87 Final Project

Elliot Cornfeld

11/18/2019

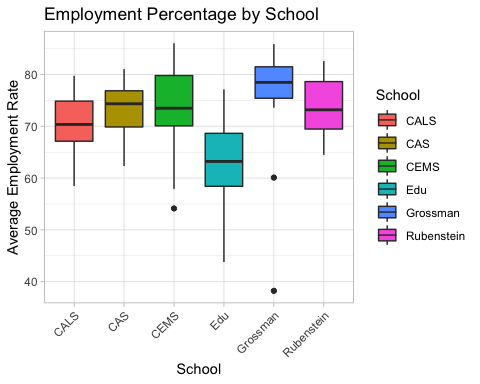
## Introduction

The data set “all-ages.csv” includes a list of all college majors, their category, and the corresponding number of employed/unemployed, unemployment rates, and summary data about salaries. I am interested in looking at the best and worst majors in terms of employment rate and median salary and comparing the percentages of employment to median income. The data came from the American Community Survey 2010-2012. I don’t expect that there is any substantial bias in the data because it is generally common knowledge that specific majors tend to have greater incomes and higher employment rates than others. Also, I found that the sampling the survey used, according to the 2017 PUMS Technical Documentation, is a 1% randomized sample group from every town and region in the country. From this randomization, I can infer that the population sample includes graduates of all ages from all kinds of different backgrounds in order to ensure the absence of any bias. This data was of interest to me because I am close to joining the workforce once I graduate from college. The insights extracted from the data could be relevant to some of the choices I have to make in the coming years.

# Read in dataset  
A <- read.csv("~/Google Drive/COLLEGE/Junior Year UVM/CS087 Scripts/all-ages(1) (1).csv")  
# Has no NA values  
  
# Putting Major Categories into Schools  
CALS <- c("Social Science", "Humanities & Liberal Arts", "Health")  
CAS <- c("Arts", "Psychology & Social Work", "Communications & Journalism",   
 "Law & Public Policy")  
Grossman <- c("Industrial Arts & Consumer Services", "Business")  
Edu <- c("Education")  
CEMS <- c("Engineering", "Computers & Mathematics", "Physical Sciences")  
Rubenstein <- c("Agriculture & Natural Resources", "Biology & Life Science")  
  
A <- A %>% mutate(CALS = ifelse(Major\_category %in% CALS, 1, 0),  
 CAS = ifelse(Major\_category %in% CAS, 1, 0),  
 Grossman = ifelse(Major\_category %in% Grossman, 1, 0),  
 Edu = ifelse(Major\_category %in% Edu, 1, 0),  
 CEMS = ifelse(Major\_category %in% CEMS, 1, 0),  
 Rubenstein = ifelse(Major\_category %in% Rubenstein, 1, 0))  
A <- A %>% mutate(School = ifelse(CALS == 1, "CALS",   
 ifelse(CAS == 1, "CAS",  
 ifelse(Grossman == 1, "Grossman",  
 ifelse(Edu == 1, "Edu",  
 ifelse(CEMS == 1, "CEMS",  
 ifelse(Rubenstein == 1, "Rubenstein",NA)))))))  
  
# No NA values for school  
A <- A %>% filter(!is.na(School))  
  
# Making percentages of employment:   
A <- A %>% mutate(PercEmpl = 100\*Employed/Total, PercEmplYr = 100\*Employed\_full\_time\_year\_round/Total,  
 PercYREmpl = 100\*Employed\_full\_time\_year\_round/Employed)

## Graph 1

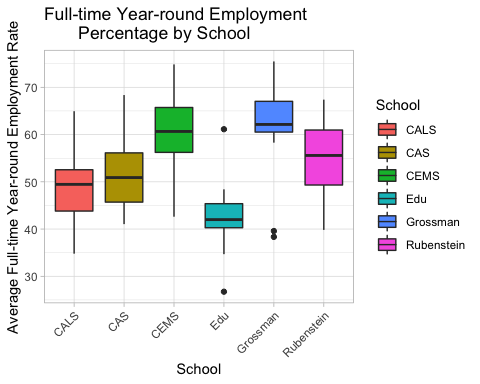
# Boxplot of Employment Percentages by Schools  
ggplot(data = A,  
 mapping = aes(x = School, y = PercEmpl, fill = School))+  
 geom\_boxplot() +  
 labs(title = "Employment Percentage by School",  
 x = 'School',  
 y= 'Average Employment Rate',  
 color = 'School') +  
 theme\_light() +  
 theme(axis.text.x=element\_text(angle=45,hjust=1))



Graph 1 shows the relationship between the school the major categories are part of and the average employment rates. The school that has the highest mean employment percentages is the Grossman School of Business. I was suprised to see that CEMS, where math, science, and engineering are studied, was not also higher. The other schools all have similar mean employment rates except for Education. The average employment rate of education students is between 60% and 65%.

