COMPARATIVE ANALYSIS OF SINGLE SAMPLE HYPOTHESIS TESTING: CRITICAL EVALUATION OF FREQUENTIST APPROACHES AND BAYESIAN INFERENCE ON SIMULATED DATA

Pardomuan Robinson Sihombing[1]*
[1]Badan Pusat Statistik, Indonesia
‡Korespondensi Penulis: robinson@bps.go.id

ARTIKEL INFO

Article history:
Received 24 Nov, 2025
Revised 09 May, 2026
Accepted 20 June, 2026
Published 30 June, 2026

Abstract

Introduction/Main Objectives: The validity of statistical inference is a key pillar of data-driven decision-making. **Background Problems:** Despite the importance, the inappropriate selection of methods for non-ideal data often threatens the validity of statistical inference. **Novelty:** This study evaluates the performance of single-sample hypothesis testing methods by comparing the frequentist paradigm (Student's t-test, Wilcoxon signed-rank test, sign test) and the Bayesian paradigm (Bayes factor). **Research Methods:** Using Monte Carlo simulation data generated in R Studio (10,000 iterations), this study investigates statistical power, Type I error rate, and the accuracy of effect size estimates (Cohen's d, Rank-Biserial Correlation, Cohen's g) under Normal, Heavy-tailed (t-Student), and Skewed (Log-normal) distribution conditions with sample variations n=30, 50, 100. **Finding/Results:** The results show that under the t-Student distribution (df=3), the Wilcoxon test consistently outperforms the T-test in terms of Power (0.514 versus 0.416 at n=30). Another crucial finding is the bias in Cohen's d estimation for Log-normal data, which tends to underestimate the true impact of location relative to Rank-Biserial Correlation. The Bayesian approach proved to be more conservative but provided better inference stability in large samples.

Keywords:

Bayes Factor; Bayesian Inference; Cohen's d; Monte Carlo Simulation; One-Sample Test; Statistical Power

1. Introduction

Modern inferential statistics often faces an epistemological dichotomy between the frequentist and Bayesian paradigms. In the context of single-sample testing, researchers often get stuck in routine procedures without considering data distribution in depth. The dominance of *Null Hypothesis Significance Testing* (NHST), with its reliance on p-values, has long been criticized for being frequently misinterpreted as the probability that the hypothesis is true (Wasserstein & Lazar, 2016). This misinterpretation is highly problematic because it often leads to false-positive conclusions, driving the replication crisis in science and potentially resulting in misguided public policies when applied to empirical socioeconomic research. The core of this issue lies in the fundamental mathematical definition of the p-value: it strictly represents the probability of observing the data given that the null hypothesis is true, $P\left(D\middle|H_0\right)$, not the probability that the

null hypothesis is true given the data, $P\left(H_0 \middle| D\right)$. This subtle yet profound logical inversion, often termed the conditional probability fallacy, leads researchers to overstate the significance of marginal findings heavily. This data drives the need for alternative methods that can quantify evidence more directly, such as the Bayes Factor in a Bayesian framework (Morey & Rouder, 2011; Rouder et al., 2009). Unlike the binary threshold of p-values, the Bayes Factor acts as a continuous metric of evidence. It quantifies exactly how much the empirical data shifts our prior beliefs to posterior odds, thereby providing a more nuanced, logically sound foundation for scientific claims.

On the other hand, in the frequentist framework, the choice between parametric (T-test) and non-parametric (Wilcoxon, Sign Test) tests is often based on outdated rules of thumb. Although the Student's t-test is known as *the Uniformly Most Powerful* (UMP) test for normal data, its validity is questioned for data with *outliers* or extreme skewness (Blair & Higgins, 1980). Previous studies, such as Yap and Sim (2012), highlight the importance of selecting the appropriate normality test, with Shapiro-Wilk often superior to Kolmogorov-Smirnov. However, the impact of failing to detect these non-normality on *the Power* of location tests and *effect size* bias has not been extensively explored simultaneously in a comprehensive study.

While seminal works have extensively compared the T-test and the Wilcoxon test (Bridge & Sawilowsky, 1999) or evaluated normality tests (Yap & Sim, 2012), they predominantly focus on either frequentist power or Type I error rates in isolation. This methodological gap becomes particularly critical when analyzing non-ideal regional data, such as economic growth, income disparity, or poverty indicators in East Nusa Tenggara (NTT). In macroeconomic monitoring and public policy evaluation, such as assessing regional Gross Domestic Product (GDP) fluctuations, the Human Development Index (HDI), or Gini ratios across provinces, data structures naturally suffer from extreme positive skewness and severe outliers. Relying on fragile parametric assumptions when analyzing these specific macro-indicators does not merely result in academic inaccuracy; it translates directly into flawed resource allocation. Such real-world socioeconomic data are notoriously skewed and heavy-tailed. Applying robust statistical inference is paramount to accurately evaluate regional intervention programs and ensure better policy-making for a developing NTT. This study aims to fill this gap by conducting a comparative analysis of the performance of the t-test, the Wilcoxon test, the Sign test, and the Bayesian t-test using simulated data that reflects real conditions (Normal, Heavy-tailed, and Skewed). The primary focus is to evaluate *the Power* of the test and the accuracy of the effect size estimator, specifically comparing *Cohen's d* with *Rank-Biserial Correlation*, which is often overlooked (Kerby, 2014). Specifically, this paper addresses the following research questions:

1. How do heavy-tailed and skewed distributions affect the statistical Power of frequentist versus Bayesian one-sample tests?

2. To what extent do parametric effect sizes (Cohen's d) distort the magnitude of effects in non-normal data compared to non-parametric alternatives?

2. Methodology

This study employs a Monte Carlo simulation design. The systematic steps of this research include: (1) defining the theoretical population distributions, (2) generating simulated datasets, (3) applying both frequentist and Bayesian hypothesis tests across iterations, and (4) extracting power and effect size metrics for comparative evaluation.

The data utilized in this research are entirely synthetic, generated through computer-based simulations rather than collected from real-world empirical observations. The rationale for utilizing simulated data is to establish an absolute 'ground truth' regarding population parameters (such as the exact mean, effect size, skewness, and kurtosis), which is practically impossible to achieve with observational field data. By strictly controlling these underlying data structures via pseudo-random number generation, this study can objectively isolate and evaluate the mathematical behaviour, Type I error rates, and statistical Power of each hypothesis test without the interference of unobserved confounding variables.

The simulations were conducted in R Studio with K=10,000 iterations to ensure a Monte Carlo Standard Error (MCSE) of less than 0.005, a threshold recommended for highly reliable and valid power estimates (Morris et al., 2019).

. The population distributions used include:

- *Normal:*$\mathcal{N}\left(\mu,\sigma^2\right)=\mathcal{N}(0,1)$.
- *Heavy-Tailed (t-Student):*$t(df=3)$, representing symmetric data with extreme *outliers* (high kurtosis).
- *Skewed (Log-normal):* $LN(0,1)$, representing right-skewed data (such as financial/income data).

Effects ($Effect\ Size$): $\delta = 0.5$ (Moderate Effects) are added to the data to test *Power*.

The methods used include:

- Parametric test: Student's t-test (Walpole, 2012) :

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \quad (1)$$

Where t is the test statistic, $\bar{x}$ is the sample mean, μ_0 is the hypothesized population mean, s is the sample standard deviation, and n is the sample size.

Effect Size (Cohen's d): The bias-corrected estimator (Hedges' g) is often recommended, but standard Cohen's d is used for baseline comparisons:

$$d = \frac{\left(\bar{x} - \mu_0\right)}{s} \text{(Cohen, 1988)} \quad (2)$$

Where d represents the standardized mean difference.

- Non-parametric test: The Wilcoxon Signed-Rank test is used to test the population median (θ) with signed rank statistics$W$ (Siegel, 1997)

$$z = \frac{W - E(W)}{\sigma_W}, \quad \text{where } E(W) = \frac{n(n+1)}{4} \text{ and } \sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}} \quad (3)$$

Where z is the standard normal deviate, W is the sum of the positive signed ranks, $E(W)$ is the expected value of W, and $\$\sigma_W$ is the standard deviation of W.

Effect Size (Rank-Biserial Correlation -r_{rb}), calculated using the difference in proportions formula (Kerby, 2014) :

$$r_{rb} = \frac{4\left|W - (n(n+1)/4)\right|}{n(n+1)} \quad (4)$$

Where r_{rb} indicates the proportion of data pairs that support the hypothesis, W is the sum of the positive signed ranks, and n is the number of samples.

- Non-parametric test: Sign Test

Tests the probability of success p in a Binomial distribution. Effect size using Cohen's g:

$$g = P_{pos} - 0.5 \text{(Cohen, 1988)} \quad (5)$$

Where g is the distance between the observed proportion of positive signs (P_{pos}) and the expected null proportion (0.5).

- Bayesian Inference

Bayesian Inference is computed using the BayesFactor package in R (Morey & Rouder). The Bayes Factor (BF_{10}) was calculated via numerical integration (Gaussian quadrature) using a default JZS Cauchy prior on the standardized effect size ($r = 0.707$). Interpretation: $BF_{10} > 3$ is considered moderate evidence supporting H_1 (Jeffreys, 1961).

3. Result and Discussion

Data distribution validation (Table 1) shows that the generated data aligns with theoretical characteristics. The t-Student data exhibits extreme kurtosis (13.78), indicating very thick tails compared to Normal (Kurtosis≈ 3). The Log-normal data show a very high positive skewness (Skewness 6.91).

Table1 . Data Characteristics (Simulation Results $N_{pop} = 1.000$)

Distribution	Mean	Median	SD	Skewness	Kurtosis
Normal	-0.001	0.000	1.000	-0.027	2.993

Student-t ($df = 3$)	-0.015	-0.008	1.614	-0.285	13.784
Log-normal	0.685	0.032	2.229	6,918	132.35

The results in Table 2 confirm the findings of Razali & Wah (2011) that the Shapiro-Wilk test has far superior Power compared to the Kolmogorov-Smirnov test. This result occurs because the Shapiro-Wilk test is a regression-based test that is highly sensitive to the tails of a distribution. In contrast, the Kolmogorov-Smirnov test relies on the empirical cumulative distribution function, making it less sensitive to extreme outliers. The low Power of the KS test risks leading researchers to mistakenly assume the data is normal.

Table 2. Power of Normality Tests (Rejection Rate of H_0)

Distribution	Test	N=30	N=50	N=100	Conclusion
Student-t	Shapiro-Wilk	0.444	0.672	0.876	SW is more sensitive in detecting *heavy tails.*
	Kolmogorov-Smirnov	0.325	0.500	0.730	KS often fails to detect abnormalities.
Log-normal	Kolmogorov-Smirnov	*0.942*	*0.998*	*1.000*	Highly accurate for skewed data
	Shapiro-Wilk	0.994	1.000	1.000	Highly accurate for skewed data.
Normal	Shapiro-Wilk	*0.047*	*0.047*	*0.052*	*Normal*
	Kolmogorov-Smirnov	*0.052*	*0.048*	*0.059*	Very accurate on skewed data.

Table 3 presents the main findings regarding the Power of the test in detecting the effect of$(\delta = 0.5)$. Superiority of Wilcoxon on Thick-Tailed Data: In the Student-t distribution, the T-test experiences a drastic decrease in *Power* (only 0.416 at$N = 30$) compared to the Wilcoxon test (0.514). This result is due to sample variance inflation (s^2) caused by *outliers* in the t-distribution, which reduces the value of the$t$ statistic. The Wilcoxon test, a rank-based test, mitigates the impact of these *outliers.* These findings are consistent with the study by Bridge & Sawilowsky (1999).

