Statistical inference with the GSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(data.table)
```

Load data

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

Part 1: Data

All data were collected from random surveys through the United States, with missing values being removed. As random sampling were used, the data should be able to be generalized to the whole US population. However, as certain questions might be sensitive to fairly large population, thus the answers to those questions tend to be missing and the corresponding variables in the data set getting removed, the data here could be potentially biased to miss out certain areas of interest.

As no random assignment was performed, no causality can be formed from researches based on this data set. And the research from mine below certainly will not make any cause and result conclusions.

Part 2: Research question

Interest: As we know the environment a person grew up in could affect him/her later in life, and could in turn have some social impact, it would be interesting to find out how a person’s background might affect his/her financial satisfaction later.

Research Question:Is there relationship between how a person feels about his/how current financial situation and how his/her family’s financial position was when he/she was 16?

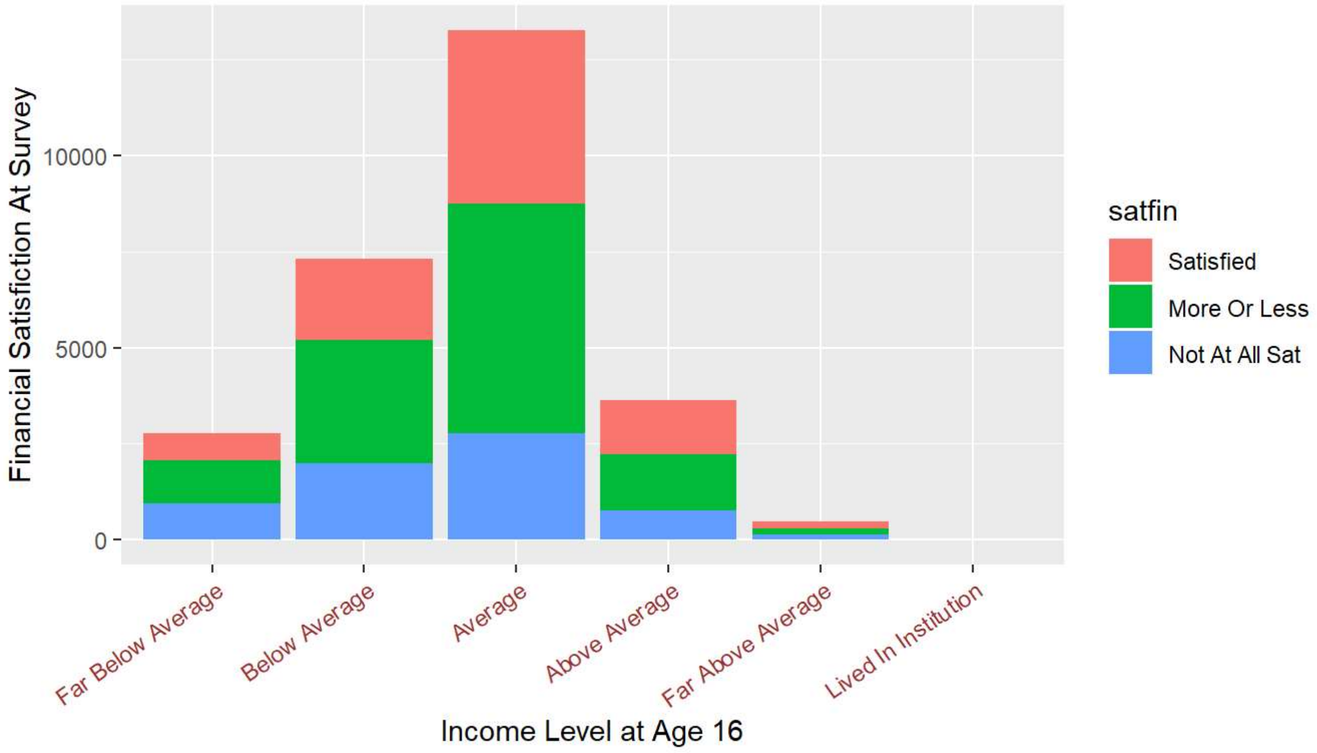
Part 3: Exploratory data analysis

My data and data processing:

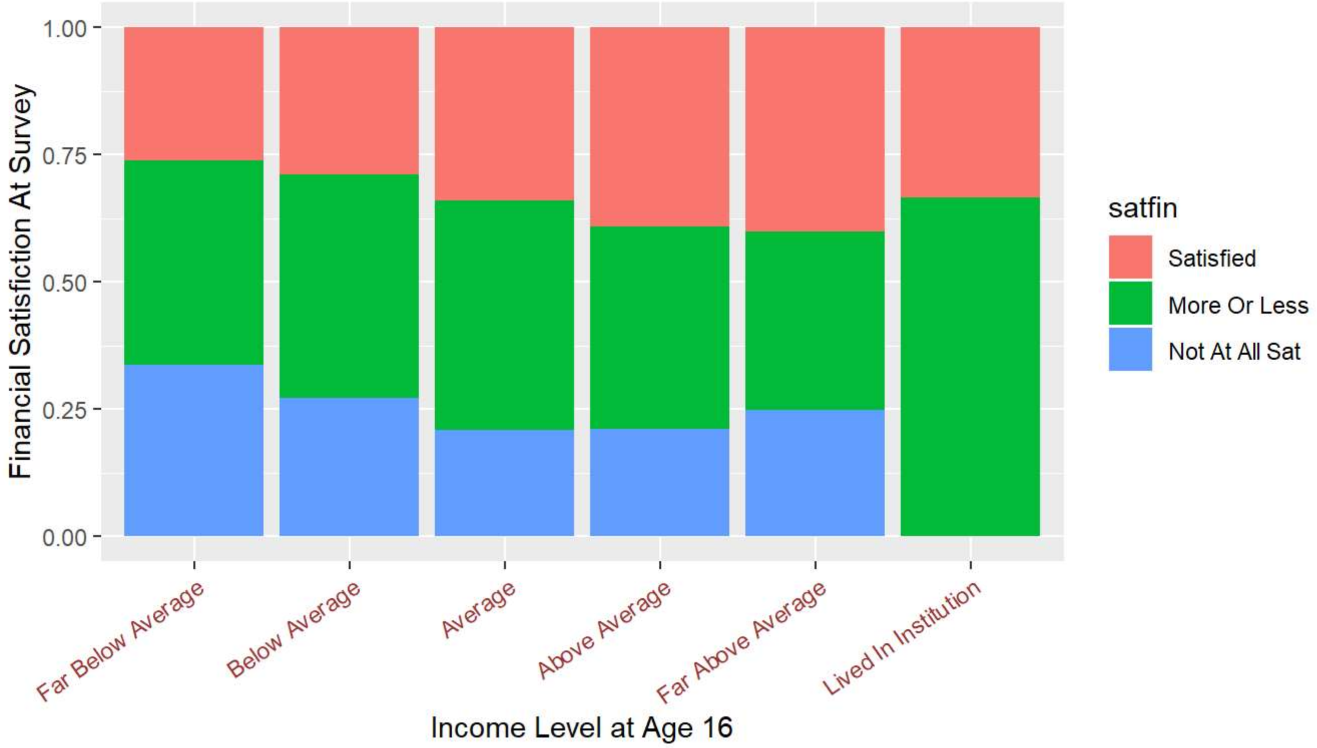
In early life stage (age 18 - 34), a person, in general, has not been settled down yet, and his/her financial situation could change dramatically one month after the survey. This variability could hurt the reliability of data, specifically how satisfied a person is about his/her current financial situation. Therefore, to eliminate its effect, I filtered for people over 35 years old in all my research data below, to get a more consistent satisfaction expression in a more stabled life stage.

incom16, representing family financial situation at age of 16, satfin, representing satisfaction level about current financial situation, and age, are the variables that we focus on in this research.

