ISSN 2808-2605 EISSN 2808-4497 *Forum Analisis Statistik*
Desember 2025, 5 (2): 60-70

Methodological Comparison of Frequentist and Bayesian Two-Sample Tests on Regional Socioeconomic Disparities in Indonesia

Pardomuan Robinson Sihombing[1*]

[1]BPS-Statistics Indonesia
Jl. Dr. Sutomo 6-8, Jakarta
Correspondence Writer. e-mail: robinson@bps.go.id

ABSTRACT

The socioeconomic gap between Western Indonesia and Eastern Indonesia is a persistent challenge in the national development agenda. This study has two main objectives: (1) to empirically compare poverty levels and income inequality (Gini ratio) between the western region (Sumatra, Java, Bali) and the eastern/other regions of Indonesia using hypothetical data from 38 provinces in 2025; and (2) to conduct a systematic comparative analysis of five independent two-sample statistical tests (frequentist and Bayesian tests) to evaluate their consistency and applicability. The results confirm statistically and substantively significant disparities, particularly in rural poverty and inequality, with eastern regions exhibiting much higher levels. The comparative analysis shows a high degree of convergence among the existing statistical tests; most methods produce the same substantive conclusions, reinforcing the validity of the findings. However, methods robust to assumption violations, such as the Brunner-Munzel Test, proved to provide more reliable results theoretically. Through Bayes Factor calculations, the Bayesian approach offers a more nuanced measure of evidence strength than p-value-based binary decisions, allowing for the quantification of evidence for both alternative and null hypotheses.

Keywords: bayesian statistics, gini ratio, nonparametric test, poverty, regional disparities, t-test.

INTRODUCTION

Since independence, Indonesia's economic development has faced a fundamental structural challenge: regional development disparities. Historically, development has tended to be concentrated in the western part of the country, particularly on the islands of Sumatra, Java, and Bali, while the eastern regions, which include Nusa Tenggara, Kalimantan, Sulawesi, Maluku, and Papua, have shown slower rates of development (Firdaus, 2013) ; (Ningsih et al., 2024). This gap is not merely an economic issue, but a multidimensional phenomenon that includes disparities in access to education, health services, economic opportunities, and digital technology (Ningsih et al., 2024). Various empirical studies consistently show that regional inequality in Indonesia has emerged significantly since the mid-1990s and continues to day (Firdaus, 2013). This phenomenon is rooted in differences in natural resource content and diverse geographical conditions, which affect regions' ability to drive the development process (Sjafrizal, 2008). As a result, a polarization has formed between developed and underdeveloped regions, perpetuating a cycle of mutually reinforcing problems (Sjafrizal, 2008).

In measuring and analyzing regional disparities, the two most fundamental socioeconomic indicators are poverty rates and income inequality, which are generally measured using the Gini ratio. Poverty is a significant issue that serves as a benchmark for a nation's dignity and is explicitly mandated by the constitution to be addressed by the state (Moniyana & Pratama, 2021) (Mu'minah & Tjenreng,

2025). Various studies in Indonesia have confirmed the close relationship between regional inequality, economic growth, and poverty, where economic growth alone is not sufficient to effectively reduce poverty if it is not accompanied by fair and inclusive distribution policies (Agussalim et al., 2024); (Akhmad et al., 2018) (Safrita et al., 2021). The Gini ratio serves as the primary proxy for measuring income distribution equality, and high levels of inequality are significantly positively correlated with poverty levels (Akhmad et al., 2018); (Safrita et al., 2021). Furthermore, gap analysis often focuses on the urban-rural dimension, where the poverty gap between the two is often more pronounced than the gap between provinces.

