Statistical Inference: Course project 1

Rachel Smith

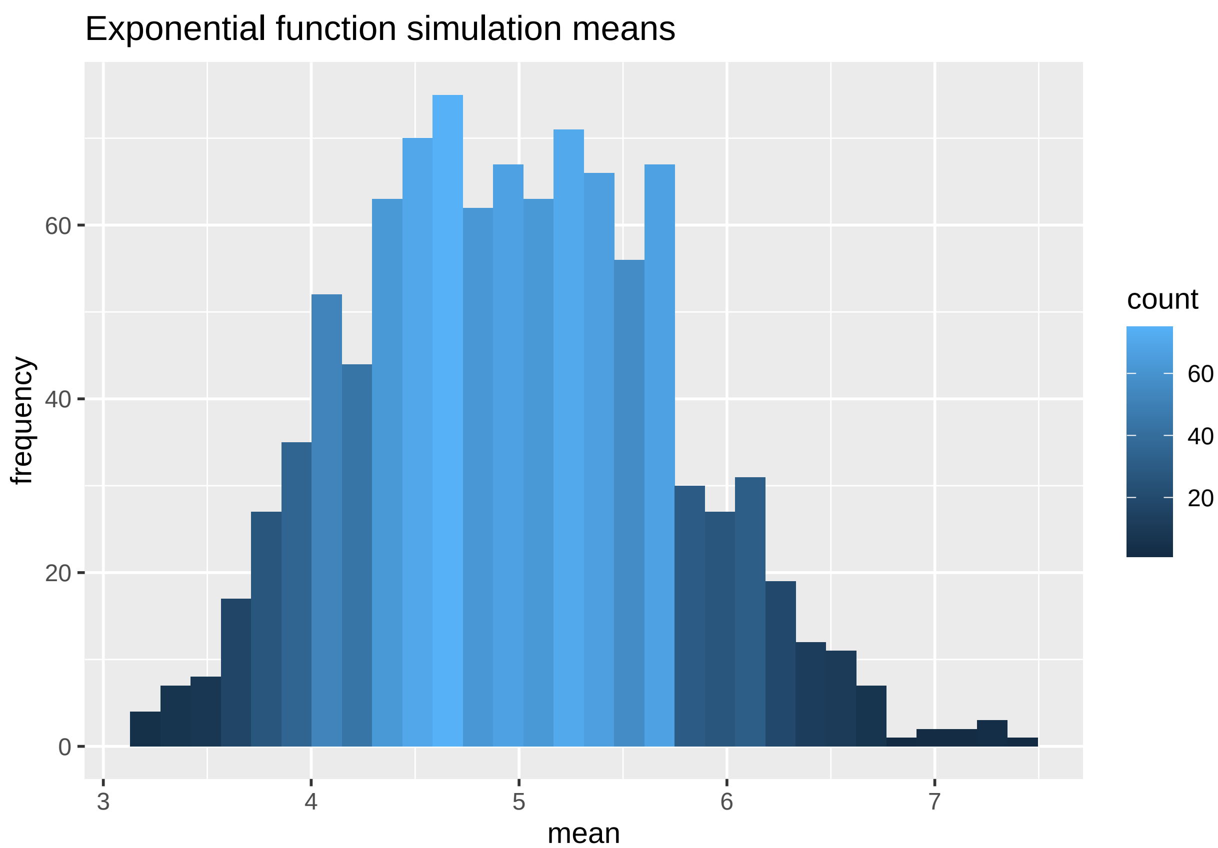
August 4, 2020

Overview

This document is a completion of part 1 for the Coursera Statistical Inference Course Final Project. Here, we investigate the exponential distribution in R and compare it with the Central Limit Theorem. For the following simulations, the rate parameter (lambda) is 0.2, and we investigate the distribution of averages of 40 exponentials over 1000 simulations.

Simulations

```r
# set seed for reproducibility

       set.seed(1234)

# set parameters according to instructions

       n <- 40
       lambda <- 0.2
       sims <- 1000

# run simulations

       m_sims <- replicate(sims, rexp(n, lambda))

# calculate sample mean

       m_means <- apply(m_sims, 2, mean)

# plot histogram

       data.frame(m_means) %>% ggplot() +

              geom_histogram(aes(x = m_means, y = ..count.., fill = ..count..)) +
              xlab("mean") +
              ylab("frequency") +
              ggtitle("Exponential function simulation means")
```



Sample mean vs. theoretical mean

The mean of an exponential distribution is 1/lambda. Therefore, in this instance, the theoretical mean = 1/0.2 = 5.

```
# compare sample and theoretical means

      data.frame(sample_mean = mean(m_means), theoretical_mean = 1/lambda)
```

```
##   sample_mean theoretical_mean
## 1    4.974239                5
```

```
# 95% confidence interval on sample mean

      t.test(m_means)[4]
```

```
## $conf.int
## [1] 4.927362 5.021116
## attr(,"conf.level")
## [1] 0.95
```

The sample mean is very close to the theoretical mean, with a 95% confidence interval between 4.927362 and 5.021116.

Sample variance vs. theoretical variance

The theoretical variance of an exponential distribution is ((1/lambda)^2)/n, while the standard deviation is (1/lambda)/sqrt(n).

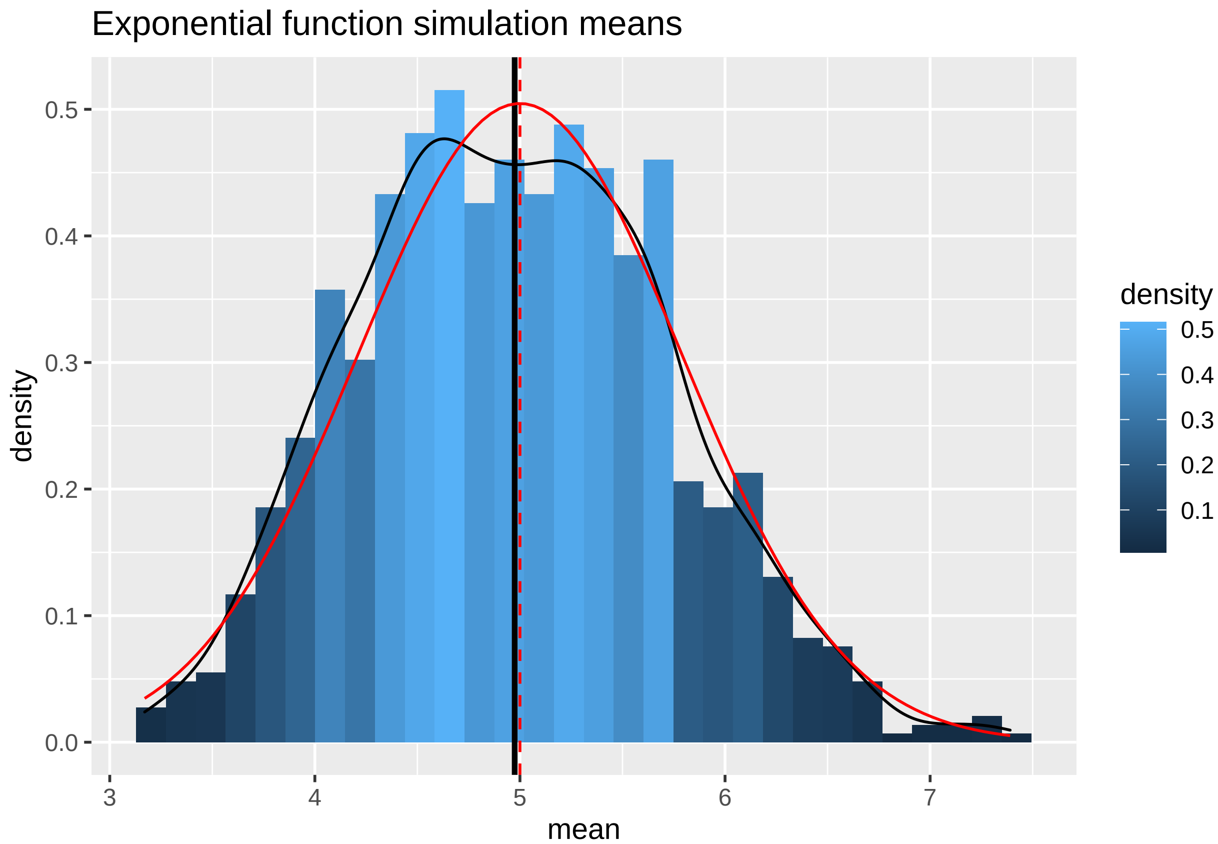
```
df_var <- data.frame(sample = c(var(m_means), sd(m_means)),
                     theoretical = c( (((1/lambda)^2)/n), ((1/lambda)/sqrt(n))))

row.names(df_var) <- c("variance", "stdev")

df_var
```

```
##             sample theoretical
## variance 0.5706551   0.6250000
## stdev    0.7554171   0.7905694
```

As shown by these data, the sample variance closely follows the theoretical.

Distribution

```
data.frame(m_means) %>% ggplot(aes(x = m_means)) +

       geom_histogram(aes(y = ..density.., fill = ..density..)) + # plot sample data
       geom_vline(aes(xintercept = mean(m_means)), lwd = 1, col = "black") + # sample mean
       geom_density() + # sample distribution density
       stat_function(fun = dnorm,
                     args = list( mean = 1/lambda, sd = df_var[2,2]),
                     color = "red") + # theoretical distribution density
       geom_vline(xintercept = 1/lambda, color = "red", linetype = "dashed") + # theoretical mean

       xlab("mean") +
       ylab("density") +
       ggtitle("Exponential function simulation means")
```



As shown in this plot, the distribution of means of our simulations closely follows a normal distribution, due to the Central Limit Theorem.

For these simulations, we assume sampling without replacement.