Proceedings of the 1991 Winter Simulation Conference
Barry L. Nelson, W. David Kelton, Gordon M. Clark (eds.)

EMPIRICAL INPUT DISTRIBUTIONS: AN ALTERNATIVE TO STANDARD INPUT DISTRIBUTIONS IN SIMULATION MODELING

Aarti Shanker
W. David Kelton

Department of Operations and Management Science
Carlson School of Management
University of Minnesota
Minneapolis, Minnesota 55455

ABSTRACT

We investigate the effect of input-distribution specification on the validity of output from simple queueing models. In particular, the use of various kinds of empirical distributions for approximating service-time distributions is studied.

1 INTRODUCTION

Among the various sources of error that make the output from a simulation less valid is modeling the "wrong" distribution for various input quantities like arrival times of jobs to a job shop, service or repair times of machines, etc. A commonly encountered problem in simulation modeling is the specification of a suitable input distribution for the observed data. These data are specific realizations of some underlying distribution that can be regarded as the "true" distribution. A prevalent practice is to approximate this true distribution by a fitted distribution from a standard family (e.g., exponential, uniform, etc.). In many situations, this approximation may not adequately represent the observed data, and may introduce significant error in the input that may adversely affect the validity of the output.

For example, consider a "real" single-server system with lognormal ($\mu = -0.35$, $\sigma^2 = 0.69$) interarrival times (which has mean 1 and variance 1) and exponential service times (with mean 0.9). The system performance was simulated for 250 delays in queue with empty-and-idle starting conditions. The true expected average delay in queue based on 100 replications for this simulation is 6.1 time units. If, instead of lognormal, one erroneously used exponential interarrival times (with the same mean and variance as the lognormal distribution), and the service-time distribution as before, the expected average delay in queue for this system is 5.3 time units. Even though the first two moments of the true service-time distribution were matched by the chosen approximation, an error of 13% was introduced due to wrong distribution selection.

In simulation modeling, the true input distribution is approximated using either a standard distribution or some nonstandard (e.g., empirical) distribution. Any estimator of the true distribution should be able to generate variates beyond the range of the observed data, i.e., beyond the smallest and the largest of the observations. This makes the estimator ***generalizable*** because the range of the estimation is not dependent on that particular realization of the observed data vector. It also should have the capability of modeling a variety of distributional shapes, i.e., it should be ***flexible***. Flexibility is of great importance since it improves the quality of the fit by approximating the true distribution as closely as possible.

In general, there are three methods of specifying an input distribution:

1. Use a "standard" parametric distribution: These include distributions such as uniform, exponential, Weibull, etc., that have a known (closed-form or otherwise) cumulative distribution function (CDF). Such standard distributions are generalizable but not necessarily flexible.

2. Use an empirical distribution: Here the observed data themselves are used in some way to form a distribution function. Actual values of the individual observations or grouped data can be used to come up with a distribution function. There are several ways of constructing such distribution functions. For example, define

$$F(x) = \frac{i}{n} \text{ if } X_{(i-1)} < x \leq X_{(i)}$$

where $X_{(1)}, \ldots, X_{(n)}$ are the order statistics of the observed data (Bratley, Fox, and Schrage 1987, p. 150). Most empirical distribution functions apply only to the observed data, i.e., the variates generated in this way cannot take on values beyond the smallest and the largest of the observations (Law and Kelton 1991, pp. 350-353). In this case, the distribution is, by design, flexible but not generalizable.

3. Use a flexible parametric family: Such a parametric family supplies a flexible distribution function that is an approximation of the true distribution function (Johnson 1949). The distribution function does allow variate generation beyond the observed data (as in 1). This alternative can be viewed as a compromise between the first two as it is perhaps both generalizable and flexible.

Another important aspect of distribution selection is the ease with which one can generate variates from the specified cumulative distribution function (CDF). There are many exact algorithms for variate generation. Some of these fall into the classes generally known as inverse-transform, composition, convolution and acceptance-rejection (Law and Kelton 1991, pp. 465-484). Among all these methods, the method of inverse transform proves to be of great advantage because it can be used to facilitate and strengthen many variance-reduction techniques (Bratley, Fox, and Schrage 1987, pp. 44-59).

