International Journal of Multidisciplinary Approach and Studies

ISSN NO:: 2348 – 537X

Sample Size Effects on Normality Test Performance: A Monte Carlo Analysis

Ralph Jay M. Magsalay, PhD

Lourdes College, Inc., Cagayan de Oro City, Philippines, 9000

ABSTRACT

The assumption of normality is key for many statistical analyses, and how well normality tests work is vital for good research. This study used Monte Carlo simulations in R software to check the Shapiro-Wilk (SW), Kolmogorov-Smirnov (KS), and Anderson-Darling (AD) tests across different group sizes (n = 10 to 1000) by looking at their error rates using both normal and non-normal data. The results showed that SW and AD were too sensitive with larger sample size (n ≥ 200), producing more false results. For large sample size (n > 60), KS showed low rates of both Type I and II errors. However, for smaller sample size, KS often failed to identify true non-normality, though it rarely gave Type I error. In contrast, SW and AD were better at detecting non-normality in these smaller sample size providing a good balance between false results and detection ability for samples sizes around 20 to 60. Below 20 samples, SW and AD became less reliable at detecting non-normality, suggesting to use visual inspection of Q-Q plots. Note that these findings assume a clean data, and the result might differ with outliers or other data issues.

KEYWORDS: *Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, Sample Size, Monte Carlo Simulation.*

INTRODUCTION

The normality of data is a very important aspect for many basic statistics tests. These tests include t-tests, ANOVA, and linear regression. The results from these tests would be trustworthy and easy to understand if the data has a normal distribution. Findings that are not accurate can result if the data does not have a normal distribution, and this can cause scientific ideas that are incorrect in fields like health, engineering, and social studies (Demir, 2022). Checking data for a normal spread is important for reliable research. Researchers often apply tests such as Shapiro-Wilk to see if numbers are normally distributed. The Shapiro-Wilk test is effective at identifying various deviations from a normal pattern (Shapiro & Wilk, 1965). Kolmogorov-Smirnov checks if data fits a normal spread well (Kolmogorov, 1933), while Anderson-Darling looks closely at the data ends. This helps Anderson-Darling see if data is too pointed or leaning (Anderson & Darling, 1954). These tests give a p-value that helps decide if there is enough proof to state that the data is normally distributed.

A major problem with applying these normality tests is that the sample size can change its accuracy. Recent research showed that even very tiny differences from a normal data can make the tests incorrectly identify as non-normal if you have large sample size, and this might be a real problem. On the other hand, the tests might not find real differences from normal if you have too few samples, which would lead to a false positive. One study found that the Kolmogorov-Smirnov test was not very dependable if you had less than ninety

samples. But the Anderson-Darling test was usually reliable if you had fifty or more (Biu et. al., 2020). This showed that the sample size can change how well these normality tests work. There was not much new research that compared how sensitive the SW, KS, and AD tests were when the data was not normal in different ways and when there were varied sample sizes even though older studies discussed about this problem. It was important for researchers to really understand how these tests act with different amounts of data so they could determine the appropriate test and understand what the results mean. Studies showed that normality tests might not work well if you only have few data. One study looked at how well normality tests worked with small sample size and found that the Shapiro-Wilk test shows good accuracy. But this test could give false results if there were only thirty data points, and this could cause problems for medical guidelines (de Souza et. al., 2023).

This study looked at how reliable the Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling tests were for different sample sizes. This study made different sets of data with different sizes that were normal and skewed using computer simulations. Then, the SW, KS, and AD tests were applied to this data. Following this, the study quantified how frequently the tests incorrectly indicated non-normality for data that was actually normal, and how effectively they detected non-normality for data that was skewed, for each sample size. Finally, it looked at how the test results changed when the amount of data changed. This study revealed how the Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling normality tests behaved with different size of datasets using Monte Carlo Simulation. The goal is to give simple advice to researchers on how to choose and understand normality tests so that their results are reliable.

METHOD

This study applied Monte Carlo Simulation using R software to check three normality tests. These tests were Shapiro-Wilk (SW), Kolmogorov-Smirnov (KS), and Anderson-Darling (AD). The study tested these on different sample sizes. Monte Carlo Simulation means repeating random sampling many times to see how well a statistical test works (Anastasiou, 2020). These three tests were chosen because they are often applied when checking if data is normal. The study had two parts. The first part checked how often the tests wrongly said normal data was not normal (Type I error). The second part checked how often the tests wrongly said non-normal data was normal (Type II error).

Phase 1: Evaluation with Normally Distributed Data (Type I Error)

In this part, the computer made 10,000 sets of normal data for each sample size. The sample sizes were from 10 to 1000. Each set was checked using the Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling tests. The computer counted how many times each test gave a p-value less than 0.05. This showed how often the test wrongly said the data was not normal. These numbers were changed into percentages. The percentages showed the chance of a Type I error.

Phase 2: Evaluation with Non-Normally Distributed Data (Type II Error)

In this phase, the same steps were used, but the data came from a skewed Weibull distribution. This type of data is common in simulations because it shows non-normal shapes well (Wijekularathna et. al., 2020). Since normality tests often do not work well on small

sample sizes, this phase tested sample sizes from 10 to 150. For each sample size, the computer created 10,000 sets of skewed data. Then, the same three tests were applied. The computer counted how many times each test gave a p-value of 0.05 or higher. This showed how often the test wrongly said the data was normal. These wrong results were counted for each sample size and test. Then, the number was changed to a percentage. This percentage showed the chance of a Type II error.

DATA ANALYSIS

The number of Type I and Type II errors for each normality test was put into tables for all the different sample sizes that were looked at. These numbers were then changed into percentages to show the estimated chances of making these errors. To help see how these chances changed with different group sizes, line graphs were made. These graphs showed how the chance of a Type I error (for all sample sizes) and a Type II error (for smaller sample sizes up to 150) changed as the sample size increases for each of the tests.

RESULTS AND DISCUSSION

Table 1 shows how often each test made mistakes with large sample size. A mistake means the test said the normal data was not normal. The tests are Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling. For sample size 300, Shapiro-Wilk made 5191 mistakes (51.91%). Kolmogorov-Smirnov made 87 mistakes (0.87%). Anderson-Darling made 4031 mistakes (40.31%). For sample size 700, Shapiro-Wilk made 8514 mistakes (85.14%). Kolmogorov-Smirnov made 464 mistakes (4.64%). Anderson-Darling made 7579 mistakes (75.79%). For sample size 1000, Shapiro-Wilk made 9420 mistakes (94.20%). Kolmogorov-Smirnov made 1051 mistakes (10.51%). Anderson-Darling made 8861 mistakes (88.61%). The results show that Shapiro-Wilk made the most mistakes. Kolmogorov-Smirnov made the fewest. Anderson-Darling was in the middle. All tests made more mistakes when the sample size got bigger.

Table 1. *Comparison of Type I Error Rates for Large Sample Sizes*

Sample Size	Shapiro-Wilk Test		Kolmogorov-Smirnov		Anderson-Darling	
	Frequency	%	Frequency	%	Frequency	%
200	3859	38.59	35	0.35	2898	28.98
300	5191	51.91	87	0.87	4031	40.31
400	6365	63.65	152	1.52	5126	51.26
500	7258	72.58	219	2.19	6008	60.08
600	7926	79.26	351	3.51	6862	68.62
700	8514	85.14	464	4.64	7579	75.79
800	8972	89.72	670	6.70	8158	81.58
900	9196	91.96	835	8.35	8501	85.01
1000	9420	94.20	1051	10.51	8861	88.61

This information is also shown in Figure 1, which displays the "Chance of Committing Type I Error" on the vertical axis and "Sample Size" on the horizontal axis. The graph includes three lines, each representing one of the statistical tests: the Shapiro-Wilk Test (blue line), the Kolmogorov-Smirnov Test (orange line), and the Anderson-Darling Test (grey line). For a sample size of 300, the Shapiro-Wilk Test incorrectly identified the data as non-normal 51.91% of the time, as shown by the blue line above 50% at the sample size of 300.

