International Journal of Mathematics and Statistics
ISSN (Online): 2208-2212
Volume 11 Issue 03 September 2025
DOI: 10.69980/84g63842

STATISTICAL ASSESSMENT OF SAMPLING VARIABILITY AND CONFIDENCE INTERVAL COVERAGE USING BOOTSTRAP METHODS

Dr. Emily Carter[1], Dr. Lukas Schneider[2*], Dr. Sofia Laurent[3]

[1]Department of Statistics and Probability, Northbridge Institute of Mathematical Sciences, United Kingdom
[2]Department of Applied Mathematics, Rhine Valley Institute of Technology, Germany*
[3]Department of Statistical Modelling and Data Science, Westlake Research Institute, Canada

Corresponding Author: Dr. Lukas Schneider

Abstract

Bootstrap resampling is widely applied to statistical inference to estimate the sampling variability and derive confidence intervals. In small-scale studies, however, its reliability may be affected by the sample size, probability distribution and estimation uncertainty. The purpose of this study was to assess the sampling variability, estimation accuracy and empirical coverage of the confidence intervals from sample-based statistical powers in the different simulation conditions. The study is of quantitative, simulation type with secondary data that is publicly available from Zenodo. Thirty-six conditions—three probability distributions (normal, skewed, and ceiling/discrete), four effect sizes, and three different pilot sample sizes (10, 15, and 20 per group) were analysed. There were 400 stand-alone pilot simulations with 400 bootstrap replications per pilot simulation per condition. The statistical performance was evaluated by standard deviation, estimation bias, root mean squared error (RMSE) and empirical coverage of confidence intervals. Results indicated that sampling variability (standard deviation between pilots) decreased as the number of samples collected in a pilot increased (0.2856 to 0.2179). The RMSE of Bootstrap also decreased from 0.3281 to 0.2363, indicating better estimation accuracy. However, its empirical mean coverage was as low as 14.37% as opposed to the nominal level of 95%. These results highlight the need to consider the various sources of uncertainty and improve the calibration of confidence intervals to obtain more reliable statistical inference using the bootstrap in small-sample studies.

Keyword: - *Bootstrap Resampling; Sampling Variability; Confidence Interval Coverage; Monte Carlo Simulation; Statistical Inference.*

INTRODUCTION

Making inferences about the population from observations of samples is to obtain information about the properties of the population and make meaningful inferences about the population based on observations of samples. However, samples might differ from each other (because of sampling error) in the estimates they provide, if the samples are small. The variability in samples has a direct impact on the precision, stability and reproducibility of statistical results. Williams (2022) noted that statistical relationships and expected replicability should take into consideration sampling fluctuations. The confidence interval gives an idea of how uncertain estimation is, and they are used by researchers to check the accuracy of statistics (Jané et al., 2024). Furthermore, complex sampling designs pose challenges in creating reliable confidence intervals due to potential inadequacies in conventional estimation methods to capture uncertainties (Lin et al., 2024). Therefore, it is important to be aware of sampling variability and its performance of confidence intervals for the sake of better statistical inference.

Today, bootstrap resampling is a valuable computational method for calculating sampling distributions without making stringent assumptions about the shape of the underlying population distribution. This technique consists of resampling data again using the same sampling method but with replacement to simulate the sampling variation of the statistical estimates. Rousselet et al., (2023) showed how useful bootstrap methods are to perform data-driven statistical inference. Likewise, Mokhtar et al. (2023) talked about Bootstrap Confidence Intervals and its relevance in Statistical Applications. Bootstrap procedures (Aldawsari, 2023) can also be used for predictive inference and for assessing the statistical reproducibility. Additionally, Maitner et al. (2023) demonstrated the usefulness of bootstrapping in estimating the characteristics of a distribution that are applicable in ecology. The examples illustrate how one can use bootstrap techniques in different statistical situations.

The proportion of intervals that contains the true population parameter in repeated sampling experiments will be called confidence interval coverage probability. The coverage of nominal 95% confidence intervals should be approximately 95% when the procedure is repeated many times. Actual coverage, however, may vary due to small sample sizes, distributional asymmetry, estimation bias and limitations of interval construction methods. Mokhtar et al. (2024) studied confidence interval estimation for normal and skewed distribution focusing on the importance of distributional characteristics. Differences in the performance of bootstrap confidence intervals have also been studied in comparative simulation studies (Justus et al., 2024). In addition, the Bayesian bootstrap methods have been studied for estimating the mean under small-sample sizes (Tongmual & Yonwilad, 2025). These studies suggest that prior to assuming any statistical reliability of nominal confidence levels, it is important to assess the empirical coverage.

