and for valuable editing suggestions.

Z. T. Daniels, MS, MBA, for formatting WesVar output and text tables and for locating Burundi population data on the web.

Graham Kalton, PhD and Ibrahim Yansaneh, PhD, for valuable organizational and editing suggestions.

James Chromy, PhD, for careful reading of manuscript drafts and valuable editing suggestions.

Several anonymous referees for careful reading of manuscript drafts and for valuable editing suggestions.

Paul Weiss, MS, for instruction and valuable hints in navigating WesVar.

Any errors in this chapter are the sole responsibility of the author.

References

Brogan, Donna (1998 and in press). Software for sample survey data: misuse of standard packages. Invited chapter in *Encyclopedia of Biostatistics*, Peter Armitage and Theodore Colton, eds.-in-chief. New York: John Wiley, vol. 5, pp. 4167-4174. Revised chapter in press for 2nd ed. *Encyclopedia of Biostatistics*, to be published in 2004.

Brogan, Donna, and others (1994). Increasing the accuracy of the expanded programme on immunization's cluster survey design. *Annals of Epidemiology*, vol. 4, No. 4, pp. 302-311.

Carlson, Barbara L. (1998). Software for sample survey data. In *Encyclopedia of Biostatistics*, vol. 5, Peter Armitage and Theodore Colton, eds.-in-chief, New York: John Wiley and Sons, pp. 4160-4167.

Cochran, William G. *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons.

Expanded Programme on Immunization (EPI) (1996). Estimating tetanus protection of women by serosurvey. *Weekly Epidemiological Record.* (World Health Organization), vol. 71, pp. 17-124.

Hansen, Morris H., William N. Hurwitz and William G. Madow (1953). *Sample Survey Methods and Theory*, vol. I, *Methods and Applications*. New York: John Wiley and Sons.

Judkins, D. (1990). Fay's method for variance estimation. *Journal of Official Statistics,* vol. 6, pp. 223-240.

Kish, Leslie (1965). *Survey Sampling*. New York: John Wiley and Sons.

_____, and M. R. Frankel (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B,* vol. 36, pp. 1-37.

Korn, Edward L., and Barry I. Graubard (1999). *Analysis of Health* Surveys. New York: John Wiley and Sons.

Krotki, Karol P. (1998). Sampling in developing countries. In *Encyclopedia of Biostatistics*, vol. 5, Peter Armitage and Theodore Colton, eds.-in-chief. New York: John Wiley and Sons, pp. 3939-3944.

Levy, Paul S., and Stanley Lemeshow (1999). *Sampling of Populations: Methods and Applications*, 3rd ed., New York: John Wiley and Sons.

Lohr, Sharon L. (1999). *Sampling: Design and Analysis*. Pacific Grove, California: Duxbury Press, Brooks/Cole Publishing.

Rao, J.N.K., and A. J. Scott (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*, vol. 76, pp. 221-230.

_____ (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics,* vol. 12, pp. 46-60.

Rust, K.F., and J.N.K. Rao (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, vol. 5, pp. 283-310.

Shah, Babubhai V. (1998). Linearization methods of variance estimation. In *Encyclopedia of Biostatistics*, vol. 3, Peter Armitage and Theodore Colton, eds.-in-chief, New York: John Wiley and Sons, pp. 2276-2279.

Som, R.K. (1995). *Practical Sampling Techniques*, 2nd ed. New York, Basel and Hong Kong: Marcel Dekker.

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

Software references:

CENVAR Variance Calculation System: IMPS Version 3.1: User's Guide, 1995. Bureau of the Census, United States Department of Commerce, Washington, D.C. Available from http://www.census.gov/ipc/www/imps/download.htm.

Epi-Info. Available from http://www.cdc.gov/epiinfo/ for the software and documentation.

IVEware. Available from http://www.isr.umich.edu/src/smp/ive/ for the software and documentation.

PC CARP (1986, 1989). *User's Manual,* Wayne Fuller and others, eds. Statistical Laboratory, Iowa State University, Ames, Iowa. Available from http://cssm.iastate.edu/software.

Research Triangle Institute (2001). *SUDAAN User's Manual, Release 8.0*. Research Triangle Park, North Carolina: Research Triangle Institute. Available from www.rti.org/sudaan.

SAS/STAT. Available from http://www.sas.com/technologies/analytics/statistics/stat/index.html for information on SAS/STAT software that includes procedures for sample survey data.

STATA. Available from http://www.stata.com for STATA, from http://www.stata.com/help.cgi?svy for a discussion of the svy commands in STATA, and from http://www.stata.com/bookstore/ for reference manual availability.

WesVar 4.2 User's Guide (2002). Rockville, Maryland: Westat. See also the web site http://www.westat.com/WesVar/about/.

Annex:

This chapter includes an Annex (English only) containing illustrative and comparative analyses of data from the Burundi Immunization Survey using five statistical software packages. The contents of the CD-ROM, including program codes and output for each of the software packages, may be downloaded directly from the UN Statistics Division website (http://unstats.un.org/unsd/hhsurveys/) or the CD-ROM may be made available upon request from the UN Statistics Division (statistics@un.org).