Coursera Statistics Inference Assignment

Keith Bailey

February 4, 2017

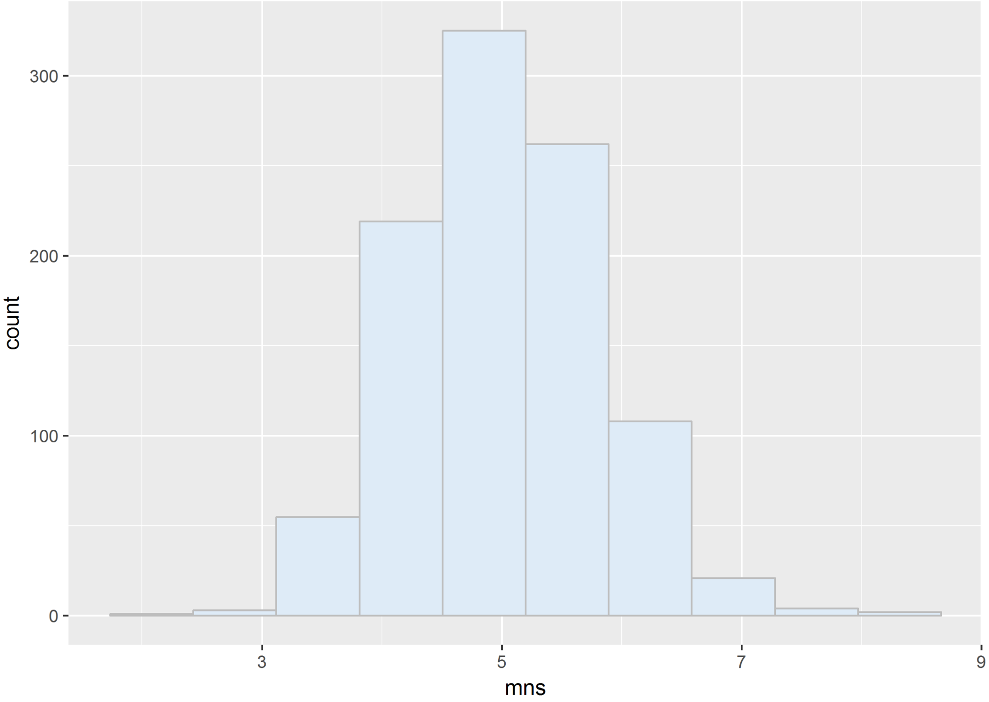
Part 1

For part 1 of this assignment we will be demonstrating, based on the exponential distribution, that the sample mean is a rerpesentative of the population from which we are sampling by using the Central Limit Theorm. We will achieve this by looking at samples and building to establish, for sample sizes of 40, using 1000 simulations, how close we are to the known popualation mean of 1/lambda and standard deviation of 1/lambda.

For the exercise, we will be using lambda = 0.2, so for the populationm we know that the mean and standard deviationi are both 5.

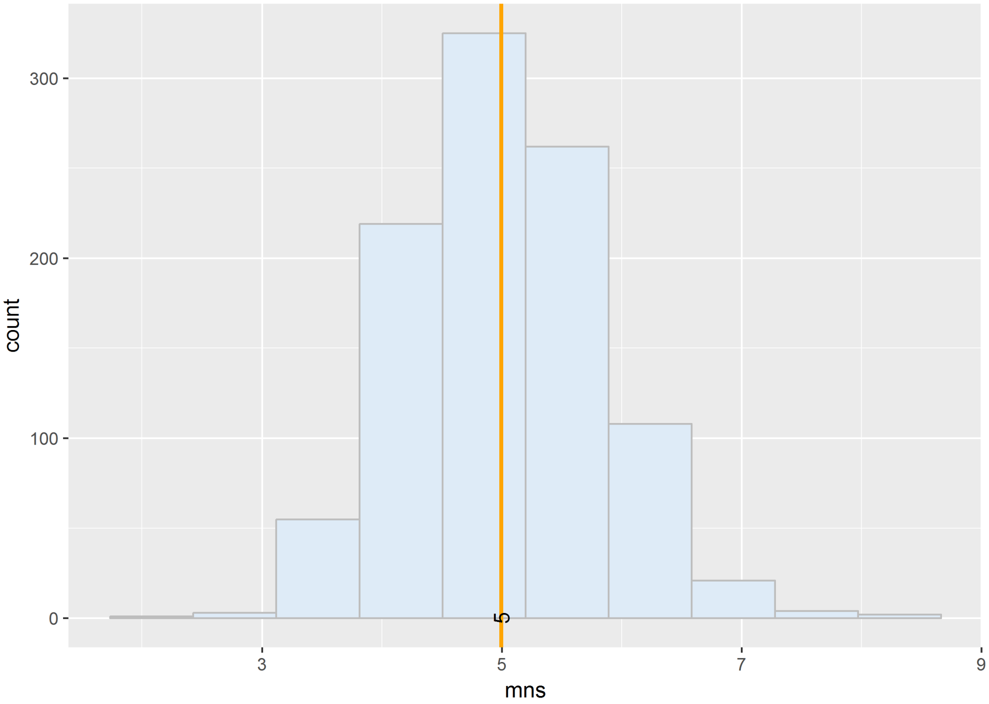
Show the sample mean and compare it to the theoretical mean (5) of the distribution.

```
lambda <- 0.2
simulations <- 1000

#Actual mean & sd of exp distribution
exp_mu<-1/lambda
exp_sd<-exp_mu

mns = NULL
vrs = NULL
for (i in 1 : simulations) {
  temp <- rexp(40, lambda)
  mns <- c(mns, mean(temp))
  vrs <- c(vrs, var(temp))
}

ggplot() + aes(mns)+ geom_histogram(bins=10,colour="#bdbdbd", fill="#deebf7")
```



