Statistical Inference course project

Ben de Haan

22 Aug 2015

Introduction

In this project, I will investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). The CLT states that iid variables should have a normal distribution, even if these variables are not normally distributed.

Firstly, we need to load a library for plotting:

```
library(ggplot2)
```

Secondly, we need to set a seed to ensure reproducibility. Furthermore, for all simulations, lambda is set to 0.2. The number of iterations is 1000. I will draw 40 values of exponentials for each iteration in the simulation.

```
set.seed(123456)
lambda = 0.2
n = 40 # Number of values drawn for each iteration
reps = 1000 # Number of iterations
```

Simulation code

First we need to initialise two vectors for storing the means and variances, respectively. Then we can

```
means <- c() # Init vector for storing means
variances <- c() # Init vector for storing variances
# Simulate 1000 reps
for (i in 1:reps){
    values <- rexp(n, lambda) # Get n values
    means[i] <- mean(values) # Store the mean for each iteration
    variances[i] <- var(values) # Store the variance for each iteration
}
```

Data analysis

Theoretical mean vs. sample mean

The theoretical mean of the exponential distribution is $1/lambda$. This means that the theoretical mean is:

```
theoretical.mean <- 1 / lambda
theoretical.mean
```

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

The sample mean can be inferred by taking the mean from the vector of means. Its 95% confidence interval bounds can be derived by taking the 2.5% and 97.5% values from the distribution, which are upper and lower bound, respectively.

```
sample.mean <- mean(means)
sample.mean
```

```
## [1] 5.023
```

```
sample.mean.bounds <- quantile(variances, probs = c(0.025, 0.975))
sample.mean.lowerBound <- sample.mean.bounds[1]
sample.mean.upperBound <- sample.mean.bounds[2]
sample.mean.lowerBound
```

```
##  2.5%
## 9.767
```

```
sample.mean.upperBound
```

```
## 97.5%
## 55.13
```

The theoretical variance is 5 and the sample variance is 5.0229, which means that there is a 0.4562% deviation.

Theoretical variance vs. sample variance

The theoretical standard deviation of the exponential distribution is $1/lambda$. Since $var = sd^2$, the variance is:

```
theoretical.sd <- 1 / lambda
theoretical.var <- theoretical.sd^2
theoretical.var
```

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

The sample variance can be inferred by taking the mean from the vector of variances. Its 95% confidence interval bounds can be derived by taking the 2.5% and 97.5% values from the distribution, which are upper and lower bound, respectively.

```
sample.var.mean <- mean(variances)
sample.var.mean
```

```
## [1] 25.24
```

```
sample.var.bounds <- quantile(variances, probs = c(0.025, 0.975))
sample.variance.lowerBound <- sample.var.bounds[1]
sample.variance.upperBound <- sample.var.bounds[2]
sample.variance.lowerBound
```

```
##  2.5%
## 9.767
```

```
sample.variance.upperBound
```

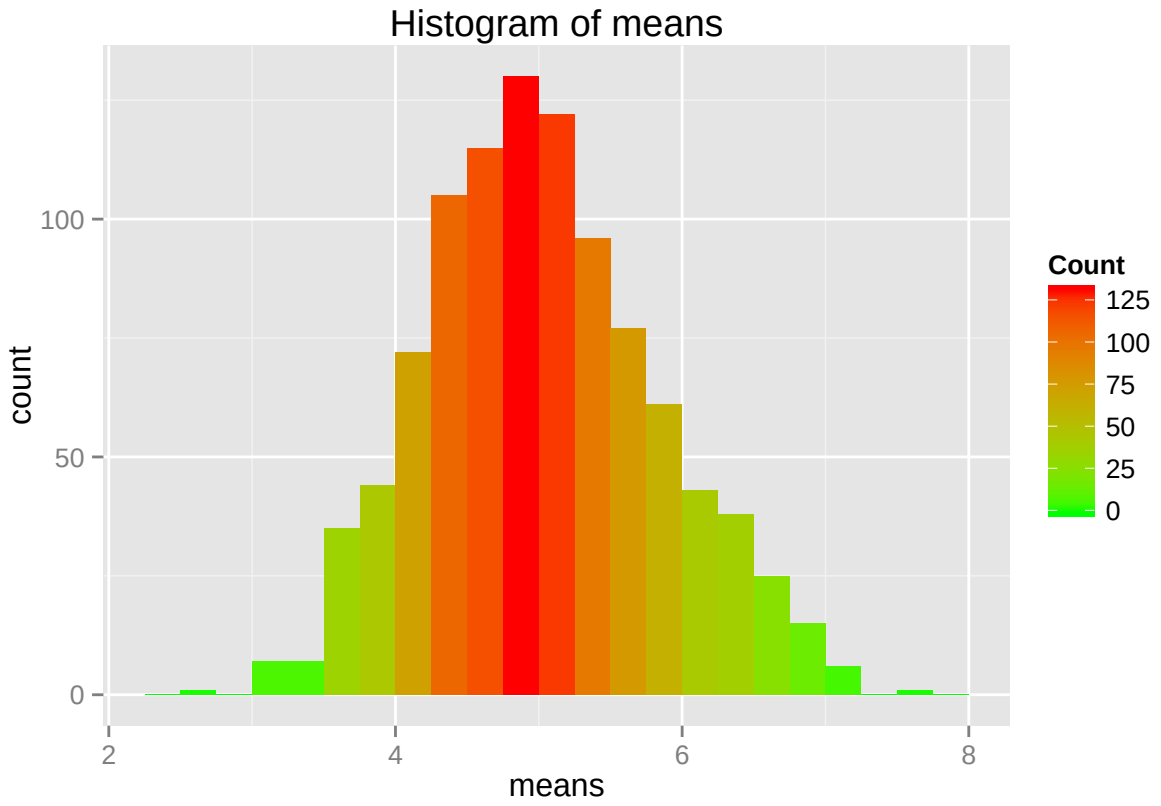
```
## 97.5%
## 55.13
```

The theoretical variance is 25 and the sample variance is 25.2425, which means that there is a 0.9606% deviation.

Distribution

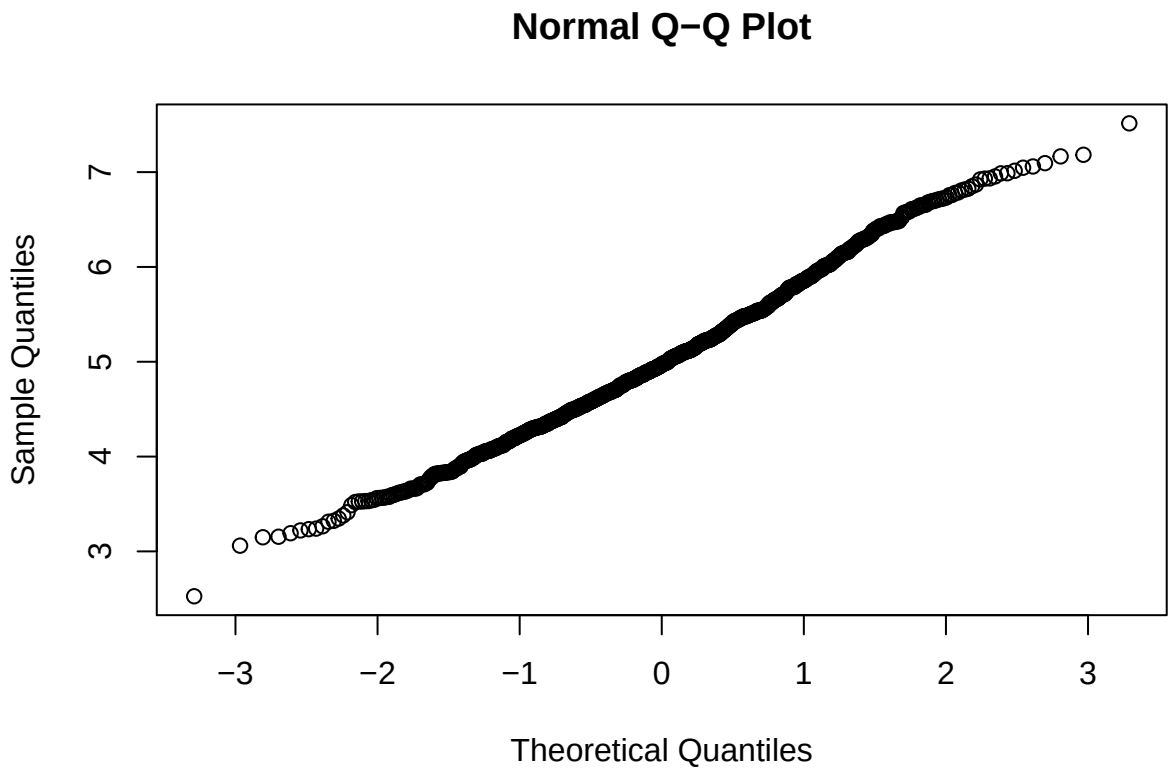
The means are distributed as follows:

```
qplot(means, binwidth = 0.25, fill=..count.., main = "Histogram of means") +
    scale_fill_gradient("Count", low = "green", high = "red")
```

