IPUMS Data Training Exercise:

An introduction to IPUMS USA

(Exercise 1 for SPSS)

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:

  o 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)

IPUMS USA

DATA CART
YOUR DATA EXTRACT
1 VARIABLE
3 SAMPLES

VIEW CART

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
☑	FARM	Farm status	H	codes	X	X	X
⊕	MIGFARM5	Farm status 5 years ago	P	codes	.	X	.
ⓘ	MIGFARM1	Farm status 1 year ago	P	codes	.	.	.
ⓘ	FARMSCHD	Farm schedule	H	codes	.	.	.
⊕	FARMPROD	Sales of farm products	H	codes	X	.	.
ⓘ	FBUSINC	Business income of other family members	H	codes	.	.	.
ⓘ	QFARMPRO	Flag for Farmprod, Farm	H	codes	ⓘ	.	.
ⓘ	INCBUS	Non-farm business income	P	codes	.	.	.
ⓘ	INCBUS00	Business and farm income, 2000	P	codes	.	.	.
ⓘ	NPROFMGR	Employed professional and kindred workers, and managers and administrators -- except farm -- age 16+ (total employed persons age 16+)	H	codes	.	.	.
ⓘ	NAGRIC	Employed farmers, farm managers, farm laborers, and farm foremen age 16+ (total employed persons age 16+)	H	codes	.	.	.
ⓘ	QACREPRO	Flag for Acreprop, Farm	H	codes	ⓘ	.	.
ⓘ	PFARMSCH	Person received farm schedule	P	codes	.	.	.
ⓘ	INCFARM	Farm income	P	codes	.	.	.
⊕	QFARM	Flag for Farm	H	codes	ⓘ	X	.
⊕	INCBUSFM	Business and farm income	P	codes	X	.	.
ⓘ	OCCHISCO	Occupation, HISCO classification	P	codes	.	.	.
⊕	OCC1950	Occupation, 1950 basis	P	codes	X	X	X
ⓘ	URBRURAL	Detailed Urban/Rural Classification	H	codes	.	.	.
ⓘ	WRKLSTWK	Worked last week	P	codes	ⓘ	.	.

Add to cart	Variable	Variable Label	Type	Codes	1960 1pct	1940 1pct	1860 1pct
ⓘ	INCEARN	Total personal earned income	P	codes	.	.	.

Review and submit extract #1

- 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 SPSS. 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 SPSS 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.
- 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 folder contains three files that begin with "usa_###".

Read in the data

- Double click on the ".sps" file, which should automatically have been named "usa_000...".

- The first two lines should read:

```
cd ".".
data list file = "usa_000..."/
```

- Change the first line to read: cd (location where you've been saving your files). For example: `cd "C://Documents".`

- Change the second line to read:

```
data list file = "C: /Documents//usa_000...dat"/
```

- Under the "Run" menu, select "All"; an output viewer window will open. Use the Syntax Editor for the SPSS code below, highlight the code, and choose "Selection" under the Run menu.

SPSS Code to Review

Code	Purpose
compute	Creates a new variable
freq	Displays a simple tabulation and frequency of one variable
crosstabs	Displays a cross-tabulation for up to 2 variables and a control
~=	Not equal to

Common Mistakes to Avoid

- Not fully decompressing the data

- Giving the wrong filepath to indicate the dataset

- Excluding cases you don't mean to. Avoid this by turning off weights and select cases after use, otherwise they will apply to all subsequent analyses

- Terminating commands prematurely or forgetting to end commands with a period (.) Avoid this by carefully noting the use of periods in this exercise

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?

    _____

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

    _____

```
weight by perwt.

crosstabs

    /tables = farm by year

    /cells=count row.

execute.
```

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. You will need to apply the household weight (HHWT) and filter so that PERNUM = 1.

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?

_____

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

_____

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

_____

_____

_____

```
weight by hhwt.

compute filter_$=(pernum = 1).

variable labels filter_$ 'pernum = 1 (filter)'.

value labels filter_$ (f1.0).

filter by filter_$

crosstabs

    /tables=farm by year

    /cells=count column.

execute.

filter off.
```

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 had a research question regarding 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?

       _____

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

    a.   Employed in 2000?

       _____

    b.   Not part of the labor force in 2000?

       _____

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

   _____

   _____

```
*Under Data/Select Cases/If, enter "VETSTAT > 0 & EMPSTAT>0".

weight by perwt.

crosstabs

  /tables= vetstat by empstat by year

  /cells = count row.

execute.
```

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

_____

_____

_____

```
*Under Data/Select Cases/If, enter "VETSTAT > 0 & EMPSTAT>0 &
EMPSTAT ~= 3&YEAR=2000".

weight by perwt.

crosstabs

    /tables= vetstat by empstat

    /cells = count row.

execute.
```

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		

```
compute car = 1.

if (carpool=0) car = 2.

if (carpool=1) car = 0.

execute.
```

```
*Under Data/Select Cases/If, enter "YEAR=1980&METRO~=0
&METRO~=4".

weight by perwt.

crosstabs

   /tables = car by metro

   /cells = count column.

execute.
```

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.?

_____

_____

_____

_____

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? 7,075,921 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:
    a. Employed in 1980? Non - veterans 54.32%, Veterans 76.06%
    b. Not part of the labor force in 1980? Non - veterans 41.70%, Veterans 20.09%

11. What percent of veterans and non-veterans were:
    a. Employed in 2000? Non - veterans 61.82%, Veterans 54.50%
    b. 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	24. 64	8.52
Central City	22.68	7.05
Outside Central City	31.30	8.70

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.