

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 the 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.

Sample means vs. theoretical mean

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

We calculate and plot a histogram of the relative difference between our theoretical mean and our 1000 sample means $((\mu - \bar{X}_i)/\mu$ for $i = 1, 2, \dots, 1000$).

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

```
## [1] "Empirical mean: 4.98"
```

While it is pretty clear that the empirical mean, \bar{X} , is close to the theoretical mean, μ , we can also calculate the relative difference between the empirical and theoretical means, $(\bar{X} - \mu)/\mu$, to convince ourselves further that these values truly are close to each other.

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

Sample variances vs. theoretical variance

The theoretical variance this exponential distribution is also $\sigma^2 = 1/\lambda = 1/0.2 = 5$, so we expect our simulated sample variances to be close to this value.

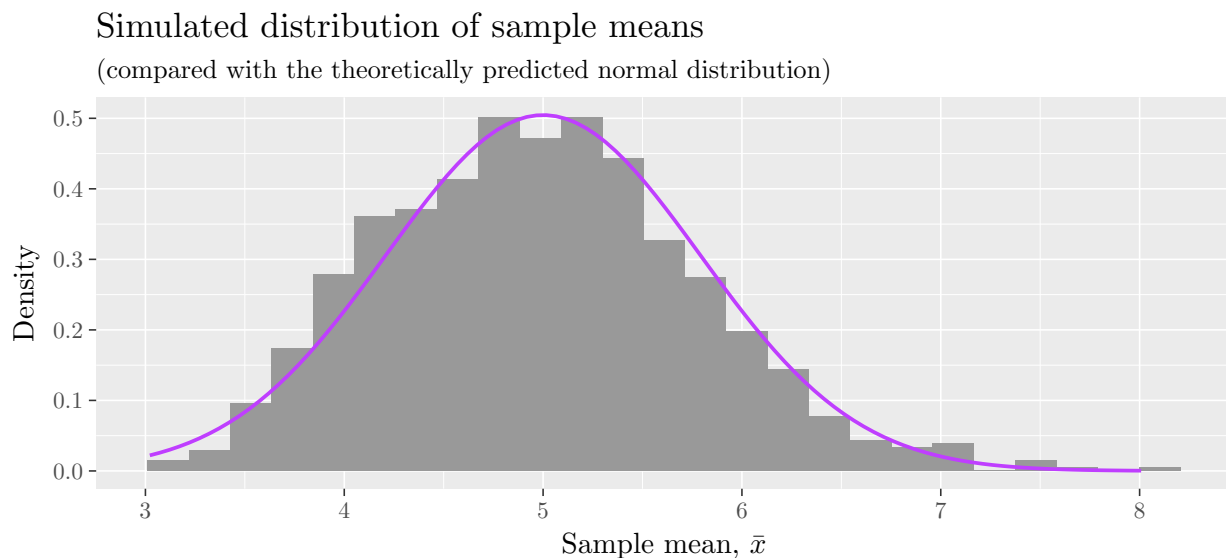
We calculate and plot a histogram of the relative difference between our theoretical variance and our 1000 sample variances $((\sigma^2 - s_i^2)/\sigma^2$ for $i = 1, 2, \dots, 1000$).

Again, the relative difference between the sample variances and the theoretical variance accumulate around zero, *i.e.* the sample variances accumulate around the theoretical variance. However, the sample variances vary more from their associated theoretical value than the sample means; the majority of sample variances are within 100% of the theoretical variance.

Distribution of sample means

According to the Central Limit Theorem, 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 these parameters.

To illustrate this point, we plot this histogram of sample means and overlay a normal distribution with mean μ and variance $\sigma^2/40$.



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

Appendix

The following is all the code used to generate the figures in this report.

```
## Set default knitr chunk options
knitr::opts_chunk$set(dev = "tikz", echo = FALSE,
  fig.height = 3)

## Load necessary libraries
library(purrr)
library(ggplot2)

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

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

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

## Calculate theoretical mean
mu <- 1/lambda

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

## Print results
print(paste0("Theoretical mean: ", mu))
print(paste0("Empirical mean: ", round(xbar,
  2)))
## Calculate relative difference
rel_diff <- (xbar - mu)/mu

## Print results
print(paste0("Relative difference: ", round(rel_diff,
  5)))
## Set theoretical variance
sigma2 <- 1/lambda^2
```

```

## Calculate relative differences in means
## from mu (compared to mu)
diff_vars <- data.frame(diffs = (s2 - sigma2)/sigma2)

## Plot histogram of differences
p <- ggplot(data = diff_vars, mapping = aes(x = diffs)) +
  geom_histogram(bins = n_bins, fill = fill) +
  labs(x = "$(\sigma^2 - s^2)/\sigma^2$",
       y = "Count", title = "Relative difference of sample variances",
       subtitle = "(compared to the theoretical variance)") +
  geom_vline(xintercept = 0, colour = "dodgerblue",
             size = 1.2)

print(p)

## 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
p <- 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 = "Density", title = "Simulated distribution of sample means",
                 subtitle = "(compared with the theoretically predicted normal distribution)") +
  stat_function(fun = dnorm, args = list(mean = mu,
                 sd = sqrt(sigma2/n_draw)), colour = "darkorchid1",
                 size = 1.2) + NULL

print(p)

```