Statistic Inference Project

Wednesday, December 17, 2014

This document will describe the steps done to proper process the data (in this case, we're asked to generate a exponential distribution of about 40 observations). In order to proper analyze the information, 1000 simulations will be run as requested.

After we're done with the above, we have to proper answer the folloing points:

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.
2. Show how variable it is and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

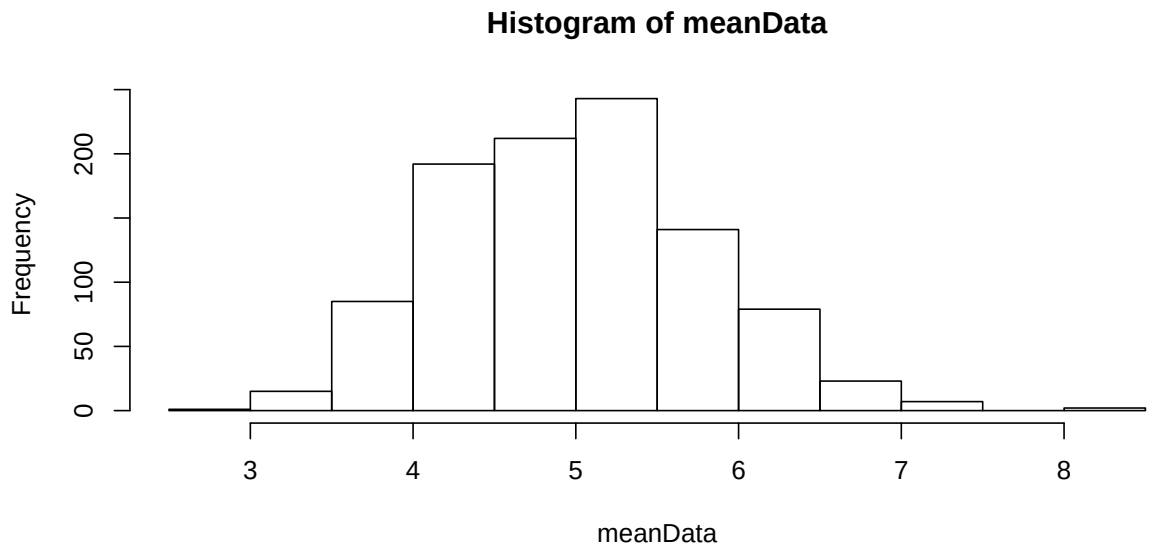
```
set.seed(31)
lambda <- 0.2
nsim <- 1000
data <- array(nsim)
for (i in 1 : nsim) { ## 1000 simulations
  data[i] <- data.frame(rexp(40, lambda)) ## 40 observations
}

meanData <- sapply(data, mean) ## take mean of each of the simulations
```

Now, we're going to focus on responding the item 1.

Let's take a look at the distribution of the means (histogram):

```
hist(meanData)
```



According to the exponential distribution, its theoretical mean is $\frac{1}{lambda}$, which is:

```
1/lambda
```

```
## [1] 5
```

The distribution of sample means is centered at:

```
mean(meanData)
```

```
## [1] 4.993867
```

The central limit theorem states that given a distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean μ and $\frac{\sigma^2}{N}$ as variance, when N - the sample size, increases. Both should converge and they did, as expected.

Now, we're going to focus on responding the item 2.

According to the exponential distribution, its theoretical standard deviation is $\frac{1}{lambda}$, which is:

```
1/lambda
```

```
## [1] 5
```

So, we can conclude that $(\frac{1}{lambda})^2$ is its theoretical variance:

```
(1/lambda)^2
```

```
## [1] 25
```

The variance of the sampling distribution of the mean is computed as $\frac{\sigma^2}{N}$, where sigma is the standard deviation of the distribution and N is the sample size.

So, the variance of the sample means should be:

```
(1/lambda)^2/40
```

```
## [1] 0.625
```

If the calculate the actual variance, we can see:

```
var(meanData)
```

```
## [1] 0.6291041
```

The central limit theorem states that given a distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean μ and $\frac{\sigma^2}{N}$ as variance, wheh N - the sample size, increases.

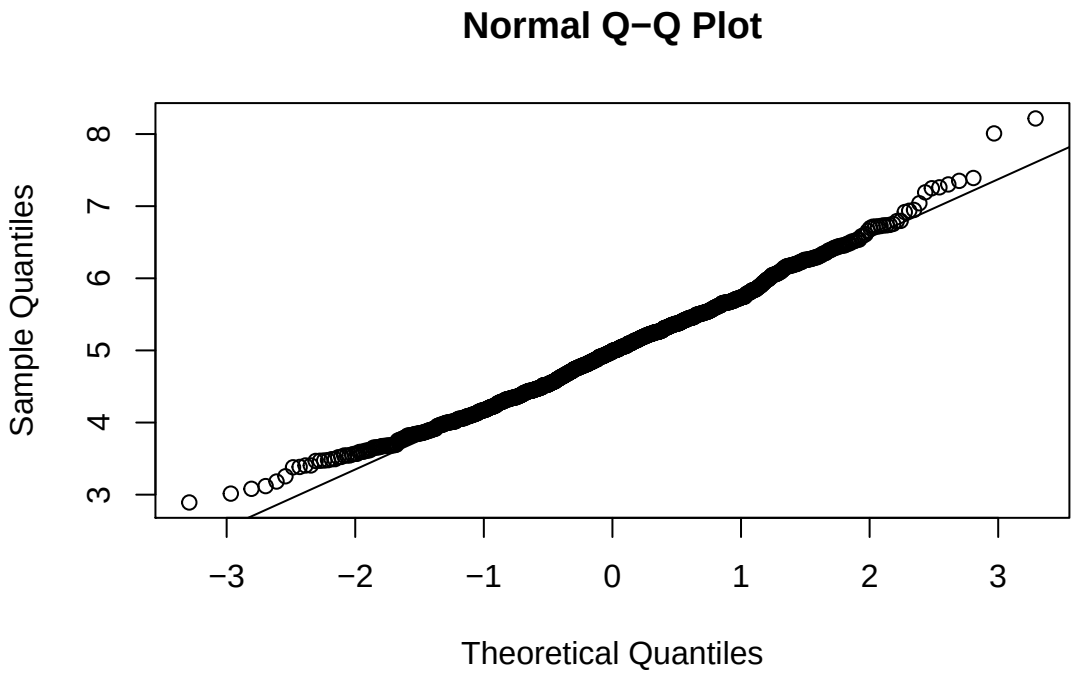
Both should converge and they did, as expected.

Now, we're going to focus on responding the item 3.

Again, the sampling distribution of the mean should approach a normal distribution, according to CLT. Let's check this.

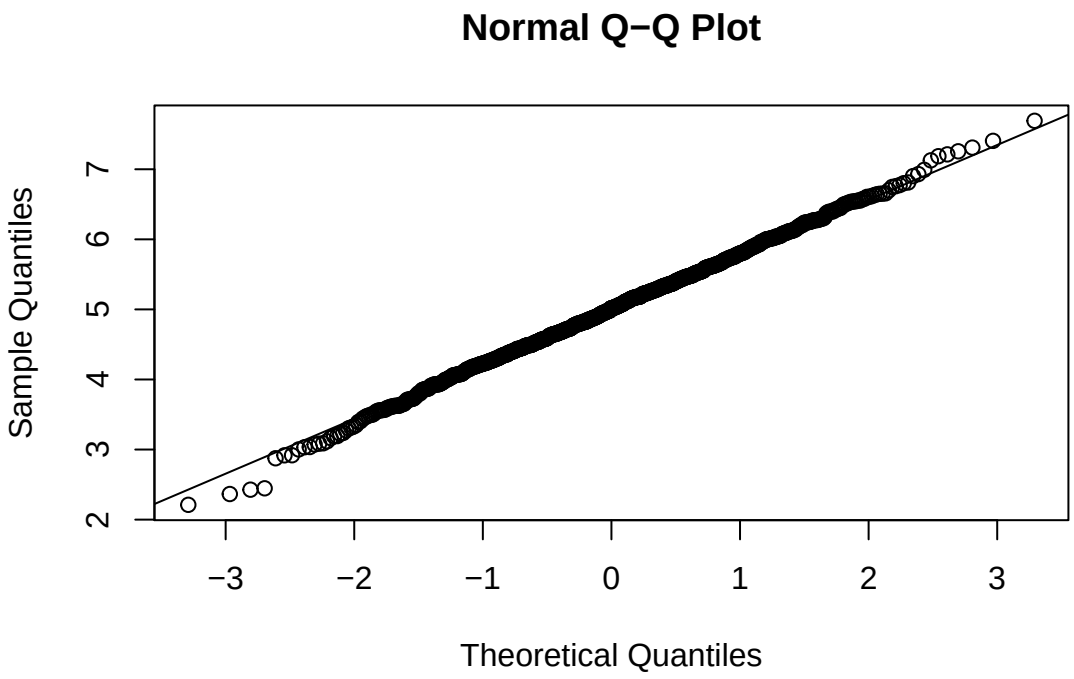
The QQ Plot can help us identify if this is likely to be a normal-like distribution:

```
qqnorm(meanData)
qqline(meanData) ## creates a line that goes from 2nd to the 4th quantile.
```



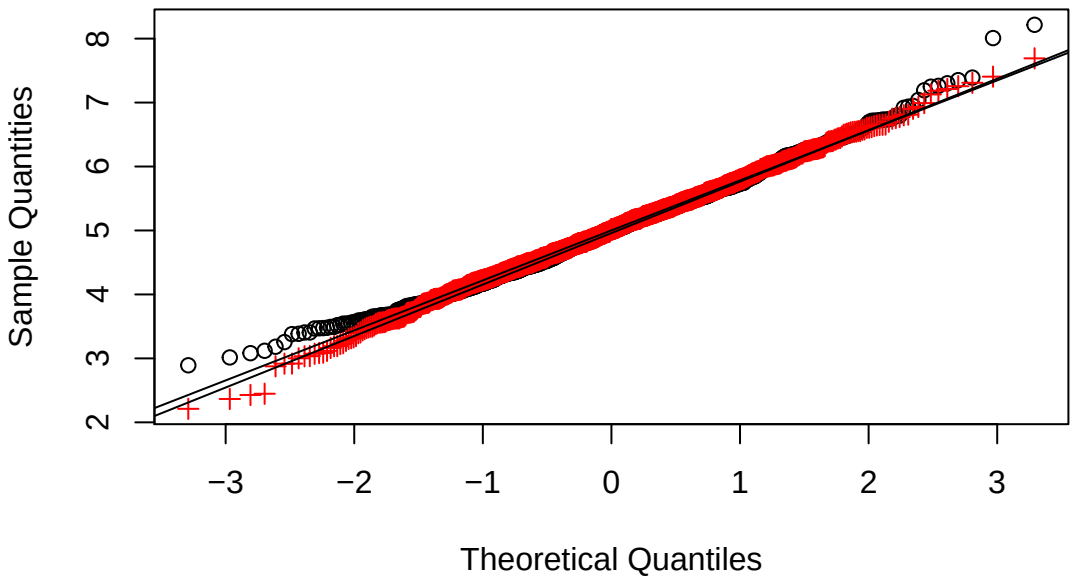
For comparison, let's see a normal distribution with the same parameters as this sample:

```
norm<-rnorm(1000, mean=mean(meanData), sd=sd(meanData))
qqnorm(norm)
qqline(norm) ## creates a line that goes from 2nd to the 4th quantile.
```



If we put all in one graph, we can see they are very similar:

```
q1 <- qqnorm(meanData, plot.it = FALSE)
q2 <- qqnorm(norm, plot.it = FALSE)
plot(range(q1$x, q2$x), range(q1$y, q2$y), type = "n",
     ylab="Sample Quantities", xlab="Theoretical Quantiles")
points(q1)
points(q2, col = "red", pch = 3) ## normal
qqline(meanData) ## quantile line, going from 0.25 to 0.75 quantile
qqline(norm) ## quantile line, going from 0.25 to 0.75 quantile
```

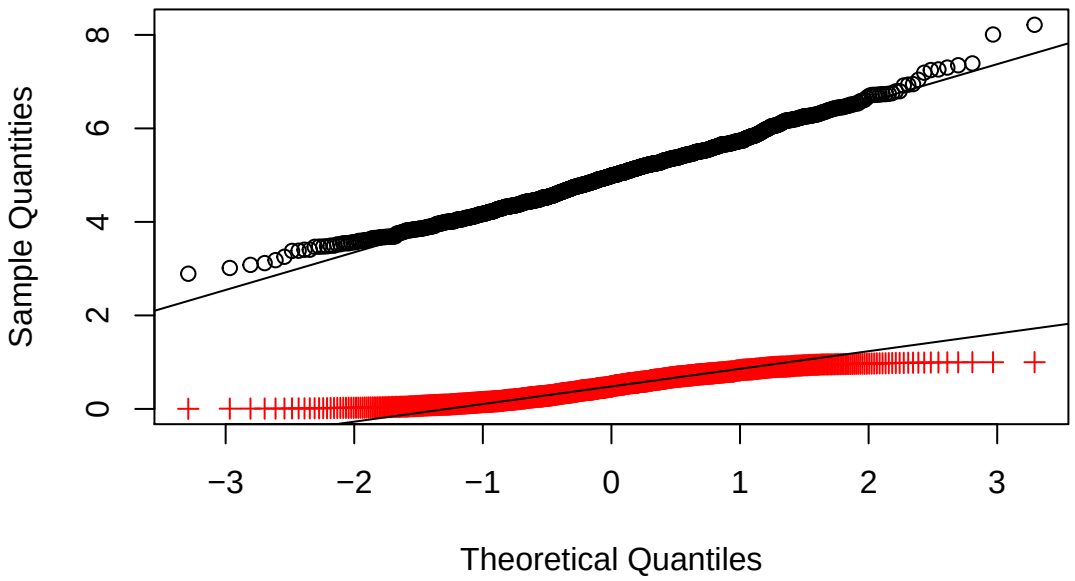


The red plot denotes the normal distribution. The other one is our sample.

As the sample size grows, the mean distribution will get more 'normal-like' as stated by CLT.

If we compare the quantiles of our sample with a uniform distribution (as requested by the exercise, in red in graph below), the following will happen:

```
unif<-runif(1000) ## a 1000 uniforms
q1 <- qqnorm(meanData, plot.it = FALSE)
q2 <- qqnorm(unif, plot.it = FALSE)
plot(range(q1$x, q2$x), range(q1$y, q2$y), type = "n",
     ylab="Sample Quantities", xlab="Theoretical Quantiles")
points(q1)
points(q2, col = "red", pch = 3) ## uniform
qqline(meanData) ## quantile line, going from 0.25 to 0.75 quantile
qqline(unif) ## quantile line, going from 0.25 to 0.75 quantile
```



Despite the fact that the means and other factors are different, we can see the difference in the graph above: a uniform distribution will have points in bothe sides of the quantile slope due to its nature of being uniform (and symmetric as we can see) while a normal one will have the points very close to the line.