Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(2) 2010 129-135

Role of Weights in Descriptive and Analytical Inferences from Survey Data: An Overview

J.N.K. Rao[1*], M. Hidiroglou[2], W. Yung[2] and M. Kovacevic[2]
[1]*Carleton University, Ottawa, ON, Canada*
[2]*Statistics Canada, Ottawa, Canada*

Received 06 May 2010; Accepted 10 May 2010

SUMMARY

Statistical agencies generally collect data from samples drawn from well defined finite populations and using complex sampling procedures that may include stratification, clustering and multi-stage sampling and using unequal probabilities of selection. Sample design weights, defined from the sampling procedures, are often adjusted to account for non-responding units and to calibrate to known population totals of auxiliary variables. Once adjusted, these 'final' weights are included on the survey datasets. There has been some discussion on the necessity of using these weights in the estimation of descriptive statistics and to perform analysis of data from these surveys. In this paper, we discuss the role of weights in descriptive and analytical inference.

Keywords: Calibration, Design Weights, Re-sampling methods, Multi-level models, Survey data analysis.

1. USE OF WEIGHTS FOR DESCRIPTIVE STATISTICS

The development of sample survey theory progressed more or less inductively. Strategies (design and estimation) that appeared reasonable were entertained and relative properties were carefully studied by analytical and/or empirical methods, mainly through comparisons of mean squared errors, and sometimes also by comparing anticipated mean squared errors or variances under plausible super-population models. Unbiased estimation under a given design was not insisted upon because it "often results in much larger mean squared error than necessary" (Hansen *et al.* 1983). Instead, design consistency was deemed necessary for large samples in the sense that the estimator approaches the population value as the sample size increases.

For a given probability sampling design, Narain (1951) and Horvitz and Thompson (1952) used the "design weights" $d_i = \pi_i^{-1}$ (>0) to construct a design-unbiased estimator

$$\hat{Y}_{NHT} = \sum_{i \in s} d_i y_i$$

of a population total Y, where s denotes the sample and π_i is the inclusion probability for unit i in the population ($i = 1, \ldots, N$). More importantly, the design weights d_i ensure design consistency of $\hat{Y}_{NHT}$ in large samples. Stratified random sampling based on size stratification and near optimal sample allocation is commonly used to handle highly skewed populations such as business survey populations. As a result, the design weights d_i differ across strata and the design-based inferences associated with $\hat{Y}_{NHT}$ are asymptotically valid. On the other hand, model-based methods that ignore the above design weights could lead to erroneous inference in large samples, even under minor model misspecification that cannot be detected easily, as demonstrated by Hansen *et al*. (1983).

*Corresponding author : J.N.K. Rao
E-mail address : jrao34@rogers.com

The Narain-Horvitz-Thompson (NHT) estimator is highly efficient under the above stratified design. It is also efficient when the probabilities π_i are proportional to size measures x_i that are closely related to the variable of interest, y_i. In fact, NHT proposed such designs, called probability proportional to size (PPS) designs. On the other hand, the NHT estimator with π_i proportional to x_i can be very inefficient in surveys with multiple characteristics when a characteristic y is unrelated or weakly related to the size measure x (such as poultry count y and farm size x in a farm survey). Rao (1966) proposed efficient alternative estimators for such cases that ignore the NHT weights. Basu (1971) constructed a "bad" design with π_i unrelated to y_i and then demonstrated that the associated NHT estimator leads to absurd estimates.

In practice, it is often desirable to use a common weight for all variables y. Beaumont (2008) addressed this issue and proposed the use of a "smoothed" common weight that attempts to maintain efficiency across specified variables. The smoothed NHT weights are obtained by modeling the design weights d_i, $i \in s$, using the observed values y_i, $i \in s$ associated with a selected set of characteristics y. Further research in this direction would be useful.

In the field of social statistics, stratified multi-stage sampling is commonly used. Hansen and Hurwitz (1943) developed the basic theory of stratified two-stage sampling with one primary sampling unit (PSU) within each stratum drawn by PPS sampling and then sub-sampled at a rate that ensures self-weighting (equal overall probabilities of selection) within strata. This approach provides approximately equal interviewer work loads which is desirable in terms of field operations. It also leads to significant variance reduction by controlling the variability arising from unequal PSU sizes without actually stratifying by size and thus allowing stratification on other variables to reduce the variance. Given the equal overall probabilities of selection, and hence the equivalence of the design weights d_i, one can ignore the design for point estimation but it cannot be ignored during variance estimation. Variance estimation must take into account the sample design and thus the weights at each stage of sampling play an important role. For instance, in the case of a self-weighting stratified multi-stage design the overall probabilities of selection are defined as a product of the probabilities of selection at the different stages of the design. Although the overall probabilities are equal, the selection probabilities at each stage are almost certainly different. The selection probabilities at the different stages are needed to produce valid variance estimators.

