INDIRECT CYCLE-TIME QUANTILE ESTIMATION USING THE CORNISH-FISHER EXPANSION

Jennifer M. Bekki[1] John W. Fowler[2*] Gerald T. Mackulak[3] Barry L. Nelson[4]

[1] Department of Industrial Engineering, Arizona State University
Tempe, Arizona, 85287-5906, U.S.A
IIE Student Member

[2] Department of Industrial Engineering, Arizona State University
Tempe, Arizona, 85287-5906, U.S.A

[3] Department of Industrial Engineering, Arizona State University
Tempe, Arizona, 85287-5906, U.S.A
IIE Fellow

[4] Department of Industrial Engineering and Management Sciences, Northwestern University
Evanston, Illinois, 60208-3119, U.S.A
IIE Senior Member

[*] To whom correspondence should be addressed:
e-mail: john.fowler@asu.edu, tel: 1-480-965-3727, fax: 1-480-965-8692

Abstract

This paper proposes a technique for estimating steady-state quantiles from discrete-event simulation models, with particular attention paid to cycle-time quantiles of manufacturing systems. The technique is justified through an extensive empirical study and supported with mathematical analysis. The Cornish-Fisher expansion is used as a basis for this estimation, and it is shown that for an M/M/1 system, a system of 5 tandem M/M/1 queues, and a full factory simulation model, the technique provides precise, accurate estimates for the most commonly estimated quantiles with minimal data storage. The performance of the Cornish-Fisher expansion is compared to both traditional direct quantile estimation using order statistics and indirect quantile estimates obtained from four-parameter distributions. Based on these evaluations, the Cornish-Fisher expansion is found to be the most appealing quantile estimation technique of the three approaches because it provides accurate results for a variety of systems and has the advantages of being easy to implement and having extremely low data-storage requirements. It also provides the capability for estimating all quantiles of a given random variable from a single set of multiple replications at a given design point, negating the need to know in advance which quantile estimates are desired.

1 Introduction

The ability to quote accurate delivery times is crucial in any service-driven industry, including manufacturing. In fact, on-time delivery is the key metric for assessing the overall customer service level of a production facility (Gordon, 1993). An essential component of generating good lead times is producing accurate estimates of cycle time parameters such as the mean cycle time or quantiles of the cycle time distribution. The mean cycle time value provides the decision maker with a general feel for the expected production time, but production managers are typically not measured against "average" values. Instead, they prefer to know the percentage of items produced within a particular cycle time window. Cycle time quantiles provide the decision maker with a complete picture of the cycle time distribution, allowing delivery time quotes to be made with varying levels of confidence.

Discrete-event simulation (DES) models have been commonly used to generate estimates of average cycle time, with research focusing on reducing the simulation run time or on increasing the estimation accuracy. For instance, Roeder *et al.* (2002) and Schruben and Roeder (2003) discuss a resource-driven simulation methodology which executes significantly more quickly than traditional job-driven simulation methodologies. However, even with decreasing run times, there are not yet efficient and easily implemented methods for obtaining precise and accurate estimates of cycle time quantiles. A significant reason for this is that quantiles are often more difficult to estimate than simple means and can require excessive data storage. The data storage requirements have led many industry practitioners to use crude quantile estimation techniques that require fewer computing resources. For example, indirect estimates of cycle time quantiles are often made by collecting the first two sample moments (equivalently sample mean and variance) of the cycle time distribution from a simulation model and assuming the distribution to be normal. These estimates, while extremely easy to generate, do not consider the skewness or kurtosis of the actual cycle time distribution and, therefore, risk being grossly inaccurate for all cycle time distributions in which skewness or kurtosis differ substantially from those of the normal distribution.

We define a *direct quantile estimate* to be one in which the estimate is obtained by inverting an empirical cumulative distribution function (cdf). Order statistics are traditionally used for this purpose. To generate an estimate of a cycle time quantile using order statistics, a DES model is run, and the individual cycle time observations are stored. Once all observations have been collected, the observations are sorted, and the appropriate quantile is selected from the sorted values. For example, if 1000 observations were collected, an estimate of the 0.70 cycle time quantile would be the 700^{th} largest value. Smoothing and interpolation are also regularly used to improve quantile estimates based on order statistics. With a sufficient quantity of data, order statistics provide very accurate and precise results. However, order statistics also present the significant drawback of requiring substantial data storage. Even with rapidly increasing computing power, sorting and storing the millions of samples required to estimate some quantiles is still one of the fundamental issues in quantile estimation (Chen and Kelton, 2001).

To address the data storage issue, Jain and Chlamtac (1985) developed the P^2 algorithm that yields direct quantile estimates without requiring the storage of individual cycle time observations.

However, if estimates of several quantiles of the same variable are required, the efficiency and accuracy of the algorithm degrades. Heidelberger and Lewis (1984) developed the maximum transformation technique for reducing data storage needs when estimating quantiles from dependent, stationary data, but the parameters of their algorithm change based on the desired quantile to estimate, requiring simulations to be run for each quantile.

Jin, Fu, and Xiong (2003) propose a stratified quantile estimator as their own direct quantile estimator and demonstrate that the error probability for this estimator approaches zero with appropriate sample size, but the sample size grows exponentially as the problem dimension increases. Avramadis and Wilson (1998) propose a technique for reducing bias and variance of quantile estimators through the use of correlation induction. Hesterberg and Nelson (1998) exploit control variates with known quantiles to reduce the variance in estimating selected quantiles of a distribution. Their technique showed significant reduction in MSE for extreme quantiles (0.9, 0.95, and 0.99), but has the disadvantage of requiring the use of control variates with known quantiles. Also of note is the fact that both Hesterberg and Nelson (1998) and Avramadis and Wilson (1998) require i.i.d. outputs. Chen and Kelton (2006) developed both the zoom-in algorithm for direct quantile estimation, which uses the concepts of order statistics and ϕ-mixing to provide bounds on the desired quantile at each iteration of the algorithm, and the quasi-independent procedure which collects samples of simulation output data at long lags so that the data are independent as measured by the runs-up test. Similarly, Chen and Kelton (2008) show how to estimate a well-chosen set of quantiles in order to estimate the empirical distribution of a stationary-simulation generated process. Both Chen and Kelton (2006) and Chen and Kelton (2008) are based on order statistics and, therefore, have the desirable quality that their estimation accuracy does not depend on the shape of the distribution from which the quantiles are being estimated. However, both approaches also have the drawback of requiring long run lengths, especially for highly correlated systems, which are more likely than not within a DES model. The average run length for their quasi-independent algorithm, for example, applied to an M/M/1 system at 90% traffic intensity, for example, is 14,792,893 customer waiting times (Chen and Kelton, 2006).

An *indirect quantile estimate* is one which is obtained without building an empirical cdf. Yang et al. (2005), for example, propose such an indirect estimation technique based on the Generalized Gamma Distribution. Indirect estimation techniques have the advantage of reduced data storage requirements, but the estimators may have less precision or less accuracy than direct estimators. An indirect quantile estimation technique that provides high accuracy, low variability, and which is easy to implement would be extremely useful. This paper proposes such a method for indirectly generating steady-state cycle time quantile estimates from DES models of manufacturing systems. The method relies on the assumption that the DES output process, $X_t, t = 1, 2, 3, \ldots$ converges weakly to X, a limiting random variable having continuous support and finite moments of order at least four, thus ensuring that X has well-defined quantiles. This assumption will be approximately valid for many manufacturing simulations. Further, we assume that the raw sample moments, $n^{-1} \sum_{t=1}^{n} X_t^k$, converge almost surely to $\mathrm{E}[X^k]$ as $n \to \infty$ for $k = 1, 2, 3, 4$. A thorough empirical study of the technique is presented, along with some supporting mathematical analysis.

The suggested technique utilizes the first four terms of the Cornish-Fisher expansion (Cornish and Fisher, 1937) and takes into account the first four moments of the cycle-time distribution, allowing accurate and precise quantile estimates to be generated for even extremely non-normal distributions. The remainder of the paper is organized as follows. In the next section, the Cornish-Fisher expansion is discussed in more detail, and an analysis of its ability to approximate quantiles from both the normal and exponential distributions is shown. Section 3 formally presents the suggested steady-state quantile estimation procedure, and Section 4 discusses the sources of bias in the procedure. Section 5 presents experimental results. Results are given for three systems which represent a variety of cycle-time distributions: an M/M/1 system, a system of 5 tandem queues, and a full factory model. Section 6 presents alternative methods for indirectly estimating quantiles and compares these to the estimates obtained using the Cornish-Fisher expansion. Finally, Sections 7 and 8 provide conclusions and areas for future work.

