Statistical Inference Course Project

Part 1: Simulation Exercise

Irena Papst

Overview

We investigate the Central Limit Theorem (CLT), which essentially states that a distribution of sample means (calculated from different samples of the same population) tends toward a normal distribution as the number of sample means collected tends to infinity. What is remarkable about this theorem is that it holds even when the observations sampled are *not* normally distributed themselves. In fact, we will simulate draws from an *exponential* distribution to illustrate the CLT.

Simulating draws from an exponential distribution

We sample 40 observations from an exponential distribution with $\lambda = 0.2$ and then take the sample means and variances. We repeat this process 1000 times.

```r
## Load necessary libraries
library(purrr)
library(ggplot2)
library(gridExtra)
```

```r
## Set seed for reproducible results
set.seed(15)

## Set parameters
lambda <- 0.2
n_draw <- 40
n_sim <- 1000

## Simulate draws
x <- replicate(n_sim, rexp(n_draw, lambda), simplify = FALSE)
## Calculate sample means
sample_means <- map_dbl(x, mean)
## Calculate sample variances
s2 <- map_dbl(x, var)
```

Empirical vs. theoretical mean

The theoretical mean of this exponential distribution is $\mu = 1/\lambda = 1/0.2 = 5$, so we expect the mean of our simulated sample means (*i.e.* the empirical mean) to be close to this value.

We calculate and display both the theoretical and empirical means.

```r
## Calculate theoretical mean
mu <- 1/lambda

## Calculate mean of sample means
xbar <- mean(sample_means)

## Print results
print(paste0("Theoretical mean: ", mu))
```

```
## [1] "Theoretical mean: 5"
```

```r
print(paste0("Empirical mean: ", round(xbar,3)))
```

```
## [1] "Empirical mean: 4.981"
```

To fully convince ourselves that the empirical mean, $\bar{X}$, is close to the theoretical mean, μ, we can quantify just how close these two values are by calculating the relative difference between the empirical and theoretical means, $(\bar{X} - \mu)/\mu$.

```r
## Calculate relative difference
rel_diff <- (xbar-mu)/mu

## Print results
print(paste0("Relative difference: ", round(rel_diff, 5)))
```

```
## [1] "Relative difference: -0.00389"
```

This relative difference indicates that the empirical mean is only 0.389% smaller than the sample mean: a very small difference.

Empirical vs. theoretical variance

The variance of this exponential distribution is $\sigma^2 = 1/\lambda = 1/0.2 = 5$, so we expect the variance of our simulated sample means (*i.e.* the empirical variance) to be close to the theoretical variance of $\sigma^2/n = 1/(n\lambda^2)$.

We calculate and display both the theoretical and empirical variances.

```r
## Calculate theoretical variance
sigma2 <- 1/(n_draw*lambda^2)
```

```r
## Calculate variance of sample means
s2 <- var(sample_means)

## Print results
print(paste0("Theoretical variance: ", sigma2))
```

```
## [1] "Theoretical variance: 0.625"
```

```r
print(paste0("Empirical variance: ", round(s2,3)))
```

```
## [1] "Empirical variance: 0.619"
```

The relative difference between the empirical variance, s^2, and theoretical variance is given by $(s^2 - \sigma^2/n)/\sigma^2/n$.

```r
## Calculate relative difference
rel_diff <- (s2-sigma2)/sigma2

## Print results
print(paste0("Relative difference: ", round(rel_diff, 5)))
```

```
## [1] "Relative difference: -0.01013"
```

This relative difference indicates that the empirical variance is only 1.013% smaller than the sample variance: again, a small difference.

Distribution of sample means

According to the CLT, the distribution of sample means should be normal (with mean μ and variance σ^2/n) as the number of samples, n, goes to infinity. Thus the distribution of sample means for our 40 draws should be *approximately* normal with those parameters.

To illustrate this point, we plot this histogram of sample means and overlay a normal distribution with mean $\mu = 5$ and variance $\sigma^2/n = 5^2/40 = 0.625$ in figure 1.

```r
## Convert sample mean data to a data frame
data <- data.frame(sample_means = sample_means)

## Set histogram parameters
n_bins <- 25
fill <- "grey60"

## Plot histogram of sample means with CLT predicted
## normal distribution overtop
subtitle <- "(compared with the theoretically predicted normal distribution)"
p <- ggplot(data = data,
```

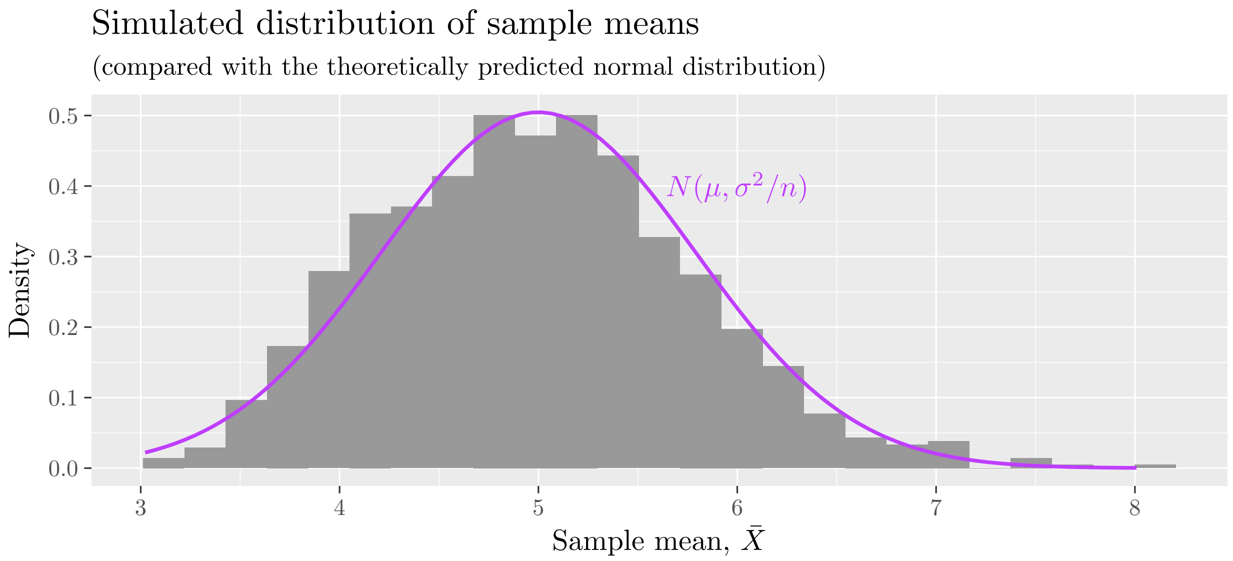

Figure 1: The simulated distribution of 10,000 sample means of 40 exponential random variables with $\lambda = 0.2$, compared with the normal distribution predicted by the CLT (mean $\mu = 5$ and variance $\sigma^2 = 0.625$. The simulated sample mean distribution matches very well with the CLT-predicted normal.

```
            mapping = aes(x = sample_means)) +
  geom_histogram(aes(y=..density..),
                 bins = n_bins,
                 fill = fill) +
  labs(x = "Sample mean, $\\bar{X}$",
       y = "Density",
       title = "Simulated distribution of sample means",
       subtitle = subtitle) +
  stat_function(fun = dnorm,
                args = list(mean = mu, sd = sqrt(sigma2)),
                colour = "darkorchid1",
                size = 1.2) +
  annotate("text", x = 6, y = 0.4,
           label = "$N(\\mu, \\sigma^2/n)$",
           colour = "darkorchid1")

print(p)
```

We see that the theoretically predicted normal distribution matches quite well with the simulated distribution, and so we conclude that the simulated distribution is approximately normal.

Note that the distribution of many random exponentials is *not* the same as the distribution of many *sample means* of 40 exponentials, as illustrated in figure 2.

```r
## Format simulations of random exponentials to work with
## ggplot
exp_draws <- data.frame(x = unlist(x))

## Plot histogram of many random exponentials
p1 <- ggplot(data = exp_draws,
             mapping = aes(x = x)) +
  geom_histogram(aes(y = ..density..),
                 bins = n_bins,
                 fill = fill) +
  labs(x = "Exponential draw",
       y = "Density",
       title = "\\small{Distribution of 40000 exponentials}")

## Plot histogram of sample means of 40 exponentials
p2 <- ggplot(data = data,
            mapping = aes(x = sample_means)) +
  geom_histogram(aes(y = ..density..),
                 bins = n_bins,
                 fill = fill) +
  labs(x = "Sample mean, $\\bar{X}$",
       y = "",
       title = "\\footnotesize{Distribution of 10000 sample means}",
       subtitle = "(of 40 exponentials)")

## Arrange plots in a grid
grid.arrange(p1, p2, nrow = 1)
```

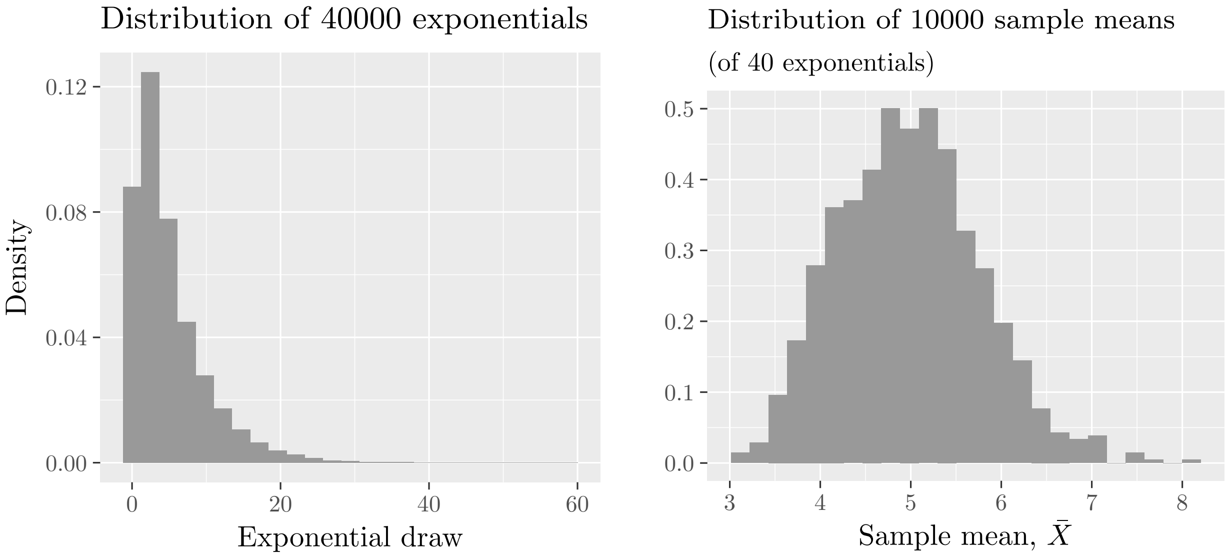

Figure 2: The distribution of 40,000 random variables drawn from an exponential distribution with $\lambda = 0.2$ compared to the distribution of 10,000 sample means of 40 exponentials (using the same data). We see that the distribution of exponentials is *exponential*, as expected, but the distribution of *sample means* of the same exponentials is *normal*, as predicted by the CLT.