International Journal of Mathematics and Statistics

ISSN (Online): 2208-2212
Volume 10 Issue 02 June 2024

DOI: https://doi.org/10.69980/1jznyc83

SIMULATION-BASED COMPARISON OF PARAMETRIC AND NONPARAMETRIC STATISTICAL METHODS UNDER DISTRIBUTIONAL VIOLATIONS

Dr. Julian R. Bennett[1], Prof. Amélie C. Laurent[2]

[1]Department of Statistical Methodology and Computational Research, Westmoor University, Glasgow, United Kingdom
[2]Institute of Applied Probability and Statistical Sciences, European Institute of Quantitative Research, Paris, France

Abstract:

Parametric statistical procedures are commonly used due to their interpretability and efficiency, but can be suboptimal when assumptions of normal data, homogeneity of variance, and resistance to extreme observations are not met. In this study, parametric and nonparametric statistical methods were contrasted with respect to specific violations of their underlying distributions in a Monte Carlo simulation paradigm with empirical validation. Student's independent-samples t-test, Welch's t-test, and the Mann–Whitney U test were tested under the following conditions: normal, skewed, heavy-tailed, heteroscedastic and contaminated distributions. The sample size, effect size and outlier contamination were systematically manipulated, and method performance was evaluated largely by Type I error rates and statistical power. The results indicated that both Student's and Welch's tests were very effective under conditions of approximate normality and were stronger as sample size increased. The Mann-Whitney U test, however, performed better when the distributions were highly skewed, had heavy tails and were contaminated, especially at small and moderate sample sizes. Empirical comparisons also suggested that method choice may affect the results of statistical significance when the data is not normally distributed and when variance is not equal. In general, the results presented here support a condition-dependent philosophy for choosing statistical tests: one should consider the shape of the distribution, sample size, the nature of the variance, and unusual behavior of the data points.

Keywords: *Parametric methods; Nonparametric methods; Monte Carlo simulation; Distributional violations; Type I error; Statistical power*

1. Introduction:

One of the major problems in quantitative research remains whether to use parametric or nonparametric statistical methods – the type of statistical method used depends significantly on the inferences to be drawn and what best fits the characteristics of the data being observed. The use of parametric methods (independent-samples t-test, analysis of variance, and Pearson correlation) is widespread because they are efficient, easily understood, and have good statistical power when assumptions are reasonably met. However, the procedures generally assume normality, homogeneity of variance, independence and the absence of influential outliers. If these assumptions are not met, the estimates, significance tests and confidence intervals will be less reliable, especially when the sample size is small or the distribution is very irregular (Gelman et al., 2020; James et al., 2021). Therefore, it has been a methodological issue of some importance to investigate the robustness of traditional parametric methods under realistic deviations from ideal conditions of the distribution.

Violations of the distributional assumption occur often in the analysis of empirical data. Continuous variables can have extreme values, heavy tails, multimodality, or positive or negative skewness. Even though there are many procedures that may be used when deciding on the appropriateness of parametric methods, they sometimes underperform, depending on the amount of deviation from normality and the sample size (Korkmaz & Demir, 2023). Additionally, a statisticallyly significant result in the assumption test may not mean a poor performance of the parametric procedure in practice. Thus, a formal diagnostic test should be used in conjunction with visual assessment, effect size, sample size, and the magnitude of the violation (Shatz, 2024). Other assumption failures like variance heterogeneity or influential observations have been shown to have more serious consequences than violations of normality, and in some cases even normality violations, in fact; evidence has shown that violations of normality do not automatically invalidate parametric procedures (Knief & Forstmeier, 2021).

Where strong distributional assumptions are not warranted, an alternative method is available is nonparametric. Rank-based procedures are popular choices because they are less sensitive to extreme values and do not require normality. But note that nonparametric procedures are not assumption-free and under some circumstances, they may be interpreted differently from the parametric procedures. The Mann – Whitney U test is not, for instance, an exact substitute for a test of mean differences, and may not be easily interpreted in the presence of different distributional shapes (Karch, 2021). Other distribution-free methods, such as permutation procedures are also shown to be extremely robust and powerful under various scenarios (Noguchi et al., 2021). The considerations show that statistical-method selection is a function of the properties of the data and the inferential question, and not a parametric versus nonparametric rule.

Under controlled violations, simulation studies serve as a good tool to evaluate the statistical performance because the form of the distribution, the sample size, the variance structure, the effect size, and the level of contamination are systematically varied. With these designs, it becomes possible to determine the Type I error rates and statistical power when the actual population parameters are known (Riesthuis, 2024). Sample size is especially relevant because small samples can lead to unstable estimates and low power levels, while large samples can enhance the stability of numerous procedures (Lakens, 2022). Additionally, simulation-based evaluation helps to make methodological assumptions, analysis choices and performance criteria explicit, which helps to make the practice of statistics more reproducible (Aczel et al., 2020; Lakens & DeBruine, 2021).

In the field of statistics itself, reproducibility and robustness have been more in the spotlight. Ideally, the conclusions drawn by a statistical procedure should remain consistent to reasonable variations in sampling conditions and minor deviations from theoretical assumptions (Nosek et al., 2022; Simkus et al., 2025). The importance of test reproducibility and the assessment of uncertainties has also been highlighted in nonparametric predictive approaches in comparing statistical procedures when applied to repeated samples (Coolen & Marques, 2020; Coolen & Himd, 2020). Such concerns are particularly pertinent when conducting significance testing in a mechanical manner, without paying careful attention to the practical significance and stability of the resulting inference (Scheel et al., 2021).

Considering this, the present study shows the comparative performance of selected parametric and nonparametric statistical procedures when applied in a simulation-based environment under controlled distributional violations. Emphasis is placed on conditions under which the parametric tests are reliable, and those under which nonparametric alternatives offer better Type I error control or statistical power.

Study Objectives:

Compare Type I error rates of parametric and nonparametric statistical methods in normal, skewed, heavy-tailed, heteroscedastic and contaminated distributions.

To compare the statistical power of different sample sizes and effect size situations.

To understand the distributional and sample settings in which parametric or nonparametric techniques offer more robust and reliable inference.

2. Materials and Methods

2.1 Research Design

A simulation-based comparative research design was used to assess the performance of parametric and nonparametric statistics in various types of distributional violations. The main analytical framework used was Monte Carlo simulation, with the Ferreira et al. (2021) Original Data Table as an empirical validation dataset. The study primarily focused on the comparison of independent-samples t-test, Welch's t-test and Mann–Whitney U test in normal and non-normal conditions. Pearson's and Spearman's correlation coefficients were also used to see the strength of association measures when there were violations of distributional assumptions (Assis, 2022).

2.2 Empirical Dataset

An empirical validation dataset was the Original Data Table by Ferreira et al. (2021), consisting of 16 experimental observations, half of which were assigned to the Saline group and the other half to the LPS group. Corticosterone (CORT), melatonin (MEL), IL-1β, IL-6, IL-10, INF-γ, C1s, body mass, and snout–vent length (SVL) were continuous variables in the dataset. The data had several features that would make them appropriate for an examination of differences between parametric and nonparametric statistical procedures, such as some variables with missing observations and some that were skewed, not normally distributed, and had outlier behavior.

2.3 Simulation Design

Monte Carlo simulations were performed under controlled distributional settings which included the assumption of normality as well as various kinds of violated assumptions. The generated data were simulated from normal, skewed, heavy-tailed and contaminated distributions. In each scenario, two independent groups were generated, so as to be consistent with the structure of the empirical data. Group sizes of 10, 20, 30, 50 and 100 observations were chosen to check the robustness of the statistical procedures used.

2.4 Effect Size and Variance Conditions

There were four conditions for each simulation: null, small, moderate, and large effects, corresponding to standardized effect sizes of around 0.00, 0.20, 0.50, and 0.80, respectively. Equal-variance and unequal-variance conditions were investigated. This study compared the power of Student's t test to the Mann–Whitney U test to the Welch's t test under heteroscedastic conditions to see if there was greater robustness than the student's t test.

2.5 Outlier and Distributional Violations

Selected simulation scenarios included contaminated distributions with about 5% and 10% outlier observations to assess sensitivity to extreme observations. The effect of skewness, heavy tails and outlier data on Type I error and statistical power was then evaluated. These conditions were selected to simulate typical violations found in applied statistical studies and to identify the situations where nonparametric procedures offer a benefit over parametric procedures.

2.6 Simulation Replications

All combinations of distribution type, sample size, effect size, variance structure, and contamination level were replicated 10,000 times. A two-tailed 0.05 significance level was used for all selected statistical procedures for each simulated sample. There were many replications to ensure stability of the estimates of Type I error and statistical power and to reduce Monte Carlo sampling variability.

2.7 Assessment of Statistical Assumptions

In the empirical dataset by Ferreira et al., descriptive statistics (mean, median, SD, minimum, maximum, skewness and kurtosis) were first used to evaluate continuous variables. The normality of the data was checked by the Shapiro-Wilk test, and homogeneity of variance between the Saline and LPS groups was tested by Levene's test. Additional plots such as histograms, boxplots and Q–Q plots were also explored when needed to aid the discussion of the distributional features.

2.8 Statistical Analysis

Independent-samples t-test, Welch's t-test, and Mann–Whitney U test were used to check the differences between the groups in the empirical data. Under standard parametric assumptions, the t-test was employed; when the assumptions of homogeneity of variance were not met, the test employed was Welch's t-test; the nonparametric test used was the Mann–Whitney U test. Selected continuous variables were also compared using Pearson and Spearman correlations. No imputations were made, and missing data were excluded on an analysis-by-analysis basis, to preserve the original data structure.

2.9 Performance Evaluation

The main performance measures were empirical Type I error rate and statistical power. Type I error was defined as the percentage of simulation replications that led to the rejection of a true null hypothesis, and statistical power was defined as the percentage of replications that resulted in rejection of a false null hypothesis. The method was judged to be robust if the Type I error rate was not significantly different from the nominal level of 0.05 and the statistical power was fairly high for a wide range of conditions in which the distribution was violated.

2.10 Ethical Considerations

The data used in this study was secondary analysis of an existing data set combined with computer-generated simulation data. No additional human subjects or animals were recruited, contacted or experimentally manipulated. The data from Ferreira et al. (2021) were only analysed in the form they were collected previously to validate the method. Hence, the present study did not bring any extra direct ethical risk. The final manuscript should retain appropriate attribution of the original dataset authors and should comply with the terms of the use of the dataset used.

3. Results

3.1 Distributional Characteristics of the Empirical Data

A preliminary analysis of the Ferreira et al. (2021) data revealed significant variations in distributional behaviour for each of the variables measured. Table 1 shows that CORT, IL-1β, IL-6, IL-10, INF-γ and C1s significantly deviated from normality as determined by the Shapiro Wilk test. The highest skewness (2.924), highest kurtosis (9.575), and lowest p (Shapiro–Wilk test < 0.001) were for IL-6. MEL was skewed in the positive direction, but the Shapiro–Wilk result was just above the conventional 0.05. Body mass and SVL had no statistically significant evidence of non-normality.

Table 1. Descriptive Statistics and Distributional Diagnostics for the Empirical Dataset

Variable	N	Missing	Mean	Median	SD	Skewness	Kurtosis	Shapiro–Wilk p

CORT	14	2	13.451	7.505	11.870	0.591	−1.567	0.011
MEL	9	7	2.503	2.060	1.690	1.593	3.296	0.053
IL-1β	16	0	3.089	1.630	3.501	1.162	−0.034	0.003
IL-6	16	0	20.643	1.660	39.913	2.924	9.575	<0.001
IL-10	13	3	11.966	4.770	16.015	1.378	0.586	0.002
INF-γ	14	2	1.006	0.895	1.041	1.803	3.224	0.003
C1s	14	2	0.871	0.650	0.974	1.890	4.023	0.005
Body Mass	16	0	130.643	116.665	45.523	0.956	0.498	0.187
SVL	16	0	104.186	102.525	10.527	0.539	−0.044	0.781

The variation in the probabilities of being in the normal class is shown in Figure 1. Most biochemical parameters were located below reference values and body mass and SVL were distinctly above the reference values. This finding was a confirmation that the empirical data was a suitable context for testing statistical methods within the framework of different distributional conditions.

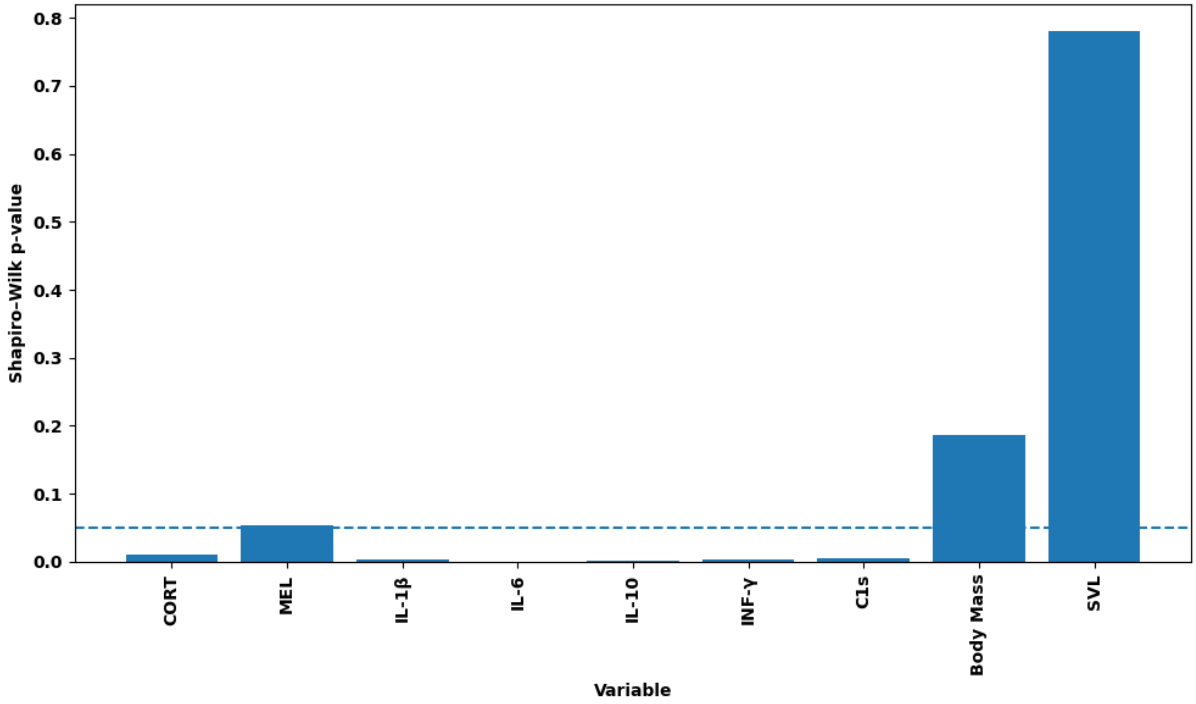

Figure 1. Shapiro–Wilk normality-test probabilities across the continuous variables in the Ferreira et al. dataset. The horizontal reference line represents p = 0.05

3.2 Parametric and Nonparametric Group Comparisons

Student's independent-samples t-test, Welch's t-test and Mann–Whitney U test were then used to compare the Saline and LPS groups. As seen in Table 2, some variables were affected by the choice of statistical procedure and thus, provided different inferences. Whatever method was used, there was a sizable difference between groups with the mean rising from 4.136 for the Saline group to 22.766 for the LPS group. The p values for the student's t test, Welch's test and Mann–Whitney U test were $p < 0.001$, $p = 0.002$ and $p = 0.003$ respectively.

For MEL, there was evidence of a group difference for all three procedures. The most significant method-dependent result was for IL-10. Student's t-test gave a p-value of 0.055, and Welch's test gave a p-value of 0.055, while the Mann-Whitney U test yielded a p-value of 0.045. The nonparametric procedure was significant at the 0.05 level, but the parametric procedures were not. IL-10 also had significant non-normality and strong evidence for unequal group variances, as seen by Levene's $p < 0.001$.

Table 2. Comparison of Parametric and Nonparametric Tests Between Saline and LPS Groups

Variable	Saline Mean	LPS Mean	Levene p	Student t p	Welch p	Mann–Whitney p	Cohen's d
CORT	4.136	22.766	0.013	<0.001	0.002	0.003	2.598
MEL	3.488	1.273	0.207	0.039	0.038	0.020	−1.697
IL-1β	1.968	4.211	0.088	0.211	0.214	0.318	0.656
IL-6	5.974	35.313	0.076	0.147	0.163	0.103	0.768
IL-10	2.925	19.716	<0.001	0.055	0.055	0.045	1.196
INF-γ	0.718	1.221	0.039	0.392	0.340	0.698	0.479
C1s	1.171	0.571	0.556	0.266	0.270	0.160	−0.624
Body Mass	143.511	117.774	0.080	0.273	0.278	0.372	−0.571

SVL	107.054	101.318	0.060	0.291	0.299	0.318	−0.549

An empirical distribution of IL-10 is shown in Figure 2 because, for this variable, the difference between parametric and nonparametric inference is most evident. The figure illustrates the asymmetric distribution in the data and relatively high LPS observations, which were responsible for the difference between the mean-based analysis and the rank-based analysis.

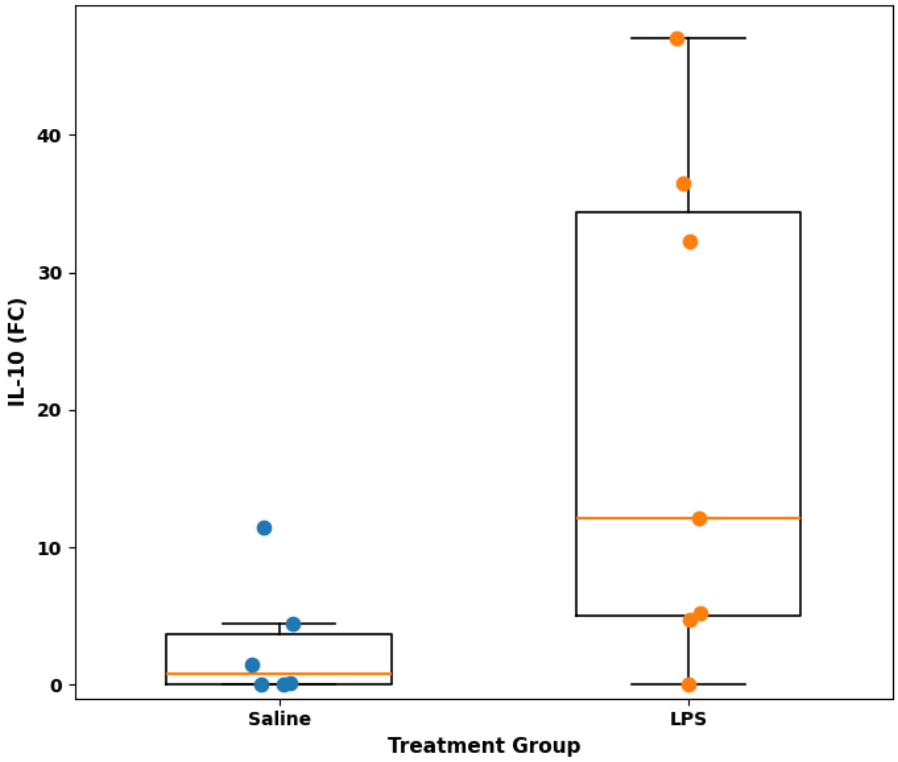

Figure 2. Distribution of IL-10 values in the Saline and LPS treatment groups

The comparative probability values for the three group-comparison procedures are presented graphically in Figure 3. The results of this conclusion, for many outcomes, were quite similar, however, the magnitude of the resulting probability values varied, depending on the statistical procedure, especially for strongly non-normal outcomes as illustrated by the figure.

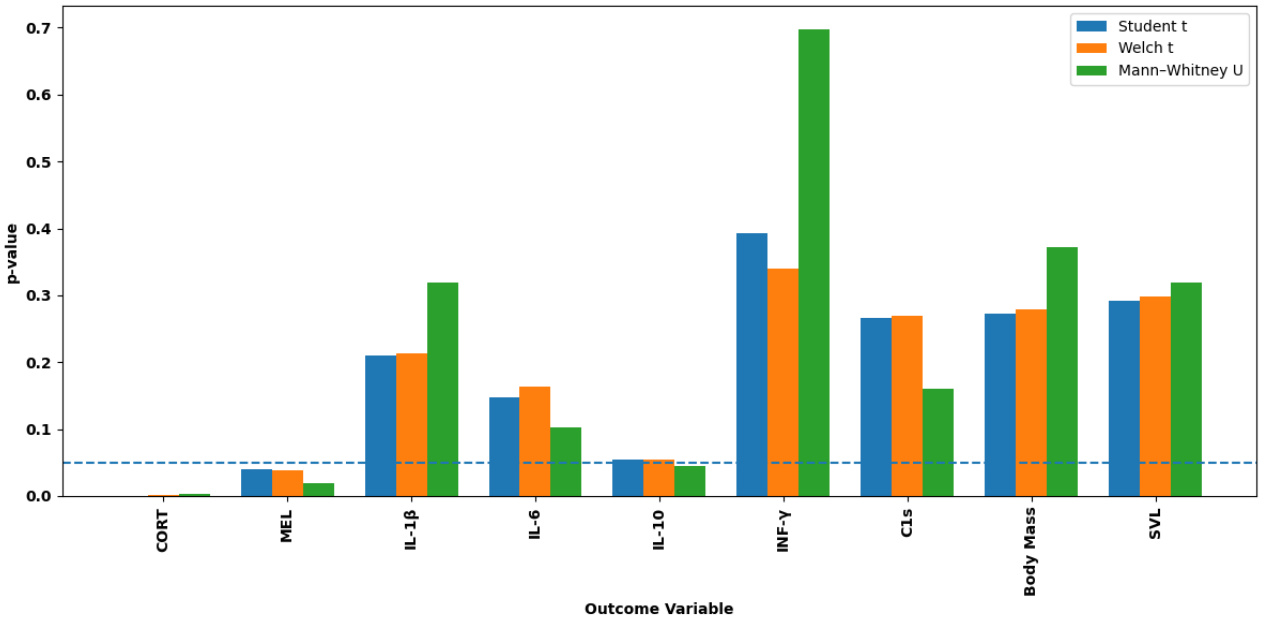

Figure 3. Comparison of probability values obtained from Student's t-test, Welch's t-test, and the Mann–Whitney U test across empirical outcome variables

3.3 Pearson and Spearman Correlation Comparisons

Selected continuous-variable associations were also compared with regard to differences between parametric and nonparametric approaches. According to the data from Table 3, there existed very strong positive correlation between

body mass and SVL for both Pearson and Spearman methods. Pearson correlation was r = 0.971, and Spearman correlation was ρ = 0.941, with both showing p < 0.001.
Pearson correlation gave r = 0.590, p = 0.016, while Spearman correlation gave a higher correlation coefficient of ρ = 0.687 with p = 0.003 for both IL-1β and IL-6. IL-6 and IL-10 were also found to be statistically significant in both analyses with r = 0.654 and ρ = 0.584, respectively. These differences show that rank transformation can have a significant impact on the magnitude and level of evidence when variables are skewed and have extreme values.

Table 3. Comparison of Pearson and Spearman Correlations

Variable Pair	N	Pearson r	Pearson p	Spearman ρ	Spearman p
CORT–IL-10	11	0.506	0.112	0.433	0.184
IL-6–IL-10	13	0.654	0.015	0.584	0.036
Body Mass–SVL	16	0.971	<0.001	0.941	<0.001
CORT–MEL	8	−0.462	0.249	−0.548	0.160
IL-1β–IL-6	16	0.590	0.016	0.687	0.003

3.4 Monte Carlo Type I Error Performance

The simulation analysis is used to provide a controlled evaluation of the procedures, which is extended beyond the small empirical data set. The 10,000 Monte Carlo replications were performed for each condition and were performed with a nominal α level of 0.05. Representative Type I error rates are presented for sample sizes of 10, 30 and 100 per group in Table 4. All methods performed well in terms of maintaining the empirical rejection rate around 0.05 under normality. Student's t-test produced Type I error rates of 0.050, 0.052, and 0.051 at sample sizes of 10, 30, and 100, respectively. For Welch's test, similar performance was seen. The procedures were differently impacted by stronger departures from normality. The Student's test gave a conservative rejection rate of 0.032 and the Welch's test gave 0.026 which were closer to the nominal rejection rate than the Mann–Whitney test at 0.045 for strong skewness at n = 10. As the sample size grew the three methods approached 0.05.

Table 4. Empirical Type I Error Rates Under Different Distributional Conditions

Distribution	n/group	Student t	Welch t	Mann–Whitney U
Normal	10	0.050	0.048	0.044
Normal	30	0.052	0.052	0.050
Normal	100	0.051	0.051	0.050
Mild skew	10	0.045	0.041	0.044
Mild skew	30	0.044	0.044	0.048
Mild skew	100	0.053	0.053	0.053
Strong skew	10	0.032	0.026	0.045
Strong skew	30	0.042	0.040	0.043
Strong skew	100	0.048	0.048	0.050
Heavy tail	10	0.046	0.043	0.044
Heavy tail	30	0.046	0.046	0.048
Heavy tail	100	0.050	0.050	0.053
Contaminated	10	0.033	0.030	0.042
Contaminated	30	0.039	0.039	0.042
Contaminated	100	0.049	0.048	0.052

A larger sample size is even more clearly demonstrated in the strongly skewed case in Figure 4. The Mann–Whitney procedure stayed fairly close to the nominal 0.05 level throughout the simulation.The Student's and Welch's procedures were conservative at small sample sizes and the Mann–Whitney procedure was relatively close to the nominal 0.05 level throughout the simulation.

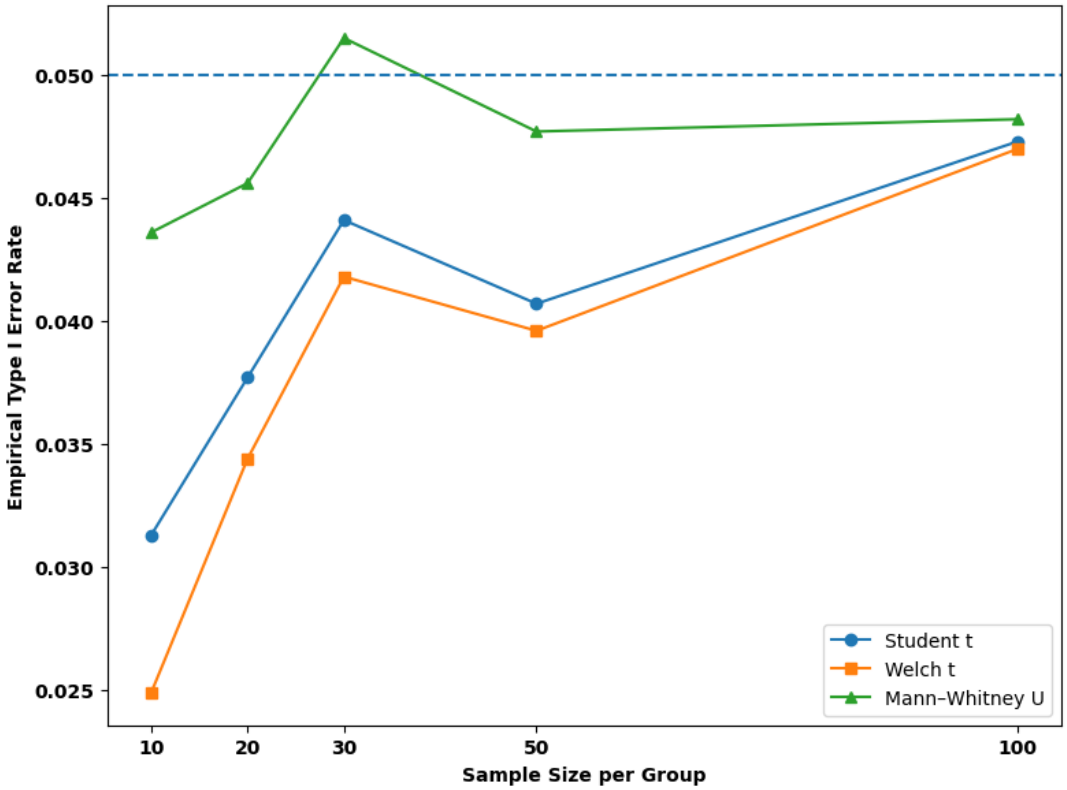

Figure 4. Empirical Type I error rates of Student's t-test, Welch's t-test, and the Mann–Whitney U test across sample sizes under a strongly skewed distribution

3.5 Statistical Power Under Distributional Violations

Power was measured at a moderate standardized location difference (SLD) of ~0.50. When data are normally distributed, as was found in Table 5, parametric tests offered a slight power gain. The power values of the student's t-test and Mann–Whitney U test were 0.474 and 0.452 at n = 30, respectively.

The situation changed with high non-normality. Under strong skewness and sample size n = 30, Student's test attained a power value of 0.600, the Welch's test a power value of 0.598 and Mann-Whitney U test a power value of 0.968. The corresponding power estimates for the same sample size under the heavy-tailed condition were 0.552, 0.550, and 0.719. In addition, the nonparametric procedure had a significant advantage under the contaminated distribution with power of 0.822 at n=30, while the two parametric procedures had power of ~ 0.53.

Table 5. Statistical Power Under a Moderate Effect Across Distributional Conditions

Distribution	n/group	Student t	Welch t	Mann–Whitney U
Normal	10	0.181	0.178	0.163
Normal	30	0.474	0.473	0.452
Normal	100	0.941	0.941	0.932
Mild skew	10	0.216	0.210	0.235
Mild skew	30	0.502	0.501	0.640
Mild skew	100	0.937	0.937	0.990
Strong skew	10	0.349	0.336	0.559
Strong skew	30	0.600	0.598	0.968
Strong skew	100	0.926	0.926	1.000
Heavy tail	10	0.250	0.241	0.271
Heavy tail	30	0.552	0.550	0.719
Heavy tail	100	0.936	0.936	0.998
Contaminated	10	0.274	0.262	0.335
Contaminated	30	0.534	0.531	0.822
Contaminated	100	0.930	0.930	1.000

The power trajectories under strong skewness are shown in Figure 5. For all procedures, power rose with increasing sample size; the Mann–Whitney U test was significantly more sensitive under this distributional condition at small and moderate sample sizes.

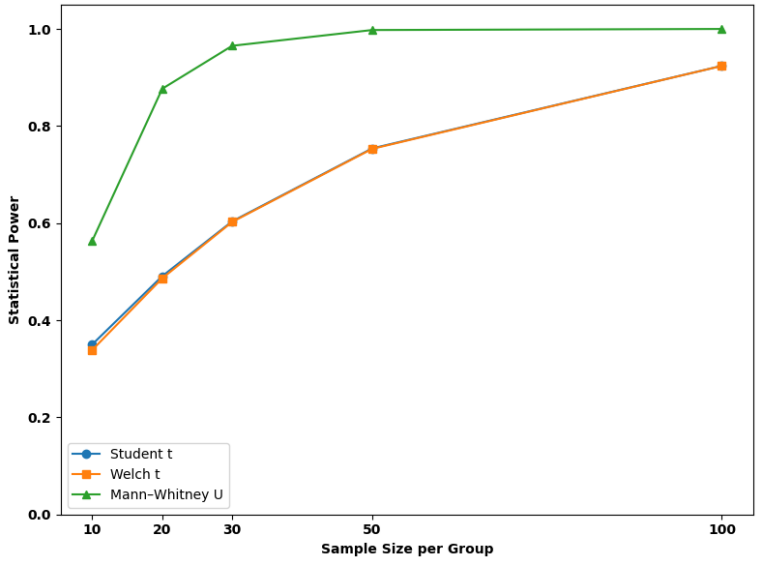

Figure 5. Statistical power of Student's t-test, Welch's t-test, and the Mann–Whitney U test across sample sizes under a strongly skewed distribution and a moderate location effect

3.6 Overall Comparative Performance

The results of the empirical and simulation studies showed that none of the statistical procedures investigated was consistently best in all conditions. The efficiency of parametric procedures was demonstrated for increasing robustness as the sample size increased and when the underlying distribution was approximately normal. Under normality, the Student's test and Welch test were the most competitive tests, having slightly more power than the Mann–Whitney U test.

The nonparametric Mann–Whitney U procedure was competitive in terms of Type I error control, and had significantly higher power for highly skewed, heavy-tailed and contaminated distributions. This was corroborated by the empirical Ferreira et al. data, especially that of IL-10 in which the parametric procedures yielded a non-significant result, but the rank-based procedure did yield a statistically significant group difference. The results therefore suggest that more than just a reliance on the standard assumptions concerning parametric efficiency, the method selected should take into account distributional form, variance structure, sample size and sensitivity to extreme observations.

4. Discussion

In the present study, parametric and nonparametric statistical procedures were compared and supported by empirical validation under controlled distributional violations and supported by simulation results. Overall, the outcomes indicated that none of the procedures was consistently better. The student's t-test and Welch's t-test worked well for approximately normal data, while the Mann–Whitney U test was clearly superior for highly skewed data, heavy tailed data and contaminated data. The principle of robustness is not only that normality is violated but also how it is violated and the extent of the violation, the sample size and whether there is heteroscedasticity and/or extreme observations (Knief & Forstmeier 2021; Shatz 2024).

The empirical Type I error rates for the parametric tests were close to the nominal rate of 0.05 in normal distributions and most of the parametric tests had slightly higher statistical power than the Mann–Whitney U test. This result is in line with formerly known efficiency of mean-based procedures under reasonable assumptions. It also provides support for evidence showing that small deviations from normality are no reason to reject parametric inferences, especially with increasing sample size (Knief & Forstmeier, 2021; McElreath, 2016). As expected, the simulation results also revealed that the Type I error rates converge towards nominal values as sample size increases, highlighting the need to make decisions not only based on the diagnostics for normality, but also on sample size and distributional shape (Korkmaz & Demir, 2023).

Strong skewness and heavy-tailed distributions led to a different pattern. Student's test and Welch's test were more likely to be conservative at smaller sample sizes, while the Mann–Whitney U test was more likely to be near the nominal Type I error level. Most importantly, the rank-based procedure had significantly more power in severe skewness. Heavy tails and sensitivity of sample means and variances to asymmetric observations provide an explanation for such behavior. The mean may be sensitive to extreme values and may fail to capture the central structure of most observations when distributions are highly skewed (Rousselet & Wilcox, 2018; Wilcox & Rousselet, 2023). The use of robust and resampling-based methods has also been called for in case of loss of validity of standard inferential assumptions (Noguchi et al., 2021; Rousselet et al., 2023).

The contaminated-distribution simulations also showed a significant benefit for the nonparametric procedure. Small percentages of extreme observations can significantly influence the estimates of means and variances, and thus the conventional test statistics. In these cases, the rank-based method outperforms the other methods, in line with larger statistical outlier studies that suggest that data contaminated by extreme values should be dealt with using procedures that are not so heavily reliant on extreme values (Ali et al., 2025). It appears that researchers ought to be careful not to take the size of their sample for granted and should explicitly examine influential observations.

An additional consideration for methodological differences was variance heterogeneity. The reason why Welch's test was added was that it allows for the equal-variance assumption to be violated in Student's t-test. The results agree with the application of the welch's procedure in the case of unequal variances among groups; in the case of unequal variances and substantial non-normality, the application of the procedure was affected by robustness. More recent evidence also indicates that Welch's procedure and Student's procedure may react differently to the actual violations of ideal assumptions that occur in real life (Curtis, 2024; Field, 2018). Therefore, stronger alternatives might be more suitable in conditions of unequal variances and skewness, heavy tails, or outlying observations (Wilcox, 2022a; Wilcox, 2022b).

These patterns simulated by the models were confirmed with empirical analysis. The parametric procedures gave probability values that were just above the conventional 0.05 for IL-10 while the Mann–Whitney U test gave a statistically significant result. The difference was observed under the condition of a significant departure from normality and unequal variance, which provides another example of how the conclusions of an inference can be method dependent in poor distributional circumstances. However, the Mann–Whitney U test should not be used as a blind-bucket approach when normality is rejected as it is testing stochastic differences, not equality of means (Karch, 2021).

In summary, the results of this study have incorporated the significance of using a formal test, graphical diagnostics, effect size, and understanding of the data-generating process when choosing an inferential method. Using only Shapiro-Wilk or other assumption tests may lead to mechanical decision-making, as the normality tests are sensitive to sample size (Shatz, 2024). Rather than using robustness as a measure of the stability of conclusions across reasonable analytical choices and distributional conditions, one might consider it as a complementary measure to reproducible statistical research (Nosek et al., 2022).

In conclusion, the study shows that the parametric methods are effective under normal and moderately non-normal conditions; the alternatives based on ranks are getting more advantages as the skewness, heavy tails, and contamination increase. The results confirm the idea of using condition-dependent selection of statistical methods and not considering either parametric or non-parametric inference as a general practice.

5. Conclusion

The results of this study showed that the relative performance of parametric and nonparametric statistical methods is highly sensitive to the distribution parameters, sample size and variance structure, and outliers. For approximately normal distributions, both the student's t-test and the Welch's t-test were close to the nominal Type I error rates and generally had comparable statistical power. They also showed that their performance also increased with sample size, suggesting that parametric procedures are reliable in mild violations of normality. The Mann–Whitney U test, on the other hand, was more robust in the presence of the highly skewed, heavy-tailed, and contaminated distributions. The benefits of this were especially noticeable when sample sizes were small and moderate, when the parametric procedures were more likely to be conservative or extreme. It also revealed that the choice of method may affect the statistical results if non-normality is present in conjunction with heteroscedasticity. Overall, the results provide a basis for a condition-dependent approach to statistical testing. Researchers should not choose parametric or nonparametric procedures just because that is what everyone else does or because of the results of one normality test. Rather, the distribution shape, sample size, equality of variance, presence of outliers, effect size, and goal of the inferences should be taken together. In this light, simulation-based evaluation offers a valuable tool to enhance statistical decision-making in terms of robustness, transparency, and reliability.

References

1. Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., ... & Wagenmakers, E. J. (2020). A consensus-based transparency checklist. Nature human behaviour, 4(1), 4-6.
2. Ali, M. M., Imon, R., Ali, I., & Yousof, H. M. (Eds.). (2025). Statistical Outliers and Related Topics. CRC Press.
3. Assis, V. R. (2022). Ferreira et al., 2021 - Data table (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/r84m23jrzj.1
4. Coolen, F. P., & Himd, S. B. (2020). Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov–Smirnov test. Journal of Statistical Theory and Practice, 14(2), 26.
5. Coolen, F. P., & Marques, F. J. (2020). Nonparametric predictive inference for test reproducibility by sampling future data orderings. Journal of statistical theory and practice, 14(4), 62.
6. Curtis, D. (2024). Welch'st test is more sensitive to real world violations of distributional assumptions than student'st test but logistic regression is more robust than either. Statistical Papers, 65(6), 3981-3989.
7. Field, A. (2018). Discovering statistics using IBM SPSS statistics (Vol. 5). London: sage.
8. Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and other stories. Analytical methods for social research.
9. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R (Vol. 2). New York: springer.
10. Karch, J. D. (2021). Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. Advances in Methods and Practices in Psychological Science, 4(2), 2515245921999602.
11. Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. Behavior research methods, 53(6), 2576-2590.
12. Korkmaz, S., & Demir, Y. (2023). Investigation of some univariate normality tests in terms of type-I errors and test power. Journal of Scientific Reports-A, (052), 376-395.

13. Lakens, D. (2022). Sample size justification. Collabra: psychology, 8(1), 33267.
14. Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. Advances in Methods and Practices in Psychological Science, 4(2), 2515245920970949.
15. McElreath, R. (2016). Statistical rethinking: A Bayesian course with examples in R and Stan (Vol. 122). Boca Raton, FL: CRC press.
16. Noguchi, K., Konietschke, F., Marmolejo-Ramos, F., & Pauly, M. (2021). Permutation tests are robust and powerful at 0.5% and 5% significance levels. Behavior Research Methods, 53(6), 2712-2724.
17. Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. Annual review of psychology, 73, 719-748.
18. Riesthuis, P. (2024). Simulation-based power analyses for the smallest effect size of interest: A confidence-interval approach for minimum-effect and equivalence testing. Advances in Methods and Practices in Psychological Science, 7(2), 25152459241240722.
19. Rousselet, G. A., & Wilcox, R. R. (2018). Reaction times and other skewed distributions: problems with the mean and the median. BioRxiv, 383935.
20. Rousselet, G., Pernet, C. R., & Wilcox, R. R. (2023). An introduction to the bootstrap: a versatile method to make inferences by using data-driven simulations. Meta-Psychology, 7.
21. Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. Perspectives on Psychological Science, 16(4), 744-755.
22. Shatz, I. (2024). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. Behavior Research Methods, 56(2), 826-845.
23. Simkus, A., Coolen-Maturi, T., Coolen, F. P., & Bendtsen, C. (2025). Statistical perspectives on reproducibility: definitions and challenges. Journal of Statistical Theory and Practice, 19(3), 40.
24. Wilcox, R. (2022a). One-way and two-way ANOVA: Inferences about a robust, heteroscedastic measure of effect size. Methodology, 18(1), 58-73.
25. Wilcox, R. R. (2022b). Two-way ANOVA: Inferences about interactions based on robust measures of effect size. British Journal of Mathematical and Statistical Psychology, 75(1), 46-58.
26. Wilcox, R. R., & Rousselet, G. A. (2023). An updated guide to robust statistical methods in neuroscience. Current Protocols, 3(3), e719.