## Graph 2

# Boxplot of Year Round Employment Percentages by Schools  
ggplot(data = A,  
 mapping = aes(x = School, y = PercEmplYr,  
 fill = School))+  
 geom\_boxplot() +  
 labs(title = "Full-time Year-round Employment   
 Percentage by School",  
 x = 'School',  
 y= 'Average Full-time Year-round Employment Rate',  
 color = 'School') +  
 theme\_light() +  
 theme(axis.text.x=element\_text(angle=45,hjust=1))

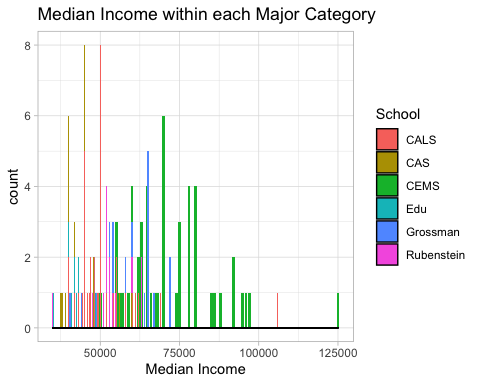


Graph 2 shows the average rate of employment into full-time, year-round jobs for each school. It is clear that the majority of schools have similar rates, however CEMS, Grossman, and Education all stand out. CEMS and Grossman have higher rates of employment into year-round jobs than the other schools. Education has the opposite relationship, it has lower rates of employment into full-time, year-round jobs. This graph is not quite the same as the first graph comparing overall employment rates of each school, which is interesting. There appears to be more spread among the average rates in this graph than the first one. The means are also shifted downward from the first one. This is probably because some of the employed have part-time jobs. It would be interesting to see the difference in proportions of full-time employment rates to total employment rates for each school.

## Graph 3

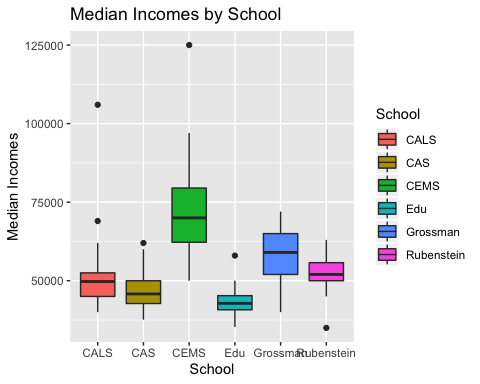
ggplot(data = A, mapping = aes(x = Median, fill = School)) +  
 geom\_bar() + geom\_density(alpha=.5) + theme\_light() +  
 labs(title = "Median Income within each Major Category",  
 x = "Median Income",  
 fill = "School")

## Warning: position\_stack requires non-overlapping x intervals



Graph 3 shows the relationship between college majors (by category) and median income. Each major category is color filled and corresponds to certain bars on the graph. As evident by the data, there is a slight skew to the right, which is understandable because some of the majors make more money than the center of the graph relative to the underachieving majors. The graph allows me to see that those majors which exceed in median income tend to be in the engineering and computers/mathematics categories. In contrast, I can see that some of the majors that have lower median incomes are among the arts and psychology/social work categories.

ggplot(data = A,  
 mapping = aes(x = School, y = Median, fill = School))+  
 geom\_boxplot() +  
 labs(title = "Median Incomes by School",  
 x = "School",  
 y = "Median Incomes",  
 fill = "School")



Graph 3 shows the relationship between colleges and median income. Each school is color filled. This graph shows which schools have higher incomes on average. CEMS has significantly higher median incomes than all the other schools. Grossman also has higher median incomes, though not as high as CEMS. This means that major categories such as engineering and business have higher median incomes than major categories such as education, art or journalism. These differences are understandable considering the jobs our society places value on.

## Confidence Intervals of Salaries

#Rubenstein:  
AvgR1 <- A %>% filter(School == 'Rubenstein')  
AvgR2 <- (AvgR1$Median)  
PE <- mean(AvgR2)  
print(paste("Point Estimate:", PE))

## [1] "Point Estimate: 52562.5"

BootstrapMeansR <- c()  
for (i in 1:1000) {  
 BootSam <- sample(x=AvgR2, size=length(AvgR2), replace=TRUE)  
 BootstrapMeansR <- c(BootstrapMeansR, mean(BootSam))}  
  
SE <- sd(BootstrapMeansR)  
ME <- 2\*SE  
UpLim <- PE + ME  
LowLim <- PE - ME  
UpLim <- round(UpLim, 1)  
LowLim <- round(LowLim, 1)  
print(paste("I am 95% confident that the true mean median",  
 "income for those in the Rubenstein school is between",  
 LowLim, "to", UpLim))

## [1] "I am 95% confident that the true mean median income for those in the Rubenstein school is between 50026.7 to 55098.3"

#Grossman:  
AvgR3 <- A %>% filter(School == 'Grossman')  
AvgR4 <- (AvgR3$Median)  
PE <- mean(AvgR4)  
print(paste("Point Estimate:", PE))

## [1] "Point Estimate: 57825"

BootstrapMeansR <- c()  
for (i in 1:1000) {  
 BootSam <- sample(x=AvgR4, size=length(AvgR4), replace=TRUE)  
 BootstrapMeansR <- c(BootstrapMeansR, mean(BootSam))}  
  
SE <- sd(BootstrapMeansR)  
ME <- 2\*SE  
UpLim <- PE + ME  
LowLim <- PE - ME  
UpLim <- round(UpLim, 1)  
LowLim <- round(LowLim, 1)  
print(paste("I am 95% confident that the true mean median",  
 "income for those in the Grossman school is between",  
 LowLim, "to", UpLim))

## [1] "I am 95% confident that the true mean median income for those in the Grossman school is between 53589.7 to 62060.3"

#Education:  
AvgR5 <- A %>% filter(School == 'Edu')  
AvgR6 <- (AvgR5$Median)  
PE <- mean(AvgR6)  
print(paste("Point Estimate:", PE))

## [1] "Point Estimate: 43831.25"

BootstrapMeansR <- c()  
for (i in 1:1000) {  
 BootSam <- sample(x=AvgR6, size=length(AvgR6), replace=TRUE)  
 BootstrapMeansR <- c(BootstrapMeansR, mean(BootSam))}  
  
SE <- sd(BootstrapMeansR)  
ME <- 2\*SE  
UpLim <- PE + ME  
LowLim <- PE - ME  
UpLim <- round(UpLim, 1)  
LowLim <- round(LowLim, 1)  
print(paste("I am 95% confident that the true mean median",  
 "income for those in the Education school is between",  
 LowLim, "to", UpLim))

## [1] "I am 95% confident that the true mean median income for those in the Education school is between 41292.8 to 46369.7"

#CEMS:  
AvgR7 <- A %>% filter(School == 'CEMS')  
AvgR8 <- (AvgR7$Median)  
PE <- mean(AvgR8)  
print(paste("Point Estimate:", PE))

## [1] "Point Estimate: 72160"

BootstrapMeansR <- c()  
for (i in 1:1000) {  
 BootSam <- sample(x=AvgR8, size=length(AvgR8), replace=TRUE)  
 BootstrapMeansR <- c(BootstrapMeansR, mean(BootSam))}  
  
SE <- sd(BootstrapMeansR)  
ME <- 2\*SE  
UpLim <- PE + ME  
LowLim <- PE - ME  
UpLim <- round(UpLim, 1)  
LowLim <- round(LowLim, 1)  
print(paste("I am 95% confident that the true mean median",  
 "income for those in CEMS is between",  
 LowLim, "to", UpLim))

## [1] "I am 95% confident that the true mean median income for those in CEMS is between 68228.4 to 76091.6"

#CAS:  
AvgR9 <- A %>% filter(School == 'CAS')  
AvgR10 <- (AvgR9$Median)  
PE <- mean(AvgR10)  
print(paste("Point Estimate:", PE))

## [1] "Point Estimate: 46584.6153846154"

BootstrapMeansR <- c()  
for (i in 1:1000) {  
 BootSam <- sample(x=AvgR10, size=length(AvgR10), replace=TRUE)  
 BootstrapMeansR <- c(BootstrapMeansR, mean(BootSam))}  
  
