Simulation of averages from exponential distribution

Jan Herman

October 25, 2015

Overview

In this short document, we focus on showing the validity of Central limit theorem in case of drawing iid samples from exponential distribution with parameter $\lambda = 0.2$. In particular, we show that a distribution of averages of $n = 40$ iid random variables drawn from this exponential distribution is approximately normal $N(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$.

Theory

The exponential distribution

The exponential distribution $\text{Exp}(\lambda)$ is defined by its probability density function

$$f(x;\lambda) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

or equivalently by its cumulative distribution function

$$F(x;\lambda) = \begin{cases} 1 - \lambda \mathrm{e}^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

We mention its mean and variance without proof. The mean is given by $\mathrm{E}[X] = \frac{1}{\lambda}$ and the variance by $\mathrm{Var}[X] = \frac{1}{\lambda^2}$.

The central limit theorem

We won't go into much detail here. Rougly speaking, the Central limit theorem says, that the distribution of averages of n iid random variables taken from distribution with mean μ and variance $\sigma^2 < \infty$ can be approximated by $N(\mu, \frac{\sigma^2}{n})$ for n large enough.

More formally, as $n \to \infty$, the $\sqrt{n}(S_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, where $S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $X_1, X_2, \ldots, X_n$ are iid random variables from distribution with mean μ and variance $\sigma^2 < \infty$.

Expected results

If we join both theory paragraphs, we can deduce that for n large enough, the distribution of averages of n iid random variables from $\text{Exp}(\lambda)$ should be approximately $(N(\frac{1}{\lambda}, \frac{1}{n\lambda^2}))$.

In particular, if we set $\lambda = 0.2$ and $n = 40$, we expect that the resulted distribution of averages can be approximated by $N(5, 0.625)$.

Simulation in R

We will simulate 1000 cases of taking average of 40 iid variables from Exp(0.2).

At first we set the parameters and the random seed (to make the simulation reproducible).

```
sim_count <- 1000
lambda = 0.2
n <- 40
set.seed(2)
```

Then we sample 40,000 values from Exp(0.2) and organize them into matrix with 1000 rows and 40 columns. Each row represents one simulation.

We compute row means – these are our averages whose distribution we want to examine.

```
samples <- matrix(rexp(sim_count * n, lambda),
                  nrow = sim_count,
                  ncol = n)
sample_means <- rowMeans(samples)
```

At first we compute the two basic statistics – sample mean and sample variance – and compare them to the predicted theoretical values $mu_{\mathrm{T}} = \frac{1}{\lambda} = 5$ and $\sigma^2_{\mathrm{T}} = \frac{1}{n\lambda^2} = \frac{1}{40 \cdot 0.2^2} = 0.625$.

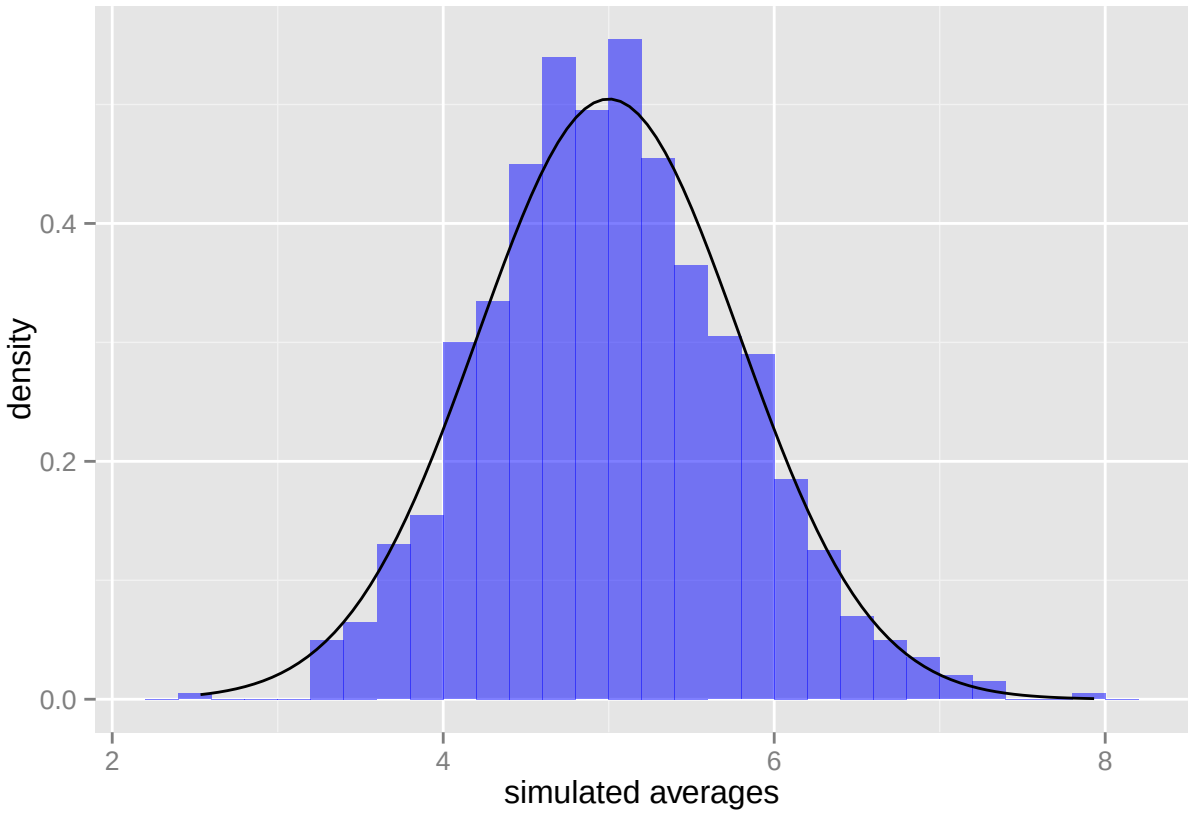
```
overall_mean <- mean(sample_means)
overall_var <- var(sample_means)
c(overall_mean, overall_var)
```

```
## [1] 5.0163562 0.5776132
```

We see, that the mean is pretty close ($|5 - 5.016| = 0.016$), while the variance is a little off ($|0.625 - 0.578| = 0.047$); both rounded to 3 decimal places. We will briefly discuss it later on.

And let the last evidence be the graphics – the histogram of simulated averages together with the density of normal distribution with previously computed parameters (black line).

```
library(ggplot2)
g <- ggplot() +
    aes(sample_means) +
    geom_histogram(alpha = .5, binwidth = .2, aes(y = ..density..), fill = "blue") +
    labs(x = "simulated averages") +
    stat_function(fun = dnorm, args = list(mean = 1/lambda, sd = 1/(sqrt(n)*lambda)))
g
```



We see that the histogram quite corresponds to the normal distribution. A little more discussion takes place in dedicated section.

Discussion

The mean of simulated averages is almost exactly the mean of discussed normal distribution $N(5, 0.625)$, but this is no surprise. When computing the sample mean of averages of iid variables from a distribution with mean 5, we cannot get nothing else that the mean of this distribution (and with the Law of large numbers in our backpack we are pretty sure that 40,000 samples – 40 for each of 1000 simulations – is well enough).

Regarding variance, the story is a little more complicated. The CLT states only that the distribution of averages should converge to the normal distribution as n tends to infinity. But is $n = 40$ large enough for the variance? The personal opinion of the author (as he is not very experienced in statistics) is that for most practical purposes the approximation with discussed normal distribution is suitable.

There remain some open questions, since the mean and the variance do not describe the distribution in any case fully. As far as author can say – from the included plot is apparent that the approximation with normal distribution is pretty accurate (in the sense of the density curve). But the proper analysis of this statement is beyond the scope of this little document (and beyond the ability of the author, too).

Notice

The knitr source code of this analysis is available at github.