Statistical Inference Project - Simulation

Bowen Zhang

6/4/2020

Overview

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.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Simulation

Defining some parameters

```
set.seed(123) #setting the seed for reproducibility
n <- 40
nosim <- 1000
lamba <- 0.2
```

Perform our simulation of 1000 by creating a matrix from the 40 samples drawn from the distribution

```
sim_matrix <- matrix(rexp(n*nosim, rate = lamba), nosim, n)
```

Calculate the mean across 40 values for each of the 1000 simulations

```
sim_mean <- rowMeans(sim_matrix)
```

Comparing Sample vs Theoretical (Means)

Now we can compare the mean of the means of simulations we obtained with the theoretical means which is 1/lamba.

Sample Mean:

This is the average sample mean of 1000 simulations of 40 random samples from an exponential distribution

```
samp_mean <- mean(sim_mean)
samp_mean
```

```
## [1] 5.011911
```

Theoretical Mean:

This is the theoretical mean of an exponential distribution (which is 1/lamba)

```r
theo_mean <- 1/lamba
theo_mean
```

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

Difference of means

This is the total difference between the means.

```r
abs(samp_mean - theo_mean)
```

```
## [1] 0.01191128
```

Findings: The total difference is very small. Sample mean is centered around 5.0119113 while the theoretical mean is centered around 5. The difference between those are very small. Thus the simulated and theoretical (expected) means are pretty close.

Comparing Sample vs Theoretical (Variance)

Sample Variance:

This is the variance of the 1000 simulations of 40 random samples from an exponential distribution

```r
samp_var <- var(sim_mean)
samp_var
```

```
## [1] 0.6088292
```

Theoretical Variance:

This is the theoretical variance of the distribution which is the standard deviation 1/lamba squared and divided by the sample size

```r
theo_var <- (1/lamba)^2 / n
theo_var
```

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

Difference of variances

This is the total difference between the variances

```r
abs(samp_var - theo_var)
```

```
## [1] 0.01617077
```

Findings: Again, the difference in variances is very small. Similar to the means, the sample variance is very close to the theoretical (expected) variance.

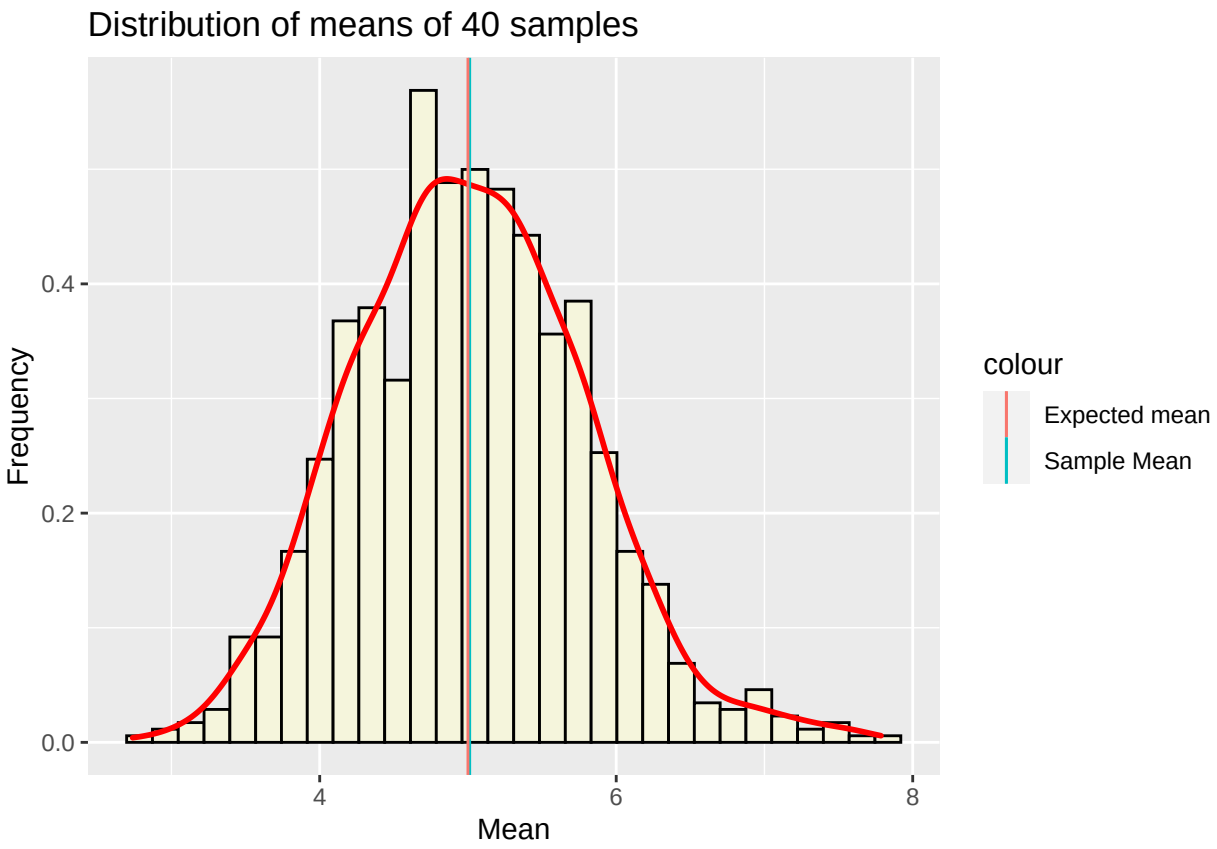
Distribution

Here, we will show that the sample distribution is approximately normal.

```r
g <- ggplot(data.frame(sim_mean), aes(x=sim_mean))

#plot the distribution
g + geom_histogram(aes(y=..density..), colour = "black", fill = "beige") +
        geom_density(colour = "red", size = 1) +
        labs(title = "Distribution of means of 40 samples", x = "Mean", y = "Frequency") +
        geom_vline(aes(xintercept = samp_mean, color = "Sample Mean")) +
        geom_vline(aes(xintercept = theo_mean, color = "Expected mean"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The plot shows a close to normal distribution (bell curve). The theoretical and sample mean indicated by the vertical lines are so close that they basically overlap.

Sample Confidence Interval:

```
round(samp_mean + c(-1,1)*1.96*sqrt(samp_var)/sqrt(n), 2)
```

```
## [1] 4.77 5.25
```

Theoretical Confidence Interval:

```
round(theo_mean + c(-1,1)*1.96*sqrt(theo_var)/sqrt(n), 2)
```

```
## [1] 4.76 5.24
```

Findings: The sample confidence interval and the theoretical (expected) confidence interval at 95% are very close to each other.

Conclusion

From our investigation, we find that the simulated sample distribution does in fact have similar means, variance, and confidence intervals. This demonstrates the Central Limit Theorem in application. Our graph shows that the distribution is approximately normal.