Research and Methodology Directorate

A History of the Survey of Income and Program Participation and Disclosure Avoidance

By Laura McKenna

Issued April 2019

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

INTRODUCTION[1]

The U.S. Census Bureau conducts the Survey of Income and Program Participation (SIPP) under Title 13, U.S. Code, Section 9 mandate to not "use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports (13 U.S.C. § 9 (2007))." The Census Bureau applies Disclosure Avoidance (DA) techniques to its publicly released statistical products in order to protect the confidentiality of its respondents and their data. None of the information in this paper is confidential.

HISTORY OF THE SIPP

The SIPP is a national, longitudinal survey of households that collects monthly data and provides information that can be used to analyze the economic situation of the U.S. population, <www.census.gov/history/www/programs/demographic/survey_of_income_and_program_participation.html>. It offers data on income, taxes, assets, liabilities, and transfer programs. It is useful in evaluating the effectiveness of federal, state, and local programs, <https://en.wikipedia.org/wiki/Survey_of_Income_and_Program_Participtation>. SIPP also collects information on family dynamics, educational attainment, housing expenditures, health insurance, disability, and child care.

SIPP began in 1975 in the Department of Health, Education, and Welfare (HEW). A few years later, HEW collaborated with the Census Bureau to develop it into a longitudinal survey. The first test of this was in 1978. In 1979, there was a representative national sample of 8,200 households. In 1981, HEW lost funding for the survey, but the Census Bureau received funds to continue SIPP in 1983.

The original design of SIPP was a nationally representative sample of households, where all people aged 15 or older were interviewed every 4 months over a period of 32 months following the initial interview. Every February, a new group of respondents was interviewed so there was an overlapping series of survey panels. Funding became a problem in the late 1980s. Some panels had to be grouped with others, and panel sizes decreased. SIPP was redesigned in 1996, <http://nap.edu/12715>. The Census Bureau then began interviewing a single, larger panel over a 4-year period, interviewing each household three times each year. Computer Assisted Interviewing also began in that year allowing for faster production.

In 2014, SIPP was again reengineered, <http://nap.edu/24864>. The goal was to reduce respondent burden and program costs and improve accuracy, timeliness, accessibility, and relevance. The sample size is now approximately 53,000 households with a panel length of 4 years. Respondents are interviewed annually with questions about the previous calendar year. Interviewers conduct SIPP interviews with personal visits, using computer-assisted personal interviewing. Previously, core questions about income and transfer programs appeared in every wave, and additional "topical modules" included questions on things such as child care and disability. With the reengineering in 2014, this was changed, and a portion of topical module content was integrated into the main body of the SIPP interview, <www.census.gov/programs-surveys/sipp/about/sipp-introduction-history.html>.

DA FROM THE 1990S TO THE PRESENT

Tabular Data

Due to the small sample size with large weights, SIPP tabular data are published for very large areas for data quality reasons. No additional DA techniques were deemed necessary to protect the data.

Microdata

Removal of Direct Identifiers

The Census Bureau removes direct identifiers from the file such as name, address, phone number, etc.

Geographic Threshold

All geographic areas identified must have a population of 250,000 or more. When calculating this population, all geography related variables on the file are cross-tabulated to obtain the final population count of an area that can be identified as a piece of geography. This population threshold is larger than the 100,000 threshold for the Current Population Survey and the American Housing Survey. The Microdata Review Panel, predecessor of the Disclosure Review Board[2] (DRB), determined in the early 1990s that SIPP should

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

[2] Census Bureau data products must be approved before dissemination by the DRB.

have a higher population threshold than some other surveys because the Public Use Microdata File (PUF) contained a very large amount of detailed income items not available on most surveys, and the panel was concerned about an inflated disclosure risk.

After that ruling, the Census Bureau could not publish a SIPP PUF for all states because Primary Sampling Units (PSUs) were public information, and when crossing those areas with all geographic variables on the PUF, some states did not make the 250,000 threshold and had to be combined. A few years later, the Census Bureau decided not to reveal PSUs for any survey to the public, so all states can now be identified.

Topcoding and Bottom-Coding

The Census Bureau uses topcoding and bottom-coding to eliminate outliers in a file for continuous variables such as wages and salary. A topcode (cutoff) is in place for 0.5 percent of all values or 3 percent of all nonzero values, whichever is the larger of the two. Bottom codes are the same except on the other side of the distribution. A bottom code might be applied to gross income. For variables that are part of a sum, the individual summands are topcoded prior to their summation. Originally, all topcoded values were replaced with the topcode cutoff itself.

Beginning in 1996, the topcoded values were replaced with the mean of the topcoded values. There are a small number of variables that still use the topcode cutoff itself (rather than the mean). At least three values from three different respondents must be topcoded or the topcode is lowered to meet this requirement. Bottom codes are the same except on the other side of the distribution. A bottom code might be applied to gross income or years received for a given benefit. For variables that are part of a sum, the individual summands are topcoded prior to the summation.

A small number of variables, such as benefits from SNAP, TANF, SSI, and Unemployment Compensation, use the maximum benefit amount as the topcode.

Rounding/Recoding

Each category of a categorical variable must contain at least 10,000 weighted people or households (depending on the universe for that particular variable) nationwide. If a category does not meet this threshold, it must be combined with other categories until it does. Some variable's categories, such as language spoken at home and immigration status, are combined even more.

Dollar amounts must follow one of two rounding/recoding schemes.

Round to two significant digits, or use this recoding scheme:

- Zero rounds to zero.
- 1 to 7 rounds to 4.
- 8 to 999 rounds to the nearest multiple of 10.
- 1,000 to 49,999 rounds to the nearest multiple of 100.
- 50,000 and greater rounds to the nearest multiple of 1,000.

Any totals or other derivations are calculated using the rounded numbers.

Noise Infusion

Noise infusion is used to hide very unusual characteristics of a person or household at a given point in time. For example, consider a woman with sextuplets or a 10-year-old in college or a household with 13 people in it. Such unusual circumstances are often well known and sometimes in the news. Census Bureau editing procedures capture and alter many, but not all, of these types of unusual circumstances.

Unlike many Census Bureau surveys and censuses, SIPP collects month of birth, but noise is added to this variable. In addition, SIPP is longitudinal and changes in personal or household characteristics can often be found in public records. For example, a birth, death, marriage, or divorce would be reflected in the SIPP while a given household is in sample.

The Census Bureau does not publicly release the details of how noise is added to protect these types of data that pose a disclosure risk.

PROMISING CONFIDENTIALITY

There were two related problems that affected the Census Bureau's promise of confidentiality to a subset of respondents, both stemming in a way from SIPP respondents.

First, in 1996, President Clinton signed The Personal Responsibility and Work Opportunity Reconciliation Act (also known as the Welfare Reform Act). One section of the Act charged the Census Bureau with continuing the collection of data from the 1992 and 1993 SIPP panels to evaluate the impact of the law with a focus on welfare and children. The Census Bureau then developed the Survey of Program Dynamics (SPD) to carry out this mandate, <www.census.gov/history/www/programs/demographic/survey_of_program_dynamics.html>. Part of the survey collected data about whole households and part of it (Self-Administered Questionnaire or SAQ) collected data about adolescents 12 to 17 years of age in those households, <www.census.gov/srd/papers/pdf/sm98-08.pdf>. When staff presented both microdata files (one from each part) to the DRB, the Board was very concerned. The adolescents had been told that their answers to the survey questions (several about sex, alcohol consumption, and drug use) would be confidential. However, due to overlapping variables in the adolescent survey and the household survey, the two microdata files could be easily linked, meaning that parents in the households could identify their children's responses. The DRB denied the request, and staff that had worked on the adolescent part of the survey were very upset. The DRB took the issue to the executive staff. It was decided that the adolescent SPD microdata file could not be publicly released, but would be available at the Federal Statistical Research Data Centers only, a great disappointment for many people.

Second, in 1997, staff became concerned about a situation that occurred during a SIPP interview. A female member of the household had given previous interviews, answering questions for the whole household. When she was unavailable for a subsequent interview, her husband was asked to be interviewed, and Census Bureau staff reminded him of his wife's responses in order to reduce respondent burden by shortening the length of the interview and making the questions easier to answer for the new respondent. This was common practice. Unfortunately, in previous interviews, the wife had revealed that she had been previously married. She had never told her husband about the previous marriage, and he was very upset. This lead to the implementation of the Respondent Identification Policy, <www.researchgate.net/publication/237521296>.

Because of these two incidents, the Census Bureau then had to rewrite the documentation on confidentiality protection that accompanies all household surveys and censuses, explaining that confidentiality between members of the same household cannot be protected.

THE FUTURE

Recently, the Census Bureau has embarked on an aggressive effort to replace its legacy DA methods with modern DA techniques based on formal privacy methods, <https://privacytools.seas.harvard.edu/formal-privacy-models-and-title-13>. Current methods will gradually change with the introduction of formal privacy (Nissim et al., 2018). Most of the current Census Bureau's DA research is focused on formal privacy for all types of data (Nissim et al., 2007). An algorithm operating on a private database of records satisfies formal privacy if its outputs are insensitive to the presence or absence of any single record in the input (Dwork, 2006). The DRB is quickly learning about formal privacy and how it protects Census Bureau data products.

REFERENCES

C. Dwork, "Differential Privacy," International Colloquium on Automata, Languages, and Programming (ICALP), 2006, pp. 1–12.

K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing, 2007, pp. 75–84.

K. Nissim, T. Steinke, A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, D. O'Brien, and S. Vadhan, "Differential Privacy: A Primer for a Non-technical Audience (Preliminary Version), Harvard University Privacy Tools for Sharing Research Data, 2018, <http://privacytools.seas.harvard.edu>.