Minnesota Population Center
Training and Development

IPUMS – Int.I Extraction and Analysis

Exercise 1

OBJECTIVE: Gain an understanding of how the IPUMS dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the IPUMS to explore demographic and population characteristics of Mexico and Uganda.

10/24/2012

IPUMS-I Training and Development

Research Questions

What are the differences in urbanization, literacy, and occupational participation between Mexico and Uganda?

Objectives

- Create and download an IPUMS data extract
- Decompress data file and read data into SPSS
- Analyze the data using sample code
- Validate data analysis work using answer key

IPUMS Variables

- URBAN: Household location
- SEX: Sex
- EMPSTAT: Employment status
- OCCISCO: Employment category
- FLOOR: Flooring material
- LIT: Literacy
- AGE: Age

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

Review Answer Key (page 12)

Common Mistakes to Avoid

1 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

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

Registering with IPUMS

Go to http://international.ipums.org, click on User Registration and Login and Apply for access. On login screen, enter email address and password and submit it!

Step 1

Make an Extract

- Go back to homepage and go to Select Data
- Click the Select Samples box and check the box for the 2000 sample for Mexico and 2002 for Uganda
- Click the Submit sample selections box
- Using the drop down menu or search feature, select the following variables:

  URBAN: Household location

  SEX: Sex

  EMPSTAT: Employment status

  OCCISCO: Employment category

  FLOOR: Flooring material

  LIT: Literacy

  AGE: Age

• • •

Step 2

Request the Data

- Click the green VIEW CART button under your data cart
- Review variable selection
- Click the green Create Data Extract button
- Review the 'Extract Request Summary' screen, describe your extract and click Submit Extract
- You will get an email when the data is available to download
- To get to page to download the data, follow the link in the email, or follow the Download and Revise Extracts link on the homepage

Getting the data into your statistics software

The following instructions are for SPSS. If you would like to use a different stats package, see: http://cps.ipums.org/cps/extract_instructions.shtml

Step 1

Download the Data

- Go to http://cps.ipums.org and click on Download or Revise Extracts
- Right-click on the data link next to extract you created
- Choose "Save Target As..." (or "Save Link As...")
- Save into "Documents" (that should pop up as the default location)
- Do the same thing for the SPSS link next to the extract

•••

Step 2

Decompress the Data

- Find the "Documents" folder under the Start menu.
- Double-click on the ".dat" file
- In the window that comes up, press the Extract button
- Double-check that the Documents folder contains three files starting "ipumsi_000…"
- Free decompression software is available at http://www.irnis.net/soft/wingzip/

•••

Step 3

Read in the Data

- Double click on the ".sps" file, which should automatically have been named "ipumsi_000….."
- The first two lines should read:

  cd ".".

  data list file = 'ipumsi_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\ipumsi_000…dat"/

- Under the "Run" menu, select "All" and 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

Analyze the Sample – Part I Variable Documentation

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

Section 1

Analyze the Variables

A) Under "Household" and subcategory "Geography", select the URBAN variable. What constitutes an urban area:

i. In Mexico in 2000? _____

ii. In Uganda in 2002? _____

B) What are the codes for URBAN?

_____

C) Find the variable EMPSTAT (employment status). Is the reference period of work the same for these two samples?

_____

D) What is the universe for EMPSTAT:

i. In Mexico 2000? _____

ii. In Uganda 2002? _____

Analyze the Sample – Part II Frequencies

Section 1

Analyze the Data

A) Website: Find the codes page for the SAMPLE variable and write down the code values for Mexico 2000 and Uganda 2002.

_____

B) How many individuals are in the Mexico 2000 sample extract?

_____

C) How many individuals are in the Uganda 2002 sample extract?

_____

```
freq sample.
```

D) How many individuals in the sample lived in urban areas?

Mexico 2000 _____ Uganda 2002 _____

E) What proportion of individuals in the sample lived in urban areas? Mexico 2000 _____ Uganda 2002 _____

```
crosstabs
/tables = urban by sample
/cells = count column.
```

Section Continues Below…

Analyze the Sample - Part II Frequencies (WTPER)

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

Section 2

Weighting the Data

F) Using weights, what is the total population of each country?

Mexico 2000 _____ Uganda 2002 _____

G) Using weights, how many individuals lived in urban areas?

Mexico 2000 _____ Uganda 2002 _____

H) Using weights, what proportion of individuals lived in urban areas? Mexico 2000 _____ Uganda 2002 _____

```
weight by wtper.
crosstabs
/tables = urban by sample
/cells = count column.
```

• • •

Section 3

Weighting Explanation

When to use the 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, under the Data menu, click "Select Cases", choose "If condition is satisfied", and click "If". In the top box type "PERNUM = 1" and select Continue and then Ok.

In addition to using the "weight by" command, you can also click the data tab, select "Weight Cases", then "Weight cases by" to choose a weight.

Analyze the Sample – Part III Trends in the Data

Section 1

Analyze the Data

A) Using weights, which occupational category has the highest percentage of workers from each country?

Mexico 2000 _____ Uganda 2002 _____

```
crosstabs
/tables = occisco by sample
/cells = count column.
```

B) Which occupational category has the highest percentage of female workers in each country?

Mexico 2000 _____ Uganda 2002 _____

```
crosstabs
/tables = occisco by sex by sample
/cells = count column.
```

Section Continues Below…

Compare the distribution of occupational activity among people in the labor force

Section 2

Compare the Variables

Note that 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."

A) What is the labor force participation distribution by gender in each country? Mexico 2000 %:_____ Uganda 2002 %:_____

```
crosstabs
/tables=empstat by sex by sample
/cells=count column.
```

Section Continues Below…

From Part 1, you found that employment questions were only asked of persons above a certain age. Filter the data to include only employed persons who are 15 or older (EMPSTAT = 1 and AGE ≥ 15).

Section 2

Compare the Variables

In Select Cases, enter "age ≥ 15 and empstat = 1" and click "continue."

B) What percentage of women *within the labor force* is working:

i. In Agriculture; Mexico 2000:_____ Uganda 2002:_____

ii. In Service; Mexico 2000:_____ Uganda 2002:_____

```
crosstabs
/tables=occisco by sex by sample
/cells=count column.
```

Analyze the Sample – Part IV Graphical Analysis

Section 1

Graph the Data

A) What percent of the population is literate in each sample?

_____

B) How are universe differences seen on the graph?

_____

```
graph
/bar(grouped)=pct by lit by sample.
```

Recode literacy to look at literacy rates across age

```
recode lit (0=sysmis) (9=sysmis) (1=0) (2=1) into literate.
variable labels  literate 'Literate binary'.
execute.

freq lit literate.

graph
/line(multiple)=mean(literate) by age by sample.
```

• • •

Section 2

Recode the Data

Analyze the Sample – Part IV Graphical Analysis, Age/Literacy

Section 3

Analyze Recoded Data

A) Which country has higher overall literacy?

_____

B) At (approximately) which ages are literacy rates highest?

Mexico 2000 _____ Uganda 2002 _____

C) How are universe differences seen on the graph?

_____

D) In which country are literacy rates nearly equal for men and women? _____

```
graph
/bar(grouped)=mean(literate) by sex by sample.
```

E) What type of floor material is most common in Uganda 2002?

_____

•••

Section 3

Weighting Explanation

```
graph
/bar(grouped)=pct by floor by sample.
```

•••

Complete! Validate Your Answers

ANSWERS: Analyze the Sample – Part I Variable Documentation

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

Section 1

Analyze the Variables

A) Under "Household" and subcategory "Geography", select the URBAN variable. What constitutes an urban area:

i. In Mexico in 2000? 2,500+ people

ii. In Uganda in 2002? 2,000+ people

B) What are the codes for URBAN? 1 Rural 2 Urban

C) Find the variable EMPSTAT (employment status). Is the reference period of work the same for these two samples? Both samples use a reference week.