2. WEIGHT ADJUSTMENTS

Design weights are commonly adjusted for many different reasons such as compensating for unit nonresponse or for calibrating to known totals of auxiliary variables. In what follows, we assume that nonresponse is not present and concentrate on weight adjustments brought on by calibrating to known totals. Calibration weights $w_i(s)$ that ensure consistency with user-specified auxiliary totals $\mathbf{X}$ are obtained by adjusting the design weights $d_i = \pi_i^{-1}$ to satisfy the benchmark constraints $\sum_{i \in s} w_i(s)\mathbf{x}_i = \mathbf{X}$. Estimators that use calibration weights are called calibration estimators and they use a single set of weights $\{w_i(s)\}$ for all the variables of interest, y. We note that the model assisted Generalized Regression (GREG) (Särndal *et al*. 1992) estimator is a calibration estimator, but a calibration estimator may not be model-assisted in the sense that it could be model-biased under a working model unless the x-variables in the model exactly match the variables corresponding to the user-specified totals. For example, suppose the working model suggested by the data is a quadratic in a scalar variable x while the user-specified total is only its total $\mathbf{X}$. The resulting calibration estimator can perform poorly even in fairly large samples unlike the model-assisted GREG estimator based on the working quadratic model that requires the population total of the quadratic variables x_i^2 in addition to $\mathbf{X}$ (Rao *et al*. 2003).

Unified approaches to calibration, based on minimizing a suitable distance measure between calibration weights and design weights subject to benchmark constraints, have attracted the attention of users due to their ability to accommodate arbitrary number of user-specified benchmark constraints, for example, calibration to the marginal counts of several post-stratification variables. For instance, a chi-squared distance measure leads to a GREG estimator. However, the resulting calibration weights may not satisfy desired range restrictions, for example some weights may be negative or too large especially when the number of constraints is large and the variability of the design weights is large. Huang and Fuller (1978) proposed a scaled modified chi-squared distance measure and obtained the calibration weights through an iterative solution that satisfies the benchmark constraints at each iteration. However, a solution that satisfies benchmark constraints and range restrictions may not exist, so sometimes there must be a trade off between satisfying the benchmark constraints or allowing the range restrictions to be relaxed. Alternative methods propose to change the distance function (Deville and Särndal 1992) or drop some of the benchmark constraints (Bankier *et al*. 1992). Skinner and Nascimento Silva (1997) considered alternative approaches to selecting benchmark variables based on the minimization of the estimator of the variance of the regression estimator. Rao and Singh (1997, 2009) proposed a "ridge shrinkage" iterative method that ensures convergence for a specified number of iterations by using a built-in tolerance specification to relax some benchmark constraints while satisfying range restrictions; see also Chen *et al*. (2002) and Beaumont and Bocci (2008) for similar ridge calibration methods.

3. VARIANCE ESTIMATION

In survey sampling, variance estimation is typically performed using either Taylor linearization or, more recently, re-sampling methods. In both methods, the design or adjusted weights play an important role. Demnati and Rao (2004) derived Taylor linearization variance estimators for a general class of calibration estimators with weights $w_i(s) = d_i F(\mathbf{x}_i^T \hat{\lambda})$ where the LaGrange multiplier $\hat{\lambda}$ is determined by solving the calibration constraints. The choice $F(a) = 1 + a$ leads to GREG weights $w_i(s) = d_i g_i(s)$, where

$$g_i(s) = 1 + (\mathbf{X} - \hat{\mathbf{X}}_{NHT})^T \left(\sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k^T\right)^{-1} \mathbf{x}_k$$

Taylor linearization variance estimators are obtained by 'operating' on the weights d_i by taking derivatives of the estimators with respect to these weights. Demnati-Rao approach suffers from the same disadvantage as the other linearization type variance estimators in that analytical work is needed for each and every new estimator. Re-sampling methods, on the other hand, remove the human labor and replace it with computer labor. They play a vital role in developing methods that take account of survey design in the analysis of data. All one needs is a data file containing the observed data, the adjusted survey weights and the corresponding adjusted weights for each pseudo-replicate generated by the re-sampling method. Software packages that take account of survey weights in the point estimation of parameters of interest can then be used to calculate correct estimators and standard errors. More details on two commonly used re-sampling methods, the bootstrap and the jackknife, will be discussed in the following section.

4. ANALYSIS OF SURVEY DATA

Standard methods of data analysis are generally based on the assumption of simple random sampling, although some software packages do take account of survey weights and provide correct point estimates. However, application of standard methods to survey data, ignoring the design effect due to clustering and unequal probabilities of selection, can lead to erroneous inferences even for large samples. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated, type I error rates of tests of hypotheses can be much bigger than the nominal levels, and standard model diagnostics, such as residual analysis to detect model deviations, are also affected. Rapid progress has been made over the past 20 years or so in developing suitable methods. In particular, re-sampling methods have allowed users of survey data to perform complex analyses themselves very easily using standard software packages.

Re-sampling methods in the context of large-scale surveys using stratified multi-stage designs have been studied extensively. For inference purposes, sample PSUs are treated as if drawn with replacement within strata. This leads to over-estimation of variances but it is small if the overall PSU sampling fraction is negligible. Let $\hat{\theta}$ be the survey-weighted estimator of a "census" parameter of interest, θ_c, computed from the final weights $w_i(s) = w_i$ and let the corresponding weights for each pseudo-replicate r generated by the re-sampling method be denoted by $w_i^{(r)}$. The estimator based on the pseudo-replicate weights $w_i^{(r)}$ is denoted as $\hat{\theta}^{(r)}$ for each $r = 1, ..., R$.

In the above discussion, we let $\hat{\theta}$ denote the estimator of a "census" parameter, θ_c. The census parameter θ_c is often motivated by an underlying super-population model and the census is regarded as a sample generated by the model, leading to census estimating equations whose solution is θ_c. The census estimating functions $U_c(\theta)$ are simply population totals of functions $u_i(\theta)$ with zero expectation under the assumed model, and the census estimating equations are given by $U_c(\theta) = \sum_{i \in A} u_i(\theta) = 0$ (Godambe and Thompson 1986) where $\sum_{i \in A}$ denotes summation over all units in the finite population A. Kish and Frankel (1974) argued that the census parameter makes sense even if the model is not correctly specified. For example, in the case of linear regression, the census regression coefficient could explain how much of the relationship between the response variable and the independent variables is accounted by a linear regression model. Noting that the census estimating functions are simply population totals, survey weighted estimators of the census estimating functions, $\hat{U}(\theta) = \sum_{i \in s} w_i u_i(\theta)$ from the full sample and $\hat{U}^{(r)}(\theta)$ from each pseudo-replicate, are obtained. By solving the corresponding estimating equations $\hat{U}(\theta) = 0$ and $\hat{U}^{(r)}(\theta) = 0$, we obtain the estimators $\hat{\theta}$ and $\hat{\theta}^{(r)}$ respectively.

A re-sampling variance estimator of $\hat{\theta}$ is of the form

$$v(\hat{\boldsymbol{\theta}}) = \sum_{r=1}^{R} c_r (\hat{\boldsymbol{\theta}}^{(r)} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(r)} - \hat{\boldsymbol{\theta}})^T$$

for specified coefficients c_r determined by the re-sampling method. Commonly used re-sampling methods include (a) delete-a-cluster (delete-PSU) jackknife and (b) the Rao and Wu (1988) bootstrap. Jackknife pseudo-replicates are obtained by deleting each sample cluster $r = (hj)$ in turn, leading to jackknife design weights $d_i^{(r)}$ taking the value 0 if the sample unit i is in the deleted cluster, $d_i n_h / (n_h - 1)$ if i is not in the deleted cluster but in the same stratum h, and unchanged if i is in a different stratum. The jackknife design weights are then adjusted for unit non-response and calibration, leading to the final jackknife weights $w_i^{(r)}$. The jackknife variance estimator is given as above with $c_r = (n_h - 1) / n_h$ for $r = (hj)$. The delete-a-cluster jackknife method has two possible disadvantages: (1) When the total number of sampled PSUs, $n = \sum_h n_h$, is very large, R is also very large because $R = n$. (2) It is not known if the delete-a-cluster jackknife variance estimator is design-consistent in the case of non-smooth estimators $\hat{\theta}$, for example the survey-weighted estimator of the median. For simple random sampling, the jackknife is known to be inconsistent for the median or other quantiles.