Although there have been many studies that attempt to quantify the development gap between Western and Eastern Indonesia, there is a methodological gap that is often overlooked. Many of these studies apply standard inferential statistical tests, such as the Student's t-test, without being preceded by an in-depth discussion of the fulfillment of the underlying statistical assumptions, such as the normality of data distribution and homogeneity of variance. Socioeconomic data at the aggregate level is particularly vulnerable to violations of these assumptions. Using statistical tests that are not appropriate for the characteristics of the data can lead to erroneous conclusions. Applying the Student's t-test to data with heterogeneous variance, for example, can increase the risk of Type I errors (Ruxton, 2006); (Welch, 1947), while the use of nonparametric tests such as the Mann-Whitney U test on data with different distribution shapes can also produce inaccurate results (Karch, 2023). Failure to select the appropriate method creates a "methodological blind spot" that has the potential to result in policy recommendations based on fragile statistical evidence. Recognizing this gap, this study was designed with two complementary main objectives: (1) to provide an up-to-date empirical assessment of socioeconomic disparities between western and eastern Indonesia, and (2) to conduct a systematic methodological comparison of the results of four frequentist tests (Student's t-test, Welch's t-test, Mann-Whitney U test, Brunner-Munzel test) and one Bayesian test (Bayesian t-test) to evaluate the consistency of the results and highlight the practical implications of methodological choices in socioeconomic analysis.

METHODOLOGY

This study uses hypothetical cross-sectional secondary data representing the conditions of 38 provinces in Indonesia in 2025. This data is designed to reflect Indonesia's existing economic and social structure. The entire data analysis process used Jamovi statistical software.

The independent variable in this study is region, a binary categorical variable that divides the 38 provinces into two independent sample groups: Western Region (n=17). This group includes all provinces located on the Sumatra, Java, and Bali islands. This region has historically been the center of economic activity and development in Indonesia. Eastern Region/Others (n=21): This group comprises provinces in the Nusa Tenggara Islands, Kalimantan Island, Sulawesi Island, the Maluku Islands, and Papua Island. There are six dependent variables analyzed. All of them are ratio-scale data representing poverty and inequality indicators for urban, rural, and total areas.

The analytical framework in this study was designed to compare five different statistical approaches to test the differences between two independent groups. These approaches were chosen to cover scenarios in which statistical assumptions are met or violated.

Frequentist Parametric Test

Parametric tests assume that the sample data comes from a population distributed according to a specific probability distribution, which is generally normal.

- The Student's t-test is a classic method for comparing the means of two independent groups (Student, 1908) (Walpole, 2012). This test operates under three main assumptions: (1)

independence of observations, (2) normality of data distribution in both groups, and (3) homogeneity of variance (homoscedasticity), which means that the population variance of both groups is assumed to be the same. The test statistic is calculated using the formula:

$$t_{stat} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{01} - \mu_{02})}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (1)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means,n_1 and n_2 are the sample sizes, and s_p is the pooled standard deviation, calculated as: $s_p^2 = \frac{(n_1-1)s_1^2+(n_1-2)s_2^2}{n_1+n_2-2}$, where s_1^2 and s_2^2 are the sample variances of each group. The degrees of freedom (df) for this test are $df = n_1 + n_2$

- Welch's t-test

The Welch's t-test is a modification of the Student's t-test designed for situations where the assumption of homogeneity of variance is violated (heteroscedasticity) (Karch, 2023) (Walpole, 2012). This test is considered more robust and recommended as the primary choice by many statisticians due to its good performance even when the variances are homogeneous (Ruxton, 2006). The Welch test statistic is calculated as:

$$t_{stat} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{01} - \mu_{02})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (2)$$

The main difference lies in the denominator, which does not use the combined standard deviation. The degrees of freedom for the Welch test are not always integers and are approximated using the Welch-Satterthwaite equation (Welch, 1947):

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} \quad (3)$$

To measure the magnitude of the difference between two groups in practical terms, Cohen's d effect size is used (Cohen, 1988). Cohen's d quantifies the difference in means in terms of standard deviations. For the Student's t-test, the formula for Cohen's d is (Cohen, 1988):

$$d = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p} \quad (4)$$

For Welch's t-test, the denominator is modified using the square root of the mean variance of the two groups. The conventional interpretation for Cohen's d is: d≈0.2 (small effect), d≈0.5 (moderate effect), and d≈0.8 (significant effect) (Cohen, 1988).