2 Cornish-Fisher expansion

The Cornish-Fisher expansion (CFE) is an asymptotic series used to approximate normalized quantiles from any distribution, given a quantile from the standard normal distribution and the distribution's cumulants (Cornish and Fisher, 1937). When sample cumulants are used in place of theoretical cumulants, the approximation becomes an estimator, and, consequently, it becomes important to have consistent estimators of the cumulants. Equation (1) gives a truncated version of the expansion, including only the first four terms. In this equation, z_α is the α quantile from the standard normal distribution, κ_3 and κ_4 are the third and fourth cumulants of the distribution, σ is the standard deviation, μ is the mean, and x_α is the approximation of the α quantile.

$$x_\alpha = \mu + 1/2(\sigma^2 + 1)z_\alpha + 1/6(z_\alpha^2 - 1)\kappa_3 + 1/24(z_\alpha^3 - 3z_\alpha)\kappa_4 - 1/36(2z_\alpha^3 - 5z_\alpha)\kappa_3^2 \tag{1}$$

Equation (1) can also be written in terms of distributional moments (rather than cumulants). This version of the CFE is shown in Equation (2), in which γ_1 is the standardized central skewness, and γ_2 is the standardized central excess kurtosis. To arrive at Equation (2), the distribution is standardized such that the expansion is in terms of x'_α, where $x'_\alpha = (x_\alpha - \mu)/\sigma$. The standardized mean becomes 0 and the standardized variance becomes 1. As a result, the first term of Equation (1) drops out and the second term becomes only z_α, giving $x'_\alpha = z_\alpha + 1/6(z_\alpha^2 - 1)\kappa_3 + 1/24(z_\alpha^3 - 3z_\alpha)\kappa_4 - 1/36(2z_\alpha^3 - 5z_\alpha)\kappa_4^2$. Then, given the relationships between moments and cumulants ($\mu_2 = \kappa_2$, $\mu_3 = \kappa_3$, and $\mu_4 = \kappa_4 + 3\kappa_2^2$, where μ_i is the i^{th} central moment and κ_i represents the i^{th} cumulant), the cumulants κ_3 and κ_4 can be replaced by the skewness and the excess kurtosis, respectively. Since μ_2 is equal to 1 for a standardized distribution, the moment-cumulant relationships show that $\kappa_4 = \mu_4 - 3$, which is equal to the excess kurtosis. Similarly, μ_3 is equivalent to the skewness. Formulas for calculating the standardized skewness (γ_1) and the standardized excess kurtosis (γ_2) are $\gamma_1 = \mu_3/\mu_2^{3/2}$, and $\gamma_2 = \mu_4/\mu_2^2 - 3$ (Stuart and Ord, 1987).

Therefore, by replacing the cumulants with the standardized skewness and standardized excess kurtosis, the quantile from the standardized distribution is given by $x'_\alpha = z_\alpha + 1/6(z_\alpha^2 - 1)\gamma_1 + 1/24(z_\alpha^3 - 3z_\alpha)\gamma_2 - 1/36(2z_\alpha^3 - 5z_\alpha)\gamma_1^2$. This value must then be transformed back to the quantile from the original unstandardized distribution. After performing the transformation, the final quantile estimate from the sample distribution is found using Equation (2). This equation is equivalent to Equation (1) and is the form of expansion used in this research. In Equation (2), x_α represents the quantile approximation from the standardized distribution, and Y^α represents the quantile approximation from the original sample distribution.

$$Y^{\alpha} = \mu + \sigma x'_{\alpha} \text{ where } x'_{\alpha} = z_{\alpha} + 1/6(z_{\alpha}^2 - 1)\gamma_1 + 1/24(z_{\alpha}^3 - 3z_{\alpha})\gamma_2 - 1/36(2z_{\alpha}^3 - 5z_{\alpha})\gamma_1^2 \quad (2)$$

The principle behind the Cornish-Fisher expansion is that if a set of moments of a true and fitted distribution agree, the quantiles of the fitted distribution can be regarded as an approximation to the quantiles from the true distribution. In the case of the CFE, the quantiles from the fitted distribution are expressed as an asymptotic series, which is a function of the corresponding quantiles of the standard normal distribution. The terms of the expansion are polynomial functions of the corresponding standardized quantile of the normal distribution, and the coefficients are functions of the standardized moments (or cumulants, as shown in Equation 1) of the distribution of interest. The terms are arranged so that for terms beyond the third term, the r^{th} coefficient is of the order $O\left(n^{a-(r/2)}\right)$, where a is a constant and n is a parameter assumed for practical applications to be very large. Because of this property, in a valid application of the CFE, each consecutive term is of smaller and smaller order, resulting in convergence for fixed r. If coefficients up to the r^{th} are used, convergence is guaranteed as n approaches infinity. However, for finite n, the best approximation is not necessarily obtained with larger r; for example, we will show that four terms is best for the exponential distribution. Typically, the CFE is used for estimation of quantiles of a random variable, which is itself a sum of other m random variables, when the sample size m is not so large that the Central Limit Theorem (CLT) makes the distribution normal. In these cases, the CLT guarantees that the distribution of interest is asymptotically normal and that the cumulants ensure that the CFE coefficients meet the $O\left(n^{a-(r/2)}\right)$ property. The CFE is then used to improve the normal approximation. However, while the CFE is justified as n approaches infinity, in practice it is used with finite n, making the CFE actually a moment-based correction for a non-normal random variable. Examples of this type of application include Chan and Cui (2003), who use the CFE to estimate control limits for a standardized control chart of a skewed process distribution, and Miguel and Olave (2002), who use it to obtain quantiles of the distribution representing forecast error. This paper uses the CFE as a moment-based correction to estimate quantiles from the non-normal random variable representing cycle time. No specific assumptions about the cumulants of the cycle time distributions can be made, but we argue that the distribution of the cycle time random variable observed in most realistically sized manufacturing systems will be reasonably close to normal (Rose, 1999 and Mittler *et al.*, 1995), particularly in systems with FIFO dispatching. Consequently, the adjustments made by the CFE to correct for non-normality will not be extreme, and results from experimentation tailored to the type of distributions found in manufacturing systems, shown in Section 5, demonstrate that the approach produces both accurate and precise results.

2.1 Selection of number of terms

Since the CFE is an asymptotic series, it is important to determine the appropriate number of terms to include. Adding additional terms to the CFE requires estimating progressively higher moments, which are, in turn, more and more difficult to estimate since higher moments have increased sensitivity to sampling fluctuation (Stuart and Ord, 1987). Moreover, for the inverse cdf function of the random variable estimated by the CFE to be monotone (which it should be), the order of the expansion must be even (Jaschke, 2002). Therefore, we seek to identify a small, fixed, even number of terms that will provide accurate results while requiring the estimation of the least number of moments from the cycle time distribution.

For a quantile estimator to be useful in estimating cycle time quantiles for manufacturing settings, it should provide accurate and precise estimates for a variety of cycle time distributions. It is important for a quantile estimation technique to have the capability of accurately generating quantile estimates of systems that are at least approximately normally distributed. However, our empirical evidence shows that as the utilization of a manufacturing system approaches 100%, the distribution of the cycle times, even for complicated systems, becomes more skewed. Consequently, the same system at different traffic intensities can have cycle time distributions ranging from approximately normal to more skewed. The ideal quantile estimation technique would be able to produce accurate quantile estimates for both types of distributions.

For any number of terms, the CFE perfectly approximates quantiles from the normal distribution. This is clear when it is observed that all terms in Equation (2), except the z_α term, are zero in such a case. Consequently, to determine how the number of terms used in the expansion affects the accuracy of the quantile approximations for non-normal distributions, the exponential distribution was used. Figure 1 demonstrates the effect that the number of terms used in the expansion has on quantile approximations from an exponential distribution. Theoretical moment values (instead of empirically estimated via simulation) were used to generate this figure. Figure 1 shows that using the first four terms of the expansion clearly yields the best results for the widest variety of quantiles from the exponential distribution. The sixth order expansion also performs well across all quantiles, but requires estimation of the fifth moment, making it a less desirable choice than the four-term expansion. Figure 1 also highlights the following property of the CFE: For a given function, approximations from the CFE do not necessarily improve in accuracy as additional terms are added (Jaschke, 2001). For the exponential distribution, it turns out that the improvements in accuracy stop after four terms. We conjecture that as the distributions get closer to normal, the corrections required by the CFE will be less, and a number of terms that works well for the exponential distribution will likely work well for distributions closer to normal.

[Figure 1 Here]

To further support the selection of four terms, Figure 2 shows the relative percent difference between the true quantiles for an exponential distribution and the approximated quantiles generated from the first four terms of the CFE using the distribution's true moments. For most quantiles of interest to manufacturers, the first four terms of the CFE produce a reasonably accurate approximation for the exponential distribution. This behavior is not entirely surprising, however, since the CFE is known to produce less and less reliable approximations as the probability associated with the quantile being estimated approaches 0 or 1 (Jaschke, 2001). For all quantiles above 0.45, though, the approximated value is within 5% of the theoretical value, and in a manufacturing setting, it is unlikely that estimates of the lower cycle time quantiles would be useful for quoting lead times. Customer delivery dates generated based on an estimate of the 0.3 quantile, for example, would result in only 30% of product being delivered on time.

Also, it should be noted that, for a given quantile, the CFE does not produce approximations at the same level of accuracy for all non-normal distributions. Estimates of the same quantile from two random variables (with nonidentical distributions) could have different levels of accuracy. As the distribution from which quantiles are estimated gets closer to normal, however, the quality of the estimates improves for all quantiles. We use the fact that the exponential distribution is as severely non-normal a cycle time distribution as is likely to be found in manufacturing to represent an outer limit of the behavior that the CFE is likely to produce for cycle time quantile estimation. Accordingly, we assume that a given number of terms that works well for the exponential distribution will also be likely to work well for distributions closer to normal. Based on this assumption and the results of Figures 1 and 2, the selection of four terms is reasonable.

