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.

Sample means 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.

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

```
## [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, $(\bar{X} - \mu)/\mu$.

```
## [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.

Sample variances 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.

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

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

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

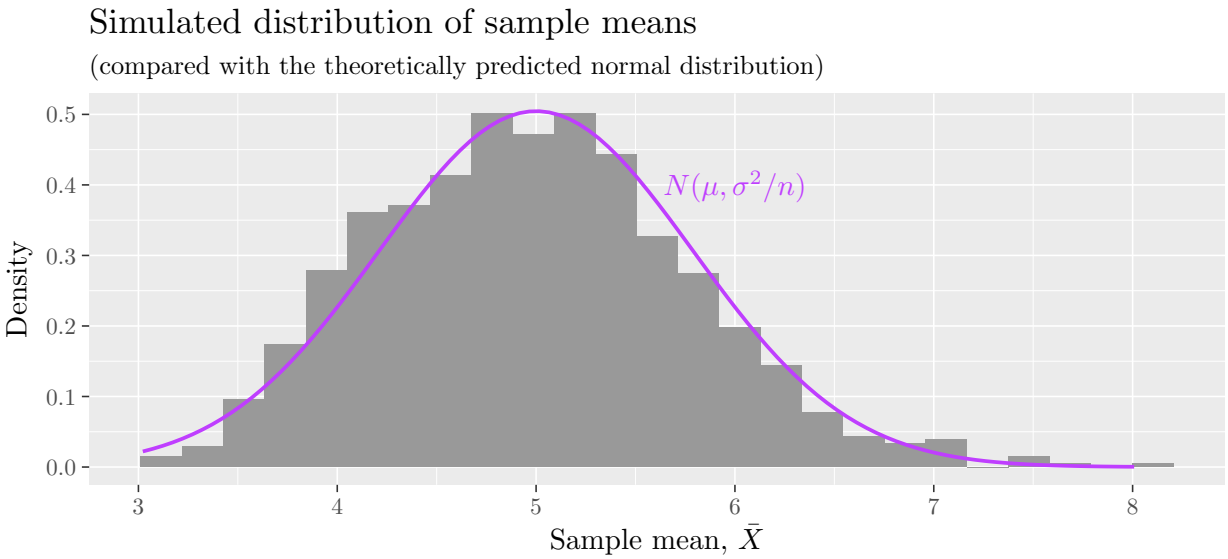
```
## [1] "Relative difference: -0.01013"
```

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 $\sigma^2/n = 5^2/40$.



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.

Appendix A: Code

The following code was used to generate all of the results presented 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, 3)))

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

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

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

## Calculate variance of sample means
```

```r
s2 <- var(sample_means)

## Print results
print(paste0("Theoretical variance: ", sigma2))
print(paste0("Empirical variance: ", round(s2, 3)))

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

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

## 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, 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)

sessionInfo()
```

Appendix B: Session Info

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_2.2.1 purrr_0.2.4
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.16     knitr_1.20       magrittr_1.5     munsell_0.4.3
##  [5] colorspace_1.3-2 rlang_0.2.0      filehash_2.4-1   stringr_1.3.0
##  [9] plyr_1.8.4       tools_3.4.4      grid_3.4.4       tikzDevice_0.11
## [13] gtable_0.2.0     tinytex_0.5      htmltools_0.3.6  yaml_2.1.18
## [17] lazyeval_0.2.1   rprojroot_1.3-2  digest_0.6.15    tibble_1.4.2
## [21] formatR_1.5      evaluate_0.10.1  rmarkdown_1.9    labeling_0.3
## [25] stringi_1.1.7    pillar_1.2.1     compiler_3.4.4   scales_0.5.0
## [29] backports_1.1.2
```