Using Simulation to Explore the Central Limit Theorem

Connor Claypool

18 June 2018

Introduction

One of the most important statistical principles, being key for fundamental inferential techniques such as calculating confidence intervals and testing hypotheses, is the Central Limit Theorem (CLT). The CLT states that the mean of n independent and identically distributed (IID) random variables follows a normal distribution when n is large enough, regardless of the underlying distribution of the individual random variables. 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 test this using simulation, we will generate 1000 samples of size 40 from a population defined by an exponential distribution, calculate the mean of each sample, and compare the theoretical properties of the distribution of sample means with those empirically observed.

Simulation

To simulate a distribution of sample means, we will generate 1000 samples of size 40 from the population defined as following an exponential distribution with rate 0.2, and then calculate the mean of each sample. To accomplish this in R, we populate a matrix of 1000 rows and 40 columns with 40,000 random exponentials each with rate 0.2, taking the mean of each row to calculate the sample means. Note that we first set the random seed to ensure the results are reproducible.

```
set.seed(3791)

samples <- 1000
n <- 40
rate <- 0.2

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

Observed Mean vs Theoretical Mean

In theory, 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 our sample means 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 around 0.0088. A small difference can be expected given the finite number of samples used in our simulation.

Observed Variance vs Theoretical Variance

The variance of 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 distribution of our 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"
```

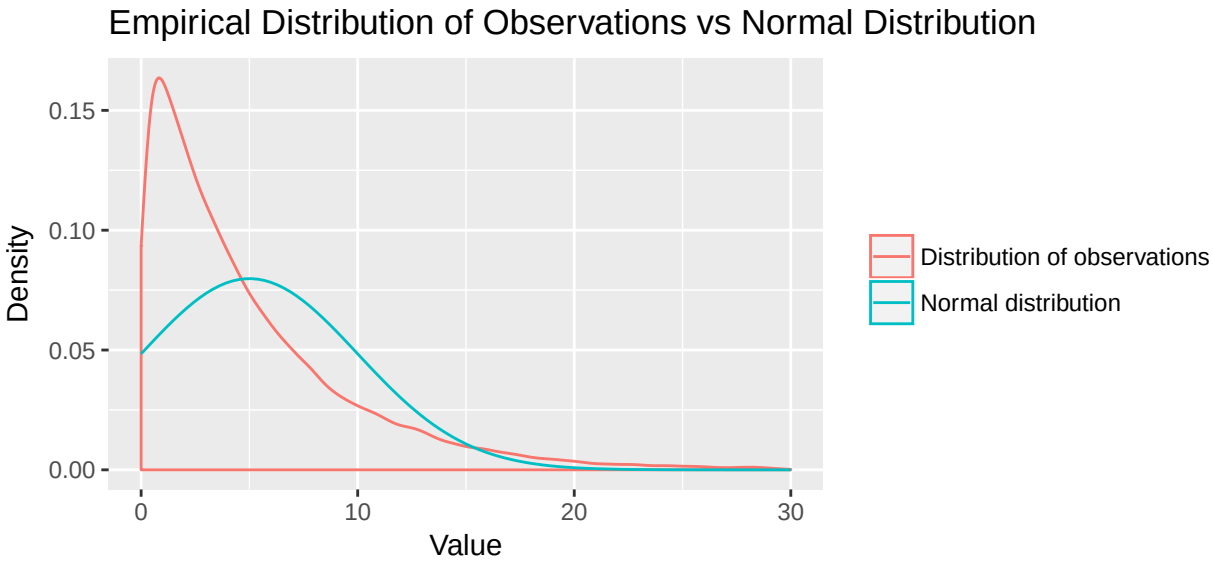
```
print(paste("Observed variance of sample means:", observed_variance))
```

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

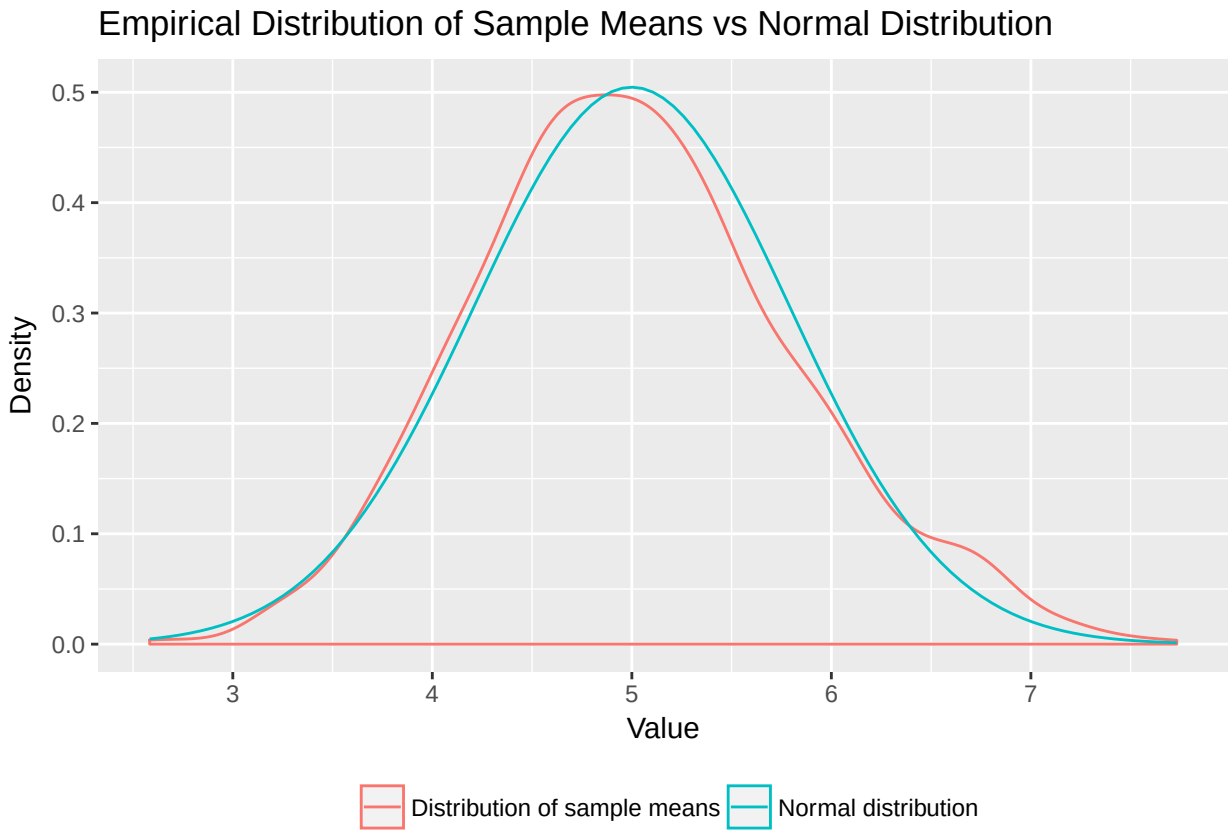
So the difference between the theoretical and observed values this time is about 0.0246.

Distribution of Sample Means vs Distribution of Observations

While the observed mean and variance of our sample means are close to their theoretical values, this gives us no information about the exact shape of the distribution. Before we ascertain whether our sample means are normally distributed, however, it is worth visually comparing the underlying distribution of our observations, i.e. the exponential distribution, with the normal distribution having the same mean and variance, in order to appreciate their dissimilarity.



Clearly, our individual observations are far from normally distributed. Next, we will compare the distribution of our sample means with the normal distribution which has the theoretical mean and variance calculated previously.



This plot shows that the means of samples of 40 exponentials follow a distribution closely approximated by a normal "bell curve", a drastically different distribution from that of the observations themselves. Thus, in this situation at least, the CLT is shown to be an excellent approximation of reality.

Appendix: Code for Reproducing the Plots

Empirical Distribution of Observations vs Normal Distribution

```r
library(ggplot2)
ggplot(data = data.frame(rv = as.vector(simulated_data))) +
    geom_density(aes(x = rv, color = "Distribution of observations")) +
    stat_function(fun = dnorm, args = list(mean = 1/rate, sd = 1/rate),
                  aes(color = "Normal distribution")) +
    xlim(0, 30) +
    theme(legend.position = "right", legend.title = element_blank()) +
    xlab("Value") +
    ylab("Density") +
    ggtitle("Empirical Distribution of Observations vs Normal Distribution")
```

Empirical Distribution of Sample Means vs Normal Distribution

```r
library(ggplot2)
ggplot() +
    geom_density(data = data.frame(sm = sample_means),
                 aes(x = sm, color = "Distribution of sample means")) +
    stat_function(fun = dnorm, args = list(mean = theoretical_mean,
                                           sd = sqrt(theoretical_variance)),
                  aes(color = "Normal distribution")) +
    theme(legend.position = "bottom", legend.title = element_blank()) +
    xlab("Value") +
    ylab("Density") +
    ggtitle("Empirical Distribution of Sample Means vs Normal Distribution")
```