Voluntary and Volunteer Government Surveys in the US

Phillip S. Kott
RTI International
Email: pkott@rti.org

Abstract

Although many US government surveys are mandatory, most are voluntary and more than a few can be considered volunteer. We will discuss the repercussions of this on the quality of those surveys and what can be done to assure the results are defensible.

Keywords: mandatory, nonresponse bias, calibration weighting, selection bias, prediction model, double protection.

1. Introduction

The United States does not have a centralized government statistical agency, like Statistics Canada, Statistics Sweden, or the Australian Bureau of Statistics. Important government surveys are administered by the the Bureau of Labor Statistics (BLS), the National Center of Health Statistics (NCHS), the Energy Information Administration (EIA), the Federal Reserve Board (FRB), the National Center for Education Statistics (NCES), the National Agricultural Statistics Service (NASS), the US Census Bureau, and a host of other agencies. Complicating matter even further, only the Census Bureau and NASS run their own surveys. The other government agencies contract out their surveys, sometimes to private contractors like RTI International and Westat, sometimes to the Census Bureau, and sometimes to state government agencies. To a certain extent, that is what NASS does but their relationships with state agricultural agencies are so close that the distinction between NASS and those state agencies is (in my view) little more than a technicality when it comes to running surveys.

US law requires the administering agency to tell a individual or establishment selected for a survey whether their participation is mandatory under penalty of law. Although examples of penalties for refusing to respond to a mandatory US survey are rare, there is strong evidence that the simple statement that a survey is mandatory has a dramatic impact on its response rate (Tulp *et al*., 1991; Navarro *et al.* 2011).

Determining from the outside which of the many US government surveys run by the diversified federal statistical system is mandatory is not a trivial exercise. Most surveys of individuals are not mandatory. Obvious exceptions are the decennial population census and the American Community Survey, administered by the Census Bureau. A few other surveys of specialized populations, like the National Inmate Survey of the Bureau of Justice Statistics (BJS), are also mandatory.

Examples of voluntary surveys of individuals include BLS's Current Population Survey, which is used to estimate politically sensitive unemployment rate and various health surveys like the National Health Interview Survey (NHIS), the National Health and Nutrition Examination Survey (NHANES), and the National Survey of Drug Use and Health (NSDUH). The first two of these are administered by NCHS, which is part of the Center for Disease Control. The later by the Substance Abuse and Mental Health Services Administration of the Public Health Service. The Agency for Healthcare Research and Quality also conducts voluntary surveys of individuals (e.g., the Medical Expenditure Panel Survey) as does the BJS (the National Crime Victimization Survey).

Many establishment surveys (i.e., surveys of businesses, farms, or institutions) are mandatory. These include the various five-year economic censuses, like the Census of Agriculture (NASS) and the Census of Manufactures (the Census Bureau, which runs all the US economic censuses other than the Census of Agriculture), the Quarterly Financial Report (FRB), and the Survey of Occupational Injuries and Illnesses (BLS). Most EIA surveys are mandatory. NASS surveys are not except related to the Census of Agriculture. Some NCES school surveys, such as those belonging to the Integrated Postsecondary Education Data System (IPEDS), are mandatory. In fact, school that fail to participate in IPEDS surveys have been fined, an exception to the general rarity of penalties for mandatory-survey noncompliance. Others NCES school surveys are voluntary.

Some establishment surveys rely on a panel of willing participants. This is effectively the case with the Census Bureau's "M3 Survey" (Manufacturers' Shipments, Inventories, and Orders), which plays a central role in the index of leading economic indicators, and literally the case with the US Department of Agriculture's Pesticide Data Program, which for a number of years produced national estimates of pesticide residue of fruits and vegetables based on data collected in nine volunteer states.

The existence of survey nonresponse does not necessarily imply that there is nonresponse bias (see, for example, Groves and Peytcheva 2008), only that there is a potential for nonresponse bias that would not be otherwise. Since mandatory surveys often suffer from nonresponse, mandatory surveys like voluntary surveys can have nonresponse bias.

Nonresponse bias is one of many sources of survey error. Others include measurement error resulting from erroneous survey responses, which may as much be the consequence of poor questions as poor answers, and coverage error due to imperfections in the frame (list of units) from which the survey sample was drawn. Even censuses have can have frame errors either because the government is unaware of the existence of some units that should be on the frame or because the government fails to detect that some units are contained on the frame multiple times.

Section 2 lays out some theory on nonresponse bias and discusses a general approach for reducing its potential impact. The calibration-weighting method used to adjust for potential nonresponse bias in Section 2 is generalized to selection bias in Section 3, which includes potential biases from coverage errors and from nonrandom sampling. Section 4 provides some concluding remarks.

