1/18/01

Why Sampling is not Polling

There is a considerable amount of controversy over the potential adjustment of the 2000 Census. It is therefore important to understand what the census is attempting to do, how sampling can help to improve the accuracy of the census and how both the census and sampling are different from polling.

Most polls are collected when somebody needs information quickly and cheaply. A good poll is based on a scientific sample, i.e., persons are selected randomly and interviewers are told to interview these randomly selected people and no one else. A good poll also has a carefully written questionnaire that the interviewer follows to the letter. One of the best-known uses of polling is to predict the outcome of an election. Polls are also used to obtain opinions on subjects of national interest such as whether healthcare reform is needed or the federal government should do something about global warming. They are also used to test reactions to certain products or marketing campaigns.

Polls usually have rather small sample sizes. A survey with 1,000 respondents will typically have a "sampling error" of "plus or minus 3 percentage points" which is an everyday way of saying that had the survey included everybody, the likely result would be within 3 percentage points of outcome obtained from the sample. This is sufficient precision for most purposes, even though the 1,000 respondents are a tiny fraction of the total population.

Errors due to sampling are not the largest source of error in a survey, however. One very important problem that most polls face is the fact that many people simply will not participate. For a good poll, taken by a high quality survey research organization, the interviews are likely to contact over 1,500 potential respondents in order to get 1,000 to be interviewed. Many others refuse even to answer their telephone. This hypothetical survey is likely to have another 2,000 to 3,000 numbers classified as "non-contacts," meaning that the interviewer did not get past an answering machine, or did talk to an actual person who told the interviewer to call back later. In sum, it might take 3,000 to 4,000 residential telephone numbers to get 1,000 actual respondents. Well under half of all potential respondents even agree to be interviewed. After the interviews are completed, and the polltaker prepares to publish the survey results, (s)he needs to find a way to account for the differences between those who do and do not agree to be interviewed.

The biggest problem in a poll is respondents providing misinformation. Many respondents never think about the questions the interviewer asks, and the questions are not generally on their minds when the telephone rings. As a result, respondents frequently give answers that are "off the top of their heads" and have not been thought about very much.

Often times the behavior that the survey is concerned about differs from the attitudes expressed in the survey. Democrats turn into Republicans when it actually comes time to vote – many others who state a candidate preference don't even cast their votes. To cite another example, respondents who state great optimism when asked about the future of the economy may become very conservative and risk-averse when actually developing their own investment strategies. The divide between stated attitudes and actual behavior is perhaps the greatest problem facing survey takers these days.

The 2000 Census

The census is quite different from a poll. Every ten years the Census Bureau attempts to gather information about each household in the United States. The questions asked on the census are not opinion questions, but rather questions that pertain to the number of people in the household and the age and sex of each member. While every household in America should respond to the census, many do not get forms, do not return their forms or are otherwise not correctly counted. Because some people are omitted, or left out, of the census, while others are erroneously counted, the Census Bureau needs a method to test the accuracy of their data.

The use of sampling is intended as a quality control check on the entire census. The Census Bureau has selected a demographically representative sample of 314,000 households[1], and it hired a team of expert interviewers to determine who should have been counted at each of the households. The Census Bureau is using this sample in two ways – to calculate rates of omission and counting error. Omissions are people who should have been counted, but in fact were not. The Census Bureau believes that there were about 8 million of these in 1990. Counting errors occur in three main ways. First, some people get counted twice at the same address, and these are called duplicates. Second, other people get counted at a place that is not their main residence, e.g., at a vacation home. Third, census enumerators sometimes get incorrect information at addresses where no one answers the door. The neighbors who do supply the information may get it wrong.

The Census Bureau subtracts the number of counting errors from the number of omissions to estimate the net undercount – the difference between the number of people actually counted and the number that should have been counted. It will calculate separate estimates of the net undercount for 64 different groups of people. These groups are defined by variables such as race, Hispanic origin, residential location, and how difficult the census was to take in their local neighborhoods.

In this manner, the Census Bureau will be able to assess its level of error when taking the 2000 Census. In particular, it will be able to assess whether or not the differential undercount – the difference in the net undercounts (of minorities and non-Hispanic Whites) was as large as it has been in all the other most recent censuses. Because this quality check has nothing to do with opinions, preferences, or voting

---

[1] A sample size of 1,000 is typical for many opinion surveys. The Census Bureau has decided to survey 300 times this number to ensure more accurate results.

patterns, it should not be referred to as "polling." Instead, it is a quality control check on the accuracy of census data collection. As such, it follows the same logic as the quality control check of a manufacturing process intended to calculate the percentage of items that are defective.

Eugene Ericksen is a Professor of Sociology and Statistics at Temple University and a Special Consultant to National Economic Research Associates (NERA). He has taught undergraduate and graduate courses on statistics and survey research. His research has investigated census taking, as well as the general topics of survey and statistical methodologies.