Errors in Census and American Community Survey Data Files

Jaime Trujillo

Carnegie Mellon University, Pittsburgh, Pennsylvania

Summer 2012 National Science Foundation Research Training Group, Advisor: Professor William F. Eddy

Introduction

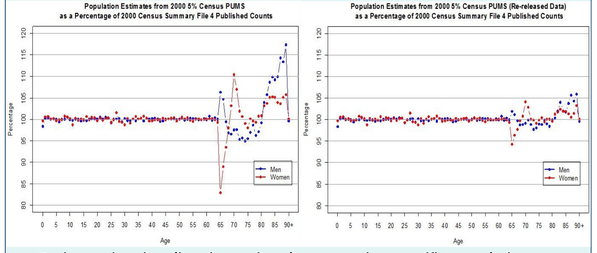
Goal: To find errors in Census and American Community Survey (ACS) data files and examine their impact in analyses which incorporate such data. The data files that were used are available online to the public through the Census Bureau's website and the Integrated Publice Use Microdata Series (IPUMS) website. Some of these data files are called Public Use Microdata Series files, often referred to as PUMS; they are a sample of the actual responses from the ACS and include most population and housing variables found in the full, original dataset. The surveys listed below were all conducted by the Census Bureau.

Surveys:
2000 5% Census
2006 1-Year ACS
2007 1-Year ACS
2008 1-Year ACS
2009 1-Year ACS
2010 1-Year ACS

IPUMS:
Integrated Public Use Microdata Series is a project operated by the Minnesota Population Center which consists of microdata samples from the United States and international census records. The IPUMS database for the U.S. comprises of samples from fifteen censuses between 1850 and 2000 and ACS samples from 2000 to 2010. Such data files are already available on the Census Bureau website; however, IPUMS aims to provide a user-friendly, data extraction system that enables users to combine these samples and select only the variables they require. Since IPUMS does not manipulate the data values in the Census Bureau files, values from IPUMS files should match exactly to the values in the equivalent Census Bureau files.

Consequences of Disclosure Avoidance Techniques

2000 5% Census: 1-in-20 national random sample of the population. This is a weighted sample, meaning the records in the national file will have an average original weight of 20 since it's a 5% census.



•Both graphs visualize the national age-gender specific population estimates from the 2000 5% Census PUMS as a proportion of the 2000 Census Summary File 4 (SF4) published counts. However, the right graph uses re-released PUMS data.

•Each point is calculated by the quotient: $\frac{\text{2000 5\% Census PUMS estimate}}{\text{2000 5\% Summary File 4 published estimate}}$

Ex.) Left graph, 895,052 is the PUMS estimate for 65-year-old women in 2000 and 1,079,328 is the SF4 estimate: 895,052 ÷1,079,328 = 82.927%

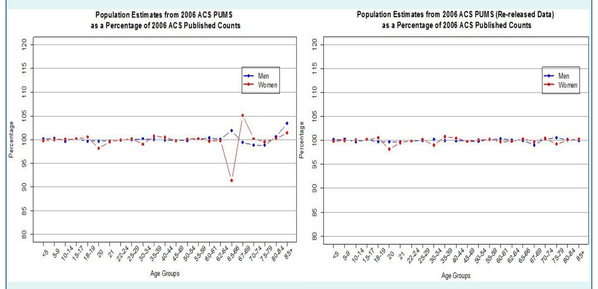
•Left graph: replicated graph of published findings[1] shows for men and women starting from age 65 and up, there are substantial differences in population estimates from their respective published counts.

•Right graph: re-released PUMS data greatly reduces the effects of disclosure avoidance techniques but still see differences as much as 5% for elderly men and women.

1. http://www.nber.org/papers/w15703

(cont'd)

2006 1-Year ACS: 1-in-100 national random sample of the population. This is also a weighted sample so the records in the national file will have an average original weight of 100 .

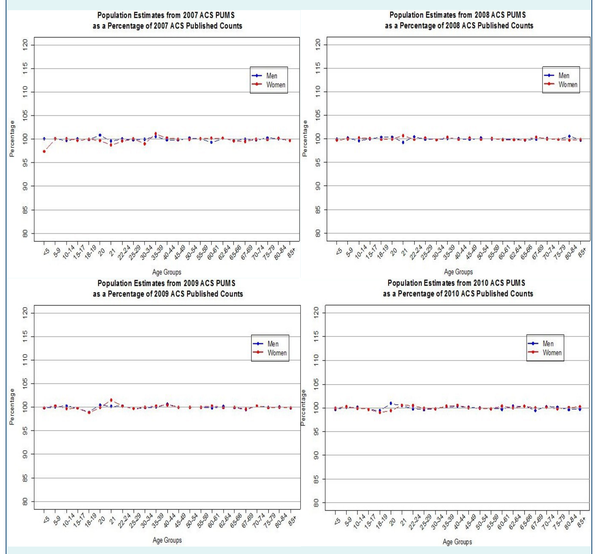


•Left graph: replicated graph of published findings[1] again shows for men and women starting from age 65 and up, there are significant differences in age-sex group specific population estimates from their respective published counts, especially for women.

•Right graph: re-released PUMS data here eliminates the effects of disclosure avoidance techniques (i.e. age perturbation).

Testing for Similar Age and Sex Data Errors

2007, 2008, 2009, 2010 1-Year ACS: 1-in-100 national random samples of the population. These are all weighted samples.



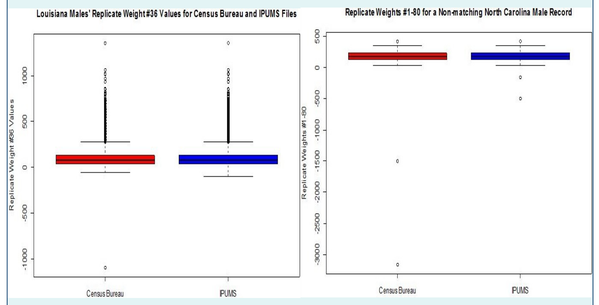
• Unlike 2006, from 2007 to 2010 the 1-Year ACS PUMS age-sex group specific population estimates do not differ significantly from their respective published counts for elderly men and women.

•Top left graph, for 2007 the largest difference between PUMS estimates and their respective published counts is about 3% at the first data point for women which represents age group "less than 5 years old."

•For 2008, 2009, and 2010 graphs, the 1-Year ACS PUMS age-sex group specific population estimates do not differ more than about 1% from their respective published counts for both men and women.

Non-matching Replicate Weights Between Sources

2010 1-Year ACS: The following graphs in this column use replicate weights from this survey to compare values between equivalent files from the Census Bureau website and the IPUMS website.



•Left graph shows a replicate weight #36 outlier from Census Bureau Louisiana data file that is not in the equivalent IPUMS file.

•Right graph shows two replicate weight low outliers from one Census Bureau North Carolina male record that do not match the lowest outliers from the equivalent IPUMS record.

Standard Error Variables from Replicate Weights

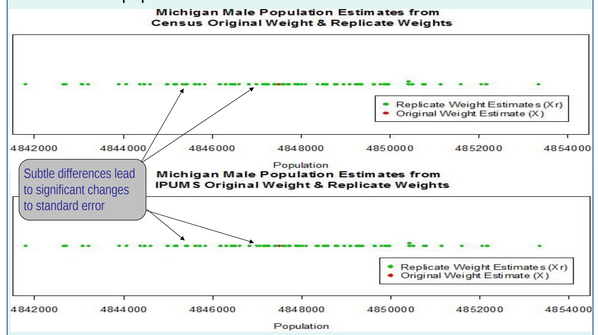
•Formula for calculating standard error:

•These strip charts plot the X and X_r values for calculating Michigan male and female population standard errors.

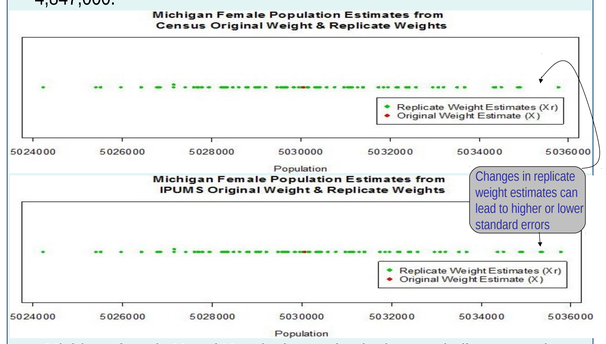
$$SE(X) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}(X_r - X)^2}$$

X = the estimate based on the original weight

X_r = the 80 individual estimates based on each of the replicate weights.



•Graphs for X and X_r values of 2010 1-Year ACS Michigam male records from Census (top) and IPUMS (bottom) look very similar except for replicate weight estimates at about 4,845,500 and 4,847,000.



• Michigan female X and X_r stripcharts also look very similar except in their higher replicate weight estimates.

Errors in Variance Estimation

2010 1-Year ACS: The following tables also refer to this survey and examine the consequences of non-matching replicate weights.

Non-matching Replicate Weight Values

State	Row Number	Sex	Replicate Weight #	Census File Value	IPUMS File Value
Louisiana	15590	Male	36	-1096	-96
Louisiana	15591	Male	36	-1096	-96
Michigan	9179	Male	38	-1856	-856
Michigan	21794	Male	62	-1505	-505
Michigan	29137	Male	62	-1381	-381
Michigan	38218	Male	38	-1083	-83
Michigan	39825	Female	38	-3455	-455
Michigan	72719	Male	62	-1214	-214
Michigan	77849	Female	62	-1319	-319
Michigan	79962	Male	62	-1684	-684
Michigan	84828	Male	38	-1274	-274
North Carolina	41554	Male	2	-3160	-160
North Carolina	41554	Male	40	-1499	-499

Standard Errors (SE) and Margin of Errors (MOE) for Affected State-Gender Populations

State	Population Estimate	Census File SE	IPUMS File SE	Census MOE	IPUMS MOE
Louisiana Males	2,226,234	3028.461	3055.057	4981.818	5025.569
Michigan Males	4,847,509	4884.339	4762.097	8034.738	7833.65
Michigan Females	5,030,065	4869.109	5031.851	8009.684	8277.395
North Carolina Males	4,656,342	5559.939	5725.497	9146.1	9418.443

•Row numbers given for specific-state ACS PUMS files from Census Bureau website and value is the same for equivalent IPUMS ACS file.

•The first table shows all of the Michigan replicate weight values that did not match were either from replicate weight #38 or #62.

•Louisiana had both of its non-matching values from replicate weight #36 in successive rows and North Carolina had two non-matching values in a single row (record).

•All of the replicate weight values that did not match were negative and the digits from the IPUMS file values were a subset of the digits in the correct Census file values.

•For Michigan males, the errors in replicate weights decreased the standard error for its population estimate while for Michigan females, the opposite developed.

•Margin of errors are calculated for a 90% confidence interval.

Conclusions

•The re-released 2000 5% Census and 2006 1-Year ACS data eliminate or greatly reduced the effects of disclosure avoidance techniques evident in the original data.

•Such errors in sex and age data were not found in 1-Year ACS PUMS files for the years: 2007, 2008, 2009, and 2010.

•2010 1-Year ACS PUMS files from IPUMS contained replicate weight values that did not match those in the corresponding Census Bureau PUMS files for the states of Louisiana, Michigan, and North Carolina.

•These non-matching replicate weight values led to non-matching replicate weight estimates.

•Replicate weight estimates are used to calculate standard errors of population estimates and thus, data gathered from IPUMS led to incorrect standard errors.

•IPUMS must correct these errors because researchers that use 2010 1-Year ACS PUMS replicate weights (from Louisiana, Michigan, and/or North Carolina) from their site to calculate standard errors for population estimates may end up drawing incorrect inferrences.

Acknowledgements

I would like to thank Professor Bill Eddy, the Carnegie Mellon Statistics Department, and the National Science Foundation for giving me the opportunity and resources to work on this research and for their help and guidance in my analysis.

I would also like to thank both the Census Bureau[2] and the Minnesota Population Center[3] at the University of Minnesota for providing the PUMS files and the Census Bureau for conducting the surveys.

2. http://www.census.gov/
3. http://usa.ipums.org/usa/