Statistical Inference Course Project Part 1

Mark Nicholls

24/04/2021

Statistical Inference Course Project Part 1

Exponential Distribution

The aim of this analysis is to investigate the exponential distribution in R through sampling simulation. I will collect 1000 samples of 40 random exponentials with rate 0.2, and take the mean and standard deviations. Then i will compare this to the theoretical mean, theoretical variance and show how the distribution of the means is approximately normal.

First lets set up the lambda (rate) and calculate the theoretical mean and variance.

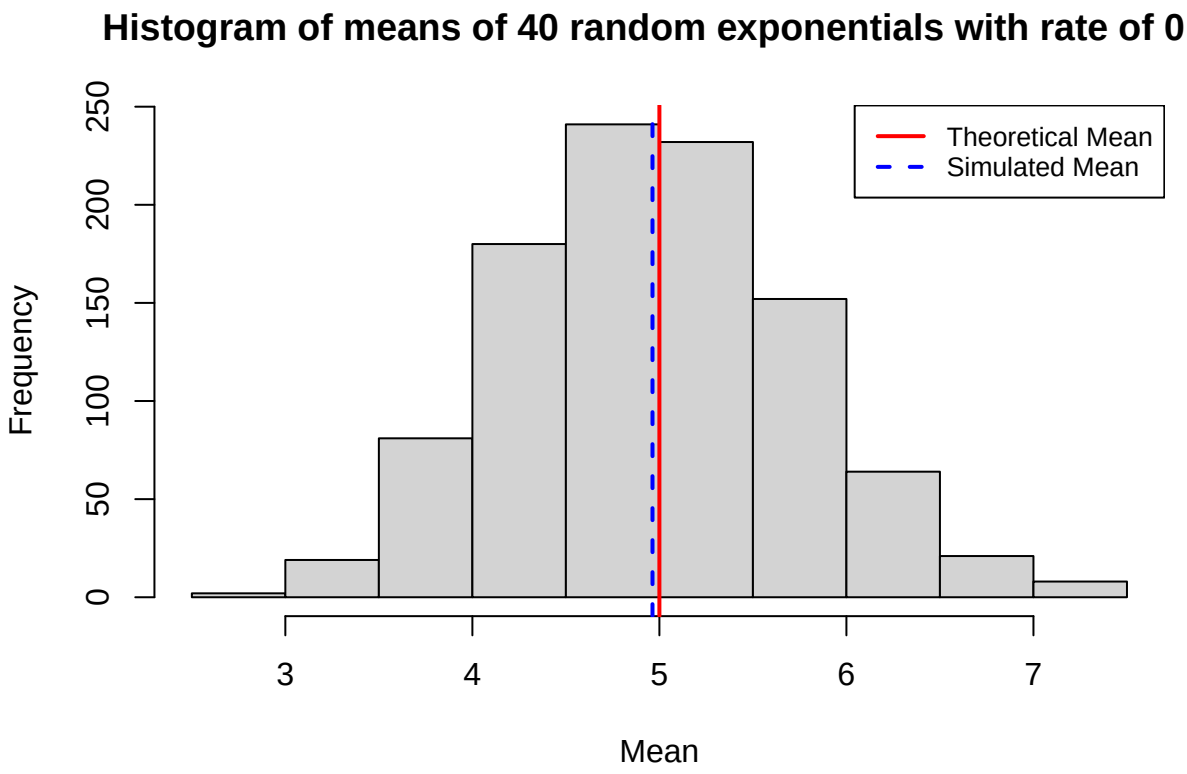
```
# set rate parameter
lambda = 0.2
# calculate theoretical mean / sd
theo_mean =  1/lambda
theo_sd = 1/lambda
```

Simulations

Sample Mean versus Theoretical Mean

In the first simulation I collect 1000 means of samples of 40 random exponential with rate 0.2. and plot this in a histogram. Then add vertical lines at the theoretical mean (1/lambda) and mean of the simulations you can see they are close.

