# CS544Final_Heng

Kyle Heng

2024-02-27

## Importing Data

The data being used for this analysis is a list of college majors with information pertaining to sex, earnings, and type of employment of the people who studied each major.

```
recent.grads <- read.csv("~/BU School Files/CS544//recent-grads.csv", header=T)
```

### Importing Process:

- Downloaded raw csv file from source on Github
- Imported file from my directory using read.csv
- Set header = TRUE since column names are included in the original file.

## Categorical Variable Analysis

### This is a histogram of the frequencies of each Major Category being represented in the data.

```
library(plotly)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
fig <- plot_ly(y = recent.grads$Major_category, type = "histogram")
fig
```

From the bar plot, we can see an overwhelmingly large amount of representation from engineering majors. Types of majors that are under represented in this data can be categorized under communications/journalism or interdisciplinary majors.

## Numerical Variable Analysis

### This is a bar plot of unemployment rates for each of the majors represented in the data.

```
fig <- plot_ly(x = recent.grads$Major,y = recent.grads$Unemployment_rate, type = "bar")
fig
```

At a glance, we can see that majors with the highest rates of unemployment in the data include Nuclear Engineering, Clinical Psychology, Computer Networking/Telecommunications, and Public Administration.

### Majors of Top 10 Unemployment Rates:

```
top <- tail(sort(recent.grads$Unemployment_rate),10)
topsub <- subset(recent.grads, Unemployment_rate %in% top)
select(topsub, c('Major', 'Unemployment_rate'))
```

```
##                                       Major Unemployment_rate
## 2               MINING AND MINERAL ENGINEERING         0.1172414
## 6                           NUCLEAR ENGINEERING         0.1772264
## 30                               PUBLIC POLICY         0.1284263
## 54    COMPUTER PROGRAMMING AND DATA PROCESSING         0.1139826
## 59                                ARCHITECTURE         0.1133319
## 80                                   GEOGRAPHY         0.1134586
## 85  COMPUTER NETWORKING AND TELECOMMUNICATIONS         0.1518498
## 90                        PUBLIC ADMINISTRATION         0.1594906
## 106                 COMMUNICATION TECHNOLOGIES         0.1195115
## 171                        CLINICAL PSYCHOLOGY         0.1490482
```

### Majors of Lowest 10 Unemployment Rates:

```
bottom <- head(sort(recent.grads$Unemployment_rate),10)
bottomsub <- subset(recent.grads, Unemployment_rate %in% bottom)
select(bottomsub, c('Major', 'Unemployment_rate'))
```

```
##                                       Major Unemployment_rate
## 1                          PETROLEUM ENGINEERING       0.018380527
## 15   ENGINEERING MECHANICS PHYSICS AND SCIENCE       0.006334343
## 20                               COURT REPORTING       0.011689692
## 53            MATHEMATICS AND COMPUTER SCIENCE       0.000000000
## 65                              GENERAL AGRICULTURE       0.019642463
## 74                          MILITARY TECHNOLOGIES       0.000000000
## 84                                       BOTANY       0.000000000
## 113                                SOIL SCIENCE       0.000000000
## 120              MATHEMATICS TEACHER EDUCATION       0.016202835
## 121 EDUCATIONAL ADMINISTRATION AND SUPERVISION       0.000000000
```

# Bivariate Analysis

## Analysis of Relationships between Major Category/Sex and Major Category/Unemployment Rate:

Box plots of Proportion of Women per Major Category

```
fig <- plot_ly(recent.grads, y = ~ShareWomen, color = ~Major_category, type = "box")
fig
```

```
## Warning: Ignoring 1 observations
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for palette Set2 is 8
## Returning the palette you asked for with that many colors
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for palette Set2 is 8
## Returning the palette you asked for with that many colors
```

From the box plots, we can see that the types of majors with the highest average percentage of women include Arts, Health and Education. Computer/Mathematics and Engineering majors have much lower female representation in the data. Industrial Arts/Consumer Services majors also have lower female representation but they have the largest standard deviation of proportions.

Box plots of Unemployment Rate per Major Category

```
fig <- plot_ly(recent.grads, y = ~Unemployment_rate, color = ~Major_category, type = "box")
fig
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for palette Set2 is 8
## Returning the palette you asked for with that many colors
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for palette Set2 is 8
## Returning the palette you asked for with that many colors
```

Major Categories with the highest average unemployment rate include Social Sciences, Arts, and Computer & Mathematics. Physical Sciences and Education Majors have less unemployment on average while Law and Public Policy majors have the largest standard deviation of unemployment rate.

## Examining Distribution of Median Earnings of Year-Round Full-Time Workers

```
x <- recent.grads$Median
fig <- plot_ly(x=x, type = "histogram")
fig
```

```
mean(x)
```

```
## [1] 40151.45
```

```
sd(x)
```

```
## [1] 11470.18
```

From the histogram, we can see that the median earnings are right skewed. Therefore, a majority of majors represented by the data will make less than the mean value of $40.151.45 with a select few making a much higher amount.

## Drawing Random Samples of Various Sizes From Median Earnings Data

### Drawing 1000 samples of sample sizes 20,40,60,80:

```
samples <- 1000
xbar20 <- numeric(samples)
xbar40 <- numeric(samples)
xbar60 <- numeric(samples)
xbar80 <- numeric(samples)

x <- recent.grads$Median
for (i in 1:samples) {
  xbar20[i] <- mean(sample(x, size = 20, replace = F))
  xbar40[i] <- mean(sample(x, size = 40, replace = F))
  xbar60[i] <- mean(sample(x, size = 60, replace = F))
  xbar80[i] <- mean(sample(x, size = 80, replace = F))
}
```

### Histograms of the Sample Means For Each Sample Size:

```
fig1 <- plot_ly(x=xbar20, type = "histogram", name = "n=20")
fig2 <- plot_ly(x=xbar40, type = "histogram", name = "n=40")
fig3 <- plot_ly(x=xbar60, type = "histogram", name = "n=60")
fig4 <- plot_ly(x=xbar80, type = "histogram", name = "n=80")
```

```
fig <- subplot(fig1, fig2, fig3, fig4, nrows = 2) %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(
           zerolinecolor = '#ffff',
           zerolinewidth = 2,
           gridcolor = 'ffff'),
         yaxis = list(
           zerolinecolor = '#ffff',
           zerolinewidth = 2,
           gridcolor = 'ffff'))
fig
```

The shape of the distribution starts off right skewed with sample size 20, similar to the original data. As larger samples are taken, the distribution becomes less skewed and follows a more symmetrical bell-curve. This is very apparent in the histograms of sample sizes 60 and 80. These findings fall in line with the Central Limit Theorem, supporting the notion that the distribution of sample means will always be normal, given that the sample size is large enough.

## Sampling with Other Types of Sampling Methods

### A sample size of 60 will be taken for all methods.

### Simple Random Sample with Replacement

```
library(sampling)
s <- srswr(60, nrow(recent.grads))

srs <- recent.grads[s != 0,]
```

### Systematic Sampling with Unequal Probabilities

```
pik <- inclusionprobabilities(recent.grads$Median, 60)

s <- UPsystematic(pik)
sys<- recent.grads[s != 0,]
```

## Histograms Comparing Whole Data with Various Samples

```
fig1 <- plot_ly(x=recent.grads$Median, type = "histogram", name = "Original Data")
fig2 <- plot_ly(x=xbar60, type = "histogram", name = "sample() function n=60")
fig3 <- plot_ly(x=srs$Median, type = "histogram", name = "Simple Random Sampling")
fig4 <- plot_ly(x=sys$Median, type = "histogram", name = "Systematic Sampling")

fig <- subplot(fig1, fig2, fig3, fig4, nrows = 2) %>%
  layout(plot_bgcolor='#e5ecf6',
```

```
         xaxis = list(
           zerolinecolor = '#ffff',
           zerolinewidth = 2,
           gridcolor = 'ffff'),
         yaxis = list(
           zerolinecolor = '#ffff',
           zerolinewidth = 2,
           gridcolor = 'ffff'))
fig
```

Given a sample size of 60, we can see that the sample() function draws samples that follow the normal distribution more closely that simple random sampling or systematic sampling. Systematic sampling seems to be the most skewed out of the four examples while simple random sampling is only slightly less skewed than the original data.

