Statistical Inference Course Project: Exponential Distributions and the Central Limit Theorem

equinaut

October 23, 2015

Overview

This report is to demonstrate Central Limit Theorem in action by analyzing a random sampling of exponentials' means and variances. We will compare these values versus the expectations derived from theoretical computation.

Environment Set Up

```
# Libraries
library(ggplot2)
library(grid)

# Settings
set.seed(1, kind = "Mersenne-Twister")
lambda <- 0.2
n <- 40
sample.size <- 1000
```

Simulations

We will be comparing two sets of values:

- `theor.*` - denotating the theoretical mean (`* = .mean`) and standard deviation (`* = .sd`) values
- `dist.e` - a set of 1000 means of random exponential samples, each of sample size 40

The code below builds both sets of values using the specified `lambda` . In a separate data frame entitled `dist.stats` we will list the mean and standard deviation of `dist.e` and compare it to the theoretical values, both of which will equal 1/lambda.

```r
# Theoretical Values
theor.mean <- 1/lambda
theor.sd <- 1/lambda

# Simulated Values
dist.e <- NULL
# Populates dist.e with the means of 1000 rexp() distributions of sample size 40
for(i in 1:sample.size) dist.e <- c(dist.e, mean(rexp(n, lambda)))

# Comparison Data Frame
dist.stats <- data.frame(
                rbind(c(theor.mean, mean(dist.e)), c(theor.sd, sd(dist.e))),
                row.names = c("mean", "sd"))
colnames(dist.stats) <- c("Theoretical", "dist.e")

dist.stats
```

```
##      Theoretical    dist.e
## mean           5 4.9900252
## sd             5 0.7817394
```

Numerically speaking the means appear to be nearly identical, while the standard deviations are vastly different at 5 and .78, respectively. The former occurence is to be expected. Assuming proper random sampling has occurred, as the sample size increases the average (*mean*) sample mean will converge upon the mean of the parent population. This "mean of means" is fundamental to the Central Limit Theorem. Although the underlying population itself may not be normally distributed, as is the case of this exponential distribution, the distribution of means with a large enough sample size will be. Similarly the standard deviation of the sample won't necessarily reflect the population; this is because the standard deviation of our particular data set, `dist.e` , is the standard deviation of the *sample means* themselves.

This concept is illustrated below. The first figure depicts the distribution of 1000 random exponential values, `dist.a` , while the second figure is the distribution of `dist.e` . Demarcated in red and blue is each distribution's mean and 68% confidence interval, respectively.

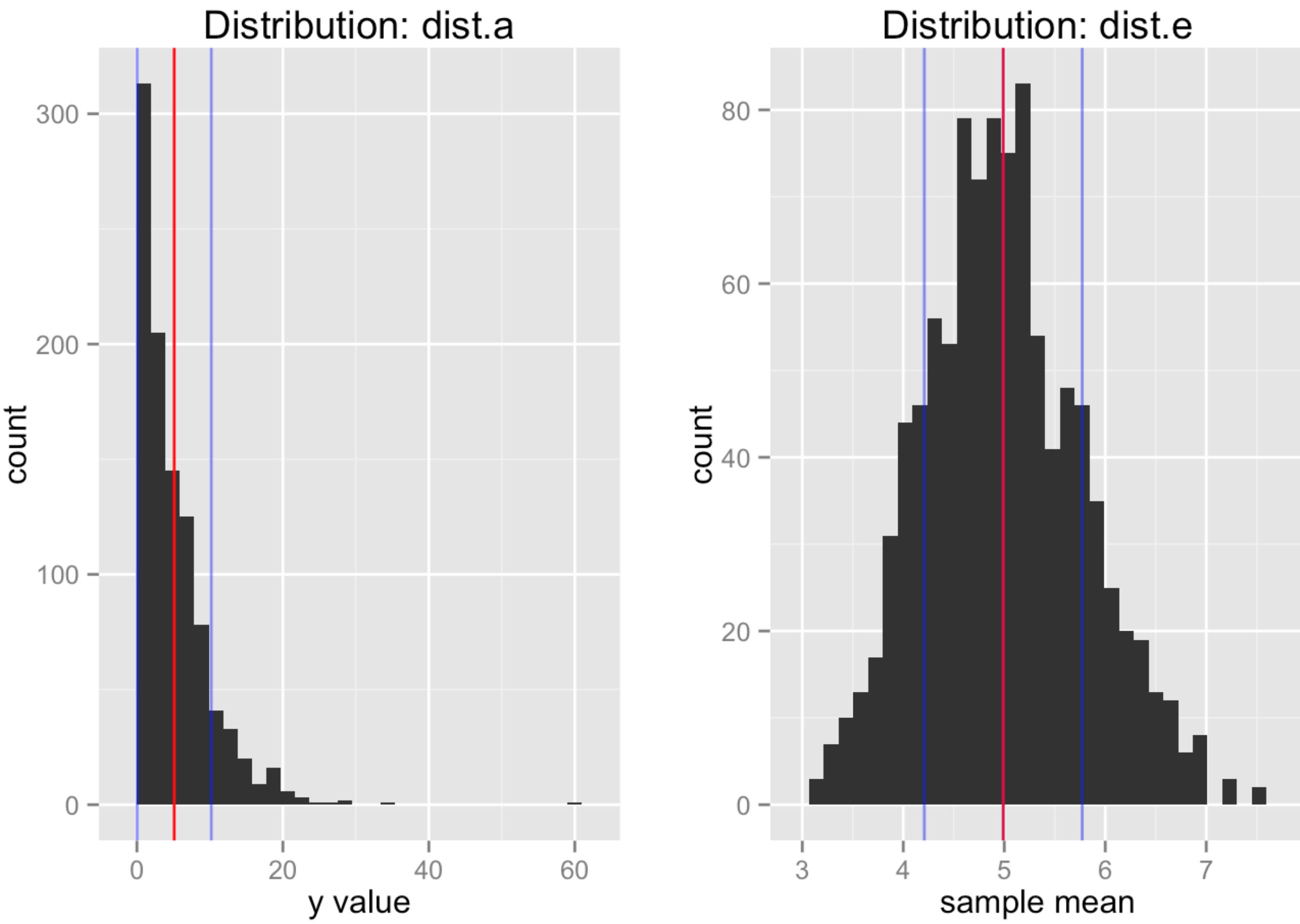
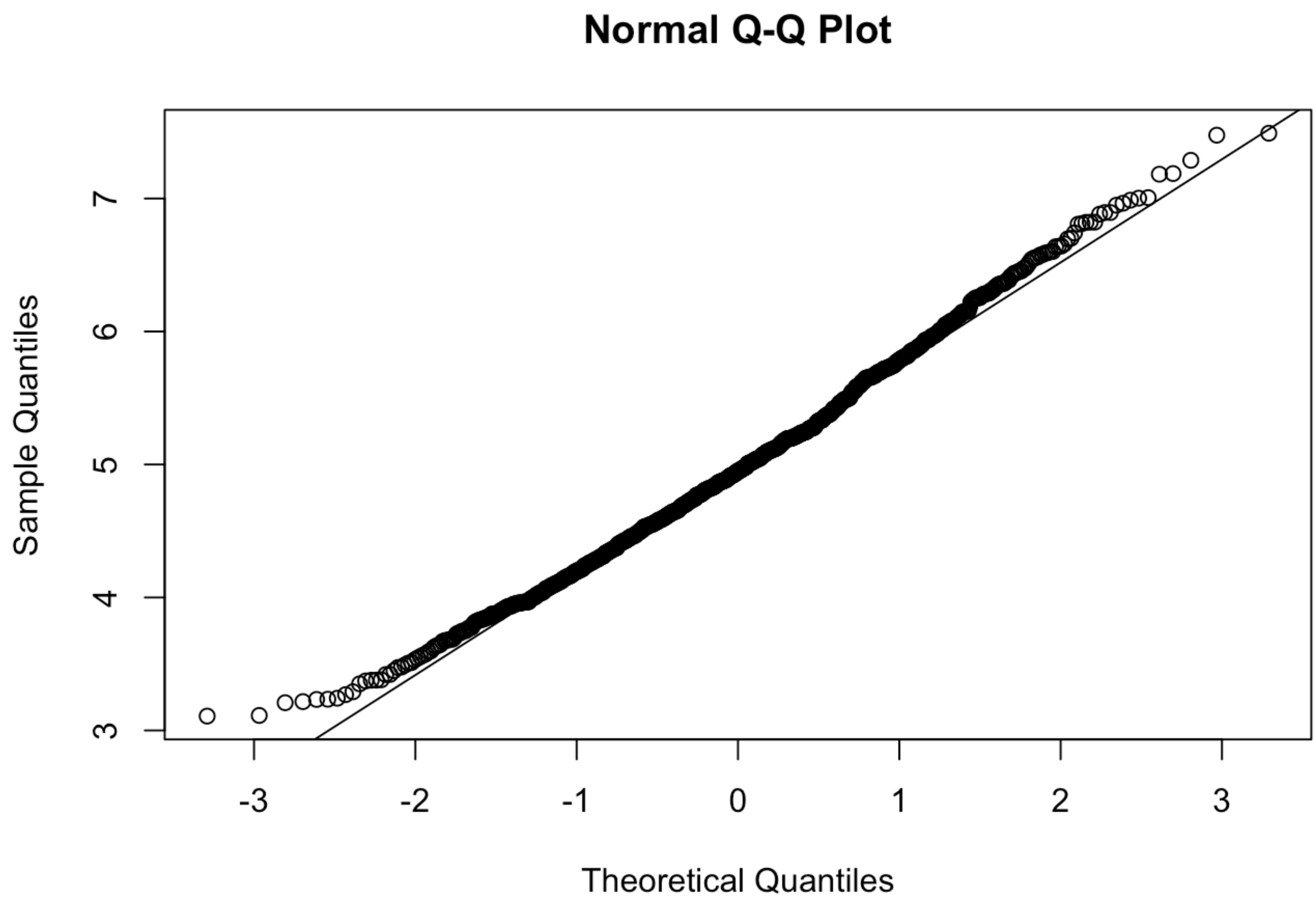


Figure two has the classic bell curve shape to it and from strictly visual analysis appears normally distributed. Conversantly figure 1 – the distribution of random exponentials – is tremendously left skewed. Yet despite this difference in distribution and variance between the plots their means both converge on 5.

