Exponential Distribution and the Central Limit Theorem

automatically generated using Rmarkdown

Hung P. Do, PhD, MSEE

September 21, 2023

Contents

List of Figures

List of Tables

1 Overview

According to *wikipedia* the central limit theorem (CLT) was stated that:

> *"In many situations, for independent and identically distributed random variables, the sampling distribution of the standardized sample mean tends towards the standard normal distribution even if the original variables themselves are not normally distributed."*

This report contains simulations of the exponential distribution and investigations of whether the distribution of the averages of exponential samples would follow the normal distribution as predicted by the Central Limit Theorem.

2 Simulations

Let's first examine the exponential distribution given by the `rexp(n,lamda)` function in R.

```
set.seed(20201201)
ss <- rexp(100000,0.2)
hist(ss, breaks = 25, xlab="x", main = "Exponential distribution")
```

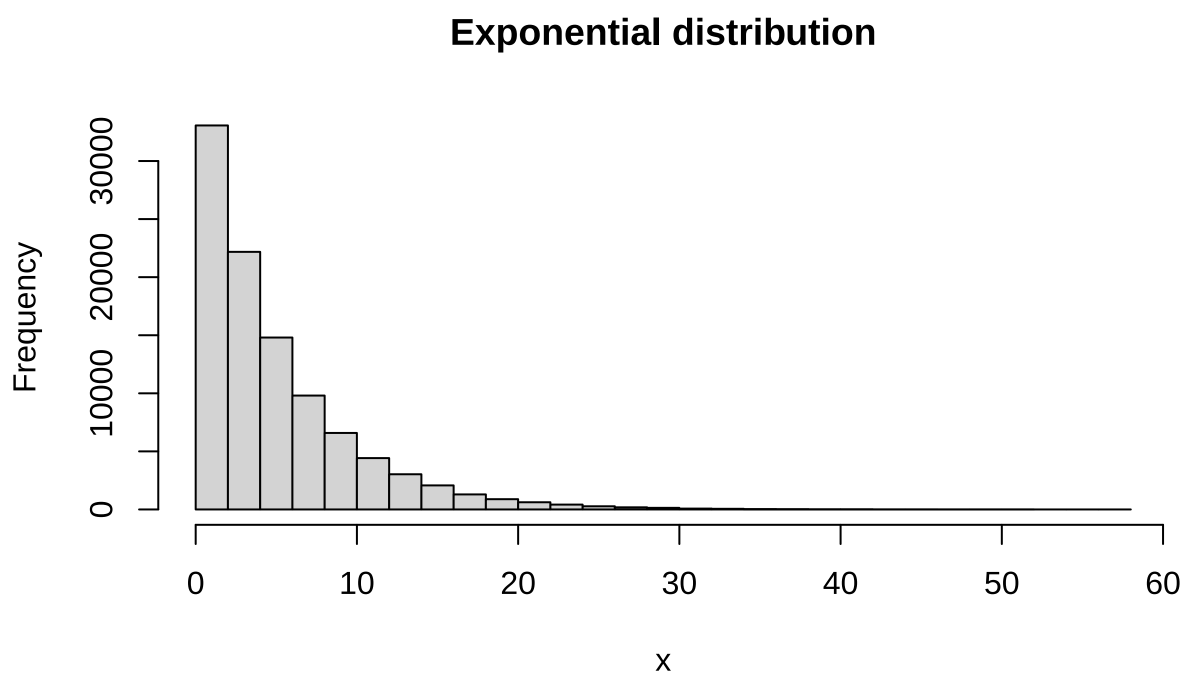

Figure 1: Histogram of 100,000 samples drawn from the exponential distribution.

As we can see from the **Figure** 1, the histogram shows exponential decay trend as the value of `x` increases.

Let's now simulate a distribution of `1,000` trials, in which each sample is an average of `25` random numbers drawn from the exponential distribution.

```r
Ntrial <- 1000
Ns <- 25
lambda = 0.2
mns = NULL
for (i in 1 : Ntrial) mns = c(mns, mean(rexp(Ns,lambda)))
bin_width <- dpih(mns)
nbins <- seq(min(mns)-bin_width, max(mns)+bin_width, by=bin_width)
hist(mns, breaks = nbins, xlab="average of 25 random numbers",
     main = "Distribution of averages of 25 random numbers")
```

