IPUMS Data Training Exercise:

An Introduction to IPUMS International (Exercise 1 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 Mexico and Uganda 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 urbanization, literacy, and occupational participation in Uganda and Mexico?" You will create a data extract that includes the variables URBAN, SEX, EMPSTAT, OCCISCO, FLOOR, LIT, and AGE; 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 2000 sample for Mexico and 2002 for Uganda
- 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 Mexico or Uganda.
    - 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 the drop down menu or search feature to add these variables to your data cart:
    - URBAN: household location
    - SEX: sex
    - EMPSTAT: Employment status
    - OCCISCO: Employment category
    - FLOOR: Flooring material
    - LIT: Literacy
    - AGE: Age

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 urbanization, literacy, and occupational participation in Uganda (2002) and Mexico (2000)" 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 ASCIIDAT", 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. Under the "Household" dropdown menu, find the "Geography" subcategory and click on the variable URBAN. What constitutes an urban area in each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____
2. What are the codes for URBAN?

    _____
3. Find the variable EMPSTAT (employment status). Is the reference period of work the same for Mexico and Uganda? _____
4. What is the universe for EMPSTAT:
    a. in Mexico in 2000? _____
    b. in Uganda in 2002? _____

Part 2: Frequencies

5. Find the codes page for the SAMPLE variable. What are the code values for each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____
6. How many individuals are in the Mexico 2000 sample extract? _____
7. How many individuals are in the Uganda 2002 sample extract? _____

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

8. How many individuals in the sample live in urban areas in each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____

9. What proportion of individuals in the sample lived in urban areas in each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____

```
proc freq;
     tables sample*urban;
run;
```

Part 3: Weighted frequencies

To get a more accurate estimate for the actual proportion of individuals living in urban areas, you will have to use the person weight.

10. Using weights, what is the total population of each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____
11. Using weights, how many individual lived in urban areas in each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____
12. Using weights, what proportion of individual lived in urban areas in each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____

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

When to use household weights (HHWT)

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 (HHWT). To identify only one person from each household, use the "where" statement to select only cases where the PERNUM equals 1.

Part 4: Trends

13. Using weights, which occupational category has the highest percentage of workers:
    a. in Mexico 2000? _____
    b. in Uganda 2002? _____

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

14. Which occupation category has the highest percentage of female workers:
    a. in Mexico in 2000? _____
    b. in Uganda in 2002? _____

```
proc freq;
     where sex = 2;
     tables occisco*sample;
     weight perwt;
run;
```

Compare the variables

In order to do your analysis, you must decide whether you are analyzing the total population or the people participating in the labor force. The previous commands yielded totals and percentages of people within an occupation among all people in the population. If you want to know how women's work is distributed among women in the labor force, you have to limit your analysis to people who are employed. To find out who is working, look at employment status category 1, "employed."

15. What is the labor force participation distribution by gender in each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____

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

16. What percentage of women within the labor force is working:
    a. in agriculture in Mexico 2000? _____
    b. in agriculture in Uganda 2002? _____
    c. in service in Mexico 2000? _____
    d. in service in Uganda 2002? _____

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

Part 5: Graphical Analysis

17. What percent of the population is literate in each country?
    a. Mexico 2000 _____
    b. Uganda 2002 _____
18. How are universe differences seen on the graph? _____
    _____

```
proc gchart;
     vbar lit / discrete type = percent;
     where cntry = 484;
run;
proc gchart;
     vbar lit / discrete type = percent;
     where cntry = 800;
run;
```

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

Next, recode literacy to explore literacy rates by age.

```
data ipums.ipumsi_###;
     set ipums.ipumsi_###;
     literate = _null_;
     if lit = 1 then literate = 0;
     if lit = 2 then literate = 1;
run;
proc freq;
     tables lit*literature;
run;
proc sgplot=ipums.ipumsi_###;
     vline age/
     response=literate
     stat=mean
     markers;
     by sample;
xaxis value = (0 to 100 by 5) integer;
     xaxis fitpolicy = staggerthin;
run;
```

19. Which country has higher overall literacy? _____
20. At (approximately) what ages are literacy rates highest in each country?
    a. Mexico 2000? _____
    b. Uganda 2002? _____
21. How are universe differences seen on the graph? _____
    _____
22. In which country are literacy rates nearly equal for men and women? _____

```
proc sgpanel data=ipums.ipumsi_###;
     panelby sample;
     vbar sex /
     response=literate
     stat=mean;
run;
```

23. What type of floor material is most common in Uganda in 2002? _____

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

Answers

Part 1: Variable Documentation

1. Under the "Household" dropdown menu, find the "Geography" subcategory and click on the variable URBAN. What constitutes an urban area in each country?
    a. Mexico 2000: 2,500+ people
    b. Uganda 2002: 2,000+ people
2. What are the codes for URBAN? 1 Rural 2 Urban
3. Find the variable EMPSTAT (employment status). Is the reference period of work the same for Mexico and Uganda? Both samples use a reference week
4. What is the universe for EMPSTAT:
    a. in Mexico 2000: Persons age 12+
    b. in Uganda 2002: Persons age 5+

Part 2: Frequencies

5. Find the codes page for the SAMPLE variable. What are the code values for in each country?
    a. Mexico 2000: 4845
    b. Uganda 2002: 8002
6. How many individuals are in the Mexico 2000 sample extract? 10,099,182 persons
7. How many individuals are in the Uganda 2002 sample extract? 2,497,449 persons
8. How many individuals in the sample live in urban areas in each country?
    a. Mexico 2000: 5,976,764
    b. Uganda 2002: 306,054
9. What proportion of individuals in the sample lived in urban areas in each country?
    a. Mexico 2000: 59.2%
    b. Uganda 2002: 12.3%

Part 3: Weighted Frequencies

10. Using weights, what is the total population of in each country?
    a. Mexico 2000: 97,014,867
    b. Uganda 2002: 24,974,490

11. Using weights, how many individual lived in urban areas in each country?
    a. Mexico 2000: 72,409,464
    b. Uganda 2002: 3,060,540
12. Using weights, what proportion of individual lived in urban areas in each country?
    a. Mexico 2000: 74.6%
    b. Uganda 2002: 12.3%

    Comparing frequencies and proportions, you can see that unweighted sample data from Mexico grossly misrepresent the population. The Mexico data was designed specifically to oversample rural areas. Weighting corrects the proportional representation of individuals or households.

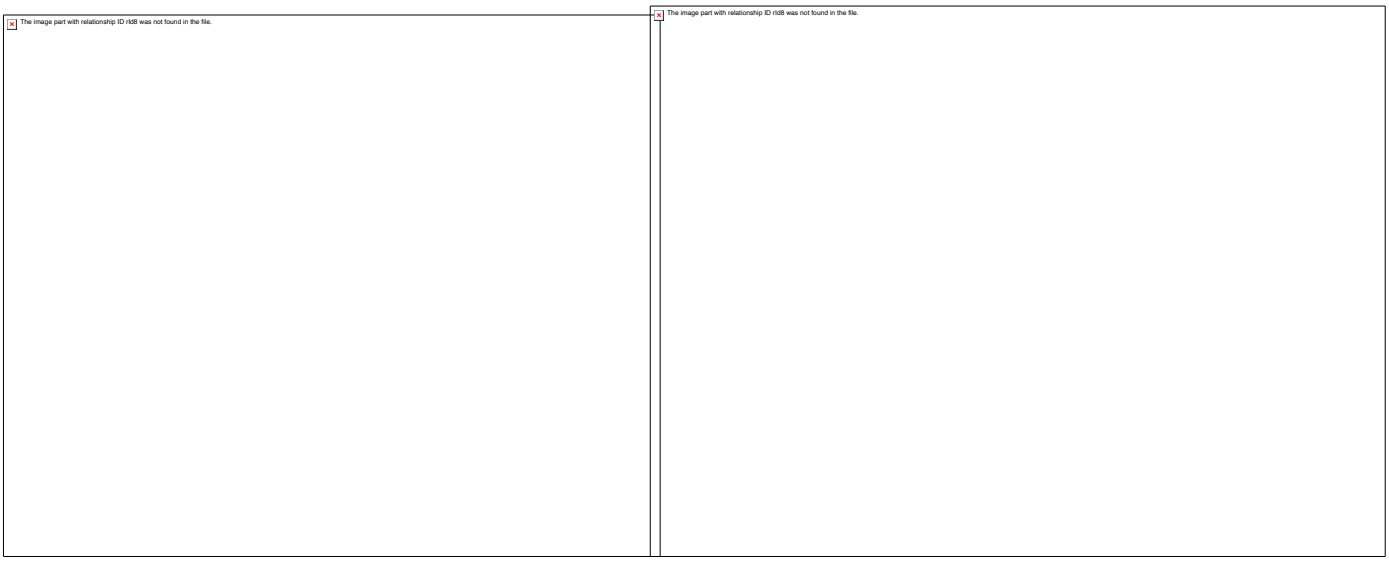
Part 4: Trends

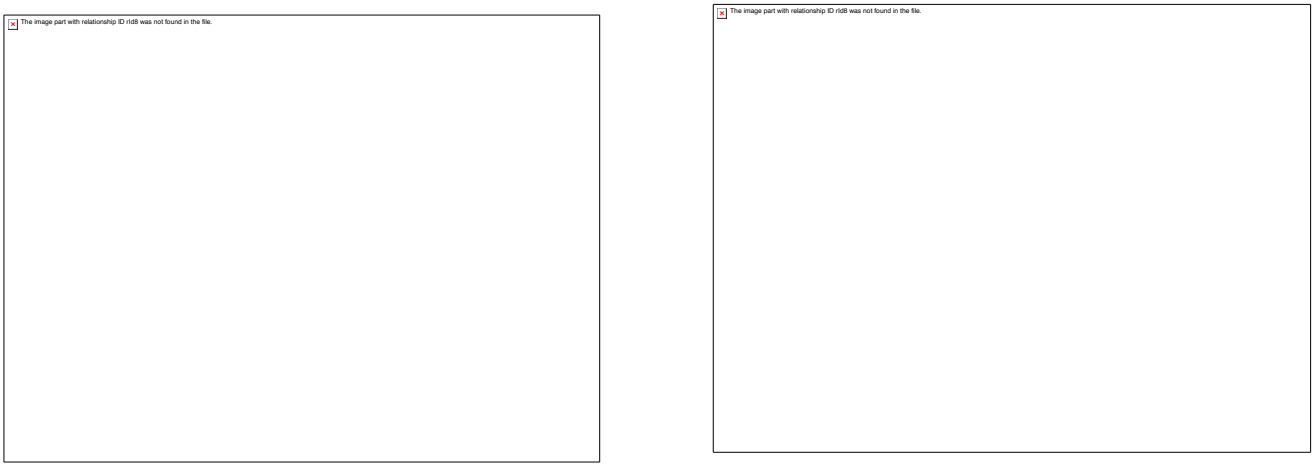
13. Using weights, which occupational category has the highest percentage of workers in:
    a. Mexico 2000: 6.5% Crafts and Related Trades
    b. Uganda 2002: 21.5% of people work in Agriculture
14. Which occupation category has the highest percentage of female workers in:.
    a. Mexico 2000: Service, shop, and market sales (5/5%)
    b. Uganda 2002: Agricultural work (21.2%)
15. What is the labor force participation distribution by gender in each country?
    a. Mexico 2000: M 50.3%; F: 22.9%
    b. Uganda 2002: M: 33.7%; F: 26.5%
16. What percentage of women within the labor force is working:
    a. in agriculture in Mexico 2000? 4.7%
    b. in agriculture in Uganda 2002? 79.7%
    c. in service in Mexico 2000: 23.9%
    d. in service in Uganda 2002: 9.0%

Part 5: Graphical Analysis

17. What percent of the population is literate in each country?
    a. Mexico 2000: ~84%
    b. Uganda 2002: ~68%

18. How are universe differences seen on the graph? NIU is included as a separate category; within universe % would be higher





19. Which country has higher overall literacy? Mexico

20. At (approximately) what ages are literacy rates highest in each country?
    a. Mexico 2000: ~12-16
    b. Uganda 2002: ~14-18
21. How are universe differences seen on the graph? Lines begin at different ages (5 in Mexico, 10 in Uganda). Apart from universe, Mexico records higher ages which are included with corresponding literacy rates in the graph.
22. In which country are literacy rates nearly equal for men and women? Mexico (2000)
23. What type of floor material is most common in Uganda in 2002? None (earth floor)