Frequentist Nonparametric Tests

Nonparametric tests do not assume a specific data distribution and are generally used when the normality assumption is violated.

- The Mann-Whitney U test (the Wilcoxon Rank-Sum test) is a nonparametric alternative to the Student's t-test (Mann & Whitney, 1947) (Walpole, 2012). This test does not compare means, but instead tests the null hypothesis that the two samples come from populations with identical distributions by comparing the medians of *the* data *ranks*. Although it does not require the assumption of normality, this test has another important assumption: the shape of the distributions

of the two groups must be similar, which implicitly means that their variances must be homogeneous. The U statistic is calculated based on the sum of ranks:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad (5)$$

where R_1 is the sum of ranks for the first group. The test statistic used is the smaller value between U_1 and U_2 (where $U_1 + U_2 = n_1 n_2$). For sufficiently large samples (usually n>20), the distribution of U can be approximated by a normal distribution using the z-statistic.

The appropriate effect size for the MWU Test is the rank-biserial correlation Kerby, 2014). This measure quantifies the difference between the proportion of pairs supporting the hypothesis and those not supporting it. The simple formula is (Kerby, 2014):

$$r = 1 - \frac{U_1}{n_1 n_2} \quad (6)$$

The r value ranges from -1 to +1, with an interpretation similar to the correlation coefficient.

- The Brunner-Munzel test is a more robust nonparametric test and is a direct alternative to the Welch's t-test (Edgar Brunner & Ullrich Munzel, 2000). The main advantage of this test is that it does not require the assumption of homogeneity of variance, making it suitable for non-normal data with heterogeneous variance (Edgar Brunner & Ullrich Munzel, 2000). This test examines the hypothesis of stochastic equality, namely whether the probability of a random observation from the first group being greater than a random observation from the second group is equal to 0.5 ($P(X > Y) = 0.5$). Due to its robustness, some researchers recommend the BM test as the default nonparametric test, replacing the MWU test (Karch, 2023). The effect size for the BM Test is the *relative effect* $\hat{p}$, defined as the probability of stochastic superiority (Karch, 2023):

$$\hat{p} = P(X < Y) + 0.5 \cdot P(X = Y) \quad (7)$$

The value of the$\hat{p}$ has an intuitive interpretation: if$\hat{p}$ = 0.5, there is no difference between the two groups. If $\hat{p}$ > 0.5, then observations from group Y tend to be larger than those from group X, and vice versa if $\hat{p}$ 0.5.

Bayesian Approach

The Bayesian approach offers a fundamental alternative to the *Null Hypothesis Significance Testing* (NHST) framework used in the frequentist approach (Rouder et al., 2009). Instead of generating p-values, Bayesian t-tests calculate a metric called *the Bayes Factor (*BF) (Kass & Raftery, 1995). *The Bayes Factor* quantifies the strength of evidence provided by the data for one hypothesis compared to another (Jeffreys, 1961). Specifically, the *Bayes Factor* (BF_{10} *)* is defined as the ratio of *the likelihood* of the data under the alternative hypothesis (H_1) to the likelihood of the data under the null hypothesis (H_0):

$$BF_{10} = \frac{P(data \mid H_1)}{P(data \mid H_0)} \quad (8)$$

The interpretation of theBF_{10} is straightforward: aBF_{10} =10 means that the observed data is 10 times more likely to occur if the alternative hypothesis is true than if the null hypothesis is true. Conversely, a value ofBF_{10} = 0.2 (orBF_{01} = 5) means that the data is 5 times more likely to occur under the null hypothesis (Rouder et al., 2009). A commonly used interpretation scale, adapted from Jeffreys (1961), is as follows:

$1 < BF_{10} < 3$: Anecdotal evidence for H_1

$3 < BF_{10} < 10$: Moderate evidence for H_1
$10 < BF_{10} < 30$: Strong evidence for H_1
$BF_{10} > 30$: Powerful/extreme evidence for H_1.

The main advantage of this approach is its ability to quantify the evidence supporting the null hypothesis, which cannot be done by p-values (Rouder et al., 2009).

