PSTAT 105 - Assignment 1

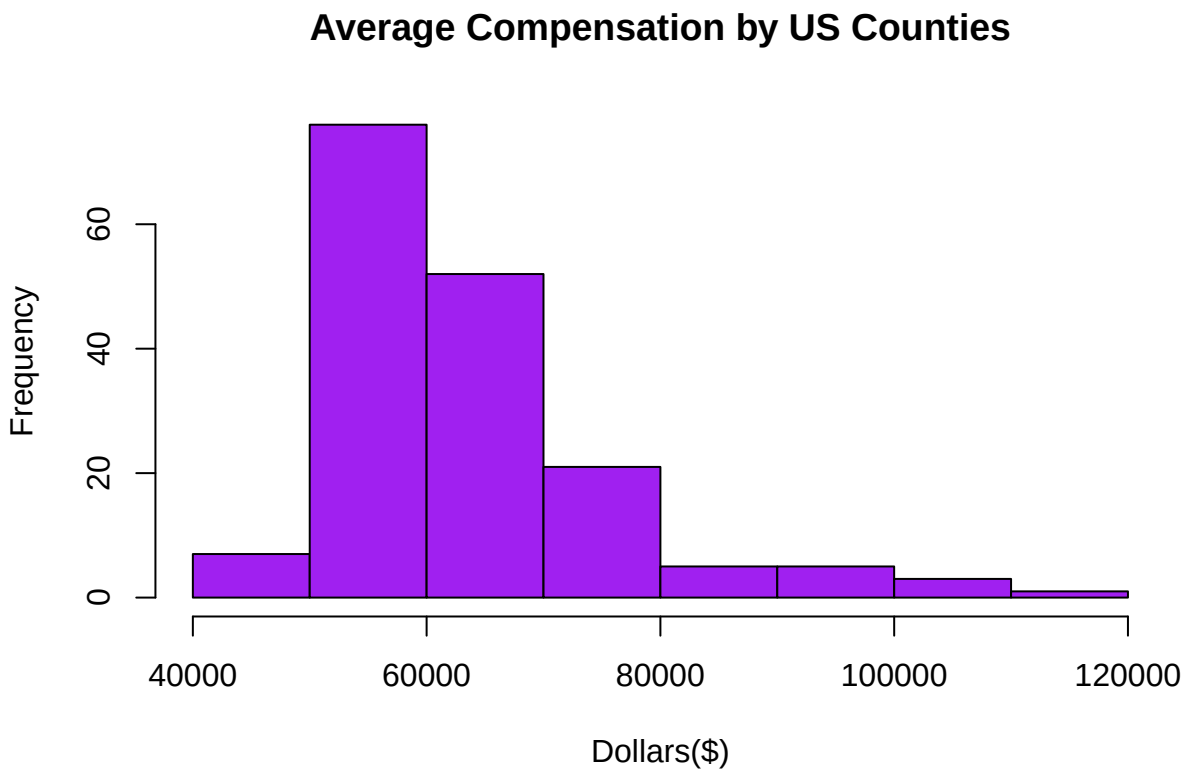
Roshan Mehta

January 31, 2022

The file `compensation.txt` contains data about average compensation in a sample of counties in the US.

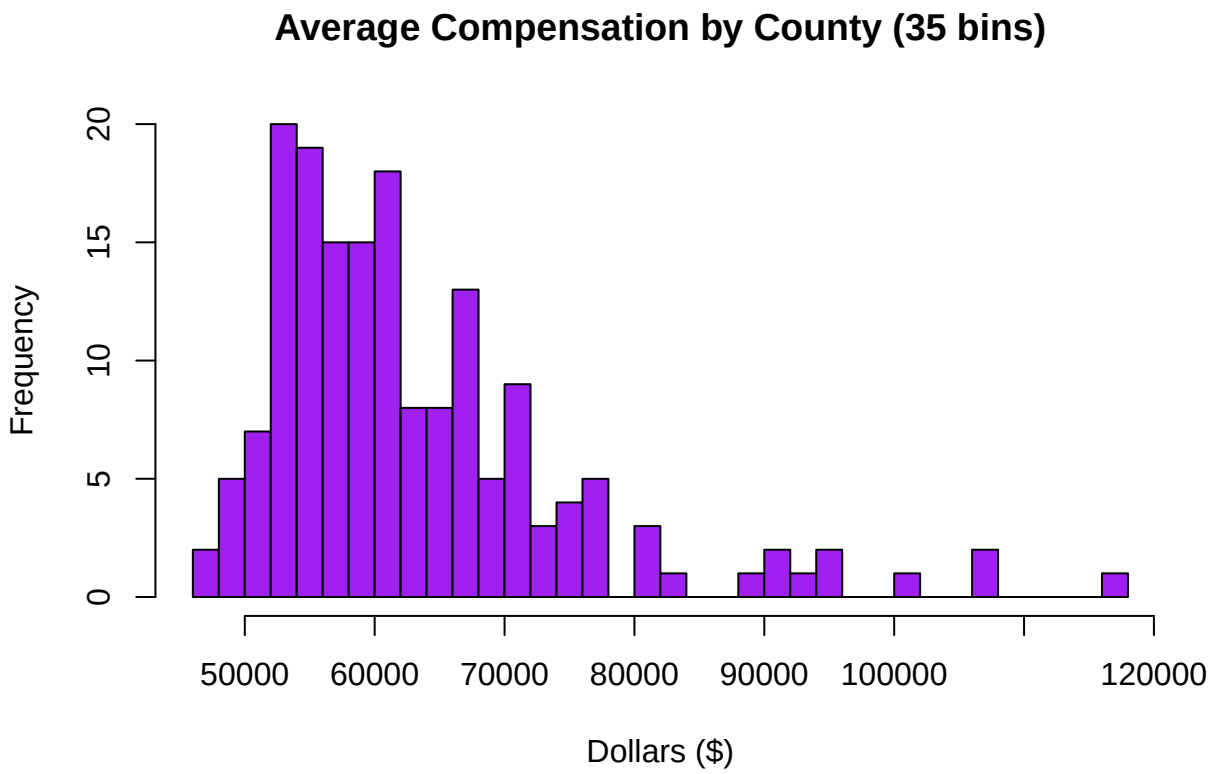
a) Load the data into R and draw a histogram of the data. The histogram given the default specifications is

```
setwd("/Users/roshanmehta/Downloads/PSTAT/PSTAT 105/pstat105-nonparametric-methods/data_files")
compensation <- scan("compensation.txt", n=170, skip=3)
hist(compensation, col="purple",main="Average Compensation by US Counties", xlab="Dollars($)")
```

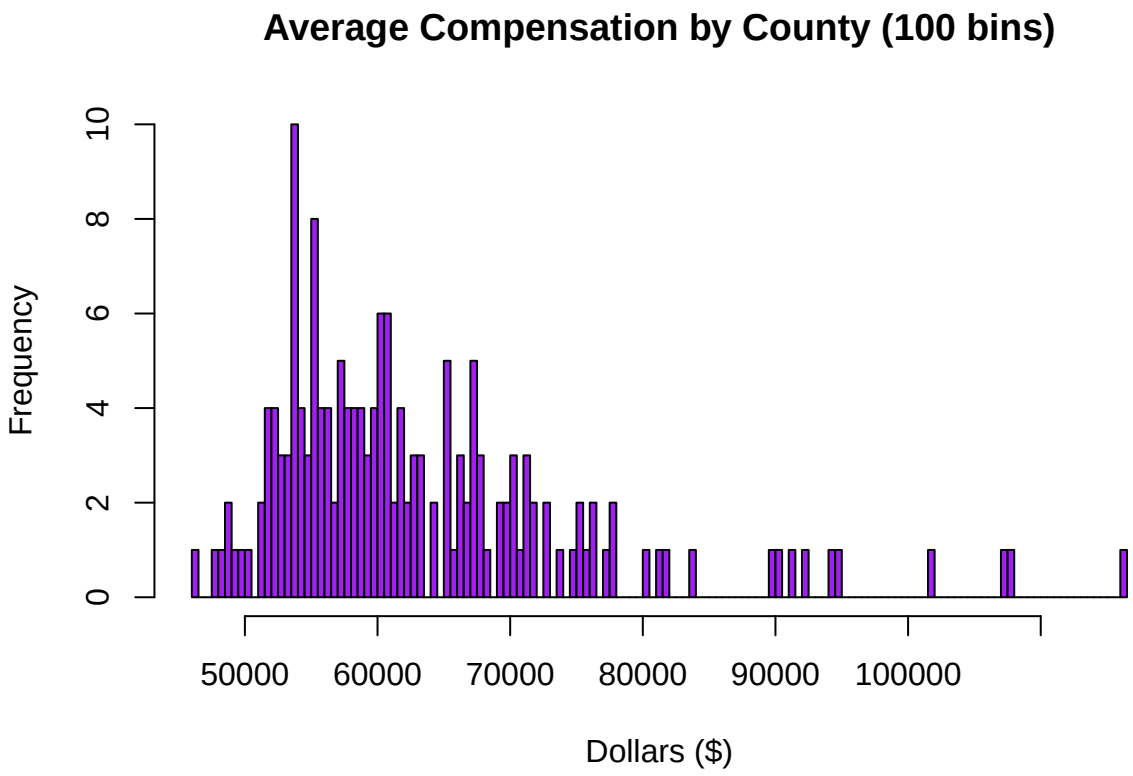


b) The histogram produced by default settings in R always seems to me to have too few bars. Plot additional histograms where the number of bars is increased to 35 and 100. Which histogram do you think does the best job of illustrating the data set and why?

```
hist(compensation, main="Average Compensation by County (35 bins)",xlab="Dollars ($)", breaks=35, col="
```




```
hist(compensation, main="Average Compensation by County (100 bins)",xlab="Dollars ($)",breaks=100, col=
```



I think that the histogram with the 35 bins finds the happy medium between a graph that washesout interesting aspects of the data by having too few breaks, and a histogram that is too noisybecause it has too many bars.

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)
ci <- t.stat$conf.int
ci
```

```
## [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.

```
hist(compensation, main="Average Compensation by County (35 bins)",xlab="Dollars ($)", breaks=35, col="
abline(v=ci, col='red', lwd=3)
```

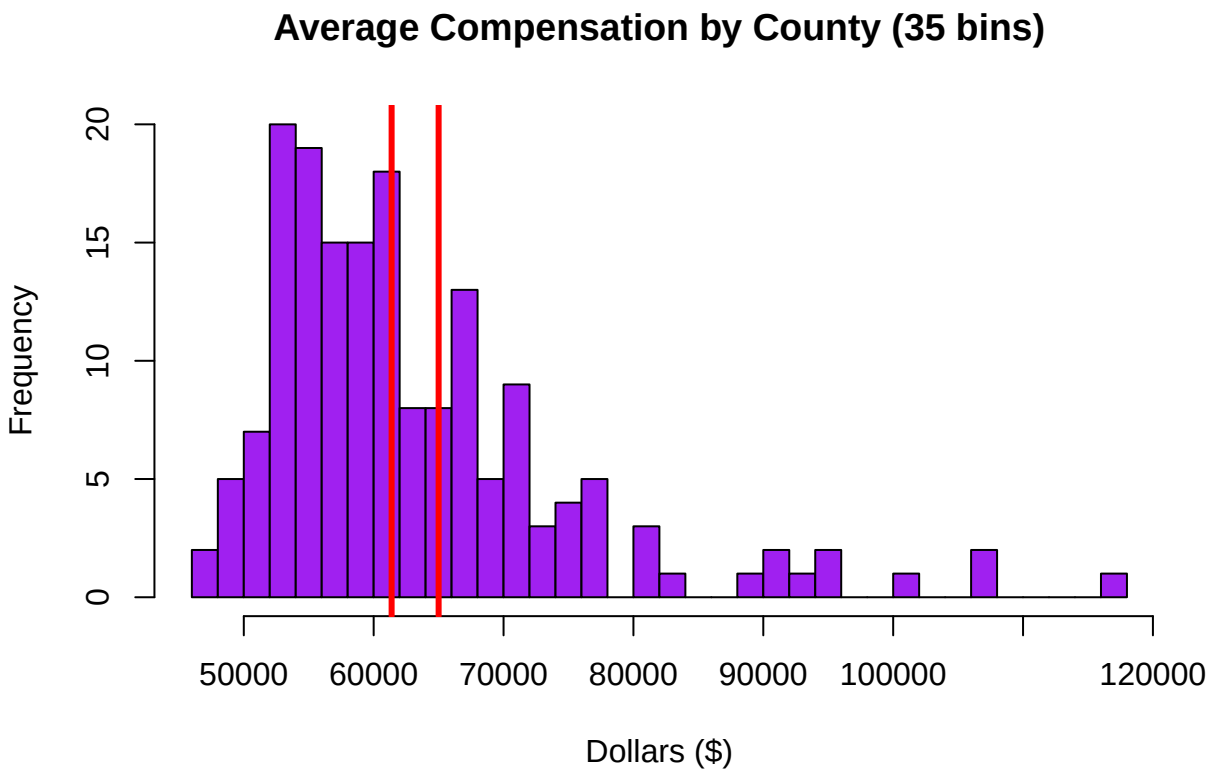


Figure 1: The histogram of the data showing the upper and lower bounds of the confidence interval in red.

Looking at the histogram, I am worried about the fact that the distribution of the data looks skewed. In may just be the existence of a couple of outliers that give the histogram this asymmetricshape. In either case, if one side of the histogram stretches out more than the other then this can pull the estimate in this direction, and also cause the estimate of the variance to be too large. If you look at figure 1, you will notice that the interval is further to the right than we might expect. It is being pulled that way because of a few large observations. It is not the case that we should expect this interval to cover 95% of the data. This interval is meant to indicate a range of values for the unknown mean parameter. We can compare this interval to what we might expect to see from normal data where the interval is more centrally located.

d)

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

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

The command above a trick to the proportion of a logical vector that is `TRUE`. There are 7 observations inthe data set that are no bigger than 50,000. The mean function tries to sum up a logical outcome from the inequality which it converts to 1's and 0's. Summing up a binary vector just tells us how many 1's are in it. Then, dividing by length of the vector is like dividing by n. We can use it to help us find estimates of the probability that the compensation is less than 50,000, and more generally, to find the empirical CDF by using it on other values.

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.

The hypothesis in this test are: $H_0 : \mathbb{P}\{X \leq 80,000\} = 0.75$ vs. $H_1 : \mathbb{P}\{X \leq 80,000\} \neq 0.75$

We can designate the probability of being less than 80,000 as θ. Our estimate of θ is the sample proportion:

```
(theta.hat <- mean(compensation <= 80000))
```

```
## [1] 0.9176471
```

For the binomial experiment, our approximately normal test statistic is

$$Z = \frac{\hat{\theta} - 0.75}{\sqrt{0.75(0.25)/n}}$$

```
(theta.hat-0.75)/sqrt(0.75*0.25/170)
```

```
## [1] 5.048005
```

This is certainly statistically significant. We would reject our null hypothesis that $\theta = 0.75$. The P-value is

```
2*pnorm((theta.hat-0.75)/sqrt(0.75*0.25/170),lower.tail = FALSE)
```

```
## [1] 4.464478e-07
```

Alternatvely, we could have used `R` to calculate a binomial probability

```
pbinom(170*theta.hat,size=170,p=0.75, lower.tail=FALSE) + pbinom((2*170*.75 - 170*theta.hat),size=170,p=
```

```
## [1] 1.27965e-06
```

Both of these are small.