Statistical Inference Project Part 1

Camille Tolentino

10/7/2020

Introduction

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with `rexp(n, lambda)` where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Preparing the simulation

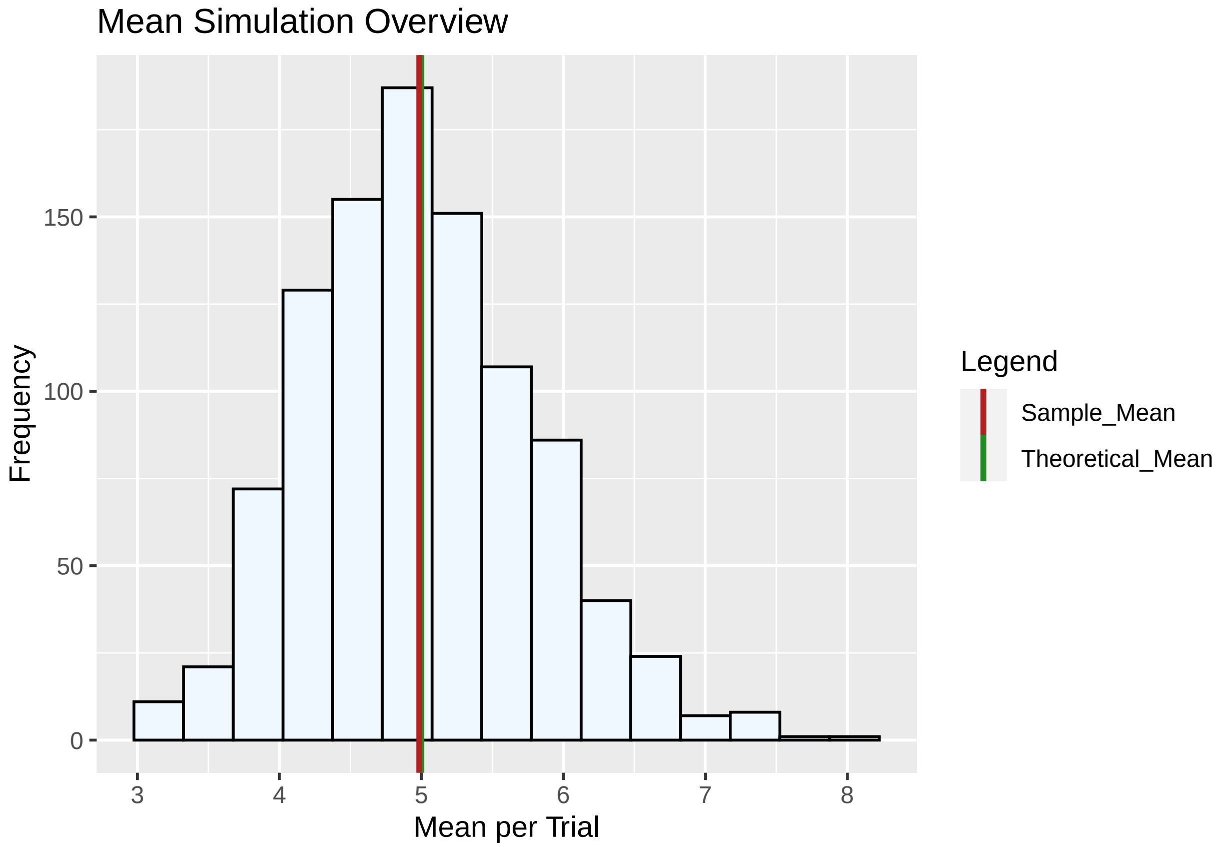
Using the the premise of `lambda = 0.2` and `n = 40` with number of simulations at 1000, we can generate the sample matrix using random numbers which we generate from seed. Setting the `seed = 200` should give the exact same results as the rest of this report.

```r
set.seed(200)
lambda <- 0.2
n <- 40
sim <- 1000

sampleNum <- replicate(sim, rexp(n,lambda))
sampleMat <- matrix(sampleNum, sim, n)
```

Show the sample mean and compare it to the theoretical mean of the distribution

Looking at the mean of the generated matrix, we can come up with a histogram of mean per trial and compare it to our theoretical mean. Recall that the mean for an exponential distribution is equal to `1/lambda`

```r
matMean <- rowMeans(sampleMat)
sampleMean <- mean(matMean)
theoMean <- 1/lambda

library(ggplot2)
qplot(matMean,geom="histogram",
      binwidth = 0.35,
      main = "Mean Simulation Overview",
      xlab = "Mean per Trial",
      ylab = "Frequency",
      fill=I("aliceblue"),
      col=I("black"))+
  geom_vline(aes(xintercept = theoMean, color="Theoretical_Mean"),size=1)+
  geom_vline(aes(xintercept = sampleMean, color="Sample_Mean"),size=1)+
  scale_color_manual(name = "Legend", values = c(Theoretical_Mean = "forestgreen", Sample_Mean = "fireb
```



