UNITED NATIONS SECRETARIAT **ESA/STAT/AC.93/6**
Statistics Division **03 November 2003**

English only

Expert Group Meeting to
Review the Draft Handbook on
Designing of Household Sample Surveys
3-5 December 2003

D R A F T

Estimation of sampling errors for complex survey data [*]

by

Ibrahim S. Yansaneh[**]

[*] This document is being issued without formal editing.

[**] The views expressed in this paper are those of the author and do not imply the expression of any opinion on the part of the United Nations Secretariat.

Table of contents

Abstract

This chapter provides a brief overview of the various methods used for estimating sampling errors for survey data. It is emphasized that standard statistical software packages underestimate the sampling errors, leading to wrong conclusions regarding the parameters of interest to the survey. This problem is solved by the use special statistical software packages that take full account of the complex nature of the design that generated the data being analyzed. Several of these software packages are described and compared. The preparation of data files for analysis is also discussed.

Key Words. Complex survey design; Sampling errors; Stratification; Clustering; Unequal probability sampling

Chapter Six: Estimation of sampling errors for complex survey data

6.1. Introduction

1. The analysis of data from household surveys in developing countries is frequently restricted to basic tabular analysis, with estimates of means and totals, without any indication of the precision or accuracy of these estimates. Even in national statistical offices with an extensive infrastructure in statistical data collection and processing, one often finds a surprising lack of expertise in detailed analysis of micro-level data. Many analysts are often surprised to learn, for instance, that the clustering of elements introduces correlations among the elements that reduce the precision of the estimates relative to the simple random samples they are accustomed to analyzing; or that the use of weights in the analysis generally inflates the sampling errors; or that the standard software packages they routinely use in their work do not account for these losses in precision appropriately.

2. One of the key measures of precision in sample surveys is the sampling error, an indicator of the variability introduced by choosing a sample instead of enumerating the whole population, assuming that the information collected in the survey is otherwise exactly correct. For any given survey, an estimator of this sampling error can be evaluated and used to indicate the accuracy of the estimates. For sample designs for household surveys, which often involve stratification, clustering, and unequal probability sampling, the forms of these estimators are often complex and very difficult to evaluate. The calculation of sampling error for household survey data requires procedures that take into account the complexity of the sample design that generated the data, and that employ appropriate computer software. This chapter provides a brief overview of methods of computing estimates of sampling error for household survey data, including an evaluation and comparison of some publicly available specialized software for sampling error estimation.

6.2. The pitfalls of using standard statistical software packages to analyze household survey data

3. The analytical objectives of well-designed household surveys have in recent times moved beyond basic summary tables of counts or totals of parameters of interest. Analysts are now also interested in hypothesis generation and testing or model building. For instance, instead of simply estimating the proportion of a population in poverty or with secondary or higher education, analysts now want to evaluate the impact of policies, or explore the way in which a key response variable, e.g. academic performance of a school-going child, or the poverty level of a household, is affected by factors such as region, socio-economic status, gender, and age. Answering these types of questions requires sophisticated analyses at the household or person level, in other words, micro-level analyses.

4. Appropriate analyses of household survey data require that sampling errors of estimates be computed in a manner that takes into account the complexity of the design that generated the

data. This includes stratification, clustering, unequal-probability sampling, non-response, and other adjustments of sample weights (see chapter 5 for details). Standard statistical software packages do not account for these complexities because they typically assume that the sample elements were selected from the population by simple random sampling. As demonstrated in chapter 5, point estimates of population parameters are impacted by the sample weights associated with each observation. These weights depend upon the selection probabilities and other survey design features such as stratification and clustering. Because they ignore the sample weights, standard packages yield biased point estimates. Doing a weighted analysis with these packages reduces the bias in the point estimates somewhat, but even then, the sampling errors of point estimates are often grossly underestimated because the variance estimation procedure typically does not take into account such other design features as stratification and clustering. This means that inferences drawn from such analyses would be misleading. For instance, differences between groups might be erroneously declared significant or hypotheses might be erroneously rejected. Wrong inferences from the analyses of household data could have significant implication for resource allocation and policy formulation at the national and regional levels.

5. We now use data from the Demographic and Health survey series to illustrate the fact that the use of standard statistical software packages can lead to biased point estimates, inappropriate standard errors and confidence intervals, and misleading tests of significance.

(Insert DHS example here)

6.3. Methods of sampling error estimation

6. In this section, we briefly describe some conventional methods for estimating sampling errors for estimates based on survey data. Interested readers can obtain more details from such references as Kish and Frankel (1974), Wolter (1985), or Lehtonen and Pahkinen (1995). Methods for the estimation of sampling errors for household survey data can be classified into two broad categories:

a. Taylor series linearization.
b. Replication.

