IPUMS Data Training Exercise:

IPUMS International
Data Extract and Analysis
(Exercise 1 for R)

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.

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 R. 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 for the DDI link next to the extract
- (Optional) Do the same thing for the R script
- You do not need to decompress the data to use it in R

Install the IPUMSR package

- Open R from the Start menu
- If you haven't already installed the ipumsr package, in the command prompt, type the following command: `install.packages ("ipumsr")`

Read in the data

- Set your working directory to where you saved the data above by adapting the following command (Rstudio users can also use the "Project" feature to set the working directory. In the menubar, select File->New Project->Existing Directory and then navigate to the folder):

  `setwd("~/") # "~/"` goes to your Document director on most computers

- Run the following command from the console, adapting it so it refers to the extract you just created (note the number may not be the same depending on how many extracts you've already made):

```
library(ipumsr)
ddi <- read_ipums_ddi("ipums_00001.xml")
data <- read_ipums_micro(ddi)
```

#Or, if you downloaded the R script, the following is equivalent:

`# source("ipumsi_00001.R")`

- This tutorial will also rely on the dplyr and ggplot2 packages, so if you want to run the same code, run the following command (but if you know other ways better, feel free to use them):

```
library(dplyr)
library(ggplot2)
```

- To stay consistent with the exercises for other statistical packages, this exercise does not spend much time on the helpers to allow for translation of the way IPUMS uses labelled values to the way base R does. You can learn more about these in the value-labels vignette in the R package. From R run command:

```
vignette("value-labels", package + "ipumsr")
```

R Code to Review

This tutorial's sample code and answers use the so-called "tidyverse" style, but R has the blessing (and curse) that there are many different ways to do almost everything. If you prefer another programming style, please feel free to use it. For your reference, these are some quick explanations for commands that this tutorial will use:

Code	Purpose
%>%	The pipe operator, which helps make code with nested function calls, is easier to read. When reading code, read as "and then". The pipe make it so that the code like ingredient %>% stir() %>% cook( ) is equivalent to cook (stir(ingredients)) (read as "take *ingredients* and then *stir* and then *cook*")
as_factor	Converts the value labels provide for IPUMS data into a factor variable for R
summarize	Summarize a datasets observations to one or more groups
group_by	Set the groups for the summarize function to group by
filter	Filter the dataset so that it only contains these variables
mutate	Add on a new variable to a dataset
ggplot	Make graphs using ggplot
weighted.mean	Get the weighted mean of the "a" variable

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 in:
    a. Mexico 2000? _____
    b. Uganda 2002? _____

Part 2: Frequencies

5. Find the codes page for the SAMPLE variable. What are the code values for:
    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? _____

```
data %>%

  group_by(SAMPLE = as_factor(lbl_clean(SAMPLE), levels =
"both")) %>%
summarize(n = n())
How many individuals in the sample live in urban areas in each
country?
```

    a. Mexico 2000 _____

b. Uganda in 2002 _____

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

```
data %>%

  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    URBAN = as_factor(URBAN)
  ) %>%
  summarize(n =n()) %>%
  mutate(pct = n / sum(n))
```

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.

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

```
data %>%
  group_by(SAMPLE = as_factor(lbl_clean(SAMPLE), levels =
"both")) %>%
  summarize(n + sum(PERWT))
```

```
data %>%
  group_by(SAMPLE = as_factor(lbl_clean(SAMPLE)),URBAN =
as_factor(URBAN)
  ) %>%
  summarize(n + sum(PERWT)) %>%
  mutate(pct = n / sum(n))
```

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

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

```
data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, desc(pct)) %>%
  top_n(3, pct)
```

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

```
data %>%
  filter(SEX == 2) %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, desc(pct)) %>%
  top_n(3, pct)
```

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."

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

```
data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    SEX = as_factor(SEX)
  ) %>%
  summarize(pct = weighted.mean(EMPSTAT == 1, PERWT))
```

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

```
agg_and_service <- c)
  "Skilled agricultural and fishery workers",
  "Service workers and shop and market sales"
  )

data %>%
  filter(EMPSTAT == 1 & SEX == 2) %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, OCCISCO) %>%
  filter(OCCISCO %in% agg_and_service)
```

Part 5: Graphical Analysis

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

    _____

```
data_summary <- data %>%
  group_by(
    SAMPLE = as_factor(SAMPLE),
    LIT = as_factor(LIT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))

ggplot(data_summary, aes(x = LIT, y = pct)) +
  geom_col() +
  facet_wrap(~as_factor(SAMPLE)) +
  theme(axis.text.x = element_text(angle = 20, hjust = 1))
```

Next, recode literacy to explore literacy rates by age.

```
data <- data %>%
  mutate(
    LIT_BIT = LIT %>%
      lbl_na_if(~.lbl %in% c)"NIU (not in universe)",
"Unknown/missing")) %>%
      {. == 2}
  )

data_summary <- data %>%
  filter(AGE < 999 & !is.na(LIT_BIN)) %>%
  group_by(SAMPLE = as_factor(SAMPLE, AGE = as.numeric(AGE)) %>%
  summarize(MEAN_LIT = weighted.mean(LIT_BIN, PERWT))

ggplot(data_summary, aes(x = AGE, y = MEAN_LIT, color = SAMPLE,
group = SAMPLE))+
  geom_line()
```

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

```
data_summary <- data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(MEAN_LIT = weighted.mean(LIT_BIN, PERWT, na.rm =
TRUE))

  ggplot(data_summar, aes(x = SAMPLE, y = MEAN_LIT, fill = SEX))
+ scale_fill_manual(values = c("#7570b3", "#e6ab02")) +
  geom_col(position = "dodge")
```

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

```
data_summary <- data %>%
  filter(as_factor(SAMPLE)  == "Uganda 2002") %>%
  group_by(FLOOR = as_factor(FLOOR)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))

data_summary
```

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 in 2000? Persons age 12+
    b. in Uganda in 2002? Persons age 5+

Part 2: Frequencies

5. Find the codes page for the SAMPLE variable. What are the code values for:
    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 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 were 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:
    a. in Mexico 2000? 6.5% Crafts and Related Trades
    b. in Uganda 2002? 21.5% of people work in Agriculture
14. Which occupation category has the highest percentage of female workers:
    a. in Mexico 2000? Service, shop, and market sales (5/5%)
    b. in 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 in 2000? 4.7%
    b. in agriculture in Uganda in 2002? 79.7%
    c. in service in Mexico in 2000? 23.9%
    d. in service in Uganda in 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)