Minnesota Population Center

Training and Development

IHIS Extraction and Analysis

Exercise 1

OBJECTIVE: Gain an understanding of how the IHIS dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the IHIS dataset to explore basic frequencies of flu vaccination, health insurance coverage, and educational attainment, and the relationship between overall health status and employment status.

11/13/2017

Research Questions

What is the distribution of insurance coverage and educational attainment in the United States? How many people in the US receive a flu shot every year?

Objectives

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

IHIS Variables

- HINOTCOVE: Health Insurance Status
- EDUCREC2: Education attainment
- EMPSTAT: Employment status
- HEALTH : Self-reported health status
- VACFLUSH12M: Flu vaccination within the past 12 months

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
- **weighted.mean** - Get the weighted mean of the a variable

Review Answer Key (At End)

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 IHIS

Go to http://www.ihis.us, click on User Registration and Login and Apply for access. Log in if you are a registered user. If you are a first time user, enter an email address and password, then submit your user information so you can create IHIS data extracts.

Step 1: Make an Extract

- Return to the homepage and click on Browse and Select Data.
- Click the Select Samples box, and check the box for the 2010 sample. Click the Submit sample selections box.
- ï‚§Using the drop down menu or search feature, select the following variables and add them to your data cart using the plus symbol to the left of the variable:
  - HINOTCOVE: Health Insurance Status
  - EDUCREC2: Education attainment
  - EMPSTAT: Employment status
  - HEALTH: Self-reported health status
  - VACFLUSH12M: Flu vaccination within the past 12 months

Step 2: Request the Data

- Click the green VIEW CART button under your data cart.
- Review variable selection. Note that additional variables are in your data cart. The data extract system automatically supplies variables that indicate the sample (YEAR), are needed for variance estimation (SERIAL, PERNUM), and are used for weighting the variables and years selected. Click the green Create Data Extract button.
- Review the 'Extract Request Summary' screen, describe your extract, and click Submit Extract.

- Note: there are three different data extracts required to complete the exercises included in this tutorial; you may create the data extracts as you go, or may want to look ahead and create all three extracts before beginning the exercises.
- You will receive an email when the data is available to download.
- To access the page to download the data, follow the link in the email, or click on the Download or 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: https://www.ihis.us/ihis/extract_instructions.shtml

Step 1: Download the Data

- Go to http://www.ihis.us/ 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("ihis_00001.xml")
data <- read_ipums_micro(ddi)
```

```
# Or, if you downloaded the R script, the following is equivalent:
#    source("ihis_00001.R")
```

- This tutorial will also rely on the dplyr package, 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)
```

- 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 Frequencies

Section 1: Analyze the Data

A) On the website, find the universe page for the HINOTCOVE variable and write down the universe statement, which indicates who was asked this specific question. _____

B) How many people in 2010 sample report being uninsured?

_____

C) What proportion of the 2010 sample report being uninsured?

_____

```
data %>%
  group_by(HINOTCOVE = as_factor(HINOTCOVE)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
```

Section 2: Weighting the Data: Using person weights (PERWEIGHT)

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

A) Using weights:

i. How many people were uninsured in 2010? _____

ii. What proportion of the population was uninsured in 2010? _____

```
data %>%
  group_by(HINOTCOVE = as_factor(HINOTCOVE)) %>%
  summarize(n = sum(PERWEIGHT)) %>%
  mutate(pct = n / sum(n))
```

B) On the website, examine the variable description for EDUCREC2 and write down the universe statement. _____

C) Using weights, how many people had a 4 year college or Bachelorâ€™s degree as their highest educational attainment? _____

D) Using weights, what proportion of the population had a 4 year college or Bachelorâ€™s degree as their highest educational attainment?

_____

```
data %>%
  group_by(EDUCREC2 = as_factor(EDUCREC2)) %>%
  summarize(n = sum(PERWEIGHT)) %>%
  mutate(pct = n / sum(n))
```

Analyze the Sample â€“ Part II Relationships in the Data

These questions require you to create a second data extract using the 1972, 1981, 1997, and 2010 samples and the HEALTH variable.

