



IPUMS Data Training Exercise: An introduction to IPUMS USA (Exercise 1 for Stata)



IPUMS
USA

Learning goals

- Understand how IPUMS USA dataset is structured
- Create and download an IPUMS data extract
- Decompress the data and read it into your statistical package

Summary

In this exercise, you will gain basic familiarity with the IPUMS USA data exploration and extract system to answer the following research questions: What proportion of the U.S. population lives on farms? Is there an association between veteran status and labor-force participation? What is the trend in carpooling over time by metropolitan area status? You will create a data extract that includes the variables FARM, EMPSTAT, VETSTAT, METRO, CARPOOL, STRATA, and CLUSTER; then you will use the sample code to analyze these data. After completing this exercise, you will have experience navigating the IPUMS USA website and should be able to leverage these data to explore your own research interests.

Register for an IPUMS Account

Go to <https://usa.ipums.org/usa/> click on Login at the top, and apply for access. On login screen, enter email address and password and submit it!

Make data extracts

- Navigate to the IPUMS USA homepage and click on "Browse Data."

Select Samples - Extract #1: Farm Population

- Go to the homepage and click SELECT DATA located at the top of the page.
- On the following webpage, click SELECT SAMPLES
- Choose the 1860, 1940, and 1960 1% samples by "check marking" the radio box to the left of each sample name.
- Once checked, click SUBMIT SAMPLE SELECTIONS

Select Variables - Extract #1: Farm Population

- Return to the SELECT DATA page. Using the variable table or search feature, find the variables:
 - FARM: Household Farm Status
- Using the search feature: Click SEARCH and input 'FARM' for the search term and click SEARCH. The default search criteria will be sufficient. The resulting page will return a list of related variables to the search terms. Once you have located FARM, click the radio button 'Add to cart' on the left side of the page. This selects FARM to be included in the data extract. The radio button should then change from a '+' to a checkmark to confirm selection (see below)



SELECT HARMONIZED VARIABLES

GROUP ALPHABETICAL SEARCH

[CHANGE SAMPLES](#)

HARMONIZED VARIABLES ⓘ
 SOURCE VARIABLES

[HELP](#)
[DISPLAY OPTIONS](#)

'farm' found 56 time(s)

Add to cart	Variable	Variable Label	Type	Codes	1960 1pct	1940 1pct	1860 1pct
<input checked="" type="checkbox"/>	FARM	Farm status	H	codes	X	X	X
<input type="checkbox"/>	MIGFARM5	Farm status 5 years ago	P	codes	.	X	.
<input type="checkbox"/>	MIGFARM1	Farm status 1 year ago	P	codes	.	.	.
<input type="checkbox"/>	FARMSCHD	Farm schedule	H	codes	.	.	.
<input type="checkbox"/>	FARMPROD	Sales of farm products	H	codes	X	.	.
<input type="checkbox"/>	FBUSINC	Business income of other family members	H	codes	.	.	.
<input type="checkbox"/>	QFARMPRO	Flag for Farmprod, Farm	H	codes	1	.	.
<input type="checkbox"/>	INCBUS	Non-farm business income	P	codes	.	.	.
<input type="checkbox"/>	INCBUS00	Business and farm income, 2000	P	codes	.	.	.
<input type="checkbox"/>	NPROFMGR	Employed professional and kindred workers, and managers and administrators -- except farm -- age 16+ (total employed persons age 16+)	H	codes	.	.	.
<input type="checkbox"/>	NAGRIC	Employed farmers, farm managers, farm laborers, and farm foremen age 16+ (total employed persons age 16+)	H	codes	.	.	.
<input type="checkbox"/>	QACREPRO	Flag for Acreprop, Farm	H	codes	1	.	.
<input type="checkbox"/>	PFARMSCH	Person received farm schedule	P	codes	.	.	.
<input type="checkbox"/>	INCFARM	Farm income	P	codes	.	.	.
<input type="checkbox"/>	QFARM	Flag for Farm	H	codes	1	X	.
<input type="checkbox"/>	INCBUSEM	Business and farm income	P	codes	X	.	.
<input type="checkbox"/>	OCCHISCO	Occupation, HISCO classification	P	codes	.	.	.
<input type="checkbox"/>	OCC1950	Occupation, 1950 basis	P	codes	X	X	X
<input type="checkbox"/>	URBRURAL	Detailed Urban/Rural Classification	H	codes	.	.	.
<input type="checkbox"/>	WRK1STWK	Worked last week	P	codes	1	.	.
Add to cart	Variable	Variable Label	Type	Codes	1960 1pct	1940 1pct	1860 1pct
<input type="checkbox"/>	INCEARN	Total personal earned income	P	codes	.	.	.

Review and submit extract #1: Farm Population

- Click on the "View Cart" button underneath your data cart.
- Review your variable and sample selection to ensure your extract is complete.
 - You may notice a number of additional variables you did not select are in your cart; IPUMS preselects a number of key technical variables, which are automatically included in your data extract.
- Add additional variables or samples if they are missing from your extract, or click the "Create Data Extract" button.
- Review the Extract Request screen that summarizes your extract; add a description of your extract (e.g., "USA Exercise 1") and click "Submit Extract".



- You will receive an email when your data extract is available to download.

Select Samples - Extract #2: Veteran and Labor Force Status

- Go to the homepage and click SELECT DATA located at the top of the page.
- On the following webpage, click SELECT SAMPLES
- Choose the 1980 (5% state) and 2000 (1%) samples by “check marking” the radio box to the left of each sample name.
- Once checked, click SUBMIT SAMPLE SELECTIONS

Select Variables - Extract #2: Veteran and Employment Status

- Return to the SELECT DATA page. Using the variable table or search feature, find the variables:
 - VETSTAT: Veteran Status
 - EMPSTAT: Employment Status
- Once you have located the variables, click the radio button `Add to cart` on the left side of the page. This selects them to be included in the data extract. The radio button should then change from a `+` to a checkmark to confirm selection.
- Review and provide a short description for the extract and click SUBMIT EXTRACT. You will receive an e-mail when the data is available for download.

Select Samples - Extract #3: Carpooling and Metropolitan Status

- Go to the homepage and click SELECT DATA located at the top of the page.
- On the following webpage, click SELECT SAMPLES
- Choose the 2010 (ACS 1-year) and 1980 (5% state) samples by “check marking” the radio box to the left of each sample name.
- Once checked, click SUBMIT SAMPLE SELECTIONS



Select Variables - Extract #3: Carpooling and Metropolitan Status

- Return to the SELECT DATA page. Using the variable table or search feature, find the variables:
 - CARPOOL: Mode of carpooling
 - METRO: Metropolitan Status
- Once you have located the variables, click the radio button `Add to cart` on the left side of the page. This selects them to be included in the data extract. The radio button should then change from a `+` to a checkmark to confirm selection.
- Review and provide a short description for the extract and click SUBMIT EXTRACT. You will receive an e-mail when the data is available for download.

Getting the data into your statistics software

The IPUMS USA extract builder provides raw ASCII data files and the command files necessary for reading the raw data into a stats package. Note that these instructions are for Stata. If you would like instructions for a different stats package, see <https://www.ipums.org/exercises.shtml>.

Download the data

- Follow the link in the email notifying you that your extract is ready, or by clicking on the "Download and Revise Extracts" link on the left-hand side of the IPUMS USA homepage.
- Right-click on the data link next to the extract you created.
- Choose "Save Target As..." or "Save Link As..."
- Save into your preferred working directory. This tutorial assumes you will save the file into "Documents" (which should pop up as the default location).
- Do the same thing to save the Stata command file.



Decompress the data

- All IPUMS extracts are compressed. There are many applications available for decompressing files. We recommend [7zip](#) for Windows users. Macs can open these types of files without additional software.
- Find the "Documents" folder under the Start menu (or a different preferred working directory).
- Double click on the ".dat" file.
- In the window that pops up, press the "Extract" button.
- After the extract has completed, confirm that the Documents (or preferred working directory) folder contains a .dat and .do file that begin with "usa_###".

Read in the data

- Open Stata from the Start menu.
 - In the "File" menu, choose "Change working directory..."
 - Select "Documents" (or preferred working directory), click "OK".
 - In the "File" menu, choose "Do..."
 - Select the *.do file.
- You will see "end of do-file" when Stata has finished reading in the data.