[Figure 2 Here]

3 Quantile estimation procedure

The results presented in the remainder of this paper use the first four terms of the CFE as an indirect steady-state quantile estimator for DES models. To implement the approach, the following procedure is used. Step 1 gives the procedure setup, while Step 2 details the moment estimation process, which takes place within each simulation replication. We use the moment estimates from each replication to obtain a single quantile estimate (Step 3), and we obtain quantile estimates across replications so that they are independent and we can build confidence intervals (Steps 4 – 5). The procedure is based on independent simulation replications, but batch means can also be used (see the comments after the procedure).

The output process for the procedure is defined by $\{x_{ik}, i = 1, 2, \ldots, n, k = 1, 2, \ldots, l\}$, where where i is the replication index, and k is an index representing the number of observations within a given replication. For each i, $x_{i1}, x_{i2}, \ldots$ is a stationary stochastic process; we assume that the user of this procedure will handle the initialization bias independently of this procedure.

1. Select a value for n, the number of independent simulation replications, l, the number of observations in each replication, and m, a variable used for grouping quantile estimates to

form confidence intervals, where $n \geq 10$ and an integer divisible by m. We recommend $m \geq 5$. Suggestions for appropriate simulation run lengths can be found in Whitt (1989), and suggestions for determining the number of replications can be found in Banks, *et al.* (2004) or Law and Kelton (2000).

2. For $i = 1$ to n:

   (a) Run simulation replication i and calculate estimates of the sample mean ($\hat{\mu}_i$), the sample standard deviation, ($\hat{\sigma}_i$), the sample central, standardized skewness ($\hat{\gamma}_{1i}$), and the sample central, standardized, excess kurtosis ($\hat{\gamma}_{2i}$). The formulas used to calculate the sample moments are given in Equations (3)–(6) (Kenny and Keeping, 1954). In these equations, l represents the total number of sample data points, and to minimize data storage requirements, the only values that should be stored during replication i are running totals of x_{ik}, x_{ik}^2, x_{ik}^3, and x_{ik}^4.

$$\hat{\mu}_i = \frac{1}{l}\sum_{k=1}^{l} x_{ik} \tag{3}$$

$$\hat{\sigma}_i^2 = \frac{1}{l-1}\left(\sum_{k=1}^{l} x_{ik}^2 - 2\hat{\mu}_i \sum_{k=1}^{l} x_{ik} + \hat{\mu}_i^2\right) \tag{4}$$

$$\hat{\gamma}_{1i} = \frac{l}{(l-1)(l-2)\hat{\sigma}_i^{3/2}}\left(\sum_{k=1}^{l} x_{ik}^3 - 3\hat{\mu}_i \sum_{k=1}^{l} x_{ik}^2 + 3\hat{\mu}_i^2 \sum_{k=1}^{l} x_{ik} - \hat{\mu}_i^3\right) \tag{5}$$

$$\begin{aligned}\hat{\gamma}_{2i} \;=\; & \frac{l^2}{(l-1)(l-2)(l-3)\hat{\sigma}_i^2}\left[\left(\frac{l+1}{l}\right)\left(\sum_{k=1}^{l} x_{ik}^4 - 4\hat{\mu}_i \sum_{k=1}^{l} x_{ik}^3 + 6\hat{\mu}_i^2 \sum_{k=1}^{l} x_{ik}^2\right.\right. \\ & \left.\left. -\; 4\hat{\mu}_i^3 \sum_{k=1}^{l} x_{ik} + \hat{\mu}_i^4\right) - \frac{3\,(l-1)}{l^2}\left(\sum_{k=1}^{l} x_{ik}^2 - 2\hat{\mu}_i \sum_{k=1}^{l} x_{ik} + \hat{\mu}_i^2\right)^2\right]\end{aligned} \tag{6}$$

(b) Obtain a point estimate of the α quantile, Y_i^α .

$$Y_i^\alpha = \hat{\mu}_i + \hat{\sigma}_i x'_\alpha \text{ where } x'_\alpha = z_\alpha + 1/6(z_\alpha^2 - 1)\hat{\gamma}_{1i} + 1/24(z_\alpha^3 - 3z_\alpha)\hat{\gamma}_{2i} - 1/36(2z_\alpha^3 - 5z_\alpha)\hat{\gamma}_{1i}^2$$

3. Calculate the sample mean estimate of the α quantile, Y^α.

$$\overline{Y}^\alpha = \frac{1}{n}\sum_{i=1}^{n} Y_i^\alpha n$$

4. Calculate the mean, $\overline{Y}_j^\alpha$, of $v = n/m$ subsets of Y_i^α values.

$$\overline{Y}_j^\alpha = \frac{1}{m}\sum_{i=(j-1)m+1}^{jm} Y_i^\alpha \quad j = 1, 2 \ldots, v$$

5. Obtain an approximate $1 - \beta$ confidence interval around the $\overline{Y}_j^\alpha$ values on the α cycle time quantile, where S is the sample standard deviation of all $\overline{Y}_j^\alpha$ values.

$$\overline{Y}^{\,\alpha} \pm t_{\beta,v-1}\frac{S}{\sqrt{v}}$$

Note that the Central Limit Theorem supports the approximation that the means of the subsets used in Steps 4 and 5 are becoming closer to normal as m gets sufficiently larger, and the value of m can be increased if concerns about normality are present. Additionally, the subsets are generated only to improve normality for building a confidence interval; if a point estimate and standard error are sufficient, choosing $m = 1$ is best. Also note that the batch means approach could easily be substituted in place of independent replications. In this case, n would represent the number of batches rather than the number of independent replications, and l would represent the batch size rather than the replication length. From each batch, moment estimates and corresponding steady-state quantile estimates would be calculated (Steps 2a and b). Using the quantile estimates from each batch, and assuming that the batches were large enough to yield independent quantile estimates, Steps 4 - 5 would be identical to those presented for the independent replications approach. For more information on the batch means approach, see Banks, *et al.* (2004) or Law and Kelton (2000).

4 Estimator bias

Attention should be given to the fact that there are several sources of bias present in the algorithm in Section 3: initialization bias, moment-estimation bias, and bias resulting from the CFE approximation. Initialization bias is inherent in all random, steady-state simulation models and goes away with run length. The moment-estimation bias results from the fact that simulation output data are autocorrelated: however, we will show that the standardized central moments used in this paper are consistent, despite the presence of bias due to non-i.i.d. data.

THEOREM 1. *Suppose that:*

The process $X_t, t = 1, 2, 3, \ldots$ is stationary,

the estimator of $E[(X-\mu)^k]$ is $\hat{\mu}'_k = n^{-1}\sum_{t=1}^{n}(X_t - \overline{x}_n)^k$ for $k \geq 2$, where $\overline{x}_n = n^{-1}\sum_{t=1}^{n} X_t$, and $E[X_t^k] < \infty$.

Then, if $n^{-1}\sum_{t=1}^{n} X_t^k \stackrel{n\to\infty}{\longrightarrow} E[X^k]$, a.s. for $k = 1, 2, 3, \ldots, m$, then $\hat{\mu}'_k \to \mu_k$ for $k = 1, 2, 3, \ldots, m$.

PROOF. Expanding the estimator for $\hat{\mu}'_k$, we get

$$\hat{\mu}'_k = \frac{1}{n}\sum_{t=1}^{n}\left[\sum_{j=0}^{k}\binom{k}{j}X_t^j\,\overline{X}_n^{\,k-j}(-1)^{k-j}\right] = \sum_{jl=0}^{k}(-1)^{k-j}\binom{k}{j}\left(\frac{1}{n}\sum_{t=1}^{n}X_t^j\right)\overline{X}_n^{\,k-j}$$

We then compare this estimator with the following expansion of $E[(x-\mu)^k]$.

$$E[(X-\mu)^k] = \sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}E[X^j]\mu^{k-j}.$$

Then, by the Continuous Mapping Theorem, $\hat{\mu}'_k \to E[(x-\mu)^k]$. $\square$

Corollary. *For the moment estimators, $\hat{\gamma}_1$ and $\hat{\gamma}_2$, given in Section 3, $\hat{\gamma}_k \to \gamma_k$ for $k = 1, 2$.*

Proof. The two sets of moments are related as follows:

$$\hat{\sigma} = \hat{\mu}'_2,\ \hat{\gamma}_1 = \frac{\hat{\mu}'_3}{\hat{\sigma}^{3/2}},\ \hat{\gamma}_1 = \frac{\hat{\mu}'_4}{\hat{\sigma}^2}$$

Since $\hat{\mu}'_k \to \mu_k$ for $k = 1, 2, 3, \ldots, m$, by the Continuous Mapping Theorem, $\hat{\gamma}_1 \to \frac{\mu'_3}{\sigma^{3/2}}$ and $\hat{\gamma}_2 \to \frac{\mu'_2}{\sigma^4}$. □

REMARK. The assumptions made by this theorem are not especially restrictive. Dai and Meyn (1995), for instance, show that the conditions hold for open multiclass queueing networks. Additionally, any stationary processes that is also ergodic meets the conditions, having sample moments that converge to the true population moments (Robinson, 1996). An example of such a process is a $\phi - mixing$ process, which is commonly used to characterize the output process from discrete-event simulation models; see Glynn and Iglehart (1990), Chen and Kelton (2000), and Steiger and Wilson (2002), for example.

Consequently, the terms of the CFE are asymptotically consistent. However, it should also be noted that there is no method for knowing *a priori* how long a run should be to ensure low biased moment estimates. Typically, the more heavily utilized a system is, the longer the bias takes to diminish.

In summary, in the limit both the moment-estimation bias and the initialization bias disappear with increasing run length. The bias resulting from the CFE approximation remains, but the results shown in the following section indicate that it does not significantly impact the accuracy of the quantile estimates.

5 Results

To evaluate the accuracy and precision of the CFE in estimating quantiles from DES models when estimates of sample moments are all that are available, three systems were examined: a single M/M/1 queue, 5 identical M/M/1 queues in tandem, and a full factory model based on a semiconductor manufacturing process. In all systems, First-In-First-Out (FIFO) dispatching was employed at all workstations. For the tandem queuing systems, all experimentation was performed using a discrete event simulator written in C++ , while experimentation for the full factory model was done using the commercial simulation package, Factory Explorer (Chance, 1995). Additionally, for all experimentation, n was set to 30, m was set to 5, while l was set at 1,000,000 for traffic intensities of 0.9 or lower and 2,000,000 for traffic intensities above 0.9. For all precision-related results, confidence intervals on indirect estimates were built using the procedure described in Section 3. Similarly, all confidence intervals around direct estimates were built based on order statistic estimates obtained from independent simulation runs. The independent estimates were grouped, and the average of each group was calculated, and a standard t-confidence interval was built around the group averages.

5.1 M/M/1 system

Empirical results demonstrate that both the direct (using order statistics) and indirect (using the CFE) estimates provide very accurate results for predicting the 0.90 cycle time quantile; neither estimation technique provides an estimate further than 2% from the true value at any traffic intensity between 0.5 and 0.97. Figure 3 shows the 95% confidence interval surrounding the 0.90 quantile obtained using both of the estimation techniques. Results show that the confidence interval surrounding the direct estimates is slightly narrower across all traffic intensities, indicating that in this experimentation there is less variability when using order statistics to estimate quantiles from an M/M/1 system than when using the CFE to indirectly produce estimates. For a more in depth discussion of the M/M/1 case, see McNeill *et al.* (2003).

[Figure 3 Here]

5.2 System of five tandem queues

The cycle time distribution from a system of 5 identical M/M/1 queues in tandem is significantly closer to normal than the distribution from the M/M/1 system. As a result, the CFE should produce better quantile estimates for this system than for the M/M/1 system. Figure 4 shows the 95% confidence interval on the difference between the estimated and theoretical 0.90 cycle time quantile for the system of 5 tandem M/M/1 queues. Figure 4 illustrates that neither the direct nor the indirect quantile estimates were off by more than 2% from the theoretical value at any of the simulated points. The direct estimates maintain a very similar level of accuracy and precisior to what they showed for the M/M/1 system, where accuracy indicates the closeness of the point estimate to the theoretical value and precision indicates the width of the confidence interval.

The indirect estimates, however, produce noticeably more accurate results with greater precision than they did for the M/M/1 system. Additionally, the mean percentage deviation from theoretical for the direct estimates seems to depend more on the traffic intensity of the system than do the percentage deviations for the indirect estimates.

[Figure 4 Here]

In addition to comparing the confidence intervals of the direct and indirect quantile estimation techniques, Figure 5 shows the robustness of the CFE in estimating different cycle time quantiles from the system of five tandem queues. Unlike the results from the M/M/1 system, the accuracy of the CFE does not degrade severely as the quantile to be estimated gets lower. Figure 5 demonstrates that for all quantiles estimated, from 0.30 to 0.98, the CFE produces estimates within 1.5% of the theoretical value. This improvement in estimation accuracy and robustness over the M/M/1 system can be explained by the cycle time distribution being closer to normality.

[Figure 5 Here]

Generating a chart analogous to Figure 5 for estimates obtained using order statistics would have been significantly more difficult both in terms of storage and post-processing requirements. For example, direct estimation would have required storage of 1 million observations at a time, 180 times (6 design points * 30 runs), to generate Figure 5, while using the CFE required storage of only 4 moment estimates and the number of observations for each of the 180 runs. Additionally, using the CFE, all the quantile estimates shown in Figure 5 were generated from a single set of simulation runs at each design point. Once the moments at any traffic intensity were estimated, the only necessary step in obtaining multiple quantile estimates at that traffic intensity using the CFE was adjusting z_α in Equation (2) to the appropriate value. Also, should data beyond the initial sample have been collected, it would be substantially easier to use this data for updating moment estimates than it would have been to update the sorted list of individual cycle time values.

Table 1 summarizes these differences between the two quantile estimation techniques. It illustrates the benefits in computer storage and computing effort of using the CFE vs. order statistics at a single design point. These advantages are magnified as the number of design points of interest increases. Of note is the fact that Table 1 assumes the most traditional application of order statistics. Procedures exist (i.e. Jain and Chlamtac, 1985), which modify traditional order statistics by requiring the storage of only a small subset of the observations. Assuming the desired quantiles were known in advance, using such procedures would reduce the storage requirements for direct estimation reported in Table 1.

Table 1: Computational requirements for direct and indirect quantile estimation for the system of 5 tandem M/M/1 queues.

Estimation Technique	**Number of Data Values Stored**	**Post-Processing**	**Effort to Obtain a New Quantile Estimate**
Indirect Estimation (CF Expansion)	5	Evaluate the CF Expansion (virtually instantaneous using a spreadsheet)	Change z_α value in the CF expansion (also done using simple spreadsheet)
Direct Estimation (Order Statistics)	1,000,000	Sort data and retrieve value (requires computer program)	Retrieve new value from the sorted list

5.3 Full factory model

Finally, in addition to testing the CFE on a simple M/M/1 queuing system and on the system of 5 M/M/1 queues in tandem, experimentation was also performed on a more complex model representing a full semiconductor manufacturing factory. The model was taken from the website of the Modeling and Analysis of Semiconductor Manufacturing (MASM) lab at Arizona State University, `www.fulton.asu.edu/~ie/research/labs/masm`. Testbed Data Set #1 was used as the basis for the full factory model; details about the system are given in Table 2.

An example of the cycle time distribution for the full factory model run at a 60% traffic intensity at the bottleneck (the most heavily utilized workstation) is shown in Figure 6. As expected, at this moderate traffic intensity, the distribution looks closer to normal (though with a longer right tail) than even the system of 5 tandem queues. Consequently, the expectation is that for

Table 2: Description of full factory model.

Product type	Non-volatile memory
Number of products	1
Number of processing steps	232
Number of tool groups	83
Number of operator groups	32
Rework modeled?	Yes
Machine breakdown modeled?	Yes
Machine loading/unloading?	Yes

lower traffic intensities, the CFE will provide very accurate quantile estimates for this system. Figure 6 also gives an example of the cycle time distribution for the same system at 97% traffic intensity. Clearly, at this traffic intensity the system is less normally distributed, and, therefore, estimation of higher moments plays more of a role in generating accurate quantile estimates. Since the theoretical quantiles are unknown, however, assessment of the indirect estimation technique is made by comparing with direct estimates obtained using order statistics. It is assumed that after simulating for a very long time, the direct estimates provide a good estimate of the true quantiles. Therefore, an indirect estimation technique that provides results similar to the direct estimates is assumed to be accurate as well.

[Figure 6 Here]

For estimates of the 0.9 cycle time quantile at traffic intensities between 0.5 and 0.98, the direct and indirect quantile estimates for this system were never more than 2% different from each other. Also, the width of the 95% confidence interval surrounding the 0.9 quantile was not more than 3% different than the directly estimated quantile at any of the simulated points. For a variety of other quantiles, the percentage difference between the direct and indirect estimates for this system is never more than 4%, and there does not appear to be any relationship between the quantile being estimated and the quantity by which the two estimates differ, indicating that the CFE is capable of producing equally accurate results for both high and low quantiles.

Figure 7 illustrates one of the benefits to the decision maker of generating multiple quantile estimates from a single set of simulation runs. This plot was generated from the output of a single set of runs and gives the complete range of cycle time quantiles for the full factory model at 90% traffic intensity. The figure gives an estimate of the inverse of the cycle time distribution's cdf at this traffic intensity and provides a valuable tool for quoting lead times. For example, it illustrates that if the factory was being run at 90% traffic intensity, a lead time of 1400 hours could be met approximately 90% of the time, while a quoted lead time of 1300 hrs could only be met approximately 50% of the time. Additional discussion on the full factory model can be found in McNeill *et al.* (2003).

[Figure 7 Here]

6 Other quantile estimation techniques

The CFE directly translates sample moments from a distribution into quantile estimates. Another technique is to use these moment estimates to generate an estimate of the cycle-time distribution's cdf. From the cdf, quantile estimates can then be generated by taking the inverse. Since both the skewness and kurtosis are important in describing a distribution, and since estimates of the first four moments were already available from previous experimentation with the CFE, four parameter distributions were chosen for this experimentation. Additionally, distributions were sought that could accurately capture the shape of the normal distribution and the shape of distributions with more prominent skewness. Specifically, the following four-parameter distributions were investigated: the Pearson family of distributions, the Johnson family of distributions, and the Generalized Lambda Distribution. The Pearson and Johnson families were ruled out quickly based on experimentation with the exponential distribution, where they generated quantile estimates that were very poor.

