Statistical Inference Course Project - Part 1

Norman Norris

May 22, 2015

Exponential Distribution Averages and the Central Limit Theorem

Overview

This project investigates the properties of the distribution of averages of exponentials, with a sample size of 40, and compares it to the Central Limit Theorem (CLT). The CLT states the distribution of averages of independent and identically distributed (iid) variables (properly normalized) becomes that of a standard normal as the sample size increases. Therefore, the project explores whether averages of exponential distributions become normally distributed when simulated many times.

Objectives

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should 1. Show the sample mean and compare it to the theoretical mean of the distribution. 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. 3. Show that the distribution is approximately normal.

Simulations

R was used to simulate the exponential distribution with a sample size of 40 and lambda of 0.2, calculate the average of this distribution, and repeat this 1000 times. A histogram of these averages was plotted along with the calculation of the sample mean, standard deviate and variance, which were compared with the theoretical mean, standard deviate and variance.

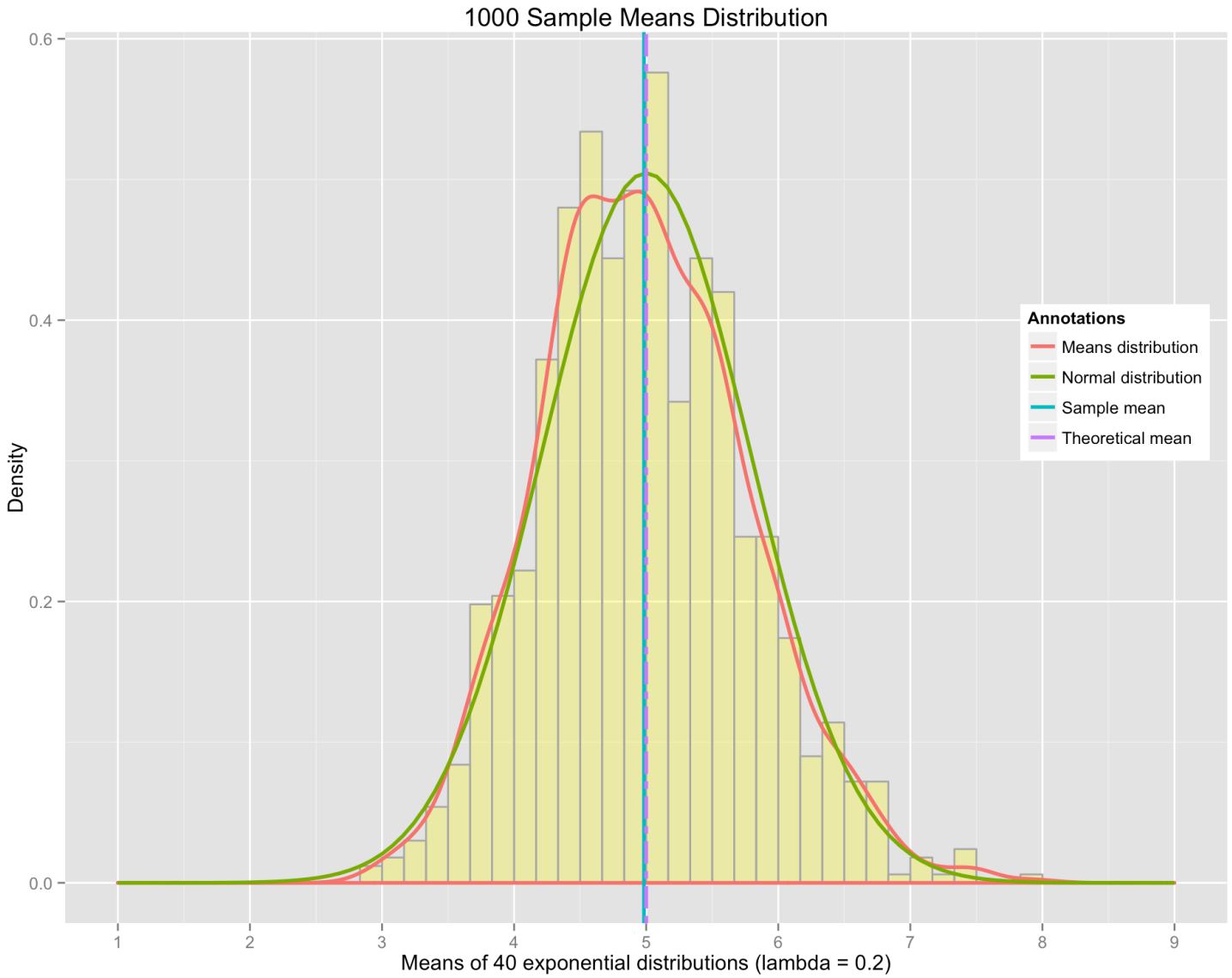
Results

Objectives 1 and 2:

The table below summarizes the results for objectives 1 and 2. The R code to run the simulation, to create the table and plots below, and to calculate the sample/theoretical means and variances appears in the appendix.

```
##              Variable Theoretical Sample
##                  Mean           5  4.983
##  Standard Deviation       0.791  0.779
##              Variance       0.625  0.606
```

To 3 decimal places, the sample mean (4.983) is very close to the theoretical mean of 5. The sample standard deviation (0.779) and variance (0.606) are also very close to their theoretical values of 0.791 and 0.625 respectively. The plot below shows the sample means distribution, the normal distribution, as well as the sample mean and the theoretical mean, all of which are labelled.

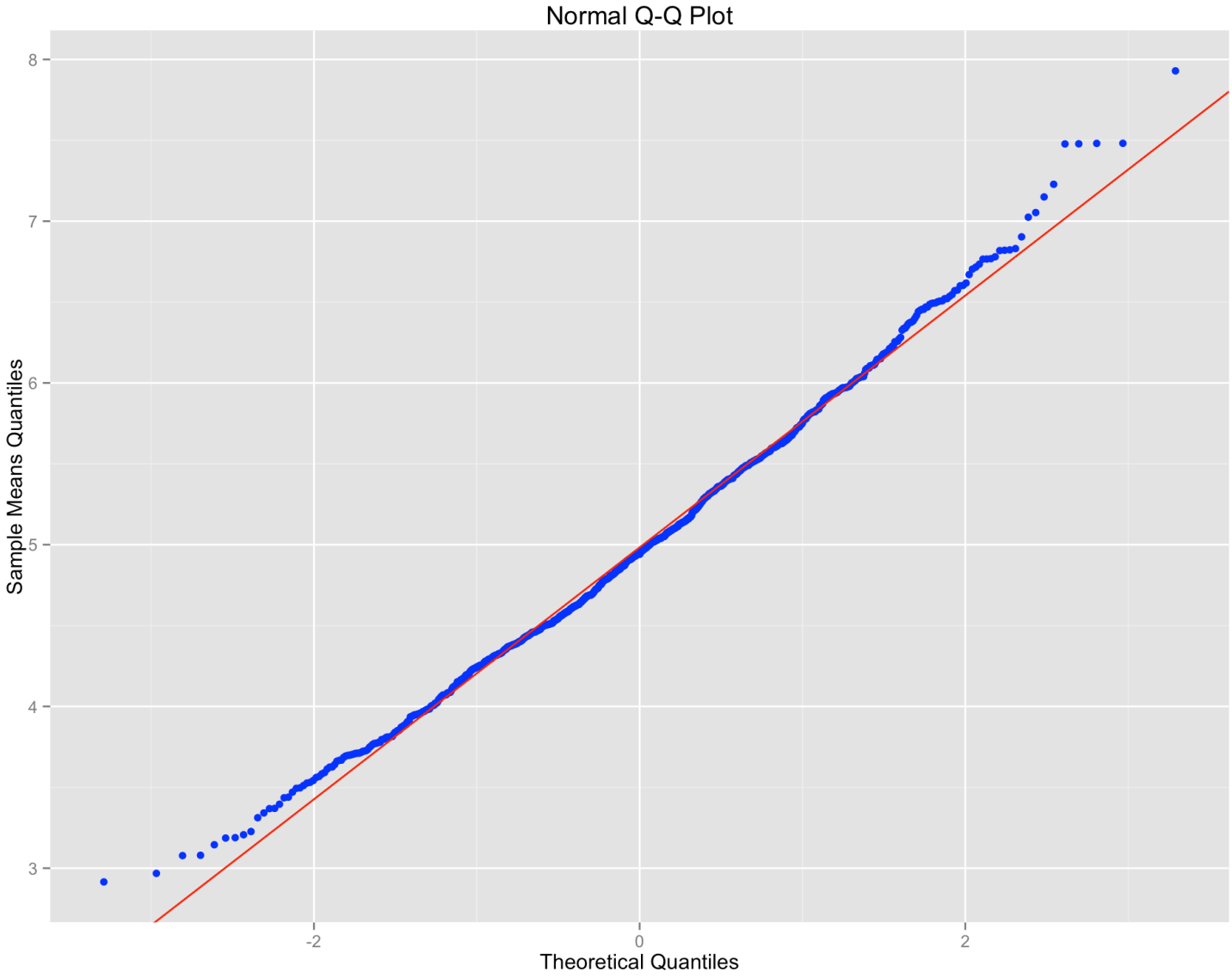


Objective 3:

Objective 3 asks to show that the sample means distribution is approximately normal. In order to do this, I produced a Q-Q plot (http://en.wikipedia.org/wiki/Q–Q_plot (http://en.wikipedia.org/wiki/Q–Q_plot)), which provides a visual guide of how normal a distribution is.