Testing for Normality

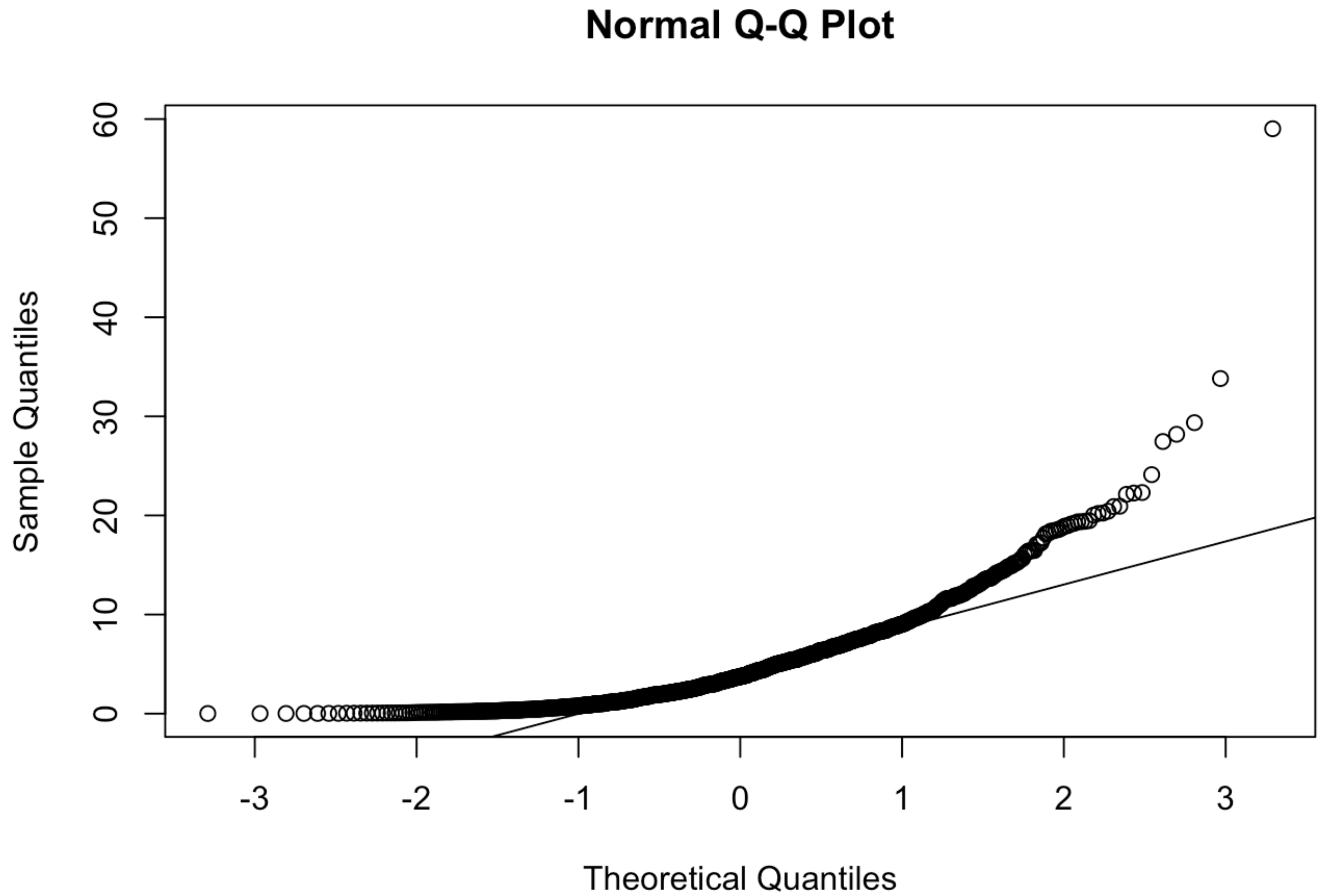
To extend the basic visual assessment above several methods are at our disposal to determine normality. We will employ only one, namely a normal Q-Q plot. A normal distribution will appear as a linear line from the lower left of the plot to the upper right with little to no deviation of the points.[1]

```
## Q-Q Plot
qqnorm(dist.e)
qqline(dist.e)
```



The Q-Q plot indicates that `dist.e` is indeed normally distributed but is light tailed, as evidenced by the large adherence to the normal line with divergence in the corners.

As a final ode to the power of the Central Limit Theorem, below is a Q-Q plot of `dist.a`, which again is a random distribution of 1000 exponentials. Unlike the sampling of means, this Q-Q is quite obviously not normal.



References

[1] How to interpret a QQ plot - http://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot (http://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot)