Census 2000: Count on Controversy
Author(s): MARK SCHILLING
Reviewed work(s):
Source: *Math Horizons*, Vol. 6, No. 2 (Nov. 1998), pp. 20-22
Published by: Mathematical Association of America
Stable URL: http://www.jstor.org/stable/25678191
Accessed: 06/11/2011 18:00

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Mathematical Association of America is collaborating with JSTOR to digitize, preserve and extend access to *Math Horizons.*

http://www.jstor.org

Census 2000: Count on Controversy

One of a statistician's main jobs is to estimate unknown quantities. "Basic training" in the science of estimation occurs in the typical undergraduate course in mathematical statistics, which focuses on problems of the following kind: A random sample of size n is drawn from a large population of individuals or items. The goal is to estimate the value of some important characteristic of the population, referred to as a parameter. For example, one might wish to estimate the mean lifetime of a certain brand of light bulb, or the proportion of bulbs that will fail before the time given in the manufacturer's guarantee. Values of the variable of interest are measured for each member of the sample, then used to form the parameter estimate.

The subject of mathematical statistics addresses the question of how to efficiently use sample measurements to estimate parameters. A sample is represented mathematically by n independent random variables with a common probability distribution. This distribution models the variation in a certain variable (e.g., lifetimes of light bulbs) throughout the population.

As normally only a portion of the population is sampled, an estimate will almost certainly deviate from the actual parameter value. The resulting difference is known as the *sampling error*. By analyzing the performance of various estimators mathematically it is often possible to determine an optimal procedure that minimizes the likely size of the sampling error.

MARK SCHILLING is Professor of Mathematics at California State University, Northridge.

Not a Classroom Problem

Counting the number of residents of the United States on April 1, 2000, the next official census date, is an entirely different sort of estimation problem from the type described above. First, the parameter of interest is the unknown size of the population itself. And the word *census* means a complete enumeration of the population—that is, the *sample* is the *entire population*, not a proper subset of it. Clearly with a full enumeration, there is no sampling error, and in principle no error whatsoever in the result. This is why censuses receive scant mention in most mathematical statistics courses, where the main goal is to quantify and minimize the error that arises due to sampling.

The United States Census takes place, however, not in the classroom but in the real world, and there is indeed a considerable error in the result. Counting all U.S. residents is a daunting task that has become harder with each census. Here are some of the reasons:

- At close to 270 million people, the U.S. population is larger and more diverse than ever before. The first census, directed by Thomas Jefferson, found less than 4 million residents.
- The Bureau conducts the initial and primary phase of the census by mail. However, the rate of voluntary response to the mailed census questionnaire (before further prompting) has declined significantly, from approximately 85% in 1970 to 75% in 1980 to only 63% in 1990.
- Large numbers of people are homeless or live in rented garages, trailers, etc. Clearly these individuals are extremely difficult to find and count. Many, such as criminals and illegal aliens, do not *want* to be found.
- More people enter and leave the country than ever before. This includes undocumented workers such as crop pickers who migrate according to seasonal patterns.

With a projected cost for the 2000 Census of approximately $4 billion dollars, the United States Census is the most expensive estimation problem in the world. To appreciate the difficulties in obtaining an accurate count, consider the following analogy provided by Tommy Wright, chief of the Statistical Research Division of the U.S. Bureau of the Census [1]: Suppose someone asks you to count the number of people in attendance at a local high school basketball game, and to make the count during half-time. Many spectators will go for refreshments or to the restroom, while some will leave the arena altogether. Some fans will switch seats. Players, coaches and referees will be in the locker rooms. Could you just use the ticket count? No, because some persons are admitted without tickets and some who purchased tickets do not come to the game. Any counting procedure you attempt certainly will miss many people while counting others more than once.

The difference between the actual population and the estimated population is known as *measurement error*. One of the Census Bureau's main goals is to minimize this error for the United States as a whole. But there is more: When Congress established the United States decennial census in 1790, its main purpose was to provide a basis for a fair apportionment of congressional seats to each state in the Union. In this century

Illustration by Tom Mann.

the census has also become used for dividing the enormous amounts of federal funds (billions of tax dollars per year) that are distributed to the states. Thus the Bureau attempts to obtain counts that are as accurate as possible for the United States as a whole as well as for each of the many thousands of local regions that together comprise the U. S., all while attempting to limit costs.

Sampling: More From Less?

Although the national census is conducted only once every ten years, the Census Bureau continually conducts surveys and analyses for other purposes. This makes it possible to obtain an idea of the measurement error in a given decennial census. For the 1990 Census, the official count was 248,709,873 people, while evidence from other surveys and demographic analyses indicated a true population of approximately 253,000,000. The 1990 Census therefore produced an *undercount* of around 4 million people. Each census since at least 1940, and possibly all of them, have evidently resulted in a sizable number of persons being missed. However, the 1990 Census was the first to have an estimated undercount larger than that of the preceding one, despite being the most expensive census on record.

How might the Bureau reverse this trend? Surprisingly, improving the accuracy of the census (reducing the measurement error, i.e., the undercount) may require utilizing a certain amount of selective sampling (thus introducing some sampling error). A more accurate count might be obtained by *not trying* to count everybody. Accordingly, here is what the Census Bureau has proposed for the 2000 Census:

The United States population is divided into approximately seven million *census blocks*—essentially, neighborhoods, each containing perhaps 15 to 50 housing units. A *tract* is a contiguous collection of such blocks. The Bureau has formulated a plan to employ conventional counting methods in the first phase, then use sampling to bring the count up to at least 90% of the total population (as estimated by prior surveys by the Bureau and other agencies and by demographic projections) in each tract. That is, within each tract whose initial response rate is below 90%, a random sample of the nonresponding units would undergo follow-up efforts to contact enough individuals to reach the 90% quota. (Initially, the Bureau had planned to let census enumerators use their own discretion in deciding how to achieve the 90% quota, but received criticism for the subjectivity and potential biases that enumerators might introduce into the process.)

Finally, the Bureau would make a statistical estimate of the number of remaining persons missed, using the information obtained from the units that responded in the follow-up sample. Totaling the conventional count with the follow-up count and the estimate of those missed by both counts would then give the overall estimated population of each tract. The Bureau would then estimate the total U.S. population by summing over all of the tracts in the United States.

Census Bureau officials have also devised a strategy for using sampling to provide a quality check on the accuracy of the estimated counts. What's more, this plan offers the potential to actually *improve* the original count: Entirely independently of the main census count, the Bureau proposed selecting a random sample of approximately 25,000 blocks nationwide. Census workers would visit each housing unit within these blocks and attempt to determine the populations of these blocks. As in the main census, some people will again be missed.

The Bureau would then have *two* numbers for each region of the country and for the country as a whole, both expected to be wrong—specifically, too small. How can this information be used to create one potentially very good count?

The answer lies in a time-tested statistical procedure known as the *capture-recapture method.* Imagine that you wish to estimate the number of perch in a certain lake. By fishing simultaneously in several areas carefully distributed within the lake your research team catches, tags and releases 40 perch (capture). Now you send your team out to fish again. This time they capture 50 perch, of which 8 are tagged. Since one-fifth (8 out of 40) of the tagged fish caught are tagged (recapture), one could conclude that approximately one-fifth of *all* of the perch in the lake were caught this time. Thus the total number of perch in the lake is estimated as 50 × 5 = 250. (This assumes, of course, that tagged fish were neither *more* likely nor *less* likely to be caught than untagged fish. Conceivably, tagged fish could be more gullible than other perch in the lake. Conversely, they could be more cautious than untagged fish during the recapture phase, having been caught once already.)

The capture-recapture estimate above can be written as (40 × 50) ÷ 8. Similarly, then, the Census Bureau's capture-recapture estimate of the human population of a given region would be the product of the two census counts described above divided by the number of people found in both counts. The reliability of these new figures will depend on assumptions analogous to those involving the likelihood of tagged perch being recaptured. Specifically, it is certainly plausible that people who were hard to find by the mail portion of the survey and the follow-up to achieve 90% coverage will also be hard to find by the door-to-door survey of 25,000 blocks.

Will They or Won't They?

The Census Bureau believes that sampling has the potential to produce a more accurate census at lower cost. The National Academy of Sciences, the American Statistical Association, and several other independent agencies support that view. Yet it now looks as if sampling may not be used at all for Census 2000. The United States Census, unfortunately, has become a political football.

The reason is that the level of undercounting varies dramatically by income status and ethnicity, being much higher for poor communities and for people of color. Partly as a result, the undercount also tends to be greater for large urban centers. Potential voters in these categories are more likely to be Democratic than Republican. Thus it is not surprising that the President and most Democratic governors and legislators support the use of sampling to enumerate missing individuals. Conversely, Republicans almost universally oppose sampling.

Although political disputes over the census are not new—George Washington issued the first presidential veto in disagreement with Congress' interpretation of the 1790 census figures—the contemporary level of controversy over how the census should be conducted is unprecedented. Numerous lawsuits were heard over the last several years regarding the 1990 Census. The issue was whether to use the original census count or a different number that attempts to correct for the undercount. The Supreme Court ultimately ruled in favor of the original, unadjusted count [2].

The 2000 Census is embroiled in a similar controversy. On August 24th of this year, a special panel of the federal appeals court ruled that the Census Bureau may not use statistical sampling in the next census. The Justice Department has appealed this decision to the Supreme Court, which will hear arguments at the end of this month. A final ruling is expected no later than March 1999.

References

1. Tommy Wright, "Sampling and Census 2000: The Concepts", *American Scientist*, May-June 1998, 495–524.

2. Margo Anderson and Stephen E. Fienberg, "An Adjusted Census in 1990: The Supreme Court Finally Decides", *Chance*, Vol. 9, No. 3, 1996, 4–9.

Bowling Green State University / Department of Mathematics and Statistics

The mathematics and statistics program at BGSU is nationally ranked and one of the strongest in Ohio. We are large enough to offer a wide array of courses and research opportunities, excellent library and computing facilities, and yet small enough to provide our students individual attention in a liberal arts atmosphere. We have 30 faculty and over 60 fully supported graduate students. Strong research areas include Algebra, Analysis, Applied Mathematics, Probability, and Statistics. We offer the Master of Arts in Mathematics and Statistics, the Master of Arts in Teaching Mathematics, the Master of Science in Applied Statistics, and the Doctoral degree in Mathematics and Statistics. Please consider visiting our campus. For further information, contact:

BOWLING GREEN STATE UNIVERSITY 1910

Graduate Secretary
Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, OH 43403-0221

(419) 372-2637
FAX (419) 372-6092
cpatter@math.bgsu.edu
http://www.bgsu.edu/departments/math/