# Assignment 1:Verification of the Central Limit Theorem and its Convergence Rate

yifan

## Background

We have already learned the Central Limit Theorem in our textbook: For a sequence of i.i.d. random variables $\{X_i\}, i = 1, 2, ..., n$ with finite second moments, we have:

$$(\sum_{i=1}^{n} X_i - nEX_1)/\sqrt{nVarX_1} \xrightarrow{d} N(0,1) \tag{1}$$

We want to verify this theorem, as well as its convergence rate, using a sequence of i.i.d. discrete random variables and a sequence of i.i.d. continuous random variables.

## Verification of Central Limit Theorem

We choose a sequence of random variables that follow the Bernoulli distribution as the representative of discrete random variables. Here, we set $P(X_i = 0) = 0.8$ and $P(X_i = 1) = 0.2, i = 1, 2, ..., n$. Apparently, we have $EX_1 = 0.2$ and $VarX_1 = 0.16$.

We write an R function: with n, m (the number of simulations), and d (the interval length for the histogram) as inputs, draw a histogram of the data after m simulations and compare it with the density function of the standard normal distribution. We also set the random seed to ensure the reproducibility of the results.

```
library(ggplot2)
```

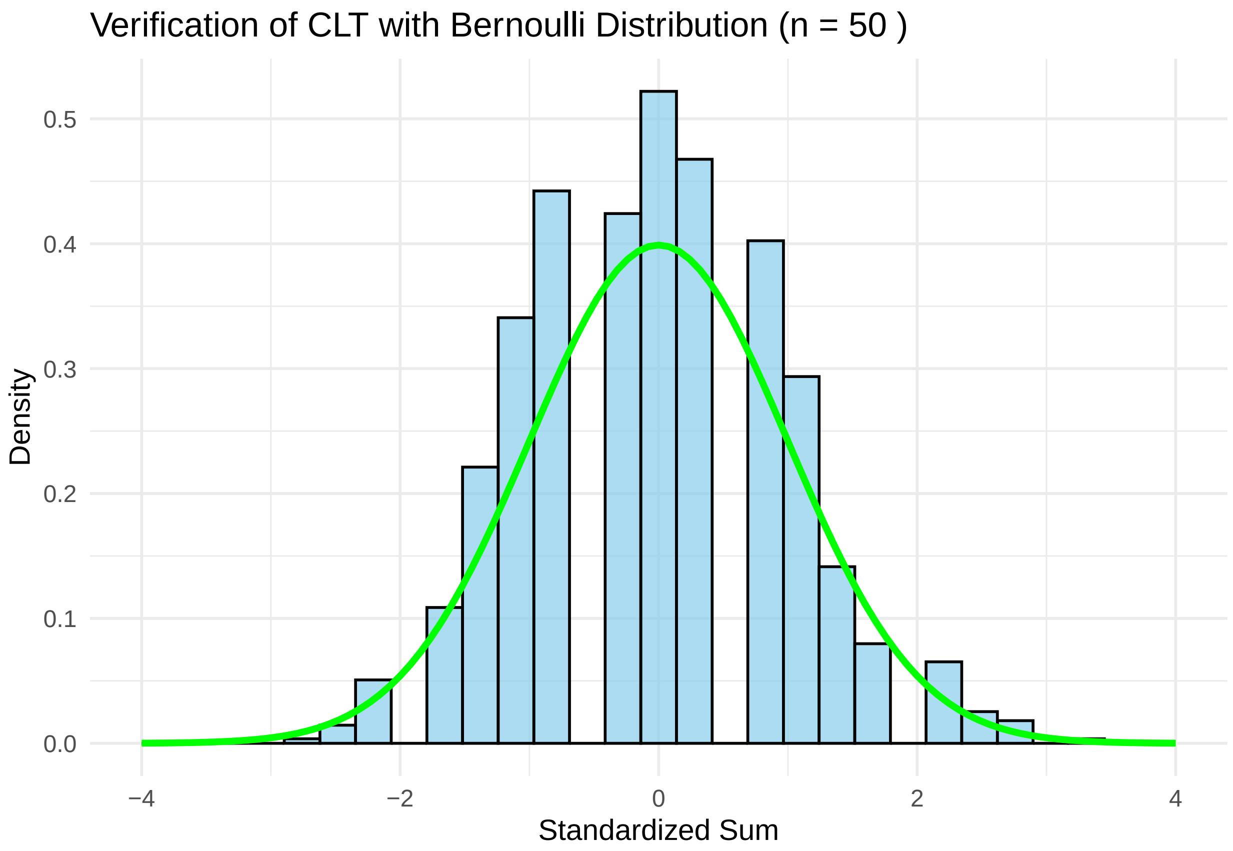
```
## Warning:   'ggplot2' R 4.3.3
```

```
clt_verification1 <- function(m, n, d, seed = 123) {
  # Set the random seed for reproducibility
  set.seed(seed)

  results <- numeric(m)

  for (i in 1:m) {
    X <- rbinom(n, 1, 0.2)  # Bernoulli distribution with probability 0.2

    # The expectation and variance of X_i
    EX <- 0.2
    VarX <- 0.16

    # Calculate the standardized value
    results[i] <- (sum(X) - n * EX) / sqrt(n * VarX)
  }

  # Plot the histogram
  hist_data <- data.frame(results)
  p <- ggplot(hist_data, aes(x = results)) +
    geom_histogram(aes(y = after_stat(density)), bins = 30, fill = "skyblue",
                   color = "black", alpha = 0.7) +
    stat_function(fun = dnorm, args = list(mean = 0, sd = 1), color = "green",
                  linewidth = 1.2) +
    labs(title = paste("Verification of CLT with Bernoulli Distribution (n =", n, ")"),
         x = "Standardized Sum", y = "Density") +
    theme_minimal() +
    xlim(c(-4, 4))

  print(p)
}
```

