Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

UNIFORM AND BOOTSTRAP RESAMPLING OF EMPIRICAL DISTRIBUTIONS

Russell R. Barton

Department of Industrial and
Management Systems Engineering
The Pennsylvania State University
University Park, PA 16802

Lee W. Schruben

School of Operations Research and
Industrial Engineering
Cornell University
Ithaca, NY 14853

ABSTRACT

Stochastic simulation models are used to predict the behavior of real systems whose components have random variation. The simulation model generates artificial random quantities based on the nature of the random variation in the real system. Very often, the probability distributions occurring in the real system are unknown, and must be estimated using finite samples. This paper presents two ways to estimate simulation model output errors due to the errors in the empirical distributions used to drive the simulation. These approaches are applied to simulations of the M/M/1 queue with an empirically sampled interarrival time. They capture components of variance in the estimate of mean time in the system that are ignored when the empirical distribution is treated as the true distribution.

1 INTRODUCTION

Stochastic simulation models are used to predict the behavior of real systems whose components have random variation. The simulation model generates artificial random quantities based on probability distributions that represent the nature of the random variation in the real system. In most practical situations, the probability distributions occurring in the real system are unknown, and must be estimated using finite samples. The distribution that is fitted to the observed samples may be either a parametric distribution or an empirical distribution.

Any finite sample leads to a distribution estimate with some error. For some simple models, Shanker and Kelton (1991) find that empirical distributions generally did as well as the best fitted parametric distributions. The nature of the error in the empirical distribution's approximation to the true distribution function is well understood, yet this error is typically ignored in the analysis of simulation output (e.g., in determining confidence intervals for W for an M/M/1 queue) when the output is based on empirical distribution approximations.

This paper presents two ways to estimate simulation model output errors due to errors in the empirical distribution's approximation to the true distribution. First, we argue that traditional empirical cdf estimates are in error, but that the distribution of the error is known. Then we show how the cdf approximation may be varied over multiple simulation runs to capture uncertainties due to the ecdf approximation. Two methods are presented, uniform resampling of the distribution value and bootstrap samples of the empirical cdf (Efron and Tibshirani 1986). These approaches are illustrated using simulations of the M/M/1 queue with an empirically sampled interarrival time.

2 CALCULATING THE EMPIRICAL CDF

The empirical distribution function for a set of data may specified in a number of ways. We describe the observed data, ordered from smallest to largest, as $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$. For convenience, we consider the approximation as a two step process: i) estimate the cdf at the observed values, say $\hat{F}(x_{(1)})$, $\hat{F}(x_{(2)})$, ..., $\hat{F}(x_{(n)})$, and ii) approximate the cdf between the observed values. Several common approximations for the first step are:

$$\begin{aligned} &\text{a)} \quad \hat{F}(x_{(i)}) = i/n, \\ &\text{b)} \quad \hat{F}(x_{(i)}) = i/n+1, \text{ and} \\ &\text{c)} \quad \hat{F}(x_{(i)}) = (i-.5)/n. \end{aligned} \tag{1}$$

Choosing among these approximations (or others) amounts to choosing the probability integrals assigned to

the intervals 1: $-\infty < x \le x_{(1)}$, 2: $x_{(1)} < x \le x_{(2)}$, , n+1: $x_{(n)} \le x < \infty$. For the approximations above, (1a) assigns 1/n to intervals 1, 2, ..., n and 0 to interval n+1, (1b) assigns 1/(n+1) to every interval, and (1c) assigns .5/n to intervals 1 and n+1 and 1/n to the other intervals. We will use 1b in the examples below. The choice is not critical to the results that are presented, however.

Given estimates for the CDF at the observed values $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$, there are many choices for extending the pointwise approximation to a complete approximation, including linear interpolation and kernel smoothing. The results of our discussions below will hold for many of these, but we will use the linear interpolation method discussed in Banks and Carson (1984). Given estimates for the cdf, $\hat{F}(x_{(1)})$, $\hat{F}(x_{(2)})$, ..., $\hat{F}(x_{(n)})$, the estimated value of F for other x values is determined by:

$$\hat{F}(x) = \alpha \hat{F}(x_{(\lambda)}) + (1-\alpha) \hat{F}(x_{(\lambda+1)}) \qquad (2)$$

where $x_{(\lambda)} = \max(x_{(i)} \mid x_{(i)} \le x)$ and $\alpha = (x_{(\lambda+1)}-x)/(x_{(\lambda+1)}-x_{(\lambda)})$. The linear interpolation method requires two artificial points, $x_{(0)}$ and $x_{(n+1)}$ as upper and lower bounds on the distribution, with $\hat{F}(x_{(0)}) = 0$, $\hat{F}(x_{(n+1)}) = 1$. This approach is used in the M/M/1 example below, with $x_{(0)} = 0$ and $x_{(n+1)} = x_{(n)} + (x_{(n)} - x_{(n-1)})$. This approximation strategy is illustrated in Figure 1.

3 ERRORS IN THE CDF APPROXIMATION

The pointwise approximations in (1) are consistent estimators of F, but for any finite sample size n, they approximate the true cdf with some error. Simulations that use approximations based on (1) and (2) in place of the true but unknown F will produce results that are in error for two reasons:

e_1) F may have been (incorrectly) discretized to values $x_{(1)}, x_{(2)}, ..., x_{(n)}$.

e_2) The values for $\hat{F}(x)$ only approximate the values of the true F at points $x_{(1)}, x_{(2)}, ..., x_{(n)}$.

This paper discusses ways to estimate the error in simulation model outputs that are caused by input distribution errors of type e_2, but not e_1. Presumably, using the linear approximation in (2) helps to reduce errors of type e_1.

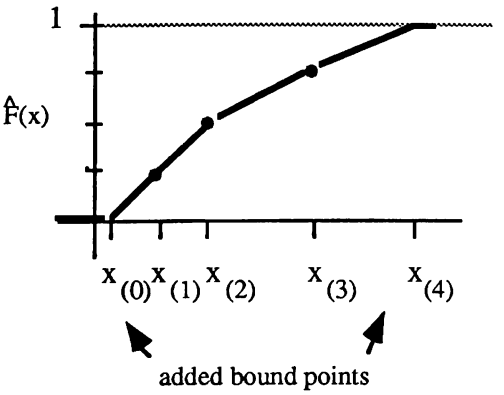

Figure 1: Interpolation-Based Empirical Distribution.

4 RESAMPLING FOR $\hat{F}(X_{(k)})$, k=1, ..., n

The joint distribution of (X, F(X)) is a multivariate distribution with all of the probability mass is concentrated on the line (x, F(x)). To generate random quantities from the distribution described by F, we generate a U(0,1) value and find the conditional distribution of X | F, which is a degenerate distribution that identifies a single value. When the true F is unknown, we cannot do this. Suppose we have a sample of values $x_{(1)}, x_{(2)}, ..., x_{(n)}$ from a distribution F. Given the corresponding distribution values $F(x_{(1)}), ..., F(x_{(n)})$, one could develop a cdf approximation without any error of type e_2. The typical approach in simulation studies is to approximate F at these points by one of the formulas in (1). The resulting approximations are fixed for the duration of the simulation, and extended via one of the two techniques in (2) or some other technique. *The resulting simulation results do not include variability due to uncertainty in the F estimate arising from the finite sample.*

A correct statistical analysis of simulation output should include an assessment of the errors that result from using finite sample estimates for probability distributions used in the simulation. This could be done by collecting several samples of data, and performing the simulation study separately for each data set. The results could then be combined in an Analysis of Variance that included the sample (on which the probability models were based) as a random effect. A standard mixed-effects model could be used to identify whether the sample-to-sample differences in the distribution estimates produce significant variations in the simulation output.

While correct, this approach is a costly one. One

might argue that instead of making say, five replications with five different empirical samples, one could produce a more accurate simulation by combining all five sets of empirical data to fit the required simulation distributions. In the remainder of this section, two methods are presented which allow the use of the combined sample to estimate distributions, yet still provide data that allows a mixed effects ANOVA.

4.1 Bootstrap Resampling

We propose two alternate methods to generate random quantities that includes the uncertainty in $\hat{F}(\bullet)$ due to errors of type e_2. They are based on sampling $\hat{F}(x_{(k)})$ values from an appropriate distribution, rather than using the same values for $\hat{F}(x_{(k)})$ for each simulation run.

The bootstrap technique (Efron and Tibshirani, 1986) has been used to characterize the sampling distribution of complex statistics. For simulating random quantities, one might think of estimating the cdf by sampling, with replacement, k values from the observed set $\{x_{(1)}, x_{(2)}, ..., x_{(n)}.\}$ and assigning distribution values based on (1a), say. Typically, k = n for bootstrap samples. If k –> ∞, then the corresponding probabilities will converge a.s. to i/n for each $\hat{F}(x_{(i)})$ by the Strong Law of Large Numbers.

In a Monte Carlo experiment, then, the uniform resampling strategy is implemented as follows:

i) Sample n values from the set $\{x_{(1)}, x_{(2)}, ..., x_{(n)}.\}$ with replacement. Call these values $v_{(1)}, v_{(2)}, ..., v_{(n)}$.

ii) Use (1) and (2) to compute an empirical cdf based on the v sample.

iii) Repeat this process t times and run the simulation with each of the resulting distribution(s) to collect t bootstrap (re)samples of the simulation process. Perform a mixed effects Analysis of Variance as described above.

(3)

The bootstrap technique can be applied in situations where very little is known about the underlying distributions. The technique below takes advantage of the known form of the sampled cdf.

4.2 Uniform Resampling

While the true $F(x_{(k)})$ values are unknown in the typical simulation model, their joint distribution is known. If X_i are i.i.d. with cdf $F(\bullet)$ then $F(X_i) \sim U(0,1)$. The joint distribution of $F(X_{(1)}), ..., F(X_{(n)})$ corresponds to that of n order statistics of from a uniform distribution, and the marginal distribution for $F(X_{(k)})$ is beta(k, n-k+1) (Lehmann, 1975, p.344). Thus, $F(x_{(1)}), ..., F(x_{(n)})$ can be thought of as a sample from a multivariate uniform distribution. The standard approach to choosing the $\hat{F}(x_{(k)})$ values based on an empirical sample is to choose a consistent estimator for the *expected value* of $F(X_{(k)})$ for the estimate $\hat{F}(x_{(k)})$, for example k/(n+1). Instead for the uniform resampling approach, the value chosen to estimate $F(x_{(k)})$ will be a *sample from the distribution* of $F(X_{(k)})$. In contrast, bootstrap resampling can be thought of as a *sample from the empirical distribution* of $F(X_{(k)})$.

In a Monte Carlo experiment then, the uniform resampling strategy is implemented as follows:

i) Sample n values from a U(0,1) distribution, with ordered values $u_{(1)}, ..., u_{(n)}$. These can be sampled and sorted, or the order statistics themselves can be sampled directly from the appropriate beta distribution.

ii) Set $\hat{F}(x_{(1)}) = u_{(1)}$, and $\hat{F}(x_{(2)}) = u_{(2)}$, and so forth. This assigns new probabilities to the break points in the empirical distribution function.

iii) Repeat this process t times and run the simulation with each of the resulting distribution(s) to collect t uniform (re)samples of the simulation process. Perform a mixed effects Analysis of Variance as described above.

(4)

5 UNIFORM RESAMPLING VS. BOOTSTRAP RESAMPLING

How do these resampling estimates differ? The bootstrap estimate has

$$\text{Prob}(F_b(X_{(i)}) \le m/n) = F_{\text{bin}(n,\, i/n)}(m),$$

while the uniform resampling estimate has

$$\text{Prob}(F_u(X_{(i)}) \le m/n) = F_{\beta(i,\, n+1-i)}\,(m/n),$$

where F_{bin} and F_β refer to binomial and beta cdfs, respectively. For large n, i, m such that m, i, n-m all > 25, normal approximations give:

$$F_{bin(n,\, i/n)}(m) \approx F_{N(0,1)}((m-i)/\sqrt{(i\,(1-(i/n)))} \text{ and}$$
$$F_{bin(n,\, 1-(m/n))}(n-i) \approx F_{N(0,1)}((m-i)/\sqrt{(m(1-(m/n)))}.$$

Since $((m-i)/\sqrt{(i\ (1-(i/n)))} \neq ((m-i)/\sqrt{(m(1-(m/n)))}$, unless m=i, the bootstrap and uniform samples for $\hat{F}(X_{(i)})$ will have different distributions.

6 AN EMPIRICAL INVESTIGATION

Consider the variability of a performance estimate for the simulation of a queueing system, say W: the average time in the system per customer. We wish to estimate the variability in the estimate for W that is caused by using a finite sample empirical distribution in place of the true probability distribution function. One can consider making t sets of r simulation runs, each set with a different $\hat{F}(\bullet)$ based a new sample from the true distribution. The resulting n = t*r runs produce estimates for W with r replications for each of the t resampled ecdfs. As described in Section 4, the Analysis of Variance could be used to estimate the components of variance, modeling the ecdf resample as a random effect. The hypothesis test for no significant ecdf effect could also be employed. Similar analyses could be performed with one long run for each of the t empirical distribution samples and r batches within each run, computing the estimates for W from batch means (Seila 1990).

Alternatively, the empirical distribution could be artificially resampled using (3) or (4). Again, r replications could be run for each resampled empirical distribution, or r batches could be constructed from a single run for each empirical sample.

The simple example considered here shows that the variation in simulation output due to empirical sampling can be significant. This is true even for relatively large (100 samples) empirical distributions. In this example, the M/M/1 queue is modeled with traffic intensity .8. The service time is sampled directly from an exponential distribution with rate $\mu = 1.0$. We consider several cases for sampling the interarrival time. In each case, the interarrival distribution is based on an empirical sample of size 10 or 100 from an exponential distribution with rate $\lambda = 0.8$. The experiment consists of 10 replications (runs) with 10 batches of size 1000 in each replication.

The distribution that is used from run to run for the first case is just the original sampled ecdf. For the other three cases, the second through tenth runs use modified ecdf's, in an attempt to capture the variability in the estimate of W that is introduced by the use of a finite-sample ecdf.

Case A. The ecdf is not changed from run to run .
Case B. The ecdf is resampled from the exponential (0.8) distribution.
Case C. The ecdf F values are sampled from the U(0,1) distribution as in (4).
Case D. The ecdf F values are determined by bootstrap resampling as in (3).

Case A represents the typical fashion in which empirical distributions are used in simulation, i.e. there is no attempt to estimate the variability introduced by the finite sample ecdf. Case B estimates this variation correctly by directly resampling the ecdf from the true distribution for each run. In actual applications, it may not be practical to collect these additional samples. Cases C and D are attempts to capture the variability without requiring additional samples from the true population. The experiment frame is described as follows.

Traffic Intensity:	.8
Batch Size:	1000
Number of Batches:	10
Number of Resamples:	10
Strategy for Incorporating ECDF Error:	none, true resample, uniform, bootstrap
Number of Data Sampled for the ECDF:	10, 100

Figure 2 shows the batch means estimates for W for each resample. Figure 2A shows the assumed relationship when the empirical distribution is used as the true distribution. The actual uncertainty in W is much larger for an ecdf based on ten samples, however, as shown in Figure 2B. Figures 2C and 2D show that the uniform resampling and bootstrap methods produce variations similar to the results for true resampling.

Figure 3 shows the batch means for each run when the empirical distribution is based on 100 samples. Figure 3B shows that the variation in the batch mean estimates for W due to the finite sample ecdf is reduced, but still nontrivial. The empirical distribution that is used for resampling strategies (3) and (4) is the first one in Figure 3B, which produces unusually low batch means. So it is not unexpected that the resampled values in Figures 3C and 3D do not reproduce the full variation

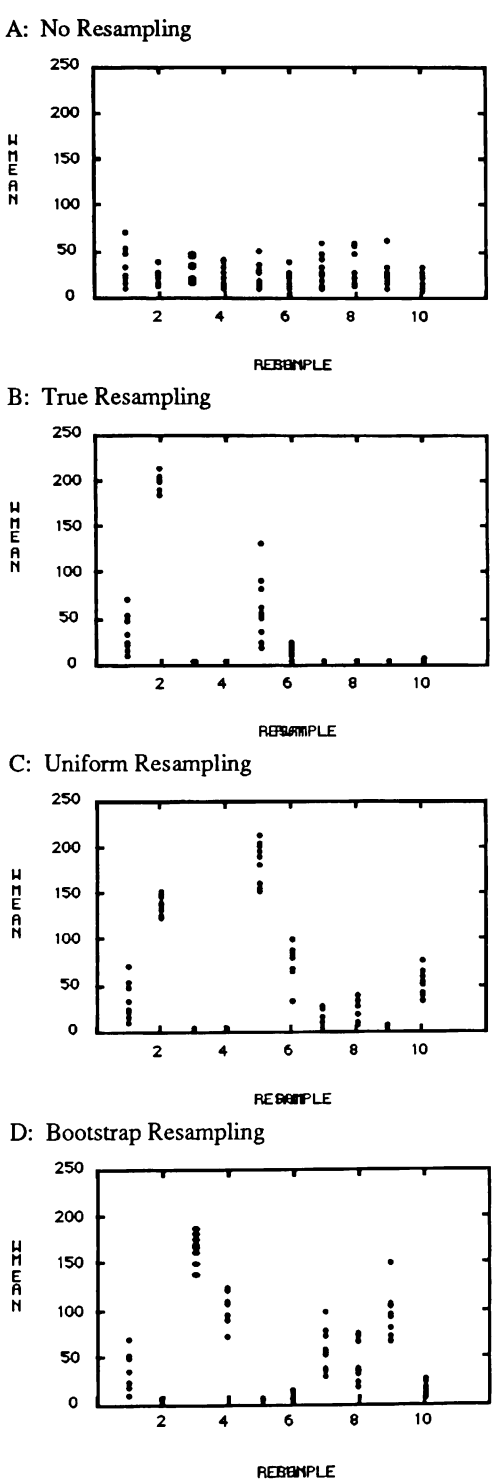

Figure 2: Batch Means vs Resamples, Using 10 Sample ecdf.

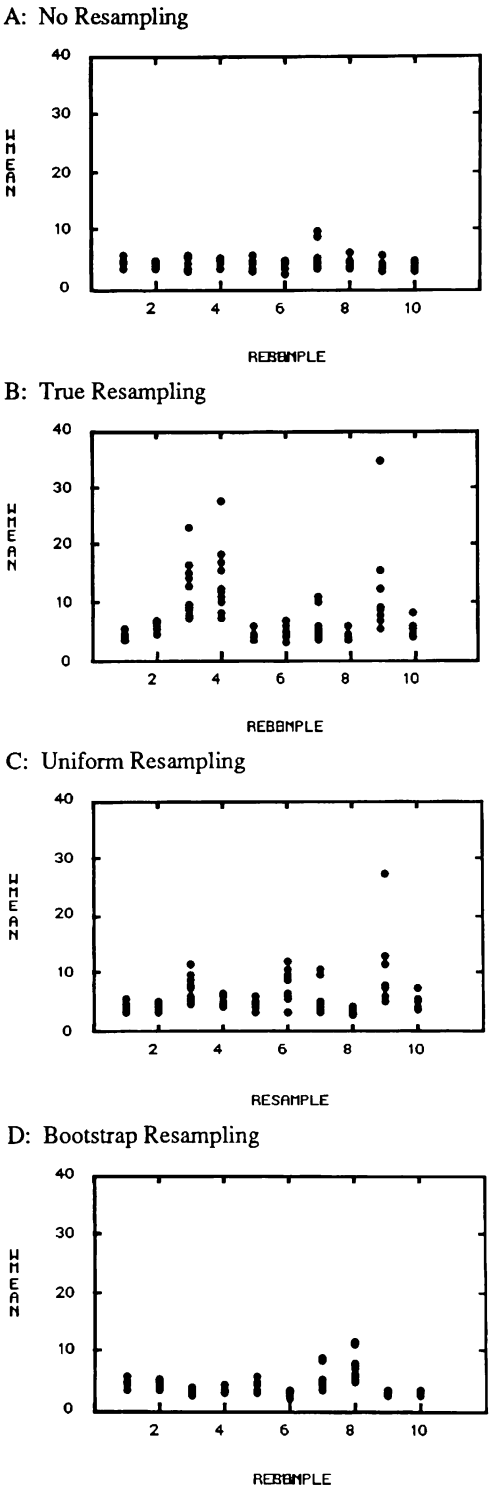

Figure 3: Batch Means vs Resamples, Using 100 Sample ecdf.

seen in the later samples in Figure 2b. In spite of this unusual base sample, the uniform resampling strategy produces a substantial run to run variation.

7 SUMMARY

The run to run variations from the simple example above show that, even for large (100 sample) empirical distributions, the distribution sampling error can have an effect on parameter estimates that is more significant than the errors due to the finiteness of the simulation runs. Ideally, several empirical samples should be taken and a mixed effects analysis of variance conducted to estimate the size and significance of the empirical cdf random effect. The uniform and bootstrap resampling methods provide an inexpensive resampling methods that give indications of the true estimation error. This provides an improvement over the standard approach, which assumes the ecdf to be a true representation of the probability law.

ACKNOWLEDGMENTS

The authors thank Professor Barry L. Nelson of The Ohio State University for his helpful comments.

REFERENCES

Banks, J. and J. S. Carson. 1984. *Discrete Event System Simulation.* Englewood Cliffs, NJ: Prentice-Hall.

Efron, B. and R. Tibshirani 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54-77.

Lehmann, E. L. 1975. *Nonparametrics: Statistics Based on Ranks*. San Francisco: Holden-Day.

Seila, A. F. 1990. Output analysis for simulation. *Proceedings of the 1990 Winter Simulation Conference* (O. Balci, R.P. Sadowski, R.E. Nance, eds.), 49-54.

Shanker, A. and Kelton, W. D. 1991. Empirical input distributions: an alternative to standard input distributions in simulation modeling. *Proceedings of the 1991 Winter Simulation Conference* (B. L. Nelson, W. D. Kelton, and G. M. Clark (eds.), 978-985.

AUTHOR BIOGRAPHIES

RUSSELL R. BARTON is an Associate Professor in the Department of Industrial and Management Systems Engineering at The Pennsylvania State University. He received a B.S. in electrical engineering from Princeton University in 1973 and a Ph.D. in operations research from Cornell University in 1978. His current research interests include graphical methods for experiment design, optimization and design of experiments for simulation models, and statistical issues in modeling manufacturing yield.

LEE W. SCHRUBEN is a Professor in the School of Operations Research and Industrial Engineering at Cornell University. He received his undergraduate degree in Engineering from Cornell University and a Masters degree from the University of North Carolina. His Ph.D. is from Yale University. His research interests are in the design and analysis of large scale simulation experiments. He is a principal developer of the SIGMA simulation system, an event-graph based simulation software package. Three of his papers have received outstanding publication awards from the TIMS College on Simulation and the Chemical Division of the American Society for Quality Control.