7. Most estimates of interest in household surveys are non-linear. Some examples are the average body mass index of school-age children in a country, the proportion of income spent on housing costs in a given city, etc. In the linearization method, such non-linear estimates are linearized using a Taylor series expansion. This involves expressing the estimate in terms of a Taylor's series expansion, and then approximating the variance of the estimate by the variance of the first-order or linear part of the Taylor series expansion. This method requires the assumption that all higher-order terms are of negligible size. If this assumption is correct, then the variance approximation works well; otherwise, serious biases in the estimates may result. Note that, with the linearization approach to variance estimation, a separate formula for the linearized estimate must be developed for each type of estimator.

8. The replication approach refers to a class of methods that involve the taking of repeated subsamples, or *replicates*, from the data, re-computing the weighted survey estimate for each replicate, and the full sample, and then computing the variance as a function of the deviations of these replicate estimates from the full-sample estimate. For instance, suppose k replicates are created from a sample, each with estimates $\hat{q}_1, \hat{q}_2, ..., \hat{q}_k$ of a parameter q, and suppose based on the full sample is $\hat{q}_0$, then the replication-based estimate of the variance is given by

$$Var(\hat{q}) = c\sum_{r=1}^{k}(\hat{q}_r - \hat{q}_0)^2$$

where c is a constant, which depends on the estimation method. The most commonly used replication techniques are the *balanced repeated replication* (BRR) method and the *jackknife* method. Other techniques less commonly used for this purpose are bootstrapping, the random group method (see Wolter (1985), and various modifications of these methods. With balanced repeated replication, forming a replicate involves dividing each sampling stratum into two primary sampling units (PSUs), and randomly selecting one of the two PSUs in each stratum to represent the entire stratum. The jackknife method involves removing one stratum at a time to create each replicate.

9. The important practical difference between the two approaches is that a Taylor series approximation must be developed analytically for each statistic, while repeated replication uses the same basic estimation method regardless of the statistic being estimated. Replication techniques are computer-intensive, mainly because they require the computation of a set of replicate weights, which are the analysis weights, re-calculated for each of the replicates selected so that each replicate appropriately represents the same population as the full sample. Unlike the Taylor Series method, replication methods do not require the derivation of variance formulas for each estimate because the approximation is a function of the sample, not of the estimate. The two approaches do not produce identical estimates of sampling error, but empirical investigations (Kish and Frankel, 1974) have shown that for many statistics, the differences are negligible. In the next section, we discuss the features of a survey dataset required for appropriate data analysis by all computer software packages.

6.4. Preparation of data files for analysis

10. A common problem with survey data collected in developing countries is that they are not amenable to analysis beyond basic frequencies and tabulations. There are several reasons for this. First, there is usually very limited or no technical documentation of the sample designs for surveys. Second the data files often do not have the format, structure, and the requisite information that would allow any sophisticated analysis. Third, there is sometimes a lack of appropriate computer software and technical expertise.

11. In order for sample survey data to be analyzed appropriately, the associated database must contain all the information reflecting the sample selection process. In particular, the database should include appropriate labels for the sample design strata, primary sampling units, secondary sampling units, etc. Furthermore, sample weights should be developed for each

sampling unit in the data file. These weights should reflect the probability of selection of each sampling unit as well as compensate for survey non-response and other deficiencies in the sample. The sample weights and the labels for the design variables are required for the appropriate estimation of the variability of the survey estimates. As mentioned in chapter 5 and in the preceding sections of this chapter, sample weights are important not only for generating appropriate survey estimates, but also for the estimation of the sampling errors of those estimates. Therefore, it is essential that all information on weights be incorporated into the data files. In particular, whenever non-response, post-stratification, or other types of adjustments are made, the survey documentation must contain a description of these adjustment procedures.

6.5. Computer software for sampling error estimation

12. The two methods of sampling error estimation have been in use for a long time in developed countries, implemented by customized computer algorithms developed by government statistical agencies, academic institutions, and private survey organizations. Recent advances in computer technology have led to the development of several software packages for implementing these techniques. Many of these software packages are now available for public use on mainframe computers as well as personal computers. The software packages use one or the other of the two general approaches to sampling error estimation discussed in section 6.3. Most survey data analysis software packages produce the most widely used estimates, such as means, proportions, ratios, and linear regression coefficients. Some software packages also include approximations for a wide range of estimators, such as Cox proportional hazards and logistic regression coefficients.

13. In this section, we present a catalogue of some publicly available software for the estimation of sampling errors for household survey data. This is not an exhaustive list of all available programmes and packages. We restrict attention only to those packages that are currently available on personal computers, and only commercial or documented free-ware statistical software packages that are currently available for use by the general survey data analyst. Each software package is briefly reviewed. No attempt is made to provide the technical and computational procedures underlying the packages. Such details can be obtained from the websites of the packages, and other contact information provided below.

14. The seven packages reviewed here are CENVAR, Epi Info, PC CARP, SAS, STATA, SUDAAN, and WesVarPC. Most of these packages use Taylor series approximations to compute sampling error estimates. The repeated replication programs in the list offer many of the basic methods, except the bootstrap. We now provide a brief overview of each software package, including basic contact and cost information.

