Data Analysis and Statistical Inference - Project Family Size and Education Level

Mariame M

October 16, 2014

1. Introduction

Only children have been the subjects of numerous studies, sometimes stigmatized as spoiled brats and other times as high achievers. Parents of only children can often themselves be stigmatized as being selfish for not giving siblings to their only child. When it comes to the decision of having one or more children, parents try to assess the impact it would have on the current children. They ask themselves whether a bigger family could hurt or not their children financially, emotionally, etc. One of the important impacts is on their chldren's education: will having another child take time away from their current children and impact their studies (and therefore their future)?

In order to give one element of response, we will analyze in this project the relationship between the number of siblings a person has or had, and her/his level of education. The question we will try to answer is thus the following:

Are people with few siblings better educated than people with larger families?

In other word, we are trying to find out whether family size matters when it comes to education.

2. Data Processing

2.1. Background

To answer that question, we will use data from the General Social Survey (GSS) Cumulative File 1972-2012, which provides a sample of selected indicators on contemporary American society.
Detailed information on this file can be found at https://d396qusza40orc.cloudfront.net/statistics%2Fproject%2Fgss1.html.

GSS is a sociological survey, conducted between 1972 and 2012, and where data on demographic characteristics and attitudes of Americans was collected by randomly selected households in the the United States. The dataset contains a total of 57,061 cases (i.e. respondents) and 114 variables. There does not seem to be any bias in the data.

2.2. Study and Scope

Using the GSS data, we will conduct in this document an observational study as the data was collected in a way that does not directly interfere with how the data arose and respondents were randomly selected. Also, this is a retrospective study as we will be using data from the past (1972 to 2012).

In our study, the population of interest is the population of the United States. Since random sampling was used in the survey, we should be able to generalize the results of our study to the population of the United States (at least for the period 10972-2012).

Note that because this is an observational study, only association can be established between the variables, but not causation. To infer causality, we would need to conduct an experiment where we would use random assignment. Given the nature of our variables (education level, number of siblings), it would be very difficult

to lead such an experiment, as researchers cannot control for the education or the number of siblings one person has.

2.3. Variables

From the GSS data, we will use the following variables:
- **sibs**: respondent's number of brothers and sisters - numeric variable;
- **degree**: respondent's highest degree - categorical (ordinal) variable.

Please note, that in the next section of this document, we will introduce a new binary categorical variable ***educ*** in place of the multi-level variable ***degree*** for simplification purposes. It will indeed be easier to first assess whether there is a difference in terms of college versus non-college education. If we see that there indeed a difference, we will then dig further and look wether there is also a difference within a group (college group) as well.

2.4. Data Cleaning

We first load the required libraries:

```
setwd("~/Repositories/Coursera_DataAnalysis_Duke/Project/")

library(reshape2)
library(ggplot2)
library(plyr)

source("http://bit.ly/dasi_inference")
```

We then load the data (and cache it) and get some summary statistics:

```
load(url("http://bit.ly/dasi_gss_data"))

# only keep count of siblings, degree and constant family income
data <- gss[,c("sibs","degree")]

# get statistics
summary(data)
```

```
##       sibs                 degree
##  Min.   : 0.0   Lt High School:11822
##  1st Qu.: 2.0   High School   :29287
##  Median : 3.0   Junior College: 3070
##  Mean   : 3.9   Bachelor      : 8002
##  3rd Qu.: 5.0   Graduate      : 3870
##  Max.   :68.0   NA's          : 1010
##  NA's   :1679
```

We see that both variables contain missing values (NA). In order to prevent this data to create bias in our analysis, we will in those missing values using the following strategy: - ***sibs***: the median number of siblings being 3, we use that number to fill in the NAs in that variable;
- ***degree***: most people have responded "High School" as the highest degree they obtained, we use that value to fill in the NAs in that variable.

```r
# use median number of siblings
sibs.default <- median(data$sibs, na.rm = TRUE)
data[is.na(data$sibs),"sibs"] <- sibs.default

# use "High School"
degree.default <- "High School"
data[is.na(data$degree),"degree"] <- degree.default
```

