# An Illustrative Example of the Central Limit Theorem

*James C. Dembowski*

**Overview**

We provide an illustrative example of the Central Limit Theorem. 1000 random samples are obtained and the mean of each sample is calculated. The distribution of those means are then compared to the normal distribution both graphically and with the Shapiro-Wilk.

**Central Limit Theorem:**

*Consider a sample of size n taken from a population with mean $\mu$ and standard deviation $\sigma$. As n increases, the distribution of the sample mean $\bar{x}$ converges to the normal distribution, with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$, regardless of the probability distribution of the population from which the sample was taken.*

**Methodology**

1000 random samples, each of size $n = 40$ and with parameter $\lambda = 0.2$, are obtained from an exponential distribution. For each sample, the mean is calculated and saved, and the original samples are discarded. The mean and variance of the sample means are calculated and compared to the theoretical mean and variance, respectively. The distribution of the sample means is compared to the distribution from which the samples were obtained, as well as to the normal distribution. A Q-Q plot is created, and the Shapiro-Wilk test is performed.

**Mean and Variance: Sample vs. Theoretical**

We calculate the mean $\overline{\bar{x}}$ of our sample means, obtaining $\overline{\bar{x}} \approx 4.979$. The theoretical mean of the distribution of sample means is equal to the mean of the exponential distribution from which the samples are taken, $\frac{1}{\lambda} = 5$.

We next calculate the sample variance $s^2$ of our sample means, obtaining $s^2 \approx 0.642$. The theoretical variance is the variance of the exponential distribution divided by our sample size, $\frac{\sigma^2}{n} = \frac{1}{40\lambda^2} = 0.625$. (We use $n = 40$ instead of $n = 1000$ because we are interested in the variance of the sample mean for a sample size $n = 40$.)

Table 1: Means and Variances

| | Theoretical | Actual |
|---|---|---|
| Mean | 5.000 | 4.979 |
| Variance | 0.625 | 0.642 |

The statistics calculated from the sample and the theoretical parameter values are reasonably close.

**Exponential Distribution vs. Distribution of Sample Means from Samples of Exponentially Distributed Random Variables**

Figure 1 shows a histogram of a random sample, size $n = 1000$, taken from a population with an exponential distribution, $\lambda = 0.2$, with the density curve overlaid. The asymmetric shape of exponential distribution is clear.

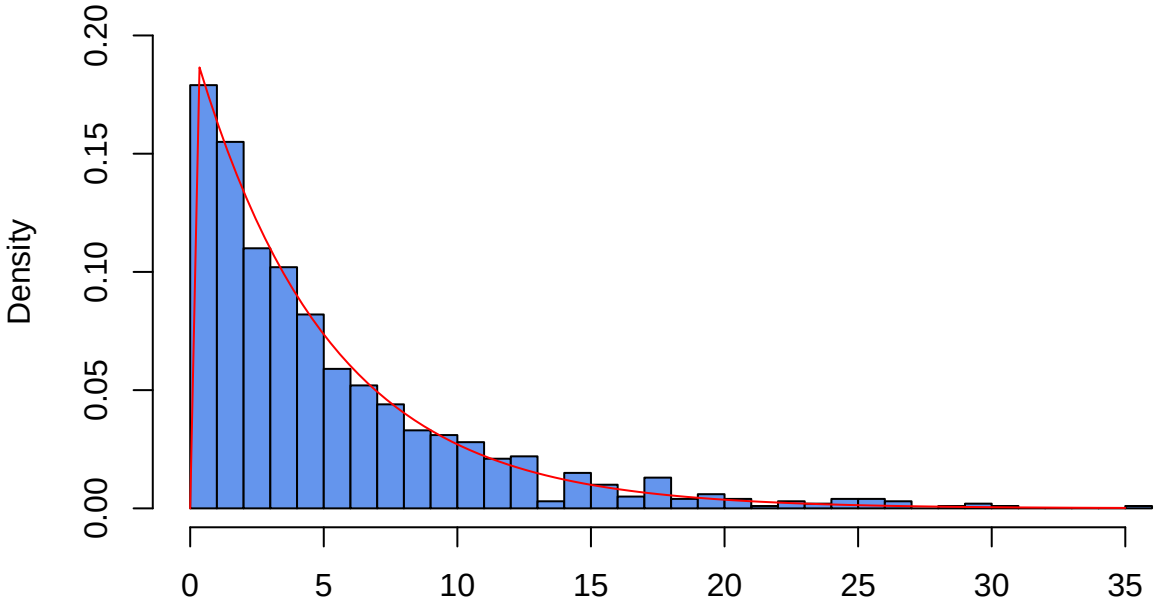


*Figure* 1: *Random sample from exponential distribution*, $n = 1000$, $\lambda = 0.2$.

Figure 2 shows a histogram of our sample means, with the density curve of the normal distribution with mean $\mu = \frac{1}{\lambda} = 5$ and variance $\frac{\sigma^2}{n} = \frac{1}{40\lambda^2} = 0.625$ overlaid. The distribution of the sample means is symmetric and approximately follows the normal curve.

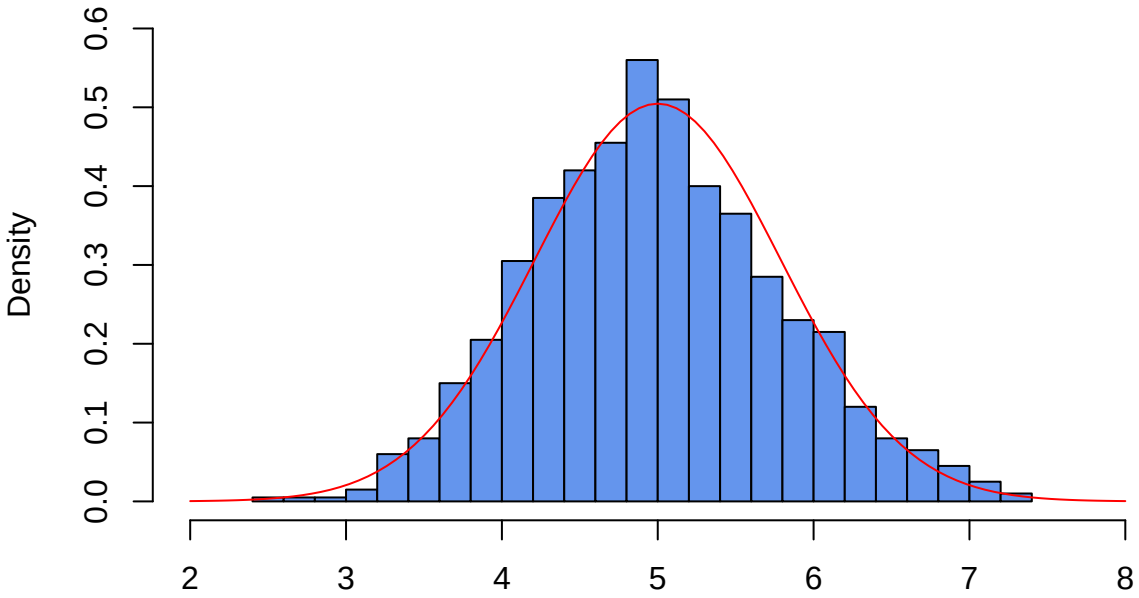


*Figure* 2: *Sample means of* 1000 *random exponential samples of size* $n = 40$, $\lambda = 0.2$.

We create a Q-Q plot, Figure 3, to perform an additional visual verification of normality. The ordered sample means largely fall along the diagonal of the plot, indicating their distribution is approximately normal. The extreme valued sample means lie further away from the diagonal, indicating the approximation is not perfect.

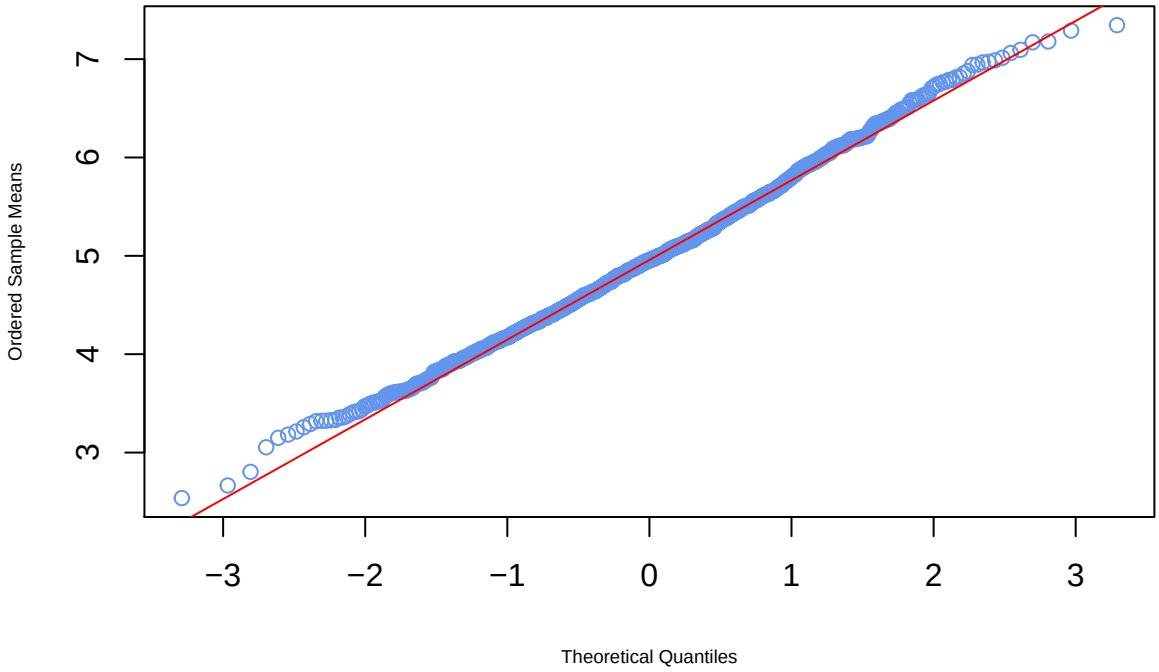


*Figure* 3: *$Q-Q$ plot*: *Ordered sample means vs. theoretical quantiles.*

We also test the distribution of sample means using the Shapiro-Wilk test, which tests the null hypothesis that the sample of means come from a normal distribution against the alternative hypothesis of non-normality. We obtain a $p$-value of 0.02, which further suggests our sample means are not *exactly* normally distributed.

We also use the Shapiro-Wilk test to see how increasing $n$ from 40 affects the normal approximation of the sampling distribution of sample means. We calculate the sample mean for each of 1000 samples of each size 100, 1000, 10,000, and 100,000. As the sample size increases, the $p$-value for the test also increases, indicating that for sufficiently large $n$, our distribution of sample means is not statistically different from the normal distribution. Results are shown in Table 2.

Table 2: Shapiro-Wilk Test Results

| | Test Statistic W | P-Value |
|---|---|---|
| n = 40 | 0.996 | 0.020 |
| n = 100 | 0.997 | 0.067 |
| n = 1000 | 0.998 | 0.188 |
| n = 10,000 | 0.998 | 0.517 |
| n = 100,000 | 0.999 | 0.910 |

**Conclusion**

Our results are consistent with the predictions of the central limit theorem. Even though we took our samples from an exponential distribution, the sample means have an approximately normal distribution. Additionally, as $n$ increases, the normal approximation gets better.

## Appendix

R Code

```
# parameters

lambda = 0.2

# We need a function definition to pass to sapply().
expSampleMean = function( rate, n) {

    mean( rexp( n, rate))

}


# Simulations

n = 40

set.seed( 11235)
    # Set seed for reproducibility

meansVector = sapply( rep( lambda, 1000), expSampleMean, n = n)




# Mean and Variance: Sample vs. Theoretical

actualMean = mean( meansVector)

expMean = 1 / lambda
  # Theoretical mean of the exponential distribution

expMean

actualMean

actualVariance = var( meansVector)

expVariance = (1 / lambda) ^ 2
    # theoretical variance of exponential distribution

theoreticalVariance =  expVariance / 40
    # theoretical variance of the distribution of sample mean

theoreticalVariance

actualVariance
```

```r
# Create table to compare actual mean & variance to theoretical mean & variance

library( knitr)

theoretical = c( expMean, theoreticalVariance)

actual = c( actualMean, actualVariance)

statisticsTable = data.frame( theoretical, actual)

rownames( statisticsTable) = c( "Mean", "Variance")

colnames( statisticsTable) = c( "Theoretical", "Actual")

kable( statisticsTable, digits = 3, caption = "Means and Variances")



# Show shape of exponential distribution

# histogram of random sample from an exponentially distributed population
set.seed( 314159)

randExpSample = rexp( 1000, rate = lambda)
    # Take random sample

hist( randExpSample, prob = TRUE, xlim = c( 0, 35), ylim = c( 0, 0.20), breaks = 30,
      col = "cornflowerblue", main = NULL)

x = randExpSample

curve( dexp( x, rate = lambda), add = TRUE, col = "red")
    # Overlay curve of density function
    #
    # This command returns error if I name 'x' something else, such as
    #   'randExpSample'. Why?



# Show shape of distribution of sample means

# histogram of sample means
hist( meansVector, prob = TRUE, xlim = c( 2, 8), ylim = c( 0, 0.6), breaks = 30,
      col = "cornflowerblue", main = NULL)

randNormSample = rnorm( 1000, mean = expMean, sd = sqrt( theoreticalVariance))

x = randNormSample
    # prevent error in following curve() call.

curve( dnorm( x, mean = expMean, sqrt( theoreticalVariance)), add = TRUE,
    col = "red")
```

```
    # Overlay curve of density function


# Visually verify that the distribution of sample means is approximately normal

qqnorm( meansVector, main = "  ", ylab = "Ordered Sample Means", cex.lab = 0.6,
      col = "cornflowerblue")

qqline( meansVector, col = "red")


# Analytic comparison of distribution of means using Shapiro-Wilks test
#   including increasing n


meansMatrix = matrix( ncol = 1000, nrow = 5)

testResults = matrix( ncol = 2, nrow = 5)

set.seed( 141421)

for (i in 2:5) {

    n = 10 ^ i

    meansMatrix[ i,] = sapply( rep( lambda, 1000), expSampleMean, n = n)

    test = shapiro.test( meansMatrix[ i,])

    testResults[ i,] = c( test$statistic, test$p.value)

}

meansMatrix[ 1,] = meansVector

test = shapiro.test( meansVector)

testResults[ 1,] = c( test$statistic, test$p.value)

shapiroWilkTable = data.frame( testResults)

rownames( shapiroWilkTable) = c( "n = 40", "n = 100", "n = 1000", "n = 10,000",
    "n = 100,000")

colnames( shapiroWilkTable) = c( "Test Statistic W", "P-Value")

kable( shapiroWilkTable, digits = 3, caption = "Shapiro-Wilk Test Results")
```