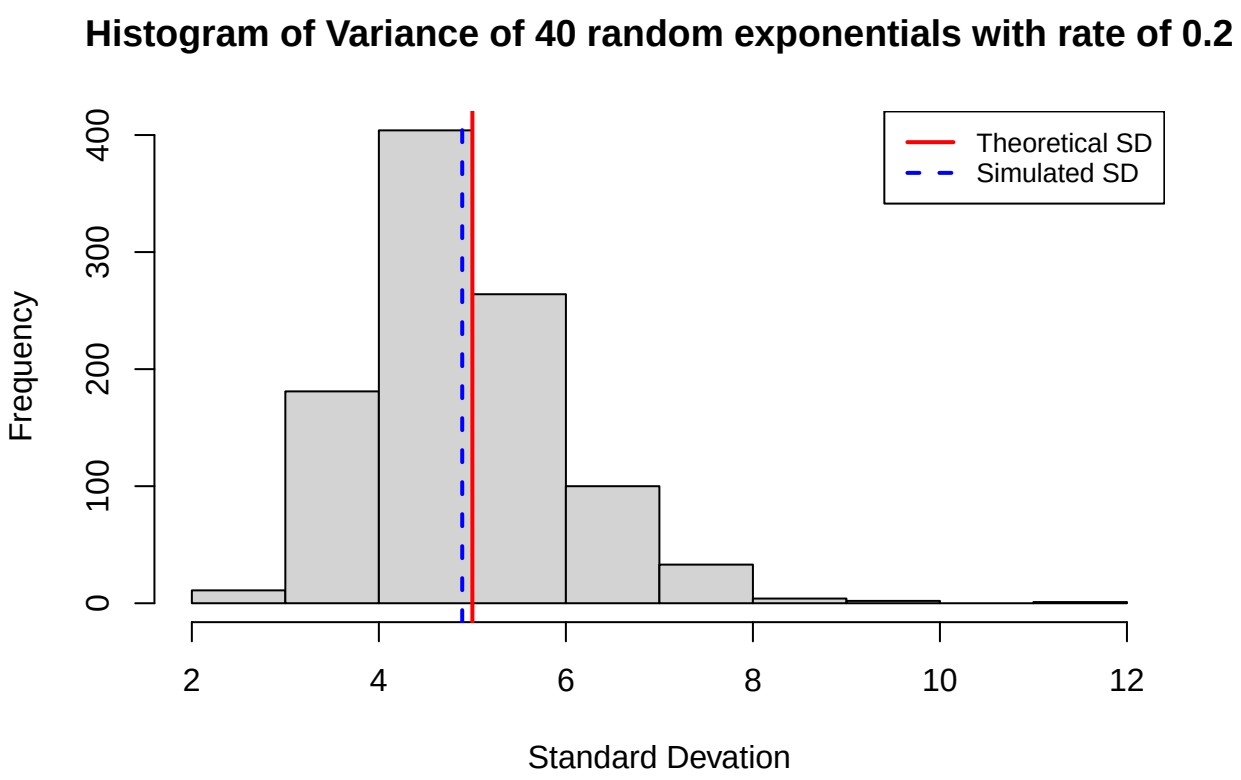
```
# simulate 1000 averages of 40 random exponentials
means = NULL
for (i in 1:1000) means = c(means, mean(rexp(40, lambda)))
# plot the histogram
hist(means, xlab = "Mean",
     main = "Histogram of means of 40 random exponentials with rate of 0.2")
# add vertical line for theoretical and simulated mean
abline(v = theo_mean, lwd = 2, col = "red")
abline(v = mean(means), lwd = 2, col = "blue", lty = 2)
# add legend
legend("topright", legend = c("Theoretical Mean", "Simulated Mean"),
       lty = c(1, 2), col = c("red", "blue"), lwd = 2, cex = 0.8)
```



Sample Variance vs Theoretical Variance

In the second simulation I collect 1000 standard deviations of samples of 40 random exponentials with rate of 0.2. I plot these in a histogram and then add vertical lines at the theoretical standard deviation (1/lambda) and the average of the sample standard deviations, you can see these lines are close, but not as close as the for the means, this is due to the distribution of the samples variances being right skewed.

```
#caluclate SD of 1000 samples of 40 random exponentials
sds = NULL
for (i in 1:1000) sds = c(sds, sd(rexp(40, lambda)))
hist(sds , xlab = "Standard Devation",
     main = "Histogram of Variance of 40 random exponentials with rate of 0.2")
#add vertical lines for theoretical and simulated
abline(v = theo_sd, lwd = 2, lty = 1, col = "red")
abline(v = mean(sds), lwd = 2, lty = 2, col = "blue")
# add legend
legend("topright", legend = c("Theoretical SD", "Simulated SD"),  lty = c(1, 2),
       col = c("red", "blue"), lwd = 2, cex = 0.8)
```

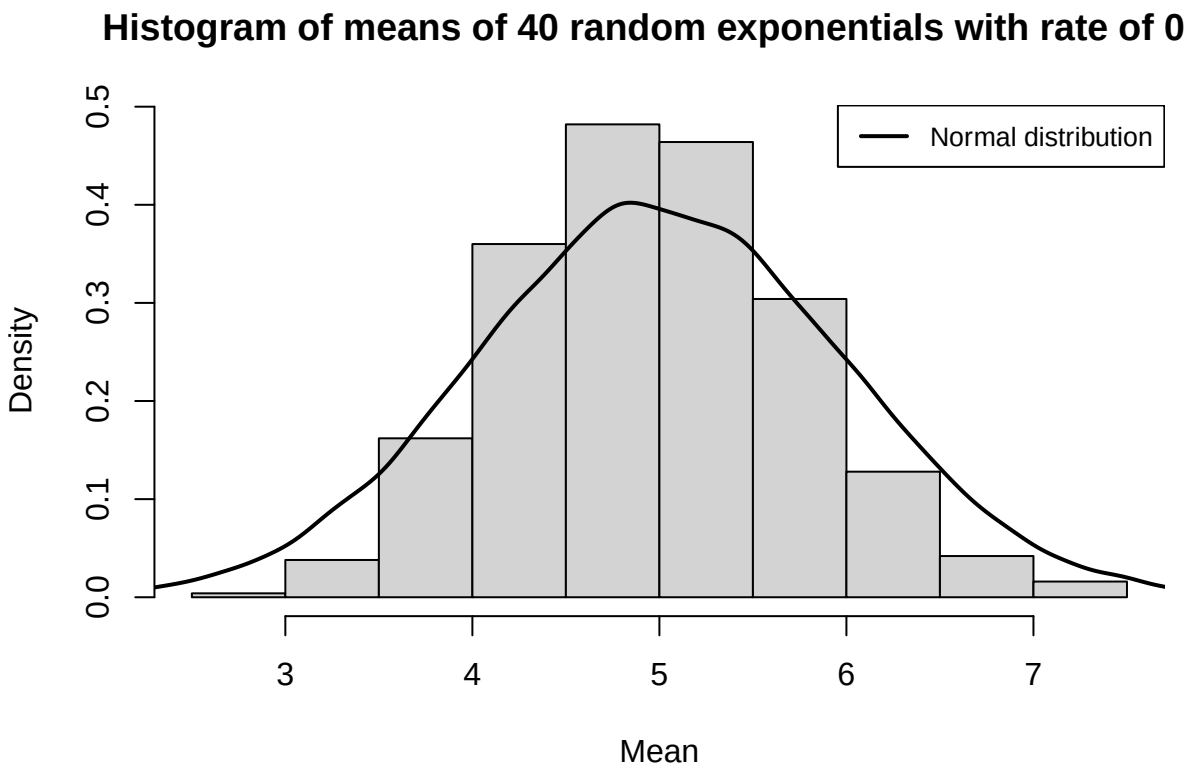


Show that the Distribution is approximately normal

Here I show that the means of the sample generated for the first part are approximately normally distributed. First i generate 40000 random normals with a mean that of the theoretical mean (1/lambda) and a standard deviation of 1. I then over lay this distribution on the histogram from the first part, and you can see that the distribution of means is approximately normal.

