Exploring the BRFSS data

Chirantan Ganguly

Setup

Load packages

```
library(ggplot2)
library(dplyr)
```

Load data

```
load("F:/Education/Data_Science/Introduction to Probability and Data/brfss2013.RData")
```

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) objective is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population. Factors assessed by the BRFSS in 2013 include tobacco use, HIV/AIDS knowledge and prevention, exercise, immunization, health status, healthy days — health-related quality of life, health care access, inadequate sleep, hypertension awareness, cholesterol awareness, chronic health conditions, alcohol consumption, fruits and vegetables consumption, arthritis burden, and seatbelt use. **Since 2011, BRFSS conducts both landline telephone- and cellular telephone-based surveys.** In conducting the BRFSS landline telephone survey, interviewers collect data from a randomly selected adult in a household. In conducting the cellular telephone version of the BRFSS questionnaire, interviewers collect data from an adult who participates by using a cellular telephone and resides in a private residence or college housing. Health characteristics estimated from the BRFSS pertain to the non-institutionalized adult population, aged 18 years or older, who reside in the US. In 2013, additional question sets were included as optional modules to provide a measure for several childhood health and wellness indicators, including asthma prevalence for people aged 17 years or younger.

- As it has been explicitly mentioned the data collected is purely **OBSERVATIONAL STUDY**, because all that we are doing is studying the past action of the respondents. There has been no random assignment, but random sampling, thus we cannot conclude to causality if the variables are associated. However the associations are generalizable atleast as far as US citizens are concerned.
- Another issue is that the interview are contacted over the phone and **participation is voluntary**, therefore there is no random assignment and also **people who do not own a telephone (landline and/or mobile) and do not live in a private residence are excluded from the study**. In fact, according the Centers for Disease Control and Prevention (CDC) website, “No direct method of accounting for non-telephone coverage is employed by the BRFSS”. From people who have these requisite, living in a private residence/college and own a landline or mobile line, of non-institutionalized adult population, aged 18 years or older, a random sample from each state has been selected, therefore this is a stratified random sample (according to the CDC website : Home telephone numbers are obtained through random-digit dialing).
- We can assume independence of the random sampling even if some of the interviews were contacted over mobile phones and therefore there might be the chances of having interviewed two people from the same household on their personal mobile phones. However we can consider this chance very small.

Part 2: Research questions

Research quesion 1:

Among non-institutionalized adults in the US, we investigate any differences in alcohol comsumption between veterans and non-veterans. The results could indicate whether veterans are at a lower or higher risk of alcohol addiction. We note that respondents are likely to underreport their alcohol consumption, leading to a possible bias in the data. The variables of interest are:

- veteran3: Are You A Veteran
- alcday5: Days In Past 30 Had Alcoholic Beverage
- avedrnk2: Avg Alcoholic Drinks Per Day In Past 30

Research quesion 2:

Among non-institutionalized adults in the US, we investigate any differences in general health condition depending on the the income level of the individual. We also try to investigate if being a smoker adversely affect the general health of the individual irrespective of the income level. The variables of interest are:

- income2: Income Level
- genhlth: General Health
- X_rfsmok3: Adults who are current smokers

Research quesion 3:

Among non-institutionalized adults in the US, Is a respondent’s Body Mass Index (BMI) affect their chances to get depressive disorders? Is there any difference between genders? This is an interesting question as it looks for linkage between opinion about their mental health to a slightly more objective measure of overall health. The difference between genders is also interesting, as one can tease out different perceptions and pressures within society.The variables of interest are:

- addepev2:(Ever told) you that you have a depressive disorder, including depression, major depression, dysthymia, or minor depression
- X_bmi5cat: Computed variable that categorizes BMI into 4 categories. BMI is derived from reported height & weight.
- sex: Reported gender

Part 3: Exploratory data analysis

Research quesion 1:

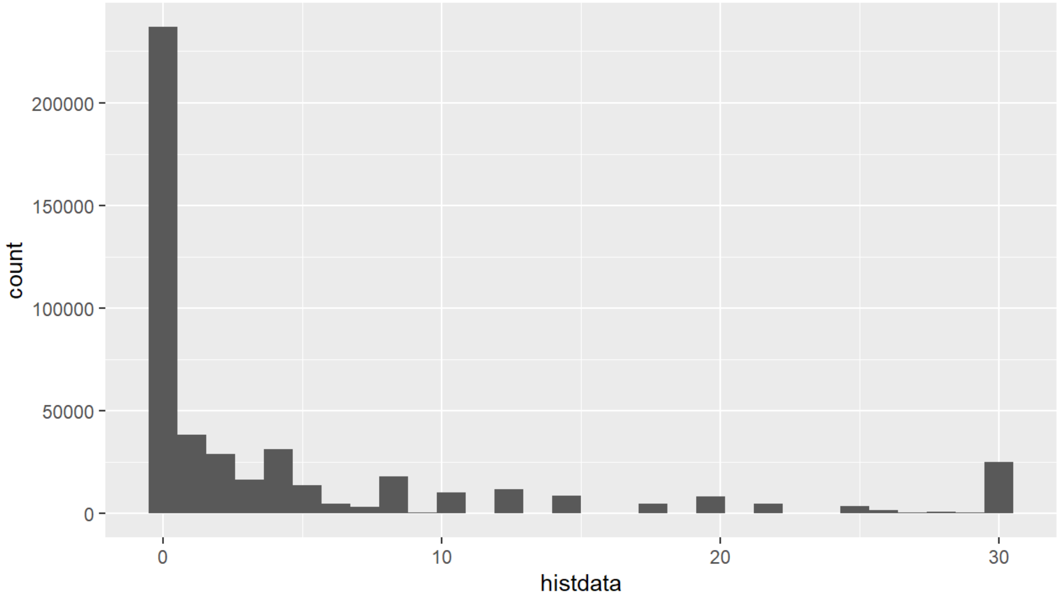
Let us first visualise the alcohol intake of the sample in general. Remember that the in the variable alcday5: The first digit denotes days per week (1) or days per month (2). The remaining digits indicate the count of days.

So we first generalise the data and add a variable alcdaysimple to just store the no. of days and plot a histogram of the data

```
brfss2013<-brfss2013%>%
  mutate(alcdayssimple=ifelse((alcday5 >= 101) & (alcday5 <= 199), (alcday5 - 100) * 30/7, ifelse((alcday5 >=

histdata<-brfss2013$alcdayssimple[!is.na(brfss2013$alcdayssimple)]
ggplot()+aes(x=histdata)+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the data it is clear that majority of people did not consume alcohol at all fpr the last 30 days, but we also see a significant no. of people have alcohol everyday.

To get the total no of drinks last 30 days we multiply alcdayssimple with avedrnk2, and store the result in totaldrinks

```
brfss2013<-brfss2013%>%
  mutate(totaldrinks=alcdayssimple*avedrnk2)
```

Let us select the variables of interest for our first research question and store in `q1`, our variables of interest are:

- `veteran3`

- `totaldrinks`

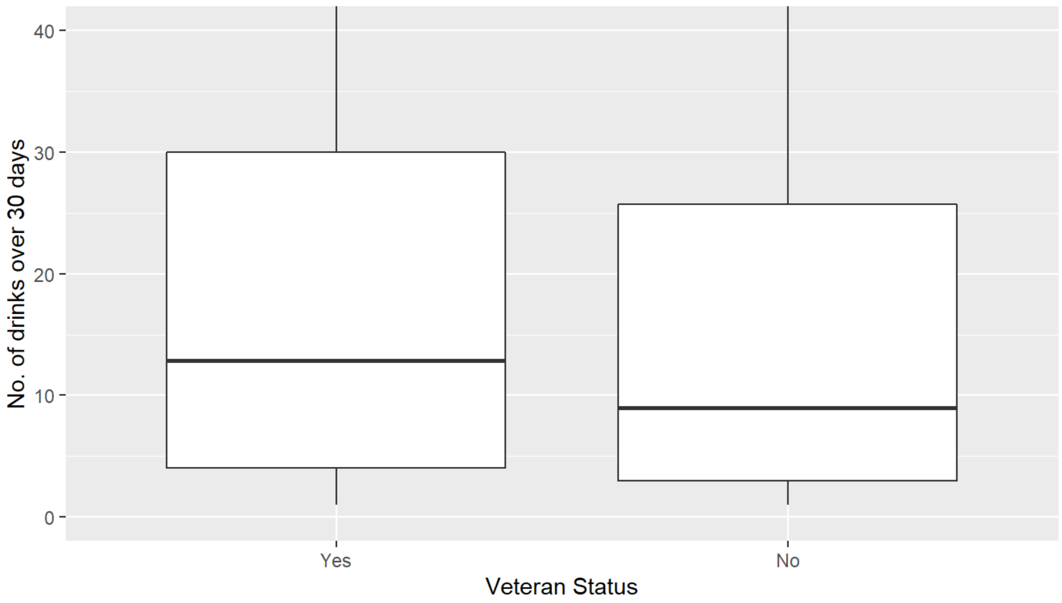
After storing the data let us summarise the data grouped on the basis of veteran3

```
q1<-select(brfss2013,veteran3,totaldrinks)%>% na.omit()
q1%>%
  group_by(veteran3)%>%
  summarise(Drinks_Mean=mean(totaldrinks),Drinks_Median=median(totaldrinks),Drinks_SD=sd(totaldrinks),Drinks_m
```

```
## # A tibble: 2 x 6
##   veteran3 Drinks_Mean Drinks_Median Drinks_SD Drinks_min Drinks_max
##   <fct>          <dbl>         <dbl>     <dbl>      <dbl>      <dbl>
## 1 Yes             29.1          12.9      54.8          1       2280
## 2 No              21.6           9        43.0          1       2280
```

Thus from the data, we see that the mean of the amount of alcoholic beverage consumed by a veteran is almost 34.25% higher than the mean of non-veterans, let us also compare boxplots of the data obtained, separated by veteran status

```
ggplot(data=q1,aes(x=veteran3,y=totaldrinks))+
  geom_boxplot()+coord_cartesian(ylim=c(0,40))+labs(x="Veteran Status",y="No. of drinks over 30 days")
```



Again we see that veterans tend to consume more alcohol. However, we cannot conclude that being a veteran causes one to drink more alcohol, because the data is randomly sampled, not randomly assigned.

Research quesion 2:

Let us select the variables of interest for our second research question and store in `q2`, our variables of interest are:

- `genhlth`
- `income2`
- `X_rfsmok3`

```
q2<-select(brfss2013,genhlth,income2,X_rfsmok3)%>% na.omit()
```

Before doing anything else let us have a look at the variables of our interest.

```
q2%>%
  group_by(genhlth)%>%
  summarise(Number_in_each_category=n())
```

```
## # A tibble: 5 x 2
##   genhlth   Number_in_each_category
##   <fct>                       <int>
## 1 Excellent                   72399
## 2 Very good                  135329
## 3 Good                       124587
## 4 Fair                        54212
## 5 Poor                        22553
```

```
q2%>%
  group_by(income2)%>%
  summarise(Number_in_each_category=n())
```

```
## # A tibble: 8 x 2
##   income2           Number_in_each_category
##   <fct>                               <int>
## 1 Less than $10,000                   24486
## 2 Less than $15,000                   25886
## 3 Less than $20,000                   33669
## 4 Less than $25,000                   40440
## 5 Less than $35,000                   47463
## 6 Less than $50,000                   59958
## 7 Less than $75,000                   63820
## 8 $75,000 or more                    113358
```

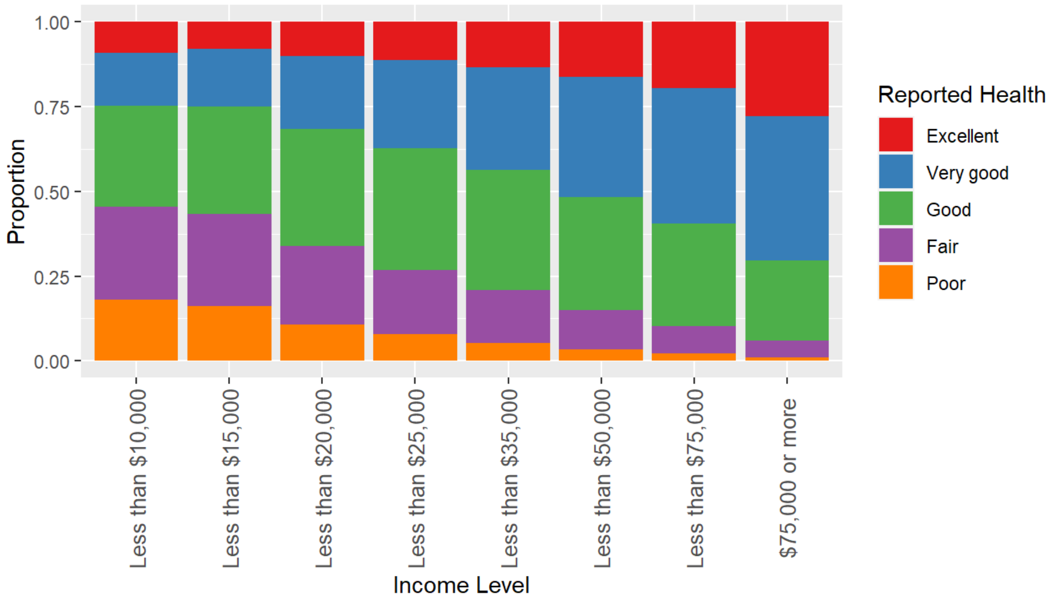
```
q2%>%
  group_by(X_rfsmok3)%>%
  summarise(Number_in_each_category=n())
```

```
## # A tibble: 2 x 2
##   X_rfsmok3 Number_in_each_category
##   <fct>                       <int>
## 1 No                         341603
## 2 Yes                         67477
```

Now lets plot the two variables income level(income2) and level of general health(genhlth) in a barplot.

```
g <- ggplot(q2) + aes(x= income2,fill=genhlth) + geom_bar(position = "fill")
g <- g + xlab("Income Level") + ylab("Proportion") + scale_fill_brewer(name="Reported Health", palette = "Set

g
```

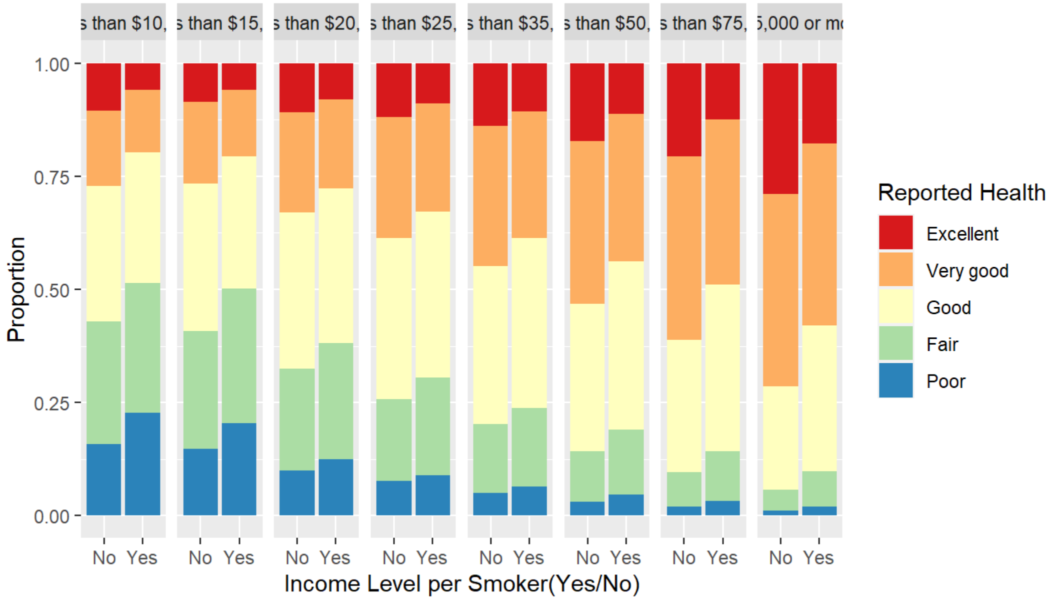


A difference in the Health condition with varying income level can be very well made out from the plot that we just generated. With increasing income level, the health conditions also improve, however we once again conclude income is a cause for good or poor health as the data collected is observational.

Let us now include the variable X_rfsmok3 in the plot and observe the changes.

```
gsmoke <- ggplot(q2) + aes(x= X_rfsmok3,fill=genhlth) + geom_bar(position = "fill") + facet_grid(.~income2)
gsmoke <- gsmoke + xlab("Income Level per Smoker(Yes/No)") + ylab("Proportion") + scale_fill_brewer(name="Repo

gsmoke
```



We clearly see a detoriation in the health conditon of smokers irrespective of the income level the person belongs in. Thus this data shows a clear assosiation of smoking with detoriation in the health condition. However we cannot conclude smoking as the cause for detoriating health as the data is observational.

Research quesion 3:

Let us select the variables of interest for our third research question and store in `q`, our variables of interest are:

- `addepev2`
- `X_bmi5cat`
- `sex`

```
q3<-select(brfss2013,addepev2,X_bmi5cat,sex)%>% na.omit()
```

Before doing anything else let us have a look at the variables of our interest.

```
q3%>%
  group_by(addepev2)%>%
  summarise(Number_in_each_category=n())
```

```
## # A tibble: 2 x 2
##   addepev2 Number_in_each_category
##   <fct>                      <int>
## 1 Yes                        91249
## 2 No                        371851
```

```
q3%>%
  group_by(X_bmi5cat)%>%
  summarise(Number_in_each_category=n())
```

```
## # A tibble: 4 x 2
##   X_bmi5cat     Number_in_each_category
##   <fct>                           <int>
## 1 Underweight                      8202
## 2 Normal weight                  154253
## 3 Overweight                     166425
## 4 Obese                          134220
```

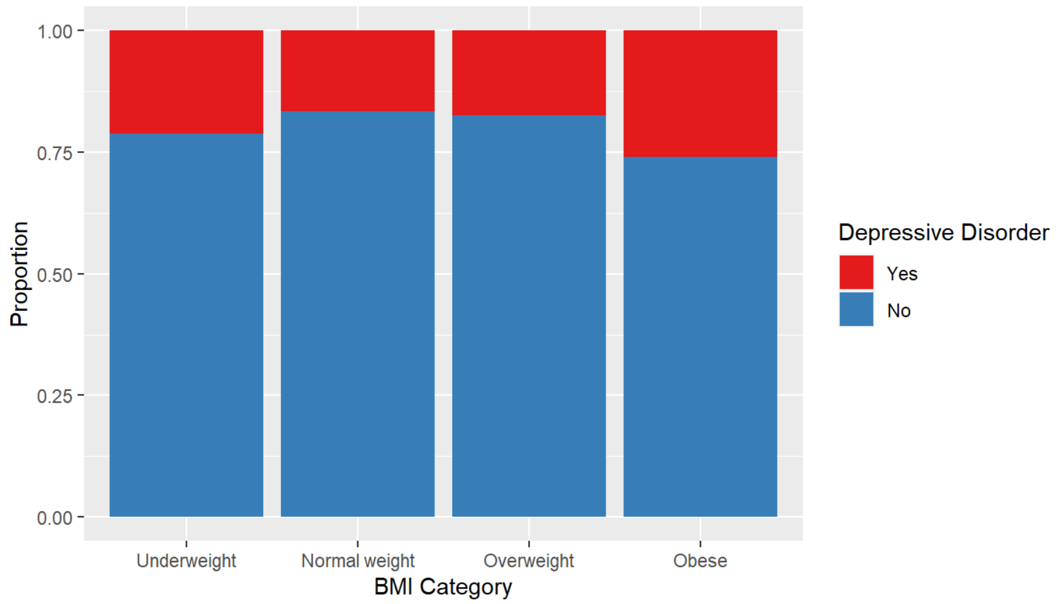
```
q3%>%
  group_by(sex)%>%
  summarise(Number_in_each_category=n())
```

```
## # A tibble: 2 x 2
##   sex    Number_in_each_category
##   <fct>                    <int>
## 1 Male                    196221
## 2 Female                  266879
```

Now lets plot the two variables BMI Category(X_bmi5cat)) and level of depression disorder(addepev2) in a barplot.

```
g <- ggplot(q3) + aes(x= X_bmi5cat,fill=addepev2) + geom_bar(position = "fill")
g <- g + xlab("BMI Category") + ylab("Proportion") + scale_fill_brewer(name="Depressive Disorder", palette = '

g
```

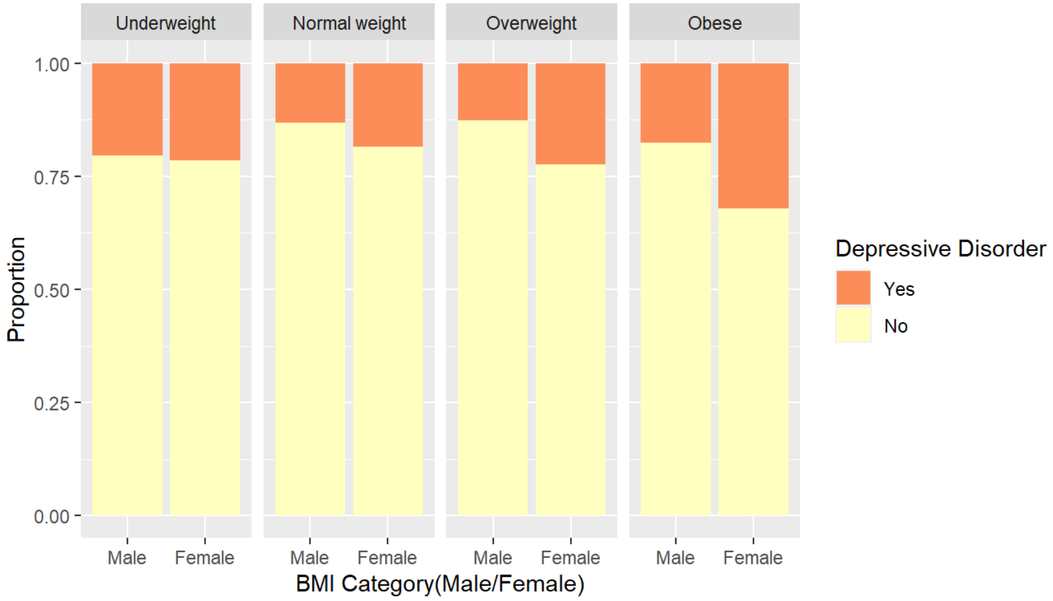


A difference in the no of people diagnosed with depression clearly varies with the BMI category the person lies in. Clearly underweight and Obese person have a larger probability to be diagnosed with some form of depression mainly because of the prevelant socail norms in the society. However the correlation between BMI and depression disorder cannot be concluded to be causal as the data obtained is observational and not experimental.

Let us now include the variable sex in the plot and observe the changes.

```
gsex <- ggplot(q3) + aes(x= sex,fill=addepev2) + geom_bar(position = "fill") + facet_grid(.~X_bmi5cat)
gsex <- gsex + xlab("BMI Category(Male/Female)") + ylab("Proportion") + scale_fill_brewer(name="Depressive Di

gsex
```



We clearly see an increase in probabilty of being diagnosed with depression for females irrespective of the BMI category the person belongs to. This is probably because of the wide spread old social norms we live in which tend to judge females based on their body type. Thus this data shows a clear assosiation of gender with increase in probability of being diagnosed with depression. However we cannot conclude gender as the cause for increase in probabiity of being diagnosed by depression, as the data is observational not experimental.