Investigation of Data Release Rules for Medians and Zero Estimates for the American Community Survey

Karen King, Michael Starsinic
U.S. Census Bureau
Washington, DC 20233

Keywords: American Community Survey, Data Quality/Reliability Indicator, Data Release Rules

Abstract

The American Community Survey (ACS) uses various techniques to improve the quality of the 1-year and 3-year data products it releases based on a series of data release rules. The release rules use the size of an estimate's coefficient of variation (CV) as a quality/reliability indicator. Products with too many estimates having large CVs fail the release criteria and are not published. This research is in response to feedback from ACS users that the current release rules are too conservative, especially those applied to certain types of estimates. We look into alternatives to the current release rules as they are applied to ACS 1-year and 3-year medians and zero estimates.

Introduction[1]

The American Community Survey (ACS) is a continuous survey that collects the data historically collected by the decennial census long form sample. Full implementation of the ACS began in January 2005, with the sample expanding to a size of approximately 2.9 million housing unit addresses, with sample in all counties and county equivalents in the 50 states, the District of Columbia, and Puerto Rico.

One major design goal of the ACS has been to produce useful estimates of high reliability. Reliability concerns arise when estimates are subject to high sampling variability because this variability limits the usefulness of the data. As sampling variability increases, the reliability of the estimates decreases. It is up to the ACS program to decide what data are released to the public. The methods used by the Census Bureau to improve the reliability of published ACS data include:

- Minimal population publication thresholds,
- Data reliability rules that removes specific data products with high levels of sampling variability, and
- to some extent the design of ACS data products.

This research focuses on the current data reliability checks and looks into possible changes to these checks as applied to the 1-year and 3-year detailed tables, specifically for two types of estimates: medians and zero counts estimates. This work is in response to some of the criticism from external ACS data users that the current data reliability rules are at best too conservative or at worst questionable for these two types of estimates (Navarro and Garrett). Subject matter analysts within the Census Bureau have also

[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.

pressed for changes to the current rules. Some suggest that more liberal rules are desirable for zero estimates which would increase the quantity of released table data. This would allow users to decide for themselves (with help from provided measures of sampling variability) if estimates are reliable enough for their purposes. Others feel the current reliability rules are inappropriate for median estimates. A few alternative rules for each were examined and are discussed below.

Background

The ACS program takes on the responsibility for deciding if survey estimates are sufficiently reliable to be released to the public. Not all users of survey estimates have sufficient statistical knowledge to understand how statistics and estimates are produced. In addition, they may not necessarily understand the things that may affect the quality of estimates, that is, sampling and non-sampling error. One distinguishing feature of the set of ACS data products is that an estimate of the associated sampling error is published with each estimate. For 2005 and earlier products, the 90 percent confidence intervals were provided, and the margin of error (MOE) has been provided since 2006. Measures of non-sampling error are also published and referred to as Data Quality Measures. These measures are presented in the American Fact Finder (AFF) in the B98 table series. The general user often refers to census data or survey estimates for information to base funding and policy or business decisions. Most of the time these users take the information provided at face value and don't consider the reliability of those estimates.

A single year's worth of sample in the ACS is not adequate to publish statistically reliable estimates for all geographic areas for which Census 2000 long form estimates were published. Instead, single-year estimates are published only for geographic areas with a population size of at least 65,000. For smaller areas, multiple years of ACS sample are pooled together to create "period" estimates. The first estimates based on three years of pooled ACS data were published in 2008 for all areas with a population size of at least 20,000 using data collected from January 2005 through December 2007. All geographic areas, including Census tracts and block groups, will be published using five years' worth of pooled ACS data. The five-year data will first be published in 2010 for data collected for the years 2005-2009.

For the 1-year and 3-year ACS releases, about 1,500 data products are created with some containing hundreds of individual estimates, for thousands of different geographic areas - over 6,000 areas for 1-year data and over 13,000 areas for 3-year data. That adds up to hundreds of millions of estimates released each year. The Census Bureau realizes that despite population size thresholds, not all of the estimates that are produced are of high reliability - many may be questionable as they are based on only a handful of sampled observations, and others may be the result of not having any sample cases in that geographic area having those characteristics.

The most detailed set of ACS estimates are released in a set of data products called detailed tables. Initially ACS data products, particularly the detailed tables, were initially designed to be "… comparable with the Census 2000 Summary File 3 to allow comparisons between data from Census 2000 and the ACS. However, when Census 2000 users indicated certain changes they wanted in many tables, ACS managers saw the years 2003 and 2004 as opportunities to redefine ACS products based on users' advice". (US Census Bureau, 2009, pages 13-3).

The ACS has chosen to address the problem of estimates with low reliability by instituting several data release rules which identify tables with unacceptable levels of estimates with low reliability and prevents their publication. Many highly detailed tables have a simpler predefined version where some of the estimates are collapsed together making it easier to meet the reliability requirements. These "collapsed" tables were developed in order to provide some data for a given topic when the full detailed table fails. Unfortunately, even collapsed tables may not meet reliability requirements when checked.

About 90 percent of all detailed tables are count data tables. This means each line in the table presents an estimate of the total number of people with a particular characteristic within a particular area (such as a state). As a result, the data reliability rules were optimized for testing count tables. In the early years (2000 through 2004) of the ACS, count tables were removed from publication based on a reliability rule that required the count table to be supported by a weighted count of at least 500 and an un-weighted average of 2 cases per cell. (U.S. Census Bureau, 2006) This reliability rule attempted to both minimize the disclosure risk and reduce the number of published estimates with high levels of sampling error. It was revised because the rule was found to be biased against small geographies. It suppressed too many good estimates fit for most uses of the data for small governmental units while letting too many tables of poor quality be released for large governmental units.

Starting with the 2005 ACS release, new data reliability rules were applied to each eligible detailed table for publication. These rules incorporated a measure of the reliability of each estimate in the table. Each estimate is subject to sampling variability that can be estimated by the coefficient of variation (CV). The CV is defined as the standard error (SE) of the estimate divided by the estimate itself. Higher CV values are associated with low reliability estimates. The inspection of a detailed table begins with the coefficient of variation (CV) being calculated for each estimate or line in the table. If the median CV, and thus, more than half of the CVs of all *detailed* lines in the table (those that are not the total line or a subtotal line) are greater than 0.61, then the whole table fails and it will not be published for a particular geographic area. There are a few caveats and exceptions to the rule which will be part of our discussion later.

The cutoff value is set to 0.61 because, at that value (1/1.645 rounded to two decimal places), the 90 percent margin of error is equal to the estimate itself, and for larger CVs, the margin of error is larger than the estimate. In other words, for estimates with CVs of 0.61 or higher, the estimate is not significantly different from zero at the 90 percent confidence level. We are not attempting an actual statistical test here. If there is at least one case with a characteristic, then the population count for that characteristic must be nonzero. This is simply a means of identifying – and giving a plausible statistical justification for – a reasonable cutoff value.

Fewer detailed tables and estimates are released under the current reliability rules than the previous rules. The current reliability rule more efficiently identifies tables with the greatest data reliability problems. The current rule fails about 37 percent of count tables annually. However, since the operation was designed to target "whole" tables, it cannot ensure that all estimates in tables that are released are reliable. At the same time, in some instances reliable estimates are not released when they are included in a table that contains a majority of questionable estimates. Fortunately very few reliable estimates suffer this fate.

Table 1: The CV Distribution for ACS Estimates Withheld Due to Reliability Issues

Year	Total Estimates Withheld	cv<.1	.1<cv<.2	.2<cv<.3	.3<cv<.4	.4<cv<.5	.5<cv<.61	cv>.61	est=0*
2007	72,928,466	3.0%	4.9%	5.2%	5.1%	4.8%	5.2%	23.1%	48.7%
2005-2007	146,739,018	3.2%	4.6%	4.9%	4.8%	4.6%	5.0%	22.5%	50.3%

- "est=0*" contains zero estimates and a few special cases for median and ratio estimates where either the estimate or the standard error could not be calculated.
- Starsinic, 2009

Table 1 shows the distribution of the CVs of estimates that were withheld from the public after applying the data reliability checks. As we see about half are zero estimates (est=0*), roughly a quarter are estimates with CVs greater than 0.61, and only three percent are estimates with a CV less than 0.1. The release rules attempt to strike a balance between minimizing the number of questionable estimates being released to the public and maximizing the number of reliable estimates being released to the public.

