Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

TWO-PHASE QUANTILE ESTIMATION

E. Jack Chen

BASF Corporation
3000 Continental Drive - North
Mount Olive, NJ 07828-1234, U.S.A.

ABSTRACT

This paper discusses the implementation of a two-phase procedure to construct confidence intervals for a simulation estimator of the steady-state quantiles of stochastic processes. We compute sample quantiles at certain grid points and use Lagrange interpolation to estimate the p quantile. The algorithm dynamically increases the sample size so that quantile estimates satisfy the proportional precision at the first phase and the relative or absolute precision at the second phase. We show that the procedure gives asymptotically unbiased quantile estimates. An experimental performance evaluation demonstrates the validity of using grid points and the quasi-independent procedure to estimate quantiles.

1 INTRODUCTION

Simulation studies have been used to investigate system characteristics, such as the mean and the variance of certain system performance of the system under study. In this paper, we propose a method to construct an empirical distribution and estimate quantiles of the parameter of interest. For $0 < p < 1$, the *p quantile* (percentile) of a distribution is the value at or below which $100p$ percent of the distribution lies. Related to quantiles, a *Histogram* is a graphical estimate of the underlying probability density (mass) function and reveals all the essential distributional features of random variables analyzed by simulation. A histogram can be constructed with a properly selected set of quantiles.

Wood and Schmeiser (1995) have experimented on overlapping batch quantiles, and concluded that large sample sizes and batch sizes are needed to obtain reliable standard error estimators with overlapping batch quantiles, even when data are independent and identically distributed (i.i.d.). We propose a simple *Quasi-Independent* (QI) algorithm (see Chen and Kelton 2000) to determine the simulation run length and use grid points to construct a 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 i.i.d. and ϕ*-mixing* sequences, sample quantiles will be asymptotically unbiased if certain conditions are satisfied. Intuitively, a stochastic process is ϕ-mixing if its distant future event is essentially independent of its present and past events (Billingsley 1999). However, in practical situations, simulation experiments are restricted in time, and the required simulation run length for the estimator to be unbiased is not known in advance. Moreover, the variance of the quantile estimator needs to be estimated in order 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.

Chen and Kelton (2001) propose a histogram approximation based on a QI procedure to construct a proportional half-width (see Section 2.2) confidence interval (CI) of quantile estimates for stationary simulation outputs. In this paper, we extend the procedure to construct absolute or relative precision half-width CI. The algorithm will sequentially determine the simulation run length so that quantile estimates satisfy the proportional precision at the first phase and the relative or absolute precision at the second phase. The proposed procedure produces asymptotically valid quantile estimates. The asymptotic validity of our QI procedure occurs as the sequence appears to be independent, as determined by the *runs-up* test (see Knuth 1998).

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 would then require 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-experimental 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, quantile and histogram estimation, and proportion estimation.

2.1 Order Statistics Quantile Estimators

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):

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

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

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 $\| L_n \| \le k|x_n|$ for each $n \ge N$. $\| L_n \|$ denotes the Euclidean norm of L_n.

Let

$$\sigma_p^2(n) = \frac{p(1-p)}{nf^2(x_p)}.$$

Then

$$\frac{(\hat{x}_p - x_p)}{\sigma_p(n)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

as $n \to \infty$, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2, and $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

Quantile estimation based on order statistics not only can be used 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. However, 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 $\mathrm{SSVC}(x_p)/p(1-p)$ (Sen, 1972), where

$$\mathrm{SSVC}(x_p) = C_0 + 2\lim_{n\to\infty}\sum_{k=1}^{n-1}(1-k/n)C_k \tag{3}$$

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

2.2 Quantile and Histogram Estimation

Let

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

and ϵ be the proportional half-width of the $1-\alpha_1$ confidence interval. The proportional half-width ϵ is dimensionless; it is a proportion value with no measurement unit and must be between 0 and $\max(p, 1-p)$, $0 < p < 1$.

