Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

QUANTILE AND HISTOGRAM ESTIMATION

E. Jack Chen

Mathematical Modeling Group
BASF Corporation
3000 Continental Drive
Mount Olive, New Jersey 07828, U.S.A.

W. David Kelton

Department of Quantitative Analysis and
Operations Management
University of Cincinnati
Cincinnati, Ohio 45221, U.S.A.

ABSTRACT

This paper discusses implementation of a sequential procedure to construct proportional half-width confidence intervals for a simulation estimator of the steady-state quantiles and histograms of a stochastic process. Our *quasi-independent* (QI) procedure increases the simulation run length progressively until a certain number of essentially independent and identically distributed samples are obtained. We compute sample quantiles at certain grid points and use Lagrange interpolation to estimate the p quantile. It is known that order statistics quantile estimator is asymptotically unbiased when the output sequences satisfy certain conditions. Even though the proposed sequential procedure is a heuristic procedure, it does have strong basis. Our empirical results show that the procedure gives quantile estimates and histograms that satisfy a pre-specified precision requirement. An experimental performance evaluation demonstrates the validity of using the QI procedure to estimate the quantiles and histograms.

1 INTRODUCTION

Simulation studies have been used to investigate the characteristics of the system under study, for example the mean and the variance of certain system performance. In this paper, we propose a method to construct an empirical distribution of the parameter of interest. For $0 < p < 1$, the *p quantile* (percentiles) of a distribution is the value at or below which $100p$ percent of the distribution lies. Related to quantile, a *histogram* is a graphical estimate of the underlying probability density (mass) function and reveals all the essential distributional features of output random variables analyzed by simulation. A histogram can be constructed with a properly selected set of quantiles. We propose a simple *Quasi-Independent* (QI) algorithm (see Chen and Kelton 2000a,b) to determine the simulation run length and use grid points to construct the histogram (multiple quantiles). Iglehart (1976), Seila (1982a,b) and Hurley and Modarres (1995) have developed quantile estimation algorithms based on grid points. However, their procedures require that users enter the values of the grid points. For an overview of quantile estimation procedures see Law and Kelton (2000).

It is known that for both independent and identically distributed (i.i.d.) and *ϕ-mixing* sequences, see Section 2, sample quantiles will be asymptotically unbiased if certain conditions are satisfied. The asymptotic validity is reached as the sample size or simulation run length get large. However, in practical situations simulation experiments are restricted in time and it is not known in advance what is the required simulation run length for the estimator to be unbiased. Moreover, estimating the variance of the quantile estimator is needed to evaluate the precision of the quantile estimator. Therefore, a workable finite-sample size must be determined dynamically for the precision required of a simulation.

We propose a histogram approximation based on a QI procedure (see Section 3) for estimating quantiles from a stationary simulation output. The proposed procedure will sequentially determine the simulation run length so that the quantile estimate satisfies a pre-specified precision requirement. The asymptotic validity of our QI procedure occurs as the subsequence appears to be independent, as determined by the *runs-up* test (see Section 2).

The main advantage of our approach is that by using grids to approximate the underlying distribution, we avoid storing and sorting all the observations. However, the savings come at a cost, using interpolation to obtain quantile estimates introduces bias. Fortunately, the bias can be reduced by specifying finer grid points, which of course requires longer execution time. The QI procedure computes the number of required independent samples at the beginning of the procedure making implementation a relatively simple task.

In Section 2 we discuss some theoretical basis of quantile estimation in the context of simulation output analysis. In Section 3 we present our methodologies and proposed procedure for quantile and histogram estimation. In Section

4 we show our empirical-experiment results of quantile and histogram estimation. In Section 5 we give concluding remarks.

2 BACKGROUND

This section presents the theoretical basis of our quantile estimation: order statistics quantile estimators, ϕ-mixing, and runs-up test.

Let $X_1, X_2, \cdots, X_n$, be a sequence of i.i.d. random variables from a continuous distribution $F(x)$ with probability density function $f(x)$. Let x_p $(0 < p < 1)$ denote the $100p^{th}$ percentile or the p quantile, which has the property that $F(x_p) = Pr(X \le x_p) = p$. Thus, $x_p = \inf\{x : F(x) \ge p\}$. If $Y_1, Y_2, \ldots, Y_n$, are the order statistics corresponding to the X_i's from n independent observations, (i.e. Y_i is the i^{th} smallest of X_1, X_2, ..., X_n) then a point estimator for x_p based on the order statistics is the sample p quantile $\hat{x}_p$,

$$\hat{x}_p = y_{\lceil np \rceil} \tag{1}$$

where $\lceil z \rceil$ denotes the integer ceiling (round-up) of the real number z.

For data that are i.i.d., the following properties of $\hat{x}_p$ are well known (David 1981):

$$E(\hat{x}_p) = x_p - \frac{p(1-p)f'(x_p)}{2(n+2)f^3(x_p)} + O(1/n^2);$$

$$\text{Var}(\hat{x}_p) = \frac{p(1-p)}{(n+2)f^2(x_p)} + O(1/n^2).$$

We say that L_n *is large order of* x_n (as $n \to \infty$) and write $L_n = O(x_n)$ if there exists a constant $k > 0$ and N such that $\parallel L_n \parallel \le k|x_n|$ for each $n \ge N$. $\parallel L_n \parallel$ denotes the Euclidean norm of L_n.

