Central Limit Theorem Study

Statistical Inference Coursera Project - Part 1

Magdalena Paszko

This paper is a part of the final project in the Coursera's Statistical Inference course (a part of Data Science specialization by Johns Hopkins University) in which we investigate the Cetral Limit Theorem. By simulating repeated sample draws from exponential distribution we show that sampling means tend toward a normal distribution. The complete R code is available in the appendix.

Introduction

Central Limit Theorem states that statistics calculated on samples of iid random variables tend to follow normal distribution as the sample size gets larger, independently of the distribution of the random variable itself. Specifically, when a random variable X comes from a distribution with mean μ and variance σ^2, the mean $\bar{X}$ of sample of size n tend to follow normal distribution with mean μ and variance σ^2/n:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Here we show this empirically using random draws from **exponential distribution**.

Simple sample from exponentil distribution

First let's look at simple draws from exponential distribution with $\lambda = 0.2$.

In Figure 1 we plot a histogram of 1000 such draws versus real exponential density curve (the orange shape). The empirical distribution of randow draws seems to quite closely fit the theoretical curve.

Exponential distribution with $\lambda = 0.2$ has mean $\mu = 1/\lambda = 5$ and variance $\sigma^2 = 1/\lambda^2 = 25$. The sample estimations of these two values are quite acurate: 5.18 for the mean and 25.83 for the variance. The sample and theoretical mean have been additionally plotted as a gray and an orange vertical line, respectively.

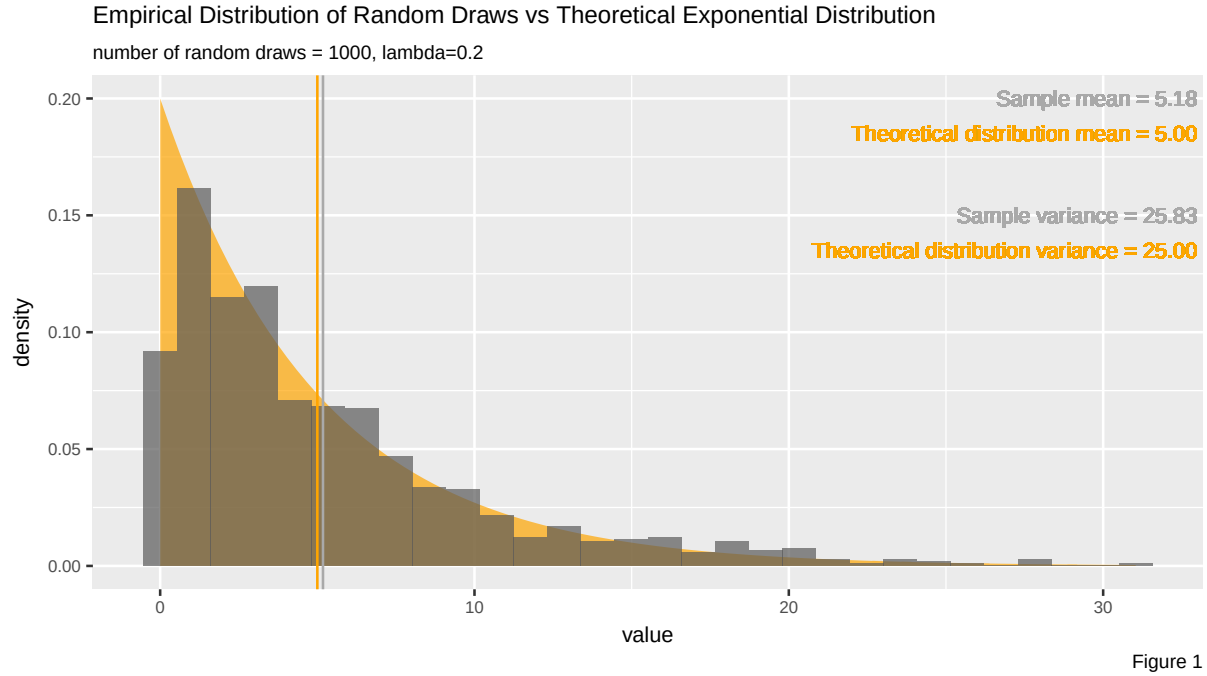

Figure 1

Repeated samples from exponential distribution

Now we simulate drawing 1000 samples of size $n = 40$ from the same exponential distribution with $\lambda = 0.2$. For each sample we calculate the mean and we plot the empirical distribution of those means (Figure 2).

According to the Central Limit Theorem, the sample mean should tend to follow normal distribution with mean $\mu = 1/\lambda = 5$ and variance $\sigma^2/n = 1/\lambda^2 n = 0.625$.

In Figure 2 we plot this theoretical normal distribution (orange shape). The empirical histogram fits the normal curve quite closely. Also the sample mean (5.02) is very close to the theoretical mean (gray and orange vertical lines in the plot respectively). Accordingly, the empirical variance (0.578) approaches the theoretical one. The difference might come from the fact that the empirical values are always non-negative (due to the properties of exponential distribution) and thus the variance is smaller than that of a normal distribution which includes negative values (although they account for a very small density mass).

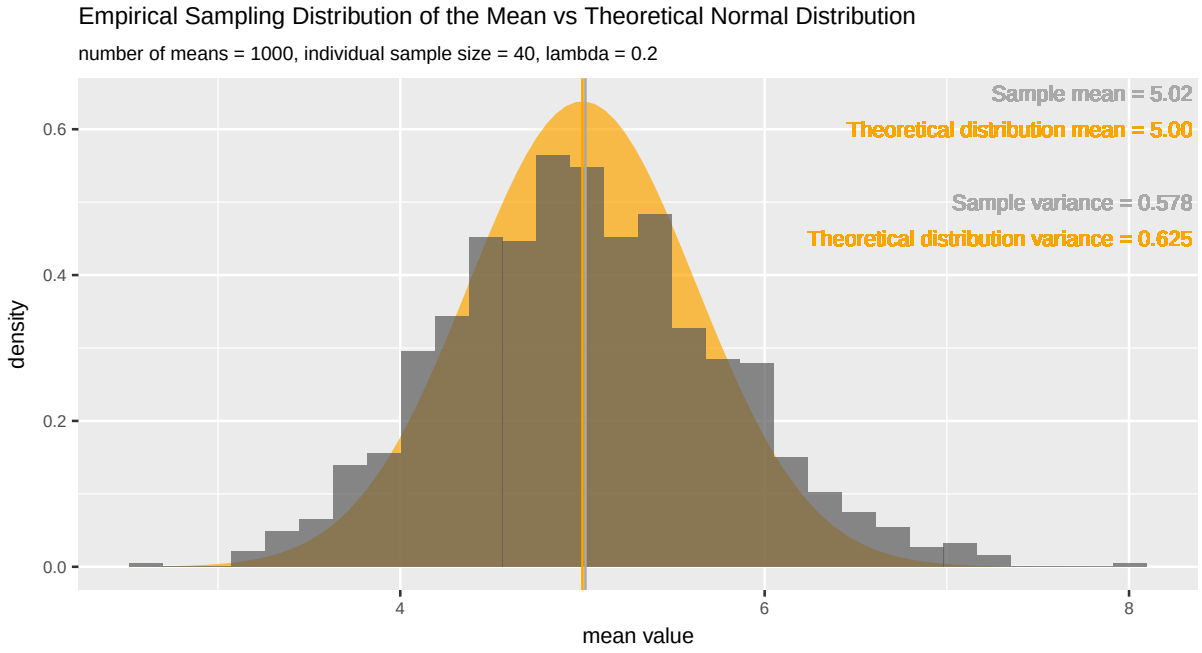

Figure 2

Conclusion

We conclude that a simple exercise with a relatively small sample size (40) illustrates quite accurately the Central Limit Theorem. The empirical distribution of means from exponential samples does follow a bell-curve shape and fits quite closely in its theoretical normal distribution given by the CLT.

