Statistical Inference Course Project Part 1: Simulation Exercise

Hedley Stirrat

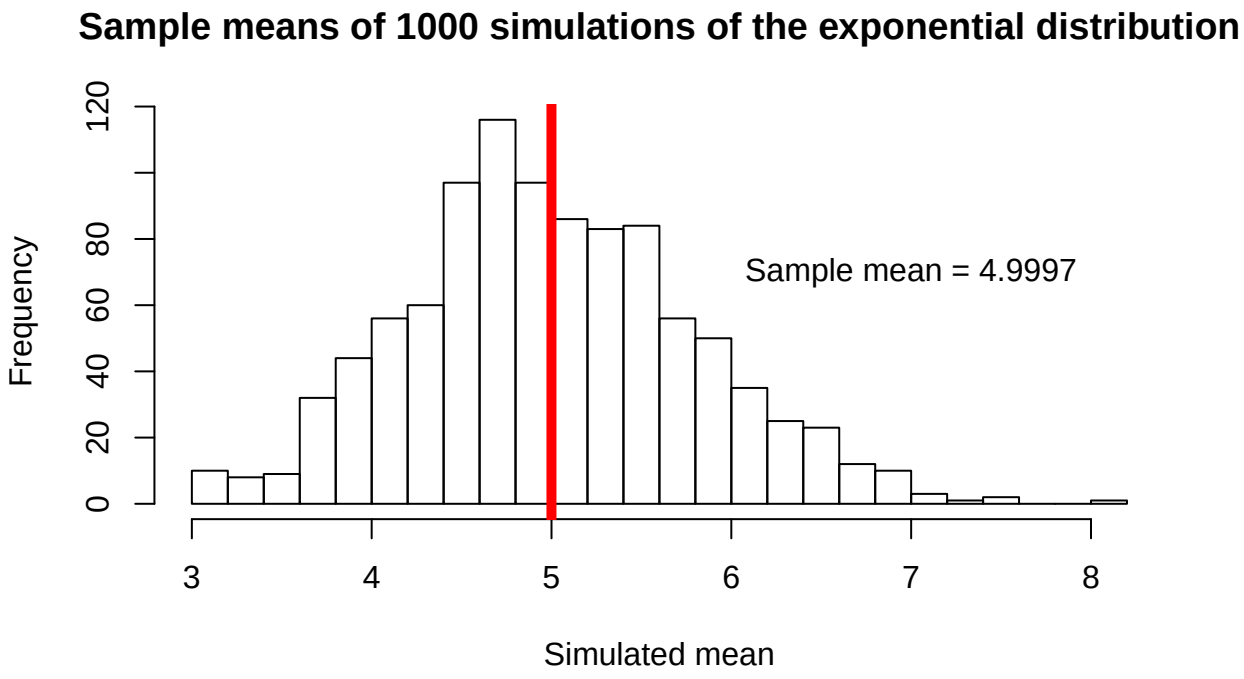
17 May 2020

Overview

- In this section, we investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). The CLT states that the distribution of averages of independent, identically distributed rangom variables becomes that of a standard normal as the sample size increases.
- To explore the CLT, we will run 1000 simulations of the exponential distribution with parameters `n = 40` and `rate = 0.2`.
- The raw results, containing our collection of 1000 simulations of 40 random exponentials, will be stored in the vector `results`. We will also take the mean of each simulation and store it in the vector `mean_results`.

