Minnesota Population Center
Training and Development

IPUMS – Int.I Extraction and Analysis

Exercise 2

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 Cambodia, Ireland, and Uruguay.

10/24/2012

Research Questions

What are the differences in water supply, internet access, car ownership, and age distribution among Cambodia, Uruguay, and Ireland?

Objectives

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

IPUMS Variables

- WATSUP: Water supply
- SEX: Sex
- INTRNET: Internet Access
- AUTOS: Automobiles available
- EDATTAN: Educational Attainment
- AGE: Age
- WTHH: Household weight technical variable

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 specified cases to include in a procedure

Review Answer Key (page 9)

Common Mistakes to Avoid

1 Not fully decompressing the data

2 Giving the wrong filepath to indicate the dataset

3 Forget to close a procedure with "run;"

4 Forget to terminate a command with a semicolon ";"

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:

  WATSUP: Water supply

  SEX: Sex

  INTRNET: Internet Access

  AUTOS: Automobiles available

  EDATTAN: Educational Attainment

  AGE: Age

  WTHH: Household weight technical variable

•••

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 the 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 SAS. 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://international.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 SAS 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

- Open the "ipumsi_000##.sas" file
- In the do file 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_000##.dat";
- Choose Submit under the Run file menu

Analyze the Sample – Part I Variable Documentation

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

Section 1

Analyze the Variables

A) Find the codes page for the SAMPLE variable and write down the code values for:

i. Cambodia 2008? _____

ii. Ireland 2006? _____

iii. Uruguay 2006? _____

B) Are there any differences in the universe of WATSUP among the three samples? _____

C) What is the universe for EMPSTAT:

i. Cambodia 2008? _____

ii. Ireland 2006? _____

iii. Uruguay 2006? _____

Analyze the Sample – Part II Frequencies

Section 1

Analyze the Data

...

Section 2

Weight the Data

A) How many individuals are in each of the sample extracts?

_____

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

When to use the person weights (WTPER)

To get a more accurate estimation of demographic patterns within a county from the sample, you will have to turn on the person weight.

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

Cambodia 2008 _____

Ireland 2006 _____

Uruguay 2006 _____

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

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

Cambodia 2008 _____

Ireland 2006 _____

Uruguay 2006 _____

```
proc freq;
        tables watsup*sample;
        weight wtper;
run;
```

Analyze the Sample - Part II Frequencies (WTHH)

Suppose you were interested not in the number of people with or without water supply, but in the number of households – you will need to use the household weight.

Section 3

Weight the Data

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 "if" statement to select only cases where the PERNUM equals 1.

D) What proportion of households in each country did not have access to piped water?

Cambodia 2008 _____

Ireland 2006 _____

Uruguay 2006 _____

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

E) In which country do individuals have the most access to the internet? _____

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

Section Continues below…

Analyze the Sample - Part II Frequencies (WTHH)

Section 3

Analyze the Data

F) In that country, what proportion of households have both access to internet and at least one car? _____

Note: First you'll 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;
```

G) 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 wtper;
run;
```

Analyze the Sample – Part III Graphical Analysis

Suppose you want to compare age distribution across countries.

Section 1

Graph the Data

A) Approximately what percent of Uruguay's population is around 50 years old? _____

B) Compare the age distributions of Cambodia and Ireland. Is this a pattern that could be observed in other developed and developing nations? _____

C) 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.