SE <- sd(BootstrapMeansR)  
ME <- 2\*SE  
UpLim <- PE + ME  
LowLim <- PE - ME  
UpLim <- round(UpLim, 1)  
LowLim <- round(LowLim, 1)  
print(paste("I am 95% confident that the true mean median",  
 "income for those in CAS is between",  
 LowLim, "to", UpLim))

## [1] "I am 95% confident that the true mean median income for those in CAS is between 44224.7 to 48944.5"

#CALS:  
AvgR11 <- A %>% filter(School == 'CALS')  
AvgR12 <- (AvgR11$Median)  
PE <- mean(AvgR12)  
print(paste("Point Estimate:", PE))

## [1] "Point Estimate: 51325"

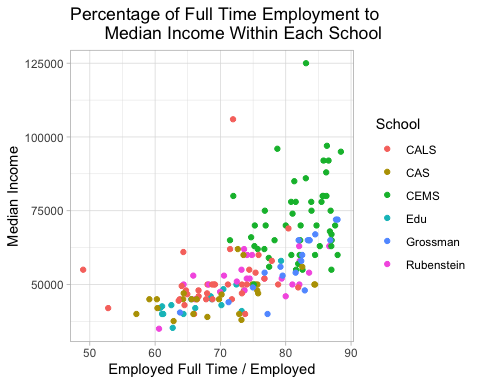
BootstrapMeansR <- c()  
for (i in 1:1000) {  
 BootSam <- sample(x=AvgR12, size=length(AvgR12), replace=TRUE)  
 BootstrapMeansR <- c(BootstrapMeansR, mean(BootSam))}  
  
SE <- sd(BootstrapMeansR)  
ME <- 2\*SE  
UpLim <- PE + ME  
LowLim <- PE - ME  
UpLim <- round(UpLim, 1)  
LowLim <- round(LowLim, 1)  
print(paste("I am 95% confident that the true mean median",  
 "income for those in CALS is between",  
 LowLim, "to", UpLim))

## [1] "I am 95% confident that the true mean median income for those in CALS is between 47606.3 to 55043.7"

These Confidence Intervals further enforces the differences seen between college’s median incomes. CEMS has the highest confidence interval, with its lowest value of median income being greater than the highest value of median income for the schools: Rubenstein, Grossman, Education, CAS, and CALS. Education has the lowest confidence interval, with its highest value of median income being less than the lowest value of median income for the schools: Rubenstein, Grossman, CEMS, and CALS.

## Graph 4

A <- A %>% filter(PercYREmpl <= 100)  
  
ggplot(data = A, mapping = aes(x = PercYREmpl, y = Median, color = School))+  
 geom\_point()+  
 labs(title = "Percentage of Full Time Employment to   
 Median Income Within Each School",  
 x = "Employed Full Time / Employed",  
 y = "Median Income",  
 color = "School")+  
 theme\_light()



Graph 4 shows the insights about how the full-time employment / total employed percentage might correlate with median income. I did a scatterplot to plot each major within each School at UVM and then color-coded it according to the code. By looking at the scatterplot, you can see that Schools such as CAS tend to have lower rates of full-time employment as well as lower median incomes. On the other side of the spectrum, Schools such as CEMS tend to have higher full-time employment rates as well as higher median incomes. These insights from this graph are not surprising since majors within CEMS are in high demand while people with majors in CAS may have a harder time finding a full-time gig.

## Conclusions

In the beginning of my analysis of the “all-ages.csv” dataset, I believed that more technically inclined majors such as engineering and computer science, belonging to CEMS, would have higher median incomes on average as well as higher employment rates. According to the data and the graphs, my assumptions are correct as each one shows that CEMS, whose majors are more technically inclined, have higher incomes and higher employment rates. The reason for this could be because there is more demand for talented people in these professions as well as more financial backing to pay higher incomes to these people. Business students that belong to Grossman also had the highest employment rates and median incomes.

## Limitations and Recommendation

An example of a limitation in my study is that the number of years out of college affects almost all of the variables in the dataset. The more years since graduation, the higher the income you will have and the greater chance you will have of securing a full time job. I do not know if the researchers who compiled the data accounted for a somewhat equal amount of years out of college for each major category. This means that there could be discrepancies in the data if they took information from recent grads in one category and grads 5-10 years out of college in another. An idea for future research that would build on this project would be to figure out what the factors are that drive the high median incomes and high employment rates for CEMS majors and vice versa for CAS majors.