Journal of Official Statistics, Vol. 17, No. 1, 2001, pp. 103–106

Comment

Lynne Stokes[1]

Platek and Särndal have presented a thorough discussion of challenges that National Statistical Agencies face in producing high quality data and in making information available to users about its quality. My remarks will focus mostly on one specific aspect of their discussion: the absence of practical knowledge about evaluating survey accuracy among users and researchers.

I agree with the authors that a focus on customer satisfaction as a way to evaluate the quality of data a National Statistical Agency produces is dangerous. This practice is akin to universities using student course evaluations to assess teaching quality. On the positive side, it is a method that is cheap, easy, and politically correct (''Are the users or students not paying for it?''). It also provides clear-cut rankings that can be compared and monitored for change, which is useful for management. Indeed, the customer is best positioned to evaluate certain characteristics of the process of data delivery, like timeliness and accessibility, just as a student is to evaluate certain aspects of course delivery; such as how well the professor delivers lectures and how consistently he or she keeps office hours. Without some minimal quality in these areas, there would be few customers for the data or students in the course. However, high quality in these areas does not imply high quality in the content of the data or the course.

The typical data user is *not* in a position to evaluate the quality of the data itself. If they have been good students of statistics, then they may have some knowledge of precision. But virtually no user can assess data accuracy. One reason, as Platek and Särndal point out, is that information needed for the assessment is often not collected, or if it is collected, it is not provided to the user. But perhaps more importantly, we as theoreticians and practitioners have not done a good job of describing the relationships between measures of quality that are available, such as nonresponse rates, and estimator behavior. And we have done an even worse job of teaching these concepts.

When we train consumers of statistics, even those in the most rudimentary courses, we expend much time and effort on conveying knowledge about variability, both in a population and of estimators. In fact, emphasis on this concept is considered to be the modern approach to teaching of statistics. We make sure our students understand the relationship between sample size and confidence interval length, and we even develop animations they can watch to help them absorb the idea (see, e.g., www.math.uah.edu/stat/sample/index.html **or** www.stat.berkeley.edu/users/stark/Java/SampleMean). As a profession, we have worked hard on educating the masses about the meaning of margin of error (see, e.g., ''What is a Margin of Error?'' 1998 published by the American Statistical

[1] University of Texas at Austin, TX, U.S.A.

Association), and have been so successful that it is now common to hear news reporters describing polls with small differences between candidates as a ''statistical dead heat.''

But we have not made an understanding of the impact of bias on estimation a part of the core curriculum for introductory statistics. Perhaps we should work harder at explaining this concept to the masses. It could be made very concrete and of practical use to consumers of statistics. It is easy to imagine an applet allowing the student to explore the impact of the level of nonresponse and the difference between respondents and nonrespondents on confidence interval coverage, for example. For that matter, the topic of the effects of bias on estimation is not even covered in most sampling courses. Two pleasant exceptions are the text of one of the authors (Särndal, Swensson, and Wretman 1992, Section 5.2) and the computer program SURVEY (Chang, Lohr, and McLaren 1992). But suppose I wanted to give a problem to my students on how to assess the bias that would arise from undercoverage in a particular telephone survey of residents of South Texas, or to calculate a ''generalized design effect'' (one that includes bias) of an alternative dual frame survey of the region. I would have no resources to use for teaching such a topic in my sampling text, though comparisons of sampling error of equally complicated designs are well covered. Even the elementary conceptual idea that bias is of relatively less importance in small (compared to large) geographic areas is something that is not emphasized, as far as I know, in courses at any level. So why would the typical data user have an appreciation for this fact?

One reason that we do not include much about the effects of bias on estimators in our sampling curriculum is that the theoretical development leading to the results is not elegant, and some parts of it are not even particularly statistical. It is not that fun to teach. Also, the students in a sampling course cannot use the information to assess performance of a real survey because the information required to carry out the accuracy assessment must come from experimental data that is not available from the sample. In fact, much of the data that would be needed is not available at all, and most of us doing the teaching are not knowledgeable about what is available anyway.

I agree with Platek and Särndal that many of the practical issues related to assessing the affect of nonsampling errors on accuracy of estimators is under-researched. However, I do not share their belief that the lack of a unifying theory is the major problem holding us back. Rather, I think a more serious problem may be the restriction of what we accept as appropriate methods for producing knowledge. The assessment of bias from many types of nonsampling errors (such as nonresponse or undercoverage) does not require elegant mathematical development, but does require careful experimentation and considerable cost. Though experimentation is considered a worthy endeavor in many technical disciplines, it is not considered so in our field, at least not as worthy as mathematical proof. If all one does is establish empirically that one type of adjustment for telephone noncoverage for a certain category of questions is better than another in various populations, this is considered a result interesting enough for our journals only if one or another of those methods requires some interesting technical development. This is not necessarily bad, but it does mean that the practical issue of when one can expect a simple method to help accuracy and by how much is not a topic that researchers are encouraged to think about.

This contrasts with the situation that prevails in some other technical disciplines. For example, journals in database systems, a subarea of computer science/engineering, are

filled with papers that rely on empirical results to establish the goodness of new procedures. This methodology is so well established that there are industry standard databases, along with queries on these databases, which come from a university-industry coalition called the Transaction Processing Council (TPC). These databases are synthetic, and are built to mimic those collected by various types of business decision support systems. For example, the TPC-C benchmark is a database meant to behave like one from bank teller machines. Though originally developed for marketing purposes by companies who sell database systems, they are now used by industrial and academic researchers to test new algorithms. Industrial researchers also have access to and use real databases in experimentation. If an algorithm performs well on a real database, it is considered to be a higher form of ''proof'' than performing well on synthetic ones, but they are usually used only within a company or institution, and they sometimes have to be encrypted to be cited in a paper. The field certainly does not seem to have suffered from this approach, as evidenced by the fact that new developments are occurring at a very rapid pace. It would be useful to many in our research community if benchmark data sets were available. They could be used, for example, for comparing methods of adjusting for the bias of various kinds of nonsampling errors.

Of course one difference between the database systems field and ours is that if an algorithm that performed well in a test database does not perform well on the user's database, the user will recognize a problem; for example, he or she may notice a long response time to a certain type of query. Also, the user is better able to recognize the *gain* in performance that is realized from improved algorithms. Our data users have no way to observe improvements in accuracy except in special circumstances, such as if the data are part of a series, or if they have data on similar topics from different sources. Thus, our users may be less concerned about and less willing to pay for improvements in the data, so we are not so pushed by market pressures to improve our methods as database system developers are.

I believe that we will find, however, that the uses to which our data are put are changing in a way that makes the user more aware of poor data quality. Nowadays, many data vendors sell survey data (much of which is repackaged government data) to companies for merging with their transaction data. The purpose is to enrich the set of variables available for building predictive models, one of the activities in the collection of methods known since recently as data mining. If the survey data, or the transaction data for that matter, is of poor quality, it should be exposed by poorer predictions. It will be easy for a company to compare vendors on the basis, for example, of how much ''lift'' their variables supply to the prediction. I expect these users will have more concrete methods to discriminate between data producers than a general feeling of trust, which the authors suggest to be the major source of users' confidence in data now.

Because missing data and selection bias are problems that are commonly encountered in data mining, I believe that it will not be long before researchers in this field, many of whom come from the fields of computer science and machine learning, will begin producing research about how to deal with nonsampling errors. There are already many papers in the field dealing with sampling errors, addressing such questions as how to tell when a sample is big enough to produce a stable model (e.g., Provost, Jensen, and Oates 1999). This literature also includes papers on how to tell when a database has changed enough that a model needs to be refit, which is really a question about bias

due to coverage (e.g., Kelly, Hand, and Adams 1999). In fact bias is generally of bigger concern than variance in these applications because of the size of the data. The results produced by these researchers are empirically based and very practical, and a computer code for implementing them is frequently made widely available. So the good news is that we may have some practical answers to some of our problems about data accuracy before long. But the bad news is that we may not be producing the answers. I think this news is bad because researchers in our field do have insights to offer about methods for handling nonsampling errors, but future commercial software systems (at least in data mining products) may not incorporate the best of our ideas.

In summary, I too believe that we as statisticians working in survey methods are not yet doing all we can to provide information about the quality of the data we produce, nor to educate users to understand the information we do have. One reason is that we do not have all the answers about how to evaluate accuracy. Even if our approach is only to continue to accumulate best practices, I believe we can make progress in this area.

References

American Statistical Association (1998). What Is Margin of Error? ASA Series ''What Is a Survey?''

Chang, T.S., Lohr, S., and McLaren, C.G. (1992). Teaching Survey Sampling Using Simulation. The American Statistician, 46, 232-237.

Kelly, M.G., Hand, D.J., and Adams, N.M. (1999). The Impact of Changing Populations on Classifier Performance. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 367-371.

Provost, G., Jensen, D., and Oates T. (1999). Efficient Progressive Sampling. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99), 23-32.

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

Received December 2000