Analysis Procedure

The data analysis process follows a systematic and structured workflow:

1. Descriptive Analysis: Descriptive statistics (N, mean, median, standard deviation, minimum, and maximum values) were calculated for the six dependent variables, stratified by region group variable.
   Assumption Testing:
2. Normality: The Shapiro-Wilk test was applied to each variable for each group. A p-value < 0.05 was interpreted as a violation of the normality assumption.
   Homogeneity of Variance: Levene's test was used to compare the variance between the two groups for each dependent variable. A p-value < 0.05 was interpreted as evidence of heteroscedasticity (non-homogeneous variance).
3. Comparative Testing: Each dependent variable is tested for differences between the Western Region and Eastern/Other Regions groups using the five statistical procedures described above.
   Significance Level: The significance level (alpha) was set at α=0.05 for all frequency tests, data collection process, and data analysis.

RESULT AND DISCUSSION

The first step in the analysis was to explore the characteristics of the data through descriptive statistics and test the assumptions underlying the inferential tests. Table 1 presents descriptive statistics for the six research variables, separated by region. From this table, a clear pattern emerges. For all poverty variables (urban, rural, and total), the average percentage in the Eastern/Other regions is consistently higher than in the Western region. The most dramatic gap is seen in rural poverty, where the average for the Eastern/Other Region (21.01%) is more than double the average for the Western Region (8.92%). The standard deviation for poverty variables is also much greater in the Eastern/Other Region, indicating higher variability between provinces in this group. For the Gini ratio, the pattern is less uniform. The average urban Gini is slightly higher in the Western Region, while rural Gini and total Gini show higher average values in the Eastern/Other Region.

Table 1. Descriptive statistics of research variables by region

Variable	Region	N	Mean	Median	SD	Min	Max
	Other	21	6.48	5.51	2.59	3.43	13.00
Urban Poverty	Sumatra, Java, Bali	17	6.94	7.00	2.61	3.27	12.34
	Total	38	6.68	6.19	2.57	3.27	13.00
Rural Poverty	Others	21	21.01	20.80	7.73	4.97	38.47
	Sumatra, Java, Bali	16	8.92	8.24	2.82	4.97	14.44
	Total	37	15.67	12.93	9.47	4.25	38.47
Total Poverty	Others	21	13.01	10.92	7.73	3.84	30.03
	Sumatra, Java, Bali	17	7.64	7.19	2.77	3.72	12.33
	Total	38	10.61	9.49	6.55	3.72	30.03
Urban Gini Ratio	Others	21	0.320	0.310	0.050	0.210	0.430
	Sumatra, Java, Bali	17	0.360	0.352	0.055	0.232	0.441
	Total	38	0.338	0.332	0.057	0.207	0.441
Rural Gini Ratio	Others	21	0.326	0.323	0.077	0.248	0.511
	Sumatra, Java, Bali	16	0.263	0.264	0.042	0.199	0.334
	Total	37	0.299	0.275	0.071	0.199	0.511
Total Gini Ratio	Others	21	0.335	0.333	0.043	0.261	0.412
	Sumatra, Java, Bali	17	0.336	0.330	0.059	0.222	0.441
	Total	38	0.336	0.332	0.050	0.222	0.441

Source: BPS-Statistics Indonesia (processed)

Furthermore, the results of the normality and variance homogeneity assumptions tests are presented in Table 2. These results are crucial because they form the basis for determining the most appropriate theoretical frequency test for each variable.

Table 2. Results of normality assumption test (Shapiro-Wilk) and variance homogeneity test (Levene)

Assumption	**Variable**	**Poverty**			**Gini Ratio**		
		Urban	**Rural**	**Total**	**Urban**	**Rural**	**Total**
(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Normality	statistics	0.934	0.959	0.934	0.981	0.920	0.978
Shapiro Wilk	p	0.026	0.183	0.027	0.744	0.011	0.648
Homogeneity	F	0.049	37.864	13.957	0.055	4,279	1,730
Levene	p	0.827	<.001	<.001	0.817	0.046	0.197
Conclusion	Assump-tion	Not Normal, Homoge-neous	Normal, Heteroge-neous	Non-Normal, Heteroge-neous	Normal, Homoge-neous	Non-Normal, Heteroge-neous	Normal, Homoge-neous
	Statistical Test	Whitney U	Welch	Brunner Munzel	Student's t	Brunner Munzel	Student T

Source: BPS-Statistics Indonesia (processed)

The results in Table 2 highlight the complexity of real-world socioeconomic data. No single statistical test is suitable for all variables. The data shows all four possible combinations of assumption test results:

- Normal and Homogeneous (Urban Gini, Total Gini): The ideal scenario where the Student's t-test is the most appropriate choice.
- Normal and Heterogeneous (Rural Poverty): Conditions where the Welch's t-test is specifically designed to be used.
- Non-Normal and Homogeneous (Urban Poverty): A classic situation for applying the Mann-Whitney U test.
- Non-Normal and Heterogeneous (Total Poverty, Rural Gini): The most challenging scenario in which the most robust test, the Brunner-Munzel test, is the most appropriate choice.

The diversity of the results of these assumption tests is an important finding. It validates the research premise that a "one size fits all" approach to statistical analysis is inadequate. A flexible methodological toolkit and a deep understanding of the conditions under which each test should be applied are required. This finding transforms the study from a mere theoretical exercise into a practical demonstration of the importance of methodological rigor.

Table 3 presents the results of the five statistical tests applied to the six dependent variables. This table allows for a direct comparison between the frequentist and Bayesian approaches and parametric and nonparametric tests under various assumption fulfillment conditions.

Table 3. Comparative Results of Frequentist and Bayesian Independent Two-Sample Tests

Variable		Test		Statistic	p	Effect Size	Conclusion
(1)		(2)		(3)	(4)	(5)	(6)
Urban Poverty	Frequentist	Parametric	Student T	-0.541	0.592	-0.176	Not Significantly Different
			Welch	-0.540	0.593	-0.176	
		Nonparametric	U Mann-Whitney	157,000	0.538	0.120	
			Brunner-Munzel Test	0.598	0.555	0.560	
	Bayesian	Bayes factor$_{10}$		0.355			
Rural Poverty	Frequentist	Parametric	Student T	2.9270[a]	0.006	0.971	Significantly Different
			Welch	3.308	0.003	1.031	
		Nonparametric	U Mann-Whitney	102,500	0.046	-0.390	
			Brunner-Munzel Test	-2.180	0.038	0.305	
	Bayesian	Bayes factor$_{10}$		7.436			
Total Poverty	Frequentist	Parametric	Student T	2.7224	0.01	0.888	Significantly Different
			Welch	2.960	0.006	0.926	
		Nonparametric	U Mann-Whitney	100,500	0.023	-0.437	
			Brunner-Munzel Test	-2.573	0.015	0.282	
	Bayesian	Bayes factor$_{10}$		5.018			
Urban Gini Ratio	Frequentist	Parametric	Student's t-test	-2.299	0.027	-0.750	Significantly Different
			Welch	-2.293	0.028	-0.749	
		Nonparametric	U Mann-Whitney	99,500	0.021	0.443	
			Brunner-Munzel Test	2.595	0.014	0.721	
	Bayesian	Bayes factor$_{10}$		2.346			

Variable		Test		Statistic	p	Effect Size	Conclusion
(1)		(2)		(3)	(4)	(5)	(6)
Rural Gini Ratio	Frequentist	Parametric	Student's t-test	2.9635	0.005	0.983	Significantly Different
			Welch	3.195	0.003	1.021	
		Nonparametric	U Mann-Whitney	80,000	0.007	-0.524	
			Brunner-Munzel Test	-3.306	0.002	0.238	
	Bayesian	Bayes factor$_{10}$		8.001	±	3.02E-08	
Total Gini Ratio	Frequentist	Parametric	Student T	-0.070	0.945	-0.023	Not Significantly Different
			Welch	-0.068	0.946	-0.023	
		Nonparametric	U Mann-Whitney	175,500	0.941	-0.017	
			Brunner-Munzel Test	-0.084	0.934	0.492	
	Bayesian	Bayes factor$_{10}$		0.317			

