Statistical Inference Course Project Part 1

JSchneyer

1/3/2021

Introduction

The purpose of this project is to investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). We will simulate data from the exponential distribution, calculate the sample mean and variance and compare to the theoretical values. Finally, we will plot the distribution of sample means to show that it is approximately normal.

Part 1: Simulation Exercise with the Exponential Distribution

The exponential function contains one (1) parameter, lambda, known as the "rate parameter." The mean and standard deviation of the exponential distribution is 1/lambda.

In this part we will:

- sample (n = 40) from the exponential distribution 1,000 times and compare the sample mean to the theoretical mean of the distribution
- measure the sample variance and compare to the theoretical variance of the distribution
- Show that the distribution of the sample mean is approximately normal.

The first step is to define out seed so that our analysis and results can be reproduced.

```
set.seed(1234)
```

Next, we will define the distribution parameters (lambda), sample size (n) and number of simulations (sims). We will also calculate the theoretical mean (mu) and standard deviation (sd).

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

mu <- 1/lambda
sd <- 1/lambda
```

With the parameters of the theoretical distribution set, we can simulate samples from the distribution. We will sample 40 variables from the distribution 1,000 times, and create a matrix (named sample.distr) with 40 columns and 1,000 rows. Each row is a simulation and the columns are the sample variables.

```
sample.distr <- matrix(data = rexp(n*sims,lambda), nrow = sims)
```

Sample mean vs. theoretical mean Now that we have our simulated data, we will calculate the mean of the 40 variables for each of the 1,000 simulations. This vector will be called sample.means and contain 1,000 values (mean of each simulation).

```
sample.means <- apply(sample.distr, 1, mean)
```

We know that the theoretical mean of the exponential distribution is 1/lambda. We calculated this value (mu) earlier. If we compare that theoretical mean to our mean value of our sample means, we see that the values are very close.

```
mu; mean(sample.means)
```

```
## [1] 5
```

```
## [1] 4.974239
```

The absolute difference between our theoretical and sample mean is:

```
abs(mu-mean(sample.means))
```

```
## [1] 0.02576123
```

Sample variance vs. theoretical variance Similarly, we can compute the sample mean variance and compare to the theoretical variance. We will calculate the variance of the sample means.

```
sample.mean.var <- var(sample.means)
sample.mean.var
```

```
## [1] 0.5949702
```

Now we will compute the theoretical variance of the exponential distribution. The theoretical variance of the exponential distribution is equal to 1 divided by lambda times the square root of the number of samples (n = 40) squared.

```
theoretical.var <- 1/((lambda*sqrt(n)))^2
```

When we compare our sample mean variance to that of the theoretical variance of the exponential distribution, see that the values are close.

```
theoretical.var; sample.mean.var
```

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

```
## [1] 0.5949702
```

The absolute difference between our theoretical and sample mean variance is:

```r
abs(theoretical.var-sample.mean.var)
```

```
## [1] 0.03002984
```

Distribution of sample means

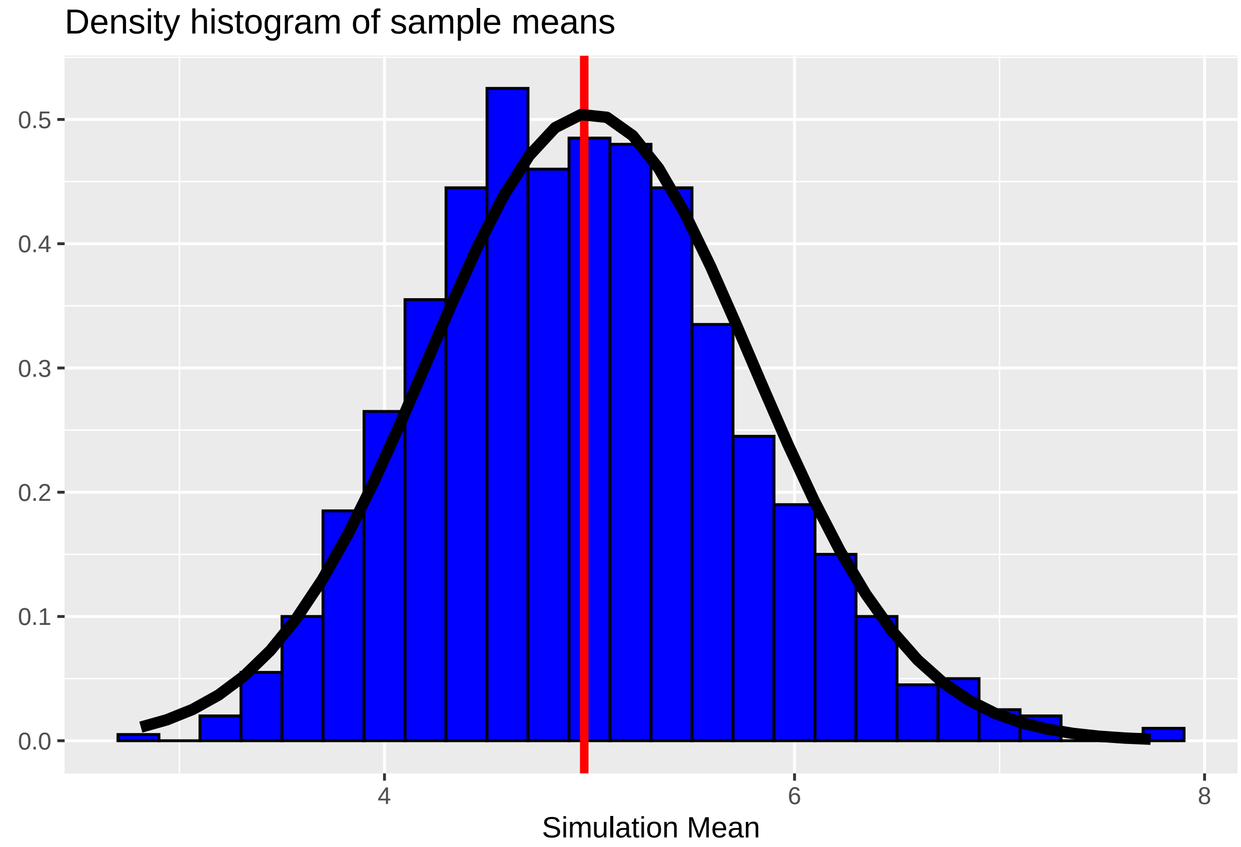
We know, by the CLT, that the distribution of sample means should be approximately normal, with a mean similar to the theoretical mean of the exponential distribution. We proved that the sample and theoretical means are close already.

A density histogram of the sample means is presented below. The vertical red line represents the mean of the distribution of sample means (4.974239). We can see that the distribution is approximately normal with a mean near 5 (the theoretical mean, mu, or 1/lambda). In addition, we will overlay a normal distribution curve using the theoretical mean and standard deviation parameters to show that this distribution is approximately normal.

We will use the ggplot2 package to plot.

```r
library(ggplot2)

ggplot(data.frame(y = sample.means), aes (x = y)) +
  geom_histogram(aes(y =..density..), binwidth = 0.2, fill = "blue", color = "black") +
  geom_vline(xintercept = mean(sample.means), color = "red", size = 1.5) +
  stat_function(fun = dnorm, n = 40, args = list(mean = 1/lambda, sd = 1/(lambda*sqrt(n))), size = 2) +
  labs(title = "Density histogram of sample means",
       x = "Simulation Mean",
       y = NULL)
```


Density histogram of sample means
0.5
0.4
0.3
0.2
0.1
0.0
4
6
8
Simulation Mean