Journal of Official Statistics
Vol. 3, No. 4, 1987, pp. 339–342

Quantifying Nonsampling Errors and Bias

Robert D. Tortora[1]

1. Introduction

One of the most challenging areas of statistical research is the quantification of nonsampling errors and biases. These errors and biases occur at every stage of the survey process, making the task even more formidable. Error profiles such as those by Brooks and Bailar (1978) and Beller (1979) illustrate the large number of sources of error in a survey. For a continuing survey, error profiles reflect what is known about nonsampling errors and biases, and help statisticians to systematically address the measurement problem.

In their discussion of standards, Gonzalez et al. (1978) recommended that "nonsampling errors should also be discussed and the user made aware that the total error is larger than the estimated sampling errors." Gonzalez et al. characterize nonsampling errors in the following way:

"... smaller for estimates of month-to-month relatives than for estimates of monthly levels ..."

"... minor for most general statistics estimates and somewhat greater for the product class estimates ..."

"... wider margin of relative error and response to variability in data for small areas than for large areas ..."

Only one example specifies the magnitude of the possible nonsampling error, giving the percent of data imputed. But without a measure of the magnitude of the errors and biases, data users[2] are in a precarious position. They must interpret magnitudes expressed in terms like smaller, minor, wider, etc., and surely their interpretations differ. In addition, for continuing surveys, the data user must interpret the survey's performance and over time. Are the errors measured at time l equal to the measures of error at time $(l+k)$?

It is almost impossible to list all the potential nonsampling errors and biases associated with a survey or census, and even more difficult to quantify these errors and biases. Two proposed research areas that address these problems are (1) generalized models to measure nonsampling errors and biases, and (2) a process quality control system to measure survey performance. Ideally, the data user should receive, along with the survey performance measures, three additional values for each survey statistic, namely, a measure of: sampling error, nonsampling error, and bias. In the remainder of this paper I would like to outline these areas of research.

[1] Director, Research & Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250.

[2] Data users are broadly defined and can include the producers of survey statistics.

2. General Nonsampling Error and Bias Models

A generalized nonsampling error model and a generalized bias model apply to surveys where one can obtain at least a proxy for the "true" values. Kish (1965) classifies survey error into variable error and bias. This model is useful in understanding the sources of survey error. Anderson et al. (1979) apply Kish's model to study specific variable nonsampling errors and biases for a health survey. To obtain true values, they used: hospital records, physician office records, insurance company records, and employer records. But as error profiles point out, it is difficult, at best, to quantify all of the variable nonsampling errors and biases associated with a survey. Data users have a great deal of information to absorb and still might not understand the total error in a survey.

One way to quantify nonsampling errors and biases is by developing generalized nonsampling error models and generalized bias models, which are analogous to generalized variance functions (GVFs) (Wolter (1983)). Of course, the key to developing these models is the existence of a true value. In many surveys it is possible, though expensive, to obtain true values through validation studies, use of administrative records, and highly accurate reinterviews.

The purpose of generalized nonsampling error and bias models is different from that of GVFs. GVFs are usually used to minimize the computations when variances are computed for a large number of variables. The generalized nonsampling error models and generalized bias models are proposed here to help the data user understand and interpret survey results. Until recently, the justification for GVFs has been almost entirely empirical, not theoretical (see Valiant (1987)). The same case can be made for generalized nonsampling error models and generalized bias models, but Faulkenberry and Tortora (1983) illustrate that a theoretical framework for these models can be developed.

Suppose $\hat{\theta}(x)$ and $\hat{\theta}(y)$ are two estimators for a parameter θ, where $y=(y_1,\ldots, y_n)$ denotes a sample with units measured with no bias or random error, and where $x=(x_1,\ldots, x_n)$ denotes observations that may contain both bias and random errors. y denotes true values and x denotes values with nonsampling error. If we have a sampling plan so that $\hat{\theta}(y)$ is unbiased for θ and if we can observe y, then the coefficient of variation of y, $\mathrm{CV}(\hat{\theta}(y))$, is a measure of the relative precision of estimation of θ. If we, however, observe x rather than y, then the relative mean square error of x, $(\mathrm{E}(\hat{\theta}(x)-\theta)^2)^{1/2}/\theta=(\mathrm{MSE}(\hat{\theta}(x))^{1/2}/\theta$, is a measure of how close we expect $\hat{\theta}(x)$ to be to θ. Comparing $\mathrm{CV}(\hat{\theta}(y))$ and $(\mathrm{MSE}(\hat{\theta}(x))^{1/2}/\theta$ shows the total effect of nonsampling error on the estimation of θ.

