Statistical inference with the GSS data

Setup

Load packages

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

Load data

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

Part 1: Data

From the subset variable lists of Abstract for the GSS Cumulative File, in 1972-2012, we can see GSS questions cover a diverse range of issues, including national spending priorities, marijuana use, crime and punishment, race relations, quality of life, confidence in institutions, and sexual behavior. Those variables and data are randomly sampled, considering the great variety of respondents. However, we are unknown about whether the whole assignment is randomly arranged. Thus, it is fair to say this experiment is generalized but not causal.

Part 2: Research question

Education plays a fairly significant role in our evolutionary process, and much emphasis has been laid on, especially in our modern days. However, with education becoming more and more industrialized, we often hear about the news that many kids have to drop out of school due to the overwhelming burden education has imposed on their families. Out of this primary concern, we may want to explore the relationship between respondents' education level and their family income levels (Specifically, when respondents were 16 years old).

Research question: The relationship between respondents' education levels and their family income levels.

Part 3: Exploratory data analysis

Step 1

We can use summarize functions to have a general understanding of the respondents' education duration time. Remember to exclude those NAs by using filter().

```
gss %>%
  filter(!is.na(educ)) %>%
  summarise(educmean = mean(educ), educmedian = median(educ), educsd = sd(educ), mineduc = min(educ), ma
```

```
##   educmean educmedian   educsd mineduc maxeduc
## 1 12.75359         12 3.181642       0      20
```

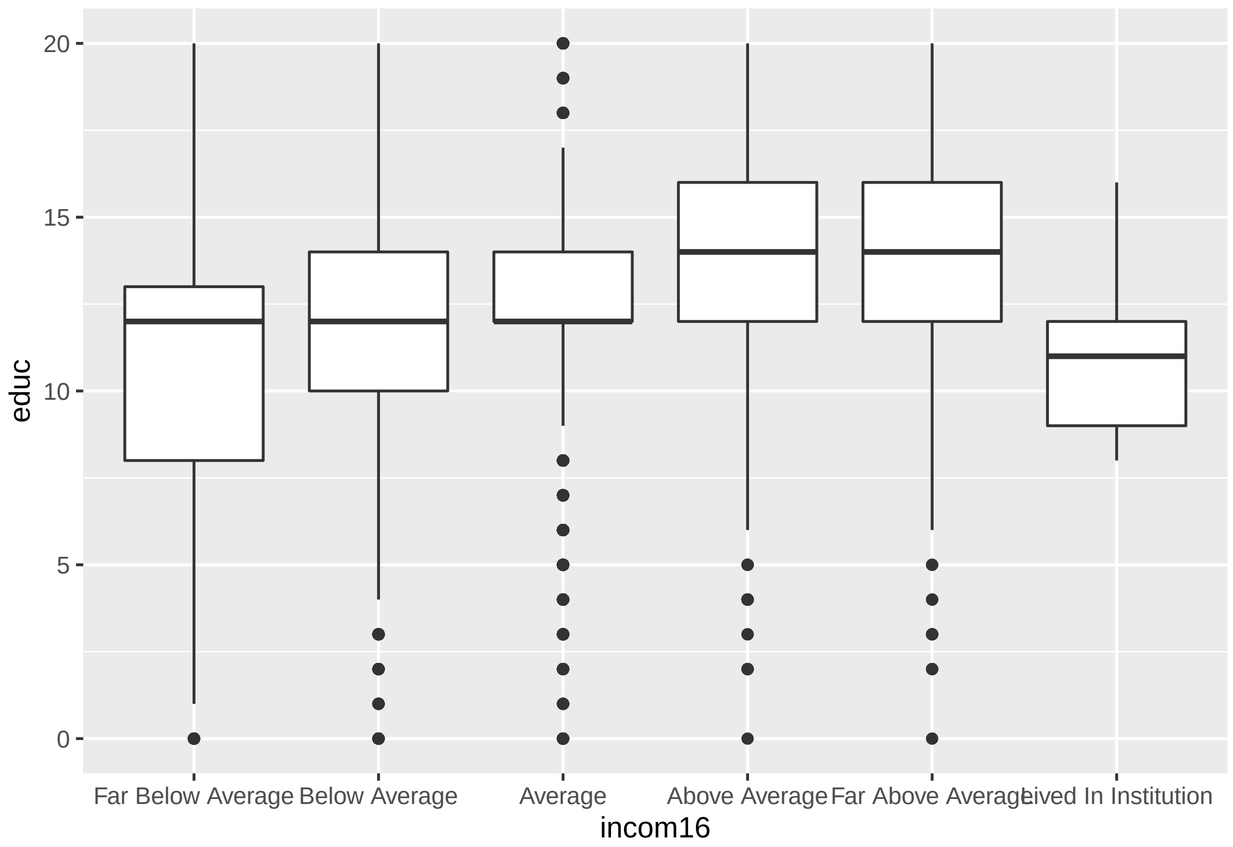
From the result, it's clear that the median of respondents' education time is relatively lower than the mean one, which indicates the distribution of respondents' education time is right-skewed - a large number of extreme positive observations included in this survey.

Step 2

To visualize the two variables better, we use the boxplot to explore the relationship between respondents' education levels and their family income levels. Also, we need to use filter() to exclude NAs in the variable "incom16".

```
gss %>%
  filter(!is.na(incom16)) %>%
  ggplot(gss, mapping = aes(x = incom16, y = educ)) +
  geom_boxplot()
```

```
## Warning: Removed 91 rows containing non-finite values (stat_boxplot).
```



From this box plot, generally, we can tell that the overall distribution of people's education duration time varying from their family income levels is left-skewed. This distribution indicates a few positive variables, and less than 50% of the education duration time of respondents from different home income levels are smaller than the average one. This survey is good news to identify those significant respondents not left behind due to their family's economic problems.

Specifically, the distribution of the education duration time of respondents from an average-income-level family is more uniform. In contrast, the distribution of those who come from a far-below-level family is more variable and lower than other groups. Hopefully, society is paying attention to bring more educational opportunities for those relatively impoverished people to chase their dreams and realize their values.

Part 4: Inference

Step 1: State hypotheses

According to the EDA process we have conducted above, we have already had a preliminary understanding of the relationship between the respondents' education duration time and their family income levels. To focus on a specific group, we choose respondents who have received education higher than the average level (12 years) to estimate the overall education level. Firstly we need filter() to sort out those who satisfy our research goal.

```
gss %>%
  filter(educ >= 12) %>%
  summarise(count = n())
```

```
##   count
## 1 43484
```

Here we get the result that the total number of respondents who qualify our goal is 43,484, now we use dim() to get to the total number of respondents who have attended this survey.

```
dim(gss)
```

```
## [1] 57061   114
```

According to the result, now we know that this data set has 57,061 rows and 114 columns, which also indicate 57,061 respondents and 114 variables. Here we use R to calculate the proportion of our goal respondents in this survey.

```
43484/56879
```

```
## [1] 0.7645001
```

Note: Here, we changed 57,061 to 56,879 based on the result from the following inference process. To eliminate unnecessary deviations, we changed it purposedly to be more accurate and convincing.

Therefore, we get to know that the GSS survey has found that roughly 76% of 56,879 randomly sampled Americans have received education for more than 12 years. Does this provide convincing evidence that most Americans have also received an education that lasts continuously for more than 12 years? As a result of this, we state our hypotheses as below:

H0: $p = 0.5$

HA: $p > 0.5$

Step 2: Check conditions

1. Independence:

The total number of our respondents, 56,879, is less than 10% of Americans as its population. And according to our EDA process, we have already known that the samples are randomly assigned. Therefore, whether one American in the sample has had education higher than the average level is independent of another.

2. Sample size / skew :

To evaluate the success condition:

```
56879*0.5
```

```
## [1] 28439.5
```

Clearly this result is much greater than 10. Therefore, the success condition is met.

To evaluate the failure condition:

```r
56879*(1-0.5)
```

```
## [1] 28439.5
```

Still, this result is much greater than ten as well. Therefore, the S-F condition is met, which means this sample is a nearly normal sampling distribution.

All in all, the requirements of independence and sample size / skew-ness are met.

Step 3: State the methods to be used and why and how.

From the box plot of our EDA process and the p-hat (roughly 76%), we can visualize the proportion here is not small. We won't use the simulation method mainly applicable to small balances. Here we would like to conduct a hypothesis test and report the associated p-value and the conclusion. Also, we would calculate the confidence interval for the parameter of interest.

Step 4: Perform inference

H0: $p = 0.5$, HA: $p > 0.5$

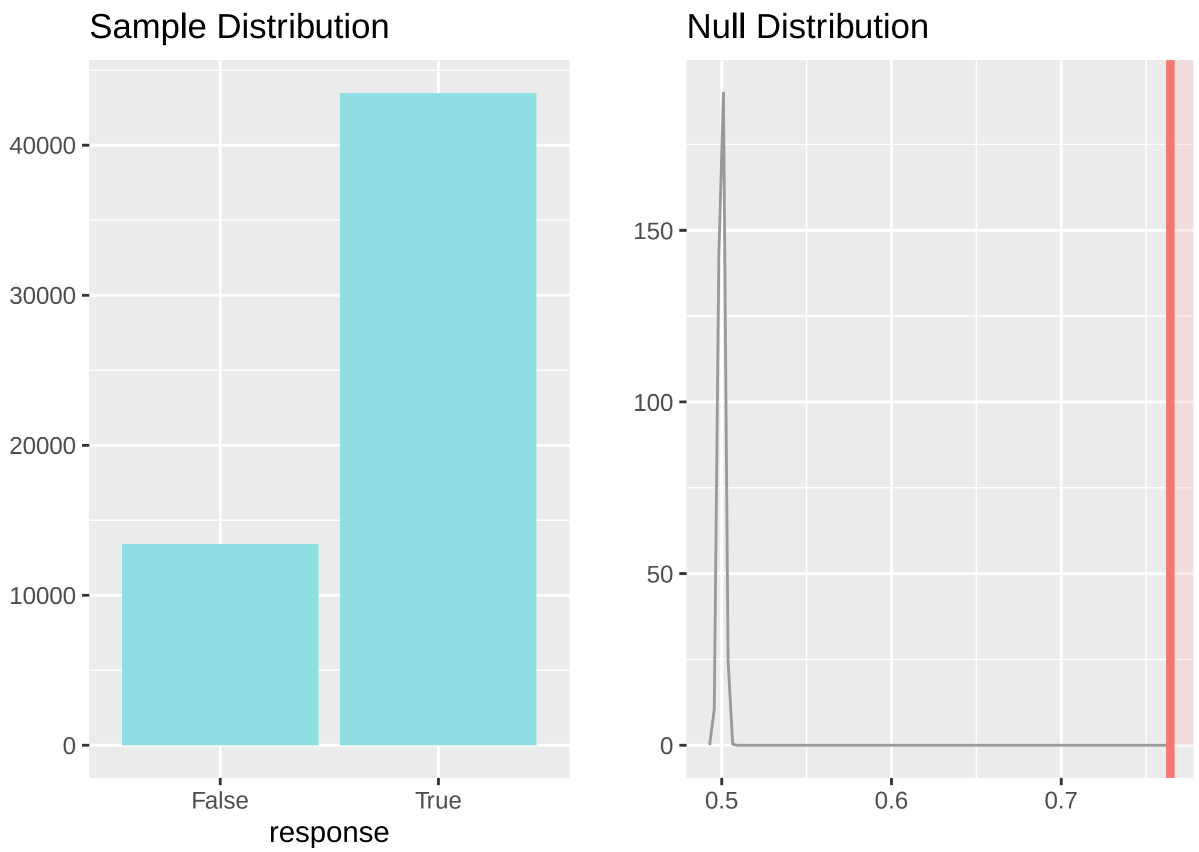
Here we want to set up a new variable in the data set "gss" to conduct our inference function; respondents whose education time is longer than 12 years are marked as "True," otherwise marked as "False."

```r
gss <- gss %>%
  mutate(response = ifelse(educ>=12,"True","False"))
```

Now we can get down to our inference test by using the inference function. Note that this test is not a two-sided one. We need to choose alternative = "greater."

```r
inference( y=response, data = gss, statistic = "proportion", type = "ht", null = 0.5,
           alternative = "greater", method = "theoretical", success = "True")
```

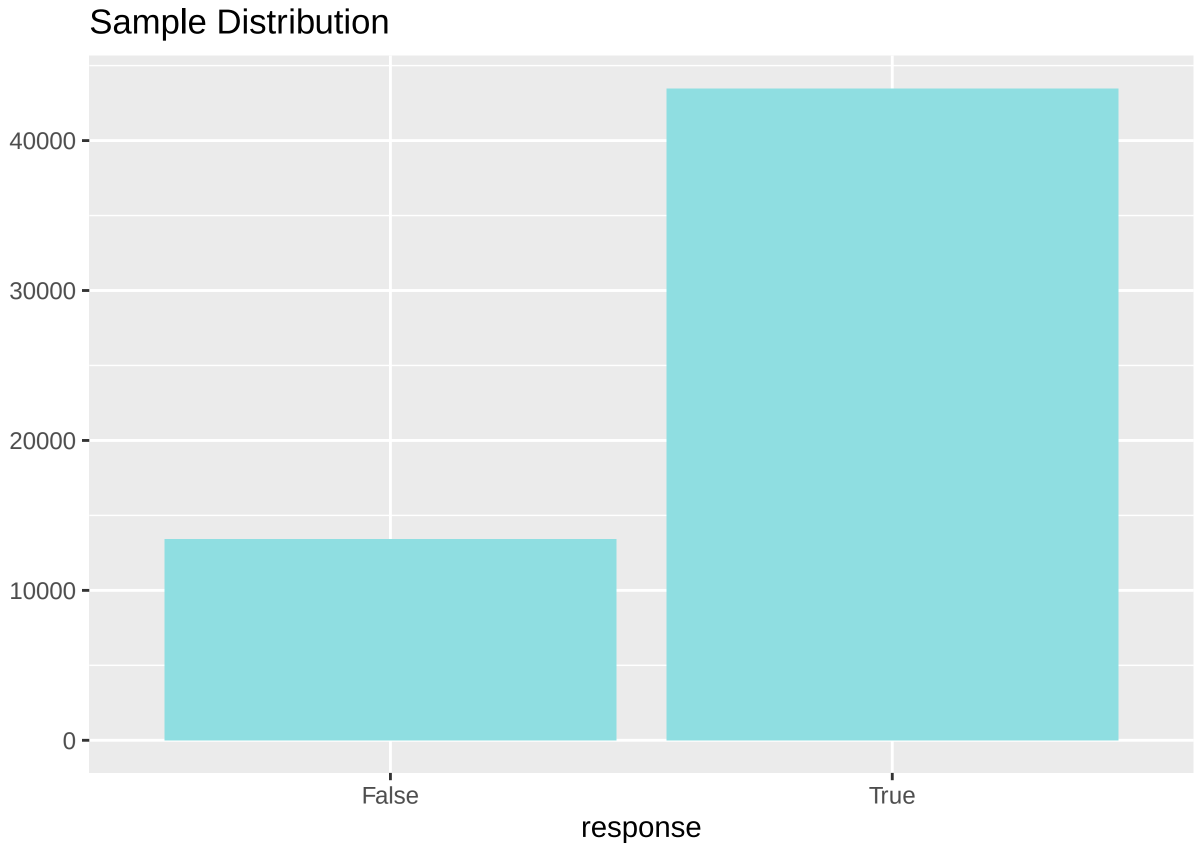
```
## Single categorical variable, success: True
## n = 56897, p-hat = 0.7643
## H0: p = 0.5
## HA: p > 0.5
## z = 148.5029
## p_value = < 0.0001
```



Besides the hypotheses test result, we also need the inference function to get to know the confidence intervals of our test:

```
inference(y= response, data = gss, statistic = "proportion", type = "ci",
          method = "theoretical",
          success = "True")
```

```
## Single categorical variable, success: True
## n = 56897, p-hat = 0.7643
## 95% CI: (0.7608 , 0.7677)
```



Now we've got all we need, the CI and p-value will help us better interpret the test results.

Step 5: Interpret results

Since the p-value is less than 0.0001, the conclusion is to reject the H0: There is almost 0% chance of obtaining a random sample of 56,897 where roughly 46% of people's education duration time is 12 years or more if, in fact, 50% Americans have received education for 12 years or more.

As for the CI's interpretation, we can conclude that 95% of random samples of 56,897 Americans will yield confidence intervals (45.27%,46.09%) that contain the proper proportion of Americans who have received education for 12 years or more. Namely, we are 95% confident that 45.27% to 46.09% of all Americans have had access to education for 12 years or more.

There's an urgent need to promote the popularization of higher education in this country. The government and society are also expected to pay more attention to dropout problems due to family economic crisis or the scarcity of higher education access opportunities. As the foundation of a democratic society, education would be the best and final prevention against social evils.