From this we can see that the mean of our 1000 samples appears to be between 4 and 6. Lets find out what it is and plot it on the chart.

```
mean_mns = mean(mns)

ggplot() + aes(mns)+ geom_histogram(bins=10,colour="#bdbdbd", fill="#deebf7") +
  geom_vline(aes(xintercept=mean_mns),
             linetype="solid", size=1, colour="orange") +
  annotate("text", x = mean_mns, y = 0,  angle = 90, label = round(mean_mns,2), parse = TRUE)
```



This looks very close to our population mean of 5, but we don’t know the variability in our data, We can assess this by considering the variance of our means of our 1000 samples of 40 observations and establish our standard error of our sample mean.

Lets do that and then plot it so we can see the interval where we would have 95% confidence, based on the means of each of our samples (i.e. 1000 samples of 40 observations), that the population mean would fall within.

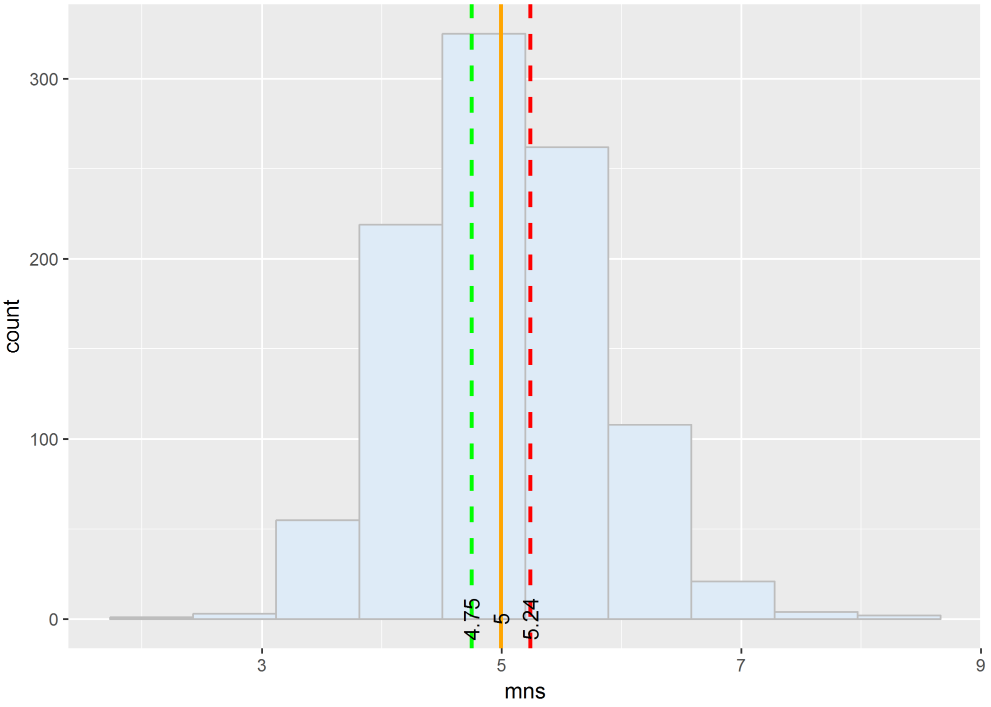
```
se=sqrt(var(mns)/40)

#standard error
se
```

```
## [1] 0.12704
```

```
#95% confidence interval +/- the following
se.95 = se*1.96

#confidence interval therefore
mean_mns +c(-1,1)*se.95
```

```
## [1] 4.746466 5.244462
```

```
ggplot() + aes(mns)+ geom_histogram(bins=10,colour="#bdbdbd", fill="#deebf7") +
  geom_vline(aes(xintercept=mean_mns),
             linetype="solid", size=1, colour="orange") +
  annotate("text", x = mean_mns, y = 0,  angle = 90, label = round(mean_mns,2), parse = TRUE)+
  geom_vline(data=cToothGrowth, aes(xintercept=round(mean_mns+se.95,2)),
             linetype="dashed", size=1, colour="red") +
  annotate("text", x = round(mean_mns+se.95,2), y = 0,  angle = 90, label = round(mean_mns+se.95,2), parse
 = TRUE)+
  geom_vline(data=cToothGrowth, aes(xintercept=round(mean_mns-se.95,2)),
             linetype="dashed", size=1, colour="green") +
  annotate("text", x = round(mean_mns-se.95,2), y = 0,  angle = 90, label = round(mean_mns-se.95,2), parse
 = TRUE)
```



So our 95% confidence interval, and we can say that there is only a 5% chance that the range 4.75 to 5.24 excludes the mean of the population.

How variable is our sample data?

Lets check the variance and compare it to the theoretical population variance of (1/lamda)^2 = 25. As part of our simulations we also noted the variance of each sample. The average of these is

```
mean_vrs<-mean(vrs)

mean_vrs
```

```
## [1] 24.9366
```

This is very close to the actual variance of the population. Just as with the mean, and infact any statistic we can use the standard error to establish a confidence interval for our statistic of interest.

```
se=sqrt(var(vrs)/40)

#standard error
se
```

```
## [1] 1.699209
```

```
#95% confidence interval +/- the following
se.95 = se*1.96

#confidence interval therefore
mean_vrs +c(-1,1)*se.95
```

```
## [1] 21.60615 28.26705
```

We could have decided not to simulate the variances, but instead calculated the variance of the simulated mean vs the theortical variance. The theoretical variance = (1/lambda)^2/n = (1/0.2)^2/40 = 25/40 = 0.625.

```
var(mns)
```

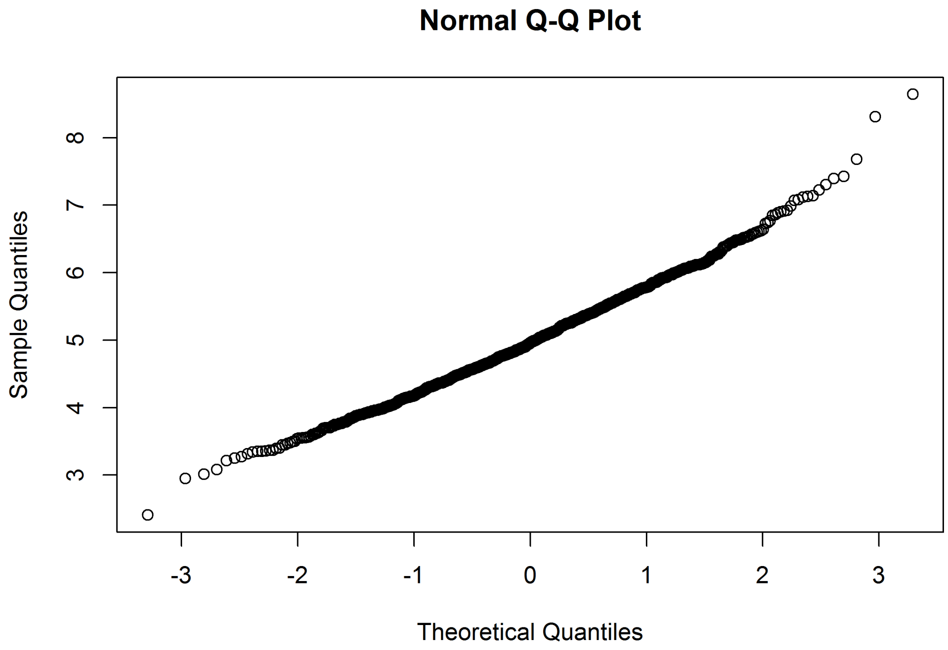
```
## [1] 0.6455665
```

Which is remarkably close.

Is the distribution approximately normal?

We establish this by looking at our data by using a quantile quantile plot. If our data plots in a straight line, we can say that it is indeed normally distributed.

```
qqnorm(mns)
```



From this, we can see that it is normally distributed.