

# **American Community Survey (ACS) Public Use Microdata Samples (PUMS) Webinar Transcript**

February 17, 2016

Marisa Hotchkiss

## **American Community Survey Office**

This presentation was presented as a webinar to the General Public on February 17, 2016 by Marisa Hotchkiss of the U.S. Census Bureau's American Community Survey Office. A link to the recording is available at

<https://census.webex.com/census/ldr.php?RCID=4c5c4edd1e0bb5a1dd9a28229ab783ac>

The transcript of the webinar follows. Slide references and links have been added to the spoken text as appropriate.

Coordinator: Welcome and thank you for standing by. Your lines are in listen only mode until the question and answer session.

If you would like to que up to ask a question, you may do so by pressing Star 1 on your phone.

Today's conference is being recorded. If you have any objection, you may disconnect at this time. I would like to turn over the call to (Marisa Hotchkiss) - you may begin.

### **Slide 1 – Title Slide**

(Marisa Hotchkiss): Alright, hi everyone. I'm really pleased with the large number of participants. This is really exciting - or at least I'm excited. This is going to be an introduction to using American Community Survey Public Use Microdata Sample files.

Not always exciting, but I'm excited about it, so hopefully I can get you all excited about it as well. This presentation is **not** an overview of the American Community Survey - the ACS.

Hopefully, you're already familiar with how that survey is collected and how to find the more than a thousand standard pre-tabulated data tables that are on American FactFinder.

If not, I would absolutely recommend one of our other webinars that would just be an intro to the ACS. So we're mostly just going to talk about PUMS today.

If you put Census Data Products on a scale of easiest to use to most complicated, the PUMS files are going to be near the more-complicated end.

Working with PUMS data is generally going to involve downloading large datasets onto a local computer and analyzing that data with a statistical software package - something like R, SPSS, STATA, SAS, or becoming familiar with the Census Bureau's DataFerret tool.

So I don't want you to hang up at this point -- I just want to warn you that using PUMS can be complicated, and this presentation is not going to turn you into an expert user overnight.

## **Slide 2 - Outline**

But, hopefully after going through a PUMS overview, an explanation of the geography, kind of some tips for accessing PUMS data, going over some comments, speedbumps that people run into, and then showing you where to find documentation and guidance -- I can get you to the point where you're excited to start exploring PUMS on your own.

## **Slide 3 - Why Use PUMS**

So hopefully everybody is still on the line after that. Let me give you a quick sales pitch for why anybody would be interested in using PUMS.

If you're comfortable with the ACS already, you know that with the more than 11 billion estimates that we produce annually, we meet the needs of every data user, right?

Oh, not quite. So the U.S. Census Bureau produces the Public Use Microdata Sample (PUMS) files precisely so that data users with needs that are not met by our standard products can conduct their own analysis.

So on the screen are some examples of why a data user might turn to PUMS, and I'll give you a real life use case.

So let's say I have observed in my metro area that a lot of jobs traditionally held by teenagers seem to be filled by adults and seniors. So maybe I want to investigate this issue in the data.

The standard tables get very detailed. So I can get this table which is sex by school enrollment, by educational attainment, by employment status for the population 16 to 19 years.

Wow, is that detailed - that's great! But, maybe I want to do a more sophisticated analysis. I might have a theory that 16 to 17 year olds are employed at different rates than adult teenagers, 18 to 19 year olds.

So maybe I want to split that data and I want narrower age categories—that's not supported by standard tables. Or, maybe I have a theory that teenagers who speak a language other than English, are more likely to work while they are in school. So I could use the data to investigate whether there is a correlation.

Alternatively, maybe I want to test a theory that when there are other adults in households who have lower median earnings on their occupations, maybe the teenager in that household is more likely to have a job. So now

I'm combining household and person variables to make a new variable and that kind of creation of new measures is something that I can do in the PUMS but not really in any other standard product. So, all interesting ideas that you can really only explore with PUMS.

## **Slide 4 – What are PUMS files?**

What are these PUMS files that I'm talking about? They are Public Use Microdata Sample files.

The "Public Use" implies that these files have been created specifically for the public. They do differ from the survey microdata, though identifying information has been removed, some categories have been modified, either the extreme values have been grouped together, which we call top and bottom coding, or the categories are broader in general.

However, being public use means that these files also come with a suite of downloadable data user guidance, reference materials, and of course they are provided to you like other Census Bureau data products at no cost.

The "Microdata" aspect. These are individual records of survey responses, with the identifying information removed. But this means that users create the estimates, tables, margins of error, all of that processing on their own.

"Sample." These files don't include every record of every person who responded to the ACS - only a select few that in turn are representative of the population. So this does mean the PUMS estimates will not exactly match the American FactFinder estimates, and I'll talk about this a little bit later.

A 1-year ACS PUMS file contains about one percent of all U.S. households—we commonly call it a 1 percent file. To give you a sense of how that stacks up, in the 2011 data, for example, ACS collected about five million person records to create the official estimates that are on American

FactFinder. 3.1 million of those records are included in the one year ACS PUMS file. And then those in turn represent 312 million persons in the estimate.

Then the 5-year, of course, is equivalent to five 1-year files that would include about five percent of all households.

## **Slide 5 – Summary Data vs. Microdata (What's the Difference?)**

Okay, so what does microdata mean? We'll get into a little bit more detail. This is an example of what would make microdata different from summary data that you might see in American Fact Finder or other sources.

So in an aggregated table the individual records are categorized, they're weighted, they're meant to represent an estimate of the larger population. The statisticians of the Census Bureau have taken the records of all the respondents who lived in Maryland, were 18 years or over, male, naturalized citizens, group them together, weighted them to create an estimate of all the adult males in Maryland who are naturalized citizens - that's the number you see here.

By contrast, the microdata provides a sample of the records that those statisticians used. So here you can see one person who responded to the ACS in Maryland is male, he is 58, and he's a naturalized citizen. So then to create an estimate using that microdata, you would have to take the raw material and do the work of weighting, calculating margins of error, etc., that the statisticians have already done to the summary data.

## **Slide 6 – Summary Data vs. Microdata (Pros and Cons)**

So, it sounds like a lot of work and I just want to kind of recap here for a second. I can point out both the benefits and the limitations so you'll have a balanced view of what you would be getting into.

Benefits. The summary file data is going to be a little bit easier to use - we do have many tables that have already been produced. All of the margins of error, and the weighting and everything have already been done. And the summary file data is also available for very small geographies (down to the tract and block group level, if you're not already familiar with that).

The microdata is going to involve more work, but I do want to sell you on the fact that the detail that you can get from the microdata is really worth it. The person-level files include about 250 variables per record, and the housing unit files include about 200 variables. They also include many useful constructed variables - things like poverty status, and subfamily identification.

Limitations. Summary data - we're not going to be able to meet the needs of every data user cause we just aren't always going to know what people will need. We can do our best to kind of pair our variables together that make sense, like education and employment. But there are definitely other combinations that we just can't really foresee.

Microdata—definitely more complex. It's although a smaller sample, and unfortunately, fewer geographies, which I will get into. Microdata is also going to have some edits to protect privacy. It will have collapsed codes, broader categories for some variables - things like race, Hispanic origin, ancestry, place of birth, industry, occupation - that kind of thing.

## **Slide 7 – PUMS Availability**

One more note about the differences between summary data and microdata.

We release the microdata a little later than the summary data.

Typically it's going to be available about a month after we release the summary data. This means that the 1-year PUMS, in the past, has been released in October - one month after the September release of the pre-tabulated 1-year data. Then, 5-year PUMS files are released in January - about a month after the December 5-year release on AFF.

You could prepare yourself to process these files early, however, because we begin releasing the documentation about a week before the release of the actual files. So you can get your programs updated and up to date.

## **Slide 8 – Multiyear (5-year) PUMS Files**

One extra note about the 5-year PUMS files - you may have caught this earlier when I mentioned that the 5-year PUMS is the equivalent of five 1-year files (so it includes about five percent of all U.S. households). You might be asking yourselves why a person would wait until January for the 5-year PUMS when they can just merge the five most recent 1-year files in October.

So the short answer is that you certainly can if you want to, but we do do some nice standardizations for the five year PUMS files that might be worth waiting for - depending on your use. What we do is produce new weights for all the records so that the weighted estimates match the latest population estimates. We put in some adjustment factors for any dollar amount so that we can standardize all of them to the current year. That way no one is comparing apples to oranges.

And then other coding schemes are updated. An example would be codes like ancestry, occupation, industry, birth place - things that have a lot of categories. We're typically going to take a lowest common denominator approach, so that all the years have codes that are the same.

## **Slide 9 - Outline**

Next, I am going to talk about PUMS geography - this is another thing that is a little bit different and I want to make sure everybody's got a handle on it.

## **Slide 10 – Limited Geographic Detail**

There is limited geographic detail, basically to ensure the confidentiality of ACS respondents, the Census Bureau has to balance geographic detail with detail in the data. As I mentioned before, there are more than 250 variables on a PUMS person records. This means that we can't identify as many small geographies in the PUMS as users might hope.

Typically, we're going to give out regions, divisions and states on the file, but the only other geography is something that we call a Public Use Microdata Area and we call it a PUMA.

PUMS is not really designed for statistical analysis of small geographic areas, but the PUMAs can still be used for focus analysis and planning in areas of 100,000 or more. So this means many of the metro areas - but not all.

For example, Baltimore city, with a population of over a million is subdivided into six separate PUMAs, but not all places will be easy to analyze like this.

## **Slide 11 – Public Use Microdata Area (PUMA)**

What is a PUMA? It's basically an area with 100,000 or more population. That means that it is large enough to meet most of our disclosure avoidance requirements.

They are identified by five digit codes that are unique within each state and then they nest within the states or equivalent entities. They cover the entirety of the U.S. - so they're geographically contiguous and they're defined after each census.

So in most states the state data centers define boundaries of the PUMA - in some states the Census Bureau regional geography staff are involved in defining the PUMAs, and it takes a little while for us to incorporate these.

So PUMAs redefined after the 2010 census, were first used in the 2012 ACS PUMS file. This is going to be important later - I'm going to talk about what to do with multi-year files that contain two different PUMAs - so two different *vintages*, is what we call them, of PUMAs.

And that's - problem right now with the 5-year that we're releasing. So if you're using the most current 5-year PUMS files, you're going to run into this problem, but I'll give you some advice on how to deal with it.

They're mostly built—these PUMAs are built—on census tracts and counties, and they can be combined to create kind of rough approximations of different geographical areas.

When you look at the PUMA maps on the Census website, you also might want to know about the two different types of PUMAs mentioned. So if you go to TIGERweb to explore PUMA maps, you might see something called a one percent PUMA (or a super-PUMA) and a five percent PUMA. The PUMA used by the ACS PUMs is the same as the five percent PUMA that's used by the decennial census. (Just in case you run into that name for a PUMA.)

## **Slide 12 – Public Use Microdata Area (PUMA) Maps**

As with many geographic concepts, I think seeing a PUMA on a map can help you understand them better. Believe it or not, these two maps are at the same scale and these blue borders are PUMA borders. I think it's easiest to understand with these maps that it is really based on population, and not on geographic size.

So you can see that most of the PUMAs in Wyoming, with a population of more than about 500,000, are larger than the PUMAs in New Jersey, which has a population approaching 9 million.

The Missouri Census Data Center, the little link that you see on the slide (<http://mcdc.missouri.edu/websas/geocorr14.html>) has a fantastic geographic correspondence engine that they call MABLE and it can match PUMAs to other geographies of interest. So if your first thought is – well, how do I tell which PUMA includes my city of interest? Or, my city of interest is *this* and how many different PUMAs are overlapping in that city? This is a fantastic tool. They do such a great job.

You can also use the links at the bottom of the slide to explore the PUMAs in and around your geography of interest. That first link is a link to static reference maps (<http://www.census.gov/geo/maps-data/maps/reference.html>), and the second link, of course, is to TIGERweb (<http://tigerweb.geo.census.gov/tigerweb/>) which is a great mapping application if you haven't used it.

## **Slide 13 - Outline**

Next, I am going to talk about accessing PUMS data and then, if there is some interest, we'll see if we can take a couple of interim questions.

## **Slide 14 – American Factfinder**

Let me kind of walk you through how to get to this data.

If you are familiar with American FactFinder --hopefully you are, if you're also familiar with ACS data—it's FactFinder.Census.gov (<http://factfinder.census.gov>). There are a few different ways to get to the PUMS data and pull it down and download it.

I typically go the super-lazy route of just typing the acronym P-U-M-S into that yellow bar where there's a search field that says "topic or table name." (That's what I'm illustrating here.) You can type that in - it'll go in your Selections box up in that upper left hand corner and then it will offer you the most recent PUMS files in csv and SAS format.

You can also search on the left side. If you were to open that Topics (sort of grayish) rectangle on the left-hand side. You can expand something called “Product Type” and pick “Public Use Microdata Sample Files.”

## **Slide 15 – American Factfinder (Cont’d)**

Both of them are going to bring you to this next screen. You are going to be able to choose to download the entire US file - either the Housing Unit File or the Population File. You can download just the state, you can download a few states. Either way when you click on one of these blue links, you’re going to open a zip file.

And in that zip file there are going to be two different things - there’s going to be a ReadMe pdf, and there’s going to be the actual PUMS file. Please, please read the ReadMe. Contrary to normal Census Bureau acronyms, we did not give this a crazy long name that’s hard to tell what it is. We gave it the name PUMS ReadMe because we think it’s really important to read this before you get too deep into a file. So please read the ReadMe. If that’s the only thing you get out of this webinar, that would be great. It’s in every single zip file, it’s also on our website. You will always see it here so feel free to read it, download it.

When you open the csv file - that one with the Excel acronym - this is what you get. It looks very similar to that screenshot of the microdata that I had a few slides back.

## **Slide 16 – Census Bureau FTP Site**

Another way to get to the PUMS data is through our FTP sites. You can go through our website to get to the FTP sites, but it’s basically going to work the same. It’s just a little bit less clear exactly what you’re getting when you get into these.

So you would go through that URL at the bottom

(<http://www2.census.gov/programs-surveys/acs/data/pums/>) that goes to our FTP site and then specifically to the PUMS data. The year on that left image pertains to the data year.

So if you're looking for a 2014 one-year ACS PUMS file, you would open 2014 and you would get an image like what's on the right hand side. This is as if I'm looking for csv files - so that's what those first three letters are. Then an underscore H would represent a housing unit file. P would represent a population file or a person record file and then those last two letters are going to match a state acronym. So that first file (csv\_hak.zip) there would be the housing unit file for Alaska. Hopefully, that makes sense.

Then we go back one second - after you click on one of these csv files, it's going to be the same as what you would pull down from FactFinder. (It's going to have that ReadMe - please, please read it - and the actual data in whatever method you picked - SAS or csv.)

## **Slide 17 - DataFerrett**

The third option would be DataFerret. If you don't have statistical software or if you're not comfortable using it, learning how to use it, Census Bureau DataFerrett software is a great extra little product that we offer.

This tool searches and retrieves PUMS data - it can recode variables, it can create complex tabulations. It also - this is really nice - will allow you to download only specific variables. So if you don't have the space to store one of these huge files, going through DataFerret is a great way to do it.

## Slide 18 – DataFerrett Assistance

(Technical difficulties.) DataFerrett does require a little bit of time to learn how to use it so there are some really good resources.

This first one that I'd like to point out is on our website--the Census.gov/ACS website (<http://www.census.gov/acs>). You can go look at these PUMS DataFerrett video tutorials that will walk you through how to create a table in DataFerrett.