2. Nonresponse Bias

In this section, we assume that the frame is perfect and that there are no measurement errors in our survey. If, in addition, there were no nonresponse, then an unbiased estimator for a population total, $T_y = \sum_U y_k$, under probability-sampling theory is $t_y = \sum_S d_k y_k$, where U denote the population, S the sample, and d_k the sampling weights of population unit k. By design $d_k = 1/\pi_k$, and π_k is the probability that unit k is selected for the sample. Similarly, a (nearly) unbiased estimator for the corresponding population mean, $M_y = \sum_U y_k / \sum_U 1$, under probability-sampling theory is $m_y = \sum_S d_k y_k / \sum_S d_k$. For many, but not all, establishment surveys and some human-population surveys, $\sum_S d_k \equiv N$ and the modifying "nearly" can be removed from "unbiased."

Now suppose there is unit nonresponse to a survey. One obvious way of handling nonresponse in a quasi-probability framework is to assume that each unit k has a probability of response denoted by ρ_k or more precisely $\rho_k\,|S$. If this value were known, than a nearly unbiased estimators for T_y and M_y would be $t_{y,\rho}$ = $\sum_R (d_k/\rho_k)y_k$ and $m_{y,\rho}$ = [$\sum_R (d_k/\rho_k)y_k$]/ $\sum_R (d_k/\rho_k)$, respectively, where R is the respondent subsample. Unfortunately, the ρ_k are unknown. Instead, they must be estimated or somehow proxied. Two obvious and well know proxies are 1 and r/n, the sample response rate (which is sometimes replaced by its weighted analogue: $\sum_R d_k/\sum_S d_k$).

Whatever choice is made, and we will discuss better choices later in this section, the biases in $t_{y,r}$ and $m_{y,r}$ caused by replacing the unknown ρ_k with r_k can be expressed theoretically as

$$\begin{aligned} E\left(t_{y,r}-T_y\right) &= -E\left(\sum\nolimits_R d_k y_k \frac{r_k-\rho_k}{r_k}\right), \quad \text{and} \\ E\left(m_{y,r}-M_y\right) &= -E\left(\sum\nolimits_R \frac{d_k}{r_k}\left[y_k-M_y\right]\left[r_k-\rho_k\right]\right). \end{aligned} \tag{1}$$

The biases in equation (1) can be simplified if we assume that the probabilities of response are independent of selection (i.e., $\rho_k = \rho_k|S$). Even without that assumption, it is easy to see that, if r_k where consistently higher (or lower) than ρ_k, $t_{y,r}$ would clearly be biased downward (upward). The same, however, could not be said about $m_{y,r.}$. In fact, $m_{y,r}$ is only biased if r_k is in some sense correlated with $y_{k.}$. If N were known, the same could be said about $t_{y,r}$ if each r_k were replaced by $r_k' = r_k[(\sum_R (d_j/r_j)/N)$. This is because we have forced $\sum_R (d_k/r_k')$ to equal N and $t_{y,r'}$ to equal $Nm_{y,r}$.

It is clear from equation (1) that if each r_k is an unbiased estimator for ρ_k, then there would be no nonresponse bias. It is therefore tempting to assume that unit response is a function, say a logistic function, of a vector of variables $\mathbf{x}_k$ with values known for all units in the full sample and then to estimate the r_k using logistic regression.

Suppose there is a vector of variables $\mathbf{z}_k$ such that either the population total, $\mathbf{T}_\mathbf{z} = \sum_U \mathbf{z}_k$, or weighted full-sample total, $\mathbf{t}_\mathbf{z} = \sum_S d_k\mathbf{z}_k$, can be treated as known, and there is a vector $\boldsymbol{\lambda}$ such that $\sum_U c_k\boldsymbol{\lambda}^T\mathbf{z}_k = 1$ for some c_k. Let $\boldsymbol{\beta} = (\sum_U c_k\mathbf{z}_k\mathbf{z}_k^T)^{-1}\sum_U c_k\mathbf{z}_k y_k$ so that $T_y = \mathbf{T}_\mathbf{z}^T\boldsymbol{\beta}$. Some or all of the components of $\mathbf{z}_k$ may coincide with the components of the vector $\mathbf{x}_k$ postulated above, but that is not required. I

Now suppose we can determine values for the r_k such that $\mathbf{t}_{\mathbf{z},r} = \sum_S (d_k/r_k)\mathbf{z}_k$ equals $\mathbf{t}_\mathbf{z} = \sum_S d_k\mathbf{z}_k$ or $\mathbf{T}_\mathbf{z} = \text{v}\sum_U \mathbf{z}_k$. Then it is not hard to show that

$$E\left(t_{y,r}-T_y\right) = -E\left(\sum\nolimits_R \frac{d_k}{r_k}\left[y_k-\mathbf{z}_k{}^T\boldsymbol{\beta}\right]\left[r_k-\rho_k\right]\right). \tag{2}$$

The analogous expression for $m_{y,r}$ includes the factor $1/([(\sum_R (d_j/r_j)]$ within the parentheses on the right-hand size of equation (2). One practical advantage is that knowledge of population and full-sample means of the components of $\mathbf{z}_k$ replaces knowledge of their totals.