```
gss%>%
  filter(!is.na(incom16),!is.na(satfin), age >= 35)%>%
  ggplot(aes(x=incom16, fill = satfin)) +
  geom_bar() +
  theme(axis.text.x = element_text(color = "#993333", angle = 35, hjust = 1)) +
  labs(
    x = "Income Level at Age 16",
    y= "Financial Satisfiction At Survey"
  )
```



```
gss%>%
  filter(!is.na(incom16),!is.na(satfin), age >= 35)%>%
  ggplot(aes(x=incom16)) +
  geom_bar(aes(fill = satfin), position = "fill") +
  theme(axis.text.x = element_text(color = "#993333", angle = 35, hjust = 1)) +
  labs(
    x = "Income Level at Age 16",
    y= "Financial Satisfiction At Survey"
  )
```



From the bar count plot and the bar percentage plot above, We can see differences (both in proportion and in absolute counts) among groups with different family financial background at 16. A very clear trend is that, as family background moved from Far Below Average to Far Above, the percentage of people satisfied about their current financial situation gradually increased. The More Or Less Satisfied percentage increased, maxed at the group with Average family, then decreased. Whereas, the Not At All Satisfied percentage went the exact opposite way, decrease, minimized at the group with Average family, then increased back up. This changes indicates that the more extreme of the family background, the more extreme that group of people could feel about their current financial situation.

But are any of the differences so statistically significant that we can conclude about a clear relationship between feeling about someone's current financial position and his/her family financial status at 16?

We also could see that we had unequal sample sizes among groups, fortunately, Chi-square test is set exactly to fit in this circumstance.

Part 4: Inference

```
# to create another data table that contains only people this research is interested in
fingss=subset(gss, !is.na(incom16)&!is.na(satfin)& age >= 35)
```

```
# to create a contengency table
fin16.table<-table(fingss$incom16,fingss$satfin)
addmargins(fin16.table)
```

```
##
##                         Satisfied More Or Less Not At All Sat   Sum
##   Far Below Average           723         1121            934  2778
##   Below Average              2110         3203           1988  7301
##   Average                    4505         5971           2777 13253
##   Above Average              1424         1439            769  3632
##   Far Above Average           192          168            119   479
##   Lived In Institution          2            4              0     6
##   Sum                        8956        11906           6587 27449
```

```
                                                        Contengency table 1
```

Condition check for all the tests below:

1. independence within and between groups: As the survey was randomly taken, and it was unlikely someone would report to fall into two different financial status at 16, and our sample sizes in any group would be less than 10% of total US population falling into that group, we could reasonably believe that all surveyors were independent from each other, within or between groups.
2. Sample sizes/skewness: Per Contengency table 1, except for the Lived In Institution group, we had fairly large sample sizes in all groups, and had over 10 people in any of the three satisfaction levels (easily met the 10-successes-and-failures condition) , therefore this condition was met by all but the LII group, which would be excluded from our research anyway due to comparability.

Two conditions were both met, we will be able to proceed to tests below (chi-square test, and z score test through confidence interval and hypothesis testing).

```
chisq.test(fin16.table)
```

```
## Warning in chisq.test(fin16.table): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fin16.table
## X-squared = 370.97, df = 10, p-value < 2.2e-16
```

H0: Satisfaction level about current financial situation has nothing to do with the surveyor's family financial position when he/she was 16. Therefore we would expect the distribution of satisfaction status to be roughly equal among different family financial background.

HA: Satisfaction level about current financial situation is somewhat dependent on the surveyor's family financial position when he/she was 16.

The X-square is huge and p-value is almost zero. This means that there were significant differences among the data, and hence we would reject H0 and do some further dig in.

```
# to create data frame for being far below average and below average at 16 in order to compare t
  he two
# groups using chi-square function
df_fbvsba <- setNames(data.frame(matrix(ncol = 3, nrow = 2)),c("Satisfied", "More Or Less", "Not
  At All"))
row.names(df_fbvsba) = c("Far Below Average", "Below Average")
df_fbvsba[1,1] = 723
df_fbvsba[1,2] = 1121
df_fbvsba[1,3] = 934
df_fbvsba[2,1] = 2110
df_fbvsba[2,2] = 3203
df_fbvsba[2,3] = 1988
```

```
# chi-square test to see if one's feeling about his/her job is linked to if his/her family was f
  ar
# below average or below average in financial position when he/she was 16
chisq.test(df_fbvsba)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_fbvsba
## X-squared = 40.081, df = 2, p-value = 1.979e-09
```

H0: Satisfaction level about current financial situation has nothing to do with the surveyor's family financial position when he/she was 16. Therefore we would expect the distribution of satisfaction status to be roughly equal between the Far Below Average group and the Below Average group.

HA: Satisfaction level about current financial situation is somewhat dependent on the surveyor's family financial position when he/she was 16, specifically Far Below Average or Below Average.

The X-square is big and p-value is tiny. This means that there were significant differences among the data, and hence we would reject H0, and be 95% confident that how people feel about their current financial situation is related to their family background at 16, specifically, from a far below average family or a below average family.

```
# to create data frame for being below average and at average at 16 in order to compare
df_bavsav <- setNames(data.frame(matrix(ncol = 3, nrow = 2)),c("Satisfied", "More Or Less", "Not
  At All"))
row.names(df_fbvsba) = c("Below Average", "Average")
df_bavsav[1,1] = 2110
df_bavsav[1,2] = 3203
df_bavsav[1,3] = 1988
df_bavsav[2,1] = 4505
df_bavsav[2,2] = 5971
df_bavsav[2,3] = 2777
```

```
# chi-square test to see if one's feeling about his/her job is linked to if his/her family was
# below average or at average in financial position when he/she was 16
chisq.test(df_bavsav)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_bavsav
## X-squared = 119.37, df = 2, p-value < 2.2e-16
```

H0: Satisfaction level about current financial situation has nothing to do with the surveyor’s family financial position when he/she was 16. Therefore we would expect the distribution of satisfaction status to be roughly equal between the Below Average group and the Average group.

HA: Satisfaction level about current financial situation is somewhat dependent on the surveyor’s family financial position when he/she was 16, specifically Below Average or Average.

The X-square is huge and p-value is almost zero. This means that there were significant differences among the data, and hence we would reject H0, and be 95% confident that how people feel about their current financial situation is related to their family background at 16, specifically, from a below average family or an average family. Moreover, the difference between this group is a lot more significant than the difference between the Below Average group and the Average group.

```
# to create data frame for being at average and above average at 16 in order to compare
df_avvsaa <- setNames(data.frame(matrix(ncol = 3, nrow = 2)),c("Satisfied", "More Or Less", "Not
  At All"))
row.names(df_fbvsba) = c("Average", "Above Average")
df_avvsaa[1,1] = 4505
df_avvsaa[1,2] = 5971
df_avvsaa[1,3] = 2777
df_avvsaa[2,1] = 1424
df_avvsaa[2,2] = 1439
df_avvsaa[2,3] = 769
```

```
# chi-square test to see if one's feeling about his/her job is linked to if his/her family was
# at average or above average in financial position when he/she was 16
chisq.test(df_avvsaa)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_avvsaa
## X-squared = 41.323, df = 2, p-value = 1.064e-09
```

H0: Satisfaction level about current financial situation has nothing to do with what surveyor’s family financial position when he/she was 16. Therefore we would expect the distribution of satisfaction status to be roughly equal between the Average group and the Above Average group.

HA: Satisfaction level about current financial situation is somewhat dependent on the surveyor’s family financial position when he/she was 16, specifically at Average or Above Average.

The X-square is big and p-value is tiny. This means that there were significant differences among the data, and hence we would reject H0, and be 95% confident that how people feel about their current financial situation is related to their family background at 16, specifically, from an average family or an above average family.

```
# to create data frame for being above average and far above average at 16 in order to compare
df_aavsfa <- setNames(data.frame(matrix(ncol = 3, nrow = 2)),c("Satisfied", "More Or Less", "Not
  At All"))
row.names(df_fbvsba) = c("Above Average", "Far Above Average")
df_aavsfa[1,1] = 1424
df_aavsfa[1,2] = 1439
df_aavsfa[1,3] = 769
df_aavsfa[2,1] = 192
df_aavsfa[2,2] = 168
df_aavsfa[2,3] = 119
```

```
# chi-square test to see if one's feeling about his/her job is linked to if his/her family was
# above average or far above average in financial position when he/she was 16
chisq.test(df_aavsfa)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_aavsfa
## X-squared = 4.9605, df = 2, p-value = 0.08372
```

H0: Satisfaction level about current financial situation has nothing to do with the surveyor’s family financial position when he/she was 16. Therefore we would expect the distribution of satisfaction status to be roughly equal between the Above Average group and the Far Above Average group.

HA: Satisfaction level about current financial situation is somewhat dependent on the surveyor’s family financial position when he/she was 16, specifically Above Average or Far Above Average.

X-square is somewhat small, and p-value > 0.05, therefore, the test failed to reject H0. That being said, we are 95% confident that there is no significant differences as to feelings of current financial position between people whose family was above average and whose family was far above average at age of 16.

Further research:

We will further investigate the three pairs of comparisons where significant differences were noticed. As what could be a social problem is that people do not feel satisfied about their current financial situation at all, but satisfied or more or less satisfied would not make such a difference, we would further group those two into one level called “Not Unsatisfied”, and compare the new combined group to the unsatisfied group.

Method used in further research:

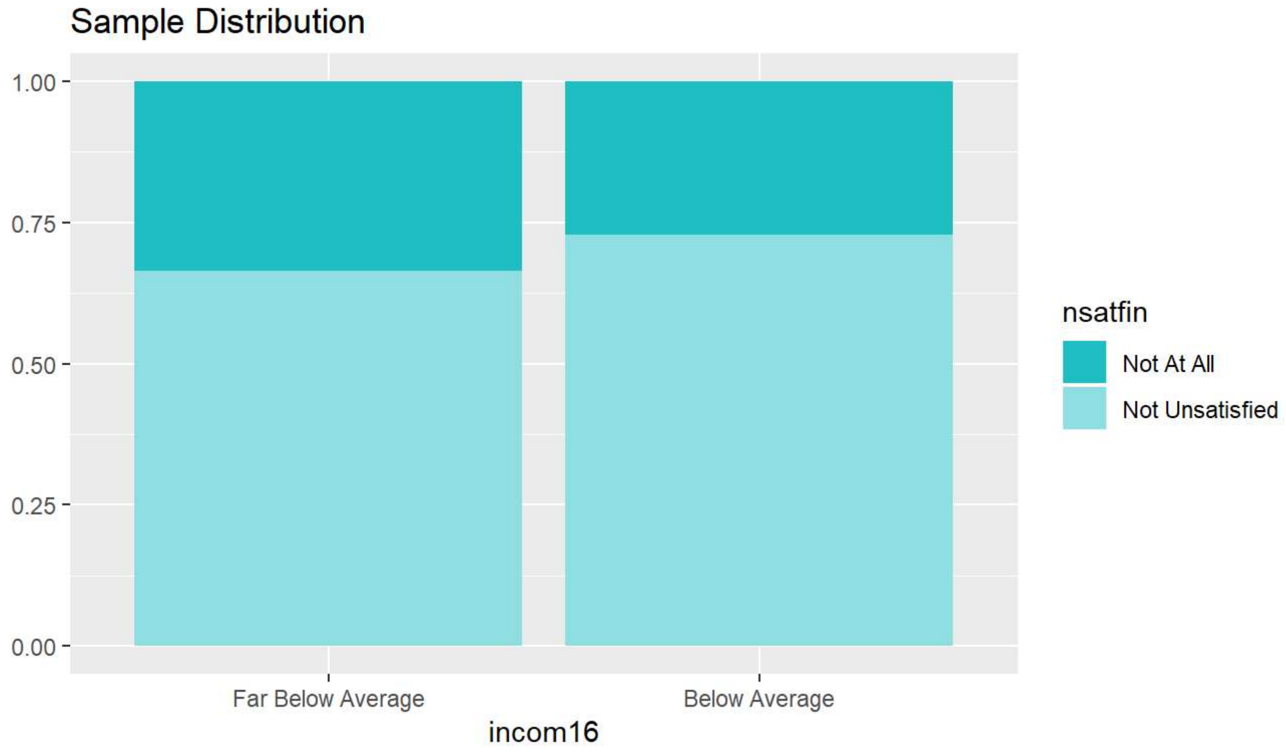
By further grouping categorial variable levels, and by comparing only two family background levels at a time, we are able to eliminate both x and y variables to only two levels, and therefore, we will be able to form confidence interval and use z score hypothesis testing to test potential differences between two categorical variables’ proportion expressions.

```
# to create a new variable
fingss<-fingss%>%
  mutate(nsatfin = ifelse(satfin == "Satisfied" | satfin == "More Or Less", "Not Unsatisfied",
  "Not At All"))
```

```
# to create a subset table with the new variable and the data we are interested in
fingss_fbvsba=subset(fingss, fingss$incom16 == "Far Below Average" | fingss$incom16 == "Below Av
  erage")
fingss_fbvsba$incom16<-factor(fingss_fbvsba$incom16)
```

```
# create inference
inference(data = fingss_fbvsba, y = nsatfin, x = incom16, type = "ci",
          statistic = "proportion", method = "theoretical",
          success = "Not At All", alternative = "twosided", conf_level = 0.95)
```

```
## Response variable: categorical (2 levels, success: Not At All)
## Explanatory variable: categorical (2 levels)
## n_Far Below Average = 2778, p_hat_Far Below Average = 0.3362
## n_Below Average = 7301, p_hat_Below Average = 0.2723
## 95% CI (Far Below Average - Below Average): (0.0436 , 0.0842)
```

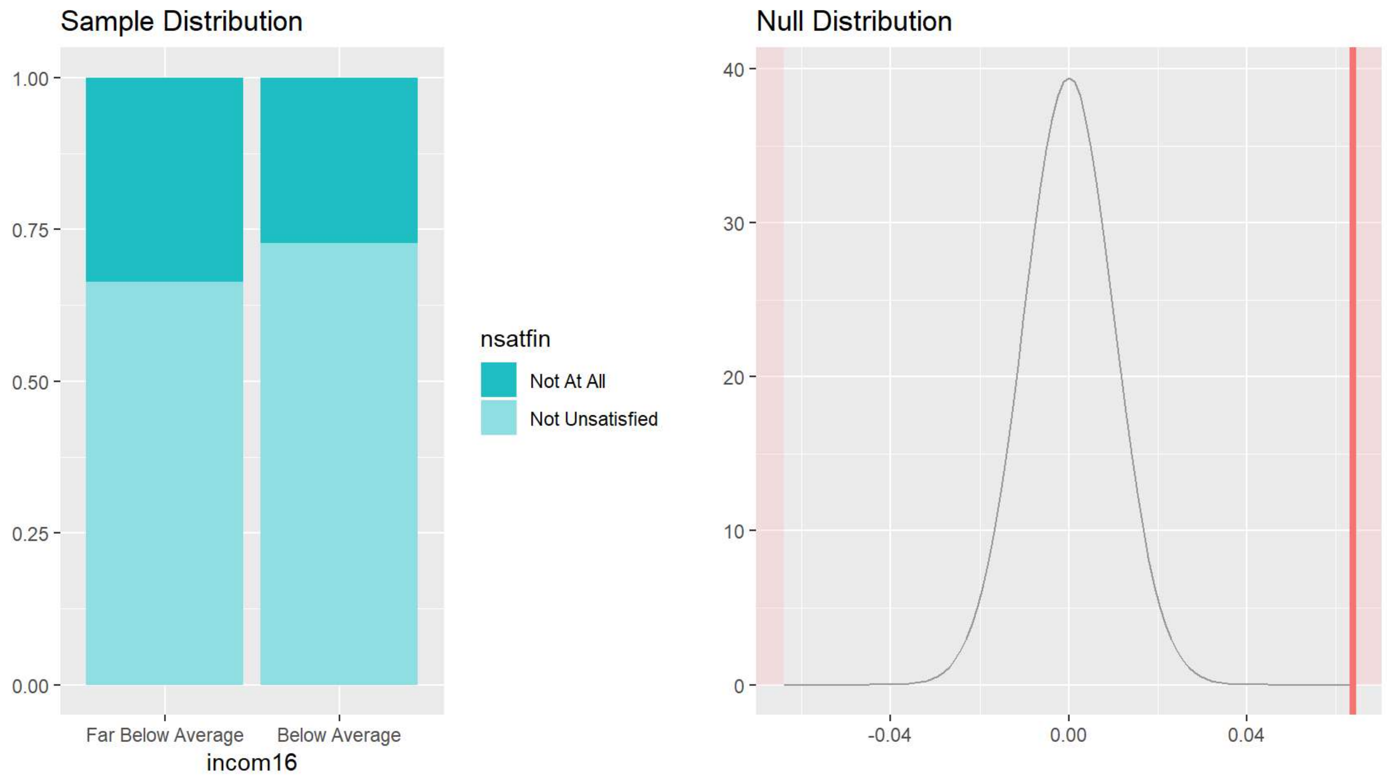


Confidence Interval Method Result Interpretation:

The 95% confidence interval for the differences between the Far Below and Below Average group is (0.0436, 0.0842). As the confidence interval does not include 0, we can 95% confident to conclude that if people feel satisfied about their current financial condition at all has a positive relationship with leveling up from Far Below to Below Average family at the age of 16.

```
# create inference
inference(data = fingss_fbvsba, y = nsatfin, x = incom16, type = "ht",
          statistic = "proportion", method = "theoretical",
          null = 0, success = "Not At All", alternative = "twosided", conf_level = 0.95)
```

```
## Response variable: categorical (2 levels, success: Not At All)
## Explanatory variable: categorical (2 levels)
## n_Far Below Average = 2778, p_hat_Far Below Average = 0.3362
## n_Below Average = 7301, p_hat_Below Average = 0.2723
## H0: p_Far Below Average =  p_Below Average
## HA: p_Far Below Average != p_Below Average
## z = 6.3199
## p_value = < 0.0001
```



Z score Hypothesis Testing Method Result Interpretation:

H0: Satisfaction level about current financial situation has nothing to do with the surveyor’s family financial position when he/she was 16. Therefore we would expect the distribution of satisfaction status to be roughly equal between the Far Below Average group and the Below Average group.

HA: Satisfaction level about current financial situation is somewhat dependent on the surveyor’s family financial position when he/she was 16, specifically Far Below Average or Below Average.

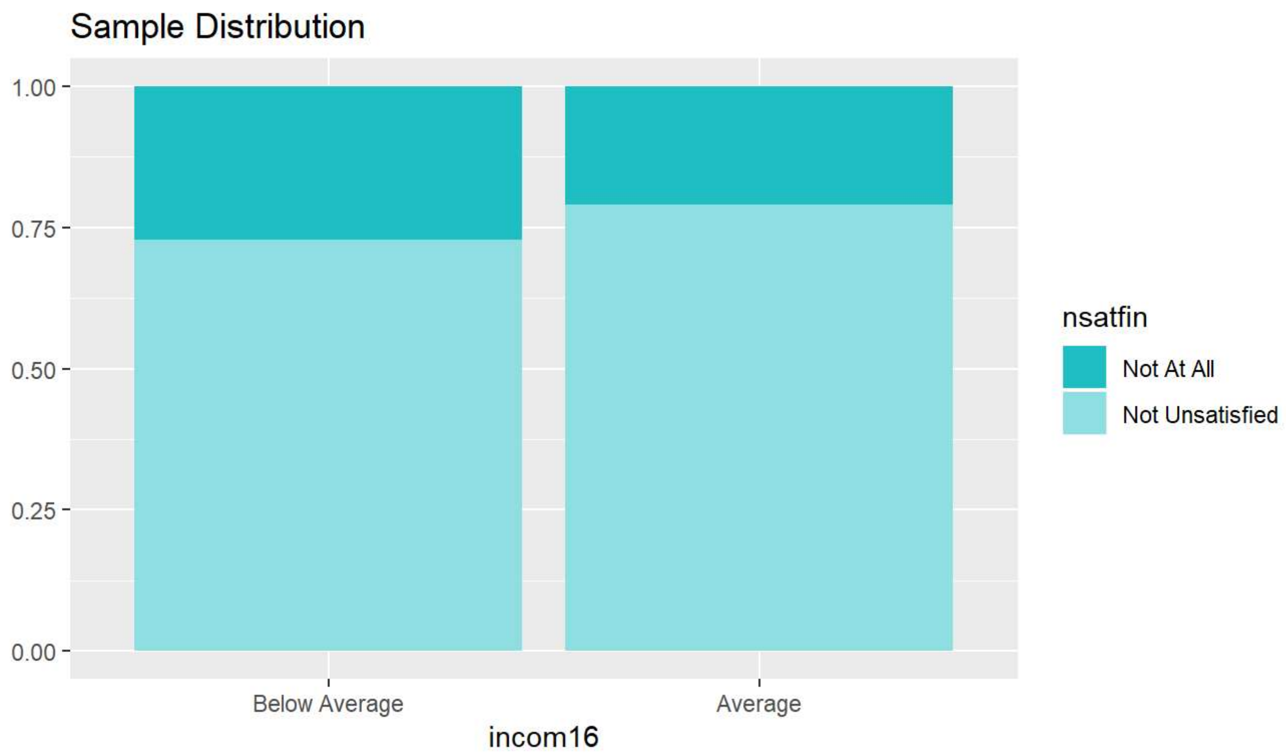
Z score is positive and fairly big, and p-value is small. This means that there were significant differences among the data, and hence we would reject H0, and be 95% confident that if people feel satisfied about their current financial condition at all has a positive relationship with leveling up from Far Below to Below Average family at the age of 16.

This result agrees with the CI method result above.

```
fingss_bavsav=subset(fingss, fingss$incom16 == "Below Average" | fingss$incom16 == "Average")
fingss_bavsav$incom16<-droplevels(fingss_bavsav$incom16)
```

```
# create inference
inference(data = fingss_bavsav, y = nsatfin, x = incom16, type = "ci",
          statistic = "proportion", method = "theoretical",
          success = "Not At All", alternative = "twosided", conf_level = 0.95)
```

```
## Response variable: categorical (2 levels, success: Not At All)
## Explanatory variable: categorical (2 levels)
## n_Below Average = 7301, p_hat_Below Average = 0.2723
## n_Average = 13253, p_hat_Average = 0.2095
## 95% CI (Below Average - Average): (0.0504 , 0.0751)
```

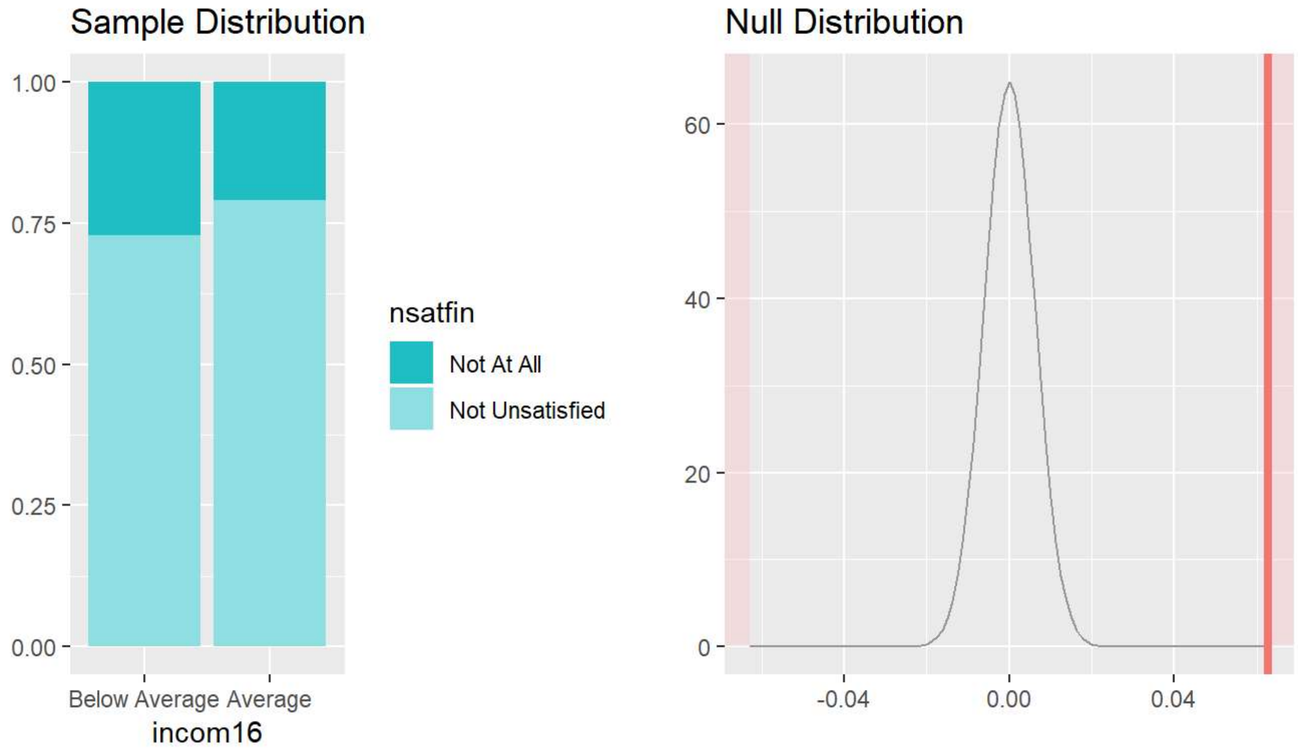


Confidence Interval Method Result Interpretation:

The 95% confidence interval for the differences between the Below Average and Average group is (0.0504, 0.0751). As the confidence interval does not include 0, we can 95% confident to conclude that if people feel satisfied about their current financial condition at all has a positive relationship with leveling up from Below Average to Average family at the age of 16.

```
# create inference
inference(data = fingss_bavsav, y = nsatfin, x = incom16, type = "ht",
          statistic = "proportion", method = "theoretical",
          null = 0, success = "Not At All", alternative = "twosided", conf_level = 0.95)
```

```
## Response variable: categorical (2 levels, success: Not At All)
## Explanatory variable: categorical (2 levels)
## n_Below Average = 7301, p_hat_Below Average = 0.2723
## n_Average = 13253, p_hat_Average = 0.2095
## H0: p_Below Average =  p_Average
## HA: p_Below Average != p_Average
## z = 10.203
## p_value = < 0.0001
```



Z score Hypothesis Testing Method Result Interpretation:

H0: Satisfaction level about current financial situation has nothing to do with the surveyor’s family financial position when he/she was 16. Therefore we would expect the distribution of satisfaction status to be roughly equal between the Below Average group and the Average group.

HA: Satisfaction level about current financial situation is somewhat dependent on the surveyor’s family financial position when he/she was 16, specifically Below Average or At Average.

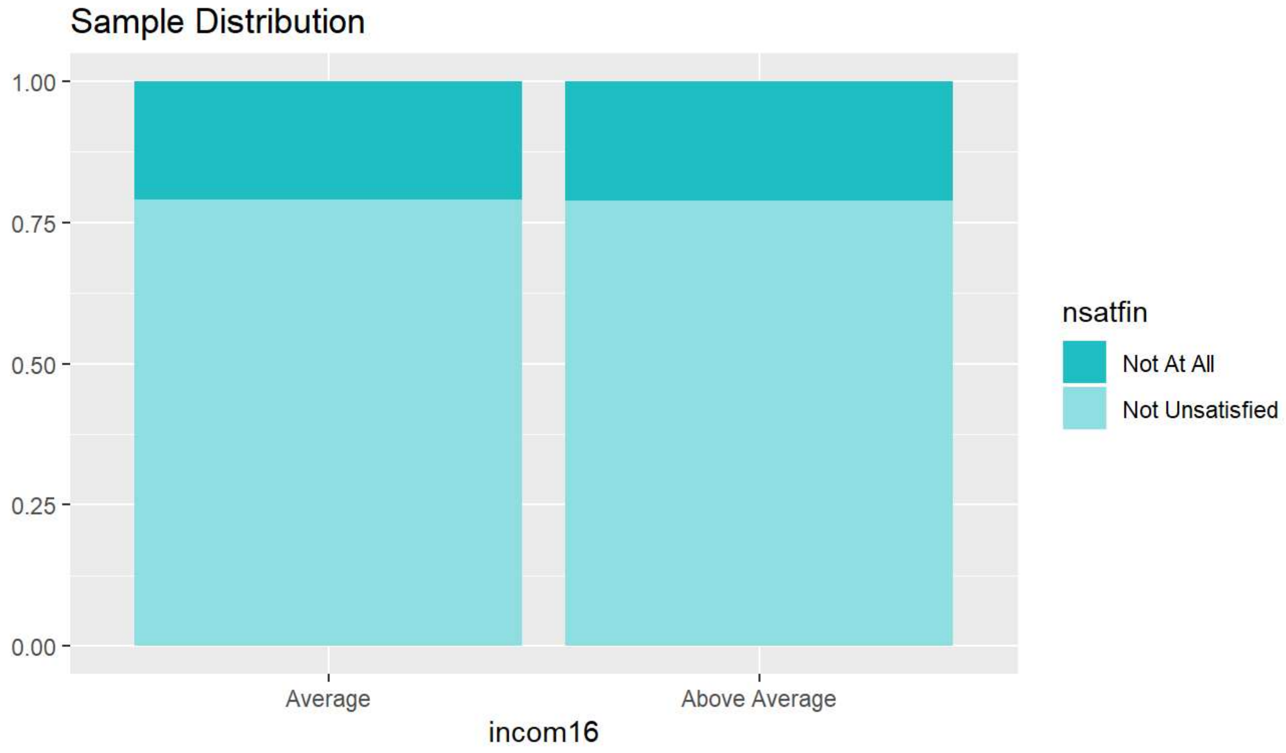
Z score is positive and fairly big, and p-value is small. This means that there were significant differences among the data, and hence we would reject H0, and be 95% confident that if people feel satisfied about their current financial condition at all has a positive relationship with leveling up from Below Average to Average family at the age of 16.

This result agrees with the CI method result above.

