Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

BOOTSTRAP CONFIDENCE INTERVALS FOR SIMULATION OUTPUT PARAMETERS

Russell R. Barton

Department of Supply Chain and Information Systems
The Pennsylvania State University
210 Shortlidge Road
University Park, PA 16802, USA

Luke A. Rhodes-Leader

Department of Management Science
Lancaster University
Lancaster, LA1 4YR, UK

ABSTRACT

Bootstrapping has been used to characterize the impact on discrete-event simulation output arising from input model uncertainty for thirty years. The distribution of simulation output statistics can be very non-normal, especially in simulation of heavily loaded queueing systems, and systems operating at a near optimal value of the output measure. This paper presents issues facing simulationists in using bootstrapping to provide confidence intervals for parameters related to the distribution of simulation output statistics, and identifies appropriate alternatives to the basic and percentile bootstrap methods. Both input uncertainty and ordinary output analysis settings are included.

1 INTRODUCTION

The interest in simulation models as *digital-twins* requires using input probability distributions that correctly capture those of the real system. Input distribution mischaracterizations happen for many reasons. There are issues of lack of stationarity, non-representative sampling methods, and limited sample size that must be addressed. Here we focus on the issue of mischaracterization due to finite sample size, assuming that we have input distributions that are stationary, and that truly random samples have been collected. Any finite sample size is insufficient to perfectly characterize an input probability distribution. When the objective is prediction rather than insight, it is important to give a quantitative characterization of the uncertainty in simulation output results caused by imperfect input probability distributions. We often seek an estimate of a parameter θ that characterizes the performance of the simulated system. Examples for θ might include expected throughput, the .95-quantile for waiting time, etc. We want to characterize the effect of errors in input distributions on errors in an estimated parameter, $\hat{\theta}$. This is the *input uncertainty* problem.

The use of parametric distribution families such as exponential, lognormal, normal, or Weibull permit a reduction in the uncertainty of probability model characterization, but they have their own drawbacks. Most parametric distributions used by the simulation community have infinite support, yet that is not the case for the distributions for real systems. Further, characteristics such as the exponential distribution family's continuously increasing density as the argument approaches zero is not representative of real systems. The use of parametric probability models can lead to unrealistic characterization of congestion. Difficulties in validating parametric probability distribution assumptions have, in part, motivated simulation modelers and software developers to use nonparametric empirical distributions based on samples of data from the real system. The issues and approaches discussed in this paper apply to both parametric and non-parametric input distribution characterizations.

Denote the simulation output (using input distribution F) from the rth replication as $Y_r(F) = \mathrm{E}(Y_r(F)|F) + \varepsilon_r(F), r = 1,2,\ldots,R$, where $\varepsilon_r(F)$ represents the run-to-run deviation from the mean of the output statistic. We assume that $\varepsilon_r(F)$ has zero mean and finite variance that can depend on the input probability distribution F: $\sigma^2(F) = \mathrm{Var}(\varepsilon_r(F)|F)$. Note that F can be multivariate, and here we let it represent the entirety of input

distributions used to drive the simulation. Our objective is to estimate some parameter $\theta(F^c)$ characterizing the probability distribution of $Y_r(F^c)$ where F^c represents the true (c for correct) input probability distributions. For example, $\theta(F^c) = \mathrm{E}(Y_r(F^c))$ or $\theta(F^c) = q_{.95}(Y_r(F^c))$. The parameter estimate $\hat{\theta}(\hat{F^c})$ is generally a straightforward calculation using the outputs $\{Y_r(\hat{F^c})\}$, where $\hat{F^c}$ represents the estimated input distributions based on finite samples.

Consider characterizing the uncertainty in $\hat{\theta}(\hat{F^c})$ as an estimate of $\theta(F^c)$ by constructing a $100(1-\alpha)\%$ confidence interval for $\theta(F^c)$. To construct the interval it is important to understand two different sources of error in $\hat{\theta}(\hat{F^c})$.

$$\hat{\theta}(\hat{F^c}) = \theta(F^c) + \left(\hat{\theta}(\hat{F^c}) - \theta(F^c)\right) + \psi(\varepsilon_r(\hat{F^c})). \tag{1}$$

Uncertainty in the value of $\theta(F^c)$ comes from the second and third terms in (1). The second term captures random variation that comes from the difference between the parameter estimate using the estimated input distributions in $\hat{F^c}$ (which depend on random samples) and the true parameter value using the true distribution F^c. This source is called *extrinsic error* since it arises from error in the input distribution, determined externally to the simulation model. The third term represents the impact of the finite duration of the stochastic simulation on the estimated parameter. It arises from the internal stochastic nature of the simulation and called *intrinsic error*. Even with the true input distributions, this error would remain: the stochastic simulation runs for only finite time. This source of error leads us to write $\hat{\theta}(\hat{F^c})$ rather than $\theta(\hat{F^c})$. When θ is an expected value, the intrinsic error will have zero mean, but this need not be the case for other parameters.

The two sources of error in (1) lead to complications in finding valid confidence intervals for $\theta(F^c)$ addressed by the delta method (Cheng and Holland 1997) and special versions of the bootstrap (Barton, Lam, and Song 2018). In this paper the focus is on bootstrap methods to characterize the uncertainties caused by extrinsic uncertainty in (1) when the intrinsic uncertainty is negligible. These issues are also present for the bootstrap applied to ordinary simulation output analysis: when the extrinsic uncertainty in (1) is negligible. The key finding is that the performance of any particular bootstrap method depends on the nature of the simulation response, and that two response characteristics that arise in simulation require special care when using bootstrap methods.

Our objective is to construct *valid* (i.e., correct coverage probability) and *tight* (i.e., small) confidence intervals for some characterization of a simulation output statistic, such as a mean or quantile. Bootstrap methods can be effective in this setting, but care must be taken to avoid incorrect application of the bootstrap. Section 2 reviews basic and percentile bootstrap methods and highlights their strengths and weaknesses. The following section identifies two situations in which simulation output poses problems for these methods. Section 4 presents alternative bootstrap methods that can perform well in these situations. Section 5 presents two small computational studies comparing bootstrap method performance. The paper concludes with recommendations on bootstrapping for input uncertainty characterization, and identifies issues that remain open.

2 BASIC AND PERCENTILE BOOTSTRAPPING

The bootstrap has been an important method for characterizing input uncertainty for thirty years (Barton and Schruben 1993). It continues to be an active area of research (Corlu et al. 2020; Barton et al. 2022). Here we focus on bootstrap methods for constructing confidence intervals for some function of a simulation output, typically a run-average statistic Y_r for the r^{th} replication. There are many flavors of the bootstrap for this setting. To describe each, we simplify notation to and write F^c as F and $\hat{F^c}$ as F_n, where F_n is based on a (random) sample of size n.

The bootstrap concept was popularized and characterized in Efron (1979). Consider a parameter estimate $\hat{\theta}(F)$ based on a sample of $X = \{X_1, X_2, \ldots, X_n\}$ drawn from F. The randomness in $\hat{\theta}(F)$ comes from the randomness in $\{X_1, X_2, \ldots, X_n\}$. One might imagine characterizing the distribution of $\hat{\theta}(F)$ by repeated real-world samples X, but there is only observed sample set $x = \{x_1, x_2, \ldots, x_n\}$, and one resulting θ estimate, $\hat{\theta}(F|X = x)$, or just $\hat{\theta}(x)$.

Confidence intervals are constructed by inverting probability statements about $\hat{\theta}(F) - \theta(F)$. The fundamental idea behind bootstrap methods is that the distribution of $\hat{\theta}(F) - \theta(F)$ can be approximated by the distribution of $\hat{\theta}^*(F_n) - \hat{\theta}(x)$ where $\hat{\theta}^*(F_n)$ is the statistic computed from a random bootstrap resample $\{X_1^*, X_2^*, \ldots, X_n^*\}$ drawn from the empirical distribution F_n (usually by drawing samples with replacement from the original sample $x = \{x_1, x_2, \ldots, x_n\}$). And $\hat{\theta}(x)$ is known - it is the value of the statistic computed from the original sample.

The distribution of $\hat{\theta}^*(F_n)$ is often estimated empirically by taking repeated (with index b) so-called *bootstrap* samples $\{X_{b,1}^*, X_{b,2}^*, \ldots, X_{b,n}^*\}$ drawn randomly with replacement from the original sample $x = \{x_1, x_2, \ldots, x_n\}$ and computing an estimate $\hat{\theta}_b^*$ from each bootstrap sample, $b = 1, 2, \ldots, B$. For a perfect characterization of the bootstrap distribution for $\hat{\theta}^*(F_n)$, an analytic form (or $B = \infty$) is required. Typically an analytic form for the distribution $\hat{\theta}^*(F_n)$ is not known, and the empirical distribution of a large number of bootstrap resamples $\{\hat{\theta}_b^*, b = 1, 2, \ldots, B\}$, say $B = 1000$, approximates the true bootstrap distribution of $\hat{\theta}^*(F_n)$. For a *parametric* bootstrap, F_n is a parametric distribution fitted to the n samples. For the *nonparametric* bootstrap, F_n is the empirical distribution function, and the bootstrap samples are drawn from the original sample with replacement. Once we have the approximate distribution for $\hat{\theta}^*(F_n)$ we can construct confidence intervals by inverting probability statements about $\hat{\theta}^*(F_n) - \hat{\theta}(x)$ (or about $\hat{\theta}^*(F_n)$) in the usual way, as shown in the next two sections.

When thinking about bootstrapping in a simulation context, it is important to determine what is the sample and what is the statistic.

For input uncertainty:

- the sample data $\{X_1, X_2, \ldots, X_n\}$ drawn from F are the data used to fit the input distributions, and
- the statistic computed is some function of a set of R simulation replications $\{Y_1, Y_2, \ldots, Y_R\}$ - e.g., mean, variance, or quantile.

For output analysis:

- the sample data $\{Y_1, Y_2, \ldots, Y_R\}$ are drawn from the distribution of outputs having cdf F, and
- the statistic computed is some function of R simulation replications $\{Y_1, Y_2, \ldots, Y_R\}$ - e.g., mean, variance, or quantile.

When examining input uncertainty with both extrinsic and intrinsic error, both n and R are important in the analysis. For output analysis only R is important.

The two simplest and most common bootstrap methods are the *basic bootstrap* and the *percentile bootstrap*. Each of these can be applied in the parametric or nonparametric setting, and the resulting confidence intervals can be based on a standard error estimate or can be based on the empirical distribution of the output statistic.

2.1 The Basic Bootstrap Confidence Interval

Approximation of the distribution of $\hat{\theta}(F) - \theta(F)$ by $\hat{\theta}^*(F_n) - \hat{\theta}(x)$ is used directly in the basic bootstrap. $\mathrm{P}(L \le \hat{\theta}(F) - \theta(F) \le U) \approx \mathrm{P}(L \le \hat{\theta}^*(F_n) - \hat{\theta}(x) \le U)$ permits L and U calculation as $L = q(\hat{\theta}^*(F_n) - \hat{\theta}(x), \alpha/2)$ and $U = q(\hat{\theta}^*(F_n) - \hat{\theta}(x), 1 - (\alpha/2))$, where $q(x, p)$ is the quantile function for data x with probability value p. Using these values for L and U and, since $\hat{\theta}(x)$ is fixed, $q(\hat{\theta}^*(F_n) - \hat{\theta}(x), p) = q(\hat{\theta}^*(F_n), p) - \hat{\theta}(x)$ gives the basic bootstrap $100(1-\alpha/2)$ confidence interval:

$$\mathrm{CI}_{basic} = [2\hat{\theta} - U^*, 2\hat{\theta} - L^*], \tag{2}$$

where L^* and U^* are the quantiles for $\hat{\theta}^*(F_n)$.

2.2 The Percentile Bootstrap Confidence Interval

The percentile interval takes the simpler form:

$$\mathrm{CI}_{percentile} = [L^*, U^*], \tag{3}$$

where the quantiles for $\hat{\theta}^*(F_n)$ are used directly in construction of the confidence interval. The percentile bootstrap appears to be backwards, but Efron and Tibshirani (1994) show that the percentile bootstrap can be valid. First, it is important to understand that if the distribution of the statistic is symmetric, then the values L^* and U^* are interchangeable. When the distribution is not symmetric, both methods can have problems, but less so for the percentile bootstrap. That is due to two properties of the percentile bootstrap. First, the percentile bootstrap has the *range-preserving property*. Because of the way it is constructed, confidence interval bounds are always within the range of the original data. This is not necessarily true for the basic bootstrap. Second, the percentile bootstrap has the *transformation-preserving property*. Suppose that a monotonic transformation h exists that gives a symmetric distribution for $h(\hat{\theta}(F))$ (e.g. the logarithm for a lognormally distributed $\hat{\theta}(F)$). One expects better performance on the transformed statistic. Back-transforming to establish a percentile bootstrap confidence interval for $\hat{\theta}(F)$ from the percentile bootstrap interval for $h(\hat{\theta}(F))$ results in the same interval that would be found without transformation. Thus the percentile bootstrap interval has some robustness to statistics having asymmetric distributions, as we will see in Section 5.

3 LIMITATIONS OF THE BASIC AND PERCENTILE BOOTSTRAP METHODS

The basic bootstrap is sensitive to deviations from a symmetrically distributed sample statistic, resulting in bias in the approximating distributions and confidence intervals with incorrect coverage. In spite of the flexibility provided by the range-preserving and transformation-preserving properties, bias can remain for the percentile bootstrap. Further, the range-preserving property can lead to difficulties in some simulation settings. In this paper the focus is on two characteristics that one might find in simulation output statistics: asymmetry and boundedness of the bootstrap distribution. Each of these are discussed below. When the bootstrap is used to characterize input uncertainty with both intrinsic and extrinsic uncertainty present, bootstrap methods require special modifications that are not covered here. See for example, (Barton et al. 2014) and (Barton et al. 2018).

3.1 Statistics with Asymmetric Distributions

Simulations of queues under high utilization produce asymmetrically distributed output statistics. Bootstrap-based confidence intervals for output analysis must be chosen with this in mind. For example, Figure 1 shows the distribution of daily waiting time for the five-teller bank simulation in Law (2015), implemented in `simmer` (Ucar 2022). The figure is based on a sample of 20,000 average daily waiting times, and the mean and .95-quantiles are indicated by vertical lines. Computing a confidence interval for daily average waiting time is problematic for the basic bootstrap when the sample size is small. Computing a confidence interval for a quantile can be problematic for the basic bootstrap with samples of size 200. We revisit this example in Section 5.

3.2 Bounded Statistics

In a digital-twin simulation one might expect that the real system has been designed for approximately optimal performance. Consider for example the service queuing network represented by the Jackson network in Figure 2. Jobs of different types arrive at the input nodes, and are routed to processing stations that have different processing capabilities. In this simple example one might assume that the route assignments have been made to balance the load, approximately minimizing mean sojourn time. Of course the balance will not likely be perfect. The arrival rates λ_i and service rates μ_j have been estimated from finite samples and

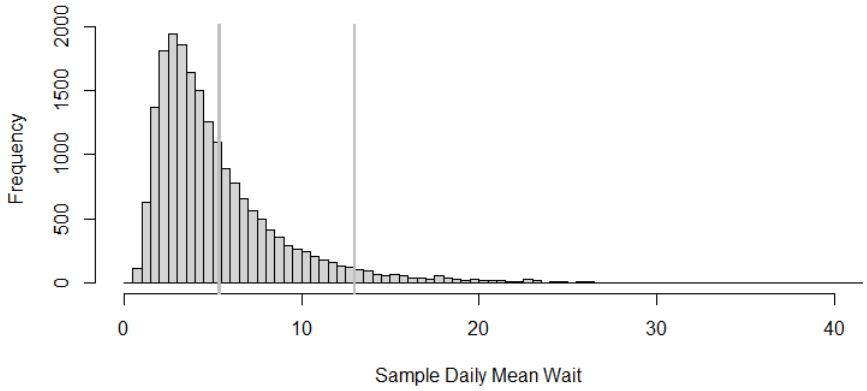

Figure 1: Daily average waiting times for the 5-teller bank.

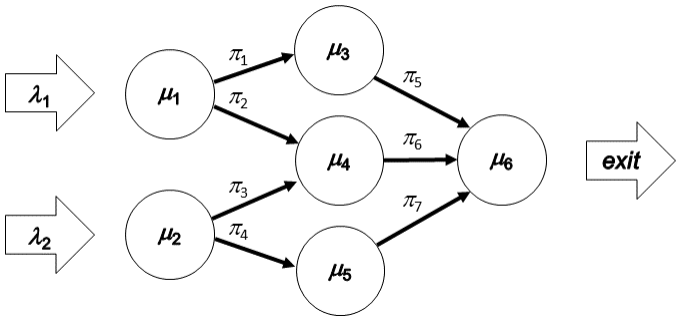

Figure 2: A service network digital twin.

so have sampling error. In this situation, bootstrapping for input uncertainty characterization is unlikely to generate a system response lower than the result from the model (nearly) optimized to the available data. This leads to difficulties in constructing input uncertainty confidence intervals for mean sojourn time. We consider a similar but simpler example in Section 5.

4 ALTERNATIVE BOOTSTRAP METHODS

The shortcomings of the basic and percentile bootstrap methods have been recognized almost since their development. Many enhancements have appeared in the literature. The books Davison (1997) and Efron and Tibshirani (1994) describe many of these. Here we focus on the BCa, double bootstrap, and subsample methods. They have arisen for different reasons and so are discussed separately in the next two sections.

4.1 Bootstrap Methods for Statistics with Asymmetric Distributions

Asymmetry in the distribution of the sample statistic leads to bias in determining bootstrap confidence intervals, and undercoverage for both asymmetric and bounded statistics. The BCa and double bootstrap methods have been shown to be effective in estimating and removing or reducing bias. Each has its shortcomings.

4.1.1 The BCa Bootstrap

Efron's *bias corrected and accelerated* (BCa) bootstrap adjusts the p-values of the bootstrap quantiles to compensate for bias. The form is:

$$\text{CI}_{percentile} = [q(\hat{\theta}^*(F_n), \alpha_1), q(\hat{\theta}^*(F_n), \alpha_2)]. \tag{4}$$

The complication is in how the adjusted p-values α_1 and α_2 are determined. Assume that a transformation exists, say $\phi = h(\theta)$ so that $h(\hat{\theta}) \sim N(\phi - w\sigma(\phi), \sigma^2(\phi))$ with $\sigma(\phi) = 1 + a\phi$. Then w characterizes skewness and a characterizes heteroskedasticity. For a nominal quantile probability α the BCa adjustment uses

$$\tilde{\alpha} = \Phi\left(w + \frac{w + z_\alpha}{1 - a(w + z_\alpha)}\right) \tag{5}$$

where the bias is estimated as $w = \#(\hat{\theta}_b^* \leq \hat{\theta})/(n+1)$, and a is estimated as one sixth the standardized skewness using a linear approximation to the maximum likelihood estimate for $\hat{\theta}$. See Section 5.1 of (Davison 1997) for the detailed derivation. The advantage is an asymptotically accurate adjustment for bias and heteroscedasticity in $\hat{\theta}^*(F_n)$. The disadvantage is that small samples can lead to poor estimates of the variance and skewness and resulting in $\alpha_1 < 1/(n+1)$ or $\alpha_2 > n/(n+1)$, i.e. outside the range of observed bootstrap statistics, with a deleterious impact on coverage.

4.1.2 The Double Bootstrap

The double bootstrap provides a different method to correct for bias (Davison 1997). Originally requiring a pivot statistic, Shi (1992) showed a version that can be applied when an estimate of variance is not available to construct a pivot, further illustrated by McCullough and Vinod (1998). Shi's double bootstrap also adjusts the p-values for the $\hat{\theta}^*(F_n)$ quantiles using the form in (4), but the computations for α_1 and α_2 do not require pivot construction, nor a normal approximation, nor do they require estimating variance and skewness. Further, the computations are somewhat easier to understand. The method adjusts the confidence interval quantile α_1 and α_2 values by bootstrapping each bootstrapped sample K times for each outer bootstrap sample b. The revised α_1- and α_2-values are based on Q_b, the fraction of the K inner-bootstrap statistic estimates from the b^{th} bootstrap resample $\{X_{b,1}^*, X_{b,2}^*, \ldots, X_{b,n}^*\}$ that are less than or equal to the bootstrap statistic estimate:

$$Q_b = \frac{1}{K}\#\left(\hat{\theta}_{bk}^{**} \leq \hat{\theta}_b^*\right). \tag{6}$$

The revised α_1- and (α_2)-values are computed as $\alpha_1^* = q(Q_b, \alpha/2)$ and $(\alpha_2)^* = q(Q_b, (1-\alpha/2))$, where $q(x,p)$ is the quantile function for data x with probability value p. Letson and McCullough (1998) discussed optimal selection of B and K, and gave one example with $B = 1999$ and $K = 246$. For use in input uncertainty characterization, the double bootstrap as defined above may not be practical: BK simulations (491,754 for the case just given!) are required as opposed to only $2B$ for the BCa method, not counting replications. The effort per resample would be greatly reduced for metamodel-resampling input uncertainty (Barton, Nelson, and Xie 2014). An extension of Lam's 'cheap bootstrap' might be applied in some direct-resampling cases (Lam 2022). Unfortunately, its application in the case of bounded or highly skewed sample statistics is problematic, particularly when a pivot is not available. When a pivot statistic is available and only intrinsic uncertainty must be characterized, there are other ways to reduce the computational effort (Davison 1997). The added computation is less concerning for simulation output analysis, since no new simulations need be run.

4.1.3 Software for Computing BCa and Double Bootstrap Confidence Intervals

Functions for basic, percentile and BCa bootstrap confidence interval calculations are available in the `boot` package in `R`. Software for double bootstrap confidence interval calculation is not readily available, but Davison (1997) give example R code that uses the `boot` function from the `boot` package.

4.2 Bootstrap Methods for Statistics with Bounded Values

The basic and percentile bootstrap methods can fail to be consistent for statistics estimating bounded parameters or statistics that are bounded themselves. Davison (1997) present a frequently cited example, constructing a confidence interval for θ using a sample $X \sim U(0,\theta)$. Andrews (2000) identify many such examples from the literature, including their own: $X \sim N(\theta,1); \theta \geq 0$. They present four methods that are effective in this setting. Here we focus on one that is well-researched: subsampling. The double bootstrap can also be effective in this setting.

4.2.1 The Subsample Bootstrap

Subsampling was defined by Politis and Romano (1994) and its asymptotic properties evaluated in Politis, Romano, and Wolf (2001). This method takes repeated resamples $\{X_1^*, X_2^*, \ldots, X_m^*\}$ from X where $m < n$. The confidence interval calculations are analogous to (2) and (3) but the resulting statistics and distributions must be adjusted for the difference in sample size. While the asymptotic characteristics are demonstrated for a subsample of all $\binom{n}{m}$ subsets of size m, they later show the results hold asymptotically for a random sample of B subsets of $\{X_1^*, X_2^*, \ldots, X_m^*\}$ with or without replacement. Lee (1999) develops further enhancements using the double bootstrap that we do not explore here. Asymptotic consistency requires that $B \to \infty$, $n \to \infty$ and $m/n \to 0$. Yet the latter does not seem critical since jackknife estimates have good asymptotic properties and correspond to $m = n-1$. In the computational examples below, large m (e.g., $m = .9n$) gave the best coverage.

Software for m_out_of_n bootstrap confidence interval calculation is available in the `distillery` package in `R`.

5 COMPUTATIONAL EXAMPLES

In this section we compare the performance of these bootstrap methods on two representative examples. The first example applies bootstrap methods to output analysis, where the output (average waiting time) exhibits the high asymmetry arising from a congested queueing system. The second example applies bootstrap methods to input uncertainty in a case with negligible intrinsic uncertainty. It imitates the time-in-system for an example like that shown in Figure 2, where the system is designed to minimize the time-in-system for values that are near the estimated parameter values. Bootstrapping for input uncertainty results in deviations from these parameters and so almost always poorer performance.

5.1 Bank Waiting Time Simulation

We examine the distribution of daily average waiting time for a simple bank simulation described in Law (2015). A detailed example of the simulation model built in `simmer` is described in the tutorial Web pages Garmonsway (2022a) and Garmonsway (2022b). The version used here is a simple case, with only one entity type: *customer*, and all customers have identical characteristics. There is only one trajectory for all customers, unlike a real bank, where some customers would go to the safety deposit box, some to a cubicle to apply for a loan with a loan officer, etc.. Only one resource type is employed: *teller*. The number of tellers can be chosen, but all tellers are identical in their characteristics. The simulation runs for one day, with the bank doors open for 8 hours. There may be customers waiting and in service at closing time. In order to process these, no simulation end time is specified, but the generation of arrivals ends at time 480 minutes (= 8 hours). Effectively, the bank doors close, but tellers continue to service those already inside. The simulation ends when there are no events in the future events list.

We retain the interarrival time and service time characteristics used in the tutorial. The model description below shows that interarrival and service times have exponential distributions with rates 1 and 1/4.5 respectively, suggesting that utilization will be higher than 1 unless there are 5 tellers or more. For

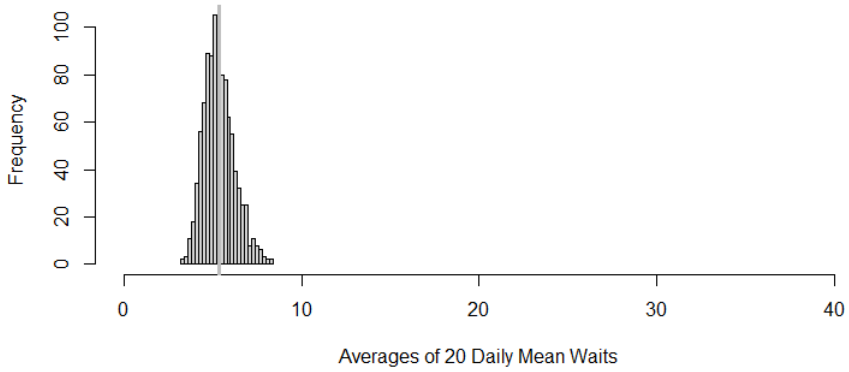

Figure 3: Distribution of 20-day averages.

example, with four tellers the combined service rate would be 4/4.5, less than the arrival rate of 1. We study the waiting time (time in system - time in service) averaged over all customers in a day.

5.2 Mean Waiting Time

Suppose that we have a sample of 20 simulated days, and wish to compute a confidence interval for the expected value of the average waiting time. Figure 3 shows a histogram for 1000 such averages over 20 days along with the true mean daily average. The central limit theorem certainly has had an effect; the asymmetry and non-normality are greatly reduced. This is not the situation requiring attention. The standard interval based on the sample standard deviation and the critical t_{19} value works well in this case.

5.3 A Waiting Time Quantile

Now consider the estimation of the 95% quantile of the daily average wait. The distribution of estimated quantiles over 100 samples of 200 days is shown in Figure 4. Confidence intervals for quantiles are more difficult to determine correctly. As for the mean, the usual confidence interval for a quantile θ is based on a normal approximation. The standard error is estimated by $\sqrt{p(1-p)/(n\hat{f}^2)}$ where $\hat{f}$ is an estimate of the density at $\hat{\theta}$.

The estimated coverage and average confidence interval widths are shown in Table 1. The table includes the standard confidence interval for a quantile using the normal approximation.

Table 1: 95% quantile confidence interval coverage and widths.

	Normal	Basic	Percent	BCa	Subsample	Double
Coverage	0.89	.83	.93	.93	.95	.94
Width	4.6	5.5	5.5	5.6	5.7	5.9

The results show inferior coverage of the normal interval and the basic bootstrap. The percentage bootstrap and BCa perform similarly, and are only slightly inferior to the more complex subsample and double bootstraps. The subsample basic bootstrap interval had inferior performance and is not shown. The normal interval has an advantage over the basic bootstrap because it is based on a pivotal statistic, but still suffers from bias. While its interval width is smaller than the others, its coverage is inferior. The other widths are somewhat comparable. Figure 5 shows one of the 100 computations of confidence intervals in this experiment. For this instance, both the normal and basic fail to cover. Their bias is clear. It is clear

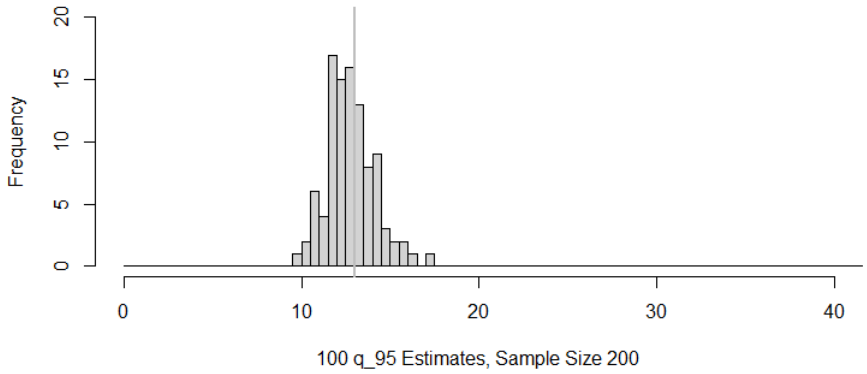

Figure 4: Distribution of 95% sample quantiles, 100 samples of 200 days each.

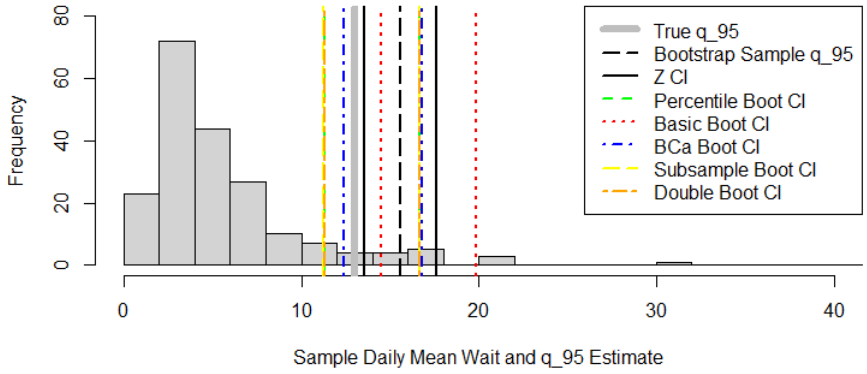

Figure 5: Waiting times and CIs for one real sample of 200 days.

that the least biased intervals for this sample are the subsample and double bootstrap intervals. All interval widths are similar except for the normal (Z) based interval, which is smaller.

Note that while we used quantile confidence interval estimation to illustrate the performance of bootstrap methods for statistics with asymmetric distributions, there are methods superior to the bootstrap that can be used for this problem. See for example the quantile interpolation method (Hettmansperger and Sheather 1986; Nyblom 1992) implemented in the `eqnpar` function in the `EnvStats` package (Millard and Kowarik 2022), described by Höhle (2016).

Next, we compare the performance of these methods for constructing confidence intervals for a bounded statistic.

5.4 A Bounded Statistic

This input uncertainty example is stylized to represent a system whose output parameter of interest is near optimal for the current parameter settings. For this example we assume that intrinsic uncertainty is negligible: for fixed input distribution parameter values, expected values of simulation output are measured without error as $\theta(x) = 100(x-1)^2$. So the simulated system has a minimum performance with input

parameter $x = 1$. The two cases we examine have true input parameter values $x = .60$ and $x = .99$. The input parameters are estimated with normally distributed input uncertainty. The two input uncertainty probability models are $X \sim \mathrm{N}(.60, .2)$ for a true parameter value $x = .60$ and $X \sim \mathrm{N}(.99, .2)$ for a true input parameter value $x = .99$. We considered sample sizes of 50, 100, and 200 and compute the 95% confidence interval for θ by various methods. Tables 2 and 3 show the results. The tables include the standard confidence interval for a mean using the t distribution.

Table 2: Mean confidence interval coverage and widths, E(X)=.60, $\theta = 16$.

	Sample Size	t	Basic	Percent	Subsample	Double
Coverage	50	.65	.92	.93	.94	.93
Width	50	.11	8.8	8.8	9.2	8.9
Coverage	100	.37	.95	.95	.96	.95
Width	100	.08	6.2	6.2	6.5	6.3
Coverage	200	.07	.95	.96	.96	.94
Width	200	.06	4.4	4.4	4.6	4.5

Table 3: Mean confidence interval coverage and widths, E(X)=.99, $\theta = .01$.

	Sample Size	t	Basic	Percent	Subsample	Double
Coverage	50	0.0	.82	.96	.96	.96
Width	50	.11	.69	.69	.74	.71
Coverage	100	0.0	.72	.97	.97	.97
Width	100	.08	.34	.34	.36	.35
Coverage	200	0.0	.71	.96	.97	.96
Width	200	.06	.19	.19	.20	.19

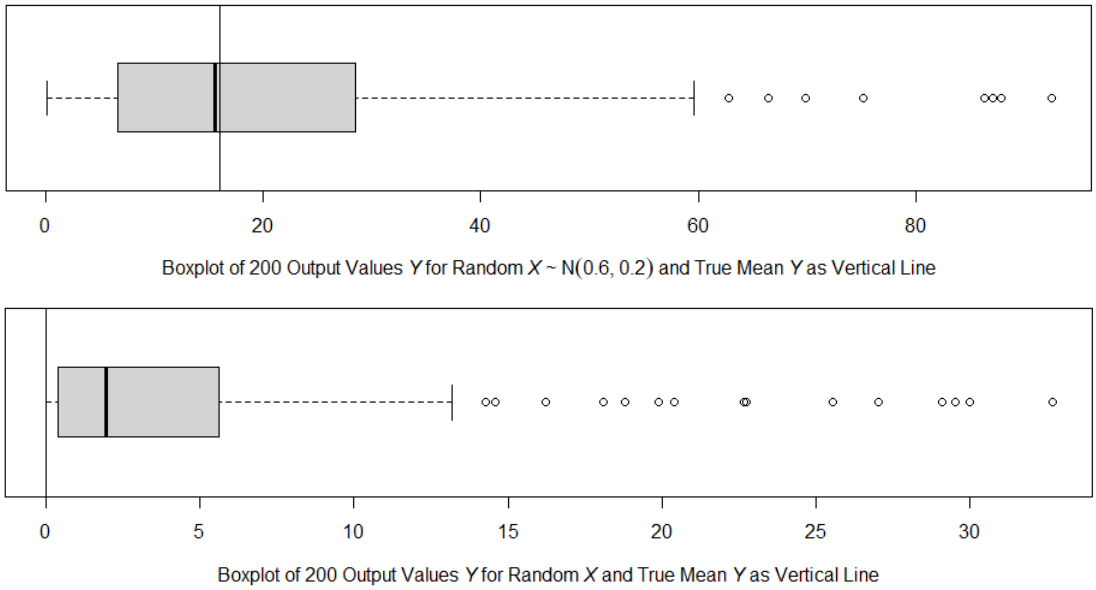

Figure 6: Boxplots for $\hat{\theta}$ values, $x = .60$ (top) and $x = .99$ (bottom).

The subsample basic bootstrap interval had inferior performance and is not shown. The BCa procedure frequently failed to solve for its parameters and so did not return an interval. As in the bank waiting time

example, the percentile, subsample, and double bootstraps all performed well, with similar confidence interval widths. Without compensation for bias, both the standard and basic bootstrap intervals show decreasing coverage with increasing sample size since their confidence interval sizes capture variance but not bias.

Figure 6 shows boxplots for 200 resampled $Y = \hat{\theta}$ values. Both show the extreme bias in values due to the near-optimal E(X) values. The upper plot is for E(X) = .60 ($\theta = 16$), where E(X) is two standard deviations from the value (zero) yielding the minimum (bounded) output statistic. The lower plot for E(X) = .99, very near the value yielding minimum simulation output performance. In this case the bias is extreme, and the coverage of the standard confidence interval is approximately zero.

6 SUMMARY

The nature of simulation output makes the basic bootstrap a risky choice for constructing confidence intervals. Output statistics commonly have asymmetric or bounded distributions, which can make confidence interval construction difficult. The percentile bootstrap worked reasonably well in the two computational examples for asymmetric and bounded statistics. The subsample percentile and double bootstrap methods had slightly better performance. Given the extreme computational cost, the subsample percentile bootstrap appears to be the best choice. To learn more about the bootstrap for confidence intervals in a variety of settings, see (Chernick 2008) and (Mammen and Nandi 2012).

ACKNOWLEDGMENTS

Discussions with Eunhye Song and Henry Lam helped initiate this investigation. We would like to thank the anonymous reviewers for their very helpful corrections and suggestions.

REFERENCES

Andrews, D. W. K. 2000. "Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space". *Econometrica* 68(2):399–405.

Barton, R. R., H. Lam, and E. Song. 2018. "Revisiting Direct Bootstrap Resampling for Input Model Uncertainty". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1635–1645. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., H. Lam, and E. Song. 2022. "Input Uncertainty in Stochastic Simulation". In *The Palgrave Handbook of Operations Research*, edited by S. Salhi and J. Boylan. Cham: Springer International Publishing.

Barton, R. R., B. L. Nelson, and W. Xie. 2014. "Quantifying Input Uncertainty via Simulation Confidence Intervals". *INFORMS Journal on Computing* 26(1):74–87.

Barton, R. R., and L. W. Schruben. 1993. "Uniform and Bootstrap Resampling of Input Distributions". In *Proceedings of the 1993 Winter Simulation Conference*, edited by G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles, 503–508. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Cheng, R. C., and W. Holland. 1997. "Sensitivity of Computer Simulation Experiments to Errors in Input Data". *Journal of Statistical Computation and Simulation* 57(1-4):219–241.

Chernick, M. 2008. *Bootstrap Methods: A Guide for Practitioners and Researchers*. 2nd ed. New York: Wiley.

Corlu, C. G., A. Akcay, and W. Xie. 2020. "Stochastic Simulation under Input Uncertainty: A Review". *Operations Research Perspectives* 7:100162.

Davison, A. C. 1997. *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife". *The Annals of Statistics* 7(1):1–26.

Efron, B., and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.

Garmonsway, D. 2022a. "The Bank Tutorial: Part I". https://r-simmer.org/articles/simmer-04-bank-1.html, accessed 14th April 2023.

Garmonsway, D. 2022b. "The Bank Tutorial: Part II". https://r-simmer.org/articles/simmer-04-bank-1.html, accessed 14th April 2023.

Hettmansperger, T. P., and S. J. Sheather. 1986. "Confidence Intervals Based on Interpolated Order Statistics". *Statistics & Probability Letters* 4(2):75–79.

Höhle, M. 2016. "EnvStats: Package for Environmental Statistics, Including US EPA Guidance version 2.7.0 from CRAN". https://rdrr.io/cran/EnvStats/, accessed 3rd April 2023.

Lam, H. 2022. "A Cheap Bootstrap Method for Fast Inference". arXiv eprint 2202.00090.
Law, A. M. 2015. *Simulation Modeling and Analysis*. 5th ed. New York: McGraw Hill Higher Education.
Lee, S. M. S. 1999. "On a Class of m out of n Bootstrap Confidence Intervals". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 61(4):901–911.
Letson, D., and B. D. McCullough. 1998. "Better Confidence Intervals: The Double Bootstrap with No Pivot". *American Journal of Agricultural Economics* 80(3):552–559.
Mammen, E., and S. Nandi. 2012. "Bootstrap and Resampling". In *Handbook of Computational Statistics: Concepts and Methods*, edited by J. E. Gentle, W. K. Härdle, and Y. Mori, 499–527. Berlin: Springer.
McCullough, B. D., and H. D. Vinod. 1998. "Implementing the Double Bootstrap". *Computational Economics* 12(1):79–95.
Millard, S. P. and Kowarik, A. 2022. "Better Confidence Intervals for Quantiles". url-https://staff.math.su.se/hoehle/blog/2016/10/23/quantileCI.html, accessed 3rd April 2023.
Nyblom, J. 1992. "Note on Interpolated Order Statistics". *Statistics & Probability Letters* 14(2):129–131.
Politis, D. N., and J. P. Romano. 1994. "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions". *The Annals of Statistics* 22(4):2031–2050.
Politis, D. N., J. P. Romano, and M. Wolf. 2001. "On the Asymptotic Theory of Subsampling". *Statistica Sinica* 11(4):1105–1124.
Shi, S. G. 1992. "Accurate and Efficient Double-Bootstrap Confidence Limit Method". *Computational Statistics & Data Analysis* 13(1):21–32.
Ucar, I. 2022. "simmer: Discrete-Event Simulation for R". https://CRAN.R-project.org/package=simmer, accessed 4th February 2023.

AUTHOR BIOGRAPHIES

RUSSELL R. BARTON is Distinguished Professor of Supply Chain and Information Systems in the Smeal College of Business and Professor of Industrial Engineering at the Pennsylvania State University. He recently finished his term as chair of the Computer Simulation Archive Advisory Committee, described at https://d.lib.ncsu.edu/computer-simulation/. His research interests include applications of statistical and simulation methods to system design and to product design, manufacturing and delivery. His email address is rbarton@psu.edu and his homepage is https://sites.psu.edu/russellbarton/.

LUKE A. RHODES-LEADER is a Lecturer in Management Science at Lancaster University. His research interests include applications of simulation optimization and methodological aspects of digital twins. His email address is l.rhodes-leader@lancaster.ac.uk, and his website is https://www.lancaster.ac.uk/lums/people/luke-rhodes-leader.