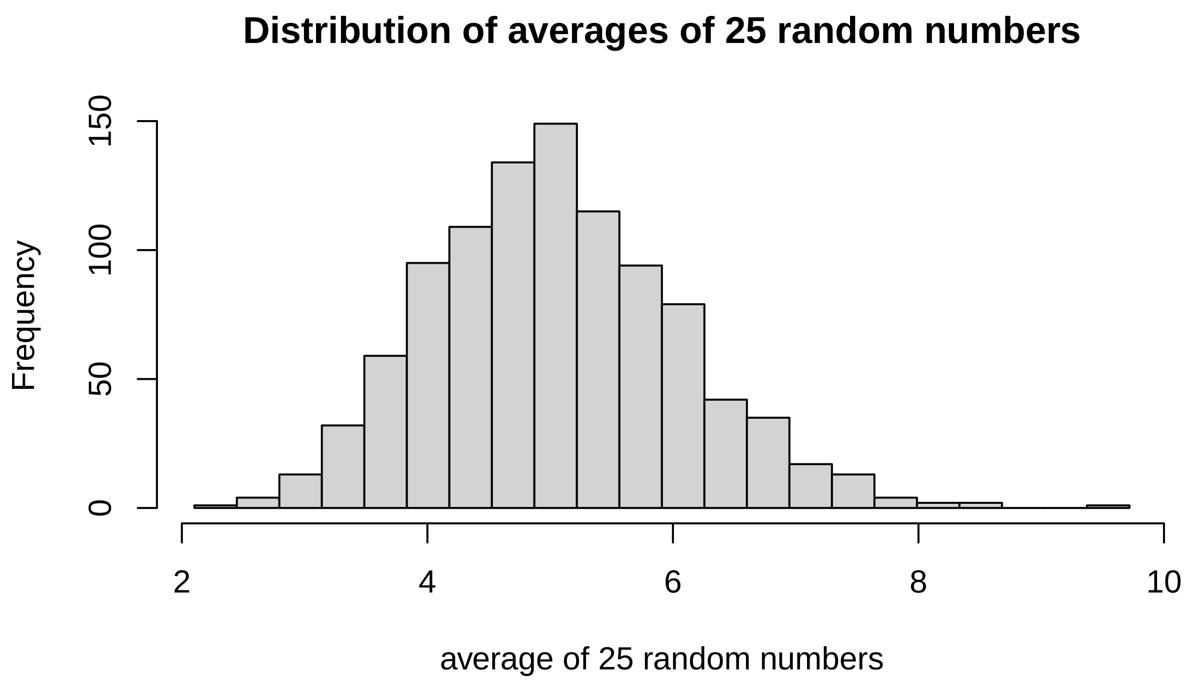

Figure 2: Histogram of 1,000 trials, in which each is an average of 25 random numbers drawn from the exponential distribution.

As we can see from the **Figure** 2, the histogram of `1,000` averages of `25` random numbers drawn from the exponential distribution looks bell-shaped appearance (i.e. looks like shape of the normal distribution). As the number of samples used to calculate average are increased the resulting distribution is closer to the ideal normal distribution as seen in the **Figure** 3. Oppositely, the resulting distribution will deviate from the normal distribution if the number of samples to be averaged is too small as seen in the **Figure** 4.

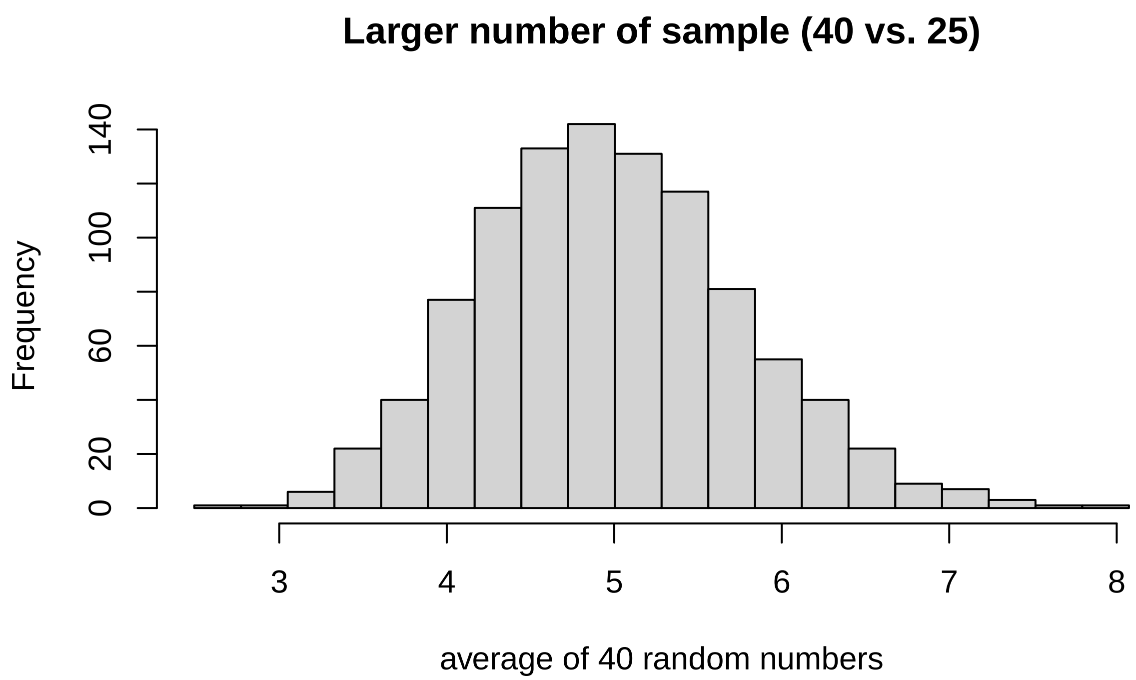

Figure 3: Histogram of 1,000 trials, in which each is an average of 40 random numbers drawn from the exponential distribution.

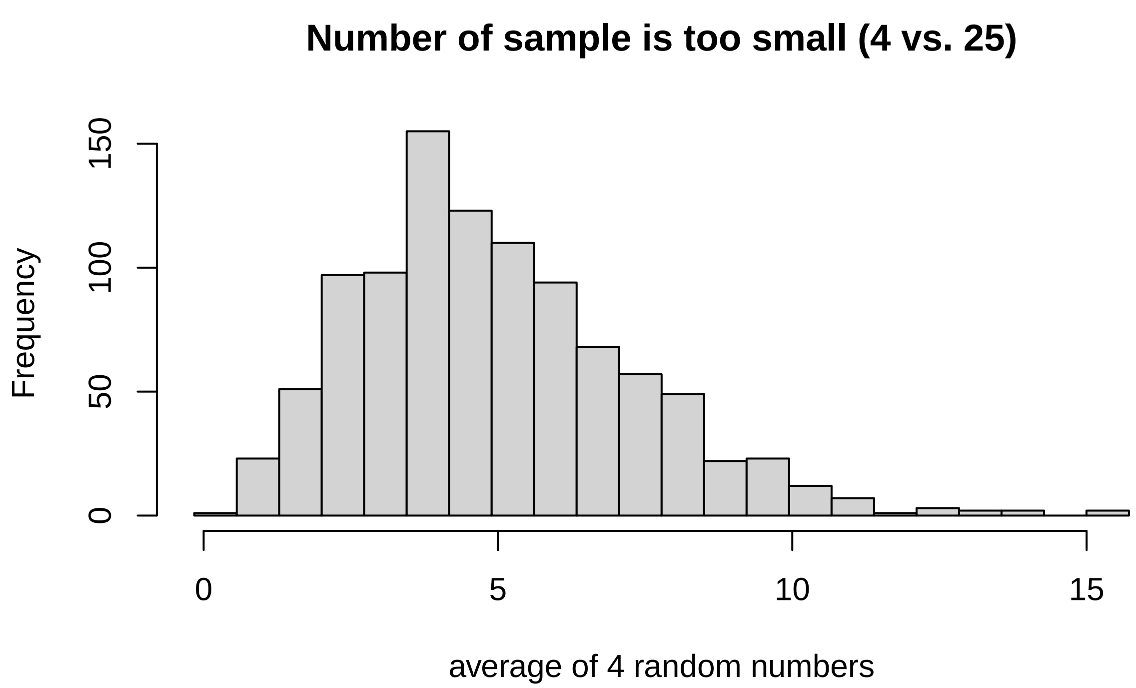

Figure 4: Histogram of 1,000 trials, in which each is an average of 4 random numbers drawn from the exponential distribution.