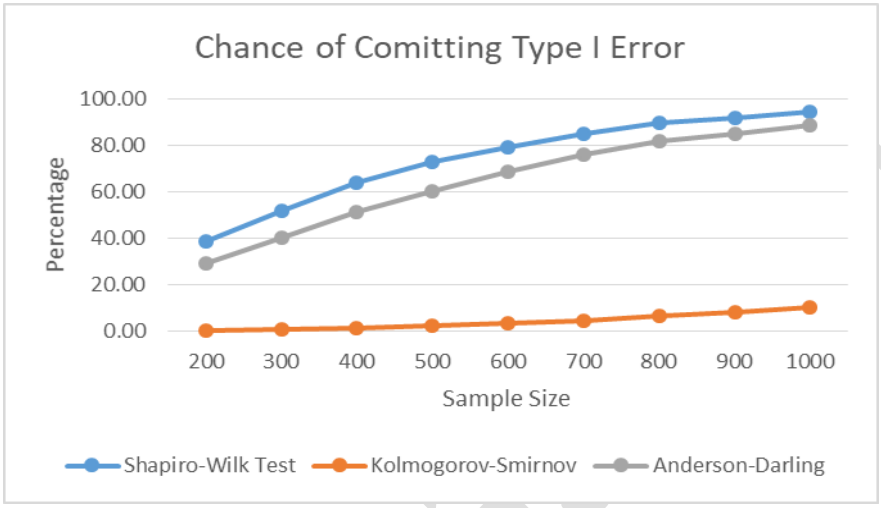

Figure 1. *Chance of Committing Type I Error for SW, KS, and AD Tests*

For the same sample size, the Kolmogorov-Smirnov Test made only 0.87% errors. This is shown by the orange line staying low. The Anderson-Darling Test had a 40.31% error rate, and the grey line is between the other two. As the sample size gets bigger, the figure shows that all three tests make more Type I errors. At sample size 700, the Shapiro-Wilk Test made 85.14% errors. The blue line is very high at this point. The Kolmogorov-Smirnov Test had a 4.64% error rate, which is still low. The Anderson-Darling Test had a 75.79% error rate, which is between the other two. At sample size 1000, the Shapiro-Wilk Test made 94.20% errors. The Kolmogorov-Smirnov Test made 10.51%, and the Anderson-Darling Test made 88.61%. These results match Chen and Genton's (2023) study which states that the Kolmogorov-Smirnov Test is less sensitive to small changes in normal data, so it makes fewer errors. The Shapiro-Wilk Test is more sensitive, so it makes more errors as sample size gets larger. Figures 2, 3, and 4 show how often each test makes both kinds of errors when the sample size is small.

Table 2 shows how often the Shapiro-Wilk Test makes mistakes with small sample size. A Type I error means it says normal data is not normal. A Type II error means it says non-normal data is normal. The table shows how the test works under different sample sizes. This helps researchers know if the Shapiro-Wilk Test gives good results when the data is small.

Looking at the Type I error rates in the table, when the sample size is only 10, the Shapiro-Wilk Test incorrectly flags normal data as not normal 632 times, and this equates to 6.32%. As the sample size becomes larger, this error rate tends to increase. For a sample size of 50, the Type I error rate rises to 15.11%, and further increases to 22.62% when the sample size is 100. By the time the sample size reaches 150, the test incorrectly identifies normal data as not normal 3086 times, or 30.86% of the time. Turning to the Type II error rates, when the sample size is a very small 10, the test fails to recognize non-normal data 3749 times, and this

is a high rate of 37.49%. However, as the sample size becomes larger, the occurrence of this error drops significantly. For a sample size of 20, the Type II error rate falls to 4.76%. Notably, for sample sizes of 60 and above, the Type II error rate becomes zero, indicating that the test correctly identifies non-normal data as non-normal in all those simulated cases.

