Analysis of the Exponential Distribution and its relation to the Central Limit Theorem

Elmar Langholz

April 19, 2015

Overview

In this document, we will analyze the exponential distribution (see Appendix: Exponential distribution properties) and its relation with the central limit theorem. We will compare, through simulation, the sample mean and variance with the theoretical mean and variance. Similarly, we will show that the distribution of averages is approximately normal.

Simulations

In order to perform the exponential distribution simulations lets first define a helper function that will:

1. Generate `sampleSize * simulationCount` exponentials with a rate of `lambda`.
2. Organize these in `data.frame` of `simulationsCount` rows and `sampleSize` columns.
3. Determine the `mean`, `var` and `sd` for each row.

The return value will be a data frame of `simulationsCount` rows by `sampleSize + 3` columns. Each row is a simulation and the corresponding column indexes [1, sampleSize] (named X1 through XsampleSize) are the exponentials. Column indexes `sampleSize + 1`, `sampleSize + 2` and `sampleSize + 3` is the mean (named mean), variance (named var) and standard deviation (named sd) accordingly.

```
exponentialDistributionSimulation <- function (lambda, sampleSize, simulationsCount = 1) {
    simulations <- rexp(sampleSize * simulationsCount, rate = lambda)
    simulations <- matrix(simulations, simulationsCount)
    indexes <- 1:sampleSize
    simulationsMean <- apply(simulations[, indexes], 1, mean)
    simulationsVariance <- apply(simulations[, indexes], 1, var)
    simulationsStandardDeviation <- apply(simulations[, indexes], 1, sd)
    simulations <- data.frame(simulations,
                              mean = simulationsMean,
                              var = simulationsVariance,
                              sd = simulationsStandardDeviation)
    simulations
}
```

For the simulations, lets assume that the rate is $\lambda = \frac{1}{5}$ (0.2). We will perform one thousand simulations (1,000) each with a sample size of fourty (40) exponentials.

```
seed <- 31337
set.seed(seed)              # make this reproducible by others
lambda <- 1/5               # the exponential rate
sampleSize <- 40            # the sample size
simulationsCount <- 1000    # the number of simulations
simulations <- exponentialDistributionSimulation(lambda, sampleSize, simulationsCount)
```

Sample Mean versus Theoretical Mean

Given $\lambda = \frac{1}{5}$, the theoretical mean of the exponential distribution is defined as $E[X_i] = \frac{1}{\lambda} = \frac{1}{\frac{1}{5}}$.

```
meanTheoretical <- 1 / lambda
```

The simulated sample mean is defined as the mean of the sample mean of all simulations.

```
meanSample <- mean(simulations$mean)
```

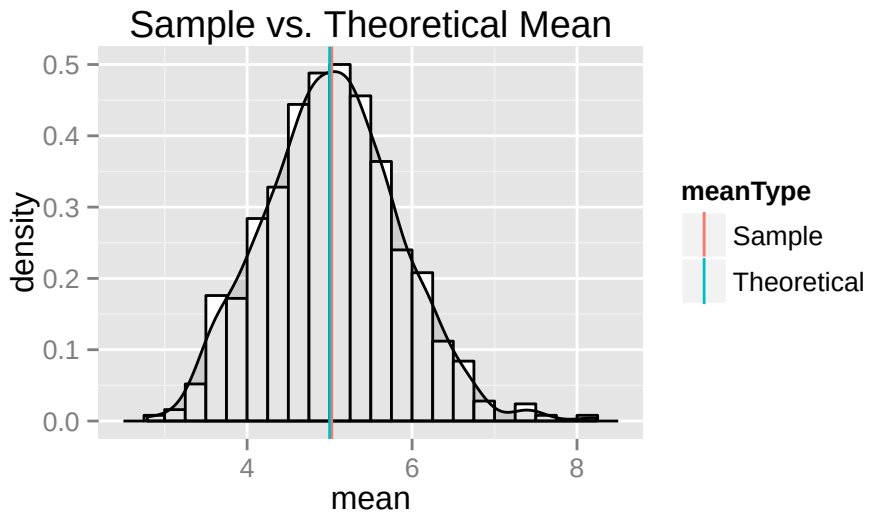

Figure 1: The simulated sample histogram and density with the sample and theoretical mean

When we compare these two, we are able to see that while the **theoretical mean is 5**, the **sample mean is 5.026**. Just as shown in *Figure 1*, the theoretical and sample mean are very close together and in fact **these two diverge by 0.026**.

Sample Variance versus Theoretical Variance

Since standard deviation is easier to compare to the mean because it is on the same scale, we will use this instead of the variance. Given $Var[X_i] = \frac{\sigma^2}{n}$ and $\sigma = \frac{1}{\lambda}$, the theoretical standard deviation of averages is defined as $Sd[X_i] = \sqrt{\frac{\sigma^2}{n}} = \frac{\frac{1}{\lambda}}{\sqrt{n}} = \frac{1}{\lambda\sqrt{n}}$

```
standardDeviationTheoretical <-  1 / (lambda * sqrt(sampleSize))
```

The simulated sample standard deviation of the mean sample distribution is defined as the standard deviation of the sample mean of all simulations.

```
standardDeviationSample <- sd(simulations$mean)
```

When we compare these two, we are able to see that while the **theoretical standard deviation is 0.7906**, the **sample standard deviation is 0.813**. Through *Figure 2* we are able to compare these through a boxplot and demonstrate that the theoretical and sample standard deviation (and therefore variances) are very close together and in fact **these two diverge by 0.0224**.

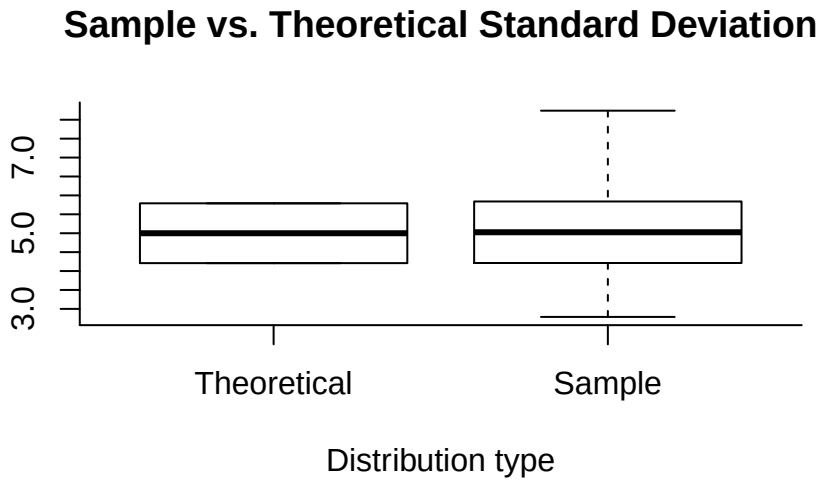

Figure 2: The simulated sample and theoretical mean and mean (SD) [mean - sd, mean + sd] comparison

For completeness, the theoretical variance can be calculated as $Var[X_1] = \frac{\frac{1}{\lambda^2}}{n} = \frac{1}{n\lambda^2}$ and the sample variance is the variance of all the sample means.

```
varianceTheoretical <-  1 / (sampleSize * lambda^2)
varianceSample <- var(simulations$mean)
```

The **theoretical variance is 0.625** and the **sample variance is 0.6609**. They **diverge by 0.0359**

Distribution

To validate that the distribution of the sample means is well-modeled by a normal distribution, lets compare the mean histogram with the sample mean distribution and the theoretical distribution $N(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$.

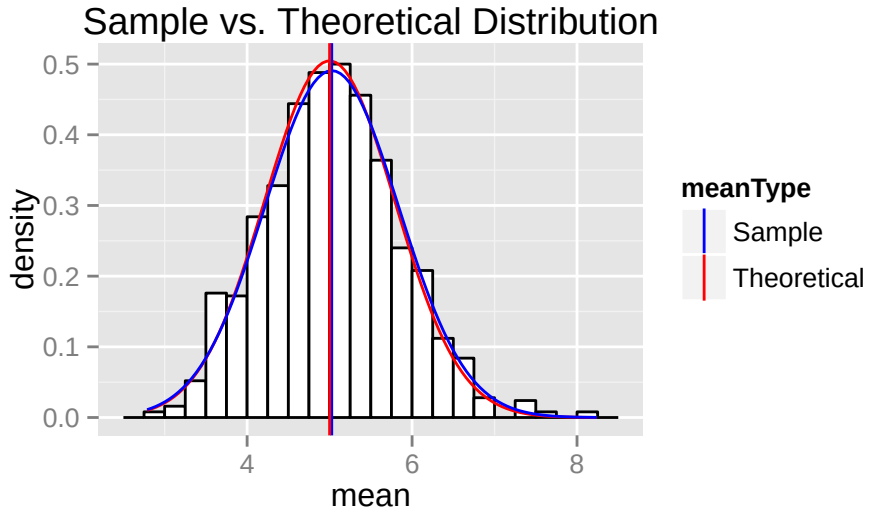

Figure 3: The sample vs. the theoretical distribution shows that these are very close to the normal distrubtion

Visually comparing these, both curves and the histogram match up. Therefore we can say that *the sampling distribution of the sample mean approximates the normal distribution.* To further land this point we also include a Q-Q plot in Appendix: Q-Q Plot for the sampling distribution.

Appendix

Exponential distribution properties

- The mean μ is defined as $\frac{1}{\lambda}$.
- The variance σ^2 is defined as $\frac{1}{\lambda^2}$.
- The standard deviation σ is defined as $\frac{1}{\lambda}$

The exponential distribution can be simulated in R with `rexp(n, lambda)` where lambda is the rate parameter.

Q-Q Plot for the sampling distribution

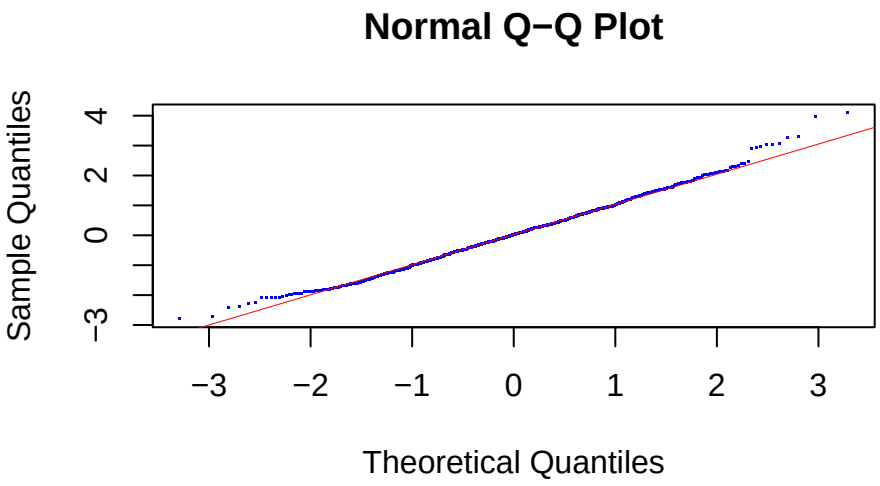

Figure 4: The Q-Q plot of the sampling mean in which the linearity of the points suggests that the data is normally distributed