Stat 630 Lab5

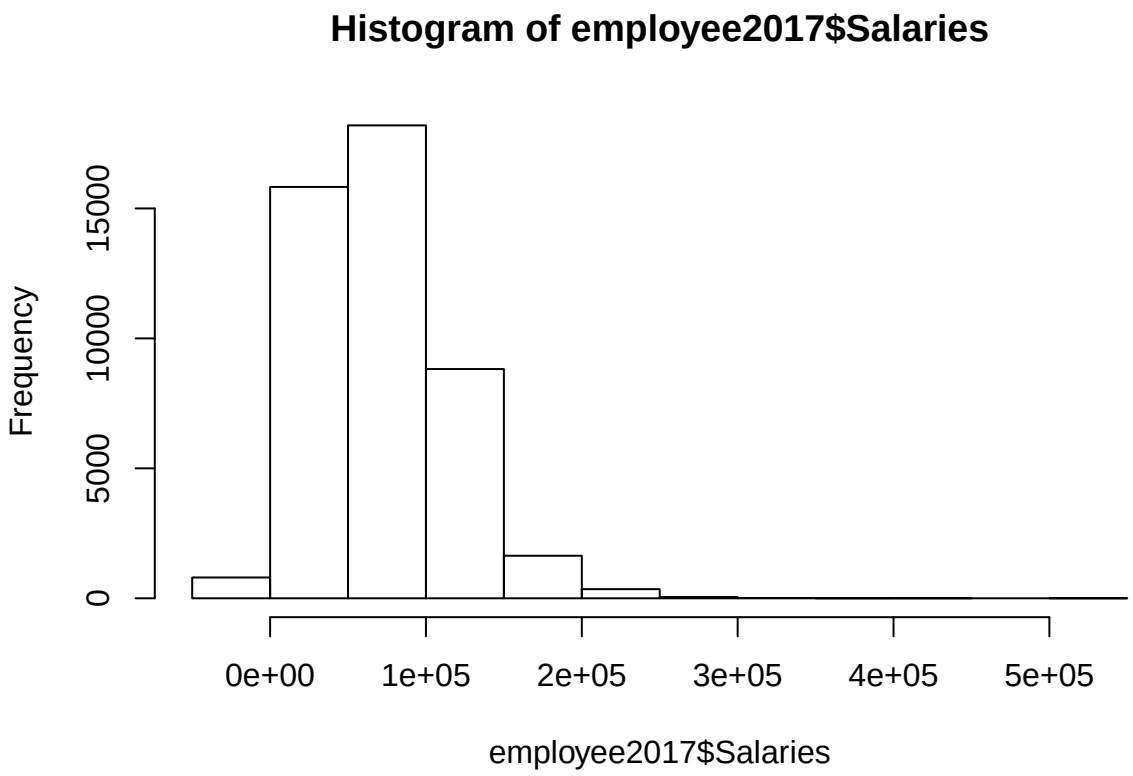
Chris Kalra aa6389

9/27/2018

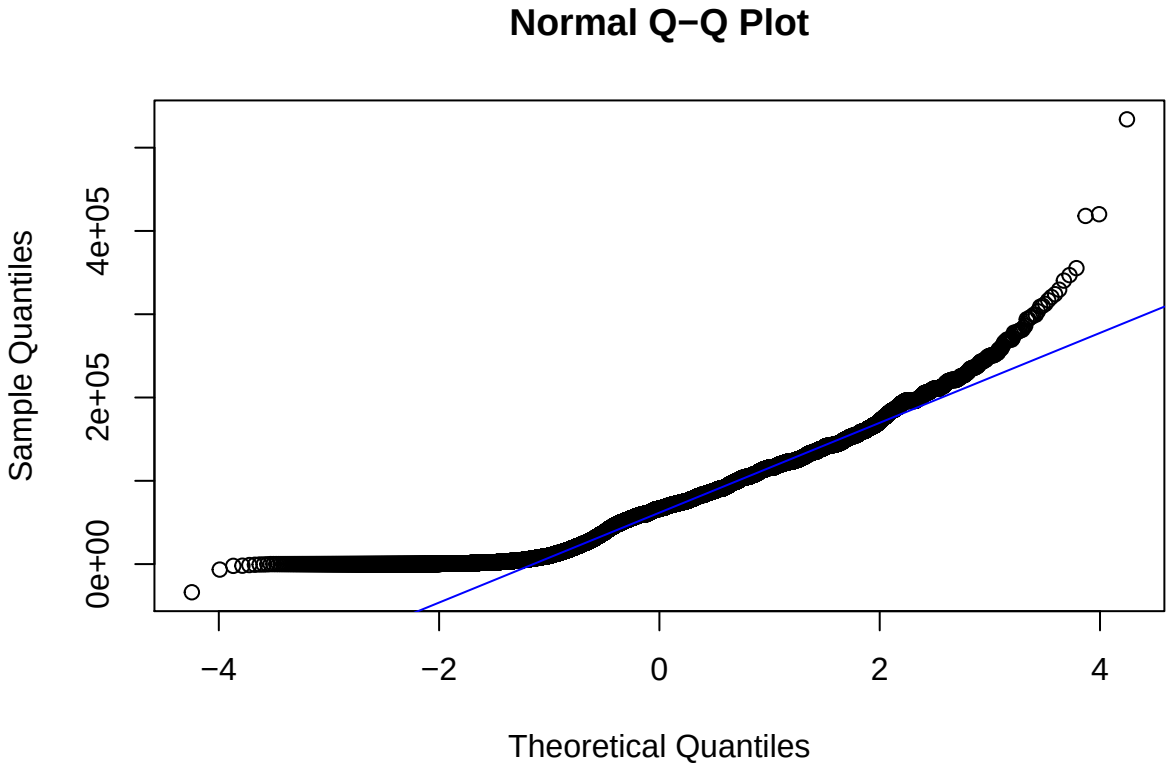
Exercise 1

```
data_url <- "https://github.com/ericwfox/stat630data/raw/master/Employee_Compensation.csv"
employee <- read.csv(data_url, header = TRUE)
employee2017 <- subset(employee, Year == 2017)

#a)
hist(employee2017$Salaries)
```




```
qqnorm(employee2017$Salaries)
qqline(employee2017$Salaries, col='blue')
```




```
mean(employee2017$Salaries)
```

```
## [1] 67062.38
```

```
sd(employee2017$Salaries)
```

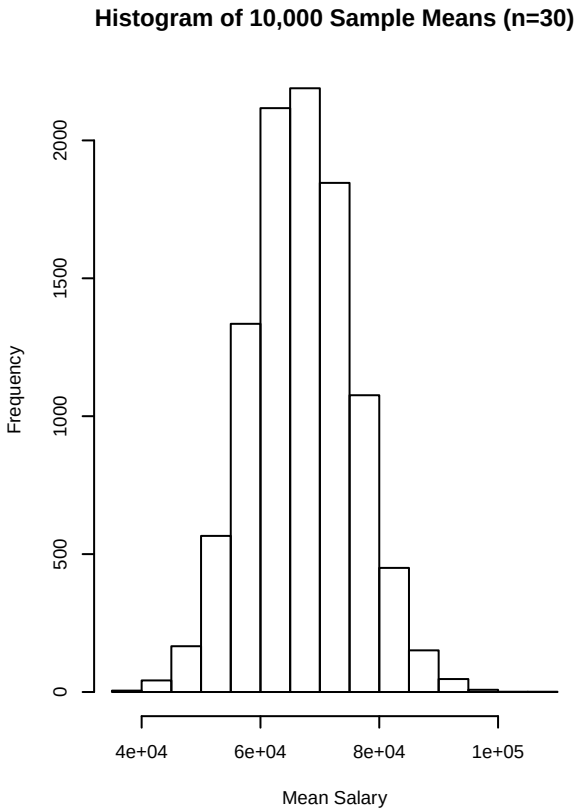
```
## [1] 47586.65
```

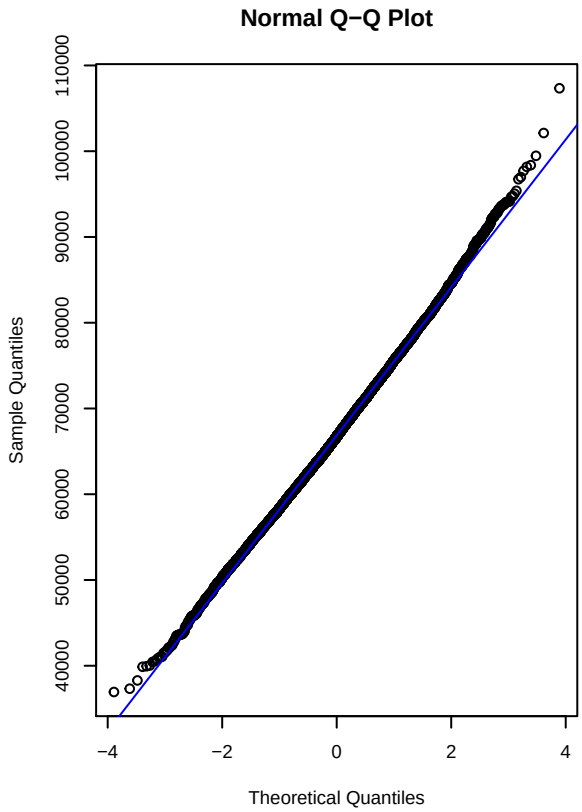
```
#b)
set.seed(630)
x=sample(employee2017$Salaries, 30)
mean(x)
```

```
## [1] 69817.59
```