Roughly speaking, the sequence $X_1, X_2, \cdots, X_n$ is ϕ-mixing if X_i and X_{i+j} become essentially independent as j becomes large; see Billingsley (1999) for formal definition. For example, the waiting time W_i of an M/M/1 delay-in-queue is ϕ-mixing, because W_i and W_{i+j} become essentially independent as j becomes large (Daley, 1968). A broad class of dependent stochastic processes possess this ϕ-mixing property.

Quantile estimation can be computed using standard nonparametric estimation based on order statistics, which can be used not only when the data are i.i.d. but also when the data are drawn from a stationary, ϕ-mixing process of continuous random variables. It is shown in Sen (1972) that quantile estimates, based on order statistics, have a normal limiting distribution and are asymptotically unbiased, if certain conditions are satisfied.

For the case of ϕ-mixing sequences, quantile estimation is much more difficult than in the independent case. The usual order-statistic point estimate, $\hat{x}_p$, is still asymptotically unbiased; however, its variance is inflated by a factor of SSVC$/p(1-p)$ (Sen, 1972), where

$$\text{SSVC} = C_0 + 2\lim_{n\to\infty}\sum_{k=1}^{n-1}(1-k/n)C_k$$

is the steady-state variance constant and $C_k = \text{Cov}[X_m, X_{m+k}]$ is the lag k covariance of the process.

Order statistics quantile estimators are asymptotically unbiased, however, in practice we must use a workable finite-sample size. Chen and Kelton (2000a,b) propose using the runs-up test on the output sequence to determine the simulation run-length. The runs-up test looks solely for independence and has been shown to be very powerful (Knuth 1998). The runs-up test is used in our procedure to determine the lag l, so that observations at least $l-1$ observations apart will be independent in statistical sense. We set α to 0.1 in the runs-up test of independence at our procedure.

3 METHODOLOGIES

This section presents the methodologies we will use for our quantile and histogram estimation. Although asymptotic results are often applicable when the amount of data is "large enough," the point at which the asymptotic results become valid generally depends on unknown factors. An important practical decision must be made regarding the sample size n required to achieve the desired precision.

3.1 Determine the Simulation Run Length

We propose a quasi-independent procedure to sequentially estimate the required sample size, and use histogram approximation to estimate quantile. The QI procedure will increase simulation run length progressively until a subsequence of n samples (taken from the original output sequence) appears to be independent, as determined by the runs-up test. We accomplish this by *systematic sampling*, i.e., select a number l, choose that observation and then every l^{th} observation thereafter. Here l will be chosen sufficiently large so that samples are statistically independent. This is possible because we assume the underlying process satisfies the ϕ-mixing conditions. We compute n the required number of independent samples and l for our systematic sampling. The minimum required sample size is then $N = nl$, i.e., the total simulation run length. In the proposed method, the variance of the quantile estimator is estimated directly from multiple quantile estimators. To avoid storing and sorting the whole output sequence, we compute sample quantiles only at certain grid points and use (four points) Lagrange interpolation (Knuth 1998) to compute the p quantile.

Chen and Kelton (1999) propose controlling the precision of quantile estimates by ensuring that the p quantile estimator

$$\hat{x}_p \in x_{[p\pm\epsilon]_0^1} \qquad (2)$$

where

$$[P]_0^1 = \begin{cases} P & \text{if } 0 \leq P \leq 1, \\ 0 & \text{if } P < 0, \\ 1 & \text{if } P > 1. \end{cases}$$

That is, we will have $1-\alpha_1$ confidence that the p quantile estimator $\hat{x}_p$ is between the $[p-\epsilon]_0^1$ and $[p+\epsilon]_0^1$ quantiles, i.e.,

$$\Pr[|F(\hat{x}_p) - p| \leq \epsilon] \geq 1 - \alpha_1,$$

where ϵ is the maximum proportional half-width of the confidence. The proportional half-width ϵ is dimensionless; it is a proportion value with no measurement unit and must be between 0 and $\max(p, 1-p)$, $0 < p < 1$.

Using the this precision requirement (i.e. equation (2)), the required sample size n_p for a fixed-sample-size procedure of estimating the p quantile of an i.i.d. sequence is the minimum n_p that satisfies

$$n_p \geq \frac{z^2_{1-\alpha_1/2}p(1-p)}{\epsilon^2}, \qquad (3)$$

where $z_{1-\alpha_1/2}$ is the $1-\alpha_1/2$ quantile of the standard normal distribution, ϵ is the maximum proportional half-width of the confidence interval, and $1-\alpha_1$ is the confidence level. For example, if the data are independent and we would like to have 95% confidence that the 0.5 quantile estimator has no more than $\epsilon = 0.005$ deviation from the true but unknown quantile, the required sample size is $n_p \geq \frac{1.960^2 0.5(1-0.5)}{0.005^2}$. So $n_p = 38,416$.

3.2 Histogram Approximation

We propose a simple quasi-independent algorithm and use grid points to construct histogram (multiple quantiles). We require users enter the value of the required parameters ϵ and α. The number of main grid points is computed by $GRID = 1/\epsilon$, where ϵ is the desired proportional half-width. The number of auxiliary grid points is $GRID2 = 2 \times A \times GRID + 3$, where $0 < A < 1$. Based on our experiment, we recommend $A \geq 10\%$. The total number of grid points is $G = GRID + GRID2$. For the following discussion, the value of ϵ is set to 0.01 and $A = 10\%$. Therefore, $GRID = 100$, $GRID2 = 23$, and the number of total grid points is $G = 123$. The value of the grid points $g_0, g_1, \ldots, g_{122}$ will be constructed as followings: g_0 and g_{122} are set to $-\infty$ and ∞ (i.e., the minimum and maximum of the underlying computer) respectively.

If the analyst knows what may be the minimum or maximum values of the distribution, those values should be used. For example, the waiting-time of any queuing systems can not be negative, the analyst should enter 0 as the minimum. Grid point g_{11} is set to the minimum of the initial n or $2n$ samples, depending on whether the data appear to be independent, as determined by the runs-up test. Grid points g_{i+11}, $i = 1, 2, \ldots, 100$, are set to the $i\%$ quantile of the initial n or $2n$ samples, depending on whether the data appear to be independent. We will set grid points g_1 through g_{10} and g_{112} through g_{121} to appropriate values so that g_1 through g_{12} will have the same segment length and g_{110} through g_{121} will have the same segment length. Therefore, the grids will be dense where the probability distribution has high values and will be sparse where the probability distribution has low values. A corresponding array of $n_0, n_1, \ldots, n_{122}$ is used to store the number of observations between two consecutive grid points. For example, the number of observations between g_{i-1} and g_i is stored in n_i.

The initial sample size n is computed according to equation (3), however, if $n < 4000$ then $n = 4000$ will be used. The simulator will generate $n \geq 4000$ (the minimum recommended sample size for the runs-up test) observations initially (0^{th} iteration). For the following discussion, we assume $n = 4000$. We allocate a buffer A with size $t = 12,000$ ($3 \times n$) to store our QI sequence. We then carry out a runs-up test with these 4000 samples $y_1, y_2, y_3, \ldots, y_{4000}$, as our initial ($0^{th}$) iteration. See chen and Kelton (2000b) for detail on how sample sizes are determined.

The following shows the incremental sample sizes at each iteration:

0	1_A	1_B
4000(4000)	4000(8000)	4000(12000)
$0:1_B$	2_A	2_B
12000(6000)	4000(8000)	8000(12000)
$0:2_B$	3_A	3_B
24000(6000)	8000(8000)	16000(12000)
...		

The equation inside the parenthesis shows the number of samples stored in buffer A. For example, at the beginning of the 2_A^{th} iteration. There are 6000 samples in the buffer and each sample is the lag 2 observations, thus, there are 12000 observations in total. At the 2_A^{th} iteration, we generate another 4000 observations and store 2000 samples that are lag 2 observations in the buffer. Therefore, at the end of the 2_A^{th} iteration there are 8000 samples in the buffer. Note that the entire observations are used to compute the quantile estimates. We discard samples in the QI sequence so that the size of the QI sequence will be no more than

t. Samples in the QI sequence are used by the runs-up test to determine sample sizes and are not used to compute the quantile estimates. The information in the grid points will be updated each time a new observation x_i is generated. The number stored in n_i is increased by one if $g_{i-1} < x_i \le g_i$.

Theoretically the sample quantile of the QI subsequence will be an unbiased quantile estimate. Our experimental results confirm this. Because samples in the QI subsequence are statistically independent, consequently, the variance of a mean estimator can be computed indirectly from the variance of individual samples. One disadvantages of using only the QI subsequence for quantile estimation is that we will waste lots observations for highly correlated data. Moreover, for extreme quantiles ($p > 0.95$), the quantile variance estimators are biased low. This maybe because the extreme values are not captured as frequently as they happened.

Once the QI algorithm have determined the sample size is large enough for the required precision, we can then compute the point quantile estimator by Lagrange interpolation of the quantile at four grid points. The array $n_i, i = 1, 2, \ldots, G-1$ stores the number of observations between grid points g_{i-1} and g_i, therefore, the quantile of g_i at the grid point i can be computed by $p_i = \sum_{j=1}^{i} n_j/N$, for $i = 1, 2, \ldots, G-1$, where $N = \sum_{j=1}^{G-1} n_j$ is the number of all observations. Thus, for some k such that $p_{k-1} < p \le p_k$, the p quantile point estimator can be computed as follows: Let

$$w_j = \prod_{j'=1, j' \ne j}^{4} \frac{p - p_{k+j'-3}}{p_{k+j-3} - p_{k+j'-3}}, \text{ for } j = 1, 2, 3, 4,$$

then

$$\hat{x}_p = \sum_{j=1}^{4} w_j g_{k+j-3}. \tag{4}$$

In two extreme cases, $p_0 < p \le p_1$ or $p_{G-2} < p \le p_{G-1}$, linear interpolation will be used.