Table 2. *Shapiro-Wilk Test Type I and Type II Error Rates for Small Sample Sizes*

Sample Size	Type I Error		Type II Error	
	Frequency	Percentage	Frequency	Percentage
10	632	6.32	3749	37.49
20	886	8.86	476	4.76
30	1059	10.59	29	0.29
40	1321	13.21	1	0.01
50	1511	15.11	0	0
60	1648	16.48	0	0
70	1780	17.8	0	0
80	1929	19.29	0	0
90	2143	21.43	0	0
100	2262	22.62	0	0
110	2353	23.53	0	0
120	2610	26.1	0	0
130	2775	27.75	0	0
140	2953	29.53	0	0
150	3086	30.86	0	0

Figure 2 shows these error rates in a visual way, and the blue line on the graph shows the "Chance of Committing Type I Error" across different small sample sizes. The blue line begins at a lower percentage for smaller sample sizes, and it goes up slowly as the sample size gets bigger from 10 to 150, which visually shows that a Type I error is more likely to happen.

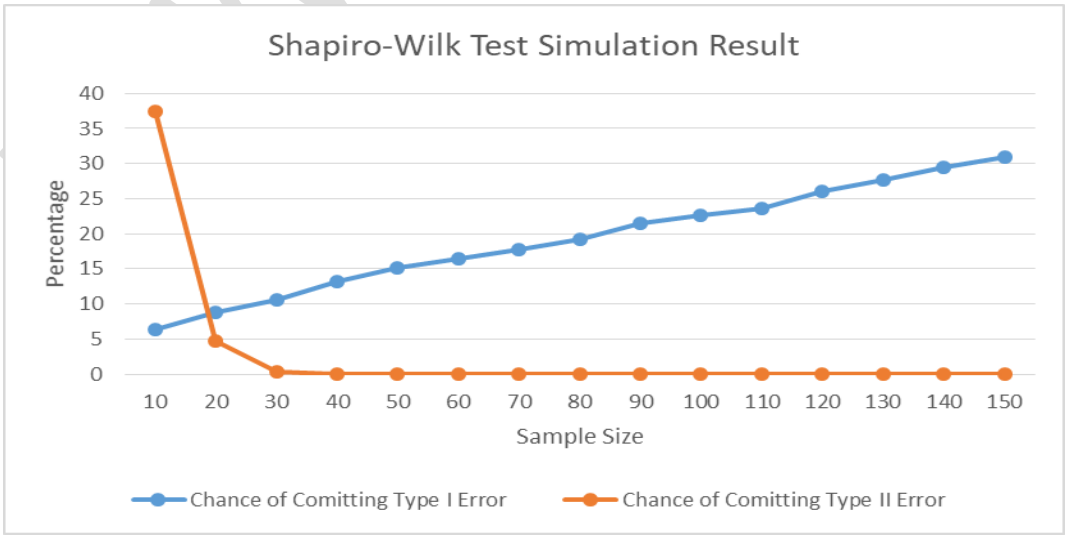

Figure 2. *Chance of Committing Type I and Type II Error for Shapiro-Wilk Test*

The orange line in Figure 2 shows the "Chance of Committing Type II Error." This line starts high for a sample size of 10 and drops quickly as the sample size gets bigger. After sample size 60, the line flattens close to zero. This matches what is shown in Table 2, where there are no Type II errors for larger small sample sizes. Looking at both types of errors, the Shapiro-Wilk Test performs better between sample sizes 20 and 60. In this range, the chance of missing non-normal data (Type II error) becomes very small, and the chance of incorrectly marking normal data as non-normal (Type I error) stays lower than it does for larger sample sizes.

Table 3 shows the Type I and Type II error rates for the Kolmogorov-Smirnov Test with small sample sizes. For example, with a sample size of 10, the test wrongly flags normal data as not normal only once out of 10,000, or 0.01%.

