A COMPARATIVE ANALYSIS OF BOOTSTRAP AND MAXIMUM LIKELIHOOD ESTIMATION METHODS FOR ASSESSING RELIABILITY INDEX

Imasuen Kennedy (Ph.D.) and George Obed Samuel[2]

[1]Institute of Education, University of Benin, Benin City, Nigeria.

[2]Department of Mathematics, University of Benin, Benin City, Nigeria.

[*]Corresponding Author's Email: kennedy.imasuen@uniben.edu; Tel.: + 2348109670163

Cite this article:

Imasuen, K., George, O. S. (2025), A Comparative Analysis of Bootstrap and Maximum Likelihood Estimation Methods for Assessing Reliability Index. African Journal of Mathematics and Statistics Studies 8(1), 159-166. DOI: 10.52589/AJMSS-6SOFLRUI

Manuscript History

Received: 13 Jan 2025

Accepted: 27 Feb 2025

Published: 10 Mar 2025

ABSTRACT: *This study compared bootstrap and maximum likelihood estimation methods for assessing the reliability index using scores from the 2022 National Business and Technical Examination Board (NABTEB) Economics examination. Cronbach's Alpha reliability statistic was applied across various sample sizes (50, 100, 200, 500, 1000, and greater than 1000) to assess measurement reliability. Five confidence interval (CI) estimation methods were utilized: Wald, Profile Likelihood, Bootstrap Percentile, Bias-Corrected and Accelerated (BCa), and Studentized. Findings revealed that SE decreases as sample size increases, demonstrating greater precision with larger samples. The Wald confidence interval, though effective for large samples, proved unreliable for small ones due to its assumption of normality. The Profile Likelihood confidence interval, slightly wider than the Wald confidence interval, better accounted for non-normality. The Bootstrap Percentile confidence interval, a nonparametric approach, provided robust estimates when population distribution assumptions were violated. The BCa method improved accuracy by adjusting for bias and skewness, while the Studentized confidence interval offered conservative estimates, accounting for sample variability. Reliability estimates also increased with sample size. It was therefore recommended that for large samples, use Wald CI; for small samples or skewed data, opt for Profile Likelihood or Bootstrap CIs.*

KEYWORDS: Sample size, Standard error, Confidence intervals, Reliability, Bootstrap.

INTRODUCTION

Reliability estimation is crucial in psychometrics and measurement, ensuring that assessment tools consistently produce stable and accurate scores. Among the numerous statistical techniques for estimating reliability, Maximum Likelihood Estimation (MLE) and the Bootstrap method stand out due to their robust mathematical frameworks and applicability in different contexts. MLE is a parametric method that assumes a known probability distribution and seeks to find parameter estimates that maximize the likelihood function (Kim & Lee, 2023). The Bootstrap method, a non-parametric alternative, relies on resampling techniques to estimate the sampling distribution of a statistic, offering advantages in cases where parametric assumptions may not hold (Efron & Hastie, 2021).

Recent advancements in computational statistics have heightened interest in comparing these two methods, particularly in educational and psychological testing. MLE has been widely applied in Classical Test Theory (CTT) and Item Response Theory (IRT) frameworks, where large-sample, normally distributed data are assumed (Baker & Kim, 2022). In contrast, the Bootstrap method provides a flexible approach, particularly useful in small samples and skewed data distributions (Xie & Wang, 2022). The selection of an appropriate estimation method is crucial, as it directly influences the accuracy and precision of reliability indices.

This study aimed to compare the Bootstrap and Maximum Likelihood Estimation methods in estimating reliability indices, analyzing their performance under different sample sizes and data distributions. The findings would contribute to an improved understanding of the most suitable estimation technique for various psychometric contexts.

Reliability and Its Estimation Methods

Reliability refers to the consistency and stability of measurement instruments across different conditions and testing instances (Tavakol & Dennick, 2019). Traditional reliability estimation methods include Cronbach’s alpha, split-half reliability, and test-retest reliability, all of which are influenced by sample size, data distribution, and underlying model assumptions (Fan & Thompson, 2020).

Maximum Likelihood Estimation (MLE) Model

MLE is a statistical method used to estimate parameters by maximizing the likelihood function of a given observed data. For reliability estimation, the MLE approach assumes a known probability distribution, commonly a normal or logistic distribution. In a psychometric setting, given a set of observed test scores $X = (X_1, X_2, \dots, X_n)$, the likelihood function for a normal distribution with mean μ and variance σ^2 is:

$$L\left(\mu, \frac{\sigma^2}{X}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\ (-\frac{(X_i-\mu)^2}{2\sigma^2})$$

The MLE estimates for μ and σ^2 are obtained by solving:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

Reliability estimation using MLE is often applied in Item Response Theory (IRT), where item parameters (difficulty, discrimination, guessing, and carelessness) are estimated via loglikelihood functions (van der Linden, 2021). However, MLE has limitations, particularly in small samples, where variance estimates can be biased (Meng, 2023).

Bootstrap Estimation Model

The Bootstrap method, introduced by Efron in 1979, is a resampling-based technique that estimates the sampling distribution of a statistic by repeatedly drawing samples (with replacement) from the original dataset. The estimated reliability index using Bootstrap is derived by computing the statistic of interest across multiple resampled datasets.

The Bootstrap procedure follows these steps:

1. Generate B resamples from the observed data set $X = (X_1, X_2, \ldots, X_n)$ by randomly selecting n observations with replacement.
2. Compute the reliability index for each resample.
3. Calculate the Bootstrap estimate of reliability as the average across all B resamples:

$$\hat{R}_{Boot} = \frac{1}{B}\sum_{b=1}^{n} R^b$$

where $R^{(b)}$ represents the reliability estimate from the b^{th} resample.

The Bootstrap method is particularly effective in small-sample conditions and when data are non-normally distributed (Mooney & Duval, 2022). However, its computational intensity and potential for overestimating reliability in highly skewed data warrant careful application (Xie & Wang, 2022).

Several studies have compared MLE and Bootstrap methods in psychometric and educational measurement contexts. Xie and Wang (2022) found that while MLE is more efficient in large samples with normally distributed data, the Bootstrap method produces more stable estimates in small-sample settings. Similarly, Fan and Thompson (2020) demonstrated that the Bootstrap method yields better confidence intervals for reliability estimates, making it a preferable choice in exploratory research.

Maximum Likelihood Estimation (MLE) - Wald Confidence Interval

Using the asymptotic normality of MLE, the Wald confidence interval is given by:

$CI = \hat{R} \pm z\alpha/2 \times SE(\hat{R})CI$

where:

$\hat{R}$ = estimated reliability coefficient

SE($\hat{R}$)= standard error of $\hat{R}$ estimated as:

SE($\hat{R}$) $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{R} - \underline{R})^2}$

Article DOI: 10.52589/AJMSS-6SOFLRUI
DOI URL: https://doi.org/10.52589/AJMSS-6SOFLRUI

where:

n = sample size

$\hat{R}$ = estimated reliability coefficient for sample ii

$\underline{R}$ = mean of all estimated reliability coefficients

$Z_{\frac{\alpha}{2}}$= critical value from the standard normal distribution (e.g., 1.96 for a 95% confidence interval)

Bootstrap Confidence Interval

The Bootstrap Standard Error (SE) is computed from bootstrap resampling:

$$\text{SE}(\hat{R}\ ) = \sqrt{\frac{1}{B}\sum_{b=1}^{nB}(\hat{R}^{*} - \underline{R}^{*})^{2}}$$

where:

B = number of bootstraps resamples

$\hat{R}^{*}$ = estimated reliability coefficient from bootstrap sample bb

$\underline{R}^{*}$ = mean of bootstrap reliability estimates

Bootstrap Confidence Interval Formulas

Percentile Bootstrap CI:

$$CI = (\hat{R}^{*}(_{0.025B}), \hat{R}^{*}(_{0.975B})$$

This is obtained by **s**orting the bootstrap estimates $\hat{R}_{b}{}^{*}$and selecting the 2.5th percentile and 97.5th percentile as the CI bounds.

Studentized Bootstrap CI (Uses SE)

$$CI = \hat{R}\ \pm t_{(0.975,B)} \times SE^{*}$$

where $t_{(0.975,B)}$ is the bootstrap t-statistic from the bootstrap distribution.

Maximum Likelihood Estimation (MLE) and Confidence Intervals for Reliability Coefficients

MLE estimates reliability coefficients by maximizing the likelihood function, but it also allows us to compute confidence intervals for these estimates using asymptotic theory or the profile likelihood method.

MLE Confidence Interval Estimation Approaches

Wald Confidence Interval (Standard Error Approach)

The Wald method requires large sample sizes for valid inference, and assumes the sampling distribution is normal, which may not hold in small or skewed samples. After estimating the reliability coefficient $\hat{R}$, the standard error (SE) is computed using the Fisher information matrix.

$$SE(\hat{R}\,)= \sqrt{Var(\hat{R}\,)}$$

The Wald confidence interval is then given by:

$$CI = \hat{R}\ \pm Z_{\frac{\alpha}{2}} \times SE(\hat{R}\,)$$

where $Z_{\frac{\alpha}{2}}$ is the critical value from the standard normal distribution.

Profile Likelihood Confidence Interval

Instead of using a normal approximation, this method constructs confidence intervals by finding the range of values for which the likelihood function remains within a given threshold. This method is more accurate for small samples compared to Wald intervals.

Mathematically, it finds values R_{low} and R_{high} such that:

$$2logL\left(\hat{R}\,\right) - 2LogL(R) \leq \chi^2{}_{1,\alpha}$$

where $\chi^2{}_{1,\alpha}$ is the chi-square critical value for a given confidence level.

Bootstrap Confidence Interval Estimation

Bootstrap resampling allows us to estimate confidence intervals without relying on normality assumptions.

Bootstrap Confidence Interval Approaches

Percentile Bootstrap CI

The percentile bootstrap confidence interval is simple to implement and does not assume normality. However, it can be biased if the original sample size is small. The steps are:

Resample the data B times.

Compute the reliability coefficient $R^*{}_b$ for each bootstrap sample b.

Sort the bootstrap estimates and take the 2.5th percentile and 97.5th percentile as the lower and upper bounds, respectively. $CI = (R^*({}_{0.025B}), R^*({}_{0.975B})$

Bias-Corrected and Accelerated (BCa) Bootstrap CI

It is more accurate than percentile CI, and recommended for skewed or non-normal data. However, it requires more computation than percentile CI. The Bias- Corrected and

Article DOI: 10.52589/AJMSS-6SOFLRUI
DOI URL: https://doi.org/10.52589/AJMSS-6SOFLRUI

Accelerated bootstrap adjusts for bias and skewness in the bootstrap distribution. The adjusted confidence limits are given by:

$$CI = (R^*(\Phi(z_0 + z^*{}_{0.025}), (R^*(\Phi(z_0 + z^*{}_{0.975}))$$

where z_0 corrects for bias and z^*corrects for acceleration (curvature of the bootstrap distribution).

Studentized Bootstrap CI

The studentized bootstrap confidence interval computes bootstrap standard errors for each sample and adjusts the confidence interval accordingly:

$$CI = (\hat{R} \ - t_{(0.975,B)} \cdot SE^*, \hat{R} \ + t_{(0.025,B)} \cdot SE^*)$$

It is more robust in small samples, but requires estimating standard error (SE) for each bootstrap sample, increasing computational cost.

RESULTS AND DISCUSSION OF FINDINGS

The study analyzed scores from the 2022 National Business and Technical Examination Board (NABTEB) Economics examination. To assess reliability, Cronbach's Alpha statistic was employed across various sample sizes, including 50, 100, 200, 500, 1000, and a final category exceeding 1000. Five different methods were utilized in the analysis: the Wald and Profile Likelihood confidence intervals, derived from Maximum Likelihood Estimates, along with the bootstrap confidence interval method, which includes the Percentile, Bias-Corrected and Accelerated (BCa), and Studentized approaches. The results are presented in Table 1.

Table 1: Confidence Interval for Reliability Estimates

N	50	100	200	500	1000	>1000
Mean	36.88	36.00	33.42	33.18	34.02	33.51
Std Dev	2.80	2.47	6.26	5.99	5.56	6.34
SE	0.40	0.25	0.44	0.27	0.18	0.19
Reliability estimate	0.64	0.75	0.81	0.82	0.84	0.86
Wald CI (95%)	(36.10, 37.66)	(35.51, 36.49)	(32.56, 34.28)	(32.65, 33.71)	(33.67, 34.37)	(33.14, 33.88)
Profile Likelihood CI (Approx.)	(36.05, 37.60)	(35.45, 36.55)	(32.50, 34.30)	(32.60, 33.75)	(33.62, 34.40)	(33.10, 33.90)
Bootstrap Percentile CI	(36.02, 37.50)	(35.40, 36.50)	(32.45, 34.35)	(32.55, 33.80)	(33.60, 34.42)	(33.05, 33.95)
Bias-Accelerated (BCa) CI	(35.98, 37.55)	(35.35, 36.55)	(32.40, 34.40)	(32.50, 33.85)	(33.55, 34.45)	(33.00, 34.00)
Studentized CI	35.95, 37.60)	(35.95, 37.60)	(32.35, 34.45)	(32.45, 33.90)	(33.50, 34.50)	(32.95, 34.05)

Table 1 highlights the inverse relationship between standard error (SE) and sample size, illustrating that as the sample size increases, SE decreases. For instance, when N=50N = 50, SE is 0.40, whereas at N=1000N = 1000, SE drops to 0.18, indicating greater precision with a larger sample. This supports the statistical principle that larger samples produce more accurate estimates of the population mean (Kelley & Maxwell, 2023; Xu & Dang, 2022).

Each confidence interval (CI) method estimates the range within which the true population mean is expected to fall 95% of the time, though they handle variability differently. The Wald Confidence Interval performs well with large samples but is less reliable for small ones due to its assumption of normality, which can lead to inaccuracies under non-normal conditions (Agresti, 2021). The Profile Likelihood Confidence Interval is slightly wider than the Wald CI, reflecting greater uncertainty. For example, at N=50N = 50, the Profile Likelihood CI was (36.05, 37.60), demonstrating its flexibility in handling non-normality (Bolker, 2023).

The Bootstrap Percentile Confidence Interval**,** based on resampling techniques, is particularly useful when the population distribution is unknown. At N=50N = 50, Bootstrap Percentile CI = (36.02, 37.50), making it a robust nonparametric alternative when traditional assumptions do not hold (Efron & Hastie, 2022). The Bias-Corrected and Accelerated (BCa) Confidence Interval refines the standard bootstrap by adjusting for bias and skewness, enhancing accuracy in non-normal distributions. At N=50N = 50, BCa CI was (35.98, 37.55), offering improved interval estimation, particularly for small samples (Davison & Hinkley, 2023). The Studentized Confidence Interval is more conservative, often yielding slightly wider intervals to account for sample variability. At N=50N = 50, the Studentized CI was (35.95, 37.60), providing more reliable estimates by incorporating an SE adjustment (Wilcox, 2023).

Reliability estimates improve with increasing sample size, leading to more stable measurements. For example, at N=50N = 50, the reliability estimate is 0.64, whereas for N>1000N > 1000, it rises to 0.86. This aligns with Classical Test Theory (CTT**)**, which states that reliability strengthens as sample size increases (McNeish & Wolf, 2022).

Overall, larger sample sizes reduce SE, enhancing precision. The Bootstrap and Profile Likelihood confidence interval**s** offer more robust estimates for skewed or non-normal data, whereas the Wald confidence interval, though easy to compute, may underestimate uncertainty in small samples (Hox, Moerbeek, & van de Schoot, 2023). The BCa and Studentized confidence interval**s** improve on bootstrap methods by adjusting for bias and variability, resulting in more reliable confidence intervals (Efron & Hastie, 2022). As sample size increases, reliability also improves, ensuring greater stability and consistency in measurements. This underscores the importance of larger samples in generating precise and dependable data (Zumbo & Hubley, 2023).

CONCLUSION

Based on the findings of the study it was concluded that as sample size increases, standard error (SE) decreases, leading to more precise estimates of the population mean. This principle is evident across various CI methods, where larger samples yield narrower and more accurate confidence intervals. The Wald confidence interval performs well with large samples but is less reliable for small samples or non-normal data. The Profile Likelihood confidence interval provides a slightly wider interval, incorporating uncertainty more effectively. Bootstrap-based methods (Percentile, Bias-Corrected and Accelerated, and Studentized confidence intervals) offer robust alternatives for non-normal distributions, with the BCa method improving accuracy by adjusting for bias and skewness.

RECOMMENDATIONS

Based on the findings of the study, it was recommended as:

1. To enhance reliability and reduce standard error, prioritize larger sample sizes in research.

2. For large samples, use Wald CI; for small samples or skewed data, opt for Profile Likelihood or Bootstrap CIs.

3. Future research should investigate advanced CI methods, especially in high-dimensional or machine learning-based studies.

REFERENCES

Agresti, A. (2021). *Statistical Methods for the Social Sciences* (5th ed.). Pearson.

Baker, F. B., & Kim, S. H. (2022). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press.

Bolker, B. M. (2023). *Generalized Linear Mixed Models: A Practical Guide*. CRC Press.

Davison, A. C., & Hinkley, D. V. (2023). *Bootstrap Methods and Their Application*. Cambridge University Press.

Efron, B., & Hastie, T. (2021). *Computer age statistical inference: Algorithms, evidence, and data science* (2nd ed.). Cambridge University Press.

Efron, B., & Hastie, T. (2022). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press.

Fan, X., & Thompson, B. (2020). Confidence intervals for reliability coefficients. *Educational and Psychological Measurement, 60*(2), 225–239.

Hox, J., Moerbeek, M., & van de Schoot, R. (2023). *Multilevel Analysis: Techniques and Applications*. Routledge.

Kelley, K., & Maxwell, S. E. (2023). *Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation*. Routledge.

Kim, J., & Lee, Y. (2023). Maximum likelihood estimation in psychometric analysis: Principles and applications. *Journal of Educational Measurement, 60*(1), 45–67.

McNeish, D., & Wolf, M. G. (2022). *Reliability and Validity in Measurement: A Quantitative Perspective*. Springer.

Meng, X. (2023). The challenge of small-sample inference in maximum likelihood estimation. *Psychometrika, 88*(3), 512–530.

Mooney, C. Z., & Duval, R. D. (2022). *Bootstrapping: A nonparametric approach to statistical inference* (2nd ed.). SAGE Publications.

Tavakol, M., & Dennick, R. (2019). Making sense of Cronbach’s alpha. *International Journal of Medical Education, 2*, 53–55.

van der Linden, W. J. (2021). *Handbook of item response theory* (Vol. 1). Chapman & Hall/CRC Press.

Wilcox, R. R. (2023). *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. CRC Press.

Xie, Q., & Wang, L. (2022). A comparison of bootstrap and maximum likelihood estimation in reliability analysis. *Psychological Methods, 27*(4), 599–615.

Xu, W., & Dang, S. (2022). *Principles of Statistical Inference: A Computational Approach*. Wiley.

Zumbo, B. D., & Hubley, A. M. (2023). *Understanding and Investigating Validity in Social Science Research*. Springer.