Milestone 1: CDC's National Survey of Family Growth

Julianne; Julianne.Ammirati@emory.edu

February 08, 2017

```
# loading tidyverse and haven
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(haven)
library(shiny)
```

The code used to generate this report is available from my GitHub Account:

https://github.com/JulianneA/N741_Milestone1

Overview and Motivation:

Prior to returning to nursing school, I served as a clinical coordinator for a psychiatry lab at UCSF researching schizophrenia and the prodromal period. I was most intrigued by the clinical interviews that I conducted with the younger women, who were often navigating not only their new psychiatric diagnosis, but the social implications of their diagnosis combined with the "everyday" pressures of being a young woman. My work in psychiatry sparked an interest in women's health, and my while my PhD research will focus on psychological health in the perinatal period, I have interests in the public health aspects of access and utilization of care by female patients.

Project Objectives:

Given the variability of women's access to care in the current healthcare system, I am interested in examining patterns of contraceptive utilization as a function of various socio-demographic factors and how the utilization has changed over time. The National Survey of Family Growth (NSFG) is the largest publicly accessibly dataset I could locate that surveys women regarding their contraceptive use. I do not plan to analyze the entire dataset, but rather, I plan to conduct this exploratory analysis by focusing mostly on family planning variables (coded as Section E of the dataset in the three most recent iterations). A lot of work has been done in this area, but I think that (1) it could use some updated visualization methods to make it more easily digestible, and (2) by breaking it down into various groups based on race/age/demographics, I think the data can tell a better story about the groups that are being left behind as contraceptive methods improve.

Ideally, for this project I would like to reducing the number of variables to strictly sociodemographic variables and contraceptive variables, but expand the number of cases to include the survey results from additional years (downloading the NSFG data for all surveys dating back to 1973). This would better allow me to follow trends over time, to examine how certain contraceptive options have become more or less utilized.

Data:

The CDC has allowed open access to the NSFG survey results from 1973-2015:
https://www.cdc.gov/nchs/nsfg/index.htm.

These data are available for download from a CDC operated FTP site. For the sake of space, I have only provided links to the most recent iteration of the survey, 2013-2015.
The original data for the 2013-2015 survey is available here:
ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NSFG/2013_2015_FemRespData.dat

The CDC's SAS setup file for this dataset is available here:
ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NSFG/sas/2013_2015_FemRespSetup.sas

CDC guidance for the use of this data (ie, variable names, past surveys, etc.) is available here:
https://www.cdc.gov/nchs/nsfg/nsfg_2013_2015_puf.htm

Data Wrangling:

The results of the NSFG projects are available in a .dat file with setup programs that can be run in SAS, STATA and SPSS to properly import and code variables. In order to import the data into R, I chose to download both the setup file (.SAS) and the data file (.dat) from the CDC to my personal drive at Emory. I opened the .SAS program and then updated the directory to read in the .dat file. Finally, I exported to a CSV file using:

```
dm "dexport new 'H:\SAS 9.4 Temporary Files\2013_2015_FemRespData.csv' ";
```

I attempted this process with the most recent NSFG data (2013-2015), and was successful. This was a particularly arduous way of reformatting the dataset, but I don't yet trust my R-abilities to import the .SAS setup file and the dat file without destroying it. Any guidance on this would be appreciated.

Note: To save space, all R-chunk code in the markdown has been printed to PDF but eval has been set to FALSE, so no output is produced.

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   NOWPRGDK = col_character(),
##   OTHKDRAC2 = col_character(),
##   KDBSTRAC = col_character(),
##   OKDISABL2 = col_character(),
##   OTHKDRAC7 = col_character(),
##   KDBSTRAC2 = col_character(),
##   OKDISABL6 = col_character(),
##   OTHKDSPN3 = col_character(),
##   OTHKDRAC11 = col_character(),
##   OTHKDRAC12 = col_character(),
##   KDBSTRAC3 = col_character(),
##   OKBORNUS3 = col_character(),
##   OKDISABL9 = col_character(),
##   OKDISABL10 = col_character(),
##   TRYADOPT4 = col_character(),
##   OTHKDFOS4 = col_character(),
##   OTHKDSPN4 = col_character(),
##   OTHKDRAC16 = col_character(),
##   OTHKDRAC17 = col_character(),
##   KDBSTRAC4 = col_character()
##   # ... with 459 more columns
## )
```

```
## See spec(...) for full column specifications.
```

Following my initial success, I repeated this process with the 2006-2010 and 2011-2013 datasets, and continued to download the .dat and .SAS files for the previous six cycles. The first six NSFG cycles (1973-2002) are much less standardized in their data structure and SAS program setup file syntax. Thus far, I've only imported/re-exported the 1995 dataset as a .csv file.

At this point, I'll need to append the respondent data for each of the datasets that I am able to convert to csv to the current 2013-2015 file. Prior to appending the new cases to the current dataset, each individual csv file will need to be cleaned to remove variables that I do not intend to analyze. The datasets are large on their own (ie: 2013-2015 has 5699 cases and 3207 variables), but they are fairly sparse. While I would ideally merge the datasets prior to the removal of variables from each to allow me to more easily expand my analysis in the future, I am unsure whether my computer could (happily) handle 4 decades of NSFG data. Below is information regarding the 9 NSFG surveys and their pre-processing status.

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   MOSCURRP = col_double(),
##   OTHKDRAC7 = col_character(),
##   KDBSTRAC2 = col_character(),
##   OKDISABL6 = col_character(),
##   OKDISABL10 = col_character(),
##   OKDISABL14 = col_character(),
##   OKDISABL18 = col_character(),
##   OKDISABL22 = col_character(),
##   DATKDCAM_M7 = col_character(),
##   DATKDCAM_Y7 = col_character(),
##   CMOKDCAM7 = col_character(),
##   OTHKDFOS7 = col_character(),
##   OKDDOB_M7 = col_character(),
##   OKDDOB_Y7 = col_character(),
##   CMOKDDOB7 = col_character(),
##   OTHKDSPN7 = col_character(),
##   OTHKDRAC31 = col_character(),
##   OTHKDRAC32 = col_character(),
##   KDBSTRAC7 = col_character(),
##   OKBORNUS7 = col_character()
##   # ... with 386 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: Missing column names filled in: 'X3313' [3313], 'X3314' [3314],
## 'X3315' [3315]
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   NOWPRGDK = col_character(),
##   MOSCURRP = col_double(),
##   OTHKDRAC2 = col_character(),
##   OTHKDRAC3 = col_character(),
##   KDBSTRAC = col_character(),
##   OKDISABL2 = col_character(),
##   OKDISABL3 = col_character(),
##   OTHKDSPN2 = col_character(),
##   OTHKDRAC6 = col_character(),
##   OTHKDRAC7 = col_character(),
##   OTHKDRAC8 = col_character(),
##   KDBSTRAC2 = col_character(),
##   OKBORNUS2 = col_character(),
##   OKDISABL5 = col_character(),
##   OKDISABL6 = col_character(),
##   OKDISABL7 = col_character(),
##   OTHKDSPN3 = col_character(),
##   OTHKDRAC11 = col_character(),
##   OTHKDRAC12 = col_character(),
##   OTHKDRAC13 = col_character()
##   # ... with 901 more columns
## )
## See spec(...) for full column specifications.
```

```
## Warning: 12279 parsing failures.
## row col     expected       actual
##   1  -- 3315 columns 3741 columns
##   2  -- 3315 columns 3741 columns
##   3  -- 3315 columns 3741 columns
##   4  -- 3315 columns 3741 columns
##   5  -- 3315 columns 3741 columns
## ... ... ........... ...........
## See problems(...) for more details.
```

```
## Warning: Missing column names filled in: 'X3716' [3716]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   CASEID = col_integer(),
##   AGE = col_integer(),
##   BDAYCENM = col_integer(),
##   MYSCHOLX = col_integer(),
##   MYSCHOL = col_integer(),
##   COLSTOP6 = col_integer(),
##   COLNEXT7 = col_integer(),
##   COLSTOP7 = col_integer(),
##   COLNEXT8 = col_integer(),
##   COLSTOP8 = col_integer(),
##   COLNEXT9 = col_integer(),
##   COLSTOP9 = col_integer(),
##   WHENEAR0 = col_integer(),
##   WHENEAR1 = col_integer(),
##   WHENEAR2 = col_integer(),
##   WHENEAR3 = col_integer(),
##   WHENEAR4 = col_integer(),
##   VOCSTART = col_integer(),
##   VOCSTOP2 = col_integer(),
##   VOCNEXT3 = col_integer()
##   # ... with 193 more columns
## )
## See spec(...) for full column specifications.
```

```
## Warning: 11748 parsing failures.
## row col     expected       actual
##   1  -- 3716 columns 5754 columns
##   2  -- 3716 columns 5754 columns
##   3  -- 3716 columns 5754 columns
##   4  -- 3716 columns 5754 columns
##   5  -- 3716 columns 5754 columns
## ... ... ........... ...........
## See problems(...) for more details.
```

```
## [1] 5699
```

```
## [1] 5601
```

```
## [1] 12279
```

```
## [1] 10847
```

