Exponential Distribution and Central Limit Theorem

Thomas Fischer

May 19, 2018

Abstract

This document provides the assignment 'Course Project Part 1' for Coursera's Statistical Inference Class in the Coursera Data Science series. Replication files are available on the author's Github account (https://github.com/tomfischersz).

```
knitr::opts_chunk$set(echo = TRUE, fig.pos= "h", out.extra = '')
```

1. Synopsis

The aim of this report is to investigate the sampling distribution for $\bar{X}_n$, derived as the mean from samples with size n from an exponential distribution. The Central Limit Theorem (CLT) states, that for large n we should except our sampling distribution to be approximately normal with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (also called standard error of the mean).

2. Simulation

The first step is to simulate 1000 random variables each calculated as the mean of $n = 40$ random exponential variables with population parameter $\lambda = 0.2$ (rate). To generate random exponential variables we use the R function rexp(). The resulting 1000 sample means, which themselves are random variables, are stored in a vector (script can be found here).

3. Sample Mean versus Theoretical Mean

The theoretical mean is the population mean of our exponential distribution with $\lambda = 0.2$ and is given as $\mu = \frac{1}{\lambda}$. According to the Law of Large Numbers we also know that the sampling distribution for $\bar{X}_n$ is centered at μ, therefore $\mu_{\bar{X}} = \frac{1}{\lambda}$. We can now compare this theoretical mean with the actual mean of the simulation of 1000 means from samples of size 40 (Code).

The following table shows, that the two calculated values are very close to each other:

Theoretical Mean	Sample Mean	Deviation from Theoretical
5	5.0034469	0.07 %

4. Sample Variance versus Theoretical Variance

We now have a closer look at the variability or spread of our sampling distribution of means. The theoretical variance of the mean of samples with size n from iid random exponential variables is given as $VAR(\bar{X}) = \frac{\sigma^2}{n}$, where σ^2 is the variance of the population we sampled from which is $\frac{1}{\lambda^2}$. We therefore can rewrite $VAR(\bar{X}) = \frac{1}{\lambda^2 n}$. We calculate the theoretical variance and the actual variance (Code) and show it as a table:

Theoretical Variance	Actual Variance	Deviation from Theoretical
0.625	0.6551711	4.83 %

5. Distribution

Finally we want to examine, if the distribution of our simulation of 1000 averages of 40 random exponentials is approximately normal. In figure 1 we plot the histogram of our simulation data together with the theoretical normal distribution we would except according to the CLT. In figure 2 we plot a Q-Q (Quantile-Quantile) Plot, where we compare the theoretical quantiles with observed quantiles.

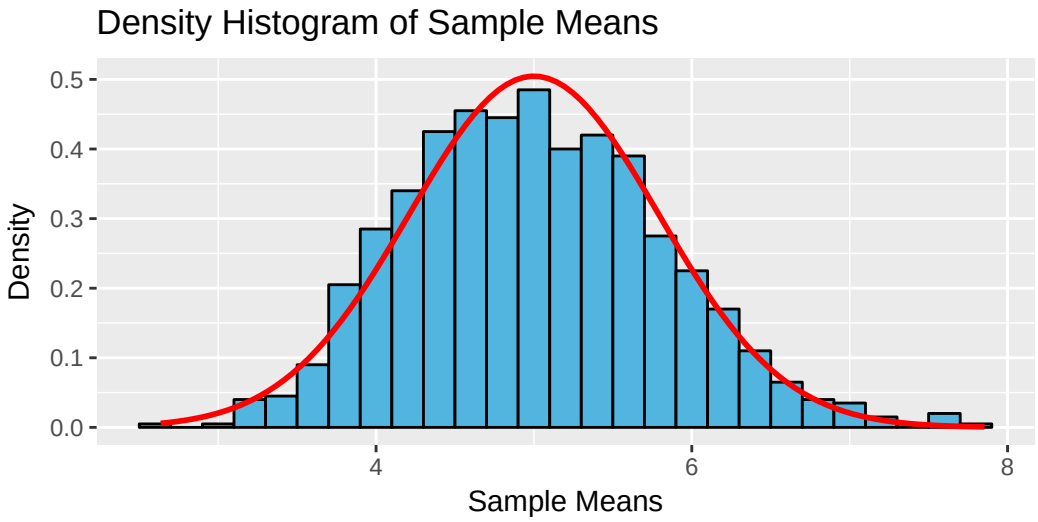

Figure 1: Sampling distribution of the sample mean. The red line shows the standard normal pdf we would except according the CLT.

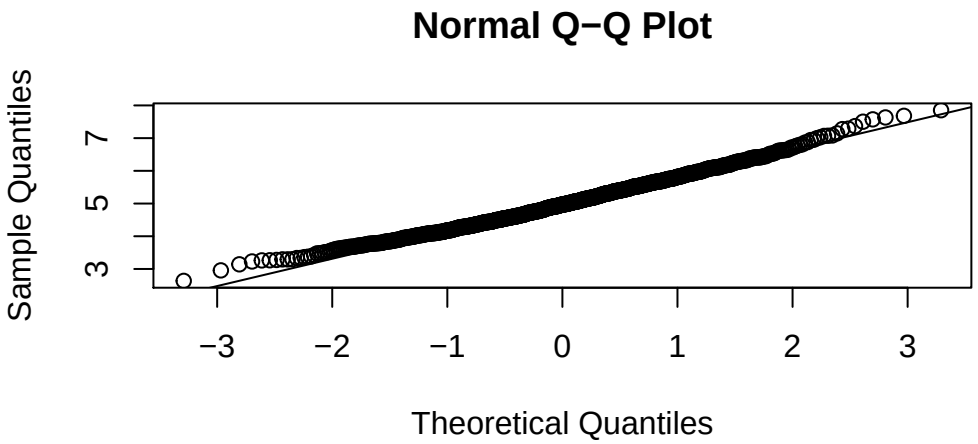

Figure 2: Quantile-Quantile Plot of the observed sampling distribution of the sample mean versus a normal distribution with mean = 5 and standard deviation = 0.79.

6. Conclusion

Both figure 2 , showing a nearly linear plot and figure 1 , showing a density distribution that is quite normal, let us conclude that in fact our simulated distribution of 1000 averages of 40 random exponential variables

follows quite good a normal gaussian distribution. That is what we expected from the Central Limit Theorem.

Appendix

1. Code for creating our simulation:

```
lambda <- 0.2 # rate parameter
n <- 40 # sample size
no_simulations <- 1000 # number of samples
set.seed(1347)
sample_means <- replicate(n = no_simulations, mean(rexp(n, rate = lambda)))
```

2. Calculating the theoretical mean and the mean of the 1000 sample means:

```
theoretical_mu <- 1/lambda
mu_sample_means <- mean(sample_means)
dev_mean <- paste(round((mu_sample_means - theoretical_mu) /
                            theoretical_mu * 100, 2),'%')
```

3. Histogram 1:

```
hist_1 <- ggplot(data = data.frame(sample_means), aes(x=sample_means))
hist_1 <- hist_1 + geom_histogram(color = 'black', binwidth = 0.2, fill = '#51B5E0')
# g <- g + theme_minimal()
hist_1 <- hist_1 + geom_vline(xintercept = theoretical_mu, color = '#F77F00', size = 1)
hist_1 <- hist_1 + geom_vline(xintercept = mu_sample_means, color = 'black', size = 1)
hist_1 <- hist_1 + labs(x = "Sample Means",
                        y = "Count",
                        title = "Histogram of Sample Means")
hist_1 <- hist_1 + annotate(geom = 'text',
                            label = paste('Theoretical mean = ', theoretical_mu,
                                          "\n Actual mean = ", round(mu_sample_means,2)),
                            x = 6, y = 95)
```

4. Calculating theoretical and actual variance:

```
theoretical_var <- 1/(lambda^2 * n)
var_sample_means <- var(sample_means)
dev_variance <- paste(round((var_sample_means - theoretical_var) /
                                theoretical_var * 100, 2),'%')
```

5. Histogram 2:

```
hist_2 <- ggplot(data = data.frame(sample_means), aes(x=sample_means))
hist_2 <- hist_2 + geom_histogram(color = 'black', binwidth = 0.2,
                                  fill = '#51B5E0', aes(y=..density..))
hist_2 <- hist_2 + labs(x = "Sample Means",
                        y = "Density",
                        title = "Density Histogram of Sample Means")
```

```r
hist_2 <- hist_2 + stat_function(fun = dnorm,
                                 args = list(mean = theoretical_mu,
                                             sd = sqrt(theoretical_var)),
                                 colour = "red", size=1)
```

6. Q-Q Plot

```r
qqnorm(sample_means, main = "Normal Q-Q Plot")
qqline(sample_means)
```