Section 1: Analyze the Data

A) Determine the proportion of the population that reported excellent health status over time. Note: Youâ€™ll want to exclude the unknown responses for HEALTH, so use a filter in R to exclude them. On the website, check the codes for HEALTH.

- 1972: _____
- 1981: _____
- 1997: _____
- 2010: _____

```
unknown_labels <- c("Unknown-refused", "Unknown-not ascertained", "Unknown-
don't know")
data %>%
  mutate(HEALTH = HEALTH %>% lbl_na_if(~.lbl %in% unknown_labels) %>%
as_factor()) %>%
  group_by(YEAR) %>%
  summarize(health_ex = weighted.mean(HEALTH == "Excellent", PERWEIGHT, na.rm
= TRUE))
```

Section 2: Thinking Critically

B) An initial glance may lead you to conclude that excellent health has declined since 1972. This interpretation is complicated by a change in the data collection during this time period.

Using the website, navigate to the HEALTH variable description and find the year that this variable changed from a four-point scale to a five-point scale.

_____

Analyze the Sample â€“ Part III Relationships in the Data

These questions require you to create a third extract using samples of years 1997 through 2010, and the VACFLUSH12M variable.

Section 1: Analyze the Data

A) Examine the documentation for the flu shot variable (VACFLUSH12M) and write down the universe statements from 1997 to 2010. _____

B) Suppose you want to examine trends in the proportion who reported Influenza vaccination during the past 12 months using the extracted data. Since this variable was only for a sample person we will use the sample weight (SAMPWEIGHT) instead of the person weight. Also, exclude respondents who did not answer yes or no using the code â€œfilter(VACFLUSH12M == 1 | VACFLUSH12M == 2)â€•. Which survey years had the highest and lowest percentage receiving the vaccine within the past 12 months?
Highest: _____
Lowest: _____

```
data %>%
  mutate(
    flu_bin = VACFLUSH12M %>%
      lbl_na_if(~.lbl %in% c("NIU", "Refused", "Not Ascertained", "Don't
know")) %>%
      as_factor() %>%
      {. == "Yes"}
    ) %>%
  group_by(YEAR) %>%
  summarize(had_flu_shot = weighted.mean(flu_bin, PERWEIGHT, na.rm = TRUE))
%>%
  filter(had_flu_shot %in% range(had_flu_shot))
```

ANSWERS Analyze the Sample â€“ Part I Frequencies

Section 1: Analyze the Data

A) On the website, find the universe page for the HINOTCOVE variable and write down the universe statement, which indicates who was asked this specific question.
2010: All persons.

B) How many people in 2010 sample report being uninsured?
16029

C) What proportion of the 2010 sample report being uninsured?
17.8%

```
data %>%
  group_by(HINOTCOVE = as_factor(HINOTCOVE)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 3 x 3
#>              HINOTCOVE     n         pct
#>                 <fctr> <int>       <dbl>
#> 1     No, has coverage 73225 0.813828132
#> 2 Yes, has no coverage 16029 0.178147506
#> 3   Unknown-don't know   722 0.008024362
```

Section 2: Weighting the Data: Using person weights (PERWEIGHT)

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

A) Using weights:

i. How many people were uninsured in 2010?
48,311,184

ii. What proportion of the population was uninsured in 2010?
15.9%

```
data %>%
  group_by(HINOTCOVE = as_factor(HINOTCOVE)) %>%
  summarize(n = sum(PERWEIGHT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 3 x 3
#>              HINOTCOVE         n         pct
#>                 <fctr>     <dbl>       <dbl>
#> 1     No, has coverage 253627732 0.833955861
#> 2 Yes, has no coverage  48311184 0.158852483
#> 3   Unknown-don't know   2187170 0.007191655
```

B) On the website, examine the variable description for EDUCREC2 and write down the universe statement.
Persons age 5+

C) Using weights, how many people had a 4 year college or Bachelorâ€™s degree as their highest educational attainment?
40,229,764

D) Using weights, what proportion of the population had a 4 year college or Bachelorâ€™s degree as their highest educational attainment?
13.23%

```
data %>%
  group_by(EDUCREC2 = as_factor(EDUCREC2)) %>%
  summarize(n = sum(PERWEIGHT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 13 x 3
```

```
#>                                EDUCREC2        n          pct
#>                                  <fctr>    <dbl>        <dbl>
#>  1                                  NIU 21484842 0.0706445221
#>  2    Never attended/kindergarten only  9986359 0.0328362461
#>  3                  Grade 1, 2, 3, or 4 18528078 0.0609223571
#>  4                     Grade 5, 6, or 7 17159651 0.0564228187
#>  5                              Grade 8  7671179 0.0252236798
#>  6                  Grade 9, 10, or 11 27232931 0.0895448705
#>  7                             Grade 12 68783532 0.2261678138
#>  8             1 to 3 years of college 67331814 0.2213944055
#>  9   4 years college/Bachelor's degree 40229764 0.1322798860
#> 10                 5+ years of college 21271386 0.0699426553
#> 11                     Unknown-refused  2433885 0.0080028814
#> 12             Unknown-not ascertained   121908 0.0004008469
#> 13 Unknown (1996 forward - Don't know)  1890757 0.0062170168
```

ANSWERS Analyze the Sample â€“ Part II Relationships in the Data

These questions require you to create a second data extract using the 1972, 1981, 1997, and 2010 samples and the HEALTH variable.

Section 1: Analyze the Data

A) Determine the proportion of the population that reported excellent health status over time. Note: Youâ€™ll want to exclude the unknown responses for HEALTH, so use a `lbl_na_if()` in R to set them to missing On the website, check the codes for HEALTH.

- 1972: *51.8%*
- 1981: *49.3%*
- 1997: *38.3%*
- 2010: *35.2%*

```
unknown_labels <- c("Unknown-refused", "Unknown-not ascertained", "Unknown-
don't know")
data %>%
  mutate(HEALTH = HEALTH %>% lbl_na_if(~.lbl %in% unknown_labels) %>%
as_factor()) %>%
  group_by(YEAR) %>%
  summarize(health_ex = weighted.mean(HEALTH == "Excellent", PERWEIGHT, na.rm
= TRUE))
#> # A tibble: 4 x 2
#>    YEAR health_ex
#>   <int>     <dbl>
#> 1  1972 0.5182957
#> 2  1981 0.4931913
```

```
#> 3  1997 0.3825985
#> 4  2010 0.3520339
```

Section 2: Thinking Critically

B) An initial glance may lead you to conclude that excellent health has declined since 1972. This interpretation is complicated by a change in the data collection during this time period.
Using the website, navigate to the HEALTH variable description and find the year that this variable changed from a four-point scale to a five-point scale.
1872

ANSWERS Analyze the Sample â€“ Part III Relationships in the Data

These questions require you to create a third extract using samples of years 1997 through 2010, and the VACFLUSH12M variable.

Section 1: Analyze the Data

A) Examine the documentation for the flu shot variable (VACFLUSH12M) and write down the universe statements from 1997 to 2010.
1997-2004: Sample adults age 18+ & 2005-2010: Sample adults and sample children

B) Suppose you want to examine trends in the proportion who reported Influenza vaccination during the past 12 months using the extracted data. Since this variable was only for a sample person we will use the sample weight (SAMPWEIGHT) instead of the person weight. Also, exclude respondents who did not answer yes or no using the code â€œfilter(VACFLUSH12M == 1 | VACFLUSH12M == 2)â€•.
Which survey years had the highest and lowest percentage receiving the vaccine within the past 12 months?
Highest: *2010*
Lowest: *2005*

```
data %>%
  mutate(
    flu_bin = VACFLUSH12M %>%
      lbl_na_if(~.lbl %in% c("NIU", "Refused", "Not Ascertained", "Don't
know")) %>%
      as_factor() %>%
      {. == "Yes"}
    ) %>%
  group_by(YEAR) %>%
  summarize(had_flu_shot = weighted.mean(flu_bin, PERWEIGHT, na.rm = TRUE))
%>%
  filter(had_flu_shot %in% range(had_flu_shot))
#> # A tibble: 2 x 2
```

```
#>    YEAR had_flu_shot
#>   <dbl>        <dbl>
#> 1  2005    0.2142348
#> 2  2010    0.3595962
```