Comparison of exponential distribution with Central Limit Theorem

Ricardo Leon

February 23, 2020

Overview

Central Limit Theorem is a very useful tool for statistical inference. Through exponential distribution sampling, it is going to be exposed empirical proofs of it effectiveness. 1000 samples with 40 observations each will be used to this purpose.

Simulations

The first step is to set initial values and calculate theorical mean and standard deviation.

```
lambda = 0.2
distribution_mean = 1 / lambda
distribution_variance = 1 / (lambda * lambda)
```

Next, using `rexp` R function, we will generate two dataframes containing the mean and standard deviation for each sample.

```
mns = NULL
variances = NULL
for (i in 1 : 1000) {
    sample <- rexp(40, lambda)
    mns <- c(mns, mean(sample))
    variances <- c(variances, var(sample))
}
means_data <- data.frame(means=mns)
vars_data <- data.frame(vars=variances)
```

Once we get the results, calculate the estimated population mean and standard deviation:

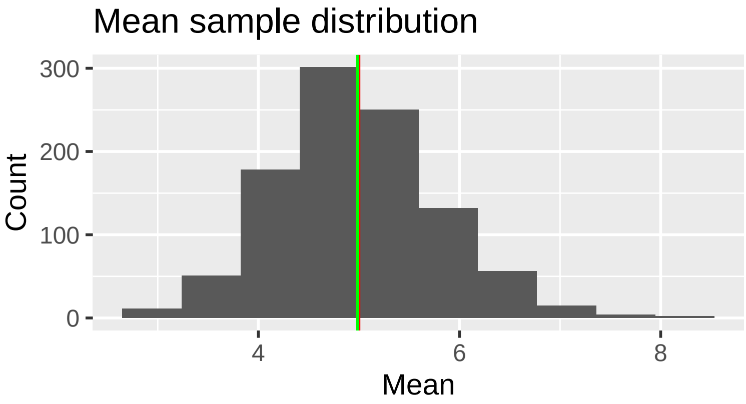
```
estimated_population_mean <- mean(mns)
estimated_population_variance <- mean(variances)
```

Then, estimated population mean is 4.9865083 and estimated variance 25.165914.

Sample Mean versus Theoretical Mean

In the plot below, we can see the sampling distribution of the sample means. The red vertical line represents the theoretical distribution mean (5) and the green one the estimated sample mean (4.9865083).

```
ggplot(data=means_data, aes(x = means)) +
    geom_histogram() +
    stat_bin(bins = 10) +
    xlab('Mean') +
    ylab('Count') +
    ggtitle('Mean sample distribution') +
    geom_vline(xintercept = distribution_mean, color = 'red') +
    geom_vline(xintercept = estimated_population_mean, color = 'green')
```



```
mean_diff <- 100 * (distribution_mean - estimated_population_mean) / distribution_mean
```

There is a difference of 0.2698337% so we can conclude the Central Limit Theorem provides a good assertion for the estimation of the population mean.

Sample Variance versus Theoretical Variance

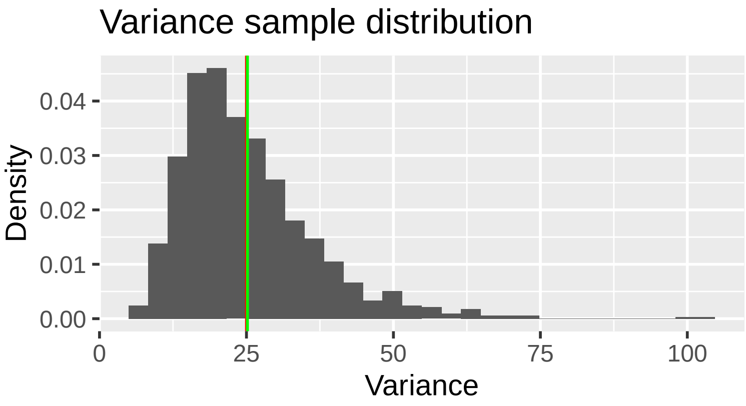
Applying the same concept from above, we have a sample variace of 25.165914 and theoritical variance of 25.

```
variance_diff <- 100 * (distribution_variance - estimated_population_variance) / distribution_variance
```

The difference between the two variables is -0.6636558%. We reach the same conclusion for this statistic.

Graphic representation is provided below.

```
ggplot(data=vars_data, aes(x = vars)) +
    geom_histogram(aes(y = ..density..)) +
    xlab('Variance') +
    ylab('Density') +
    ggtitle('Variance sample distribution') +
    geom_vline(xintercept = distribution_variance, color = 'red') +
    geom_vline(xintercept = estimated_population_variance, color = 'green')
```

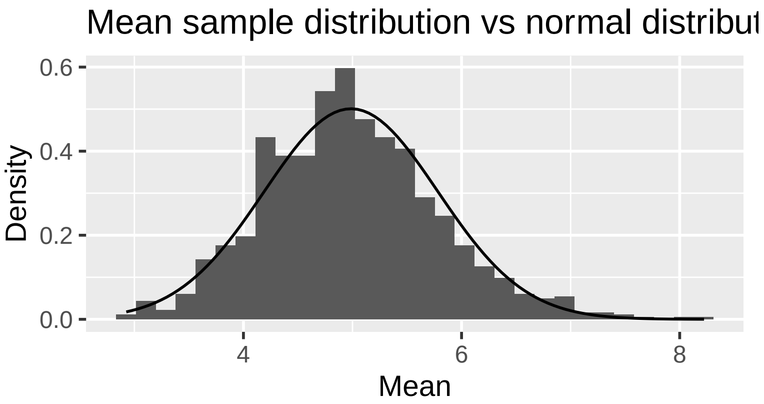


Distribution

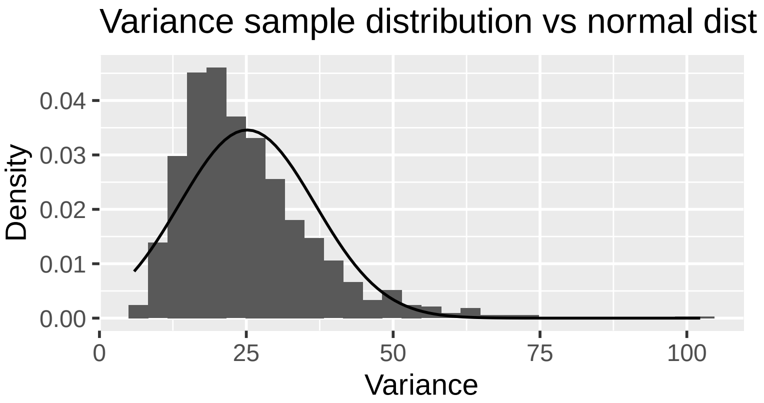
Now we will compare both distributions with the normal distribution.

As we can see in the following plots, mean distribution is approximately normal. The same happens with the variance distribution but in this case, this distribution is right skewed. In both cases, for larger samples they will be getting closer to the "bell shape".

```r
ggplot(data=means_data, aes(x = means)) +
    geom_histogram(aes(y = ..density..)) +
    xlab('Mean') +
    ylab('Density') +
    ggtitle('Mean sample distribution vs normal distribution') +
    stat_function(fun = dnorm, args = list(mean = mean(mns), sd = sd(mns)))
```



```r
ggplot(data=vars_data, aes(x = vars)) +
    geom_histogram(aes(y = ..density..)) +
    xlab('Variance') +
    ylab('Density') +
    ggtitle('Variance sample distribution vs normal distribution') +
    stat_function(fun = dnorm, args = list(mean = mean(variances), sd = sd(variances)))
```


Variance sample distribution vs normal dist
Density
Variance