Appendix - R code

Preliminary preparations:

```
library(ggplot2)
lambda <- 0.2
n <- 40
B <- 1000
```

Simple sample generation. We draw B=1000 random exponentials with lambda=0.2:

```
set.seed(2)
sample_x <- rexp(B, lambda)
```

Plotting Figure 1:

```
#prepare annotations
sample_mean <- paste("Sample mean =", round(mean(sample_x),2))
dist_mean <- paste("Theoretical distribution mean =", format(1/lambda, nsmall=2))
sample_var <- paste("Sample variance =", round(var(sample_x),2))
dist_var <- paste("Theoretical distribution variance =", format(1/lambda^2, nsmall=2))

ggplot(data.frame(sample_x), aes(sample_x)) +
  stat_function(fun=dexp, args=c(lambda), geom="area",fill="orange", alpha=0.7) +
  geom_histogram(aes(y = ..density..), alpha=0.7) +
  theme(text = element_text(size=9),
        plot.title = element_text(size = 10)) +
  geom_text(label=sample_mean, x=33, y=0.2, hjust=1, color="darkgray", size=3) +
  geom_text(label=sample_var, x=33, y=0.150, hjust=1, color="darkgray", size=3) +
  geom_text(label=dist_mean, x=33, y=0.185, color="orange", hjust=1, size=3) +
  geom_text(label=dist_var, x=33, y=0.135, color="orange", hjust=1, size=3) +
  geom_vline(xintercept = mean(sample_x), color="darkgrey") +
  geom_vline(xintercept = 1/lambda, color="orange") +
  labs(title="Empirical Distribution of Random Draws vs Theoretical Exponential Distribution",
       subtitle="number of random draws = 1000, lambda=0.2",
       x="value",
       y="density",
       caption="Figure 1")
```

Repeated samples generation. We draw B=1000 samples of size n=40 from exponential distribution with lambda=0.2

```
set.seed(2)
x <- matrix(rexp(n*B, lambda), B, n)
x_bar <- apply(x, 1, mean)
```

Plotting Figure 2:

```
#prepare annotations
sample_mean2 <- paste("Sample mean =", round(mean(x_bar),2))
dist_mean2 <- paste("Theoretical distribution mean =", format(1/lambda, nsmall=2))
sample_var2 <- paste("Sample variance =", round(var(x_bar),3))
dist_var2 <- paste("Theoretical distribution variance =", format(1/lambda^2/n, nsmall=3))

ggplot(data.frame(x_bar), aes(x_bar)) +
  stat_function(fun=dnorm, args=c(1/lambda, 1/lambda^2/n), geom="area",fill="orange", alpha=0.7) +
  geom_histogram(aes(y = ..density..), alpha=0.7) +
  geom_text(label=sample_mean2, x=8.35, y=0.65, hjust=1, color="darkgray", size=3) +
  geom_text(label=sample_var2, x=8.35, y=0.5, hjust=1, color="darkgray", size=3) +
```

```
  geom_text(label=dist_mean2, x=8.35, y=0.6, color="orange", hjust=1, size=3) +
  geom_text(label=dist_var2, x=8.35, y=0.45, color="orange", hjust=1, size=3) +
  geom_vline(xintercept = mean(x_bar), color="darkgrey") +
  geom_vline(xintercept = 1/lambda, color="orange") +
  labs(title="Empirical Sampling Distribution of the Mean vs Theoretical Normal Distribution",
       subtitle="number of means = 1000, individual sample size = 40, lambda = 0.2",
       x="mean value",
       y="density",
       caption="Figure 2") +
  theme(text = element_text(size=9),
        plot.title = element_text(size = 10))
```