Statistical Inference - Course Project: A Comparision of Means

Randy Armknecht

Saturday, August 22, 2015

Overview

In this paper we examine the exponential distribution, and compare the theoretical and actual means when taking samples. Further, we compare the theoreticl variance against that of our sample population. In the end we show that the distribution is approximately normal.

Simulations

The first simulation I ran was to generate the samples for the exponential distribution. I used a seed so that the samples generated will be reproduciable. The function replicate is useful for performing a certain task an explicit number of times. I used it to generate an exponential distribution *rexp* of 40 samples using a lambda of 0.2 as requested in the project description. I ran this simulation 1000 times and stored the results in a 40x1000 item array named *edist*

```
lambda <- 0.2
samples <- 40
interations <- 1000

set.seed(1984)
edist <- replicate(n=interations, expr=rexp(samples, lambda))
```

Sample Mean vs. Theoretical Mean

Now we can calculate the mean of the samples, and the theoretical mean, which we know is 1/lambda.

```
samp_mean <- mean(colMeans(edist))
theo_mean <- 1/lambda
c(samp_mean, theo_mean)
```

```
## [1] 4.981324 5.000000
```

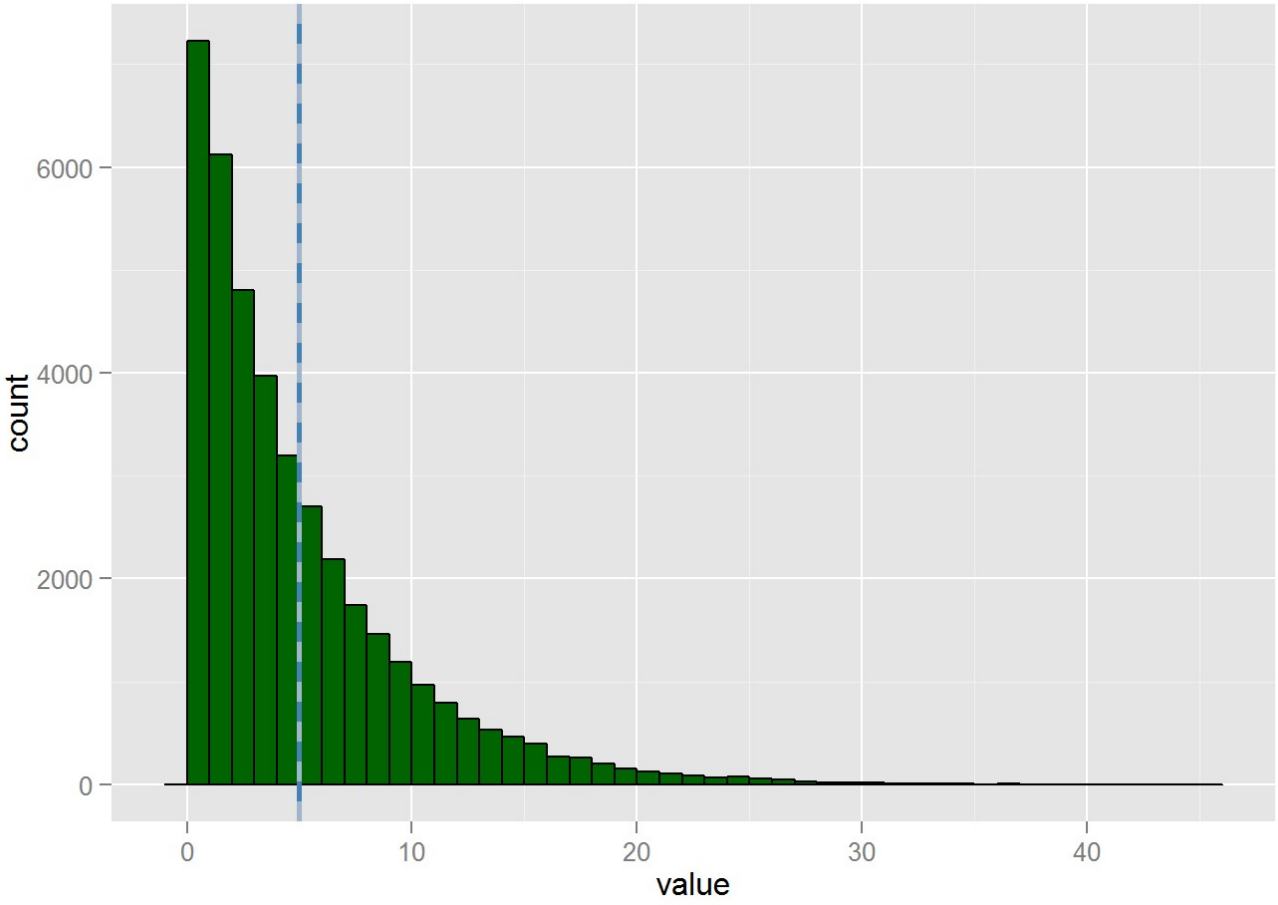
```
theo_mean - samp_mean
```

```
## [1] 0.01867599
```

Here we can see that the difference is relatively small at just 0.0186.

Let's take a look at these on a graph (blue = theoretical mean, grey = mean of samples):

```
library(ggplot2)
gg <- qplot(as.vector(edist), type="histogram", binwidth=1.0)
gg <- gg + geom_histogram(color="black", fill="darkgreen", binwidth=1.0)
gg <- gg + geom_vline(x=theo_mean, color="steelblue", size=1)
gg <- gg + geom_vline(x=samp_mean, color="slategray3", size=1, linetype="dashed")
gg <- gg + labs(x="value", y="count")
#gg <- gg + geom_rug(color="steelblue", alpha=0.1)
gg
```



Sample Varience vs. Theoretical Varience

In order to determine the varience of the sample, we must first determine the Standard Diviation of the sample by applying R's *sd* function to the means of all 1000 simulations of 40 samples. Then, we simply square the calculated Standard Diviation to determine the Varience. Variance = SquareRoot(StandardDiviation)

```
samp_sd <- sd(colMeans(edist))
samp_var <- samp_sd^2
c(samp_sd, samp_var)
```

```
## [1] 0.8038905 0.6462399
```

We can take the same approach from the Theoretical side if we recall that the Standard Diviation of a population is 1/lambda * 1/SquartRoot(SampleSize). We'll then square that, and arrive at our answer.

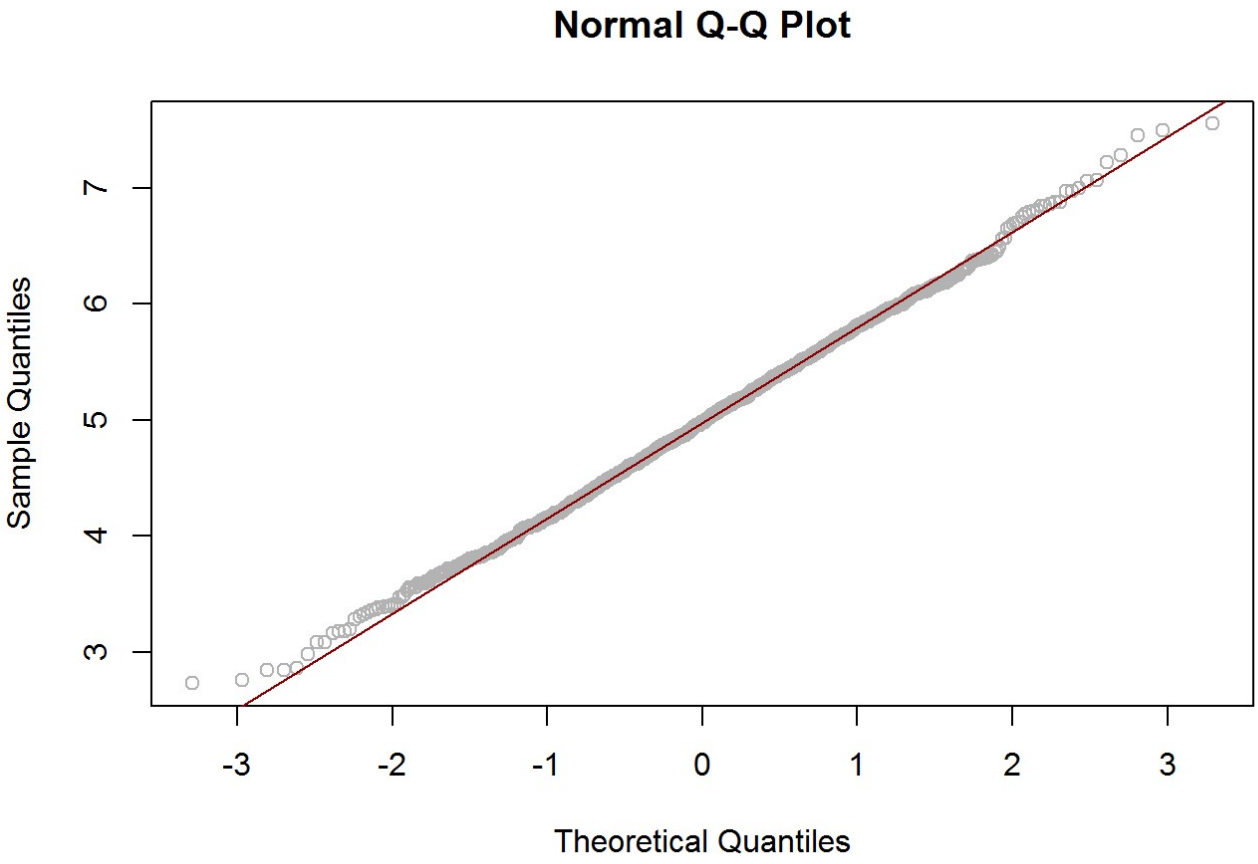
```
theo_sd <- 1/lambda * (1/sqrt(samples))
theo_var <- theo_sd^2
c(theo_sd, theo_var)
```

```
## [1] 0.7905694 0.6250000
```

Distribution is Normal

To determine this, I used the Q-Q Plot. The *qqline* command draws a line for the theoretical normal through the first and third quantiles. By also using the *qqnorm* command, I'm able to generate a QQ plot of the means of the sample. By looking at this graph, we are able to visually see that the disrtribution is relatively normal.

```
qqnorm(colMeans(edist), col="grey70")
qqline(colMeans(edist), col="darkred")
```



References

1. http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/ (http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/)
2. http://stackoverflow.com/questions/23709060/change-r-markdown-plot-width (http://stackoverflow.com/questions/23709060/change-r-markdown-plot-width)