Statistical Inference Project - Part 1

Paulo Viana

In this project you will 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, rate)` where rate is the lambda parameter. The mean of exponential distribution is $1/\lambda$ and its 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.

Preliminary Simulations

Let us start simulating 1000 randomly sampled distributions and then calculate the mean for each.

```r
set.seed(50)
s <- 1000;
n <- 40;
lambda <- 0.2

simulation <- rexp(n * s, rate = lambda) %>% # generate exp dist with 40000 values
    matrix(nrow = s, ncol = n) %>%      # create a 1000 X 40 matrix
    rowMeans %>%                        # take the mean of all rows
    data.frame(values = .) %>%      # create data.frame with means
    tbl_df
```

Mean and Variance Comparison

The values for the mean and the variance of a theoretical exponential distribution with $\lambda = 0.2$ and $n = 40$ values are, respectively,

$$\mu = 1/\lambda = 5$$

and

$$Var = (1/\lambda * 1/\sqrt{n})^2 = 0.625$$

Calculating the values for our simulations and we obtain $\mu_{\bar{X}} = 4.9691419$ and $Var_{\bar{X}} = 0.6188363$.

CLT

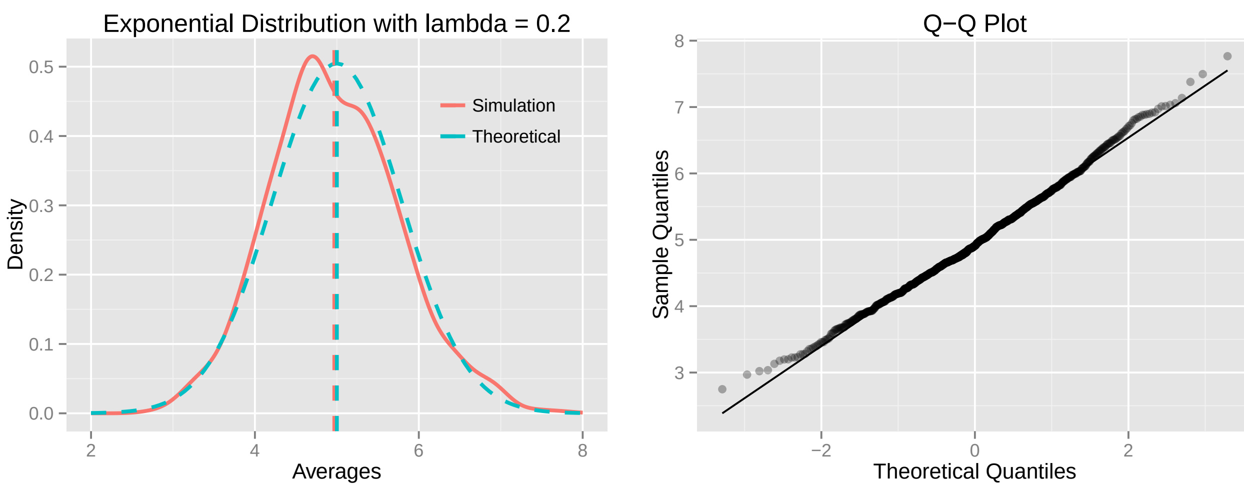
Due to the Central Limit Theorem we expect the distribution of the average of exponentials to follow the normal distribution. Below we plot two figures that support this claim: the first showing both distributions and their respective means, and the second comparing its quantiles.

```r
g1 <- ggplot(simulation, aes(x = values)) +
    stat_density(aes(colour = 'Simulation'), geom = 'line', size = 1) +
    stat_function(data = data.frame(x = c(2, 8)), aes(x = x, colour = 'Theoretical'),
                  geom = 'line', linetype = 'dashed', size = 1,
                  fun = dnorm, args = list(mean = 5, sd = sqrt(0.625))) +
    geom_vline(data = data.frame(x = mean(simulation$values), y = 5) %>% stack,
               aes(xintercept = values, colour = ind),
               linetype = 'dashed', size = 1) +
    scale_color_manual(values = c('Simulation'='#F8766D', 'x'='#F8766D',
```

```
                                    'Theoretical'='#00BFC4', 'y'='#00BFC4'),
                        breaks = c('Simulation', 'Theoretical')) +
    labs(title='Exponential Distribution with lambda = 0.2', x='Averages', y='Density') +
    plot.theme

g2 <- ggplot(simulation$values %>% qqnorm(plot=F) %>% as.data.frame, aes(x = x, y = y)) +
    geom_smooth(method = 'lm', se=F, colour = 'black') +
    geom_point(alpha = 0.3, size = 2.2) +
    labs(title = 'Q-Q Plot', x = 'Theoretical Quantiles', y = 'Sample Quantiles')

suppressMessages(grid.arrange(g1, g2, nrow = 1))
```



Coverage of the Confidence Intervals

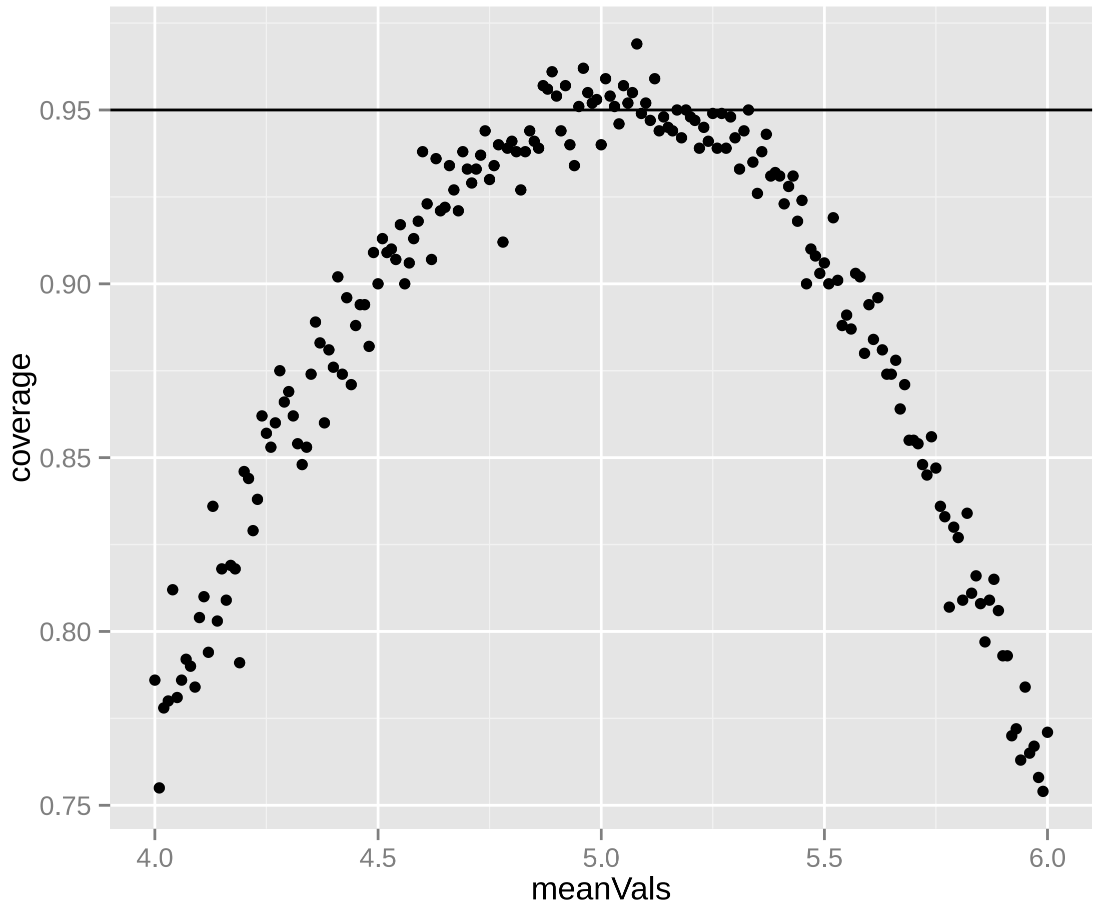
Now let analyze the 95% confidence interval for the distribution, calculated as $1/\lambda = \bar{X} \pm 1.96\frac{S}{\sqrt{n}}$.

```
meanVals <- seq(4, 6, by = 0.01)
coverage <- sapply(meanVals, function(val) {
        mhats <- rexp(n * s, rate = lambda) %>% matrix(nrow = s, ncol = n) %>% rowMeans
        ll <- mhats - qnorm(.975) * sqrt(1/lambda^2/n)
        ul <- mhats + qnorm(.975) * sqrt(1/lambda^2/n)
        mean(ll < val & ul > val)
})
coverage[meanVals == 5]
```

```
## [1] 0.94
```

The plot below shows the result of 200 simulations with various hypothetical values for the mean, and their coverage, or the proportion of the values included between 2 standard deviations (our value of interest). The "true" value for μ, 5, is included within this interval at least 95% of time.

```
qplot(meanVals, coverage) + geom_hline(yintercept = 0.95)
```


0.95
0.90
0.85
0.80
0.75
coverage
4.0
4.5
5.0
5.5
6.0
meanVals