Bridging the Gap between the Theory and Practice of Analysis of Data from Complex Surveys - Some Statistics Canada Experiences

Georgia Roberts[1], Milorad Kovacevic[1], Owen Phillips[1], and Jane Gentleman[2]
[1]Data Analysis Resource Centre, Statistics Canada, Ottawa, Canada
[2]National Centre for Health Statistics, Washington, USA

Abstract

So as not to publish misleading results, subject matter analysts at Statistics Canada are urged to take account of the complexities of the survey design when doing analysis using data from Statistics Canada's surveys. While commercial software packages that incorporate methods for controlling for features of the sample design are becoming more readily available and more efficient to use, analysts without some background in survey theory still have difficulty in knowing how to proceed.

Statistics Canada has a small unit called the Data Analysis Resource Centre (DARC) whose purpose is to provide specialized services in analysis of statistical data. One of the major activities of DARC is the support of subject matter analysts who are using data from surveys with complex designs. This paper will present some of our experiences in DARC with assisting analysts in doing their research and some "tips and traps" that we have identified.

Because of the practice in many publications, due to space restrictions, of presenting descriptive estimates without their corresponding variance estimates, new analysts of survey data are frequently surprised that they require more data about the survey design than just the final weights in order to produce acceptable variance estimates. These analysts, who are generally secondary users of survey data rather than having been involved in the implementation of the survey, welcome assistance with possible approaches to accounting for the actual survey design and tips on different software packages that can implement the approach that best suits their needs. In order to communicate our advice on these topics, DARC has found it useful to develop illustrations of real analysis using real data from the specific surveys being analysed by a group of researchers - particularly the new Statistics Canada longitudinal surveys on health, youth, and labour and income dynamics. Each illustration makes use of many different software packages so that the traps and advantages of each can be pointed out. Through these illustrations, appropriate methods of variance estimation can be introduced and discussed.

DARC staff members have found that this direct provision of support to analysts is also an excellent forum for identifying methodological problems - such as ideosyncracies in a survey design that could be modified in the future, limitations in software, and analytical questions still requiring theoretical research.

1. Introduction

Subject matter analysts at Statistics Canada are urged to take account of the complexities of the survey design when doing analysis using data from Statistics Canada's surveys, so as not to publish misleading results,. While commercial software packages that incorporate methods to accommodate for the sample design are becoming more readily available and more efficient to use, analysts without some background in survey theory still have difficulty in knowing how to proceed. The Data Analysis Resource Centre (DARC) is a small unit at Statistics Canada that provides specialized services in analysis of statistical data. One of the major recent activities of DARC is the support of subject matter analysts who are using data from surveys with complex designs. This paper will present some of our experiences in DARC with assisting analysts in doing their research and some “tips and traps” that we have identified. In Section 3, the issue of what data an analyst requires for doing design-based analysis is briefly presented. This is followed, in Section 4, with a description of what features analysts are likely to want in the software tools that they use for design-based analysis. Section 5 explores two topics that have impact on how analysis is carried out with survey data - (I) doing analysis through scaling of weights only and (ii) effective degrees of freedom for variance estimation. The paper finishes, in Section 6, with some illustrations of analytical examples, including the use of different software packages.

2. What data are required for carrying out analysis

Because of the practice in many publications of presenting descriptive estimates without their corresponding variance estimates, new analysts of survey data are frequently surprised that they require more details about the survey design than just the final weights in order to produce accurate variance estimates. These analysts are generally secondary users of survey data rather than having been involved in the implementation of the survey.

Survey weights are needed in order to obtain reasonably unbiased estimates of population parameters, whether these be descriptive statistics or model parameters. The sampling weight can be thought of as the number of units in the population represented by a sample member. It is thus obvious to the analyst that these weights are essential for estimating a population total. And comparison of a histogram of the sample values of a continuous variable with an estimated probability mass function for the same variable, using the weights, generally convinces the analyst that the weights are needed for other estimates too, provided that there is some variability in the weights due to design features such as unequal selection probabilities.

The sampling weights do contain all the information necessary to construct point estimates. However, the weights alone do not give the extra information required to estimate variances, which are needed for inference. As an extreme example, for a design where each unit has equal probability of selection, the weights cannot reveal anything about the stratum memberships of the sample units; yet, with a stratified design, it is desirable to estimate the variance separately within each stratum. What additional information is required for variance estimation depends on the actual design of the survey and suitable approximations to that design. For many Statistics Canada surveys, which have complex multistage designs, the most common approximation is to treat the data as coming from a stratified design with sampling of psu’s with replacement at the first stage; thus, provided that it is not considered necessary to also take account of weight adjustments in the estimation of variance, knowledge of the weights and identification of stratum and psu membership of each sample unit is sufficient for variance estimation. Where it is considered necessary to account for weight adjustments, additional information about the particular adjustments done, such as the adjustment classes and benchmarking totals used, would also need to be obtained and software that can do the required adjustments must be employed.

3. Desirable features in software

Some desirable features of software packages for survey data analysis are described below, together with some of our observations on whether selected commercial packages that we have used at Statistics Canada actually have or lack these features. We restrict our comments to our experiences with SAS Version 6.12, SUDAAN Release 7.5.2 and WesVar Complex Samples 3.0. In discussing desirable software features, we are assuming that the analysts are not restricted by confidentiality as to what data are available to them.

For analysts to make the transition to doing design-based analysis, it is easier if the new software tool that he chooses be relatively *straightforward to use*, compared to whatever tools they are familiar with for doing data management or for carrying out analysis without accommodating for the design. It is also important that it be straightforward to transfer data from the usual package to the new one. At Statistics Canada, where SAS is the most prevalent package currently being used for data management and analysis, a new tool should be easy to “pick up” for SAS users. The structure of SUDAAN procedures looks very familiar to a SAS user. However, a major frustration for a SAS user beginning to use SUDAAN is the restriction to numeric variables, often requiring recoding of many variables before using SUDAAN.

It is desirable that a software package have *good documentation* containing explanations and a table of contents, and that the writing be understandable to the analyst who is not well-versed in sampling theory. Both SUDAAN and WesVar Complex Samples are attempting to bridge that gap.

It is desirable that one package be able to *accommodate a range of analyses* generally of interest to many analysts. At Statistics Canada, this means that the package should be able to (I) produce descriptive statistics (means, totals, proportions, quantiles) and standard errors of these quantities, (ii) do simple categorical analyses (test of independence, contingency table comparisons), and (iii) do model fitting (linear, logistic, proportional hazards) and model testing. Also desirable is that the package be able to handle some of the specialized analyses that are important to subgroups of analysts. Some such specialized areas at Statistics Canada are gross flows for economic quantities (for example, flows into and out of poverty), hierarchical modelling of education data, and standardized rates (for example, for prevalences of health conditions). A major difficulty that has to be pointed out to analysts wishing to use certain specialized analyses is that extension of some specialized methods to survey data has not yet been done; therefore, there is no possibility that a software package can properly accommodate them.

It is essential that a software package *correctly use survey weights* when producing design-based point estimates. WesVar Complex Samples and SUDAAN, which were created particularly to handle survey data, do produce design based estimates. In a wide range of situations, an analyst can produce the same estimates in SAS, through use of a WEIGHT statement, with the final survey weight identified as the weight variable.

It is also essential, for most analytical work, that the software package give standard error estimates that *adequately account for the survey design*. Related to this, the software package should also give useful test statistics for the most frequently used hypotheses (in addition to simple t or χ^2 statistics accompanying individual estimates) - for example, a test of independence in a 2-way contingency table and a test of nested hypotheses when fitting most popular models to data. A very desirable feature at Statistics Canada is that the software *accommodate different approaches* to design-based variance estimation, since different surveys have different design information available to the analyst, and analysts have different computer capacity and statistical background for doing analysis. Some surveys at Statistics Canada, may have just stratum and psu identifiers readily available to users (which, if with-replacement sampling of psu's is a reasonable approximation to the design, would allow a Taylor linearization approach or a replication approach - such as jackknife or bootstrap - but would not give enough information for accommodating for weight adjustments), while other surveys may have a file containing sets of jackknife weight variables or bootstrap weight variables, where these weight variables each derive from one jackknife or bootstrap sample, and resultant weight adjustments. (If a survey has previously-generated replicate weights, we would usually prefer to use them, due to increased speed of processing, and due to the fact that these weights could contain more weight adjustments that were incorporated into the final weights than a commercial software package is likely to be able to accommodate.) SUDAAN 7.5.2 can use either a Taylor linearization or replication method for variance estimation, but for its jackknife option, it must generate the jackknife weights rather than read them from an external file. WesVar Complex Samples restricts itself to replication methods and can make use of replication weights supplied from an external source. In the SAS procedures that we were using for analysis, design-based standard errors were not produced, since these procedures were not made for design-based analysis.