Table 3. *Kolmogorov-Smirnov Test Type I and Type II Error Rates for Small Sample Sizes*

Sample Size	Type I Error		Type II Error	
	Frequency	Percentage	Frequency	Percentage
10	1	0.01	9856	98.56
20	4	0.04	8751	87.51
30	6	0.06	6626	66.26
40	7	0.07	4138	41.38
50	12	0.12	2174	21.74
60	9	0.09	878	8.78
70	9	0.09	280	2.8
80	9	0.09	93	0.93
90	10	0.1	17	0.17
100	16	0.16	4	0.04
110	14	0.14	2	0.02
120	24	0.24	0	0
130	18	0.18	0	0
140	18	0.18	0	0
150	31	0.31	0	0

The Type I error rate goes up to 0.12% when the sample size is 50, and to 0.16% when the sample size is 100. At the sample size of 150, the test makes a mistake 31 times, or 0.31%, by saying normal data is not normal. The Kolmogorov-Smirnov Test has a very low Type I error rate for all small sample sizes. The Type II error rate is very high at 98.56% when the sample size is 10, meaning the test does not find non-normal data 9856 times. As the sample size gets bigger, this error rate decreases, but it stays high for smaller sample sizes. For example, at a sample size of 30, the Type II error rate is 66.26%. At a sample size of 110, it drops to 0.02%. For sample sizes of 120 or more, the Type II error rate is 0, meaning the test always finds non-normal data correctly. These results are similar to what Ramazanov and Senger found (2023), which shows that the Kolmogorov-Smirnov Test has more Type I errors when the sample size is bigger. They also say that researchers should think about both types of errors when picking a normality test, especially with small sample sizes.

Figure 3 shows the error rates for the Kolmogorov-Smirnov Test. The blue line in the figure shows the chance of making a Type I error for different sample sizes. The line stays close to zero, which shows that the chance of making a Type I error is very low.

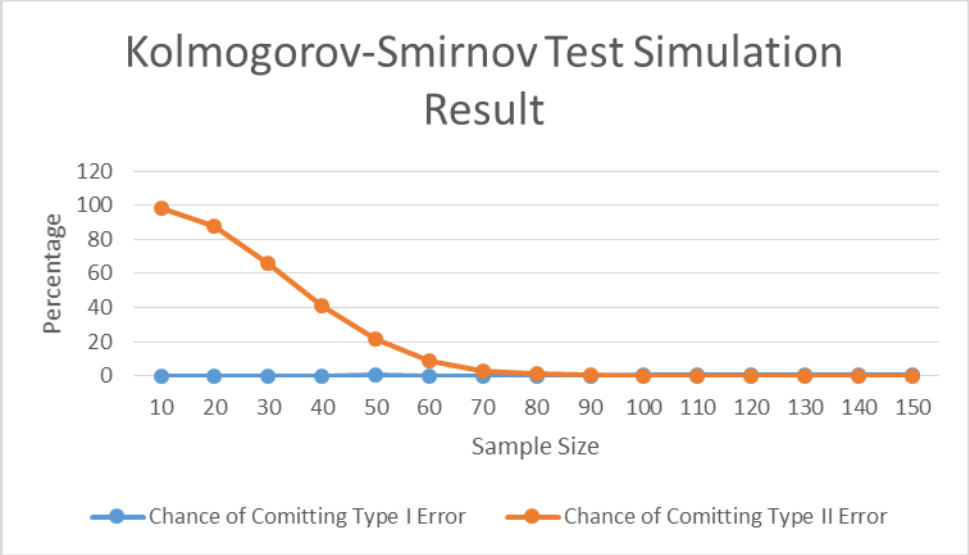

Figure 3. *Chance of Committing Type I and Type II Error for Kolmogorov-Smirnov Test*

The orange line in Figure 3 shows the Chance of Committing Type II Error. This line starts at a very high percentage when the sample size is 10. The line decreases sharply as the sample size increases because the error becomes less likely. It becomes close to zero around a sample size of 120. This confirms that the Type II error becomes very low or zero when the sample size becomes larger. The Kolmogorov-Smirnov Test shows a lower chance of identifying normal data as not normal for all small sample sizes. However, this test needs a larger small sample size, more than 60, to reduce the chance of identifying non-normal data as normal. The balancing of Type I error and Type II error is supported by this. The Shapiro-Wilk Test, however, shows a better balance between the two errors for sample sizes ranging from 20 to 60.

Table 4 presents the Type I error and Type II error rates for the Anderson-Darling Test across different small sample sizes. The Type I error rate is 6.17% for a sample size of 10, meaning the test incorrectly classifies normal data as non-normal 617 times. The Type I error increases as the sample size increases. The rate reaches 12.65% for a sample size of 50 and 17.2% for a sample size of 100. At a sample size of 150, the test identifies 2332 instances of Type I error, or 23.32%. The Type II error rate is 41.41% for a sample size of 10, indicating that the test fails to correctly identify non-normal data 4141 times. The error becomes smaller as the sample size increases. The rate becomes 0.92% for a sample size of 30. When the sample size becomes 60 or more, the test makes no Type II error. These results match the findings of Baumgartner and Kolassa (2023) since their study shows that the Anderson-Darling Test shows more Type I error as the sample size becomes larger.

Table 4. *Anderson Darling Test Type I and Type II Error Rates for Small Sample Sizes*

Sample Size	Type I Error		Type II Error	
	Frequency	Percentage	Frequency	Percentage
10	617	6.17	4141	41.41
20	796	7.96	708	7.08
30	913	9.13	92	0.92
40	1086	10.86	6	0.06
50	1265	12.65	0	0
60	1281	12.81	0	0
70	1346	13.46	0	0
80	1498	14.98	0	0
90	1642	16.42	0	0
100	1720	17.2	0	0
110	1789	17.89	0	0
120	2008	20.08	0	0
130	2139	21.39	0	0
140	2190	21.9	0	0
150	2332	23.32	0	0

Figure 4 provides a visual representation of these error rates for the Anderson-Darling Test. The blue line on the graph illustrates the "Chance of Committing Type I Error" across different small sample sizes. The blue line starts at a relatively low percentage for smaller sample sizes and gradually increases as the sample size increases from 10 to 150, visually showing the increasing likelihood of a Type I error.

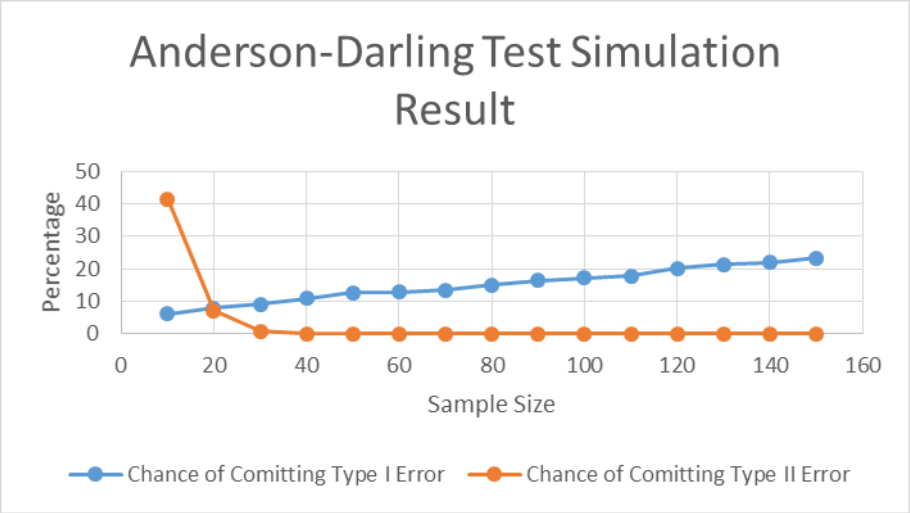

Figure 4. *Chance of Committing Type I and Type II Error for Anderson-Darling Test*

Figure 4 shows the error rates for the Anderson-Darling Test. The blue line shows the Chance of Committing Type I Error for different small sample sizes. This line starts at a low percentage when the sample size is small. The line increases gradually as the sample size

increases from 10 to 150. This shows that the chance of Type I error becomes higher with a larger sample size. The orange line in Figure 4 shows the Chance of Committing Type II Error. The line starts at a high percentage when the sample size is 10. The line drops sharply as the sample size increases because the error becomes less likely. After the sample size reaches about 60, the orange line becomes flat at zero. This confirms that there is no Type II error when the sample size is larger. The Anderson-Darling Test shows better performance for sample sizes between 20 and 60. In this range, the Type II error decreases clearly, and the Type I error stays lower compared to larger sample sizes.

CONCLUSION AND RECOMMENDATION

This simulation study shows that the Shapiro-Wilk Test, the Kolmogorov-Smirnov Test, and the Anderson-Darling Test work differently depending on the sample size. Researchers should think about their usual sample size when choosing a normality test. For sample sizes over 60, the Kolmogorov-Smirnov Test is better because it keeps the Type I error rate low. It also gets better at finding non-normal data as the sample size increases. For sample sizes between 20 and 60, the Shapiro-Wilk Test and the Anderson-Darling Test offer a good balance between power and error rates. For sample sizes under 20, it is better to use Q-Q plots for a visual check because the tests do not have much power. These suggestions apply to clean data. If the data has outliers or other problems, the balance between Type I error and Type II error might change.

Conflicts Of Interest

The author declares that there are no conflicts of interest related to the conduct and publication of this study.

Ethical Guidelines

This study did not involve human participants or personal data. However, the research procedures involving simulated datasets were reviewed and approved by the Lourdes College Research Ethics Committee.

Funding Sources

This study did not receive external funding. It was independently conducted and institutionally supported by Lourdes College.

REFERENCES

i. Anastasiou, A., Karagrigoriou, A., & Katsileros, A. (2020). Comparative evaluation of goodness of fit tests for normal distribution using simulation and empirical data. Biometrical Letters, 57(2), 237-251.

ii. Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. Journal of the American statistical association, 49(268), 765-769.

iii. Baumgartner, D., & Kolassa, J. (2023). Power considerations for Kolmogorov–Smirnov and Anderson–Darling two-sample tests. Communications in Statistics-Simulation and Computation, 52(7), 3137-3145.

iv. Biu, E. O., Nwakuya, M. T., & Wonu, N. (2020). Detection of non-normality in data sets and comparison between different normality tests. Asian J. Probab. Stat, 5(4), 1-20.

v. Chen, W., & Genton, M. G. (2023). Are you all normal? It depends!. International Statistical Review, 91(1), 114-139.

vi. de Souza, R. R., Toebe, M., Mello, A. C., & Bittencourt, K. C. (2023). Sample size and Shapiro-Wilk test: An analysis for soybean grain yield. European Journal of Agronomy, 142, 126666.

vii. Demir, S. (2022). Comparison of normality tests in terms of sample sizes under different skewness and Kurtosis coefficients. International Journal of Assessment Tools in Education, 9(2), 397-409.

viii. Kolmogorov, A. (1933). Sulla determinazione empirica di una legge didistribuzione. Giorn Dell'inst Ital Degli Att, 4, 89-91.

ix. Ramazanov, S., & Senger, Ö. (2023). Küçük, eşit ve büyük örnek hacimlerinde Wald-Wolfowitz, Kolmogorov-Smirnov ve Mann-Whitney testlerinin I. tip hata oranlarının karşılaştırılması. Ardahan Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 5(2), 137-144.

x. Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3-4), 591-611.

xi. Wijekularathna, D. K., Manage, A. B., & Scariano, S. M. (2020). Power analysis of several normality tests: A Monte Carlo simulation study. Communications in Statistics-Simulation and Computation, 51(3), 757-773.