Even with the methodological advances, uncertainty estimation is difficult in case the sampling conditions are complex or if statistical procedures are based on repeated computational estimation. Mayen et al. (2024) also raised methodological issues about combining complex survey data, highlighting the need for statistical methods. Kluger et al. (2025) provided an experiment on inference in the context of non-uniform sampling and imputing covariates, as other examples of uncertainty assessment. It is also worth considering computational reproducibility since the results of statistical inference based on machine learning can depend on the random selection of the seed value (Schader et al., 2024). Furthermore, out-of-sample predictive performance estimation and inference was considered by Hawinkel et al. (2024), which further underscores the importance of proper estimation of the uncertainty. The present study tries to answer these methodological questions and concentrates on the statistical power estimates resulting from the bootstrapping method whose results were publicly available by Monte Carlo simulation. It is used to compare the sampling variability, estimation bias, root mean squared error and coverage of the empirical confidence intervals for different sample sizes, effect sizes and probability distributions. Independently generated pilot samples are considered, and the variability between is considered as to whether it is well covered by the confidence intervals. The study looks at the coverage compared to the nominal 95% confidence level to offer evidence of the reliability of bootstrap-based uncertainty estimation under small sample size.

Research Objectives

- To assess sampling variability and estimation accuracy across different sample sizes, effect sizes, and probability distributions using bootstrap-based simulation results.
- To evaluate empirical confidence interval coverage and determine deviations from the nominal 95% confidence level under different simulation conditions.
- To examine the influence of sample size and distributional characteristics on bootstrap estimation bias, standard error, and overall statistical reliability.

2. Materials and Methods

2.1. Study Design

The research design applied in this study was quantitative and utilized simulations with secondary data to analyse sampling variability and confidence interval coverage in terms of bootstrapping. Effect size, sample size and probability distribution were some of the variables that have been analysed with respect to their influence on estimation accuracy and reliability through the Monte Carlo simulation technique. Statistical power, sampling variability, estimation bias and confidence interval coverage have been used as measures of analysis in the simulations.

2.2. Data Source and Description

The secondary data was used in this study, and it was sourced from an open-access dataset developed by da Silva et al. (2026). The number of simulation conditions in this data was 36, while the total number of variables in the main file

sim_results.csv was 16; including the probability distribution, sample size, effect size, statistical power, statistical power estimate, standard deviation, bias, root mean squared error (RMSE), and confidence interval coverage. There is more information included in an additional file, extra_results.json. These files contained the statistical information that was used to assess the performance of the bootstrap estimation.

2.3. Simulation Conditions and Study Variables

Three probability distributions were used: normal, skewed and bounded discrete distributions with ceiling effects. There were four standardized effect sizes (0.0, 0.3, 0.6, and 0.9) and three pilot sample sizes (10, 15, and 20 per group) resulting in 36 simulation conditions. The planned target population size was 20 per group. There were 400 independent pilot simulations, with 400 bootstrap repetitions per pilot for each condition. Estimated statistical power, between-pilot standard deviation, estimation bias, RMSE and empirical coverage of the CI's were the main variables.

2.4. Bootstrap Resampling Procedure

Bootstrap resampling was carried out by repeatedly sampling from independently generated pilot samples with replacement. Groups were resampled separately and a two-sided Mann–Whitney U test was performed at a significance level of 0.05. The statistical power was calculated as the percentage of significant results for 400 bootstrap replications. Each condition was sampled with 40,000 independent Monte Carlo samples and the reference power values were obtained from those. The resulting estimates were compared with the reference values to consider the accuracy and sampling variability.

2.5. Assessment of Sampling Variability and Estimation Accuracy

The standard deviation of the power estimates obtained from independent pilot samples was used to measure the sampling variability. Their distributions were summarized by their values at the mean, median, and percentiles. Estimation bias was defined as the difference between the mean estimated power and the reference power and RMSE indicated the merge of the bias and variability. The absolute bias and RMSE were lower, thus showing higher accuracy. The effects of these measures were compared using various sample sizes, effect sizes and probability distributions.

2.6. Confidence Interval Coverage Assessment

Empirical confidence interval coverage was evaluated using the percentage of 95% confidence intervals that covered the reference statistical power. Power was estimated and normal-approximation intervals around the estimated power were formed using the Monte Carlo standard error of the estimated power. The percentages of coverage were obtained over multiple repeated pilot simulations and compared with the nominal 95% coverage percentage. The interval reliability and the impact of between-pilot sampling variability were assessed by comparing nominal and empirical coverage.