It is fairly close to our true mean (as it is only about 4% larger), so while it is not perfectly accurate, it is by no means a horrible estimate either

```
#c)
set.seed(999)
xbars = rep(NA, 10000)
for(i in 1:10000) {
  samp_i = sample(employee2017$Salaries, 30)
  xbars[i] = mean(samp_i)
}
par(mfrow=c(1,2), cex=0.6)
hist(xbars, xlab='Mean Salary', main='Histogram of 10,000 Sample Means (n=30)')
qqnorm(xbars)
qqline(xbars, col='blue')
```





Despite the fact that our initial distribution was fairly right skewed, it is clear that our sampling distribution of Xbar is normal. This is due to the CLT taking effect, as we have a sufficiently large sample (each sample is of size 30)

```
#d)
set.seed(999)
mean(xbars)
```

```
## [1] 66922.49
```

```
mean(employee2017$Salaries)
```

```
## [1] 67062.38
```

```
sd(employee2017$Salaries)/sqrt(30)
```

```
## [1] 8688.093
```

```
sd(xbars)
```

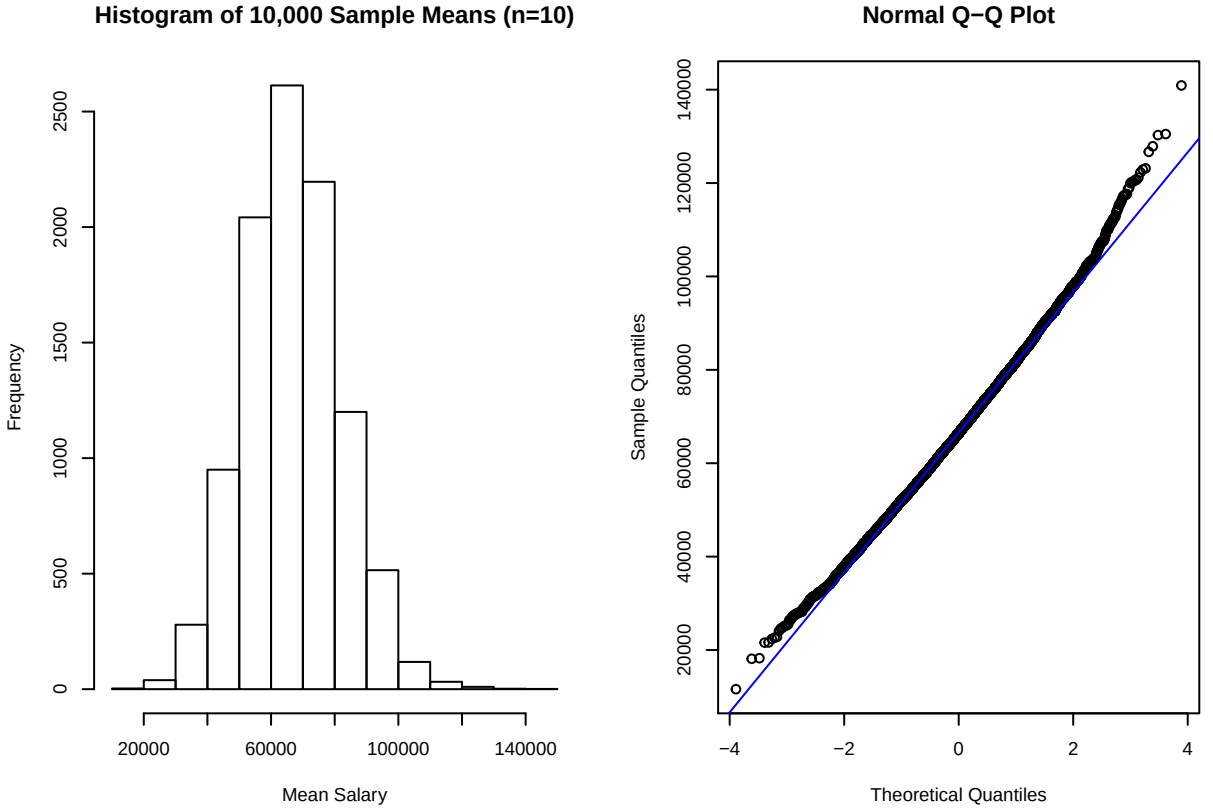
```
## [1] 8620.206
```

The means are fairly close; the sample mean is only about .26% smaller than our population mean, and the standard error of Xbar is less than 1% larger than the true standard deviation of (our sample / sqrt(30)), so the estimates are extremely close

```
#e
```

