Research and Methodology Directorate

A History of the Current Population Survey 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 monthly Current Population Survey (CPS) for the Bureau of Labor Statistics 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 CPS

In the 1930s, there was a great need to accurately determine the nation's monthly unemployment due to the Great Depression. There had been prior efforts to measure unemployment that ranged from simple guesses to enumeration counts.

The research staff at the Work Progress Administration (later, the Work Projects Administration or WPA) began developing techniques for measuring unemployment in the late 1930s. They ran local tests and then went national with the Enumerative Check Census taken as part of the 1937 Census of Unemployment. That was the first attempt to estimate nationwide unemployment using probability sampling. This experience lead to the WPA's monthly Sample Survey of Unemployment in March 1940, <www.census.gov/history/www/programs/demographic/current_population_survey.html>. This survey is also often called the Monthly Report of Unemployment, <www.bls.gov/respondents/cps/history.htm>. The survey attempted to classify people as working, looking for work, or not in the labor force.

In August 1942, the responsibility for this survey was given to the Census Bureau, and the survey's name was later changed to the Current Population Survey. The Census Bureau substantially revised the sample methodology in October 1943. The households were now in 68 Primary Sampling Units (PSUs) that included 125 counties and independent cities. By 1945, there were 25,000 housing units in sample and 21,000 interviewed. In 1954, the number of PSUs increased to 230, but the number of households in sample (and the Census Bureau's budget for the survey) did not increase. The redesigned sample was the result of a more efficient field operation and lead to more accurate estimates.

The CPS is now a monthly survey of about 60,000 eligible households representing the civilian noninstitutional population. It is conducted by the Census Bureau for the Bureau of Labor Statistics (BLS). It classifies people aged 15 and older as employed, unemployed, or not in the labor force. The survey is employment-focused, enumerator-conducted (by phone or in person), continuous, and cross-sectional. Households are in sample for 4 months, then out for 8 months, then in for another 4 months. The households and the people in them can be linked together for the 8 total months that they are in sample. The latest increase in sample size was the addition of 10,000 households in July 2001, <https://en.wikipedia.org/wiki/Current_Population_Survey>.

There is a basic CPS accompanied by different monthly supplements. The largest supplement is known as the Annual Social and Economic Supplement (ASEC, formerly known as the March supplement) which collects data on income and health insurance. Other supplements focus on topics such as school enrollment and food security. The DA techniques described below are applied to the basic CPS and all supplements as required by what types of variables are on each file.

DA FROM THE 1990S TO THE PRESENT

Tabular Data

The Census Bureau and the BLS have a memo of understanding that lets BLS employees with Special Sworn Status (SSS) access the Census Bureau CPS internal files. Those employees create national-level tables from the basic CPS. Any tables that go below the national level for the basic CPS are created using model-based estimates. The Census Bureau's Disclosure Review Board[2] (DRB) does not review these tables. Tables from CPS supplements are created from the public-use microdata files and are at the national level. Due to the small sample size with large weights,

[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.

CPS 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 100,000 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.

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. Originally, all topcoded values were replaced with the topcode cutoff itself. Later in 1996, the topcoded values were replaced with the mean of the topcoded values. At least three values 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. For variables that are part of a sum, the individual summands are topcoded prior to their summation.

In 2011, CPS ASEC topcoded values began being replaced with values generated from a technique called Rank Proximity Swapping, <https://cps.ipums.org/cps/inctaxcodes.shtml>. The technique preserves the distribution of values, while maintaining adequate DA, <www2.census.gov/programs-surveys/cps/techdocs/cpsmar17.pdf>. People/households with values above the topcode are sorted and ranked by those values from lowest to highest, and those values are swapped between the people/households within a given rank interval. All values must be swapped. The bounded interval is large enough to include many people/households in order to protect the data and small enough to ensure that the swapped values are within proximity of each other. The parametric details of this are confidential.

In addition, all of the values are rounded to two significant digits, <https://www2.census.gov/programs-surveys/demo/datasets/income-poverty/time-series/data-extracts/pu-swaptopcodes-readme.docx>.

Rounding/Recoding

Each category of a categorical variable must contain at least 10,000 weighted people or households (depending on the universe of the variable) for that particular variable nationwide. If a category does not meet this threshold, it must be combined with other categories until it does.

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.

In addition, CPS 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 CPS 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.

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>.