The Rao-Wu bootstrap is valid for arbitrary n_h (≥ 2) and it can also handle non-smooth $\hat{\theta}$. Each bootstrap replicate is constructed by drawing a simple random sample of PSUs of size $n_h - 1$ from the n_h sample clusters, independently across the strata. The bootstrap design weights are given by

$d_i^{(r)} = [n_h/(n_h-1)]\, m_{hi}^{(r)} d_i$ if i is in stratum h and replicate r, where $m_{hi}^{(r)}$ is the number of times sampled PSU (hi) is selected, $\sum_i m_{hi}^{(r)} = n_h - 1$. The weights $d_i^{(r)}$ are then adjusted for unit non-response and calibration to get the final bootstrap weights and the estimator $\hat{\theta}^{(r)}$. Typically, $R = 500$ bootstrap replicates are used in the bootstrap variance estimator. Several surveys at Statistics Canada have adopted the bootstrap method for variance estimation because of the flexibility in the choice of R and wider applicability.

Note that the resampling variance estimators are designed to estimate the variance of $\hat{\theta}$ as an estimator of the census parameters θ_c but not of the model parameters θ. Under certain conditions, the difference can be ignored but in general we have a two-phase sampling situation, where the census is the first phase sample from the super-population and the sample is a probability sample from the census population. For the census parameter θ_c Taylor linearization methods provide asymptotically valid variance estimators for general sampling designs, unlike re-sampling methods, but they require a separate formula for each estimator $\hat{\theta}$. Binder (1983), Rao *et al*. (2002) and Demnati and Rao (2004) have provided unified linearization variance formulae for estimators defined as solutions to estimating equations. Demnati and Rao (2010) extended the Demnati-Rao approach to obtain linearization variance estimators for model parameters, θ.

Pfeffermann (1993) discussed the role of design weights in the analysis of survey data. If the population model holds for the sample (i.e., if there is no sample selection bias), then model-based unweighted estimators will be more efficient than the weighted estimators and lead to valid inferences, especially for data with smaller sample sizes and larger variation in the weights. However, for typical data from large-scale surveys, the survey design is informative and the population model may not hold for the sample. As a result, the model-based estimators can be seriously biased and inferences can be erroneous. Pfeffermann and his colleagues initiated a new approach to inference under informative sampling; see Pfeffermann and Sverchkov (2003) for recent developments. This approach seems to provide more efficient inferences compared to the survey weighted approach, and it certainly deserves the attention of users of survey data. However, much work remains to be done, especially in handling data based on multi-stage sampling. Excellent accounts of methods for analysis of complex survey data are given in Skinner *et al*. (1989), Chambers and Skinner (2003) and Lehtonen and Pahkinen (2004).

5. SURVEY WEIGHTS IN MULTI-LEVEL MODELS

If some coefficients in a regression model are assumed to vary randomly over the clusters, we arrive at models with random effects. A typical model of this kind is a two-level model where individuals are nested within clusters and the scientific interest is to study the regression relationships at different levels simultaneously including the cross-level interactions (Snijders and Bosker, 1999, page 9). In multilevel modeling the parameters of interest include, in addition to fixed-effects, the variance components which define the distribution of random effects.

In the case of two-level models there is a need to use survey weights at both levels since the sample design may be informative at any level (Pfeffermann *et al*. 1998). Thus the use of weights at only one level, typically at the level of ultimate unit, is not sufficient to compensate for possible biasing of sample design. This implies a basic condition for multi-level modeling of survey data: to have survey weights available at each level.

Survey datasets are commonly produced for single-level analyses, thus having survey weights only at one level. Generally, a multi-level modeling of such data would not be possible. In some cases, however, weights at one of the levels can be assumed to be equal to 1, and together with available weights, they would provide sufficient weighting information. A good example of such analysis is a simple growth curve analysis where repeated measurements nested within an individual have weight of one (Skinner and Holmes 2003). Alternatively, one can approximate unavailable weights using available information on sample designs at different levels, the cluster membership of ultimate sampling units, and the available weights (Kovacevic and Rai 2003).

Several methods are available for estimation of parameters of multi-level models from survey data: Probability Weighted Iterative Generalized Least Squares (Pfeffermann *et al*. 1998), and different flavors of Pseudo Maximum Likelihood Method (Kovacevic and Rai 2003, Asparouhov 2004, Grilli and Pratesi 2004, Rabe-Hesketh and Skrondal 2006). All of these methods provide approximately unbiased estimators of the parameters when the sample sizes are 'large' for all nested levels for which weighting is applied. This means that in the two level case, when weighting is applied at both levels, we need large number of clusters as well as large within-cluster samples.

ACKNOWLEDGEMENTS

This paper was presented as an invited paper at the International Statistical Institute 2009 meetings in Durban, South Africa. J.N.K. Rao wishes to acknowledge the Natural Sciences and Engineering Research Council of Canada for supporting his research. J.N.K. Rao also wishes to acknowledge Dr. G.R. Seth's generosity, advice and mentorship while Rao was working in New Delhi, 1963-64. A joint research paper on the comparison of simple random sampling with and without replacement also appeared in *Sankhya*, **A26**, 1964, 85-86.

REFERENCES

Asparouhov, T. (2004).Weighting for unequal probability of selection in multilevel modeling, Mplus Web Notes No. 8 available from http://www.statmodel.com/

Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Methodology Branch Working Paper, Census Operations Section, Social Survey Methods Division, Statistics Canada, Ottawa, Canada.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. In: *Foundations of Statistical Inference* (Eds. V.P. Godambe and D.A. Sprott). Holt, Rinehart and Winston, Toronto, 203-242.

Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, **95**, 539-553.

Beaumont, J.-F. and Bocci, C. (2008). Another look at ridge calibration. *Metron*, **66**, 5-20.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Internal. Statist. Rev.*, **51**, 279-292.

Chambers, R.L. and Skinner, C.J. (Eds) (2003). *Analysis of Survey Data*. Wiley, Chichester.

Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators in surveys. *Biometrika*, **89**, 230-237.

Demnati, A. and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17-26.

Demnati, A. and Rao, J.N.K. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology.* In press.

Deville, J.-C. and Särndal, C.E. (1992). Calibration estimation in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376-382.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *Internal. Statist. Rev.*, **54**, 127-138.

Grilli, L. and Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology*, **30**, 93-103.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.*, **14**, 333-362.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *J. Amer. Statist. Assoc.*, **78**, 776-793.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663-685.

Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section*, *Amer. Statist. Assoc.*, 300-305.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples. *J. Roy. Statist. Soc.*, **B36,** 1-37.

Kovacevic, M.S., and Rai, S. (2003) A pseudo-maximum likelihood approach to multilevel modelling of survey data. *Comm. Statist.-Theory Methods*, **32**, 103-121.

Lehtonen, R. and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Wiley, Chichester.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *J. Ind. Soc. Agril. Statist.*, **3**, 169-174.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Internal. Statist. Rev.,* **61**, 317-337.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. In: *Analysis of Survey Data* (Eds. R.L. Chambers and C.J. Skinner), Wiley, Chichester, 175-195.

Pfeffermann, D., Skinner, C.J., Holmes D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J. Roy. Statist. Soc.,* **B60**, 123-140.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *J. Roy. Statist. Soc.,* **A169**, 805-827.

Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, **A28**, 47-60.

Rao, J.N.K and Singh, A.C. (1997). A ridge shrinkage method for range-restricted weight calibration in survey sampling. *American Statistical Association 1997 Proceedings of the Section on Survey Research Methods*, 57-65.

Rao, J.N.K. and Singh, A.C. (2009). Range restricted calibration for survey data using ridge regression. *Pak. J. Statist.,* **25**, 371-384.

Rao, J.N.K., Jocelyn, W. and Hidiroglou, M.A. (2003). Confidence interval coverage properties for regression estimators in uni-phase and two-phase sampling. *J. Official Statist.*, **19**, 17-30.

Rao, J.N.K., Yung, W. and Hidiroglou, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, **A64**, 364-378.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.*, **83**, 231-241.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Skinner, C.J. and Nascimento Silva, P.L.D. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, **23**, 23-32.

Skinner, C.J. and Holmes, D.J. (2003). Random effects models for longitudinal survey data. In: *Analysis of Survey Data* (Eds. Chambers, R.L. and Skinner, C.J.), Wiley, Chichester, 205-219.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. John Wiley and Sons, Inc, New York.

Snijders, T.A.B. and Bosker, R.J. (1999). *Multilevel Analysis*. Sage Publications, London.