IPUMS Data Training Exercise:

An introduction to IPUMS NHIS

(Exercise 2 for Stata)

IPUMS HEALTH SURVEYS

Learning goals

- Create and download an NHIS data extract
- Decompress data file and read data into Stata
- Analyze the data 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 associations among BMI, poverty, health status, sleep, and frequency of exercise. You will create a data extract that includes the variables AGE, SEX, POORYN, HEALTH, BMI, HRSLEEP, and VIG10FWK; 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 2004 through 2010 samples. 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 sleep under the "Health Behaviors" 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:
    - AGE: Age
    - SEX: Sex
    - POORYN: Above or below poverty threshold
    - HEALTH: Health status
    - BMI: Body Mass Index
    - HRSLEEP: Usual hours of sleep per day
    - VIG10FWK: Frequency of vigorous activity (10+ min) per week

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.

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: Group Means

1. On the website, find the codes page for the HRSLEEP and HEALTH variables. What code values for HRSLEEP should be excluded to avoid skewing the average number of hours slept? How would you restrict the code values for HEALTH to eliminate unknown responses? _____

   _____

   _____

2. Suppose you wanted to study the relationship between hours of sleep and health status. Determine the average reported hours of sleep per night by health status. On average, how many hours does an individual with excellent health in this sample sleep per night? _____

```
mean hrsleep if health<6 & hrsleep>0 & hrsleep<25, over(health)
```

3. Is there a noticeable trend between health status and hours of sleep using this sample? _____
4. Does the trend change for people under 60 in this sample? _____

   _____

```
mean hrsleep if health<6 & hrsleep>0 & hrsleep <25 & age<60,
over(health)
```

Part 2: Weighting the Data

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

5. Without weights, what proportion of people in this sample was below the poverty threshold in 2010? _____

```
tab year pooryn, row
```


6. Using weights, what proportion of the population was below the poverty threshold in 2010? _____

```
svyset psu [pweight=perweight], strata(strata)
svy: tab year pooryn, row
```

7. Using the household weight (and you must exclude all but one individual from a household), what proportion of households was below the poverty threshold in 2010? _____

```
svyset, clear
svyset psu [pweight=hhweight], strata(strata)
svy: tab year pooryn if pernum==1, row
```

Part 3: Generating Variables

Generate a variable that is 0 when an individual exercises less than 3 times a week, and 1 when an individual exercises 3 or more times a week. ****NOTE: Graphing procedures (in Part 4 of this exercise) across statistical packages do not consistently allow for all weight and variance estimation options. Because this section generates variables used in the graphing section of this exercise, neither Part 3 nor Part 4 of this exercise use weights.***

8. Check the output of the do file in the Log window to find the codes for VIG10FWK. Which code means "Never"? _____
   Note: You'll have to exclude codes above 28 when defining when exer3 is greater than 3 times a week.
9. What is the average difference in BMI for an individual in this sample who exercises at least 3 times a week compared to someone who exercises fewer than 3 times per week? _____
   Remember to restrict the codes for BMI so unknown and missing codes are excluded.

```
gen exer3 = 0
replace exer3 = 1 if vig10fwk>=3 & vig10fwk<=28
mean bmi if bmi>0 & bmi<99, over(exer3)
```

10. What percent of more frequent exercisers report excellent health? _____
11. What percent of less frequent exercisers report excellent health? _____

```
tab health exer3 if health<6, col
```

Part 4: Graphing

Create a graph to show the mean BMI over age for males and females.

12. How does the universe for BMI appear on this graph? _____
13. Approximately at what age does BMI peak:

    for women? _____ for men? _____

```
egen meanbmim = mean(bmi) if sex==1 & bmi>0 & bmi<99, by(age)
egen meanbmif = mean(bmi) if sex==2 & bmi>0 & bmi<99, by(age)
twoway(line meanbmim meanbmif age, sort)
```

Introduce the variable POORYN

14. Create a graph to show how an associated effect of poverty status on BMI differs with gender, controlling for frequent exercise. Comment on three apparent trends.