2.7. Statistical Analysis

Descriptive statistics included mean, standard deviation, median, percentile range and percentage. Between-pilot standard deviations were used to assess sampling variability, and bias and RMSE were used to assess the accuracy of the estimates. The empirical coverage of the intervals was carried out for 36 simulation conditions, according to three factors: sample size, effect size, and probability distribution. Statistically patterns were displayed in tables and figures. Comparisons were mostly descriptive due to the lack of numerical results for each of the different simulation conditions.

2.8. Ethical Considerations

Secondary simulation data was used for the study and retrieved from Zenodo. No human participants were recruited, and no personal or identifiable information collected. The numerical data were only used for academic analysis, and proper credit was given to the original source. Since simulated and non-identifiable data was used for the study, institutional ethical clearance was not needed.

3. Results

A total of 36 simulation conditions were considered, based on three types of probability distribution (normal, skewed, and ceiling/discrete), four standardized effect sizes (0.0, 0.3, 0.6, and 0.9), and three pilot sample sizes (10, 15, and 20 observations per group), from the uploaded Zenodo reproducibility archive. There were 400 independent pilot simulations for each condition, and 400 bootstrap replications per pilot. Sampling variability, estimation bias, root mean squared error (RMSE) and empirical confidence interval coverage were evaluated. The mean between-pilot standard deviation was 0.2504, mean bootstrap RMSE was 0.2796, and mean empirical coverage was 14.37% which is well below the nominal 95% confidence level for all conditions.

3.1. Characteristics of Simulation Conditions

The simulation structure contained 12 settings for each of the probability distributions, where the settings were the same in terms of pilot sample size and standardized effect size. The balanced design allowed comparisons to be made among the three distributional settings for the bootstrap performance as shown in Table 1. Four effect-size conditions were used for each combination of distribution and pilot sample size to illustrate the distribution of simulation conditions as a function of the pilot sample size (Figure 1).

Table 1. Distribution of Simulation Conditions and Statistical Parameters

Probability distribution	Simulation conditions	Pilot sizes	Effect-size levels	Pilot simulations per condition
Normal	12	3	4	400
Skewed	12	3	4	400
Ceiling/discrete	12	3	4	400
Total	36	3 unique sizes	4 unique levels	14,400 in total

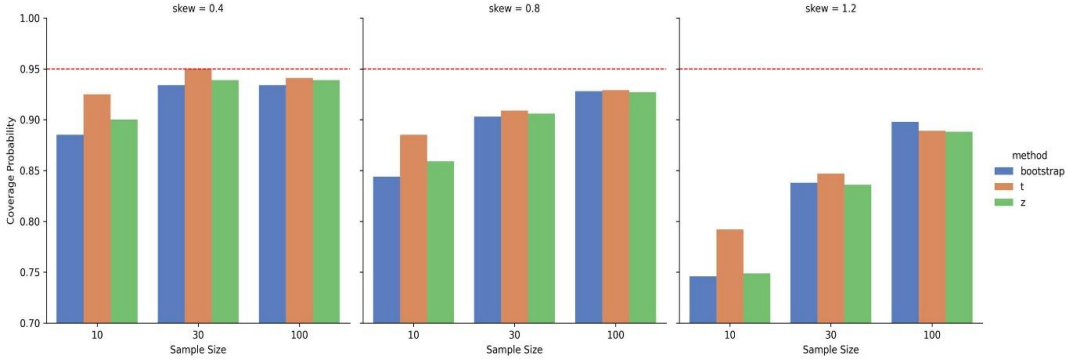


Figure 1. Distribution of simulation conditions across probability distributions and pilot sample sizes

3.2. Sampling Variability Across Simulation Conditions

As the sample size of the pilot increased, the sampling variability decreased consistently among the three probability distributions. The mean between-pilot standard deviation decreased from 0.2856 for a pilot size equal to 10 pilots to 0.2179 for a pilot size equal to 20 pilots (a 23.7% decrease) as shown in Table 2. It was the skewed distribution which was the least variable in terms of average at each sample size, with the normal and the ceiling distribution having a greater variability. Figure 2 shows the between-pilot standard deviations of the outcomes for all successive pilot sample sizes, suggesting that the standard deviations are more stable to be estimated with larger pilot sample sizes.

