Statistical Inference Course Project Part 1: Exponential Distribution and the Central Limit Theorem

Daniela Gracia

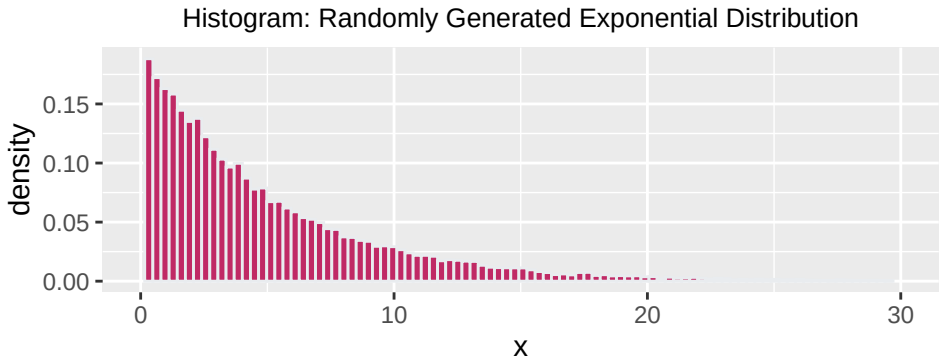
Synopsis

In this project we will investigate the exponential distribution in R. Our goal is to answer the question: **¿Does the distribution of means of 40 exponentials behave as predicted by the Central Limit Theorem (CLT)?**. The structure for the analysis consists of a *(1) Simulation* section, an *(2) Analysis* section and a *(3) Conclusions* section.

Simulations

In this section we will generate all the data necessary for the analysis. First we simulate and exponential distribution of 40000 observations and store it in the rows of a 1000x40 matrix we name "data". We use a fixed rate parameter: $\lambda = 0.2$ for all simulations. We also calculate the mean of each row and store it in the matrix "means". Below is a histogram of the exponential distribution we simulated:

```
# Here we simulate an exponential distribution using rexp()
# define the rate parameter (lambda), the mean, the standard deviation and the number of simulations
lambda <- 0.2
theoretical_mean <- 1/lambda
s <- 1/lambda
N <- 1000
n <- 40
# set a seed to make our data reproducible
set.seed(1111)
# simulate an exponential distribution of n*N observations and store it in a 1000 x 40 matrix
data <- matrix(rexp(n*N, rate=lambda),ncol=n,nrow=N)
```



Analysis

Now, we will use the data we simulated to study whether the distribution of means of 40 exponentials behaves a normal distribution as predicted by the CLT.

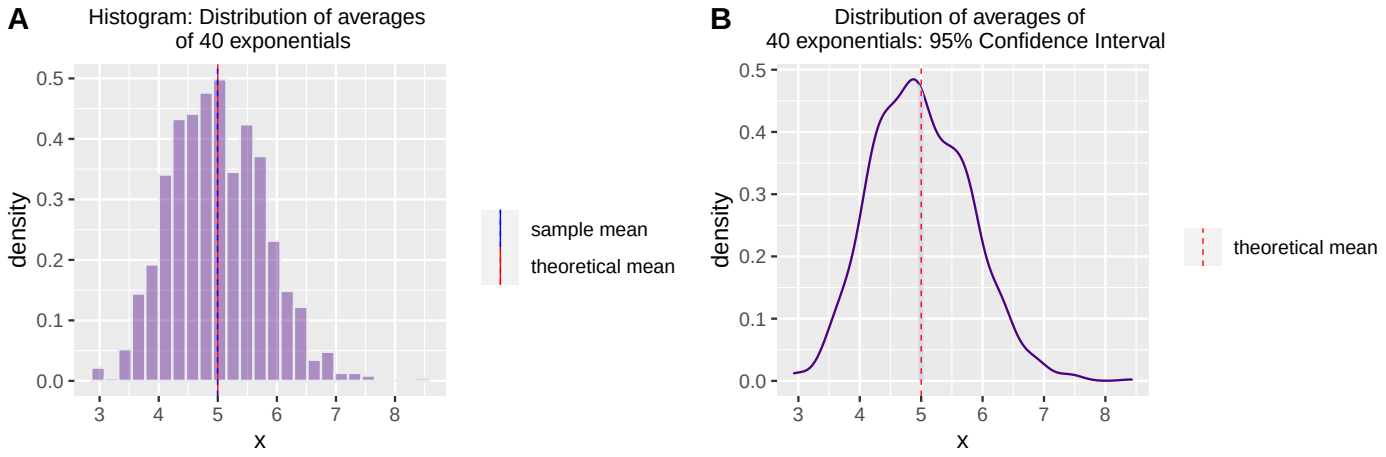
Sample Mean VS Theoretical Mean We will evaluate the sample mean vs the theoretical mean to investigate whether the hypothesized mean is supported by our data. Reversely, we want to know if our sample mean approaches the theoretical mean. In order to do this we will test the following null and alternative hypotheses:

Null Hypothesis: $H_0 : \mu = \mu_0$ and we know that $\mu_0 = \frac{1}{\lambda}$, if we use $\lambda = 0.2$: $H_0 : \mu = 5$.
Alternative Hypothesis: $H_a : \mu \neq \mu_0 \neq 5$. Now we will use the t.test() function to run a two sided test, calculate the 95% confidence interval, p-value, t-statistic and the sample mean.

```
##
##  One Sample t-test
##
## data:  means
## t = -0.20422, df = 999, p-value = 0.8382
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.945711 5.044055
## sample estimates:
## mean of x
##  4.994883
```

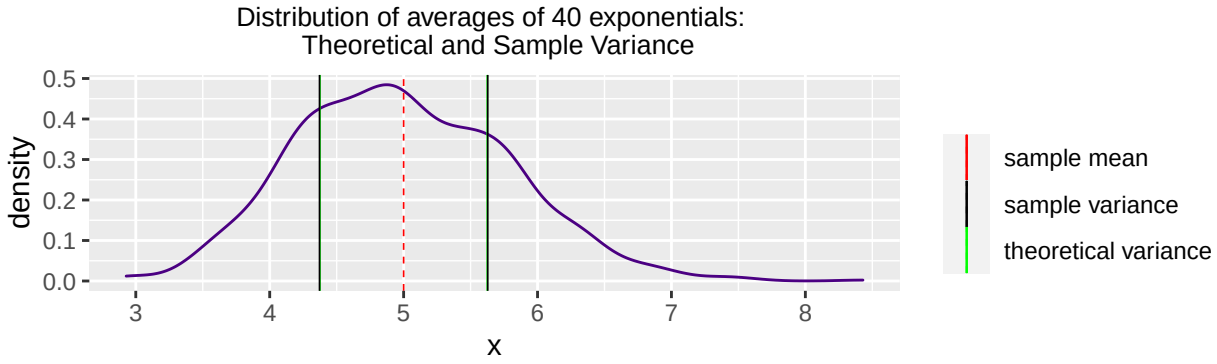
When evaluating the two sided test we can see that the absolute value of the t-statistic is *0.204* which is less than the x_975 t-quantile *1.962*. Therefore we fail to reject the null hypothesis H_0 and accept it with a (higher than alpha) p-value of *0.838* which also supports this. Additionally the theoretical mean is well within the 95% confidence interval which means it is supported.

To illustrate we will display our distribution of averages in a histogram (A). The red vertical line represents the theoretical mean and the blue line is the sample mean. As we can see they are very close to each other and hard to discern. This confirms that the sample mean is approaching the population mean as the sample size increases and that we have 95% certainty that the theoretical mean is indeed the population mean. Additionally we will show that the theoretical mean (as well as the sample mean) are within the confidence interval. We have shaded the 95% confidence region in a darker shade of gray in the density plot below (B).

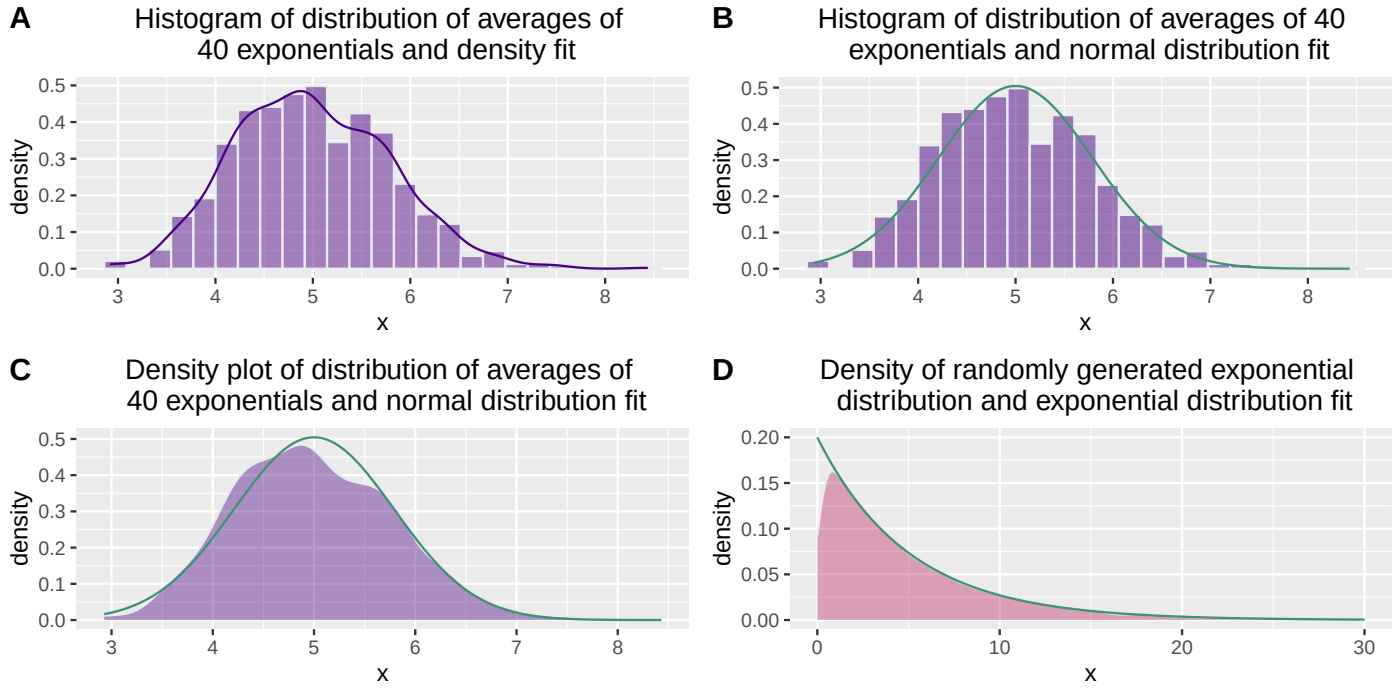


Sample Variance VS Theoretical Variance We will now evaluate the sample variance vs the theoretical variance to continue to investigate whether the distribution of averages of 40 exponentials behaves as predicted by the CLT. We know that the sample variance is an estimate of the population variance and that the distribution of sample variances is centered at the population variance. We also know that the square of the standard error of the mean corresponds to the theoretical variance.

We know that $\sigma_0^2 = \frac{population\ variance}{n}$, the population variance in and exponential distribution is $\frac{1}{\lambda}$ and in our simulation we are using $\lambda = 0.2$, so: $Theoretical\ Variance : \sigma^2 = 0.625$. The sample variances is *0.628* and the theoretical variance is *0.625*. These two values are very close to each other, so the data is consistent with the theory and we conclude that the sample variance is indeed approaching the theoretical variance as n increases. In the plot below both variances are displayed and as it happened with the mean they are very close and difficult to discern from each other:



Distribution In order to prove that the distribution is approximately normal we are going to compare it to a normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 0.625$. We will also pay attention to the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.



Plot A: Histogram and density plot of the distribution of means of 40 exponentials with $\lambda = 0.2$ that we generated.
Plot B: Histogram of the distribution of means of 40 exponentials with $\lambda = 0.2$ that we generated and normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 0.625$.
Plot C: Density plot of the distribution of means of 40 exponentials with $\lambda = 0.2$ that we generated and normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 0.625$.
Plot D: Density plot of the random exponential distribution of 40000 observations we generated with $\lambda = 0.2$ and exponential distribution fit with rate $\lambda = 0.2$.

From the plots above we can clearly see that the distribution of averages of 40 exponentials is normally distributed. We can see this in plots **B** and **C** where we overlay a normal distribution with the data we simulated. They both follow the same bell shape and trend. When looking at plot **C** where the exponential distribution we generated is displayed, we see that it follows an exponential trend and that if we compare the shape to the one of the distribution of averages they are completely different.

Conclusions

Assuming our variables are IID with we conclude that our simulations behave as predicted by the CLT. Our analysis showed the following: 1) The mean and variance from our distribution of averages are consistent with the theoretical

values, 2) the T test we ran to study the mean gave us a near 1 p-value as well as a 95% confidence interval for the mean 3) our distribution is bell shaped and when compared to a normal distribution it is clear that they are very similar. Therefore, we conclude that the distribution of averages of 40 exponentials is normally distributed.

Appendix

Below is all the code used to make calculations and create figures in this report.

```
# using the simulated data we now create a matrix of the average of each distribution
means <- matrix(data=NA, nrow=N, ncol=1)
for (i in 1:N) {
        means[i,1] <- mean(data[i,])
}
```

```
# load necessary packages and create dataframe
library(ggplot2)
library(ggpubr)
data_0 <- as.vector(data)
data_0 <- data.frame(data_0)
colnames(data_0) <- "data"

# create histogram of the exponential distribution plot using ggplot2
bw0 <- 2 * IQR(data_0[,1]) / length(data_0[,1])^(1/3)
plot0 <- ggplot(data_0, aes(x=data)) +
        geom_histogram(aes(y=..density..),binwidth = bw0,
        fill="#C02965", color="#e9ecef", alpha=1) +
        labs(x="x",y="density",title="Histogram: Randomly Generated Exponential Distribution") +
        theme(plot.title = element_text(hjust = 0.5, size =10)) +
        scale_x_continuous(breaks = seq(0, 30, by=10), limits=c(0,30))
plot0
```

```
#  run t.test()
sample_mean <- round(mean(means), digits=3)
quant_95 <- round(qt(.975,999), digits=3)
t_statistic <- abs(round(t.test(means, alternative = "two.sided", mu=5, paired=FALSE)$statistic, digits=3))
p_val <- round(t.test(means, alternative = "two.sided", mu=5, paired=FALSE)$p.value,digits=3)
t.test(means, alternative = "two.sided", mu=5, paired=FALSE)
```

```
# load necessary packages
library(ggpubr)

# create dataframe
meansdata <- as.data.frame(means)
colnames(meansdata)<-"average"

# create histogram using ggplot2
bw <- 2 * IQR(meansdata[,1]) / length(meansdata[,1])^(1/3)
plot1 <- ggplot(meansdata, aes(x=average)) +
        geom_histogram(aes(y=..density..),binwidth = bw,
        fill="#4B0082", color="#e9ecef", alpha=0.4) +
        geom_vline(aes(xintercept=theoretical_mean,color="theoretical mean"),
        show.legend=TRUE, size=0.3) +
        geom_vline(linetype="dashed",aes(xintercept=sample_mean,
        color="sample mean"), show.legend=TRUE,size=0.3) +
        labs(x="x",y="density",title="Histogram: Distribution of averages \n of 40 exponentials") +
        theme(plot.title = element_text(hjust = 0.5, size=10)) +
        scale_colour_manual(name="", values = c("blue","red"))

# create density plot using ggplot2
```

```r
plot2 <- ggplot(meansdata, aes(x=average)) +
        geom_density(color="#4B0082",fill="#d8d9e6", alpha=0.01)

d <- ggplot_build(plot2)$data[[1]]

plot2 <- plot2 +
        geom_area(data = subset(d, x >4.946396 & x<5.043370),
        aes(x=x, y=y), fill="#d8d9e6", alpha=1) +
        labs(x="x",y="density",title="Distribution of averages of \n 40 exponentials: 95% Confidence Interval
        geom_vline(linetype="dashed",aes(xintercept=theoretical_mean,
        color="theoretical mean"), show.legend=TRUE, size=0.3) +
        theme(plot.title = element_text(hjust = 0.5, size=10)) +
        scale_colour_manual(name="", values = c("red"))

figure1 <- ggarrange(plot1, plot2, ncol=2,nrow=1,labels = c("A", "B"))
figure1
```

```r
# calculate the sample and theoretical variances
exponential_sd <-(1/lambda)
exponential_var <- exponential_sd^2
n <- 40
theoretical_var <- round(exponential_var/n, digits=3)
sample_var <- round(var(means),digits=3)
```

```r
# create density plot using ggplot2
plot3 <- ggplot(meansdata, aes(x=average)) +
        geom_density(color="#4B0082",fill="#d8d9e6", alpha=0.01) +
        labs(x="x",y="density",title="Distribution of averages of 40 exponentials: \n Theoretical and Sample
        geom_vline(aes(xintercept=theoretical_mean+theoretical_var,
        color="theoretical variance"), show.legend=TRUE, size=0.3) +
        geom_vline(aes(xintercept=theoretical_mean-theoretical_var,
        color="theoretical variance"), show.legend=TRUE, size=0.3) +
        geom_vline(aes(xintercept=theoretical_mean+sample_var,
        color="sample variance"), show.legend=TRUE, size=0.3) +
        geom_vline(aes(xintercept=theoretical_mean-sample_var,
        color="sample variance"), show.legend=TRUE, size=0.3) +
        geom_vline(linetype = "dashed", aes(xintercept=theoretical_mean,
        color="sample mean"), show.legend=TRUE, size=0.3) +
        theme(plot.title = element_text(hjust = 0.5, size=10)) +
        scale_colour_manual(name="", values = c("red", "black","green"))
plot3
```

```r
# create plot A: histogram of distribution of means and density of the same distribution
plot4 <- ggplot(meansdata, aes(x=average)) +
        geom_histogram(aes(y=..density..),binwidth = bw,
        fill="#4B0082", color="#e9ecef", alpha=0.5) +
        geom_density(color="#4B0082",fill="#d8d9e6", alpha=.1) +
        labs(x="x",y="density",
        title="Histogram of distribution of averages of \n 40 exponentials and density fit") +
        theme(plot.title = element_text(hjust = 0.5))

# create plot B: histogram of distribution of means and normal distribution
plot5 <- ggplot(meansdata, aes(x=average)) +
        geom_histogram(aes(y=..density..),binwidth = bw,
        fill="#4B0082", color="#e9ecef", alpha=0.5) +
        labs(x="x",y="density",
        title="Histogram of distribution of averages of 40 \n exponentials and normal distribution fit") +
        theme(plot.title = element_text(hjust = 0.5)) +
        stat_function(fun = dnorm, n = 1000, args = list(mean = 5, sd = sqrt(theoretical_var)),
        color = "#3F9274")
```

```r
# create plot C: density of distribution of means and normal distribution
plot6 <- ggplot(meansdata, aes(x=average)) +
        geom_density(color="#e9ecef",fill="#4B0082", alpha=.4) +
        labs(x="x",y="density",
        title="Density plot of distribution of averages of \n 40 exponentials and normal distribution fit") +
        theme(plot.title = element_text(hjust = 0.5)) +
        stat_function(fun = dnorm, n = 1000, args = list(mean = 5, sd = sqrt(theoretical_var)),color = "#3F92

# create plot D: density of exponential distribution and exponential distribution
plot7 <- ggplot(data=data_0, aes(x=data)) +
        geom_density(color="#e9ecef",fill="#C02965", alpha=.4) +
        labs(x="x",y="density",
        title="Density of randomly generated exponential \n distribution and exponential distribution fit") +
        theme(plot.title = element_text(hjust = 0.5)) +
        stat_function(fun = dexp, n = 1000, args = list(rate = 0.2),color = "#3F9274") +
        scale_x_continuous(breaks = seq(0, 30, by=10), limits=c(0,30))

figure2 <- ggarrange(plot4, plot5, plot6, plot7, ncol=2,nrow=2,labels = c("A", "B", "C", "D"))
figure2
```