Chen and Kelton (1999) propose controlling the precision of quantile estimates by ensuring that the p quantile estimator $\hat{x}_p$ is between the $[p-\epsilon]_0^1$ and $[p+\epsilon]_0^1$ quantiles with a desired confidence, i.e.,

$$\Pr[\hat{x}_p \in x_{[p\pm\epsilon]_0^1}] \ge 1-\alpha_1, \tag{4}$$

or equivalently

$$\Pr[|F(\hat{x}_p) - p| \le \epsilon] \ge 1-\alpha_1.$$

Using this precision requirement (i.e. equation (4)), 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

$$n_p \ge \frac{z^2_{1-\alpha_1/2}p(1-p)}{\epsilon^2} \tag{5}$$

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.

When data are correlated, the QI procedure will progressively increase simulation run length until a sequence of n samples (taken from the original output sequence) appears to be independent, as determined by the runs-up tests. Briefly, a runs up is a monotonically increasing subsequence, and the length of a runs up is known as the run length. We examine the proportion of different run lengths to check if the sequence appears to be independent. We obtain QI samples by *systematic sampling*, i.e., select a number l, choose that observation and then every l^{th} observation thereafter. Here, the chosen l will be sufficiently large so that samples are statistically independent. This is possible because we assume the underlying process satisfies the ϕ-mixing conditions.

Chen and Kelton (2001) detail the steps of the quantile and histogram estimation. Briefly, they compute n the required number of independent samples with equation (5) and l for the systematic sampling. The minimum required sample size is then $N = nl$, i.e., the total simulation run length. The starting value of l is 1, and it increases by 1 at the first two iterations; thereafter, it is doubled at every two iterations. To avoid storing and sorting the whole output sequence, they compute sample quantiles only at certain grid points and use (four points) Lagrange interpolation (Knuth, 1998) to compute the p quantile. The grid points are strategically allocated such that every grid should contain no more than ϵ of the distribution. That is, the grid will be small where the probability density is high and will be large where the probability density is low.

There are a certain number of main and auxiliary grid points. The number of main grid points is computed by $G_m = \lceil 1/\epsilon \rceil$, where ϵ is the desired proportional half-width. Let $b+1$ be the index of the first main grid points. Grid point g_b is set to the minimum of the initial n, $2n$, or $3n$ samples, depending on the degree of the correlation of the sequence, as determined by the runs-up test. Grid points g_{b+i}, $i = 1, 2, \ldots, G_m$, are set to the i/G_m quantile of the initial n, $2n$, or $3n$ samples, depending on the degree of correlation of the sequence.

2.3 Proportion Estimation

Chen (2001) modified the quantile estimation procedure in the previous section to estimate proportions of correlated sequences. When a proportion is estimated, we are interested in the probability p that the random variable X belongs to a pre-specified field ω: $p = \Pr\{X \in \omega\}$. An estimate of p is based on a transformation of the output sequence $\{X_j\}$, $j = 1, 2, \ldots, n$:

$$\hat{p} = \frac{1}{n}\sum_{j=1}^{n} I_j,$$

where

$$I_j = \begin{cases} 1 & \text{if } X_j \in \omega, \\ 0 & \text{otherwise.} \end{cases}$$

For data that are i.i.d., the following properties of I_j are well known (Hogg and Craig, 1995, pp. 116-117): $E(I_j) = p$ and $Var(I_j) = p(1-p)$. Moreover, $E(\hat{p}) = p$ and $Var(\hat{p}) = p(1-p)/n$. Thus, an exact confidence interval for the estimated proportion $\hat{p}$ can be obtained using the binomial distribution. However, we cannot assume that the I_j's are independent. Instead, we assume that the sequence $\{I_j\}$ is covariance stationary. In this case

$$\mathrm{SSVC}(I_j) = p(1-p)(1 + 2\lim_{n\to\infty}\sum_{k=1}^{n-1}(1-k/n)\rho_k),$$

and

$$\mathrm{SSVC}(p) = \mathrm{SSVC}(I_j)/n, \tag{6}$$

where $\rho_k = \mathrm{Cor}(I_j, I_{j+k})$ is the correlation coefficient between I_j and I_{j+k}, see Law and Kelton (2000, pp. 246-252).

Since $\hat{p}$ is based on the mean of the random variable I_j, we can use any method developed for estimating the variance of the mean to estimate $Var(\hat{p})$. The variance estimator of equation (6) reflects the transformation from estimating x_p to p. Furthermore, let $C_0 = p(1-p)/n$ and $C_k = C_0\rho_k$, then equation (6) becomes equation (3).

3 METHODOLOGIES

This section presents the methodologies we will use to construct CI of the quantile estimates such that it satisfy the absolute or relative half-width requirements, i.e.,

$$\Pr[\hat{x}_p \in x_p \pm \epsilon'] \geq 1-\alpha_2,$$

or

$$\Pr[\hat{x}_p \in x_p \pm r|x_p|] \geq 1-\alpha_2, \tag{7}$$

where $\epsilon' > 0$ and $1 > r > 0$ are, respectively, the absolute and the relative precision.

For i.i.d. sequences, if $n \geq n_p$ (of equation (5)), then the quantile estimator should satisfy the precision requirement of equations (4) or

$$\Pr[\hat{p} \in p \pm \epsilon] \geq 1-\alpha_1,$$

where $\hat{p} = F(\hat{x}_p)$. From equation (2), asymptotically

$$\text{Var}(\hat{x}_p) \approx \frac{p(1-p)}{nf^2(x_p)}.$$

Therefore, when data are i.i.d.

$$\frac{\text{Var}(\hat{x}_p)}{\text{Var}(\hat{p})} \approx 1/f^2(x_p). \tag{8}$$

Thus, if the sample size

$$n' \geq \frac{n}{f^2(x_p)},$$

then

$$\Pr[\hat{x}_p \in x_p \pm \epsilon'] \geq 1 - \alpha_1,$$

where ϵ' has the same numerical value as ϵ and has the same unit of x_p. Moreover, if ϵ'' is the desired absolute precision and

$$n' \geq \frac{n}{f^2(x_p)}(\frac{\epsilon'}{\epsilon''})^2,$$

then

$$\Pr[\hat{x}_p \in x_p \pm \epsilon''] \geq 1 - \alpha_1.$$

Note that $1 > \epsilon > 0$, and $\epsilon'' > 0$. Furthermore, if r is the desired relative precision and

$$n' \geq \frac{n}{f^2(x_p)}(\frac{\epsilon'}{r\hat{x}_p})^2,$$

then approximately

$$\Pr[\hat{x}_p \in x_p \pm r|x_p|] \geq 1 - \alpha_1.$$

Here, $|x|$ is the absolute value of x.

Theoretically, when n is large, the value of $f(x_p)$ can be approximated by finite forward differences:

$$f(x_p) \approx \frac{1}{n(\hat{x}_{p+1/n} - \hat{x}_p)} \tag{9}$$

because

$$F'(x_p) = \lim_{x \to x_p} \frac{F(x) - F(x_p)}{(x - x_p)} \approx \frac{F(x_{p+1/n}) - F(x_p)}{(x_{p+1/n} - x_p)}.$$

Alternatively, the value of $f(x_p)$ can be approximated by finite central differences:

$$f(x_p) \approx \frac{2}{n(\hat{x}_{p+1/n} - \hat{x}_{p-1/n})}$$

because

$$F'(x_p) \approx \frac{F(x_{p+1/n}) - F(x_{p-1/n})}{(x_{p+1/n} - x_{p-1/n})}.$$

However, in practice for this approximation to be good, the required size for n is not known. Chen (2002) evaluates the performance of derivative estimation with finite differences using the empirical distribution constructed with the method of Chen and Kelton (2001). Generally, the results are excellent in terms of CI coverage and relative precision. The observed relative precision of the derivative estimators of i.i.d. sequences in our experiments are all within 4%. Since the method requires that the simulation run length increases as the correlation become stronger, the observed relative precision of the derivative estimators become smaller. Furthermore, there is no significant difference between the performance of forward differences and central differences when using the histogram approximation to estimate the derivative.

