Minnesota Population Center

Training and Development

IPUMS â€“ International 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.

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 R
- 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
- EDATTAIN: Educational Attainment
- AGE: Age
- HHWT: Household weight technical variable

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. But, for your reference, these are some quick explanations for commands that this tutorial will use:

- **%>%** - The pipe operator which helps make code with nested function calls easier to read. When reading code, it can be read as "and then". The pipe makes it so that code like `ingredients %>% 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 values
- **mutate** - Add on a new variable to a dataset
- **ggplot** - Make graphs using ggplot2
- **weighted.mean** - Get the weighted mean of the a variable

Review Answer Key (At end of document)

Common Mistakes to Avoid

1) Not changing the working directory to the folder where your data is stored
2) Mixing up = and == ; To assign a value in generating a variable, use "<-" (or "="). Use "==" to test for equality.

Note: In this exercise, for simplicity we will use "weighted.mean". For analysis where variance estimates are needed, use the survey or srvyr package instead.

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 2008 sample for Cambodia, 2006 for Ireland and 2006 for Uruguay
- 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
  - EDATTAIN: Educational Attainment
  - AGE: Age
  - HHWT: Household weight technical variable

Step 2: Request the data

- Click the blue VIEW CART button under your data cart
- Review variable selection

- Click the blue Create Data Extract button
- Review the â€˜Extract Request Summaryâ€™ 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 R. If you would like to use a different stats package, see: http://international.ipums.org/international/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 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

Step 2: 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")
```

Step 3: 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 Documents directory 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("ipumsi_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-labes vignette in the R package. From R run command: `vignette("value-labels", package = "ipumsr")`

Analyze the Sample â€“ Part I Variable Documentation

Section 1: Analyze the Variables

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

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 â€“ Part II Frequencies

Section 1: Analyze the Data

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

____

```
data %>%
  group_by(SAMPLE = as_factor(SAMPLE, level = "both")) %>%
  summarize(n = n())
```

Section 2: When to use the person weights (PERWT)

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 ________________

```
data %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(n = sum(PERWT))
```

C) Using weights, what proportion of individuals in each country did not have access to piped water? Cambodia 2008 _______________
Ireland 2006 __________________
Uruguay 2006 ________________

```
data %>%
  mutate(
    NOT_PIPED = WATSUP %>%
      lbl_collapse(~ .val %/% 10) %>%
      as_factor() %>%
      {. != "Yes, piped water"}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(NOT_PIPED = weighted.mean(NOT_PIPED, PERWT, na.rm = TRUE))
```

Section 3: When to use Household Weights (HHWT)

Suppose you were interested not in the number of people with or without water supply, but in the number of households â€“ you will 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 â€œfilterâ€• 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 ________________

```
data %>%
  filter(PERNUM == 1) %>%
  mutate(
    NOT_PIPED = WATSUP %>%
      lbl_collapse(~ .val %/% 10) %>%
      as_factor() %>%
      {. != "Yes, piped water"}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(NOT_PIPED = weighted.mean(NOT_PIPED, HHWT, na.rm = TRUE))
```

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

```
data %>%
  mutate(
    HAVE_INTERNET = INTERNET %>%
      as_factor() %>%
      {. == "Yes"}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(HAVE_INTERNET = weighted.mean(HAVE_INTERNET, PERWT, na.rm =
TRUE))
```

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

Note: First youâ€™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 %>%
  filter(as_factor(SAMPLE) == "Ireland 2006") %>%
  mutate(
    HAVE_INTERNET = INTERNET %>%
      as_factor() %>%
      {. == "Yes"},
    HAVE_AUTO = AUTOS %>%
    {. > 0 & . < 8}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(HAVE_BOTH = weighted.mean(HAVE_INTERNET & HAVE_AUTO, PERWT, na.rm
= TRUE))
```

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: ________________

```
data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(
    HAVE_SEC = weighted.mean(EDATTAIN == 3, PERWT, na.rm = TRUE),
    HAVE_UNIV = weighted.mean(EDATTAIN == 4, PERWT, na.rm = TRUE)
  )
```

Analyze the Sample â€“ Part III Graphical Analysis

Section 1: Graph the Data

Suppose you want to compare age distribution across countries.

A) Approximately what percent of Uruguayâ€™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?

____

```
ggplot(data, aes(x = as.numeric(AGE), y = ..prop.., weight = PERWT)) +
  geom_bar() +
  facet_wrap(~as_factor(SAMPLE), ncol = 1)
```

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 _______

```
data_summary <- data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(age_med = median(AGE))

ggplot(data_summary, aes(x = SAMPLE, y = age_med, fill = SEX)) +
  geom_col(position = "dodge", width = 0.8) +
  scale_fill_manual(values = c(Male = "#7570b3", Female = "#e6ab02"))
```

ANSWERS Analyze the Sample â€“ Part I Variable Documentation

Section 1: Analyze the Variables

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

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

i. Cambodia 2008? *116200801*

ii. Ireland 2006? *372200601*

iii. Uruguay 2006? *858200621*

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 â€“ Part II Frequencies

Section 1: Analyze 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

```
data %>%
  group_by(SAMPLE = as_factor(SAMPLE, level = "both")) %>%
  summarize(n = n())
#> # A tibble: 3 x 2
#>                       SAMPLE       n
#>                       <fctr>   <int>
#> 1 [116200801] Cambodia 2008 1340121
#> 2  [372200601] Ireland 2006  440314
#> 3  [858200621] Uruguay 2006  256866
```

Section 2: When to use the person weights (PERWT)

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 *13401210*
Ireland 2006 *4403140*
Uruguay 2006 *3065604*

```
data %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(n = sum(PERWT))
#> # A tibble: 3 x 2
#>          SAMPLE        n
#>          <fctr>    <dbl>
#> 1 Cambodia 2008 13401210
#> 2  Ireland 2006  4403140
#> 3  Uruguay 2006  3065604
```

C) Using weights, what proportion of individuals in each country did not have access to piped water? Cambodia 2008 *85.68%* Ireland 2006 *5.61%*
Uruguay 2006 *3.22% Note, we have treated NIU/Unknown as lacking water, but it would also be reasonable to treat them as missing.*

```
data %>%
  mutate(
    NOT_PIPED = WATSUP %>%
      lbl_collapse(~ .val %/% 10) %>%
      as_factor() %>%
      {. != "Yes, piped water"}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(NOT_PIPED = weighted.mean(NOT_PIPED, PERWT, na.rm = TRUE))
#> # A tibble: 3 x 2
#>          SAMPLE  NOT_PIPED
#>          <fctr>      <dbl>
#> 1 Cambodia 2008 0.85681218
#> 2  Ireland 2006 0.05610314
#> 3  Uruguay 2006 0.03218061
```

Section 3: When to use the household weights (HHWT)

Suppose you were interested not in the number of people with or without water supply, but in the number of households â€“ you will 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 â€œfilterâ€• 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 *86.51*
Ireland 2006 *9.90%*
Uruguay 2006 *3.28% Note, we have treated NIU/Unknown as lacking water, but it would also be reasonable to treat them as missing.*

```
data %>%
  filter(PERNUM == 1) %>%
  mutate(
    NOT_PIPED = WATSUP %>%
      lbl_collapse(~ .val %/% 10) %>%
      as_factor() %>%
      {. != "Yes, piped water"}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(NOT_PIPED = weighted.mean(NOT_PIPED, HHWT, na.rm = TRUE))
#> # A tibble: 3 x 2
#>          SAMPLE NOT_PIPED
#>          <fctr>     <dbl>
#> 1 Cambodia 2008 0.8651485
#> 2  Ireland 2006 0.0990099
#> 3  Uruguay 2006 0.0328401
```

E) In which country do individuals have the most access to the internet?
Ireland (53.1% Yes - again including NIU/Unknown as not having access)

```
data %>%
  mutate(
    HAVE_INTERNET = INTERNET %>%
      as_factor() %>%
      {. == "Yes"}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(HAVE_INTERNET = weighted.mean(HAVE_INTERNET, PERWT, na.rm =
TRUE))
#> # A tibble: 3 x 2
#>          SAMPLE HAVE_INTERNET
#>          <fctr>         <dbl>
#> 1 Cambodia 2008     0.0024162
#> 2  Ireland 2006     0.5306758
#> 3  Uruguay 2006     0.1438366
```

F) In that country, what proportion of households have both access to internet and at least one car?
50.6% (again including NIU/Unknown as not having access)

```
data %>%
  filter(as_factor(SAMPLE) == "Ireland 2006") %>%
  mutate(
    HAVE_INTERNET = INTERNET %>%
      as_factor() %>%
      {. == "Yes"},
    HAVE_AUTO = AUTOS %>%
    {. > 0 & . < 8}
  ) %>%
  group_by(SAMPLE = as_factor(SAMPLE)) %>%
  summarize(HAVE_BOTH = weighted.mean(HAVE_INTERNET & HAVE_AUTO, PERWT, na.rm
= TRUE))
#> # A tibble: 1 x 2
#>         SAMPLE HAVE_BOTH
#>         <fctr>     <dbl>
#> 1 Ireland 2006 0.5061524
```

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 (18.7%/19.8%; 4.0%/4.2%)*

    - Least equal completion rates: *Cambodia (4.7%/2.4%; 1.3%/0.6%) (again including NIU/Unknown as not that level of education)*

```
data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(
    HAVE_SEC = weighted.mean(EDATTAIN == 3, PERWT, na.rm = TRUE),
    HAVE_UNIV = weighted.mean(EDATTAIN == 4, PERWT, na.rm = TRUE)
  )
#> # A tibble: 6 x 4
#> # Groups:   SAMPLE [?]
#>           SAMPLE    SEX   HAVE_SEC   HAVE_UNIV
#>           <fctr> <fctr>      <dbl>       <dbl>
#> 1 Cambodia 2008   Male 0.04758998 0.013574355
#> 2 Cambodia 2008 Female 0.02442265 0.006043526
#> 3  Ireland 2006   Male 0.28476800 0.128888929
#> 4  Ireland 2006 Female 0.30387437 0.146710838
#> 5  Uruguay 2006   Male 0.18683797 0.039929928
#> 6  Uruguay 2006 Female 0.19761573 0.042343716
```

ANSWERS Analyze the Sample â€“ Part III Graphical Analysis

Section 1: Graph the Data

Suppose you want to compare age distribution across countries.

A) Approximately what percent of Uruguayâ€™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 Cambodiaâ€™s population is 25 or younger, while the mean age of Irelandâ€™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?
â€œAll Ireland samples provide single years of age through 19 and 5-year age intervals thereafter, top-coded at 85+â€ From the Comparability Tab on the website.

```
ggplot(data, aes(x = as.numeric(AGE), y = ..prop.., weight = PERWT)) +
  geom_bar() +
  facet_wrap(~as_factor(SAMPLE), ncol = 1)
```

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*

```
data_summary <- data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(age_med = median(AGE))

ggplot(data_summary, aes(x = SAMPLE, y = age_med, fill = SEX)) +
  geom_col(position = "dodge", width = 0.8) +
  scale_fill_manual(values = c(Male = "#7570b3", Female = "#e6ab02"))
```