IPUMS Data Training Exercise:

An introduction to IPUMS NHIS

(Exercise 1 for Stata)

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 Stata
- 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.

Stata Code to Review

Code	Purpose
generate	Creates a new variable, "replace" specifies a value according to cases
mean	Displays a simple tabulation and frequency of one variable
tabulate	Displays a cross-tabulation for up to 2 variables
!=	Not equal to

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 "=". Use "==" to specify a case when a variable is a desired value using an *if* statement.
- Not using svy suite of commands for appropriate variance estimation.

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 Stata. 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 Stata link next to the Data link.

Decompress the data

- All IPUMS extracts are compressed. There are many applications available for decompressing files. Windows users may consider WinZip and WinRAR; MacGZIP and Stuffit Expander are applications for Macs.
- Find the "Documents" folder under the Start menu.
- Right click on the ".dat" file.
- Use your decompression software to extract the .dat files.
- Double-check that the Documents folder contains three files starting with "nhis_000...".

Read the data

- Open Stata from the Start menu.
- In the "File" menu, choose "Change working directory..."

  Select "Documents", click "OK".

- In the "File" menu, choose "Do..."

  Select the *.do file.

- You will see "end of do-file" when Stata has finished reading in the data.

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? _____

```
tab hinotcove
```

Using person weights (PERWEIGHT)

To get a more accurate estimation of demographic patterns, you will have to utilize the person weight and variance estimation variables. To account for the complex sample design of NHIS data, you should use the *svy* suite of commands for appropriate variance estimation when analyzing NHIS data in Stata.

To set up your svy commands to use the person weight and public use primary sampling unit and strata variables, use the following syntax:

```
svyset psu [pweight=perweight], strata(strata)
```

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

```
svy: tab hinotcove, count format(%14.3gc)
```

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

```
svy: tab hinotcove
```


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? _____

```
svy: tab educrec2, count format(%14.3gc)
```

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

```
svy: tab educrec2
```

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. Be sure to reset your svy commands using the previous syntax if you cleared your commands in Stata.

8. Determine the proportion of the population that reported excellent health status over time. Note: You'll want to exclude the unknown responses for HEALTH, so u se a conditional command in Stata to exclude them. On the website, check the codes for HEALTH.

```
svy: tab year health if health < 7, row
```

1972: _____ 1997: _____

1981: _____ 2010: _____

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. _____

Because of the universe for VACFLUSH12M, you must use sampweight instead of perweight when setting up your svy commands:

```
svyset, clear
svyset psu [pweight=sampweight], strata(strata)
```

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. Exclude respondents who did not answer yes or no using the code `"if vacflush12m ==1|vacflush12m ==2".`

    Which survey years had the highest and lowest percentage receiving the vaccine within the past 12 months?

    Highest: _____ Lowest: _____

```
svy: tab year vacflush12m if vacflush12m ==1|vacflush12m ==2,
row
```


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
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