D) What is the universe for EMPSTAT:

i. In Mexico 2000? Persons age 12+

ii. In Uganda 2002? Persons age 5+

ANSWERS: Analyze the Sample – Part II Frequencies

Section 1

Analyze the Data

A) Website: Find the codes page for the SAMPLE variable and write down the code values for Mexico 2000 and Uganda 2002.
Mexico 2000: 4845; Uganda 2002: 8002

B) How many individuals are in the Mexico 2000 sample extract?
10,099,182 persons

C) How many individuals are in the Uganda 2002 sample extract?
2,497,449 persons

```
freq sample.
```

D) How many individuals in the sample lived in urban areas?

Mexico 2000 5,976,764 Uganda 2002 306,054

E) What proportion of individuals in the sample lived in urban areas? Mexico 2000 59.2% Uganda 2002 12.3%

```
crosstabs
/tables = urban by sample
/cells = count column.
```

Section Continues Below…

ANSWERS: Analyze the Sample - Part II Frequencies (WTPER)

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

Section 2

Weighting the Data

F) Using weights, what is the total population of each country?

Mexico 2000 97,014,867 Uganda 2002 24,974,490

G) Using weights, how many individuals lived in urban areas?

Mexico 2000 72,409,464 Uganda 2002 3,060,540

H) Using weights, what proportion of individuals lived in urban areas? Mexico 2000 74.6% 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.

```
weight by wtper.
crosstabs
/tables = urban by sample
/cells = count column.
```

• • •

Section 3

Weighting Explanation

When to use the 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, under the Data menu, click "Select Cases", choose "If condition is satisfied", and click "If". In the top box type "PERNUM = 1" and select Continue and then Ok.

In addition to using the "weight by" command, you can also click the data tab, select "Weight Cases", then "Weight cases by" to choose a weight.

ANSWERS: Analyze the Sample – Part III Trends in the Data

Section 1

Analyze the Data

A) Using weights, which occupational category has the highest percentage of workers from each country?

Mexico 2000 6.5% Crafts and Related Trades

Uganda 2002 21.5% of people work in Agriculture

```
crosstabs
/tables = occisco by sample
/cells = count column.
```

B) Which occupational category has the highest percentage of female workers in each country?

Mexico 2000 Service, shop and market sales 5.5%

Uganda 2002 Agricultural work 21.1%

```
crosstabs
/tables = occisco by sex by sample
/cells = count column.
```

Section Continues Below…

ANSWERS: Compare the distribution of occupational activity among people in the labor force

Section 2

Compare the Variables

Note that 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."

A) What is the labor force participation distribution by gender in each country?

Mexico 2000%: 50.3% of males and 22.9% of females are employed

Uganda 2002%:33.7% of males and 26.5% of females are employed

```
crosstabs
/tables=empstat by sex by sample
/cells=count column.
```

Section Continues Below…

From Part 1, you found that employment questions were only asked of persons above a certain age. Filter the data to include only employed persons who are 15 or older (EMPSTAT = 1 and AGE ≥ 15).

Section 2

Compare the Variables

In Select Cases, enter "age ≥ 15 and empstat = 1" and click "continue."

B) What percentage of women *within the labor force* is working:

i. In Agriculture; Mexico 2000: 4.7% Uganda 2002: 79.7%

ii. In Service; Mexico 2000: 23.9% Uganda 2002: 9.0%

```
crosstabs
/tables=occisco by sex by sample
/cells=count column.
```

ANSWERS: Analyze the Sample – Part IV Graphical Analysis

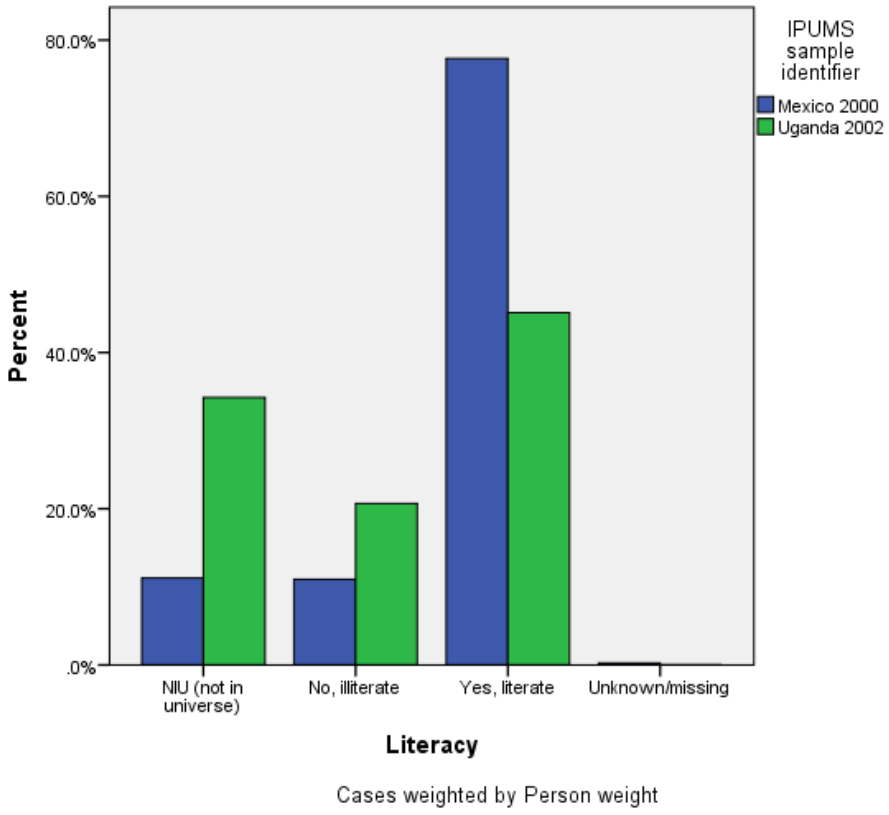
Section 2

Graph the Data

A) What percent of the population is literate in each sample? Mexico 2000 ~78% Uganda 2002 ~45%

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

```
graph
/bar(grouped)=pct by lit by sample.
```



ANSWERS: Analyze the Sample – Part IV Graphical Analysis, Age/Literacy

Recode literacy to look at literacy rates across age

Section 2

Graph the Data

A) Which country has higher overall literacy? Mexico 2000

B) At (approximately) which ages are literacy rates highest?

Mexico 2000 ~13-25

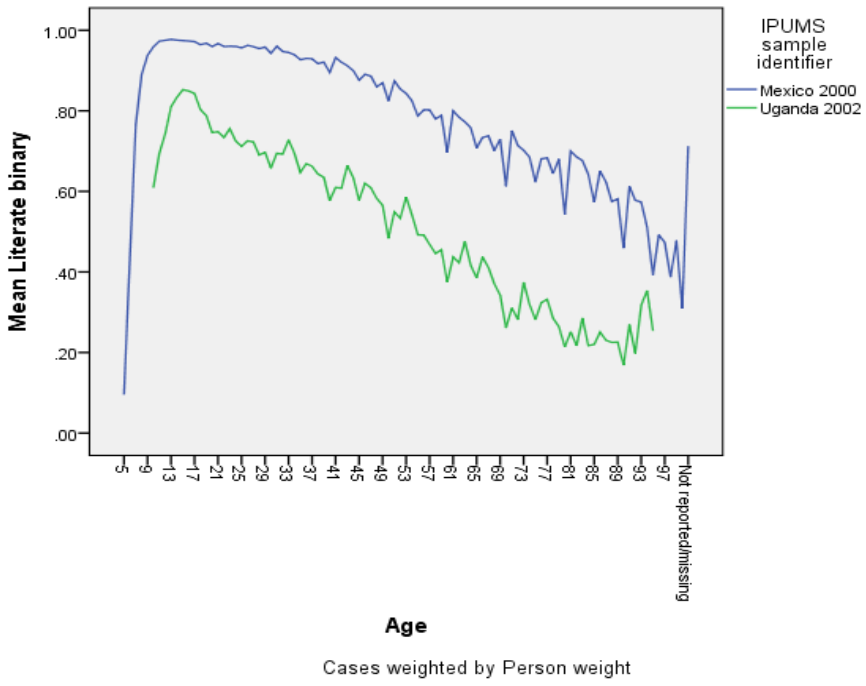
Uganda 2002 ~14-18

C) 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.

```
recode lit (0=sysmis) (9=sysmis) (1=0) (2=1) into literate.
variable labels  literate 'Literate binary'.
execute.

freq lit literate.

graph
/line(multiple)=mean(literate) by age by sample.
```

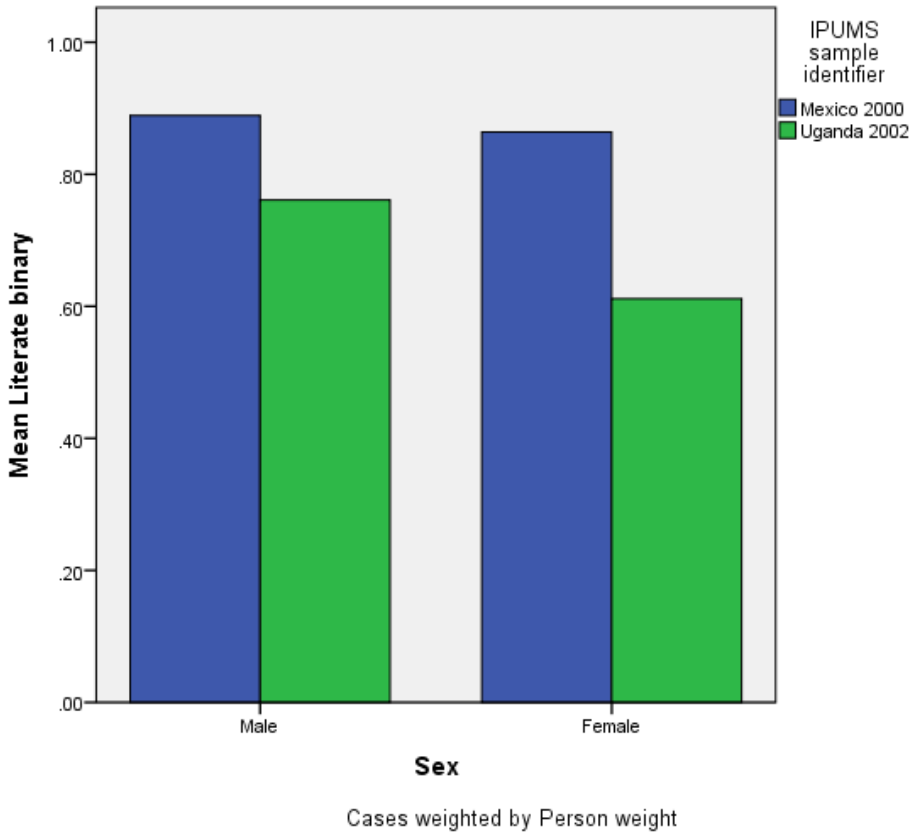


Section 3

Analyze Recoded Data

D) In which country are literacy rates nearly equal for men and women? Mexico 2000

```
graph
/bar(grouped)=mean(literate) by sex by sample.
```



Section 3

Graph the Data

E) What type of floor material is most common in Uganda 2002?
None (earth floor)

```
graph
/bar(grouped)=pct by floor by sample.
```