Briefly, a Q-Q plot looks at the various quantiles of the sample and compares them to the normal quantiles (and plots these). If the sample distribution is (approximately) normal, we would expect the points plotted to fall along the “y = x” line.

The Q-Q plot below shows the sample means distribution is definitely (approximately) normal, with the majority of points lying close to the “y = x” line (sample quantiles equal normal/theoretical quantiles), which I have added in red for reference. The sample means do start to deviate beyond +/- 2 standard deviations, which is understandable, given they are means for a sample size of only 40.



Appendix

Here is the R code to run the simulations and product the table and plots.

```
## This R script generates the answers for the Statistical Inference Course Project - Part 1

## Load packages
library(ggplot2)
```

```
## Set the seed value so the results are reproducible
set.seed(7)

## Exponential distribution sample parameters
num_sims <- 1000
lambda <- 0.2
sample_size <- 40

## Calculate the theoretical mean, standard deviation, and variance
## Theoretical mean = 1 / lambda
theoretical_mean = 1 / lambda

## Theoretical standard deviation = theoretical mean / sqrt(sample_size)
theoretical_sd = theoretical_mean / sqrt(sample_size)

## Theoretical variance = theoretical standard deviation ^ 2
theoretical_variance = theoretical_sd ^ 2

## Create a matrix of the theoretical values which will be added to a table for print
out
theoretical_values <- matrix(c(theoretical_mean, round(theoretical_sd, 3), round(theo
retical_variance, 3)))

## Creates a matrix (1000 rows, 40 columns)
## Each row represents an exponential distribution with a sample size of 40 and lambd
a of 0.2
## The exponential distribution is simulated 1000 times, 1 in each row
exponentials <- matrix(rexp(num_sims * sample_size, rate = lambda), num_sims, sample_
size)

## The mean of each exponential/row is calculated and stored in sample_means (1000 ro
ws by 1 column)
sample_means <- rowMeans(exponentials)

## Calculate the sample mean, standard deviation, and variance
sample_rowMean <- mean(sample_means)
sample_rowSD <- sd(sample_means)
sample_rowVariance <- var(sample_means)

## Create a matrix of the sample values which will be added to a table for printout
sample_values <- matrix(c(round(sample_rowMean, 3), round(sample_rowSD, 3), round(sam
ple_rowVariance, 3)))

## Create a matrix of the row names which will be added to a table for printout
variable_names <- matrix(c("Mean", "Standard Deviation", "Variance"))

## Create a summary table (data frame) to print out theoretical vs sample data
summary_table <- data.frame(cbind(variable_names, theoretical_values, sample_values))
colnames(summary_table) <- c("Variable", "Theoretical", "Sample")
```

```
print(summary_table)

## Create a histogram of the means of the sample distributions
## Adds lines for the distribution of sample means, the normal distribution, the samp
le mean, and theoretical mean
## Customizes colours, axes labels, x-axis scale, and adds a legend
means.df <- data.frame(Means = sample_means)

## Open the png device
png(file = "./plot1.png", width = 800, height = 600)

g1 <- ggplot(means.df, aes(x = Means)) +
    geom_histogram(aes(y = ..density..), fill = "yellow", binwidth = 1/6, color = "da
rkgrey", alpha = 1/3) +
    geom_density(aes(color = "Means distribution"), size = 1, show_guide = FALSE) +
    stat_function(fun = dnorm, arg = list(mean = theoretical_mean, sd = theoretical_s
d), aes(color = "Normal distribution"), size = 1) +
    geom_vline(aes(xintercept = sample_rowMean, colour = "Sample mean"), size = 1) +
    geom_vline(aes(xintercept = theoretical_mean, colour = "Theoretical mean"), size
= 1, linetype = "twodash") +
    theme(legend.justification = c(1.15,-1.4), legend.position = c(1,0.5)) +
    labs(title = "1000 Sample Means Distribution", x = "Means of 40 exponential distr
ibutions (lambda = 0.2)", y = "Density") +
    scale_x_continuous(limits = c(1, 9), breaks = 1:9) +
    scale_color_discrete(name = "Annotations")

print(g1)

## Close device
dev.off()

## Create Q-Q Plot, which plots sample means quantiles vs theoretical/normal quantile
s
## Adds (red) line where sample and theoretical quantiles are equal
## Open the png device
png(file = "./plot2.png", width = 800, height = 600)

g2 <- ggplot(means.df, aes(sample = Means)) + stat_qq(color = "blue", alpha = 1) +
    geom_abline(intercept = mean(means.df$Means), slope = sd(means.df$Means), color =
"red") +
    labs(title = "Normal Q-Q Plot", x = "Theoretical Quantiles", y = "Sample Means Qu
antiles")

print(g2)

## Close device
dev.off()

## End of file
```