Because we are estimating quantiles of stochastic systems, it is unreliable to make inference based on only one output sequence. Therefore, we will run R (we use 3 in our algorithm) replications to get R quantile estimators. Let $\hat{x}_{p,r}$ denote the estimator of x_p in the r^{th} replication. We use

$$\bar{\hat{x}}_p = \frac{1}{R} \sum_{r=1}^{R} \hat{x}_{p,r} \tag{5}$$

as a point estimator of x_p. Assuming the asymptotic approximation is valid with the simulation run length determined by our procedure, then each $\hat{x}_{p,r}$ has a limiting normal distribution, by the central limit theorem a confidence interval (CI) for x_p using the i.i.d. $\hat{x}_{p,r}$'s can be approximated using standard statistical procedures. That is, the ratio

$$T = \frac{\bar{\hat{x}}_p - x_p}{S/\sqrt{R}}$$

would have an approximate t distribution with $R-1$ d.f. (degrees of freedom), where

$$S^2 = \frac{1}{(R-1)} \sum_{r=1}^{R} (\hat{x}_{p,r} - \bar{\hat{x}}_p)^2$$

is the usual unbiased estimator of $\sigma_p^2(n)$, the variance of x_p. This would then lead to the $100(1-\alpha_2)\%$ CI, for x_p,

$$\bar{\hat{x}}_p \pm t_{R-1,1-\alpha_2/2} \frac{S}{\sqrt{R}}, \tag{6}$$

where $t_{R-1,1-\alpha_2/2}$ is the $1-\alpha_2/2$ quantile for the t distribution with $R-1$ d.f. ($R \ge 2$).

This confidence interval estimator is approximately valid when the sample size N becomes large because the quantile estimator $\hat{x}_{p,1}, \hat{x}_{p,2}, \ldots, \hat{x}_{p,R}$ become almost normally distributed (from the theorem of Sen (1972) for ϕ-mixing sequences) and become almost independent (since the process satisfy the ϕ-mixing conditions). Our QI procedure addresses the problem of determining the simulation run length that is required to satisfy the assumptions of independence and normality of the quantile estimate. Theoretically, if these assumptions are satisfied, then the actual coverage of the CI's should be close to the pre-specified level. However, we are not sure whether the asymptotic approximation is valid, therefore, the CI constructed by equation (6) may have coverage less than specified. On the other hand, the quantile estimators should satisfy the precision requirement of equation (2).

Let the half-width

$$H = t_{R-1,1-\alpha_2/2} \frac{S}{\sqrt{R}},$$

the final step in the QI procedure is to determine whether the CI meets the user's half-width requirement, a maximum absolute half-width ϵ' or a maximum relative fraction r of the magnitude of the final point quantile estimator $\bar{\hat{x}}_p$. If the relevant requirement

$$H \le \epsilon', \tag{7}$$

or

$$H \le r|\bar{\hat{x}}_p| \tag{8}$$

for the precision of the confidence interval is satisfied, then the QI procedure terminates, return the point quantile estimator $\hat{x}_p$ and the CI with half-width H. If the precision requirement (7) or (8) is not satisfied with R replications, then the QI procedure will increase the number of replications by one. This step can be repeated iteratively until the pre-specified half-width is achieved.

For large sample sizes, it becomes impractical to store and sort the entire sequence. These limitations can be overcome by using the proposed histogram approximation, which computes quantiles only at grid points and uses quasi-independent algorithm to determine the required simulation run length. Savings in storage and sorting are substantial for our method. The proposed histogram approximation method can estimate multiple quantiles simultaneously without much extra effort. However, the simulation run length and run time will grow quickly because the *Bonferroni inequality* is used. Because histogram was used to estimate quantiles, we can use the histogram to generates an empirical distribution of the output sequence. We can then estimate the $1-\alpha_2$ confidence interval of any quantile separately, i.e., without claiming that all CIs are satisfied with $1-\alpha_2$ confidence simultaneously.

The quasi-independent algorithm uses runs-up test to determine the simulation run length, which has strong theoretical basis. The QI algorithm is easy to implement because the sample sizes (not the simulation run length) is determined at the beginning of the procedure. This method works well and is a solid practical procedure.

4 EMPIRICAL EXPERIMENTS

In this section we present some empirical results obtained from simulations using the quasi-independent procedure proposed in this paper. The purpose of the experiments was not only to test the methods thoroughly, but also to demonstrate the interdependence between the correlation of simulation output sequences and simulation run lengths, and the validity of our methods. We tested two stochastic models: AR(1), and M/M/1 delay-in-queue processes. In these experiments, no relative precision or absolute precision were specified, therefore, the half-width of the CI is the result of the default precision. In all experiments, we conservatively set the required parameters of determining the simulation run length (i.e. equation (2)) with $p = 0.5$, $\epsilon = 0.5\%$, and $\alpha_1 = 0.05$. The confidence level α_2 of the quantile CI (i.e. equation (6)) is set to 0.1.

A stochastic model that is covariance stationary and admits an exact analysis of performance criteria is the *first-order auto-regressive* (AR(1)) process, generated by the recurrence relation

$$X_i = \mu + \varphi(X_{i-1} - \mu) + \epsilon_i \quad \text{for} \quad i = 1, 2, \ldots,$$

where

$$E(\epsilon_i) = 0, \quad E(\epsilon_i \epsilon_j) = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases},$$

$$0 < \varphi < 1,$$

and X_0 is specified to be some constant x_0. The ϵ_i's are commonly called *error terms.*

The AR(1) process shares many characteristics observed in simulation output processes, including first- and second-order stationarity, and autocorrelations that decline exponentially with increasing lag. If we make the additional assumption that the ϵ_i's are normally distributed, since we have already assumed that they are uncorrelated, they will now be independent as well, i.e., the ϵ_i's are i.i.d. $\mathcal{N}(0, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. It can be shown that X has asymptotically a $\mathcal{N}(0, \frac{1}{1-\varphi^2})$ distribution, and the steady-state variance constant of the AR(1) process is $1/(1-\varphi)^2$.

We tested two AR(1) models with φ set to 0.75, and 0.90 respectively. μ is set to zero for these two models. In order to eliminate the initial bias, X_0 is set to a random variate drawn from the steady-state distribution. The summary of our experimental results of the AR(1) process is listed in Tables 1 and 2. Each design point is based on 100 replications. The p row lists the quantile we want to estimate. The *quantile* row lists the true p quantile value. The *cover p* row lists the percentage of the estimators satisfy equation (2). The *coverage* row lists the percentage of the CIs that cover the true p quantile value. The *avg. rp* row lists the average of the relative precision of the quantile estimators. Here, the relative precision is defined as $rp = \text{abs}(\hat{x}_p - x_p)/\hat{x}_p$, where $\text{abs}(x)$ is the absolute value of x. The *stdev rp* row lists the standard deviation of the relative precision of the quantile estimators. The *avg. hw* row lists the average of the confidence interval half-width. The *stdev hw* row lists the standard deviation of the CI half-width. The *avg. samp* row lists the average of the sample size of each independent run. The *stdev samp* row lists the standard deviation of the sample size.

All quantile estimators satisfy the precision requirement of equation (2). The CI coverage of these four design points are around the specified 90% confidence level. For the 0.5 quantile estimates, the average relative precision is 1.0 and the standard deviation of the relative precision is 0, because the parameter under investigation $x_{0.5}$ is 0. The simulation run length increases as the correlation coefficient φ of the AR(1) process increases. For the AR(1) process, the PDF (Probability Distribution Function) values decrease as the quantile increases from 0.50 to 0.95. Therefore, the average half-width increases as quantile increases. This is because with the same amount of quantile deviation, the coverage deviation is smaller where PDF is smaller.

Table 1: Coverage of 90% Confidence Quantile Estimators for the AR(1) Process with $\varphi = 0.75$

Precision	Correlation Coefficient φ 0.75	
p	0.50	0.95
quantile	0.0	2.487326
cover p	100%	100%
coverage	92%	90%
avg. rp	1.0	0.002022
stdev rp	0.0	0.001467
avg. hw	0.013044	0.018933
stdev hw	0.007366	0.010497
avg. samp	247352	
stdev samp	43389	

Table 2: Coverage of 90% Confidence Quantile Estimators for the AR(1) Process with $\varphi = 0.90$

Precision	Correlation Coefficient φ 0.90	
p	0.50	0.95
quantile	0.0	3.774373
cover p	100%	100%
coverage	90%	89%
avg. rp	1.0	0.002007
stdev rp	0.0	0.002157
avg. hw	0.020172	0.026635
stdev hw	0.011936	0.015294
avg. samp	689197	
stdev samp	118128	

Figure 1 shows the empirical distributions of the AR(1) process with $\varphi = 0.90$, generated from the first run of our experiments. The theoretical steady-state distribution of this AR(1) process is $\mathcal{N}(0, 1/0.19)$. This graph shows that the histogram estimate provides an excellent approximation of the underlying steady-state distribution. The empirical distribution reveals all the essential distributional features of output random variables under estimation and can provide valuable insights.

The average lag of the AR(1) process with $\varphi = 0.75$ to appear independent is around 6.44. The required sample size to estimate 0.50 quantile of an independent sequence is 38416. Consequently, the average simulation run length is around 38416 × 6.44 = 247399, which is much smaller than the theoretical required sample size of 2458624 ($\geq \frac{1.960^2 \text{SSVC}}{0.005^2}$, where SSVC $= 1/(1-\varphi)^2 = 16$). Despite our simulation run length is much smaller than the theoretical sample size, we still get good results. We believe this is because the theoretical sample size is for the worst-case scenario and too conservative.

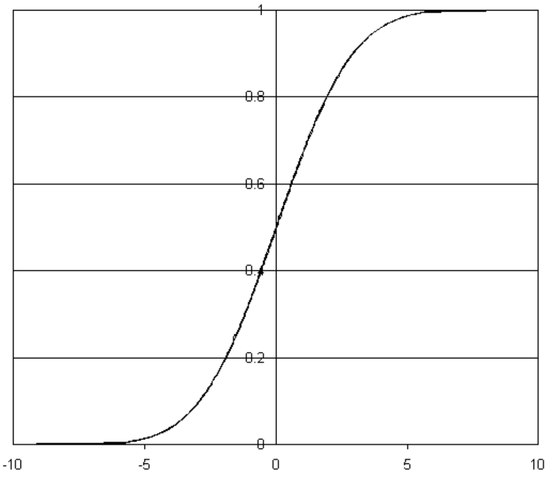

Figure 1: The Empirical Distribution of the AR(1) Process with $\varphi = 0.90$

Queuing systems are usually positively correlated and often strongly so. Furthermore, the skewness of the exponential distribution causes exponential-servers queuing models to yield output data that might be considered "worst case." Some feel that if an output-analytic method works well for an exponential-servers model, it is likely to work well in practice. We tested three M/M/1 queuing models. The service rate (ν) is set to 1.0 per period for these three models. The arrival rate (λ) is set to 0.50, 0.75, and 0.90 per period respectively. The steady-state variance constant of the waiting time of the M/M/1 delay-in-queue process is SSVC $\approx C\left(\frac{(1+\rho)}{(1-\rho)} + \frac{2\rho(3-\rho)}{(2-\rho)(1-\rho)^2}\right)$, where $C = \frac{\rho^3(2-\rho)}{\lambda^2(1-\rho)^2}$ and $\rho = \lambda/\nu$ is the traffic intensity (Daley, 1968).

Let $\{A_n\}$ denote the interarrival-time i.i.d. sequence and $\{S_n\}$ denote the service-time i.i.d. sequence. Then the waiting-time sequence $\{W_n\}$ is defined by

$$W_{n+1} = \ (W_n + S_n - A_{n+1})^+ \quad \text{for} \quad n \geq 1,$$

where $w^+ = \max(w, 0)$. In order to eliminate the initial bias, w_1 is set to a random variate drawn from the steady-state distribution. Because the waiting-time distribution function of a stationary M/M/1 delay-in-queue is $F(x) = 1 - \frac{\lambda}{\nu}e^{-(\nu-\lambda)x}$, the quantiles for M/M/1 delay-in-queue are applicable only when the estimated quantiles are large than or equal to $(1-\lambda/\nu)$. The waiting time of a stationary M/M/1 delay-in-queue distribution function $F(x)$ is discontinuous at $F(x) = 1-\lambda/\nu$, (i.e. $x = 0$). Therefore, it is useful before

Table 3: Coverage of 90% Confidence Quantile Estimators for the M/M/1 Delay-in-Queue Process with $\rho = 0.50$

Precision	Traffic Intensity ρ	
	0.50	
p	0.50	0.95
quantile	0.00	4.605170
cover p	100%	100%
coverage	88%	94%
avg. rp	1.0	0.006305
stdev rp	0.0	0.004085
avg. hw	0.006622	0.093919
stdev hw	0.004485	0.041834
avg. samp	172839	
stdev samp	31686	

Table 4: Coverage of 90% Confidence Quantile Estimators for the M/M/1 Delay-in-Queue Process with $\rho = 0.75$

Precision	Traffic Intensity ρ	
	0.75	
p	0.50	0.95
quantile	1.621860	10.832200
cover p	100%	100%
coverage	87%	90%
avg. rp	0.005018	0.004979
stdev rp	0.003736	0.003977
avg. hw	0.024738	0.178979
stdev hw	0.015167	0.099925
avg. samp	938536	
stdev samp	196939	

Table 5: Coverage of 90% Confidence Quantile Estimators for the M/M/1 Delay-in-Queue Process with $\rho = 0.90$

Precision	Traffic Intensity ρ	
	0.90	
p	0.50	0.95
quantile	5.877865	28.90370
cover p	100%	100%
coverage	93%	92%
avg. rp	0.003136	0.004953
stdev rp	0.002631	0.003790
avg. hw	0.058634	0.446812
stdev hw	0.032374	0.243081
avg. samp	7355313	
stdev samp	1678455	

conducting an informative experiment to know whether a desired quantile is attainable.

The summary of our experimental results of the M/M/1 delay-in-queue process is summarized in Tables 3 through 5. Again, all quantile estimators satisfy the precision requirement of equation (2) and CI coverages are above or close to the specified 90%. We experienced some problems when estimating the 0.50 quantile of the M/M/1 queuing process with $\rho = 0.5$ ($\rho = \lambda/\nu$ is the traffic intensity), because the distribution is not continuous at the true quantile value 0. For example, if the histogram indicates that more than 50% of the distribution is less than or equal to zero, the algorithm will interpolate between $-\infty$ and 0 to estimate the 0.5 quantile. Therefore, some of the 0.5 quantile estimators of the M/M/1 queuing process with $\rho = 0.5$ will have very large negative values. To correct this problem, we assume that the analyst knows the underlying distribution is non-negative and use 0 instead of $-\infty$ as the minimum in the algorithm for this estimation. This is a very reasonable assumption because the waiting-time can not be negative. Furthermore, for the 0.5 quantile estimates of the M/M/1 delay in queue with $\rho = 0.5$, the average relative precision is 1.0 and the standard deviation of the relative precision is 0, because the parameter under investigation $x_{0.5}$ is 0.

The average sample size increases as the traffic intensity increases. Because the PDF values of the steady-state distribution of M/M/1 delay-in-queue also decrease as quantile increases from 0.50 to 0.95, the rest of the results have the same implications as the results from the AR(1) process. The average CI half-width of the 0.95 quantiles of the M/M/1 delay in queue are much larger than those of the 0.50 quantiles, this is because the quantile under estimation has larger value. Figure 2 shows the empirical distributions of the M/M/1 delay-in-queue process with $\rho = 0.90$, generated from the first run of our experiments. The theoretical steady-state distribution of this M/M/1 queuing process is $1 - 0.9e^{-0.1x}$, where $x \geq 0$. Again, our experimental results show that the histogram estimate provides an excellent approximation of the underlying steady-state distribution.

The average lag of the M/M/1 delay-in-queue sequence with $\rho = 0.75$ to appear independent is around 24.43. The required sample size to estimate 0.50 quantile of an independent sequence is 38416. Consequently, the average simulation run length is about $38416 \times 24.43 = 938502$, which is much smaller than the theoretical required sample size of 115708992 ($\geq \frac{1.960^2\text{SSVC}}{0.005^2}$, SSVC = 753).

5 CONCLUSIONS

We have presented an algorithm for estimating the histogram and quantile x_p of a stationary process. Some quantile es-

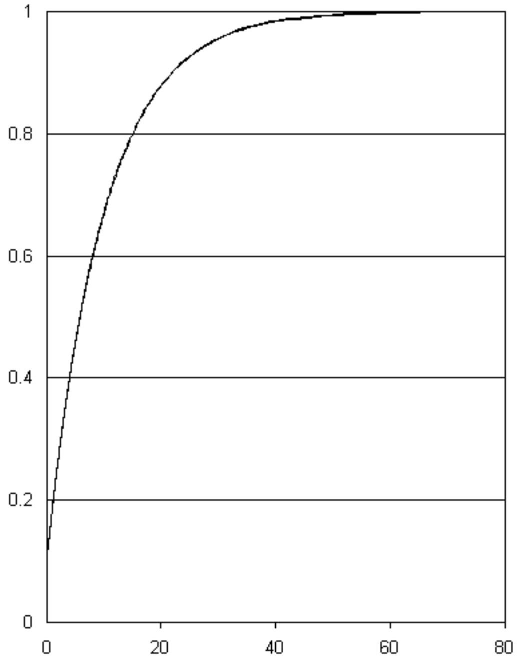

Figure 2: The Empirical Distribution of the M/M/1 Queuing Process with $\rho = 0.90$

timates require more observations than others before the asymptotics necessary for quantile estimates to become valid. Our proposed quasi-independent algorithm works well in determining the required simulation run length for the asymptotic approximation to become valid. The results from our empirical experiments show that the procedure is excellent in achieving the pre-specified accuracy. However, the variance of the simulation run length from our sequential procedure is large when estimating highly correlated sequence. This is not only because of the randomness of the output sequence, but also because we double the lag length l every two iterations. Because the sample size grows rapidly at later iterations, further research is then to develop new algorithms so that the simulation run length does not need to be doubled every two iterations.

Our proposed histogram approximation algorithm computes quantiles only at grid points and use Lagrange interpolation to estimate p quantile. The algorithm also generates an empirical distribution (histogram) of the output sequence, which can provide insights of the underlying stochastic process. Savings in storage and sorting are substantial for our method. Our approach has the desirable properties that it is a sequential procedure and it does not require users to have *a priori* knowledge of values that the data might assume. This allows the user to apply this method without having to execute a separate pilot run to determine the range of values to be expected, or guess and risk having to re-run the simulation. Either of these options represents potentially large costs to the user because many realistic simulations are time-consuming to run. The main advantage of our approach is that by using a straightforward runs-up test to determine the simulation run length and obtain quantiles at grid points, we can apply classical statistical techniques directly and do not require more advanced statistical theory, thus making it easy to understand, simple to implement, and fast to run. The simplicity of this method should make it attractive to simulation practitioners and software developers.

Moreover, the proposed procedure can be used to estimate proportions (Chen 2001). For example, let p_q denote the proportion of customers wait no more than q minutes in a queue before they get served, we can find some k such that $g_{k-1} < q \le g_k$, the proportion of customers wait no more than q minutes can be estimated by: Let

$$w_j = \prod_{j'=1, j' \neq j}^{4} \frac{q - g_{k+j'-3}}{g_{k+j-3} - g_{k+j'-3}}, \text{ for } j = 1, 2, 3, 4,$$

then $\hat{p}_q = \sum_{j=1}^{4} w_j p_{k+j-3}$, where g_i and p_i are computed as discussed in Section 3.

Chen and Kelton (2000b) point out that the modified runs-up test also works well for discrete distributions. Thus, our method can also be used to determine the required simulation run length to obtain discrete distributed quantile estimate that satisfies a pre-specified precision. However, the quantile estimate is computed through interpolation so it may not be a valid value of the underlying discrete distribution. If the output data can be read through again, then a valid value can be estimated. In the first phase we obtained lower and upper bound of the quantile. When we read the data again in the second phase, we will count the number of observations that are less than the lower bound, record the values that are between lower and upper bounds and count the number of observations in each of those values. For example, if there are N observations in total, n_0 observations are less than the lower bound, k values (X_i, $i = 1, 2, \ldots, k$) are between the lower and upper bound and their corresponding number of observations are n_i (i.e., n_i is the number of observations having the value X_i), then the p quantile will be the value X_j such that $Np \le \sum_{i=0}^{j} n_i$.

The QI algorithm can also be used to estimate the variance and batch size of batch means methods for estimating the mean of stochastic process output sequences, for detail see Chen and Kelton (2000a,b).

ACKNOWLEDGMENTS

We would like to thank Dr. Shane Henderson for his suggestions, which enlighten this research.

REFERENCES

Billingsley, P. 1999. *Convergence of Probability Measures*. 2nd ed. New York: John Wiley & Sons, Inc.

Chen, E. J. 2001. Proportion Estimation of Correlated Sequences. *Simulation*. To appear.

Chen, E. J., and W. D. Kelton. 1999. Simulation-Based Estimation of Quantiles. *Proceedings of the 1999 Winter Simulation Conference*, ed. P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, 428–434. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Chen, E. J., and W. D. Kelton. 2000a. Mean Estimation Based on Phi-Mixing sequences. In *Proceedings of the* 33*rd* *Annual Simulation Symposium*, ed. D.C. Young, 237–244. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Chen, E. J., and W. D. Kelton. 2000b. A Stopping Procedure Based on Phi-Mixing Conditions. *Proceedings of the 2000 Winter Simulation Conference*, ed. J.A. Joines, R. Barton, P. Fishwick, and K. Kang. 617–626. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Daley, D. J. 1968. The Serial Correlation Coefficients of Waiting Times in a Stationary Single Server Queue. *The Journal of the Australian Mathematical Society*, Vol. 8, Part 4, 683–699.

David, H.A. 1981. *Order Statistics*. 2nd ed. Wiley, New York.

Hurley, C., and R. Modarres. 1995. Low-Storage Quantile Estimation. *Computational Statistics*. 10:311–325.

Iglehart, D. L. 1976. Simulating Stable Stochastic Systems; VI. Quantile Estimation. *J. Assoc. Comput. Mach.* 23:347–360.

Knuth, D. E. 1998. *The Art of Computer Programming*. Vol. 2. 3rd ed. Reading, Mass.:Addison-Wesley.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. 3rd ed. New York:McGraw-Hill.

Seila, A. F. 1982a. A Batching Approach to Quantile Estimation in Regenerative Simulations. *Management Science*. 28. No. 5:573–581.

Seila, A. F. 1982b. Estimation of Percentiles in Discrete Event Simulation. *Simulation*. 39. No. 6:193–200.

Sen, P. K. 1972. On the Bahadur Representation of Sample Quantiles for Sequences of ϕ-mixing Random Variables. *Journal of Multivariate Analysis*. 2. No. 1:77–95.

AUTHOR BIOGRAPHIES

E. JACK CHEN is a Senior Staff Specialist with BASF Corportation. He received an M.S. in computer science from Syracuse University, an M.B.A. from Northern Kentucky University, and a Ph.D. degree from the University of Cincinnati. His research interests are in the area of computer simulation. His email and web addresses are `<chenej@basf.com>` and `<www.econqa.cba.uc.edu/~chenj>`.

W. DAVID KELTON is a Professor in the Department of Quantitative Analysis and Operations Management at the University of Cincinnati. He received a B.A. in mathematics from the University of Wisconsin-Madison, an M.S. in mathematics from Ohio University, and M.S. and Ph.D. degrees in industrial engineering from Wisconsin. His research interests and publications are in the probabilistic and statistical aspects of simulation, applications of simulation, and stochastic models. He is co-author of *Simulation Modeling and Analysis* (3d ed., 2000, with Averill M. Law), and *Simulation With Arena* (1998, with Randall P. Sadowski and Deborah A. Sadowski), both published by McGraw-Hill. Currently, he serves as Editor-in-Chief of the *INFORMS Journal on Computing*, and has been Simulation Area Editor for *Operations Research*, the *INFORMS Journal on Computing*, and *IIE Transactions*, as well as Associate Editor for *Operations Research*, the *Journal of Manufacturing Systems*, and *Simulation*. From 1991 to 1999 he was the INFORMS co-representative to the Winter Simulation Conference Board of Directors and was Board Chair for 1998. In 1987 he was Program Chair for the WSC, and in 1991 was General Chair. His email and web addresses are `<david.kelton@uc.edu>` and `<www.econqa.cba.uc.edu/~keltond>`.