What is left is a method for determining the r_k so that the *calibration equation* $\sum_S (d_k/r_k)\mathbf{z}_k = \sum_S d_k\mathbf{z}_k$ or $\sum_U \mathbf{z}_k$ is satisfied (the right-hand side of this equation can in practice) contain a mix of population and full-sampled weighted totals). When the ρ_k are assumed to have the form $\rho_k = \rho(\mathbf{x}_k{}^T\boldsymbol{\gamma})$, and the number of components in $\mathbf{z}_k$ and $\mathbf{x}_k$ are the same, then one can often use Newton's method to find a consistent estimator $\mathbf{g}$ for $\boldsymbol{\gamma}$ that satisfies the calibration equation (sometimes no solution exists). After that, one simply sets $r_k = \rho(\mathbf{x}_k{}^T\mathbf{g})$. The combined weight $w_k = d_k/r_k$ is called a "calibration weight" (Deville and Särndal 1992).

If, in fact, the *response model* for ρ_k has the form we assumed, then computing the r_k as described above removes the potential for nonresponse bias in $t_{y,r}$ (and $m_{y,r}$) when the sample size is large (most quasi-probability sampling results are asymptotic). Alternatively, if the prediction model for y_k is linear in $\mathbf{z}_k$ and the model error is the same given $\mathbf{z}_k$ whether or not k responds, then $t_{y,r}$ is unbiased under the combination of the prediction model and the sampling design no matter what the response model. This property of calibration weighting that when *either* an assumed response or prediction model holds, nonresponse bias is (nearly) eliminated is called "double protection" (see, for example, Kott and Liao, 2012). Equation (2) shows us that even if neither model holds, the resulting estimator can be unbiased, as long as the r_k as an estimator for ρ_k is in some sense uncorrelated with the population residual $y_k - \mathbf{z}_k{}^T\boldsymbol{\beta}$.

One problem with nonresponse-bias evaluation when the assumed response model doesn't hold is that it depends on the variable total or mean being estimating. Most surveys are designed to estimate a number of population totals (or means). There may be no bias when estimating one variable total but bias when estimating a different total.

SUDAAN 11 (RTI 2011) allows the user to create calibration weights when the response model is assumed to be logistic and the $\mathbf{z}_k$ and $\mathbf{x}_k$ vectors differ. In practice, however, determining calibration weights may not require the use of specialized software. Commonly, the population is divided into P mutually exclusive model groups, and $\mathbf{z}_k = \mathbf{x}_k$ is simply a vector of group identifiers. When the population sizes of the groups are used in the calibration, this is called "poststratification." When the weighted sample sizes are used, "weighting-class adjustment." The hard work of nonresponse adjustment is creating the groups. For human-population surveys, this often involves the cross-classification of categorical variables like race, sex, and age group.

Many establishment surveys employ a simple variant of poststratification/weighting-class adjustment when there is a measure of size, q_k, associated with every member of the frame or full sample (e.g., sales in a previous Census or survey). In this variant, $\mathbf{z}_k$ is set equal $q_k\mathbf{x}_k$ rather than $\mathbf{x}_k$. Thus, a simple ratio estimate (y over q) is computed within each group and then combined using the population q-totals (off full-sample-weighted q-totals) for each group.

It is in this context that the c_k in $\boldsymbol{\beta} = (\sum_U c_k\mathbf{z}_k\mathbf{z}_k{}^T)^{-1}\sum_U c_k\mathbf{z}_k y_k$ are not equal to 1. They are $1/q_k$, and $\boldsymbol{\beta}$ is a P-vector the p^{th} component of which is $\sum_{Up} y_k/\sum_{Up} z_k$, where Up is that part of population governed my model group p. The implicit response model is that every unit in a group is equally likely to respond. Similarly, a separate prediction model holds with each group: y_k is a linear function of $\mathbf{q_k}$ that goes through the origin regardless of whether k responds. If either of these models hold, the estimate is nearly unbiased in some sense.

A popular method of calibration weighting when the components of $\mathbf{z}_k = \mathbf{x}_k$ are binary is iterative proportional fitting or raking (Deming and Stephan 1940). This corresponds to the response model of the form $\rho(\mathbf{x}_k{}^T\boldsymbol{\gamma}) = \exp(\mathbf{x}_k{}^T\boldsymbol{\gamma})$, which allows r_k to exceed 1. Less popular, but allowing components of $\mathbf{z}_k = \mathbf{x}_k$ to be continuous, is simple linear calibration in which implicitly $\rho(\mathbf{x}_k{}^T\boldsymbol{\gamma}) = 1/(1 + \mathbf{x}_k{}^T\boldsymbol{\gamma})$, and some r_k can exceed 1 while others fall below zero. Fuller *et al* (1994) used this approach in the now-defunct Continuing Survey of Food Intakes by Individuals administered by an also defunct agency of the US Department of Agriculture.

SUDAAN 11 allows the response model to have the general exponential form (Folsom and Singh 2000) form:

$$\rho_k = \frac{(u-c)+(c-\ell)\exp(\boldsymbol{\gamma}^T\mathbf{x}_k)}{\ell(u-c)+u(c-\ell)\exp(\boldsymbol{\gamma}^T\mathbf{x}_k)}, \tag{3}$$

where $\infty \geq u > c > \ell \geq 0$, which constrains the estimated probabilities of selection between $1/u$ and $1/\ell$. Logistic response is a special case ($\ell = 1$, $c = 2$, $u = \infty$).

3. Selection Bias

It is a simple matter to extend the theory developed in the last section to coverage adjustment. Instead of ρ_k being the probability of response, it is the expected number of times k is on the frame, which can exceed unity if there is duplication in the frame. The values of the ρ_k do not depend on the sample actually drawn. When there is the possibility of frame duplication, ℓ in equation (3) can be set below 1.

The Census of Agriculture uses a truncated version of linear calibration to adjust for the undercoverage of its list frame (Fetter 2009). The **z**-totals come from an area-based probability survey.

The theory likewise extends to nonprobability surveys, which are becoming increasing popular. In that context ρ_k becomes the probability that unit k self-selects itself for the survey. Although it is unlikely that we can successfully model this response probability with a simply function, equation (2) nonetheless provides a method for discussing the potential for selection bias. Often some of the totals or means for the **z-**vector come from a probability survey (see, for example, DiSogra *et al.* 2011) or some other outside source.

4. Concluding Remarks

In theory, calibration weighting employs that the same weight regardless of the survey variable, *y*. Often in practice when some form of post-stratification or separate group ratios is used, the calibration weighting is implicit, and there can be different implicit calibration weights for different survey variables.

The effectiveness of calibration-weighting method rely on the accuracy of **z-**totals (or means). Ideally, they should be perfect. Failing that, nearly unbiased. In the US, which does not have nearly complete population and business registers, these values often comes from other surveys. Thus, even though (in my view) the use of nonprobability principles may sometimes result in estimates with little to no selection bias, gold-standard government probability surveys will always be needed. Moreover, in order to limit their potential for nonresponse bias, it would be advisable for them to have mandatory collection authority.

Acknowledgements

Many folks kindly responded to my internet query about US mandatory surveys. Among them were Carol Blumberg, Rachel Caspar, Jeffrey Gonzalez, Mark Harris, David Judkins, Arthur Kennickell, and Marilyn Seastrom.

References

Deming, W. and Stephan, F. (1940). "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known,"" *Annals of Mathematical Statistics* , 427–444.

Deville, J. and Särndal, C. E. (1992). "Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, 376-382.

DiSogra, C., Cobb, C., Chan, E. and Dennis, J. (2011). "Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics," *ASA Proceedings of the Survey Research Methods Section*, http://www.amstat.org/sections/srms/proceedings/y2011f.html.

Fetter, M. (2009). "An Overview of Coverage Adjustment for the 2007 Census of Agriculture," *ASA Proceedings of the Section on Government Statistics*, http://www.amstat.org/sections/srms/proceedings/y2009/ Files/304352.pdf.

Folsom, R. and Singh, A. (2000). "The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification," *ASA Proceedings of the Survey Research Methods Section*, http://www.amstat.org/sections/srms/Proceedings/y2000f.html.

Fuller, W. Loughin, M., and Baker, H. (1994). "Regression Weighting for the 1987–88 National Food Consumption Survey, *Survey Methodology*, 75–85.

Groves, R. and Peytcheva, E. (2008). "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis," *Public Opinion Quarterly*, 168-189.

Kott, P. and Liao D. (2012). "Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-Weighting Routine," *Survey Research Methods*, 105-111.

Navarro, A., King, K., and Starsinic, M. (2011). *Comparison of the American Community Survey Voluntary Verse Mandatory Estimates: Final Report*, http://www.census.gov/acs/www/Downloads/library/2011/2011_Navarro_01.pdf.

RTI International. (2012). *SUDAAN language manual, Release 11.0.* Research Triangle Park, NC: RTI International.

Tulp, D., Hoy, E., Kusch, G., and Cole, S. (1991). "Nonresponse under Mandatory Vs. Voluntary Reporting in the 1989 Survey of Pollution Abatement Costs and Expenditure," *ASA Proceedings of the Survey Research Methods Section*, http://www.amstat.org/sections/srms/Proceedings/papers/1991_044.pdf.