Set $m = 1000, d = 0.2$, and $n = 50, 100, 1000$, then run the function:

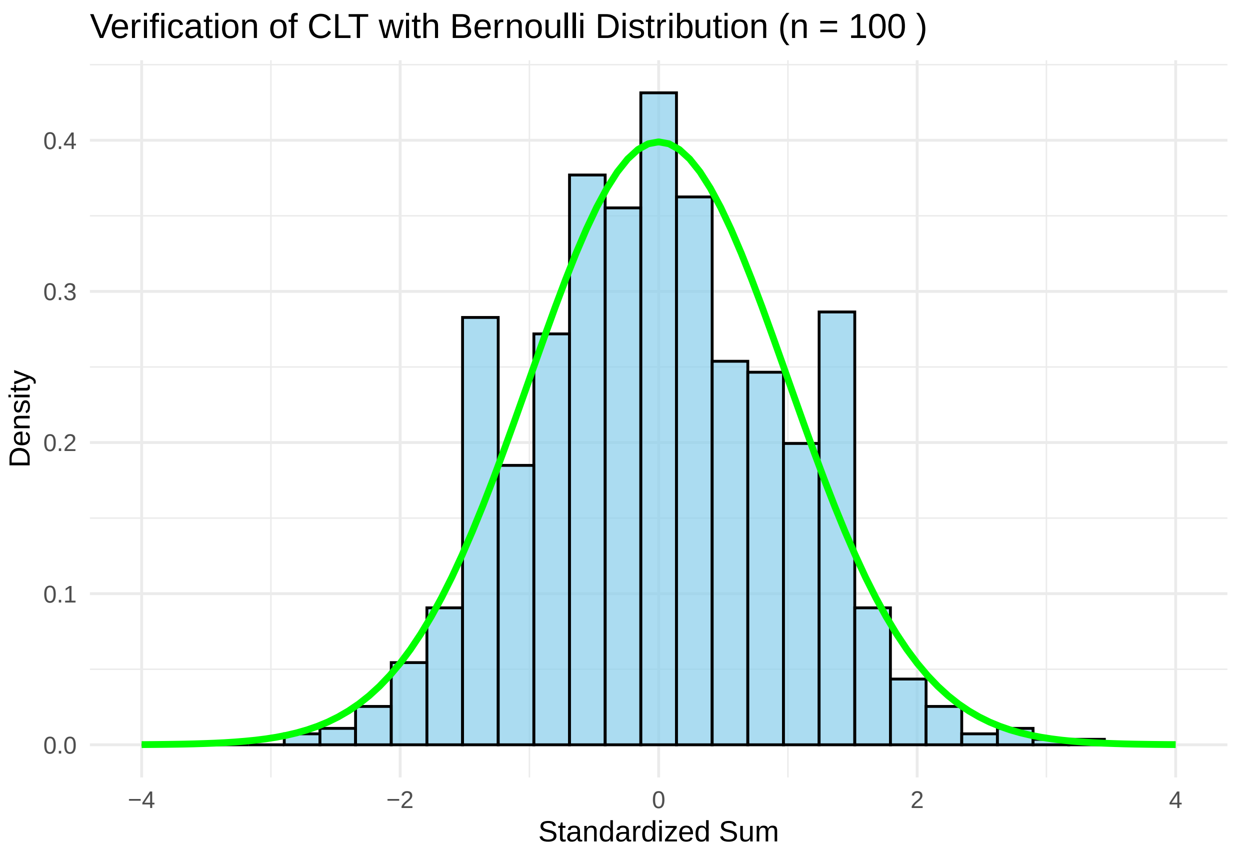
```
clt_verification1(m = 1000, n = 50, d = 0.2)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



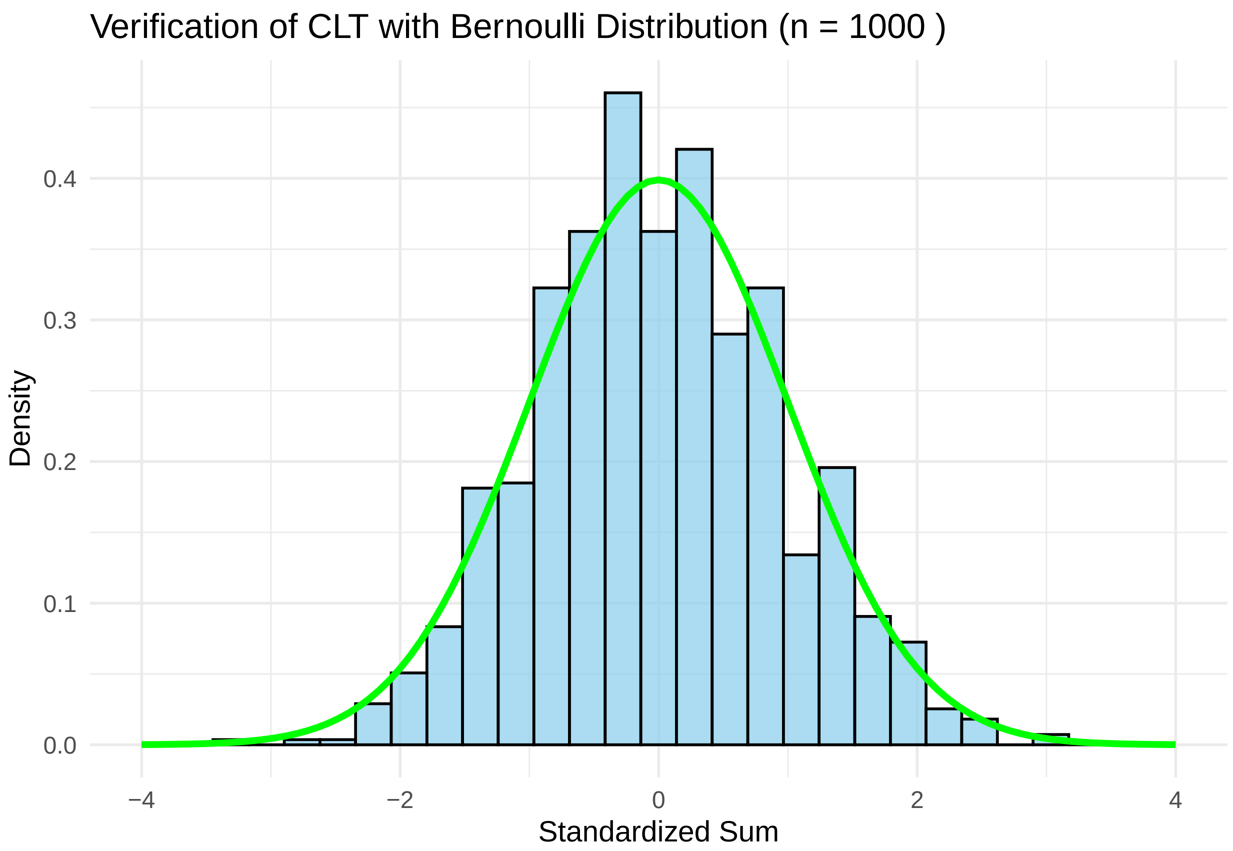
```
clt_verification1(m = 1000, n = 100, d = 0.2)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



```
clt_verification1(m = 1000, n = 1000, d = 0.2)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



We can see that as n increases, the histogram converges to the standard normal distribution.

Furthermore, we can perform a normality test for the case of m=1000 and n=1000, using Shapiro-Wilk test:

```r
# Set the random seed for reproducibility
set.seed(123)

  results <- numeric(1000)

  for (i in 1:1000) {
    X <- rbinom(1000, 1, 0.2)  # Bernoulli distribution with probability 0.2