Analysts want to get on with their analysis. Thus, it is desirable that software accommodating for the design can *do the analysis in an "acceptable" length of time*, as compared to when the design is totally ignored. As well, it is essential that the software can *accommodate "large" data files* produced by some of the analytical surveys at Statistics Canada. The degree of "largeness" of the data file can be a function of the approach to variance estimation. As an example, if jackknifing is the method of variance estimation to be used, the program must be able to handle the number of psu's in the sample. We have found occasions where both SUDAAN and WesVar Complex Samples have not performed acceptably with respect to these two properties, but these occasions have been rare.

A desirable feature for software is that it produce *useful diagnostics* - particularly graphical diagnostics. At the moment, SUDAAN and WesVar Complex Samples do not have this feature.

For the more sophisticated analyst, it is desirable that the software *provide the full covariance matrix of a selection of estimates*, which may be used for generating test statistics not automatically produced by the software. We have encountered situations with both SUDAAN and WesVar Complex Samples where we have not been able to obtain a desired covariance matrix.

4. Some particular issues

4.1 Scaling the weights

For some analysts, it is very popular to scale the final weights to sum to the total sample size. These scaled weights are used for preparing descriptive estimates, fitting models and testing a variety of hypotheses regarding population or model parameters. Analysts very often use scaled weights combined with an assumption of simple random sampling (SRS) from an infinite population (as assumed in most traditional statistical packages) as a substitute for appropriate handling of the complexity coming from the sampling design.

It is generally true, however, that by scaling the weights, one obtains the same point estimates for ratio-type parameters as with unscaled weights; using only weights, scaled or unscaled, and assuming a simple random sampling, one obtains the correct point estimates of the parameters but generally underestimates the standard errors of the resulting estimates if other aspects of the survey design are ignored. Variance estimators of scale-invariant estimators are scale-invariant too. Non-ratio type statistics, on the other hand, estimate different parameters with scaled weights than with unscaled weights.

Some test statistics like the Pearson $\boldsymbol{X}^2$ or the likelihood ratio appear to take on reasonable values when calculated with scaled weights. However, for data coming from complex samples, it is not appropriate to compare these statistics to the critical values of a central χ^2 distribution, since these statistics have a different distribution due to the complexities of the survey design. A number of approaches have been proposed to account for the survey design when testing for model fit. (See Rao and Scott 1981, 1987; Thomas and Rao, 1987; Rao and Thomas, 1988; Thomas, Singh and Roberts, 1996). One approach consists of "correcting" the scaled-weight weighted $\boldsymbol{X}^2$ statistic so that the corrected statistic more closely approximates a central χ^2 distribution.Thus, relying only on the scaled weights when calculating the $\boldsymbol{X}^2$ test statistic is not sufficient when data come from a complex sample.

4.2 Effective number of degrees of freedom

When dealing with survey data, analysts are mostly occupied with appropriate estimation of the variances of the statistics of interest. Another important issue is usually less emphasized and to some extent neglected: calculation of the effective number of degrees of freedom (*df*). A common mistake is to say that because the sample size is large the number of *df* is very large. The importance of the number of *df* comes into the picture when confidence intervals are estimated or hypotheses are tested. Here we sketch a way of calculating the number of *df* based on Satterthwaite's (1946) formula:

Assuming that $\bar{y}_h$, $h = 1,...,L$ are independently and normally distributed, the effective number of degrees of freedom for the estimated variance of $\bar{y} = \sum W_h \bar{y}_h$ is approximately given by

$$\hat{df} = (\sum_h W_h s_h^2)^2 \ / \ \sum_h \left[W_h^2 s_h^4 / (n_h - 1) \right],$$

where $s_h^2 = \Sigma_i (y_{hi} - \bar{y}_h)^2 / (n_h - 1)$, $h = 1,...,L$. It is also known (Cochran, 1977) that $\hat{df}$ is always between the smallest of the $(n_h - 1)$'s and their sum $\Sigma n_h - L$. In the case of a multi-stage design, $y_{hi} = \Sigma w_{hij} y_{hij}$. Usually we assume that for, the entire survey, $\hat{df} = \Sigma n_h - L$. However, the number of *df* is influenced by several factors. Most importantly, it depends on the way the variance is estimated. If the variance is estimated using a delete-one -PSU jackknife method, the number of PSU's reduced for the number of strata is roughly the number of *df*. However, if a repeated half sample method of variance estimation is used, then the number is closer to the number of strata. In general, for consistent estimates of variances, $\hat{df}$ can be approximated by

$$\hat{df} \approx 2[var(\hat{y})]^2 / var[var(\hat{y})] .$$

We now discuss the importance of an accurate assessment of the number of degrees of freedom. A smaller number of *df* implies a larger value of the $1-\alpha/2$ percentile of the t-distribution and results in a wider confidence interval and more conservative tests. Also, the number of *df* measures the stability of a variance estimate: the smaller the number of *df* the more unstable the variance estimator. As an example, if we express the stability of the estimator by its relative variance, then

$$\frac{var[var(\hat{y})]}{[var(\hat{y})]^2} = \frac{2}{\hat{df}} .$$

For the acclaimed *df*=30 the coefficient of variation for the variance estimate is approximately 26%.

On the other hand, it is also important to note that if the bias of the variance estimate is small and the number of *df* is large (*df*>30) then the use of normal intervals instead of *t* intervals will give approximate $1-\alpha$ coverage, although the estimation of the variance is not necessarily very precise.

Usually loglinear modelling of survey data is based on the application of the generalized least squares method (Grizzle, Starmer and Koch (1969), and Koch, Freeman and Freeman (1975)). The assumption of availability of a consistent estimate of the covariance matrix of the estimated cell counts (or proportions) is essential and most often, the Wald statistic is used for testing. However, due to the low precision of covariance estimation for complex surveys, especially when the number of cells in the table is large, the Wald statistic may not perform adequately, resulting in a high rate of rejection under a null hypothesis. A good study of this phenomenon is given in Hidiroglou and Rao (1987).

5. Illustrations of some concepts and "tricks"

For the purpose of illustration, we will focus on data from two of Statistics Canada's flagship surveys, namely the Survey of Labour and Income Dynamics (SLID) and the National Population Health Survey (NPHS). These surveys are longitudinal household surveys, that is, they track individuals from selected households over a period of time in an effort to gain insight into changes in labour market activity and income, and the state of health, respectively, in Canada's provinces and territories.

SLID's goal is to gather and provide data on labour market activity and family income stability by tracking households over a period of six years. A panel of SLID respondents consists of approximately 15,000 households, and a new panel is implemented every three years, in order to provide reliable cross-sectional estimates. Individuals aged 15 and over in a selected household are given a preliminary interview in January of the year in which the panel was introduced, and labour and income information is collected on these individuals in subsequent years.

The NPHS covers several aspects related to the health of Canadians. Demographic information and general information pertaining to health are collected about each member of a selected household, and more detailed health questions are asked of a randomly selected individual (the longitudinal respondent) within the household. Follow-up interviews are conducted every two years, and an individual may remain in-sample for up to twenty years. In 1994/1995, the survey had 17,626 longitudinal respondents.

As previously mentioned, the survey design can have a strong impact on the variances of estimates and should be reflected in how variances are calculated. Theory is often well ahead of practice. The area of analysis of survey data is no exception. While the theory has made great strides since the early 1980's, it may not be possible to exactly implement the theory for your particular problem with the software that you have at hand. However, we have found that with knowledge of what you are trying to accomplish and of what your software can do, you may be able to find a workable approximation. In this section, again, we will make reference to our experiences with SAS Version 6.12, SUDAAN Release 7.5.2 and WesVar Complex Samples 3.0.

For many surveys, the design aspect that has the greatest impact on both the estimates of interest and on their variances is unequal selection probabilities. These unequal probabilities are reflected in the unequal weights. A "trick" which many SAS users employ in order to produce estimates and their variances that are influenced by this aspect of the survey design is to include a WEIGHT statement in the procedures being used and to specify as the WEIGHT variable the final survey weight which has been standardized to have an average value of 1 in the subsample of the domain being studied. This "trick" is taking account of the inequality in the selection probabilities and is also recognizing the fact that many SAS procedures treat the sum of weights, rather than the number of observations, as the sample size. The scaling has no impact on variance estimates for ratio-type estimates. This "trick", however, can also be a "trap", such as in the following instances. First of all, it does not account for other design aspects that influence variances, such as stratification and clustering. Second, it is not a possibility in those SAS procedures where the weight information can only be supplied through a FREQ statement rather than a WEIGHT statement, since this may result in the truncation of standardized weight values to integers and deletion of observations with standardized weight values less than 1. Third, the scaling will yield incorrect estimates of totals. As well, even in SAS procedures where a WEIGHT statement is allowed, the weight information may not be used for all estimates, such as in the estimation of quantiles in PROC UNIVARIATE.

To illustrate some of these points, we look at an example from the NPHS. The data are from those 1996 longitudinal respondents who were, at one time, smokers. The size of the population of former smokers was estimated, along with the mean and its standard error and the median for number of years since having quit smoking (QUITYRS). The estimates were produced, by sex, using four approaches: SAS with no WEIGHT statement; SAS with the final survey weight as the variable in a WEIGHT statement ; SAS with a scaled final weight as the variable in a WEIGHT statement; and, SUDAAN using 500 bootstrap weight variables. The results are given below, with the SUDAAN values being considered as the standard for comparison since the unequal weighting, stratification and clustering of the NPHS design have all been appropriately accounted for by this software.

QUITYRS (Number of years since having quit smoking)
Descriptive statistics, by sex
Population: former smokers

		Method*			
		SAS: no weights	**SAS: final weights**	**SAS: scaled final weights**	**SUDAAN: bootstrap**
Both Sexes	**Sample Size**	2,903	2,903	2,903	2,903
	Est. Popln Size	2,903	4,973,066	2,903	4,973,066
	Mean	14.11	13.70	13.70	13.70
	SE(Mean)	0.21	0.20	0.20	0.23
	Median	12	12	12	11.68
Males	**Sample Size**	1,532	1,532	1,532	1,532
	Est. Popln Size	1,532	2,814,302	1,532	2,814,302
	Mean	15.01	14.28	14.28	14.28
	SE(Mean)	0.29	0.28	0.28	0.33
	Median	13	13	13	12.88
Females	**Sample Size**	1,371	1,371	1,371	1,371
	Est. Popln Size	1,371	2,158,764	1,371	2,158,764
	Mean	13.10	12.94	12.94	12.94
	SE(Mean)	0.30	0.29	0.29	0.34
	Median	10	10	10	10.49

*SAS results were obtained with PROC SUMMARY and PROC UNIVARIATE

The use of weights in SAS - scaled or unscaled - produces the same mean estimates as SUDAAN. An unscaled weight, however, is required to obtain the appropriate estimate of the size of the population of former smokers. The approach of using a WEIGHT statement in SAS provides standard error estimates for the means that are of a similar order of magnitude as SUDAAN. While the formulae used by SAS to produce standard errors, when a WEIGHT statement is present, are generally not the formulae that you would ideally wish to compute, the results are frequently sufficiently in the desired range to be useful when still in the exploratory stage of your analysis. The bootstrap method of standard error estimation accounts for aspects of the design not already contained in the weighting. As seen for the SE(mean) values in the above table, by comparing the results from the two WEIGHT approaches with SAS to the SUDAAN results, the consequence of ignoring these additional design aspects, particularly the clustering, is often the underestimation of standard errors. Note that the median estimates are the same for all of the three approaches using SAS because PROC UNIVARIATE does not make use of the weight variable when calculating percentiles.

A “trick” for using design-based software to approximate a method for which it was not intended came to light when we were wanting to use data from SLID and a jackknife method of variance estimation with SUDAAN. SUDAAN is a software package specifically for the analysis of survey data. To properly carry out jackknifing, the survey weights should undergo the same weight adjustments each time a psu is dropped as when the final weights were created for the full sample. SLID final weights contain many adjustments, but SUDAAN (Release 7.5.2), which currently must create the jackknife weights if the “JACKKNIFE” variance estimation method is chosen, will only carry out a final benchmarking each time a psu is dropped. However, we had available to us previously-generated SLID jackknife weights containing all weight adjustments; as well, we knew that SUDAAN will allow the input of externally-generated weights for its "BRR" method of weight adjustment. The "trick" was to notice the similarity in the forms of the formulae for jackknife and BRR variance estimates when the survey design has two psu's per stratum, and to notice that 206 of the 221 strata in SLID contain two psu's. This then meant that a multiplication of the variance provided by SUDAAN by a factor of 458/2 (where 458 is the number of SLID psu's sampled) would provide an estimate approximately equal to what was desired.

As an illustration of this "trick",consider a logistic regression using SLID data. Various socio-economic and demographic variables, such as age, family composition and education, were used to predict the probability of moving out of the state of low earnings between 1993 and 1995. Standard errors for twenty-six regression coefficients were calculated using the approximation to the jackknife described above, and again using the correct form of the jackknife formula (as calculated using WesVar Complex Samples 3.0). Absolute relative differences in the standard error estimates produced by the two approaches were calculated in order to compare them. These are summarized in the table below.

	Frequency Distribution of Absolute Relative Difference (ARD)			
	ARD < 0.5%	0.5% < ARD < 1.0%	1.0% < ARD < 1.5%	1.5%< ARD
Frequency	15	7	3	1

Of the standard errors obtained for the twenty-six regression coefficients using the SUDAAN trick, only four differed by more than 1.0% from the value obtained using the correct form of the jackknife in WesVar and, among those four, no difference exceeded 1.7%. Hence, it would seem that this "trick" works well. In contrast, by using SAS with final weights as the WEIGHT variable, the average absolute relative difference was 27.64%.

Such tricks as those shown above have been handy in allowing analysts to use software packages currently at hand to carry out the analyses that they wish. Yet, through our ongoing direct provision of support to analysts, we should be able to identify such methodological difficulties which could lead to corrections of such software limitations.

References:

Cochran, W.G. (1977). *Sampling techniques*. John Wiley and Sons Inc. 3rd edition.

Grizzle, J.E., C.F. Starmer and G.G. Koch (1969). Analysis of Categorical Data by Linear Models. *Biometrics* 25, 489-504.

Hidiroglou, M.A. and J.N.K. Rao (1987). Chi-Squared Tests with Categorical Data from Complex Surveys. Part I. *Journal of Official Statistics*, 3, 117-132.

Koch, G.G., D.H. Freeman and J.L. Freeman, (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. *International Statistical Review*, 43, 59-78

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*, 76, 221-230.

Rao, J.N.K. and A.J. Scott (1984). On Chi-Squared Tests for Multiway Contingency Tables With Cell Proportions Estimated From Survey Data. *The Annals of Statistics*, 12, 46-60

Rao, J.N.K. and A.J. Scott (1987). On Simple Adjustments to Chi-Square Tests With Sample Survey Data. *The Annals of Statistics*, 15, 385-397.

Rao, J.N.K. and D.R. Thomas (1987). Chi Squared Tests for Contingency Tables. In *Analysis of Complex Surveys*, (Eds. Skinner, C.J., D.Holt, and T.M.F.Smith) John Wiley and Sons Ltd. Chichester

Rao, J.N.K. and D.R. Thomas (1988). The Analysis of Cross-Classified Categorical Data From Complex Sample Surveys. *Sociological Methodology*, 18, 213-270.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114

Thomas, D.R.,A.C. Singh, and G.R. Roberts (1996). Tests of Independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review* 64: 295-311.