Source: BPS-Statistics Indonesia (processed)

Table 4. Summary of results based on selected tests

Variable	Statistical Test	Statistics	p	Effect Size	BF$_{10}$
Urban Poverty	U Mann-Whitney	157,000	0.538	0.120 (Biserial rank r)	0.355
Rural Poverty	Welch	3.308	0.003	1.031 (Cohen's d)	7,436
Poverty Total	Brunner-Munzel	-2.573	0.015	0.282 (Relative Effect)	5.018
Gini Ratio Urban	Student T	-2.299	0.028	-0.750 (Cohen's d)	2.346
Gini Ratio Rural	Brunner-Munzel	-3.306	0.002	0.238 (Relative Effect)	8.001
Gini Ratio Total	Student T	-0.070	0.945	-0.023 (Cohen's d)	0.317

Source: BPS-Statistics Indonesia (processed)

The following is a narrative of the test results for each variable:

- Urban Poverty: The most appropriate test is the Mann-Whitney U test, which yields p = 0.538. This result indicates no statistically significant difference in urban poverty levels between the two regions. The Student's t-test also yields the same conclusion (p = 0.592). The Bayesian approach reinforces these findings with BF$_{10}$ = 0.355 (or BF$_{01}$ = 1/0.355 ≈ 2.82), which provides anecdotal to moderate evidence *for the null hypothesis* (no difference).
- Rural Poverty: The most appropriate test is the Welch's t-test. The results show a statistically significant difference (t(29.27) = 3.308, p = 0.003). Cohen's d effect size = 1.031 indicates a huge difference. This conclusion is consistently supported by all other tests: Student's t-test (p < .001), *Brunner-Munzel* test (p = 0.010), and Bayesian t-test yielding BF$_{10}$ = 7.436, providing decisive evidence of a difference.
- Total Poverty: The most appropriate test is the Brunner-Munzel Test. The results show a statistically significant difference (p = 0.015). *The Relative Effect* size of 0.282 indicates that the probability of a province from the Eastern Region having a higher level of total poverty than a province from the Western Region is very large. The Student's t-test (p = 0.005) and Welch's test (p = 0.006) also support this conclusion. BF$_{10}$ = 5.018 provides moderate evidence for the existence of a difference.
- Urban Gini Ratio: The most appropriate test is the Student's t-test, which shows a significant difference (t(36) = -2.299, p = 0.028), with a higher level of inequality in the Western Region. The

Mann-Whitney U test (p = 0.021) provides a similar conclusion. BF_{10} = 2.346 provides anecdotal evidence for a difference.

- Rural Gini Ratio: The most appropriate test is the Brunner-Munzel test, which shows a highly significant difference (p = 0.002). *A Relative Effect* of 0.238 indicates a strong stochastic dominance of the level of inequality in the Eastern Region. The Student's t-test (p = 0.003) and Welch's test (p = 0.003) are highly consistent. BF_{10} = 8.001 provides moderate to strong evidence for a difference.
- Total Gini Ratio: The most appropriate test is the Student's t-test, which found no significant difference (p = 0.945). The Mann-Whitney U test (p = 0.941) and Bayesian t-test (BF_{10} = 0.317, or $BF_{01} \approx 3.15$) also support the conclusion that there are no differences, with the Bayesian approach providing moderate evidence for the null hypothesis.

Discussion

One of the most striking findings of this analysis is the high level of convergence among the five different statistical methods. For each variable, the substantive conclusion whether there is a significant difference remains consistent across the spectrum of testing, from classical parametric to robust nonparametric and Bayesian. This consistency has important implications. It suggests that the differences (or, in the case of the lack of differences) detected in the data are not statistical artifacts arising from a particular method, but rather a reflection of a robust pattern in the data itself. When the Student's t-test, Welch's t-test, Brunner-Munzel test, and Bayesian t-test all point in the same direction, as in the case of rural poverty, confidence in the validity of the findings increases substantially.