Section 2

Analyze the Data

D) 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;
```

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 online to answer each question.

Section 1

Analyze the Variables

A) Find the codes page for the SAMPLE variable and write down the code values for:

i. Cambodia 2008? 1,162

ii. Ireland 2006? 3,728

iii. Uruguay 2006? 8,585

B) 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.

C) What is the universe for EMPSTAT:

i. Cambodia 2008? All persons.

ii. Ireland 2006? Non-absent persons age 15+.

iii. Uruguay 2006? Persons age 14+.

ANSWERS: Analyze the Sample – Part II Frequencies

Section 1

Analyze the Data

• • •

Section 2

Weight the Data

A) How many individuals are in each of the sample extracts?
Cambodia 2008: 1,340,121; Ireland 2006: 440,314; Uruguay 2006: 256,866

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

When to use the person weights (WTPER)

To get a more accurate estimation of demographic patterns within a county from the sample, you will have to turn on the person weight.

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

Cambodia 2008 13,401,210

Ireland 2006 4,403,140

Uruguay 2006 3,065,604

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

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

Cambodia 2008 84.12%

Ireland 2006 14.25%

Uruguay 2006 3.22%

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

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

Suppose you were interested not in the number of people with or without water supply, but in the number of households – you will need to use the household weight.

Section 3

Weight the Data

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 "if" statement to select only cases where the PERNUM equals 1.

D) What proportion of households in each country did not have access to piped water?

Cambodia 2008 83.91%

Ireland 2006 12.59%

Uruguay 2006 3.28%

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

E) In which country do individuals have the most access to the internet? Ireland 2006 (53.1% Yes)

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

Section Continues below…

Section 3

Analyze the Data

F) In that country, what proportion of households have both access to internet and at least one car? 40.7%

Note: First you'll 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;
```

G) In which country is educational attainment (Secondary and University in particular) between men and women most equal? Least equal?

Most equal completion rates: Uruguay (18.68/19.76%; 3.99/4.23%)

Least equal completion rates: Cambodia (4.76/2.44%; 1.36/0.6%)

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

ANSWERS: Analyze the Sample – Part III Graphical Analysis

Suppose you want to compare age distribution across countries.

Section 1

Graph the Data

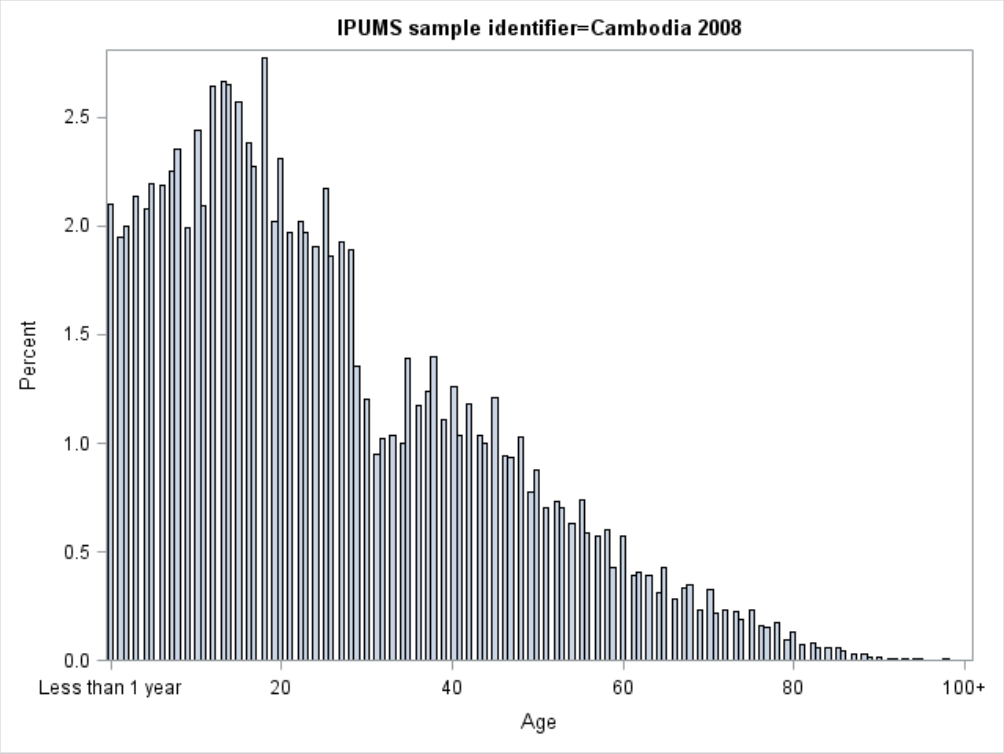
A) Approximately what percent of Uruguay's population is around 50 years old? ~2.4%

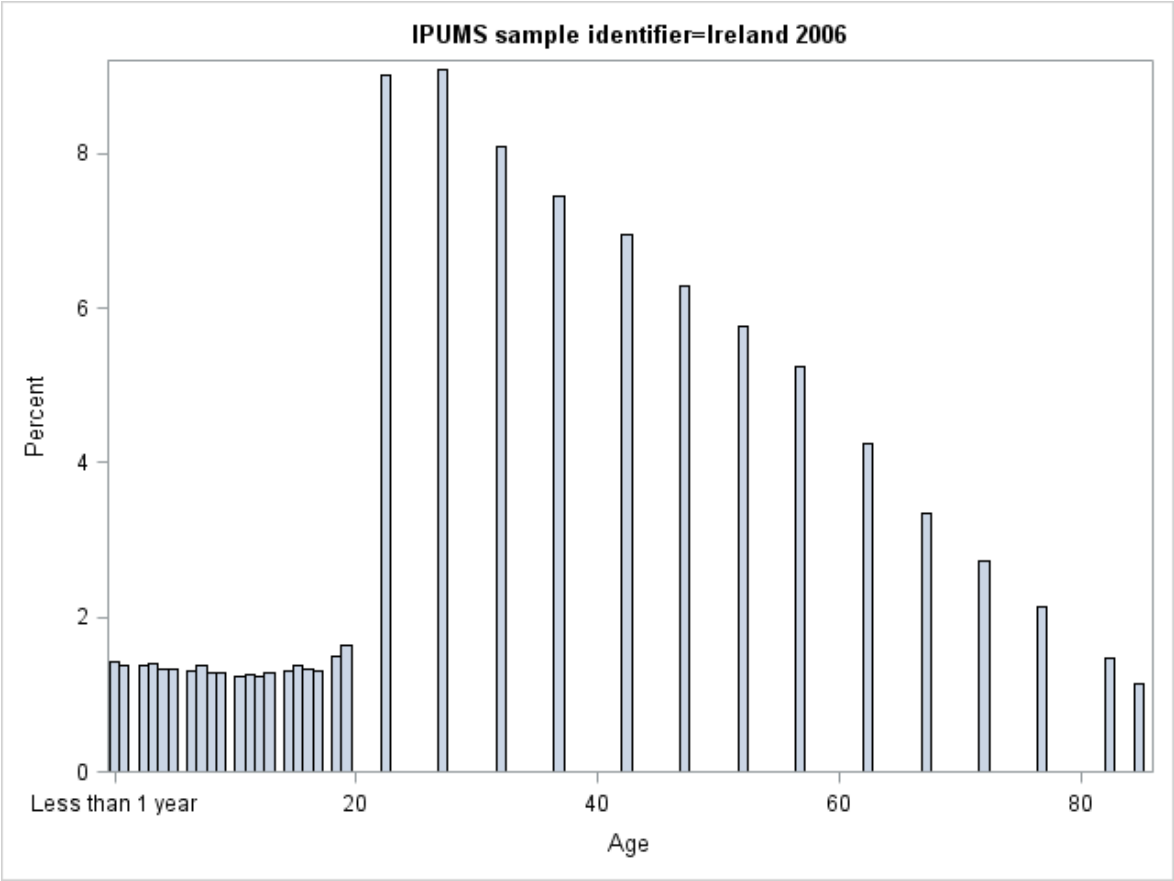
B) Compare the age distributions of Cambodia and Ireland. Is this a pattern that could be observed in other developed and developing nations? A large proportion of Cambodia's population is 25 or younger, while the mean age of Ireland's population seems a bit older.

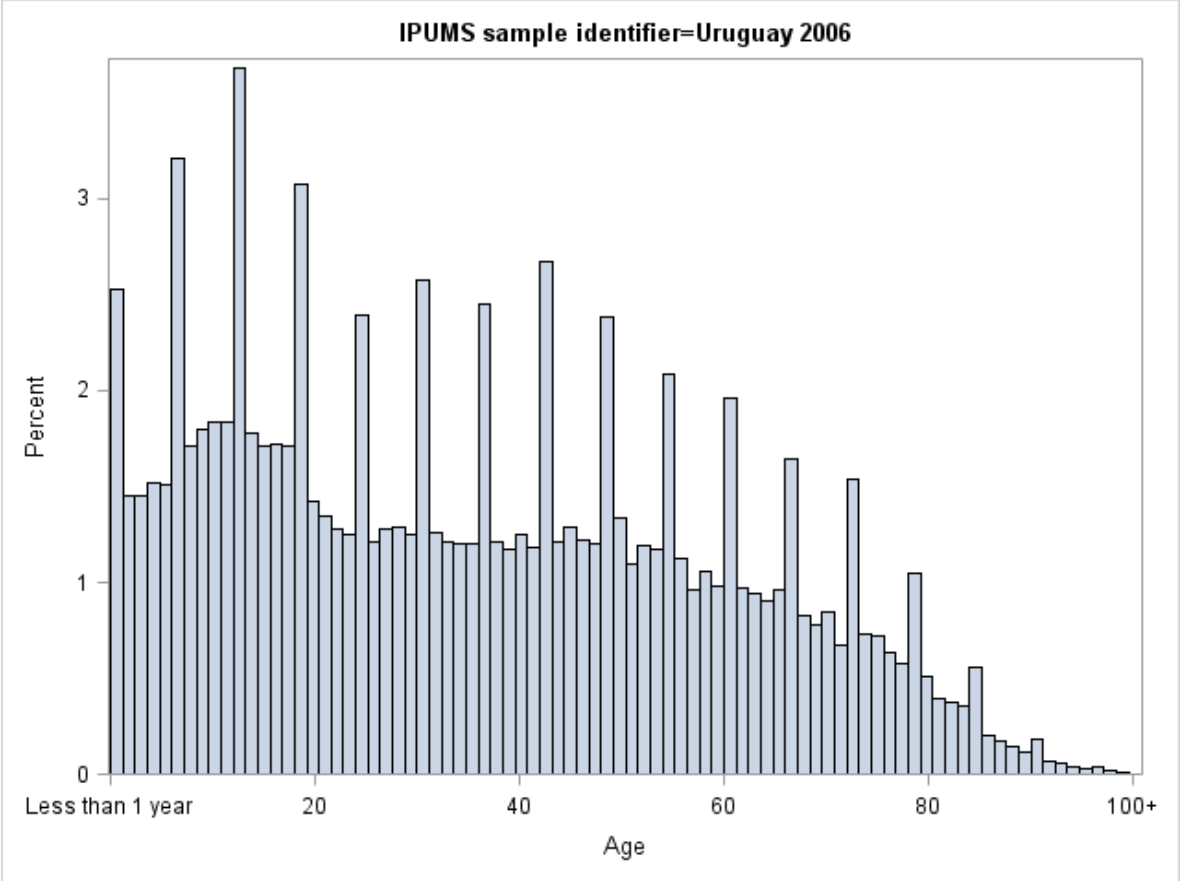
C) Can the shape of the histogram of Ireland compared to the other countries indicate anything about the differences in data collection? "All Ireland samples provide single years of age through 19 and 5-year age intervals thereafter, top-coded at 85+" From the Comparability Tab on the website.

```
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.




IPUMS sample identifier=Ireland 2006
Percent
0
2
4
6
8
Less than 1 year
20
40
60
80
Age


IPUMS sample identifier=Uruguay 2006
Percent
0
1
2
3
Less than 1 year
20
40
60
80
100+
Age

Section 2

Analyze the Data

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

Women:

Cambodia 2008 23 Ireland 2006 32 Uruguay 2006 35

Men:

Cambodia 2008 20 Ireland 2006 32 Uruguay 2006 32

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