How To: Download IPUMS Data & Open it in Stata

Carolin Ferwerda, Wellesley College
Revised October 2012

This document describes the steps required to access the Public Use Microdata Samples from the U.S. Census. All other data available from IPUMS have similar extraction procedures. Some of these steps will be relevant for other data available online as well, such as the Demographic Health Survey.

Step 1: Go to www.ipums.org

This is a website where you can access many years of micro-data. The IPUMS staff have integrated the micro-data samples so that you can easily select the variables you need for your analysis.

Step 2: For U.S. Census Data, click on "IPUMS-USA"

On the IPUMS-USA page, you will see menus in the sidebar that allow you to explore Data or Documentation. It is a good idea to first become familiar with the documentation. Start by reading the "User's Guide" and "What is IPUMS-USA?" in the FAQ.

NOTICE: For **IPUMS-International** data, you must apply for access. Sometimes this can take several weeks. I strongly suggest that you apply immediately using a description of your project as it is now (even if you haven't completely finalized your topic or approach). From the IPUMS site:

> "Access to the [IPUMS-International] documentation is freely available without restriction; however, users must apply for access to the data. The application system requires a description of an applicant's proposed research, and asks for the user's institutional affiliation and other information to verify identity. Every application is individually reviewed by project staff. Applicants are required to agree to a number of conditions to use the data. Access to the system enables a user to extract data from any country in the database. To apply for access go vist http://international.ipums.org/international-action/menu "

Step 3: Familiarize Yourself with which Variables are Available

After reading the above documentation, click on "Browse & Select Data" in the Data sidebar menu. This will bring you to a webpage where you can browse variables by Household, Person, alphabetical listing, or via a keyword search.

The Census data are collected by sending a questionnaire to households that are selected to be in the sample. The questionnaire includes questions about the household and about the people in the household. Thus, there are different ways of organizing the data. If you only want information about households (e.g., what fraction of households had a flush toilet in 1960?) then you could select only variables defined at the household level and get a dataset that had one observation per household in the sample.

On the other hand, if you want information about individuals (e.g., what fraction of people in 1960 had a high school degree?), you would need to access individual level data. You would want a data set that has one observation per person in the sample, rather than one observation per household in the sample.

It is a good idea to spend some time clicking on the variables described under the Household Record list and Person Record list. This will familiarize you with what is available in each year.

For example:

1. Click on “Group Quarters” in the Household Record menu. This will bring you to a table with various permutations of the group quarters variable down the page and years across the top. An “X” indicates that the information is available for that year.
2. Click on “GQTYPE” in the “Variable” column for a description of the variable
3. Go back to the previous webpage.
4. Click on the “codes” column for “GQTYPE.”
    a. This will bring you to a page that describes the level of detail available for the type of group quarters that people might be housed in. The level of detail varies from year to year. Note, for example, that in 1990 and 2000, one can identify whether the household is in an institution or a non-institutional group quarters setting. In 1980, more detail was available, so one can identify if the “household” is a correctional institution, a mental institution, or college dorm.
5. Go back to the main “Browse & Select Data” page.
6. Click on “Education” in the Person Record menu.
    a. There are several variables that pertain to education. Notice that different censuses have had different ways of getting at the question of “what is the highest level of schooling attended.” In the 1800s, for example, the Census just asked about literacy and whether the person was currently attending school.
    b. The IPUMS staff have created a variable, called “EDUC,” that reconciles the information about highest level of school attained for all Censuses since 1940.
7. Click on the “codes” column for “EDUC.” This shows you for which years the data are available and what the codes mean.

Step 4: Decide which Sample You Need

Once you determine which variables you want, you have to decide which sample you need. In part, this is based on what variables are available in each sample. As discussed above, some data are available in some years and not in others. So, that is part of choosing your sample.

In addition, there are differently sized samples available in some years. Go back to the main “Browse & Select Data” page, and click on the blue “Select Samples” button.

In 2000, there is a 1% and a 5% sample available. Because the 1% sample has fewer people in it, the Census is more concerned about confidentiality issues. Thus, less geographic detail is available for the smaller samples. For example, the variable “PUMA” is only available in the 5% sample of the 2000 Census. So, if you want to analyze how metropolitan area policies affect individual level outcomes, you will need the 5% sample. Note that the ACS (American Community Survey) is available for more recent years and that the percentage sampled varies by year (1% some years, 3% other years). Click on the links for each sample to find out more about the specific sample size.

Go back to the previous page and click on “Geographic” in the Household menu. If you click on the link in the “Codes” column for “PUMA” you will note that there are lots of metropolitan areas that are not identifiable in the 2000 Census. Only large ones (>100,000 residents) are identified.

In addition to the availability of different variables, the obvious difference between a 1% and a 5% sample is that the latter is 5 times bigger. If you want to analyze labor market outcomes for Cape Verdean immigrants in Massachusetts, you will want to get a sample from the 5% Census because you are interested in a small group.

For many projects, the 1% sample will be large enough (and I would recommend sticking to the 1% sample if you do not need geographic identifiers or data on a small group). Just check to make sure that the critical variables for your project are available in that sample.

Step 5: Choose the Sample(s) and Variables for your Extract

1. Go back to the "Browse & Select Data" page. The first step is to select your sample. Click the blue "Select Samples" button. Which sample you select depends on the variables you need and the type of analysis you are proposing – see Step 4.
    a. Uncheck "Default U.S. Sample from Each Year" to deselect all the samples. Then put a check next to the sample(s) that you want. For this example, choose "**2000 1% sample**."
    b. Click "Submit Sample Selections" to add the sample(s) to your cart.
    c. You should see 1 sample listed in your Data Cart at the top of the page.
2. Now it's time to select variables to include in your data extract.
    a. Under "Geographic" in the Household menu, choose STATEFIP by clicking on the yellow plus sign in the left hand column. When a variable is selected a green checkmark will appear next to it
    b. Go to "Demographic" in the Person menu. Choose AGE, SEX, MARST.
    c. Go to "Race, Ethnicity, and Nativity" in the Person menu and choose BPL.
    d. NOTE: There are some variables that you will need every time you create an extract. Be sure to select the following variables:
        i. Choose "**GQ**" under Household | Group Quarters. It's important to select this variable because most studies want to exclude people living in group quarters. It doesn't make very much sense to try to analyze wages for people in institutions, so the group quarters variable (**GQ**) is needed so that these people can be excluded.
        ii. Choose "**PERWT**" under Person | Technical—note that this variable has **[preselected]** next to it. That's because this variable is so vital to using the data that the IPUMS people preselect it for every data extract. IT IS VERY IMPORTANT THAT YOU TAKE THE "**PERWT**" VARIABLE. THESE DATA ARE NOT NATIONALLY REPRESENTATIVE UNLESS YOU USE THE APPROPRIATE WEIGHT. In some samples, weights are not necessary, but they are in the 2000 and 1990 Census data. The general documentation tells you about whether or not weights are necessary for the sample you choose.
        iii. Choose "**PERNUM**" under Person | Technical and "**SERIAL**" under Household | Technical. These variables will allow you to match people and households across different extracts from the same sample. This is useful if you have to go back and extract additional variables. Note that PERNUM is not automatically selected for you, though SERIAL is.

Step 6: Create the Extract

1. The Data Cart should now contain 1 sample and 5 variables (preselected variables don't count). Click "View Cart."
    a. Review the selected variables and sample(s). If you need to make changes, click Add More Variables/Samples.
    b. Click "Create Data Extract" to continue.
2. You will come to the Extract Request summary page. Here you can revise your sample and variable selection, as well as set up some additional options.

3. The third row allows you to select a data format. The default is a raw data file with a corresponding set of commands to read in the data to the software of your choice (Stata, SPSS, SAS, etc). You can also choose to download the data in Stata format.
    a. Click on the "Change" link next to Data Format.
    b. From the list, choose "Stata (.dta)" as the data format. Leave Rectangular as the default.
    c. Click "Submit"
4. In the fourth row on the Extract Request page, you will see "Structure: Rectangular". This refers to the structure of the dataset. If you are analyzing *people*, rectangular structure (the default) is fine because it will give you a person-level dataset. If you chose some household-level characteristics they will be attached to the persons file. If you *only* want to analyze *household* characteristics with no person-level information, click the "Change" link for Structure and change the dataset structure to "Household Records Only." For this example we are interested in person-level data, so just leave the default of "Rectangular."
5. Further down on the page, there is an Options section. Here you can set up some special options for certain variables.
    a. Click "Attach Characteristics" in order to attach data about other members of the household to the individual's record. For this example, click on "Attach Characteristics" and choose "Father" and "Mother" for MARST. This will attach information about the individual's parents' marital status. These variables will be named MARST_POP and MARST_MOM in the extract. The variables will have a missing value if the individual's parents are not living in the household, but this does not mean the individual does not have parents! Click "Submit" when you have finished attaching characteristics.
    b. You can use "Select Cases" to create a subset of the data extract that matches specific values for the selected variables. This is useful if you know that you are only interested in, for example, people who are younger than 25. Use this option with caution, though, because you may inadvertently exclude part of your sample! For this example, do not use this option.
6. Last, you will need to include a brief description of the data. Something like "Person level data from the 1% Sample of U.S. 2000 Census for Econometrics Class project" should suffice.
    a. After double-checking that you have what you need, click the "Submit Extract" button at the bottom of the page.
7. You will be asked to register if you are a new user or log in if you have been to the site before. Go ahead and do that.
8. This will bring you to a "Download or Revise Extracts" page with instructions about what to do next, plus a list of any previously created data extracts. You may need to wait until the next day before you get an email telling you that your extract is available to download, but often you will receive the email within just a few minutes.

Step 7: Download the data, codebook, and command files

When you get the email from IPUMS there will be a link. Click on the link to go to the download page.

Download or Revise Extracts

Use the links provided below to download a data extract (right-click the links for the data, command files, and codebook) or to revise an extract (that is, use a previous extract as the basis for defining a new extract). For instructions on downloading and opening an extract on your computer go here. Note: data files will be available for 72 hours, after which they are subject to deletion.

Extract Number	Date	Formatted Data	Fixed-width Text Files				Codebook		Revise Extract	Resubmit Extract	Description (click to edit)
			Data	Command Files							
22	2012-09-25	STATA	Data	SPSS	SAS	STATA	Basic	DDI	revise		Data for 203

If you would like to have more control over how you read the data into Stata:

Download the data, codebook, and command files. Be sure to download the STATA version of the command file. This will put a large zipped ascii data set, a codebook, and a STATA do-file on your computer.

1. Click on the data file to download it.
2. To download the codebook and the do-file, right-click and select **Save Link As**.

If you would like a simple Stata data file and you do not need control over how the data is read into Stata:
Simply download the formatted data by clicking on the **STATA link** under "Formatted Data." This will put a large zipped Stata dataset on your computer.

Step 8: Extract Your Data from the Compressed Files

IPUMS gives you the data in compressed format (the ".tar.gz" extension) in order to make the download faster. Before you can open the data in Stata, you must uncompress or unzip this file.

*If you are working on a **Mac***, double-click on the file to decompress it. The file will be added to your downloads folder or wherever you saved the .tar.gz file.

*If you are working on a **PC***:

1. Download "wingzip.exe" from the link http://www.irnis.net/soft/wingzip/
   a. Note: Skip this step if you already have WingZip, 7-ZIP, or WinZip on your personal computer. Lab PCs do not have WingZip.
2. Double-click on wingzip.exe and click Run. Click Install.
3. Browse (…) and choose your IPUMS ".gz" file as the input file.
4. The default name for your output file is the same as your input file without the ".gz".
   a. If you change the name of your output file, remember to make the same changes on your do-file as well.
5. Click "Start."
6. This will create a large file called "usa_#####.dat" (raw data) or usa_#####.dta (Stata dataset).

Step 9: Read Your Data into STATA

Option A. If you are working with the Stata-formatted data file (.dta)
Move the "usa_#####.dta" file to into your working directory (e.g., T:/username or My Documents\Econ203). Then double-click on the file to open it in Stata.

If you receive an error that there is not enough memory to load the data, quit all other applications except Stata and try again. If the file still won't open, revise your IPUMS extract to include fewer samples (years) or fewer variables. Another option is to subset your dataset, but you should talk with me about if this is appropriate.

Option B. If you are working with the raw data (.dat) and do-file
Note: These steps are not necessary if you are working with the Stata data file (.dta).
The file called "usa_#####.dat" contains the raw data. If you open the file with Notepad you will see large blocks of numbers. In order to make sense of these numbers, you need to use the IPUMS do-file to read the data into STATA.

The first step is to move the IPUMS ".dat" file, the do-file, and the codebook into your working directory (e.g., T:/username).

Once the files have been moved to the folder you want to work out of:

1. Start STATA and change your working directory (**cd** "pathname") to the folder containing the data file, do-file, and codebook.
2. Using the do-editor in STATA, open the do-file that you downloaded (command: **doedit** or click ).
3. In the do-editor, go to File | Open and find your IPUMS do-file. (DON'T just double click on the do-file because that will make STATA try to run it and you don't want to do that yet.) You will need to edit the file slightly and then save it.
    a. This is do-file that came with the example extract. I have indicated the changes (in **bold**) that you will need to make to it. I have also annotated the file so that you know what the commands are doing. See comments below with *** and all in CAPS.
4. Once you have edited the do-file, save it (File | Save in the do-editor). Then run the do-file by clicking the do  button or use the command, **do dofilename.do**. A STATA data set will be magically created. In this example, it will have 14 variables and almost 3 million observations.

NOTE: If you are using data from a different source but are able to download it into STATA (such as DHS data), you will have similar files so you should understand what is going on in the commands below.

MODIFIED IPUMS DO-FILE:

/* Important: you need to put the .dat and .do files in one folder/directory and then set the working directory to that folder (command: `cd` "pathname"). */

```
set more off
clear

/*YOU SHOULD CAPTURE WHAT THE PROGRAM DID IN A LOG FILE. THE FIRST
COMMAND MAKES SURE THAT YOU DON'T ALREADY HAVE A LOG FILE OPEN
AND THE 2ND COMMAND STARTS A NEW LOG.*/
capture log close
log using make_census.log, text replace

/*THIS NEXT SECTION TELLS STATA HOW TO READ IN THE DATA
FOR EXAMPLE, IT TELLS STATA THAT THE NUMBERS IN THE FIRST TWO COLUMNS OF
THE ".dat" FILE INDICATE THE YEAR OF THE CENSUS
THE INFORMATION IN COLUMNS 7-8 INDICATE THE STATE FIPS CODE, ETC.*/

infix ///
 int    year                1-4 ///
 byte   datanum               5-6 ///
 double serial                7-14 ///
 int    hhwt                15-18 ///
 byte   statefip             19-20 ///
 byte   gq                   21 ///
 int    pernum               22-25 ///
 int    perwt                26-29 ///
 int    age                  30-32 ///
 byte   sex                  33 ///
 byte   marst                34 ///
 long   bpld                 35-39 ///
 byte   marst_mom            40 ///
 byte   marst_pop            41 ///
```

```
using usa_00001.dat

/*THESE LABEL COMMANDS ARE FOR YOUR CONVENIENCE THEY WILL MAKE IT EASIER
TO SEE WHAT THE DATA ACTUALLY MEAN */
label var year `"Census year"'
label var datanum `"Data set number"'
label var serial `"Household serial number"'
label var hhwt `"Household weight"'
label var statefip `"State (FIPS code)"'
label var gq `"Group quarters status"'
label var pernum `"Person number in sample unit"'
label var perwt `"Person weight"'
label var age `"Age"'
label var sex `"Sex"'
label var marst `"Marital status"'
label var bpld `"Birthplace [detailed version]"'
label var marst_mom `"Marital status [of mother]"'
label var marst_pop `"Marital status [of father]"'

***IF YOU HAVE MORE THAN 1 YEAR OF DATA THIS NEXT BIT IS VERY USEFUL.
label define yearlbl 1850 `"1850"'
label define yearlbl 1860 `"1860"', add
label define yearlbl 1870 `"1870"', add
label define yearlbl 1880 `"1880"', add
label define yearlbl 1900 `"1900"', add
label define yearlbl 1910 `"1910"', add
label define yearlbl 1920 `"1920"', add
label define yearlbl 1930 `"1930"', add
label define yearlbl 1940 `"1940"', add
label define yearlbl 1950 `"1950"', add
label define yearlbl 1960 `"1960"', add
label define yearlbl 1970 `"1970"', add
label define yearlbl 1980 `"1980"', add
label define yearlbl 1990 `"1990"', add
label define yearlbl 2000 `"2000"', add
label define yearlbl 2001 `"2001"', add
label define yearlbl 2002 `"2002"', add
label define yearlbl 2003 `"2003"', add
label define yearlbl 2004 `"2004"', add
label define yearlbl 2005 `"2005"', add
label define yearlbl 2006 `"2006"', add
label define yearlbl 2007 `"2007"', add
label values year yearlbl


/*THE FOLLOWING LABELS ARE VERY USEFUL. THIS SERVES AS A CODEBOOK. FOR
EXAMPLE, SUPPOSE YOU WANTED TO CREATE A DUMMY VARIABLE FOR THE STATE OF
CALIFORNIA, THIS LETS YOU KNOW THAT YOU WOULD DEFINE THAT BY GIVING THE
COMMAND: gen calif=(statefip==6) */
label define statefiplbl 01 "Alabama", add
label define statefiplbl 02 "Alaska", add
label define statefiplbl 04 "Arizona", add
label define statefiplbl 05 "Arkansas", add
```

label define statefiplbl 06 "California", add
label define statefiplbl 08 "Colorado", add
label define statefiplbl 09 "Connecticut", add

**AND IT GOES ON FOR MANY PAGES…….

***IT IS A GOOD IDEA TO HAVE A LOOK AT THE DATA TO MAKE SURE IT LOOKS OK
summarize
tab age
tab statefip

**YOU WILL WANT TO SAVE YOUR DATA AT THE END OF THE DO-FILE

save example_census_2000.dta, replace