Statistical Inference Project Part 1: Simulation Exercise

Jared Kramer

1/22/2021

Introduction

In this report, I investigate the statistical properties of the mean of 40 random variables drawn from an exponential distribution. Using 1,000 simulations, I investigate the *mean* of the distribution of these means (comparing them to the theoretical mean), as well as the *variance* of this distribution (comparing the sample variance to the theoretical variance). Finally, I observe the shape of the distribution compared to a normal distribution.

Simulations

First, I set parameters to be used in the simulations: lambda will be 0.2 for all simulations. The parameter `Nexp` is set to 40 to indicate the number of exponential random variables that will be generated in each simulation, and `Nsimul` indicates that there will be 1,000 simulations run.

```
set.seed(1234)
lambda <- 0.2; Nexp <- 40; Nsimul <- 1000
```

For the simulations, I draw 1000 * 40 random numbers from an exponential distribution, and reshape into a matrix with 1000 rows and 40 columns. Each row is thus a single simulation of 40 data points. I then use `apply` to compute the mean of each row/simulation.

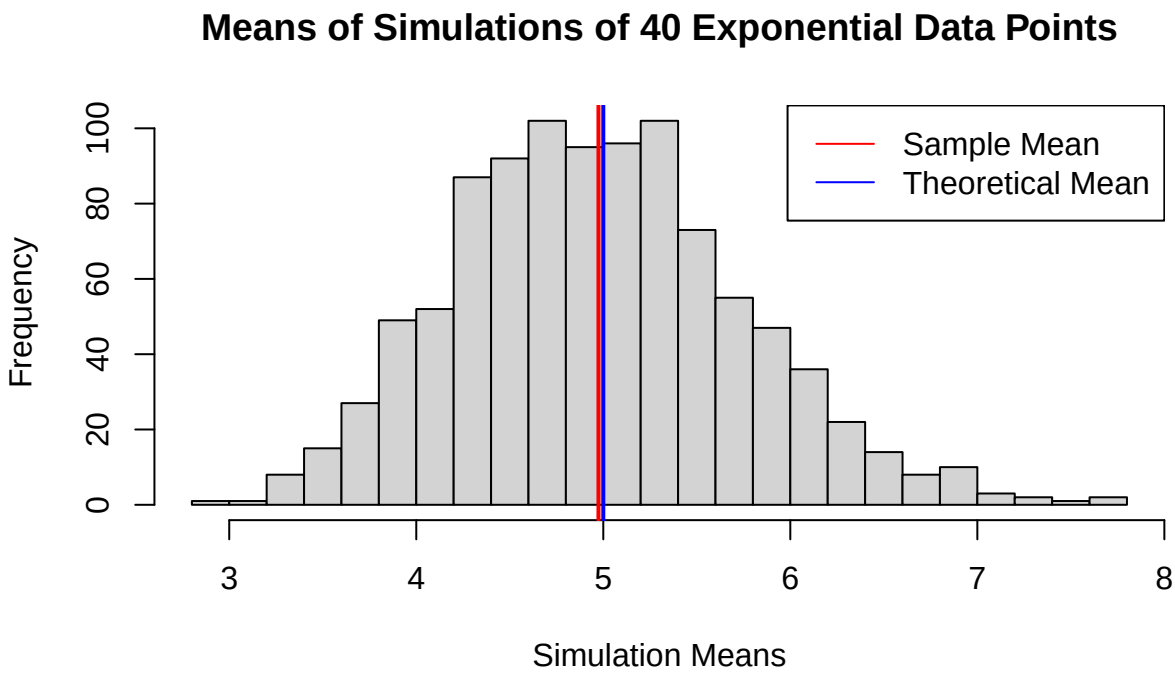
```
M <- matrix(rexp(Nsimul * Nexp, rate = lambda), nrow = Nsimul, ncol = Nexp)
means <- apply(M, 1, mean)
```

Sample vs. Theoretical Means

Theoretically, the mean of 40 random numbers drawn from a single distribution should have a mean that equals the mean of the underlying distribution. Since the underlying distribution (here, exponential) has mean `1/lambda` (here,5) the distribution of the means should also have mean `1/lambda`, i.e., 5.

In our simulations, the sample mean is very close to this theoretical value, i.e.,: 4.9742388 vs. the theoretical of 5. We can see graphically from the below histogram that this is the case as well (red vertical line shows sample mean, and blue vertical line shows theoretical).

```
hist(means, main = "Means of Simulations of 40 Exponential Data Points",
     xlab = "Simulation Means", breaks = 20)
abline(v = mean(means), col = "red", lwd = 2)
abline(v = 1/lambda, col = "blue", lwd = 2)
legend("topright", legend = c("Sample Mean", "Theoretical Mean"),
       col = c("red", "blue"), lty = 1)
```



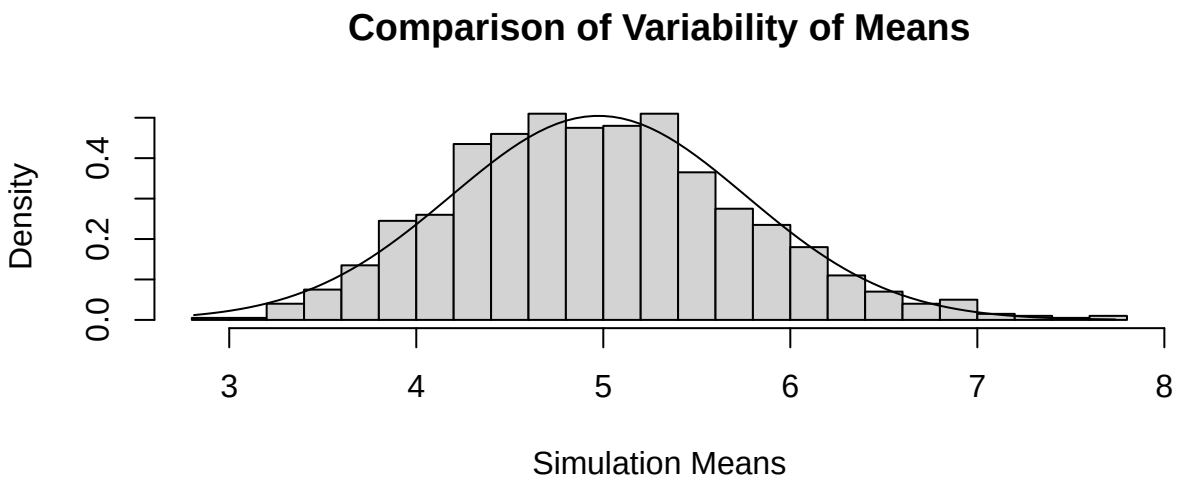
Sample vs. Theoretical Variance

Theoretically, the distribution of the mean of 40 random numbers drawn from a single distribution should have a variance that equals the variance of the underlying distribution, *divided by* the number of data points contributing to the mean.

In this case, the underlying distribution is exponential with standard deviation `1 / lambda` (i.e., 5) and variance `1 / lambda ^ 2` (i.e., 25), the theoretical variance of the means of each simulation is `1 / (lambda ^ 2 * Nexp)`, or 0.625. The simulated variance is quite close to this value, at, 0.5949702.

We can visualize this variability differential by again plotting a histogram of the simulation means, and overlaying a normally distributed curve with mean equal to the simulated mean, and variance equal to the theoretical variance.

```
hist(means, main = "Comparison of Variability of Means", prob = TRUE, breaks = 20,
     xlab = "Simulation Means")
x <- seq(min(means), max(means), length = 100)
y <- dnorm(x, mean = mean(means), sd = 1/(lambda * sqrt(Nexp)))
lines(x, y)
```



Shape of Distribution

For a large number of data points per simulation, the central limit theorem would imply that the distribution of the means will approach the shape of a normal distribution.

We can visualize this roughly from the histogram in the previous section, which seems to match reasonably well with the overlaid normal density curve. For a more precise comparison, I use a quantile-quantile plot that compares the quantiles of a normalized version of the means (subtracting the sample mean from each and dividing by the sample standard deviation), with the theoretical quantiles of a standard normal distribution.

The graph shows a reasonably good match throughout the bulk of the distribution, although in the "tails" of the the distribution the simulations diverge somewhat from normal: on the left tail, the simulated distribution is slimmer than in a normal distribution, and in the right tail, the simulated distribution is slightly "fatter".

```
normalized <- (means - mean(means))/(sd(means))
qqnorm(normalized); qqline(normalized)
```