

The Interplay Between Skewness and Sample Size in the Convergence of Sampling Distribution to Normality

Amrit Subedi

amritsubedi@usf.edu

University of South Florida <https://orcid.org/0009-0001-4311-4003>

Short Report

Keywords: Central Limit Theorem, sample size, skewness, kurtosis, simulation

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11244616/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

The Central Limit Theorem (CLT) ensures asymptotic normality, but the convergence rate is highly affected by population skewness. Common rules of thumb (e.g., $n \geq 30$) are often inadequate for asymmetric data. This study quantifies this relationship using a wide range of simulations from the Beta distribution family. We derive a conservative quadratic regression model,

$n_{cons} \approx \exp \left(1.59 + 1.41 |\gamma| + 0.61 |\gamma|^2 \right)$, to predict the minimum sample size required for normality as a function of skewness ($|\gamma|$). Validation against hold-out distributions reveals that while the model is robust for continuous, moderate-kurtosis data, convergence is significantly delayed by discreteness (lattice effects) and severe excess kurtosis, requiring sample sizes beyond standard recommendations.

1. Introduction

The Central Limit Theorem (CLT) is a foundation of statistical inference [1], yet its practical application is often governed by the "rule of thumb" that a sample size of $n \geq 30$ is sufficient for normality [3]. Theoretical work, such as the Berry-Esseen Theorem, has long established that the rate of convergence is bounded by higher-order moments, particularly skewness [2].

However, applied researchers rarely have a quantitative tool to estimate the specific n required for their data's skewness, often relying on the static $n = 30$ heuristic even when data are heavily skewed [4]. This reliance on a static threshold can lead to substantial Type I error inflation in hypothesis testing and inaccurate coverage probabilities in confidence intervals. This study bridges the gap between theory and practice by empirically modeling the relationship between population asymmetry and the convergence threshold. By deriving a formula dependent on skewness, we provide a more robust alternative to the static rule.

2. Methodology

We employed a Monte Carlo simulation using the Beta distribution family (α, β) due to its bounded support $[0,1]$ and flexibility in generating shapes ranging from symmetric ($\alpha = \beta$) to highly skewed ($|\gamma| \approx 1.89$) by varying shape parameters.

2.1. Simulation Design

We generated populations from all combinations of shape parameters $\alpha, \beta \in \{1, 2, \dots, 50\}$. For each population, 1,000 random samples were drawn at varying sizes ($n = 5, 10, \dots, 1000$). For each n , the empirical sampling distribution of the mean was assessed for normality.

2.2. Assessment of Convergence

Normality was formally tested using the Shapiro-Wilk test (W), which is known for its high power in detecting deviations from normality in small-to-moderate samples [5]. The convergence threshold was defined as the minimum sample size at which the Shapiro-Wilk test failed to reject the null hypothesis of normality at the $\alpha = 0.05$ significance level ($p > 0.05$). This conservative criterion ensures that the sampling distribution is statistically indistinguishable from a Gaussian distribution.

3. Results

3.1. The Myth of $\text{varvec}n \geq 30$

To explicitly test the validity of the common rule of thumb, we examined the sampling distributions at the static threshold $n = 30$ for two distinct populations: the symmetric **Beta(2,2)** distribution and the highly skewed **Exponential(1)** distribution ($\gamma = 2.0$).

Table 1
Normality test results at static sample size $n = 30$.

Distribution	Skewness	p-value (n = 30)	Result
Beta(2,2) (Symmetric)	0.00	0.5458	Pass
Beta(5,2) (Mild Skew)	0.60	0.7096	Pass
Beta(1,9) (High Skew)	1.47	0.0021	Fail
Exponential(1) (Severe Skew)	2.00	< 0.0001	Fail

As shown in Table 1, the symmetric Beta(2,2) distribution passes the normality test easily at $n = 30$ ($p = 0.5458$). In sharp contrast, the Exponential distribution fails significantly ($p < 0.0001$), rejecting the assumption of normality.

3.2. Baseline Convergence Patterns

Having established that $n = 30$ is inadequate for skewed data, we determined the approximate sample size required for convergence across a spectrum of distributions. Table 2 summarizes the minimum n required for approximate normality.