Another similar comparison characterizing the bias due to nonsampling error is: $(\mathrm{MSE}(\theta(x))^{1/2}/\mathrm{E}(\hat{\theta}(x))$ versus $(\mathrm{CV}(\hat{\theta}(x))$. The latter quantity is usually calculated from survey data. If a relationship were established between these two quantities, then we could predict a more realistic measure of the relative precision given $\mathrm{CV}(\hat{\theta}(x))$.

Using data from an agricultural economic survey where true values were obtained from administrative records we developed two linear regressions to predict the increase in error due to variable nonsampling errors and to bias. The generalized nonsampling error model is of the form $(\mathrm{MSE}(\hat{\theta}(x))^{1/2}/\theta=a+b(\mathrm{CV}(\hat{\theta}(y))$ and the generalized bias model is of the form $(\mathrm{MSE}(\hat{\theta}(x))^{1/2}/\hat{\theta}(x)=c+d(\mathrm{CV}(\hat{\theta}(x))$.

Generalized nonsampling error and bias models allow the data user to assess the increase in relative error due to nonsampling errors and bias.

3. Process Quality Control System

The second area of research that would improve the understanding of continuing surveys is

process quality control. Almost all of the data necessary to start a process quality control system are already available in the survey data system. The process quality control system would use various "performance" variables to determine if the survey process is in control. Examples of survey performance variables are: the amount of imputation, the nonresponse rate, the number of proxy respondents, the measures of sampling error, etc. Of course, this would be done after the tabulation stage of a survey. The results could be presented, preferably graphically, along with the publication of the usual survey report.

More specifically, a process quality control system would single out specific performance variables to measure a source of error in a survey. The key to the success of a process quality control system is, of course, identifying the correct performance variables. For example, suppose one wanted to measure the National Agricultural Statistics Service's (NASS) list sampling frame coverage of U.S. farms. (Since most NASS probability surveys are dual frame, and one of the frames is a complete area frame, this coverage measure is possible.) To calculate this type of coverage measure, one needs a dual frame survey where one of the frames is a regularly updated list frame. Using the survey data and its newly updated list frame a point estimate and its sampling error can be calculated and added to the following control charts:

a. percent of farms in the area frame and not on the list frame,
b. amount of land in farms in the area frame and not on the list frame,
c. number of farms overlap between the two frames as a percent of the list sampling frame,
d. percent of farms in the area frame and not on the list frame by type of farm (here there are actually several charts, for example, livestock, cash grain farm, etc.), and
e. percent of farms in the area frame and not on the list frame by size of farms (for several gross value of sales classes).

When analyzed over time, each one of the control charts provides information about the list frame coverage of U.S. farms. Of course, since a list frame is built for each state in the U.S., charts can be generated for each state, too. When one or more of the charts indicate that the list frame coverage is out of control, not only can corrective measures be instituted over the ensuing year to attempt to bring the list frame coverage back into control and the potentiel for increasing variable errors and biases is reduced.

4. Summary

The results of the process quality control system can complement the generalized nonsampling error models and the generalized bias models. The latter provide measures of the total nonsampling error and bias, while process quality control systems provide information about a source of a nonsampling error or bias or indicate a deterioration in the survey process over time.

5. References

Anderson, R., Kasper, J., Frankel, M.R., and Associates (1979): Total Survey Error. Jossey-Bass Publishers, San Francisco.

Beller, N.D. (1979): Error Profile for Multiple-Frame Surveys. Economics, Statistics, and Cooperative Services Report, ESCS-63, U.S. Department of Agriculture, Washington, D.C.

Brooks, C.A. and Bailar, B.A. (1978): An Error Profile: Employment as Measured by the Current Population Survey. Statistical Policy Working Paper 3. U.S. Department of Commerce, Washington, D.C.

Faulkenberry, G.D. and Tortora, R.D. (1983): A Case Study of Nonsampling Error. National Agricultural Statistics Service Report, U.S. Department of Agriculture, Washington, D.C.

Gonzalez, M.E., Ogus, J.L., Shapiro, G., and Tepping, B.J. (1975): Standards for Discussion and Presentation of Errors in Survey and Census Data. Journal of the American Statistical Association, 70, 351, Part II.

Kish, L. (1965): Survey Sampling. Wiley, New York.

Valiant, R. (1987): Generalized Variance Functions in Stratified Two-Stage Sampling. Journal of the American Statistical Association, 82, pp. 499–508.

Wolter, K.M. (1983): Introduction to Variance Estimation. Springer-Verlag, New York.

Received December 1987