IPUMS Data Training Exercise:

CPS Extraction and Analysis

(Exercise 2 for R)

Learning goals

- Gain an understanding of how the IPUMS dataset is structured and how it can be leveraged to explore your research interests.
- 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

Summary

This exercise will use the IPUMS dataset to explore associations between parent and child health, and analyze relationships between disability variables and marital status to answer the following research question: "Is there an association between parent and child health? What are the trends in disabilities and marital status?" You will create a data extract that includes the variables AGE, SEX, MARST, HEALTH, DIFFHEAR, and DIFFEYE; then you will use 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 which helps make code with nested function calls is easier to read. When reading code, read as “and then”. The pipe make it so that the code like **ingredient %>% 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 variables
mutate	Add on a new variable to a dataset
ggplot	Make graphs using ggplot2
weighted.mean	Get the weighted mean of the “a” variable

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.
3. Not including missing values when needed. The attached characteristics have missing values when the person doesn't have the relationship, but sometimes you want to treat that as a "No", not as a missing value.

Note: In this exercise, for simplicity we will use "weighted.mean". For analysis where variance estimates that take the survey design into consideration, use either the "survey" or "srvyr" package instead.

Registering with IPUMS

Go to http://cps.ipums.org, click on Register with IPUMS and apply for access. On login screen, enter email address and password and submit it!

Creating and downloading an IPUMS data extract

Make an Extract

- Go to the homepage and go to Select Data
- Click the Select/Change Samples box, check the box for the 2010 and 2011 ASEC samples, then click Submit Sample Selections
    - Using the drop down menu or search feature, select the following variables:
        - AGE: Age
        - SEX: Sex
        - MARST: Marital status
        - HEALTH: Health status
        - DIFFHEAR: Hearing difficulty
        - DIFFEYE: Vision difficulty

Request the Data

- Click the VIEW CART button under your data cart
- Review variable selection. Click the Create Data Extract button
    - Click on ‘Attach Characteristics’
        - The following screen will allow you to select who you would like to attach variables for. Make sure the "Spouse" boxes are checked for all variables and that HEALTH also has the boxes for "Father" and "Mother" checked.
    - Describe your extract and click Submit Extract
- You will get an email when the data is available to download
    - To get to the 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://cps.ipums.org/cps/extract_instructions.shtml

Download the Data

- Go to http://cps.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

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 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("cps_00001.xml")

data <- read_ipums_micro(ddi)

# Or, if you downloaded the R script, the following is
equivalent:

#    source("cps_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

Part I: Creating New Variables

1. What are the names of the attached variables (can be found on extract request screen, or in the data)?

   ____

2. On the website, find the FAQ entry for attaching characteristics. What value will the respondents without a parent or spouse present have for the attached variables? ________________________________
3. What are the MARST codes for married respondents?

   ____

4. Create a variable for married men equal to the difference in spouses' age.

```
data <- data %>%

    mutate(AGE_DIFF = ifelse(SEX == 1 & (MARST %in% c(1, 2)),
    AGE - AGE_SP, NA))
```

5. What is the mean age difference between married men and their spouses?

   ____

   a. For men aged 30 and under? ____________________
   b. For 50 and over? ____________________________

```
data %>%

     summarize(mn_all_married_men = weighted.mean(AGE_DIFF,
     ASECWT, na.rm = TRUE))

data %>%

     filter(AGE <= 30) %>%

     summarize(mn_under30 = weighted.mean(AGE_DIFF, ASECWT,
     na.rm = TRUE))

data %>%

     filter(AGE >= 50) %>%

     summarize(mn_over50 = weighted.mean(AGE_DIFF, ASECWT, na.rm
     = TRUE))
```

Part II Relationships in the Data

6. What is the universe for DIFFEYE and DIFFHEAR? What is the Code for NIU (Not in Universe)? ____
7. What percent of the population (in the universe) is deaf or has a serious hearing difficulty? ____
    a. What percent of the population (in the universe) is blind or has serious sight difficulties? ____

```
data %>%

     filter(DIFFHEAR != 0) %>%

     summarize(DIFFHEAR = weighted.mean(DIFFHEAR == 2, ASECWT))

data %>%

     filter(DIFFEYE != 0) %>%

     summarize(DIFFEYE = weighted.mean(DIFFEYE == 2, ASECWT))
```

8. What percent of the deaf population is married with a spouse present?

    ____

```
data %>%

     filter(DIFFHEAR == 2) %>%

     summarize(MARST = weighted.mean(MARST == 1, ASECWT))
```

9. What percent of the deaf population is married to a spouse who is also deaf?

   ____

```
data %>%

     filter(DIFFHEAR == 2) %>%

     mutate(DIFFHEAR_SP_BIN = !is.na(DIFFHEAR_SP) & DIFFHEAR_SP
     == 2) %>%

     summarize(COUPLE_DEAF = weighted.mean(DIFFHEAR_SP_BIN,
     ASECWT, na.rm = TRUE))
```

Part III Relationships in the Data

10. What ages of respondents have their parents identified through the attach characteristics? (hint: see variable descriptions for MOMLOC and POPLOC).

    ____

11. Does there seem to be a relationship between parents and children's health?

    ____

```
data_summary <- data %>%

     filter(!is.na(HEALTH_MOM)) %>%
     group_by(HEALTH = as_factor(HEALTH), HEALTH_MOM =
     as_factor(HEALTH_MOM)) %>%
     summarize(n = sum(ASECWT)) %>%
     mutate(pct = n / sum(n))

ggplot(data_summary, aes(x = HEALTH_MOM, y = pct,
     fill = HEALTH)) + geom_col(position = "dodge")
```

12. What other tests could you do to examine this relationship?

    ____

13. Could there be a sampling issue affecting the relationship between children and parent's health? ________________________________

Answers

Part I: Creating New Variables

1. What are the names of the attached variables (can be found on extract request screen, or in the data)? AGE_SP, age of spouse; HEALTH_MOM, health of mother; HEALTH_POP, health of father; HEALTH_SP, health of spouse; DIFFHEAR_SP, hearing disability of spouse; DIFFEYE_SP, vision disability of spouse
2. On the website, find the FAQ entry for attaching characteristics. What value will the respondents without a parent or spouse present have for the attached variables? A missing code
3. What are the MARST codes for married respondents? 1 Married, spouse present; 2 Married, spouse absent
4. Create a variable for married men equal to the difference in spouses' age.

```
data <- data %>%

     mutate(AGE_DIFF = ifelse(SEX == 1 & (MARST %in% c(1, 2)),
     AGE - AGE_SP, NA))
```

5. What is the mean age difference between married men and their spouses? 2.3
   a. For men 30 and under? -.16
   b. For 50 and over? 3.2

```
data %>%

     summarize(mn_all_married_men = weighted.mean(AGE_DIFF,
     ASECWT, na.rm = TRUE))

#> # A tibble: 1 x 1

#>   mn_all_married_men

#>          <dbl>

#> 1        2.266375


data %>%

     filter(AGE <= 30) %>%

     summarize(mn_under30 = weighted.mean(AGE_DIFF, ASECWT,
     na.rm = TRUE))

#> # A tibble: 1 x 1

#>   mn_under30

#>          <dbl>

#> 1        -0.1605179


data %>%

     filter(AGE >= 50) %>%

     summarize(mn_over50 = weighted.mean(AGE_DIFF, ASECWT, na.rm
     = TRUE))

#> # A tibble: 1 x 1

#>   mn_over50

#>          <dbl>

#> 1        3.230359
```

Part II: Relationships in the Data

6. What is the universe for DIFFEYE and DIFFHEAR? What is the Code for NIU (Not in Universe)? Persons age 15+, 0 is the NIU code
7. What percent of the population (in the universe) is deaf or has a serious hearing difficulty? 3.1%
    a. What percent of the population (in the universe) is blind or has serious sight difficulties? 1.7%

```
data %>%
     filter(DIFFHEAR != 0) %>%
     summarize(DIFFHEAR = weighted.mean(DIFFHEAR == 2, ASECWT))
#> # A tibble: 1 x 1
#>   DIFFHEAR
#>        <dbl>
#> 1      0.03093566

data %>%
     filter(DIFFEYE != 0) %>%
     summarize(DIFFEYE = weighted.mean(DIFFEYE == 2, ASECWT))
#> # A tibble: 1 x 1
#>   DIFFEYE
#>        <dbl>
#> 1      0.01666509
```

8. What percent of the deaf population is married with a spouse present? 49.7%

```
data %>%
     filter(DIFFHEAR == 2) %>%
     summarize(MARST = weighted.mean(MARST == 1, ASECWT))
#> # A tibble: 1 x 1
#>   MARST
#>        <dbl>
#> 1      0.4965242
```

9. What percent of the deaf population is married to a spouse who is also deaf? 7.7%

```
data %>%
     filter(DIFFHEAR == 2) %>%
     mutate(DIFFHEAR_SP_BIN = !is.na(DIFFHEAR_SP) & DIFFHEAR_SP
     == 2) %>%
     summarize(COUPLE_DEAF = weighted.mean(DIFFHEAR_SP_BIN,
     ASECWT, na.rm = TRUE))
#> # A tibble: 1 x 1
#>   COUPLE_DEAF
#>        <dbl>
#> 1      0.07663705
```

Part III: Relationships in the Data

10. What ages of respondents have their parents identified through the attach characteristics? (hint: see variable descriptions for MOMLOC and POPLOC). Children under age 19
11. Does there seem to be a relationship between parents and children's health? Parent's health and children's health seem to be directly correlated

```
data_summary <- data %>%

     filter(!is.na(HEALTH_MOM)) %>%

     group_by(HEALTH = as_factor(HEALTH), HEALTH_MOM =
     as_factor(HEALTH_MOM)) %>%

     summarize(n = sum(ASECWT)) %>%

     mutate(pct = n / sum(n))

ggplot(data_summary, aes(x = HEALTH_MOM, y = pct, fill =
     HEALTH)) + geom_col(position = "dodge")
```

12. What other tests could you do to examine this relationship? Correlation matrix, covariance analysis, regression analysis
13. Could there be a sampling issue affecting the relationship between children and parent's health? Yes, parents are reporting children's health