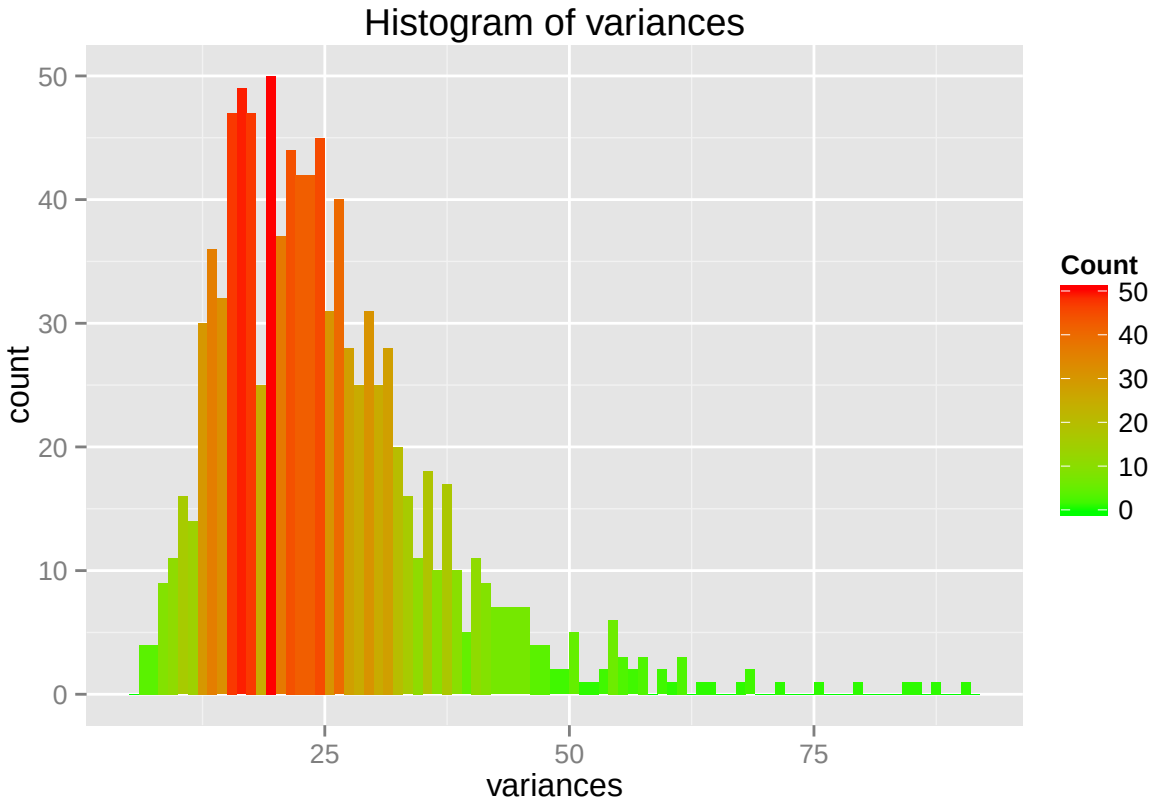


The Q-Q normality plot can grant us additional insight as to whether the means are normally distributed.

```
qqnorm(means)
```



Variances are distributed as depicted by the following histogram:



We can see from both the sample mean and variance histograms and Q-Q plot that the distribution is right skewed.

The Shapiro-Wilk test of normality can test the means vector for a normal distribution. The null hypothesis here is that the data are normally distributed.

```
shapiro.test(means)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  means
## W = 0.9933, p-value = 0.0001679
```

Since $p < 0.05$, the test reveals the means drawn from the Exponential distribution are not normally distributed, which is not expected since the CLT would predict so. Since it is unlikely the Central Limit Theorem is false, the test could be redone using a higher number of values drawn and/or more simulations.