3. Exploratory Analysis

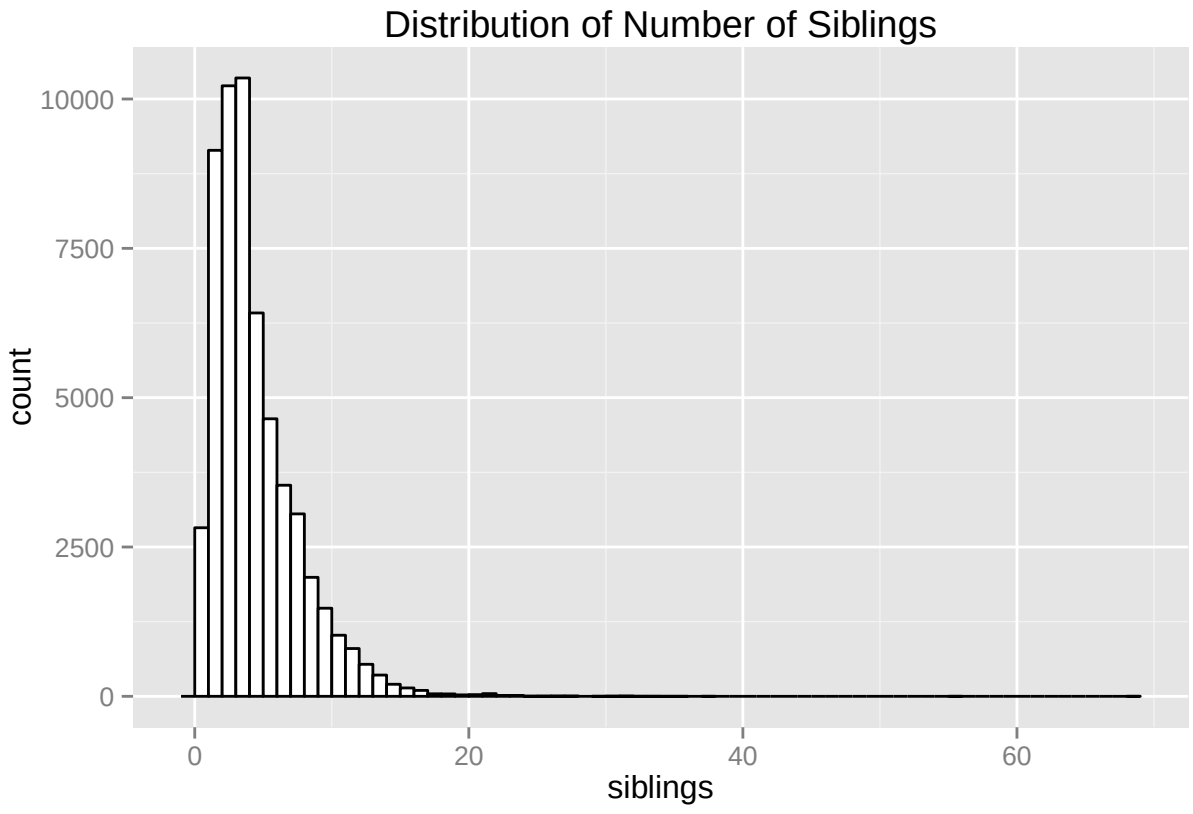
In this section we will draw a few exploratory plots to get a first impression of each variable and the relationships between variables.

3.1. Single Variable Analysis

3.1.1. Number of Siblings

We observe that the distribution for the number of siblings is right-skewed and bounded at zero on the left. Both of these observations are expected as one cannot have a negative number of siblings. Similarly, we do expect the count of respondents to decrease as the number of siblings increases:

```r
ggplot(data, aes(x=sibs)) +
    geom_histogram(binwidth=1, colour="black", fill="white") +
    xlab("siblings") +
    ggtitle("Distribution of Number of Siblings")
```



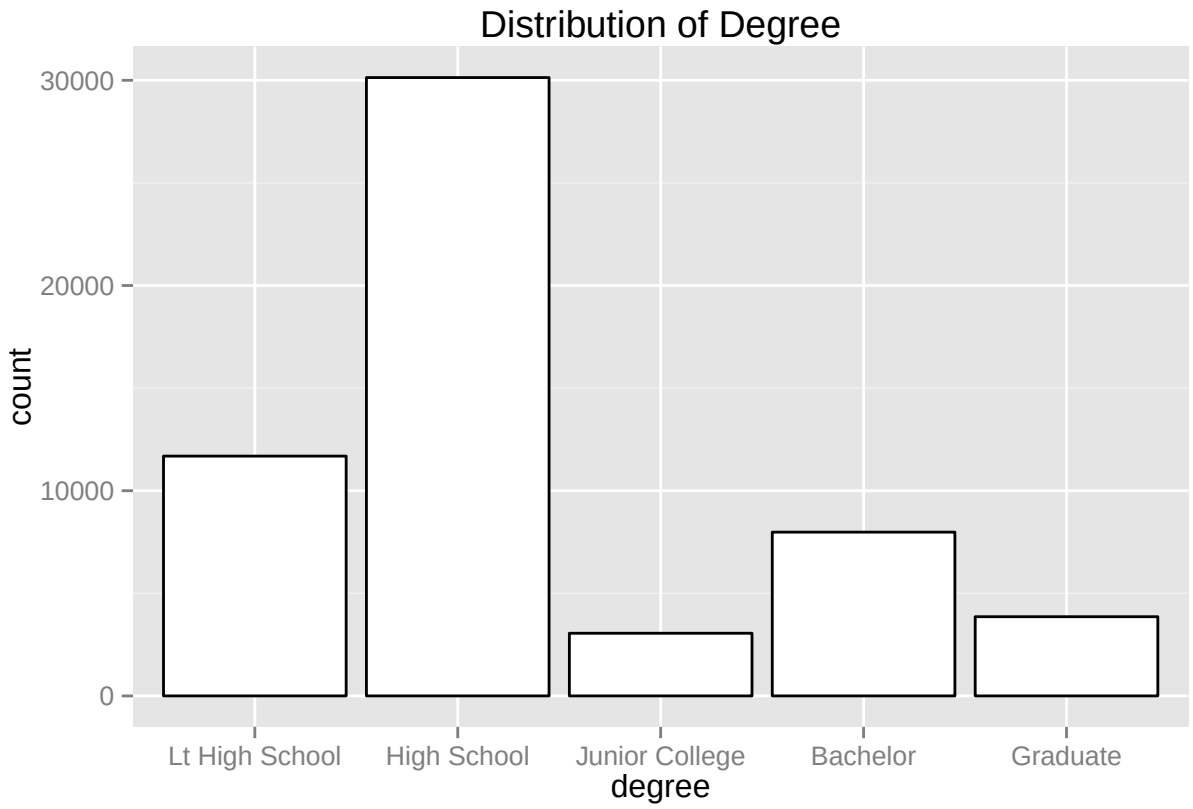
We see that there is a very small number of respondents who have extreme numbers of siblings (well over 60). These data points represents outliers which are likely to skew the resuts of our analysis. Therefore, we remove these entries from our data and will focus on respondents who have 15 or less siblings:

```
data <- data[data$sibs <= 15,]
```

3.1.2. Highest Degree

Most respondents hold a high school diloma as their highest degree:

```
ggplot(data, aes(x=degree)) +
    geom_histogram(color="black", fill="white") +
    ggtitle("Distribution of Degree")
```

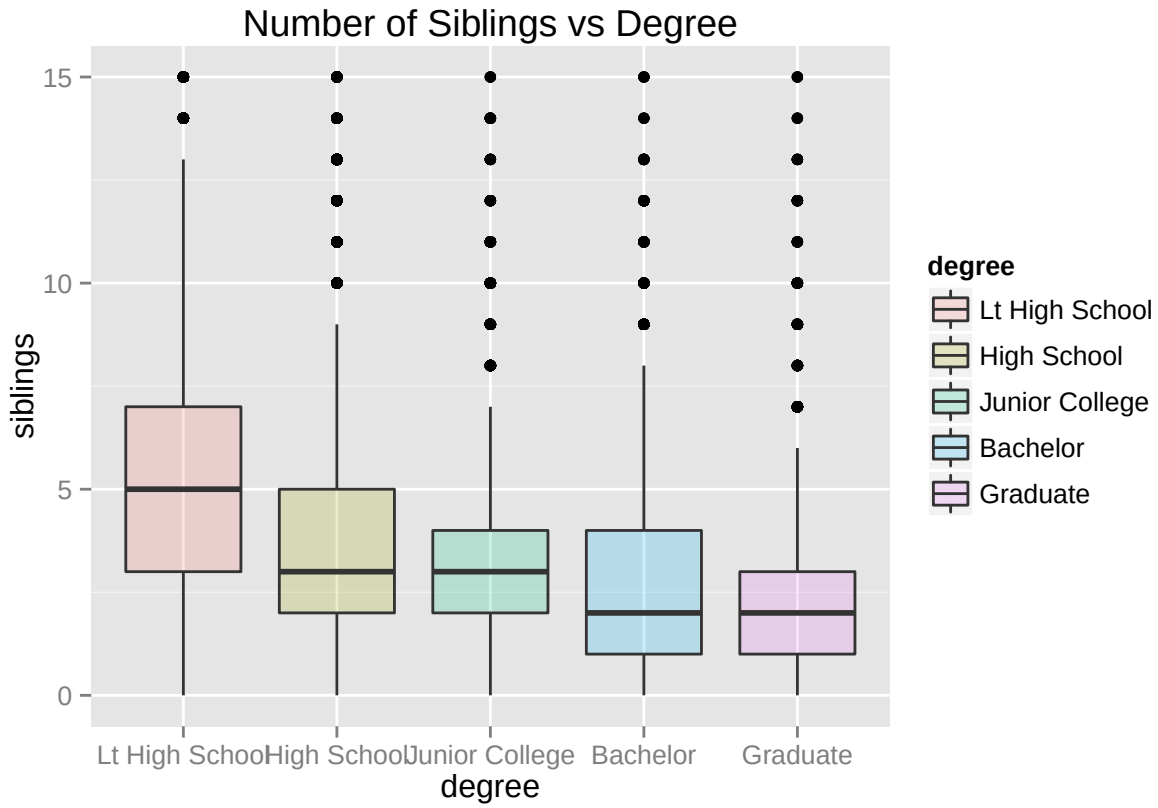


3.2. Two Variable Analysis

3.2.1. Number of Siblings vs Degree

On the boxplots below, we see that as the level of education increases, the number of siblings decreases. For the two highest degrees (bachelor and graduate), we see that while their median number of siblings is similar, their IQR clearly differs, showing a lower IQR as the degree level is higher:

```
ggplot(data, aes(x=degree, y=sibs, fill=degree)) +
    geom_boxplot(alpha=0.2) +
    xlab("degree") +
    ylab("siblings") +
    ggtitle("Number of Siblings vs Degree")
```



This graph seems to show a negative relationship between number of siblings and level of education achieved (i.e. the less siblings, the higher the degree).

3.2.2. Transformation

The variable ***degree*** has 5 levels ranging from 'Lower Than High-School' to 'Graduate'. In order to simplify our analysis, we will transform this multi-level variable into a new binary variable ***educ*** with the possible values 'College' and 'No College'. We will then use this variable in our subsequent statistical inference analysis.

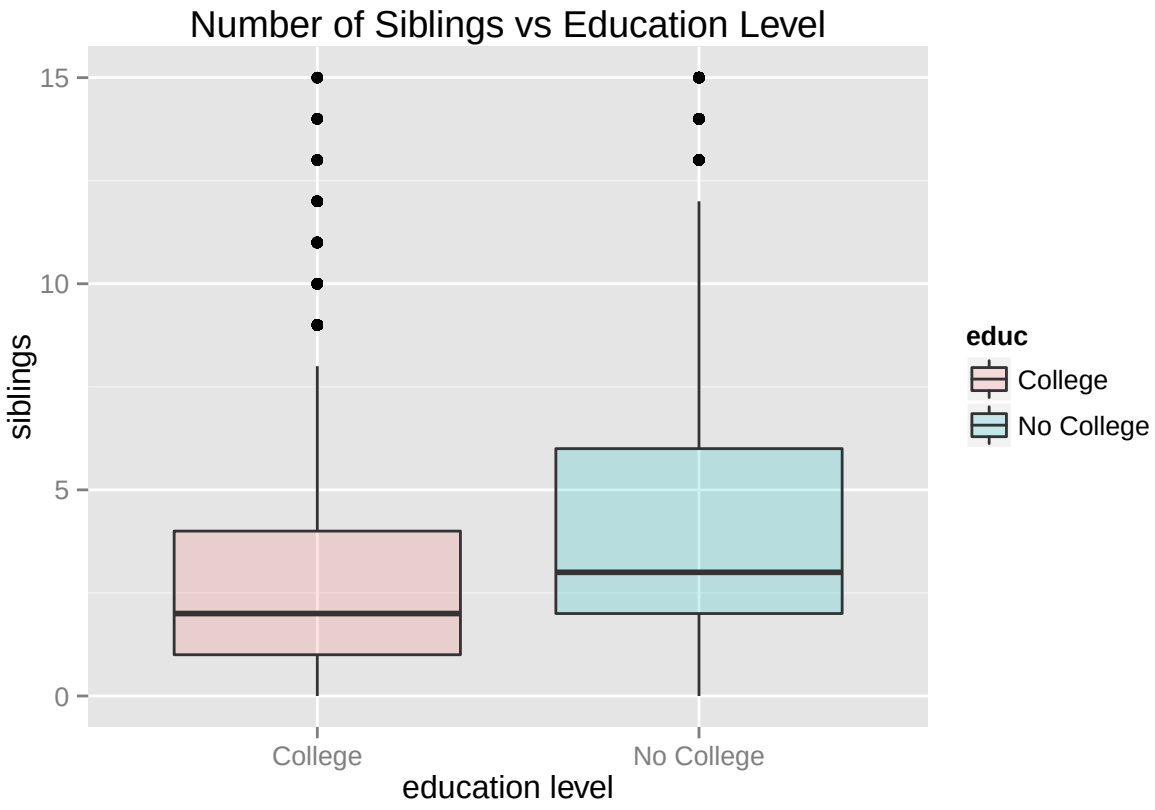
```
# create a new binary variable
college <- c('Junior College','Bachelor','Graduate')
data[data$degree %in% college,"educ"] <- 'College'
data[!(data$degree %in% college),"educ"] <- 'No College'
data$educ <- as.factor(data$educ)

# get updated statistics
summary(data)
```

```
##       sibs                  degree             educ
##  Min.   : 0.00   Lt High School:11686   College   :14894
##  1st Qu.: 2.00   High School   :30132   No College:41818
##  Median : 3.00   Junior College: 3054
##  Mean   : 3.81   Bachelor      : 7981
##  3rd Qu.: 5.00   Graduate      : 3859
##  Max.   :15.00
```

The negative relatinship between family size and education achievement seems more obvious on the following boxplots where we use the binary variable ***educ***:

```r
# box plot the number of siblings against the binary educ variable
ggplot(data, aes(x=educ, y=sibs, fill=educ)) +
    geom_boxplot(alpha=0.2) +
    xlab("education level") +
    ylab("siblings") +
    ggtitle("Number of Siblings vs Education Level")
```

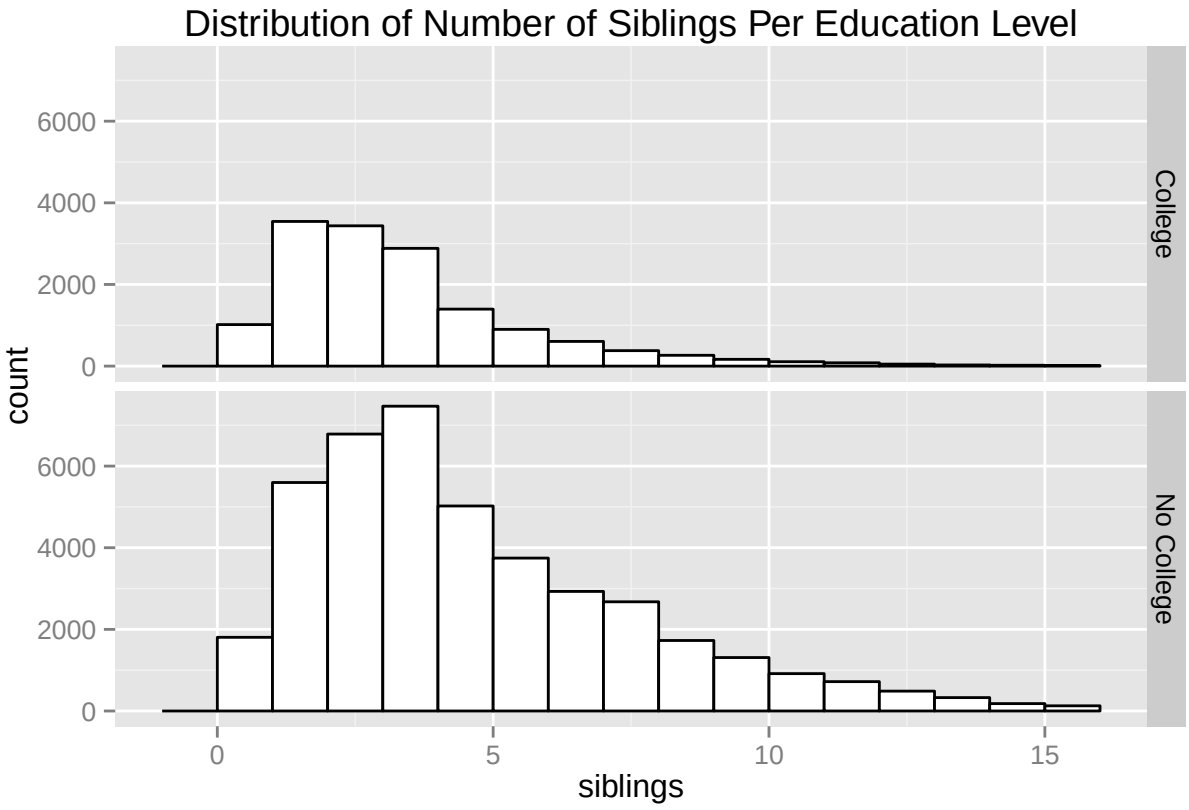


4. Statistical Inference

In this section, we will perform the following statistical inference analysis:
- first by computing 95% confidence intervals for the average number of siblings for non-college and college graduates,
- and second by running a hypothesis test at a 5% significance level with regards to whether there is a difference in number of siblings between non-college and college graduates.

Prior to each computations, we will list and check that each condition for inference is met for the following two distributions:

```r
ggplot(data, aes(x=sibs)) +
    geom_histogram(binwidth=1, colour="black", fill="white") +
    xlab("siblings") +
    ggtitle("Distribution of Number of Siblings Per Education Level") +
    facet_grid(educ ~ .)
```



4.1. Confidence intervals

4.1.1. Average number of siblings for college graduates

Before computing the confidence interval, we need to verify that the conditions for the Central Limit Theroem are met:

- *Independence*: we are dealing with a random sample (as the respondents to the GSS survey were randomly selected) of 14,894 college graduates, which is definitely less than 10% of all Americans with a college degree;
- *Sample size/skew*: our sample distribution is right-skewed (which makes sense since a person can not have a negative number of siblings and we expect to see fewer and fewer respondents as the number of siblings increases), however our sample size is greater than 30.

Now that the conditions for inference are verified, we can compute our confidence interval:

```
# get a subset of college graduates only
data_col <- subset(data, data$educ == 'College')

# determine the sample size n, the sample mean xbar and the standard deviation sd
n_col <- dim(data_col)[1]
xbar_col <- mean(data_col$sibs)
sd_col <- sd(data_col$sibs)

# compute the standard error
se_col <- sd_col / sqrt(n_col)
```

```r
#finally the confidence interval
ci_col <- xbar_col + c(-1, 1) * qnorm(0.975) * se_col
ci_col
```

```
## [1] 2.815 2.888
```

Based on the above results, we conclude that we are 95% confident that Americans with college degree have on average 2.82 to 2.89 siblings.

4.1.2. Average number of siblings for non-college graduates

Before computing the confidence interval, we need to verify that the conditions for the Central Limit Theroem are met:

- *Independence*: we are dealing with a random sample (as the respondents to the GSS survey were randomly selected) of 41,818 non-college graduates, which is definitely less than 10% of all Americans without a college degree;
- *Sample size/skew*: here again our sample distribution is right-skewed, however our sample size is greater than 30.

Now that the conditions for inference are verified, we can compute our confidence interval:

```r
# get a subset of non-college graduates only
data_noc <- subset(data, data$educ == 'No College')

# determine the sample size n, the sample mean xbar and the standard deviation sd
n_noc <- dim(data_noc)[1]
xbar_noc <- mean(data_noc$sibs)
sd_noc <- sd(data_noc$sibs)

# compute the standard error
se_noc <- sd_noc / sqrt(n_noc)

#finally the confidence interval
ci_noc <- xbar_noc + c(-1, 1) * qnorm(0.975) * se_noc
ci_noc
```

```
## [1] 4.130 4.187
```

Based on the above results, we conclude that we are 95% confident that Americans without a college degree have on average 4.13 to 4.19 siblings.

4.2. Hypothesis testing for two means

4.2.1. Hypotheses

The confidence intervals we just computed tend to show that college graduates have on average a lower number of siblings than non-college graduates (about 3 versus 4). In this section, we want to run a hypothesis test at a 5% significance level to prove that college graduates have **less** siblings than non-college graduates.

Our hypotheses are then the following:

- $H_0 : \mu_{col} - \mu_{noc} = 0$
- $H_A : \mu_{col} - \mu_{noc} < 0$

Where μ_{col} is the population mean for the number of siblings college graduates have and μ_{noc} is the population mean for the number of siblings non-college graduates have.

4.2.2. Conditions

Prior to running our test, however, we need to verify that the conditions for inference for comparing two independent means are met:

- *Independence within groups*: the two groups are random samples (as the respondents to the GSS survey were randomly selected) of respectively 14,894 college graduates and 41,818 non-college graduates, which are both respectively less than 10% of all Americans with and without a college degree;
- *Independence between groups*: we are dealing with non-paired data (as the highest level of education attained by a person cannot be college and non-college at the same time), therefore the two groups are independent from each other;
- *Sample size/skew*: both sample distributions are right-skewed, however both groups have a sample size greater than 30.

4.2.3. Test

Now that we checked that our conditions are met, we can directly compute the standard error for the difference of our means, the critical value Z and the associated p-value, using the numbers calculated in section **4.2.**:

```
# compute the standard error of the difference
xbar_col_noc <- xbar_col - xbar_noc
se_col_noc <- sqrt(sd_col^2 / n_col + sd_noc^2 / n_noc)
z <- (xbar_col_noc - 0) / se_col_noc
z
```

```
## [1] -55.19
```

We see that z is very high, so we expect to have a very small p-value:

```
p.value <- pnorm(z, lower.tail = TRUE)
p.value
```

```
## [1] 0
```

Based on this result (i.e. p-value $< 5\%$), we conclude that there is convincing evidence that college graduates have on average fewer siblings than non-college graduates.

4.3. Hypothesis testing for two means

4.3.1. Hypotheses

Now that we have evidence that there is a difference, we can go one step further and find out whether we also see a difference in terms of hisghest degree achieved based on the number of siblings. Becaue here we will be dealing with a 3-level categorical variable (degree will take the values 'Junior College', 'Bachelor' or 'Graduate'), we will need to use the statistical method ANOVA.

Our hypotheses are then the following:

- $H_0 : \mu_{juc} = \mu_{bac} = \mu_{grd}$
- H_A : at least one pair of means are different from each other among the means $\mu_{juc}, \mu_{bac}, \mu_{grd}$

Where:

- μ_{juc} is the population mean for the number of siblings people with a junior college degree have;
- μ_{bac} is the population mean for the number of siblings people with a bachelor degree have;
- μ_{grd} is the population mean for the number of siblings people with a graduate degree have.

4.3.2. Conditions

Prior to running our test, we need to verify that the conditions for inference for comparing multiple independent means are met:

(i) Independence:

- *Within groups*: the three groups are random samples (as the respondents to the GSS survey were randomly selected) of respectively 3,054 junior college degree holders, 7,981 bachelor degree holders and 3,859 graduate degree holders, which are all respectively less than 10% of all Americans with such degrees;
- *Between groups*: again, we are dealing with non-paired data, therefore the three groups are independent from each other;
- *Sample size/skew*: all sample distributions are right-skewed, however all groups have a sample size greater than 30.

(ii) Approximate normality:

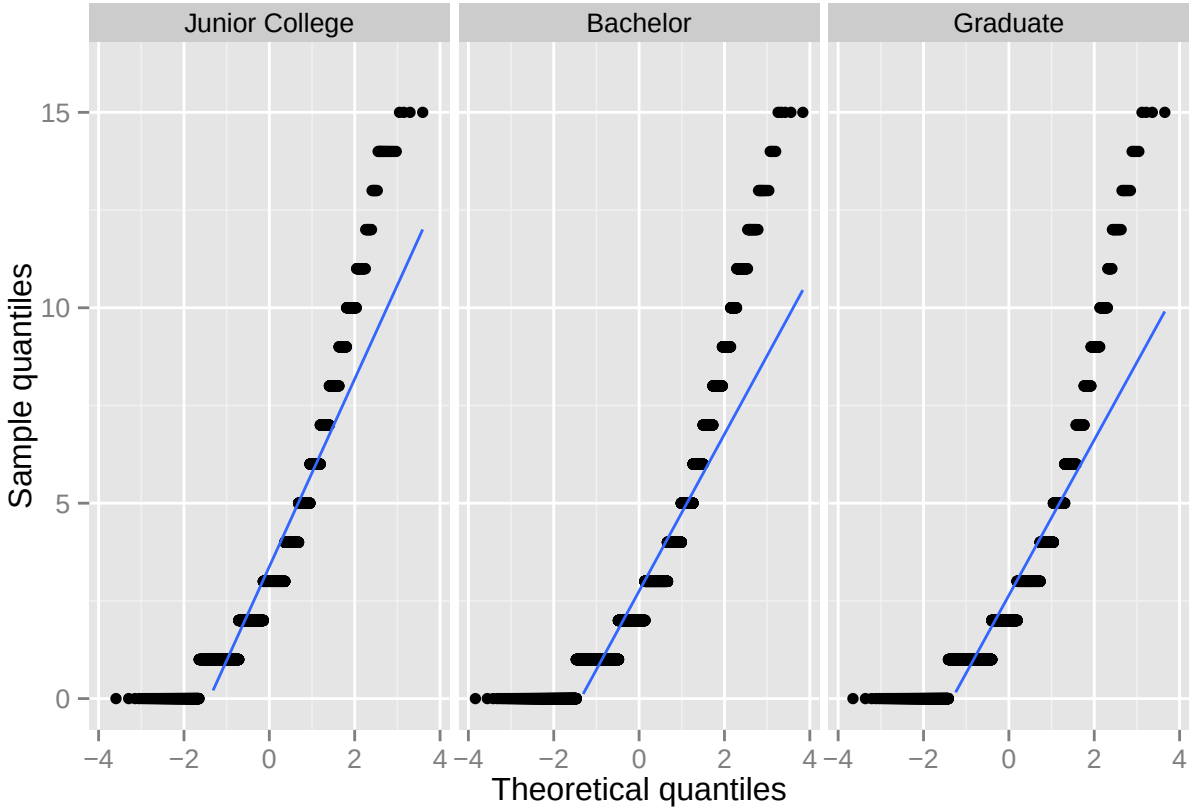
To show normality, we can plot for each group the sample quantiles against the theoritical quantiles and show that both quantities are approximately following a line:

```
# create a new factor to remove unnecessary levels (non-college)
data_col$degree <- factor(data_col$degree, levels = c('Junior College','Bachelor','Graduate'))

# calculate the normal theoretical quantiles per group
data_ply <- ddply(.data = data_col, .variables = .(degree), .fun = function(dat){
    q <- qqnorm(dat$sibs, plot = FALSE)
    dat$theo.sibs <- q$x
    dat
})

# plot the sample values against the theoretical quantiles
ggplot(data = data_ply, aes(x = theo.sibs, y = sibs)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    xlab("Theoretical quantiles") +
    ylab("Sample quantiles") +
    ylim(0, 16) +
    facet_grid(. ~ degree)
```

```
## Warning: Removed 25 rows containing missing values (geom_path).
## Warning: Removed 26 rows containing missing values (geom_path).
## Warning: Removed 26 rows containing missing values (geom_path).
```



(iii) Equal variance:

Finally, as seen on the boxplot in section 3.2.1., variability looks consistent between junior college and graduate levels. We do observe slightly more variability in the bachelor degree group.

While the quantiles plots above do not exactly follow a line and the bachelor group shows a little more variability than the other two groups, we will assume, in the remaining of this study, that our groups are approximately normal and that we have constant variability among groups so that we can perform ANOVA.

4.3.3. Test

In the appendix, we will use the inference method to compute the p-values between our different groups. For illustration purposes, we "manually" compute the hypothesis testing results comparing the mean number of siblings for college degree holders. Using the Bonferroni correction, we run our test at a significance level of 1.67%:

$\alpha^* = \frac{2*\alpha}{k*(k-1)} = (2 * 0.05)/(3 * 2) = 0.05/3 \approx 0.0167$

First, we need to compute a few statistics:

```
n <- dim(data_col)[1]
k <- 3

# compute the global mean number of siblings
ybar <- mean(data_col$sibs)

# gather statistics by group
```

```r
means <- aggregate(data_col$sibs, by = list(data_col$degree), FUN=mean)
colnames(means) <- c("degree","mean")
counts <- count(data_col, c("degree"))
stats <- merge(means, counts)
```

Below, we then go through all the steps allowing us to compute a p-value:

```r
# sum of squares total
sst <- sum((data_col$sibs - ybar)^2)
sst
```

```
## [1] 77200
```

```r
# sum of squares group
ssg <-sum(stats$freq * (stats$mean - ybar)^2)
ssg
```

```
## [1] 1089
```

```r
# sum of squares error
sse <- sst - ssg
sse
```

```
## [1] 76111
```

```r
# degrees of freedom total
dft <- n - 1
dft
```

```
## [1] 14893
```

```r
# degrees of freedom group
dfg <- k - 1
dfg
```

```
## [1] 2
```

```r
# degrees of freedom error
dfe <- dft - dfg
dfe
```

```
## [1] 14891
```

```r
# mean squares group
msg <- ssg / dfg
msg
```

```
## [1] 544.7
```

```r
# mean squares error
mse <- sse / dfe
mse
```

```
## [1] 5.111
```

```r
# F statistic
f <- msg / mse
f
```

```
## [1] 106.6
```

```r
# p-value
pval <- pf(f, dfg, dfe, lower.tail = FALSE)
pval
```

```
## [1] 1.117e-46
```

We see that the p-value is very very small, smaller than our adjusted significance level, therefore we conclude that there is convincing evidence that there is at least one mean among the average numbers of siblings for junior college, bachelor and graduate degrees holders that is different from the other two.

5. Conclusion

In this project, we tried to answer the question : Are people with few siblings better educated than people with larger families? In our statistical inference analysis, we found that indeed people with fewer siblings tend to have a higher level of education than people with more siblings.

While our analysis allows us to conclude that there is a relationship between our two variables, it is important to note that this study does **not** allow us to imply causation between the variables, so we cannot conclude that having a low number of siblings will cause a person to have a better education level.

An interesting study to lead next would be to determine if there is a relationship between levels of happiness and education. As, beyond education achievement, what most parents aspire to is for their children to be happy ...

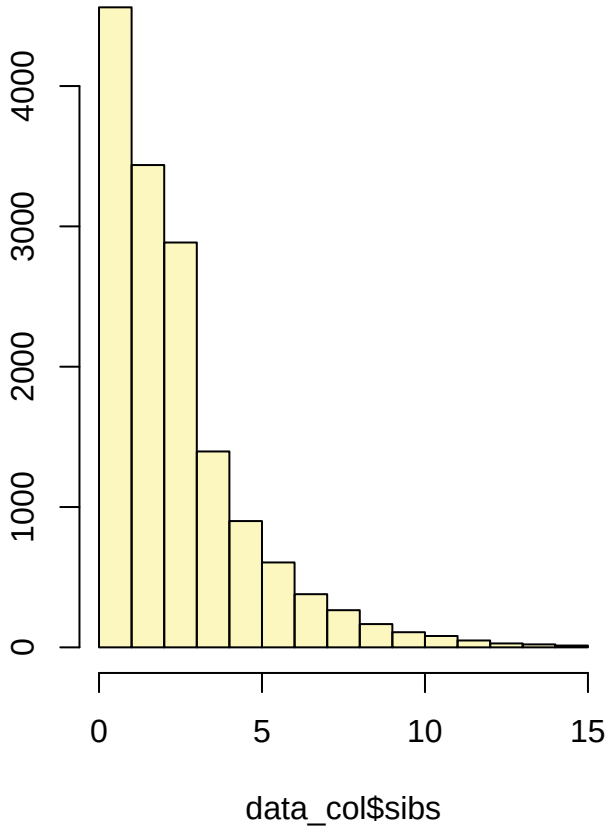
Appendix

In this appendix, we are checking the results of our calculations using the **inference** method:

```r
inference(data_col$sibs, est = "mean", type = "ci", null = 0,
          method = "theoretical")
```

Confidence interval for college degrees

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

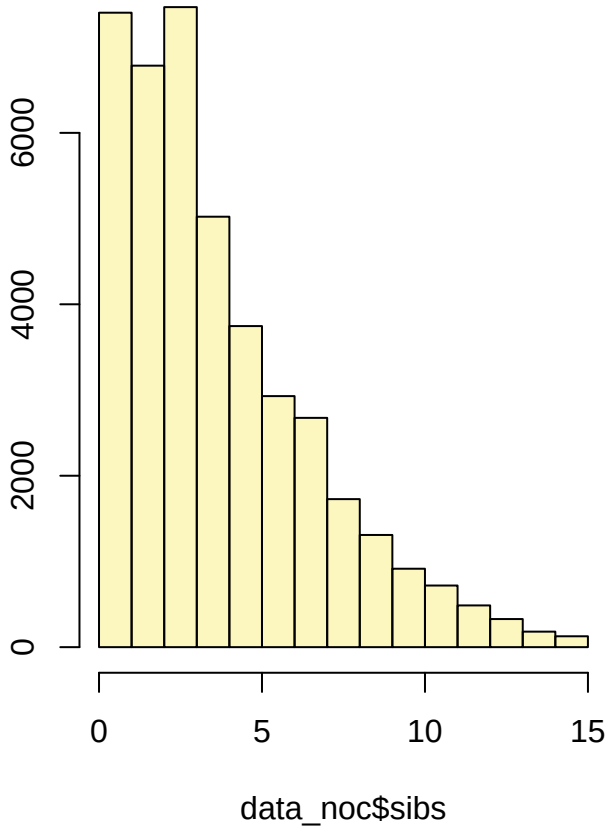



```
## mean = 2.8512 ;  sd = 2.2768 ;  n = 14894
## Standard error = 0.0187
## 95 % Confidence interval = ( 2.8147 , 2.8878 )
```

```
inference(data_noc$sibs, est = "mean", type = "ci", null = 0,
          method = "theoretical")
```