To provide a deeper mathematical exposition on the power behaviors observed in Table 3, the drastic decline of the T-test's Power under the Student-t distribution (0.416 at $N = 30$) can be formally attributed to the behaviour of the sample variance (s^2) in the presence of heavy tails. In a Student-t distribution with low degrees of freedom (df=3), the theoretical variance is highly unstable, and empirical realizations frequently produce extreme outliers. Because the parametric t-statistic computes the standard error using the sample standard deviation ($s/\sqrt{n}$), a single extreme outlier in the numerator is mathematically squared in the denominator. This variance inflation artificially inflates the standard error, thereby compressing the calculated t-value toward zero and severely inflating the Type II error rate (the probability of failing to reject a false null hypothesis).

The Bayesian method with a default Cauchy prior ($r = 0.707$) yields lower statistical Power in small samples compared to frequentist methods. However, rather than being inherently 'conservative', this behaviour reflects how the Bayes Factor operates: it heavily penalizes the alternative hypothesis for being too vague. The default Cauchy prior spreads the probability mass widely, thereby demanding substantially stronger empirical evidence to shift beliefs away from the null hypothesis. Furthermore, to bridge the interpretative gap between these paradigms, it is crucial to recognize the equivalent bounds of evidence. Based on the calibration by Sellke et al (Sellke et al., 2001), a frequentist p-value of 0.05 corresponds to a Bayes Factor (BF_{10}) of at most 2.44, which constitutes merely anecdotal evidence. This theoretical calibration explains why the Bayes Factor threshold of $BF_{10} > 3$ appears stricter, highlighting the epistemological danger of over-interpreting borderline p-values in standard NHST.

Table3 . Comparison of Statistical Test Power (alpha=0.05, Bayes BF>3)

Distribution	*N*	T-test	Wilcoxon	Sign Test	Bayesian
Normal	*30*	*0.776*	*0.752*	*0.572*	*0.610*
Student-t	*30*	*0.416*	*0.514*	*0.448*	*0.250*
Log-normal	*30*	*0.974*	*0.992*	*0.798*	*0.884*
Normal	*50*	*0.928*	*0.918*	*0.756*	*0.816*
Student-t	*50*	*0.584*	*0.728*	*0.638*	*0.360*
Log-normal	*50*	*0.998*	*1.000*	*0.960*	*0.986*
Normal	*100*	*1,000*	*1,000*	*0.962*	*0.988*
Student-t	*100*	*0.848*	*0.984*	*0.936*	*0.662*
Log-normal	*100*	*1.000*	*1.000*	*0.998*	*1,000*

Table 4 highlights the distortion that occurs when using parametric metrics on non-normal data. In t-Student and Log-normal data, Cohen's d provides an underestimate because of the enlarged standard deviation (s). In the Student-t case, Rank-Biserial (0.425) provides a stronger and more stable indication of effect than Cohen's d (0.313). Researchers who report Cohen's d only on non-normal data risk reporting a more negligible intervention effect than the true one (Tomczak & Tomczak, 2014).

Table 4. Accuracy of Effect Size Estimates (Average Estimates)

Distribution	Sample_N	Est_Cohen_d	Est_RankBis	Est_Cohen_g
Normal	30	0.525	0.522	0.199
Student-t	30	0.327	0.409	0.169
Log-normal	30	0.629	0.748	0.250
Normal	50	0.507	0.516	0.193
Student-t	50	0.317	0.417	0.175
Log-normal	50	0.617	0.764	0.256
Normal	100	0.508	0.522	0.194
Student-t	100	0.313	0.425	0.173
Log-normal	100	0.594	0.769	0.256

Table 5 presents statistical test results for a set of random samples from a normal population with the same effect, but with different sample sizes (N=30, 50, and 100). These results empirically highlight the fundamental weakness of relying solely on p-values and the strength of the Bayesian method in providing a richer interpretation.

Sensitivity of p-values to sample size in small samples (n=30), the T-test yields p=0.363 and the Wilcoxon test yields p=0.410, both of which fail to reject H_0: μ =0 (for α=0.05, rejected). However, in a large sample (n=100) with similar distribution characteristics, the p-value drops dramatically to p<0.001 (highly significant). This phenomenon confirms the criticism raised by Fordellone et al. (2025), who argue that in large samples, even trivial clinical effects can yield misleading statistical significance (i.e., statistically significant but not practically significant). This result underscores that the binary "reject/accept" decision in NHST (Null Hypothesis Significance Testing) is highly susceptible to sample size bias.

Table 5. Statistical Test Results

N	Method	Statistics	P_Value	Effect Size Estimate
30	T-test	t = 0.92	0.3635	d = 0.17
	Wilcoxon	V = 273	0.4107	r_{rb} = 0.17
	Sign Test	S = 17	0.5847	g = -0.16
	Bayesian	BF10 = 0.29	-	delta = 0.16
50	T-test	t = 1.73	0.0897	d = 0.17
	Wilcoxon	V = 801	0.1156	r_{rb} = 0.17
	Sign Test	S = 32	0.0649	g = -0.16
	Bayesian	BF10 = 0.61	-	delta = 0.16
10	T-test	t = 4.16	0.0001	d = 0.17
	Wilcoxon	V = 3616	0.0002	r_{rb} = 0.17
	Sign Test	S = 69	0.0002	g = -0.16
	Bayesian	BF10 = 267.69	-	delta = 0.16

The most striking difference is seen in the Bayesian column. At n=30, BF_{10}= 0.29 is obtained. On the Jeffreys scale (Jeffreys, 1961), this value (BF < 1/3) provides substantial evidence in favour of the Null Hypothesis (H_0). This result means that the data are 3.4 times more likely to occur if there is no effect. Frequentist methods (p-values) cannot express this; they can only state "failed to reject," which is often misinterpreted as "no effect." At n=100, $BF_{(10)}$= 267.69 is obtained, which is extreme evidence supporting $H_{(1)}$. The transition from support for H(0) to H(1) as the evidence from the data increases demonstrates the self-correcting nature of consistent Bayesian inference, as explained by Wagenmakers et al.

Unlike fluctuating p-values, effect size estimates are more stable. Cohen's d: Ranges from 0.17 (n=30) to 0.42 (n=100). Rank-Biserial ranges from 0.17 to 0.43. The consistency between Cohen's d and Rank-Biserial in this normal data supports Kerby's (2014) findings that Rank-Biserial is a robust estimator and equivalent to Cohen's d when the normality assumption holds. However, it has the advantage of a more

intuitive interpretation (percentage of data pair dominance). Therefore, reporting effect sizes is mandatory to provide context for significant results (Setiawan & Sukoco, 2021).

4. Conclusion and Suggestion

This comprehensive simulation study critically evaluated the comparative efficacy of frequentist and Bayesian single-sample hypothesis testing methods under both ideal and non-ideal distributional assumptions. This study concludes that heavy-tailed and skewed distributions significantly degrade the statistical Power of the parametric T-test, whereas the Wilcoxon Signed-Rank Test maintains high robustness and superiority. Specifically, when confronted with extreme kurtosis (t-Student distribution), the parametric T-test suffers a substantial penalty in Power due to sample variance inflation. In contrast, the rank-based Wilcoxon test successfully mitigates the impact of these outliers, confirming its status as a highly reliable tool for heavy-tailed datasets. Furthermore, the evaluation of normality tests reiterates that researchers should strongly favour the Shapiro-Wilk test over the Kolmogorov-Smirnov test, as the latter often fails to detect severe distributional abnormalities, potentially leading to flawed downstream methodological choices.

Furthermore, relying on parametric effect sizes, such as Cohen's d, for non-normal data severely underestimates the true magnitude of an effect, making Rank-Biserial Correlation a much more accurate alternative. The distortion of Cohen's d is particularly alarming in skewed log-normal distributions, where the inflated standard deviation artificially depresses the standardized mean difference. The Rank-Biserial Correlation, conversely, demonstrates remarkable stability across our simulations and provides a highly intuitive metric of effect magnitude that remains resistant to extreme values. While the Bayesian approach exhibits lower Power in small samples, it provides a vital safeguard against false positives in large-sample inferences. By requiring substantially stronger empirical evidence to shift the probability mass away from the null hypothesis, the Bayes Factor effectively counters the fundamental weakness of standard p-values, which often flag trivial effects as highly significant when N is large. Therefore, the Bayesian framework acts not merely as an alternative testing mechanism but as a crucial epistemological check against the replication crisis fueled by over-reliance on Null Hypothesis Significance Testing (NHST).

Based on these comprehensive empirical findings, the author proposes several targeted recommendations to improve the rigour of statistical reporting and analytical practices:

- Methodological Selection: For researchers and policymakers, particularly those analyzing highly skewed regional socioeconomic data (such as poverty gaps, regional GDP, or public welfare metrics in developing regions like NTT), the author strongly advises to abandon routine T-tests. Utilizing the Wilcoxon test alongside

Rank-Biserial correlation will provide a much more accurate evaluation of regional development interventions.

- Mandatory Effect Size Reporting: Academic journals and peer reviewers should require the reporting of non-parametric effect sizes (e.g., the Rank-Biserial correlation or Cohen's g) whenever non-parametric tests are used. Relying solely on p-values or reporting parametric effect sizes for skewed data should be actively discouraged in empirical literature.
- Integration of Bayesian Metrics: Furthermore, integrating Bayes Factors is recommended to ensure that public policies are driven by substantial empirical evidence rather than mere statistical artefact. The author strongly recommends that statistical software training in higher education to begin incorporating Bayesian analysis as a standard curriculum component to familiarize future researchers with evidence quantification beyond the binary 'reject/accept' paradigm.
- Future Research Directions: Future studies should expand upon this simulation framework by investigating the comparative performance of these tests under multivariate scenarios and missing-data conditions, and by evaluating how these non-normal distributions affect effect-size estimates in more complex structures, such as longitudinal or panel-data models.

Ethics Approval

This study utilizes computer-generated simulation data; therefore, formal ethical approval from an institutional review board was not required.

Competing Interests

The author declares that there are no conflicts of interest regarding the publication of this paper.

Bibliography

[1] Blair, R. . C., & Higgins, J. J. (1980). A Comparison of the Power of Wilcoxon ' s Rank-Sum Statistic to That of Student ' s t Statistic under Various Nonnormal Distributions. *Journal of Educational Statistics*, *5*(4), 309–335.

[2] Bridge, P. D., & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon Rank-Sum test in small samples applied research. *Journal of Clinical Epidemiology*, *52*(3), 229–235.

[3] Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. In *Educacao e Sociedade* (2nd ed.). Lawrence Erlbaum Associates. http://www.biblioteca.pucminas.br/teses/Educacao_PereiraAS_1.pdf%0Ahttp://www.anpocs.org.br/portal/publicacoes/rbcs_00_11/rbcs11_01.htm%0Ahttp://re

positorio.ipea.gov.br/bitstream/11058/7845/1/td_2306.pdf%0Ahttps://direitoufma2010.files.wordpress.com/2010/

[4] Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

[5] Kerby, D. S. (2014). The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation. *Comprehensive Psychology*, *3*(1), 1–9. https://doi.org/10.2466/11.it.3.1

[6] Morey, R. D., & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, *16*(4), 406–419. https://doi.org/10.1037/a0024377

[7] Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

[8] Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

[9] Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71.

[10] Setiawan, E. P., & Sukoco, H. (2021). Exploring First Year University Students' Statistical Literacy: A Case on Describing and Visualizing Data. *Journal on Mathematics Education*, *12*(3), 427–448.

[11] Siegel, S. (1997). *Statistik Nonparametrik untuk Ilmu-ilmu Sosial*. PT Gramedia Pustaka.

[12] Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, *1*(21), 19–25.

[13] Walpole, R. E. (2012). *Probability & Statistics for Engineers & Scientists*. Pearson.

[14] Wasserstein, R. L., & Lazar, N. A. (2016). The ASA ' s statement on p-values : context , process , and purpose. *The American Statistician*, *70*(2), 70(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108