Statistical Inference Course Project - Part 1

Overview

The Central Limit Theorem states that the distribution of normalized iid variables becomes that of a standard normal as the sample size increases, i.e.

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

becomes that of a standard normal for large n, and one can say that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$.

This report comprises an investigation of this result for the case of the Exponential Distribution.

The Exponential Distribution is parameterized by λ, and is such that the population mean, μ is equal to $1/\lambda$, as is the population standard deviation, σ.

It will be shown that, for a large sample size, the following three facts hold:

- The mean of the sample means approximates μ
- The variance of the sample means approximates σ^2/n
- The distribution of the normalized sample means approximates the standard normal

For the duration of the report a sample size, n, of 40 will be used, alongside λ of 0.2. 1000 simulations will be performed.

Simulations

One can sample from an Exponential Distribution in R using the rexp function, parameterized by sample size and λ, as follows:

```
rexp(12, 0.2)
```

```
##  [1]  0.9916841  3.3044763  1.4174552  0.1909595  2.3658831  7.3181357
##  [7]  1.5699229  2.0506478  5.9579891  3.5743124  6.7235780 12.0434221
```

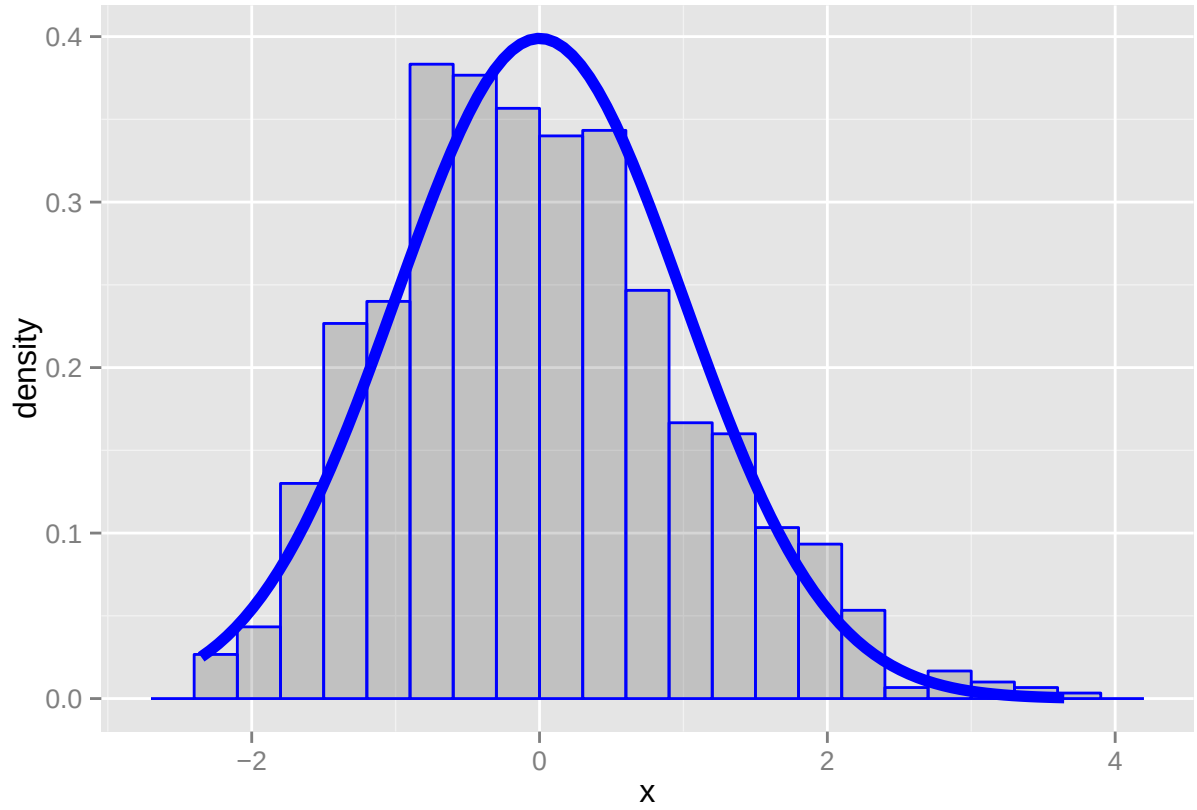
Appendix A lists the simulate function, used to generate the samples using rexp, and the calculateMeans function, which calculates the mean of each sample.

Sample Mean versus Theoretical Mean

As described in the overview, the population mean of the exponential distribution is equal to $1/\lambda$. Since we are setting $\lambda = 0.2$, we find that our population mean, μ, is equal to 5.

Sample means were calculated for each of our 1000 simulations and their mean compared to that predicted by the CLT, as listed in Appendix B. As can be seen, the mean of the sample means (4.988) does indeed approximate the population mean.

We can also standardize the distribution of means (with parameters per the CLT) and compare its plot to that a standard normal, as in the chart overleaf (code in Appendix B). It is clear that our standarized mean is close to 0.