Table 2. Sampling Variability Across Probability Distributions and Pilot Sample Sizes

Distribution	Pilot size	Mean estimated power	Between-pilot SD	Mean P10	Mean P90
Normal	10	0.4621	0.2905	0.0943	0.8589
Normal	15	0.4123	0.2509	0.1104	0.7619
Normal	20	0.3879	0.2208	0.1254	0.6844
Skewed	10	0.5922	0.2734	0.2348	0.9269
Skewed	15	0.5509	0.2338	0.2612	0.8317
Skewed	20	0.5591	0.2061	0.3149	0.8064
Ceiling	10	0.4468	0.2928	0.0935	0.8714
Ceiling	15	0.4242	0.2587	0.1143	0.7740
Ceiling	20	0.4032	0.2268	0.1316	0.7029

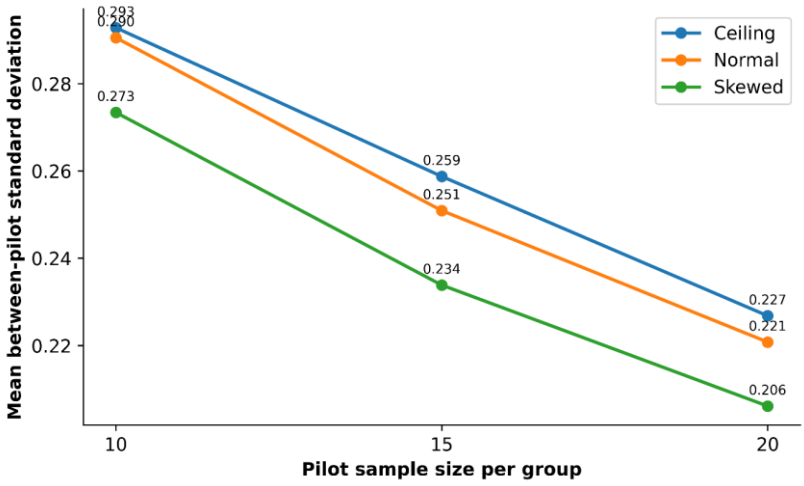


Figure 2. Changes in sampling variability across pilot sample sizes and probability distributions

3.3. Bootstrap Estimation Bias and Accuracy

The accuracy of the bootstrap estimation was dependent on effect size and the probability distribution. As reported in Table 3, all three distributions showed positive estimation bias at β= 0.0 and 0.3, and negative estimation bias at β= 0.9,

in general. The mean bootstrap RMSE for the skewed distribution was lowest at 0.2662 followed by the normal at 0.2820 and the ceiling at 0.2907. The average bootstrap RMSE was lower compared to the parametric estimator under skewed conditions and smaller differences were obtained for normal and ceiling distributions. Figure 3 shows the difference in means and RMSEs of distributions.

Table 3. Bootstrap Estimation Bias and Root Mean Squared Error Across Probability Distributions

Distribution	Effect size	Reference power	Bootstrap power	Bias	Bootstrap RMSE	Parametric RMSE
Normal	0.0	0.0488	0.2104	0.1615	0.2469	0.2375
Normal	0.3	0.1451	0.2873	0.1423	0.2859	0.2871
Normal	0.6	0.4350	0.4809	0.0459	0.3122	0.3164
Normal	0.9	0.7706	0.7044	−0.0662	0.2829	0.2786
Skewed	0.0	0.0489	0.2334	0.1845	0.2789	0.2682
Skewed	0.3	0.3356	0.4282	0.0925	0.3093	0.2721
Skewed	0.6	0.8001	0.7171	−0.0830	0.2871	0.4118
Skewed	0.9	0.9670	0.8910	−0.0761	0.1895	0.3599
Ceiling	0.0	0.0480	0.2160	0.1680	0.2567	0.2540
Ceiling	0.3	0.1426	0.3055	0.1629	0.3080	0.3081
Ceiling	0.6	0.4272	0.4867	0.0595	0.3121	0.3154
Ceiling	0.9	0.7607	0.6907	−0.0700	0.2858	0.2805

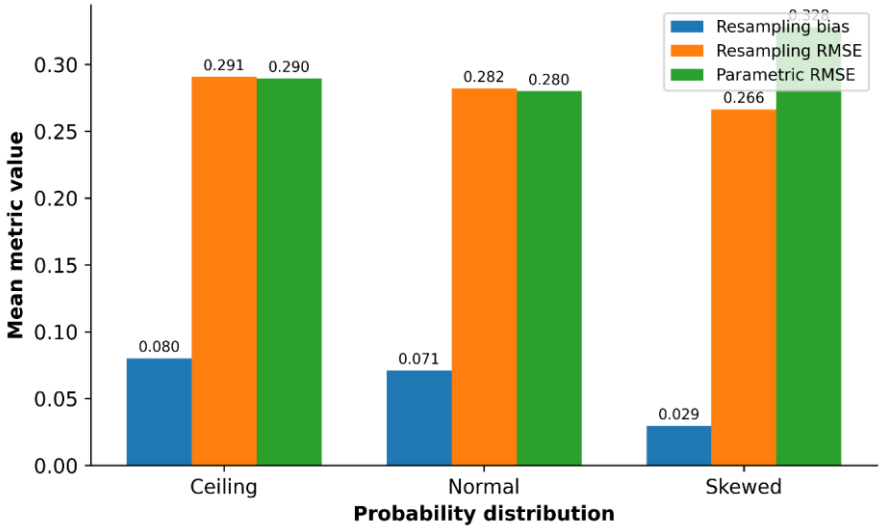


Figure 3. Distribution-wise comparison of bootstrap estimation bias and root mean squared error

3.4. Empirical Confidence Interval Coverage

Empirical confidence interval coverage remained substantially below the nominal 95% level under all simulation conditions. As shown in Table 4, average coverage ranged from 10.62% for ceiling-distributed observations with a pilot sample size of 10 to 18.00% for skewed observations with a pilot sample size of 20. Across the 36 individual conditions, coverage ranged from 5.50% to 35.00%. Increasing the pilot sample size generally improved coverage, although none of the conditions approached the nominal confidence level. Figure 4 illustrates the distribution- and sample-size-specific coverage patterns. These values refer specifically to the existing normal-approximation Monte Carlo intervals surrounding bootstrap-based power estimates.

Table 4. Empirical 95% Confidence Interval Coverage Across Simulation Conditions

Distribution	Pilot size	Mean coverage (%)	Minimum (%)	Maximum (%)	Deviation from 95% (percentage points)
Normal	10	11.62	5.50	20.75	−83.38
Normal	15	14.62	8.75	26.50	−80.38
Normal	20	17.62	9.00	35.00	−77.38
Skewed	10	12.00	7.00	19.25	−83.00
Skewed	15	13.06	7.25	24.75	−81.94
Skewed	20	18.00	10.00	32.00	−77.00
Ceiling	10	10.62	6.25	17.00	−84.38
Ceiling	15	14.19	7.25	27.00	−80.81
Ceiling	20	17.56	11.00	28.75	−77.44

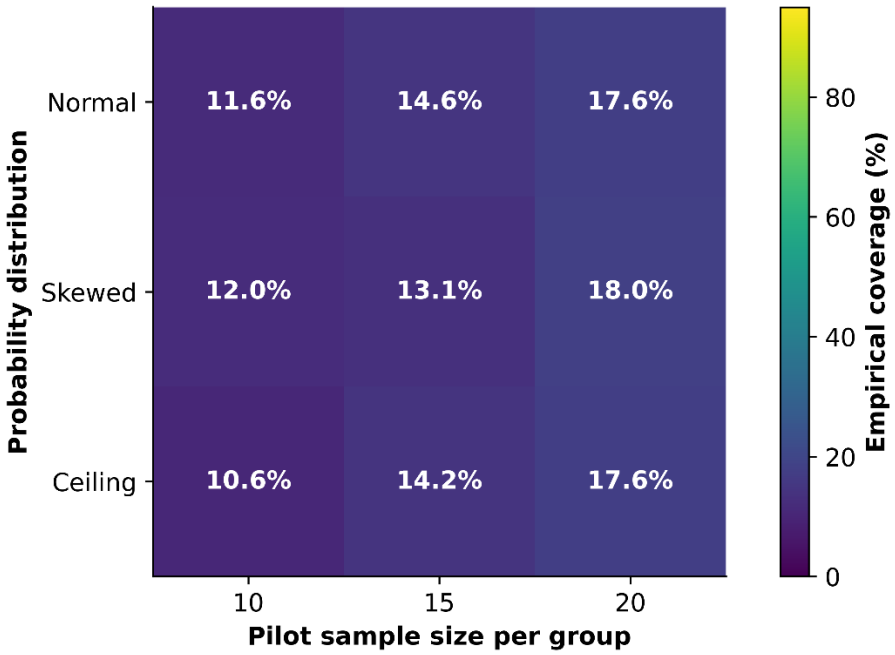


Figure 4. Empirical confidence interval coverage across probability distributions and pilot sample sizes

3.5. Influence of Sample Size on Overall Statistical Reliability

The assessment of sampling variability, estimation accuracy, and coverage of confidence intervals showed improvement as the number of pilot sample units increased. The mean absolute biases (as summarized in Table 5) decreased from 0.1438 for a pilot size of 10 to 0.0806 for a pilot size of 20. In addition, Bootstrap RMSE was also reduced from 0.3281 to 0.2363, and Average empirical coverage was also improved from 11.42% to 17.73%. Between-pilot standard deviation was found to have an empirical coverage with a strong negative descriptive correlation across the 36 simulation conditions ($r = -0.840$). This inverse relationship is seen in Figure 5, but the correlation is based on associations among simulation conditions (not independent observations at the participant level).

Table 5. Overall Statistical Performance Across Pilot Sample Sizes

Pilot size	Mean SD	Absolute bias	Bootstrap RMSE	Parametric RMSE	Coverage (%)
10	0.2856	0.1438	0.3281	0.3414	11.42
15	0.2478	0.1037	0.2744	0.2968	13.96
20	0.2179	0.0806	0.2363	0.2593	17.73

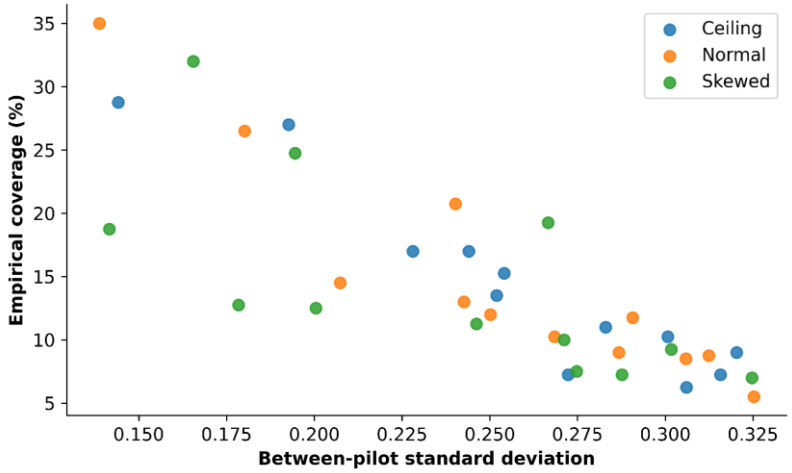


Figure 5. Relationship between sampling variability and empirical confidence interval coverage across simulation conditions

The results, overall, showed that larger pilot sample size led to a better estimation of the bootstrap with less sampling variability, absolute bias and RMSE. But the mean empirical coverage was still significantly lower than the nominal 95%, with an average of 14.37% over 36 simulation conditions. The results emphasize the importance of sample size and probability distribution in determining the reliability of the estimation, and show that the usual Monte Carlo confidence intervals are inappropriate for statistical power estimates based on the bootstrap.

4. Discussion

Sampling variability, bootstrap estimation error, and empirical confidence interval coverage were assessed in the present study for 36 conditions of the simulation. Results showed that the larger the number of pilot sample, the more stable the estimation results, and the standard deviation between the pilots was reduced from 0.2856 to 0.2179. For the same reason, mean absolute bias and root mean squared error (RMSE) also decreased with the increase in sample size. But the actual coverage of the confidence intervals was significantly less than the nominal 95% level, with a mean of 14.37%. Results show that this use of the within-pilot Monte Carlo uncertainty for providing confidence intervals was inadequate for capturing the variability between pilot samples that were independent of each other.

The smaller sampling variability for larger pilot samples is important to emphasize the significance of good sample sizes for good statistical estimation. The RMSE of the Bootstrap was reduced from 0.3281 to 0.2363 showing better accuracy of estimation. Similar issues with variance estimation methods have been studied in propensity score analyses, where bootstrap and asymptotic methods can give different estimates of uncertainty (Austin, 2022). When using dispersion measures to compare two populations, sampling variability also needs to be carefully considered (Greco et al., 2024). Additionally, the use of flexible probability models for complex count observations has revealed the significance of considering data characteristics in statistical parameter estimation (Jiang et al., 2023). The results indicate that the size of the samples and the characteristics of the statistical model have an impact on the reliability of the estimation.

The performance of the bootstrap was not identical for the data sets with normal, skewed, and ceiling distributed observations. The RMSE was comparatively lower for the skewed distribution while it was higher for the normal and ceiling distributions due to higher sampling variations. The importance of considering dependence structures when assessing estimation procedures has been highlighted in statistical studies using interrupted time series (Bottomley et al., 2023). Under temporally correlated noise conditions, also the analytical and the residual bootstrap approaches have been investigated (Innocenti et al., 2022). Additionally, adaptive rcbi's show that interval performance can vary with choice of correction procedure (Tibbe & Montoya, 2022). These studies suggest that bootstrap reliability should be judged in the light of the distributional and sampling conditions.

The most noticeable result was the significant under coverage of nominal 95% confidence intervals with the empirical coverage between 5.50% and 35.00%. This suggests that the assessed intervals were underestimates of the uncertainty due to pilot-to-pilot variation. In the realm of statistical inference, bootstrap confidence intervals have been utilized in indirect and conditional effect estimation, highlighting their significance (Cheung & Cheung, 2024). A distinction between the bootstrap and Bayesian methods also demonstrates the importance of modelling uncertainty estimation methods (Falk et al., 2024). Likewise, the threshold selection for EV analysis is of particular importance, as it is crucial to consider other sources of inferential uncertainty (Murphy et al., 2025). The current results confirm the need to consider not only nominal but also empirical level of confidence.

The results are relevant for statistical modelling and prediction of uncertainty. Uncertainty approaches based on bootstrap have been used in streamflow projections (Galavi et al., 2023), prediction intervals for photovoltaic power (Herrera-Casanova et al., 2024), prediction intervals for neural network (Nguinabé et al., 2025), and carbon price forecasting (Zhu et al., 2025). The applications have different statistics settings, but they all show the need for accurate interval estimation research across fields. Based on the present analysis, bootstrap-based confidence intervals seem to provide a good way to account for uncertainty from independent sampling, especially when small pilot samples provide statistical estimates.

One of the great strengths of this study was the systematic comparison of the various sample sizes, effect sizes, and probability distributions, and the reproducible simulation results. The analysis contained, however, only pre-specified simulation conditions, and combined bootstrap-based power estimates, with a lack of generalizability to other confidence interval procedures. An example of the challenges of estimating quantities that are not certain based on measurement is provided by research that developed methane inventories using measurements (Johnson et al., 2023). Similarly, uncertainty assessment in carbon emission estimation shows that appropriate assessment metrics are important (Lee et al., 2024).

Both larger sample sizes and other probability distributions as well as other bootstrap confidence interval procedures (e.g., percentile, studentized, bias-corrected and accelerated methods) should be investigated further for improved intervals and better statistical reliability.

5. Conclusion

The sampling variability, the accuracy of the estimation and the empirical coverage of the confidence intervals for various sample sizes (n), effect sizes (d) and probability distributions were evaluated in the present study by means of bootstrapping. Results showed that larger the sample size of pilot, more stable the estimate and smaller the statistical uncertainty. The standard deviation between pilots reduced from 0.2856 to 0.2179 while the mean absolute bias and root mean squared error (RMSE) reduced too, signifying smaller errors for larger samples. The results emphasize the need to consider sample size in the context of the reliability of statistical estimation from bootstrap samples. The estimation was very precise, but an average of the 36 simulation conditions was 14.37%, which was below the nominal 95% confidence level. This was a significant under coverage which meant that the 17 normal approximation intervals analysed did not adequately represent the uncertainty of an independent pilot sampling process. Variations between normal, skewed and ceiling distributions also showed the effects of distributional properties on sampling variability and estimation performance. The interpretation of the confidence interval should consider uncertainty both between and within the bootstraps as well as the uncertainty between the independent samples. Additional research needs should be

done for larger samples, other interval techniques and improved uncertainty estimation procedures. Results suggest that the conventional interval construction is problematic for the estimation of power using bootstrap, and may aid the construction of statistical inferences with small sample research applications.

References

1. Aldawsari, A. (2023). Parametric predictive bootstrap and test reproducibility (Doctoral dissertation, Durham University).
2. Austin, P. C. (2022). Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes. Statistics in Medicine, 41(22), 4426-4443.
3. Bottomley, C., Ooko, M., Gasparrini, A., & Keogh, R. H. (2023). In praise of Prais-Winsten: An evaluation of methods used to account for autocorrelation in interrupted time series. Statistics in medicine, 42(8), 1277-1288.
4. Cheung, S. F., & Cheung, S. H. (2024). manymome: An R package for computing the indirect effects, conditional effects, and conditional indirect effects, standardized or unstandardized, and their bootstrap confidence intervals, in many (though not all) models. Behavior Research Methods, 56(5), 4862-4882.
5. 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.
6. Falk, C. F., Vogel, T. A., Hammami, S., & Miočević, M. (2024). Multilevel mediation analysis in R: A comparison of bootstrap and Bayesian approaches. Behavior Research Methods, 56(2), 750-764.
7. Galavi, H., Mirzaei, M., Yu, B., & Lee, J. (2023). Bootstrapped ensemble and reliability ensemble averaging approaches for integrated uncertainty analysis of streamflow projections. Stochastic Environmental Research and Risk Assessment, 37(4), 1213-1227.
8. Greco, L., Luta, G., & Wilcox, R. (2024). On testing the equality between interquartile ranges: L. Greco et al. Computational statistics, 39(5), 2873-2898.
9. Hawinkel, S., Waegeman, W., & Maere, S. (2024). Out-of-sample R 2: Estimation and inference. The American Statistician, 78(1), 15-25.
10. Herrera-Casanova, R., Conde, A., & Santos-Pérez, C. (2024). Hour-ahead photovoltaic power prediction combining BiLSTM and Bayesian optimization algorithm, with bootstrap resampling for interval predictions. Sensors, 24(3), 882.
11. Innocenti, S., Matte, P., Fortin, V., & Bernier, N. (2022). Analytical and residual bootstrap methods for parameter uncertainty assessment in tidal analysis with temporally correlated noise. Journal of Atmospheric and Oceanic Technology, 39(10), 1457-1481.
12. Jané, M., Xiao, Q., Yeung, S., Ben-Shachar, M. S., Caldwell, A., Cousineau, D., ... & Feldman, G. (2024). Guide to effect sizes and confidence intervals. DOI: https://doi. org/10.17605/OSF. IO/D8C4G.
13. Jiang, R., Zhan, X., & Wang, T. (2023). A flexible zero-inflated poisson-gamma model with application to microbiome sequence count data. Journal of the American Statistical Association, 118(542), 792-804.
14. Johnson, M. R., Conrad, B. M., & Tyner, D. R. (2023). Creating measurement-based oil and gas sector methane inventories using source-resolved aerial surveys. Communications Earth & Environment, 4(1), 139.
15. Justus, V. L., Rodrigues, V. B., & Sousa, A. R. D. S. (2024). Bootstrap confidence intervals: A comparative simulation study. arXiv preprint arXiv:2404.12967.
16. Kluger, D. M., Lu, K., Zrnic, T., Wang, S., & Bates, S. (2025). Prediction-powered inference with imputed covariates and nonuniform sampling. arXiv preprint arXiv:2501.18577.
17. Lee, K., Ko, J., & Jung, S. (2024). Quantifying uncertainty in carbon emission estimation: Metrics and methodologies. Journal of Cleaner Production, 452, 142141.
18. Lin, S., Bun, M., Gaboardi, M., Kolaczyk, E. D., & Smith, A. (2024). Differentially private confidence intervals for proportions under stratified random sampling. Electronic Journal of Statistics, 18(1), 1455-1494.
19. Maitner, B. S., Halbritter, A. H., Telford, R. J., Strydom, T., Chacon, J., Lamanna, C., ... & Enquist, B. J. (2023). Bootstrapping outperforms community-weighted approaches for estimating the shapes of phenotypic distributions. Methods in Ecology and Evolution, 14(10), 2592-2610.
20. Mayen, M. S. A., Nisha, S. N., Afrin, S., Ahammed, T., Chowdhury, M. A. B., & Uddin, M. J. (2024). Evaluating the current methodological practices and issues in existing literature in pooling complex surveys: a systematic review. BMC Medical Research Methodology, 24(1), 279.
21. Mokhtar, S. F., Yusof, Z. M., & Sapiri, H. (2023). Confidence intervals by bootstrapping approach: a significance review. Malaysian Journal of Fundamental and Applied Sciences, 19(1), 30-42.
22. Mokhtar, S. F., Yusof, Z. M., & Sapiri, H. (2024). Confidence interval estimating the mean of normal distribution and skewed distribution. Malaysian Journal of Fundamental and Applied Sciences, 20(5), 1124-1135.
23. Murphy, C., Tawn, J. A., & Varty, Z. (2025). Automated threshold selection and associated inference uncertainty for univariate extremes. Technometrics, 67(2), 215-224.
24. Nguinabé, J., Rockefeller, R., Khan, N. M., & Rodrigues, P. C. (2025). Bootstrap prediction intervals for the long short term memory (LSTM) recurrent neural network. Expert Systems with Applications, 284, 127728.
25. 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.
26. Schader, L. M., Song, W., Kempker, R., & Benkeser, D. (2024). Don't let your analysis go to seed: on the impact of random seed on machine learning-based causal inference. Epidemiology (Cambridge, Mass.), 35(6), 764.

27. Tibbe, T. D., & Montoya, A. K. (2022). Correcting the bias correction for the bootstrap confidence interval in mediation analysis. Frontiers in psychology, 13, 810258.
28. Tongmual, N., & Yonwilad, W. (2025). Bayesian Bootstrap Confidence Interval for Mean based on Small Sample Sizes. Malaysian Journal of Fundamental and Applied Sciences, 21(6), 2937-2944.
29. Williams, D. R. (2022). Learning to live with sampling variability: Expected replicability in partial correlation networks. Psychological Methods, 27(4), 606.
30. Zhu, B., Wan, C., Wang, P., & Chevallier, J. (2025). Interval forecasting of carbon price with a novel hybrid multiscale decomposition and bootstrap approach. Journal of Forecasting, 44(2), 376-390.