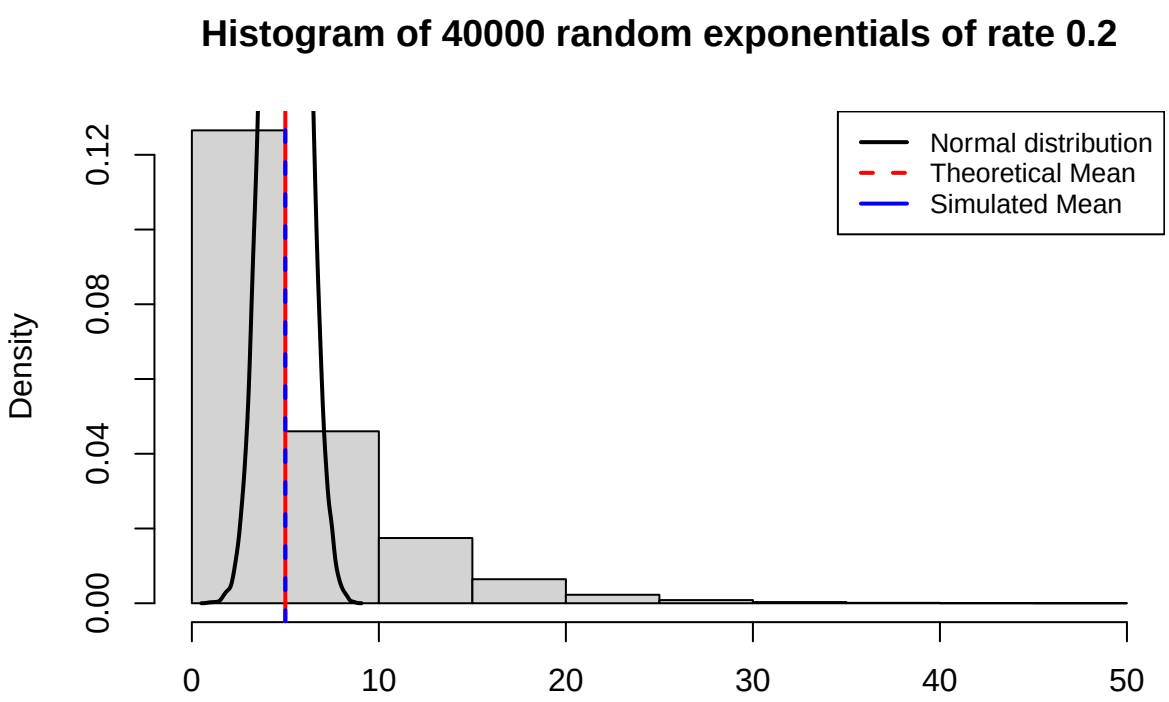
NOTE: the scale for this histogram is density, this is necessary to overlay the normal density with the density() function

```
#simulate 10,000 random normals with mean of theoretical mean, sd = 1
rnorms <- rnorm(40000, mean = theo_mean, sd = 1)
#histgram of means
hist(means, probability = TRUE, xlab = "Mean",
     main = "Histogram of means of 40 random exponentials with rate of 0.2")
#overlay an approximate normal distribution
lines(density(rnorms),  lwd = 2, col = "black")
#add legend
legend("topright", legend = c("Normal distribution"), lty = c(1),
       col = c("black"), lwd = 2, cex = 0.8)
```



Lastly I compare the process of gathering 1000 simulations of 40 random exponentials with one sample of 40000. I plot the histogram of the large sample, overlay a normal distribution (with mean 1/lambda) and plot vertical lines at the theoretic and simulated mean. Here the histogram shows the distribution of one large sample is not normal, however the simulated mean and theoretical mean are again very close

```
#generate large number of random exponentials
large_exp <- rexp(40000, lambda)
#plot histogram of them
hist(large_exp, probability = TRUE, xlab = NULL,
     main = "Histogram of 40000 random exponentials of rate 0.2")
#overlay the means and normal distribution
abline(v = theo_mean, lwd = 2, col = "red")
abline(v = mean(large_exp), lwd = 2, lty = 2, col ="blue")
lines(density(rnorms),  lwd = 2, col = "black")
#add legend
legend("topright",  legend = c("Normal distribution", "Theoretical Mean",
       "Simulated Mean"), lty = c(1, 2, 1), col = c("black", "red", "blue"),
       lwd = 2, cex = 0.8)
```


Histogram of 40000 random exponentials of rate 0.2
Density
0.00
0.04
0.08
0.12
0
10
20
30
40
50
Normal distribution
Theoretical Mean
Simulated Mean