Statistical Inference Class Project - Exponential Distribution and the Central Limit Theorem

cafkhami

Thursday, April 09, 2015

Overview

This report invetigates the exponential distribution in R and compares it with the Central Limit Theorem (CLT). For our purposes, the CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases. We will compare the theoretical exponential distribution to the mean of 40 random exponential samples taken 1000 times. Specifically exploring the resulting mean, variance and type of distribution.

Simulations

Let's first create the distribution of the mean of 40 random exponential samples taken 1000 times (with lambda =0.2):

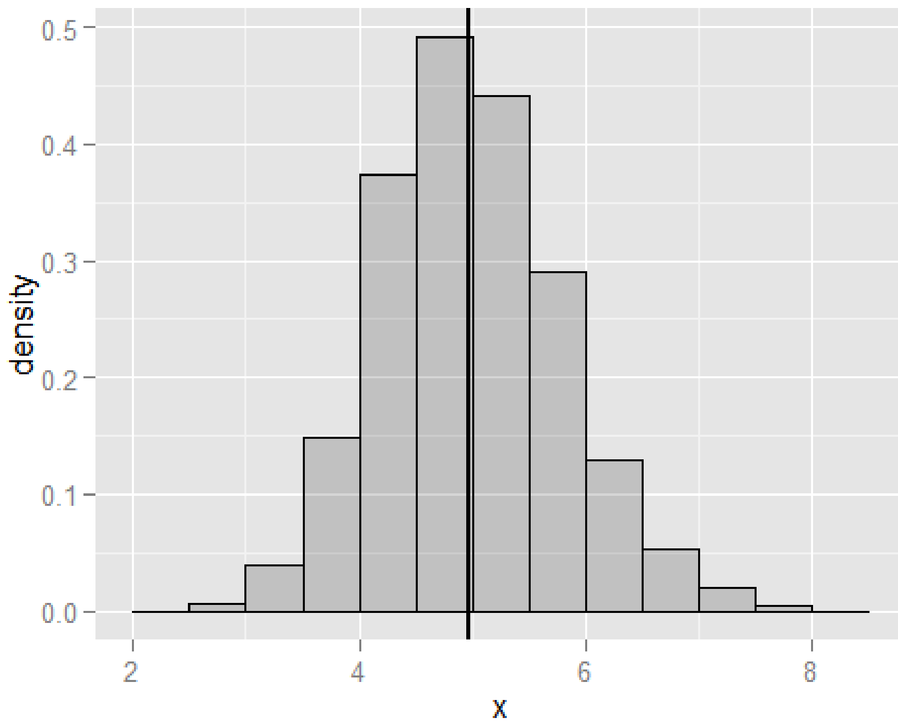
```
nosim = 1000
lambda = 0.2
mns <- NULL
mns.norm <- NULL

for (i in 1:nosim){
  mns = c(mns,mean(rexp(40,lambda)))
  mns.norm = c(mns.norm,sqrt(40)*(mean(rexp(40,lambda))-1/lambda)/(1/lambda))
}

ex <- rexp(nosim,lambda)

dat.ex <- data.frame(x = c(mns,ex), x.norm = c(mns.norm,ex), size =
factor(rep(c(40,1000), rep(nosim, 2))))
dat <- data.frame(x = mns, x.norm = mns.norm)
```

Sample Mean Vs Theoretical Mean

Let's evaluate the mean of the two pieces of data that we simulated above and plot the distributions along with their means.

```
mns.mn = mean(mns)

library(ggplot2)
g <- ggplot(dat, aes(x = x)) +
```

```
  geom_histogram(alpha = .20, binwidth=.5, colour = "black", aes(y =
..density..))
g +geom_vline(xintercept = mns.mn, size = 1)
```



The graph above show that, the mean for 1000 means of 40 samples is about 5. For an exponential distibution with an infinite number of smaples the theoretical mean should be `1/lambda` lets evaluate this theoretical mean and the calculated mean of our simulation:

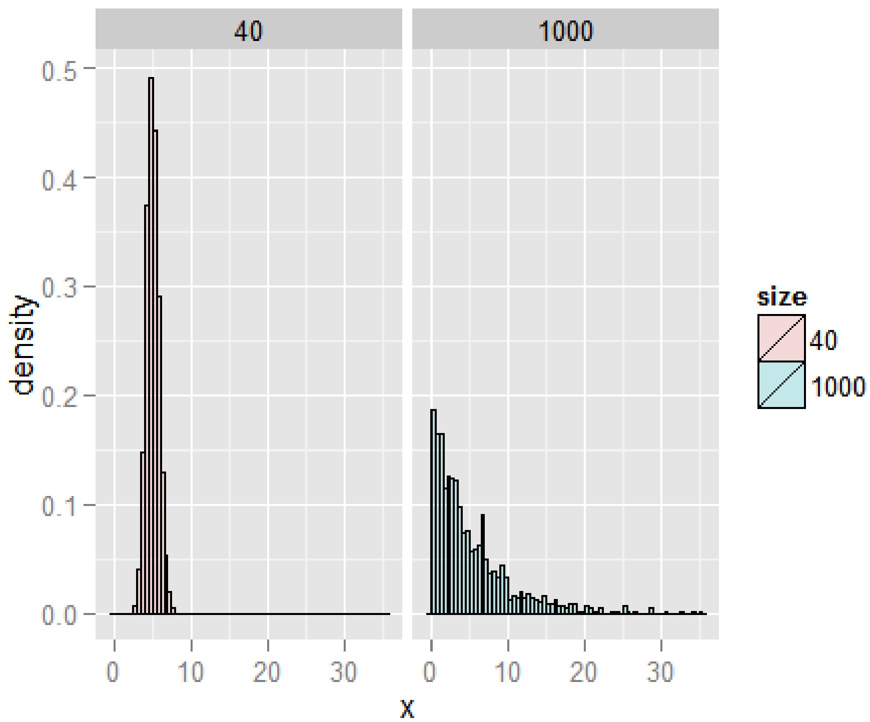
```
#Theoretical mean = 1/lambda
th.mn = 1/lambda
th.mn #mean - theoretical

## [1] 5

mns.mn #mean - 1000 means of 40 random samples

## [1] 4.978324
```

The means above are both very similar, proving that as the CLT states, `the mean of the mean of smaller samples of a distribution is equivalent to its theoretical mean.`

Sample Variance vs Theoretical Variance

Let's evaluate the variance of the distribution of 1000 means of 40 samples and the theoretical variance.

```r
g.ex <- ggplot(dat.ex, aes(x = x, fill = size)) +
  geom_histogram(alpha = .20, binwidth=.5, colour = "black", aes(y =
..density..))
g.ex + facet_grid(. ~ size)
```



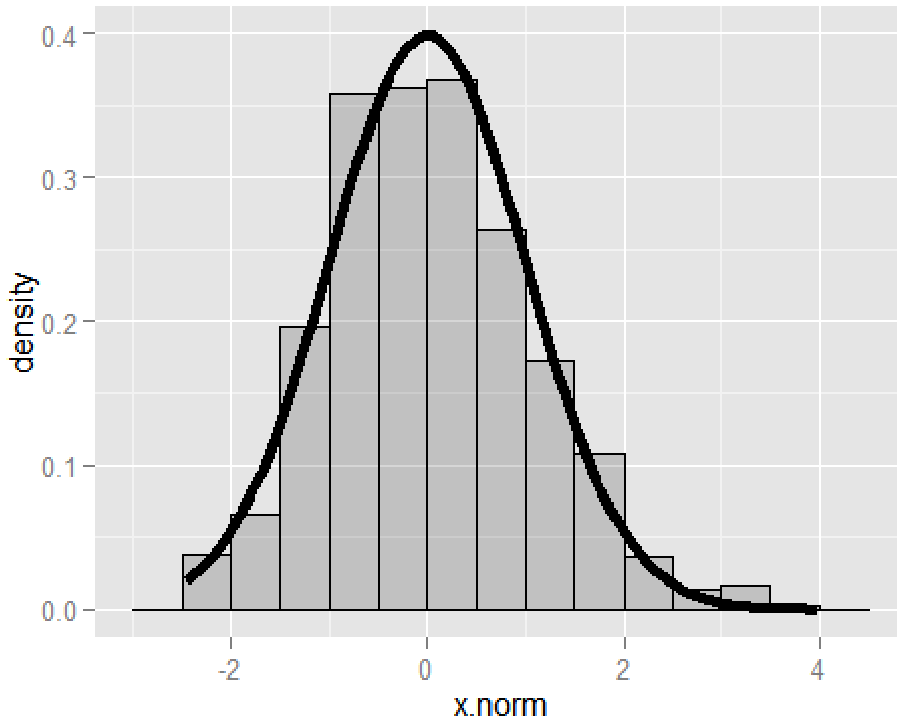

```r
#Theoretical sd = 1/lambda
th.sd = 1/lambda
th.var = th.sd^2
th.var #Variance - theoretical

## [1] 25

mns.sd <- sd(mns)
mns.var <- mns.sd^2
mns.var #Variance - 1000 means of 40 random samples

## [1] 0.6394373
```

Because we are taking the mean of 40 samples 1000 times, we find that the variance of this distribution is much smaller than the variance of the original distribution. It can be stated that 1000 means of 40 samples is a tight distribution, whereas the regular exponential distrubition is not as tightly distributed around it's mean.

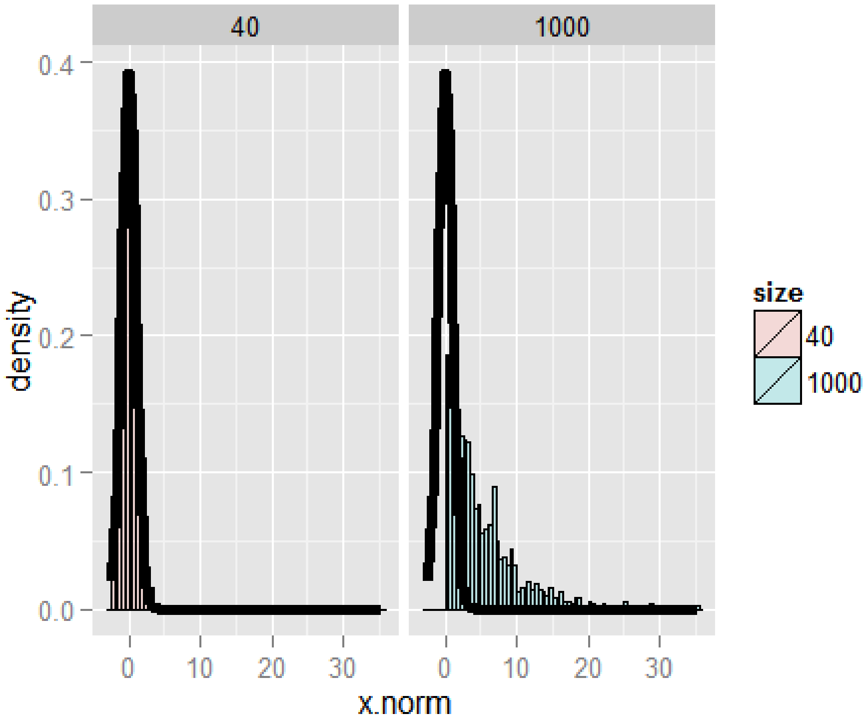
Distribution

As noted above, the variance of 1000 means of 40 samples is much smaller than the variance of the original distribution that the means were taken from. In fact, the CLT states that the distribution of the means of smaller samples will result in a normal distribution. Let's take a look at this graphically.

```
d <- ggplot(dat, aes(x = x.norm)) +
  geom_histogram(alpha = .20, binwidth=.5, colour = "black", aes(y =
..density..))
d + stat_function(fun = dnorm,size = 2)
```




```
d.ex <- ggplot(dat.ex, aes(x = x.norm, fill = size)) +
  geom_histogram(alpha = .20, binwidth=.5, colour = "black", aes(y =
..density..))
d.ex + facet_grid(. ~ size) + stat_function(fun = dnorm,size = 2)
```




```
mean(dat$x.norm)

## [1] 0.03390311

sd(dat$x.norm)

## [1] 1.037071
```

In the first plot above the data above was normalized to have mean of 0 and standarad deviation of 1. When we do this and then overlay the histogram with the standard normal distribution density curve, it is noted that the distributions are very similar. Thus, you can conclude that the distribution of 1000 means of 40 samples from an exponential distribution results in a noraml distribution, as was stated by the CLT.

Compare this to the original distribution (seen in the following second figure above), it is obvious that the exponential distribution does not follow the standard normal distribution.