**Portfolio Project: Option 1**

Hailey Tibbetts

MIS500

5/10/20

Dr. Mills

To start off, I am using the *Popular\_Baby\_Names.csv* from our course resources. I chose this dataset because I felt that I could get useful information out of the data about the genders, names, and most frequently occurring of the two. I wanted to see if I could determining the larger of the two genders with only using the t-test and then graphics. I did not want to just be able to find the length of the subsets because I wanted to test my knowledge of R. I was able to do this with little problems. I use RMarkdown to knit my data to Word which caused a few problems in the process. This is because I didn’t have the exact right information in the code set in order for RMarkdown to run it, even though the code would run in RStudio. Once I got all the bugs fixed, it was smooth sailing.

I was able to do everything I wanted to in this project: testing the boundaries of my knowledge while also recalling what I do well. I filtered my data into new data frames to make my organization cleaner. I was also able to remember how to specify the levels for my bar graph so I could have clear labels. I filtered my “Gender” column into “Male” and “Female” subsets so I could use that information in the t-test to find the greater gender group. I actually expected the female gender to me more prominent in this study just based of my quick glance through the information, so it was surprising to find the male gender is actually more pronounced.

I found it interesting that most of the top ten names were male names. While the male gender is more prominent in the study, I thought that the popularity of names would be more evenly mixed. This shows that there is less originality within the male gender names. From the Tableau bar chart, you can also see that over the years, there is less ingenuity in male names than female names from the amount of top male names.

MIS500 Portfolio Project

Hailey Tibbetts

5/10/2020

# Import Dataset

library(readr)  
Popular\_Baby\_Names <- read\_csv("Downloads/Popular\_Baby\_Names.csv")  
Parsed with column specification:  
cols(  
 `Year of Birth` = col\_double(),  
 Gender = col\_character(),  
 Ethnicity = col\_character(),  
 `Child's First Name` = col\_character(),  
 Count = col\_double(),  
 Rank = col\_double()  
)  
> View(Popular\_Baby\_Names)  
setwd("~/Downloads")

# Summary Statistics

* **Mean Name Count**

mean(Popular\_Baby\_Names$Count)  
[1] 33.52008

* **Maximum Name Count**

max(Popular\_Baby\_Names$Count)  
[1] 426

* **Minimum Name Count**

min(Popular\_Baby\_Names$Count)  
[1] 10

* **Name and Gender Data Frame (First 10 rows)**

Gender<-Popular\_Baby\_Names$Gender  
names<-Popular\_Baby\_Names$`Child's First Name`  
gender\_names<-data.frame("name"=names,"gender"=Gender)  
summary(gender\_names)  
 name gender   
 Avery : 38 FEMALE:9933   
 Riley : 33 MALE :9485   
 Ariel : 30   
 Dylan : 28   
 Jordan : 27   
 Logan : 27   
 (Other):19235   
my\_df<-data.frame(Name=names,Gender=Gender,Count=count)  
my\_df[1:10,]  
 Name Gender Count  
1 GERALDINE FEMALE 12  
2 GIA FEMALE 20  
3 GIANNA FEMALE 22  
4 GISELLE FEMALE 37  
5 GRACE FEMALE 43  
6 GUADALUPE FEMALE 68  
7 HAILEY FEMALE 74  
8 HALEY FEMALE 84  
9 HANNAH FEMALE 90  
10 HAYLEE FEMALE 97

* **Separate Gender into Male/Female Rows**

female<-my\_df[which(my\_df$gender=='FEMALE'),]  
male<-my\_df[which(my\_df$gender=='MALE'),]

* **Dataset to Matrix**

pop.matrix<-as.matrix(Popular\_Baby\_Names)  
fem.matrix<-as.matrix(female)  
mal.matrix<-as.matrix(male)

# Null and Alternative Hypothesis

H0: female.count.mean-male.count.mean=0

HA: female.count.mean-male.count.mean≠0

# Two-Sample T Test

setwd("~/Downloads")  
library(readr)  
Popular\_Baby\_Names <- read\_csv("Popular\_Baby\_Names.csv")

## Parsed with column specification:  
## cols(  
## `Year of Birth` = col\_double(),  
## Gender = col\_character(),  
## Ethnicity = col\_character(),  
## `Child's First Name` = col\_character(),  
## Count = col\_double(),  
## Rank = col\_double()  
## )

Gender<-Popular\_Baby\_Names$Gender  
names<-Popular\_Baby\_Names$`Child's First Name`  
count<-Popular\_Baby\_Names$Count  
my\_df<-data.frame(Name=names,Gender=Gender,Count=count)  
female<-my\_df[which(my\_df$Gender=='FEMALE'),]  
male<-my\_df[which(my\_df$Gender=='MALE'),]  
m.count<-male$Count  
f.count<-female$Count  
levels(my\_df$Gender)<-c('Female','Male')  
t.test<-t.test(x=f.count,y=m.count,alternative="two.sided",conf.level=0.9)  
t.test

##   
## Welch Two Sample t-test  
##   
## data: f.count and m.count  
## t = -17.4, df = 16752, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 90 percent confidence interval:  
## -10.460096 -8.653178  
## sample estimates:  
## mean of x mean of y   
## 28.85201 38.40865

* ***Conclusion of t-test:***

We reject the null hypothesis and accept the alternative hypothesis. The difference between the means are not equal to 0.

* ***Purpose of test:***

The purpose was to determine if there was the same count of males and females in the study. Since the mean of the count of males is larger, there are more males than females in the study.

# Visualizations

* **Select levels**

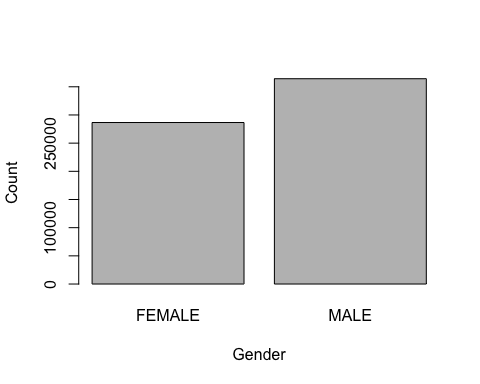
levels(my\_df$Gender)<-c('Female','Male')

* **Barplot for total count of female vs male**

setwd("~/Downloads")  
library(readr)  
Popular\_Baby\_Names <- read\_csv("Popular\_Baby\_Names.csv")

## Parsed with column specification:  
## cols(  
## `Year of Birth` = col\_double(),  
## Gender = col\_character(),  
## Ethnicity = col\_character(),  
## `Child's First Name` = col\_character(),  
## Count = col\_double(),  
## Rank = col\_double()  
## )

Gender<-Popular\_Baby\_Names$Gender  
names<-Popular\_Baby\_Names$`Child's First Name`  
count<-Popular\_Baby\_Names$Count  
my\_df<-data.frame(Name=names,Gender=Gender,Count=count)  
data.for.plot.<-aggregate(my\_df$Count,by=list(my\_df$Gender),FUN=sum)  
names(data.for.plot.)<-c("Gender","Count")  
barplot(data.for.plot.$Count,names.arg = data.for.plot.$Gender,xlab = "Gender",ylab="Count")



We can see here that there is a greater total count of males vs females. This supports the conclusion from the t.test that the means were not equal and that the mean of the male count was higher than the female’s. While the scale of the y-axis may not be helpful in noting the exact values of the populations, just being able to see the size of each bar shows which population is bigger. That is nice for a quick analysis of the information when needing to make a quick decision.

# Tableau Graphics

I wanted to have a clearer view of the data so I decided to go into tableau to create more custom graphs and charts. I made three different graphics to show different aspects of manipulating the data. The first is the “Number of Males vs Females.” This is a pie chart that is showing the proportion of males vs females so easily see the size difference. The next graphic is the “10 Most Popular Names.” This graphic is made to show the most popular names out of the male and female genders. The size and color of the rectangles show the ranking of the name. The last graphic is a graph to show the most popular names per gender, per year. So, we can look at this graph to look at a specific gender and/or year and discover the most popular name for that selection. I really enjoy using tableau, not only because it is easy to use, but because it can create clear visuals for even the most complicated data.

A screenshot of a cell phone

Description automatically generated

Figure 1: Data from Tableau

Resources

Davies, T. M. (2016). *The book of R: A first course in programming and statistics*. San Francisco, CA: No Starch Press.

Department of Health and Mental Hygiene. (2019, June 08). Popular baby Names: NYC open data. Retrieved May 10, 2020, from https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf

R. (n.d.). R Markdown Quick Tour. Retrieved May 09, 2020, from https://rmarkdown.rstudio.com/authoring\_quick\_tour.html

Shmueli, Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2018). *Data mining for business analytics: Concepts, techniques, and applications in r*. Hoboken, NJ: John Wiley & Sons.