Using Simulation to Explore the Central Limit Theorem

Connor Claypool

18 June 2018

Introduction

One of the most important statistical principles for data science, being key for inferential techniques such as calculating confidence intervals and testing hypotheses, is The Central Limit Theorem (CLT). The CLT states that the means of n independent and identically distributed (IID) random variables follow an approximately normally distribution when n is large, regardless of the underlying distribution of the individual random variables for which the means are calculated. In other words, sample means are normally distributed even if individual observations are not, as long as the sample size is sufficiently large and observations are independent. To demonstrate this principle, we will simulate a distribution of sample means by generating 1000 samples of size 40 from a population defined by an exponential distribution, and compare the theoretical and empirical properties of this distribution.

Simulation

The first step is to simulate the random samples. We generate 40,000 random exponentials with rate 0.2, and arrange them as a matrix with 1000 rows of 40 observations. To calculate the sample means, the mean of each row is taken. Note that the random seed is set first to ensure reproducibility.

```
set.seed(3791)

rate = 0.2
n = 40
samples = 1000
simulated_data <- matrix(rexp(samples * n, rate), nrow = samples, ncol = n)
sample_means <- rowMeans(simulated_data)
```

Observed Mean vs Theoretical Mean

As sample means are an unbiased estimator of the population mean, the mean sample mean should closely approximate the population mean given a large enough number of samples. We can test how well this holds for our 1000 simulated samples by comparing the mean of the means of these samples with the known population mean, 1/rate.

```
theoretical_mean <- 1/rate
observed_mean <- mean(sample_means)

print(paste("Theoretical mean of sample means:", theoretical_mean))
```

```
## [1] "Theoretical mean of sample means: 5"
```

```
print(paste("Observed mean of sample means", observed_mean))
```

```
## [1] "Observed mean of sample means 5.00881012991511"
```

So the difference between the observed and theoretical means is only 0.0088101.

Observed Variance vs Theoretical Variance

The variance of the random variable defined as the mean of n IID random variables each with variance σ^2 is known to be σ^2/n. We can use this rule to calculate the theoretical variance of the sample means, and compare this to the observed variance.

```
theoretical_variance <- ((1/rate)^2)/n
observed_variance <- var(sample_means)

print(paste("Theoretical variance of sample means:", theoretical_variance))
```

```
## [1] "Theoretical variance of sample means: 0.625"
```

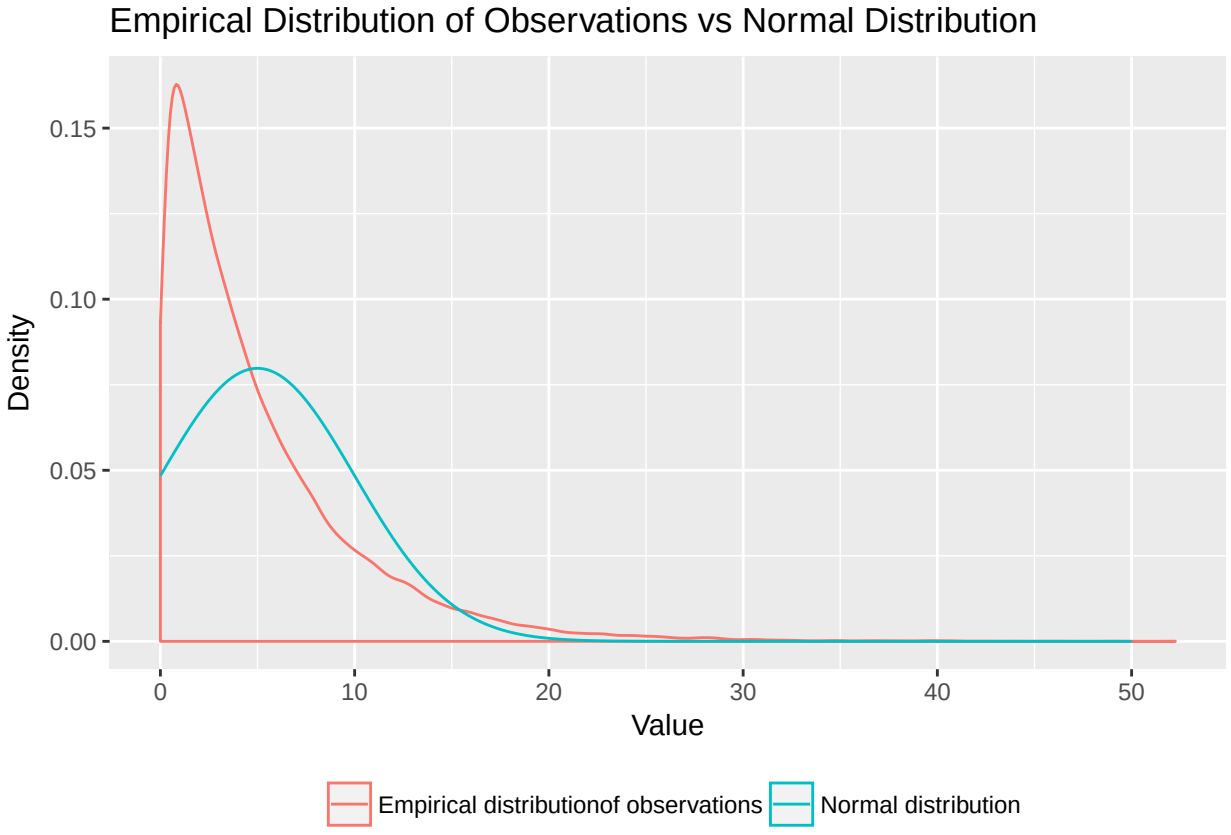
```
print(paste("Observed variance of sample means:", observed_variance))
```

```
## [1] "Observed variance of sample means: 0.649620250601754"
```

So the difference between the theoretical and observed values this time is 0.0246203.

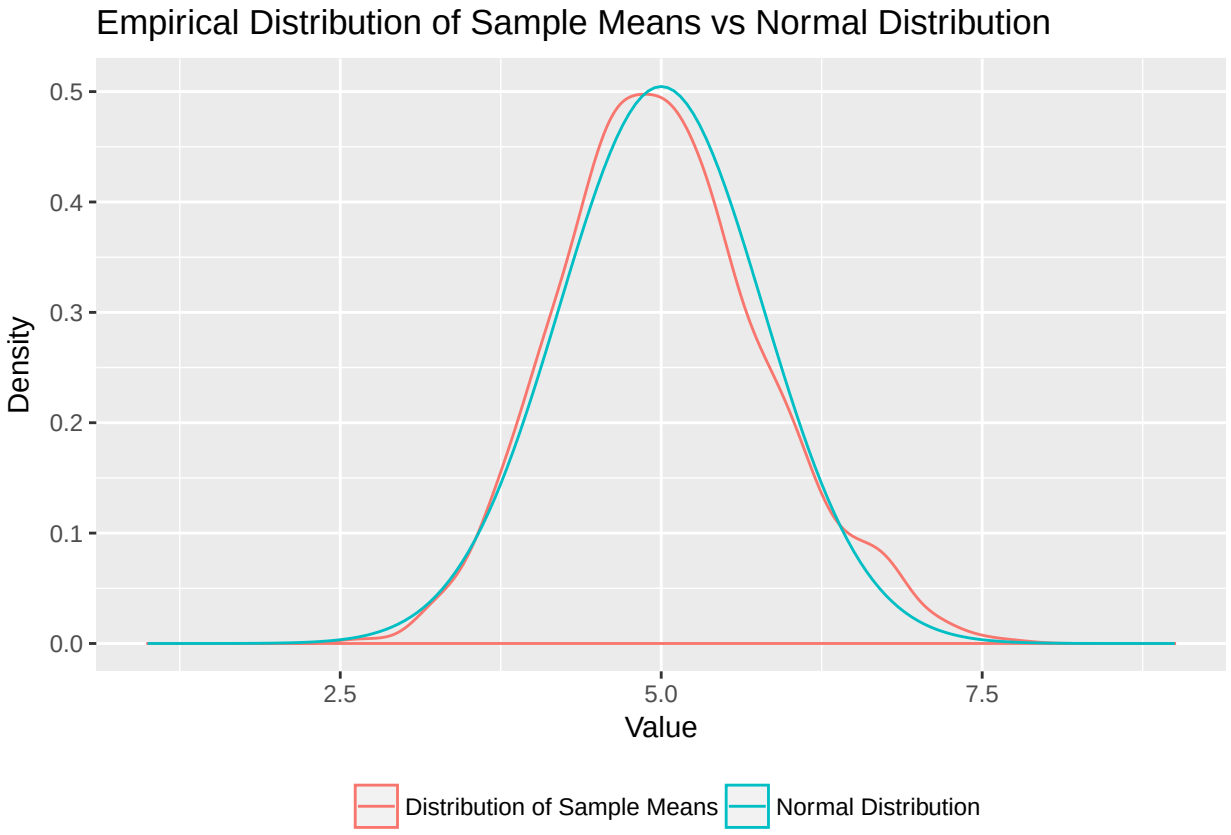
Underlying Distribution vs Sample Mean Distribution

Before examining the distribution of the sample means, we will plot the distribution of the 40,000 random exponentials, i.e. the exponential distribution, so that the difference between the two can be appreciated. The density plot below compares the underlying exponential distribution with a normal distribution with the same mean and variance.



Clearly, these distributions differ significantly. Now, we will compare the distribution of the means of samples of 40 exponential random variables with the normal distribution which has the theoretical mean and variance

calculated previously.



This plot shows that the distribution of the means of samples of exponentials, unlike the distribution of the exponentials themselves, is closely approximated by a normal "bell curve".

Appendix: Code for Reproducing the Plots

Empirical Distribution of Observations vs Normal Distribution

```
library(ggplot2)
normal_densities <- dnorm(seq(0, 50, 0.05), mean = 1/rate, sd = 1/rate)
ggplot() +
    geom_density(data = data.frame(rv = as.vector(simulated_data)),
                 aes(x = rv, color = "Empirical distributionof observations")) +
    geom_line(data = data.frame(q = seq(0, 50, 0.05), d = normal_densities),
              aes(x = q, y = d, color = "Normal distribution")) +
    theme(legend.position = "bottom", legend.title = element_blank()) +
    xlab("Value") +
    ylab("Density") +
    ggtitle("Empirical Distribution of Observations vs Normal Distribution")
```

Empirical Distribution of Sample Means vs Normal Distribution

```
normal_densities_sm <- dnorm(seq(1, 9, 0.05),
                          mean = theoretical_mean,
                          sd = sqrt(theoretical_variance))
ggplot() +
    geom_density(data = data.frame(sm = sample_means),
                 aes(x = sm, color = "Distribution of Sample Means")) +
    geom_line(data = data.frame(q = seq(1, 9, 0.05), d = normal_densities_sm),
              aes(x = q, y = d, color = "Normal Distribution")) +
    theme(legend.position = "bottom", legend.title = element_blank()) +
    xlab("Value") +
    ylab("Density") +
    ggtitle("Empirical Distribution of Sample Means vs Normal Distribution")
```