

PRESENTING A MORE COMPLETE CHARACTERIZATION OF UNCERTAINTY: CAN IT BE DONE?

Russell R. Barton

Dept. of SC&IS
Penn State University
406 Business Building
University Park, PA 16802, U.S.A.

ABSTRACT

Discrete event simulation model output analysis provides characterizations of the uncertainty in the simulation results. This allows decision makers to judge risk in decisions made based on simulation model output. These characterizations, whether they are confidence intervals, variance estimates or quantile estimates, ignore important sources of uncertainty. In particular, results obtained during the model validation phase are usually not incorporated in the uncertainties presented in the usual output analysis. Further, many validation approaches assume that the data used to fit input distributions are complete, and the output analysis does not include uncertainties caused by the use of finite samples to determine input distributions. Bootstrap methods have been used successfully in these settings, yet the formal requirement that the bootstrapped statistic be a continuous function of the data does not hold. Knowing this, it is easy to create scenarios where bootstrapping fails. This paper examines whether this shortcoming can be fixed, and presents one possible strategy.

1 INTRODUCTION

Discrete event simulation models provide a practical way for decision makers to study the behavior of real systems in order to make better decisions for real system configurations or operating policies. Generally the simulation models are inexpensive proxies that can predict real system behavior under alternate scenarios. Because discrete-event simulations mimic stochastic behavior, the decision maker's interest is usually in one or more performance measures that exhibit random variation, both in the real world and in the model.

Confusion occurs when decision makers examine the results of output analysis performed on simulation results. These output analyses present measures of uncertainty in system performance, in the form of confidence intervals, measures of variance, quantiles, or hypothesis test P-values. While output analysis experts acknowledge that

the uncertainties characterize model variation, but exclude model error, it is not clear that the decision maker understands this. Even if the understanding is correct, it is not possible to move from an uncertainty characterization of the behavior of the model to an uncertainty characterization for the real system which they are about to create or change.

This is a challenge for the simulation analysis methodology community. The decision maker surely cares about the performance of the real system, not the model. While it is unlikely that all modeling uncertainties can be characterized, it seems better to capture more if that is possible.

This paper focuses on one part of this effort: characterization of the added uncertainty (or more formally variation) that results from having incomplete knowledge of the input probability distributions that drive the simulation. The incomplete knowledge comes from having only finite samples from the input distributions. The focus here is on bootstrapped empirical samples. For other approaches to this problem see Henderson (2003) and the references therein. The next section shows the fundamental issue. The following section summarizes a bootstrap approach described in Barton and Schruben (1993, 2001), and the problem with applying the bootstrap in this setting. The next section suggests practical strategies to ensure approximately correct confidence interval coverage, illustrated by example. This can serve as a jumping-off point for the discussion and development of other strategies. The final section suggests other areas of work that would provide a fuller characterization of output uncertainty for discrete event simulations.

2 ILLUSTRATIVE EXAMPLE

To understand the problems that occur when finite samples are used to construct input distribution functions, we will focus here on empirical distributions. Similar behavior can be expected when fitting parametric distributions to finite samples. Consider the simple case of a single server queue. Suppose that the true system is a capacitated M/M/1/K queue with $K = 10$, $\lambda = .8$ and $\mu = 1.0$, and sup-

pose that the modeler chooses to estimate a confidence interval for W , the average time in the system, with a set of r simulation runs (of course, this case could be solved analytically: $W \sim 3.797$).

When the modeler knows the actual form of the interarrival and service distributions, then the simulation runs use the true exponential distributions for generating random arrival and service times. The constructed confidence interval for such an experiment, after deleting the initial transient (or beginning in steady-state), would be

$$\bar{W} \pm t_{1-\frac{\alpha}{2}, r-1} S_W / \sqrt{r} \quad (1)$$

where \bar{W} is the average time over all r replications indexed by i and S_W is the sample standard deviation of the W_i values. The values of W_i for an experiment with $r = 10$ are shown in Figure 1. They appear as the set of ten 'x' marks over the horizontal axis point marked as Experiment 1. If this experiment were repeated, i.e., another set of ten runs were made, the resulting values of W_i would be somewhat different. The values for repeated experiments, numbers 2 - 8, are shown in the figure also; all experiments produce similar results. The horizontal line indicates the true value of W , which is approximately 3.8. One would expect $100(1-\alpha)\%$ of such experiments to yield $100(1-\alpha)\%$ confidence intervals that cover the horizontal line.

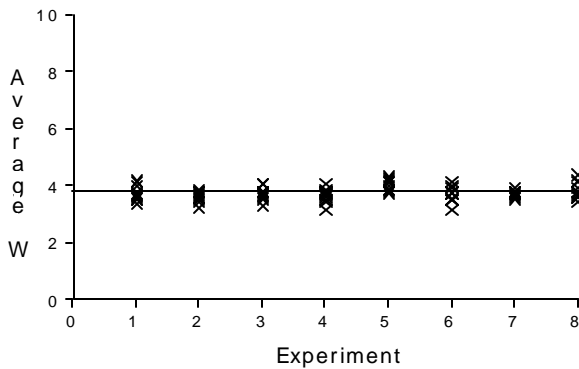


Figure 1. Average Waiting Time for Each of Ten Replications over Eight Experiments – True Distribution

The actual coverage probability of such confidence intervals depends on the distribution of the run-averages, W_i within an experiment. In particular, initialization bias and non-normality of the W_i values cause the coverage probability to deviate from the nominal level of $1-\alpha$. For long simulation runs using the true interarrival and service distributions, the distribution of the W_i values is approximately Gaussian (they are averages of individual delays and so a form of the Central Limit Theorem for dependent variables applies), and the bias approaches zero (by the

Law of Large Numbers), so the error in coverage probability will be small.

Now suppose that the modeler does not know the true arrival and service time distributions, nor even whether they are exponential distributions, and chooses to approximate them by using empirical distributions based on the observed interarrival and service times of n customers. The modeler makes r replicate runs of the simulation, with the intention of estimating W as in the case of the known probability model described above. Figure 2 shows the resulting W_i values for eight such experiments with $n = 50$ and $r = 10$. The consequence of the finite sample sizes is apparent: the coverage probability is greatly reduced, since each experiment exhibits a bias in the distribution of the W_i values. This bias varies randomly from experiment to experiment, and is due to the inaccuracy of the empirical distributions constructed from *finite* random samples of interarrival and service times. The coverage of a 90% confidence interval for this example is approximately 4%, based on the numerical experiments in Barton and Schruben (1993).

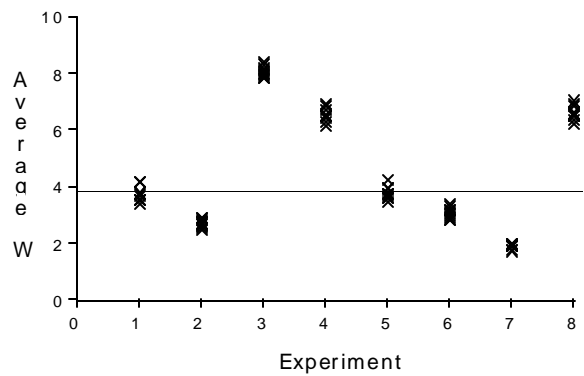


Figure 2. Average Waiting Time for Each of Ten Replications over Eight Experiments – Distribution Estimated by Finite Sample

This phenomenon was discussed by Cheng (1994). He described the two sources of error that arise when using input distributions that are fitted to empirical data. The first is bias error, due to the finiteness of the empirical sample, as seen in Figure 2, and the second as variance error, due to the finiteness of the simulation run length. He proposed a parametric bootstrap approach to estimate both components of variance across simulation runs assuming that parametric input distributions are fitted to finite input data, and the variance-covariance matrix of the estimated parameters is known. Barton and Schruben (1993, 2001) proposed bootstrap methods using smoothed empirical input distributions, and showed that they produced confidence intervals with coverage approaching the correct values for a simple capacitated queueing simulation.

3 EMPIRICAL BOOTSTRAP

There is some confusion about the use of bootstrap resampling of empirical input distributions, in terms of how the simulation runs are conducted, and how the results are analyzed. For details on the process used in this paper, see Barton and Schruben (2001). It is worthwhile highlighting some particular points, however.

First, bootstrap intervals are calculated from empirical percentiles of the output statistic, not t -based intervals using the across-run standard deviation. The reason is that the bootstrap resampling provides an estimate of the distribution function of the statistic one would expect from repeated runs with different input samples of the same size. Increasing the number of replications (that is, bootstrap resamples) provides a better estimate of the *same* distribution, rather than a tighter distribution for the statistic.

Second, the bootstrap resampling should not be done *within* a simulation run, but rather between runs. Resampling within a single run has no impact on the simulation: the resulting cdf for the input values matches the original empirical cdf. Resampling *between* runs does make a difference, however. Each simulation run is conducted with different input distributions. This component of bias described by Cheng (1994) is added to the normal run-to-run variation that occurs due to finite run length. The implication is that, without bootstrap resampling, the simulationist sees artificially small run-to-run variation, and constructs overly optimistic confidence intervals for output parameters. This is why two-step bootstrap resampling should be used: to capture the uncertainty in the predicted system performance due to finiteness of the empirical data used to determine the input distributions.

This combination of two sources of variability affects the validity of the bootstrap approach. Bootstrap methodology assumes that the statistic of interest is computed deterministically from the sample (or resample). That is, if two (re)samples are exactly the same, the computed value of the statistic will be exactly the same. But using the same empirical input distributions, two replications will produce different values for W_i (unless common random numbers are used). One may view this difference as caused by the finiteness of each simulation run. The finiteness of the simulation run length then adds a source of variability (which in conventional output analysis is the only source of variability).

In the experiments in Barton and Schruben (1993, 2001), the variation due to changes in the input distributions overwhelmed run-to-run variation due to finite run lengths, and so the coverage of the bootstrap intervals was not affected. Results presented in Barton et al. (2002) show that finite run lengths can lead to overly conservative bootstrap intervals.

4 PRACTICAL STRATEGIES

Of course, the application of the bootstrap is almost always done in violation of conditions for consistency. Representations of statistical functions on digital computers are not continuous but rather discrete. The discrepancies of the digital representation from the analytical statistical function are very small relative to the changes in values observed from the bootstrap resamples, so that the difference between the analytical calculation and the digital computer calculation is too small to affect the resulting intervals.

How small must the discrepancy from an analytical calculation be? One might expect that if the discrepancy is a small fraction of the confidence interval size that the effect on the coverage would be small. A strategy to check this is to include same-ecdf replications in the bootstrap simulation analysis. The bootstrap statistic would be the average across the same-ecdf replications. The resulting variation in the output statistic can be used to compute an estimate of the standard deviation of the output statistic due to the finite run length of the simulation.

Several metrics might be used to assess whether the bootstrap intervals are compromised by the finite-run-length variation. For example, if the same-ecdf replications show no change in the order statistics across all bootstrap distribution runs, this indicates that the variation due to finite run length is small.

A simpler strategy is to compare the standard deviation of the bootstrap statistic (based on same-ecdf replications) with the size of the bootstrap confidence interval (based on bootstrap statistics across the bootstrap ecdf resamples). This might be done in a number of ways:

1. The standard deviation for the statistic of interest might be computed for each set of same-ecdf replications, and averaged over all of the bootstrap replications.
2. The standard deviations might be computed for the sets of same-ecdf replications for only two bootstrap ecdfs: those with the largest and smallest observed statistics.

If the standard deviation is small relative to the size of the interval, then a Chebychev argument can be used to show that stochastic variation in the output statistic due to finite run length is unlikely to cause large changes in the confidence interval. One might use a coefficient of variation statistic (CV) as a figure of merit, with small values indicating that bootstrap coverage is likely to be approximately correct. What values for such a CV figure of merit might be required?

Figure 3 shows the bootstrap ('B') coverage results from Barton et al. (2002) with a plot of a coefficient of variation metric using the first type of standard deviation. Both the coefficient of variation and the coverage probabil-

ity increase as the run length decreases. While the coverage cannot generally be determined during experimentation, the coefficient of variation is easy to implement.

