Comparing Exponential Distribution to the Central Limit Theorem

Dominic Lloyd

Tuesday, October 22, 2015

Overview

In this document we will investigate the exponential distribution in R and compare it with the Central Limit Theorem.
The exponential distribution will be simulated in R with rexp.
We will investigate the distribution of averages of 40 exponentials across a thousand simulations.

Simulations

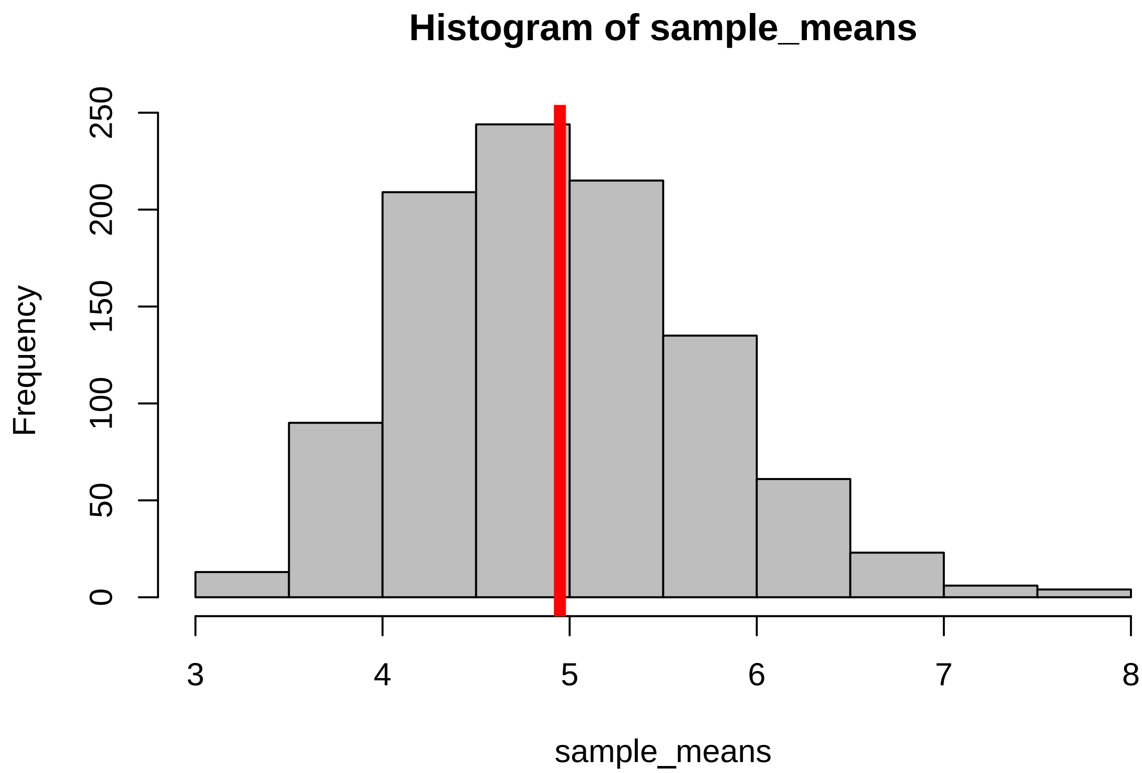
The document will illustrate as follows:
1. The sample mean compared to the theoretical mean of the distribution.
2. The variance of the sample compared to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal. (With focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials)

Sample Mean versus Theoretical Mean

In this test we are trying to show that we get similar results for mean and standard deviation for samples as we get for larger populations.
We know that when we use rexp to generate exponential numbers that the mean of these number will be 1/lambda (or 1/rate). Therefore if we use lambda = 0.2 as the rate parameter for rexp, we would expect to see a mean around 5 and a standard deviation around 5.

In this section we will generate a thousand samples of 40 random exponential numbers. In the histogram we will show the sample mean.

```
sample_means = NULL

for (i in 1 : 1000) sample_means = c(sample_means, mean(rexp(40, rate = 0.2)))

mean_mns <- mean(sample_means)

hist(sample_means, col = 'grey')

abline(v = mean_mns, col = "red", lwd = 6)
```



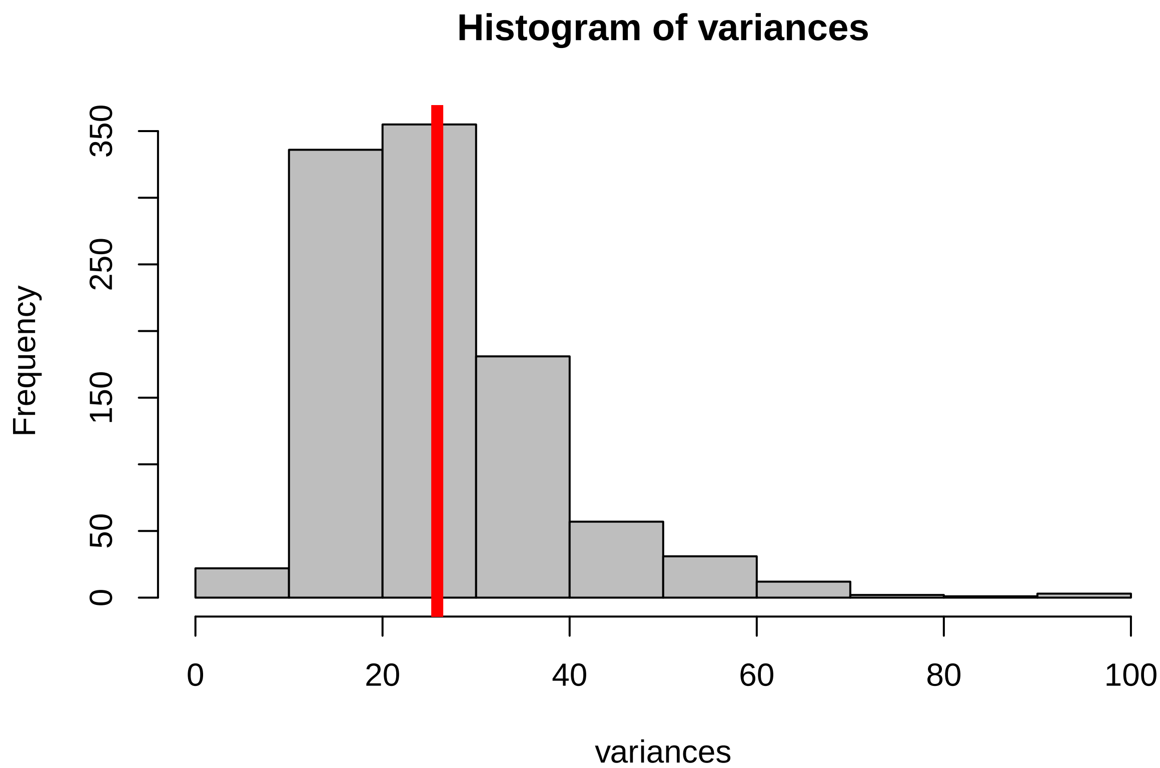
As the graph shows, from the 1000 simulations of 40 exponentials the data has become centred around the mean (red line) which shows the central limit theorem in action. The theoretical mean is 5 and the sample mean (red line) should be approximately 5.

Sample Variance versus Theoretical Variance

Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances.

The theoretical variance is the standard deviation (5) to the power $2 = 25$.
Repeating our 1000 samples of 40 exponentials, we find the standard deviation to identify the variance. We then take the mean of the variances and plot it on our graph (red line):

```
sample_sd = NULL

for (i in 1 : 1000) sample_sd = c(sample_sd, sd(rexp(40, rate = 0.2)))

variances <- sample_sd ^ 2

mean_variances <- mean(variances)

hist(variances, col = 'grey')

abline(v = mean_variances, col = "red", lwd = 6)
```



Distribution

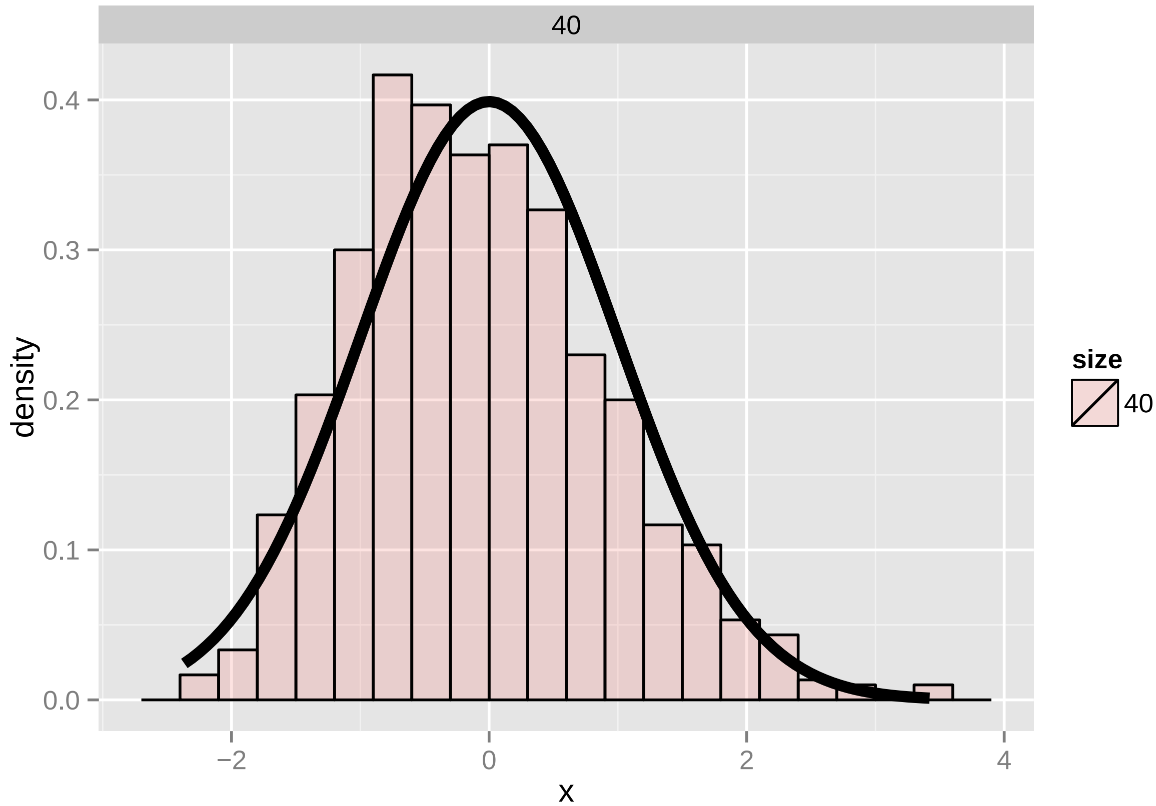
Via figures and text, explain how one can tell the distribution is approximately normal.

In this section we will take the 1000 sample means taken from a simulation of 40 random exponentials and we will plot that on a graph. We will then overlay the standard normal line. This will show us that the distribution is approximately normal.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.2
```

```
nosim <- 1000
cfunc <- function(x, n) sqrt(n) * (mean(x) - 5) / 5
dat <- data.frame(
  x = c(apply(matrix(sample_means, nosim), 1, cfunc, 40)),
  size = factor(rep(c(40), rep(nosim, 1))))
g <- ggplot(dat, aes(x = x, fill = size))
g <- g + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 2)
g + facet_grid(. ~ size)
```


40
density
0.0
0.1
0.2
0.3
0.4
−2
0
2
4
x
size
40