Journal of Official Statistics, Vol. 27, No. 1, 2011, pp. 135–138

Letter to the Editor

Letters to the Editor will be confined to discussion of articles which have appeared in the Journal of Official Statistics and of important issues facing the statistical community.

Cut-off Sampling and Total Survey Error

This is with regard to "A Framework for Cut-off Sampling in Business Survey Design" by Roberto Benedetti, Marco Bee, and Giuseppe Espa (Journal of Official Statistics, Vol. 26, No. 4, 2010, 651–671).

It was encouraging to see the subject article considering how little is published on this tool. Many only consider this a "quick-and-dirty" technique, but it actually can often be the most accurate option available for highly skewed establishment surveys, when used with a model-based ratio-type estimator.

In the Benedetti et al. article, we see three partitions of the population, U_C, U_S, and U_E, for a "take-all," "take-some," and "take-nothing" design, respectively. This is a nice approach to attempt an optimal methodology. "Cut-off sampling" usually only entails U_C, and U_E.

On page 655 of the article, the authors state that my approach in a couple of articles have been ". . . single purpose and univariate . . .," and I can understand why they thought so, as I did not explain otherwise. However, in practice, I have been applying cut-off sampling to surveys with many "purposes"/variables of interest/attributes, for many years, and I have also included multiple regression as a part of a small area estimation methodology with emphasis on total survey error, which includes simplicity of function to reduce processing and interpretation errors. An algorithm I wrote in the late 1980s for multiple-attribute coverage in a "certainty stratum" was the basis for my first cut-off sample at the U.S. Energy Information Administration (EIA) around 1990, but others have written other such algorithms, trading-off between coverage contributions of various respondents to obtain a quasi-cut-off sample. That is, if more coverage is needed for one attribute, but the potential respondent with the next largest size measure for that attribute has almost no contribution for other attributes, but a potential respondent that is almost as large for that attribute also contributes greatly to others, it may be taken instead. Considering that such measures of size are often random variables, and that coverage is only one consideration with regard to accuracy, it seems that this need not be extremely rigorous. However, for any "purpose," if coverage is too small, the relative standard error is liable to be large, and bias due to model-failure may become a factor. If coverage is large, however, we may include a great deal of nonsampling error, which is one reason why it is often helpful to use cut-off sampling at all. The smallest observations are often

Disclaimer: The views expressed are those of the author, and are not official U.S. Energy Information Administration positions unless claimed to be in an official U.S. Government document.

unreliable. In the final analysis, we have to consider many contributions to the total survey error, and how one may trade-off against another (see Knaub 2001, 2002, 2007; Royall 1970; and particularly Knaub 2010).

In the Benedetti et al. article, they take a good look at bias from an historical point of view, but there appears to be no study of variance for the current sample. Royall (1970) showed that under a model, the lowest variance can be achieved using a cut-off sample. Model-failure is the objection generally held against this, but Knaub (2010) addresses that issue. A cut-off sample basically puts the unobserved data where it can cause the least harm, especially when regression through the origin is best, which is generally the case with establishment surveys (see Brewer 2002). Regression models (ratios for regression through the origin) are very useful for providing predictions. The variance of the prediction error (Maddala and Lahiri 2009) can be made applicable to estimated totals (see Knaub 1999). The accuracy assessment made by Benedetti et al. is also an interesting contribution to understanding total survey error, in part.

On page 668 the authors note that in their example, they found that three-quarters of the population could be placed in the take-nothing part of their three-part categorization, with more than half of their sample in ". . . the completely enumerated stratum." They ascribed this to cases where ". . . a stratification variable is highly correlated with the study variable . . . ." Fortunately, it is often true that for an agency providing periodic official statistics, this will be the case. They report that this ". . . is the case here given the short time span between censuses." In my experience with official statistics, there is often high correlation, useful in regression models, with data that may even be years old, but frame changes have to be considered. Multiple regressors are also of use (Knaub 2003). Regression weights have to be considered (see Sweet and Sigman 1995; Steel and Fay 1995), but generally, the classical ratio estimator (CRE) is quite robust (Knaub 2005). However, this may need adjusting, such as a step-function for regression weights found in Knaub (2009) that is necessary because of more extreme problems with data near the origin, but above the quasi-cut-off for a given "purpose/attribute/study variable." There the problem was an inflated variance, likely because of both disproportionately large nonsampling error for the smallest observations, and outlier data that represented "start-up fuel" for electric power generation, rather than the main focus of our estimation. In a situation involving natural gas production in another survey, it may be the case that unobserved data are actually more variable, so such a step-function for regression weight may be useful, but in this latter circumstance, it may be helpful to give the unobserved cases a larger weight in that portion of the step-function. As always, attention to data groupings for purposes of modeling, and multiple regression where appropriate, may be helpful. Additional members of the Statistical Methods Team at the U.S. Energy Information Administration, and perhaps others, have made promising suggestions for that survey.