The generalized lambda distribution (GLD) is a four parameter family of distributions that has the capability of modeling both the normal and the exponential distributions quite well. The quantile function of the GLD is given below, where λ_1 and λ_2 are location and scale parameters and λ_3 and λ_4 determine the skewness of the distribution. The regions that produce a valid distribution can be found in Karian and Dudewicz (2000).

$$F^{-1}(y) = Q(y : \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{y^{\lambda_3} - (1-y)^{\lambda_4}}{\lambda_2}$$

To fit the distribution using estimates of the first four moments, a numerical search algorithm, written in Maple (Heck, 2003) and available from Karian and Dudewicz (2000), can be used. This algorithm takes as inputs the sample moment estimates and a starting point for the numerical search algorithm. The starting point for the algorithm can be obtained from a table, also available in Karian and Dudewicz (2000). The outputs of the algorithm are the distribution parameters that are used to generate the cdf, which, in turn, allows the calculation of any quantile.

Figure 8 shows estimates of various quantiles for the M/M/1 system using the GLD. The sample moment estimates used in this experimentation were identical to those used in the M/M/1 experimentation with the CFE. As a whole, the estimates appear to be slightly more accurate than those obtained using the CFE, especially at lower traffic intensities and for lower quantiles. The improvement in accuracy is not universal, however, and is not substantial enough to justify the extra work of looking up and providing a starting point for the numerical search algorithm for each simulation replication at each traffic intensity.

[Figure 8 Here]

Experimentation with the GLD was also performed on the system of 5 tandem queues to determine how well the GLD could estimate quantiles of a system with more normally distributed cycle-times. Results from this experimentation showed that the accuracy is very comparable to those obtained using the CFE.

In summary, the GLD, while capable of providing accurate quantile estimates for both the exponential and normal distributions, does not provide enough of an improvement over the CFE to warrant its use. The numerical search algorithm used to fit the distribution is fairly sensitive to the starting point, and the necessity of providing this starting point severely degrades the ease of use. As a result, neither the GLD nor the Pearson and Johnson distributions mentioned previously are recommended over the CFE for quantile estimation from DES models.

7 Conclusions

While the paper presents some mathematical support for the CFE-based quantile estimation approach, the real demonstration of the technique's value comes from the fact that we have used it successfully in extensive experiments specialized to the kind of data we expect to find in manufacturing. The results from the M/M/1 system, the system of 5 tandem M/M/1 queues, and the full factory model illustrate that the CFE can be effectively used as an indirect cycle time quantile estimator for manufacturing systems. Specifically, the experimentation supports the approach for estimating quantiles higher than 0.5 in systems operating under FIFO dispatching. Also, despite the fact that the direct estimates obtained using order statistics provided slightly more accurate quantile estimates with less variability than the indirect estimates, particularly for the M/M/1 system, the gains in accuracy were offset by the ease of implementation and data storage requirements. Similarly, fitting the GLD using its cdf to generate quantile estimates provided slightly more accurate results than the CFE for the M/M/1 system and comparable results for the system of 5 tandem queues. However, the requirement of providing a starting point for the algorithm makes the technique significantly more difficult to implement and outweighs the relatively minor improvement in accuracy.

In summary, the CFE provides an indirect steady-state quantile estimation technique that is easy to implement and has the advantage of being able to generate multiple quantile estimates from a single set of simulation runs, allowing charts like Figure 7 to be easily generated. Furthermore, perhaps equally important to the savings in computing time is the savings in computer storage required by the CFE. To make a direct quantile estimate using order statistics, storage of each cycle time observation is required. The CFE, on the other hand, requires storage of only four values used to calculate sample moments, regardless of the number of cycle time observations collected. Consequently, as the number of cycle time observations increases, so does the savings in computer memory, again highlighting the compactness of the representation of a system provided by the CFE. These low data storage requirements also make the CFE-based quantile estimation technique an attractive method to add to a commercial manufacturing simulation package such as Factory Explorer. Finally, if there is a need to collect additional data, adding this new data to moment estimates for indirect estimation is much easier than adding the data to a sorted list of cycle time observations used for direct estimation. As such, it is generally preferable to direct estimation using order statistics and indirect estimation from four-parameter distributions. Moreover, specific attention has been paid in this paper to the manufacturing industry because certain assumptions about the cycle time distribution can be made, but the application of the CFE to quantile estimation need not be limited to these applications. The method is also applicable for service systems, and, in fact, any type of queueing system.

8 Future work

The empirical results in this paper have shown that the CFE can provide accurate quantile estimates for a variety of cycle time distributions for systems operating under FIFO dispatching. An assumption of this work is that the cycle time distribution from which quantiles are desired will not be so far from the normal distribution that the CFE will have to make major adjustments. Therefore, we are very confident recommending the approach for any system operating under FIFO dispatching. However, it is unclear that this assumption will hold in the case of non-FIFO dispatching at any workstation. Preliminary experimentation indicates that the approach is promising for the Earliest Due Date (EDD) and random dispatching policies; in preliminary experimentation on a system of 5 tandem M/M/1 queues, the quantile estimation technique presented in Section 3 yielded results within 4% of the corresponding quantiles estimated using order statistics for quantiles between 0.5 and 0.99. However, the same experimenation with Last-In-First-Out (LIFO) dispatching, Shortest Processing Time (SPT) dispatching, and Critical Ratio (CR) dispatching gave much poorer results (quantile estimates were more than 300% off in some cases), indicating that an alternative approach must be used. Bekki *et al.* (2006) indicate that a power transformation, which brings the CT distribution closer to normal, may be effective in such situations. Further experimentation should be performed before applying this technique to systems which employ a dispatching policy other than FIFO at a bottleneck workstation.

References

Avramidis, N.A. and Wilson, J.R. (1998) Correlation-induction techniques for estimating quantiles in simulation experiments. Operations Research, 46, 574 – 591.

Banks, J., Carson, J., Nelson, B.L., and Nicol, D. (2004) Discrete-Event System Simulation, Fourth Edition. Prentice Hall, New Jersey.

Bekki, J., Mackulak, G.T., and Fowler., J.W. (2006) Indirect Cycle-Time Quantile Estimation for Non-FIFO Dispatching Policies, in Proceedings of the 2006 Winter Simulation Conference, eds. L.F. Perrone, F.P. Wieland, J. Liu, B.G. Lawson, D.M. Nicol, and R.M. Fujimoto, pp. 1829 - 1835.

Chan, L.K. and Cui, H.J. (2003) Skewness correction $\overline{X}$ and R charts for skewed distributions. Naval Research Logistics, 50, 555 – 573.

Chance, F. (1995) Factory Explorer, Integrated Capacity, Cost, and Cycle Time Analysis, Version 2, Level 0, Beta 2, User's Guide and Reference. Wright, Williams, & Kelly, Dublin, California.

Chen, E. J. and Kelton W.E. (2008) Estimating steady-state distributions via simulation-generated histograms. Computers in Operations Research, 35, 1003 – 1016.

Chen, E.J. and Kelton, W.D. (2006) Quantile and tolerance-interval estimation in simulation. European Journal of Operational Research, 168, 520 – 540.

Chen, E.J. and Kelton, W.D. (2001) Quantile and histogram estimation, in Proceedings of the 2001 Winter Simulation Conference, eds. B.A. Peters, J.S. Smith, D.J. Medeiros, and M.W. Rohrer, pp. 451 – 459.

Chen, E.J. and Kelton, W.D. (2000) Mean estimation based on phi-mixing sequences, in Proceedings of the 33rd Annual Simulation Symposium, pp. 237 – 244.

Cornish, E.A. and Fisher, R.A. (1937) Moments and cumulants in the specification of distributions. Revue de l'Institut International de Statistique, 5, 307 – 320.

Dai, J. and Meyn, S. (1995) Stability and convergence of moments for multiclass queueing networks via fluid limit models. IEEE Transactions on Automatic Control, 40, 1889 – 1904.

Glenn, P.W. and Iglehart, D.L. (1990) Simulation output analysis using standardized time series. Mathematics of Operations Research, 15, 1 – 16.

Gordon, V.S. (1993) A note on optimal assignment of slack due dates in single machine scheduling. European Journal of Operational Research, 70, 311 – 315.

Heck, A. (2003) Introduction to Maple, 3rd edition. New York.

Heidelberger, P. and Lewis, P.A.W. (1984) Quantile estimation in dependent sequences. Operations Research, 32, 185 – 209.

Hesterberg, T.C. and Nelson, B.L. (1998) Control variates for probability and quantile estimation. Management Science, 44, 1295 – 1312.

Jain, R. and Chlamtac, I. (1985) The P^2 algorithm for dynamic calculation of quantiles and histograms without storing observations. Communications of the ACM, 28, 1076 – 1085.

Jaschke, S.R. (2002) The Cornish-Fisher-Expansion in the context of delta-gamma-normal approximations. Journal of Risk, 4, 33 – 52.

Jin, X., Fu, M.C., and Xiong, X. (2003) Probabilistic error bounds for simulation quantile estimators. Management Science, 49, 230 – 246.

Karian, Z.A. and Dudewicz, E.J. (2000) Fitting Statistical Distributions, the Generalized Lambda Distribution and Generalized Bootstrap Methods. CRC Press, Boca Raton.

Kenney, J.F. and Keeping, E.S. (1954) Mathematics of Statistics, Part 1. D.Van Nostrand Company, Inc., New Jersey.

Law, A. and Kelton, W.D. (2000) Simulation Modeling and Analysis, Third Edition. McGraw Hill, Boston.

McNeill, J., Mackulak, G., and Fowler, J. (2003) Indirect estimation of cycle time quantiles from discrete event simulation models using the Cornish-Fisher expansion, in Proceedings of the 2003 Winter Simulation, eds. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, pp. 1377 – 1382.

Miguel, J., Olave, P. (2002) Adjusting forecast intervals in arch-m models. Journal of Time Series Analysis, 23, 587-598.

Mittler, M., Gerlich, N., and Schömig, A.K. (1995) On cycle times and interdeparture times in semiconductor manufacturing. University of Würzburg Technical Report Number No. 124, September.

Robinson, P.M. (1994) Time series with strong dependence, in Advances in Econometrics: Sixth World Congress Vol 1, ed. Sims, C.A, Cambridge University Press, New York, pp 47 – 96.

Roeder, T.M., Fischbein, S.A., Janakiram, M., and Schruben, L.W. (2002) Resource-driven and job-driven simulations, in Proceedings of the 2002 International Conference on Modeling and Analysis of Semiconductor Manufacturing, pp. 78 – 83.

Rose, O. (1999) Estimation of the cycle time distribution of a wafer fab by a simple simulation model, in Proceedings of the 1999 International Conference on Semiconductor Manufacturing, Operational Modeling and Simulation, pp. 133 – 138.

Schruben, L.W. and Roeder, T.M. (2003) Fast simulations of large-scale highly-congested systems. Simulation: Transactions of the Society for Modeling and Simulation International, 79, 115 – 125.

Steiger, N.M. and Wilson, J.R. (2002) An improved batch means procedure for simulation output analysis. Management Science, 48, 1569 – 1586.

Stuart, A. and Ord, J.K. (1987) Kendall's Advanced Theory of Statistics, Fifth Edition of Volume 1: Distribution theory. Charles Griffin & Company Limited, London.

Whitt, W. (1989). Planning queueing simulations. Management Science, 35, 1341 – 1366.

Yang, F., Ankenman, B.A. and Nelson, B.L. (2005) Estimation of perentiles of cycle time in manufacturing simulation, in Proceedings of the 2005 Winter Simulation Conference, eds. M.E. Kuhl, N.M. Steiger, F.B. Armstrong, and J.A. Joines, pp 475 – 484.

Biographical Sketches

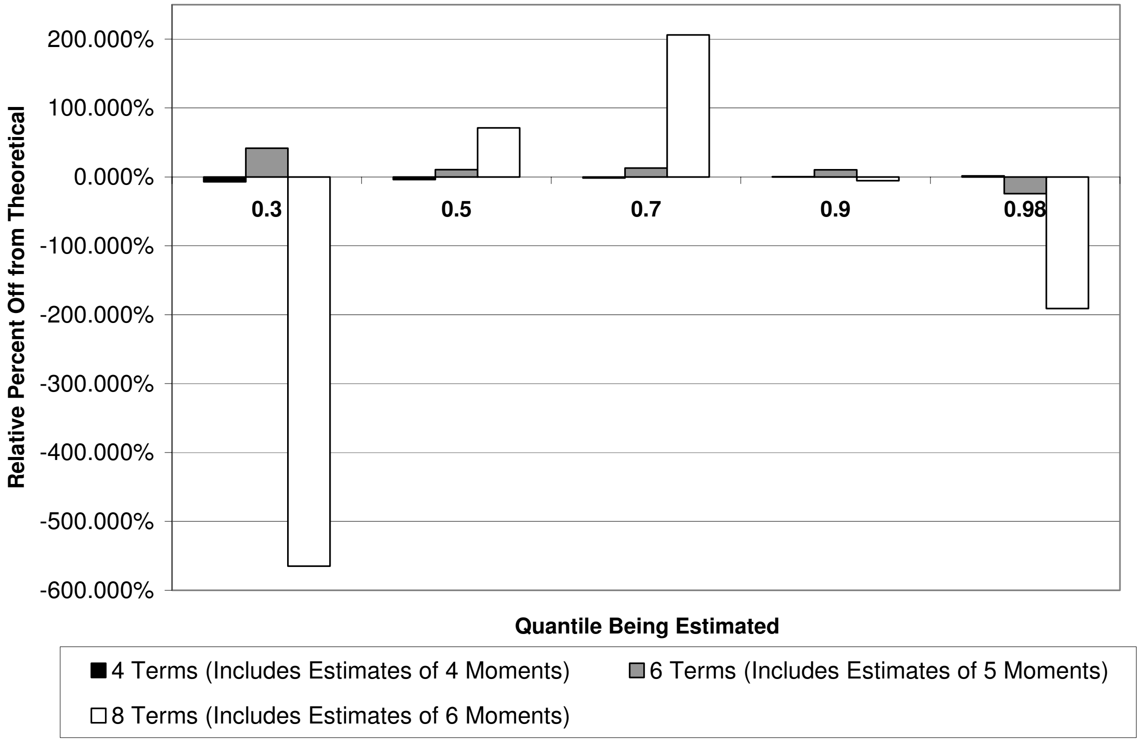
JENNIFER M. BEKKI is a Ph.D. candidate in the Industrial Engineering department at Arizona State University. Her research interests are in discrete-event simulation methodology and manufacturing applications in the semiconductor industry. Prior to beginning her Ph.D. studies, she served as an intern in the Operational Decision Support Technologies group at Intel, and she is currently funded on an SRC/Intel Doctoral Fellowship. Jennifer is a student member of both INFORMS and IIE and served as the editor of the INFORMS online student publication, *OR/MS Tomorrow*, between August, 2004 and August, 2006.

GERALD T. MACKULAK is an Associate Professor of Engineering in the Department of Industrial Engineering at Arizona State University. He is a graduate of Purdue University receiving his B.Sc., M.Sc., and Ph.D. degrees in the area of Industrial Engineering. His primary area of research is simulation applications within manufacturing with a special focus on automated material handling within semiconductor manufacturing.

JOHN W. FOWLER is a Professor of Industrial Engineering at Arizona State University (ASU) and is the Center Director for the Factory Operations Research Center that is jointly funded by International SEMATECH and the Semiconductor Research Corporation. His research interests include modeling, analysis, and control of semiconductor manufacturing systems. Dr. Fowler is a member of ASEE, INFORMS, IIE, POMS, and SCS. He is an Area Editor for SIMULATION: Transactions of the Society for Modeling and Simulation International and an Associate Editor of IEEE Transactions on Electronics Packaging Manufacturing. He is an IIE Fellow and is on the Winter Simulation Conference Board of Directors.

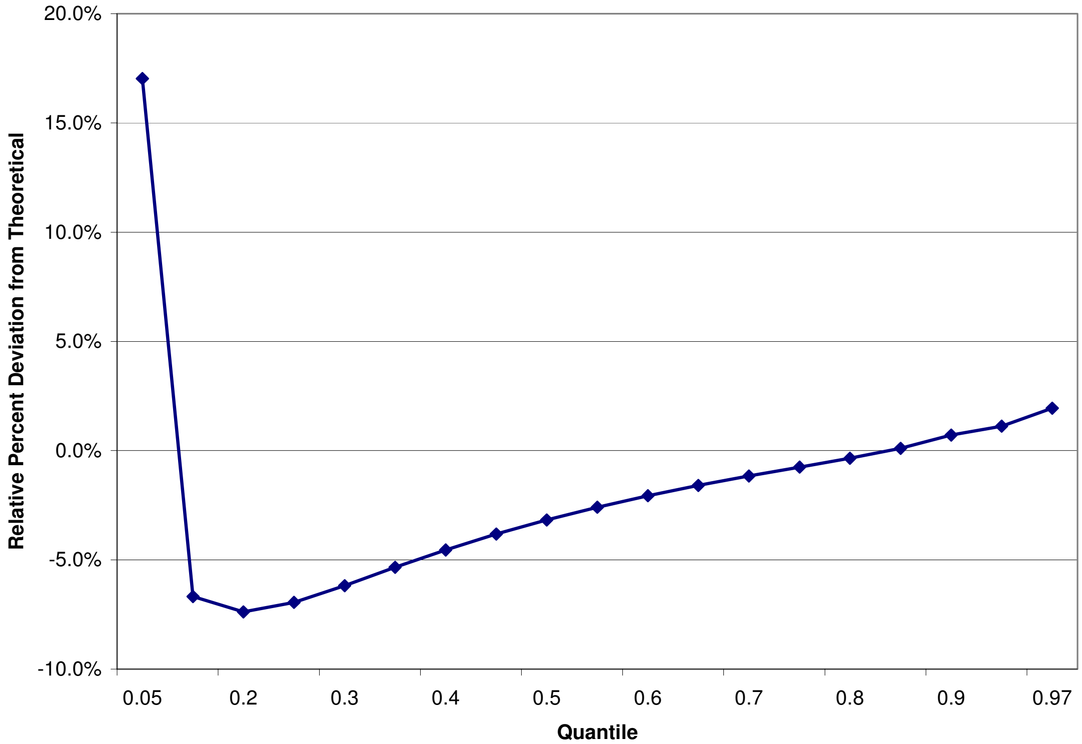
BARRY L. NELSON is the Charles Deering McCormick Professor in the Department of Industrial Engineering & Management Sciences at Northwestern University. His research interests are the design and analysis of stochastic simulation experiments, particularly issues of multivariate input modeling, optimization via simulation and metamodeling. He is currently Editor in Chief of *Naval Research Logistics*, on the Board of Directors of the Winter Simulation Conference, and is a member of IIE (senior member), INFORMS (fellow), ACM, and ASA.

Figure 1



200.000%
100.000%
0.000%
-100.000%
-200.000%
-300.000%
-400.000%
-500.000%
-600.000%
Relative Percent Off from Theoretical
0.3
0.5
0.7
0.9
0.98
Quantile Being Estimated
4 Terms (Includes Estimates of 4 Moments)
6 Terms (Includes Estimates of 5 Moments)
8 Terms (Includes Estimates of 6 Moments)

Figure 2



20.0%
15.0%
10.0%
5.0%
0.0%
-5.0%
-10.0%
Relative Percent Deviation from Theoretical
0.05
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.97
Quantile

Figure 3

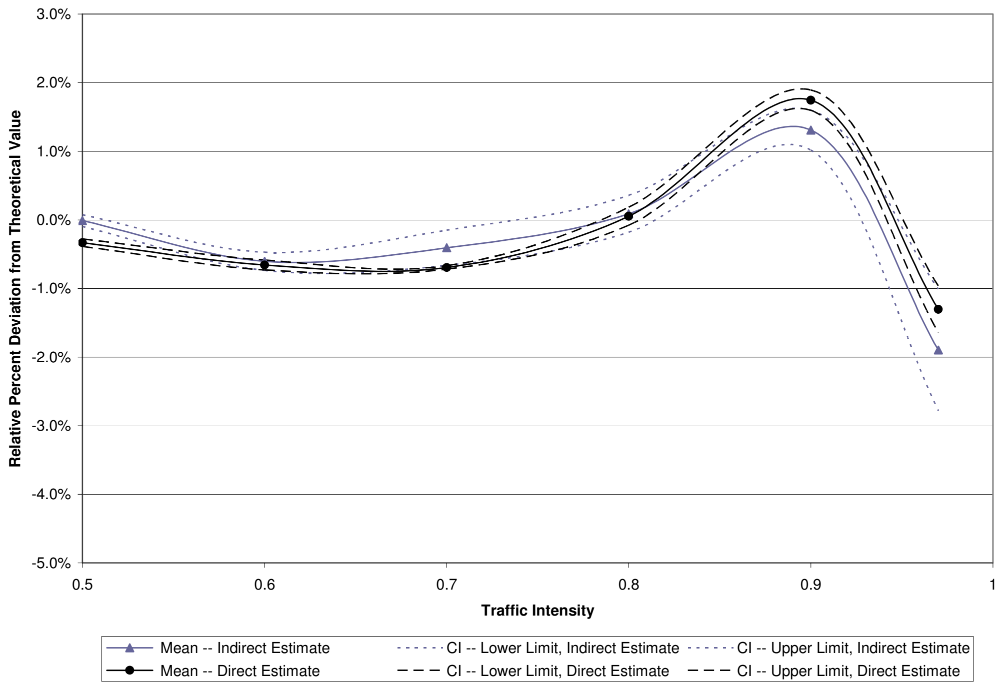

3.0%
2.0%
1.0%
0.0%
-1.0%
-2.0%
-3.0%
-4.0%
-5.0%
Relative Percent Deviation from Theoretical Value
0.5
0.6
0.7
0.8
0.9
1
Traffic Intensity
Mean -- Indirect Estimate
Mean -- Direct Estimate
CI -- Lower Limit, Indirect Estimate
CI -- Lower Limit, Direct Estimate
CI -- Upper Limit, Indirect Estimate
CI -- Upper Limit, Direct Estimate

Figure 4

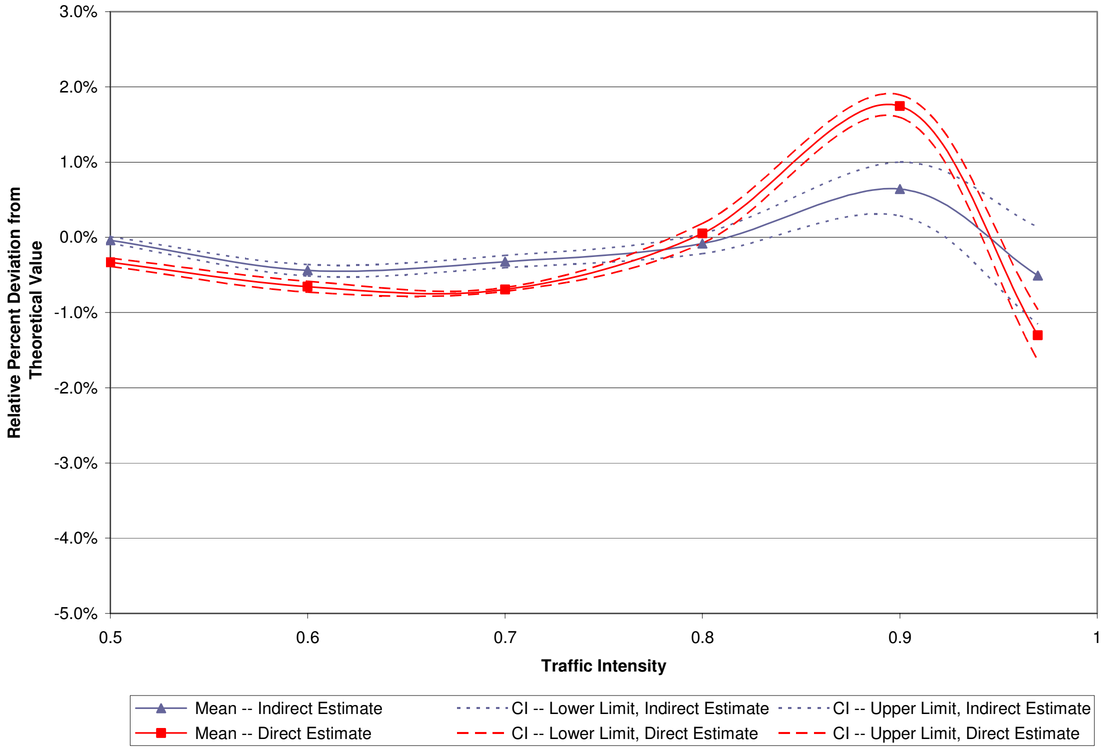

3.0%
2.0%
1.0%
0.0%
-1.0%
-2.0%
-3.0%
-4.0%
-5.0%
Relative Percent Deviation from Theoretical Value
0.5
0.6
0.7
0.8
0.9
1
Traffic Intensity
Mean -- Indirect Estimate
Mean -- Direct Estimate
CI -- Lower Limit, Indirect Estimate
CI -- Lower Limit, Direct Estimate
CI -- Upper Limit, Indirect Estimate
CI -- Upper Limit, Direct Estimate

Figure 5

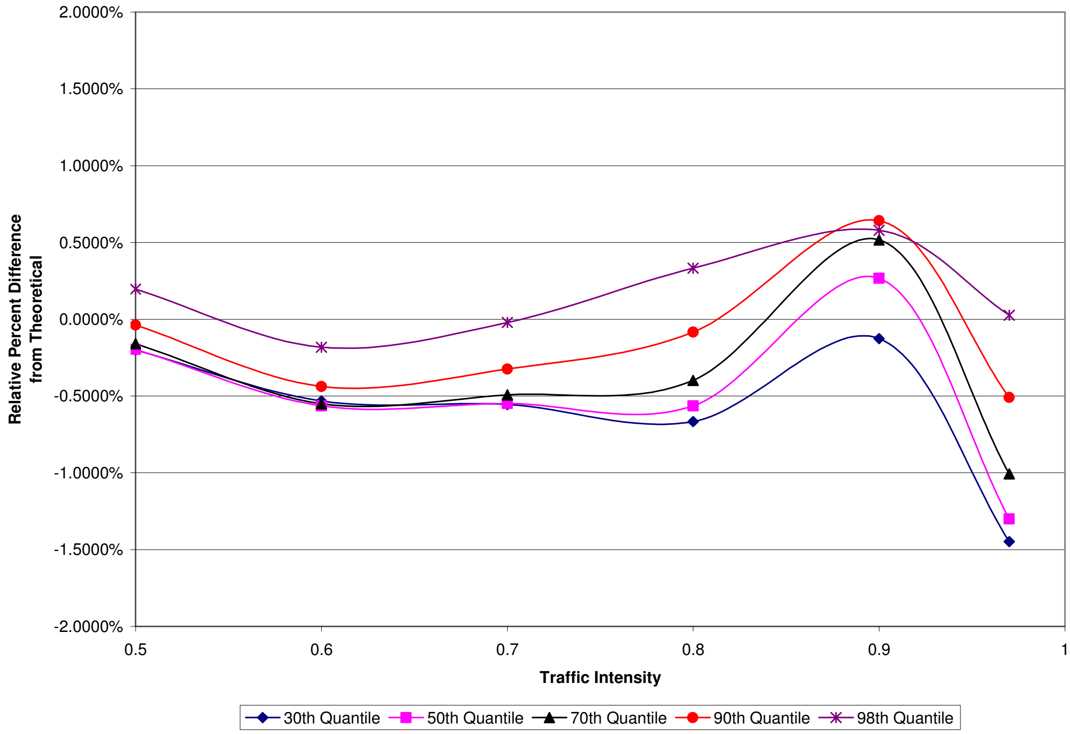

2.0000%
1.5000%
1.0000%
0.5000%
0.0000%
-0.5000%
-1.0000%
-1.5000%
-2.0000%
Relative Percent Difference from Theoretical
0.5
0.6
0.7
0.8
0.9
1
Traffic Intensity
30th Quantile
50th Quantile
70th Quantile
90th Quantile
98th Quantile

Figure 6

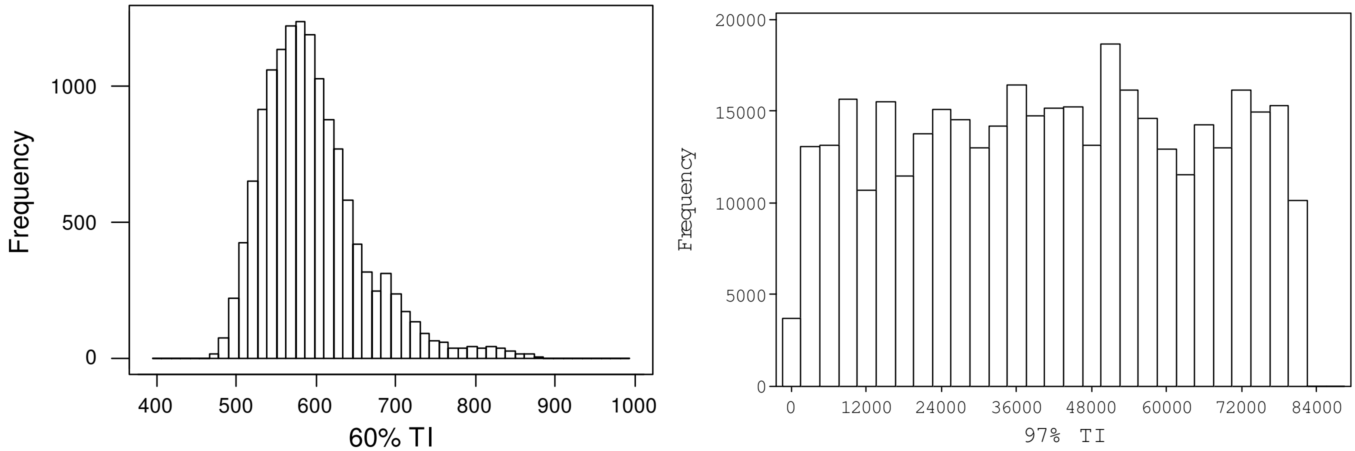

Frequency
0
500
1000
400
500
600
700
800
900
1000
60% TI
Frequency
0
5000
10000
15000
20000
0
12000
24000
36000
48000
60000
72000
84000
97% TI

Figure 7

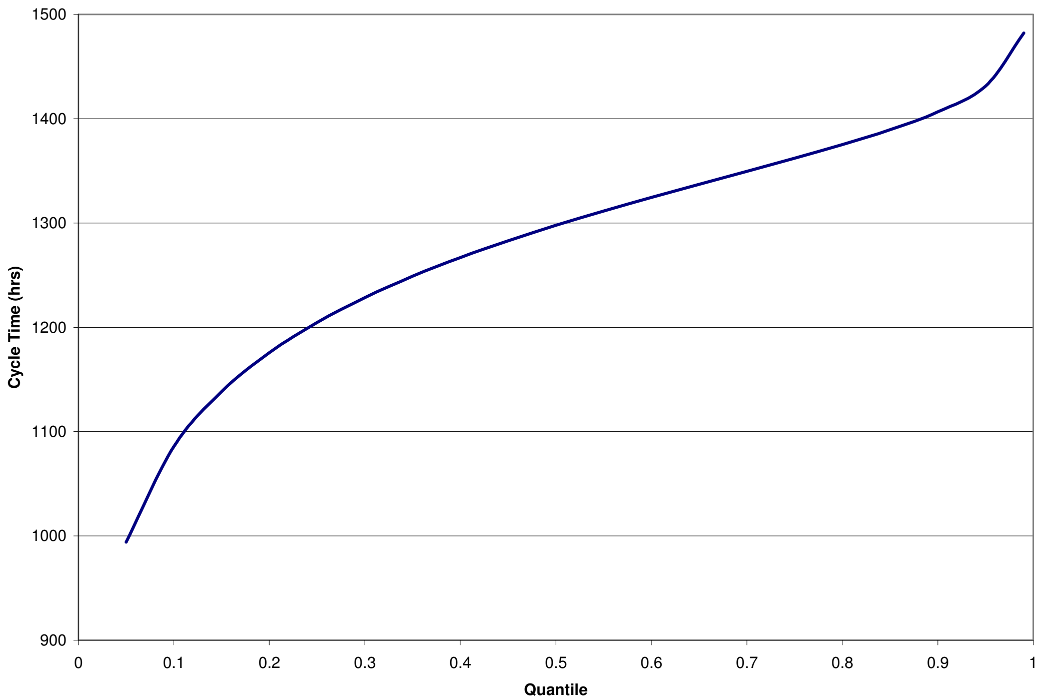
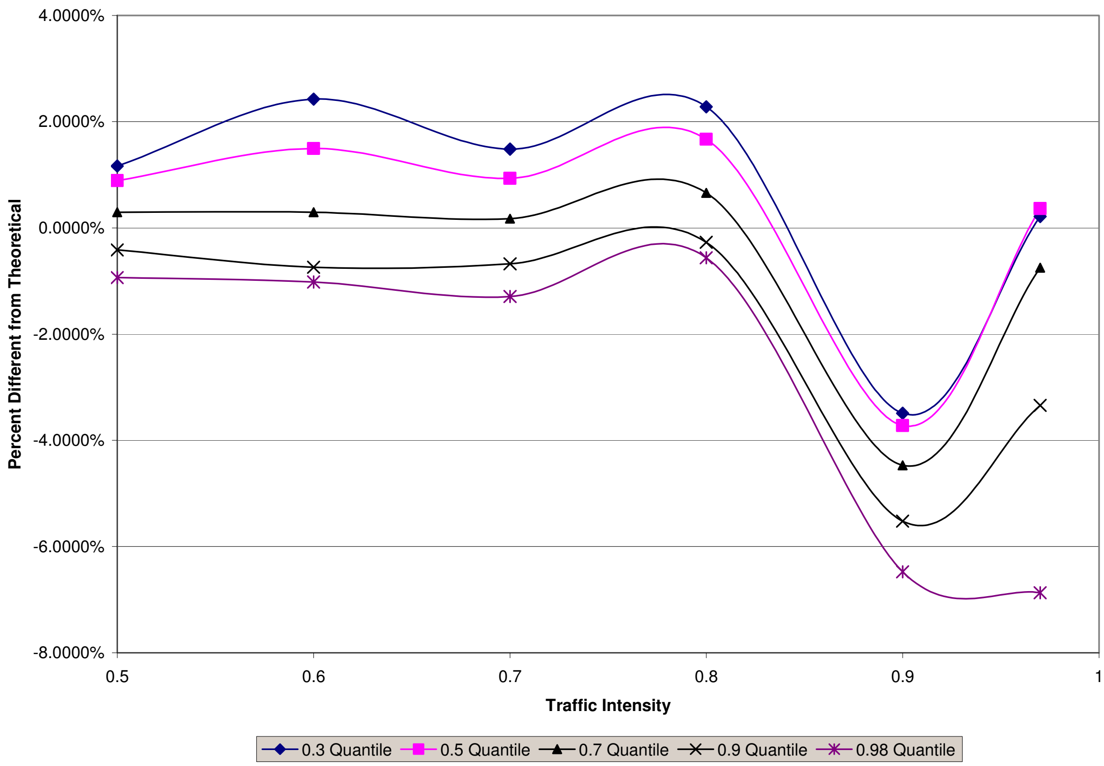

1500
1400
1300
1200
1100
1000
900
Cycle Time (hrs)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Quantile

Figure 8



4.0000%
2.0000%
0.0000%
-2.0000%
-4.0000%
-6.0000%
-8.0000%
Percent Different from Theoretical
0.5
0.6
0.7
0.8
0.9
1
Traffic Intensity
0.3 Quantile
0.5 Quantile
0.7 Quantile
0.9 Quantile
0.98 Quantile

Figure 1: Relative percent deviation between approximated and theoretical quantiles from the exponential distribution as a function of the number of terms included in the expansion. The Cornish-Fisher expansion, provided with normalized, theoretical moments, was used to compute the approximations.

Figure 2: Relative percent difference between estimated and theoretical quantiles for the exponential distribution. Estimates were computed using the first four terms of the Cornish-Fisher expansion and theoretical moments.

Figure 3: 95% Confidence interval on percentage difference between estimated and the theoretical 0.90 cycle time quantile for an M/M/1 system. Direct estimates were computed using order statistics, and indirect estimates were computed using the first four terms of the Cornish-Fisher expansion.

Figure 4: 95% Confidence interval on the relative percent difference between theoretical and estimated 0.90 cycle-time quantile for a system of 5 tandem M/M/1 queues. Direct estimates were computed using order statistics, and indirect estimates were computed using the first four terms of the Cornish-Fisher expansion.

Figure 5: Relative percent difference between indirectly estimated and theoretical quantiles for a system of 5 tandem M/M/1 queues. Indirect estimates were computed using the first four terms of the Cornish-Fisher expansion.

Figure 6: Sample cycle-time distribution from full factory model run at a traffic intensity of 60% at the bottleneck (left) and a traffic intensity of 97% at the bottleneck (right).

Figure 7: Complete range of quantile estimates at 90% traffic intensity for the full factory model. Quantile estimates were computed using the first four terms of the Cornish-Fisher expansion.

Figure 8: Relative percent difference from theoretical quantiles for an M/M/1 system. Quantile estimates were computed from the lambda distribution.