Sample Variance versus Theoretical Variance

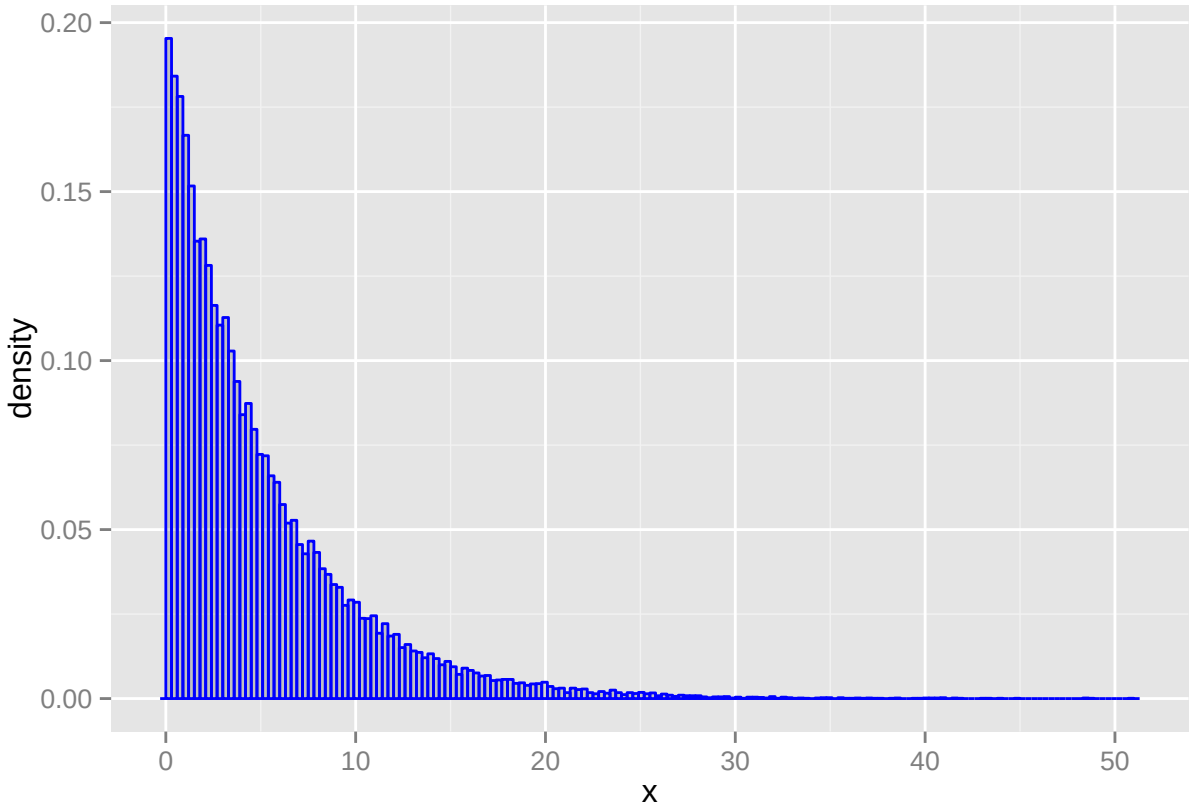
The standard deviation of the exponential distribution is equal to $1/\lambda$ and its variance therefore equal to is equal to $1/\lambda^2$. Since we are setting $\lambda = 0.2$, we find that our population variance is equal to 25. The CLT states that the variance of the sample means should be approximately the population variance divided by the sample size (σ^2/n). Since we are using a sample size of 40, we expect a variance of approximately 0.625.

The variance of the sample means was calculated as listed in Appendix C and also compared to that precicted by the CLT. Again, the variance of the sample means (0.619) well approximates the CLT's prediction.

We can also revisit the plot above and confirm that the standarized variance matches that of the standard normal closely.

Distribution

It is informative to compare the distribution of a large sample to that of the distribution of means of many samples that we have already seen (code in Appendix D). Clearly this distribution (plotted overleaf) is far from normal, very unlike the distribution we saw above.



In comparison, we have seen above that the standarized distribution of means approximates the standard normal very closely and the plot in Appendix D shows this to be true for the unstandardized distribution, as expected.

One may also use the qqplot and qqline functions, in which the quartiles of the distribution are compared to those of the standard normal. Proximity to the superimposed line may be considered a measure of similarity. Such proximity can be observed in our plots in Appendix D.

Conclusion

It has been demonstrated above that the Central Limit Theorem applies to the distribution of means of the Exponential Distribution, through comparison of the distributions' means and variances and through visual comparison of curve shape and quartiles.

Appendix A - Simulations

Generating Samples

```
simulationCount <- 1000

distribution <- function(n) rexp(n, lambda)
lambda  <- 0.2
mu <- 1 / lambda
sigma <- mu

simulate <- function(sampleSize) {
    # generate a matrix with simulationCount rows and sampleSize elements in each row
    matrix(distribution(sampleSize * simulationCount), simulationCount)
}
```

Calculating Sample Means

```
calculateMeans <- function(simulations) {
    apply(simulations, 1, mean)
}
```

Appendix B - Sample Mean versus Theoretical Mean

```
sampleSize <- 40
simulations <- simulate(sampleSize)
simulationsMeans <-calculateMeans(simulations)

mean <- mean(simulationsMeans)
print(paste("mean of sample means:", as.character(round(mean, 3))))
```

```
## [1] "mean of sample means: 5.047"
```

```
print(paste("population mean:", as.character(round(mu, 3))))
```

```
## [1] "population mean: 5"
```

```
standardizeMeans <- function(simulationsMeans, sampleSize) {
    standardize <- function(x) (x - mu) / (sigma / sqrt(sampleSize))
    sapply(simulationsMeans, standardize)
}

library(ggplot2)
standardizedMeans <- standardizeMeans(simulationsMeans, sampleSize)
g <- ggplot(data.frame(x = standardizedMeans), aes(x = x)) +
        geom_histogram(alpha = .20, binwidth=0.3, colour = "blue", aes(y = ..density..))
g + stat_function(fun = dnorm, size = 2)
```

Appendix C - Sample Variance versus Theoretical Variance

```
variance <- var(simulationsMeans)
print(paste("variance of sample means:", as.character(round(variance, 3))))
```

```
## [1] "variance of sample means: 0.619"
```

```
print(paste("sigma^2 / n:", as.character(round(sigma^2 / sampleSize, 3))))
```

```
## [1] "sigma^2 / n: 0.625"
```

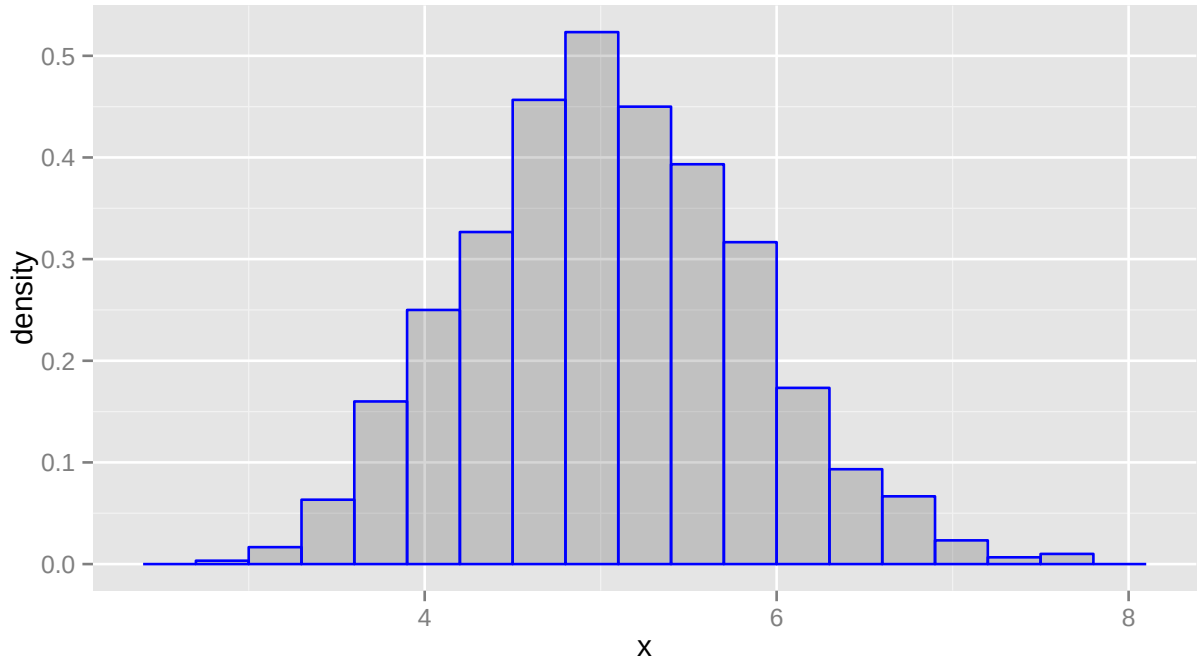
Appendix D - Distribution

Plot Large Sample

```
library(ggplot2)
ggplot(data.frame(x = x), aes(x = x)) +
        geom_histogram(alpha = .20, binwidth=0.3, colour = "blue", aes(y = ..density..))
```

Distribution of Sample Means

```
library(ggplot2)
ggplot(data.frame(x = simulationsMeans), aes(x = x)) +
        geom_histogram(alpha = .20, binwidth=0.3, colour = "blue", aes(y = ..density..))
```



Comparison of Quartiles

```
qqnorm(standardizedMeans)
qqline(standardizedMeans)
```