Recommendation for new Statistical Functionality within the American FactFinder for the American Community Survey Data Products

Nov 14, 2010

Author: Doug Hillmer

Executive Summary

This document recommends integration of eight statistical functions into the American FactFinder (AFF) to be used with the American Community Survey data products.

ACS data users are often called upon to quickly create reports containing information in response to requests from state and local governments as well as organizations using data at the state and local levels. These data users must have an intimate understanding of the ACS data, including both its strengths and weaknesses. They must be able to quickly identify the ACS estimates that are the best match to the request they have received. In many cases, they must combine ACS estimates to create new estimates that are still a good match to the request but are also more statistically reliable than the published ACS estimates. Often, these data users must also carry out statistical comparisons either to compare geographic areas for a given characteristic or to examine a characteristic for a given area over time to assess change.

All of this work requires the ability to manipulate the sampling error information provided by the Census Bureau for ACS estimates. Familiarity with sampling statistics varies widely among ACS data users. But, regardless of their statistical sophistication, these data users are called upon to quickly provide information for a request, sometimes in a matter of hours. Over a period of many years, the data user community has been trying to find more efficient and reliable ways to carry out the statistical work required for the requests they receive. That is the impetus for the recommendation described in this document. The document recommends eight statistical functions (collectively described in the data user community as a "statistical calculator") which, if integrated into a website that allows access to the ACS data (e.g., American FactFinder), can enable the data users to do much of the work they need to do quickly and reliably right in the website.

We try to address many of the details that Census Bureau analysts and software developers must consider in the implementation of these eight functions. We also have attempted to identify issues related to these statistical functions that may require more work before they can be fully addressed. Finally, we see this document as beginning a dialog between data users and Census Bureau staff on questions and issues that arise as the Census Bureau considers this recommendation and, hopefully, moves to implement the recommendation.

I. What is a Statistical Calculator and Why do ACS Data Users Need One?

A. What is a Statistical Calculator?

The phrase "Statistical Calculator" can conjure up all kinds of ideas. However, as used in this document, it really deals with some basic statistical functions that use the standard errors of sample-based estimates as their inputs. Examples include creating a standard error for an estimate that is the sum of two or more estimates; testing the difference between two estimates for statistical significance at a given level of confidence; and, creating a standard error for a ratio of one estimates to another (e.g., mean earnings for people 65 and over in the workforce). Over the years, data users have come to call any tool that performs these types of functions a "statistical calculator". In this document we use this term (and the corresponding acronym "SC") to describe the eight statistical functions (or "calculations") that are important elements in the daily work of ACSO data users. These functions are described in section II below, and the formulas needed for these functions are provided in Appendix C.

B. Massive amount of estimates for many geographic areas over many time periods

As ACS continues, it will provide users with more and more opportunities to create estimates for geographic areas of local interest and to combine annual estimates over time for a characteristic in a given geographic area to obtain an estimate that may be more statistically reliable. Of course, a major question for most users is whether change has occurred in a geographic area over time or not. Thus, users will need the ability to quickly test statistical hypotheses about change over time for a characteristic, whether it be a characteristic already in the official ACS data products or one that the user has derived by combining separate estimates in those data products. The user may also discover that analyzing the data for change in a characteristic over time can only be done after combining the estimates for small geographic areas (e.g. tracts or block groups) into larger user-defined areas. This, in turn, implies calculating standard errors for the estimates for the user-defined area. Thus, a number of statistical calculations are frequently required in the work of an ACS data user.

Appendix A contains a detailed example illustrating these types of statistical calculations. The example describes the steps a data user must go through in the current environment to create estimates for a characteristic of interest for a number of geographic areas and perform basic statistical calculations to determine if any of the estimates are different in a statistical sense. Without any tool integrated into the data access application where the employee begins his work (e.g., AFF), each of these steps is basically a manual operation which requires time (in some cases, significant amount of time) and which is prone to human error. Thus, an effective automated tool to help the data user do these types of calculations would have to make a significant reduction in the time required of the data user and in the possibility of the user making an error doing these calculations.

C. ACS sample size issues

The Census Bureau publishes statistical reliability information for every ACS estimate it publishes, the only exception being the estimates in the Narrative Profile data product. In almost all instances, this information is in the form of a margin of error (MOE) based on a 90% level of confidence. In doing so, the Census Bureau implicitly acknowledges the issues related to a “small sample size”[1] and also

[1]The phrase “small sample size” must be evaluated in relation to the actual problem the data user is trying to solve. While it is true, and has been known for a long time, that the ACS sample size is well below that of the Census 2000 long form sample, and long form samples from previous decades as well, it is still a very large sample, especially relative to all other ongoing surveys in the U.S. See http://www.prb.org/Articles/2009/2010censustestimony.aspx for a more detailed

provides data users with an important set of information that enables users to do a number of things that they could not do (or could not do very easily) with other sample-based data sets published by the Census Bureau. In particular, making the MOE available to data users enables them to

1. create estimates that are of interest to them
2. assess the statistical reliability of estimates
3. aggregate estimates for small population and housing subgroups (e.g., people born in Venezuela, American Indian and Alaska Natives living in South Carolina, etc.) to develop new estimates with a higher statistical reliability
4. perform statistical comparisons for a characteristic over time in a geographic area or between two geographic areas in a given time period.

discussion of this topic.

II. What are the problems the Statistical Calculator should solve?

There are two main problems that a Statistical Calculator tool should solve: doing the required calculations correctly; and, doing the calculations quickly. In relation to these two "problems", perhaps the most fundamental feature of such a tool is the degree to which it is truly “integrated” into the larger data access application. Such integration can yield enormous benefits, including:

A. an easy way for the user to communicate the goal of the activities, such as creating an estimate for a combination of geographic areas; comparing existing estimates among a set of geographic areas; comparing existing or user-defined estimates across time periods for a fixed geographic area; or, some combination of the above. This allows the designers of the tool to guide the user through the appropriate set of screens aimed at capturing exactly what the user wants to achieve. It also allows for early capture of the specification of the formulas needed and any relevant parameters (e.g., the confidence level for an MOE or a statistical test)

B. No transcription of any actual data values by the user; i.e., a significant reduction in errors the user could make.

C. Display of appropriate warning messages as the user works with the tool and, whenever feasible, avoiding completely a user selection that would lead to an erroneous or meaningless result. Examples of warning messages could include warnings about user selection of geographic areas that include areas of different, and possibly overlapping types. The user may have made a mistake, not considered this issue, or he or she may be fully aware of it and have good reason to proceed anyway. An example of avoiding a user selection that would lead to an erroneous result would be to prevent the user from combining cells in a table that include both detailed cells and subtotal cells containing those detailed cells.

.

This integration could also allow for requests that consist of a large volume of calculations. The integration allows us to view the tool as having two major components: a specs capture component that creates the “program” that would be run to actually do the calculations but does not actually execute the program; and, an execution component that could (eventually) include scheduling the very large requests to run in non-peak time periods so the computer resources required by the execution does not result in any noticeable degradation of performance for the other concurrent users of the data access application. (This may seem overly ambitious for the tool, and an initial version may not contain any such “scheduler” feature. However, as designers of automated tools will attest, knowing what the users would ultimately want allows programmers to design an automated tool that can accommodate these future features without major rewrites of the entire tool.) Another possible feature for the tool, one that is related to “specs capture”, would be recording and storing for later reference the different functions within the tool that people have used and the ways in which they have used these functions. This feature would eventually provide some built-in feedback about how the tool is used.

Appendix B contains screen shots representing two different scenarios demonstrating how the SC tool might be integrated into an existing data dissemination application such as AFF. The diagrams in this appendix are meant only to give the reader some idea of what is meant by an "integrated" SC capability; they are not meant to suggest an actual design for the SC tool.

NOTE: A prototype statistical calculator tool was developed by the New York SDC a few years ago, and several state SDC offices have made use of this tool, and some make it available to the public via their websites. This tool is an Excel application, and the features of the tool are covered in this requirements document. There was, of course, no way for the New York SDC to integrate this tool into AFF. Therefore, it must be used as a standalone tool, and the input data for the statistical calculations must be manually entered**.** Another standalone SC tool was developed at the University of South Florida under a grant from the U.S. Department of Transportation. This tool is much more extensive than the tool developed by the New York SDC, but it is still a spreadsheet-based tool. However the spreadsheet with the built-in application contains 17 worksheets, each with its own user documentation. The tool seems to work in a manner similar to the New York SDC tool, but it is much more extensive and deals with data from other sources as well (e.g., Census 2000). Both the tool and a document describing all the tool's capabilities can be downloaded from http://www.nctr.usf.edu/abstracts/abs77802.htm . The document has a very thorough discussion of all the statistical formulas and calculations that the user might need.

III. Primary Functions Covered by the Statistical Calculator

A. Assumptions about the environment of the Statistical Calculator

It is assumed that the SC tool will run in a computing environment that provides instant access to the input data needed for all the statistical calculations that the SC will perform. (i.e., the tool will be "integrated" into the larger application).

B. List of the functions the Statistical Calculator should perform

It should be noted that all the statistical functions described in this recommendation are already described in the "Accuracy of the Data" document that accompanies each ACS period data release. For reference, see pages 20-26 of the 2009 1-year Accuracy of the Data document which can be found at this
link: http://www.census.gov/acs/www/Downloads/data_documentation/Accuracy/ACS_Accuracy_of_Data_2009.pdf .

In the list below of eight major functions any explicit or implicit reference to an "estimate" can mean an estimate published in the ACS data products or an estimate created by the user from estimates in the ACS data products. An "implicit reference" to an estimate would be in a proportion or a ratio. The formulas for these functions are given in Appendix C. We have made every attempt to be accurate in describing these formulas. However, mistakes are always possible. Therefore, we request that that all calculations and formulas contained in Appendix C be reviewed by Census Bureau staff, corrected where necessary, and the detailed results of that review be communicated to the SDC Steering Committee in writing.

Major functions of the SC:

Creating standard errors and margins of error for user-defined estimates:

i. Create a sum or difference of two or more estimates – same geographic area(s), same time period
ii. Create a sum or difference of two or more estimates for the same characteristic in the same time period for a combination of two or more geographic areas
iii. Create a proportion
iv. Create a ratio
v. Create a product of two estimates

Using the MOEs to do other statistical calculations:

vi. Create the "coefficient of variation (CV)" for any of the estimates created by functions i)-vi)
vii. Compare estimates of the same characteristic across two or more geographic areas for a fixed time period
viii. Compare estimates of the same characteristic for the same geographic area across multiple non-overlapping time periods

B. Input/output requirements

The initial version of the SC tool will assume that all input is supplied by the user in a real-time fashion via a web-enabled interface. However, it may be beneficial in a later version of the SC to allow users to specify the calculations needed by providing a text file that is constructed using a layout convention specified in online documentation (aka "User Guide") for the SC tool.

Output from the SC tool should be available both in a web-based display format and in a downloadable form which would also be specified in the online User Guide for the SC tool.

IV. Other Recommendations and Issues to Consider

A. Information on SC Tool Usage

The Census Bureau should maintain audit trail information on each use of the tool. This information should be stored in a manner that allows for reports to be created which would characterize how the tool is being used. These reports would be important for planning later enhancements to the tool.

B. Avoiding Inappropriate Use of SC functions

The SC tool should, to the greatest extent possible, prevent a user from combining estimates which should not be combined. This includes estimates coming from two different "units of analysis" (e.g., housing units and people); detailed estimates with subtotal estimates containing those detailed estimates – this would be a form of double-counting; estimates for characteristics that may be from the same "unit of analysis" but have different "universes" (e.g., people 16 and over vs. workers 16 and over). However, there are several valid reasons for "violating" these rules. Therefore, in some cases these restrictions may be implemented only as warning messages to the user. For example, if a user wants to create a ratio estimate with a population characteristic in the numerator and a housing characteristic in the denominator (e.g., people per housing unit), the SC tool should allow the user to proceed. However, if the user attempts to create a difference between the number of housing units and the number of people in a geographic area, the SC tool should warn the user to be sure he or she has not chosen a characteristic in error.

We know that designers and programmers of the SC tool would need much more detailed specification of these restrictions before they could implement this recommendation. SDC members are glad to work with Census Bureau staff to translate this recommendation into a clear specification.

C. Restrictions and out-of-scope items

The SC tool is not meant to be a general statistical analysis tool, such as SAS, SPSS, etc. It is focused on basic statistical operations using MOEs supplied for published ACS estimates. Statistical procedures that are more involved and require more computing resources are not viewed as appropriate to integrate into a data access application that must, as its first priority, allow for efficient display and extraction of data.

D. Issues to Consider

There are several issues that developers of an SC tool must confront. We do not expect that all of these issues can be addressed in a satisfactory manner in an initial version of the SC tool. However, because these issues are important, we wanted to include them in this document.

i. missing data in 1-year and 3-year ACS data products due to data quality filtering – the SC must, at a minimum, alert the user when this occurs, possibly offering the user some alternative courses of action to consider, such as: choose larger geographic areas; choose a less detailed table that may have the input estimates required; modify/weaken the ultimate goal to allow for use of estimates less likely to be filtered out; use 5-year data, if available, since no filtering would be applied.

ii. Creation of arithmetically impossible results: division by 0 for ratios; negative result under the square root when attempting to build an MOE for a proportion; etc. Ideally, the SC tool would prevent these errors from occurring.

iii. What should be done if one of the geographic areas has an estimate that is controlled to equal the Population Estimates Program (PEP) estimate for a characteristic? Should a 0 be used as the MOE for that estimate when creating the MOE for the estimate for the combination of the geographic areas?

iv. Use of factors from other data sources. For example, the user may want to estimate the number of 3-4 year old children in a geographic area as part of estimating the potential number of head start applications in the upcoming school year. Since neither ACS nor the PEP provide estimates of this age group for local geographic areas, the user may wish to get this number as a proportion of the kids in the 0-5 age range (which ACS does estimate) from the most recent census and use that proportion as a factor by which to multiply an ACS estimate. The user must have some way of providing such an input from a separate data source, and the SC should make it clear that the tool bears no responsibility for the final results in such cases.

v. When an ACS estimate for a given geographic area is 0, a special procedure is used to derive the MOE for the estimate. Of course, the 0 could either be totally correct or just mean that the sample missed anyone with this characteristic. What if a user creates a geographic area by combining other geographic areas, say tracts, and for many of these input geographic areas the estimate is 0 for the characteristic of interest? It would seem incorrect to simply apply the formula described in Appendix C to get the MOE for this characteristic in the new geographic area. This is because that approach would lead to an over-estimate of the MOE. But, what should be done for these situations? The Census Bureau is currently investigating this issue. **For the time being, Census Bureau statisticians have said that it is acceptable for a user to simply omit the MOEs for these 0 estimates when creating the MOE for the sum of the estimates across the geographic areas.** This would imply that the function would have to include a check to see if an input estimate is 0.

