PSTAT 105 HW1

Kendall Brown 8564403

Fall 2017 Section Tuesday 2pm

Q1a: Create a vector x with the numbers 8, 9, 16, 25, and 33. Use the function c.

```
x<-c(8,9,16,25,33)
x
```

```
## [1]  8  9 16 25 33
```

Q1b: Create a vector y with values 5, 10, 15, 20, and 25 using the seq function.

```
y<-seq(5,25,by=5)
y
```

```
## [1]  5 10 15 20 25
```

Q1c: Use the function rep to create a vector z with 7 2's.

```
z<-rep(2,7)
z
```

```
## [1] 2 2 2 2 2 2 2
```

Q1d: What happens if we enter x+y? What happens if we enter y*z?

```
s<-x+y
s
```

```
## [1] 13 19 31 45 58
```

Each element of x is added to the corresponding element of y.

```
p<-y*z
p
```

```
## [1] 10 20 30 40 50 10 20
```

A seven element vector is created where the first five elements are twice the values of the y vector, nad elements 6 and 7 are twice the value of the first two elements of y.

Q1e: Try an inequality like x $> 2y$. *Explain the difierence between the following functions* $x > y2$ and (x > y)*2

```
x > 2*y
```

```
## [1] FALSE FALSE FALSE FALSE FALSE
```

A five element vector is made where the elements of the vector are the results of testing wheter or not an element of vector x is greater than twice the value of the corresponding element in vector y.

```
(x>y)
```

```
## [1]  TRUE FALSE  TRUE  TRUE  TRUE
```

```
(x > y)*2
```

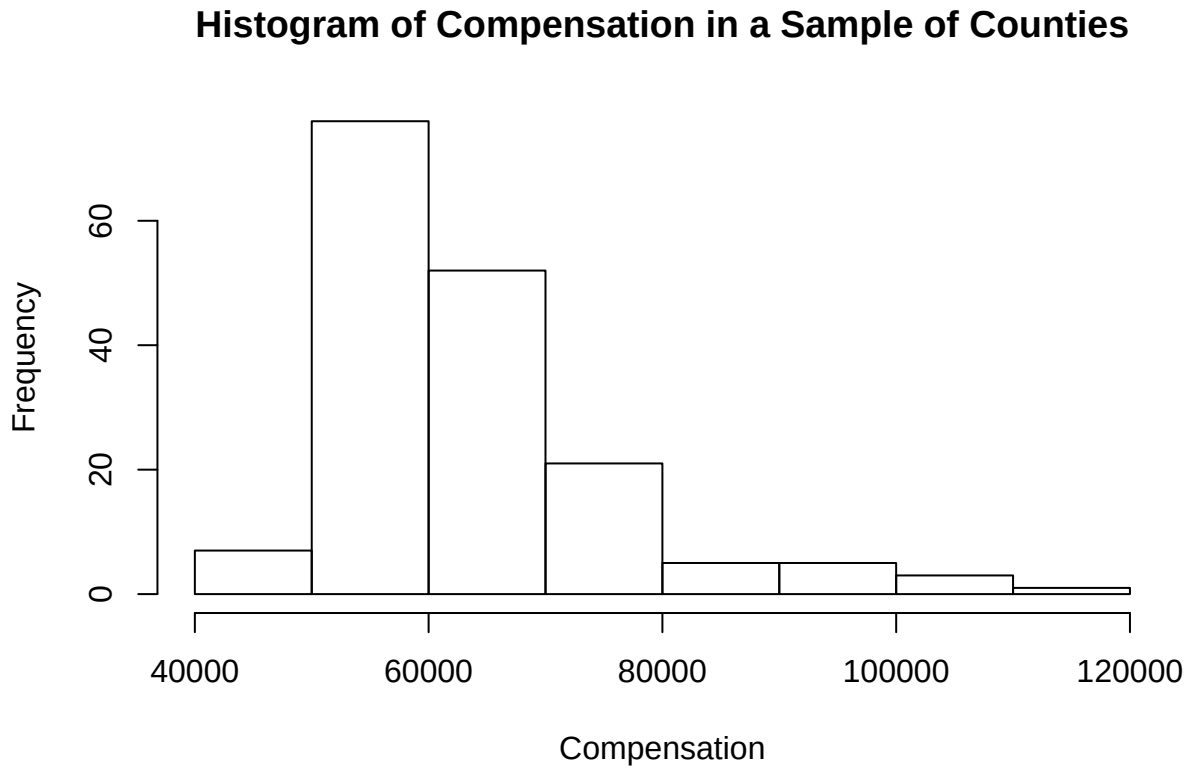
```
## [1] 2 0 2 2 2
```

A five element vector is created where the binary results of the (x>y) test (true=1, flase=0) are doubled.

The difference between x > y2 and (x > y)2 is that x>y2 tests the values of x against twice the values of y, while (x>y)*2 tests if an element of x is greater than the corresponding value of y then doubles the binary results of that test.

Q2a: Load the data into R and draw a histogram of the data using the hist function.

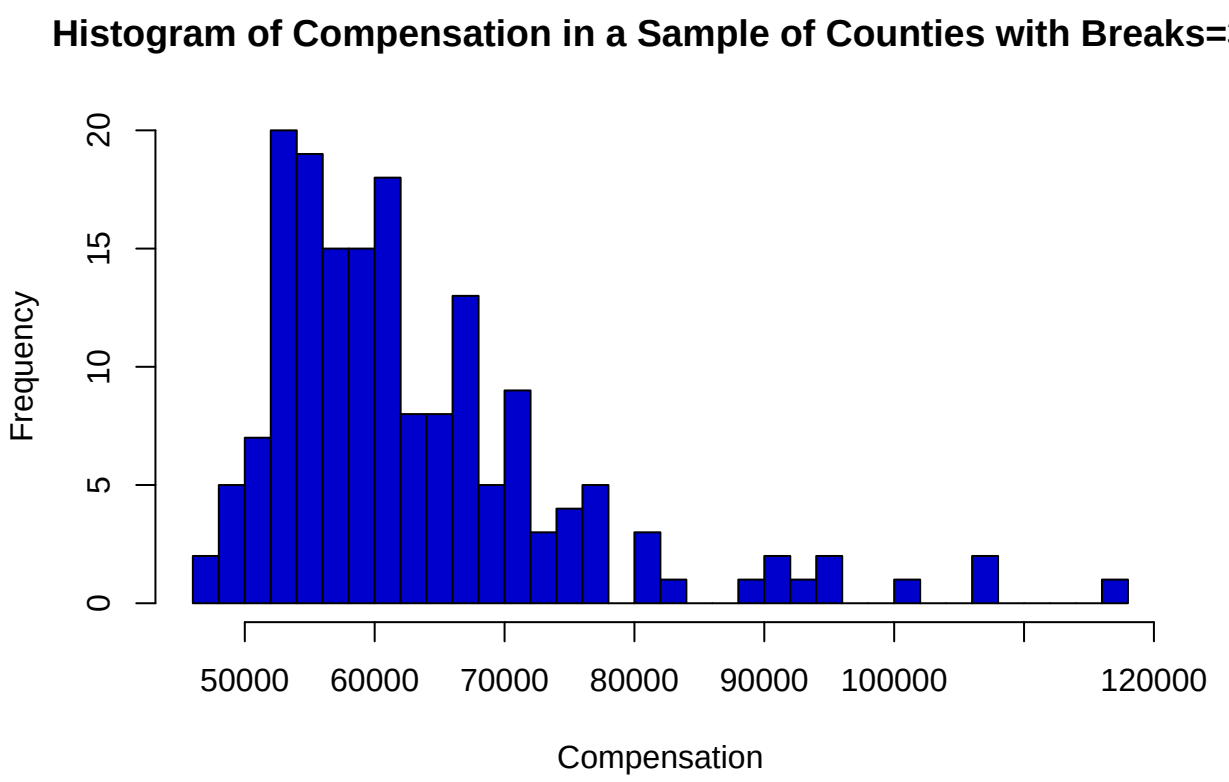
```
compensation <- scan("compensation.txt",n=170,skip=3)
hist(compensation, main="Histogram of Compensation in a Sample of Counties", xlab="Compensation")
```



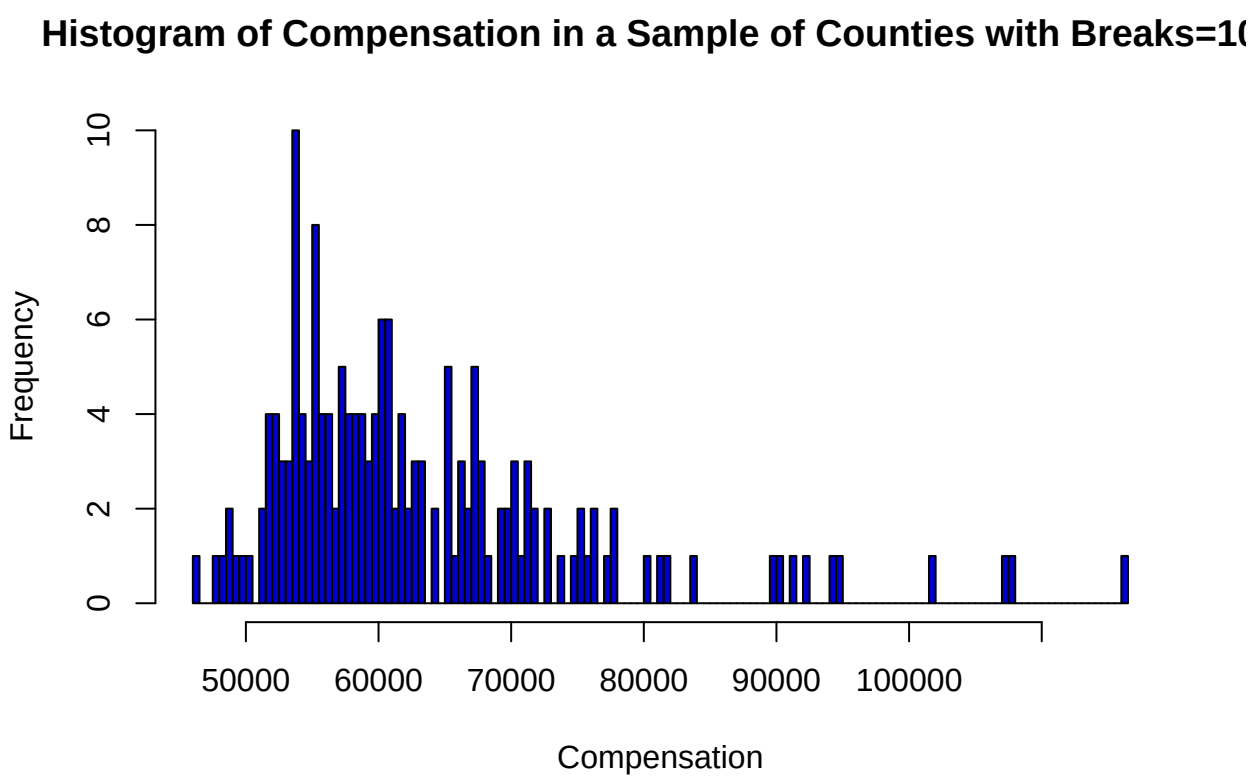
This is a histogram detailing the number of counties associated with a given range of compensation rates.

Q2b: 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.)

```
compensation <- scan("compensation.txt",n=170,skip=3)
hist(compensation, main="Histogram of Compensation in a Sample of Counties with Breaks=35", breaks=35, c
```




```
compensation <- scan("compensation.txt",n=170,skip=3)
hist(compensation, main="Histogram of Compensation in a Sample of Counties with Breaks=100", breaks=100
```



I feel that the 100 bar histogram does the best job at illustrating the dataset as the 8 and 35 bar histograms were too general and didn't accurately illustrate the rarity of a specific compensation. The 8 and 35 bar histograms grouped together multiple rates under a single bar and thus overly inflated the frequency values of rarer rates that closer to very common rates.

Q2c: 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.

There are a couple of outliers around 55000 and 57000 that appear to be skewing the confidence interval towards the lower end of the observed values by overinflating the number of observations below 60000.

Q2d: Run the command sum(compensation <= 50000). What is this doing and how could we use the result?

```
sum(compensation <= 50000)
```

```
## [1] 7
```

This is summing the number of observations bellow 50000. We can use this result to form an estimate for the probability a given county has a compensation rate below 50000 in addition to a workable chi-squared test statistic.