    _____

```
graph bar (mean) bmi if bmi>0 & bmi<99 & pooryn !=9,
over(pooryn) over (exer3) over(sex)
```

Answers

Part 1: Group Means

1. On the website, find the codes page for the HRSLEEP and HEALTH variables. What code values for HRSLEEP should be excluded to avoid skewing the average number of hours slept? How would you restrict the code values for HEALTH to eliminate unknown responses? HRSLEEP: 00 NIU; 25 Less than 1 hour; 97 Unknown-refused; 98 Unknown-not ascertained; 99 Unknown-don't know; HEALTH: 7 Unknown-refused; 8 Unknown-not ascertained; 9 Unknown-don't know

2. Suppose you wanted to study the relationship between hours of sleep and health status. Determine the average reported hours of sleep per night by health status. On average, how many hours does an individual with excellent health in this sample sleep per night? 7.2 hours

3. Is there a noticeable trend between health status and hours of sleep using this sample? There seems to be no trend at all, except perhaps Excellent and Poor health have slightly higher averages, which could indicate people in poor health sleep to improve health and people with excellent health are associated with getting more sleep.

4. Does the trend change for people under 60 in this sample? When excluding the older population (perhaps with a higher incidence of poor health), better health is associated with more hours of sleep, though the differences between averages is small.

Part 2: Weighting the Data

5. Without weights, what proportion of people in this sample was below the poverty threshold in 2010? 16.48% of the sample

6. Using weights, what proportion of the population was below the poverty threshold in 2010? 13.76% of the sample

7. Using the household weight (and you must exclude all but one individual from a household), what proportion of households was below the poverty threshold in 2010? 12.91% of the sample

Part 3: Generating Variables

8. Check the output of the do file in the Log window to find the codes for VIG10FWK. Which code means "Never"? 95 Never

9. What is the average difference in BMI for an individual in this sample who exercises at least 3 times a week compared to someone who exercises fewer than 3 times per week? 1.2 BMI (27.7-26.5)

10. What percent of more frequent exercisers report excellent health? 41.37%

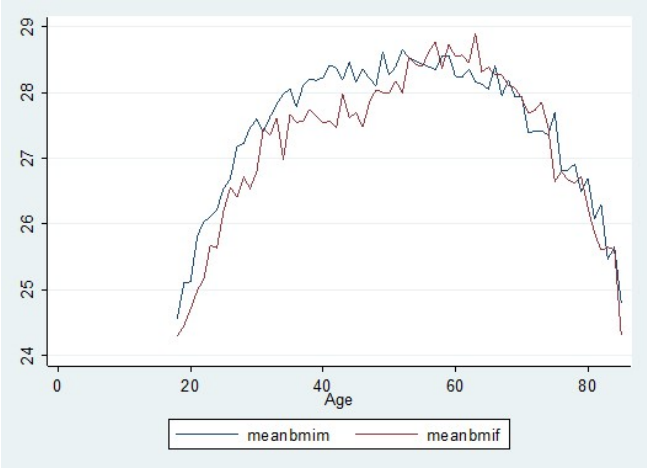
11. What percent of less frequent exercisers report excellent health 34.19%

Part 4: Graphing

12. How does the universe for BMI appear on this graph? There appears to be no BMI for individuals below 18, because the universe for BMI is only for adults older than 18.

13. Approximately at what age does BMI peak for

women? ~ 61 years old men? ~50 years old



Introduce the variable POORYN

14. Create a graph to show how an associated effect of poverty status on BMI differs with gender, controlling for frequent exercise. Comment on three apparent trends.
Women under the poverty threshold are more likely to have a higher BMI on average whether or not they exercise. Frequent exercisers have lower BMI's on average in each category. Men under the poverty threshold seem to have a lower BMI on average controlling for exercise.