```
fingss_avvsaa=subset(fingss, fingss$incom16 == "Average" | fingss$incom16 == "Above Average")
fingss_avvsaa$incom16<-factor(fingss_avvsaa$incom16)
```

```
# create inference
inference(data = fingss_avvsaa, y = nsatfin, x = incom16, type = "ci",
          statistic = "proportion", method = "theoretical",
          success = "Not At All", alternative = "twosided", conf_level = 0.95)
```

```
## Response variable: categorical (2 levels, success: Not At All)
## Explanatory variable: categorical (2 levels)
## n_Average = 13253, p_hat_Average = 0.2095
## n_Above Average = 3632, p_hat_Above Average = 0.2117
## 95% CI (Average - Above Average): (-0.0172 , 0.0128)
```

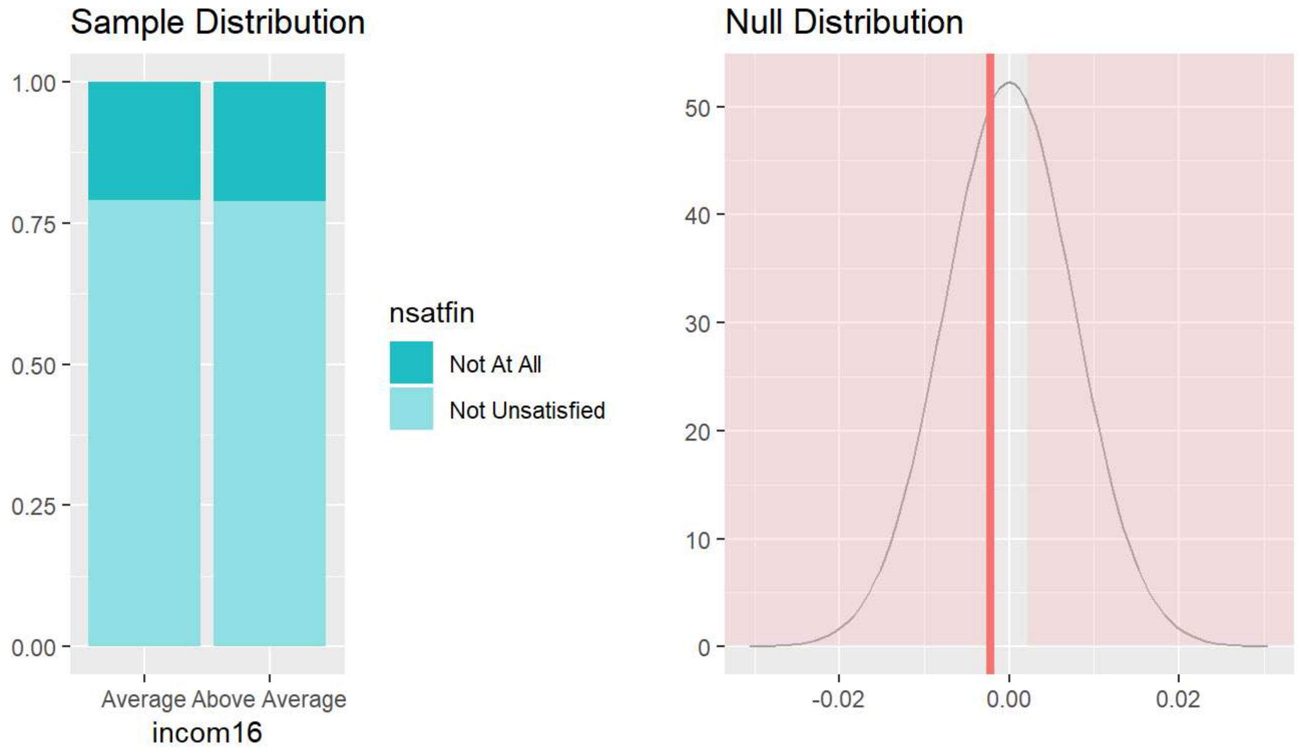


Confidence Interval Method Result Interpretation:

The 95% confidence interval for the differences between the Average and Above Average group is (-0.0172, 0.0128). As the confidence interval does include 0, we can be 95% confident to conclude that if people feel satisfied about their current financial condition at all has nothing to do with their family background of Average or Above Average at the age of 16.

```
# create inference
inference(data = fingss_avvsaa, y = nsatfin, x = incom16, type = "ht",
          statistic = "proportion", method = "theoretical",
          null = 0, success = "Not At All", alternative = "twosided", conf_level = 0.95)
```

```
## Response variable: categorical (2 levels, success: Not At All)
## Explanatory variable: categorical (2 levels)
## n_Average = 13253, p_hat_Average = 0.2095
## n_Above Average = 3632, p_hat_Above Average = 0.2117
## H0: p_Average =  p_Above Average
## HA: p_Average != p_Above Average
## z = -0.2873
## p_value = 0.7739
```



Z score Hypothesis Testing Method Result Interpretation:

H0: Satisfaction level about current financial situation has nothing to do with the surveyor’s family financial position when he/she was 16. Therefore we would expect the distribution of satisfaction status to be roughly equal between the Average group and the Above Average group.

HA: Satisfaction level about current financial situation is somewhat dependent on the surveyor’s family financial position when he/she was 16, specifically At Average or Above Average.

Z score is negative and small, and p-value is fairly big, about 0.77, which is a lot bigger than our significance level of 0.05. This means that there were no significant differences among the data, and hence we would fail to reject H0, and be 95% confident that if people feel satisfied about their current financial condition at all has nothing to do with their family background of Average or Above Average at the age of 16.

This result agrees with the CI method result above.

Conclusion:

Whether people feel satisfied, or more or less satisfied, or not satisfied at all about their current financial situation has something to do with their family financial situation at the age of 16, for those who were from a far below average, below average, at average, or above average family. But once the family financial situation became above average, regardless of how much above, the satisfaction levels became to have a consistent distribution, without significant proportion differences of the satisfaction levels.

And for the groups that yielded significant differences from their closely-lower level or closely-upper level family background, the distribution of feeling satisfied at all of people from far below average, below average, and at average family had significant differences. The upper level their family became at the age of 16, the more proportion of them became somewhat satisfied about their current financial situation. Whereas, people with at average background and people with above average background do not have this difference statistically significant anymore.