The value $f(x_p)$ has great influence on the required sample size. However, since the values $f(x_p)$ and x_p are unknown, we use the estimated value $\hat{f}(\hat{x}_p)$. To be conservative, we use the value $\hat{f}(\hat{x}_p)$ such that asymptotically $\Pr[f(x_p) \geq \hat{f}(\hat{x}_p)] \geq 0.60$.

Note that when data are correlated, sample size n will be replaced by $N = nl$. Here l is the lag in the original sequence such that the QI sequence appears to be independent, as discussed in Section 2.2.

The two-phase quantile estimation algorithm:

1. Remark: ϵ is the desired proportional half-width, r is the relative precision specified by the user.
2. Use the Quantile and Histogram estimation algorithm of Chen and Kelton (2001), as discussed in Section 2.2, to obtain proportional precision quantile estimates. Save the sample size n.
3. Use the finite forward differences, i.e., equation (9), to obtain the derivative estimate $z_p = \hat{f}(\hat{x}_p)$.
4. Let $n' = \lceil \frac{n}{z_p^2}(\frac{\epsilon'}{r\hat{x}_p})^2 \rceil$.
5. If $n' > n$, increase the sample size to n', go to step 2.
6. Otherwise, the quantile estimate should already satisfy the relative precision requirement.
7. Run R replications and compute the confidence interval of the quantile estimator according to equation (11).

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

We use

$$\bar{\hat{x}}_p = \frac{1}{R}\sum_{r=1}^{R} \hat{x}_{p,r} \tag{10}$$

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 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,

$$\bar{\hat{x}}_p \pm t_{R-1,1-\alpha_2/2}\frac{S}{\sqrt{R}}, \tag{11}$$

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

This confidence interval estimator is approximately valid when the sample size becomes large since 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). Our QI procedure addresses the problem of determining the simulation run length that is required to satisfy the assumptions of normality and independence 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.

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 the quasi-independent algorithm to determine the required simulation run length. Savings in storage and sorting are substantial for our method.

To improve the precision of the second phase quantile estimation, we can put more grid points in the grid that contains the x_p quantile estimator found in the first phase and the surrounding grids before we start the second phase. Of course, the newly set up grid points need to be based on interpolations. For example, if $g_{i-1} < \hat{x}_p \leq g_i$ and the grid between g_{i-1} and g_i contains n_i observations and approximately $100p_i\%$ of the distribution, then $k+1$ new grid points can be set up between $\hat{x}_{p-p_i/2}$ and $\hat{x}_{p+p_i/2}$. Let g'_j for $j = 0, 1, \ldots, k$ be the new grid points, then $g'_j = \hat{x}_{p-p_i/2+jp_i/k}$. The array contains the number of observations between newly set up grid points $n'_j = \lfloor n_i/k \rfloor$, for $j = 1, 2, \ldots, k-1$ and $n'_k = n_i - (k-1)n'_1$, where $\lfloor z \rfloor$ denotes the integer flooring (round-down) of the real number z.

4 EMPIRICAL EXPERIMENTS

In this section, we present some empirical results obtained from simulations using the quantile estimation 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 the proposed procedure with several i.i.d. and correlated sequences. In these experiments, we use $R = 3$ (see step 7 in the algorithm) independent replications to construct CI's. We estimated four quantile points: 0.25, 0.50, 0.75, and 0.90 for each distribution and used a relative precision of 0.05 for our experiments. We conservatively set the required parameters of determining the simulation run length (i.e. equation (5)) with $p = 0.5$, $\epsilon = 0.005$, and $\alpha_1 = 0.05$. The confidence level α_2 of the quantile CI (i.e. equation (11)) is set to 0.1. Moreover, the confidence level of the runs-up test of independent is set to 90%.

4.1 Independent Sequences

We tested two independent sequences:

- Observations are i.i.d. normal with mean 0 and variance 1, denoted as $\mathcal{N}(0,1)$.
- Observations are i.i.d. negative exponential with mean 1, denoted as expon(1).

The summary of our experimental results of the i.i.d. sequences are listed in Tables 1 and 2. Each design point is based on 100 independent simulation runs. 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 quantile estimates that satisfy equation (4), i.e., the coverage deviation of the quantile estimator is within the specified value ϵ. The *coverage* row lists the percentage of the CI's that cover the true quantile value. The *avg. rp* row lists the average of the relative precision of the x_p estimators. Here, the relative precision is defined as $rp = |\hat{x}_p - x_p|/|\hat{x}_p|$. 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 absolute half-width. The *stdev hw* row lists the standard deviation of the absolute

Table 1: Coverage of 90% Confidence Quantile Estimators for the $\mathcal{N}(0,1)$ Distribution

Item	Quantile			
p	0.25	0.45	0.75	0.90
quantile	-0.674189	-0.125381	0.674189	1.28173
cover p	100%	100%	100%	100%
coverage	94%	94%	85%	90%
avg. rp	0.004832	0.010206	0.004693	0.002997
stdev rp	0.003624	0.007715	0.003680	0.002195
avg. hw	0.010011	0.004996	0.008920	0.012771
stdev hw	0.004653	0.002447	0.005363	0.006741
avg. sp	41875	175039	41875	41875
stdev sp	6267	42596	6267	6267

Table 2: Coverage of 90% Confidence Quantile Estimators for the $expon(1)$ Distribution

Item	Quantile			
p	0.25	0.50	0.75	0.90
quantile	0.287682	0.693147	1.38629	2.30258
cover p	100%	100%	100%	100%
coverage	91%	93%	91%	92%
avg. rp	0.004594	0.003241	0.002641	0.002982
stdev rp	0.003691	0.002630	0.002318	0.002186
avg. hw	0.004180	0.007144	0.013503	0.021339
stdev hw	0.001809	0.003672	0.007390	0.011029
avg. sp	41747	41747	41747	41747
stdev sp	6204	6204	6204	6204

half-width. The *avg. sp* row lists the average of the sample size in each independent replication. The *stdev sp* row lists the standard deviation of the sample size in each independent replication.

Table 1 lists the experimental results of the $\mathcal{N}(0,1)$ distribution. For the 0.5 quantile estimates, the parameter under investigation $x_{0.5}$ is 0. Since $\hat{x}_{0.5} \approx 0$, $n' = \lceil \frac{n}{z_p^2}(\frac{\epsilon'}{r\hat{x}_p})^2 \rceil$ will be very large. For example, the sample size in the first phase is 38416 ($1.96^2 \times 0.5 \times 0.5/0.005^2$), the estimator $\hat{x}_{0.5} = -0.000146$, and $z_p = \hat{f}(\hat{x}_{0.5}) = 0.363357$. The required sample size is then $n' = \lceil \frac{38416}{0.363357^2}(\frac{0.005}{0.05\times 0.000146})^2 \rceil > 1.36\times 10^{11}$. It will require several days for common desktop computers to obtain one estimator. If users know that the true quantile value $x_p \approx 0$, then absolute precision can be used instead of relative precision. To avoid the required long execution time, we estimated the 0.45 quantile instead. The CI's coverage of 0.75 quantile is 85%, which is less than the specified nominal value of 90%. We believe this is caused by the randomness of the experiment and the half-width being too small. For example, the absolute value of the 0.25 and 0.75 quantile are the same, however, the average half-width of the 0.75 quantile estimates is only 0.008920 compares to 0.010011 of the 0.25 quantile. Furthermore, the average relative precision of the 0.75 quantile estimators is smaller than that of the 0.25 quantile estimators, those 0.75 quantile CI's that do not cover the true quantile value must miss only by a very small amount. All half-widths of the CI's, are less than $r|x_p|$, where $r = 0.05$ and are in general within 30% of $r|x_p|$.

