GSS-brfss2013_Statistical_Analysis_Report

Yi Wang

July 5, 2019

Introduction

This analysis aims to use two social survey dataset, GSS and brfss2013, to draw insights about poeple's health status and education level.

R is the main programming language that will be used. ggplot2 will be used for draw visualized insights, and hypothesis test will be used for make statistical inference.

Setup

Load packages

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

```
## Warning: package 'statsr' was built under R version 3.5.3
```

```
## Warning: package 'BayesFactor' was built under R version 3.5.3
```

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `brfss2013`. Delete this note when before you submit your work.

```
load("brfss2013.RData")
load("gss.Rdata")
```

Part 1: Datasets Summary

These two datasets can be seen as **highly generalize and not casual**.

For example, BRFSS conducts both landline telephone- and cellular telephone-based surveys. In the first way, data were collected from randomly selected adults, while in the second way were collected from participates in a private residence or college housing.

However, we may notice that there are several factors would affect the generalizability of data. This is an observation research processed based on phones. Thus, there could be two resources of sampling bias:

- **convenience samples**: they can only access people who are easy to reach;
- **non-response samples**: we can see that data in every category has null values from miss or unwillingness to reply.

Based on these highly possible factors, we can say that BRFSS's data from sampling is highly possible to be generalize, but is not 100% generalize.

Furthermore, since this dataset was produced through **observation**(they investigated interviewees), the assignment process was not random. This outcome is not casual.

Part 2: Exploratory data analysis

Personal financial condition can have a huge influence on personal health condition. Thus I would like to see the specific relationship between them. Health condition is composed of **physical** and **mental** condition, and the **daily lifestyle** can both reflect and reform a person's health condition.

To make a deep analysis, I would like to focus on females who are between 18 and 65 years old. The reason to excludeelder women(whose age is above 65) is that aging significantly is special life phase. I believe that these conditions need more professional knowledge to produce a proper and helpful analysis.

And to address these three topics I'm interested, there are 3 questions to discuss:

Research question 1: Do women with lower income would be more likely to have higher blood pressure?

- High blood pressure seems to be no harm, but in fact, it can generate serious problems to lots of organs over time. To specific physical condition, I would only discuss this topic.

```
women_adt <- brfss2013 %>% filter(sex == 'Female' & X_age65yr == "Age 18 to 64" )

women_adt %>%
  group_by(X_incomg, na.rm = T) %>%
  summarise(Percentage_of_High_Blood_Pressure = sum(bphigh4 %in% c('Yes', 'Yes, but female told only du
  arrange(desc(Percentage_of_High_Blood_Pressure))
```

```
## # A tibble: 6 x 3
## # Groups:   X_incomg [6]
##   X_incomg                     na.rm Percentage_of_High_Blood_Pressure
##   <fct>                        <lgl>                             <dbl>
## 1 Less than $15,000            TRUE                              0.413
## 2 $15,000 to less than $25,000 TRUE                              0.346
## 3 $25,000 to less than $35,000 TRUE                              0.327
## 4 $35,000 to less than $50,000 TRUE                              0.300
## 5 <NA>                         TRUE                              0.298
## 6 $50,000 or more              TRUE                              0.239
```

The table above clearly shows the relationship between blood pressure level and income level.

I group women based on income level and calculated the ratio of women with high blood pressure to women with valid blood pressure data. With this table, we can see that clear descending trend- the proportion of women with high blood pressure is descending when the income is ascending.

Thus, we can say that women with lower income would be more likely to have higher blood pressure, and these two factors have a negative correlation.

Research question 2: Do women with lower income would be more likely to be in bad mental condition?

- To be specific, I will use three serious negative feelings-hopeless, restless and depressed to address this question.

```
women_adt %>%
  group_by(X_incomg, na.rm = T) %>%
  summarise(Rate_of_Nervous = sum(mishopls %in% c('All', 'Most') & misrstls  %in% c('All', 'Most') & mis
  arrange(desc(Rate_of_Nervous))
```

```
## # A tibble: 6 x 3
## # Groups:   X_incomg [6]
##   X_incomg                     na.rm Rate_of_Nervous
##   <fct>                        <lgl>           <dbl>
## 1 Less than $15,000            TRUE           0.0639
## 2 $15,000 to less than $25,000 TRUE           0.0261
```

```
## 3 <NA>                           TRUE          0.0183
## 4 $25,000 to less than $35,000 TRUE          0.00911
## 5 $35,000 to less than $50,000 TRUE          0.00633
## 6 $50,000 or more                TRUE          0.00170
```

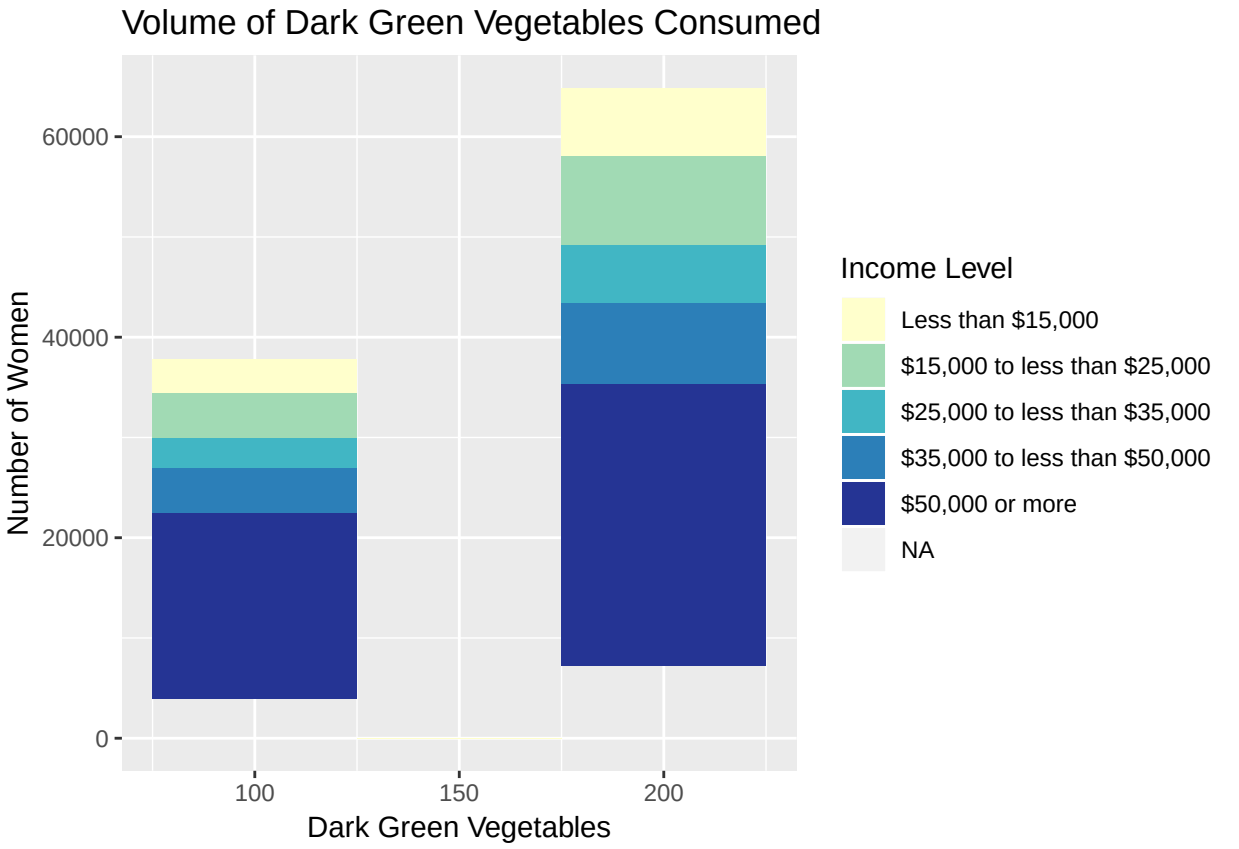
Just as the first question, the table shows a quite clear negative relationship between serious mental condition and income level of women.

I group women based on income level and calculated the ratio of women often feeling hopeless, restless and depressed (answer 'All' and 'Most' to related questions) to women with valid data. With this table, we can see that clear descending trend- the proportion of women with these feeling often is descending when the income is ascending. From the lowest income level to the highest one, the percentage even decreased by 20%.

Thus, women with lower income would be more likely to be in bad mental condition, and these two factors have a negative correlation.

Research question 3: Do women with higher income would be more likely to have a healthier eating style?

```
women_adt_veg <- women_adt %>% filter(fvgreen %in% c(101:120, 201:225))
ggplot(women_adt_veg, na.rm = TRUE,  aes(x =fvgreen , fill = X_incomg)) +
  geom_histogram(binwidth = 50) +
  labs(title = "Volume of Dark Green Vegetables Consumed",
       fill = "Income Level",
       x = "Dark Green Vegetables",
       y = "Number of Women") +
  scale_fill_manual(values = c("#ffffcc", "#a1dab4", "#41b6c4", "#2c7fb8", '#253494'))
```



The label '100' on axis X means eating dark green vegetables every day, and the label '200' means every week. We use the amount of dark green vegetables consumed to evaluate the quality of lifestyle. This histogram

only includes data of women eat dark green vegetables every day and every week. As we can see, women with income 50,000 or more consume the most vegetables, and roughly the more income women have, the more vegetable would be consumed.

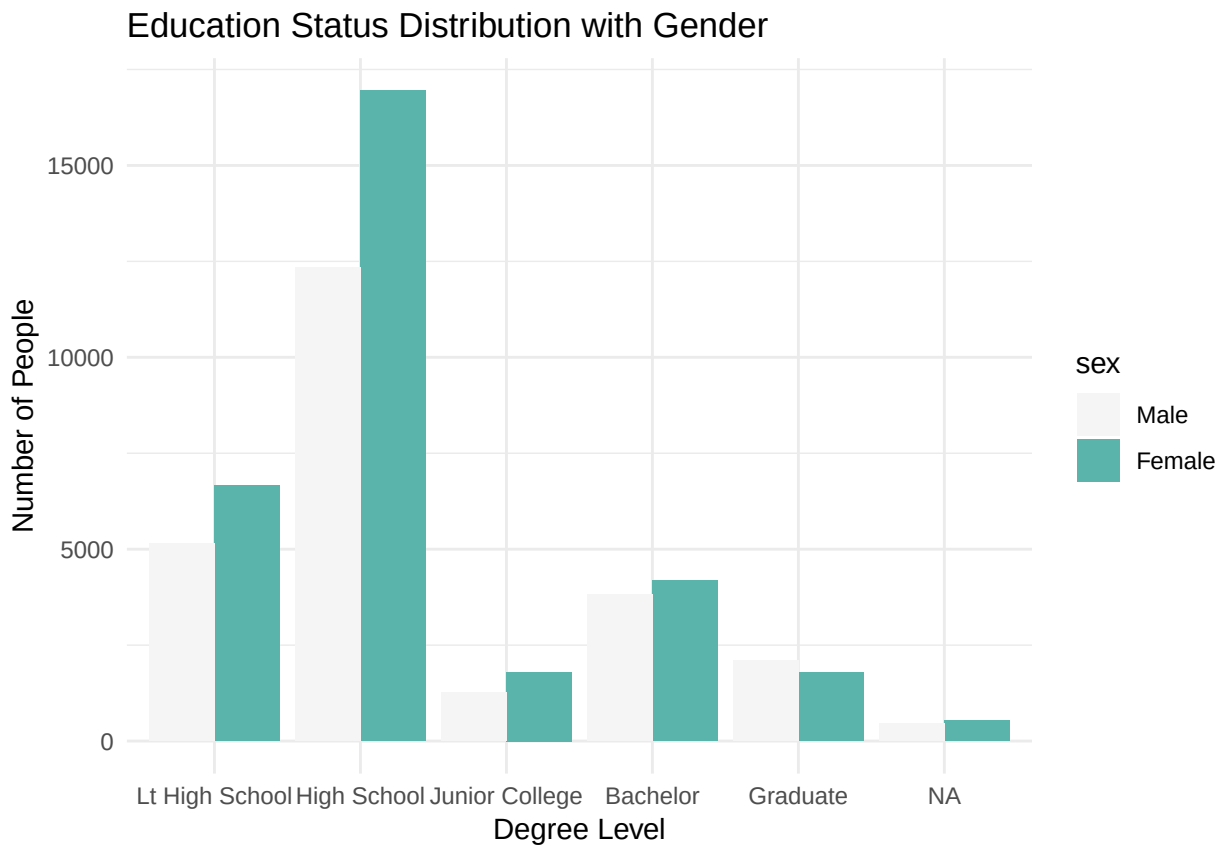
Thus, from the perspective of eating, we can say that women with higher income would be more likely to have a healthier eating style.

Hypothesis test

First, I would like to extract data needed for this research from all and visualize it to have a general understanding.

```
ggplot(gss, aes(x = degree, fill = sex)) +
  geom_histogram(stat = "count",  position="dodge") +
  labs(title = "Education Status Distribution with Gender",
      x = "Degree Level",
      y = "Number of People") +
  guides(fill = guide_legend(label = TRUE)) +
  theme_minimal()+
  scale_fill_manual(values = c("#f5f5f5", "#5ab4ac"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



From this plot above, we can see that both men and women has the same ranking about degrees. For men and women, porpotion of degrees from high to low is as follows: - Hish School - Lt High School - Bachelor - Graduate - Junior College

A slightly important trend is, while male in all other categorys has less poeple than female, male has more people in category master, which is the highest degree in this reserach. Now we have a general trend, but to

come to an conclution, we need further inference test.

Part 4: Hypothesis Test

Research question : Are there more men acquired college education than women?

Further more, I'd like to draw a hypothesis test on this problem.

```
fe_ <- gss %>% filter(sex == 'Female')%>% nrow()
ma_ <- gss %>% filter(sex == 'Male') %>% nrow()
fe_masba <- gss %>% filter(sex == 'Female' & degree %in% c("Bachelor", "Master"))%>% nrow ()
ma_masba <- gss %>% filter(sex == 'Male' & degree %in% c("Bachelor", "Master")) %>% nrow ()

fe_hat <- fe_masba/fe_ %>% round(digits=2)
ma_hat <- ma_masba/ma_ %>% round(digits=2)

degree_rate <- data.frame(fe_hat, ma_hat)
names(degree_rate) <- c("Female high degree  rate", "Male high degree rate")
degree_rate
```

```
##   Female high degree  rate Male high degree rate
## 1               0.1309729             0.1519924
```

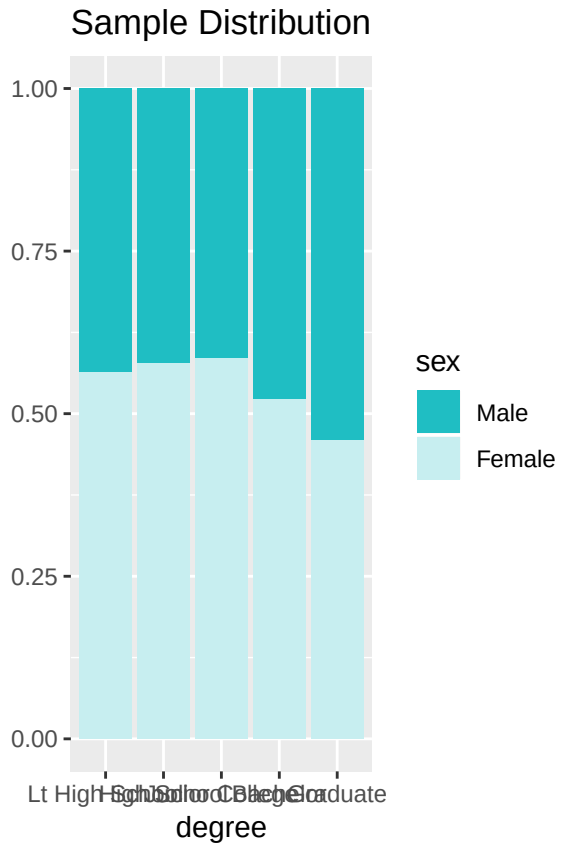
From this chart we can see that there is a slightly difference in high degree(bachelor, master) bewteen male and female. Does this significant enough to say that there is difference? We need further test.

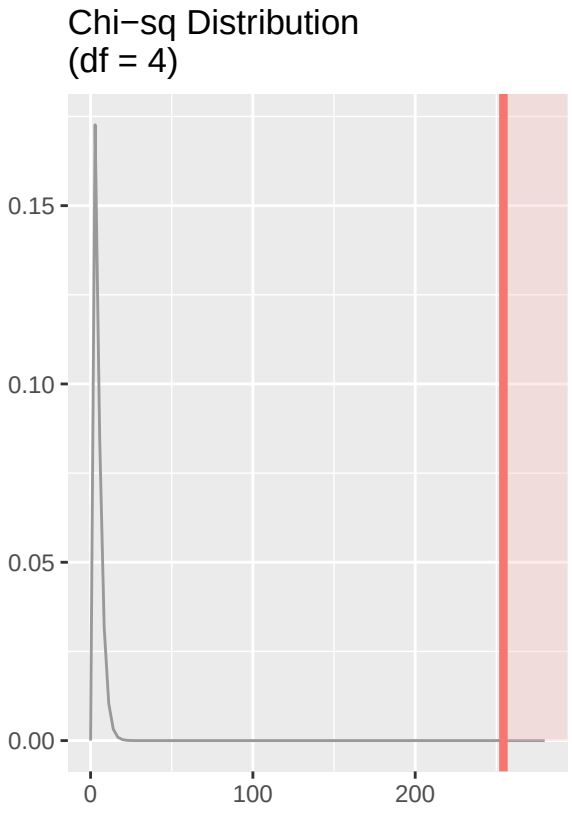
Further more I'd like to draw a Chi-Quare Independence Test. This test will compare the hig degree rate between male and female and report on possible inequality.

```
inference(y=sex, x= degree,
         data = filter(gss, !is.na(degree)),
          type = "ht", statistic = "proportion",
          success = c("Bachelor","Master"),
          method = "theoretical", alternative = "greater",
          sig_level = 0.05, conf_level = 0.95)
```

```
## Response variable: categorical (2 levels)
## Explanatory variable: categorical (5 levels)
## Observed:
##                 y
## x                 Male Female
##   Lt High School  5153   6669
##   High School    12340  16947
##   Junior College  1272   1798
##   Bachelor        3822   4180
##   Graduate        2091   1779
##
## Expected:
##                 y
## x                      Male    Female
##   Lt High School  5204.962  6617.038
##   High School    12894.410 16392.590
##   Junior College  1351.652  1718.348
##   Bachelor        3523.101  4478.899
##   Graduate        1703.874  2166.126
##
```

```
## H0: degree and sex are independent
## HA: degree and sex are dependent
## chi_sq = 254.3489, df = 4, p_value = 0
```





As we can see from the Chi-squared test result, p-value is near 0. If degree and sex are dependent, women with bachelor and graduate degree will be much more than reality. With the chi_sq value and p value, there is a significant difference bewteen male and female.

So, we can reject the H0 and say that **degree and sex are dependent**.