

How to Easily Apply the Central Limit Theorem in R

Authored by
stats writer

December 1, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Easily Apply the Central Limit Theorem in R*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=103448>

The Central Limit Theorem (CLT) stands as one of the most fundamental and profound concepts within the field of statistics. It provides crucial insights into the behavior of sample means, particularly demonstrating how these means converge toward a normal distribution, even when the underlying population distribution is decidedly non-normal, provided the sample size is sufficiently large. Implementing the power of the CLT requires a systematic approach, especially when utilizing statistical software like R.

To practically apply the CLT in R, the process generally involves several key steps. First, we must simulate or generate a large sample from a defined population. This population can follow any distribution—whether it be uniform, exponential distribution, or highly skewed. Second, we repeatedly draw smaller subsamples from this population and calculate the mean for each subsample. This iterative process generates a collection of sample means. Finally, by analyzing the distribution of these calculated means, we observe the characteristic bell curve shape that approximates the population mean, validating the theorem.

For instance, consider a scenario where we generate a sample of size 10 from the exponential distribution. Even though the exponential distribution is heavily skewed, calculating the means of numerous samples of size 10 and plotting their distribution will start to show the emergence of a normal shape centered around the true population mean. The subsequent detailed example will use a different, concrete illustration--the uniform distribution--to demonstrate this transformation visually within the R environment.

The Fundamental Principles of CLT

The core assertion of the Central Limit Theorem is elegantly simple yet profoundly impactful: regardless of the original shape of the population distribution, the sampling distribution of the sample mean will tend toward a normal distribution as the sample size increases. This tendency is typically considered robust when the sample size (n) is greater than 30, although convergence might happen earlier depending on the symmetry of the original population.

This remarkable property allows statisticians to make reliable inferences about population parameters using sample data, even when the population distribution is unknown or complex. The predictability provided by the CLT is fundamental to many advanced statistical methods, including hypothesis testing and the construction of confidence intervals. Understanding these underlying principles is paramount before diving into the computational implementation using R.

Furthermore, the Central Limit Theorem defines two critical properties that characterize the resulting sampling distribution:

The central limit theorem also states that the sampling distribution will have the following properties:

The mean of the sampling distribution will be precisely equal to the mean of the population distribution (μ):

$$\bar{x} = \mu$$

The standard deviation of the sampling distribution, often referred to as the standard error, will be equal to the population standard deviation (σ) divided by the square root of the sample size (n):

$$s = \sigma / \sqrt{n}$$

The second property is particularly important, as it explains why larger sample sizes lead to less variability in the sample means. As n increases, the standard error decreases, leading to a tighter, more peaked normal distribution around the true population mean. The following practical example vividly demonstrates how these theoretical concepts manifest when applied using statistical computing in R.

Case Study Setup: Modeling Turtle Shell Widths (Uniform Distribution)

Let us establish a practical scenario to illustrate the Central Limit Theorem. Suppose we are studying a population of turtles, and we are interested in the width of their shells. We hypothesize that the shell width follows a continuous uniform distribution with a minimum width of 2 inches and a maximum width of 6 inches. This means that if we randomly select any single turtle, the probability of its shell width being 2.5 inches is exactly the same as the probability of it being 5.5 inches, or any width in between.

In a real-world setting, this distribution is distinctly non-normal--it looks like a flat rectangle when plotted. This provides an ideal foundation for testing the CLT, as we are starting far away from a normal shape. We must first generate a sufficiently large dataset representing our entire population of measurements. The chosen population size is 1,000 turtles, ensuring we have enough data points to represent the underlying uniform distribution accurately.

The following R code block demonstrates how to create this initial dataset, ensuring the example is reproducible by setting a seed, and then visualizing the raw population distribution using a histogram. We use the `runif()` function in R, which is specifically designed to generate random numbers from a uniform distribution given minimum and maximum bounds:

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

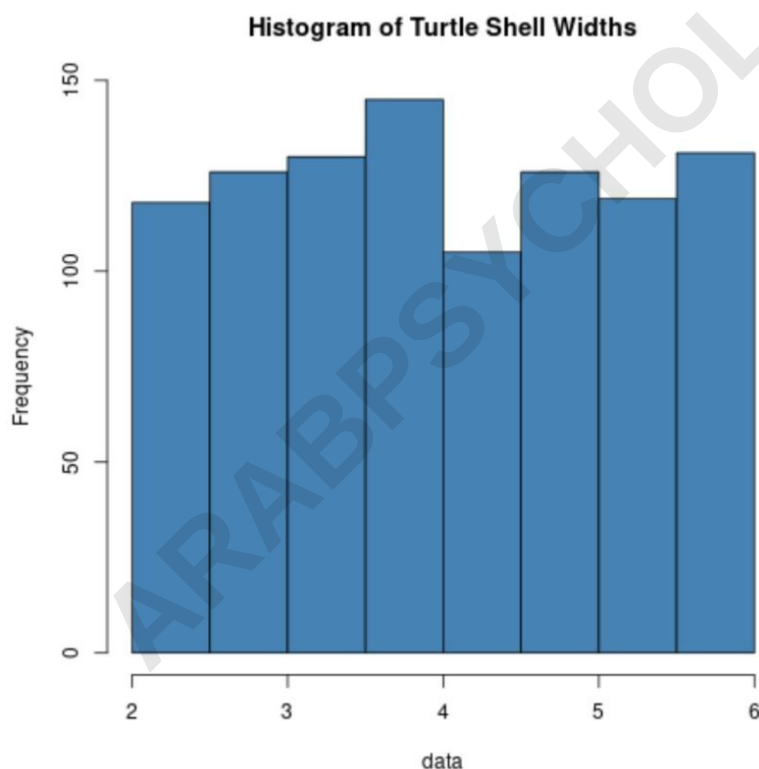
```
#create random variable with sample size of 1000 that is uniformly distributed  
data <- runif(n=1000, min=2, max=6)
```

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

Visualizing the Population Distribution (The Initial Sample)

Executing the code above yields a visual representation of our simulated population. The resulting histogram confirms the characteristics of the uniform distribution, where the frequency bars are roughly equal across the range of 2 to 6 inches, resulting in a shape that is clearly not Gaussian or bell-shaped. This confirmation of the non-normality of the initial population is vital, as it sets the stage for demonstrating the CLT's transformative power.

The image below illustrates this initial distribution. Notice the lack of a peak or central tendency; instead, the data is spread evenly across the defined interval. This visual confirms that any statistical tests relying on the assumption of normality would be inappropriate if applied directly to this raw data.



The core insight of the Central Limit Theorem hinges on the idea that even when the raw data is distributed like this flat rectangle, the distribution of the means of small subsets drawn from it will begin to coalesce into a recognizable normal curve. Our next step is to simulate the repeated sampling process required to observe this critical phenomenon in action.

Simulating the Sampling Distribution (Small Sample Size: $n=5$)

The next phase involves simulating the sampling process. Instead of measuring all 1,000 turtles, imagine we repeatedly take small, random samples of only 5 turtles ($n=5$) from the population and calculate the mean shell width for each group. We will repeat this process 1,000 times to generate a sufficient number of sample means to form a sampling distribution. This simulated distribution represents the empirical evidence for the CLT.

We expect that even with a small sample size like $n=5$, the distribution of these 1,000 sample means will begin to show a central peak and tails, deviating significantly from the flat population distribution we observed previously. The transformation occurs because extreme means (very high or very low) are less likely to occur consistently across multiple random samples than means closer to the population average.

The following R code snippet executes this simulation. We initialize an empty vector, loop 1,000 times, and within each iteration, we draw a sample of size 5 using replacement, calculate its mean, and store the result. We then calculate the statistical characteristics (mean and standard deviation) of this new distribution of means and visualize it with a histogram:

```
#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

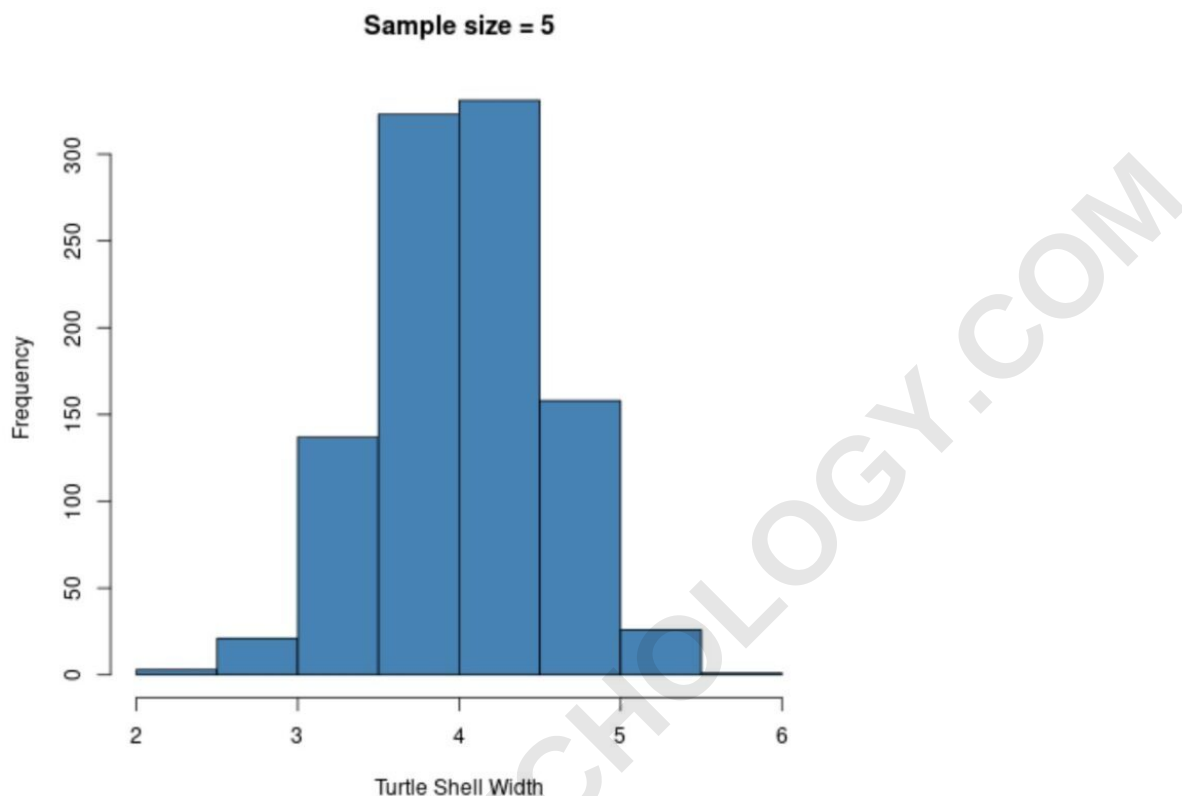
sd(sample5)

0.5171083

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

Upon reviewing the output, we immediately notice the emergence of a bell-shaped curve, confirming that the sampling distribution of sample means now approximates a normal distribution,

despite the original data being uniformly distributed. This is the hallmark outcome of the Central Limit Theorem. The distribution is centered close to 4.0, which is the theoretical mean (or expected value) of the uniform distribution between 2 and 6.



Key statistics derived from this simulation show:

Sample Mean ($\bar{x}$): 4.008 (Extremely close to the expected population mean, $\mu=4$).

Sample Standard Deviation (s): 0.517 (This represents the standard error of the mean for $n=5$).

The standard deviation here, 0.517, is a crucial metric, quantifying the spread of the sample means. Our next step is to examine how increasing the sample size impacts this measure of variability, further emphasizing the power of the CLT.

Analyzing the Impact of Increased Sample Size ($n=30$)

The Central Limit Theorem suggests that as the sample size (n) increases, the sampling distribution becomes more perfectly normal, and its standard error significantly decreases. To test this hypothesis, we now repeat the entire simulation process, but this time we increase the sample size from $n=5$ to $n=30$. The threshold of $n \geq 30$ is often cited in statistics as the point where the approximation to normality becomes highly reliable.

By increasing the sample size to 30, we are introducing more data points into each mean calculation. This inevitably reduces the random fluctuation in the means, causing them to cluster much more tightly around the true population mean ($\mu=4$). We anticipate that the shape will become even more sharply peaked, and the standard deviation (standard error) will be substantially smaller than the 0.517 observed with $n=5$.

The following R code repeats the simulation, calculating 1,000 means based on samples of size $n=30$, and then plots the corresponding histogram:

```
#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
```

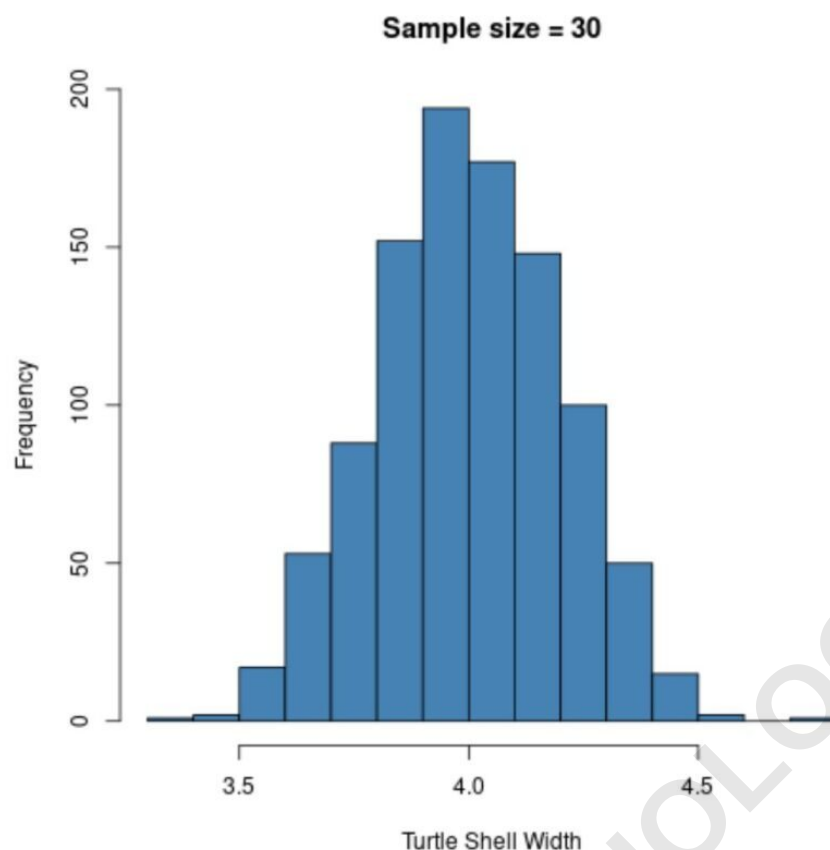
```
sd(sample30)
```

```
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 clearly displays a much narrower and taller peak compared to the $n=5$ simulation. This visual evidence confirms that the increase in sample size drastically reduces the variability of the sample means, making our estimate of the population mean more precise.



The distribution of means is now centered almost perfectly at 4.000, and the sample standard deviation has decreased dramatically:

Sample Standard Deviation (s): 0.200 (A significant reduction from 0.517).

This reduction from 0.517 to 0.200 directly demonstrates the mathematical relationship defined by the CLT: the standard error (s) is inversely proportional to the square root of the sample size (n). By increasing n, we reduce the standard error, leading to a much more reliable and concentrated estimate of the true population mean. This property is what makes the Central Limit Theorem so invaluable in statistical inference.

Interpreting the Results and Conclusion

The simulations performed in R provide empirical proof of the Central Limit Theorem. We began with a highly non-normal, continuous uniform distribution representing turtle shell widths. Through the process of repeated random sampling and calculating the mean of those samples, we observed a consistent and rapid convergence to a normal distribution.

The key takeaways from this demonstration are twofold. First, the distribution of sample means quickly adopts the shape of a bell curve, validating the CLT's promise of approximating normality,

even with relatively small sample sizes ($n=5$). Second, as the sample size increases (from $n=5$ to $n=30$), the sampling distribution tightens significantly, confirming the relationship between sample size and the standard error. Larger samples yield more precise estimates of the population mean.

This capability is why the CLT is foundational to inferential statistics. It grants us the statistical leverage to analyze the mean of virtually any population distribution, provided we collect enough data in our sample. When working with large samples, we can safely assume that the distribution of sample means is normal, allowing us to apply powerful parametric statistical tests that rely on this normality assumption, regardless of the underlying population structure.

Further Resources on the Central Limit Theorem

For readers seeking deeper theoretical understanding or alternative applications of the Central Limit Theorem, the following external resources provide comprehensive information: