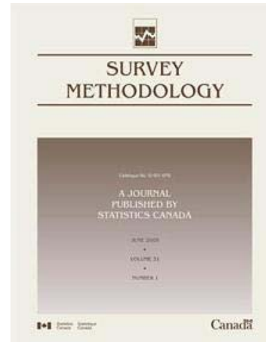
Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups



by Phillip S. Kott

June 2012

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 95-99
Statistics Canada, Catalogue No. 12-001-X

Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups

Phillip S. Kott [1]

Abstract

When there is unit (whole-element) nonresponse in a survey sample drawn using probability-sampling principles, a common practice is to divide the sample into mutually exclusive groups in such a way that it is reasonable to assume that each sampled element in a group were equally likely to be a survey nonrespondent. In this way, unit response can be treated as an additional phase of probability sampling with the inverse of the estimated probability of unit response within a group serving as an adjustment factor when computing the final weights for the group's respondents. If the goal is to estimate the population mean of a survey variable that roughly behaves as if it were a random variable with a constant mean within each group regardless of the original design weights, then incorporating the design weights into the adjustment factors will usually be more efficient than not incorporating them. In fact, if the survey variable behaved exactly like such a random variable, then the estimated population mean computed with the design-weighted adjustment factors would be nearly unbiased in some sense (*i.e.*, under the combination of the original probability-sampling mechanism and a prediction model) even when the sampled elements within a group are not equally likely to respond.

Key Words: Double protection; Prediction model; Probability sampling; Response model; Sampling phase; Stratified Bernoulli sampling.

1. Introduction

In the absence of nonresponse, it is possible to estimate the mean of a finite population from a survey sample without having to assume a statistical model which, no matter how reasonable, may not hold true. This is done by assigning each element of the population a positive probability of sample selection and creating estimators around this random-selection mechanism. Unfortunately, surveys taken in the real world often suffer from nonresponse.

Two different types of models can be used in the face of unit (whole-element) nonresponse. One is a prediction or outcome model in which the survey variable of interest is assumed to behave like a random variable with known characteristics but unknown parameters. The other is a response or selection model where the very act of an element's responding to a survey is treated as an additional phase of random sample selection.

Conventionally, survey statisticians prefer response models for two reasons. In addition to the convenience of response modeling allowing them to treat unit response as an additional phase of random sampling, a survey is usually designed to collect information on a number of variables from the sampled elements. Prediction modeling requires assuming a different model for each survey variable any one of which could fail. Response modeling, by contrast, requires only the assumption of a single model. This is no longer true when there is item (survey-variable-specific) nonresponse. Consequently, prediction modeling is more common when handling item nonresponse through imputation. That being said, item nonresponse is beyond the scope of this note.

Under an assumed response model, the element probabilities of response are treated as unknown, which means that they have to be estimated from the sample. Typically, the response mechanism is assumed to be independent across elements and not to depend on whether the element is in the sample (each element has an *a priori* probability of response which becomes operational if it is selected for the sample). The simplest and mostly commonly used response model separates the sample, and implicitly the entire population, into mutually exclusive groups, called "response homogeneity groups" by Särndal, Swensson and Wretman (1992) (the term "weighting classes" is more common; see, for example, Lohr (2009, pages 340-341)), and assumes that each element in a group is equally likely to be a unit respondent regardless of its probability of selection into the original sample, π_k. Thus, the response mechanism produces a stratified Bernoulli subsample with the groups serving as the strata.

Conditioned on the respondent sample sizes in the groups, a stratified Bernoulli subsample with unknown selection (response) probabilities is converted into a stratified simple random subsample with known selection probabilities: r_g / n_g for the elements in group g when that group has n_g sampled elements, r_g or which respond.

Although the conditional probabilities of response in group g under the stratified Bernoulli response model is r_g / n_g, we will see it is often better to multiply the design

1. Phillip S. Kott, RTI International, Suite 902, 6100 Executive Blvd., Rockville, MD 20852, U.S.A. E-mail: pkott@rti.org.

weight, $d_k = 1/\pi_k$, for a responding element in the group not by n_g / r_g, but by

$$f_g = \frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k}, \qquad (1)$$

where S_g is the original sample and R_g the respondent subsample in group g. This *adjustment factor* can be different from n_g / r_g when the d_k in group g vary.

Little and Vartivarian (2003) claim that using the f_g is what is usually done in practice. They argue, however, that incorporating design weights into the adjustment factor in this way can "add to the variance".

In section 2, we develop the notation for estimating the population mean of a survey variable. Using the n_g / r_g produces a double-expansion estimator, while using the f_g produces a reweighted-expansion estimator. We can express both using a formulation in Kim, Navarro and Fuller (2006). From that expression, it is possible to see that if the survey variable roughly behaves like a random variable with a constant mean within each group regardless of the design weights, then using the f_g will often be more efficient than using the n_g / r_g. In fact, if the survey variable behaved exactly like such a random variable, then the estimated population mean computed with the f_g would be nearly unbiased under the combination of the original sampling design and this prediction model even when the response model fails.

In Section 3, we show that empirical results in Little and Vartivarian (2003) are consistent with these arguments and offer some concluding remarks.

2. The two estimators

Suppose we want to estimate the population mean of a survey variable y_k:

$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{\sum_{g=1}^{G} \sum_{k \in U_g} y_k}{\sum_{g=1}^{G} N_g} = \frac{\sum_{g=1}^{G} N_g \bar{y}_{U_g}}{\sum_{g=1}^{G} N_g},$$

where the population U is divided into G groups, $U_1, \ldots, U_G$, each U_g contains N_g elements, and $N = N_1 + \ldots + N_G$. In the absence of nonresponse, each N_g is estimated in an unbiased fashion under probability-sampling theory by $\hat{N}_g = \sum_{k \in S_g} d_k$, and each $\bar{y}_{U_g}$ is estimated in a nearly (*i.e.*, asymptotically) unbiased fashion

$$\bar{y}_{S_g} = \frac{\sum_{k \in S_g} d_k \, y_k}{\sum_{k \in S_g} d_k}, \qquad (2)$$

under mild conditions when n_g is sufficiently large. We assume both here.

For a formal statement of the conditions under which each $\bar{y}_{S_g}$ is consistent under probability sampling theory and therefore nearly unbiased, see Fuller (2009, page 115). The interested reader is directed to Fuller whenever a result in this note depends on assumptions about the design and population as the sample size grows arbitrarily large. A more rigorous treatment of much of what is to be discussed here under the response model can be found in Kim, Navarro and Fuller (2006).

Let us label the full-sample estimator for $\bar{y}_U$ we have been discussing $\bar{y}_S = \sum^G \hat{N}_g \, \bar{y}_{S_g}$. There are more direct ways to render $\bar{y}_S$, but this version will better serve our purposes.

If we adjust for nonresponse using the f_g in equation (1), we have the reweighted-expansion estimator:

$$\hat{\bar{y}}_{rw} = \frac{\sum_{g=1}^{G} \left(f_g \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^{G} \left(f_g \sum_{k \in R_g} d_k \right)}$$

$$= \frac{\sum_{g=1}^{G} \left(\frac{\sum_{k \in S_g} d_k}{\sum_{k \in R_g} d_k} \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^{G} \sum_{k \in S_g} d_k} = \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k y_k}{\sum_{k \in R_g} d_k} \right)}{\sum_{g=1}^{G} \hat{N}_g}.$$

Technically, $\hat{\bar{y}}_{rw}$ is the ratio of two reweighted-expansion estimators, but we use the simpler terminology here.

Employing the n_g / r_g results in the double-expansion estimator:

$$\hat{\bar{y}}_{de} = \frac{\sum_{g=1}^{G} \left(\frac{n_g}{r_g} \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^{G} \left(\frac{n_g}{r_g} \sum_{k \in R_g} d_k \right)}.$$

For our purposes, this estimator can also be expressed as

$$\hat{\bar{y}}_{de} = \frac{\sum_{g=1}^{G} \left(\frac{\sum_{k \in S_g} d_k p_k}{\sum_{k \in R_g} d_k p_k} \sum_{k \in R_g} d_k y_k \right)}{\sum_{g=1}^{G} \left(\frac{\sum_{k \in S_g} d_k p_k}{\sum_{k \in R_g} d_k p_k} \sum_{k \in R_g} d_k \right)} = \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k y_k}{\sum_{k \in R_g} d_k p_k} \right)}{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k}{\sum_{k \in R_g} d_k p_k} \right)},$$

where

$$p_k = \frac{1}{d_k} \frac{\sum_{j \in S_g} d_j}{n_g} \quad \text{for } k \in S_g \tag{3}$$

(so that $\sum_{S_g} d_k p_k = \sum_{S_g} d_k = \hat{N}_g$).

Both $\hat{\bar{y}}_{rw}$ and $\hat{\bar{y}}_{de}$ can now be written in the form:

$$\hat{\bar{y}}_{S,\mathbf{q}} = \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k y_k}{\sum_{k \in R_g} d_k q_k} \right)}{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k}{\sum_{k \in R_g} d_k q_k} \right)}. \tag{4}$$

For the reweighted-expansion estimator, all $q_k = 1,$ while for the double-expansion estimator, $q_k = p_k$ as defined by equation (3).

We will soon have use of the following for our two estimators:

$$\hat{\bar{y}}_{S,\mathbf{q}} - \bar{y}_S = \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k e_k}{\sum_{k \in R_g} d_k q_k} \right)}{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k}{\sum_{k \in R_g} d_k q_k} \right)} \approx \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k \in R_g} d_k e_k}{\sum_{k \in R_g} d_k q_k} \right)}{\sum_{g=1}^{G} \hat{N}_g}, \tag{5}$$

where $e_k = y_k - \bar{y}_S$. Equation (5) holds exactly when all $q_k = 1$. When $q_k = p_k$, the near equality depends on the r_g being sufficiently large and other mild conditions.

Now assume the following response model holds: Each element k in a group has an equal, positive probability of response that does not vary with π_k or with y_k. That is to say, the response indicator ρ_k, which is 1 when k responds if sampled and is 0 otherwise, is a Bernoulli random variable with a common mean in U_g regardless of the values of π_k and y_k.

By treating unit response as a second phase of probability sampling in this way, the added variance/mean-squared-error due to nonresponse given the original sample and the r_g for both estimators can be expressed as

$$A_{\mathbf{q}} = E_\rho[(\hat{\bar{y}}_{S,\mathbf{q}} - \bar{y}_s)^2 \mid S, \{r_g\}]$$

$$\approx \frac{\sum_{g=1}^{G} \hat{N}_g^{\,2} \operatorname{Var}_\rho(\hat{\bar{e}}_{S_g,\mathbf{q}} \mid S_g, r_g)}{\left(\sum_{g=1}^{G} \hat{N}_g \right)^2}, \tag{6}$$

where $\hat{\bar{e}}_{S_g,\mathbf{q}} = \hat{\bar{y}}_{S_g,\mathbf{q}} - \bar{y}_S$, $\bar{e}_{S_g} = \bar{y}_{S_g} - \bar{y}_S$, and

$$\operatorname{Var}_\rho(\hat{\bar{e}}_{S_g,\mathbf{q}} \mid S_g, r_g)$$

$$\approx \left(\frac{n_g}{r_g} - 1 \right) \frac{\sum_{k \in S_g} d_k^{\,2} (e_k - q_k \bar{e}_{S_g})^2}{\left(\sum_{k \in S_g} d_k q_k \right)^2}$$

$$= \left(\frac{n_g}{r_g} - 1 \right) \frac{\sum_{k \in S_g} d_k^{\,2} ([y_k - \bar{y}_S] - q_k [\bar{y}_{S_g} - \bar{y}_S])^2}{\left(\sum_{k \in S_g} d_k q_k \right)^2}, \tag{7}$$

under mild conditions on the population and original sampling design we assume to hold, including (again) that the r_g are sufficiently large. These conditions make both estimators nearly unbiased under quasi-probability sampling theory (probability theory augmented with a response model) and render the distinction between large-sample variance and mean squared error moot. Quasi-probability sampling theory is also known as "quasi-design-based" and "quasi-randomization-based" sampling theory.

Looking at equations (6) and (7), we see that at one extreme $\hat{\bar{y}}_{rw}$ has an added variance due to nonresponse of (approximately) zero when all the originally sampled y_k in a group are equal, while at the other $\hat{\bar{y}}_{de}$ has an added variance of zero when all the originally sampled $d_k e_k$ (or, put another way, the $d_k [y_k - \bar{y}_S]$) in a group are equal.

Heuristically, the reweighted-expansion estimator is more efficient than the double-expansion estimator when $\bar{e}_{S_g}$ is a better predictor of e_k for $k \in S_g$ than $p_k \bar{e}_{S_g}$. Thus, when the groups were constructed as advised in Little and Vartivarian (2003) and earlier in Little (1986) so that the y_k in a group are homogeneous (as opposed to the $d_k [y_k - \bar{y}_S]$ being homogeneous), then the reweighted-expansion estimated computed with the f_g will usually be more efficient than the double-expansion estimator computed with the n_g / r_g.

The heuristic observation can be formalized with an alternative justification for using the reweighted-expansion estimator. Suppose the following prediction model holds: Each y_k in U_g is a random variable with common mean, μ_g, regardless of π_k and ρ_k. Then $\hat{\bar{y}}_{rw}$ is nearly unbiased under mild conditions with respect to the combination of the original sampling mechanism (which treats the d_k as random, where $d_k = 0$ for $k \neq S$) the prediction model (which treats the y_k as random). That is to say, $\mathrm{E}_d[\mathrm{E}_y (\hat{\bar{y}}_{rw} - \bar{y}_U \mid S)] \approx 0,$ since the double expectation of both $\hat{\bar{y}}_{rw}$ and $\bar{y}_U$ are nearly $\sum^G N_g \mu_g / \sum^G N_g$. This combined unbiasedness is exact when the design is such that $\sum_S d_k \equiv N$. Stratified, simple random sampling is an example of such a design. Unstratified sampling with unequal probabilities and many multistage designs are not.

It is not hard to see that $\hat{\bar{y}}_{rw}$ is also exactly unbiased with respect to this double expectation (*i.e.*, $\mathrm{E}_d\,[\,\mathrm{E}_y\,(\hat{\bar{y}}_{rw} - \bar{y}_U \mid S)] = 0$) when all the μ_g are equal. In fact, the prediction-model expectation of both $\hat{\bar{y}}_{rw}$ and $\hat{\bar{y}}_{de}$ equals this common mean, as does the prediction-model expectation of an estimator without any adjustment for unit nonresponse, that is, with the f_g in $\hat{\bar{y}}_{rw}$ replaced by 1. The advantage of $\hat{\bar{y}}_{rw}$ over $\hat{\bar{y}}_{de}$ under the prediction model obtains only when the μ_g vary, that is, when the survey variable has a different prediction mean across the groups.

Notice that if *either* the response model or the prediction model holds, then the reweighted-expansion estimator is nearly unbiased in some sense (*i.e.*, under the combination of the original design and the response model or under the original design and the prediction model). This property has been called "double protection" against nonresponse bias. See, for example, Bang and Robins (2005).

3. Concluding remarks

In this note, we discussed two distinct types of models. We stressed a response model, which treats the response indicators, ρ_k, as a Bernoulli random variable within each group but with unknown parameters. We also described a prediction model, which treats the survey values, y_k, as random variables with unknown means that could vary across groups but not within them.

As part of the response model, we assumed that within a group, the ρ_k do not depend on the y_k. Analogously as part of the prediction model, we assumed that within a group, the y_k do not depend on the ρ_k. When both ρ_k and y_k are treated as random variables the former assumption, that nonrespondents are *missing at random*, is equivalent to the latter assumption, that the response mechanism is *ignorable* (see, for example, Little and Rubin 1987). It should be understood, however, that the y_k need not be treated as random variables under the response model and the ρ_k need not be treated as random variables under the prediction model. The two concepts (missingness at random and ignorable nonresponse) may be equivalent in some sense but they are not identical.

The heart of Little and Vartivarian (2003) is a series of simulations featuring a binary survey variable, two potential response groups, and two original selection probabilities. Both the survey variable and response indicators are generated under five models. The expected value of each is a function of, 1, the response group only, 2, the selection probability only, 3, neither, or, 4 and 5, one of two equal combinations of response group and selection probability. This produces 25 scenarios, of which 10 are of primary interest to us. Those are the ones in which the survey variable is a function either of only the response group or of neither the response group nor the selection probability.

As our theory predicts when the survey variable is a function of neither the response group nor the selection probability, both the reweighted- and double-expansion estimators have empirical biases near zero (Table 5 in Little and Vartivarian) because both are nearly unbiased under the combination of the original sampling design and a valid prediction model: all population elements have the same mean. When the survey variable is a function of the response group *and* the response indicator is wholly or partly a function of the selection probability, only the reweighted-expansion estimator is nearly unbiased empirically since only it is unbiased under the combination of the original sampling design and a valid prediction model. As a result, $\hat{\bar{y}}_{rw}$ also has less empirical root mean squared error and significantly less average absolute error as an estimator for $\bar{y}_S$ (Tables 4 and 6 in Little and Vartivarian, respectively; the significance test treats the mean value across the simulations of $\mid \hat{\bar{y}}_{rw} - \bar{y}_S \mid - \mid \hat{\bar{y}}_{de} - \bar{y}_S \mid$ as asymptotically normal).

When both the survey variable and response indicators are functions of the response group only, the reweighted-expansion estimator has slightly less empirical root mean squared error and average absolute error than the double-expansion estimator but the latter is not significant.

It should not surprise us that the reduction in empirical root mean squared error is modest. The contribution to the variance from nonresponse under the response model mechanism expressed in equations (6) and (7) is conditioned on the original sample (technically, the contribution of non-response to the total quasi-probability variance of $\hat{\bar{y}}_{S,\mathbf{q}}$ is the expectation of $A_{\mathbf{q}}$ in equation (6) under the original sampling mechanism). In applications where the response rates are relatively large (in the simulations they averaged 0.5), this contribution can be dominated by the probability-sampling variance/mean squared error of the full-sample estimator, $\hat{\bar{y}}_U$.

Two warnings are in order. The respondent sample size within each group must be sufficiently large for the reweighted-expansion estimator to nearly unbiased under quasi-probability theory. For the double-expansion estimator, each r_g need only be positive. Moreover, that the reweighted-expansion estimator is doubly protected against nonresponse bias is only helpful when either the assumed response or prediction model is correct. If *both* the response probabilities and survey values vary with the design weights, then the reweighted-expansion estimator can be meaningfully biased. Despite the slant taken in this note, that is the take-away message Little and Vartivarian (2003) intended, and it cannot be disputed.

Acknowledgements

I thank the associate editor and two referees whose careful reading of earlier versions of the manuscript greatly improved the quality and accuracy of the resulting work. Any remaining errors in the text are wholly due to the author.

References

Bang, H., and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.

Fuller, W. (2009). *Sampling Statistics*, Hoboken, New Jersey: Wiley.

Kim, J.K., Navarro, A. and Fuller, W. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Little, R. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 139-157.

Little, R., and Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.

Little, R., and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.

Lohr, S (2009). *Sampling: Design and Analysis*, *Second Edition*, Boston: Brooks/Cole.

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