Appendix A: Tracts with low graduation rates in Franklin County, Ohio

The following is a description of a hypothetical situation that might require ACS multiyear tract-level data. Franklin County and Columbus OH which is contained within Franklin County, were chosen because no tract-level estimates are available from ACS, and Franklin County was one of the ACS test counties before ACS began full production in 2005. All test counties were sampled at the full production rate during the 2001-2005 period.

Researchers in Ohio want to study differences in high school graduation rates in Columbus. They request assistance in identifying the best data source for their work from a person in the Ohio SDC. The SDC person they contact makes them aware of the ACS data available for Franklin County during the 2001-2005 period. The Census Bureau has published a "Data Profile" for a number of geographic areas, including the tracts in Franklin County. They decide to use the section on "Educational Attainment" of Profile table 2 which is shown here for tract 001600.

Line Number	EDUCATIONAL ATTAINMENT	Estimate	Margin of Error
0	**Population 25 years and over**	**727**	**+/-191**
1	Less than 9th grade	18	+/-24
2	9th to 12th grade, no diploma	247	+/-101
3	High school graduate (includes equivalency)	258	+/-128
4	Some college, no degree	126	+/-73
5	Associate's degree	41	+/-34
6	Bachelor's degree	37	+/-43
7	Graduate or professional degree	0	+/-114
8	Percent high school graduate or higher	63.5	+/-11.8
9	Percent bachelor's degree or higher	5.1	+/-5.6

They decide to go through the following steps in processing and analysis to identify the tracts they should target in their efforts:

1. download the relevant profile data lines (shown above) from profile table 2 for all 264 tracts in Franklin County. (This step is not described further here.)
2. Sort the tracts in ascending order on the estimate in line 8, "Percent high school graduate or higher". Using that sorted output, group the tracts into several new geographic areas.
3. Re-create the estimate in line 8 for each of these new geographic areas along with the margin of error for each of the new estimates.
4. Using the Z statistic, calculate pair-wise comparisons among the estimates for these areas to see if the areas representing lower graduation rates are statistically different from the other areas.
5. If needed, re-combine the tracts into new areas and repeat steps 3 and 4.

The SDC person makes everyone aware of what is needed to calculate the margin of error for each of the new estimates. The formula for calculating the standard error (SE) is shown here

for reference. The margin of error is simply the standard error multiplied by the constant corresponding to the level of statistical confidence the Ohio State researchers which to us (90%, 95%, etc.)

$$SE(P) = 1/Y\left(\sqrt{SE(X)^2 - X^2/Y^2\,(SE(Y)^2)}\right)$$

where X is the estimate of the numerator (the number of people who are high school graduates or higher) and Y is the estimate of the denominator (all people 25 years and older).

The Ohio State staffers soon realize that step 3 above actually contains the following sub-steps:

a) Calculate the estimate of the numerator for each tract by summing up the estimates in lines 3 through 7. The denominator estimate is in line 0.

b) Calculate the margin of error for each numerator estimate calculated in sub-step a.

c) Create a numerator estimate for each of the geographic areas determined in step 2 above. Do the same for the denominator estimate. For each new estimate, the new margin of error must be calculated.

d) Create the desired proportion estimate for each geographic area and the margin of error for that estimate.

In step 4, the researchers must calculate the differences between two proportions and the standard error for that difference. Then, they can calculate the Z statistic as the difference divided by its standard error.

After a few weeks of work, the researchers complete step 4 for the seven groups they have identified in step 2. The groups of tracts are based on certain "high school graduate or higher" percentage cutoffs, the first group containing tracts with a percentage below 70% and the seventh group containing tracts with 95% or more "high school graduates or higher". The table containing the results of the Z statistics for the pair-wise comparisons is shown below. The researchers notice that only group 1 is statistically different from all the other groups at the 95% level of confidence. They are satisfied that they have identified the group of tracts where they should concentrate their initial efforts.

Z score results

Groups	Groups						
Groups	1	2	3	4	5	6	7
1	0.0000	-2.1877	-3.4018	-4.6036	-6.3888	-8.1419	-10.0343
2	2.1877	0.0000	-1.1006	-2.2731	-3.7465	-5.3402	-7.0611
3	3.4018	1.1006	0.0000	-1.0514	-2.2807	-3.6837	-5.1760
4	4.6036	2.2731	1.0514	0.0000	-1.1433	-2.5429	-4.0455
5	6.3888	3.7465	2.2807	1.1433	0.0000	-1.6005	-3.4098
6	8.1419	5.3402	3.6837	2.5429	1.6005	0.0000	-1.8685
7	10.0343	7.0611	5.1760	4.0455	3.4098	1.8685	0.0000

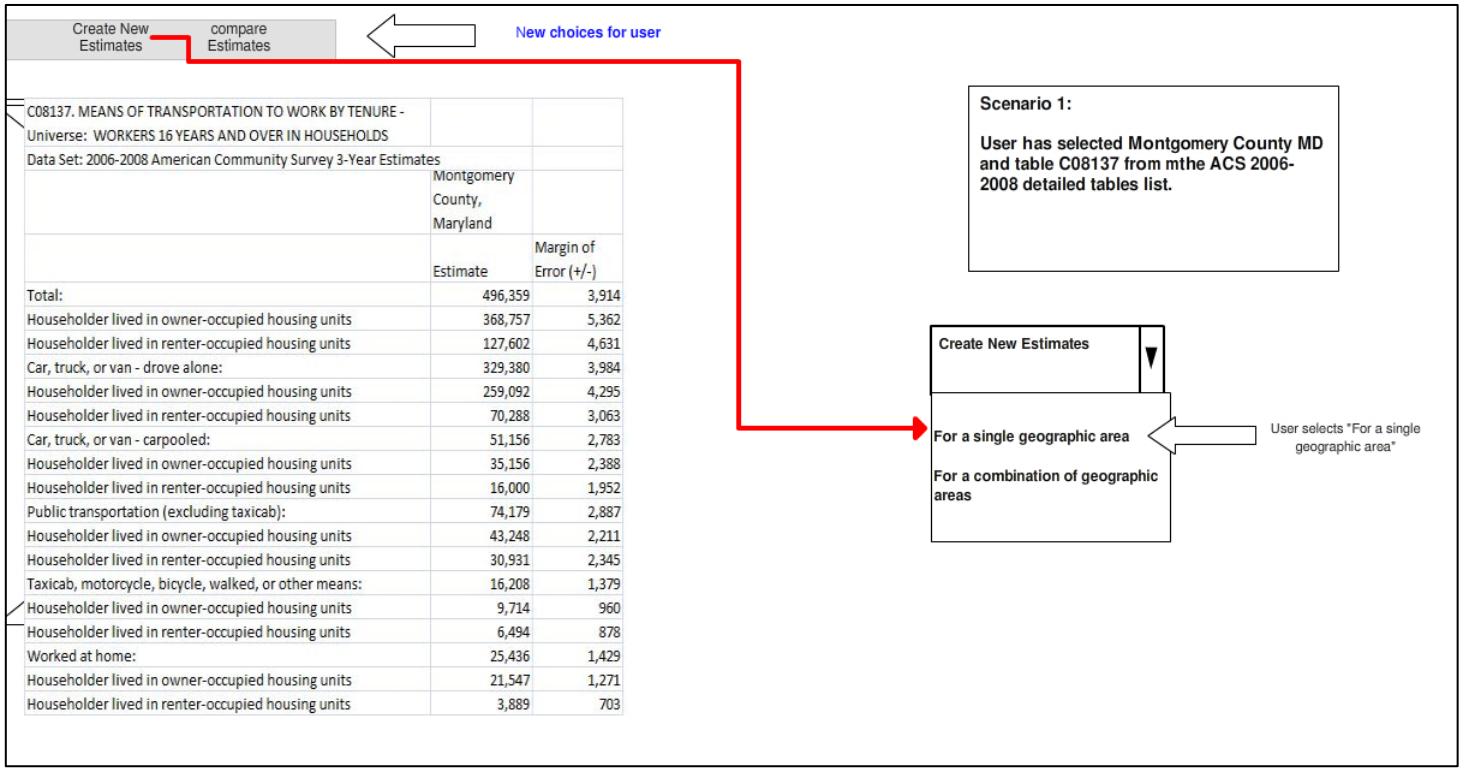
In a de-briefing meeting with the SDC person the Ohio State staffers summarize how much work this required and how error-prone each step was, meaning that a lot of verification was required. Had an SC tool such as the one described in this document existed, they might have reduced the work from weeks to a day or two, especially because the software would have performed so many of the calculations that they had to do themselves. ***Note: Scenario 2 in Appendix B provides screen shots for a possible web-based approach to testing differences for statistical significance among a number of geographic areas.***

Appendix B Screen shots illustrating two scenarios using an integrated SC tool

Below are eight separate screens that cover two scenarios of usage of an SC tool integrated into an existing ACS data dissemination application, such as the American FactFinder.

Scenario 1: Creating a new estimate by combining existing estimates.

Figure 1 Create a new estimate by combining estimates



C08137. MEANS OF TRANSPORTATION TO WORK BY TENURE - Universe: WORKERS 16 YEARS AND OVER IN HOUSEHOLDS		
Data Set: 2006-2008 American Community Survey 3-Year Estimates		
	Montgomery County, Maryland	
	Estimate	Margin of Error (+/-)
Total:	496,359	3,914
Householder lived in owner-occupied housing units	368,757	5,362
Householder lived in renter-occupied housing units	127,602	4,631
Car, truck, or van - drove alone:	329,380	3,984
Householder lived in owner-occupied housing units	259,092	4,295
Householder lived in renter-occupied housing units	70,288	3,063
Car, truck, or van - carpooled:	51,156	2,783
Householder lived in owner-occupied housing units	35,156	2,388
Householder lived in renter-occupied housing units	16,000	1,952
Public transportation (excluding taxicab):	74,179	2,887
Householder lived in owner-occupied housing units	43,248	2,211
Householder lived in renter-occupied housing units	30,931	2,345
Taxicab, motorcycle, bicycle, walked, or other means:	16,208	1,379
Householder lived in owner-occupied housing units	9,714	960
Householder lived in renter-occupied housing units	6,494	878
Worked at home:	25,436	1,429
Householder lived in owner-occupied housing units	21,547	1,271
Householder lived in renter-occupied housing units	3,889	703

Figure 2 Screen allowing user to choose estimates to combine

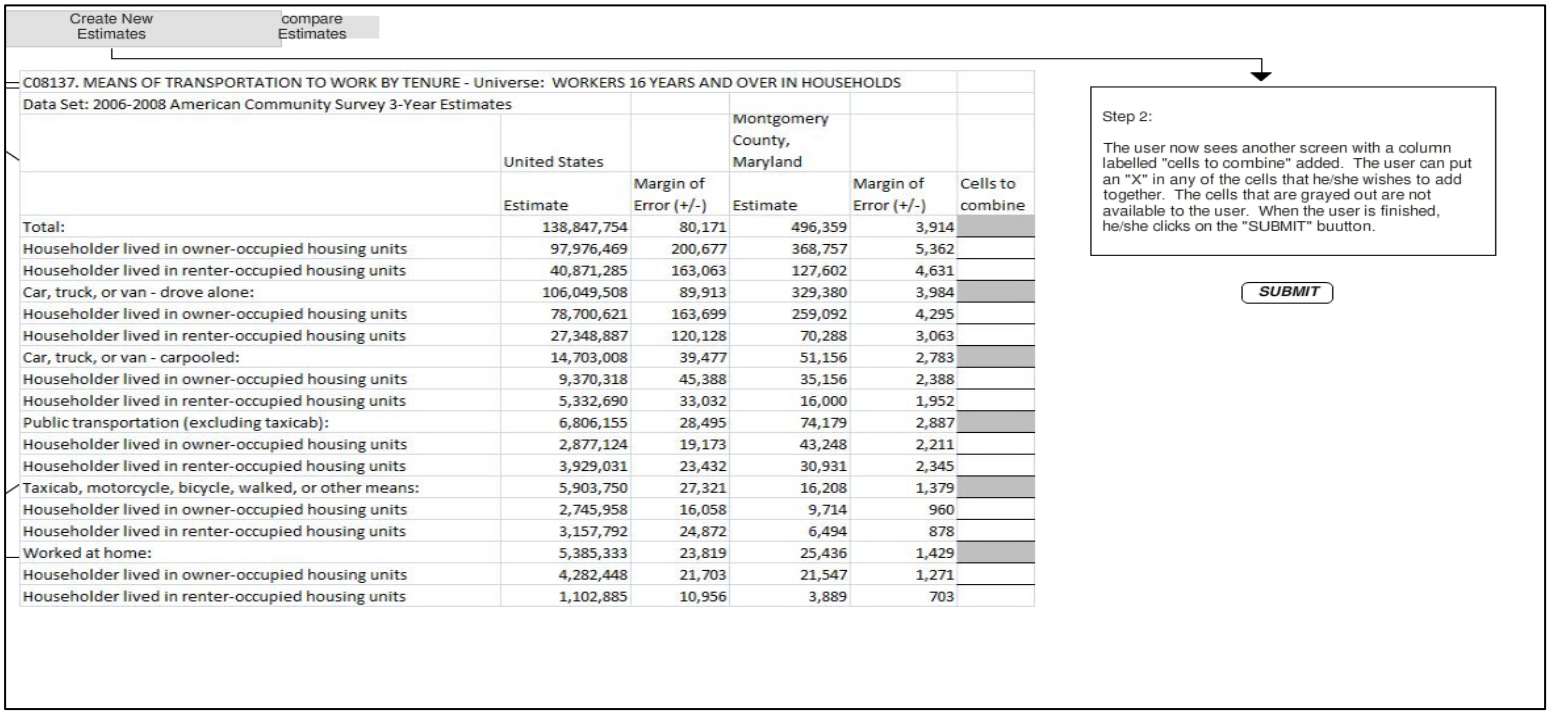


Figure 3 User selects cells to combine for new estimate

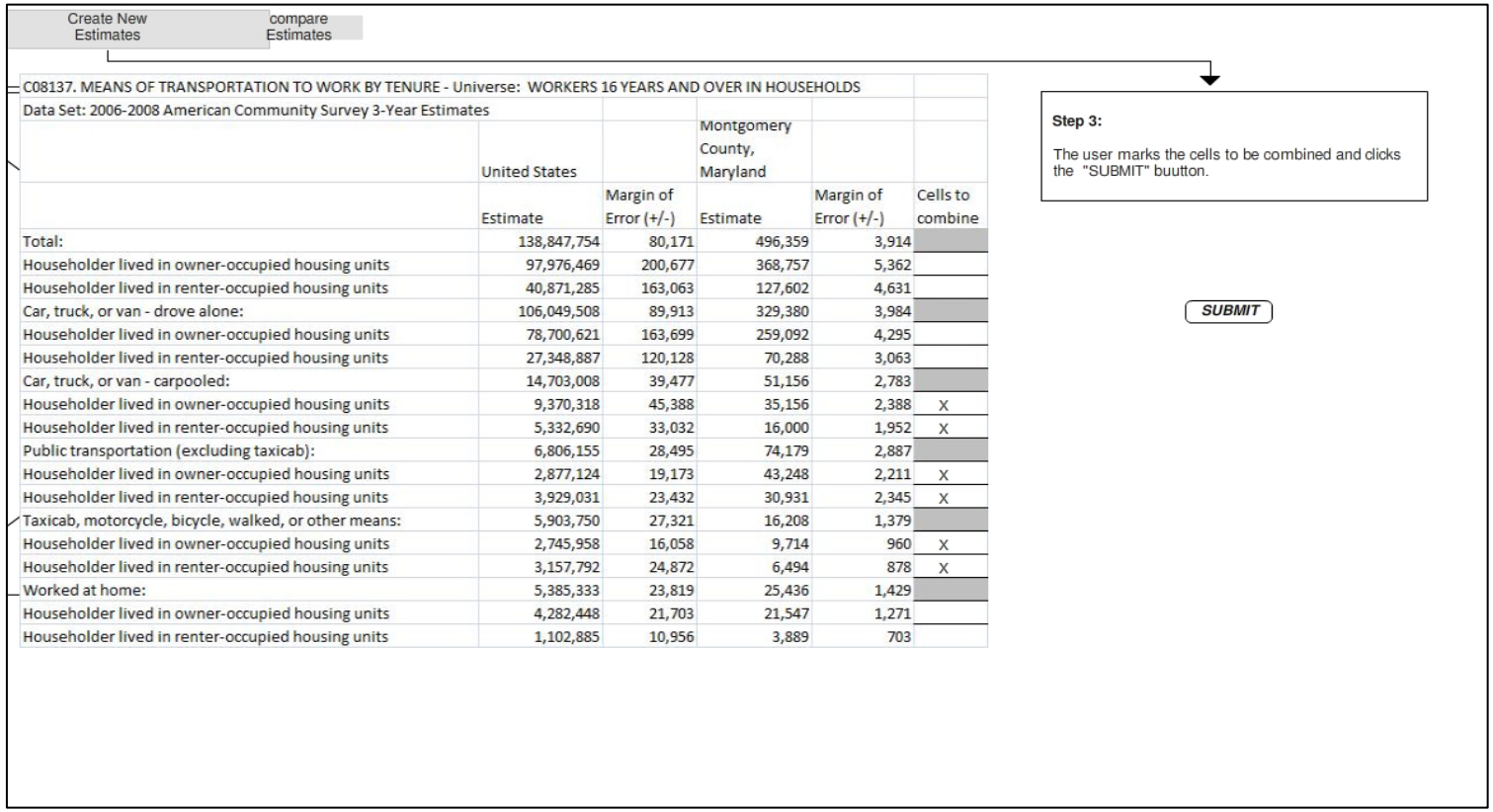
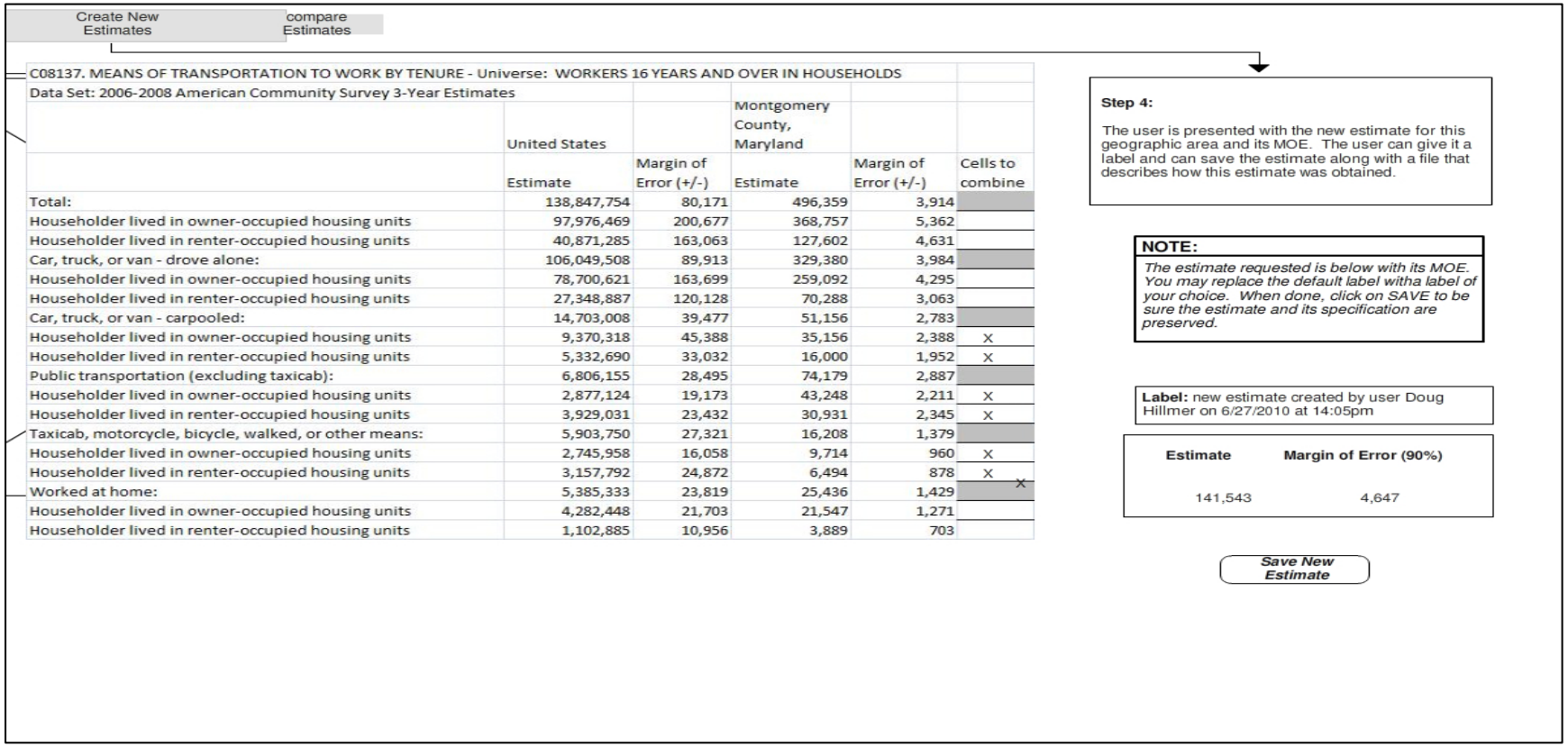
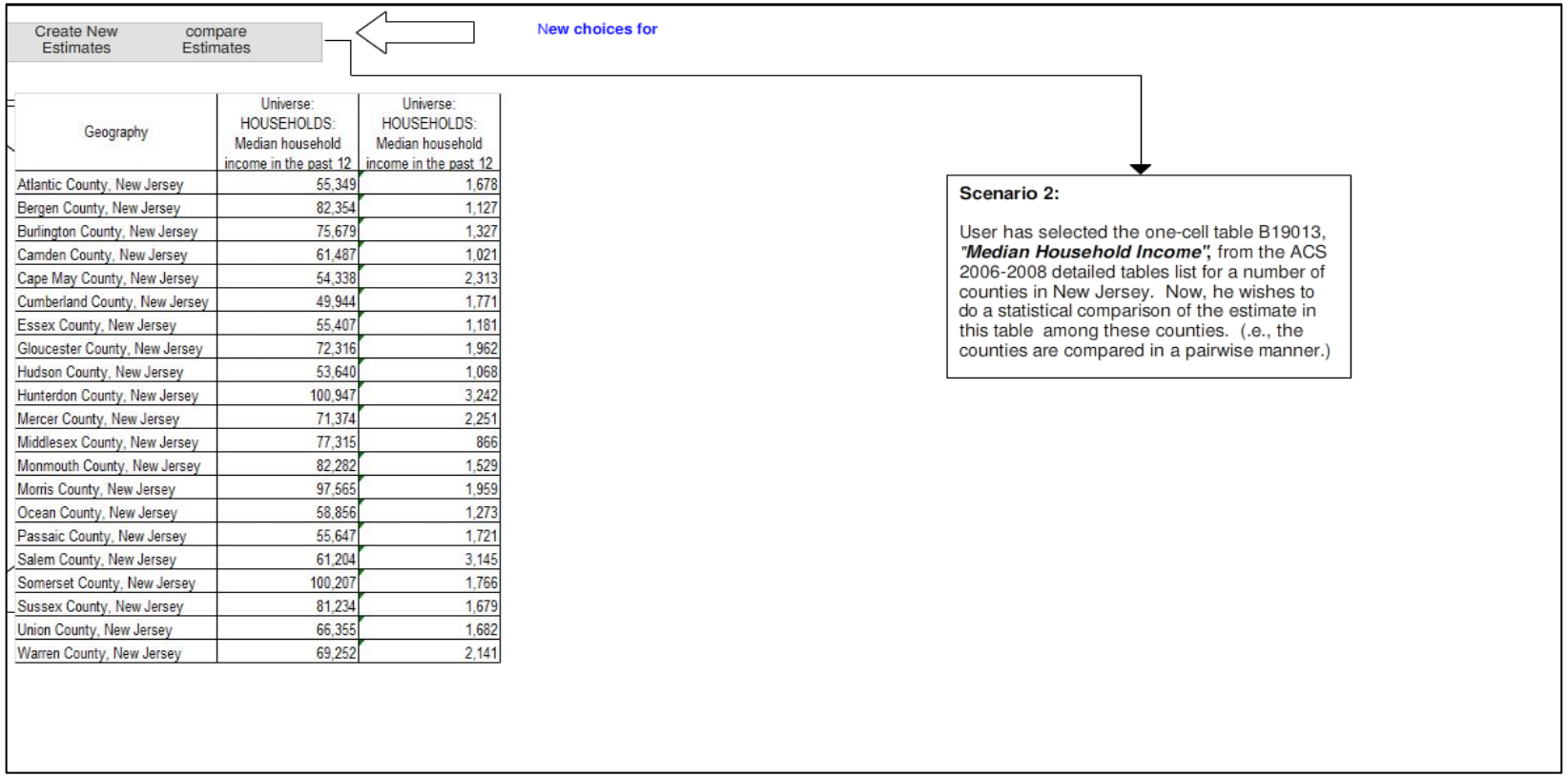


Figure 4 User can label and save new estimate



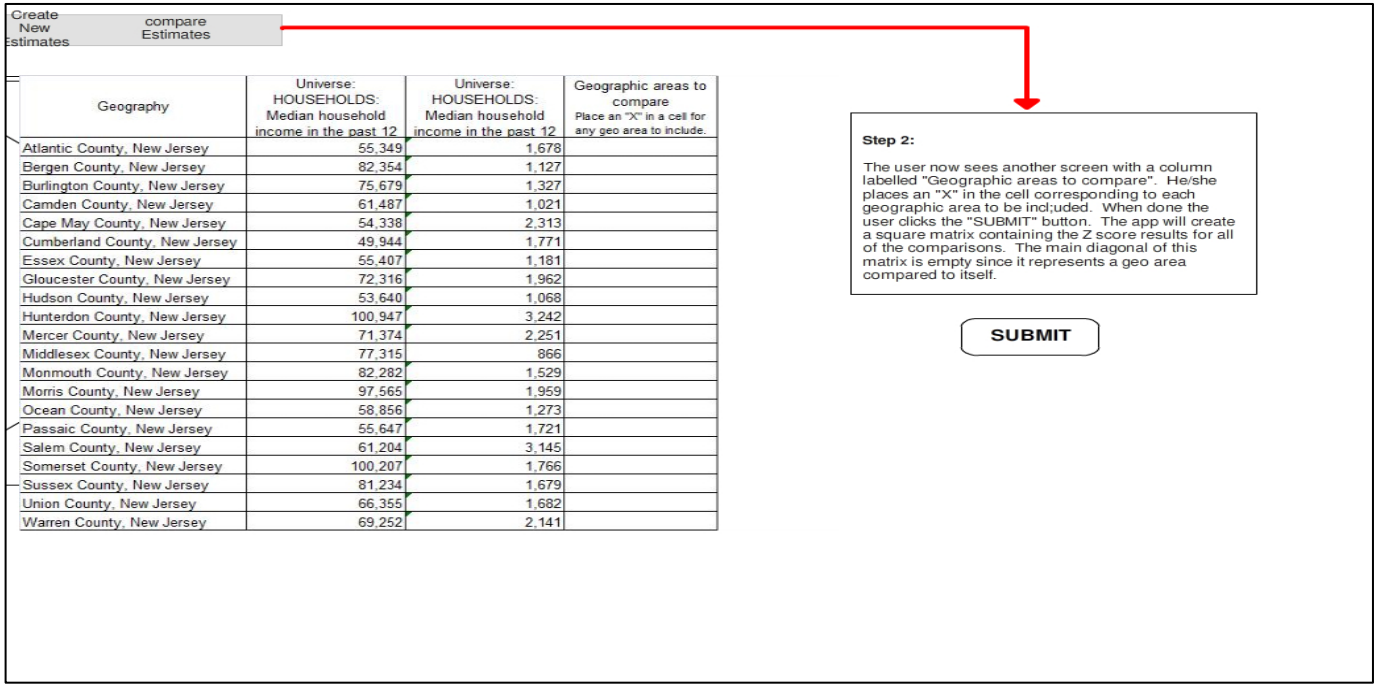
Scenario 2: Comparing a single estimate for multiple geographic areas

Figure 5 Comparing median household income estimates among New Jersey counties



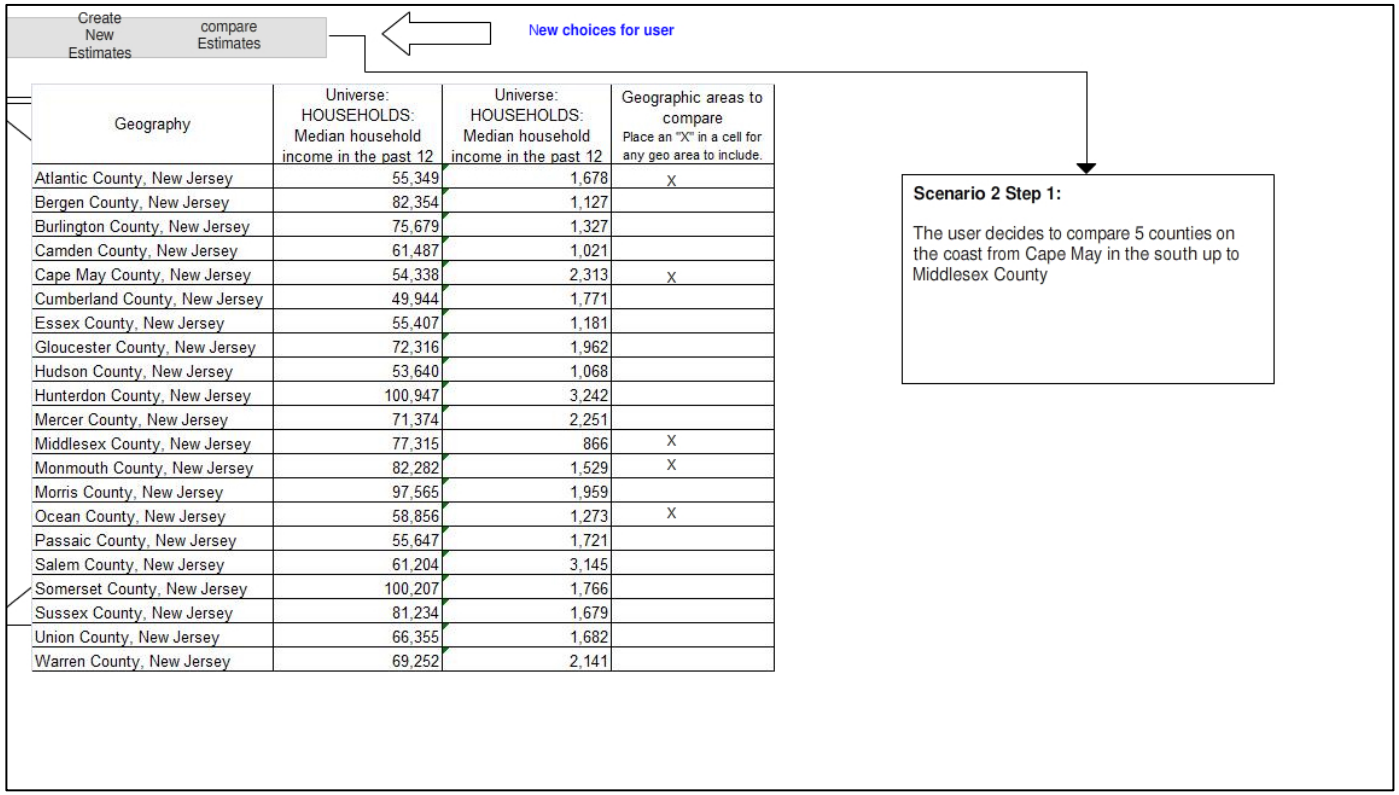
Geography	Universe: HOUSEHOLDS: Median household income in the past 12	Universe: HOUSEHOLDS: Median household income in the past 12
Atlantic County, New Jersey	55,349	1,678
Bergen County, New Jersey	82,354	1,127
Burlington County, New Jersey	75,679	1,327
Camden County, New Jersey	61,487	1,021
Cape May County, New Jersey	54,338	2,313
Cumberland County, New Jersey	49,944	1,771
Essex County, New Jersey	55,407	1,181
Gloucester County, New Jersey	72,316	1,962
Hudson County, New Jersey	53,640	1,068
Hunterdon County, New Jersey	100,947	3,242
Mercer County, New Jersey	71,374	2,251
Middlesex County, New Jersey	77,315	866
Monmouth County, New Jersey	82,282	1,529
Morris County, New Jersey	97,565	1,959
Ocean County, New Jersey	58,856	1,273
Passaic County, New Jersey	55,647	1,721
Salem County, New Jersey	61,204	3,145
Somerset County, New Jersey	100,207	1,766
Sussex County, New Jersey	81,234	1,679
Union County, New Jersey	66,355	1,682
Warren County, New Jersey	69,252	2,141

Figure 6 User selects geographic areas to be included in pairwise statistical comparison of estimates



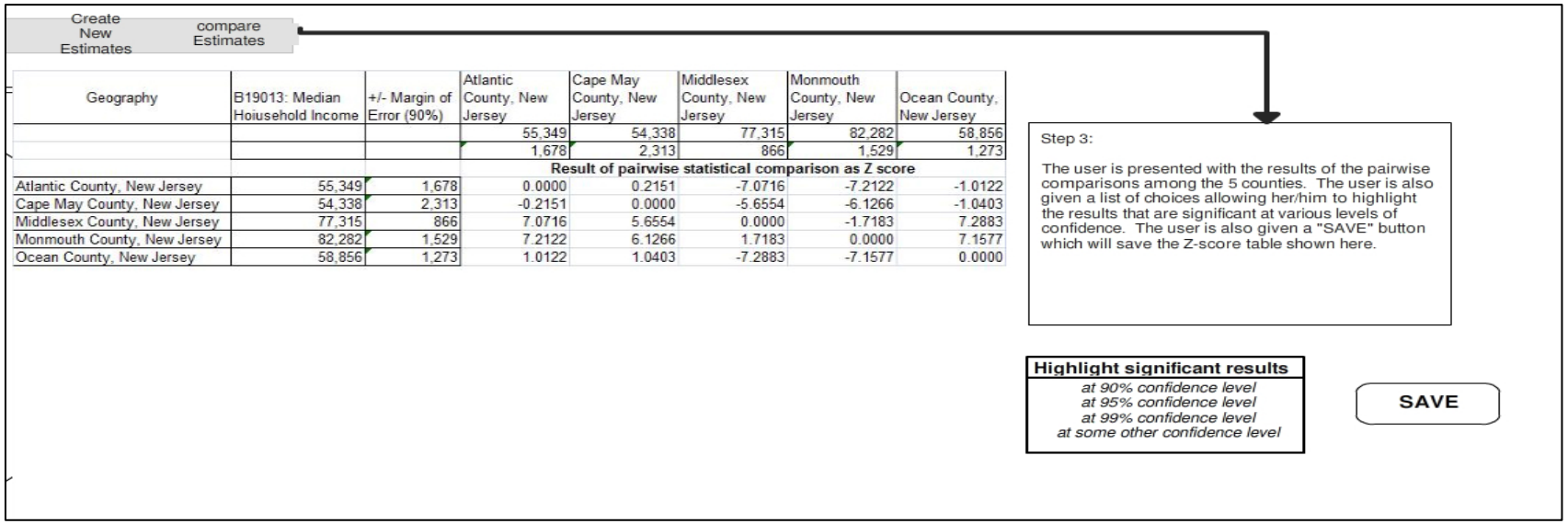
Geography	Universe: HOUSEHOLDS: Median household income in the past 12	Universe: HOUSEHOLDS: Median household income in the past 12	Geographic areas to compare Place an "X" in a cell for any geo area to include.
Atlantic County, New Jersey	55,349	1,678	
Bergen County, New Jersey	82,354	1,127	
Burlington County, New Jersey	75,679	1,327	
Camden County, New Jersey	61,487	1,021	
Cape May County, New Jersey	54,338	2,313	
Cumberland County, New Jersey	49,944	1,771	
Essex County, New Jersey	55,407	1,181	
Gloucester County, New Jersey	72,316	1,962	
Hudson County, New Jersey	53,640	1,068	
Hunterdon County, New Jersey	100,947	3,242	
Mercer County, New Jersey	71,374	2,251	
Middlesex County, New Jersey	77,315	866	
Monmouth County, New Jersey	82,282	1,529	
Morris County, New Jersey	97,565	1,959	
Ocean County, New Jersey	58,856	1,273	
Passaic County, New Jersey	55,647	1,721	
Salem County, New Jersey	61,204	3,145	
Somerset County, New Jersey	100,207	1,766	
Sussex County, New Jersey	81,234	1,679	
Union County, New Jersey	66,355	1,682	
Warren County, New Jersey	69,252	2,141	

Figure 7 User selects counties to compare



Geography	Universe: HOUSEHOLDS: Median household income in the past 12	Universe: HOUSEHOLDS: Median household income in the past 12	Geographic areas to compare Place an "X" in a cell for any geo area to include.
Atlantic County, New Jersey	55,349	1,678	X
Bergen County, New Jersey	82,354	1,127	
Burlington County, New Jersey	75,679	1,327	
Camden County, New Jersey	61,487	1,021	
Cape May County, New Jersey	54,338	2,313	X
Cumberland County, New Jersey	49,944	1,771	
Essex County, New Jersey	55,407	1,181	
Gloucester County, New Jersey	72,316	1,962	
Hudson County, New Jersey	53,640	1,068	
Hunterdon County, New Jersey	100,947	3,242	
Mercer County, New Jersey	71,374	2,251	
Middlesex County, New Jersey	77,315	866	X
Monmouth County, New Jersey	82,282	1,529	X
Morris County, New Jersey	97,565	1,959	
Ocean County, New Jersey	58,856	1,273	X
Passaic County, New Jersey	55,647	1,721	
Salem County, New Jersey	61,204	3,145	
Somerset County, New Jersey	100,207	1,766	
Sussex County, New Jersey	81,234	1,679	
Union County, New Jersey	66,355	1,682	
Warren County, New Jersey	69,252	2,141	

Figure 8 Results of comparison



Geography	B19013: Median Hoiusehold Income	+/- Margin of Error (90%)	Atlantic County, New Jersey	Cape May County, New Jersey	Middlesex County, New Jersey	Monmouth County, New Jersey	Ocean County, New Jersey
			55,349	54,338	77,315	82,282	58,856
			1,678	2,313	866	1,529	1,273
			Result of pairwise statistical comparison as Z score				
Atlantic County, New Jersey	55,349	1,678	0.0000	0.2151	-7.0716	-7.2122	-1.0122
Cape May County, New Jersey	54,338	2,313	-0.2151	0.0000	-5.6554	-6.1266	-1.0403
Middlesex County, New Jersey	77,315	866	7.0716	5.6554	0.0000	-1.7183	7.2883
Monmouth County, New Jersey	82,282	1,529	7.2122	6.1266	1.7183	0.0000	7.1577
Ocean County, New Jersey	58,856	1,273	1.0122	1.0403	-7.2883	-7.1577	0.0000

Appendix C - Required Statistical Functions and Formulas

The formulas given below are only for the Standard error (SE) for the calculated estimate. The MOEs are simply the result of multiplying the SE by a constant corresponding to the desired confidence level (1.645 ~ 90%; 1.96 ~ 95%; 2.56 ~ 99%; etc.) In the description of functions i through iv it is assumed that each input estimate described by X, Y, or Xi is simply the sum of the weights of the records from the microdata that meet the criteria for this characteristic. In other words, no estimates that are already in the form of derived measures (e.g., ratios, proportions, medians, etc.) are allowed as inputs for the formulas used in these functions.

i. Create a sum or difference of two or more estimates – same geographic area(s), same time period

Here we are assuming that the two estimates are "additive"; i.e., they represent two non-overlapping groups taken from the same population or housing universe. The simplest way to meet this requirement is to use two non-total and non-subtotal cells from a given ACS Detailed Table for a fixed geographic area and fixed ACS time period. If the user creates an estimate X as $X = \sum Xi$, then the standard error of X is given by $SE(X) = \sqrt{\sum SE(Xi)^2}$

ii. Create a sum or difference of two or more estimates for the same characteristic in the same time period for a combination of two or more geographic areas

In this case, the formula used is the same as in case i) above with the stipulation that the Xi are each the estimate of the same characteristic and time period and the index i runs across all the geographic areas that are to be combined.

iii. Create a proportion

The formula below assumes X is the estimate of the numerator characteristic which is a subgroup of the denominator characteristic, Y. The SE of the proportion cannot be calculated if Y=0. If the difference under the square root sign is negative,, the formula for the standard error of a ratio (see below) should be used instead.

$$SE(P) = 1/Y\left(\sqrt{SE(X)^2 - X^2/Y^2\,(SE(Y)^2)}\right)$$

iv. Create a ratio or percent

In this case we assume that X and Y may represent subgroups from the same larger group (aka "universe") or they may each come from a different group (e.g., the PPH ratio, people per housing unit). The formula for the SE of the ratio estimate R is

$$SE(R) = 1/Y\left(\sqrt{(SE(X)^2 + X^2/Y^2\,(SE(Y)^2))}\right)$$

v. Create a product of two estimates

Assume that X and Y are two ACS estimates, and the user wishes to create a new estimate, Z, defined as X times Y. The SE for this

product estimate is defined as follows:

$$SE(Z) = \sqrt{Y^2\,SE(X)^2 + X^2 SE(Y)^2}$$

vi. Create the "coefficient of variation (CV)" for any of the estimates created by functions i. - v.

The CV allows the user to quickly assess the statistical reliability of the estimate he or she has created. This is particularly useful when there is a rule in place stating that only estimates with a CV below a certain threshold can be used. Furthermore, the CV is inversely proportional to the sample cases underlying a given estimate. Thus, a high CV (e.g., greater than 20%) can alert the user to small sample size cases. For example, when a user has created a set of proportions, the CV can help the user concentrate on those proportions that are based on more sample. The CV calculation is simple: SE(X)/X where X is a nonzero estimate. If the estimate X has a value of 0, the CV is not defined, but the user already knows that there was no sample for this characteristic in the geographic area.

vii. Compare estimates of the same characteristic across two geographic areas

Statistical comparisons of estimates of the same characteristic for two geographic areas are done using the Z statistic. If X is the estimate for the first geographic area and Y is the estimate for the second area, then Z is calculated as

$$Z = (X - Y) / \sqrt{SE(X)^2 + SE(Y)^2}$$

Using a pre-defined level of confidence the absolute value of Z must be greater than the threshold number corresponding to that confidence level for the comparison to result in a "statistically significant difference" at this level of confidence (see the first paragraph in this appendix for the numbers to use for the most common confidence levels). This formula assumes that the estimates X and Y are statistically independent of one another, which, in this case, means that the two geographic areas are non-overlapping.

viii. Compare estimates of the same characteristic two non-overlapping time periods

The same formula is used as for function v) above with the stipulation that X and Y are each the estimate of the same characteristic and geographic area and the but for two non-overlapping time periods (e.g., 2008 1-year vs. 2007 1-year estimates or 2005-2007 vs. 2008-2010 3-year estimates).