Statistical Inference Project, Part 1

Nannette Spear

July 23, 2018

Overview

Part 1 of the course project involves applying the Central Limit Theorem to exponential distribution in R. The distribution of averages of 40 exponentials, simulated 1,000 times will be investigated.

- Lambda is 0.2
- Theoretical Mean is 1/0.2
- Theoretical Standard Deviation is also 1/0.2
- Observed data substitutes for the population
- Random sampling (simulations) corresponds to the observations

Load knitr package

```
## Warning: package 'knitr' was built under R version 3.4.4
```

Set the Working Directory

Exponential Distribution of 40 exponentials

```
x <- rexp(40, 0.20)
n <- length(x)
```

1000 Simulations

```
B <- 1000
resamples <- matrix(sample(x, n*B, replace = TRUE),B, n)
means <- apply(resamples, 1, mean)
```

Theoretical Mean of the distribution

```
tmean <- 1/0.20
tmean
```

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

Estimated Mean of all samples

```
estmean <- mean(means)
estmean
```

```
## [1] 5.03065
```

The theoretical and estimated mean are very close to each other

```
meanvar <- estmean - tmean
meanvar
```

```
## [1] 0.03064965
```

The Theroetical Standard Deviation is = 1/lambda/sqrt(n)

```
tsd <- (1/0.20)/sqrt(n)
tsd
```

```
## [1] 0.7905694
```

The variance is the standard deviation squared

```
tvar <- tsd^2
tvar
```

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

The estimated standard deviation

```
estsd <- sd(means)
estsd
```

```
## [1] 0.8247483
```

The estimated variance is the standard deviation squared

```
estvar <- estsd^2
estvar
```

```
## [1] 0.6802097
```

The theoretical and estimated variance are also very close to each other

```
varvar <- estvar - tvar
varvar
```

```
## [1] 0.05520971
```

95% Confidence Interval

```
CI <- quantile(means, c(0.025, 0.975))
CI
```

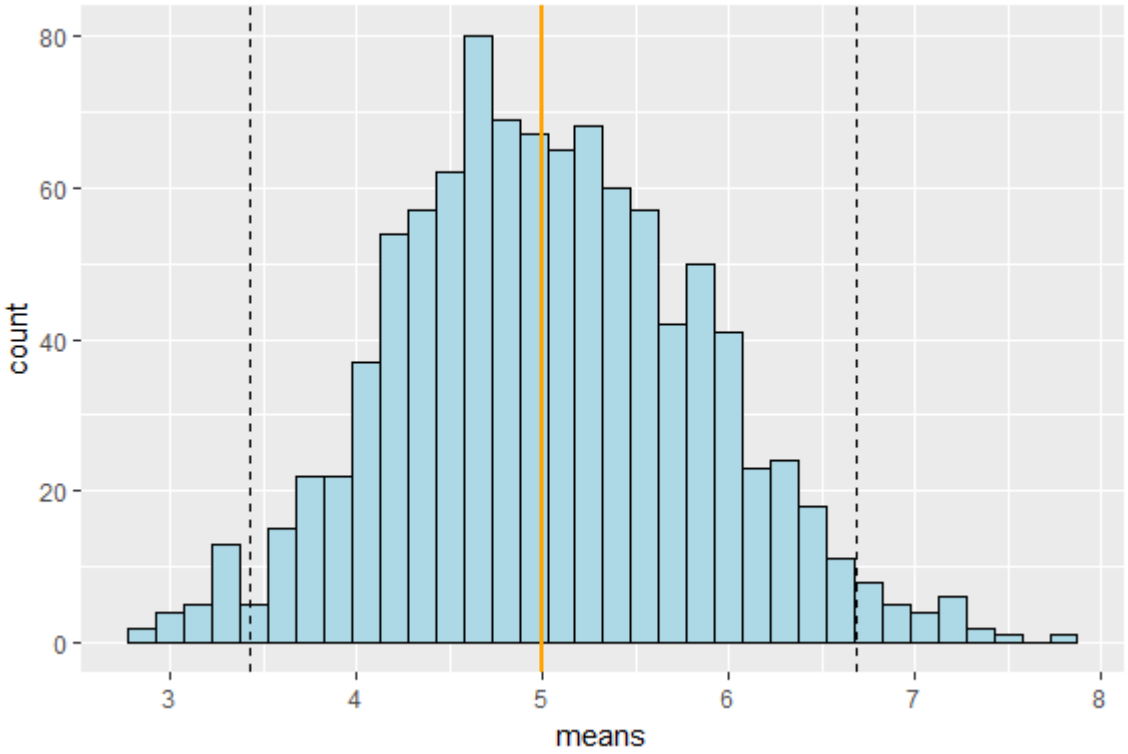
```
##     2.5%    97.5%
## 3.428833 6.684145
```

This plot is an estimate of the sampling distribution of the mean

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
g <- ggplot(data.frame(means = means), aes(x=means))
g <- g + geom_histogram(color = "black", fill = "lightblue", binwidth = 0.15)
g <- g + geom_vline(xintercept = 5, size = 1, color = "orange")
g <- g + geom_vline(xintercept = CI, size = 0.5, color = "black", linetype ="dashed")
g
```



Conclusion

- The histogram represents the simulation means distributed over a normal distribution.
- The orange vertical line at mean 5, represents the theoretical mean.
- The area between the two dashed lines represent where the estimated or sample means will fall, 95% of the time.