Statistical Inference

Schwinn(Xuan) Chen

May 28, 2017

About the Data

Since 1972, the General Social Survey (GSS) has been monitoring societal change and studying the growing complexity of American society. The data in this analysis provides cumulative GSS from 1972 to 2012. It contains 57061 samples, questioning 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.

Stratified Multistage Sampling methods were used to collect samples to ensure equal probability of respondents from population. Thus, our analysis will be based on random samples and the results can be generalized to the whole population. Because the data is collected from surveys not controlled studies, our analysis do not establish causal conclusions.

References: http://gss.norc.org/Documents/codebook/A.pdf

Load R Packages

```
library(dplyr)
library(statsr)
library(ggplot2)
source("http://bit.ly/dasi_inference")
load("gss.Rdata")
```

Part 2: Research Questions

Family status variables such as parents' level of education have been regarded as predictors of children's academic achievement. In this report, We will use data from GSS to learn whether parent's level of education has a direct association with people's academic achievement.

We will answer our research question in three parts.

1. Are people's level of education associated with their parent's?
2. Does gender of parent play a role? In other words, which parent has a stronger association, the father or mother?
3. Does race of people matter? Does white people's level of education have the same association with their mother's or father's as blacks'?

Part 3: Exploratory data analysis

Research Part 1

Are people's level of education associated with their parent's?

All respondents in the data are over 18 years old therefore the population in this report is generalized to

adults (>18 years old) in the US.

After analysing the dataset, we will use the following variables to answer this question:

‘educ’: Highest year of school completed by the respondent, a discrete variable ranging from 0 to 20.

‘paeduc’: Highest year of school completed by father, a discrete variable ranging numerical from 0 to 20.

‘maeduc’: Highest year of school completed by mother, a discrete variable ranging from 0 to 20.

If one completes more than 12 years of school, he or she has some kind of post-secondary education. We create the following categorical variables based on whether the person completed at least one year of postsecondary education:

‘educ_level’: It includes two categories.

‘yes’: the respondent completed at least one year of post-secondary education

‘no’: the respondent did not compete any year of post-secondary education

‘p_level’: It includes three categories.

‘both’: both mother and father completed at least one year of post-secondary education

‘one’: ONLY one of the parents (father XOR mother) completely at least one year of post-secondary education

‘neither’: neither parent completed at least one year of post-secondary education

We use R to create a new dataframe named ‘parent_ed’ which includes variables ‘educ’, ‘paeduc’, ‘maeduc’,‘educ_level’,and ‘p_level’.

```
parent_ed<-gss%>%
  ##remove NA data
filter(!is.na(maeduc),!is.na(paeduc),!is.na(educ))%>%

select(maeduc,paeduc,educ)%>%
  ## mutate categorical variable pa_level
mutate(pa_level=ifelse(xor(maeduc>12,paeduc>12),"one",ifelse(paeduc>12 & paeduc>12,
"both","neither")))%>%
  ## mutate categorical variable educ_level
mutate(educ_level=ifelse(educ>12,"yes","no"))

  ## data will be displayed in the order of "both" "one" "neither" in tables and ch
arts
parent_ed$pa_level=factor(parent_ed$pa_level, c("both","one","neither"))
```

Exploratory Data Analysis of Part #1

Before we do any analysis, we will learn sample statistics such as sample size (n), median, mean, sd, variance, etc. This is accomplished by the following R code.

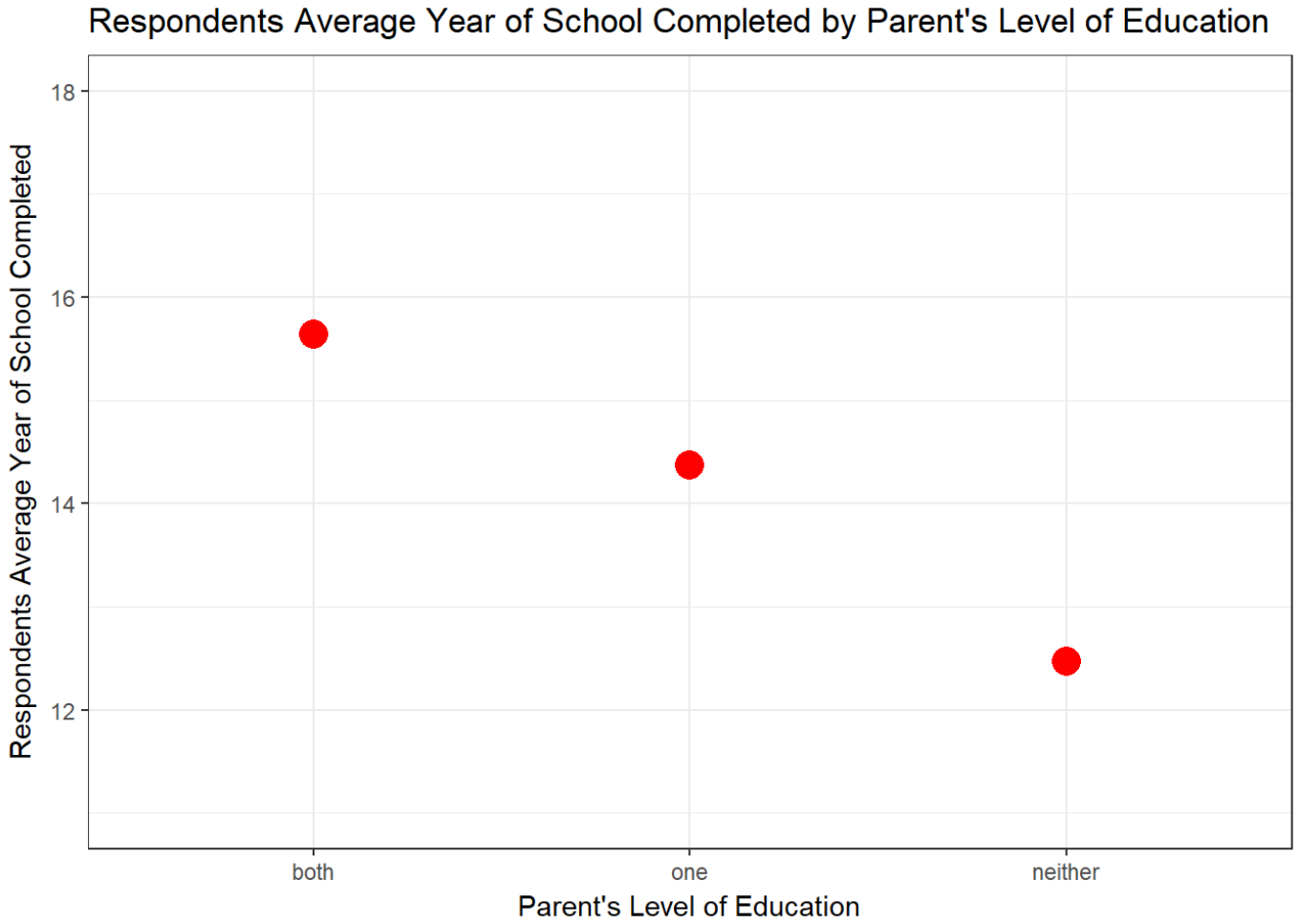
```
parent_ed %>%
summarise(n=n(),mean=mean(educ), median=median(educ),sd=sd(educ),var=var(educ))
```

```
##       n    mean median       sd      var
## 1 37423 13.3073     13 3.031783 9.191707
```

```
## contingency table show sample statiscs in 'both','one' and 'neither' categories
parent_sum <-parent_ed%>%
group_by(pa_level)%>%
summarise(count=n(), mean=mean(educ),median=median(educ),sd=sd(educ), var=var(educ)
, IQR=IQR(educ))
parent_sum
```

```
## # A tibble: 3 x 7
##   pa_level count  mean median    sd   var   IQR
##   <fct>    <int> <dbl>  <int> <dbl> <dbl> <dbl>
## 1 both      5315  15.6     16  2.28  5.18     3
## 2 one       7667  14.4     14  2.44  5.97     4
## 3 neither  24441  12.5     12  2.97  8.83     2
```

```
ggplot(parent_sum,aes(x=pa_level,y=mean))+geom_point(size=5, color='red')+theme_bw(
)+scale_y_continuous(limits=c(11,18))+labs(x="Parent's Level of Education",y='Respo
ndents Average Year of School Completed',title="Respondents Average Year of School
Completed by Parent's Level of Education")
```



From the table and scattered chart, we can find that the mean of the respondents' highest year of school completed decrease as the education level of their parents decreases.

Respondents from families that both parents completed at least one year of post-secondary education have the highest mean and median of highest year of school completed.

Respondents from families that only one parent completed at least one year of post-secondary education have the 2nd highest mean and median of highest year of school completed.

Respondents from families that neither parent completed at least one year of post-secondary education have the lowest mean and median of highest year of school completed.

Inference for Part #1

Next, we will find out if the above statements are statistically significant to the population. We will use Central Limit Theorem (CLT) to calculate the confidence intervals (CI) of the average highest year of school completed in all three categories of parent's level of education ('both', 'one', 'either').

In order to conduct CLT for the mean, the data needs to meet the following conditions:

1. Independence. Sample observations must be independent, which include random sample/assignment. This condition has been met, see introduction of data for details

If sampling without replacement, n<10% population. This condition has been met, the sample size in categories of "both", "one" and "neither" is 5315, 7667, 24441 respectively, any of which is definitely smaller than 10% of the population.

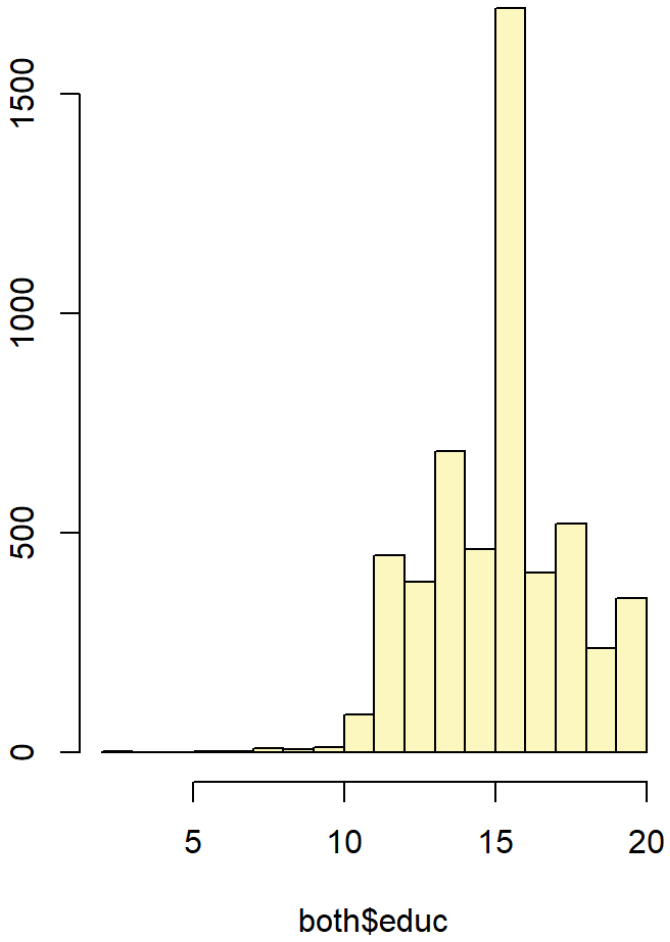
2. Either the population distribution is normal or if the population distribution is skewed, the sample size is large. The sample sizes listed in the contingency table are large enough.

Therefore, the conditions for performing CLT are met.

Now, we will calculate the CIs (at 95% confidence level) for the mean highest year of school completed by the respondents in all three categories of parent's level of education ("both", "one", and "neither").

```
both<-parent_ed %>%
filter(pa_level=='both')
inference(both$educ, est = "mean", type = "ci", null = 0, alternative = "twosided"
, method = "theoretical")
```

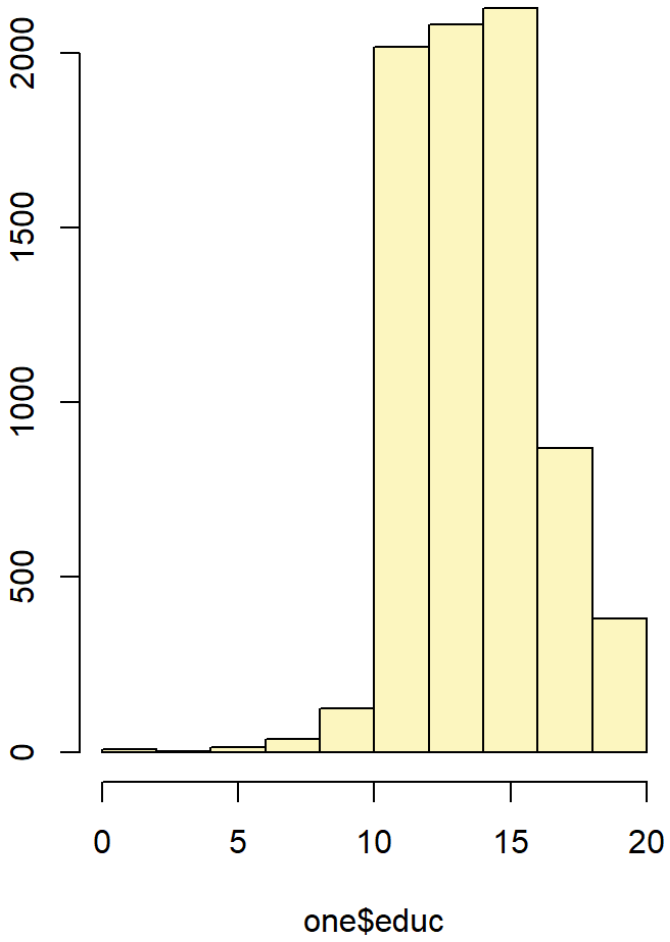
```
## Single mean
## Summary statistics:
```




```
## mean = 15.6406 ;  sd = 2.2755 ;  n = 5315
## Standard error = 0.0312
## 95 % Confidence interval = ( 15.5795 , 15.7018 )
```

```
one<-parent_ed %>%
filter(pa_level=='one')
inference(one$educ, est = "mean", type = "ci", null = 0, alternative = "twosided",
method = "theoretical")
```

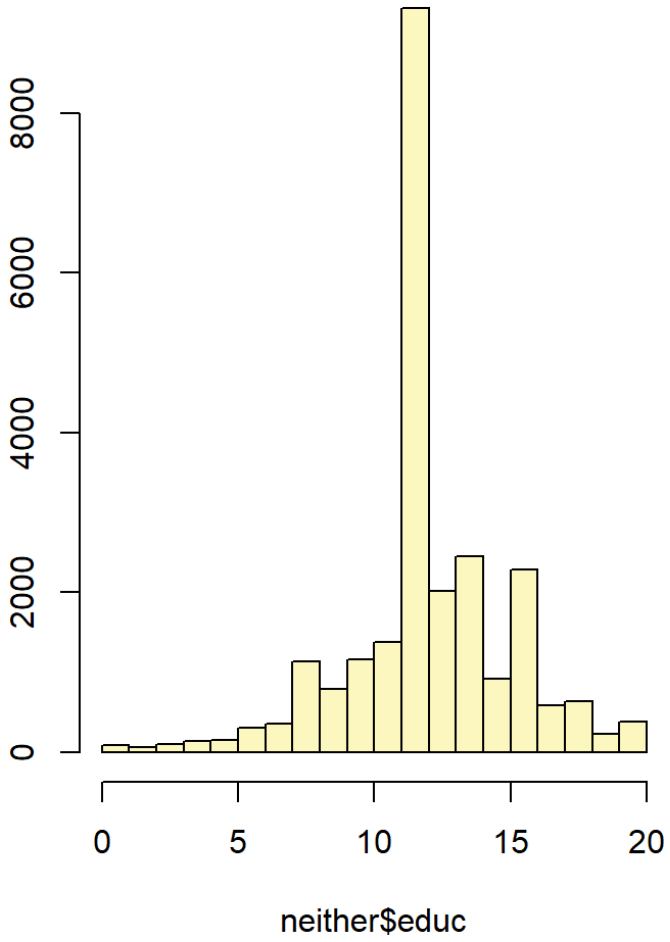
```
## Single mean
## Summary statistics:
```




```
## mean = 14.3678 ;  sd = 2.443 ;  n = 7667
## Standard error = 0.0279
## 95 % Confidence interval = ( 14.3131 , 14.4225 )
```

```
neither<-parent_ed %>%
filter(pa_level=='neither')
inference(neither$educ, est = "mean", type = "ci", null = 0, alternative = "twoside
d", method = "theoretical")
```

```
## Single mean
## Summary statistics:
```




```
## mean = 12.4672 ;  sd = 2.9722 ;  n = 24441
## Standard error = 0.019
## 95 % Confidence interval = ( 12.4299 , 12.5045 )
```

Summary of CIs for the mean of highest year of school completed from three types of families.

We are 95% confident that:

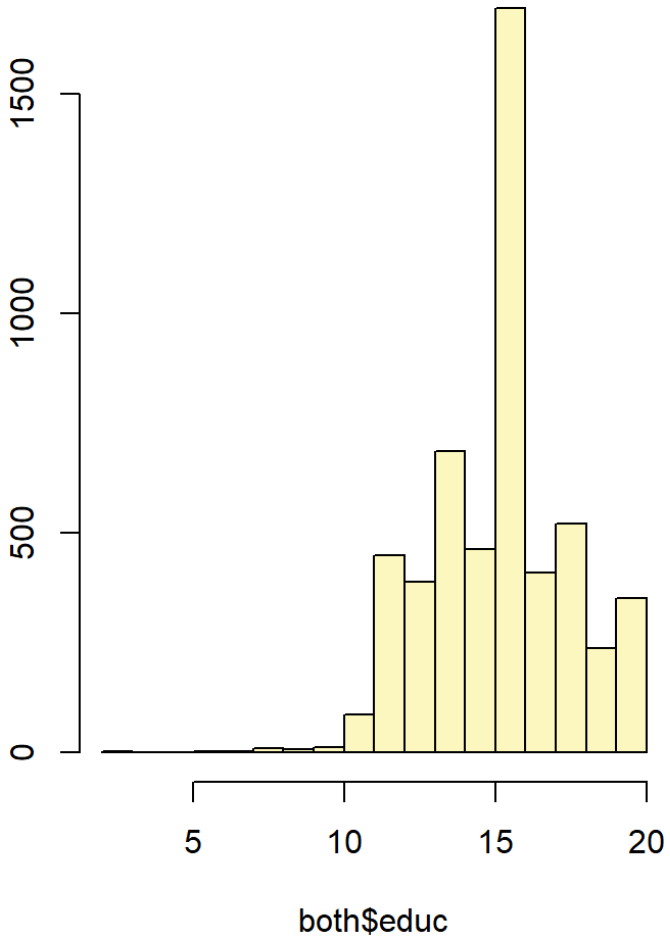
1. The average year of school completed by adults from families that both parents completed at least one year of post-secondary education is (15.5795 , 15.7018)
2. The average year of school completed by adults from families that only one parent completed at least one year of post-secondary education is (14.3131 , 14.4225)
3. The average year of school completed by adults from families that neither parent completed at least one year of post-secondary education is (12.4299 , 12.5045)

Because the three CIs do not overlap. We can probably say that the average year of school completed by adults varies with respect to their parent's level of education. The type_1 error of this conclusion is 1-0.95x0.95x0.95 = 14.26%, a quite high error. One way to reduce this error is to increase the CL. When the CL Is 98% , the type_1 error is redueced to 1-0.99x0.99x0.99 = 3%.

The following inferences calculate CIs for the three means, using 99% confident level.

```
both<-parent_ed %>%
filter(pa_level=='both')
inference(both$educ, est = "mean", type = "ci", null = 0, alternative = "twosided"
, method= "theoretical", conflevel = 0.99)
```

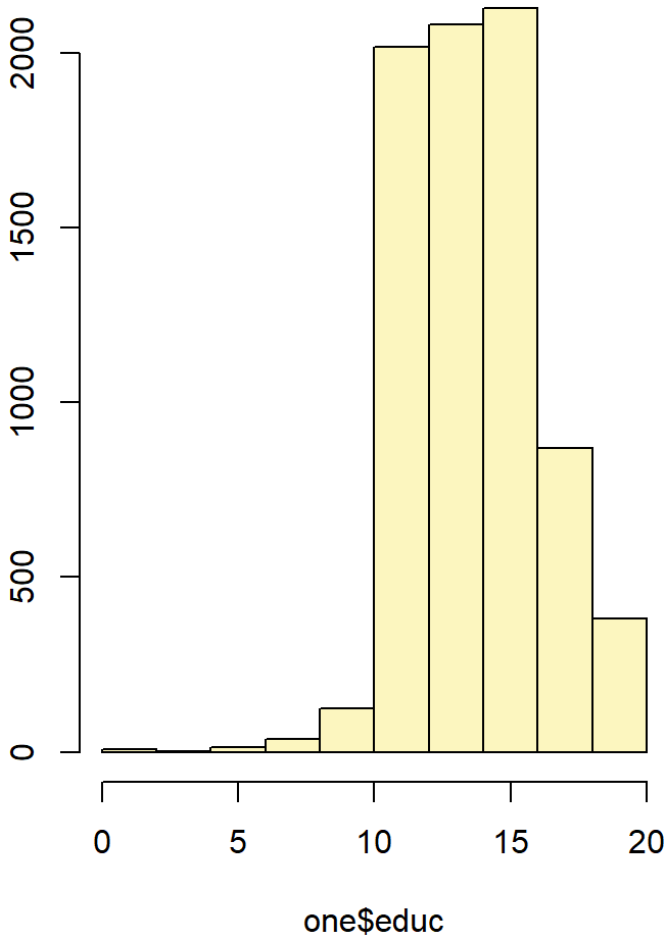
```
## Single mean
## Summary statistics:
```




```
## mean = 15.6406 ;  sd = 2.2755 ;  n = 5315
## Standard error = 0.0312
## 99 % Confidence interval = ( 15.5602 , 15.721 )
```

```
one<-parent_ed %>%
filter(pa_level=='one')
inference(one$educ, est = "mean", type = "ci", null = 0, alternative = "twosided",
method = "theoretical", conflevel = 0.99)
```

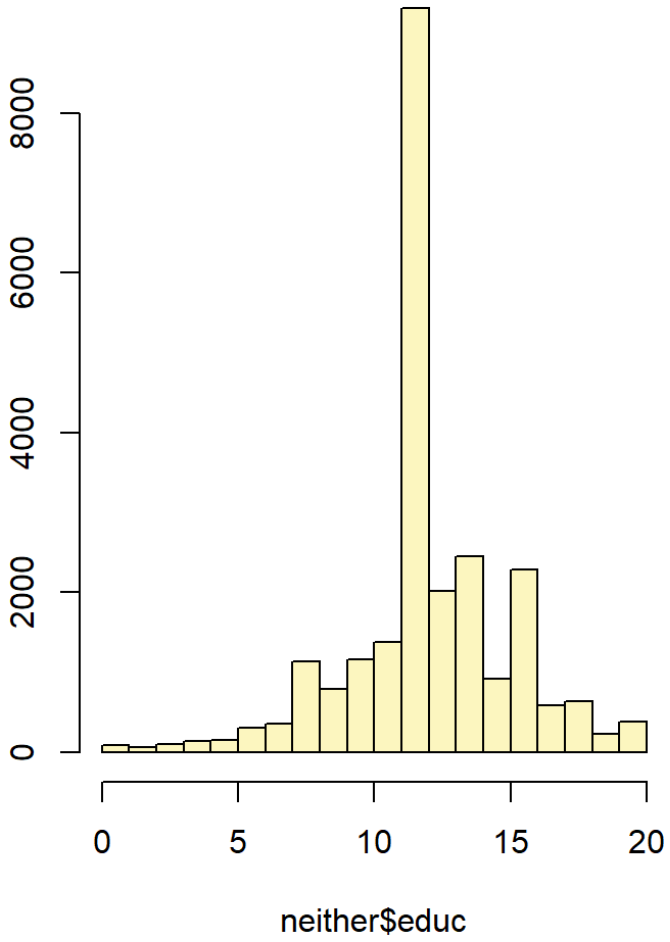
```
## Single mean
## Summary statistics:
```




```
## mean = 14.3678 ;  sd = 2.443 ;  n = 7667
## Standard error = 0.0279
## 99 % Confidence interval = ( 14.2959 , 14.4397 )
```

```
neither<-parent_ed %>%
filter(pa_level=='neither')
inference(neither$educ, est = "mean", type = "ci", null = 0, alternative = "twoside
d", method = "theoretical", conflevel = 0.99)
```

```
## Single mean
## Summary statistics:
```




```
## mean = 12.4672 ;  sd = 2.9722 ;  n = 24441
## Standard error = 0.019
## 99 % Confidence interval = ( 12.4182 , 12.5162 )
```

1. The average year of school completed by adults from families that both parents completed at least one year of post-secondary education is (15.5602 , 15.721)
2. The average year of school completed by adults from families that only one parent completed at least one year of post-secondary education is (14.2959 , 14.4397)
3. The average year of school completed by adults from families that neither parent completed at least one year of post-secondary education is (12.4182 , 12.5162)

Because the three CIs do not overlap at 99% CL in each inference. We can say that the average year of school completed by adults increase as more of the parents completed at least one year of post-secondary education. Because the analysis compares multiple categorical means at three levels, we are curious to see if an ANOVA test agrees with our finding.

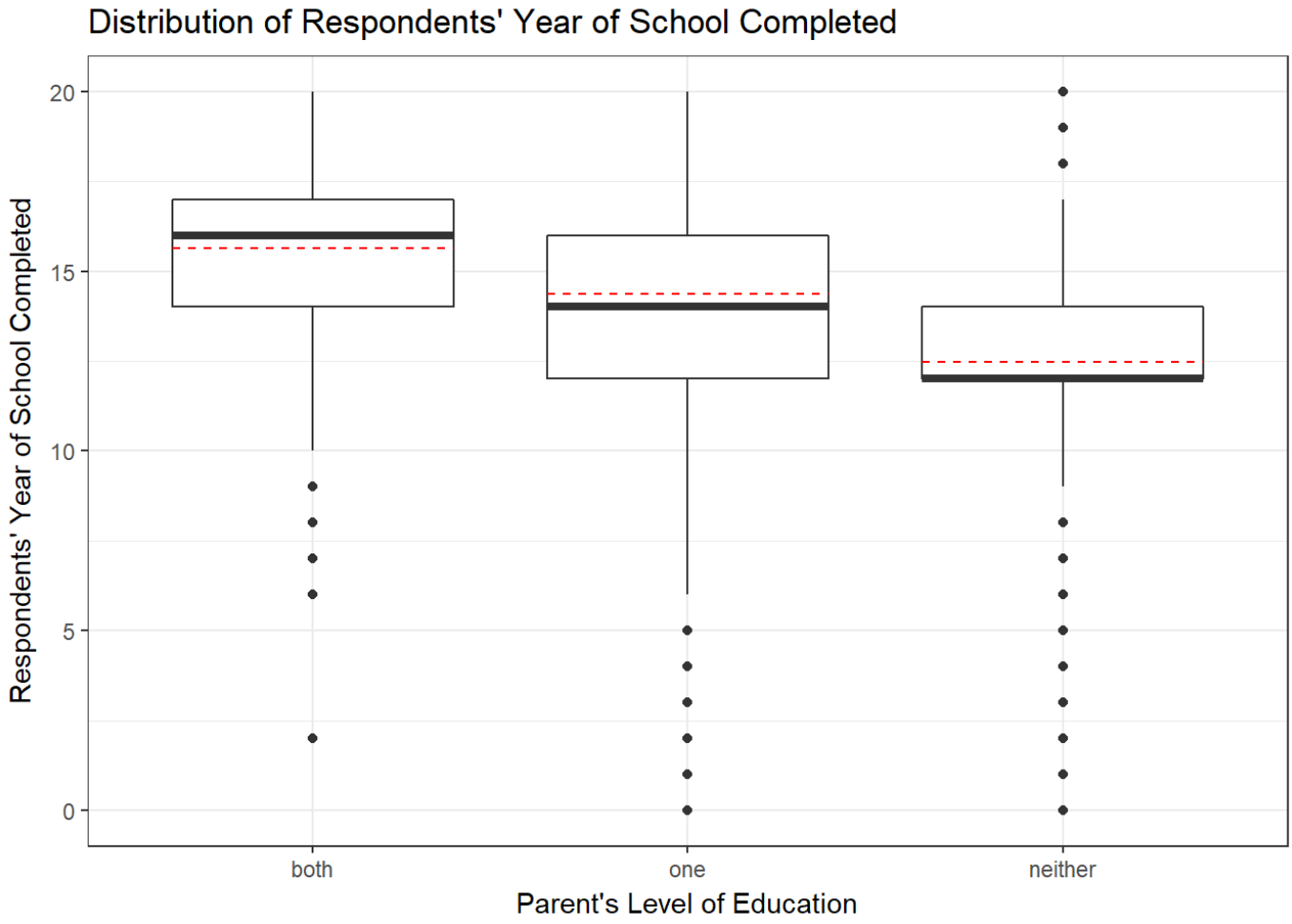
First we need to find if the conditions for ANOVA test has been met. Similar to the CTL conditions, the sample has met conditions for ANOVA such as independence within and between groups, random sample, and sample size smaller than 10% of population.

Additionally, ANOVA requires that distributions should be nearly normal within each group and groups should have roughly equal variability, especially when the sample size between groups varies.

The following R code visualizes side-by-side boxplots of the distributions of the highest year of school completed in each group.

```
##boxplots of sample distribution at all education levels of parents

ggplot(parent_ed, aes(x=pa_level,y=educ))+theme_bw()+geom_boxplot(fatten=3)+stat_summary(fun.y=mean,geom='errorbar',aes(ymax=..y..,ymin=..y..),width=0.75,linetype='dashed', color='red')+xlab("Parent's Level of Education")+ylab("Respondents' Year of School Completed")+labs(title="Distribution of Respondents' Year of School Completed")
```



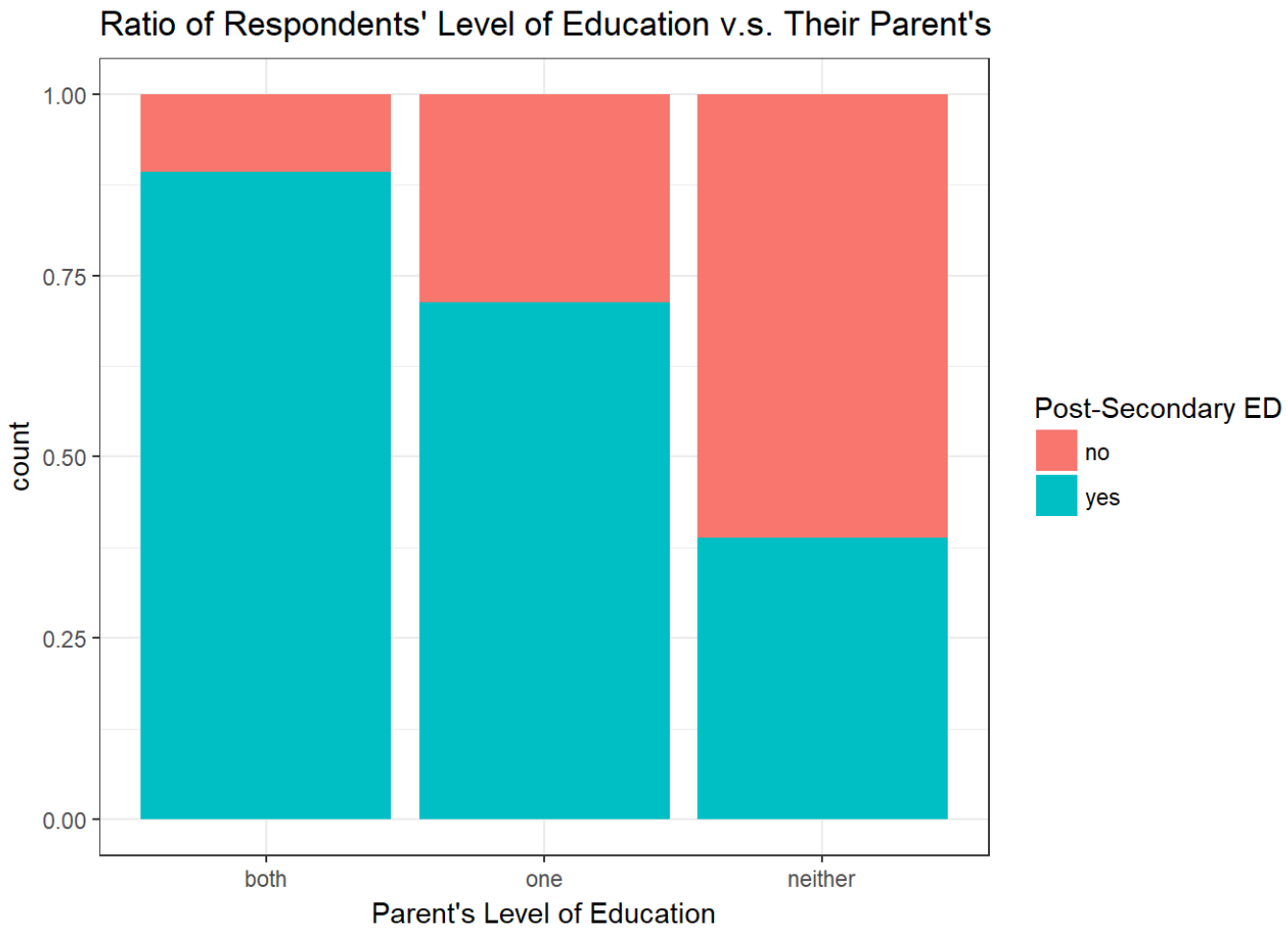
The distribution is right skewed in the 'neither' group and left skewed in the 'both' group. The variability differs a lot in the 'one' group and 'neither group'. In addition, the sample size varies greatly between groups. We do not think the criteria for an ANOVA test have been met. For future studies, if we collect more data in the 'both' and 'one' categories so the sample sizes are equal in all three groups. We should be able to conduct an ANOVA test.

Besides the analysis of the mean, we are also interested in proportions. Are adults from post-secondary educated family more likely to complete at least one year of post-secondary education?

We have created a categorical variable 'educ_level'. It classifies the respondents' year of school of completed by whether he or she finished at least one year of post-secondary school and from it we can find the proportion of respondents finishing at least one year of post-secondary education.

The filled bar chart illustrates that the proportion of respondents with post-secondary education decrease as less of their parents completed at least one year of post-secondary education. Respondents' level of education is associated with their parent's level of education.

```
ggplot(parent_ed, aes(x=pa_level,fill=educ_level))+geom_bar(position = 'fill')+theme_bw()+scale_fill_discrete(name="Post-Secondary ED")+ labs(x="Parent's Level of Education", title="Ratio of Respondents' Level of Education v.s. Their Parent's")
```



Next, we will conduct a Chi-Square Independence test to see the statistics agree with our finding. The summary of the categorical variables is listed in the following contingency table:

```
parent_percent <-parent_ed%>%
group_by(pa_level)%>%
summarise(yes=sum(educ_level=='yes'),no=sum(educ_level=='no'))
parent_percent
```

```
## # A tibble: 3 x 3
##   pa_level   yes    no
##   <fct>    <int> <int>
## 1 both      4748   567
## 2 one       5461  2206
## 3 neither   9486 14955
```

Before we perform the test, we need to fulfil the criteria required by Chi-Square Independence test.

1. Sampled observations must be indepent. As discussed earlier in this report, data is acquired from random samples and sample size is less than 10% of the population. In addition, each case only contributes to one cell in the contingency table. This requirement has also been met, because we created the ‘one’ category as ‘only one of the parent completed at least one year of post-secondary education’, which ensure that data in the ‘one’ group does not overlap with that in the ‘both’ group.
2. Each particular scenario (cell in the contingency table) must have at last 5 expected cases. Our data also meet this requirement.

Now we are ready to perform the Chi-Square Independent test. Our null and alternative hypotheses are listed below:

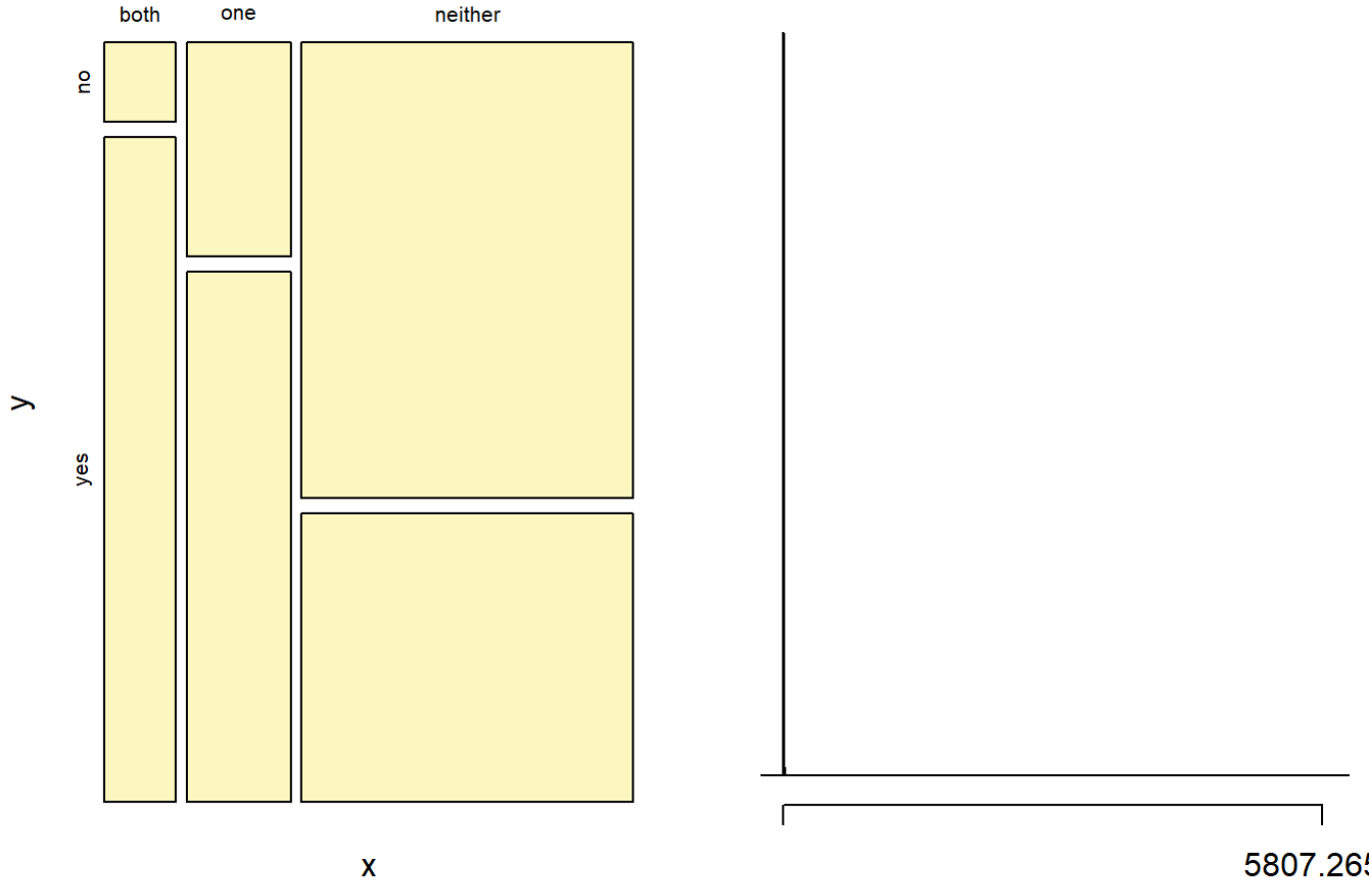
H_0: No association between adluts' and their parent' level of education H_A: adults' and their parent's level of education are associated.

We use the following R code to carry out our hypothesis test.

```
inference(parent_ed$educ_level, parent_ed$pa_level, est = "proportion", type = "ht"
,
          alternative = "greater", success="yes", method = "theoretical")
```

```
## Response variable: categorical, Explanatory variable: categorical
## Chi-square test of independence
##
## Summary statistics:
##       x
## y       both   one neither   Sum
##   no     567  2206   14955 17728
##   yes   4748  5461    9486 19695
##   Sum   5315  7667   24441 37423
```

```
## H_0: Response and explanatory variable are independent.
## H_A: Response and explanatory variable are dependent.
## Check conditions: expected counts
##       x
## y         both     one  neither
##   no  2517.82 3632.01 11578.18
##   yes 2797.18 4034.99 12862.82
##
##  Pearson's Chi-squared test
##
## data:  y_table
## X-squared = 5807.3, df = 2, p-value < 2.2e-16
```



The above Chi-Square Independence test yields a Chi-Square of 5807.3 and p-value is almost 0. Therefore, we reject the null hypothesis, thus, we have enough evidence to support that adults' and their parent's level of education are associated.

Research Part 2

We have concluded in Research Part #1 that adults'and their parent's level of education are strongly associated. The more of the parents are educated, the more their adult children.

Does gender of parent play a role? In other words, which parent has a stronger association, the father or mother?

In order to answer this question, we will use the three variables from the previous question 'educ', 'paeduc'and 'maeduc'.

We create the following two categorical variables:

'educ_level': Itcl assifies two categories. 'yes': the respondent completed at least one year of post-secondary education 'no': the respondent did not compete any year of post-secondary education

'p_one': It classifies two categories. 'father': only father completed at least one year of post-secondary education 'mother': only mother completed at least one year of post-secondary education

This variable is created to ensure independence between groups so that only one of the parent completed at least one year of post-secondary education in each group.

Exploratory Data Analysis of Part #2

The following R code creates a new dataframe called ‘compare_fa_ma’ and its contingency table.

```
compare_fa_ma<-gss%>%
  ## remove NA values
filter(!is.na(maeduc),!is.na(paeduc),!is.na(educ))%>%
select(maeduc,paeduc,educ)%>%
## mutate variable 'pa_one'
mutate(pa_one=ifelse(maeduc>12 & paeduc<=12,"mother",ifelse(paeduc>12 & maeduc<=12,
"father","discard"))) %>%
filter(pa_one=="mother" | pa_one=="father")%>%
## mutate variable 'educ_level'
mutate(educ_level=ifelse(educ>12,"yes","no"))
```

```
## contingency table
compare_fa_ma%>%
group_by(pa_one)%>%
summarise(yes=sum(educ_level=='yes'), no=sum(educ_level=='no'), prop_yes=100*sum(ed
uc_level=='yes')/n())
```

```
## # A tibble: 2 x 4
##   pa_one   yes    no prop_yes
##   <chr>  <int> <int>    <dbl>
## 1 father  3276  1193     73.3
## 2 mother  2185  1013     68.3
```

The contingency table yields that 73.3% of respondents completed at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education and that 68.3% of respondents completed at least one year of post-secondary education on the condition that their mother (only mother) completed at least one year of post-secondary education.

Inference Part #2

Based on this observation, we will conduct a hypothesis test to compare two independent proportions.

Requirements for such a test:

1. Samples are independent within groups, which means random samples and if sampling without replacement, sample size <10% of population. Same scenario as we discussed before, this condition has been met.

between groups, which means two groups must be independent of each other (not paired). This condition has been met. See the condition in which we set up the categorical variable ‘pa_one’.

2. Sample size. Each sample should have at least 10 successes and 10 failures. This condition has also been met. We have large success/failure rates.

We are ready to perform a hypothesis test to compare two independent proportions.

H_0: The proportion of adults having completed at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education is EQUAL to the proportion of those on the condition that their mother (only mother) completed at least one year of post-secondary education.

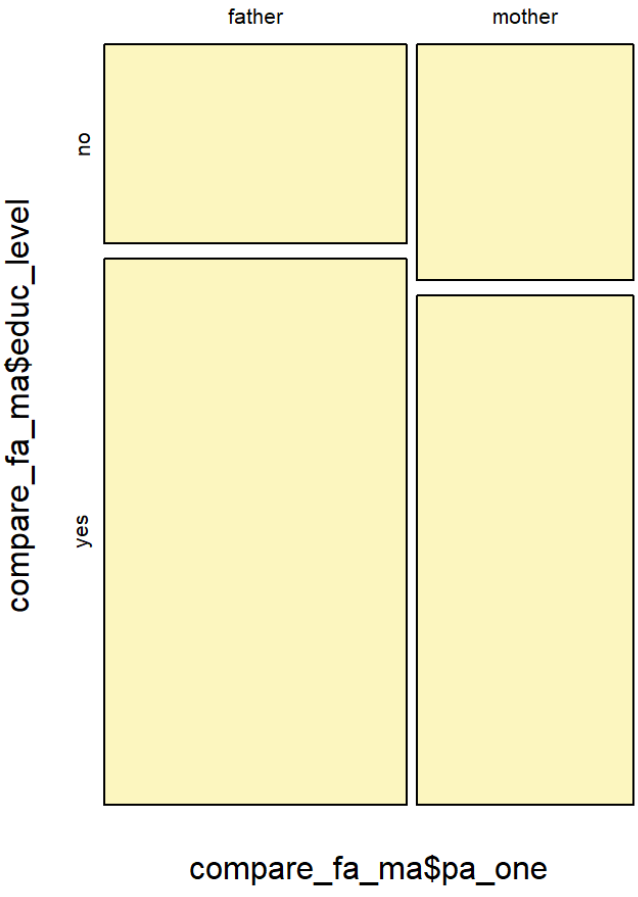
H_A: The proportion is LARGER

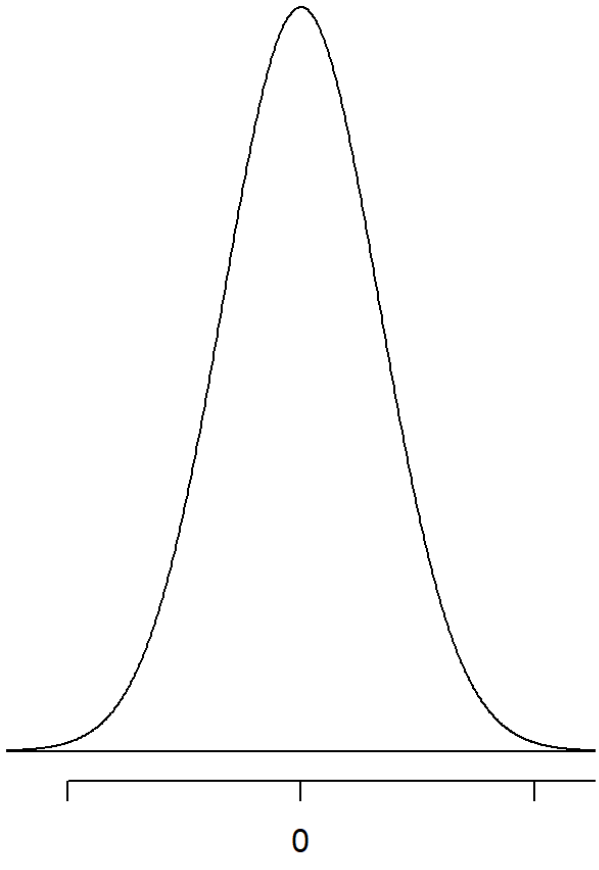
The hypotheis test is carried out using the following R code.

```
inference(compare_fa_ma$educ_level,compare_fa_ma$pa_one, est="proportion", type='ht
', method= "theoretical", null=0, alternative="greater", success="yes")
```

```
## Response variable: categorical, Explanatory variable: categorical
## Difference between two proportions -- success: yes
## Summary statistics:
##       x
## y      father mother  Sum
##   no     1193   1013 2206
##   yes    3276   2185 5461
##   Sum    4469   3198 7667
```

```
## Observed difference between proportions (father-mother) = 0.0498
## H0: p_father - p_mother = 0
## HA: p_father - p_mother > 0
## Pooled proportion = 0.7123
## Check conditions:
##    father : number of expected successes = 3183 ; number of expected failures =
1286
##    mother : number of expected successes = 2278 ; number of expected failures =
920
## Standard error = 0.01
## Test statistic: Z =  4.75
## p-value =  0
```





The hypothesis test yields a Z-score of 4.75 and a p-value near zero. Therefore, we reject the null

hypothesis. We have enough evidence to support the alternative hypotheis that the proportion of adults with at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education is LARGER than the proportion of those on the condition that their mother (only mother) completed at least one year of post-secondary education.

Research Part #3

Does race of people matter? Does white people's level of education have the same association with their mother's or father's as blacks'?

This research question addresses whether the association carries across difference races of adults.

Exploratory Data Analysis of Part #3

The 'race' categorical variable indicates the respondent is identified as White, Black or Other

We will perform analysis on white and black races.

We will need 'euduc', 'paeduc', 'maeduc','educ_level','pa_one' from previous questions.

```
#create a dataframe that the respondent is identified as White
white<-gss%>%
  ## filter respondents that are white
filter (race=='White')%>%
select (maeduc,paeduc,educ) %>%
filter(!is.na(maeduc),!is.na(paeduc),!is.na(educ))%>%
   ## mutate variable 'pa_one'
mutate(pa_one=ifelse(maeduc>12 & paeduc<=12,"mother",ifelse(paeduc>12 & maeduc<=12,
"father","discard"))) %>%
filter(pa_one=="mother" | pa_one=="father")%>%
  ## mutate variable 'educ_lev'
mutate(educ_level=ifelse(educ>12,"yes","no"))


# create a dataframe that the respondent is identified as black.
black<-gss%>%
  ## filter respondents that are black
filter (race=='Black')%>%
select (maeduc,paeduc,educ)%>%
filter(!is.na(maeduc),!is.na(paeduc),!is.na(educ))%>%
  ## mutate variable 'pa_one'
mutate(pa_one=ifelse(maeduc>12 & paeduc<=12,"mother",ifelse(paeduc>12 & maeduc<=12,
"father","discard"))) %>%
filter(pa_one=="mother" | pa_one=="father")%>%
  ## mutate variable 'educ_lev'
mutate(educ_level=ifelse(educ>12,"yes","no"))
```

Before we do any analysis, we need to have more understanding of the data. From the contingency table below, we can find that 73.3% of white respondents completed at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education and that 68.4% of those on the condition that their mother (only mother) completed at least one year of post-secondary education.

```
## contingency table for white respondents
white%>%
group_by(pa_one)%>%
summarise(yes=sum(educ_level=='yes'), no=sum(educ_level=='no'), prop_yes=100*sum(ed
uc_level=='yes')/n())
```

```
## # A tibble: 2 x 4
##   pa_one   yes    no prop_yes
##   <chr>  <int> <int>    <dbl>
## 1 father  2884  1050     73.3
## 2 mother  1854   855     68.4
```

Inference Part #3

Based on this observation, we will conduct a hypothesis test to compare two independent proportions.

Requirements for such a test:

1. Samples are independent within groups, which means random samples and if sampling without replacement, sample size <10% of population. Same scenario as we discussed before, this condition has been met

between groups, which means two groups must be independent of each other (not paired). This condition has been met. See the condition in which we set up the categorical variable 'pa_one'.

2. Sample size. Each sample should have at least 10 successes and 10 failures. This condition has also been met. We have large success/failure rates.

We are ready to perform a hypothesis test to compare two independent proportions.

H_0: The proportion of white adults with at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education is EQUAL to the proportion of those on the condition that their mother (only mother) completed at least one year of post-secondary education.

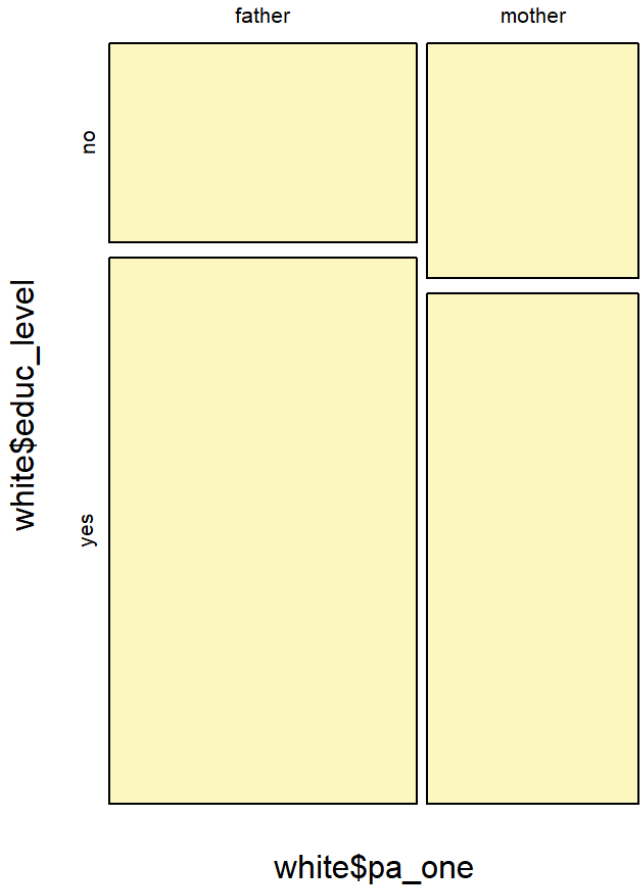
H_A: The proportion is LARGER.

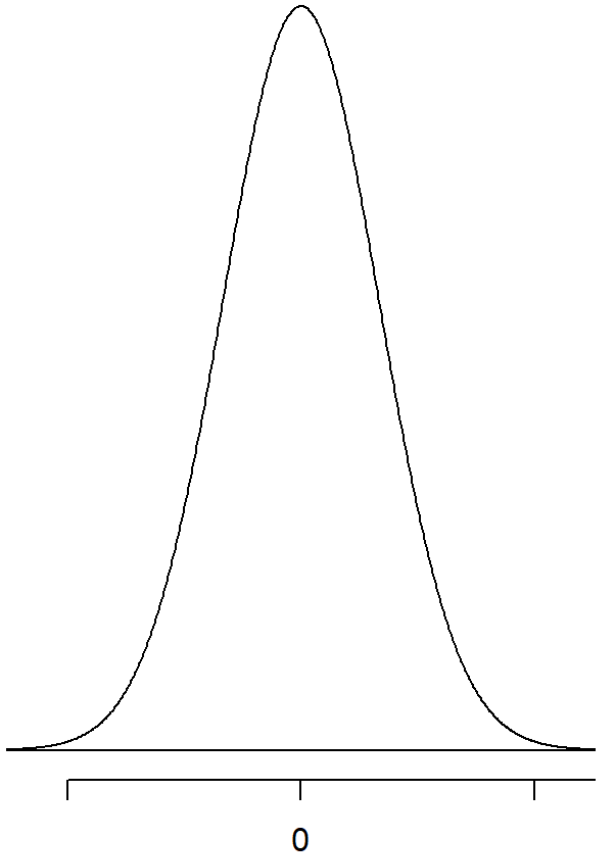
The hypothesis test is carried out using the following R code.

```
inference(white$educ_level,white$pa_one, est="proportion", type='ht', method= "theo
retical", null=0, alternative="greater", success="yes")
```

```
## Response variable: categorical, Explanatory variable: categorical
## Difference between two proportions -- success: yes
## Summary statistics:
##       x
## y      father mother  Sum
##   no     1050    855 1905
##   yes    2884   1854 4738
##   Sum    3934   2709 6643
```

```
## Observed difference between proportions (father-mother) = 0.0487
## H0: p_father - p_mother = 0
## HA: p_father - p_mother > 0
## Pooled proportion = 0.7132
## Check conditions:
##    father : number of expected successes = 2806 ; number of expected failures =
1128
##    mother : number of expected successes = 1932 ; number of expected failures =
777
## Standard error = 0.011
## Test statistic: Z =  4.314
## p-value =  0
```





The hypothesis test yields a Z-score of 4.314 and a p-value 0. Therefore, we reject the null hypothesis. We have enough evidence to support the alternative hypothes that the proportion of white adults with at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education is LARGER than the proportion of those on the condition that their mother (only mother) completed at least one year of post-secondary education.

We will perform the same analysis for black respondents. From the contingency table below, we can find that 66.3% of black respondents completed at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education and that 65.9% of those completed at least one year of post-secondary education on the condition that their mother (only mother) completed at least one year of post-secondary education.

```
## contingency table for black respondents
black%>%
group_by(pa_one)%>%
summarise(yes=sum(educ_level=='yes'), no=sum(educ_level=='no'), prop_yes=100*sum(ed
uc_level=='yes')/n())
```

```
## # A tibble: 2 x 4
##   pa_one   yes    no prop_yes
##   <chr>  <int> <int>    <dbl>
## 1 father   183    93     66.3
## 2 mother   242   125     65.9
```

Based on this observation, we will conduct a hypothesis test to compare two independent proportions.

Here are the requirements for such a test:

1. Samples are independent within groups, which means random samples and if sampling without replacement, sample size <10% of population. Same scenario as we discussed before, this condition has been met between groups, which means two groups must be independent of each other (not paired). This condition has been met. See the condition in which we set up the categorical variable 'pa_one'.
2. Sample size. Each sample should have at least 10 successes and 10 failures. This condition has also been met. We have large success/failure rates.

We are ready to perform a hypothesis test to compare two independent proportions.

H_0: The proportion of black adults with at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education is EQUAL to the proportion of those on the condition that their mother (only mother) completed at least one year of post-secondary education.

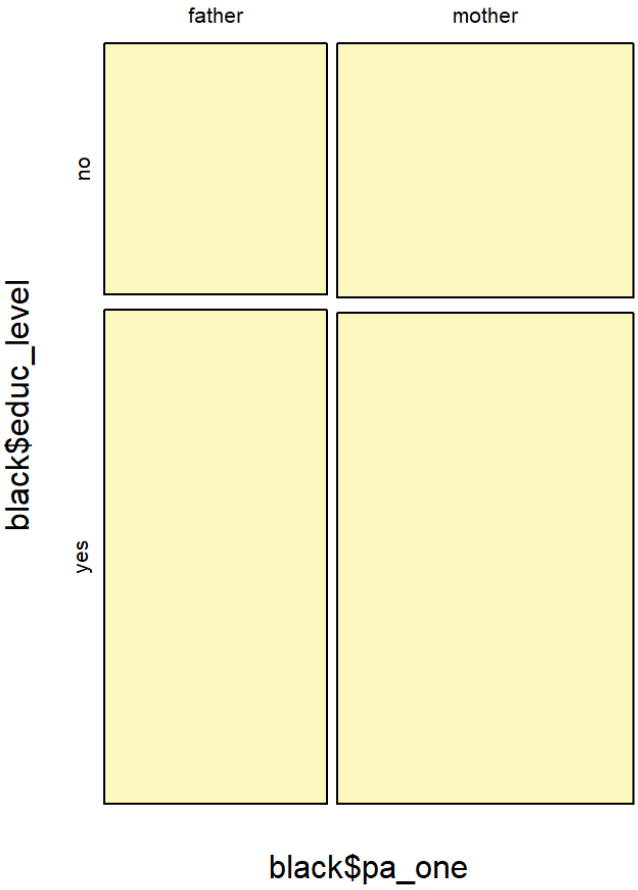
H_A: The proportion is LARGER.

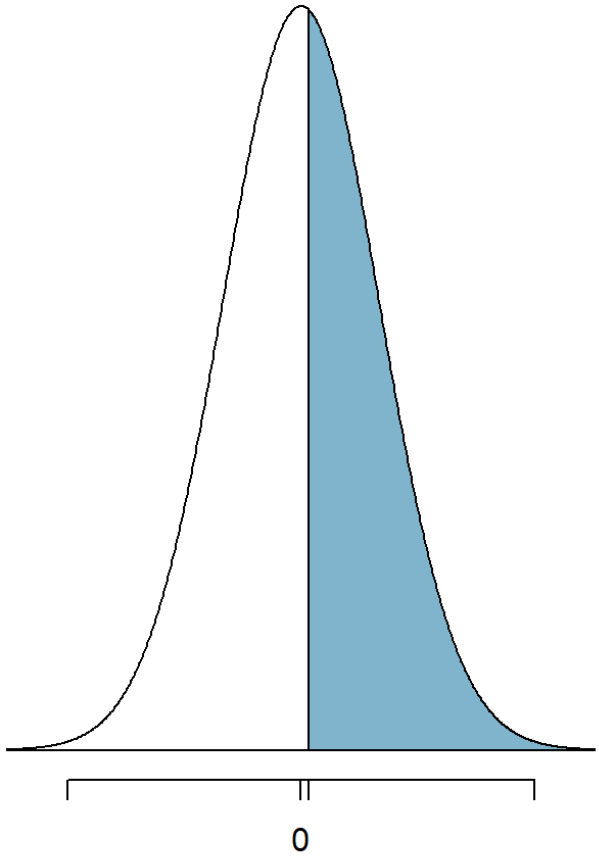
The hypothesis test is performed below:

```
inference(black$educ_level,black$pa_one, est="proportion", type='ht', method= "theo
retical", null=0, alternative="greater", success="yes")
```

```
## Response variable: categorical, Explanatory variable: categorical
## Difference between two proportions -- success: yes
## Summary statistics:
##       x
## y      father mother Sum
##   no       93    125 218
##   yes     183    242 425
##   Sum     276    367 643
```

```
## Observed difference between proportions (father-mother) = 0.0036
## H0: p_father - p_mother = 0
## HA: p_father - p_mother > 0
## Pooled proportion = 0.661
## Check conditions:
##    father : number of expected successes = 182 ; number of expected failures = 9
4
##    mother : number of expected successes = 243 ; number of expected failures = 1
24
## Standard error = 0.038
## Test statistic: Z =  0.097
## p-value =  0.4615
```





The hypothesis test yields a Z-score of 0.097 and a p-value 0.4615. Therefore, we fail to reject the null hypothesis. There is NOT enough evidence to support the alternative hypothesis that the proportion of black adults with at least one year of post-secondary education on the condition that their father (only father) completed at least one year of post-secondary education is LARGER than the proportion of those on the condition that their mother (only mother) completed at least one year of post-secondary education.

Summary of Research Question Part #1, Part #2 and Part #3

The above analysis indicates that

1. Overall People's level of education is associated with their parent's level of education
2. Overall People's level of education has a stronger positive association with their father's level of education than their mother's.
3. The stronger positive association from father applies to white adults but not to black adults.

Discussion

In the analysis of part #3, we did two seperate hypothesis tests in two dataframes (each has two categorical variables with two levels) and concluded that white adults' level of education has a stronger association with their father's level of education than their mother's, but black adults' level education do not have this trend. The probability of a type_1 error (significant level at 0.05 in each test) is 1-0.95x0.95= 9.75%, which is quite high. In order to reduce this error, we either need to design a new test that can infer variables that has subgroups, in our case, variables of gender of parents, race of respondents, with level of education as subgroups. No common inference methods can do such tests. Another option is to decrease significant levels. The probability of a type_1 error (significant level=0.02 in each test) is 1-0.98x0.98= 3.96%.

Further, the large difference between sample size might also cause some error in our conclusion. The sample size of white respondents (n=6643) is more than 10 times that of blacks (n=643). In the hypothesis test comparing two independent proportions, we used the following formula to calculate the Z-score:

$$SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$$

$$Z = \frac{\bar{X} - u}{SE}$$

We can find that the larger the sample size (n_1 and n_2), the larger the Z-score. We have n_1=3934 and n_2 =2709 in the inference of whites, n_1=276 and n_2=367 in blacks. The much greater n_1 and n_2 from white respondents yields a much larger Z-score, thus a much smaller p-value to reject the null hypothesis. We need future sampling design to assign amount of white and black. It would be very intersting to include other