IPUMS Data Training Exercise:

An introduction to IPUMS NHIS

(Exercise 1 for R)

IPUMS HEALTH SURVEYS

Learning goals

- Understand how IPUMS NHIS dataset is structured
- Create and download an NHIS data extract
- Decompress data file and read data in R
- Analyze the health insurance coverage, educational attainment, and flu shot attainment of people in the United States using sample code

Summary

In this exercise, you will gain an understanding of how the NHIS dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the NHIS dataset to explore basic frequencies of flu vaccination, health insurance coverage, educational attainment, and overall health status. You will create data extracts that include the variables HINOTCOVE, EDUCREC2, HEALTH, and VACFLUSH12M; then you will use the sample code to analyze these data.

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:

Code	Purpose
%>%	The pipe operator 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 provided for IPUMS data into a factor variable for R
summarize	Summarize a dataset's 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 variable

Common Mistakes to Avoid

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

Registering with NHIS

Go to https://nhis.ipums.org/nhis/, click on Log in and login if you are a registered user. If you are a first time user, click on Create an account, enter an email address and password, and then submit your user information so you can create NHIS data extracts.

Make a Data Extract

- Return to the homepage and click on Browse and Select Data.

Select Samples

- Click the Select Samples button, and check the box for the 2010 sample. Click the submit sample selections button.

Select Variables

- The variable drop-down menus allow you to explore variables by topic. For example, you might expect to find variables about flu shots under the "Vaccinations" group.
- The search tool allows you to search for variables. Observe the options for limiting your search results by variable characteristics or variable type.
- You may 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.
- You may view 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.
- 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 variables:

    HINOTCOVE: Health insurance status

    EDUCREC2: Educational attainment

Review and submit your extract

- 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.
- 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.

Create two additional extracts

- Create an extract using the 1972, 1981, 1997, and 2010 samples and the HEALTH variable.
- Create an extract using the samples of years 1997 through 2010 and the VACFLUSH12M variable.

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://ipums.org/support/exercises

Download the data

- Go to https://nhis.ipums.org/nhis/ and click on Download or Revise Extracts.
- Right-click on the Data link next to the extract you created.
- Choose "Save Target As..." (or "Save Link As...").
- Save into "Documents" (Documents should pop up as the default location).
- Do the same 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 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 have already made):

```
library(ipumsr)
ddi <- read_ipums_ddi("nhis_00001.xml")
data <- read_ipums_micro(ddi)
```

Or, if you downloaded the R script, the following is equivalent: source("nhis_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-labels vignette in the R package. From R, run command:

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

Analyze the Sample

Part 1: Frequencies

These questions use the first data extract with the variables HINOTCOVE and EDUCREC2 for the 2010 sample.

1. 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.

   _____

2. How many people in the 2010 sample report being uninsured? _____
3. What proportion of the 2010 sample report being uninsured? _____

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

Using person weights (PERWEIGHT)

To get a more accurate estimation of demographic patterns, you will have to utilize the person weights.

4. Using weights:
   a. How many people were uninsured in 2010? _____
   b. 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))
```

5. On the website, examine the variable description for EDUCREC2 and write down the universe statement.

   _____

   _____

6. Using weights, how many people had a 4-year college or Bachelor's degree as their highest educational attainment? _____
7. Using weights, what proportion of the population had a 4 year college or Bachelor's degree as their highest educational attainment? _____

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

Part 2: Relationships in the Data

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

8. Determine the proportion of the population that reported excellent health status over time.

1972: _____ 1997: _____ 1981: _____ 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))
```

9. 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. _____

These questions require you to use the third data extract with the VACFLUSH12M variable for the samples of years 1997 through 2010.

10. Examine the documentation for the flu shot variable (*VACFLUSH12M*) and write down the universe statements from 1997 to 2010. _____
11. 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.

    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) %>%
```

Answers

Part 1: Frequencies

1. 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.
   1988: Sample persons under age 18. 1998-2010: All persons.

2. How many people in the 2010 sample report being uninsured? 16,029 individuals in the sample

3. What proportion of the 2010 sample report being uninsured? 17.81% of the sample

Using person weights (PERWEIGHT)

4. Using weights:
   a. How many people were uninsured in 2010? 48,311,184 individuals

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

5. On the website, examine the variable description for EDUCREC2 and write down the universe statement. 1982-2010: Persons age 5+.

6. Using weights, how many people had a 4 year college or Bachelor's degree as their highest educational attainment? 40,229,764

7. Using weights, what proportion of the population had a 4 year college or Bachelor's degree as their highest educational attainment? 13.23%

Part 2: Relationships in the Data

8. Determine the proportion of the population that reported excellent health status over time. 1972: 51.8% 1997: 38.3%
   1981: 49.3% 2010: 35.2%
9. 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. 1982

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

10. 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 age 18+ and sample children under age 18.

11. Which survey years had the highest and lowest percentage receiving the vaccine within the past 12 months?
    Highest: 2010 Lowest: 2005