Confidence interval for non-college degrees

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



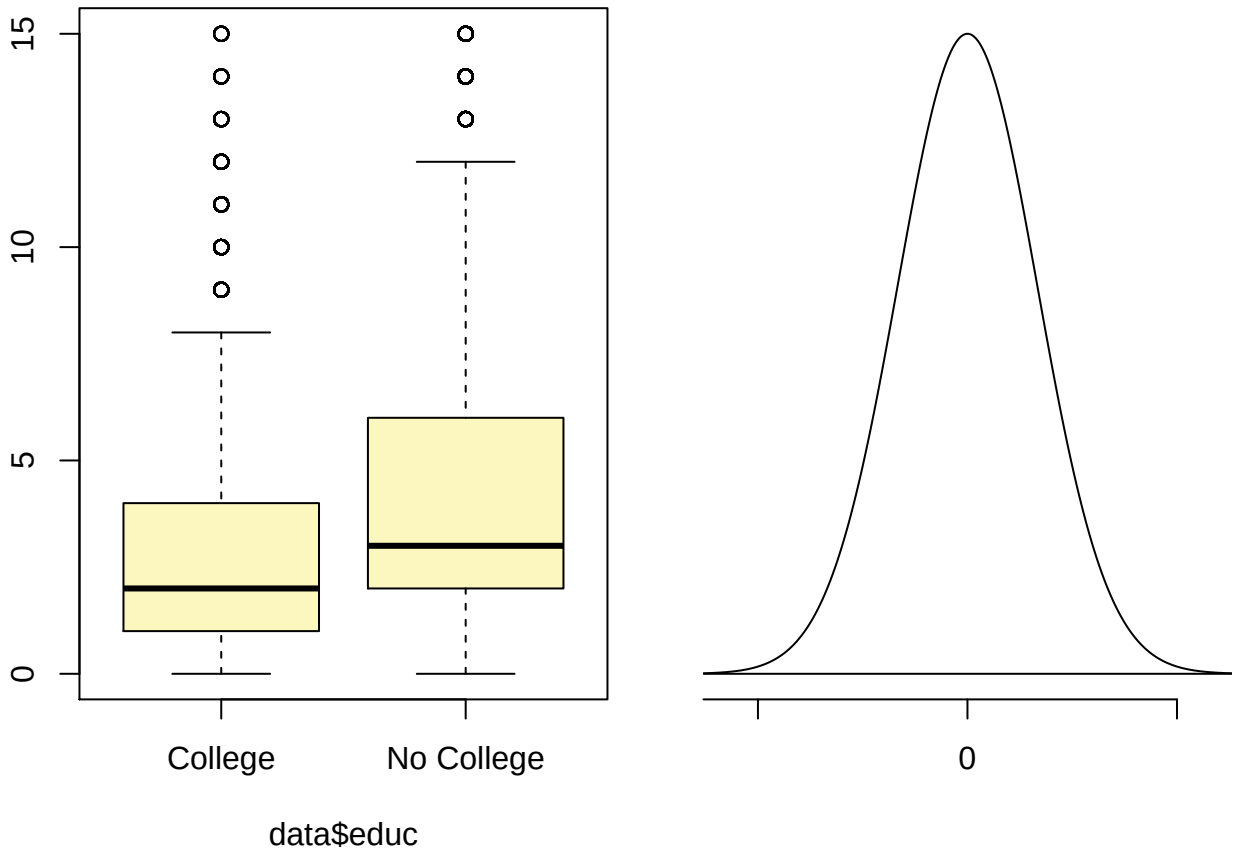

```
## mean = 4.1582 ;  sd = 2.9829 ;  n = 41818
## Standard error = 0.0146
## 95 % Confidence interval = ( 4.1296 , 4.1868 )
```

```
inference(y = data$sibs, x = data$educ, est = "mean", type = "ht",
          null = 0, alternative = "less", method = "theoretical")
```

Hypothesis testing for two means (college versus non-college)

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_College = 14894, mean_College = 2.851, sd_College = 2.277
## n_No College = 41818, mean_No College = 4.158, sd_No College = 2.983

## Observed difference between means (College-No College) = -1.307
## H0: mu_College - mu_No College = 0
## HA: mu_College - mu_No College < 0
## Standard error = 0.024
## Test statistic: Z =  -55.189
## p-value =  0
```




```
inference(y = data_col$sibs, x = data_col$degree, est = "mean", type = "ht",
          method="theoretical", alternative="greater")
```

Hypothesis testing for multiple means among college degrees

```
## Response variable: numerical, Explanatory variable: categorical
## ANOVA
## Summary statistics:
## n_Junior College = 3054, mean_Junior College = 3.376, sd_Junior College = 2.568
## n_Bachelor = 7981, mean_Bachelor = 2.751, sd_Bachelor = 2.174
## n_Graduate = 3859, mean_Graduate = 2.643, sd_Graduate = 2.176

## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##               Df Sum Sq Mean Sq F value Pr(>F)
## x              2   1089     545     107 <2e-16
## Residuals 14891  76111       5
##
## Pairwise tests: t tests with pooled SD
##          Junior College Bachelor
```

```
## Bachelor               0       NA
## Graduate               0   0.0147
```