The CLT explained via exponential distribution

Americo Costantini

19 luglio 2015

Overview

In this report we are going to investigate the so called exponential distribution, in order to explain properties of central limit theorem (CLT) and law of large numbers (LLN). Our focus will be on the distribution of 1000 averages of samples composed by 40 observations, in order to show its normality and the proximity of its statistics (mean, variance) to the population parameters.

Simulation

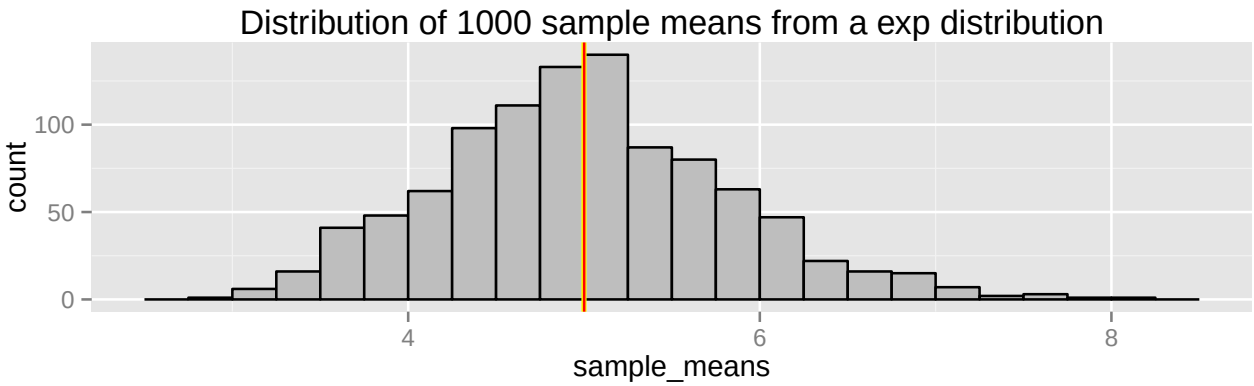
First of all, we will ask R to generate 1000 samples of 40 observation, with the rate parameter (lambda) = 0.2. Remember that the exponential distribution has mean = 1/lambda and standard deviation = 1/lambda. I will create also a sample of 100000 observations, that will be of help at the final chapter of the report.

```
set.seed(1912)
# the 1000 samples
exp_sam <- matrix(nrow = 1000, ncol = 40)
for (i in 1:1000) {
        exp_sam[i,] <- rexp(n = 40, rate = 0.2)
}
#the 100000 observations sample
exp_values <- rexp(n = 100000, rate = 0.2)
exp_dist <- data.frame(exp_values, density = dexp(exp_values))
```

Sample Mean versus Theoretical Mean

Now we calculate the mean for each of the 1000 samples, generating a distribution of sample means. We know, by CLT, that the probability distribution of the `random variable sample mean` is approximately normal, as size increases. 40 is a good size to show this, while 1000 is a good number of simulation to represent by frequency what the theorem states for probability. Please pay attention to the difference between `sample_means`, the 1000 means vector, and `samples_mean`, the mean of the 1000 `sample_means`, which must be compared to the exptected value of the exponential distribution.

```
sample_means <- apply(exp_sam, 1, mean)
samples_mean <- mean(sample_means)
theoretical_mean <- 1/0.2
ggplot(data.frame(x = sample_means), aes(x = sample_means)) +
        geom_histogram(binwidth = 0.25, colour="black", fill = "gray") +
        geom_vline(xintercept = c(theoretical_mean, samples_mean),  colour = c("yellow", "red"), size =
        ggtitle(label = "Distribution of 1000 sample means from a exp distribution")
```



You see that the mean of the samples means is 5.001, the red vertical line, while the expected value of the exponential distribution with lambda = 0.2 is 5, and it's the yellow line. The histogram is approximately normal, so:

- the distribution of the 1000 sample means is approximately normal, as the CLT states;
- the distribution is approximately centered around the expected value 5;
- this approximation would be even better if the size of the samples was bigger than 40.

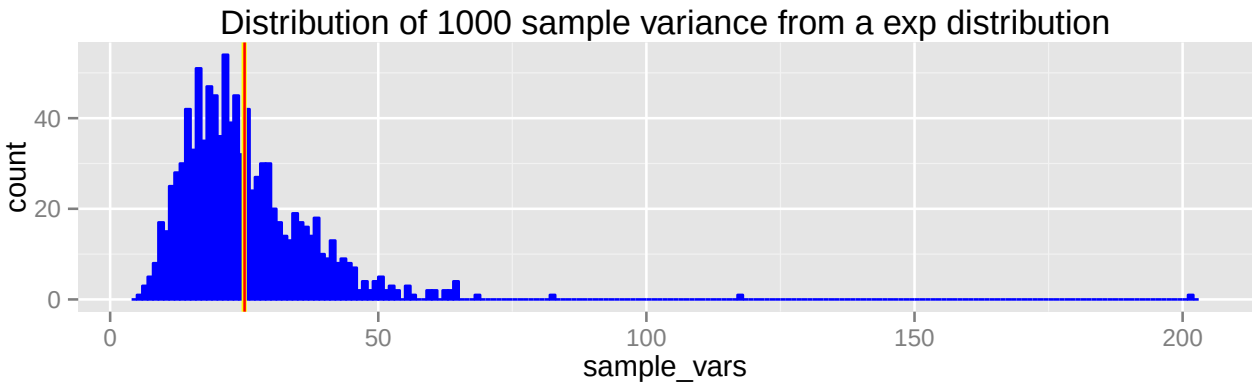
Sample Variance versus Theoretical Variance

What we know from theory is that the variance of the random variable `sample mean` is $Var(X)/n$. So we expect that the variance of the 1000 means will be approximately equal to the variance of the exponential distribution divided by the sample size, 40.

```
sample_means_var <- var(sample_means)
theoretical_mean_var <- ((1/0.2)^2)/40
```

The variance of the 1000 sample means, that in a frequentist approach can be interpreted as the variance of the random variable `sample mean` is 0.671 while the theoretical one is 0.625. That's what we expected! If we are interested in the sample variance, where size of the sample is 40, we can create a distribution of variances calculated for each of the 1000 sample and compare its mean to the population variance.

```
sample_vars <- apply(exp_sam, 1, var)
samples_var <- mean(sample_vars)
theoretical_var <- (1/0.2)^2
ggplot(data.frame(x = sample_vars), aes(x = sample_vars)) +
        geom_histogram(binwidth = 1, colour="blue", fill = "blue") +
        geom_vline(xintercept = c(theoretical_var, samples_var),  colour = c("yellow", "red"), size = c
        ggtitle(label = "Distribution of 1000 sample variance from a exp distribution")
```

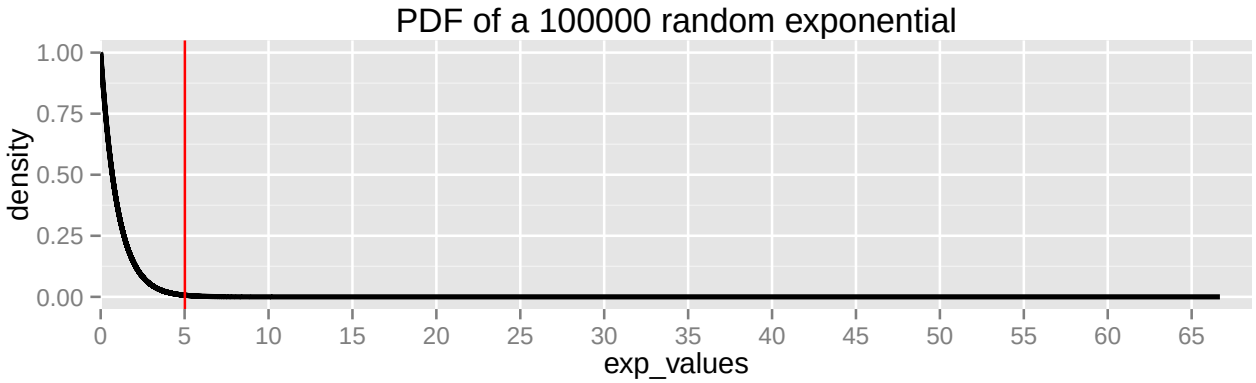


The mean of the 1000 variances is 25.07 (the yellow line) while the population parameter is 25 (the red line). So it is confirmed that sample variance is an unbiased estimator of the variance, while the histogram shows that CLT doesn't apply to sample variance (its distribution is not normal, under some assumptions is chi-squared). *Please consider that the function* `var()` *compute the sample variance, so every thing should be fine in the code.*

Distribution

The distribution of the sample means, as we have seen before, is approximately normal. This is thanks to the CLT, even if the underlying population is not normal. Let'see in depth.

```
ggplot(data = exp_dist, aes(x = exp_values, y = density)) +
        geom_line(size = 1) +
        ggtitle(label = "PDF of a 100000 random exponential") +
        scale_x_discrete(breaks = seq(0, max(exp_values), 5)) +
        geom_vline (xintercept = mean(exp_dist$exp_values), colour = "red")
```




```
dist_mean <- mean(exp_dist$exp_values)
```

This is not of course a normal population; it is exponential! So: the theoretical mean is 5; the mean of a 100000 sample is 5.015, and it is supposed to be close to 5 for the LLN. The mean of 1000 means of samples of 40 observations is 5.001, and it is supposed to be close to 5 because the sample mean is an unbiased estimator, while its distribution it is supposed to be normal because of the CLT. **Finally, every aspect of the theory is confirmed by our simulations.**