Data Release Rules for Medians

A. Current Rules and Potential Alternatives

Some ACS tables only include median values such as median age by sex or median household income. In the 2007 ACS there were about 795,000 median detailed tables containing roughly 5.6 million estimates. About six percent of these median estimates were based on either one or no sample cases. These medians are represented in the detailed tables by "-" and the margin of error by "**". In these cases, the estimate and/or the standard error of the median is zero resulting in a CV that is undefined and cannot be used to measure sampling variability. In this situation where there are no unweighted cases, it could be because the underlying characteristic to support the median is rare or there is no one with that particular characteristic. If rare, then the assumption of the underlying characteristic and associated median being unstable is true. However, if over time the estimate is consistently zero then stability is suggested.

The data reliability rule for median tables is the direct application of the table-based 0.61 median CV rule described earlier with one exception. The exception is any median with an undefined CV is ignored in the computation of the table's median CV. This means if there are four median estimates in a table and one of them was undefined, then the reliability rule would be based on the CVs of the other three median estimates. As the result of applying this rule, about 95 percent of the median tables containing 98 percent of the median estimates were published in 2007.

Application of similar data release rules for median tables and count tables implies that CVs for medians and counts behave in the same manner. However, it has been shown this is not the cases particularly for median estimates based on fewer than 10 un-weighted cases (Navarro and Garrett, 2009). We suggest that any median based on a small number of cases is unreliable no matter the calculated CV value. A CV for a median behaves differently from a CV for a count estimate, because it is influenced by the size of the characteristic that is being estimated to a greater extent, than the size of the population producing the distribution. For example, if the median income estimates of $20,000 and

$80,000 had the same standard error and were based on roughly the same number of unweighted cases, the CV for the $20,000 estimate would be four times as large as the CV for the $80,000 estimate. Thus, the CV may not be an appropriate indicator of reliability for medians.

Concerns have been expressed that the current reliability rule is allowing the release of too many questionable medians, specifically medians that are undefined, based on fewer than 10 cases, or with a CV> 0.61. Table 2 shows the relationship between the number of un-weighted cases making up the estimate and the proportion of medians and counts estimates published after the reliability checks have been applied.

Table 2. Proportion of Published Median and Counts Estimates by the Distribution of Un-weighted Cases Making up the Estimate, ACS 2007

	Number of Un-weighted Cases				
Estimate Type	2-5	6-10	11-20	21-30	>30
Medians	91.1	95.2	97.9	99.2	99.1
Counts	37.5	62.0	71.2	76.9	87.8

Beginning with the first column, around 91 percent of median estimates based on 2 to 5 cases were released to the public while the proportion of count estimates based on the same number of cases is around 38 percent. As we look across the first row, for medians we see the proportion of medians that are published continues to be over 90 percent no matter how many cases are used to calculate the estimate. In the second row, for count estimates, the current reliability check seems to be working at removing potentially questionable estimates. A little more than a third of count estimates based on small samples are published. As the number of un-weighted cases increases, the CVs tend to improve for the count estimates and more pass through the checks and are published.

Looking at it another way, Table 3 shows the proportion of median and count estimates with CVs in each range that were withheld. So the value of 0.3 percent in the cell for medians in 2007 with a CV of less than 0.3 means that of all the median estimates produced from the 2007 ACS with CVs below 0.3, only 0.3 percent were withheld as a consequence of the data release restriction. The comparable rate for counts was over 16 percent. The distributions for the 1-year and 3-year estimates are similar.

Table 3. The Percent of ACS Estimates within Each CV Range Withheld by Year and Estimate Type

Years	Estimate by Type	Total Estimates	Estimates Withheld	CV < 0.3	0.3 <CV < 0.61	CV> 0.61	Undefined CVs (zero estimates
2007	Medians	5,548,270	104,726	0.3%	2.1%	17.4%	3.5%
	Counts	142,221,921	68,489,163	16.0%	40.0%	69.4%	88.7%
2005-2007	Medians	11,585,795	202,551	0.3%	2.0%	17.1%	3.5%
	Counts	298,743,777	137,452,458	14.3%	39.0%	68.4%	88.5%

While only cutting out about 2 percent of all median estimates, the current rule fails about 17 percent of all median estimates with CV>0.61. It is not successful at cutting out undefined estimates, as only about 3 percent of them failed by the current rule. Where as for counts, overall 48 percent of estimates fail in the current rule which removes about 69 percent of cases with CV>0.61 and about 89 percent of undefined CVs. This demonstrates the poor result by the current reliability checks at removing the more questionable median estimates using the CV as measure of reliability.

Several alternative reliability rules present themselves that may produce more satisfactory results for medians:

- Option 1 – Modify the current release rule for medians by including undefined CVs by assigning them a value of 1.0. This assumes that any undefined median is unreliable. This adjustment in the current rule will tend to increase the chances of median tables failing and hopefully remove more estimates with CV > 0.61.

- Option 2 – Modify the current rule for medians by including the undefined CVs set equal to 1.0 and, for purposes of this test, assign the CVs of medians that are based on fewer than 10 un-weighted cases a value of 1.0 as well. Again because we assume that any median based on a small number of cases is unreliable no matter the calculated CV, this would tend to increase the chances of median tables failing and hopefully remove more estimates with CV > 0.61.

- Option 3 – Look at each median estimate individually and suppress medians with CV>0.61 within a table while ignoring the undefined CVs. The table itself is not tested for reliability, but would be withheld if all the estimates had CV > 0.61.

- Finally, Option 4 – Keep the current reliability rule unchanged, but find an alternative measure of the sampling variability to go into the test for just median estimates. Currently, a successive differences replicate system is used to generate standard errors for all ACS. An alternative standard error may better reflect the median estimate's reliability when used in the formation of the CV.

  Initially, two potential methods for calculating alternative standard errors were suggested. These were the Woodruff method and the Francisco-Fuller method. A review of literature identified several papers that worked with the Woodruff and the Francisco-Fuller methods. One paper actually compared these two methods to each other and to two variations of the half sample replication methods.[2] As part of their conclusion, they stated the results were comparable for the Woodruff and the Francisco-Fuller methods, but the Francisco-Fuller was far more difficult to use. For this paper, the analysis will include only the results of the Woodruff method. Any continuation of this research should include the Francisco-Fuller method in the analysis.

B. Methodology for Medians Research

[2] Dorfman and Valliant, 1993

The analysis uses median data from the 2007 ACS 1-year and the 2005-2007 ACS 3-year detailed tables. All the median detailed tables (full and collapsed versions) were included in the study. The analysis is done separately for 1-year and 3-year detailed tables. The analysis examines both the impact on the number of tables and estimates within these tables that pass and resulting quality of the published estimates by alternative options. The "current rule" refers to the 0.61 median CV table-based rule used for detailed median tables.

C. Results for Median Estimates.

Results are based on about 755,000 median detailed tables containing roughly 5.5 million estimates for 2007 and 1.6 million tables containing 11.6 million estimates for 2005-2007. Table 4 shows the simulated publication rates for 1-year and 3-year medians tables and estimates by the first three alternative release rules.

Table 4. Simulated Percent of Median Tables and Estimates Published by Option

	2007 ACS		2005-2007 ACS	
	Tables	Estimates	Tables	Estimates
Current (Undefined ignored)	94.8%	98.1%	95.3%	98.3%
CV=1 for Undefined	87.2%	95.3%	87.1%	94.9%
CV=1 for Undefined and n<10	79.4%	81.4%	79.3%	81.4%
Estimate with CV>.61	NA	93.5%	NA	94.0%

The first row gives the baseline. Under the current rule for medians, 95 percent of median tables and 98 percent of median estimates are published. The second row gives the simulated results of the first option which was assigning a CV=1 for cases with undefined CVs for the application of the current release rule. The table shows a probable drop in both the 1-year and 3-year percentages of tables published to 87 percent and estimates to about 95 percent. This is more than double the percentage of tables and of estimates withheld from the public compared to the current rule for medians. In row three, the simulated results of Option 2 – additionally assigning a CV = 1 for all medians based on fewer than 10 cases, shows the percentage of tables and estimates published as about 79 percent and about 81 percent respectively. Finally for Option 3 which fails individual estimates with CV>0.61, the results are about 93 percent of the estimates being published, more than three times the number of estimates being withheld compared with the current rule for median tables. Table 4 also shows the results for the three options using 3-year data. The failure rates are about the same as seen in the 1-year results.