Year of NFGS	Number of Respondents	Notes
2013-2015	5699	This was the first file imported as a CSV.
2011-2013	5601	Imported as CSV.
2006-2010	12279	Imported as CSV.
2002		Cycle 6, 3 SAS setup files that will take more time to work through.
1995	10847	Cycle 5, Imported as CSV.
1988		Cycle 4
1982		Cycle 3
1976		Cycle 2
1973		Cycle 1

Six previous cycles of the NSFG were conducted prior to 2006, with data dating back to 1973. Depending on the variable coding, I would like to continue to work back in time and import/append these five cycles not yet imported to my dataset, so long as they contain contraceptive use data that is pertinent to this analysis.

This data would require slight cleaning and recoding for my use. My understanding is that the data may need to be in binary (yes/no via 1s and 0s) format to allow R to count frequencies that my desired variable combination appears, so I would need to expand certain variables (ie: CONSTAT1: Current contraceptive method) from their current state in which each method is assigned a different number, to 22 variables specific to each method with 1s/0s to indicate whether that method is currently used by the respondent. Depending on how the data is most easily displayed, the contraceptive method frequency may need to be computed as a ratio (respondents using method A/total respondents using contraceptive method)

Exploratory Analysis:

After the blue bag lecture on Friday, I'm very interested in plotting the data three dimensionally to display the data as a *contraceptive method x year x demographic characteristic, such as age or race* to be able to track how the patterns are altered over time.

The interactive graph utilized by Hans Rosling in the gapminder video that displayed 2 independent variables (lifespan and income) on each axis and had the size and color of the points indicate the population and region is an interesting way to display a large number of variables at once. I am brainstorming ways of displaying method frequency, year, and demographic characterics all at once, and I believe that animating the graph to scroll through years would allowing the changing utilization to be easily visualized. Currently, the CDC has published these data as pie charts or bar charts, but these visualizations do not collapse across the datasets to look at trends over time.

Analysis:

Below is a list of goals for analysis along with course/self-imposed deadlines:

Goal 1: Data Wrangling A lot of my work will go into the initial data importing and wrangling to allow multiple NSFG datasets to be analyzed together. This step will be time-consuming, but it will need to be done meticulously to ensure that my final dataset is not flawed. I plan to have my final dataset merged and re-coded (as needed) by Sunday, February 5th.

Goal 2: I plan to analyze the data from February 5th-February 19th. The codebooks from the CDC are helpful to familiarize myself with the datasets, but these datasets are so sparse, that I need to obtain some information regarding the frequency of responses for contraceptive variables to ensure that I am not including datasets that appear to be inaccurate (ie: did respondents not disclose this information in earlier cycles?).

Goal 3: Visualization I plan on investing a great deal of time into the visualization step, since I believe that when working with so many variables, the interesting, effective dissemination of results depends on good visualization. This step will be the final week of "processing type" work before I begin writing my manuscript. Therefore, I will have functional visualizations completed before March 1st.

MILESTONE 2: Working Prototype The next milestone is due on March 15th, and I will spend from March 1st-15th working on this manuscript.

Manuscript Rewrites After discussing my project with a professor to get feedback, I plan to rewrite my manuscript and edit any analysis steps and implement any suggestions to improve the visualization of the data. The analysis changes will occur from March 15th-April1st. Visualization changes will take the next week, and from April 7th to the 15th, I will conduct the final manuscript edits.

MILESTONE 3: Final Manuscript Due Wednesday, April 19th.