1.1. Comparison of sample mean and theoretical mean

- The theoretical mean of the exponential distribution is `1/lambda`
- Here, `lambda = 0.2`, so the theoretical mean is `1 / 0.2 = 5`.
- The mean of our simulated distribution closely aligns with the theoretical mean:



1.2. Comparison of sample variance and theoretical variance

- The theoretical variance can be calculated as follows: `1 / lambda^2 / n`:

```
1/0.2^2/40
```

```
## [1] 0.625
```

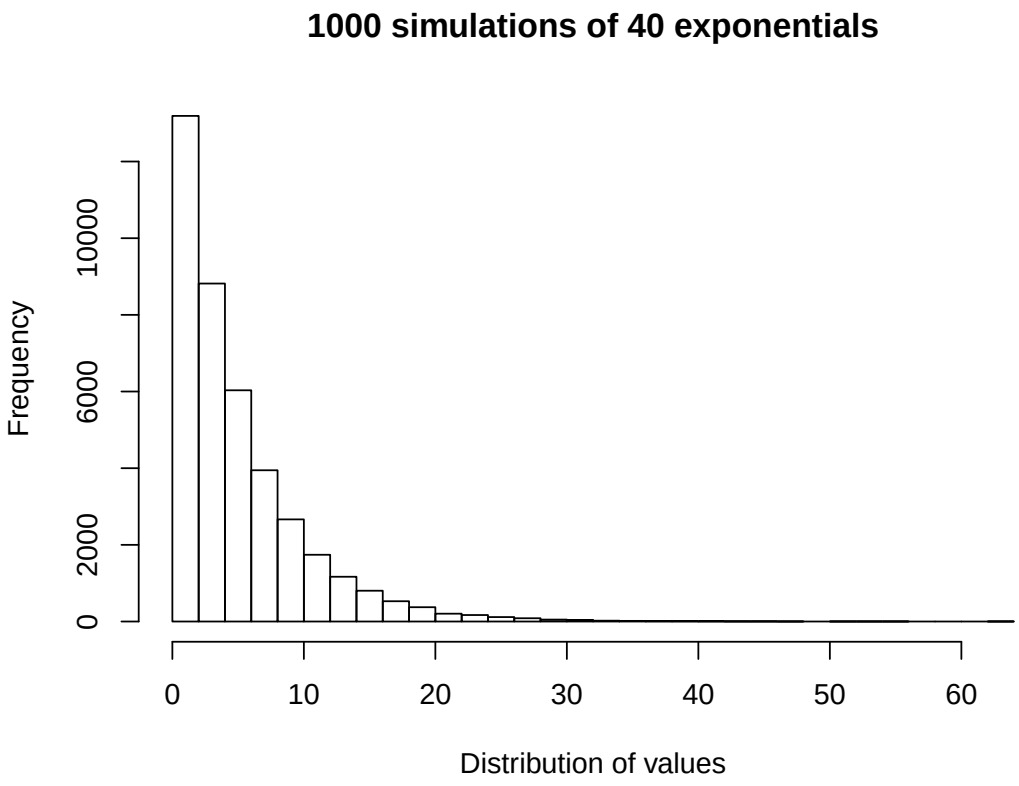
- The sample variance can be calculated in R with the `var` function, and closely aligns with the theoretical variance:

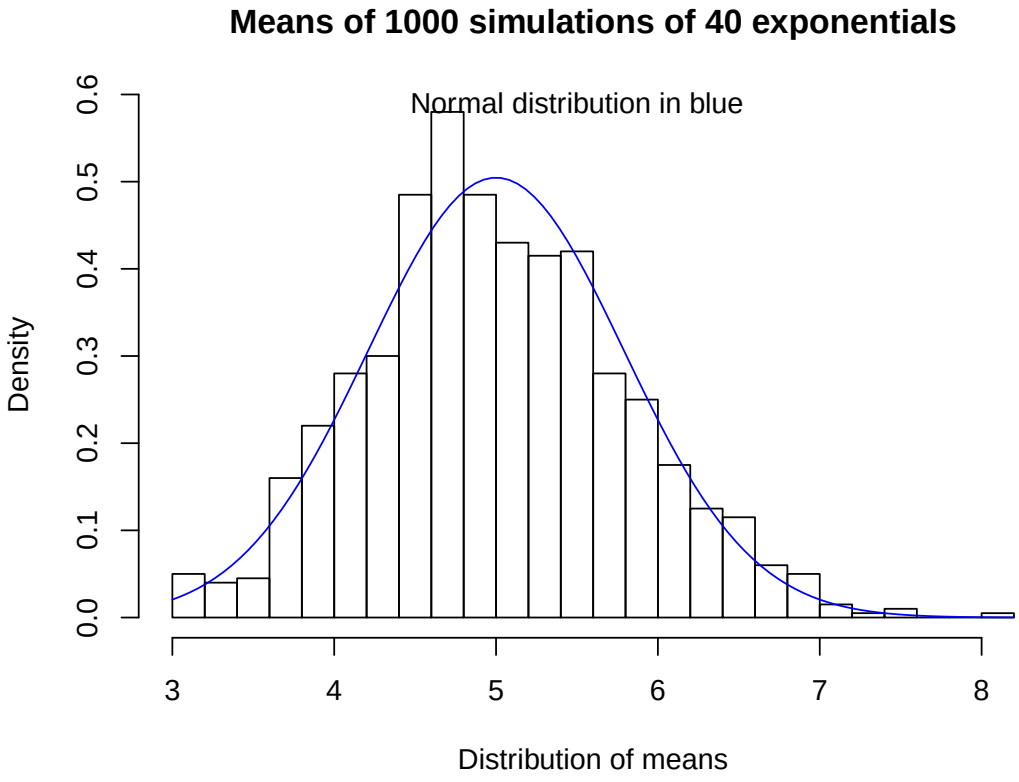
```
var(mean_results)
```

```
## [1] 0.6432442
```

1.3. Demonstration of normal distribution

- To demonstrate that our results—containing 1000 simulated means of 40 exponentials—are approximately normally distributed, we can plot a simple histogram of the simulated means (as we have done above, comparing the sample mean and the theoretical mean). However, to better understand this normal distribution, we will also plot the raw results (i.e., *not the means*) of our 1000 simulations of 40 exponentials.
- The top plot is a histogram of the distribution of the results of our 1000 simulations of 40 exponentials. As expected, the distribution is not Gaussian.
- The bottom plot is a histogram of the 1000 simulated means of our 40 exponentials. Overlaid on the histogram is a normal distribution line in blue, which has been calculated using the theoretical mean and standard deviation.
  - As shown by the plot, the 1000 simulated sample means are approximately normally distributed, and centered around the theoretical population mean of 5.W
  - Due to the CLT, if we were to increase the number of simulations, our results would appear more and more normally distributed.


1000 simulations of 40 exponentials
Frequency
0
2000
6000
10000
0
10
20
30
40
50
60
Distribution of values


Means of 1000 simulations of 40 exponentials
Normal distribution in blue
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
3
4
5
6
7
8
Distribution of means

Appendix

Code used to simulate data:

```
set.seed(100)
results <- NULL
mean_results <- NULL
for (i in 1:1000) {
    temp <- rexp(n = 40, rate = 0.2)
    results <- c(results, temp)
    mean_results <- c(mean_results, mean(temp))
}
```

Code used to create Figure 1:

```
hist(mean_results,
    breaks = 25,
    xlab = "Simulated mean",
    main = "Sample means of 1000 simulations of the exponential distribution")
abline(v = mean(results), lwd=5, col='red')
text(7,70, labels=paste0("Sample mean = ", round(mean(results),5)))
```

Code used to create Figure 2:

```
x <- seq(
    from = min(mean_results),
    to = max(mean_results),
    length = 40)
par(mfrow = c(2, 1))
hist(results,
    breaks = 25,
    main = "1000 simulations of 40 exponentials",
    xlab = "Distribution of values")
hist(mean_results,
    prob = TRUE,
    breaks = 25,
    main= "Means of 1000 simulations of 40 exponentials",
    xlab = "Distribution of means")
curve(dnorm(x, mean = 5, sd = sqrt((5/sqrt(40))^2)), add=TRUE, col = "blue")
text(x = 5.5, y = 0.59, labels = "Normal distribution in blue")
```