



## IPUMS Data Training Exercise:

### IPUMS International Data Extract and Analysis (Exercise 2 for SAS)



#### Learning Goals

- Understand how IPUMS International dataset is structured
- Create and download an IPUMS data extract
- Decompress the data file and read the data into a statistical package
- Analyze the demographic and population characteristics of Cambodia, Ireland, Uruguay using sample code
- Validate data analysis work using the answer key
- Understand how IPUMS data can be leveraged to explore research interests

#### Exercise Research Question and Variables

In this exercise, you will gain basic familiarity with the IPUMS International data exploration and extract system to answer the following research question: "What are the differences in water supply, internet access, car ownership, and age distribution among Cambodia, Uruguay, and Ireland?" You will create a data extract that includes the variables WATSUP, SEX, INTRNET, AUTOS, EDATTAN, AGE, and WTHH; then you will use the sample code to analyze these data.

## Register as an IPUMS International User

Go to <http://international.ipums.org>, click on User Registration and Login and Apply for access. On the login screen, enter email address and password and submit your application. Please note that IPUMS International user applications are reviewed by IPUMS staff, and a final decision may take 2-5 business days.

## SAS Code to Review

Code	Purpose
proc freq;	Begins a frequency procedure
proc means;	Begins a means procedure, returns the mean value of a variable.
tables	Required syntax to display frequencies
where	Selects only specific cases to include in a procedure

## Make a Data Extract

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

### Select samples

- Click on the "Select Samples" button to choose the census samples to include in your extract.
- Check the boxes for the 2008 sample for Cambodia, 2006 sample for Ireland, and 2006 for Uruguay
- Submit your sample selections by clicking the Submit sample selections box.
- Note that by selecting samples first, you will now only see variables available for either Cambodia, Ireland, or Uruguay.
  - If you would prefer to see all variables, regardless of their availability in your selected samples, click on "Display Options" from the main variable browsing page, and choose to display variables that are not available in your selected samples.

### Select variables

- The variable drop-down menus allow you to explore variables by topic. For example, you might find variables about occupational participation under the "Work" group.



- The search tool allows you to search for variables. Observe the options for limiting your search results by variable characteristics or variable type.
- Add a variable to your cart by clicking on the plus sign in the "Add to Cart" column of the topical variable list, or list of search results.
- View more information about the variable by clicking on the variable name, and navigating through the tabs that include a description of the variable, codes and value labels, the universe of persons asked the question, and information on the comparability of the variable among other pieces of information. If you are reviewing variable-specific information, you may click on the "Add to Cart" button near the top of the screen to add this variable to your data cart.
- Use either the drop down menu or the search feature to select the following variables, and add them to your data cart.
  - WATSUP: Water supply
  - SEX: sex
  - INTRNET: Internet access
  - AUTOS: Automobiles available
  - EDATTAN: Educational attainment
  - AGE: Age
  - HHWT: Household weight technical variable

### **Review data and request the extract**

- Click on the "View Cart" button underneath your data cart.
- Review your variable and sample selection to ensure your extract will be 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., "Differences in water supply, internet access, car ownership, and age distribution in Cambodia (2008), Ireland (2006), and Mexico (2006)") and click "Submit Extract".
- You will receive an email when your data extract is available to download.



## Getting the Data Into Your Statistics Software

The IPUMS International 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 SAS. 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 International 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 SAS command file (link located next to the extract).

### 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 "ipumsi\_###".

### Read in the data

- Open the "ipumsi\_###.sas" file.
- In the syntax editing window, change the first line from "libname IPUMS '." to "libname IPUMS //Documents...;" using the file directory where you saved your data files.
- After "filename ASCIIIDAT", enter the full file location, ending with ipumsi\_###.dat;
- Choose "Submit" under the Run file menu.



# Analyze the Data

## Part 1: Variable documentation

For each variable below, search through the tabbed sections of the variable description to answer each question.

1. Find the codes page for the SAMPLE variable and write down the code values for:
  - a. Cambodia 2008? \_\_\_\_\_
  - b. Ireland 2006? \_\_\_\_\_
  - c. Uruguay 2006 \_\_\_\_\_
2. Are there any differences in the universe of WATSUP among the three samples?  
\_\_\_\_\_
3. What is the universe for EMPSTAT in:
  - a. Cambodia 2008? \_\_\_\_\_
  - b. Ireland 2006? \_\_\_\_\_
  - c. Uruguay 2006? \_\_\_\_\_

## Part 2: Frequencies

4. How many individuals are in each of the sample extracts:
  - a. Cambodia 2008? \_\_\_\_\_
  - b. Ireland 2006? \_\_\_\_\_
  - c. Uruguay 2006? \_\_\_\_\_

```
proc freq;  
    tables sample;  
run;
```

## Part 3: Weighted frequencies

To get a more accurate estimate estimation of demographic patterns within a country from the sample, you will have to use the person weight.



5. Using weights, what is the total population of each country?

- a. Cambodia 2008 \_\_\_\_\_
- b. Ireland 2006 \_\_\_\_\_
- c. Uruguay 2006 \_\_\_\_\_

```
proc freq;  
    tables sample;  
    weight perwt;  
run;
```

6. Using weights, what proportion of individuals in each country did not have access to piped water?

- a. Cambodia 2008 \_\_\_\_\_
- b. Ireland 2006 \_\_\_\_\_
- c. Uruguay 2006 \_\_\_\_\_

```
proc freq;  
    tables watsup*urban  
    weight perwt;  
run;
```

### When to use household weights (WTHH)

Suppose you were interested not in the number of people living in urban areas, but in the number of households. 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 (WTHH). To identify only one person from each household, use the "where" statement to select only cases where the PERNUM equals 1.

### Part 4: Trends

7. What proportion of households in each country did not have access to piped water?

- a. Cambodia 2008 \_\_\_\_\_



- b. Ireland 2006 \_\_\_\_\_
- c. Uruguay 2006 \_\_\_\_\_

```
proc freq;
  where pernum = 1;
  tables watsup*sample;
  weight wthh;
run;
```

8. In which country do individuals have the most access to the internet? \_\_\_\_\_

```
proc freq;
  tables intrnet*sample;
  weight perwt;
run;
```

9. In that country, what proportion of households have both access to internet and at least one car? \_\_\_\_\_

Note: First you will have to generate a dummy variable that is “1” when the household has at least one car and internet, and “zero” in all other cases.

```
data ipums.ipumsi_000##;
  set ipums.ipumsi_0000##;
  autoint = _null_;
  if intrnet = 2 and autos >+1 and autos < 8 then autoint = 1
  else autoint = 0
run;
proc freq;
```



```
where sample = 3728 and pernum = 1  
tables autoint;  
weight wthh  
run;
```

10. In which country is educational attainment (Secondary and University in particular) between men and women most equal? Least equal?
- Most equal completion rates: \_\_\_\_\_
  - Least equal completion rates: \_\_\_\_\_

```
proc freq;  
tables edattan*sex;  
by sample;  
weight perwt;  
run;
```

## Part 5: Graphical Analysis

11. Approximately what percent of Uruguay's population is around 50 years old?  
\_\_\_\_\_
12. Compare the age distributions of Cambodia and Ireland. Is this a pattern that could be observed in other developed and developing nations? \_\_\_\_\_
13. Can the shape of the histogram of Ireland compared to the other countries indicate anything about the differences in data collection? \_\_\_\_\_

```
proc sgplot data = ipums.ipumsi_000##;  
histogram age;  
by sample;  
run;
```



*Note: SAS graph procedures do not allow for WEIGHT options, so graph analyses are at the sample level.*

14. What (approximately) are the median ages for men and women in each of these countries?

Women:

Cambodia 2008 \_\_\_\_\_ Ireland 2006 \_\_\_\_\_ Uruguay 2006 \_\_\_\_\_

Men:

Cambodia 2008 \_\_\_\_\_ Ireland 2006 \_\_\_\_\_ Uruguay 2006 \_\_\_\_\_

```
proc tabulate;  
  class sample sex;  
  var age;  
  table sample, sex*age*median;  
run;
```



# Answers

## Part 1: Variable Documentation

1. Find the codes page for the SAMPLE variable and write down the code values for:
  - a. Cambodia 2008: 1,162
  - b. Ireland 2006: 3,728
  - c. Uruguay 2006: 8,585
2. Are there any differences in the universe of WATSUP among the three samples?  
Cambodia 2008: Regular households, Ireland 2006: Private households in non-temporary dwellings, Uruguay 2006: All households. All have technical differences, Uruguay being most inclusive, and Ireland being the most precise.
3. What is the universe for EMPSTAT in:
  - a. Cambodia 2008: All persons.
  - b. Ireland 2006: Non-absent persons age 15+
  - c. Uruguay 2006: Persons age 14+

## Part 2: Frequencies

4. How many individuals are in each of the sample extracts?
  - a. Cambodia 2008: 1,340,121
  - b. Ireland 2006: 400,314
  - c. Uruguay 2006: 256,866

## Part 3: Weighted Frequencies

5. Using weights, what is the total population of each country?
  - a. Cambodia 2008: 13,401,213
  - b. Ireland 2006: 4,403,140
  - c. Uruguay 2006: 3,065,604
6. Using weights, what proportion of individuals in each country did not have access to piped water?
  - a. Cambodia 2008: 84.12%
  - b. Ireland 2006: 14.25%
  - c. Uruguay 2006: 3.22%

## Part 4: Trends



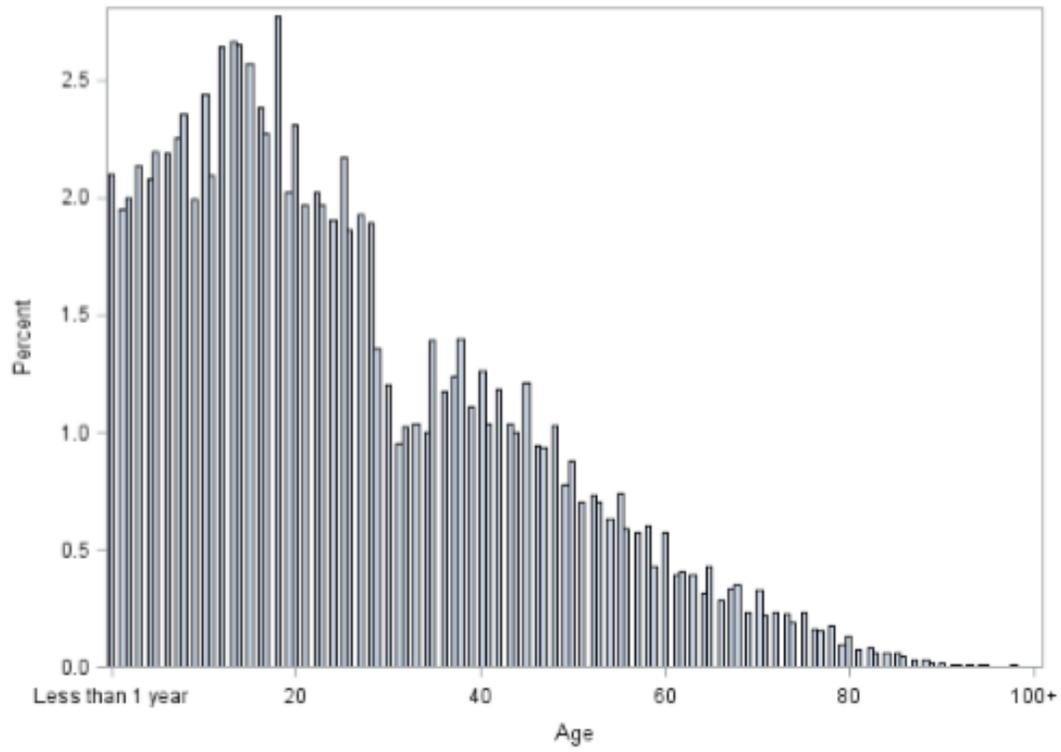
7. What proportion of households in each country did not have access to piped water?
  - a. Cambodia 2008: 83.91%
  - b. Ireland 2006: 12.59%
  - c. Uruguay 2006: 3.28%
8. In which country do individuals have the most access to the internet?  
Ireland 2006 (53.1% Yes)
9. In that country, what proportion of households have both access to internet and at least one car? 40.7%
10. In which country is educational attainment (Secondary and University in Particular) between men and women most equal? Least equal?
  - a. Most equal completion rates: Uruguay (18.68/19.76%; 3.99/4.23%)
  - b. Least equal completion rates: Cambodia (4.76/2.44%; 1.36/0.6%)

## Part 5: Graphical Analysis

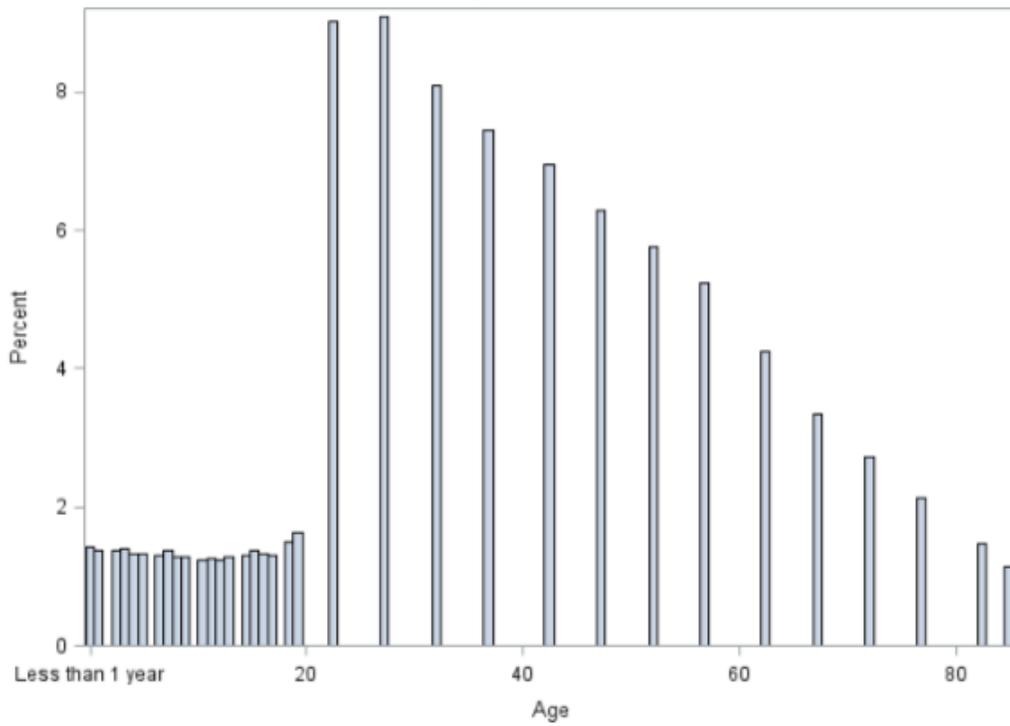
11. Approximately what percent of Uruguay's population is around 50 years old?  
~2.4%
12. Compare the age distributions of Cambodia and Ireland. Is this a pattern that could be observed in other developed and developing nations? A large proportion of Cambodia's population is 25 or younger, while the mean age of Ireland's population seems a bit older.
13. Can the shape of the histogram of Ireland compared to the other countries indicate anything about the differences in data collection? "All Ireland samples provide single years of age through 19 and 5-year age intervals thereafter, top-coded at 85+" From the Comparability Tab on the website.

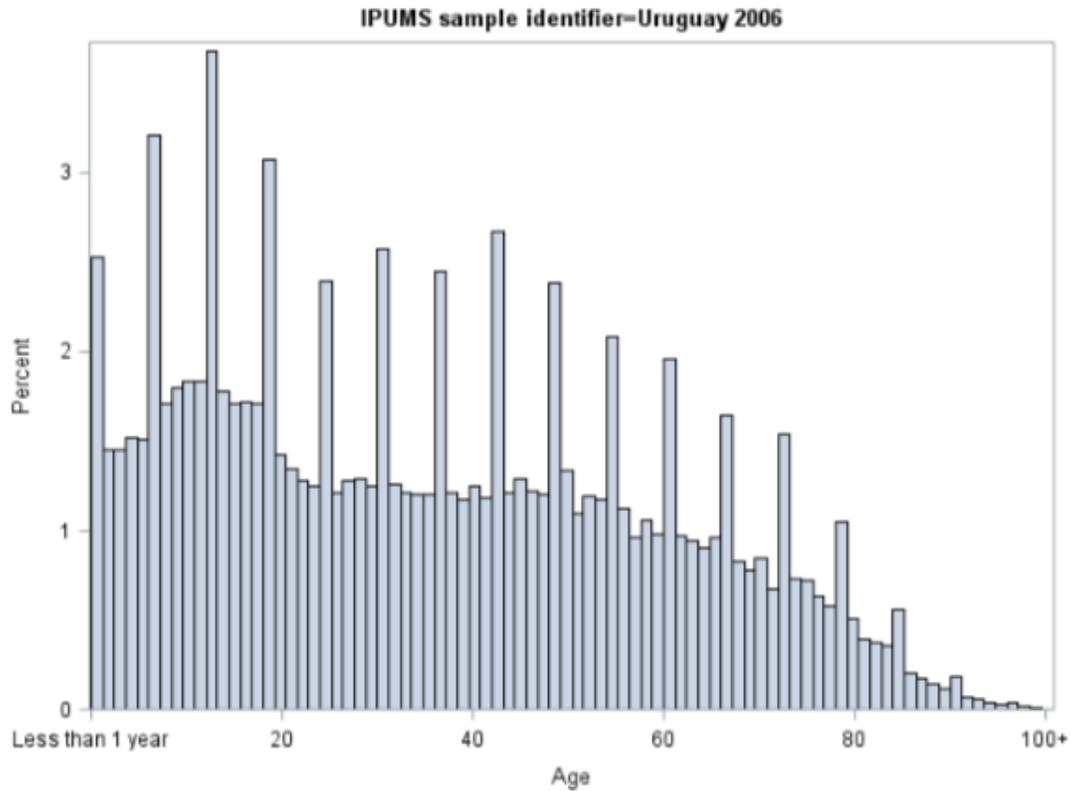
*Note: SAS graph procedures do not allow for WEIGHT options, so graph analyses are at the sample level.*

IPUMS sample identifier=Cambodia 2008



IPUMS sample identifier=Ireland 2006





14. What (approximately) are the median ages for men and women in each of these countries?

- a. Women:  
Cambodia 2008 23, Ireland 2006: 32, Uruguay 2006: 35
- b. Men:  
Cambodia 2006: 20; Ireland 2006: 32; Uruguay 2006: 32