On Sample Size Needed for Block Bootstrap Confidence Intervals to Have Desired Coverage Rates

Mathew Chandy, Elizabeth D. Schifano, & Jun Yan*

Department of Statistics, University of Connecticut, Storrs, CT

https://doi.org/10.33697/ajur.2024.101

*Student: mathew.chandy@uconn.edu**
Mentors: elizabeth.schifano@uconn.edu, jun.yan@uconn.edu

ABSTRACT
Block bootstrap is widely used in constructing confidence intervals for parameters estimated from stationary time series. Theoretically, the method should provide valid confidence intervals as the length of the time series goes to infinity. In practice, however, it is necessary to know how large of a finite sample is required for block bootstrap confidence intervals to work well. This study aims to answer this question in a simple simulation setting where the data are generated from a first-order autoregressive process. The empirical coverage rates of several commonly used bootstrap confidence intervals for the mean, standard deviation, and the lag-1 autocorrelation coefficient are compared. A quite large sample is found necessary for the intervals to have the right coverage rates even when estimating a simple parameter like the mean. Some block bootstrap methods could fail when estimating the lag-1 autocorrelation. It is surprising that the coverage property even deteriorates as the sample size increases with some commonly used block bootstrap confidence intervals including the percentile intervals and bias-corrected intervals.

KEYWORDS
Autocorrelation; Bias-Correction; Centering; Dependent Data; Percentile; Resampling; Simulation; Time Series

INTRODUCTION

Block bootstrap is a tool to construct confidence intervals (CI) to make inferences about dependent data. Essentially, it depends on correct estimation of the uncertainty in the estimation, similar to the standard bootstrap[1], but for serially dependent data. Early ideas of block bootstrap were developed not long after the standard bootstrap.[2–4] It has since been applied in various fields, for instance, econometrics and meteorology.[5, 6] Block bootstrap is especially useful for serially dependent data when the serial dependence is not specified or not of primary interest. The method is expected to produce CIs with coverage rates matching their nominal levels as the sample size grows.[7] However, when dealing with finite sample sizes, an important question is how large the sample size must be for block bootstrap CIs to have the desired coverage rates.

Lahiri[8] finds that moving block bootstrap has better performance than non-overlapping block bootstrap. Additionally, moving block bootstrap with nonrandom block sizes results in lower mean-squared errors than moving block bootstrap with random block sizes. Buhlmann and Künsch[9] notes that a drawback of block bootstrap is that it heavily depends on block size, which has to be chosen by the user of the method. Even when using the appropriate settings, as noted by Buhlmann[10] observes some general drawbacks of block bootstrap — with respect to how reasonably it imitates the data-generating process. In addition, although block bootstrap is primarily used for stationary time series, it can be outperformed by other bootstrap schemes for linear time series and categorical processes. Still, Buhlmann[10] emphasizes

that a significant advantage of block bootstrap is its simplicity. To be more specific, the resampling step of block bootstrap is not computationally more difficult than the resampling step of basic bootstrap. Furthermore, block bootstrap performs better than local bootstrap in terms of mimicking dependence structures.

For independent data, extensive research has explored the effectiveness of bootstrap standard errors in providing accurate uncertainty measures. For example, Hesterberg [11] observes that while percentile-based CIs for the mean parameter are more accurate than t-intervals for larger sample sizes, their accuracy diminishes for smaller sample sizes. The optimal parameter estimation of a distribution, according to Chernick and Labudde [12], depends on the sample size, the number of bootstrap replicates, and the confidence level. In structural equation modeling, Nevitt and Hancock [13] find that a sample size of 200–1000 is sufficient for interval estimation using standard nonparametric bootstrap. In estimating variance components, Burch [14] reports that as the sample size increases under a normal distribution, nonparametric bootstrap methods approach the coverage of a pivotal quantity, but for other distributions, the coverage can deteriorate. In estimating the correlation coefficient of bivariate normal data, Puth et al. [15] note that even for a sample size of 100 with true correlation coefficient 0, bootstrap methods are less accurate than the Fisher's transformation. The prevailing consensus highlights the necessity of a substantial sample size for bootstrap CIs to attain the desired coverage.

Limited research has offered practical guidance concerning the requisite sample size for employing block bootstrap inference with dependent data. In the context of linear regression involving dependent data, where regression errors stem from a homoscedastic autoregressive process of order-1, the investigation conducted by Goncalves and White [16] reveals that, in cases of small sample sizes, standard error estimates derived from the moving block bootstrap approach may demonstrate greater accuracy than those based on closed-form asymptotic estimates. Nonetheless, even when considering a substantial sample size of 1024, confidence intervals generated through the moving block bootstrap method still fail to adequately encompass the target parameter. The scarcity of existing literature addressing the necessary sample sizes conducive to the efficacy of block bootstrap techniques has spurred the initiation of the present study.

The goal of this paper is to provide recommendations on necessary sample size for block bootstrap with dependent data, similar to what was done for basic bootstrap in Hesterberg [11]. We consider a simple situation of a stationary time series, where the parameters of interests are the mean, standard deviation, and the first-order autocorrelation coefficient. We compare six variants of block bootstrap CIs from the literature: [17, 18] a standard normal CI, a Student's t CI, a percentile CI, a bias-corrected CI, a bias-corrected and accelerated CI, and a recentered percentile CI proposed in this article. Their empirical coverage rates at different sample sizes and dependence levels are compared in a simulation study. The results of this study suggest that recovery of temporal dependence parameters is reliant on the type of interval used.

The remainder of the paper is organized as follows. The first section reviews block bootstrap procedures and how to use block bootstrap estimates to construct CIs; a simple CI obtained by recentering at the original point estimate is proposed for comparison. The second section reports a simulation study comparing the coverage rates of six block bootstrap CIs. A discussion concludes in the final section.

BLOCK BOOTSTRAP CIS

Consider a stationary time series $\{X_t : t = 1, \ldots, n\}$ with length n. Our goal is to construct a CI for a parameter θ in the data generating model of the series. Suppose that $\hat{\theta}_n$ is a point estimator of θ based on the observed series. Bootstrap is a powerful approach to construct CIs. If the observations in the series were independent, a standard nonparametric bootstrap procedure would draw a large number B bootstrap copies of the observed data, and calculate a bootstrap point estimate $\hat{\theta}_n^{(b)}$ for each copy $b = 1, \ldots, B$. The uncertainty of $\hat{\theta}_n$ is then estimated by the empirical uncertainty of the bootstrap point estimates. When serial dependence is present, the bootstrap procedure needs to preserve the serial dependence. Block bootstrap was motivated for this situation.

Block Bootstrap

Block bootstrap preserves the serial dependence in the observed data by partitioning the data into blocks and performing bootstrap on the blocks. In particular, consider block size l and, for convenience, suppose that n is a multiple of l such that there are $k = n/l$ blocks. Each block j is $Y_j = \{X_{(j-1)l+1}, \ldots, X_{(j-1)l+l}\}$, $j = 1, \ldots, k$. Then, we sample k blocks of Y_j's from the set $\{Y_1, \ldots, Y_k\}$ with replacement and concatenate the k sampled blocks in the order they are picked to form a bootstrap sample of the data. The formation of the bootstrap sample ensures that the between-block dependence is weak and that the within-block serial dependence is preserved. Because the blocks here are non-overlapping, this bootstrap approach is known as non-overlapping block bootstrap, or simple block bootstrap.

Alternatively, block-bootstrap can be done with overlapping or moving blocks. Define moving blocks

$$Z_j = \{X_j, \ldots, X_{j+l-1}\}, \quad j = 1, \ldots, n-l+1.$$

Now we draw k blocks from the $(n-l+1)$ blocks of Z_j's with replacement and then align them in the order they were picked to form a block bootstrap sample. If n is not a multiple of l, the last block selected will be reduced in size so that the final size of the block bootstrap sample is n. It is also possible to implement moving block bootstrap while allowing blocks to wrap around the end of the series. In other words, define moving blocks (assuming $l > 1$) as:

$$Z_j = \begin{cases} \{X_j, \ldots, X_{j+l-1}\}, & \text{if } j = 1, \ldots, n-l+1, \\ \{X_j, \ldots, X_n, X_1, \ldots, X_{j-n+l-1}\}, & \text{if } j = n-l+2, \ldots, n. \end{cases}$$

This version does not require that n/l be an integer.

The block size l needs to be chosen with care. It should be large enough for each bootstrap sample to preserve the serial dependence, yet small enough for there to be a large number of blocks to give sufficient variability between each bootstrap sample. As n increases, both l and n/l should also increase. To achieve this, the order of l is often assigned a value as a function of n. A common expression that is considered optimal for the order of l is $\lceil n^{1/3} \rceil$,[9] which was adopted in this study.

Block Bootstrap CIs

Suppose that we have repeated the steps in the last subsection B times, and that for $b \in \{1, \ldots, B\}$, we have obtained a bootstrap point estimate $\hat{\theta}_n^{(b)}$ based on the bth bootstrap sample using the same method that was applied to $\{X_t : t = 1, \ldots, n\}$ to obtain $\hat{\theta}_n$. Now the question is how to construct a CI for θ using the B bootstrap point estimates $\{\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(B)}\}$. We consider six kinds of block bootstrap CIs adapted from standard bootstrap CIs.

Standard Normal CI Assuming that $\hat{\theta}_n$ is asymptotically normally distributed with θ as the mean, we just need an estimate of the standard error to construct an approximate CI.[19] Let $\widehat{\text{SE}}$ be the empirical standard error of the bootstrap point estimates $\hat{\theta}_n^{(b)}$ for $b \in \{1, \ldots, B\}$. Let $z_{(\alpha)}$ be the quantile function $F^{-1}(\alpha)$ of the standard normal distribution. A $(1-\alpha)100\%$ standard normal CI is

$$(\hat{\theta}_n - z_{(1-\alpha/2)}\widehat{\text{SE}}, \quad \hat{\theta}_n - z_{(\alpha/2)}\widehat{\text{SE}}).$$

This CI is centered by the original point estimate $\hat{\theta}_n$ and is symmetric. The standard CI is classified by Efron and Tibshirani[19] as a confidence interval based on bootstrap "tables", which essentially means it is based on an asymptotic distribution with an estimated asymptotic variance (standard error). Its validity relies on whether the distribution of $\hat{\theta}_n$ is reasonably well approximated by its asymptotic normal distribution and whether the bootstrap $\widehat{\text{SE}}$ approximates the true standard error.

Student's t CI The procedure for constructing a Student's t CI based on standard bootstrap is described in Efron and Tibshirani[19]. Let $t_{(\alpha,k)}$ be the quantile function $F^{-1}(\alpha, k)$ of a t distribution with k degrees of freedom. With block bootstrapping, a $(1-\alpha)100\%$ Student's t CI is

$$(\hat{\theta}_n - t_{(1-\alpha/2),k-1}\widehat{\mathrm{SE}}, \quad \hat{\theta}_n - t_{(\alpha/2),k-1}\widehat{\mathrm{SE}}),$$

where k is the number of blocks. This CI is centered by the original point estimate $\hat{\theta}_n$ and is symmetric. Like the standard normal interval, the Student's t CI is classified by Efron and Tibshirani[19] as a confidence interval based on bootstrap "tables". In this case, its validity relies on whether the distribution of $\hat{\theta}_n$ is reasonably well approximated by the t_{k-1} distribution with an expected value of θ and whether the bootstrap $\widehat{\mathrm{SE}}$ approximates the true standard error.

Percentile CI The percentile CI was first suggested in Efron[1]. Let $\hat{\theta}^B_{n,\alpha}$ be the empirical 100αth percentile of $\{\hat{\theta}^{(1)}_n, \ldots, \hat{\theta}^{(}_n$ A $(1-\alpha)100\%$ empirical percentile CI is

$$(\hat{\theta}^B_{n,\alpha/2}, \quad \hat{\theta}^B_{n,1-\alpha/2}).$$

This CI is not necessarily centered by the original point estimate $\hat{\theta}_n$. As will be shown in our simulation study, this approach works well for the marginal mean and standard deviation of a serially dependent process, but its coverage of the temporal dependence deteriorates as n increases, which is contrary to what one would expect.

Bias-Corrected (BC) CI The procedure for constructing a bias-corrected Bootstrap CI based on standard bootstrap is described in Carpenter and Bithell[20]. Let $\hat{z}_0 = \Phi^{-1}\{\#\{\hat{\theta}^{(b)}_n < \hat{\theta}_n\}/B\}$ for $b \in \{1, \ldots, B\}$. Define $\alpha_1 = \Phi(2\hat{z}_0 - z_{1-\alpha/2})$ and $\alpha_2 = \Phi(2\hat{z}_0 - z_{\alpha/2})$. A $(1-\alpha)100\%$ BC CI is

$$(\hat{\theta}^B_{n,\alpha_1}, \quad \hat{\theta}^B_{n,\alpha_2}).$$

Bias-Corrected and Accelerated (BCA) CI The BCA CI was first suggested in Efron[21]. Let $Z_{(i)}$ be the original sample without the ith block z_i for $i \in \{1, \ldots, k\}$, let $\hat{\theta}_{(i)}$ be the statistic of $Z_{(i)}$, and let $\hat{\theta}_{(.)} = k^{-1}\sum_{i=1}^k \hat{\theta}_{(i)}$. Let

$$\hat{a} = \frac{\sum_{i=1}^k (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^k (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2\}^{3/2}}.$$

Define

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right)$$

and

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right).$$

A $(1-\alpha)100\%$ BCA CI is

$$(\hat{\theta}^B_{n,\alpha_1}, \quad \hat{\theta}^B_{n,\alpha_2}).$$

This CI is not necessarily centered by $\hat{\theta}_n$. The BCA method corrects for bias and skewness of the B bootstrap point estimates $\{\hat{\theta}^{(1)}_n, \ldots, \hat{\theta}^{(B)}_n\}$ by including bias-correction and acceleration factors. The acceleration factor refers to the rate of change of the standard error of $\hat{\theta}_n$ with respect to θ.

Recentered Percentile CI We propose a CI that is centered at the original point estimate and uses the variation in the bootstrap estimates to construct the error bound. The motivation behind proposing such an interval was based on the simulation performance of the BC and BCA intervals, which will be discussed further in the Results section. This interval requires the computation of $\bar{\theta}_n^B = n^{-1}\sum_{b=1}^{B}\hat{\theta}_n^{(b)}$, the mean of all bootstrap point estimates. A $(1-\alpha)100\%$ CI is centered around $\hat{\theta}_n$ and can be written as

$$(\hat{\theta}_n + \hat{\theta}^B_{n,\alpha/2} - \bar{\theta}^B_n, \quad \hat{\theta}_n + \hat{\theta}^B_{n,1-\alpha/2} - \bar{\theta}^B_n).$$

It is not necessarily symmetric, as different critical values are used to compute the lower and upper bounds. It has the same width as the percentile CI.

SIMULATION DESIGN

We compared the performance of the different block bootstrap CI methods under two marginal distributions: standard normal and unit exponential.

Marginal Standard Normal Distribution

We generated time series X_t from a 1st order autoregressive (AR(1)) process:

$$X_t = \phi X_{t-1} + \epsilon_t,$$

where ϕ is an autoregressive coefficient, and ϵ_t is a series of independent errors from a normal distribution with mean zero and variance σ^2_ϵ. The strength of the serial dependence is controlled by ϕ, which was set to five levels: $\{-0.4, -0.2, 0.0, 0.2, 0.4\}$. We only used serial dependences as strong as 0.4, because we only seek to establish the general trend as the strength of the autocorrelation increases, and how it varies depending on the sign of the autocorrelation and the parameter of interest. The series X_t has mean zero and variance $\sigma^2_x = \sigma^2_\epsilon/(1-\phi^2)$, so for each value of ϕ, we set $\sigma^2_\epsilon = (1-\phi^2)$ such that $\sigma^2_x = 1$.

Three target parameters of X_t were considered: 1) $\mu = 0$, the mean of X_t; 2) $\sigma_x = 1$, the standard deviation of X_t; and 3) ϕ, the lag-1 autocorrelation coefficient. To investigate the effect of sample size n, we considered an array of values $n \in \{100, 200, 400, 800, 1600, 3200\}$. In each configuration, we generated 10,000 replicates. The block bootstrap sampling step was done with function `tsboot` from R[22] package *boot*,[23] with block size $\lceil n/l \rceil$. This function by default is an implementation of moving block bootstrap as described in the previous section, meaning that that blocks are allowed to wrap around, and we tried both $l = \lceil n^{1/3} \rceil$ and $l = \lceil 2n^{1/3} \rceil$, keeping the order of the block size constant but varying the coefficient. For each replicate, we constructed six 95% block bootstrap CIs for each parameter as described in the last section with $B = 1000$. We can estimate μ, σ_x, and ϕ by computing the sample mean, sample standard deviation, and lag-1 autocorrelation, respectively, of each bootstrap sample. Then we can construct intervals for each parameter using the appropriate procedures described in *Block Bootstrap CIs*. Then we estimated their actual coverage rates along with their 95% confidence intervals from the 10,000 replicates.

The coverage rates of the CIs were used to compare the performance of CIs. Let $\hat{\theta}_{L,r}$ and $\hat{\theta}_{U,r}$ be the lower and upper bound, respectively, for the confidence interval constructed for each replicate $r \in \{1, \ldots, R\}$, where R is the number of replicates. Then the empirical coverage rate is $\sum_{r=1}^{R} I\{\hat{\theta}_{L,r} < \theta < \hat{\theta}_{U,r}\}/R$, where $I(\cdot)$ in the indicator function. If a CI method is valid, then the coverage rate is expected to match the nominal level. Because it is unlikely for the coverage to exactly match the nominal level, we can construct a 95% Clopper-Pearson exact CI of the coverage rate,[24] which is an estimate of a proportion with $R = 10,000$. We used the R *PropCIs* package to achieve this.[25] The choice of Clopper-Pearson was motivated by the Wald interval's poor coverage as the proportion approaches 0 or 1,[26] although when we tried Wald intervals, the coverage rate intervals did not appear to have large differences. If the proportion 0.95 is included in the interval, the block bootstrap method is likely performing well. If all values in the interval are below 0.95, the results would suggest that the method either is providing inaccurate estimation, is underestimating the process' variability, or a combination of both. If all values in the interval are above .95, the results suggest that the method is

overestimating the process' variability. **Figure 1** summarizes the empirical coverage rates and the 95% confidence intervals of the real coverage for a marginal standard normal distribution using block bootstrap with $l = \lceil n^{1/3} \rceil$, generated using the R *ggplot2* package.[27]

Marginal Unit Exponential Distribution

Additionally, to investigate if the results are robust to nonnormal marginal distributions, we evaluated the performance of block bootstrap for a time series with a non-normal marginal distribution. Specifically, we estimated the mean, standard deviation, and the lag-1 autocorrelation coefficient of a stationary series with marginal unit exponential distribution. Note that we expect the CIs that are based on bootstrap "tables" to depend on the asymptotic distribution of the estimator. This asymptotic distribution depends more on the sample size than on the marginal distribution of the time series. So we expect such CIs to have similar performance for different marginal distributions when the sample sizes is large. The percentile-based CIs (Percentile, BC, BCA, Recentered Percentile) are not necessarily expected to perform better under non-normal marginal distributions. The student's t and and normal-based CIs are only noticeably different when the number of blocks is smaller than 20.

The series were generated by marginally transforming the AR(1) series X_t in the first simulation study by

$$W_t = F^{-1}[\Phi(X_t)],$$

where $F^{-1}(p)$ is the quantile function for the unit exponential distribution. The true mean (μ) and standard deviation (σ_w) parameters of W_t are 1. The lag-1 autocorrelation coefficient (ρ) is not invariant to the transformation,[28] but its value can be obtained by

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F^{-1}[\Phi(x)]F^{-1}[\Phi(y)]g_2(x,y;\phi)\mathrm{d}x\mathrm{d}y - 1,$$

where $g_2(x, y; \phi)$ is the density of a standard bivariate normal distribution with correlation parameter ϕ. We kept the configuration of $\phi \in \{-0.4, -0.2, 0.0, 0.2, 0.4\}$, and the corresponding lag-1 autocorrelation coefficients are $\rho \in \{-0.298, -0.156, 0, 0.170, 0.355\}$.

SIMULATION RESULTS

Marginal Standard Normal Distribution

For estimating the mean parameter μ, the top panel of **Figure 1** suggests that all methods eventually approach correct coverage of μ as sample size increases. Student's t CIs appear to need the smallest sample size to achieve correct coverage, except for samples with strong negative dependence, in which case, they actually over-cover μ for smaller sample sizes. For instance, for a sample with $n = 100$ and $\phi = -0.4$, the lower bound for a Students t CI's coverage of μ is greater than 95%, whereas the coverage intervals for other methods contain 95%. The standard normal, percentile, BC, and BCA, and recentered percentile CIs require similar sample sizes to recover μ at the nominal level for all combinations of n and ϕ. All methods seem to require a smaller sample to recover μ at the nominal rate when dealing with negative dependence versus positive dependence. For example, BC CIs recover μ for $n \geq 100$ when $\phi = -0.2$, but they only recover μ for $n \geq 800$ when $\phi = 0.2$. In addition, as a negative dependence gets stronger, holding everything else equal, coverage increases, which lead to the Student t CI's aforementioned over-coverage. As a positive dependence gets stronger, holding everything else equal, coverage decreases, and a larger sample is necessary to recover μ. A possible explanation for this is that if a stationary series has a positive autocorrelation, the effective sample size is decreased, whereas if a series has a negative autocorrelation, the effective sample size is increased.[29] Additionally, this seems to have a greater effect on the the estimation of the location parameter versus that of the scale parameter or temporal dependence parameter.

For estimating the standard deviation parameter σ_x, **Figure 1** suggests that every method can reach nominal coverage of σ_x if the sample is large enough, but for a given n and ϕ, coverage of σ_x will be lower than coverage of μ in general. Like μ, σ_x can be covered by Student t CIs with smaller sample sizes when compared to other methods. Unlike

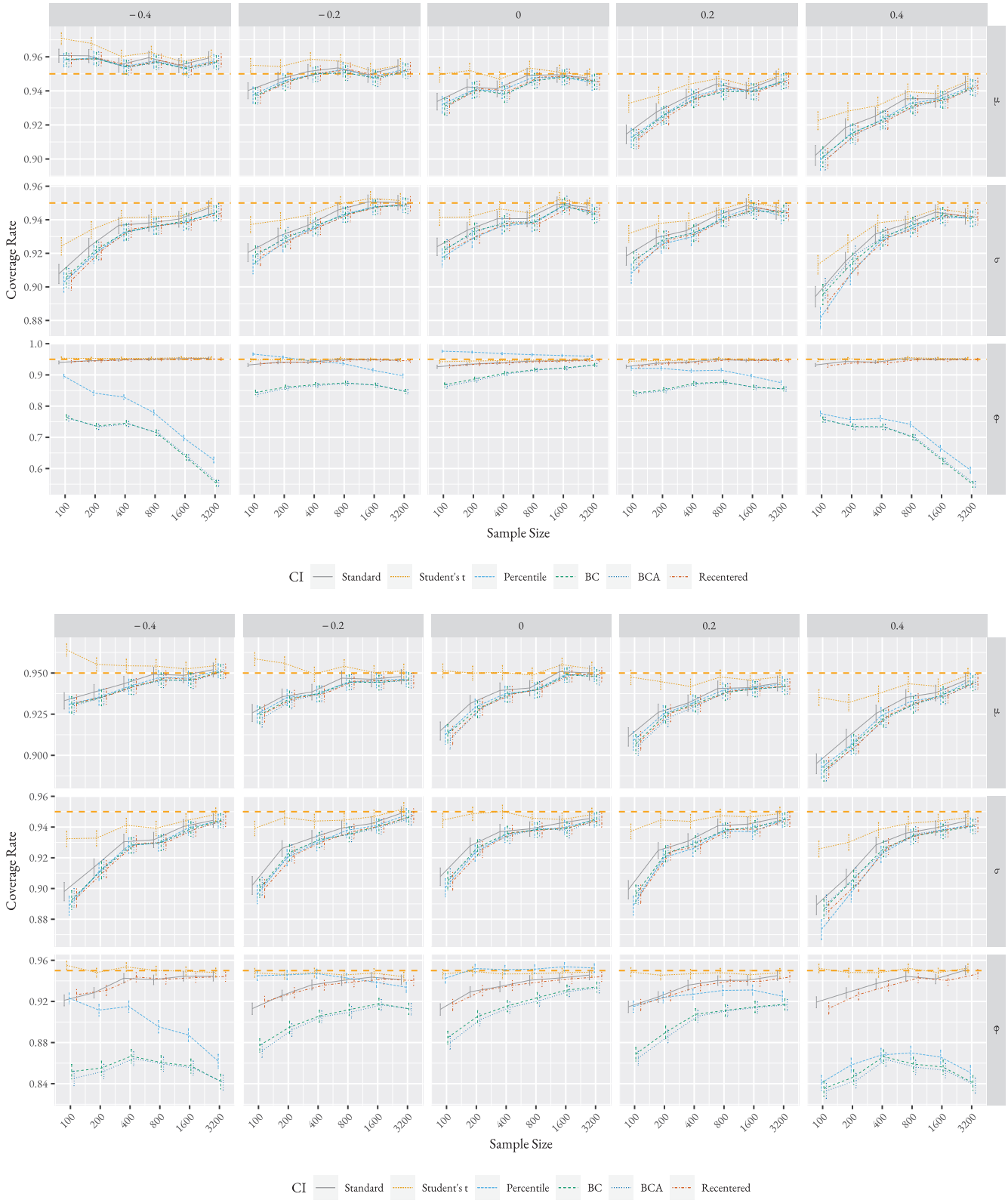

Figure 1. Empirical coverage rates of different 95% block bootstrap CIs for the marginal mean μ, the marginal standard deviation σ_x, and the first-order autocorrelation coefficient ϕ of an AR(1) process with a marginal standard normal distribution with AR coefficient $\phi \in \{-0.4, 0.2, 0, 0.2, 0.4\}$ and series length $n \in \{100, 200, 400, 800, 1600, 3200\}$ based on 10,000 replicates of block bootstrap with $l = \lceil n^{1/3} \rceil$. The error bars represent 95% CIs of the real coverage rates. Top: $l = \lceil n^{1/3} \rceil$. Bottom: $l = \lceil 2n^{1/3} \rceil$.

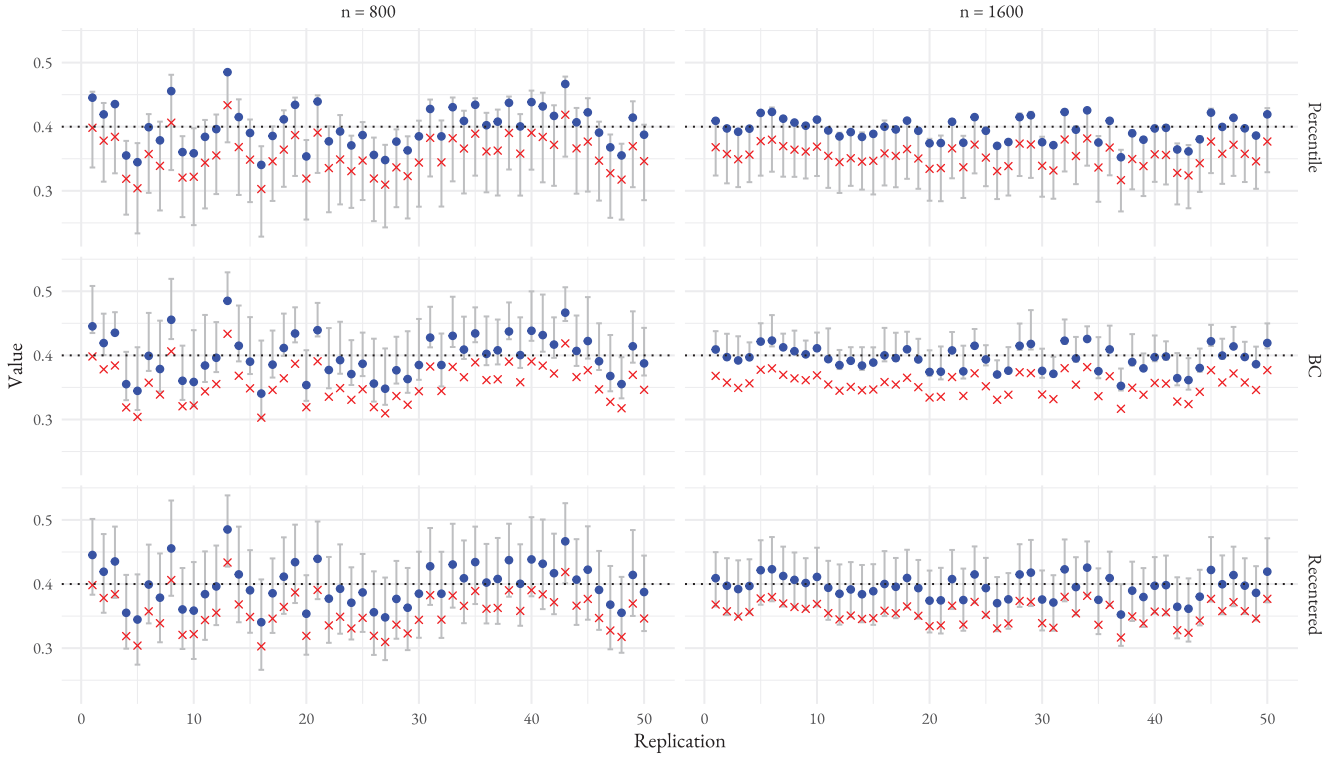

Figure 2. 50 replicate percentile, BC, and recentered percentile CIs for samples of size $n \in \{800, 1600\}$ ($l = \lceil n^{1/3} \rceil$) for the lag-1 autocorrelation of an AR(1) process with $\phi = 0.4$. For each replicate, the lower and upper bounds of the CIs are displayed, as well as $\hat{\theta}_n$ (blue circle) and $\bar{\theta}_n^B$ (red cross).

μ, there is no over-coverage issue for σ_x when $\phi = -0.4$. Standard normal, percentile, BC, BCA, and recentered percentile CIs again have similar performance. All methods seem to have slightly higher coverage of σ_x when ϕ is negative versus when ϕ is positive. Regardless of the sign, coverage of σ_x gets worse as the strength of the temporal dependence increases.

For estimating the autocorrelation parameter ϕ, **Figure 1** suggests that while standard normal, Student's t, and recentered percentile CIs do approach correct coverage as sample size increases, percentile, BC, and BCA CIs deteriorate as sample size increases, especially as the strength of the temporal dependence increases. Because of this, only standard normal, Student's t, and recentered percentile CIs should be considered as effective block bootstrap methods to estimate ϕ. Student's t CIs once again can achieve correct coverage with smaller sample sizes when compared to standard and recentered percentile CIs which perform similarly. Student's t CIs can can recover ϕ at the nominal level for $n \geq 100$ when the sample's temporal dependence is as strong as 0.4. Coverage appears to be higher for all methods when the dependence is negative rather than positive. Whether or not the dependence is negative or positive, coverage of ϕ seems to increase slightly as the absolute value increases for standard normal, Student's t, and recentered percentile CIs. For the values of ϕ observed, there are no examples of over-coverage for standard normal, Student's t, BC, BCA, or recentered percentile CIs. However, percentile CIs appear to over-cover ϕ for smaller sample sizes when $\phi = -0.2$ and when $\phi = 0$, indicating again that they should not be used.

The outcomes of the ϕ estimation raise a natural question about the lackluster performance of certain methodologies. To delve into this inquiry, a set of 50 CIs was generated for each of the percentile, BC, and recentered percentile approaches for samples of $n \in \{800, 1600\}$. Illustrated in **Figure 2**, it becomes evident that the percentile-based CIs exhibit a notable bias, predominantly manifesting as a substantial underestimation of ϕ with point estimator $\bar{\theta}_n^B$, that is, the average of B bootstrap point estimates. As the sample size increases from 800 to 1600, this bias does not vanish while the uncertainty reduces, which explains why the coverage rates deteriorate. The bias in $\bar{\theta}_n^B$ also appears to inval-

idate the bias-correction in the BC bootstrap, leading to the poor performance of the BC intervals. The BCA intervals have the same problem as the BC intervals in the bias-correction step. The root of the issue appears to be that the autocorrelation in the block bootstrap samples is somehow smaller compared to that in the original sample. On the other hand, the original point estimator $\hat{\theta}_n$ is asymptotically unbiased. Since the width is based on the uncertainty in the bootstrap point estimates $\hat{\theta}_n^{(b)}$, $b = 1, \ldots, B$, the percentile CIs recentered at the original point estimate $\hat{\theta}_n$ provide desired coverage.

To summarize, the performance of the CIs depends on the target parameter. When estimating μ and σ_x, any CI will do, although Student's t CIs perform noticeably better than the others. However, when estimating ϕ, the choice of method is of utmost importance as to avoid coverage deterioration. Coverage rates are acceptable at smaller sample sizes when phi is positive versus when phi is negative. In other words, a larger sample size is generally required to estimate a parameter for a sample with a negative ϕ versus a positive ϕ of the same magnitude. In order to know if coverage will increase as the strength of the temporal dependence increases, one need to know what the parameter of interest is, and in the case of μ, the direction of the serial dependence. The BC approach does not seem to be correcting bias appropriately when estimating ϕ. Like the percentile method, the recentered percentile method uses the spread from the bootstrap to construct the width of the CI. However, the recentered approach, does not correct from the original point estimate $\hat{\theta}_n$.

The results for $l = \lceil 2n^{1/3} \rceil$ are reported in the bottom panel of Figure 1. The performances generally seems to be inferior compared those with $l = \lceil n^{1/3} \rceil$, but importantly, the patterns in performance when varying other parameters appear to be robust to the different block size. For negative autocorrelations, the coverage rates of μ appear to be lower when using $l = \lceil 2n^{1/3} \rceil$. For example, whereas $n = 100$ or 200 would seem sufficient for most CIs when using $l = \lceil n^{1/3} \rceil$, $n = 800$ or 1600 is necessary to capture negative autocorrelations for $l = \lceil n^{1/3} \rceil$. Student's t CIs do not seem to be as affected by this change in l: for $\phi = -0.4$ and -0.2, they still over-cover μ for smaller values of n. The results for σ_x with $l = \lceil 2n^{1/3} \rceil$ look very similar to those the results for σ_x with $l = \lceil n^{1/3} \rceil$, but coverage rates of σ_x do look slightly lower especially for negative values of ϕ, although Student's t CIs are again not as influenced by this change in l. A larger sample size seems necessary when using other CIs to estimate σ_x for $l = \lceil 2n^{1/3} \rceil$. Recentered percentile and standard CIs have slightly lower coverage rates when estimating negative values of ϕ with $l = \lceil 2n^{1/3} \rceil$. Although it is still a problem, the coverage deterioration appears to be less dramatic for BCA, BC, and percentile CIs. Aside from these differences, the overall changes in performance when other experimental factors are changed are the same as when $l = \lceil n^{1/3} \rceil$.

Marginal Unit Exponential Distribution

For the scenario of marginal exponential distribution, the empirical coverage rates for μ, σ_w, and the lag-1 autocorrelation coefficient ρ using block bootstrap with $l \in \{\lceil n^{1/3} \rceil, \lceil 2n^{1/3} \rceil$, as well as 95% confidence intervals of the real coverage are displayed in **Figure 3**. Additionally, a set of 50 CIs are displayed for each of the percentile, BC, and recentered percentile approaches for exponentially distributed samples of $n \in \{800, 1600\}$ with lag-1 autocorrelation coefficient 0.355 ($\phi = 0.4$) in **Figure 4**.

It appears that a greater sample size is generally required for the bootstrap CIs to cover the mean and standard deviation parameters in the exponential margin case than in the normal margin case. However, the other trends and patterns discussed regarding the performance of various methods and diverse parameters remain unchanged. For example, Student's t confidence intervals still exhibit higher coverage rates in comparison to alternative methods. Performance continues to be more favorable when temporal dependence is negative rather than positive. Again, altering the block size results in the same changes in performance of different CIs as those in the scenario of marginal normal distribution. Of particular significance, the percentile, BC, and BCA confidence intervals still display a decline in coverage accuracy for the lag-1 autocorrelation coefficient as sample size increases as demonstrated in Figure 4. Both the percentile and BC intervals persist in manifesting the same bias issue. On the other hand, the recentered percentile confidence interval

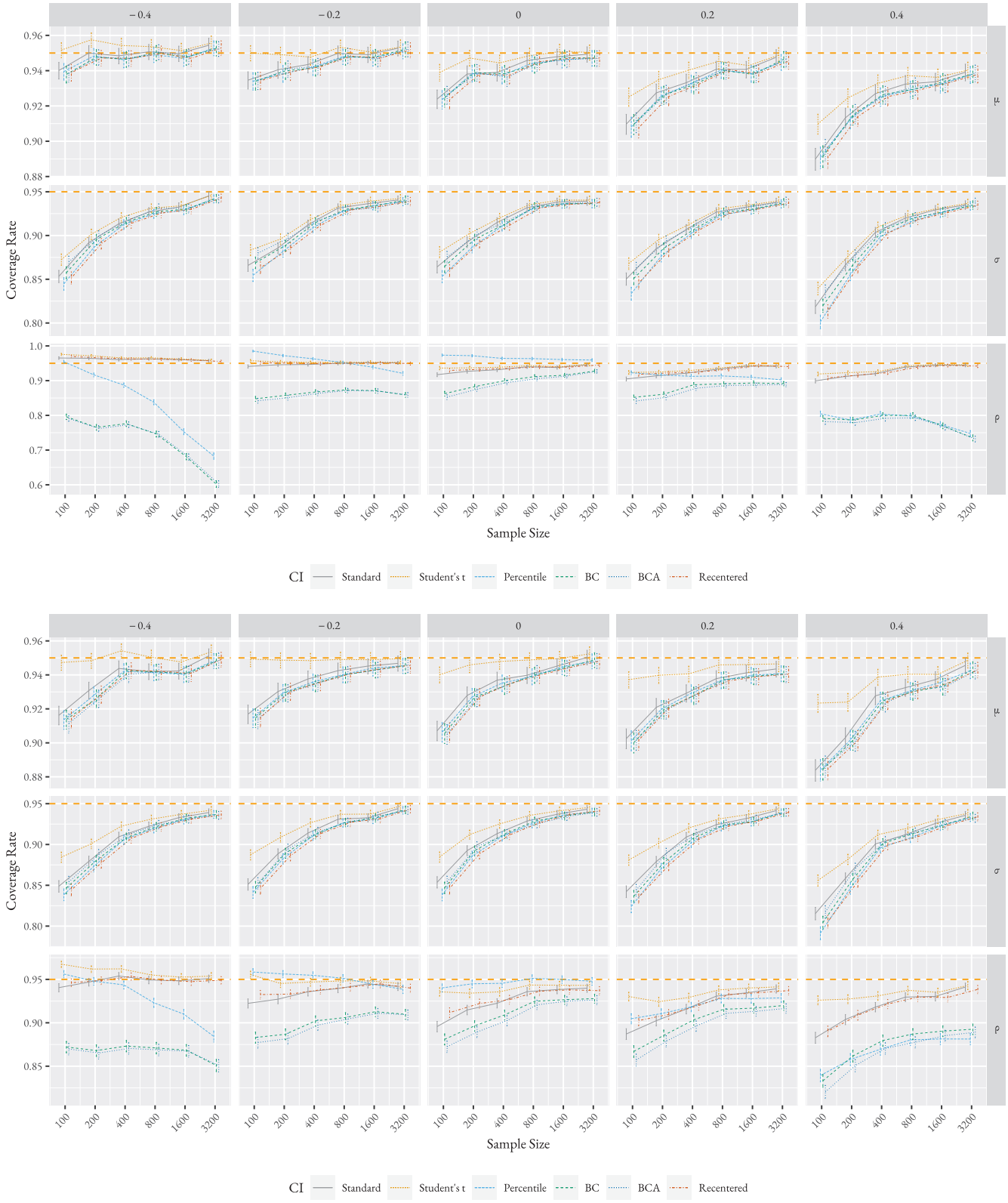

Figure 3. Empirical coverage rates of different 95% block bootstrap CIs for the marginal mean μ, the marginal standard deviation σ_w, and the first-order autocorrelation coefficient ρ of a stationary series with marginal unit exponential distribution obtained by transforming an AR(1) process with $\phi \in \{-0.4, 0.2, 0, 0.2, 0.4\}$ with series length $n \in \{100, 200, 400, 800, 1600, 3200\}$ based on 10,000 replicates replicates of block bootstrap with $l = \lceil n^{1/3} \rceil$. The error bars represent 95% CIs of the real coverage rates. Top: $l = \lceil n^{1/3} \rceil$. Bottom: $l = \lceil 2n^{1/3} \rceil$.

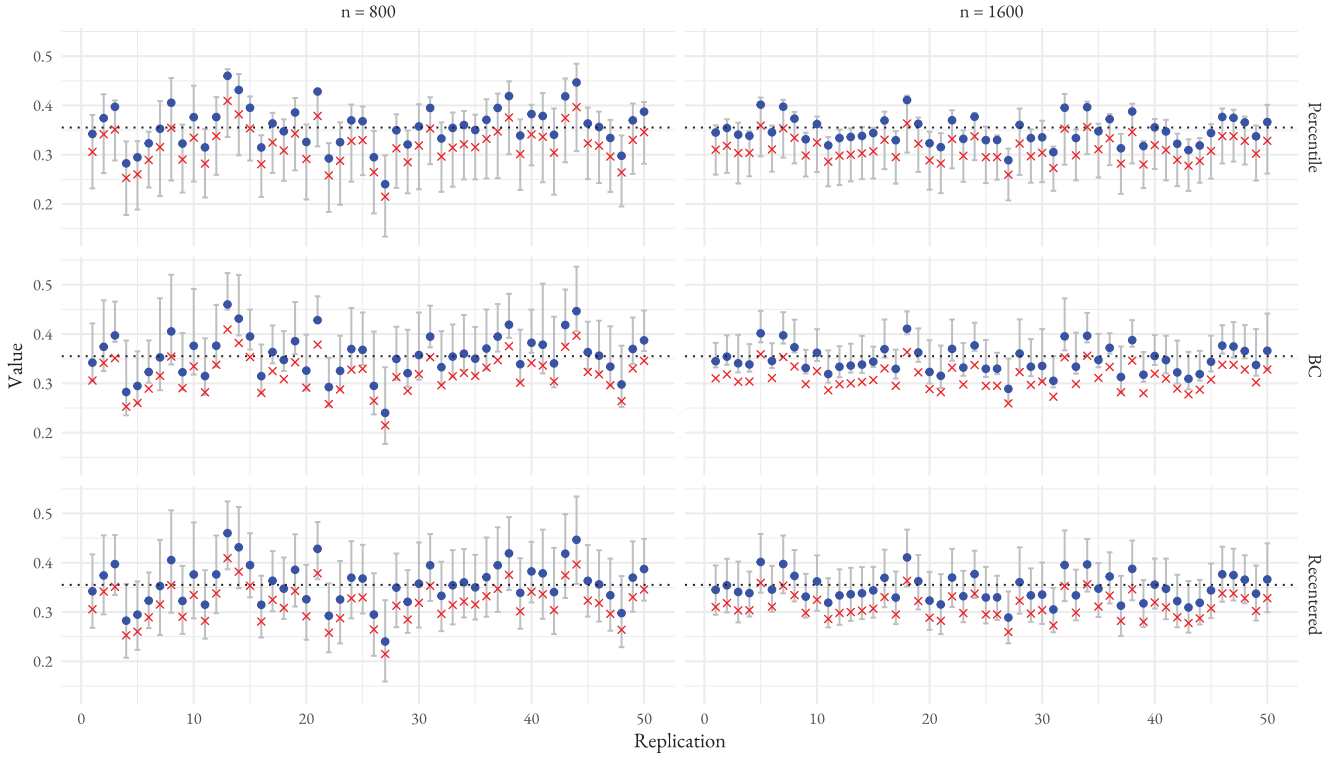

Figure 4. 50 replicate percentile, BC, and recentered percentile CIs for the lag-1 autocorrelation coefficient $\rho = 0.355$ ($\phi = 0.4$) of a stationary series with marginal exponential distribution obtained by transforming an AR(1) process with $\phi = 0.4$ with sample size $n \in \{800, 1600\}$ ($l = \lceil n^{1/3} \rceil$). For each replicate, the lower and upper bounds of the CIs are displayed, as well as $\hat{\theta}_n$ (blue circle) and $\bar{\theta}_n^B$ (red cross).

continues to be effective in estimating the temporal dependence due to the inherent unbiasedness of the original point estimator. In sum, for the most part, the findings for series that are marginally exponentially distributed closely mirror those attained for series that are marginally normally distributed.

DISCUSSION

Block bootstrap is a useful method for estimating parameters of a time series, from simple parameters like the mean to more complicated temporal dependence factors. We know theoretically that the block bootstrap procedure will cover a parameter of a time series at the nominal level given an infinitely large sample,[7] so the goal for this study was to find the smallest finite sample length n of a time series in order for the block bootstrap procedure to recover its associated parameters at an acceptable rate. Our analysis relies on the assumption that there is a size n large enough for the method to work: that is, the method's performance improves as n increases. Out of the six types of intervals used in this study, this assumption was found to hold true with respect to estimating ϕ only for standard normal, Student's t, and recentered percentile CIs, whereas percentile, BC, and BCA intervals exhibited coverage deterioration as n increased. The percentile CI's coverage deterioration can be attributed to bias that is not corrected as n increases. Specifically, as n increases, the width of the CI decreases, but because the percentile CI underestimates ϕ, the coverage decreases. The BC CI seems to correct the bias, but the width of the CI seems to be too short. The acceleration factor of the BCA CI seems to fail, as the width of the CI seems to be too short.

One of the goals of this study was to provide some practical recommendations for necessary sample sizes when using block bootstrap to estimate the parameters of serially dependent data. When using Student's t intervals and the marginal distribution and temporal dependence is unknown, the results of this study suggest that $n \geq 1600$ may be necessary for common practice to estimate μ, whereas $n > 3200$ may be necessary to estimate the standard deviation. Student's t is always preferable to Standard Normal CIs as they performs better for smaller sample sizes and performs

as good or better for larger sample sizes. Lastly, to estimate lag-1 autocorrelation, $n \geq 100$ using the Student's t method may be sufficient under a marginal standard normal distribution, whereas $n \geq 1600$ may be required under under a marginal exponential distribution. Further investigation may be necessary to see if there are other percentile-based interval corrections that fix the coverage deterioration problem for ϕ.

Although we have only used serial dependences as strong as 0.4, we have established the trends as $|\phi|$ gets larger. When estimating μ, we expect coverage rates to decrease as ϕ approaches 1 — as ϕ approaches -1, we may observe increased over-coverage. When estimating the standard deviation, we expect a larger sample size to be necessary as $|\phi|$ gets closer to 1. Lastly, when estimating ϕ for a marginal normal distribution, we expect a larger sample size to be necessary as $|\phi|$ approaches 0, assuming standard normal, Student's t, or recentered percentile CIs are used. However, when estimating the first-order autocorrelation of a marginal exponential distribution using the same methods, we expect coverage rates to respond to stronger dependences in a trend similar to that of coverage rates of μ. We expect other percentile-based CIs, which are already inadequate for relatively weak dependence structures, to perform even worse as $|\phi|$ approaches 1.

This study could be used as a guide for applied statistics courses for students to generally understand how large of a sample size is sufficient for block bootstrap to be used versus other inference methods. For undergraduate or graduate students, block bootstrap is not typically a part of curriculum, but the results of this study can easily be used to demonstrate when it is practical to use this method. This information could also prove to be useful for research using block bootstrap estimation of time series in domains such as econometrics. Future studies could investigate the n needed to make inferences about other forms of serially dependent data such as a moving average process. One could also investigate if there are types of block bootstrap interval construction such as ABC or bootstrap-t intervals[19] that could more appropriately recover the parameters of a time series. We discussed some drawbacks of block bootstrap in the introduction, which could motivate a similar simulation study for alternatives to block bootstrap, such as AR-Sieve bootstrap,[30] which Buhlmann[10] finds to be the best for linear time series. Finally, there is a need for a more in-depth exploration to comprehend the reasons behind the subpar performance of existing percentile-based CIs when estimating the autocorrelation parameter. It is crucial to conduct a thorough investigation into the specific scenarios where the proposed CI demonstrates superior performance and the conditions under which it should be recommended.

REFERENCES

1. Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 7(1), 1–26. *https://doi.org/10.1007/978-1-4612-4380-9_41*
2. Hall, P. (1985) Resampling a coverage pattern. *Stochastic Processes and their Applications* 20(2), 231–246. *https://doi.org/10.1016/0304-4149(85)90212-1*
3. Carlstein, E. (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics* 14(3), 1171–1179. *https://doi.org/10.1214/aos/1176350057*
4. Kunsch, H. R. (1989) The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* 17(3), 1217–1241. *https://doi.org/10.1214/aos/1176347265*
5. MacKinnon, J. G. (2006) Bootstrap methods in econometrics. *The Economic Record* 82(1), 2–18. *https://doi.org/10.1111/j.1475-4932.2006.00328.x*
6. Varga, L., and Zempléni, A. (2017) Generalised block bootstrap and its use in meteorology. *Advances in Statistical Climatology, Meterology and Oceanography* 3(1), 55–66. *https://doi.org/10.5194/ascmo-3-55-2017*
7. Calhoun, G. (2018) Block bootstrap consistency under weak assumptions. *Econometric Theory* 34(6), 1383–1406. *https://doi.org/10.1017/S0266466617000500*
8. Lahiri, S. N. (1999) Theoretical comparisons of block bootstrap methods. *The Annals of Statistics* 27(1), 386–404. *https://doi.org/10.1214/aos/1018031117*
9. Bühlmann, P., and Künsch, H. R. (1999) Block length selection in the bootstrap for time series. *Computational Statistics & Data Analysis* 31(3), 295–310. *https://doi.org/10.1016/S0167-9473(99)00014-6*
10. Bühlmann, P. (2002) Bootstraps for time series. *Statistical Science* 52–72. *https://doi.org/10.1214/ss/1023798998*

11. Hesterberg, T. C. (2015) What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician* 69(4), 371–386. *https://doi.org/10.1080/00031305.2015.1089789*

12. Chernick, M. R., and Labudde, R. A. (2009) Revisiting qualms about bootstrap confidence intervals. *American Journal of Mathematical and Management Sciences* 29(3-4), 437–456. *https://doi.org/10.1080/01966324.2009.10737767*

13. Nevitt, J., and Hancock, G. R. (2001) Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling* 8(3), 353–377. *https://doi.org/10.1207/S15328007SEM0803_2*

14. Burch, B. D. (2012) Nonparametric bootstrap confidence intervals for variance components applied to interlaboratory comparisons. *Journal of Agricultural, Biological, and Environmental Statistics* 17(2), 228–245. *https://doi.org/10.1007/s13253-012-0087-9*

15. Puth, M.-T., Neuhäuser, M., and Ruxton, G. D. (2015) On the variety of methods for calculating confidence intervals by bootstrapping. *Journal of Animal Ecology* 84(4), 892–897. *https://doi.org/10.1111/1365-2656.12382*

16. Gonçalves, S., and White, H. (2005) Bootstrap standard error estimates for linear regression. *Journal of the American Statistical Association* 100(471), 970–979. *https://doi.org/10.1198/016214504000002087*

17. DiCiccio, T. J., and Efron, B. (1996) Bootstrap confidence intervals. *Statistical Science* 11(3), 189–228. *https://doi.org/10.1214/ss/1032280214*

18. Rice, J. A. (2006) *Mathematical Statistics and Data Analysis*. Cengage Learning, Boston

19. Efron, B., and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton. *https://doi.org/10.1201/9780429246593*

20. Carpenter, J., and Bithell, J. (2000) Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine* 19(9), 1141–1164. *https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F*

21. Efron, B. (1987) Better bootstrap confidence intervals. *Journal of the American Statistical Association* 82(397), 171–185. *https://doi.org/10.1080/01621459.1987.10478410*

22. R Core Team. (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

23. Canty, A. (2022) *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-28.1

24. Clopper, C. J., and Pearson, E. S. (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26(4), 404–413. *https://doi.org/10.2307/2331986*

25. Scherer, R. (2018) *PropCIs: Various Confidence Interval Methods for Proportions*. R package version 0.3-0

26. Brown, L. D., Cai, T. T., and DasGupta, A. (2001) Interval estimation for a binomial proportion. *Statistical Science* 16(2), 101–133. *https://doi.org/10.1214/ss/1009213286*

27. Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York *https://doi.org/10.1007/978-0-387-98141-3*

28. Hofert, M., Kojadinovic, I., Mächler, M., and Yan, J. (2018) *Elements of Copula Modeling with R*. Springer, New York. *https://doi.org/10.1007/978-3-319-89635-9*

29. Geyer, C. J. (2011) Introduction to Markov chain Monte Carlo. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov chain Monte Carlo*, 3–48. CRC Press, Boca Raton *https://doi.org/10.1201/b10905*

30. Kreiss, J.-P. (1992) Bootstrap procedures for AR (∞)—processes. In *Bootstrapping and Related Techniques: Proceedings of an International Conference, Held in Trier, FRG, June 4–8, 1990*, 107–113. Springer, New York. *https://doi.org/10.1007/978-3-642-48850-4_14*

ABOUT STUDENT AUTHOR

Mathew Chandy is a senior majoring in both Statistics and Statistical Data Science, and he plans to graduate in the Spring of 2024.

PRESS SUMMARY

This simulation study evaluates the sample size necessary to estimate the mean, standard deviation, and lag-1 autocorrelation of a stationary time series using different block bootstrap confidence interval types. The results showed that percentile-based confidence intervals for the lag-1 autocorrelation may suffer from coverage deterioration as sample size is increased, motivating the authors to propose a new recentered percentile confidence interval which does not deteriorate in performance for greater sample sizes. The results also suggest that when using Student's t bootstrap confidence intervals, a sample size of at least 1600 may be sufficient to estimate the mean, whereas a sample size larger than 3200 may be necessary to estimate the standard deviation. The results additionally indicate that estimation of the lag-1 autocorrelation - using Student's t bootstrap confidence intervals - demands a sample size of at least 100 when the marginal distribution is standard normal and a sample size of at least 1600 when the marginal distribution is unit exponential.