The other great resource for you, if you're not sure about using DataFerrett, is what we use to call our Compass Handbook. This one is the What Public Use Micro Data Sample PUMAs data users need to know (<http://www.census.gov/programs-surveys/acs/technical-documentation/pums/dataferrett.html> ). This is from 2009, so there is certainly some screenshots in here that are a little out of date, but the DataFerrett section, which is pages 12 through 23, is actually still as right on as if I wrote it yesterday. So, great resource for you if you don't have statistical software - if you have a little time and you want to use

## Slide 19 – Outline

DataFerrett. So next I am going to get into some common speed bumps that people run into.

But I wanted to just check - operator, I don't know if you're there, but if people have any broad overview questions, or if I covered anything too fast, or I mumbled - if anybody wants to ask a quick question, we have time for a couple.

Coordinator: If you would like to que up to ask a question, please press Star 1 and record your name when prompted. Your name is used to introduce your question. Again, to ask a question, please press Star and then 1. If you would like to withdraw a question, you may press Star and then 2.

Our first question is from (name removed) your line is open.

(Question 1): Excuse me - can you hear me?

(Marisa Hotchkiss): Yes.

(Question 1): Okay, just a question on merging the two files. I saw the instructions in the documentation, but I'm used to doing things the SQL way. Is it correct to just do a simple inter-join of the two files on serial number?

(Marisa Hotchkiss): The serial number is absolutely how you would join them and for anybody that's confused about what he is asking, I'm going to talk about it in a second. I'm not super-familiar with SQL, so I might have to follow-up about that question. But in general, yes. You can just merge a person file and a housing file based on serial number match and that should get you a combined file.

(Question 1): Thank you.

Coordinator: The next question is from (name removed). Your line is open.

(Question 2): Hi, I'm just curious about the different sample sizes between the ACS and PUMS. Can you talk a little a bit about why the PUMS data doesn't use the full ACS sample?

(Marisa Hotchkiss): Yes, it's mostly for confidentiality. It's a little bit safer if not every single respondent who responds to the ACS has their record then in the PUMS file. I mean...

(Question 2): Yes, that makes sense.

(Marisa Hotchkiss): Okay.

Woman: Yes, I have a question.

(Marisa Hotchkiss): Go ahead.

Woman: So I guess two questions - the first one is - there's a place in Work PUMA and I'm wondering if the minimum criteria of the 1000 residents - usually for the residential PUMA, applies to the place of work PUMA as well for those geographies. Is there also at least 1000 residents or workers in these boundaries?

(Marisa Hotchkiss): I believe so. I'm not a hundred percent confident, so if you want to hold it to the end I can verify. There are a couple of different PUMAS - you have your regular PUMA which is based on where people live, their place of residence.

Then you have the place of work PUMA - like you were talking about. And then there's another one called a Mig PUMA that has to do with migration variables.

And I believe that they all have the same boundaries but another good way to check is when I walk through the resources, we have links to the code lists that would have the numbers for the place of work PUMAs. (You can also call it a POWPUMA, which I think is kind of a fun name.) But it would have the codes for those POWPUMAS and then you could look those up on the maps and make sure that they do correspond. (But I can double-check that for you.)

Woman: Okay, great. And the other quick question I think that's in DataFerrett, are you able to calculate standard errors with your estimates?

(Marisa Hotchkiss): You are, but it's a little bit more complicated and I believe it's discussed in the Compass Handbook. So yes, if you're comfortable with using the replicate weights, then yes.

Woman: Okay, thank you, and that would be in DataFerrett.

(Marisa Hotchkiss): Yes, and I can talk about weighting more in a second too. So if everybody is good holding their questions to the end, I'll go on with the rest of the common speed bumps and where you can go to get resources for more information. And then I can answer any additional questions that people have. I just wanted to make sure that we didn't get anybody really confused or was ready to drop out, or if I mumbled - you know what, PUMS was—and you missed that whole first slide.

So if everybody's good with that then I'll get through these next few slides and we can do more Q&A at the end.

## **Slide 20 – How Do I Put PUMS File Together?**

So this is related to the first question - how do I put PUMS files together? So you're ready to download a PUMS file - you open the zip file and you see it's in four pieces. (That's what this image on that top right is.) If you were to open the US file from a 5-year PUMS file, you would see (ss14pusa.csv, ss14usb.csv, ss14 usc.csv, ss14 usd.csv) –what do you do with that?

Or, if you have an analysis in mind and discover that one of the variables you want is on the housing unit file and the other one is on a person file, now what do you do?

Some of the PUMS files are so large that we have to release them in pieces and then unfortunately we have to ask you, the users, to put them back together before you can really use it.

And of course we separate housing and person variables because putting all of them together in one big file would be pretty unwieldy. On the page you can see some of the pseudo code—this happens to be from a SAS program -- but this is an example of how you would do this.

You would merge the files together. You can do Housing and Person files based on serial numbers (SERIALNO) - so the serial numbers should match.

Basically there would be one housing unit record with a serial number and then any of the people that lived in that housing unit, and have person records, would share that serial number.

You could also merge any of the a, b, c, d, or a, b, files together just by creating a set. This example code is also in the PUMS Read Me (I know, it's crazy!) and on the PUMS file structure page, which is a page on our website, and it's that URL at the bottom.

Also just a note if you are combining records, limiting the number of records you're processing by selecting only those of interest, will often speed up your processing speed. So if you can limit it as you're merging things together - that will make it a little faster.

## **Slide 21 – Which weight should I apply?**

Another common pitfall is which weight should I apply? So you apply a weight to an estimate that you have created with the PUMS file to make sure that that estimate is then representative of the total population.

But to make things more complicated, we have three different kinds of weights that are included on the PUMS file. So another common question is - which one am I supposed to use?

There are three that you see in bullets here are the three basic types - WGTP - is something we often call (weight-P) and this is a PUMS household weight.

So if you apply this to an estimate - your resulting number - your resulting estimate would be representative of housing units. The second one, the PUMS Person Weights - PWGTP (P-weight-P). If you apply this to an estimate, your resulting estimate would then be representative of people.

Third type is Replicate Weights - these are going to look like WGTP and PWGTP except they're going to have a number at the end of them and there are going to be 80 of each - 80 WGTP with a number after it and 80 PWGTP with a number after it. And these are what you would use to calculate standard errors.

Unfortunately that could be an entire presentation all on its own. So I would say look in the Compass Handbook that I showed earlier. It has a good discussion there. There is also some really great statistical guidance in the PUMS ReadMe and then if you need additional help, you can always contact us and we have some contact information at the end of this presentation.

## **Slide 22 – Why Don't My PUMS Estimates Match AFF?**

Another common question that comes up is - why don't my PUMS estimates match AFF? I think, ideally, by now, you picked up that the PUMS is a sample. Because it's a sample and the difference is about two million cases, it's going to cause any estimates that were produced from either one to be slightly different.

The PUMS also has some extreme values that are grouped together. (So that's that top and bottom coding.) And the categories are going to be broader and general.

So the purpose of PUMS is to create estimates that are not available on AFF, not to create AFF - I would not worry really about your PUMS estimates not matching AFF.

So I think most people know this - I think they're doing it to check their work with their weighting and checking the merging of the files. If this is your goal, we do have files that are called the PUMS Estimates for User Verification, and these are files that data users can use to see how our statisticians have created estimates using the same PUMS microdata.

So that's a better way to check that you got all your ducks in a row in the program.

### **Slide 23 – How Do I Use Dual Vintage PUMAs?**

And then the last one here is dealing with those dual-vintage PUMAS.

So what do you do when you have two different PUMAS in an area?

Let me just reiterate the multi-year PUMS files with the years before 2012 - like 2010, 2011 and those years that are after 2012, including 2012 - so 2012, 2013, 2014 are going to have two different PUMAS—it's going to have 2000 PUMAS and Census 2010 PUMAS.

Unfortunately this does not mean that each record has both PUMA codes. Rather, records from data years 2010 and 2011 have the Census 2000 PUMA codes and the records from 2012, 2013, 2014 have the Census 2010 PUMA codes.

Why don't we just put both PUMAS on every record? It's a disclosure risk, unfortunately. A data user could potentially take an old PUMA that barely overlaps with a new PUMA, figure out which records live in that sliver, and then it could potentially compromise their confidentiality.

So, unfortunately, that's more of that balance with geographic detail and case detail. There are a few solutions to this problem. The best solution is wait until the 2012-2016 5-year file – that will come out in 2017 - not an ideal solution.

Another way you can do it, is just to use state level estimates - also not the best solution. So the third solution is kind of to be comfortable with geographies that have fuzzy boundaries.

I know this isn't what everybody wants to hear but it's kind of the best way to think about it. Maybe your border is just going to be a little bit thicker than we would ideally like or that we're used to getting with Census Bureau Data products.

So how would you do this? You would determine which 2010 and 2000 PUMAS most closely approximate your area of interest. You would look at the pre-2012 records - so 2010 and 2011 - that are in that Census 2000 PUMA.

You would then look at the 2012, 2013, 2014 records that are in the Census 2010 PUMA, put them together, and then do your margin of error calculation.

One extra little note if you're trying to do this in DataFerrett, PUMA, because of this dual-vintage headache is not a selectable geography - it's in there though. So both sets of PUMA codes are in the population set of variables. So they are there.

## **Slide 24 – Use Caution...**

One final set of speed bumps that you might run into, Three million records sounds like a lot. But when you start splicing it down by different variables and PUMAS, you can quickly narrow your focus to too few cases.

So please make sure that you are calculating margins of error and you have a sense of the reliability of your results and estimates.

When the number of unweighted cases is too small, you can consider adding cases from neighboring geographies, broadening categories, or creating multi-year files on your own.

Please also keep in mind that extreme values have been grouped together. So this is not the dataset to use to determine how many millionaires, mansions, centenarians, etc. live in your state. Likewise, it's unwise to assume things like - no one ever spends more than (dot-da-da) on their rent in this area.

So if your value is on the extreme end of the range, it's a really good idea to look into the top and bottom code for each PUMS file. So this is something we publish. For example, you could go to our website and find out that in Florida, we're not going to release the age of anybody who is older than 94.

So if for example somebody was 100 years old when they responded to the ACS, we would top code them down to be an age of 94. So this does not mean that they're lying about their age, it just means that we're trying to kind of protect people that are outliers.

## **Slide 25 - Outline**

Documentation and Resources.

## **Slide 26 – ACS Main Page**

This is the [Census.gov/ACS](https://www.census.gov/acs) web page. Hopefully if you're familiar with ACS, you have been to this web page before.

The two main places that you're going to find PUMS information are in the Data section and the Technical Documentation section. (There's a little circle on the left-hand side in that navigation.)

## **Slide 27 – PUMS Data Page**

If you go to the Data section and you choose PUMS data (<http://www.census.gov/programs-surveys/acs/data/pums.html>), these are links to all of the most recent PUMS data that we have released. Anything back to 2005 is going to be available through American Fact Finder.

We do have some PUMS data for 2003-2004, however, it's not going to be available for every geography because this is during the period where ACS was in a demonstration period, trying to figure out if for a few test counties, the ACS was really feasible for the whole U.S.. So your geography might be in there - it might not. The documentation is also not as great as it is now.

If you are interested in some of the test PUMS data from 96, 97, 98, we could provide it on DVD, but it's only going to be for a few geographies.

## **Slide 28 – PUMS Technical Documentation**

If you go to technical documentation (you have the little thing there on the left), then choose PUMS Documentation - there's going to be seven different pages about PUMS. Hopefully, we will answer your questions somewhere in one of those seven pages.

Probably the most important one for users is that last page: PUMS Technical Documentation. It's also the URL at the bottom of this slide here (<http://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>).

On this page is going to be a bunch of different attachments that have pretty exhaustive information about the PUMS files. Here again, first thing you're going to see is the PUMS ReadMe. But there's also going to be a list of subjects included in the PUMS for each file. There is going to be a data dictionary (which I will show you an example of in just a second).

There are code lists - so if you're looking at what POW PUMA or what MIGPUMA had this code and what does that encompass, or if you're curious, what is occupation code 343.

All of that information is included in these very detailed code lists. Those are great if you're dealing with a variable that has a lot of different codes. That top coded and bottom coded values is going to be what I was talking about before. (So you can see which extreme values have been cut off and at which point). Accuracy of the PUMS - this is going to get into more sample design, estimation, methodologies and more statistical concepts about the PUMS.

And that last one is PUMS Estimates for User Verification. So if you want to check that you're building everything appropriately, that's a great place to go.

## **Slide 29 – PUMS Data Dictionary**

This is that PUMS Data Dictionary that I mentioned. So if you download your PUMS file and you have no idea what AGE<sub>P</sub> means, you can come to the Data Dictionary and find out that (AGE<sub>P</sub>) is the variable name for age. Similarly, if you're not sure if that person number 5 there really has an age of 0 and what on earth that could mean, you can come to the Data Dictionary and find out that that means that person is under one year of age (not in fact ageless though, I'm sure we would all like to be immortal, that's not actually what that stands for).

## **Slide 30 – PUMS Resources**

Additional Resources. As I mentioned, that Compass Handbook - What Public Use Microdata Sample Data Users Need to Know - is great. The website instructions are going to be a little bit outdated but the concepts and the text are still very sound.

There's also, on our ACS Data Users Group, they have an ACS Public User Microdata Sample Group that discusses a lot of PUMS questions. They are very knowledgeable, excellent people, so if you have questions, I would consider joining their online community.

And you can also sign up for messages through our emails that we at the Census Bureau send out. There is a specific category for Public Use Microdata Sample Files. So if we're doing anything, if we're releasing any data, or if we're having a webinar, like we're having today, we would let you know through those emails.

## **Slide 31 – Source Us!**

Finally, please source us if you're using us - it helps get new data users comfortable. It helps other people see and be inspired by what you can do with PUMS. There are a couple of different ways that people have done this.

We just ask that you cite ACS if it's a specific year that you're using, and that it is actually PUMS so we can kind of see what you're doing as well. And hey, you might get a shout-out in a webinar slide, which is always exciting.

## **Slide 32 – Continue the Conversation #ACSdata**

And if you have questions, please continue the conversation with us. You can sign up for those email alerts. You can go to our website to learn what

we're up to and to ask questions through the FAQs that are also on that website.

You can, of course, call our call center or our data user number which is that 1405 number - you can also please use our hashtag #ACSdata on Facebook, on Twitter, on Instagram, on Pinterest. You can make a video of how fun it is to use PUMS and put that on YouTube - that would also be great.

### **Slide 33 – ACS Data Users Group**

Another quick ad for the data users group. It is free and open to any ACS data users and they are very knowledgeable data users on the online community.

### **Slide 34 – Need Local Stats?**

So I think that the last thing that I wanted to mention was that you can also get a presentation in person or a workshop from our very knowledgeable data dissemination specialists located throughout the country. So if you have a need for that. It could be about anything –it doesn't have to be about PUMS - I would also contact them.

### **Slide 35 - Questions**

I think that's all I have if we want to do some more Q&As.

Coordinator: Thank you, as a reminder if you would like to ask a question, you may do so by pressing Star1 and recording your name when prompted. If you find you would like to withdraw a question, you may do so pressing Star 2.

Our next question is from (name removed). Your line is open.

(Question 4): You meant (name removed).

Coordinator: Yes, your line is open.

(Question 4): Okay, wonderful. Hi, okay I have two quick questions. The first one is about the margins of error. Do you think that everyone who uses PUMS actually goes ahead and capture to see their own margins or errors, or do you think people might just kind of (stay in) and it's like - and just sort of not do it?

I don't know if that's a really weird question and then the second question is - okay go ahead.

(Marisa Hotchkiss): I'm trying to figure out how to answer this. I'm going to say everybody *should*. What I mentioned before where you can easily run into situations where you've limited yourself to very few cases. And it's good just to have a sense of how reliable the estimate you've created is.

Because you can't always do things like add more cases or expand your geography to get something that seems a little bit more reliable.

(Question 4): Right, right, okay. Because I just feel like I - maybe some people don't do. Okay, and then the second thing really quickly is - you mentioned that the 1-year could be sort of manually added up together to basically be similar to the 5-year.

And I was wondering how that differs from ACS estimates. I don't know again if that's a clear question, but I was just seeing that I was looking at that difference). Yes, go ahead - thank you so much.

(Marisa Hotchkiss): So the short answer is that you would basically use the same method that you would use to combine a person and a housing file, or that you would use to combine different pieces of a U.S. file.

You're basically just concatenating those files. There are some things that you want to keep in mind, so if you're concatenating or merging your own

PUMS file where we have not standardized those codes, you could run into some issues with different vintages of codes. You could run into issues if you're comparing different dollar figures without adjusting them. You could also have some issues with weighting.

It's on our list to come up with some better guidance for the people that want to do this, but yes, you absolutely can.

(Question 4): Okay, is that different from like the ACS one - your estimates? Like could someone just (delete) and concatenate five years' worth of one year ACS estimates in the same way, or is there a difference in how that would work with ACS data as opposed to top status?.

(Marisa Hotchkiss): So the data that I'm talking about mostly is the ACS PUMS data.

(Question 4): Oh, the ACS PUMS data - okay, I think I meant like...

(Marisa Hotchkiss): Like the summary data?

(Question 4): Yes, that's what I meant - I'm so sorry to be not - yes, that's what I meant - summary data.

(Marisa Hotchkiss): Yes, it really wouldn't work as well then because that data has already been weighted and...

(Question 4): So I see - okay - okay - okay, great. Thank you so much.

Coordinator: The next question is from (name removed). Your line is open.

(Question 5): Yes, hello I'd like to follow-up on the matter of margin of errors. If you would spend just a couple of minutes - maybe reminding people the importance of calculating margin of error, evaluating them, especially

when it comes to cases where the margin of error is larger than the estimate.

And maybe a recommendation or suggestion to also publish margin of error so that users are clear about the reliability of an estimate.

(Marisa Hotchkiss): Yes, absolutely and it's actually a great plug because there is going to be a webinar on April 6 that is going to focus on understanding margins of error while working with ACS estimates, how to use them, how to approximate them.

So if you're listening on the line and margin of error is kind of foreign or a scary concept that would be a great webinar to also tune into.

Essentially, it's important because these are estimates. They are always going to be reliable within a range. So when the ACS produces our summary data, the data that's available on American Facts Finder, we typically use a 90% confidence interval.

So what we're saying is that we're confident that within that interval, the actual true – I'm thinking of exactly how to say this - but the actual true value would be in that range. It may not be exactly the estimate as we have presented it, but it would be within that range.

So I think that's important to understand - I think people can often feel intimidated when they see a margin of error because they assume that that means the data is wrong, and that's not really what's going on.

It's more that we're saying we're confident that, because of sampling error and non-sampling error and all the different things that can be done, it's still survey data, it's still sample data and we're not sure that our estimate is exactly what is true out in society.

It's true within a range and I think that that's just an important concept for everybody to use, even when you're using - you're creating your own estimates out of the PUMS data.

I kind of share the same feeling with you that a lot of people don't publish their margins of error and I think it's because they think that other data users will feel that that means that the data is not accurate. But it's really more... Yes, go ahead.

(Question 5): Well, I was going to follow-up with - this is even true with using ACS data that I think in the past particularly with the old Census data, we would publish one number and say that that is the number.

But especially with the transition to the ACS and the smaller sample sizes, etc. that the margins of errors in many cases did get very large and larger than the estimate itself, and people still published just the estimate.

But I think - and even in cases when the margin of error is larger and the estimate - and so I think it is important that margin of error gets published and does get calculated.

I thought that would be a good time to remind people that - that is an important part. Also evaluating it when you go to decide on which geographies you can or cannot publish estimates for.

(Marisa Hotchkiss): Right, and I think like I mentioned before, it's really important to do it because you might want to go back and choose a larger geography or add more cases to your estimates if you feel that it's not up to your standards for what you want to use.

(Question 5): Okay, well thank you very much.

Coordinator: As a reminder, if you would like to ask a question at this time, please press Star and then 1 and record your name when prompted.

The next question is from (name removed). Your line is open.

(Question 6): Thank you. I was wondering if you had any additional guidance on combining multiple years of PUMS samples. I know there used to be three-year estimates, but they were discontinued this past year.

So in terms of creating a home brewed three-year estimate, is there any information within the dataset, or also publicly available that can adjust for cost of living or inflation, or income variables or the other variables that we know change over time but might not know how much to adjust for.

(Marisa Hotchkiss): So it's a tough question and like I said, we're working on coming up with some better guidance that we can put out there. At the moment, it kind of depends what variables you're using.

I can't remember if I have it on the slide - you can email us. We have an email address that's [acso.users.support@census.gov](mailto:acso.users.support@census.gov), and if you let us know kind of which variable you are looking at, we can give you a sense of - yes, there should be no issue with those or check those out in the Data Dictionary, or, oh yeah, we changed vintages between those two years, so you're going to need to use this crosswalk. So for the time being if you let us know - kind of what analyses you're looking to do—we could give you some better, kind of personalized guidance.

But we are looking trying to figure out how to guide users because it would be ideal if you could create your own three-year PUMS file, you could certainly create a two-year file which we haven't done before.

I would—note of caution—if you start making something huge like a 10-year file, that’s going to get unwieldy, so I would stick towards the lower end here. But it’s definitely a great option to explore, and if you let us know at [acs.users.support](mailto:acs.users.support) we can try and guide you through doing that.

(Question 6): That makes sense, thanks so much. I guess I just wanted to recreate the three-year file that I was looking for but wasn’t available this year.

(Marisa Hotchkiss): Yes, that makes sense.

Coordinator: The next question is from (Question 7). Your line is open.

(Question 7): Thank you so much for this presentation. I had a follow-up question about Crosswalk PUMAs based on the 2000 Census - 2012 Census. On this very website that you referenced earlier, you actually have a Crosswalk and they allow - there’s an allocation factor going from PUMA 2000 to 2012.

So would you recommend using this method to proportion out their percentages?

(Marisa Hotchkiss):

Yeah it’s kind of going to be the same thing, like do you want to have the data be a little bit fuzzy or have the border be a little bit fuzzy. There is no like perfect match unfortunately.

I think I tend to err on the fuzzy border side but I’m sure there are people that have done that.

(Question 7): So the Census doesn’t not recommend using that approach, okay.

(Marisa Hotchkiss): So we haven't - so this is Missouri State (sic. Census) Data Center and they are great but it's definitely their product too. So we don't really have anything like that.

(Question 7): This time can I ask a specific question. For the five-year file, I found that combining the single file into five years compared to the five year ACS, that the weighted totals are actually different.

And I think - is it because in the five year file, the weights are actually of each year are actually divided by five, so that you get like the average annual estimates across the five years, versus the cumulative estimate. Is that correct in my understanding?

(Marisa Hotchkiss): Well, without seeing your code, I'm not sure that we do updates the weights and when we release a five-year PUMS file, we update the weights so that they match the most recent population estimates, and we update those on all the records.

So it's not going to be - even if you concatenated the five one-year, you probably wouldn't match exactly with our five-year PUMS file. I think - I don't remember off the top of my head, but I believe it within a couple of percentage points. So it's fairly close but it wouldn't probably match exactly.

If you also wanted to follow-up with [acso.users.support@census.gov](mailto:acso.users.support@census.gov), we could probably get you a better answer from one of our statistical methodologists or mathematical statisticians about exactly why it's different.

(Question 7): Okay, thank you.

(Marisa Hotchkiss): I think we're at 3 o'clock.

Coordinator: If you want to take further questions at this time?

(Marisa Hotchkiss): Yes, we can take a couple more questions.

Coordinator: The next question is from (name removed), your line is open.

(Question 8): Hi, thanks for taking my question. I joined a little late because of an overlap of another meeting. I just wanted to ask really quickly, are you going to circulate this presentation. You have a lot of links in it and I'm wondering if I will access to them after this.

(Marisa Hotchkiss): Yes, so on the Census Bureau website, [Census.gov/acs](https://www.census.gov/acs). We have a library link on the left-hand side of the page and then we have some educational resources. So our plan has been to update the presentations that are there.

I think this might be one of the first ones to get updated and then I believe elsewhere - like in the trainings and workshops, that educational resources page. So if you go to the Data link at the top of the page and choose Trainings and Workshops, I believe that they're also archiving recordings and presentations.

(Question 8): Okay, great, thank you.

(Marisa Hotchkiss): But also if you would like to email [acso.users.support@census.gov](mailto:acso.users.support@census.gov), I would be happy to send it to you.

(Question 8): Say it again.

(Marisa Hotchkiss): [acso.users.support@census.gov](mailto:acso.users.support@census.gov).

(Question 8): Thank you.

Coordinator: The next question is from (name removed), your line is open.

(Question 9): Hi, I had another margin of error question. Well in the tabulations that I create -- I'll create what I call a low value and a high value -- that incorporates the margin of error.

I'm just wondering is there - for example - a 90% confident that the real answer lies between those two numbers, or a 95% confidence, or what? Do we know what the probability is that the number is between the high value and the low value?

(Marisa Hotchkiss): It kind of depends on how you're creating it. There is in - I believe in the ReadMe and elsewhere in that PUMS technical documentation website, there is some guidance about how we advise people to create the standard errors using the replicate weights, and I believe they talk about the confidence interval in there.

If they don't, if you want to use that same email, I can talk to one of our mathematical statisticians and see if they can tell us.

(Question 9): Okay, you know what - I think you're right. It's just been so long since I looked at it, I had forgotten and you brought it up today and that triggered it in my mind but I do believe it is brought up in that documentation.

(Marisa Hotchkiss): I look at it almost every day and I can't exactly remember if they talk about confidence interval but I think they do.

(Question 9): Okay, well it's been years since I've looked at. Okay, well thank you very much. I appreciate the webinar too.

(Marisa Hotchkiss): Sure, thank you.

Coordinator: We have no other questions in que at this time.

(Marisa Hotchkiss): Alright. Well thank you everyone.

Coordinator: This concludes today's conference. Thank you for your attendance. You may disconnect at this time.

END