Without such an evaluation, it is difficult to determine whether short simulation run length is leading to overly conservative bootstrap estimates. The addition of the same-ecdf replications and the use of a CV assessment may be a practical strategy to determine run lengths needed for confidence intervals using bootstrap ecdfs.

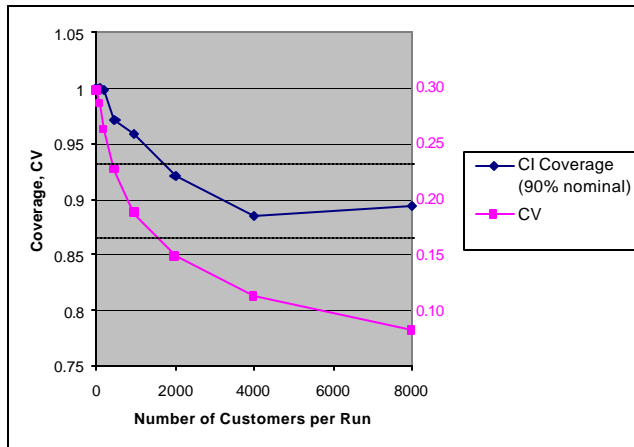


Figure 3. Short Run Lengths Give Large CV and Over-Coverage of Bootstrap Confidence Intervals

5 FUTURE DIRECTIONS

This talk has focused on capturing uncertainties from incomplete characterization of input distributions, with a focus on ecdf-type input distribution representations. An equally important effort is to capture modeling errors that were characterized during the verification and validation activities. A comprehensive overview of these activities is provided by Sargent (2005). Work by Kleijnen and co-authors on bootstrap methods trace-driven validation (Kleijnen, Cheng and Bettonvil 2001) could be extended to permit inclusion of errors in output analysis. There is also some discussion of this issue in Henderson (2003).

ACKNOWLEDGMENTS

The author acknowledges discussions with Lee Schruben and Russell Cheng that helped in the development of the concepts presented here.

REFERENCES

Barton, R. R., R. C. H. Cheng, S. E. Chick, S. G. Henderson, A. M. Law, L. M. Leemis, B. W. Schmeiser, L. W. Schruben, and J. R. Wilson. 2002. Panel on current

issues in simulation input modeling. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 353–369. Piscataway, NJ: IEEE.

Barton, R. R., and L. W. Schruben. 1993. Uniform and bootstrap resampling of empirical distributions. In *Proceedings of the 1993 Winter Simulation Conference* ed. G.W. Evans, M. Mollaghasemi, W. E. Biles, and E. C. Russell, 503-508. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Barton, R. R., and L. W. Schruben. 2001. Resampling methods for input modeling. In *Proceedings of the 2001 Winter Simulation Conference* ed. B. A. Peters, J. S. Smith, D. J. Medeiros and M. W. Rohrer, 372-378. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Cheng, R. C. H. 1994. Selecting input models. In *Proceedings of the 1994 Winter Simulation Conference*, ed. J. D. Tew, S. Manivannan, D. A. Sadowski and A. F. Seila, 184-191. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Henderson, S. G. 2003. Input model uncertainty: why do we care and what should we do about it? *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 90-100. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Kleijnen, J. P. C., R. C. H. Cheng and B. Bettonvil. 2001. Validation of trace-driven simulation models: bootstrapped tests. *Management Science* 47, 1533-1538.

Sargent, R.G. 2005. Verification and validation of simulation models. In *Proceedings of the 2005 Winter Simulation Conference*, ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 130-143. Piscataway, N.J.: Institute of Electronic and Electrical Engineers.

AUTHOR BIOGRAPHY

RUSSELL R. BARTON is a professor of supply chain and information systems at Penn State University. He received a B.S. degree in Electrical Engineering from Princeton and M.S. and Ph.D. degrees in Operations Research from Cornell. Before entering academia, he spent twelve years in industry. He is program chair for the 2007 Winter Simulation Conference. His email address is rbarton@psu.edu.