# Simulation-Based Assessment of Central Limit Theorem Convergence
# STATS 506 Final Project

Lingzhi Hao

# 1 Introduction

The Central Limit Theorem (CLT) states that, for an underlying distribution with a finite mean $\mu$ and a finite variance $\sigma^2$, the standardized sample mean $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges in distribution to a normal random variable as the sample size $n$ increases. Although $n \geq 30$ is usually considered large enough, the convergence rate varies substantially across different distributions, especially in the presence of skewness, heavy tails, and multimodality.

# 2 Research Question

The primary research question of this project is: For different underlying distributions, how large must the sample size $n$ be for the CLT to apply? I also examined how convergence differs across distributions and which distributional characteristics affect the convergence rate.

# 3 Methods

I conducted Monte Carlo simulations to evaluate the convergence of the standardized sample mean $Z$ to normality across different underlying distributions.

## 3.1 Selection of Distributions

I selected seven distributions covering key distributional features impacting the convergence under CLT, including boundedness, skewness, symmetry, discreteness, heavy tails, and multimodality. The distributions were uniform, chi-square, log-normal, Poisson, Student's $t$, Bernoulli, and a Gaussian mixture distribution.

## 3.2 Simulations

For each distribution and each sample size $n \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 100\}$, I generated independent random samples and computed the sample mean $\bar{X}$. I conducted $10,000$ Monte Carlo repetitions for this procedure to approximate the sampling distribution of $\bar{X}$. Then the sample mean $\bar{X}$ was standardized using the true mean $\mu$ and variance $\sigma^2$ set for the underlying distribution to compute

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

For distributions like the uniform distribution, where the convergence occurs at a very small sample size, I refined the sample size grid to include smaller $n$ to identify the smallest sample size $n^*$ more precisely. For distributions whose shapes are invariant to parameter change, such as the uniform distribution, I used a single parameter setting. For other distributions, I set different parameters to generate different shapes and examine how the shape affects convergence.

## 3.3 Convergence Assessment

Convergence to normality was assessed using three complementary metrics, which are the Anderson-Darling (A-D) test for normality, the absolute skewness, and the absolute excess kurtosis of the standardized sample mean. Instead of using a fixed threshold for the A-D statistic, I assessed normality by comparing the corresponding $p-$value to a 5% significance level. To avoid excessive sensitivity of the A-D test at large Monte Carlo sample sizes, the test was applied to a fixed-size random subsample of

500 standardized statistics for each $n$. The smallest sample size $n^*$ is the smallest $n$ for which all the following three conditions are simultaneously satisfied:

1. $p \geq 0.05$;
2. Absolute Skewness $|\gamma_1| \leq 0.1$;
3. Absolute Excess Kurtosis $|\gamma_2| \leq 0.2$.

In addition, histograms and Q-Q plots of the standardized sample mean $Z$ at sample size $n^*$ were plotted to visually confirm the adequacy of the normal approximation.

## 4 Results

The estimated smallest sample size $n^*$ varied substantially across distributions, reflecting the strong influence of distributional characteristics on the convergence rate of the CLT. Table 1 summarizes the $n^*$ values for all distributions and parameter settings considered.

As Table 1 shows, bounded or symmetric distributions like uniform distributions have fast convergence, requiring a very small sample size ($n^* < 10$) to apply CLT. Discrete distributions such as Poisson distributions also converge relatively quickly when the $\lambda$ is large, consistent with decreasing skewness as the distribution becomes more symmetric.

Distributions with high skewness or heavy tails require larger sample sizes. Chi-square and log-normal distributions have increasing $n^*$ as skewness becomes higher, highlighting the sensitivity of CLT convergence to tail behavior. The Student's $t$ distribution illustrates the impact of heavy tails, with smaller degrees of freedom leading to much slower convergence.

For the Bernoulli distribution, even for the $p = 0.5$, which should have the fastest convergence, $n^* = 160$ is still very large. The histogram in Figure 1 of the standardized sample mean remains visibly discrete due to the binary nature of the underlying data. However, the Q-Q plot is closely aligned with the normal line, indicating that convergence has already occurred. This example demonstrates that CLT convergence does not require the sampling distribution to be smooth, particularly for discrete distributions.

For the Gaussian mixture distribution, I focused on the multimodality. Convergence becomes slower as the mixture becomes more imbalanced. But when $w_1$ is large enough (i.e., $w_1 = 0.99$), the distribution turns out to be dominated by one single normal distribution, making the convergence much faster.

## 5 Conclusion

This project investigated the convergence of CLT across several distributions using Monte Carlo simulations. The results demonstrate that the sample size required for a reasonable normal approximation varies widely across distributions and depends strongly on distributional characteristics such as skewness, tails, discreteness, and multimodality. However, there is no universal $n^*$ to apply across all kinds of distributions or even across different parameter settings of the same distribution. These findings show that CLT and asymptotic normal approximations should be applied with careful consideration of the underlying distribution.

| NO. | **Distribution** | **Characteristics** | $n^*$ |
|---|---|---|---|
| 1 | Uniform | Bounded, symmetric | 7 |
| 2 | Chi-square ($df$) | Highly skewed, continuous, fixed skew | $df = 6$, $n^* = 70$<br>$df = 10$, $n^* = 80$<br>$df = 20$, $n^* = 35$<br>$df = 30$, $n^* = 30$<br>$df = 50$, $n^* = 15$ |
| 3 | Log-normal ($\mu = 0$, $\sigma$) | Continuous, variable skew | $\sigma = 0.01$, $n^* = 5$<br>$\sigma = 0.1$, $n^* = 15$<br>$\sigma = 0.2$, $n^* = 40$<br>$\sigma = 0.3$, $n^* = 70$<br>$\sigma = 0.5$, $n^* = 150$ |
| 4 | Poisson ($\lambda$) | Discrete, variable skew | $\lambda = 1$, $n^* = 100$<br>$\lambda = 3$, $n^* = 45$<br>$\lambda = 5$, $n^* = 20$<br>$\lambda = 15$, $n^* = 10$ |
| 5 | Student's $t$ ($\nu$) | Heavy-tailed | $\nu = 4$, $n^* = 100$<br>$\nu = 5$, $n^* = 25$<br>$\nu = 6$, $n^* = 10$<br>$\nu = 10$, $n^* = 10$ |
| 6 | Bernoulli ($p$) | Discrete, skewed | $p = 0.5$, $n^* = 160$ |
| 7 | Gaussian mixture ($w_1$, $\mu_1$, $\mu_2, \sigma_1, \sigma_2$) $\mu_1 = -\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$ | Multiple peaks, continuous | $w_1 = 0.5$, $n^* = 10$<br>$w_1 = 0.7$, $n^* = 40$<br>$w_1 = 0.8$, $n^* = 80$<br>$w_1 = 0.9$, $n^* = 200$<br>$w_1 = 0.99$, $n^* = 25$ |

Table 1: Selected distributions, their characteristics, and estimated smallest sample sizes $n^*$.

# 6 Appendix

GitHub Repository: `https://github.com/Lingzhi-Hao/STATS-506-Final-Project`

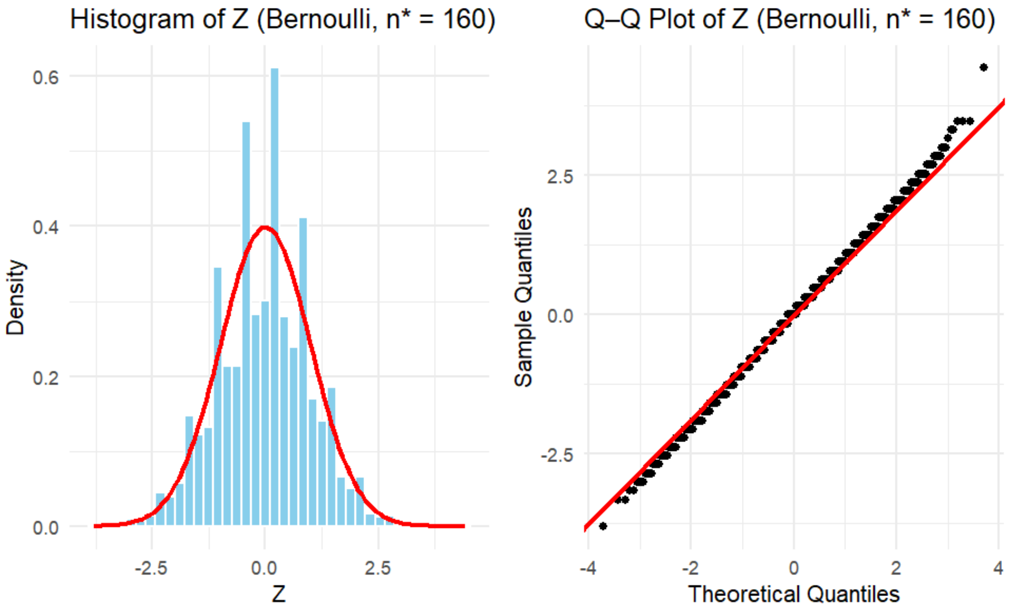

Figure 1: Histogram and Q-Q plot of Z for Bernoulli distribution