

## DISCUSSION

Eugene P. Ericksen, Temple University  
Philadelphia, PA 19122

The papers by Rossi and Quiroga and associates<sup>1</sup> illustrate problems applied statisticians often face. When sampling frames are lacking, problems of measurement are serious, and we cannot supervise data collection directly, "standard statistical procedures" provide only limited guidance. We should calculate standard errors, but realize that they won't tell us very much. We can try to estimate other, nonsampling, components of error but this is difficult. In the end, we will probably rely on a great deal of subjective judgement. To evaluate the uncertainty of estimates, we need to identify key assumptions, make reasonable changes in these assumptions, and then see how the estimates change as a result.

Research on census data (Bailar, 1976) shows that for many variables, errors due to sampling are a small part of total error. Inconsistencies of measurement combine with interviewer effects to create substantial nonsampling errors which dominate sampling errors even for small areas. Converse and Traugott (1987), in their review of errors in political polling data, show that the errors of these polls are substantially greater than the sampling errors commonly stated by the pollsters. For these types of data, the sampling errors give necessary, but minimal, guidance to understanding the uncertainty with which the estimates must be regarded. For studies of the missing, the homeless, and the victim, sampling errors are even less useful. Mean squared errors reflecting biases and nonsampling variances will be much greater than variances and these mean squared errors are difficult to estimate. The importance of the topics, though, makes the statistician want to do the best job (s)he can. This may lead to estimating the likely direction of bias and the approximate magnitude of mean squared error.

What kinds of rules should we try to apply to such data? Does the predominance of nonsampling error, often hard to measure, render the data useless? When do we decide that the estimates are good enough to use? These are questions with uncertain answers, and different statisticians will come to different conclusions for the same set of data. I would like to suggest three criteria which might help to form conclusions.

The first is that the statistician should provide as much information as possible on the structure of error. If measurements are uncertain, we need repeat measures on the same event. If we suspect interviewer effects, we need to have the repeat measures taken by different interviewers. If we are uncertain about the sampling frame, then we might want to take two different samples from different sources, perhaps merging them with a dual systems estimation strategy.

The second is that once we have evaluated the error from internal information, we should ask if our results are consistent with external information. If the social service agencies of Chicago list 10,000 homeless people that they are serving, then we would regard an estimate of 2,300 as far too low. If the estimated number of atrocities in Chile decline during a period of known unrest, then we would conclude that our intelligence was not good enough.

The third is that we should state the problem clearly and evaluate the costs of drawing erroneous conclusions. This will help us to judge the size of tolerable error, and to decide whether an overestimate is worse than an underestimate. For example, when homeless people are dying on the streets of Chicago, it is worse to underestimate their number if the underestimate will cause the authorities to reduce services for the homeless. While none of these criteria provide certain results, they do help us to decide whether our numbers are good enough to use, and this is what matters in today's uncertain world.

### Rossi's Study of the Homeless

Rossi has entered an emotional debate over the true number of homeless people in the United States. Advocates of the homeless have estimated their numbers to be in the millions, and charge that their increased numbers demonstrate the inhumanity of President Reagan's social and economic policies. The U. S. Department of Housing and Urban Development (HUD) has responded by commissioning a study of the homeless. The study concluded that the true number of homeless was much smaller than the advocates' number, and was probably between 250,000 and 350,000. Neither the advocates' nor HUD's estimates were based on very good data and neither group has even a good guess of the true number. The problem is made worse by the lack of a definition of a homeless person.

Rossi has tried to develop a method for estimating the size of one component of the homeless population - the literal homeless. For \$250,000 he has produced an estimate which is probably too low and which has only limited information on its error structure. Moreover, it provides no information on the size of the total homeless population that Chicago social workers attempt to serve.

Rossi's method demanded much of interviewers' abilities and dedication. They were asked to search a sample of city blocks in the dead of night, trying all doors which might not be locked, peeking in all parked cars, checking the possibly homeless status of all customers of all-night diners, and finding, waking, and interviewing all persons sleeping in public spaces or abandoned buildings regardless of their state of health or possible inebriation. Presumably, no attempt was made to find homeless people riding buses or subways through the sample blocks. It is not clear what was done about possibly homeless people walking through the blocks. There are many reasons to believe that interviewers, who usually worked without supervisors present, would underestimate the number of homeless. Among them are: (1) the procedure was lonely, dangerous, and time-consuming, even with the police escort, (2) it is embarrassing to quiz every patron of an all-night diner of their possibly homeless status, (3) many of the homeless go to great lengths to disguise their status, even to the point dressing up and lying to interviewers, and (4) many of the homeless are not stationary, particularly at night when it is colder.

With these obvious problems, I find it curious that Rossi invested so many resources in taking the survey twice. Even the standard errors of the first survey are likely to be substantial underestimates of the total error. Taking a second survey reduces a small component of total error only moderately and provides no information regarding nonsampling error. It would

---

<sup>1</sup>These are the only two papers I have received in written form.

be far more informative had the funds been used to estimate nonsampling errors in the first study. For example, if two interviewers went to the same block on the same night, did they get the same result? For those homeless people giving their names, how many were known to social welfare agencies? Did the presence of the police encourage interviewers to find more homeless people? What was the quality of data provided by the homeless respondents? Looking at it another way, rather than a second calculation of the standard error, it would be more informative to know how many people the interviewers encountered in the field, how many were talked to, and what were the criteria first for deciding who to talk to and who was homeless. How consistently were these criteria applied?

Next, we need to know how Rossi's numbers compare to other available estimates. If he believes the estimates of 15,000 to 25,000 preferred by advocates are too high, then what do better informed local providers of services think? If experts "guesstimates" differ from Rossi's sample estimates, we need to know whether the experts are simply making wild guesses, whether the experts have a definition of the homeless population that differs from Rossi's, or whether Rossi's estimates of the "literal homeless" have serious negative biases. Rossi merely writes he has no way of estimating the validity of his responses.

Third, Rossi gives little guidance about how his numbers might be used. Social workers help many more homeless people than those found on the street at night. Some are in hospitals or prisons; others are placed in welfare hotels or special apartments obtained for homeless people. Still others may have housing for part of the month, but go on the streets when their welfare or social security checks run out.

To help solve the problems of how to provide adequate services to the homeless, we need to know what local agencies are doing, what their costs are, and how many people they are taking care of who would otherwise be homeless. This more complete information would advise policy makers on the amount of additional resources needed to serve the homeless. Strategies for providing these services are, of course, another question.

#### The Paper by Quiroga and Associates

Quiroga and colleagues write on a subject that most of us prefer not to face. Even one atrocity is horrifying, and when their numbers are cumulated, the effect is chilling. When such statistics are presented, readers must be convinced that they are true, and extended detail increases their credibility.

It is important that data be collected in a way that minimizes double-counting. Secrecy is important to a repressive government, and readers will generally realize that many murders, disappearances, and tortures will not be counted. With care taken to prevent duplication, readers will know that the counts are underestimates, that the situation is worse than the numbers indicate, and that action must be taken. This is an example of a case where the known direction of error increases the effectiveness of the estimates.

I suggest keeping at least the following data items in a computerized data base: name of victim, age, sex, description of physical features, home address (if possible), date of human rights violation, location of violation, nature of violation, identity of informant, and whether or not confirming information was obtained from a separate source. Security considerations may necessitate blurring some of the

information in the computer. For example, rather than giving the name of the informant, it may be better to write a category such as "relation" or "friend" on the data base, and keep the informant's name in a separate place. Use of the computer, though, will enable a more powerful analysis.

When a violation is reported, checks should be made to make sure that the same violation was not previously reported. This can be done by checking reports made on the same date, but also by letting one variable fail to match and looking for violations matching all other characteristics. For example, the same violation could have been reported on a different date. If the violation is confirmed, some information on the confirmation should be reported. Is the second informant likely to have heard about it from the first informant, how similar are the two accounts, and did each observe the event directly?

Care should also be taken when reporting statistics to indicate the quality of data. Categories should indicate whether or not the event was confirmed, and whether the confirming information was consistent. It might also be a good idea to identify the number of cases where information was provided by a relative, by a close friend, by an acquaintance, or by someone else. Events known only through newspaper reports should also be identified. Again, the idea is to take care not to overestimate the number of human rights violations. The statistics have greater power when the reader realizes that they are underestimates.

Other information might also be helpful in presenting results. For example, the time series on human rights violations in Chile would perhaps have greater impact if dates of important events were also shown. The consistency of the trend in the number of violations with events such as the attempted assassination of President Pinochet increases the persuasiveness of the data. Secondly, it is likely that the completeness of reporting varies geographically, and we would expect it to be better in cities. If separate tables were shown for urban and rural areas, with the sizes of population bases also shown, we might have a more accurate picture of the level of repression.

#### Summary Comment

These two papers consider topics where good data are difficult to obtain, and it is not easy to control the quality of data, sampling frames are lacking, and measurement errors are substantial. Although one paper considers sample data and the other tries to collect information on the total population, the important sources of error on the two projects are similar. Statisticians need to develop better ways of thinking about data for situations like these. Given importance of the topics, avoiding the calculation and evaluation of imperfect estimates is not an option. Policymakers will develop policies for homeless people and human rights violations whether data are available or not.

Bailar, Barbara (1976), "Some Sources of Error and Their Effect on Census Statistics," Demography, 13:273-286.

Converse, Philip E., and Traugott, Michael W. (1986), "Assessing the Accuracy of Polls and Surveys," Science, 234:1094-1098.