According to the CLT, for a sufficiently large sample size n, the distribution of the sample mean will be normal, regardless of the population distribution. In our case, 30 seems sufficient, as our distribution of Xbar appears to be normal

```r
#f-c)
set.seed(661)
xbars10 = rep(NA, 10000)
for(i in 1:10000) {
  samp_i = sample(employee2017$Salaries, 10)
  xbars10[i] = mean(samp_i)
}
par(mfrow=c(1,2), cex=0.6)
hist(xbars10, xlab='Mean Salary', main='Histogram of 10,000 Sample Means (n=10)')
qqnorm(xbars10)
qqline(xbars10, col='blue')
```



In this case, we have a noticeable right skew to our data, meaning that a sample size of 10 is not sufficient for the Central Limit Theorem to take effect. It is important to note that our population itself also has a noticeable right skew

```r
#f-d)
set.seed(661)
mean(xbars10)
```

```
## [1] 66887.14
```

```r
mean(employee2017$Salaries)
```

```
## [1] 67062.38
```

```r
sd(employee2017$Salaries)/sqrt(10)
```

```
## [1] 15048.22
```

```r
sd(xbars10)
```

```
## [1] 15073.41
```

While the sample mean and standard error are somewhat close to the true population mean and (standard deviation / sqrt(n)) values, the estimates are not as close as the estimates from our samples of size 30

```r
#f-e)
```

The Central Limit theorem does not appear to hold true in our case for a sample of size 10, despite the fact that we have 10,000 such samples. This is because, according to the CLT, the number of samples of we have is nearly irrelevant; what matters is the size of each sample. Therefore, for this data set, a sample size of 10 is not sufficient for the CLT to take effect

Exercise 2

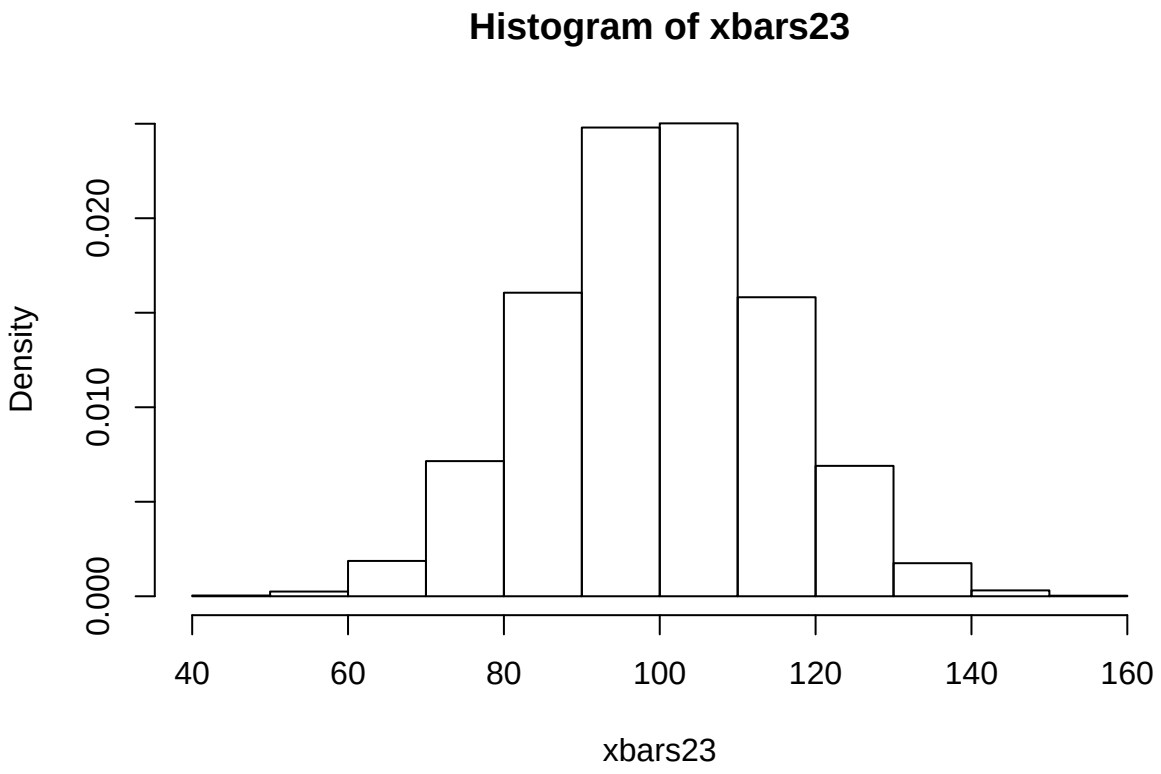
```r
set.seed(23)
xbars23=rep(NA, 10000)
for (i in 1:10000) {
  samp_i = sample(rnorm(4, 100, sd=30))
  xbars23[i] = mean(samp_i)
}
mean(xbars23)
```

```
## [1] 99.8887
```

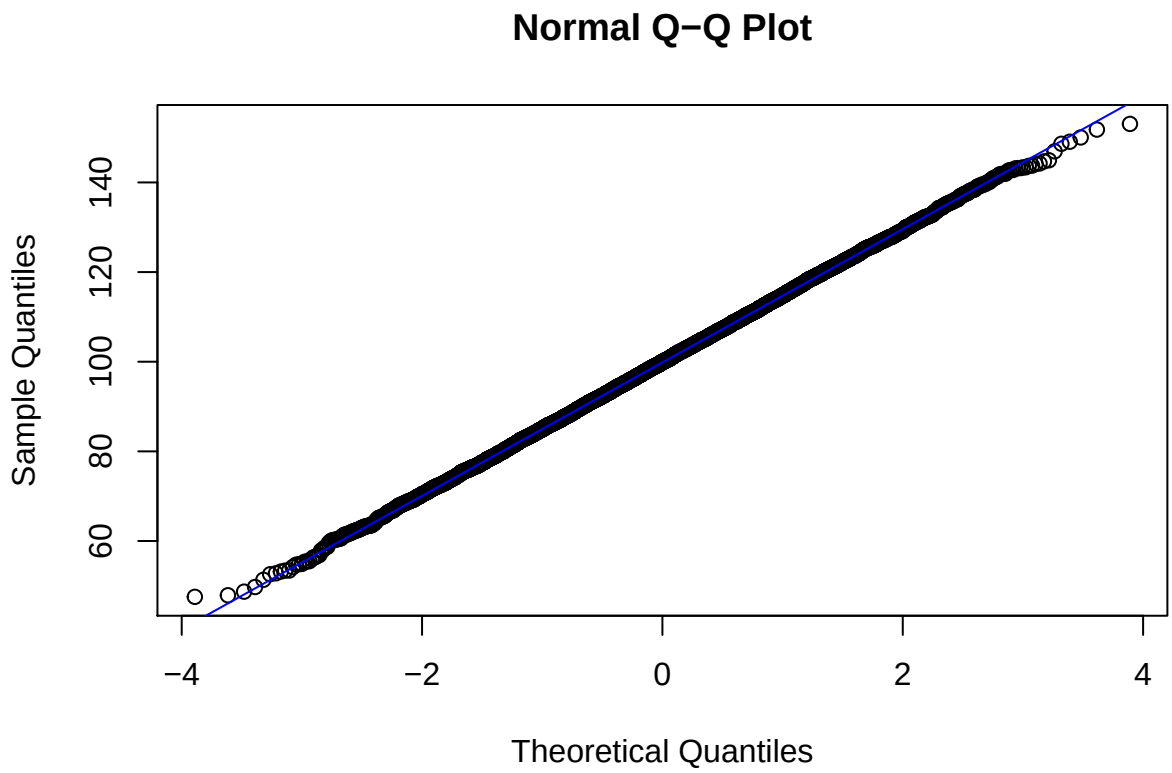
```r
sd(xbars23)
```

```
## [1] 14.85549
```

```r
hist(xbars23, freq=FALSE)
```




```
qqnorm(xbars23)
qqline(xbars23, col='blue')
```



The sampling distribution of Xbar appears normal. This is because Xbar follows a normal distribution, regardless of sample size, if our random variable X is normally distributed (which it is here). Beyond this,

the QQ plot shows us almost perfect normality

However, by superimposing the original N(100, 30) over our histogram of Xbar (as seen below), we can see that our Xbar has a standard error much less than the standard deviation of our sampling dstribution, as the tails of our distribution of Xbars are much thinner than those of a N(100, 30) population

```
hist(xbars23, freq=FALSE)
x<-seq(0,200,0.01)
curve(dnorm(x, mean=100, sd=30), add=TRUE, lwd=2, col='red')
```