6.5.1. CENVAR

(…)

6.5.2. Epi Info

(…)

6.5.3. PC CARP

(…)

6.5.4. SAS

(…)

6.5.5. STATA

(…)

6.5.6. SUDAAN

(…)

6.5.7. WesVarPC

(…)

6.6. Comparison of the software packages

(…)

6.7. Concluding remarks

15. In this chapter, we have advocated the use of special computer software for the estimation of sampling errors for survey data. We have provided examples of situations in which serious errors are committed in the estimation of sampling errors when standard statistical software packages are used. In general, the use of standard statistical packages for household survey data analysis will understate the true variability of the survey estimates. These smaller estimates of standard error can lead to the drawing of invalid inferences from the survey data, for instance erroneously declaring significant differences between the means of two groups or incorrectly rejecting a null hypothesis.

16. We have also provided a catalogue of some publicly available software packages, along with basic contact information and an overview of their application. The lack of knowledge or expertise in sampling error estimation is one of the impediments to sophisticated analysis of data in developing countries. Many analysts are not aware of the need to use specialized software or, if aware, prefer not to do so because of the need to learn a new software package. For a more extensive review of these and other software packages, including computer code and output for some of the available software, see Brogan (2003).

17. Finally, it must be recognized that with rapid advances in technology, a lot of software packages either become obsolete or are improved beyond the specifications provided in this

review, in a relatively short time. Indeed it is possible that some of these specifications will be obsolete by the time this handbook is published. It is therefore important to remember that the most accurate information regarding the software packages should be obtained from their respective manuals at the time of use.

References and further reading

An, A and Watts, D. (2001). New SAS procedures for analysis of sample survey data, *SUGI paper No. 23*, SAS Institute Inc., Cary, NC [http://support.sas.com/rnd/app/papers/survey.pdf]

Binder DA (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review* **51**, 279-92

Brick JM, Broene P, James P and Severynse J (1996). *A User's Guide to WesVarPC*, Westat, Inc., Rockville, MD.

Brogan, D. (2003): *Sampling Error Estimation for Survey Data*, Technical Report on Surveys in Developing and Transition Countries, United Nations Statistics Division, May 2003.

Burt VL and SB Cohen (1984), "A Comparison of Alternative Variance Estimation Strategies for Complex Survey Data." *Proceedings of the American Statistical Association Survey Research Methods Section.*

Carlson BL, AE Johnson, and SB Cohen (1993), "An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data," *Journal of Official Statistics* 9(4), 795-814.

Cohen SB, JA Xanthapoulos, and GK Jones (1988), "An Evaluation of Statistical Software Procedures Approriate for the Regression Analysis of Complex Survey Data." *Journal of Official Statistics* 4,17-34.

Dippo CS, RE Fay, and DH Morganstein (1984), "Computing Variances from Complex Samples with Replicate Weights." *Proceedings of the American Statistical Association Survey Research Methods Section.*

Hansen MH, WN Hurwitz, and WG Madow (1953), *Sample Survey Methods and Theory, Volume I: Methods and Applications.* New York: Wiley (Section 10.16).

Hansen MH, WG Madow, and BJ Tepping (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," *Journal of the American Statistical Association* 78(384), 776-793.

Kish L and MR Frankel (1974), "Inference from Complex Samples," *Journal of the Royal Statistical Society* B(36), 1-37.

Landis JR, Lepkowski JM, Eklund SA, and Stehouwer SA (1982). A Statistical Methodology for Analyzing Data from a Complex Survey: the First National Health and Nutrition Examination Survey. *Vital and Health Statistics, 2(92),* DHEW, Washington, DC.

Lehtonen, R., and E. J. Pahkinen (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: Wiley.

Lepkowski JM, JA Bromberg, and JR Landis (1981), "A Program for the Analysis of Multivariate Categorical Data from Complex Sample Surveys." *Proceedings of the American Statistical Association Statistical Computing Section.*

Levy, Paul S. and Stanley Lemeshow (1999). Sampling of Populations: Methods and Applications. Third edition. John Wiley & Sons, New York.

Rust KF and Rao JNK (1996). “Variance Estimation for Complex Surveys Using Replication Techniques”, Statistical Methods in Medical Research, 5, 283-310.

Rust K (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics* 1(4), 381-397.

Shah BV, Barnwell BG and Bieler GS, (1996). *SUDAAN User's Manual: Release 7.0*, Research Triangle Institute, Research Triangle Park, NC.

Tepping BJ (1968), "Variance Estimation in Complex Surveys," *Proceedings of the American Statistical Association Social Statistics Section*, pp. 11-18.

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

Woodruff RS (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association* 66(334), 411-414.

SAS Institute, Inc. (1994), *SAS System for Windows, Release 6.10 Edition*. Cary, NC.

SPSS, Inc. (1988), *SPSS/PC+ V2.0 Base Manual.* Chicago, IL.