However, behind this consistency lies an important nuance. Although the conclusion may be the same, parameters generated by tests that do not match the data assumptions (e.g., Student's t-test on heteroscedastic data) are technically unreliable. For example, in rural poverty, the Student's t-test produces more extreme t-statistics and p-values than the Welch's t-test. If the significance level is marginal (e.g., p=0.04 for Student's and p=0.06 for Welch's), a researcher relying solely on the Student's t-test could mistakenly conclude that there is statistical significance. This result underscores how fragile methods can be misleading, especially when results are around the significance threshold of α=0.05.

The Bayesian approach further enriches this analysis by shifting the paradigm from a binary "significant or not" decision to a continuous evaluation of the strength of evidence. For urban poverty, where the frequentist test fails to reject the null hypothesis, the Bayesian approach can quantify *the* evidence *supporting* the null hypothesis ($BF_{01} \approx 2.82$). This result is far more helpful than simply stating "there is not enough evidence to reject H_0". Thus, robust frequentist tests and Bayesian analysis provide the most complete and methodologically defensible picture.

Substantively, the results of this analysis paint a stark picture of inequality in Indonesia. The strongest and most consistent finding is the extreme disparity in rural poverty and rural inequality. Cohen's d effect size of 1.031 for rural poverty indicates a significant difference. In practical terms, this means that the average percentage of rural poverty in Western Region provinces is at the 85th percentile of the distribution of rural poverty percentages in Eastern Region provinces. Similarly, *the Relative Effect* of 0.238 for rural inequality indicates a high probability (around 76.2%) that a random province from the Eastern Region will have a higher rural Gini ratio than a random province from the Western Region.

These findings align with literature highlighting that the roots of development lag in Eastern Indonesia lie in dependence on low-productivity primary sectors (agriculture, fisheries, mining), geographical isolation, and chronic infrastructure and human resources deficits. Poverty in this region is concentrated in rural areas where access to markets, education, and health services is minimal (Hill et al., 2008; Vidyattama, 2013).

Interestingly, the picture becomes more complex when it comes to poverty and inequality in urban areas. No significant differences were found in urban poverty, and in terms of urban Gini, the Western Region showed a slightly higher level of inequality. This result can be explained by the "uneven growth" phenomenon in major cities in Java and Sumatra. Although on average more prosperous, these cities are also home to pockets of urban poverty and sharp inequality between very high and very low income groups. As a result, when the data is aggregated at the total provincial level (gini ratio total), the difference between the two regions becomes insignificant, as the effects of high urban inequality in the West and high rural inequality in the East cancel each other out.

CONCLUSION AND RECOMMENDATION

This study successfully achieved its two main objectives: providing empirical evidence of regional disparities in Indonesia and conducting a methodological evaluation of various statistical tests. Empirically, this analysis confirms the existence of significant socioeconomic disparities between Indonesia's Western and Eastern regions. The most acute manifestation of this disparity is found in the levels of poverty and inequality in rural areas, where the Eastern Region shows much worse conditions. This disparity is statistically significant and practically substantial, as indicated by the strong effect size. From a methodological perspective, a comparative analysis of five different statistical tests showed a high degree of convergence in substantive conclusions, providing strong confidence in the robustness of the findings. However, this study also highlights the importance of selecting appropriate methods for the data's characteristics. Robust tests for assumption violations, such as the Welch's t-test and the Brunner-Munzel’s test, provided a more solid basis for inference. Furthermore, the Bayesian approach provides added value by providing a more informative and nuanced measure of evidence strength than p-values, allowing for the quantification of evidence for both the null and alternative hypotheses.

