Statistical Inference - Course Project Part 1

Arturo Equihua

Saturday, April 18th, 2015

Overview

In this document an analysis is done on the exponential distribution to show some of its properties with regard to the approximation of the population mean and standard deviation, and also the relationship to the Central Limit Theorem.

To that end a number of simulations will be presented for an exponential distribution that has $\lambda = 0.2$, for random samples of size $n = 40$.For more details about the exponential distribution see the Wikipedia definition.

Simulations

Let X be a random exponential variable with $\lambda = 0.2$. We are going to run 1000 random samples of size $n = 40$. The following is a partial view of the dataframe **Xbardat**.

```
##          x size
## 1 5.016017   40
## 2 5.005042   40
## 3 5.401340   40
## 4 4.966949   40
## 5 5.709411   40
## 6 4.940977   40
```

Sample Mean versus Theoretical Mean

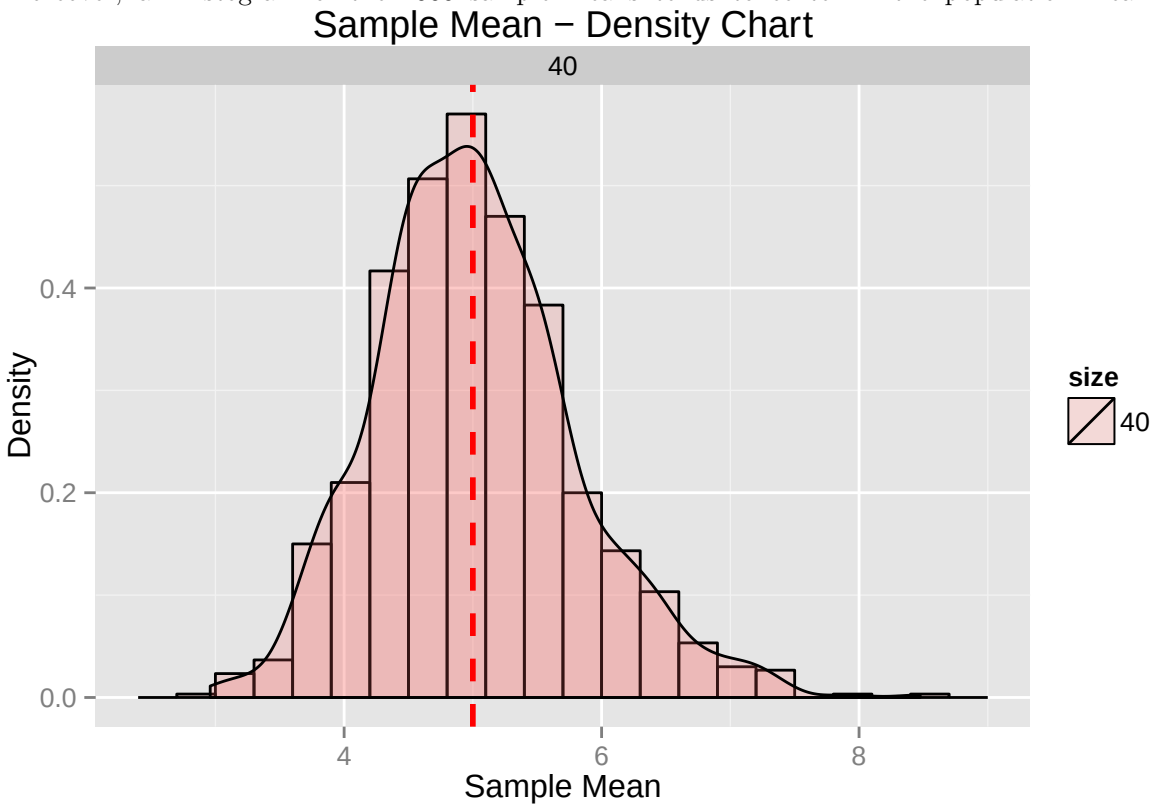
By definition, if the exponential distribution has $\lambda = 0.2$, the mean and standard deviation of the population are defined as $\mu = 1/\lambda = 5$, and $\sigma = 1/\lambda = 5$.

As per the definition of consistent estimators, it is expected that the Sample Mean from the simulation be a good approximate to the population mean. One way to show this property is by calculating the overall mean of the 1000 sample means and compare it to the population mean:

- Theoretical mean : 5
- Sample mean : 5.021

As shown above, the sample mean (`XbarMean`) is close enough to the population mean $\mu = 1/\lambda = 1/0.2 = 5$.

Moreover, an histogram of the 1000 sample means tends to center in the population mean as well:



Variance of the Sample Mean versus Theoretical Variance

It can also be shown that, for the distribution of the sample mean, its standard deviation must converge to $\sigma/\sqrt{n} = 5/\sqrt{40}$, approximately 0.7905694 :

[1] "Pop StdDev / sqrt(n) =" "0.790569415042095"
[3] " Std Dev of Simulation = " "0.77952841427613"

Since the sample size is relatively small (n=40), it is expected that for higher values of n, the histogram will tend to concentrate more to the population mean (that is, the variance of the sample mean distribution will be smaller). Still, as shown in the previous pages, the sample mean for n=40 constitutes a good estimator of the exponential population mean.

Distribution

Finally, it is relevant to contrast the actual exponential distribution with the distribution of the sample mean of the same distribution, and demonstrate the validity of the Central Limit Theorem.

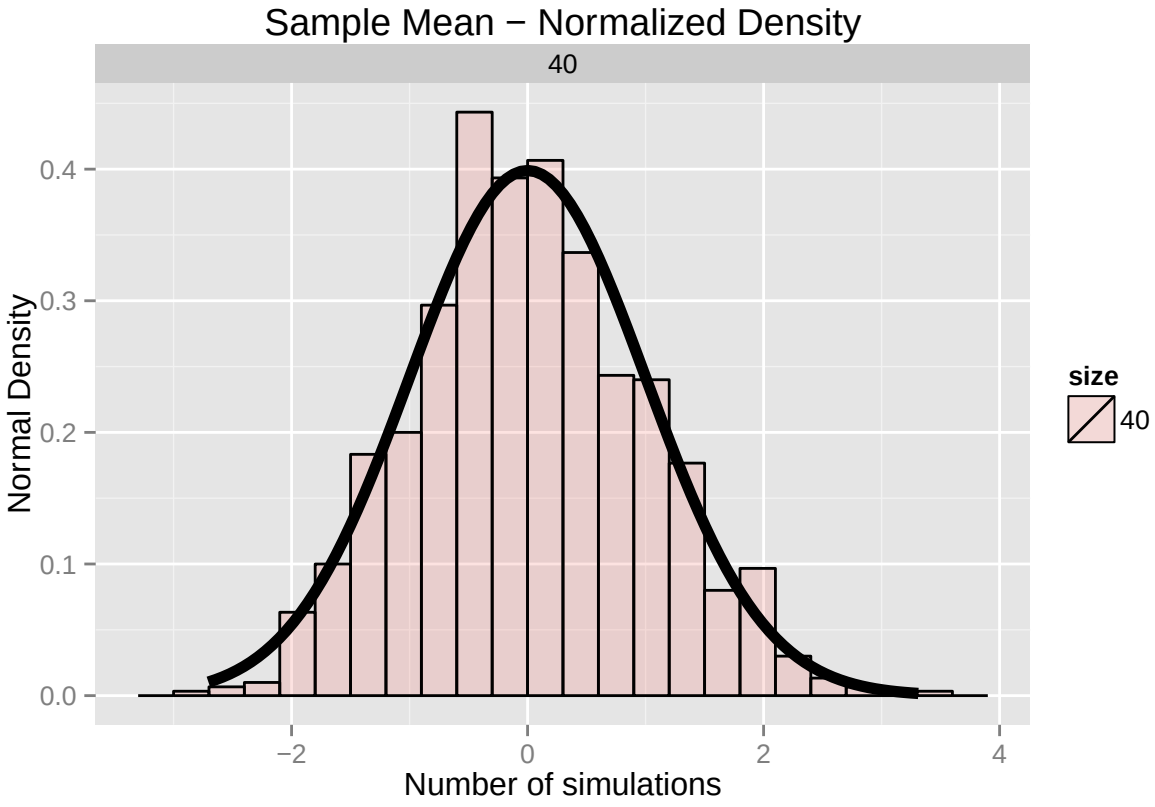
As shown below, for a simulation of 1000 random exponential values of $\lambda = 0.2$, the histogram resembles an exponential curve:

However, when shifting to the sample mean as a random variable in itself, it can be found that the density curve is very similar to that of the standard normal curve:

To simulate this situation:

- Let X_i be a random exponential variable of $\lambda = 0.2$
- Remember that $\mu = E[X_i] = 5$, $Var(X_i) = 25$, SE $\sqrt{25/n} = 5/\sqrt{n}$
- Let's take $n = 40$ random exponentials, and normalize them by taking their mean, subtracting off 5, and dividing by $5/\sqrt{40}$, and repeat this 1000 times.

```
cfunc <- function(x, n) sqrt(n) * (mean(x) - muexp) / sdexp
XbarNormdat <- data.frame(
  x = c(apply(matrix(rexp(nosim*n,lambda),
                     nosim), 1, cfunc, n)
  ),
  size = factor(rep(c(n), rep(nosim, 1))))
g <- ggplot(XbarNormdat, aes(x = x, fill = size)) + geom_histogram(alpha = .20, binwidth=.3, colour =
g <- g + labs(x = "Number of simulations", y = "Normal Density")
g <- g + ggtitle("Sample Mean - Normalized Density")
g <- g + stat_function(fun = dnorm, size = 2) + facet_grid(. ~ size)
print(g)
```



Conclusion

The previous exercises had proven useful to demonstrate the following points:

- For any given population, no matter the distribution, the sample mean and sample variance are good estimators of the population mean and standard deviation.
- These estimations will become better as the sample sizes (and/or the number of simulations) get larger
- The sample mean, if n is large enough, can be approximated to a normal distribution centered at the population mean, with a standard deviation equal to $\sigma/\sqrt{n}$.

Appendix

Code used for simulations

This is the code used for the 1000 simulations:

```
lambda=0.2
nosim <- 1000
n <- 40
Xbardat <- data.frame(
  x = c(apply(matrix(rexp(nosim*n,lambda),
                     nosim), 1, mean)
  ),
  size = factor(rep(c(n), rep(nosim, 1))))
head(Xbardat)
```

This is the code used for the plot:

```
library(ggplot2)
g <- ggplot(Xbardat, aes(x = x, fill = size)) + geom_histogram(alpha = .20, binwidth=.3, colour = "bla
g <- g + geom_density(alpha=.2, fill="#FF6666")
g <- g + geom_vline(aes(xintercept=muexp),
                    color="red", linetype="dashed", size=1)
g <- g + ggtitle("Sample Mean - Density Chart")
g <- g + labs(x = "Sample Mean", y = "Density")
g + facet_grid(. ~ size)
```

Sample Variance versus Theoretical Variance

In this section the discussion is no longer around the sample mean, but the sample variance, which is in itself a random variable that has a population mean and a population variance. According to theory, the sample variance is also a consistent estimator of the population variance. If this principle holds, then the mean of the sample variance needs to be approximately equal to the population variance $\sigma^2 = (1/\lambda)^2 = (1/0.2)^2 = 25$.

In order to analyze the properties of the sample variance, the first step is to produce a set of 1000 sample variances into the dataframe **Xvardat**:

```
lambda=0.2
nosim <- 1000
n <- 40
Xvardat <- data.frame(
  x = c(apply(matrix(rexp(nosim*n,lambda),
                     nosim), 1, var)
  ),
  size = factor(rep(c(n), rep(nosim, 1))))
```

Secondly, a calculation of the mean of the sample variance is compared to the population variance:

```
XvarMean = mean(Xvardat$x)
```

- Theoretical Population Variance : 25

- Mean of Sample Variance : 25.118

Again, the histogram of the sample variances show that these tend to center to the population variance:

```
g <- ggplot(Xvardat, aes(x = x, fill = size)) + geom_histogram(alpha = .20, binwidth=1, colour = "bla
g <- g + geom_density(alpha=.2, fill="#FF6666")
g <- g + geom_vline(aes(xintercept=sdexp^2),
                    color="red", linetype="dashed", size=1)
g <- g + ggtitle("Sample Variance - Density Chart")
g <- g + labs(x = "Sample Variance", y = "Density")
g + facet_grid(. ~ size)
```