Statistical Inference Course Project, Part 1

Cliff Hayes

February 28, 2016

Overview

In this report, the properties of a distribution of the mean of 40 exponentials will be illustrated via simulation and associated explanatory text. This exponential distribution will be investigated in R and compared with the Central Limit Theorem.

Simulations

The exponential distribution will be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of this exponential distribution will be 1/lambda and the standard deviation will also be 1/lambda. For all of the simulations, lambda will equal 0.2.

The distribution of averages of 40 exponentials will be investigated via one thousand simulations.

1. Sample Mean versus Theoretical Mean

```
lambda <- 0.2
n <- 40
count <- 1000
set.seed(41)
sim <- matrix(rexp(n*count, rate = lambda), nrow = count, ncol = n)
sim_mean <- apply(sim, 1, mean)
mean(sim_mean)
```

```
## [1] 5.018901
```

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

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

The simulation `sim` contains 40 exponential distributions with rate parameter $\lambda = 0.2$ and each distribution contains 1000 simulated values. The seed is set at 41 to make this simulation reproducible.

The object `sim_mean` is the sample mean ($\bar{X}$) from the simulations. The sample mean of $\bar{X}$ is 5.018901. This is very close to the theoretical mean, $\mu = 1/\lambda = 5$.

2. Sample Variance versus Theoretical Variance

```
var(sim_mean)
```

```
## [1] 0.6283049
```

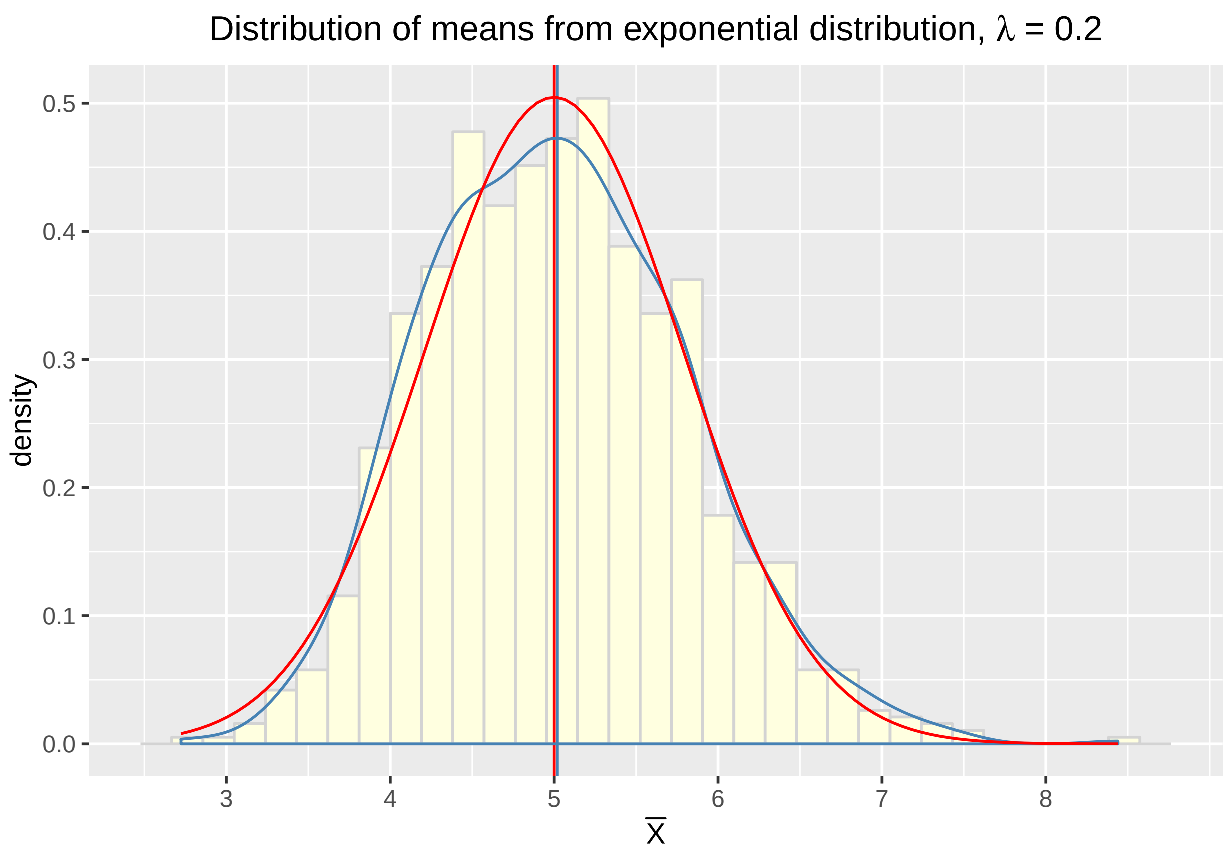
```
sigma <- (1 / lambda^2) / n
sigma
```

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

The sample variance of $\bar{X}$ is 0.628. This is very close the theoretical variance, $\sigma^2 = \frac{1/\lambda^2}{n} = 0.625$.

3. Distribution

```r
library(ggplot2)
ggplot(data.frame(sim_mean), aes (x = sim_mean)) +
    geom_histogram(aes(y = ..density..), color = "lightgray", fill = "lightyellow") +
    geom_vline(xintercept = c(mean(sim_mean), 5),
               color = c("steelblue", "red")) +
    xlab(expression(bar(X))) + geom_density(color = "steelblue") +
    ggtitle(expression(paste("Distribution of means from exponential distribution, ", lambda, " = 0.2")
    scale_x_continuous(breaks = c(seq(0, 8, 1))) +
    stat_function(fun = dnorm, arg = list(mean = 5, sd = sqrt(sigma)),
                  color = "red") +
    theme_gray()
```



The plot above shows the distribution of the means from the exponential distribution. The red density line shows the theoretical $n(5, 0.625)$ distribution. The blue density line shows the estimated distribution based on the simulation. The simulation appears to validate the Central Limit Theorem and is a normal distribution.

For further verification that the simulated distribution is a normal distribution, let's evaluate the confidence intervals for both the simulated distribution and the theoretical distribution:

```r
sim_conf_int <- round (mean(sim_mean) + c(-1,1)*1.96*sd(sim_mean)/sqrt(n),3)
thy_conf_int <- mu + c(-1,1)*1.96*sqrt(sigma)/sqrt(n)
```

The simulated 95% confidence interval is [4.773, 5.265] and the theoretical 95% confidence interval is [4.755, 5.245]. I'm calling it good!