# Further Areas of Analysis Using Data Wrangling

## Average Median Income By Major Category

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.6     v dplyr   1.1.0
## v tidyr   1.2.0     v stringr 1.5.1
## v readr   2.1.3     v forcats 1.0.0
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks plotly::filter(), stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
recent.grads %>%
  group_by(Major_category) %>%
  summarise(avgMedian <- mean(Median))
```

```
## # A tibble: 16 x 2
##    Major_category                     `avgMedian <- mean(Median)`
##    <chr>                                                   <dbl>
##  1 Agriculture & Natural Resources                         36900
##  2 Arts                                                    33062.
##  3 Biology & Life Science                                  36421.
##  4 Business                                                43538.
##  5 Communications & Journalism                             34500
##  6 Computers & Mathematics                                 42745.
##  7 Education                                               32350
##  8 Engineering                                             57383.
##  9 Health                                                  36825
## 10 Humanities & Liberal Arts                               31913.
## 11 Industrial Arts & Consumer Services                     36343.
## 12 Interdisciplinary                                       35000
```

```
## 13 Law & Public Policy                                  42200
## 14 Physical Sciences                                    41890
## 15 Psychology & Social Work                             30100
## 16 Social Science                                       37344.
```

Based on the data, Engineering majors earn the highest median income while Psychology & Social Work majors earn the least amount on average.

## Average Proportions of Fulltime/Parttime Status of Employees by Major Category

```
recent.grads %>%
  drop_na() %>%
  group_by(Major_category) %>%
  mutate(ftprop = Full_time/Total, ptprop = Part_time/Total, FtYrProp = Full_time_year_round/Total, na.
  summarise(avgFt = mean(ftprop), avgPt = mean(ptprop), avgFtYd = mean(FtYrProp))
```

```
## # A tibble: 16 x 4
##    Major_category                     avgFt avgPt avgFtYd
##    <chr>                              <dbl> <dbl>   <dbl>
##  1 Agriculture & Natural Resources    0.736 0.217   0.569
##  2 Arts                               0.577 0.338   0.409
##  3 Biology & Life Science             0.555 0.245   0.395
##  4 Business                           0.761 0.146   0.609
##  5 Communications & Journalism        0.701 0.226   0.548
##  6 Computers & Mathematics            0.722 0.169   0.546
##  7 Education                          0.726 0.196   0.530
##  8 Engineering                        0.715 0.166   0.530
##  9 Health                             0.574 0.281   0.430
## 10 Humanities & Liberal Arts          0.575 0.308   0.402
## 11 Industrial Arts & Consumer Services 0.763 0.166   0.626
## 12 Interdisciplinary                  0.653 0.258   0.507
## 13 Law & Public Policy                0.687 0.208   0.534
## 14 Physical Sciences                  0.634 0.252   0.497
## 15 Psychology & Social Work           0.617 0.272   0.469
## 16 Social Science                     0.643 0.248   0.479
```

Industrial Arts & Consumer Services have the highest proportion of full time employees out of all major categories while Biology and Life Sciences has the lowest proportion of full timers. When taking into consideration employees that are full time year round, these same findings apply as well.

Business has the lowest proportion of part time employees while the Arts major category has the highest.

## Average Proportions of Employees with a Job Requiring a College Degree by Major Category

```
recent.grads %>%
  drop_na() %>%
  group_by(Major_category) %>%
  mutate(degree = College_jobs/Total, nodegree = Non_college_jobs/Total, lowwage = Low_wage_jobs/Total,
  summarise(avgDegree = mean(degree), avgNonDegree = mean(nodegree), avgLowWage = mean(lowwage))
```

```
## # A tibble: 16 x 4
##    Major_category                    avgDegree avgNonDegree avgLowWage
##    <chr>                                 <dbl>        <dbl>      <dbl>
##  1 Agriculture & Natural Resources       0.287        0.411     0.0770
##  2 Arts                                  0.237        0.487     0.184
##  3 Biology & Life Science                0.372        0.261     0.0740
##  4 Business                              0.155        0.358     0.0916
##  5 Communications & Journalism           0.234        0.434     0.126
##  6 Computers & Mathematics               0.429        0.278     0.0594
##  7 Education                             0.582        0.224     0.0784
##  8 Engineering                           0.475        0.222     0.0510
##  9 Health                                0.369        0.335     0.0909
## 10 Humanities & Liberal Arts             0.255        0.408     0.134
## 11 Industrial Arts & Consumer Services   0.172        0.420     0.0906
## 12 Interdisciplinary                     0.421        0.317     0.0863
## 13 Law & Public Policy                   0.216        0.448     0.0979
## 14 Physical Sciences                     0.379        0.331     0.0917
## 15 Psychology & Social Work              0.370        0.351     0.0982
## 16 Social Science                        0.234        0.396     0.116
```

The major with the highest proportion of people in a job requiring a degree was Education majors. The lowest proportion studied Business.

The major with the highest proportion of people in a job not requiring a degree was Arts majors while the lowest was Engineering majors.

The major with the highest proportion of low wage earners in the data was Arts majors while the major with the lowest proportion was Engineering Majors.