Table 2

Minimum sample size (n) required for convergence to normality.

Distribution	Skewness ($ \gamma $)	Min. n	p -value
Normal (0,1)	0.00	5	0.7914
Beta (2,2) (Symmetric)	0.00	10	0.3485
Beta (2,5) (Mild Skew)	0.60	30	0.3654
Beta (1,9) (High Skew)	1.47	100	0.1271
Poisson (5)	0.45	30	0.1540
Chi-squared (4)	1.41	100	0.1353
Exponential (1)	2.00	200	0.0827

The results indicate that convergence is a function of asymmetry. While symmetric distributions converge rapidly ($n \leq 10$), the Exponential distribution requires $n \approx 200$, nearly seven times the standard heuristic recommendation.

3.3. Regression Model Development

Figure 1 illustrates the relationship between population asymmetry and the convergence threshold. Each point represents a simulated Beta distribution characterized by its skewness and the minimum sample size where the Shapiro-Wilk test fails to reject normality ($p > 0.05$). The trend reveals a clear, non-linear dependency: symmetric distributions converge with $n < 20$, whereas moderately skewed distributions ($0.5 < |\gamma| < 1.0$) require sample sizes between 20 and 60. For highly skewed distributions ($|\gamma| > 1.0$), the requirement escalates sharply, often exceeding 100 observations. The increasing dispersion at higher skewness levels suggests that convergence is influenced by higher-order moments, reinforcing the need for a conservative upper-bound model.

To create a generalized rule, we aggregated simulation data from the Beta family. We calculated the upper-bound of minimum required sample size observed for each level of skewness ($n_{cons} = \max(\min(n))$). Crucially, we filtered the training dataset to include only distributions with excess kurtosis $|\kappa| \leq 1.0$ to isolate the specific effect of skewness.

The simulations confirmed a non-linear positive relationship (Fig. 1). A quadratic regression model was fitted to predict $\log(n_{cons})$ using absolute skewness ($|\gamma|$).

The regression analysis yielded a robust predictive model (Adj. $R^2 \approx 0.63$, $p < 0.001$). The fitted equation provides a conservative estimate for the minimum sample size:

$$n_{cons} \approx \exp\left(1.5986 + 1.4080 \cdot |\gamma| + 0.6075 \cdot |\gamma|^2\right)$$

3.4. Validation and Boundaries

To assess generalizability, the formula was tested against hold-out distributions (Table 3).

Table 3
Validation of n_{cons} against Hold-Out Distributions

Distribution	Skewness	Kurtosis	Observed n	Predicted n_{cons}
Normal(0,1)	0.00	0.0	5–10	5
Chi-squared(4)	1.41	3.0	100–200	122
Beta(1,9)	1.47	2.5	≈ 200	148
<i>Failures</i>				
Poisson(5)	0.45	0.2	30–50	11
Exponential(1)	2.00	6.0	200–500	939

The model performs well for continuous distributions with moderate kurtosis. For Chi-squared(4) and Beta(1,9), the predicted values ($n = 122, 148$) align closely with the simulation thresholds. However, two distinct failure modes emerge: **Discreteness (Poisson)** and **Severe Kurtosis (Exponential)**.

4. Discussion and Conclusion

This study provides a quantitative alternative to the $n \geq 30$ heuristic. For continuous distributions with normal-like tails, Eq. (1) offers a precise, conservative benchmark. The failure of the $n = 30$ rule for the Exponential distribution ($p < 0.0001$) underscores the risk of relying on static heuristics for skewed data. Researchers working with financial returns, waiting times, or other asymmetric variables risk invalid inferential results if they do not account for skewness. Furthermore, the validation failures reveal that normality is a multi-dimensional convergence problem: skewness is the primary driver, but discreteness and heavy tails act as independent confounding factors. Future research should integrate kurtosis directly into the regression model to broaden its applicability to heavy-tailed distributions.

References

1. Casella G, Berger RL (2002) *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition
2. William Feller (1971) *An Introduction to Probability Theory and Its Applications*, volume 2. John Wiley & Sons, New York, 2nd edition
3. Barbara Illowsky and Susan Dean (2013) *Introductory Statistics*. OpenStax, Houston, TX
4. Thomas Lumley P, Diehr S, Emerson, Chen L (2002) The importance of the normality assumption in large public health data sets. *Annu Rev Public Health* 23:151–169

5. Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples).
Biometrika 52(3/4):591–611

Figures

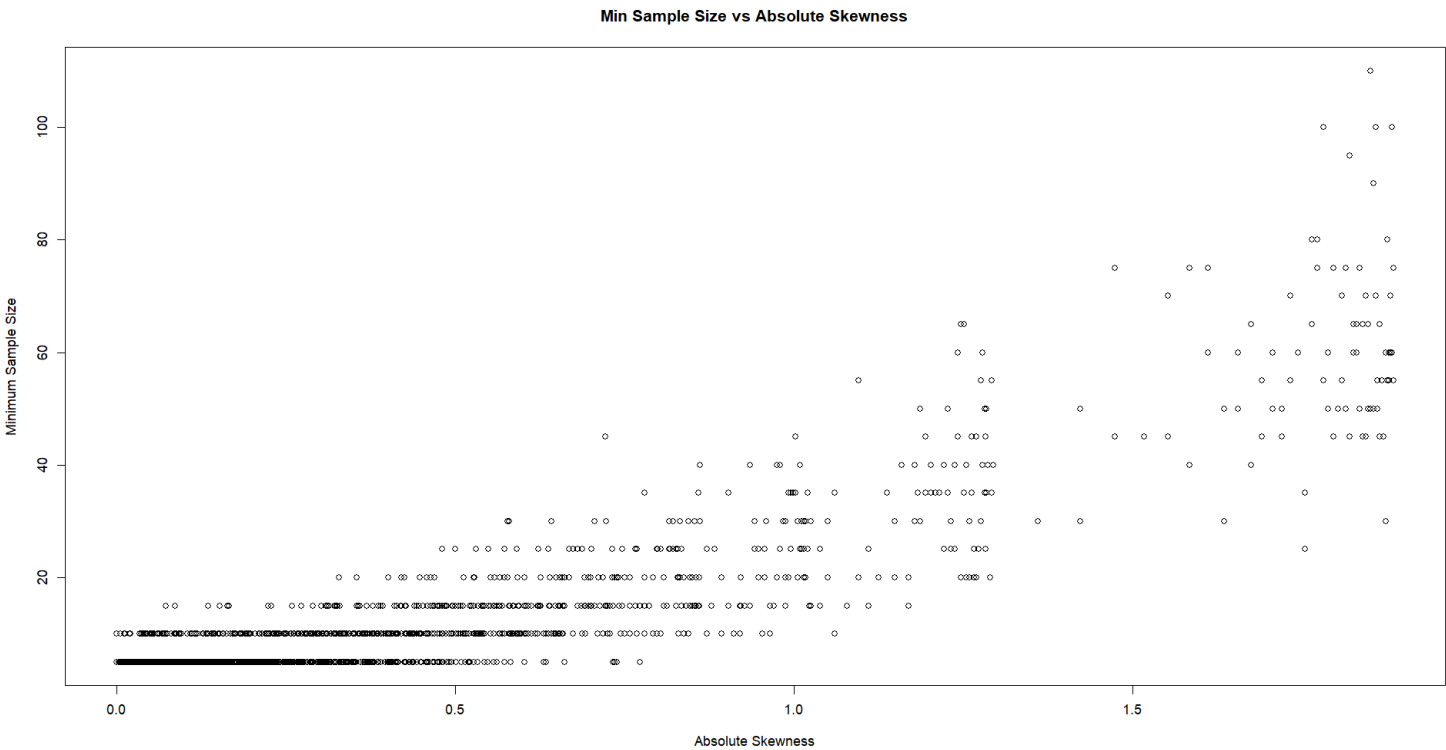


Figure 1

Skewness Vs Minimum Sample Size for Convergence to Normality