Stata Code to Review

Code	Purpose
<u>generate</u>	Creates a new variable, "replace" specifies a value according to cases
mean	Displays a simple tabulation and frequency of one variable
<u>tabulate</u>	Displays a cross-tabulation for up to 2 variables
<u>regress</u>	OLS regression



Common Mistakes to Avoid

- Not fully decompressing the data
- Giving the wrong filepath to indicate the dataset
- Mixing up = and == ; To assign a value in generating a variable, use "=". Use "==" to specify a case when a variable is a desired value using an if statement.
- Forgetting to put [weight=weightvar] into square brackets

A note on IPUMS USA and sample weighting

Many of the data samples provided by IPUMS USA are based on statistical survey techniques to obtain a nationally representative sample of the population. This means that persons with some characteristics are over-represented in the samples, while others are underrepresented.

To obtain representative statistics, users should always apply IPUMS USA sample weights for the population of interest (persons/households). IPUMS USA provides both person (PERWT) and household—level (HHWT) sampling weights to assist users with applying a consistent sampling weight procedure across data samples. While appropriate use of sampling weights will produce correct point estimates (e.g., means, proportions), it is also necessary to use additional statistical techniques that account for the complex sample design to produce correct standard errors and statistical tests.

IPUMS USA has provided the variables STRATA and CLUSTER for this purpose. While unnecessary for the following analytic exercises focused on mean and proportional estimates, a further discussion can be found on the IPUMS USA website: ANALYSIS AND VARIANCE ESTIMATION WITH IPUMS USA

https://usa.ipums.org/usa/complex_survey_vars/userNotes_variance.shtml



Analyze the Data

Part 1: Frequencies

Get a basic frequency of the FARM variable for selected historical years.

1. On the website, find the codes page for the FARM variable and write down each code value and its associated category label.

2. How many people lived on farms in the US in 1860? 1960?

```
tabstat perwt if year==1860 , s(sum) by(farm) f(%11.0fc)
tabstat perwt if year==1960 , s(sum) by(farm) f(%11.0fc)
```

3. What proportion of the population lived on a farm in 1860? 1960?

```
proportion farm [pw=perwt] , over(year)
```

Using household weights (HHWT)

Suppose you were interested not in the number of people living farms, but in the number of households that were farms. To get this statistic you would need to use the household weight. In order to use household weight, you should be careful to select only one person from each household to represent that household's characteristics (use `PERNUM = 1` as the subset). You will need to apply the household weight (HHWT).



4. What proportion of households in the sample lived on farms in 1940? (*Hint: don't use the weight quite yet*)

5. How many households were farms in 1940?

```
tab farm if year==1940 & pernum==1
```

6. What proportion of households were farms in 1940? (*use the weight now*)

```
tabstat hhwt if year==1940 & pernum==1, s(sum) by(farm)
f(%11.0fc)
```

7. Does the sample over or under-represent farm households?

```
proportion farm if year==1940 & pernum==1 [pw=hhwt]
```

Part 2: Frequencies

This portion of the exercise uses Extract #2: Veteran and Employment Status.

8. What is the universe for EMPSTAT for this sample, and what are the codes for this variable?



9. Using the variable description for VETSTAT, describe the issue a researcher would face if they wanted to research women serving in the armed forces from World War II until the present.

10. What percent of veterans and non-veterans were:

a. Employed in 1980?

b. Not part of the labor force in 1980?

```
proportion empstat if year==1980 & empstat!=0 [pw=perwt] ,  
over(vetstat)
```

11. What percent of veterans and non-veterans were:

a. Employed in 2000?

b. Not part of the labor force in 2000?

```
proportion empstat if year==2000 & empstat!=0 [pw=perwt] ,  
over(vetstat)
```

12. What could explain the difference in relative labor force participation in veterans versus non-veterans between 1980 and 2000?



13. How do relative employment rates change when non-labor force participants are excluded in 2000?

```
proportion empstat if year==2000 & !inlist(empstat,0,3)
[pw=perwt] , over(vetstat)
```

Part 3: Advanced Exercises

This portion of the exercise uses Extract #3: Carpooling and Metropolitan Status.

14. What are the codes for METRO and CARPOOL?

15. What is a limitation of CARPOOL if you are using 2010 and 1980? How could you address this limitation?



16. What are the proportion of carpoolers and lone drivers not in the metro area, in the central city, and outside the central city in 1980? First, we'll need to define a new variable from CARPOOL. Let's name it "car". If car is 0, it indicates a lone driver, if 1, it's any form of carpooling. If 2, driving to work is not applicable.

METRO	% Drive Alone	% Carpoolers
Not in Metro Area		
Central City		
Outside Central City		

```

recode carpool (1=0) (2/5=1) (0=2) , gen(car)
lab def car_lbl 0 alone 1 carpool 2 niu
lab val car car_lbl
proportion car if year==1980 [pw=perwt], over(metro)

```

17. Does this make sense?

18. Do the same for 2010. What does this indicate for the trend in carpooling/driving alone over time in the U.S.?

```

proportion car if year==2010 [pw=perwt], over(metro)

```



Answers

Part 1: Frequencies

1. On the website, find the codes page for the FARM variable and write down the code value, and what category each code represents. 0 NIU; 1 Non - Farm; 2 Farm
2. How many people lived on farms in the US in 1860? 12,931,661 people in 1860;
15,882,991 people in 1960
3. What proportion of the population lived on a farm in 1860? 1960? 47.29% of people
in 1860; 8.86% of people in 1960

Using household weights (HHWT)

4. What proportion of households in the sample lived on farms in 1940? 18.61% of
households
5. How many households were farms in 1940? 7,075,894 households
6. What proportion of households were farms in 1940? 18.32% of households,
7. Does the sample over or under-represent farm households? sample over -
represents farm households

Part 2: Frequencies

8. What is the universe for EMPSTAT for this sample, and what are the codes for this variable? Persons age 16+; 0 NIU; 1 Employed; 2 Unemployed; 3 Not in the labor
force



9. Using the variable description for VETSTAT, describe the issue a researcher would face if they had a research question regarding women serving in the armed forces from World War II until the present. Women were not counted in VETSTAT until the 1980 Census.
10. What percent of veterans and non-veterans were:
- Employed in 1980? Non - veterans 54.32%, Veterans 76.06%
 - Not part of the labor force in 1980? Non - veterans 41.70%, Veterans 20.09%
11. What percent of veterans and non-veterans were:
- Employed in 2000? Non - veterans 61.82%, Veterans 54.50%
 - Not part of the labor force in 2000? Non - veterans 34.43%, Veteran 43.11%
12. What could explain the difference in relative labor force participation in veterans versus non-veterans between 1980 and 2000? Either a growing number of aging veterans or an uptick in PTSD diagnoses in veterans.
13. How do relative employment rates change when non-labor force participants are excluded in 2000? Veterans have a higher employment rate than non - veterans. (95.79% vs 94.28% employment).

Part 3: Advanced Exercises

14. What are the codes for METRO and CARPOOL? CARPOOL : 0 N/A; 1 Drives alone; 2 Carpool; 3 Shares driving; 4 Drives others only; 5 Passenger only; METRO: 0 Not identifiable; 1 Not in metro area; 2 Central city; 3 Outside central city; 4 Central city



status unknown

15. What is a limitation of CARPOOL if you are using 2010 and 1980? How could you address this limitation? The code 2 for CARPOOL was taken for the 2010 sample, but 3, 4, and 5 are taken for the 1980 sample. They have different levels of detail for carpooling. A new variable could be defined to combine these codes. Collapsing three 1980 categories (3-5) into one (2) may fix this limitation.

16.

METRO	% Drive Alone	% Carpoolers
Not in Metro Area	<u>24.64</u>	<u>8.52</u>
Central City	<u>22.68</u>	<u>7.05</u>
Outside Central City	<u>31.30</u>	<u>8.70</u>

17. Does this make sense? Yes, commuters outside the metro area or central city are more likely to drive than those in the central city, for whom carpooling is not applicable because they could use public transportation. Commuters outside the central city might be more likely to carpool than those outside the metro area because they are likely to work within the central city and may live close to others who work in the same concentrated urban area.

18. Do the same for 2010. What does this indicate for the trend in carpooling/driving alone over time in the US? In 2010, a greater proportion of the population drove alone and a smaller proportion carpooled.