The most widely used mode of input-distribution specification, fitting from the standard families of distributions, has convenient forms for variate generation. Some of these have a closed-form CDF that can be easily inverted for variate generation (like exponential and Weibull), while some that do not enjoy such a privilege need methods like acceptance-rejection for their variate generation. But in either case, it is easy to generate the desired variates. Although variate generation may not pose a problem, these standard families may not fit the observed data well, thereby creating an error in the specification of the input distribution that may propagate through the system into the output.Hence the concerns in selection of input distributions are the ease of variate generation and the quality of the fit of the chosen distribution to the observed data.

The goal of this research is to analyze some empirical "black-box" methods existing in the literature for modeling input distributions, as an alternative to fitting standard distributions. The motivation for considering these automatic techniques is as follows:

1. These methods are designed to eliminate any judgement on the part of the analyst (which may be required in fitting standard distributions) that may lead to an erroneous decision.

2. The CDFs given by these methods have a simple inverse, and hence variate generation can be done using the inverse-transform method.

3. The distributional forms of these methods are flexible, thereby improving the possibility of a good fit.

2 EMPIRICAL DISTRIBUTIONS

The problem of input-distribution selection is inherent to simulation modeling, and has been discussed in the simulation literature for some time. In this section, we review and qualitatively contrast empirical distributions with standard distributions.

In describing these empirical distributions, we assume that $X_{n\times 1}$ is a random vector observed by the simulation analyst. We also assume that the support of these distributions is $[0,\infty)$.

The simulation literature discusses some empirical distributions for the purpose of using them as input distributions for simulation models. Among these empirical distributions is the "unadorned," which fits a piecewise-linear CDF to the entire n-vector. Defining $X_{(0)} = 0$, the CDF is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{i}{n} + \frac{x - X_{(i)}}{n(X_{(i+1)} - X_{(i)})} & \text{if } X_{(i)} \le x \le X_{(i+1)} \\ & i = 0, \ldots, n-1 \\ 1 & \text{if } x > X_{(n)} \end{cases}$$

This distribution truncates the right tail of the true distribution. A remedy suggested by Bratley, Fox and Schrage (1987, pp. 150-151) fits a piecewise-linear CDF to the smallest $n-k$ data points and an exponential tail to the remaining k points. This distribution is referred to as the *mixed empirical-exponential* distribution. Once again defining $X_{(0)} = 0$, the CDF for this distribution is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{i}{n} + \frac{x - X_{(i)}}{n(X_{(i+1)} - X_{(i)})} & \text{if } X_{(i)} \le x \le X_{(i+1)} \\ & i = 0, \ldots, n-k-1 \\ 1 - \frac{k}{n} e^{[\frac{-(x - X_{(n-k)})}{\theta}]} & \text{if } x > X_{(n-k)} \end{cases}$$

where k is the number of observations used to fit the tail $(1 \le k \le n-1)$, and

$$\theta = \frac{X_{(n-k)}/2 + \sum_{i=n-k+1}^{n} (X_{(i)} - X_{(n-k)})}{k}$$

Since the empirical distributions are formed using the data vector they are completely flexible, i.e., they change their form according to the data. But they are not always generalizable since their variate generation cannot produce variates outside the range of the observed data. An exception to this rule is the mixed empirical-exponential distribution, which can generate variates beyond $X_{(n)}$ due to its exponential tail. As discussed in the introduction, many standard distributions do not have this problem of generalizability but are seldom flexible.

The existing literature on distribution specification discusses empirical and flexible parametric distributions but does not address the implication of using ill-fitting distributions in simulation modeling, which is what a simulation analyst would be interested in. Moreover, the fits of these estimated distributions are not compared to the existing practices of fitting "standard" distributions or to the available alternatives, so that practitioners get an idea of the gains (in terms of precision and accuracy) in employing such alternatives. Thus, the approximation methods suggested have to be studied in greater detail, and in the context of some practical simulation applications.

3 ANALYSIS AND RESULTS

As mentioned earlier, input-distribution selection plays an important role because the validity of the output is of paramount importance. Hence, the approximation, however "close" to the true distribution, has to be conducive to valid and feasible output.

3.1 Propagation of Input Error

In this section, we analyze and compare the quality of the approximation of the true distribution, i.e., we study the impact of error on the output introduced by an unsuitable approximation of the input. We perform this analysis using empirical and standard distribution-specification methods. We also look at ease of implementation of these methods for simulation models.

These analyses need to be performed in the context of an application. For initial experimentation, we have chosen the $M/G/1$ queueing system. In this queueing system, the arrivals to the queue follow a Poisson process with rate λ, and the service times are from a general (unspecified continuous) distribution. In this study, the general service-time distribution has been given four known forms:

1. Exponential(μ): $\mu = 1.0$.
2. 2-Erlang(μ): $\mu = 0.5$.
3. Gamma(α, β): $\alpha = 3.5, \beta = 0.286$.
4. Weibull(α, β): $\alpha = 2.0, \beta = 1.128$.

The Poisson arrival process has rate $\lambda = 0.8$.

The effect of the chosen approximation on the output can be studied for a variety of models. The reason for selecting the $M/G/1$ system was due to the existence of simple analytical expressions for some desired output measures like steady-state waiting time in queue, given by

$$W_Q = \frac{\lambda E[X^2]}{2(1 - \lambda E[X])}$$

where $E[X]$ and $E[X^2]$ are the first and second raw moments of the general service-time distribution. This expression for the waiting time is known as the *Pollaczek-Khintchine* (P-K) formula (Ross 1989, p. 376).

To assess the goodness of the specification method, the general service-time distribution is given a known form, for which the required moments are known and thus W_Q is easily computed from the P-K formula. Now, one could get estimates of this output measure using the first two moments of the approximating distribution. Thus, the effect of the specified distribution on the output can be checked.

The comparison of the approximating methods was done as follows:

1. Compute the first and second moments of the known distributions, which will provide an exact expression for W_Q.
2. Generate a sample of size n (we took $n = 10$, 30, 50, and 100 in this study) from one of the known distributions. Pretend that the sample generated is the observed sample, i.e., its true distributional form is unknown. Then use the approximating methods on the sample (one at a time) to estimate the observed CDF. One could then compute the first and second moments of these alternative methods of specifying the service-time distribution either numerically or analytically.

In the case of empirical distributions, the first two raw moments are defined as a function of the observed data. Since an empirical distribution is formed using the observed data alone, the CDF and the moments will obviously be functions of the data vector, i.e., they will be *conditional* on the data.

For the unadorned empirical distribution, the conditional raw moments are given by

$$E[X \mid X_{(1)}, \ldots, X_{(n)}] = \bar{X} - \frac{X_{(n)}}{2n}$$

$$E[X^2 \mid X_{(1)}, \ldots, X_{(n)}] = \frac{1}{3n}[2\sum_{i=1}^{n-1} X_{(i)}^2 + \sum_{i=1}^{n-1} X_{(i)}X_{(i+1)} + X_{(n)}^2]$$

and the moments of the mixed empirical-exponential conditioned on the observed data are (Bratley, Fox and Schrage 1987, pp. 150-151),

$$E[X \mid X_{(1)}, \ldots, X_{(n)}] = \bar{X}$$

which is an unbiased estimator of the first moment, given the data, and

$$E[X^2 \mid X_{(1)}, \ldots, X_{(n)}] = \frac{1}{3n}[2\sum_{i=1}^{n-k-1} X_{(i)}^2 + \sum_{i=1}^{n-k-1} X_{(i)}X_{(i+1)} + X_{(n-k)}^2] + \frac{k}{n}[(\theta + X_{(n-k)})^2 + \theta^2]$$

For most standard distributions, the first two moments can be easily computed assuming that the distribution was specified via maximum-likelihood estimators (MLEs).

In our study, from each sample, we obtained moment estimates for each distribution-specification method. Using these estimated moments, an estimate of the required output measure, namely the P-K formula, was computed. The performance measures were the variance of the estimate and the bias in the estimate. These two measures can be combined into one single measure, the mean square error (MSE). However, we did not use this composite measure alone since one needs to understand the precision and the accuracy of these estimated separately because, in the context of this problem, accuracy may be more important than precision.

We replicated the W_Q-estimation procedure 100 times. Some approximating distributions generated negative W_Q in some replications. Those replications were discarded since the estimates were inadmissible. So for a fair comparison, the experiment was continued until we obtained 100 replicates where all the methods generated feasible W_Q's. This gave us yet another measure of performance, i.e., the total number of replications required by each specified distribution to obtain 100 "good" ones. This measure demonstrates the ability of a method to produce feasible waiting times in queue.

The results are presented in Figures 1-8. In this study, we have considered six standard distributions and six empirical distributions. The mixed empirical-exponential distribution has five different forms based on k, the number of points in the tail. In our study, k was taken as 10%, 20%, ..., 50% of the data points.

Each graph presents information on both the variance and the bias. The horizontal axis has all the estimated distributions, and the vertical axis represents both the standard deviation and the bias of the estimates. The graphs also have the root mean square error showing the combined effect of the two criteria. Although we ran the experiments for four different "observed" sample sizes, we have only included graphs for sample sizes 30 and 100 since they were representative of the other cases.

Among the specified distributions, the unadorned distribution performs well in almost all cases when MSE is the criterion. The reason for its superiority is the low variance in estimating W_Q. In terms of the bias, even though the unadorned distribution is not always better than the rest, the bias is not large in any case. The mixed empirical-exponential distribution is also consistent in its behavior. It appears to be less biased than the unadorned but is less precise. Its performance does not seem to be affected by the number of points in the tail, except when exponential is the true distribution. In that case, it seems to improve with more points in the tail, which is to be expected. In comparing the performance of the empirical distributions to that of their standard counterpart, the most striking observation is that the empirical distributions are consistent. They may not be the most accurate and precise in their estimation in all cases but their performance is consistently good. Some standard distributions like gamma and Weibull show some promise in their estimation ability but others like lognormal, triangular and uniform do not appear to be reliable alternatives. For example, when gamma was the true distribution, the variance and the bias in W_Q under the triangular distribution became worse with increasing sample size, which is counter-intuitive. This suggests a serious problem of inconsistent behavior.

3.2 Ease of Implementation

The empirical distributions considered in this study are easy to understand and implement. Variate generation is done using the inverse-transform method, which involves only a few steps. The CDFs of these distributions are simple to invert. The CDF inverse of the unadorned empirical distribution can be obtained in a single line of code, since the piecewise linear function is fitted to the entire observed vector. Obtaining a realization of this empirical CDF is simple:

```
Generate $U \sim U[0,1]$
Set $i = \lfloor nU \rfloor$
```

■ Bias ● Standard deviation – – – RMSE

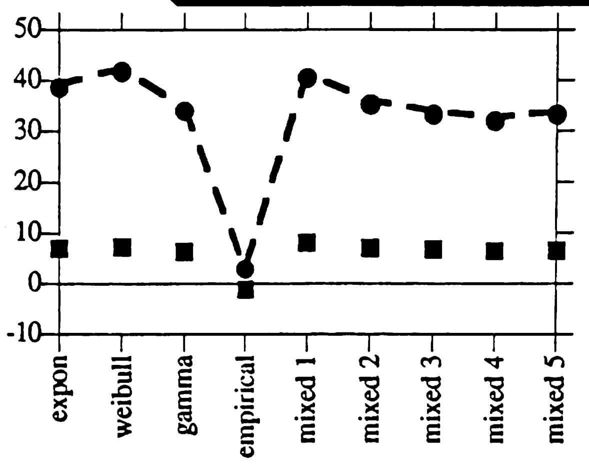

Figure 1: G =Exponential(1.0), n = 30

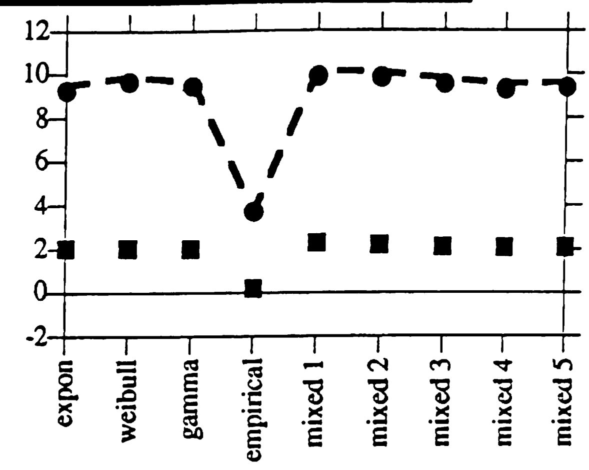

Figure 2: G =Exponential(1.0), n = 100

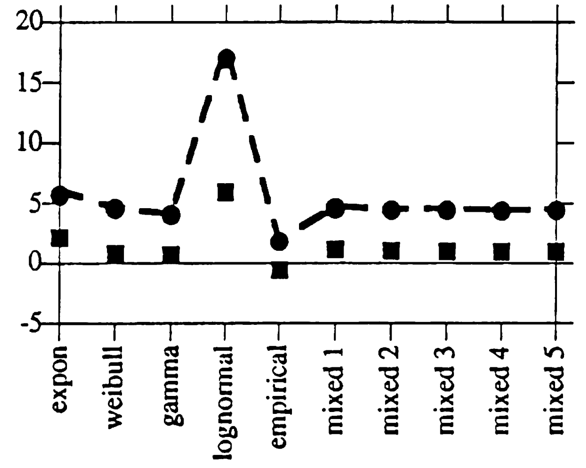

Figure 3: G =2-Erlang(0.5), n = 30

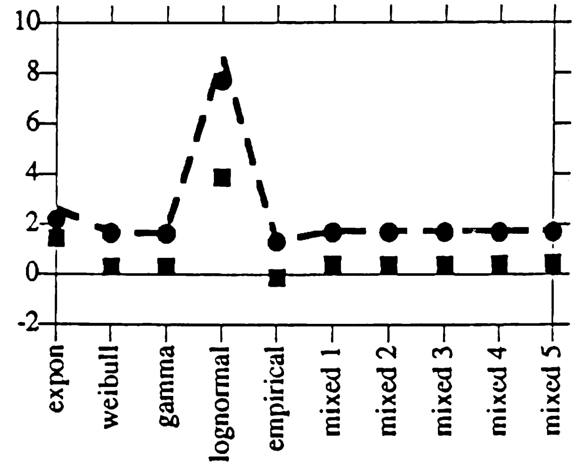

Figure 4: G =2-Erlang(0.5), n = 100

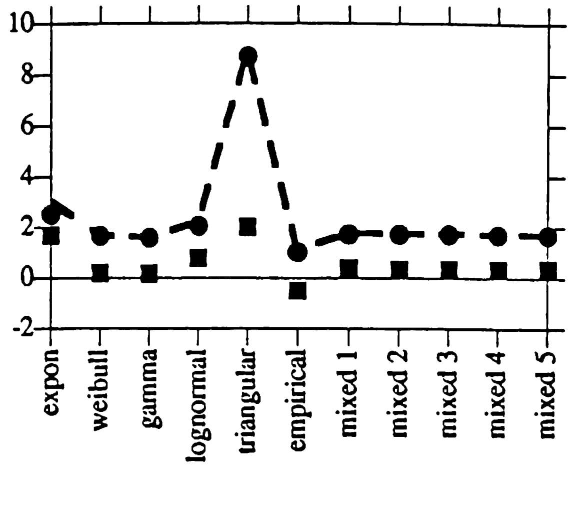

Figure 5: G =Gamma(3.5,0.286), $n = 30$

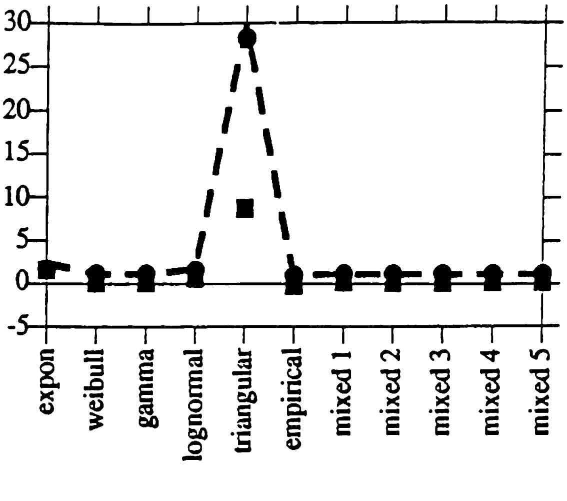

Figure 6: G =Gamma(3.5,0.286), $n = 100$

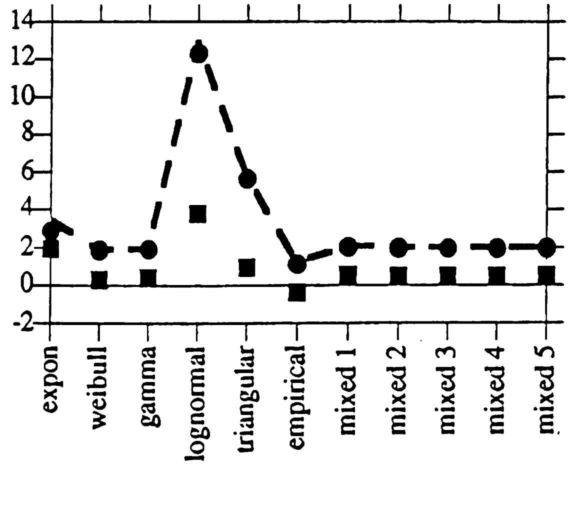

Figure 7: G =Weibull(2.0,1.128), $n = 30$

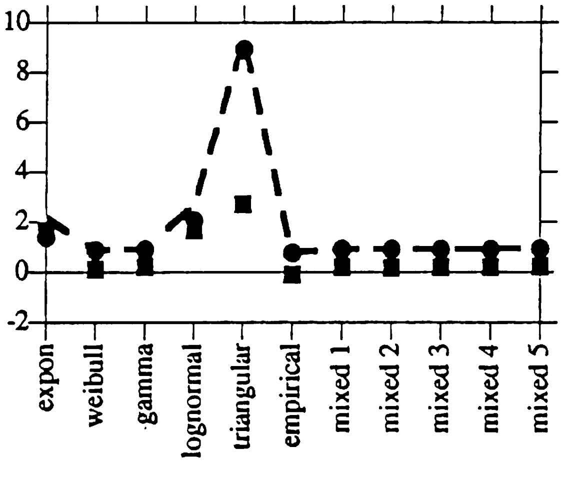

Figure 8: G =Weibull(2.0,1.128), $n = 100$

```
Return $X = n(U - \frac{i}{n})(X_{(i+1)} - X_{(i)}) + X_{(i)}$
```

The mixed empirical-exponential distribution, on the other hand, has a split inverse CDF function, one for the piecewise-linear portion fitted to the first $n - k$ data points, and the other for the shifted exponential fitted to the k largest observations. Bratley, Fox and Schrage (1987, p. 151) supply a two-step algorithm for variate generation:

```
Generate $U \sim U[0,1]$
If $U > 1 - \frac{k}{n}$ then
    Return $X = X_{(n-k)} - \theta \ln(\frac{n(1-U)}{k})$
else
    Set $i = \lfloor nU \rfloor$
    Return $X = n(U - \frac{i}{n})(X_{(i+1)} - X_{(i)}) + X_{(i)}$
```

In terms of computation, one needs to sort the data vector in both the cases, and take a logarithm in the case of the mixed empirical-exponential distribution.

As discussed in the introduction, variate generation for most standard distributions does not pose a problem. Although that is the case, it may not be as simple as in the case of empirical distributions. Some standard distributions do not have invertible CDFs. In such cases, there may be other efficient variate-generation techniques but they may be more involved than the simple inverse-transform method.

4 CONCLUSIONS

When the approximating distributions were compared on the basis of variance and bias in their estimates, the empirical distributions generally did as well as the best fitted standard distributions, and sometimes even better. For example, when Weibull was the true distribution, the fitted Weibull and gamma were the best fitting distributions among the standard distributions, with the least bias and variance. The empirical distributions were a good match where both the criteria were concerned, and in some cases, actually had lower variance and bias.

A point in favor of the empirical distributions is that their performance is consistent. This cannot be said about the standard distributions whose performance quality depends more critically upon the underlying true distribution. This robustness of an approximating method is an important issue in input-distribution specification.

Where the ability to produce feasible W_Q was concerned, most standard distributions had a high acceptance rate, i.e., very few replications were discarded due to infeasibility, except triangular and uniform which had a miserably low acceptance rate for some true distributions. The empirical distributions once again demonstrated a more consistent behavior. The graphs above include only those distributions that had a reasonable acceptance rate.

Another distinguishing factor between the empirical distributions and the standard distributions was the ease of implementation. Getting the moments of the empirical distributions was easy once a sample was generated, while for the standard distributions, computation of parameter estimates was not always easy. For example, in the case of the gamma distribution, the MLE involved using a Newton-Raphson root-finding algorithm. This led to some numerical problems, even when gamma was the true distribution.

5 FUTURE WORK

In order to give a complete treatment to the problem at hand, the following tasks need to be accomplished:

1. The true distributions considered are unimodal and continuous. The alternative specification methods need to be studied under some (not so neat) distributional forms, e.g., ones with multiple modes or some discontinuities. The queueing system studied is a single-server system with Poisson arrivals. The appeal in selecting this system was the simple analytical expression for a desired output measure, namely the waiting time in queue. The $M/G/k$ system also has an approximate expression for the waiting time in queue in terms of the first and second moments of the service-time distribution (Nozaki and Ross 1978). So the error in this output measure can be examined using the estimates of the moments obtained in this study.

2. The performance of these methods has to be investigated with more complicated systems for which no expressions for the output measure exist. For such models, one could simulate the system under study by giving the unknown distributions known forms as in the present pilot study, and get estimates of desired output measures. Then, pretend that the data generated artificially are indeed the observed data, and run the simulation using the approximating methods to get another set of estimates of the same output measure. The two sets of estimates can then be compared. The simulation study will have to be carefully designed to get accurate and precise estimates so that the comparisons have high power.

3. This study considered two extremes of the continuum, namely the empirical distributions and the standard distributions. A compromise between the two, the flexible parametric families, needs to be included as well.

ACKNOWLEDGMENTS

Christopher Bingham and Michael R. Taaffe were very helpful in formulating our approach to this study. We are thankful to the Minnesota Supercomputer Institute for computer support.

REFERENCES

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A Guide to Simulation*, Second Edition. New York: Springer-Verlag.

Johnson, N. L. 1949. Systems of frequency curves generated by methods of translation. *Biometrika* **36**: 149-176.

Law, A. M. and W. D. Kelton. 1991. *Simulation Modeling and Analysis*, Second Edition. New York: McGraw-Hill.

Nozaki, S. and S. M. Ross. 1978. Approximations in finite capacity multiserver queues with Poisson arrivals. *Journal of Applied Probability* **15**: 826-834.

Ross, S. M. 1989. *Introduction to Probability Models*, Fourth Edition. New York: Academic Press.

AUTHOR BIOGRAPHIES

AARTI SHANKER is a Ph.D. candidate in the Department of Operations and Management Science in the Carlson School of Management at the University of Minnesota. Her research interests are in the design and analysis of simulation experiments, and in applied statistics.

W. DAVID KELTON is an Associate Professor in the Department of Operations and Management Science in the Carlson School of Management at the University of Minnesota. His interests are in simulation methods, stochastic-process estimation, and queueing. He is the General Chair for the 1991 Winter Simulation Conference.