I wish to congratulate the authors for an interesting and informative article. They note that cut-off sampling has received too little theoretical attention in the past, which is the case, but as can be seen in the references, there has been some other theory applied. At the EIA we are now finding it useful to expand its application, and look forward to additional, informative results. Most of the work and innovations thus far have been done in the area of electric power sample surveys, but more is being accomplished now by a few members of the newly organized EIA Statistical Methods Team, and at least one other person in the

parent office. Benedetti et al. indicate that cut-off sampling is in widespread use and the methodology should be given additional thoughtful consideration. This would be very important for the advancement of official statistics.

References

Brewer, K.R.W. (2002). Combined Survey Sampling Inference: Weighing Basu's Elephants. Arnold: London and Oxford University Press.

Knaub, J.R., Jr. (1999). Using Prediction-Oriented Software for Survey Estimation, InterStat, August. http://interstat.statjournals.net/, partially covered in "Using Prediction-Oriented Software for Model-Based and Small Area Estimation." In Proceedings of the American Statistical Association, Survey Research Methods Section 1999, and partially covered in "Using Prediction-Oriented Software for Estimation in the Presence of Nonresponse," presented at the International Conference on Survey Nonresponse, 1999.

Knaub, J.R., Jr. (2001). Using Prediction-Oriented Software for Survey Estimation – Part III: Full-Scale Study of Variance and Bias, InterStat, June. http://interstat.statjournals.net/. (Note another version in Proceedings of the American Statistical Association, Survey Research Methods Section, 2001).

Knaub, J.R., Jr. (2002). Practical Methods for Electric Power Survey Data. InterStat, July, http://interstat.statjournals.net/. (Note another version in the Proceedings of the American Statistical Association, Survey Research Methods Section, 2002).

Knaub, J.R., Jr. (2003). Applied Multiple Regression for Surveys with Regressors of Changing Relevance: Fuel Switching by Electric Power Producers. InterStat, May, http://interstat.statjournals.net/. (Note another version in the Proceedings of the American Statistical Association, Survey Research Methods Section, 2003).

Knaub, J.R., Jr. (2005). The Classical Ratio Estimator. InterStat, October, http://interstat.statjournals.net/.

Knaub, J.R., Jr. (2007). Model and Survey Performance Measurement by the RSE and RSESP. Proceedings of the American Statistical Association, Section on Survey Research Methods, 2730–2736, http://www.amstat.org/sections/srms/proceedings/.

Knaub, J.R., Jr. (2009). Properties of Weighted Least Squares Regression for Cut-off Sampling in Establishment Surveys. InterStat, December, http://interstat.statjournals.net/.

Knaub, J.R., Jr. (2010). On Model-Failure When Estimating from Cut-off Samples. InterStat, July, http://interstat.statjournals.net/.

Maddala, G.S. and Lahiri, K. (2009). Introduction to Econometrics (Fourth Edition). Chichester, UK: John Wiley & Sons.

Royall, R.M. (1970). On Finite Population Sampling Theory Under Certain Linear Regression Models. Biometrika, 57, 377–387.

Steel, P. and Fay, R.E. (1995). Variance Estimation for Finite Populations with Imputed Data. Proceedings of the American Statistical Association, Section on Survey Research Methods, Vol. I, 374–379, http://www.amstat.org/sections/srms/proceedings/.

Sweet, E.M. and Sigman, R.S. (1995). Evaluation of Model-Assisted Procedures for Stratifying Skewed Populations Using Auxiliary Data. Proceedings of the American

Statistical Association, Section on Survey Research Methods, Vol. I, 491–496, http://www.amstat.org/sections/srms/proceedings/.

James R. Knaub, Jr.
U.S. Energy Information Administration
Statistical Methods Team
Email: James.Knaub@eia.gov