    # The expectation and variance of X_i
    EX <- 0.2
    VarX <- 0.16

    # Calculate the standardized value
    results[i] <- (sum(X) - 1000 * EX) / sqrt(1000 * VarX)
  }
shapiro.test(results)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  results
## W = 0.99807, p-value = 0.3137
```

Since $p = 0.3137$, it is convincing to believe that the data follows normal distribution.

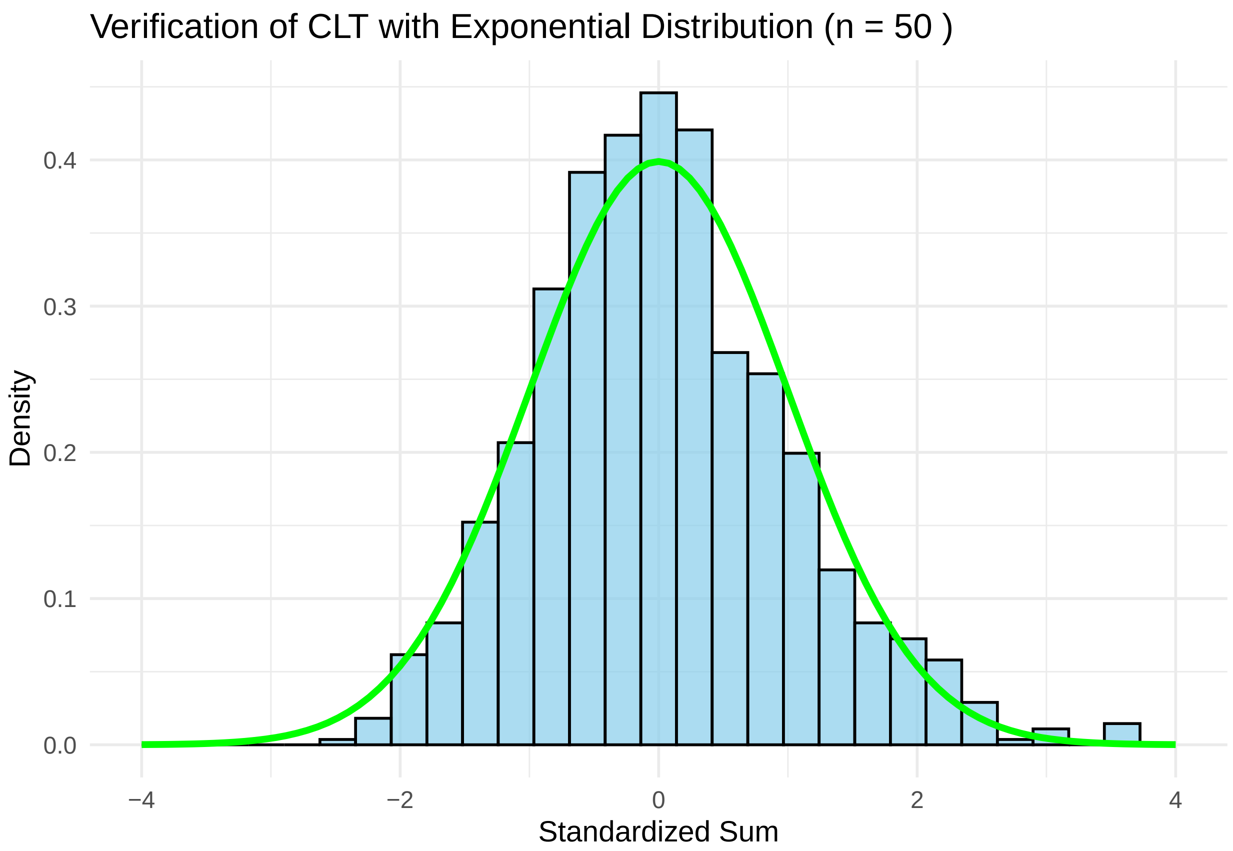
Similarly, we can define another function to verify Central Limit Theorem, using a sequence of i.i.d. continuous random variables that follow the Exponential Distribution, where $\lambda = 1$:

```r
clt_verification2 <- function(m, n, d, seed = 124) {
  # Set the random seed for reproducibility
  set.seed(seed)

  results <- numeric(m)

  for (i in 1:m) {
    X <- rexp(n, rate = 1)  # Exponential distribution with a rate of 1

    # The expectation and variance of X_i
    EX <- 1
    VarX <- 1

    # Calculate the standardized value
    results[i] <- (sum(X) - n * EX) / sqrt(n * VarX)
  }

  # Plot the histogram
  hist_data <- data.frame(results)
  p <- ggplot(hist_data, aes(x = results)) +
    geom_histogram(aes(y = after_stat(density)), bins = 30, fill = "skyblue",
                   color = "black", alpha = 0.7) +
    stat_function(fun = dnorm, args = list(mean = 0, sd = 1), color = "green",
                  linewidth = 1.2) +
    labs(title = paste("Verification of CLT with Exponential Distribution (n =", n, ")"),
         x = "Standardized Sum", y = "Density") +
    theme_minimal() +
    xlim(c(-4, 4))

  print(p)
}
```

Set $m = 1000, d = 0.2$, and $n = 50, 100, 1000$, then run the function:

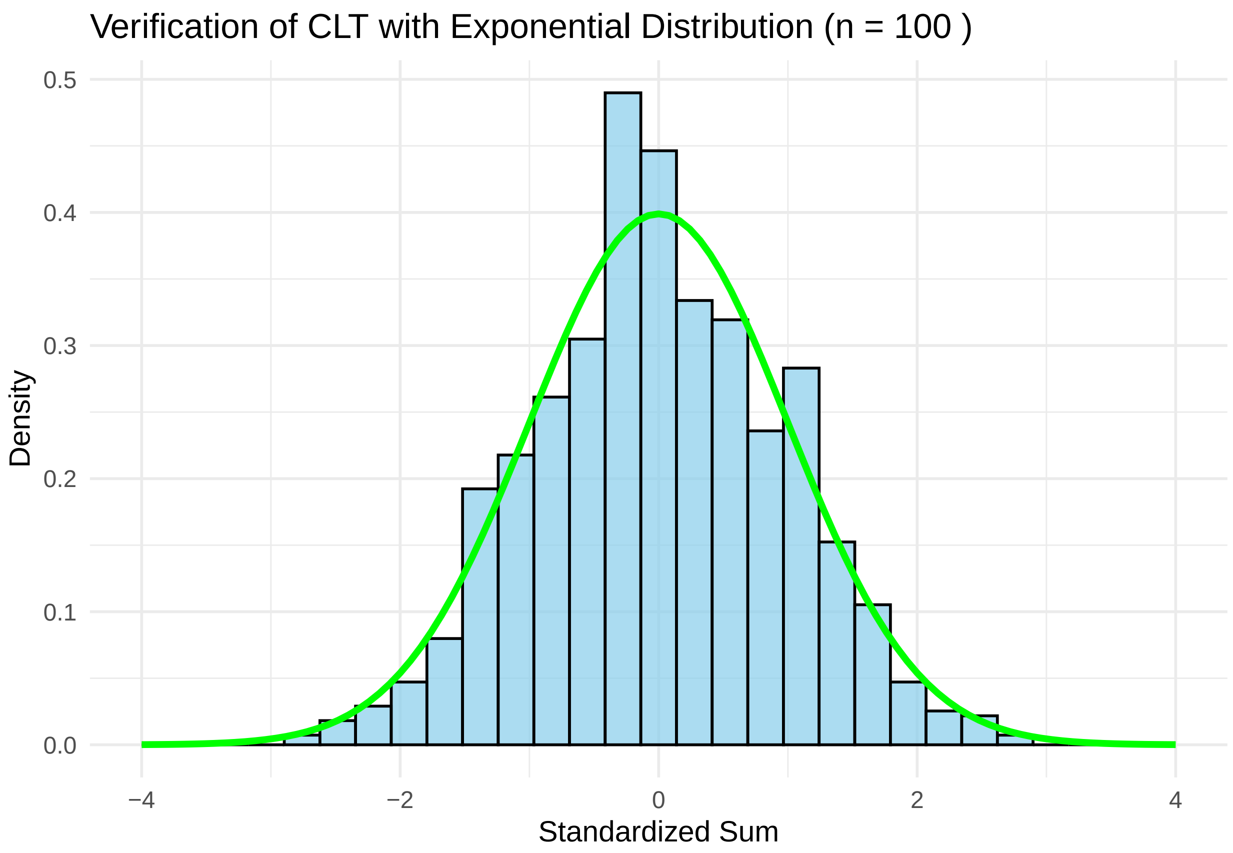
```r
clt_verification2(m = 1000, n = 50, d = 0.2)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



