BRFSS Health Survey

Justin Weltz and Andrew Brown

2/5/2018

Source of Data: The Behavioral Risk Factor Surveillance System (BRFSS) is conducted by the Centers for Disease Control (CDC) on the United States Population (supposed to capture the noninstitutionalized adult population older than 18 years residing in the United States).

Notes About the Data:

1. 486,303 observations - the observations are individuals contacted by telephone (this biases the population they are sampling from and may make inferences taken from this study non-applicable to the general US population)

2. There are 279 accessible variables (a lot of demographic information is omitted in order to preserve anonymity) on demographic characteristics, health-related risk behaviors, chronic health conditions, and use of preventative services. However, I will only be studying a subset of these dimensions.

3. In many cases, "I don't know," "None," and "Refused" are coded as multiples of 7, 8 and 9 (depending on the range of numerical responses possible). I had to take them out of the data in order to accurately analyze variables. It is also going to be important to keep track of how many 'NA's are included in my characteristics of interest.

Relevant Variables:

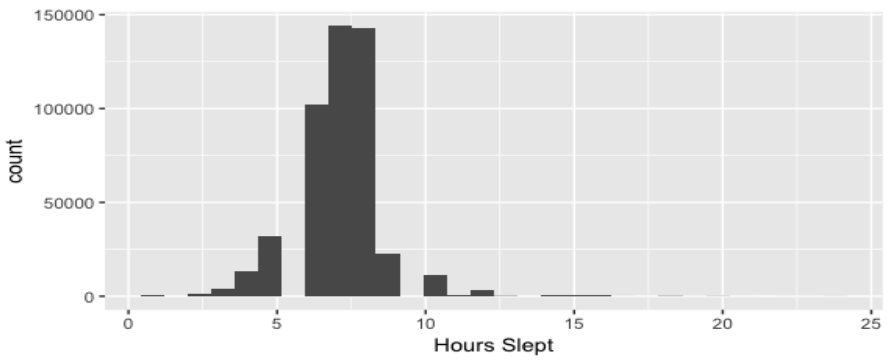
1) Sex (Categorical):

```
BRFSS <- BRFSS %>% mutate(new_sex = ifelse(sex == 9, NA, sex))
#this is indicative of the way many of the variables had to be transformed
```

Sex is pretty evenly distributed, but there are still noticeably more women than men.

2) Average Time Slept (Quantitative): Average time slept in hours over the past month

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.000   7.000   7.054   8.000  24.000    5726
```



I am slightly worried that there is a max value of 24 hours included in the response values. This observation calls into question the validity of the data since this response seems impossible. It is also interesting that the distribution seems pretty symmetrical around 7 hours. The holes in the histogram are also thought provoking.

3) Days of Bad Mental Health (Quantitative): Days of poor mental health reported for the past month

```
##            variable missing complete      n mean   sd p0 p25 median p75
##  BRFSS$new_menthlth    7955   478348 486303 3.44 7.77  0   0      0   2
##  p100     hist
##    30 ▇_______
```

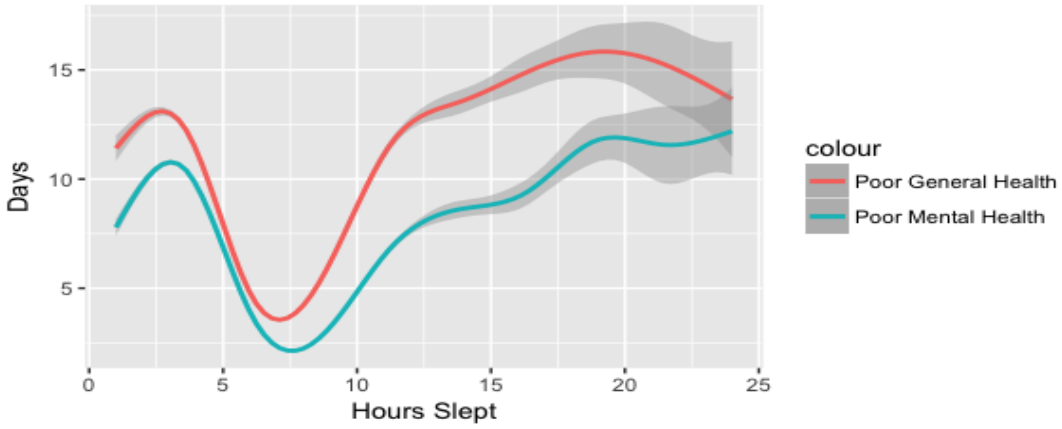
There are a large number of observations for this variable. It is interesting to note that the median of this variable is 0 while the mean is 3.44. This seems to indicate that the data is very right skewed.

4) Days of Poor Health (Quantitative): Days of poor health reported for the past month.

```
##            variable missing complete      n mean   sd p0 p25 median p75
##  BRFSS$new_poorhlth  241274   245029 486303 5.37 9.48  0   0      0   5
##  p100     hist
##    30 ▇_______
```

This variable is interesting because about half of the observations are missing. The distribution is similar to the poor mental health distribution with a higher average and standard deviation. The higher standard deviation means that the points are more dispersed (have higher variance).

Graph Day of Poor Health and Poor Mental Health vs. Average Time Slept



Generally (very descriptively), there seems to be a relationship between average hours slept and poor general and mental health days. Three interesting things pop out of this graph.

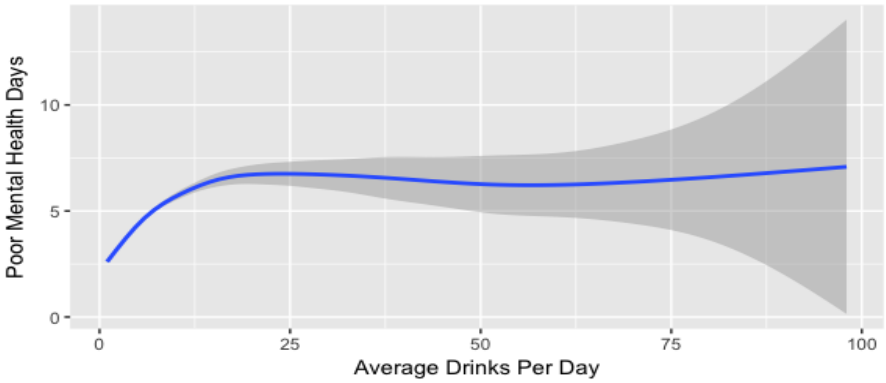
1. Poor mental health and poor general health seem to have a similar relationship to hours slept.

2. The 7-8 hour sweet spot that I often hear about in sleep studies pops out of the graph.

3. The negative relationship between hours slept and poor mental and general health days makes me think very carefully about causality. It seems very unlikely that more hours of sleep are causing people to have more "poor days." I would wager that the direction of causation goes the opposite way, but more analysis will be required

5) Average Alcholic Drinks (Quantitative): Average number of drinks per day over the past month.

```
##       variable missing complete      n mean    sd p0 p25 median p75 p100
##  BRFSS$avedrnk  249223   237080 486303 3.56 10.64  1   1      2   3   99
##      hist
##  ▇_______
```



The distribution of this variable is interesting because there are distinct outliers. When analyzing this variable it will be important to better understand the drinkers who report that they have more than 50 drinks a day. The relationship between poor mental health days and average drinks is also not as pronounced as I thought it would, especially for respondents who report very high average drinks per day (although this may be because there are just so few data points in the high range).

6) General Health (Categorical): Health status judged on a scale of: 1 - Excellent 2- Very good 3- Good 4 - Fair 5 - Poor

```
##           variable missing complete      n mean   sd p0 p25 median p75
##  BRFSS$new_genhlth    1339   484964 486303 2.58 1.09  1   2      3   3
##  p100     hist
##     5 ▆▇_▇_▃__
```

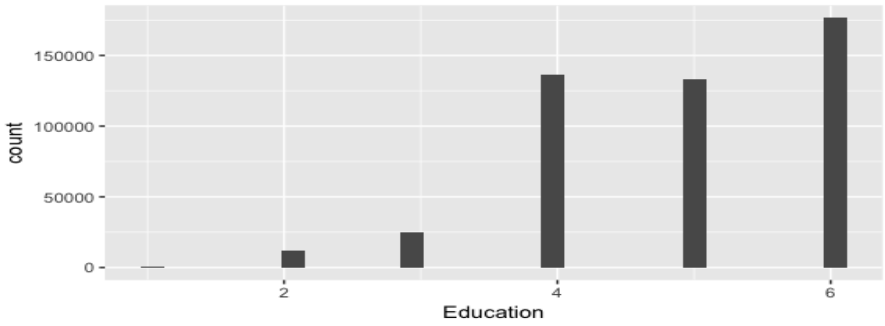
It seems that general health is distributed around "Good Health" (it is both the median and mean). It is also interesting that at least 75% of the population report that they have good to excellent health.

7) Exercise in Past 30 days (Categorical): Whether or not a participant has exercised in the last 30 days.

It is interesting how a significant portion (25.5%) of the survey participants haven't exercised at all in the last month.

8) Education (Categorical): Educational Status

1- Never attended school or only kindergarten 2- Grades 1 through 8 (Elementary) 3- Grades 9 through 11 (Some high school) 4- Grade 12 or GED (High school graduate) 5- College 1 year to 3 years (Some college or technical school) 6- College 4 years or more (College graduate)



This distribution is interesting because most participants haven't completed college and some respondents haven't completed high school. The single largest category is college graduates though.

9) Veteran (Categorical): Whether the survey participant is a veteran or not.
   1- Veteran 2- Non-Veteran

There are not that many veterans in the population proportionally (13.17% of the population). But, I am including it as a relevant variable because I am curious what it's predictive power will be on poor mental health days.

Comparison of Averages

```
     Veterans          Poor Mental Health Days
1            1                 2.814940
2            2                 3.531538
```

This is interesting because the averages do not demonstrate what I thought I would observe. Instead of the veterans having higher average poor mental health days than non-veterans (because of PTSD possibly), the table above demonstrates the opposite trend. It will be interesting to look more closely at why this might be (the biases inherent to self-reporting?).

10) Income Level (Categorical): An income variable with 8 categories ranging from less than $10,000 to $75,000 or more

```
##          variable missing complete      n mean   sd p0 p25 median p75 p100
##  BRFSS$new_income   81301   405002 486303  5.8 2.15  1   4      6   8    8
##      hist
##  ▁▂▃▃▅▆▇█
```

The income variable seems pretty symmetrically distributed around a median income category of 6 (which are participants with incomes between 35,000 to less than 50,000 dollars).

Github-

Name: BRFSS Location: https://github.com/jdnweltz/BRFSS