From here we can see that while not exactly equal, the means are approximately the same. We can also look at the actual values and take the difference to reach the same conclusion.

```
theoMean - sampleMean
```

```
## [1] 0.01587422
```

Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution

Next we look at the variance of the sample matrices and the theoretical variance of the distribution. Recall that the theoretical variance of the exponential distribution is given by $\frac{(1/lambda)^2}{n}$.

```
sampleVar <- var(matMean)
theoVar <- ((1/lambda)^2)/n
cat("Sample variance is",sampleVar,"and theoretical variance is",
    theoVar,"with difference of",abs(theoVar - sampleVar))
```

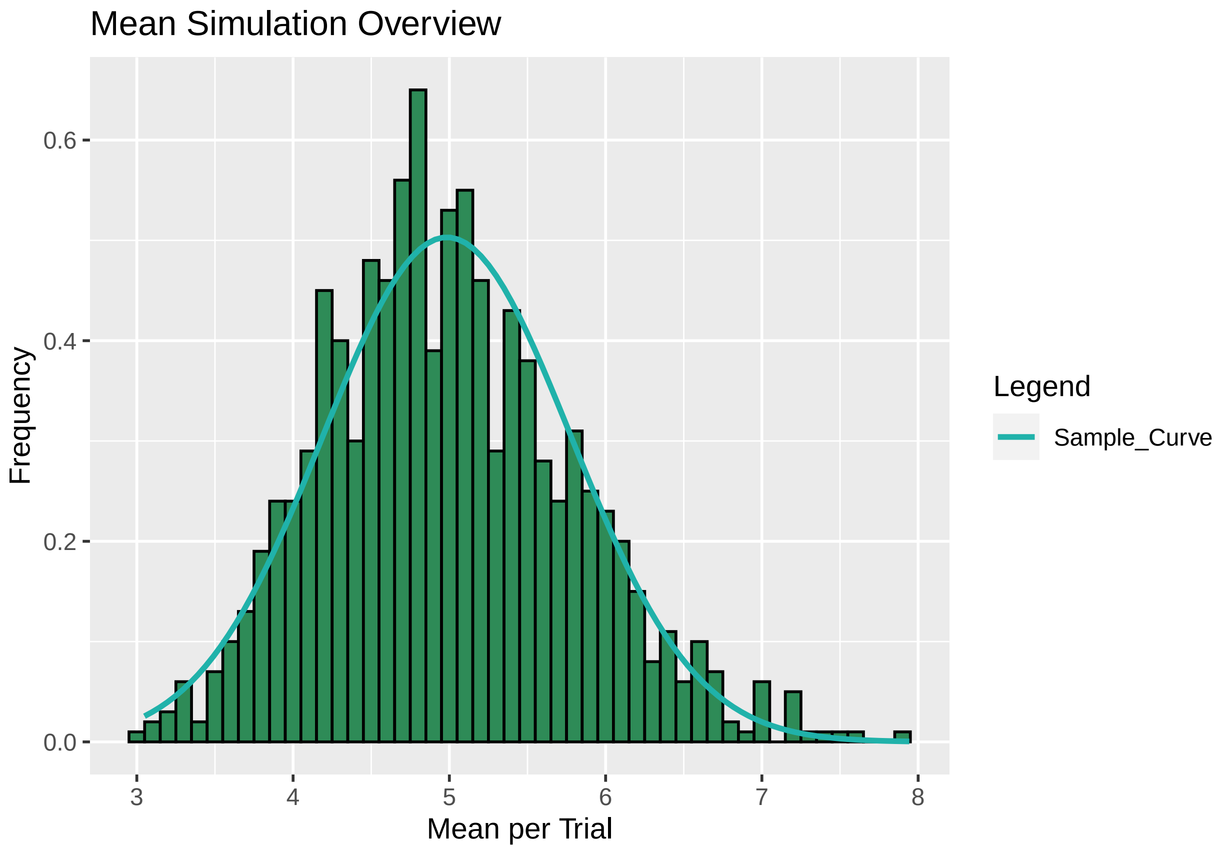
```
## Sample variance is 0.6288244 and theoretical variance is 0.625 with difference of 0.003824422
```

Similar to the mean, while the variance are not exactly equal, they are approximately the same.

