Apply the Central Limit Theorem in R (With Examples)

Authored by

Mohammed looti

November 1, 2025

The Foundational Role of the Central Limit Theorem (CLT)

The **Central Limit Theorem** (CLT) stands as one of the most critical and powerful concepts in modern statistics. It provides a vital theoretical bridge connecting the often messy reality of population data with the powerful, precise methods of statistical inference. Simply put, the CLT dictates that if we draw repeated, sufficiently large, random samples from virtually any population--irrespective of the population's original probability distribution (be it skewed, uniform, or bimodal)--the resulting distribution of the sample means will inexorably approach a **Normal Distribution**.

This phenomenon is profoundly important for practical data analysis because a vast majority of parametric statistical tests--such as T-tests and ANOVA--rely on the underlying assumption of normality. The CLT guarantees that even if our foundational data is highly non-normal or irregularly distributed, we can still confidently employ these robust statistical techniques, provided that our sample size is adequate. A sample size typically exceeding 30 observations ($N > 30$) is generally accepted as sufficient for this powerful approximation to hold true in most real-world scenarios.

Furthermore, the **Central Limit Theorem** does not just predict the shape of the resulting curve; it precisely defines how this new distribution of sample means, formally known as the **sampling distribution**, relates back to the original population parameters. Grasping these mathematical relationships is absolutely essential for conducting accurate hypothesis testing, determining p-values, and constructing reliable confidence intervals in research.

Defining the Sampling Distribution's Key Properties

When the necessary conditions for the Central Limit Theorem are satisfied (primarily randomness and sufficient sample size), the resulting **sampling distribution** of the sample means exhibits two mathematically predictable properties that link it directly to the characteristics of the source population:

The mean of the sampling distribution (the average of all sample means) will be identical to the true mean of the original population distribution. This ensures that the process of repeated sampling is unbiased.

The standard deviation of the sampling distribution, referred to specifically as the **standard error of the mean**, is inversely proportional to the square root of the sample size. This means greater sample sizes lead to less variability in the estimates.

These two properties guarantee that our statistical inferences are both accurate and efficient. The first property mathematically confirms that if we were to take an infinite number of samples, the average of those samples would perfectly center around the true population mean (μ).

x? = μ

The second property is crucial for quantifying precision. It explains why a larger sample is always preferred: as the sample size (n) increases, the variability (or spread) of the sample means decreases, leading to far more precise and reliable estimates of the unknown population mean.

s = σ / √n

Simulating the CLT in R: The Turtle Shell Example

To truly internalize the mechanics and profound implications of the Central Limit Theorem, the most effective approach is to perform a simulation. We will use the statistical programming language **R** to demonstrate how repeated sampling can transform data from a non-normal population into a perfectly behaving, normally distributed sampling distribution.

For this practical illustration, imagine we are collecting measurements of the shell width of a specific turtle species. Crucially, we hypothesize that the shell widths follow a **Uniform Distribution**. This specific distribution implies that any shell width between the minimum and maximum possible values is equally likely to occur, creating a distinctly flat, rectangular shape when plotted.

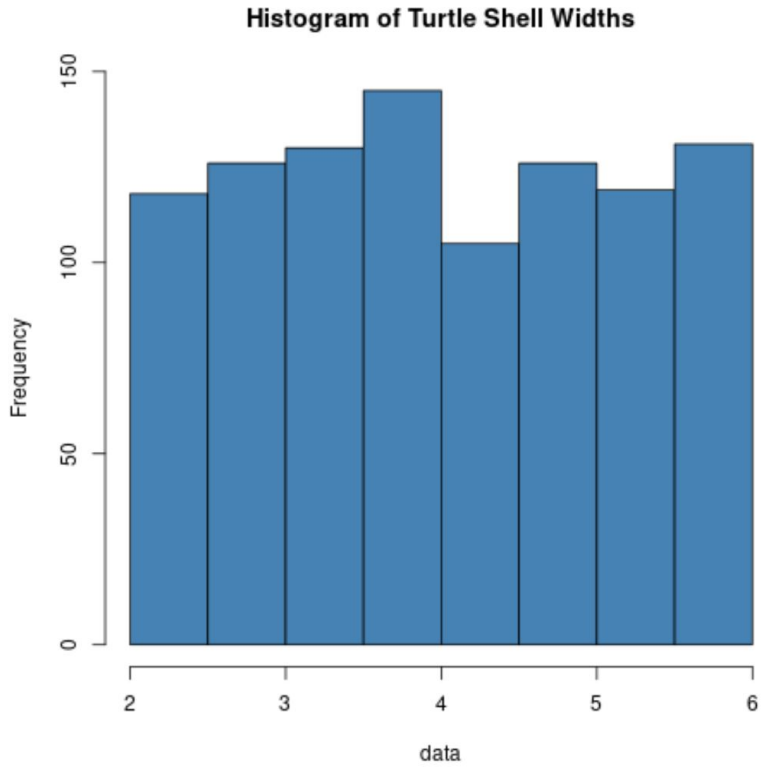
Let us establish our population parameters: the minimum shell width is 2 inches, and the maximum is 6 inches. If we randomly select one turtle, its shell width is just as likely to be 2.5 inches as it is to be 5.5 inches. Because this starting distribution is highly non-normal, it serves as an ideal and robust test case to observe the CLT in action.

Establishing the Non-Normal Population (n=1000)

Our first step in the R simulation is to generate a large dataset that accurately represents our population of turtle shell widths. We will create 1,000 measurements drawn from a uniform distribution spanning 2 to 6 inches. The following **R** code snippet executes this population generation and visually confirms its distribution using a histogram.

```
#make this example reproducible
set.seed(0)

#create random variable with sample size of 1000 that is uniformally distributed
data <- runif(n=1000, min=2, max=6)

#create histogram to visualize distribution of turtle shell widths
hist(data, col='steelblue', main='Histogram of Turtle Shell Widths')
```

The generated visualization confirms our initial hypothesis: the population distribution of turtle shell widths is emphatically non-normal. Its characteristic flat appearance perfectly matches the properties of a uniform distribution, showing no central tendency or bell shape whatsoever.



Initial Sampling Distribution Analysis (Small Sample Size: n=5)

The next phase of the simulation is the core demonstration of the CLT. We will systematically take repeated random samples from our non-normal population. For this initial test, we select a relatively small sample size: **n = 5**. We then repeat this sampling process 1,000 times, calculating the mean shell width for each individual sample and storing these 1,000 means in a new vector.

This collection of 1,000 sample means constitutes the **sampling distribution**. The R code below executes this complex simulation, computes the basic descriptive statistics for the sample means, and generates a new histogram to visualize the shape transformation.

```
#create empty vector to hold sample means
sample5 <- c()

#take 1,000 random samples of size n=5
n = 1000
for (i in 1:n){
```

```
sample5 = mean(sample(data, 5, replace=TRUE))
}

#calculate mean and standard deviation of sample means
mean(sample5)
```

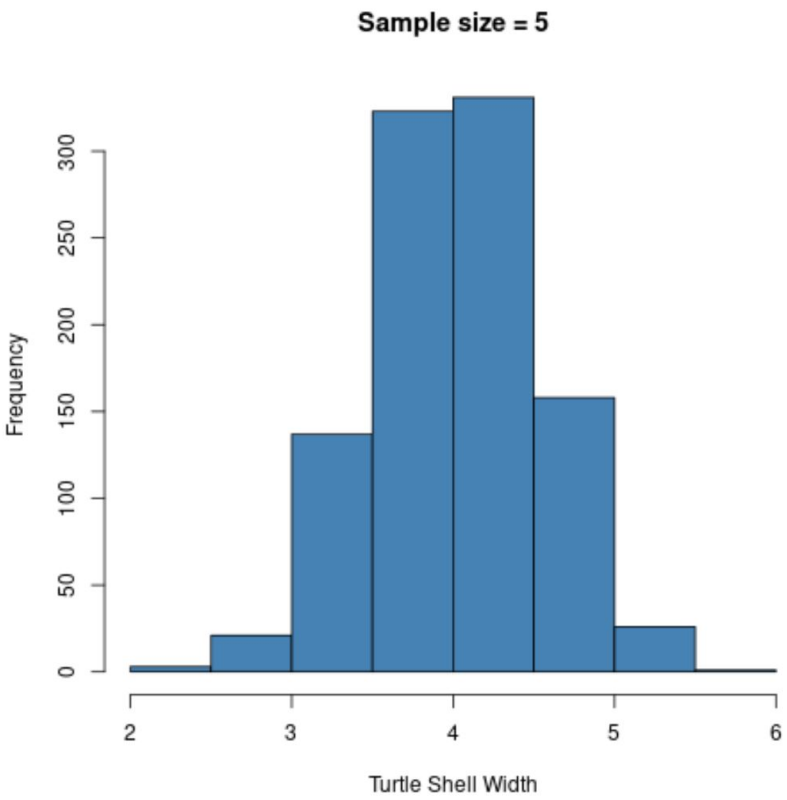
4.008103

```
sd(sample5)
```

0.5171083

```
#create histogram to visualize sampling distribution of sample means
hist(sample5, col ='steelblue', xlab='Turtle Shell Width', main='Sample size = 5')
```

A remarkable transformation is immediately apparent in the resulting histogram. Even with a small sample size (n=5), the distribution of the sample means has shed its uniform, rectangular shape and begun to form a distinct bell-shaped curve, moving decisively toward the theoretical **Normal Distribution**.



The calculated descriptive statistics for this distribution clearly illustrate the first two properties of

the CLT:

Sample Mean (x?): 4.008, which is extremely close to the true theoretical population mean of 4.0 (calculated as (6+2)/2).

Standard Error (s): 0.517, quantifying the current variability or precision of our estimations.

Demonstrating Precision: The Impact of Increased Sample Size (n=30)

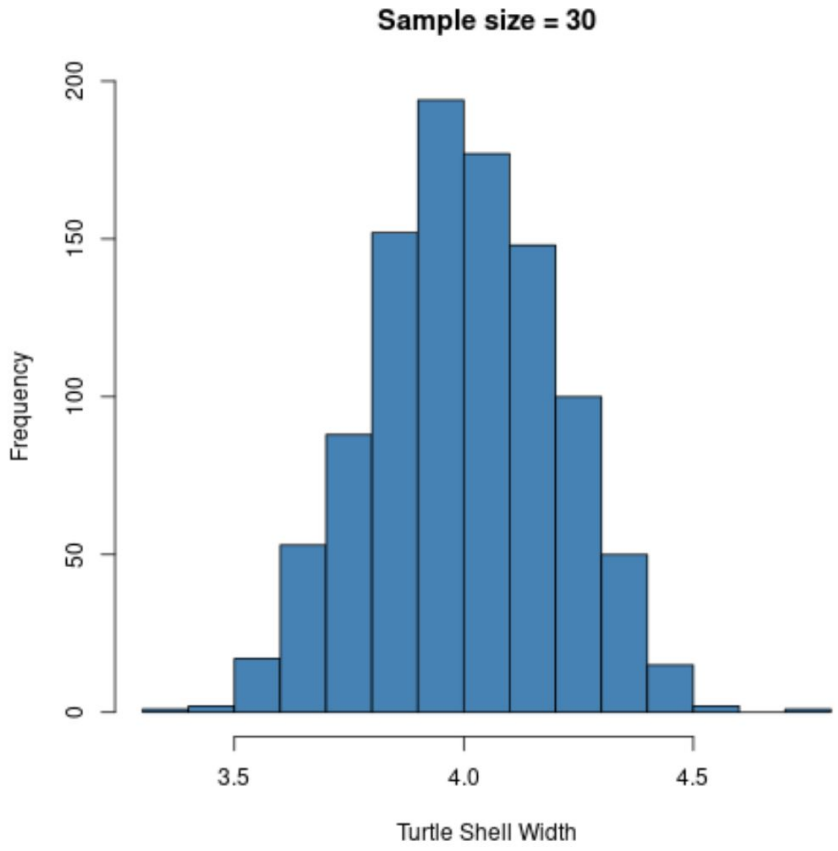
The true power of the **Central Limit Theorem** is realized when the sample size increases. The theory predicts two critical changes: first, the sampling distribution will become even more perfectly normal; and second, the standard error of the mean will decrease significantly, indicating a tighter clustering of means around the true population parameter. We now repeat the entire simulation, increasing the sample size dramatically to the conventional threshold of **n = 30**.

By increasing the sample size fivefold, we anticipate a distribution that is much narrower, more tightly clustered around the true mean (4.0), and possesses a substantially smaller standard deviation. This reduction in variability translates directly into far greater confidence and precision in any statistical inference drawn from the data.

```
#create empty vector to hold sample means
sample30 <- c()

#take 1,000 random samples of size n=30
n = 1000
for (i in 1:n){
sample30 = mean(sample(data, 30, replace=TRUE))
}

#calculate mean and standard deviation of sample means
mean(sample30)

4.000472

sd(sample30)

0.2003791

#create histogram to visualize sampling distribution of sample means
hist(sample30, col ='steelblue', xlab='Turtle Shell Width', main='Sample size = 30')
```

The resulting histogram emphatically confirms the theoretical predictions. The distribution is now

extremely narrow and highly concentrated, centering almost perfectly at the theoretical population mean of 4.0. Visually, this curve is significantly closer to a perfect **Normal Distribution** than the previous simulation with n=5.



The calculated statistics underscore the impact of increased sample size:

New Standard Error (s): 0.200

This standard error (0.200) represents a massive reduction in variability compared to the previous standard error (0.517). This powerful tightening of the distribution is the mathematical justification for why increasing sample size leads to significantly more accurate and reliable estimates in statistical research.

Conclusion and Practical Inferential Statistics

These simulations provide compelling empirical evidence for the Central Limit Theorem. We confirmed that regardless of the initial population distribution--which was the highly non-normal uniform distribution of turtle shell widths--the distribution of repeated sample means rapidly converges to a **Normal Distribution** as the sample size increases. Furthermore, we demonstrated that larger sample sizes directly result in smaller standard errors, translating to estimates of the

population mean that are both more accurate and statistically reliable.

This principle is the cornerstone of inferential statistics. It empowers researchers to move beyond simply describing sample data and to draw far-reaching conclusions about vast, often unknown, populations. By utilizing the CLT, researchers can confidently conduct rigorous statistical tests and build highly precise models, even when the underlying characteristics of the population are complex or entirely hidden from view.

The following resources offer further deep dives into the theoretical and practical application of the Central Limit Theorem: