International Journal of Mathematics and Statistics
ISSN (Online): 2208-2212
Volume 11 Issue 01 March 2025
DOI: 10.69980/twcd3g90

# SIMULATION-BASED ASSESSMENT OF STATISTICAL POWER ESTIMATION FOR NONPARAMETRIC HYPOTHESIS TESTING

**Dr. Hannah Mitchell[1*], Dr. Thomas Berger[2], Dr. Claire Thompson[3]**

*[1*]Department of Statistical Methods and Probability, Westbridge Institute of Mathematical Sciences, United Kingdom, Email: hhannah.mitchell@westbridge-ims.org*
*[2]Department of Applied Statistics and Computational Mathematics, Rhine Valley Institute of Technology, Germany*
*[3]Department of Data Analytics and Statistical Modelling, Pacific Institute of Quantitative Research, Canada*

***Corresponding Author: Dr. Hannah Mitchell***

## Abstract

*The size of a study sample, and the nature of the distribution of the population under study, play a crucial role in hypothesis testing. The accuracy and reliability of the simulation-based statistical power estimation for nonparametric hypothesis testing was assessed by Monte Carlo simulations. Normal, skewed log normal and bounded discrete distributions were simulated, including standardized effect sizes between 0.0 to 0.9 and 10, 15 and 20 pilot observations per group. Power estimations for bootstrap-based and parametric were compared with reference statistical power, using repeated Mann–Whitney U tests. Bias, standard deviation, and root mean square error (RMSE) and coverage of confidence interval were used to evaluate the estimation performance. Results showed a general bootstrap bias of 0.060, and RMSE of 0.280 and a parametric RMSE of 0.299. A comparison of the accuracy of bootstrap estimation under skewed distributional conditions showed that it performed comparatively well. The bootstrap RMSE decreased from 0.328 to 0.236 with this increase in sample size (n=20) for the pilot sample, suggesting better estimation reliability. But the empirical coverage of the confidence intervals is only 0.144 while the nominal coverage is 0.950, with the empirical coverage being well below the nominal. The study finds that distributional properties and sample size affect simulation-based power estimation. Better uncertainty estimation techniques are needed to enhance the statistical confidence and aid in rational planning of the sample size in nonparametric studies.*

**Keyword: -** *Monte Carlo Simulation; Statistical Power Estimation; Nonparametric Hypothesis Testing; Bootstrap Resampling; Mann–Whitney U Test.*

## INTRODUCTION

In hypothesis testing power is a crucial issue because it is the probability of rejecting a false null hypothesis. Appropriate sample size and statistical power enhances the validity of research findings. Nonparametric hypothesis testing is especially useful when observations are normally distributed but with some irregularities, when the data are ordinal or skewed, or when the data are not normally distributed at all. Recent advances in statistical methods have focused on flexible inferential methods for complex data such as simulation-based power estimation approaches for microbiome count data (Agronah, 2025). Even entropic based goodness of fit testing method (Cadirci, 2025) and kernel based conditional two-sample testing (Chatterjee et al., 2024) have been extended. In addition, the application of the combined nonparametric testing methods has been explored to understand the statistical power of the methods under various conditions (Fenteng, 2025). The importance of the principles of hypothesis testing for interpreting statistical decisions and their associated uncertainty is still present (Nilsson & Eckert, 2024).

When the sample sizes are small, the population parameters are unknown or the observations are not normally distributed, power calculations can be tricky. If the samples are of limited size, estimates of effect size and effect variability may be unstable, which may impact further power calculations. The simulation-based inference for finite samples has been studied to assess a statistical procedure when standard approximations are not necessarily valid (Awan & Wang, 2025). Likewise, for nonparametric tests with ordered alternatives, the determination of the sample size is important about test structure and experimental design (Bahçecitapar & Akdur, 2022). Another concern is distributional misspecification, which is the possibility that the assumptions that are built into statistical models can affect estimated power. The power of the estimators in the simulation study with correctly and mis specified distributions highlights the need to analyse the performance of the estimators under various conditions (Irmer et al., 2024).

Monte Carlo simulation offers a systematic approach to power evaluation for statistical tests and can be used to repeatedly draw observations from a given distribution and assess the likelihood of rejecting a null hypothesis for a specific statistical test. Beyond this, simulation has found extensive use in different research areas such as laboratory-based training evaluation and demonstrating its usefulness for investigating outcomes under controlled conditions (Altinbas et al., 2025). In the field of statistical methods, there are specific Monte Carlo power analysis programs that can be used to compare various hypothesis testing methods (Chalmers, 2025). To ensure reliable simulation-based inferences, suitable calibration and accounting of estimation uncertainty (Delaunoy et al., 2022) is also required. Another example of this is an application of zero-inflated count models where Monte Carlo procedures can be used to accommodate complex probability distributions for estimating statistical power (Dennis et al., 2025). Sample size calculations for recurrent-event models with competing risks have also been developed using simulation methods (Furberg et al., 2024). Further methodological aspects of designing statistical simulation studies highlight systematic experimental conditions, repeatability and transparent assessment of performance (Kelter, 2024).

However, doubts are still present on the validity of the power estimates provided by the limited sizes of pilot samples, especially for the nonparametric hypothesis testing under various distributional conditions. Although useful for estimating statistical power with less than full specification of a population distribution, the accuracy of bootstrap resampling may be sample size dependent, require large effects, and be problematic with the nature of observed values. Recently, some frameworks for power analysis based on simulation have highlighted the need to consider testing objectives and statistical decision criteria for alternative inferential procedures (Riesthuis et al., 2025). However, more investigation of bias in the estimates, variability of estimates, root mean square error and coverage of confidence intervals is warranted to better understand the limitations of small sample size bootstrap power estimation. In the current study, the current performance of simulation-based statistical power estimation for Mann – Whitney U test under normal, skewed and bounded distributional conditions is studied by Monte Carlo simulation.

### Research Objectives

- To evaluate the accuracy and variability of bootstrap-based statistical power estimates for nonparametric hypothesis testing using Monte Carlo simulation.
- To examine the influence of sample size, effect size, and distributional characteristics on statistical power estimation performance.
- To assess estimation bias, root mean square error, and confidence-interval coverage across different simulation conditions.

## 2. Materials and Methods

### 2.1. Study Design and Data Source

A simulation-based research approach was used in the present study to assess the statistical power estimation for nonparametric hypothesis testing. The publicly available PowerSim Reproducibility Package (da Silva et al., 2026) was used for the secondary simulations. The study tested power estimation based on bootstrapping for various sample sizes, effect sizes and probability distributions. Estimated power was compared to reference values for the Mann–Whitney U test, using Monte Carlo simulation. The accuracy of the estimates was evaluated by bias, standard deviation, root mean square error (RMSE) and the coverage of the confidence interval.

### 2.2. Monte Carlo Simulation Framework

Independent observations were simulated from normal, skewed log-normal and bounded discrete distributions and Monte Carlo simulations performed. Standardized effect sizes of 0.0, 0.3, 0.6 and 0.9 were explored and pilot sample

sizes consisted of 10, 15 and 20 observations per group. The desired number of observations per group was 20. There were 36 experimental conditions and 400 independent pilot samples and 400 bootstrap repetitions for each pilot sample in the main simulation. Target sample sizes between 8 and 40 were used to explore changes in statistical power in the supplementary analysis.

### 2.3. Simulation Parameters and Distributional Conditions

The influence of data characteristics on the estimation of statistical power was assessed for three distributional conditions. Symmetrically distributed continuous observations were represented by normal distributions, and positively skewed observations by log-normal distributions. To account for restricted measurements and ceiling effects, bounded discrete distributions were created on an integer scale (0-30). These conditions allowed for comparisons of the accuracy of the estimations across various distributional characteristics, effect sizes and sample sizes.

### 2.4. Nonparametric Hypothesis Testing

Mann–Whitney U test was used for comparing two independent groups of observations which are not necessarily normally distributed. The null hypothesis was that there was no difference in distribution between groups and the alternative hypothesis was that there was a difference. Throughout the simulations a two-sided significance level of .05 was used. The statistical power was defined as the percentage of the comparisons that were statistically significant for the simulated comparisons. This allowed the assessment of the nonparametric hypothesis testing performance to be consistent across the experimental conditions.

### 2.5. Bootstrap-Based Statistical Power Estimation

The number of observations was repeatedly sampled with replacement from independently generated pilot groups to obtain bootstrap power estimates. 400 bootstrap repetitions were run for each pilot sample, and the resampled groups were formulated based on the target sample size. It was used the Mann–Whitney U test for each comparison and the bootstrap power was estimated as the percentage of significant test results. This process was repeated over 400 independent pilot samples for each simulation condition to examine the stability and variability of the estimated levels of statistical power.

### 2.6. Reference Power and Parametric Comparison

The statistical power was estimated from 40,000 independent simulated comparisons for each population condition. The proportion of significant Mann–Whitney U test results were considered the reference power. A conventional two-sample t test power calculation was also conducted based on the standard mean difference calculated from the pilot observations. The accuracy and reliability of the power estimates obtained using bootstrap and parametric methods was compared to the reference power estimates.

### 2.7. Evaluation of Statistical Estimation Performance

Bias, standard deviation, RMSE, and confidence-interval coverage were used to evaluate the performance of statistical estimation. The differences between average estimated power and reference power were considered as bias and the differences among the pilot samples were considered as standard deviation. Overall estimation error was quantified using RMSE which accounted for both bias and variability. The distribution of estimated values of power was described by the 10th and 90th percentiles. Confidence intervals were constructed with a normal approximation to the binomial standard error, and their empirical coverage was determined as the proportion of intervals that covered the reference power.

### 2.8. Statistical Analysis

Means, medians, standard deviations, percentiles, and proportions were used to summarize the outcomes of the simulations. Bias, RMSE and coverage of the confidence intervals were compared across distributions, effect sizes, and pilot sample sizes. Parametric power calculations were also compared to the estimates obtained by bootstrap. Statistically power was evaluated in the supplementary analyses with respect to the increasing target sample size. Results were shown in comparative tables and graphical representations. No further inferential tests were used to compare between simulation conditions.

### 2.9. Reproducibility and Ethical Considerations

The study was based on computational simulation results that were publicly available and were simulated following a specific statistical procedure, with a fixed random seed. The reproducibility package contained simulation results and methodology specifications, to facilitate verification. Because no humans, patient information or personally identifiable records were included in the research, neither institutional ethical approval nor informed consent was required. The study was carried out with respect to the principles of reproducibility and transparency in the reporting of statistical information and in the due reference to the original source.

## 3. Results

A total of 36 simulation conditions were analysed, which involved three probability distributions, four standardized effect size and three pilot sample size. The accuracy of bootstrap method for statistical power estimation was measured

against the reference power from simulation for Mann–Whitney U test. Estimation bias, standard deviation, root mean square error (RMSE), CI coverage, and additional comparisons using target-sample size were examined.

### 3.1. Distributional Conditions and Reference Statistical Power

The statistical power for each of the three distributional conditions varied and was related to the effect sizes. For the normal distribution, the mean reference power was 0.350, for the skewed distribution, the mean reference power was 0.538 and for the bounded distribution, the mean reference power was 0.345 (see Table 1). The sloped distribution resulted in the greatest reference power for the nonzero effect sizes. With an effect size of 0.9, the resulting reference power was approximately 0.968 in the case of the skewed observations, 0.770 in the case of the normal observations, and 0.760 in the case of the bounded observations. These differences suggest that the rejection probability for Mann–Whitney U test was affected by the distributional structure.

**Table 1. Reference Statistical Power across Distributional Conditions and Effect Sizes**

| Effect size | Normal | Skewed | Bounded/ceiling |
|---|---|---|---|
| 0.0 | 0.049 | 0.050 | 0.050 |
| 0.3 | 0.147 | 0.335 | 0.144 |
| 0.6 | 0.434 | 0.799 | 0.429 |
| 0.9 | 0.770 | 0.968 | 0.760 |
| Mean | 0.350 | 0.538 | 0.345 |

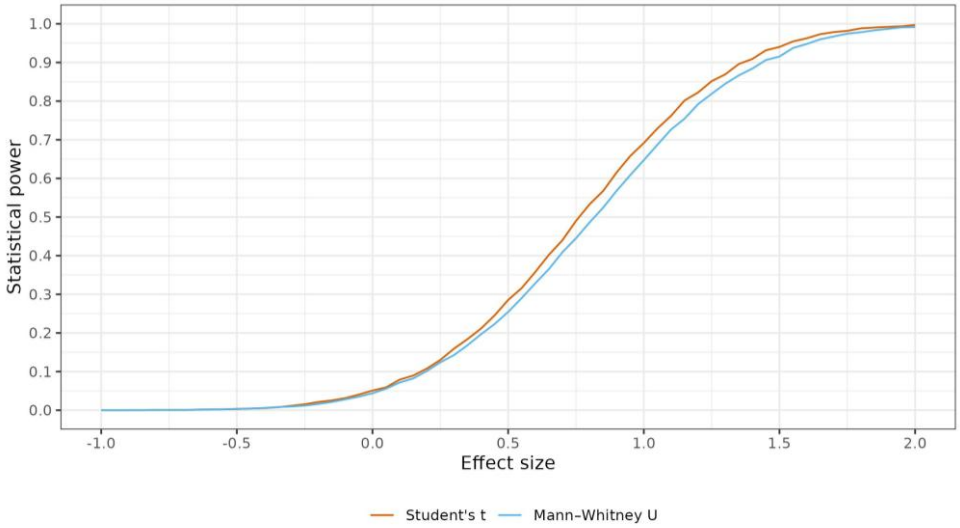



**Figure 1. Reference Statistical Power across Effect Sizes**

Figure 1 illustrates the increase in reference power with effect size across the three distributions. The largest separation between distributional conditions occurred at moderate effect sizes, particularly at 0.6, where the skewed distribution demonstrated substantially greater statistical power.

### 3.2. Comparative Performance of Bootstrap and Parametric Power Estimation

Comparative analysis revealed the differences of the accuracy of the estimation of bootstrap-based power procedures and parametric power procedures. The overall mean bootstrap bias was 0.060 whereas a parametric estimator bias of 0.026 was obtained as shown in Table 2. The mean of the bootstrap RMSE was 0.280 and the mean of the parametric RMSE was 0.299. Both methods exhibited positive average bias and similar values of RMSE under normal and bounded conditions. The skewed distribution, on the other hand, resulted in negative average parametric bias of −0.074, and positive average bootstrap estimator bias of 0.029 with a smaller RMSE.

**Table 2. Comparative Performance of Bootstrap and Parametric Power Estimation by Distribution**

| Distribution | Bootstrap bias | Bootstrap RMSE | Parametric bias | Parametric RMSE |
|---|---|---|---|---|
| Normal | 0.071 | 0.282 | 0.072 | 0.280 |
| Skewed | 0.029 | 0.266 | −0.074 | 0.328 |
| Bounded/ceiling | 0.080 | 0.291 | 0.081 | 0.290 |
| Overall mean | 0.060 | 0.280 | 0.026 | 0.299 |

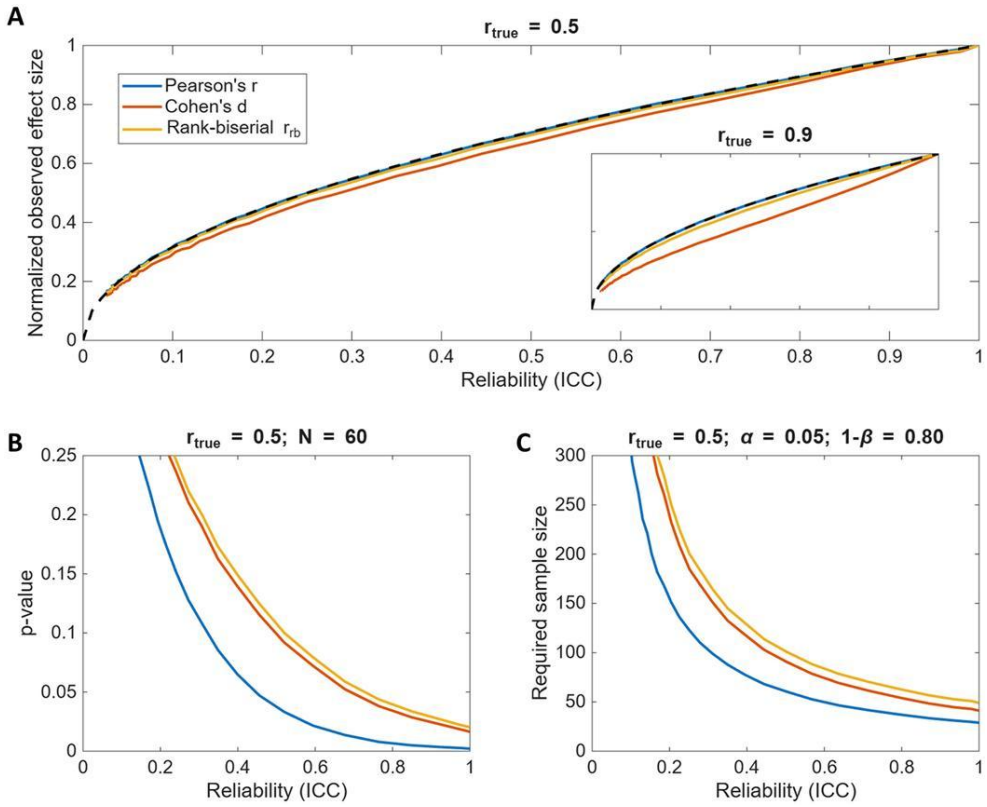



**Figure 2. Estimation Bias across Distributional Conditions**

Figure 2 compares estimation bias across effect sizes for the two methods. Both estimators exhibited positive bias under several low-effect conditions, while changes in bias direction were observed at larger effect sizes. The distribution-dependent differences were most apparent for skewed observations.

### 3.3. Influence of Pilot Sample Size on Estimation Accuracy

The variability and accuracy of the power estimates based on the bootstrap method were found to depend on the pilot sample size. Bootstrap bias, standard deviation and RMSE were reduced by increasing the size of the pilot sample from 10 to 20 observations per group, as shown in Table 3. Additionally, parametric RMSE was found to be decreasing as the size of the pilot sample increased. These results suggest that the larger pilot samples tended to yield more consistent estimates of statistical power for the conditions simulated

**Table 3. Power Estimation Performance according to Pilot Sample Size**

| Pilot sample size per group | Bootstrap bias | Bootstrap SD | Bootstrap RMSE | Parametric RMSE |
|---|---|---|---|---|
| 10 | 0.089 | 0.286 | 0.328 | 0.341 |
| 15 | 0.052 | 0.248 | 0.274 | 0.297 |
| 20 | 0.039 | 0.218 | 0.236 | 0.259 |

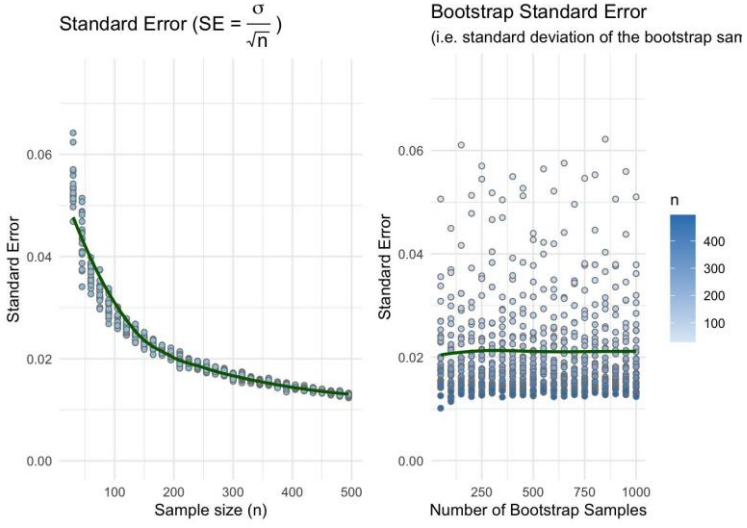



**Figure 3. Influence of Pilot Sample Size on RMSE**

Figure 3 demonstrates the progressive reduction in estimation error as pilot sample size increased. The bootstrap RMSE decreased by approximately 28.0% between pilot sample sizes of 10 and 20 per group, supporting an improvement in estimation accuracy over the examined range.

### 3.4. Confidence-Interval Coverage and Statistical Reliability

The empirical coverage of the confidence intervals was much lower than the nominal 95% in all simulation scenarios. Mean coverage rose from 0.114 to 0.177 as the number of pilot observations per group was increased from 10 to 20, as shown in Table 4. There was a comparable low level of coverage across distributional conditions ranging from 0.141 for bounded observations to 0.146 for normal observations. The overall mean coverage was 0.144 meaning that the intervals were often not able to capture the reference power obtained from the simulation.

**Table 4. Empirical Confidence-Interval Coverage across Simulation Conditions**

| Simulation condition | Mean coverage | Nominal coverage |
|---|---|---|
| Pilot size: 10 | 0.114 | 0.950 |
| Pilot size: 15 | 0.140 | 0.950 |
| Pilot size: 20 | 0.177 | 0.950 |
| Normal distribution | 0.146 | 0.950 |
| Skewed distribution | 0.144 | 0.950 |
| Bounded distribution | 0.141 | 0.950 |
| Overall mean | 0.144 | 0.950 |

Coverage represents the proportion of nominal 95% Monte Carlo intervals containing the reference statistical power. Pilot-size and distribution summaries represent different groupings of the same 36 conditions and should not be combined as independent observations.

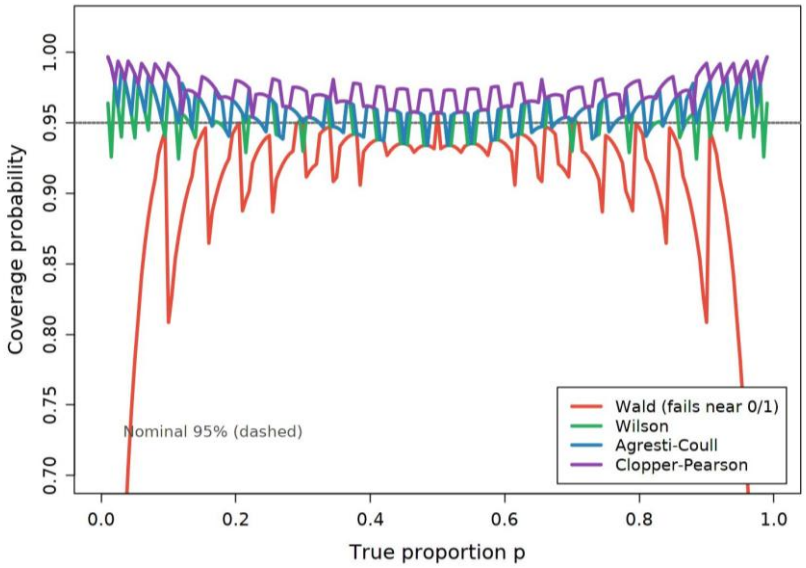



**Figure 4. Empirical Confidence-Interval Coverage**

Figure 4 illustrates the consistently low empirical coverage across standardized effect sizes. The results demonstrate substantial under coverage of the reported Monte Carlo intervals, suggesting that the intervals did not adequately capture the total uncertainty associated with estimating statistical power from small pilot samples.

### 3.5. Influence of Target Sample Size on Statistical Power

Additional simulations were conducted to determine the variation in statistical power for different target sample sizes (8 to 40 observations per group) with a standardised effect size of 0.6. As seen in Table 5, the reference power was growing as the target sample size was increasing under both the assumptions of a normal and a bounded distribution. In the normal distribution, the percentage of reference power varied from 0.191 with eight observations per group to 0.732 with 40 observations per group. Bootstrap estimates rose from 0.276 to 0.671 corresponding to this. The same trends were seen for bounded observations, but in the case of the smallest target sample sizes, the bootstrap estimates were larger than the corresponding reference power, while at the largest target sample size, the bootstrap estimates were smaller than the reference power.

**Table 5. Reference and Bootstrap-Estimated Statistical Power across Target Sample Sizes**

| Target sample size per group | Normal reference | Normal bootstrap | Bounded reference | Bounded bootstrap |
|---|---|---|---|---|
| 8 | 0.191 | 0.276 | 0.172 | 0.287 |

| | | | | |
|---|---|---|---|---|
| 12 | 0.267 | 0.355 | 0.265 | 0.357 |
| 16 | 0.349 | 0.438 | 0.349 | 0.442 |
| 20 | 0.436 | 0.479 | 0.428 | 0.474 |
| 24 | 0.499 | 0.515 | 0.498 | 0.517 |
| 30 | 0.597 | 0.598 | 0.597 | 0.600 |
| 40 | 0.732 | 0.671 | 0.730 | 0.666 |

Results from the simulation showed that there were significant differences in the power estimation of statistics among the distributions, effect sizes, and pilot sample sizes. The overall mean bias and RMSE were respectively 0.060 and 0.280 using the bootstrap based estimation, and 0.299 using the parametric estimation. For larger pilot samples, the estimation variability and RMSE decreased with the increase in sample size, and the coverage of the confidence intervals was significantly lower than the nominal 95%. Increasing reference power with target sample size was confirmed by supplementary analyses and discrepancies between estimated and reference values were also identified. The results of these two together demonstrate the need to consider distributional properties and pilot-sample uncertainty in power interpretations for the Mann–Whitney U test based on simulation.

**4. Discussion**

The reliability of the simulation-based statistical power estimation for nonparametric hypothesis testing was evaluated across varying distributional conditions, effect sizes and pilot sample sizes in the present study. Results showed significant differences between estimation accuracy, with an overall bootstrap bias of 0.060 and root mean square error (RMSE) of 0.280. The bootstrap estimation technique resulted in a lower overall RMSE, especially when conditions were skewed. As the number of pilots increased, the stability of the estimation increased, but the empirical coverage of the confidence intervals was significantly less than the nominal 95%. The results show the need to consider sampling error when calculating the power of a Mann–Whitney U test.

Moderate estimation errors were found by bootstrap-based power estimation, which was dependent on the distributional characteristics. The relative smaller bootstrap RMSE in the skewed cases indicate that resampling procedures could be beneficial as an alternative to standard parametric assumptions. New methods of statistical hypothesis testing have highlighted the need for choosing appropriate methods for underlying probability structures (Elgehady et al., 2025). Likewise, simulation studies of double sampling have shown the importance of investigating nonparametric estimation methods under different experimental settings (Hidayat et al., 2025). Nonparametric comparisons based on simulation have also been used in complex biological measurements, demonstrating a wider range of applications of computational inference (Verdier et al., 2023). Nevertheless, the estimation errors observed suggest that bootstrap power estimation is not necessarily accurate even though it does not make as many distributional assumptions.

The statistical power and the estimation performance were highly affected by the distributional characteristics. There was a positive correlation between effect size and reference power for all distributions, and the power was higher for skewed observations than normal, and for bounded observations than for observations with nonzero effects. Studies have highlighted the need for choosing statistical tools suitable for the measurement scale and distribution (Chiang et al., 2024) when using an ordinal disease severity scale. In addition, kernel-based statistical testing has been explored in the context of complex genomic observations, highlighting the importance of flexible statistical methods (Kamoljitprapa et al., 2024). Moreover, robust nonparametric inference has been generalized to other mathematical spaces such as manifold spaces (Mozaffari, 2025). The developments are useful in thinking about the distributional characteristics of comparisons made based on simulations but are not directly comparable to the findings of the present study.

The accuracy and stability of the statistical power estimates was affected by the size of the pilot sample. Bootstrap RMSE decreased from 0.328 to 0.236 as more pilot observations (20) were taken per group, providing a better estimation reliability. Monte Carlo simulation has been used along with powerful nonparametric methods to investigate the statistical performance under various market conditions (Ngurah, 2025). Likewise, statistical trend detection under different experimental conditions, indicates the importance of variation when analyzing repeated measurements (Petty, 2025). Based on the current results, it is suspected that small pilot samples can lead to an unstable power estimate, even with the use of a bootstrap resampling approach.

One of the main results was that reported confidence intervals have low empirical coverage with an average of only 0.144 as opposed to the nominal value of 0.950. This significant undercoverage suggests that the intervals were not good representations of the uncertainty in the estimated power. In nonparametric calibration, the reliable assessment of uncertainty in statistical inference has been highlighted (Sadatsafavi & Petkau, 2024). Repeated experiments have also been shown to be helpful when exploring methodological performance through simulation-based evaluation of statistical tests (Suchar & Nardin, 2022). The undercoverage observed above may be due to the way uncertainty is represented in the interval construction procedure and not indicative of a fundamental limitation of the bootstrap resampling approach.

The values found are of practical importance for the statistical planning, especially for the choice of a nonparametric test in small-sample research. They have been used in aviation optimization (Attar et al., 2024), financial risk assessment (da Cruz, 2025), training-related performance investigations (Martin Jr., 2023) and infrastructure sustainability assessment (Pulecio-Díaz, 2025). These applications highlight the versatility of the simulation-based approaches across fields but do not directly validate current power estimation approaches. The study only includes three distributional conditions, selected effect sizes, and the Mann– Whitney U test.

Moreover, its outcomes were based on an existing simulation package, not on direct observations. Further research should be conducted using other nonparametric methods, other bootstrap interval methods, and sample-size conditions to increase the generalizability of statistical power estimation.

## 5. Conclusion

In the current study, the accuracy and reliability of simulation-based statistical power estimation for nonparametric hypothesis testing was assessed by Monte Carlo simulations. Results showed that distributional properties, effect sizes, and pilot sample sizes affected the bootstrap-based power estimations. Overall, lower overall root mean square error (RMSE) was obtained from bootstrap than parametric estimation, especially under skewed distributional conditions, suggesting its utility when conventional parametric assumptions are not applicable. But there was significant estimation bias and variability across simulation conditions, reflecting the limitations of small sample size power estimates. Larger pilot sample size led to reduced bias, standard deviation, and root mean square error in estimating the population size. Additionally, statistical power also rose with the effect size and the desired sample size, but discrepancies between the estimated and the reference statistical power were observed. Importantly, the empirical coverage was much lower than the nominal 95%, indicating that the interval procedure evaluated was not capturing the uncertainty associated with estimation. These results stress the need to realize that the bootstrap re-sample is not necessarily reliable for statistical power estimation, especially if a pilot sample is small. Overall, the study highlights the importance of considering distributional assumptions, sample-size limitations, and uncertainty calibration when applying simulation-based power estimation to the Mann–Whitney U test. Further research is needed to examine other nonparametric tests, other resampling methods, and other improved confidence-interval procedures to enhance estimation accuracy, replication, and statistical decision making for other research applications.

## References


1. Agronah, M. (2025). A Novel Approach for Simulation-Based Power Estimation and Joint Modeling of Microbiome Counts (Doctoral dissertation).
2. Altinbas, B. C., Çalık, K. Y., Erdöl, E. K., Kırkbir, İ. B., Güner, S. G., Tezel, M., ... & Bulut, H. K. (2025). The effect of simulation-based laboratory training on undergraduate nursing students' clinical skill, satisfaction, and self-confidence. BMC nursing, 24(1), 1322.
3. Attar, A., Babaee, M., Raissi, S., & Nojavan, M. (2024). Airside Optimization Framework Covering Multiple Operations in Civil Airport Systems with a Variety of Aircraft: A Simulation-Based Digital Twin. Systems, 12(10), 394.
4. Awan, J., & Wang, Z. (2025). Simulation-based, finite-sample inference for privatized data. Journal of the American Statistical Association, 120(551), 1669-1682.
5. Bahçecitapar, M., & Akdur, H. T. K. (2022). Sample Size Estimation of Nonparametric Tests with Ordered Alternatives for Longitudinal Data in Randomized Complete Block Designs. Gazi University Journal of Science Part A: Engineering and Innovation, 9(4), 367-377.
6. Cadirci, M. S. (2025). Non-parametric goodness-of-fit tests using Tsallis entropy measures. Entropy, 27(12), 1210.
7. Chalmers, R. P. (2025). Spower: A general-purpose Monte Carlo simulation power analysis program. Behavior Research Methods, 57(12), 348.
8. Chatterjee, A., Niu, Z., & Bhattacharya, B. B. (2024). A kernel-based conditional two-sample test using nearest neighbors (with applications to calibration, regression curves, and simulation-based inference). arXiv preprint arXiv:2407.16550.
9. Chiang, K. S., Chang, Y. M., Liu, H. I., Lee, J. Y., Jarroudi, M. E., & Bock, C. H. (2024). Survival analysis as a basis for testing hypotheses when using quantitative ordinal scale disease severity data. Phytopathology®, 114(2), 378-392.
10. da Cruz, P. A. S. (2025). Backtesting Expected Shortfall: Historical Simulation Based Analysis of a US Diversified Portfolio.
11. da Silva, E. A., Rocha, D. S., Naves, E. L. M., & Rodrigues, T. B. (2026). Resampling-based power simulation in small samples: a Monte Carlo calibration study (reproducibility package) (Version v1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.20767954.
12. Delaunoy, A., Hermans, J., Rozet, F., Wehenkel, A., & Louppe, G. (2022). Towards reliable simulation-based inference with balanced neural ratio estimation. Advances in Neural Information Processing Systems, 35, 20025-20037.
13. Dennis, T., Rubin-McGregor, J., & O'Connell, M. J. (2025). Power analysis for zero-inflated Poisson and negative binomial generalized linear models using Monte Carlo simulation. Journal of Statistical Computation and Simulation, 95(4), 868-885.
14. Elgehady, H. S., El-Arishy, S. M., & El-Atfy, E. S. (2025). A Novel Hypothesis Testing for EBUC (mgf) Utilizing Laplace Transform Techniques with Practical Applications. Statistics, Optimization & Information Computing, 14(3), 1419-1439.
15. Fenteng, K. D. (2025). Power Analysis of Combined Nonparametric Tests: Fligner–Wolfe and Wald–Wolfowitz Approaches (Master's thesis, North Dakota State University).

16. Furberg, J. F., Andersen, P. K., Scheike, T., & Ravn, H. (2024). Simulation-based sample size calculations of marginal proportional means models for recurrent events with competing risks. Pharmaceutical statistics, 23(5), 687-708.
17. Hidayat, K., Fernandes, A. A. R., Iriany, A., Solimun, S., Hidayatulloh, M. Z., & Junianto, F. H. (2025). Nonparametric Path Modeling with Double Resampling for Waste Economic Value Utilization: Simulation-Based Performance Comparison. CAUCHY: Jurnal Matematika Murni dan Aplikasi, 10(2), 1233-1250.
18. Irmer, J. P., Klein, A. G., & Schermelleh-Engel, K. (2024). Model-implied simulation-based power estimation for correctly specified and distributionally misspecified models: applications to nonlinear and linear structural equation models. Behavior research methods, 56(8), 8955-8991.
19. Kamoljitprapa, P., Sookkhee, S., & Polsen, O. (2024, August). Statistical analysis for genome data based on multiple SNPs using kernel machine based test. In 2024 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics (RI2C) (pp. 262-266). IEEE.
20. Kelter, R. (2024). The Bayesian simulation study (BASIS) framework for simulation studies in statistical and methodological research. Biometrical Journal, 66(1), 2200095.
21. Martin Jr, F. (2023). An Experimental Mixed Methods Pilot Study for US Army Infantry Soldiers-Higher Levels of Combined Immersion and Embodiment in Simulation-Based Training Capabilities Show Positive Effects on Emotional Impact and Relationships to Learning Outcomes.
22. Mozaffari, A. (2025). Robust nonparametric inference on manifold spaces.
23. Ngurah, S. P. I. G. (2025). Evaluation of Fast Food Promotion Effectiveness Using a Hybrid Approach of Robust Non-Parametric Methods and Monte Carlo Simulation Based on Market Size Segmentation. Jurnal Akuntansi, Ekonomi dan Manajemen Bisnis, 13(1), 63-68.
24. Nilsson, P., & Eckert, A. (2024). Design principles for simulation-based learning of hypothesis testing in secondary school. Mathematical thinking and learning, 26(4), 359-381.
25. Petty, M. (2025). Trend Detection in Simulation-Based Training Performance Scores Obtained Under Varying Experimental Conditions. SAGE Publications Ltd. Petty, M. (2025). Trend Detection in Simulation-Based Training Performance Scores Obtained Under Varying Experimental Conditions. SAGE Publications Ltd.
26. Pulecio-Díaz, J. (2025). Curing sustainability assessment in concrete pavements: A 20-Year simulation-based analysis in urban road contexts. Sustainability, 17(12), 5299.
27. Riesthuis, P., Otgaar, H., & Bücken, C. (2025). Ready to ROC? A tutorial on simulation-based power analyses for null hypothesis significance, minimum-effect, and equivalence testing for ROC curve analyses. Behavior Research Methods, 57(4), 120.
28. Sadatsafavi, M., & Petkau, J. (2024). Non-parametric inference on calibration of predicted risks. Statistics in Medicine, 43(18), 3524-3538.
29. Suchar, V. A., & Nardin, L. G. (2022). Simulation-Based Assessment of the Effectiveness of Tests for Stationarity. Statistics, 10(3), 670-683.
30. Verdier, H., Laurent, F., Cassé, A., Vestergaard, C. L., Specht, C. G., & Masson, J. B. (2023). Simulation-based inference for non-parametric statistical comparison of biomolecule dynamics. PLOS Computational Biology, 19(2), e1010088.