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

Since we have established that the sample mean and variance is approximately the same as the theoretical mean and variance, we can fit our original histogram with a normal curve that follows the sample mean and variance. In the below figure, we can see that the graph does indeed seem close to a normal distribution.

```
ggplot(as.data.frame(matMean), aes(x=matMean))+
  geom_histogram(aes(y =..density..),
                 binwidth = 0.1,
                 fill=I("seagreen"),
                 col=I("black"))+
  labs(x = "Mean per Trial",
  y = "Frequency",
  title = "Mean Simulation Overview")+
  stat_function(fun = dnorm, args = list(mean=sampleMean, sd=sqrt(sampleVar)),aes(color="Sample_Curve")
  scale_color_manual(name = "Legend", values = c(Sample_Curve = "lightseagreen"))
```



Next, using a 95% confidence interval, we check the distance between the theoretical and sample observations. Recall that the calcualtion required for this is $CI = mean \pm 1.96 * SD$.

```
theoInt <- theoMean + c(-1,1)*1.96*sqrt(theoVar/n)
theoInt
```

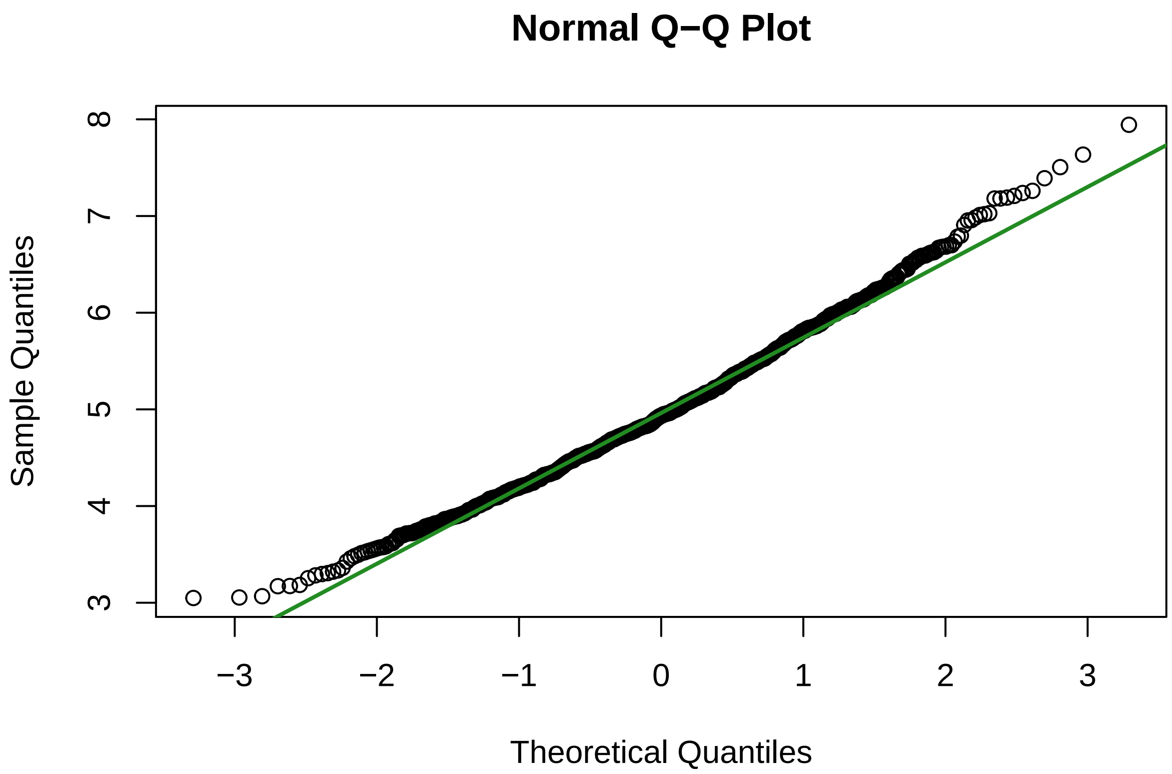
```
## [1] 4.755 5.245
```

```
sampleInt <- sampleMean + c(-1,1)*1.96*sd(matMean)/sqrt(n)
sampleInt
```

```
## [1] 4.738377 5.229874
```

There are very small discrepancies between the intervals. Finally, with the below comparison between the sample and theoretical quantiles, we can conclude that the distribution is close to normal and would be closer as n increase as per the Central Limit Theorem.

```
qqnorm(matMean)
qqline(matMean, col="forestgreen",lwd=2)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles