PSTAT 105 HW 1 Question 4

Matthew Xu (5752811)

15 Janurary 2021

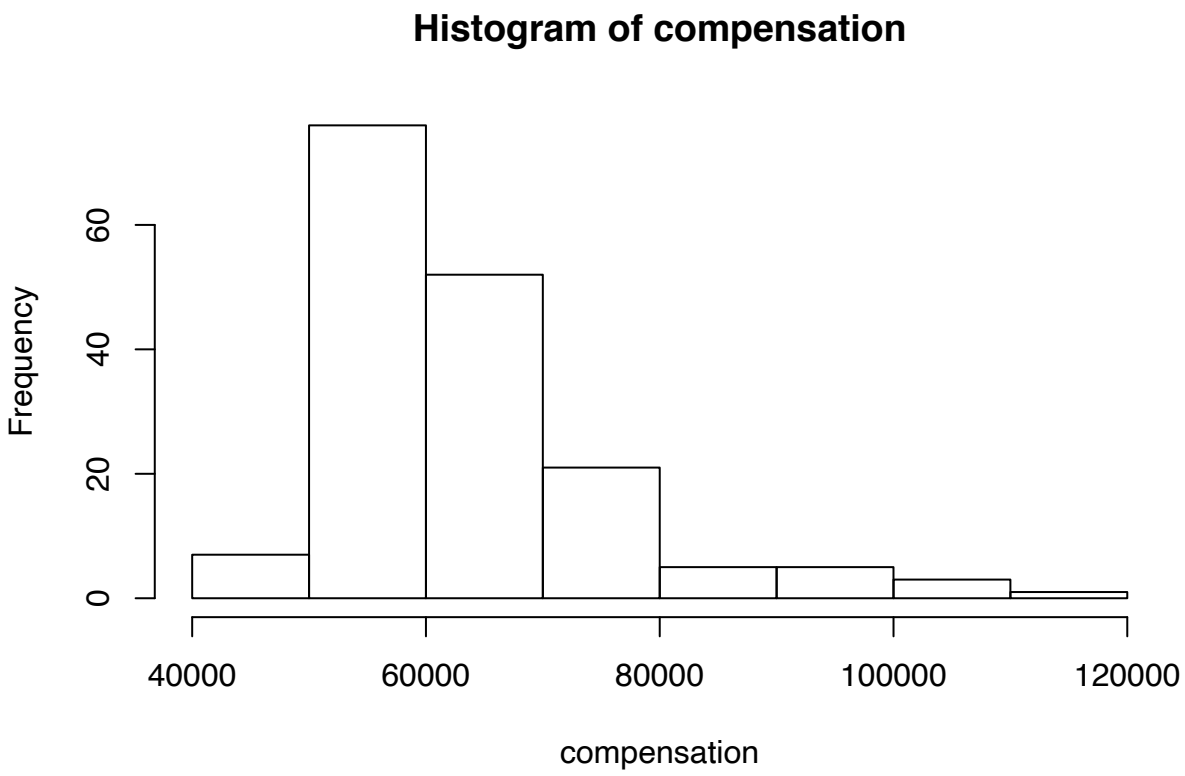
```
## set the working directory as the file location
setwd(getwd())

# read in data from file
compensation <- scan("compensation.txt", n = 170, skip = 3)
```

#4. For this question, please include typed answers to the questions along with your R input and output. The file compensation.txt contains data about average compensation in a sample of counties in the US.

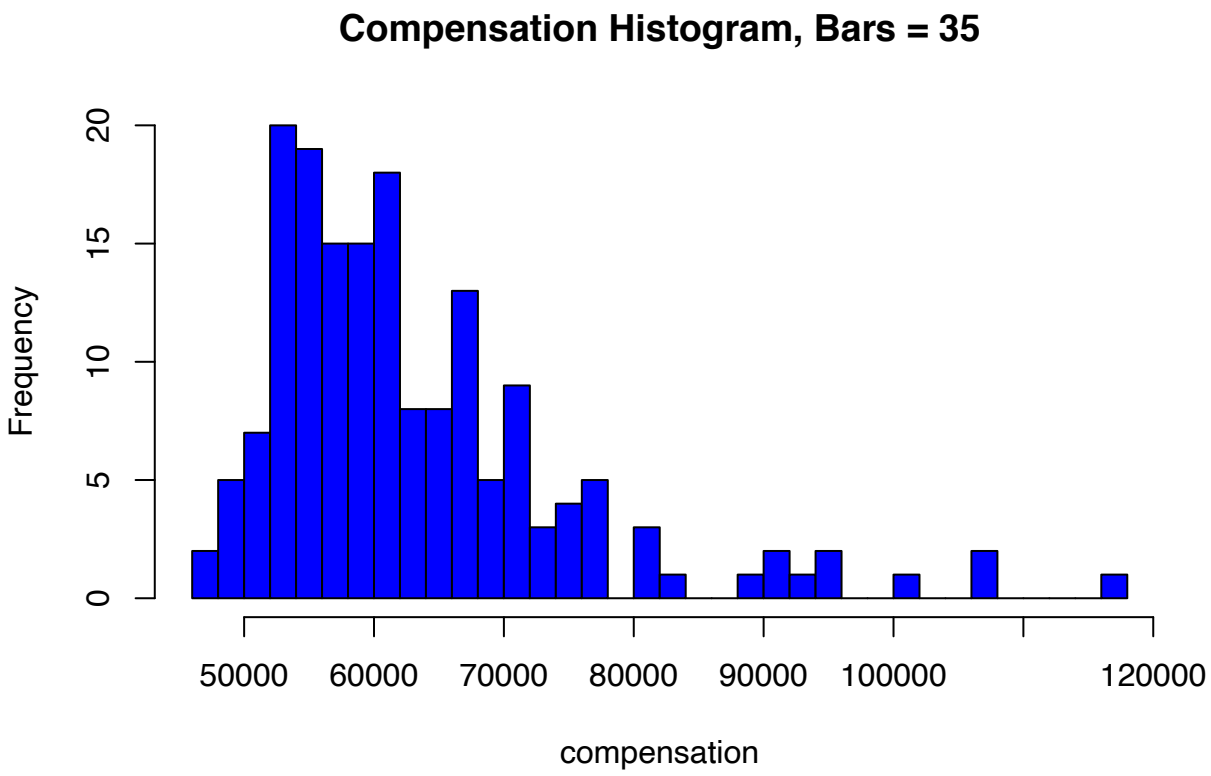
a. "PROBLEM NUMBER FOUR, PART A" Load the data into R and draw a histogram of the data. (You could use the hist function or geom_histogram.)

```
# Drawing Histrogram
hist(compensation)
```

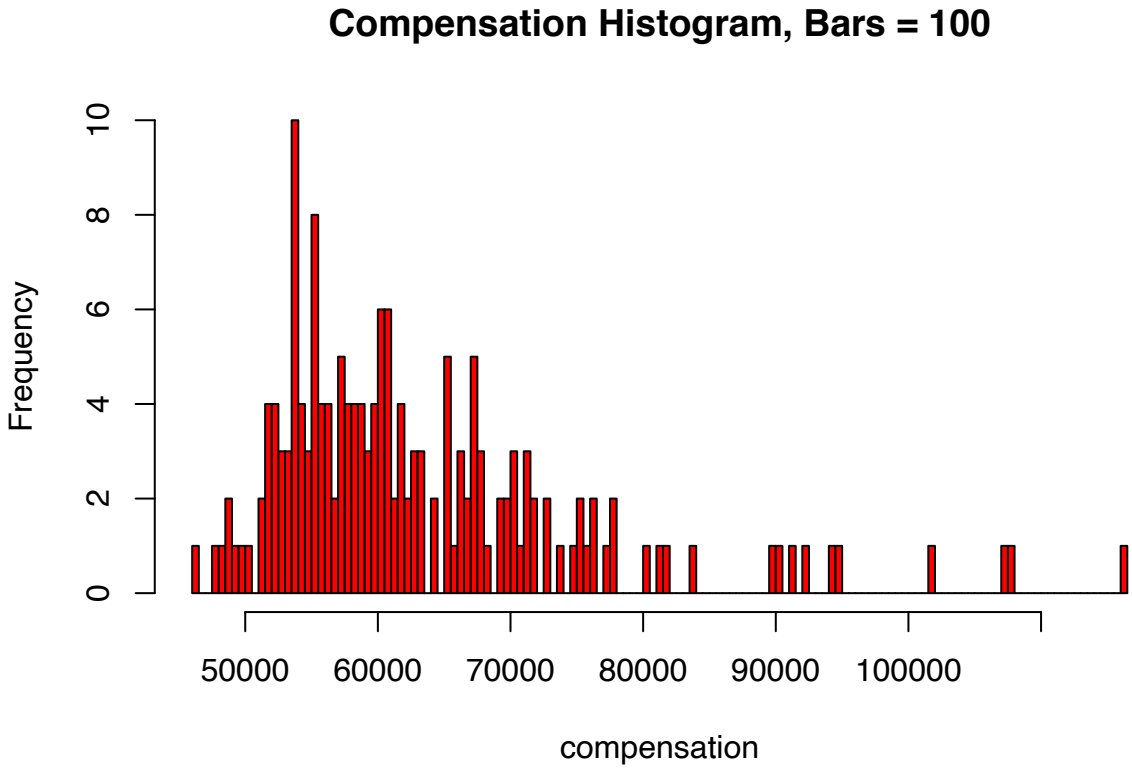


b. "PROBLEM NUMBER FOUR, PART B" The histogram produced by default settings in R always seems to me to have too few bars. Plot additional histograms where the number of bars is increased to 35 and 100. Which histogram do you think does the best job of illustrating the data set and why? (Hint: if you include a color in the call to hist, it can make it easier to see the bars.)

```
# Prof. Carter bins = bars, acceptable when bars = 0 Drawing Histrogram for bars
# = 35
hist(compensation, main = "Compensation Histogram, Bars = 35", breaks = 35, col = "blue")
```




```r
# Drawing Histrogram for bars = 100
hist(compensation, main = "Compensation Histogram, Bars = 100", breaks = 100, col = "red")
```



In both histograms, the shape decipicted by the coloring remains similar. However, the shape when bins

= 35 is clearer in shape and less noisy, while the shape in the histogram when bins = 100 is more noisy and has a slight loss in its overall shape, having many bars widely spread apart and with equivalent values within each bar.

c. "PROBLEM NUMBER FOUR, PART C" In order to find a confidence interval for the average over all the counties, I used the following R functions

t.stat <- t.test(compensation,conf.level=0.95) t.stat$conf.int [1] 61381.55 65001.70 attr(,"conf.level") [1] 0.95

Looking at the histogram, why would we be concerned about the validity of the confidence interval? Please be as specific as you can.

Although the 95% confidence interval true mean does appear to lie between the interval above, t test used is a t test which impleis that the distribution is approximetly normal. However, looking at the histrograms above, the ditribution does not appear to be normal and is skewed to the right even with a size of 170. Therefore a t test may not be sufficient nor valid.

d. "PROBLEM NUMBER FOUR, PART D" Run the command mean(compensation <= 50000). What is this doing and how could we use the result?

```
meanComp = mean(compensation <= 50000)
meanComp
```

```
## [1] 0.04117647
```

This command indicates the percentage of compensation that are less than or equal to 50000. This is a nonparametric estimation of the porportion of compensation less than or equal to 50000. Assuming validity of the t-test 95% confidence interval above, 50000 is a value outside the 95% confidence interval. Therefore, of the remaing 5% not explained by the confidence interval, 4.11% outisde of the interval is <= 50000, less than the bottom interval of the interval.

e. "PROBLEM NUMBER FOUR, PART E" Perform a test to determine whether the 75th percentile of the compensations is $80,000. Report a P value for the test, and state what you conclude specifically about the compensations.

```
percentile = quantile(compensation, 0.75)
variance = var(compensation)
sd = sqrt(variance)
zscore = (percentile - 80000)/sd/length(compensation)
cat("z-score: ", zscore, "\n")
```

```
## z-score:  -0.006117036
```

```r
pvalue2sided = 1 - 2 * pnorm(-abs(zscore))
cat("p-value (2-sided alternative): ", pvalue2sided)
```

```
## p-value (2-sided alternative):  0.004880659
```

Ho = 75th percentile of compensations is 80000. Ha = 75th percentile of compensations is not 80000.

Assume normality due to high sample size of 170. By Central Limit Theroem, we can approximate to normal distribution

The p- value is less than significance level of 0.05, therefore rejecting the null hypothesis. There is sufficient evidence to state that the 75th percentile of compensations is not 80000 and that it is statistially siginificant. This also appears to be consistent with the histograms above.