Table 5 shows the simulated distribution of published medians by the CV values for the current rule and the first three alternatives.

Table 5. The Simulated CV Distribution of Published Medians 2007 Estimates by Option

Option	Estimates Published	CV < 0.3	0.3 <CV < 0.61	CV> 0.61	Undefined
Current	5,443,544	74.4%	11.0%	5.5%	6.0%
CV = 1.0 for undefined	5,286,697	76.4%	11.3%	5.5%	4.5%
CV = 1.0 for undefined and n< 10 cases or	4,517,419	80.9%	9.6%	4.7%	3.1%
Estimate CV > 0.61	5,186,465	78.3%	11.8%	0.0%	6.5%

In the first column, we see the number of median estimates that would be published followed by the distribution of published medians by their CV values. The first row shows the results of the current rule for 2007 medians to act as a baseline. Over 5.4 million median estimates were published in 2007. Three quarters of published estimates had a CV<0.3 and as we move across the row there is a decrease in the distribution of median CVs with 6 percent undefined.

In the second row, we have the first option, assigning a CV= 1 for undefined medians and considering them in the application of the check. For this alternative, indications are that we would see a slight drop in the number of published medians compared with current level and a shift in the distribution of CVs. Those with CV<0.3 increased and those with undefined CVs dropped as expected.

The third row has the option where we assign a CV = 1 for all medians based on fewer than 10 cases in addition to all medians that are undefined. These results indicate a shift in the distribution of the estimate CVs. The proportion of medians with CV < 0.3 increased and those that were undefined dropped. With this option, there would likely be less than half as many undefined median published compared with the current rule and about a thirty percent drop for medians with a CV> 0.61.

Finally, in the last row, there are simulated results of withholding estimates individually with CV> 0.61. The indications are that we would see fewer highly unreliable estimates published because we are dropping any estimate with CV > 0.61. This would target the estimates of most concern while leaving the rest of the estimates alone.

Our fourth option was to find an alternative measure of the sampling variability for medians alone. For this paper, the analysis looked at the impact on the current rule of using the Woodruff method to calculate the standard. Table 6 shows the percent of median tables that were withheld or published crossed by the method used to calculate the standard error.

Table 6. Percent of 2007 County Level Median Tables by Publication Status and Method Used to Calculate the Standard Error.

Median Tables		Successive Difference Replicate Method		
		Withheld	Published	Total
Woodruff Method	Withheld	2.3%	1.4%	3.7%
	Published	1.2%	95.1%	96.3%
	Total	3.5%	96.5%	100.0%

Here we see some initial results of failure rates for 2007 median tables at the county level by method used to calculate the SE. The successive difference replicate method (the current method) results are on the vertical and the Woodruff method on the horizontal. Overall the percentage of median tables being published or withheld after the reliability checks is about the same for each method. There are a few tables that went from being published to being withheld and visa versa. From this view we see no advantage of using the Woodruff method, but analysis of it will continue.

Data Release Rules for Zero Count Estimates

A. Current Rules and Potential Alternatives

For smaller geographic areas, empty cells or count estimates of zero are common in detailed tables. This is often true for large heavily detailed tables even in the largest geographic areas. The CV for this type of estimate is undefined since the value in the denominator is zero. For purposes of determining whether the count table should be released, these undefined CVs are assigned a value of 1.0.[3] This practice increases the chances that the table will fail the reliability test and guarantees failure if at least half the cells have an estimate of zero. This treatment demonstrates the current assumption that zero counts are unreliable (unstable) estimates, although that is debatable, and we will address this issue later. As we have seen in earlier tables, under that assumption the current rule does a good job, i.e., about 90 percent of zero estimates are withheld (Table 3).

The assumption of instability may not be true for instances where the count is consistently zero over time because there are no individuals with that particular characteristic in that particular geographic area. For example, in the 2006, 2007, and 2008 ACS detailed tables, Table C05006 shows there was no one born in Iran, Israel, or El Salvador residing in Montana. These may be very accurate estimates.

Three alternatives were considered to the current method for zero count estimates.

- Option 1 - Apply the current rule but do not consider the undefined CVs of zero estimates. In this case, only the CVs of the nonzero estimates are used to determine whether the table passes. This is similar to the current method for median estimates, where medians with undefined estimates or standard errors are not included when calculating the table's median CV. If the current method

[3] In the published detailed tables, all zero estimates are assigned a predetermined MOE.

presumes zero estimates to be unreliable, this option offers no opinion about the reliability of zero estimates, and is satisfied to determine the table's reliability based only on nonzero estimates.

- Option 2 - Set the CV of zero estimates to a value less than 0.61. This option implies that a zero estimate is more reliable than unreliable - a zero estimate may be correct if the population total for that characteristic in that geographic area is in fact zero. Using a CV less than 0.61 for zero estimates will make it more likely that the table's median CV will be less than 0.61, than if the CVs were set to one.

- Option 3 - Assign a CV of greater than 0.61 and less than one to zero estimates. This considers zero estimates to be more unreliable than reliable, but is less severe then the current rule of assigning a CV of one.

B. Methodology for Zero Count Estimates Research

The analysis uses population, housing unit and household count data from the 2005 through 2008 ACS 1- year and the 2005- 2007 and 2006-2008 ACS 3- year estimates unless otherwise stated. All the count detailed tables and their collapsed versions were included. Analysis is done separately for 1-year and 3-year detailed tables. The frequency of zero estimates in the ACS 1- year detailed count tables over four years of production, 2005 through 2008 ACS were tabulated. Four summary levels (state, county, place, and PUMA) and five population sizes were examined. The Bayesian probability of being a zero for all years given being a zero in a given year was calculated. A similar probability was calculated of the likelihood of a zero in the fourth year, given a zero estimate in the three previous years. These probabilities may help determine which variation of the current release rules would be most promising. The analysis also examines both the impact on the possible number of tables and estimates within these tables that would be published and the resulting quality of the published estimates by alternative options. The "current rule" refers to the 0.61 median CV table-based rule used for detailed count tables.

C. Results for Zero Count Estimates

The results begin with our looking at 823 population, household, and housing unit count detail tables available for 3,435 geographies, specifically for states, counties, places, and Public Use Microdata Areas (PUMAs), for a total of 74.2 million estimates. Table 7, shows for each of the four summary levels the number of estimates that were zeroes over the 4 years. Specifically it shows the number of estimates that never had a value of zero versus those that were consistently zero across all four years and gradations in between.

Table 7. Distribution of Estimates in 2005 through 2008 ACS Detailed Tables by Geography and Frequency of Being a Zero Count

Geography	# of Estimates	Never Zero	Zero Once	Zero in Two Years	Zero in Three Years	Zero in all Four years
State	1,101,651	88.9%	3.1%	2.2%	2.3%	3.5%
County	17,064,790	63.6%	6.7%	5.9%	7.4%	16.4%
Place	11,340,525	60.8%	7.6%	6.7%	8.0%	16.9%
PUMA	44,692,469	61.2%	7.4%	6.5%	7.9%	17.1%
	74,199,435					

For counties, places, and PUMAs, a little more than 60 percent of the estimates never had a zero value, and about 17 percent had a value of zero all four years. For states, 89 percent of the estimates never had a zero value and only 3.5 percent had a value of zero all four years. This demonstrates that smaller geographic areas are more likely to have zero estimates than larger geographic areas and are more likely to have estimates that are consistently zero.

In Table 8, we see for about 800 counties a similar distribution of estimates by number of years the estimate was zero by various population sizes.

Table 8. Distribution of County Estimates in 2005 through 2008 ACS Detailed Tables by Population Size and Frequency of Being a Zero Count

Pop Size Range	Number of Counties	Number of Estimates	Never Zero	Zero Once	Zero in Two Years	Zero in Three Years	Zero in all Four Years
< 100,000	219	4,730,619	52.9%	7.6%	7.0%	9.2%	23.2%
100,000-250,000	321	6,933,921	60.9%	7.2%	6.4%	8.0%	17.4%
250,000-500,000	123	2,656,923	71.2%	6.3%	5.2%	5.9%	11.4%
500,000-1,000,000	87	1,879,287	78.9%	5.0%	3.9%	4.3%	8.0%
> 1,000,000	40	864,040	86.3%	3.4%	2.6%	2.8%	4.9%
total	790	17,064,790	63.6%	6.7%	5.9%	7.4%	16.4%

Not surprisingly, zero estimates were more common for areas with a population size under 100,000 than those with over 1 million.

In Table 9, the “Over All Four Years” row refers to the situation from the other two tables, where at least one of the four years' estimates is zero over the summary levels state, counties, places, and PUMAs.

Table 9. Distribution of County Level Estimates in 2005 through 2008 ACS Detailed Tables by Frequency of Being a Zero

Years	Not Zero	Zero in At Least One Year	Zero in Less Than four Years	Zero in All Four Years
Over All Four Years	63.6%	36.4%	55.0%	45.0%
2005	72.9%	27.1%	39.5%	60.5%
2006	73.7%	26.3%	37.8%	62.2%
2007	73.6%	26.4%	38.0%	62.0%
2008	73.6%	26.4%	37.9%	62.1%

From the data we have available currently, about 64 percent of county level estimates were never zero across this four year period. The probability of having a zero in every year given that you have a zero in one year is about 45 percent. The other four lines look at estimates which are zero in a specific year. For 2006 through 2008, the results are almost identical which is not surprising. There are some minor differences from the others in 2005.

We also learned that if a 2005-2007 3-year estimate was zero then there is 90 percent chance that the 2008 1-year estimate was also zero.

Recall there were three alternate options considered. The simulated results on the number of estimates and tables published are shown in Table 10. We found Option 3, assigning any CV value that is > 0.61 for zero estimates, gave very similar results in the percent of tables or estimates published as the current rule. Therefore, they are placed together in the first row. Likewise, choosing any value less than 0.61 for the CV in Option 2 gave very similar rates. The values in the table were obtained using a CV of zero.

Table 10. Simulated Publication Rates for 2008 and 2006-2008 ACS Detailed Tables and Estimates by Options

Options	Tables		Estimates	
	2008	2006-2008	2008	2006-2008
Current (assign CV>0.61)	62.5%	63.0%	51.9%	54.0%
Not considering	77.5%	78.7%	72.2%	73.4%
Assign CV< 0.61	95.0%	97.2%	98.1%	98.3%

In the first row, the current rule allowed the publication of about 63 percent of tables and about 52 percent of estimates in the 2008 ACS. Roughly similar results are seen for the 2006-2008 ACS 3-year products. The second row shows the simulated result of not considering (or not including) zero estimate CVs in the calculation of the reliability of the tables. About 72 percent of the estimates and about 78 percent of the tables could be published under this rule. Finally in the third row, assigning any CV value for zero estimates that is < 0.61 could result in about a 98 percent publication rate for estimates and about a 95 percent publication rate for tables.

Table 11 shows the CV distribution of published estimates under the current results and the simulated distributions using the two alternative options. In the first column, we see the number of count estimates that would be published followed by the distribution of published counts by their CV values.

Table 11. Simulated Distribution of Published 2008 ACS Estimates by Their CVs

Options	Estimates Published	CV<0.3	0.3< CV<0.61	CV >0.61	Zero Estimates
Current (assign CV > 0.61)	74,056,928	63.5%	21.3%	9.5%	5.7%
Not including	102,879,836	51.7%	20.6%	11.6%	16.1%
Assign CV < 0.61	139,849,039	39.8%	18.4%	15.5%	26.3%

We see in the first row that about 74.1 million count estimates were published in 2008. About two-thirds of these estimates had a CV<0.3. As we go across the row, there are decreasing portions of estimates with higher CV values and 5.7 percent are zero estimates.

The second row shows the simulated results if we don't consider the zero estimate CVs in the application of the check. We see an increase in the number of published estimates to 102.9 million. As we had expected there is also a shift in the CV distribution with zero counts becoming a larger portion of the total. It also shows that the less reliable estimates (CV>0.61) become more prevalent. In actual numbers, there would be roughly four times the number of zero estimates released and about a 70 percent increase for estimates with CV > 0.61.

