Coursera: Statistical Inference Course Project

Sunday, September 21, 2014

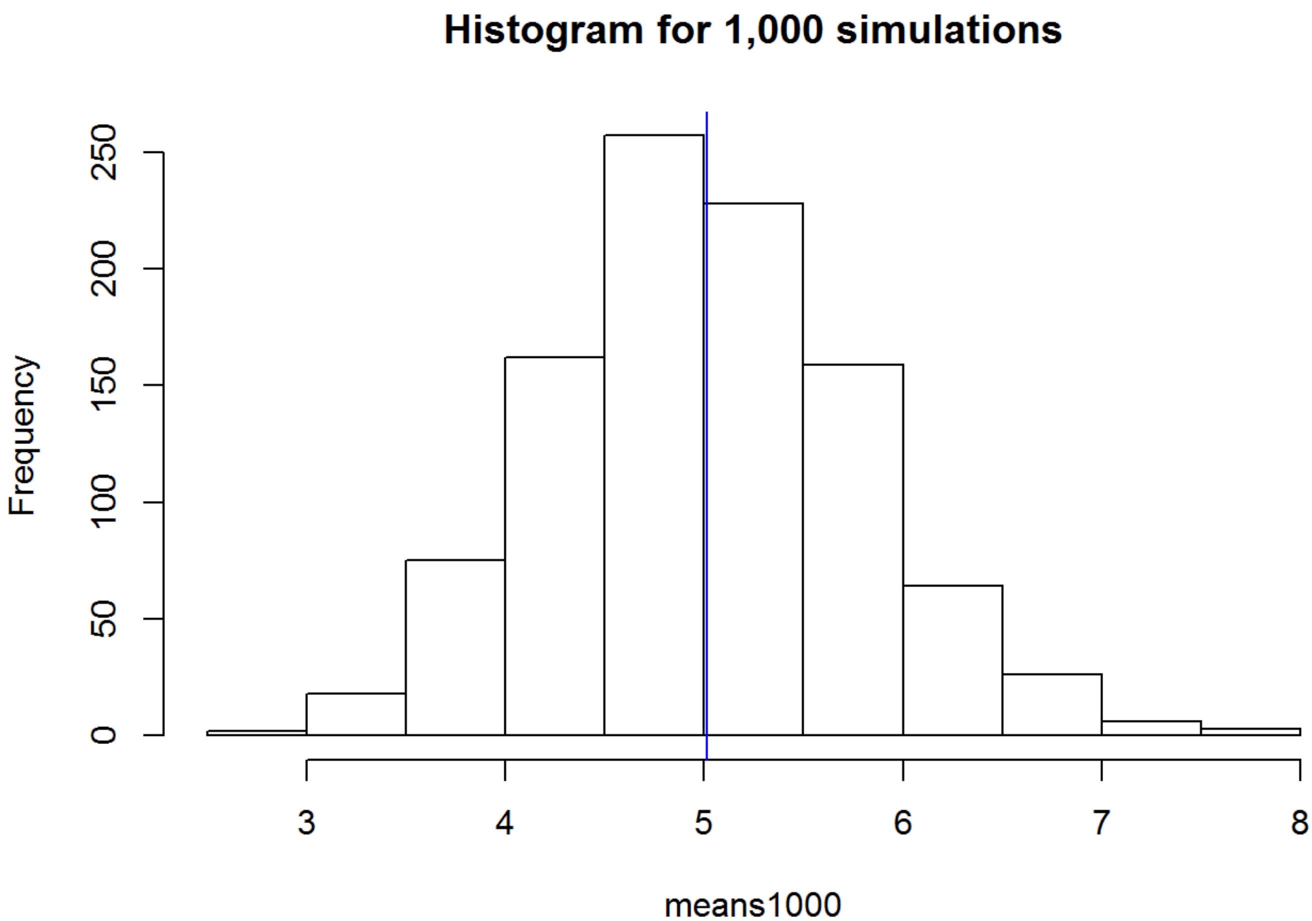
Description of the assignment

The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. In this simulation, you will investigate the distribution of averages of 40 exponential(0.2)s. Note that you will need to do a thousand or so simulated averages of 40 exponentials.Since the mean of exponential distribution is 1/lambda and lambda = 0.2, then the mean is 5.

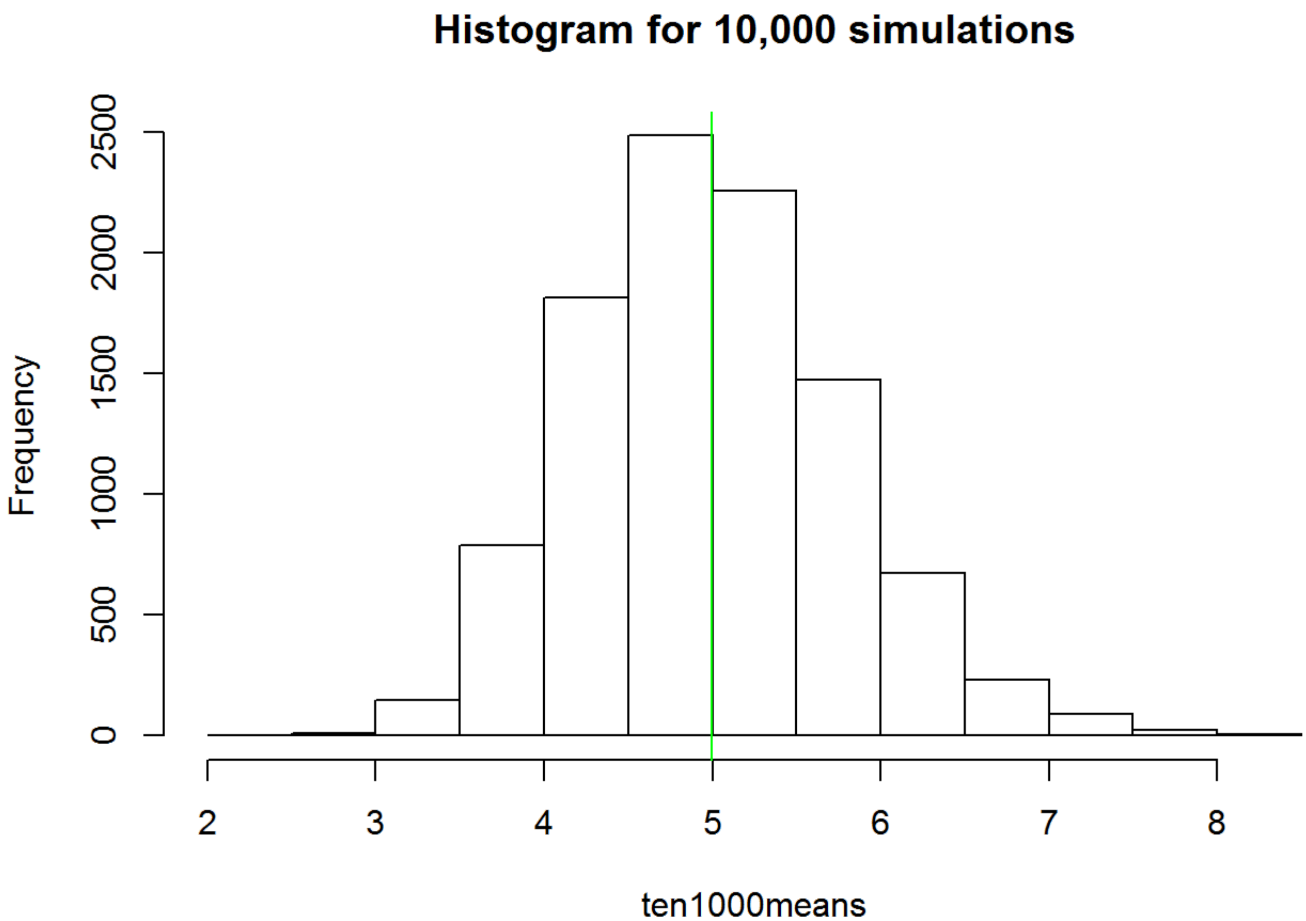
Analysis 1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.

```
simulateMean <- function(i){
set.seed(123)
vector <- c()
for(i in 1:i){
dist <- rexp(40, 0.2)
avgDist <- mean(dist)
vector <- append(vector, avgDist)
}
return(vector)
}

means1000 <- simulateMean(1000)
hist(means1000, main="Histogram for 1,000 simulations")
abline(v=mean(means1000), col = "blue")
```




```
ten1000means <- simulateMean(10000)
hist(ten1000means, main="Histogram for 10,000 simulations")
abline(v=mean(ten1000means), col = "green")
```



The vertical line (blue and green) in each histogram shows the mean. It is clear from the visualisation that these means are close to 5, which is the theoretical center of distribution.

Analysis 2. Show how variable it is and compare it to the theoretical variance of the distribution.

Variance is the square of standard deviation. The theoretical SD is 5 therefore, the theoretical variance should be 5^2 = 25.

```
var1000 <- sd(means1000)^2
var1000
```

```
## [1] 0.6005
```

```
var10000 <- sd(ten1000means)^2
var10000
```

```
## [1] 0.6117
```

The actual variance obtained above is 0.60 and 0.61 whereas our theoretical variance is 25. This demonstrates that the variance of the distribution is very variable and very different from the theoretical variance

Analysis 3. Show that the distribution is approximately normal.

For a distribution to be Normal, 68% of distribution lies between 1 SD from the mean, 95% of distribution lies between 2 SD from the mean and 97.5% of distribution lies between 3 SD from the mean. We calculate these values belwo to demostrate the above:

```
distCentre <- mean(means1000) #center of the distribution
distSD <- sd(means1000) #sd
```

We create three sets of data between 1, 2, and 3 of SD and then determine the % of data lying within each set.

```
within1SD <- means1000[means1000 >= (distCentre - distSD) & means1000 <= (distCentre + distSD)]
length(within1SD)/length(means1000)
```

```
## [1] 0.696
```

```
within2SD <- means1000[means1000 >= (distCentre - 2*distSD) & means1000 <= (distCentre + 2*distSD)]
length(within2SD)/length(means1000)
```

```
## [1] 0.95
```

```
within3SD <- means1000[means1000 >= (distCentre - 3*distSD) & means1000 <= (distCentre + 3*distSD)]
length(within3SD)/length(means1000)
```

```
## [1] 0.996
```

We see above that of values within each set are approx. 69.6%, 95%, and 99.6%, which clearly match with the distribution of values if this was a normal distribution. Hence, we see that this distribution is approximately normal.

Analysis 4. Evaluate the coverage of the confidence interval for 1/lambda

Quantile function can be used to calculate the lower and upper limit of the confidence interval.

```
q <- qt(0.975, df=length(means1000)-1)*distSD/sqrt(length(means1000))
lowerqt <- distCentre - q
upperqt <- distCentre + q
```

```
lowerqt
```

```
## [1] 4.964
```

```
upperqt
```

```
## [1] 5.06
```

The results show that the confidence interval is (4.9638, 5.0599)