3 Sample Mean vs. Theoretical Mean

Now let's calculate the theoretical mean (i.e.. `1/lambda`) and sample mean (i.e. `mean(mns)`) estimated from the simulations performed in the previous section.

```
# theoretical mean
tmean <- 1/lambda
# sample mean
smean <- mean(mns)
# standard error
se <- sd(mns)/sqrt(Ntrial)
# 95% confidence interval
ci <- smean + c(-qnorm(0.975),qnorm(0.975))*se

suppressWarnings(library(knitr))
kable(data.frame("theoreticalMean" = tmean, "sampleMean" = smean, "diff" = smean-tmean,
                 "rel_diff percent" = 100*(smean-tmean)/tmean,
                 "stderr" = se, "lower CI" = ci[1], "upper CI" = ci[2]),
      align = "ccccccc",
      digits = c(3,3,3,3,3,3,3),
      caption = "Theoretical Mean vs. Sample Mean.")
```

Table 1: Theoretical Mean vs. Sample Mean.

theoreticalMean	sampleMean	diff	rel_diff.percent	stderr	lower.CI	upper.CI
5	5.042	0.042	0.84	0.032	4.979	5.105

As we can see from the **Table** 1, the difference between sample mean and theoretical mean is approximately 0.84 (%). Additionally, the `95%` confidence interval covers the theoretical mean. Hence, based on the simulations, there is not enough evidence to reject the null hypothesis, which is that the theoretical mean is equal to the sample mean. The same conclusion could be drawn from the statistical test below, in which the sample mean is tested again the theoretical mean using one-sample Student's T-test.

```
ttest <- t.test(mns, mu = tmean, alternative = "two.sided")
print(ttest)
```

```
##
##  One Sample t-test
##
## data:  mns
## t = 1.308, df = 999, p-value = 0.1912
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.978993 5.104993
## sample estimates:
## mean of x
##  5.041993
```

For visualization purposes, we can show the bar plot of the sample mean with error bar represents the `95%` confidence interval in comparison with the theoretical mean as seen in **Figure** 5 below.

```
df <- data.frame(mean = c(tmean, smean),
                 lci = c(tmean, smean)+c(0.0, -qnorm(0.975)*se),
                 uci = c(tmean, smean)+c(0.0, qnorm(0.975)*se),
                 row.names = c("theoreticalMean", "sampleMean"))
library(ggplot2)
ggplot(df, aes(x=row.names(df), y=mean, width=0.5)) +
  geom_bar(stat="identity", fill="grey") +
  geom_errorbar(aes(ymin = lci, ymax = uci), width=0.3) +
  xlab("") +
  theme_classic()
```

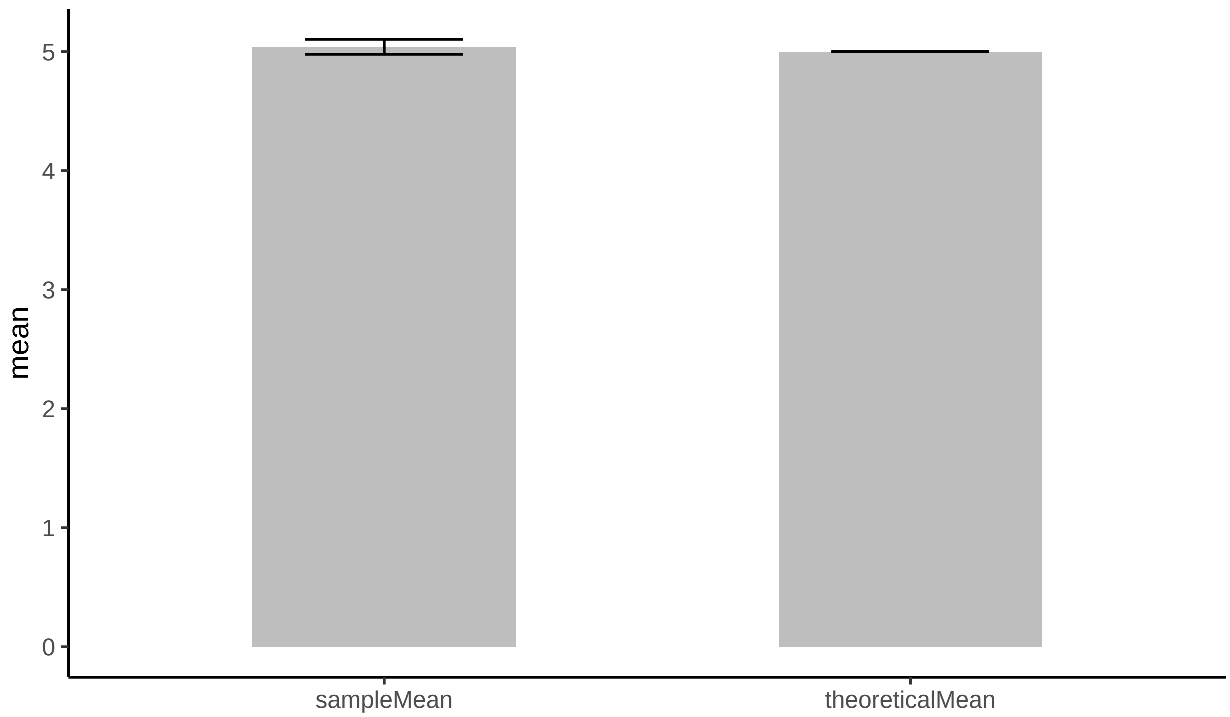

Figure 5: Bar plot of sample mean with error bar represents the 95% confidence interval in comparison with the theoretical mean.

4 Sample Variance vs. Theoretical Variance

Similar analysis could be performed to compare the sample variance and the theoretical variance.

Let's calculate the theoretical variance (i.e.. `(1/lambda^2)/Ns`) and sample variance (i.e. `var(mns)`) estimated from the simulations performed in the previous section.

```
# theoretical variance
tvar <- (1/lambda^2)/Ns
# sample variance
svar <- (sd(mns))^2
# 95% confidence interval
vci <- (Ntrial-1)*svar/qchisq(c(.975,.025), Ntrial-1)

kable(data.frame("theoryVar" = tvar, "sampleVar" = svar, "diff" = svar-tvar,
                 "rel_diff percent" = 100*(svar-tvar)/tvar,
                 "lower CI" = vci[1], "upper CI" = vci[2]),
      align = "cccccc",
      digits = c(3,3,3,3,3,3),
      caption = "Theoretical Variance vs. Sample Variance.")
```

Table 2: Theoretical Variance vs. Sample Variance.

theoryVar	sampleVar	diff	rel_diff.percent	lower.CI	upper.CI
1	1.031	0.031	3.07	0.946	1.127

As we can see from the **Table** 2, the difference between sample variance and theoretical variance is approximately 3.07 (%). Additionally, `95%` confidence interval covers the theoretical variance. Hence, based on the simulations, there is not enough evidence to reject the null hypothesis, which is that the theoretical variance is equal to the sample variance.

For visualization purposes, we can show the bar plot of the sample variance with error bar represents the 95% confidence interval in comparison with the theoretical variance as seen in **Figure** 6 below.

```
dfv <- data.frame(var = c(tvar, svar),lci = c(tvar, vci[1]), uci = c(tvar, vci[2]),
                 row.names = c("theoryVariance", "sampleVariance"))
library(ggplot2)
ggplot(dfv, aes(x=row.names(dfv), y=var, width=0.5)) +
  geom_bar(stat="identity", fill="grey") +
  geom_errorbar(aes(ymin = lci, ymax = uci), width=0.3) +
  xlab("") +
  theme_classic()
```

5 Sample Distribution vs. Normal Distribution

In this section, we will compare the sample distribution with a normal distribution.

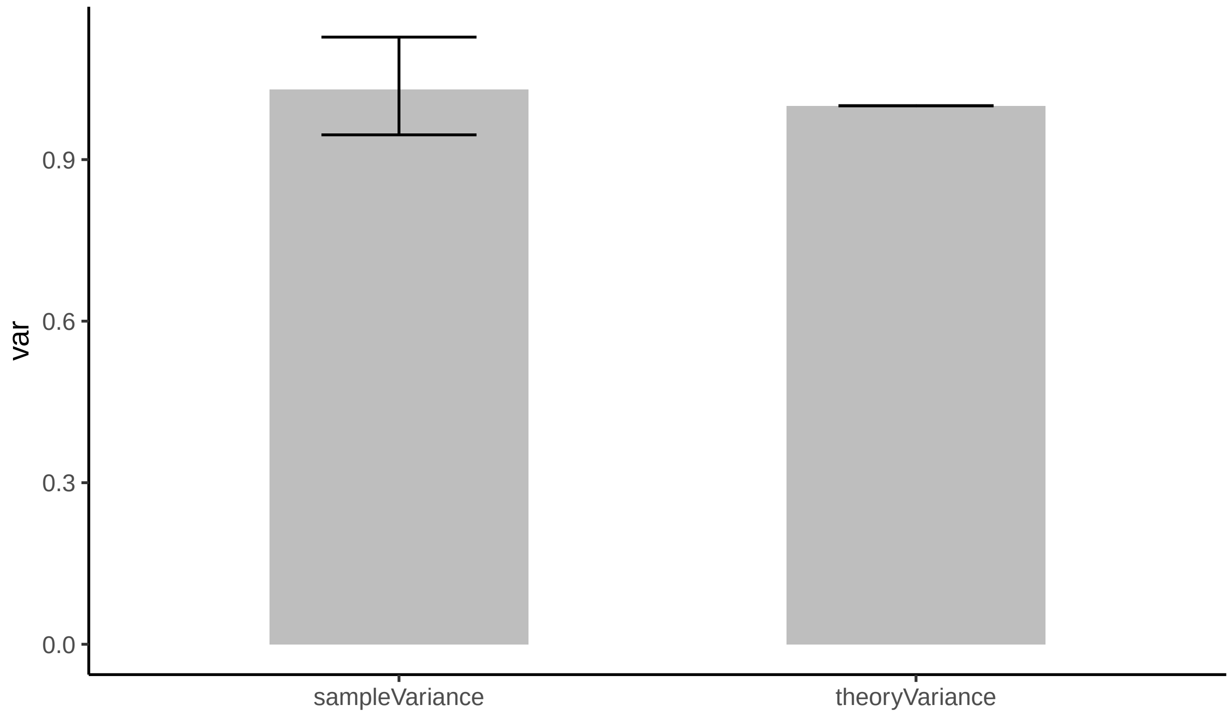

Figure 6: Bar plot of sample variance with error bar represents the 95% confidence interval in comparison with the theoretical variance.

Let's plot the density function of the sample distribution and the density of the normal distribution with theoretical mean and variance.

```
hist(mns, freq = FALSE, breaks = nbins,
     xlab="average of 25 random numbers",
     main = "Probability Density Plot")
x <- seq(0, 10, length=100)
dx <- dnorm(x, mean = tmean, sd = sqrt(tvar))
lines(x, dx, type="l", lwd=3, col="blue")
abline(v = tmean, lwd=3, lty=2, col="red")
```

As we can see from **Figure** 7, the sample distribution approximates the normal distribution.

More concretely, let's use Q-Q plot and Kolmogorov-Smirnov test to check if the distribution is normal. As seen from **Figure** 8, the sample distribution is approximate the normal distribution.

```
qqnorm((mns-mean(mns))/sd(mns), main = "QQ plot", pch=19)
qqline((mns-mean(mns))/sd(mns))
```

Kolmogorov-Smirnov test is used to test if the sample distribution and the normal distribution come from the same distribution.

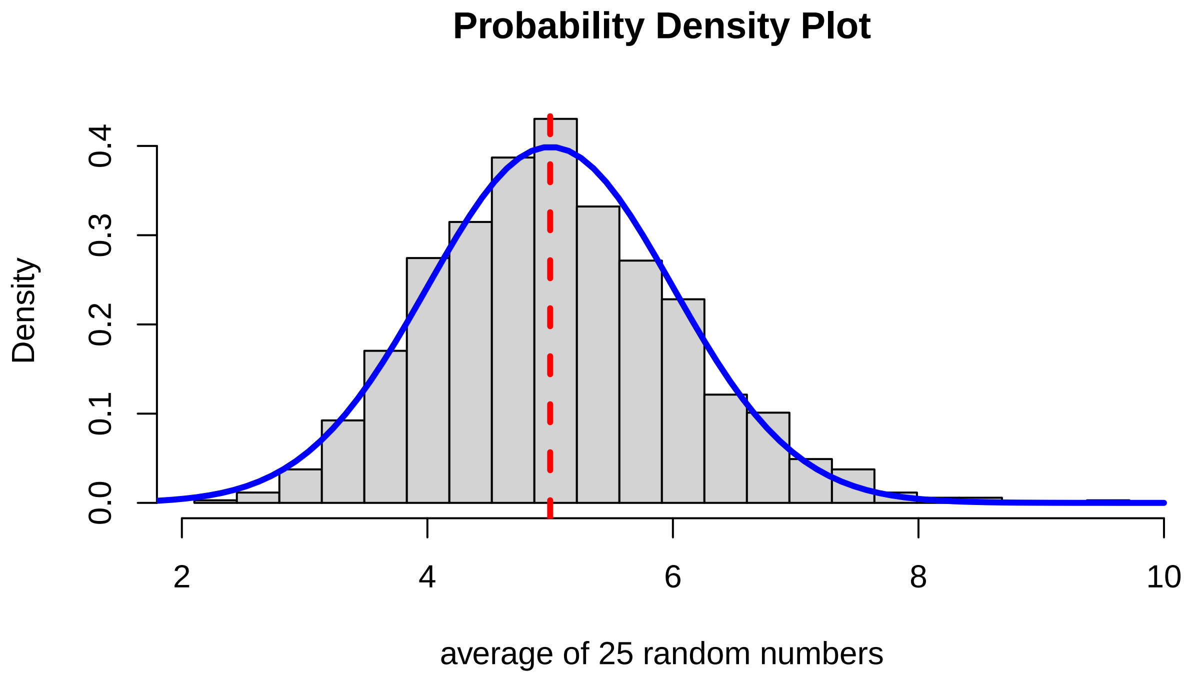

Figure 7: Density plot of the 1,000 averages of 25 samples drawn from the exponential distribution in comparison with the normal distribution with theoretical mean and theoretical variance.

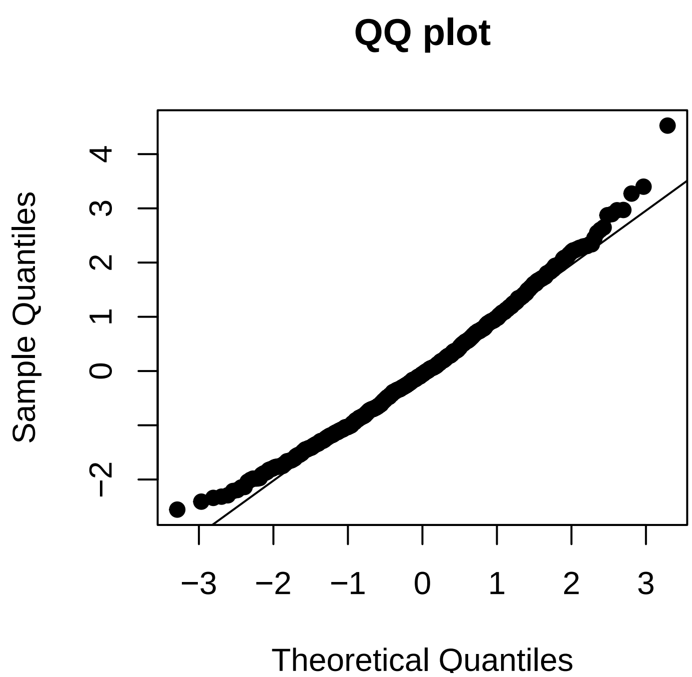

Figure 8: QQ plot of the sample distribution in comparison to the normal distribution.

```r
# Do x and y come from the same distribution?
ktest <- ks.test(mns, rnorm(Ntrial, mean=tmean, sd=sqrt(tvar)))
print(ktest)
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  mns and rnorm(Ntrial, mean = tmean, sd = sqrt(tvar))
## D = 0.03, p-value = 0.7591
## alternative hypothesis: two-sided
```

From the results of the Kolmogorov-Smirno test, the p-value is equal to 0.7591, which is larger than a typical significant level `alpha = 0.05`. It means that there is not enough evidence to reject the null hypothesis, which states that the sample distribution and the normal distribution come from the same distribution.

6 Summary

In this report, we demonstrated, through simulations and statistical analysis, that the Central Limit Theorem is applied and valid for the distribution of `1,000` averages of `25` random numbers drawn from the exponential distribution. Specifically, this report shows that a distribution of `1,000` averages of `25` random numbers drawn from the exponential distribution, which is very much different from the normal distribution, does follow the normal distribution with sample mean and sample variance are similar to the theoretical ones as predicted by the Central Limit Theorem.

7 Open Questions

7.1 Observations

One of the hypotheses was that:

> *As the number of trial (`Ntrial`) and the number of samples (`Ns`) used to average increased the resulting distribution would progress toward the ideal normal distribution.*

However, the resulting QQ-plot and KS-test show the opposite. So the open question is that:

> *Why the resulting distribution is futher deviated from the ideal normal distrubution when `Ntrial` and `Nsample` are increased.*

Let's now simulate a distribution of `100,000` trials, in which each sample is an average of `50` random numbers drawn from the exponential distribution.

```
Ntrial1 <- 100000
Ns1 <- 50
lambda = 0.2
mns1 = NULL
for (i in 1 : Ntrial1) mns1 = c(mns1, mean(rexp(Ns1,lambda)))
bin_width <- dpih(mns1)
nbins1 <- seq(min(mns1)-bin_width, max(mns1)+bin_width, by=bin_width)
hist(mns1, breaks = nbins1, xlab="average of 50 random numbers",
     main = "Larger Ntrial=100,000 & Nsample=50")
```

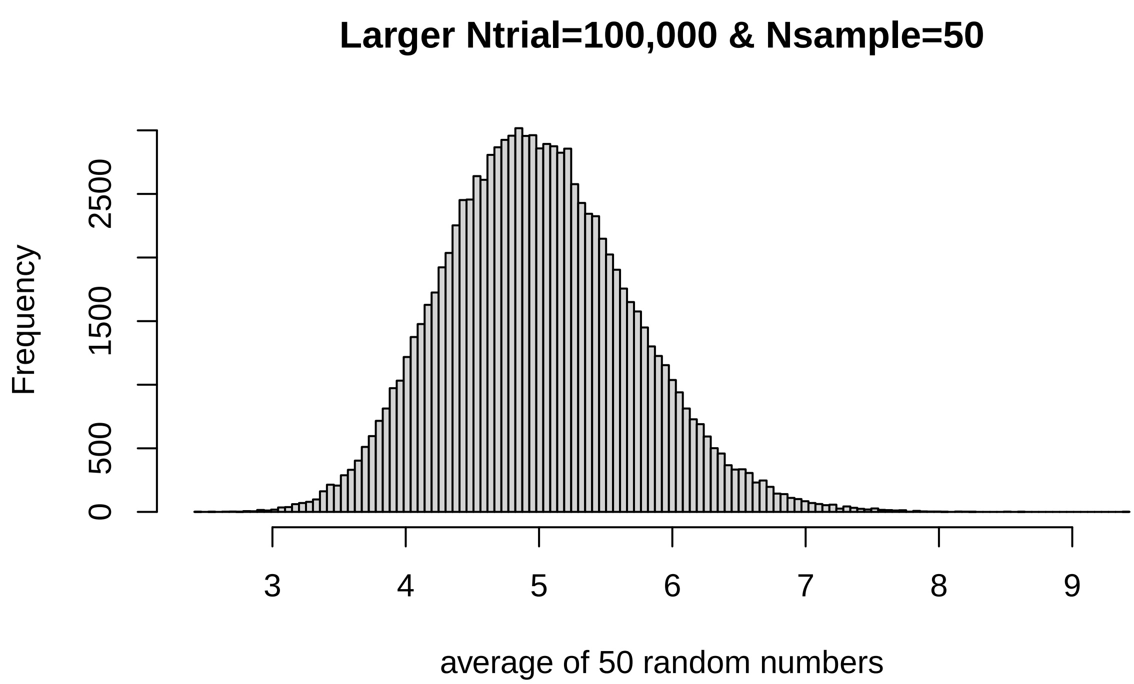

Figure 9: Histogram of 100,000 trials, in which each is an average of 50 random numbers drawn from the exponential distribution.

Let's plot the density function of the sample distribution and the density of the normal distribution with theoretical mean and variance.

```
tmean1 <- 1/lambda
tvar1 <- (1/lambda^2)/Ns1
hist(mns1, freq = FALSE, breaks = nbins1,
     xlab="average of 50 random numbers",
     main = "Larger Ntrial=100,000 & Nsample=50")
x <- seq(0, 10, length=100)
dx <- dnorm(x, mean = tmean1, sd = sqrt(tvar1))
lines(x, dx, type="l", lwd=3, col="blue")
abline(v = tmean1, lwd=3, lty=2, col="red")
```

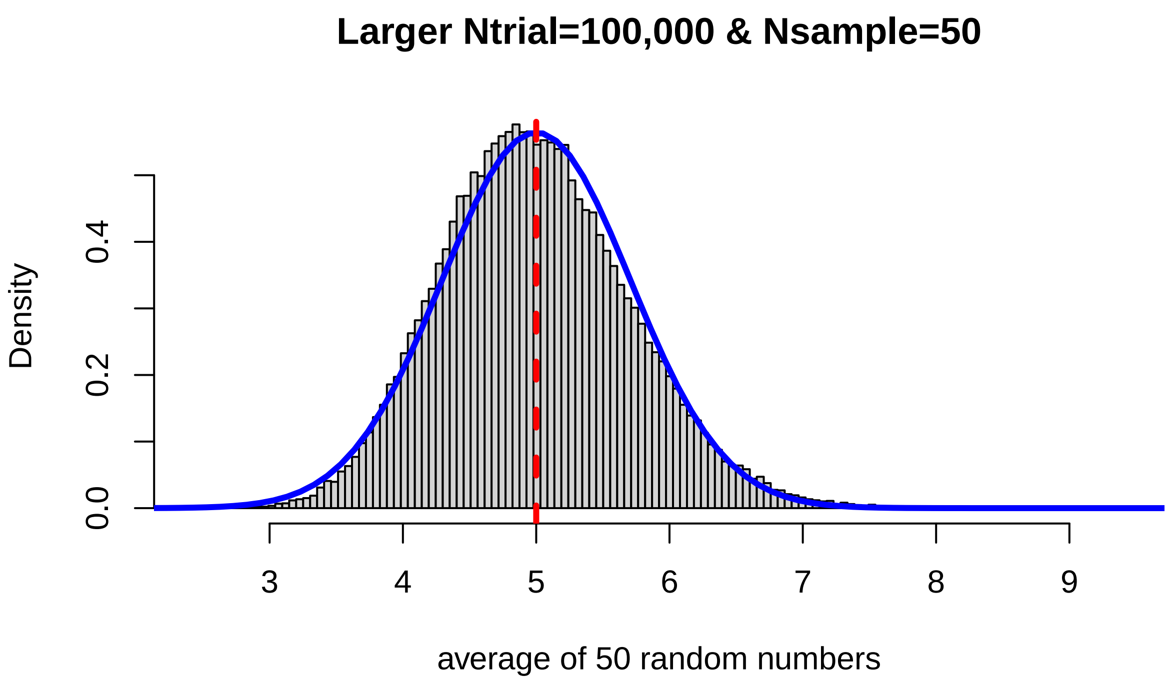

Figure 10: Density plot of the 100,000 averages of 50 samples drawn from the exponential distribution in comparison with the normal distribution with theoretical mean and theoretical variance.

As seen from **Figure** 9 and **Figure** 10, the sample distribution approximates the normal distribution.

However, as seen from the QQ-plot **Figure** 11, the sample distribution seems deviated further from the normal distribution as the samples at the two tails are shifted from the identity line.

```
qqnorm((mns1-mean(mns1))/sd(mns1), main = "Larger Ntrial=100,000 & Nsample=50", pch=19)
qqline((mns1-mean(mns1))/sd(mns1))
```

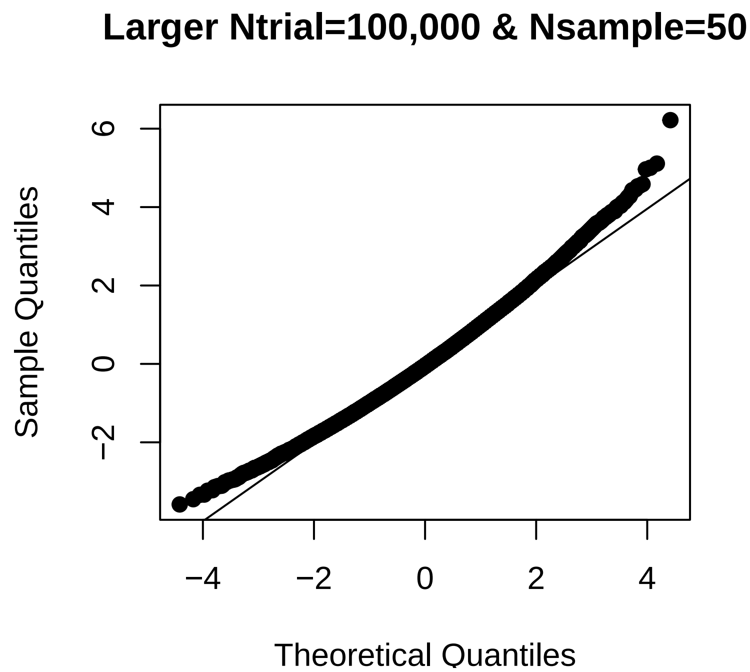

Figure 11: QQ plot of the sample distribution in comparison to the normal distribution.

```
# Do x and y come from the same distribution?
ktest1 <- ks.test(mns1, rnorm(Ntrial1, mean=tmean1, sd=sqrt(tvar1)))
print(ktest1)
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  mns1 and rnorm(Ntrial1, mean = tmean1, sd = sqrt(tvar1))
## D = 0.02076, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Kolmogorov-Smirnov test is used to test if the sample distribution and the normal distribution come from the same distribution. From the results of the Kolmogorov-Smirno test, thee p-value is equal to 0, which is **smaller** than a typical significant level `alpha = 0.05`. It means that there is enough evidence to **reject** the null hypothesis, which states that the sample distribution and the normal distribution come from the same distribution.

7.2 Sanity Check

As a sanity check, let's check if the Normal Distribution comes from the Normal Distribution. Indeed, it is as shown in the QQ-plot and the results of the Kolmogorov-Smirnov test.

Let's now simulate a distribution of `100,000` trials, in which each sample is an average of `50` random numbers drawn from the **Normal Distribution**.

```
Ntrial2 <- 100000
Ns2 <- 50
mu = 5.0
sigma = 2.5
mns2 = NULL
for (i in 1 : Ntrial2) mns2 = c(mns2, mean(rnorm(Ns2, mean = mu, sd=sigma)))
bin_width <- dpih(mns2)
nbins2 <- seq(min(mns2)-bin_width, max(mns2)+bin_width, by=bin_width)
hist(mns2, breaks = nbins2, xlab="average of 50 random numbers",
     main = "The Normal Distribution")
```

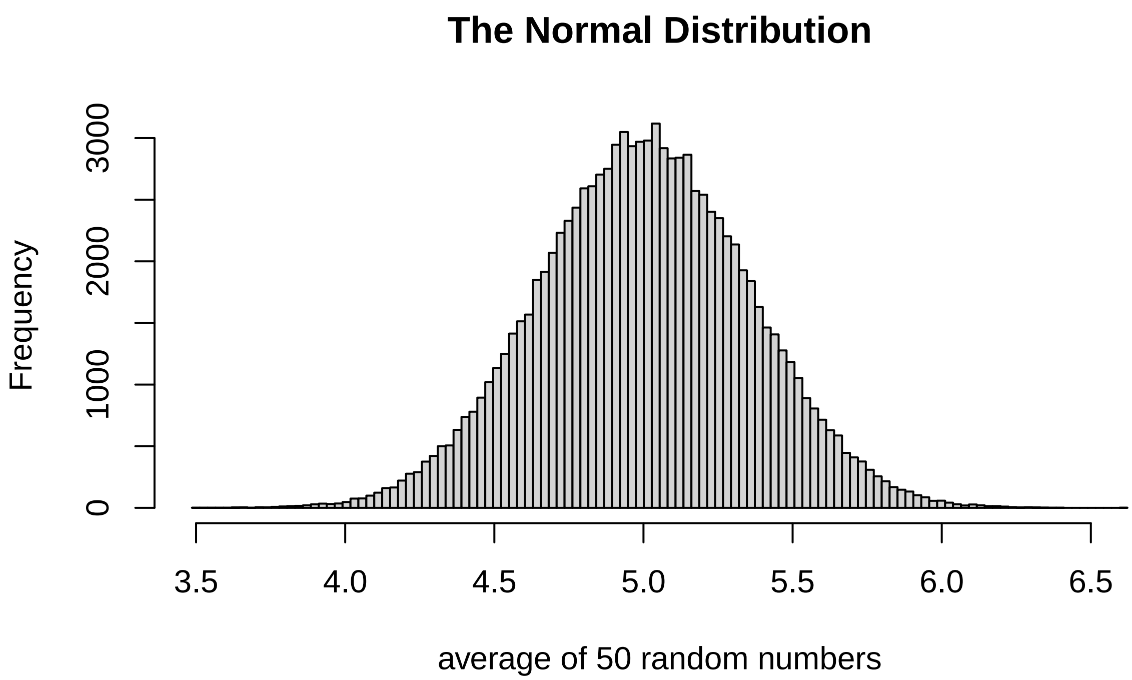

Figure 12: Histogram of 100,000 trials, in which each is an average of 50 random numbers drawn from the normal distribution.

Let's plot the density function of the sample distribution and the density of the normal distribution with theoretical mean and variance.

```
tmean2 = mu
tvar2 = sigma*sigma/Ns2
hist(mns2, freq = FALSE, breaks = nbins2,
     xlab="average of 50 random numbers",
     main = "Normal Distribution")
x2 <- seq(0, 10, length=1000)
dx2 <- dnorm(x2, mean = tmean2, sd = sqrt(tvar2))
lines(x2, dx2, type="l", lwd=3, col="blue")
abline(v = tmean2, lwd=3, lty=2, col="red")
```

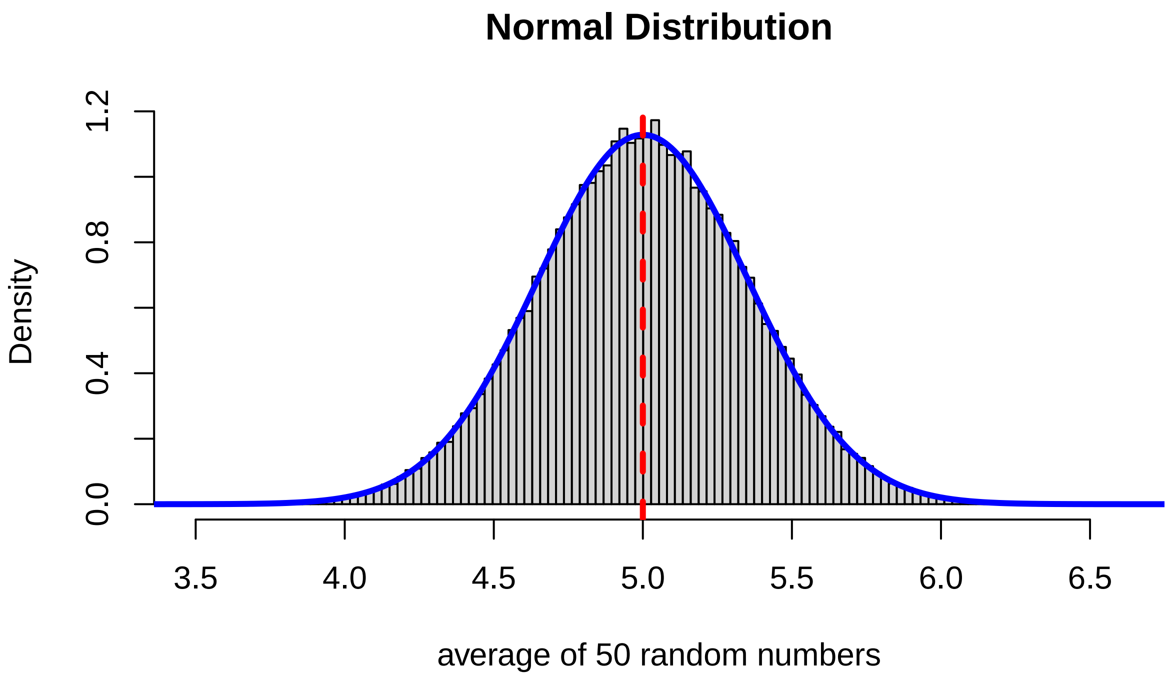

Figure 13: Density plot of the 100,000 averages of 50 samples drawn from the normal distribution in comparison with the normal distribution with theoretical mean and theoretical variance.

Both QQ plot and Kolmogorov-Smirnov test results support the fact that the sample distribution and the normal distribution come from the same distribution.

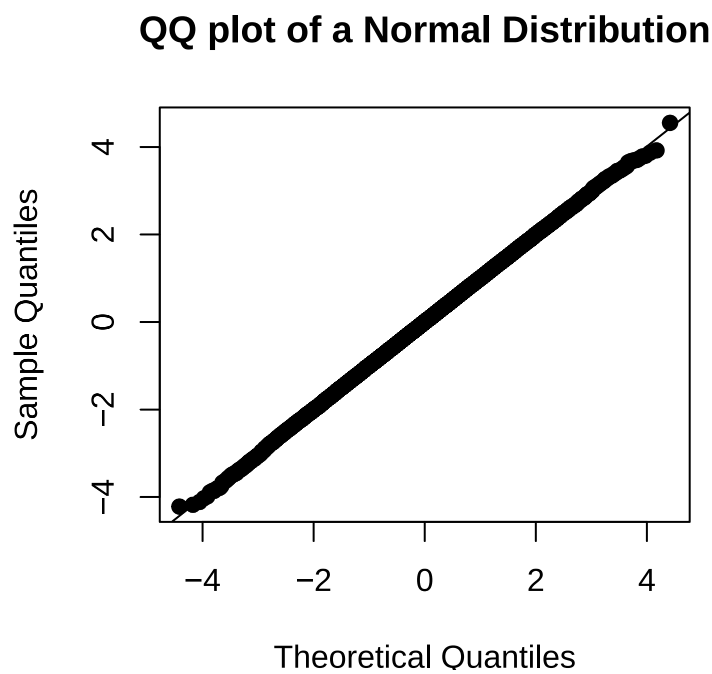

Figure 14: QQ plot of the normal distribution in comparison to the normal distribution.

```r
# Do x and y come from the same distribution?
ktest2 <- ks.test(mns2, rnorm(Ntrial2, mean=tmean2, sd=sqrt(tvar2)))
print(ktest2)
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  mns2 and rnorm(Ntrial2, mean = tmean2, sd = sqrt(tvar2))
## D = 0.00499, p-value = 0.1657
## alternative hypothesis: two-sided
```