These findings have important implications for both research practice and policy formulation. For researchers in the social and economic sciences, this study underscores the urgency of not automatically using the Student's t-test as the default method. Standard practice should include careful testing of normality assumptions and variance homogeneity. Given that socioeconomic data often violate these assumptions, adopting more robust tests, such as Welch's t-test and Brunner-Munzel's test, is advisable as standard practice. In addition, researchers are encouraged to complement their analyses with Bayesian approaches to obtain richer interpretations. From a policy perspective, the results of this study reaffirm the need for more focused and targeted development interventions to address poverty and inequality in rural Eastern Indonesia. These interventions should not only focus on fiscal transfers but also include massive investments in connectivity infrastructure, improvements in the quality of education and health services, and economic empowerment programs specifically designed for the local context.

REFERENCES

Agussalim, A., Nursini, N., Suhab, S., Kurniawan, R., Samir, S., & Tawakkal, T. (2024). The Path to Poverty Reduction: How Do Economic Growth and Fiscal Policy Influence Poverty Through Inequality in Indonesia? Economies, 12(12), 1–17. https://doi.org/10.3390/economies12120316

Akhmad, Alyas, & Amir. (2018). Effects of economic growth and income inequality on poverty in Indonesia. IOSR Journal of Economics and Finance (IOSR-JEF), 9(4), 20–26. https://doi.org/10.22136/est20191312

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. In Educacao e Sociedade (2nd ed.). Lawrence Erlbaum Associates.

Edgar Brunner, & Ullrich Munzel. (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. Biometrical Journal, 42(1), 17–25.

Firdaus, M. (2013). Ketimpangan Pembangunan antar Wilayah Di Indonesia: Fakta dan Strategi Inisiatif [Development Disparities Between Regions in Indonesia: Facts and Initiative Strategies]. In Institut Pertanian Bogor.

Hill, H., Resosudarmo, B. P., & Vidyattama, Y. (2008). Indonesia's changing economic geography. Bulletin of Indonesian Economic Studies, 44(3), 407–435. https://doi.org/10.1080/00074910802395344

Jeffreys, H. (1961). Theory of probability. Oxford University Press.

Karch, J. D. (2023). bmtest: A Jamovi Module for Brunner–Munzel's Test—A Robust Alternative to Wilcoxon–Mann–Whitney's Test. Psych, 5(2), 386–395. https://doi.org/10.3390/psych5020026

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

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

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, 18(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Moniyana, R., & Pratama, A. D. (2021). Kemiskinan dan Ketimpangan Pembangunan kabupaten/Kota di Provinsi Lampung. Jurnal Ekonomi Pembangunan, 10(1), 31–45. https://doi.org/10.23960/jep.v10i1.216

Mu'minah, S., & Tjenreng, M. B. Z. (2025). Desentralisasi dan Ketimpangan Pembangunan Antar Daerah. Scientific Journal Of Reflection : Economic, Accounting, Management and Business, 8(1), 342–351. https://doi.org/10.37481/sjr.v8i1.1053

Ningsih, U., Alpendi, & Dewi, A. S. (2024). Kesenjangan Sosial Ekonomi di Indonesia: Penyebab, Dampak, dan Solusi Kebijakan. Jurnal Sosiologi Agama Indonesia (JSAI), 5(3), 427–225. https://doi.org/10.22373/jsai.v5i3.5577

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 and Review, 16(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behavioral Ecology, 17(4), 688–690. https://doi.org/10.1093/beheco/ark016

Safrita, S., Abbas, T., & Yurina, Y. (2021). the Effect of Economic Growth and Poverty on Income Inequality in Indonesia. Journal of Malikussaleh Public Economics, 4(1), 30–37. https://doi.org/10.29103/jmpe.v4i1.4792

Sjafrizal. (2008). Ekonomi Regional: Teori dan Aplikasi. Baduose Media.

Student. (1908). The probable error of a mean. Biometrika, 6(1), 1–25. https://doi.org/10.1070/SM1996v187n11ABEH000171

Vidyattama, Y. (2013). Regional convergence and the role of the neighbourhood effect in decentralised Indonesia. Bulletin of Indonesian Economic Studies, 49(2), 193–211.

Walpole, R. E. (2012). Probability & Statistics for Engineers & Scientists. Pearson.

Welch, B. . (1947). The Generalization of `Student's' Problem when Several Different Population Variances are Involved. Biometrika, 34(1/2), 28–35.