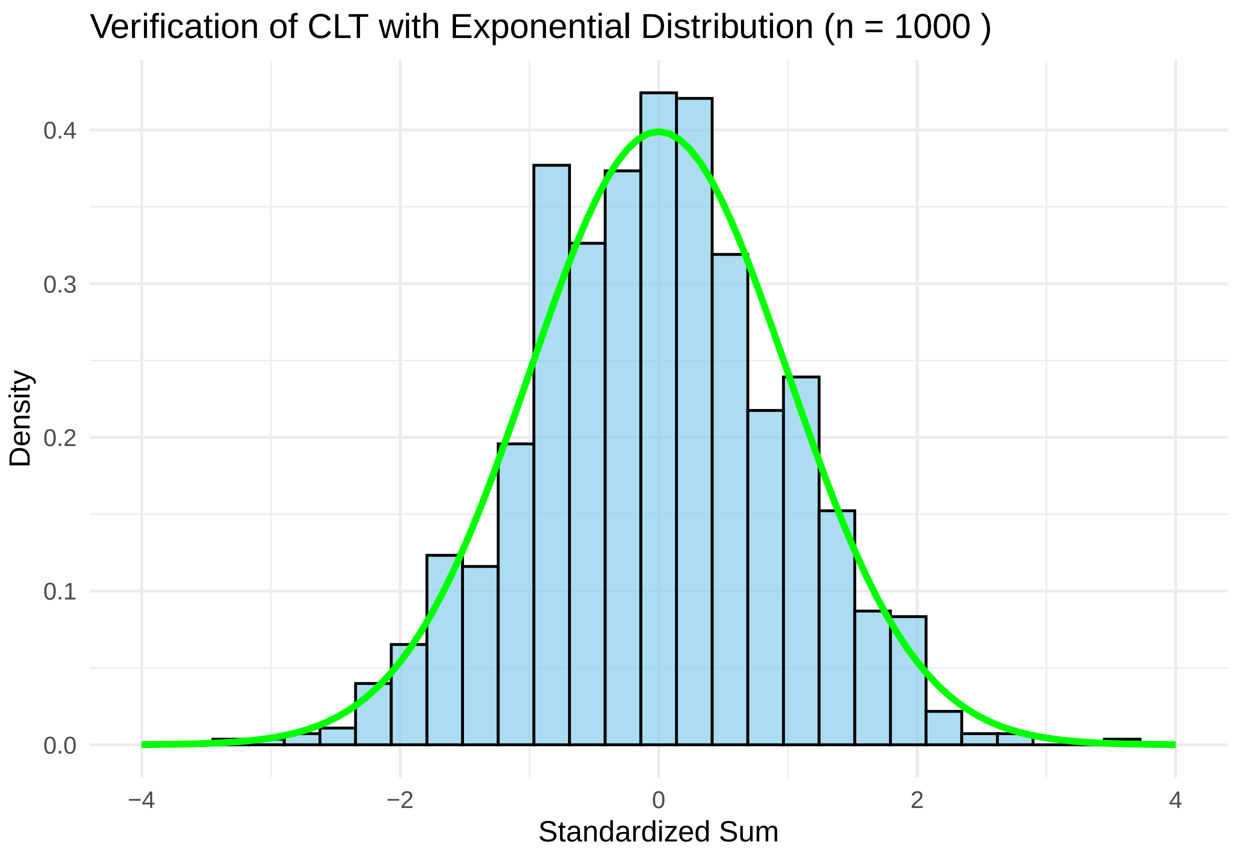
```
clt_verification2(m = 1000, n = 100, d = 0.2)
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).

## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



```r
clt_verification2(m = 1000, n = 1000, d = 0.2)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



The results are similar with the results produced by Bernoulli distribution: as n increases, the histogram converges to the standard normal distribution.

We could also perform normality test:

```r
# Set the random seed for reproducibility
set.seed(124)

results <- numeric(1000)

  for (i in 1:1000) {
    X <- rexp(1000, rate = 1)  # Exponential distribution with a rate of 1

    # The expectation and variance of X_i
    EX <- 1
    VarX <- 1

    # Calculate the standardized value
    results[i] <- (sum(X) - 1000 * EX) / sqrt(1000 * VarX)
  }
shapiro.test(results)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  results
## W = 0.99858, p-value = 0.6068
```

This time, p-value is even higher, with a value of 0.6068. Therefore, we have even more reasons to believe that the data follows normal distribution.

## Verification of the Convergence Rate of Central Limit Theorem

Furthermore, we want to verify the convergence rate of Central Limit Theorem, in a unique way. What we are trying to verify is $\frac{\sqrt{n}}{\sqrt{VarX_1}}(\overline{X}-EX_1) \xrightarrow{d} N(0,1)$. If we calculate the absolute bias of using $\overline{X}$ to estimate $EX_1$, we have:

$$\frac{\sqrt{n}}{\sqrt{VarX_1}}|\overline{X}-EX_1| \xrightarrow{d} |N(0,1)| \tag{2}$$

In other words, we can roughly say that $|\overline{X}-EX_1|$ follows the distribution: $\sqrt{VarX_1}|N(0,1)|/\sqrt{n}$, when n is large enough. Also, we know that the expectation of $|N(0,1)|$ is:

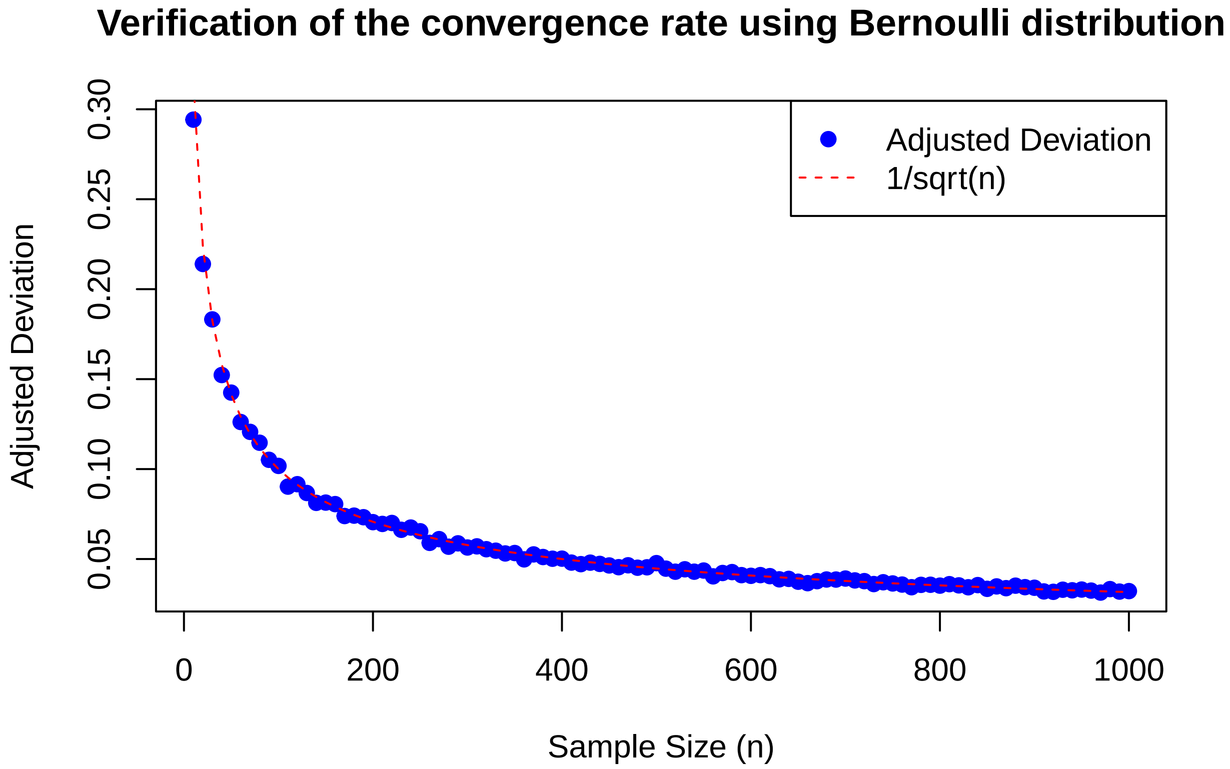
$$\int_0^\infty \frac{2x}{\sqrt{2\pi}}e^{-x^2/2}dx = \sqrt{\frac{2}{\pi}} \tag{3}$$

Therefore, as n increases, we believe that $\sqrt{\frac{\pi}{0.32}}|\overline{X}-EX_1|$ would be scattered around $1/\sqrt{n}$, if our assumption(the convergence rate of Central Limit Theorem) is indeed correct.

We can validate this using the same distributions we used earlier, the Bernoulli distribution and the exponential distribution:

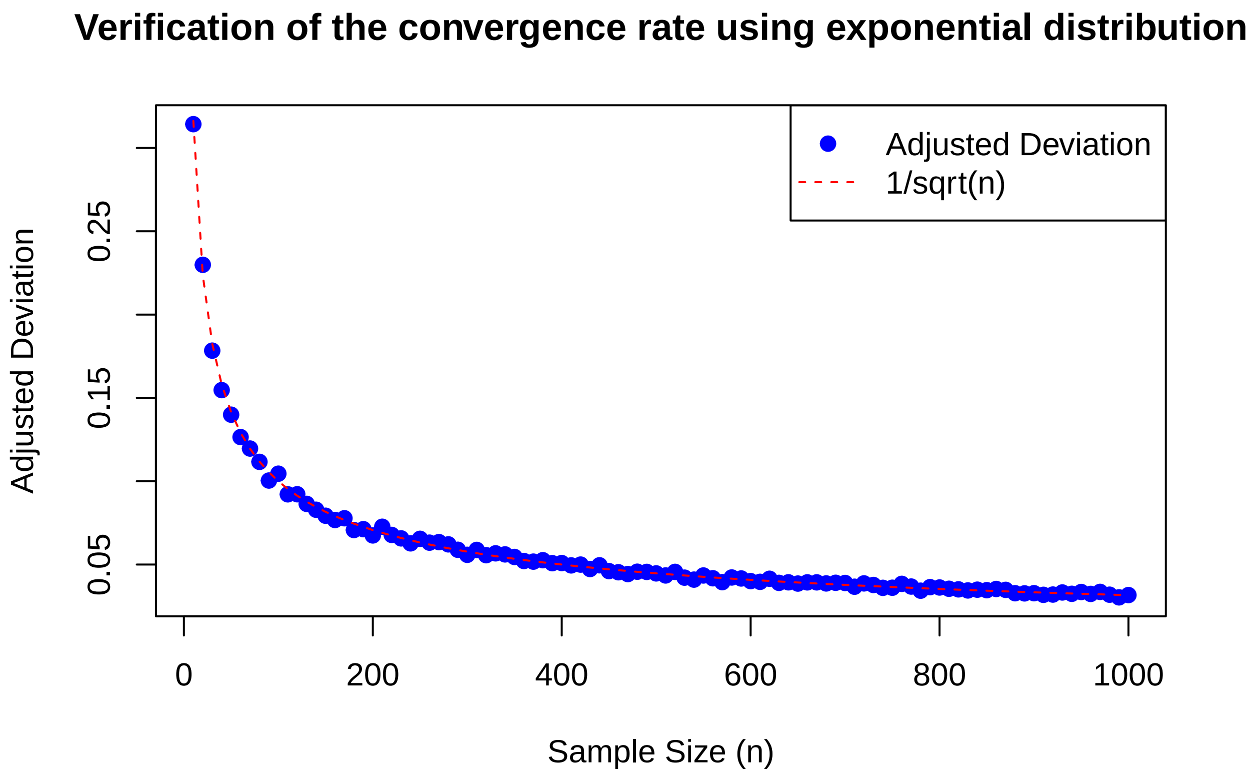
```r
set.seed(123)  # Set the random seed for reproducibility

p <- 0.2
n_simulations <- 1000
n_values <- seq(10, 1000, by = 10)
deviations <- numeric(length(n_values))

# Run the simulations
for (i in 1:length(n_values)) {
  n <- n_values[i]
  sample_means <- numeric(n_simulations)

  for (j in 1:n_simulations) {
    # Generate n i.i.d. Bernoulli random variables
    X <- rbinom(n, size = 1, prob = p)

    # Calculate the sample mean
    sample_means[j] <- mean(X)
  }

  # Calculate the absolute deviation
  deviation <- mean(abs(sample_means - p))
  deviations[i] <- deviation
}

# Adjust the absolute deviation
deviations <- deviations * sqrt(pi / 0.32)

# Calculate 1/sqrt(n) for comparison
inverse_sqrt_n <- 1 / sqrt(n_values)
```

```
plot(n_values, deviations, type = "p", col = "blue", pch = 19,
     xlab = "Sample Size (n)", ylab = "Adjusted Deviation",
     main = "Verification of the convergence rate using Bernoulli distribution")
lines(n_values, inverse_sqrt_n, col = "red", lty = 2)
abline(h = 0, col = "black", lty = 2)

legend("topright", legend = c("Adjusted Deviation", "1/sqrt(n)"),
       col = c("blue", "red"), lty = c(0, 2), pch = c(19, NA))
```



The blue dots perfectly fall on the red line, which means our assumption is correct. We can repeat this process using exponential distribution:

```
set.seed(123)  # Set the random seed for reproducibility

lambda <- 1
n_simulations <- 1000
n_values <- seq(10, 1000, by = 10)
deviations <- numeric(length(n_values))

# Run the simulations
for (i in 1:length(n_values)) {
  n <- n_values[i]
  sample_means <- numeric(n_simulations)

  for (j in 1:n_simulations) {
    # Generate n i.i.d. exponential random variables
    X <- rexp(n, rate = lambda)
    # Calculate the sample mean
    sample_means[j] <- mean(X)
  }

  # Calculate the absolute deviation
  deviation <- mean(abs(sample_means - 1))
  deviations[i] <- deviation
}

# Adjust the absolute deviation
deviations <- deviations * sqrt(pi / 2)

# Calculate 1/sqrt(n) for comparison
inverse_sqrt_n <- 1 / sqrt(n_values)

plot(n_values, deviations, type = "p", col = "blue", pch = 19,
     xlab = "Sample Size (n)", ylab = "Adjusted Deviation",
     main = "Verification of the convergence rate using exponential distribution")

lines(n_values, inverse_sqrt_n, col = "red", lty = 2)

abline(h = 0, col = "black", lty = 2)

legend("topright", legend = c("Adjusted Deviation", "1/sqrt(n)"),
       col = c("blue", "red"), lty = c(0, 2), pch = c(19, NA))
```



The result is similar to the result generated by Bernoulli distribution. In this way, we successfully verified Central Limit Theorem, as well as its convergence rate.