The third row shows the simulated results of assigning a zero estimate's CV any value that is less than 0.61 in the application of the check. We again see an increase in the number of published estimates to 139.8 million. Again there is a shift in the distribution of CVs with zero counts becoming the second largest portion as expected. Those estimates with CVs < 0.3 were hit the hardest with only a 20 percent increase in the number of estimates published. In actual numbers, there are now nine times as many zero estimates and three times as many estimates with CVs > 0.61 published compared to the current level.

Conclusions

The goal of the research is to both document the effects of the current reliability rule on median and zero estimates and demonstrate the simulated results of a few alternatives rules.

Median Estimates

We showed some evidence that the CV can be a poor measure of reliability for median estimates and its use in the data release rule can have a less than satisfactory results.

Three alternatives were proposed to remove more of the undesirable and the less reliable estimates. The first involves setting the CV = 1 for undefined estimates when applying the current rule which assumes that all such estimates are unreliable. The second is setting the CVs = 1 for medians based on less than 10 cases and the undefined under the assumption that all are unreliable. Finally, the third option specifically removes all estimates with CV > 0.61.

All the alternatives showed a probable decrease in the number of published estimates. The reduction seems to target estimates that were undefined or with CV>0.61, leaving those with smaller CVs unaffected.

Using the Woodruff method to calculate standard errors of medians doesn't seem to have much impact on the number of tables published and likely would not impact the number of estimates either.

Zero Count estimates

We showed some evidence that zero count estimates may be more stable than assumed and that the current rule may be too strict.

Two alternatives were proposed to increase the number of potentially stable zero estimates, hopefully without increasing the release of less desirable estimates with

CV>0.61. The first was not considering the undefined CVs of zero estimates and only use the CVs of the nonzero estimates to determine whether the table passes. Based on data available to us, we have about a 50 percent chance of getting an estimate that is consistently zero over time so this option offers no opinion about the reliability of zero estimates. The second involves assigning a zero estimate's CV a value less than 0.61 for testing purposes and including the CV in the test, and so implies there are no concerns regarding the reliability of zeros estimates.

Both of the alternatives examined show a probable increase in the number of estimates published. Either alternative would likely show the largest relative gain in the zero count estimates. However, they would all likely show a notable increase in the number of less reliable count estimates, i.e. those with CV>0.61. This suggests that zero count estimates and estimates with poor reliability are somewhat correlated resulting in roughly a gain of one estimate with CV>0.61 for every three zero estimates. Basing the reliability test on nonzero estimates alone doesn't have the same impact as the current rule at removing estimates with the highest CVs.

Our next steps

- We will continue looking into alternative reliability checks for medians and zero count estimates and begin our analysis of ratio estimates.
- We will present findings to ACS program managers for them to consider whether or not to make any changes in the data reliability checks.
- We will determine the feasibility of implementing alternatives.

Reference

- Dorfman, A. and R. Valliant (1993), *Quantile Variance Estimators in Complex Surveys*, Survey Research Methods Section, 1993 JSM Proceedings, pages 866 – 871
- Navarro A. and B. Garrett (2009), *Data Quality Filtering in the American Community Survey*, Prepared for the Spring 2009 meetings of the Census Advisory Committee of Professional Association.
- Starsinic, M. (2009), *Assessment of Data Release Rules on the Reliability of Multiyear Estimates in the American Community Survey Data Products*, Survey Research Methods Section, 2009 Joint Statistical Meeting Proceedings.
- U.S. Census Bureau (2006) *Research on Changes to the Current Base Tables Filtering Rules*, Memorandum for D. Hilmer from S. Baungardner, dated September 26, 2006.
- U.S. Census Bureau, (2007), *Should Zero Estimates Be Used in the Calculation of the Median CV for a Table*, draft #1.3, dated May 17, 2007.
- U.S. Census Bureau (2009) *Design and Methodology of the American Community Survey*, (ACS-DM1). http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf
- Winkler, R., J. Smith, and D. Fryback (2002), "The Role of Informative Priors in Zero-Numerator Problems: Being Conservative Versus Being Candid", *The American Statistician*, Vol. 56, No.1., February 2002,
- Wright, T. (1990), "When Zero Defectives Appear in a Sample: Upper Bounds on Confidence Coefficients of Upper Bounds", *The American Statistician*, Vol.44.,No.1. February 1990 pp. 40-41.