Table 2 lists the experimental results of the expon(1) distribution. The sample sizes are the same for all four design points because all quantile estimations have $n' = \lceil \frac{n}{z_p^2}(\frac{\epsilon'}{r\hat{x}_p})^2 \rceil < n$. Since the value $Z_p = f(X_p)$ decreases as X_p increases, sample size n does not increase as quantile value increases when ϵ' is sufficiently small and relative precision is used. On the other hand, if absolute precision is used, then sample sizes will increase as the quantile value increase. In this experiment, with relative precision $r = 0.05$, the quantile estimator obtained with the first-phase sample size should satisfy both equations (4) and (7). Again, all quantile estimators satisfy the precision requirement of equation (4), and the CI coverage of these design points are above the specified 90% confidence level. Furthermore, all half-widths are less than $r|x_p|$. Since the true p quantile value increases as p increases, the average half-width increases as p increases. Because we set the confidence level of the runs-up test of independence to be 90%, independent sequences will not pass the runs-up about 10% of the times. Consequently, the first-phase sample size for independent sequences will be around $38416 \times 1.1 = 42258$.

4.2 Correlated Sequences

We tested four correlated sequences:

- Steady-state of the *first-order moving average* process, generated by the recurrence relation

$$X_i = \mu + \epsilon_i + \theta\epsilon_{i-1} \quad \text{for} \quad i = 1, 2, \ldots,$$

  where ϵ_i is i.i.d. $\mathcal{N}(0,1)$ and $0 < \theta < 1$. This process is denoted as MA1(θ). μ is set to 0 in our experiments. It can be shown that X has an asymptotic $\mathcal{N}(0, 1+\theta^2)$ distribution.
- Steady-state of the *first-order auto-regressive* process, generated by the recurrence relation

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

  where ϵ_i is i.i.d. $\mathcal{N}(0,1)$, and

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

$$0 < \varphi < 1.$$

This process is denoted as AR1(φ). μ is set to 0 in our experiments. It can be shown that X has an asymptotic $\mathcal{N}(0, \frac{1}{1-\varphi^2})$ distribution.

- Steady-state of the M/M/1 delay-in-queue process with the arrival rate (λ) and the leaving rate ($\mu = 1$). This process is denoted as MM1(λ). Let W_i denote the waiting time of the i^{th} customer and $\rho = \lambda/\mu$ be the traffic intensity. Then, if $\rho < 1$, the theoretical steady-state distribution of this M/M/1 queuing process is $F(x) = P(W_i \leq x) \to 1 - \rho e^{-(\mu-\lambda)x}$ as $i \to \infty$ for all $x \geq 0$. 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 $w^+ = \max(w, 0)$.
- Steady-state of the M/M/s delay-in-queue process with the arrival rate (λ) and the leaving rate (μ). This process is denoted as MMS(λ). We set $s = 2$, $\lambda = 3$, and $\mu = 2$. The traffic intensity of this process is $\rho = \frac{\lambda}{s\mu} = 0.75$.

We tested the MA1 model with $\theta = 0.75$, the AR1 model with $\varphi = 0.75$, and the M/M/1 and M/M/2 models with the traffic intensity $\rho = 0.75$. In order to eliminate the initial bias, ϵ_0 and w_0 are set to a random variate drawn from the steady-state distribution. Because the true 0.5 quantile value for the tested MA1 and AR1 processes is 0, we estimated the 0.45 quantile to avoid an extremely large sample size.

The summary of our experimental results for MA1(0.75) and AR1(0.75) are listed in Tables 3 and 4. All quantile estimators satisfy the precision requirement of equation (4). Most of the CI coverage of these design points, except the coverage of the 0.75 quantile of the AR1(0.75) process, are around the specified 90% confidence level. We believe this is caused by the half-width being too small. For these two processes, $n' = \lceil \frac{n}{z_p^2}(\frac{\epsilon'}{r\hat{x}_p})^2 \rceil > n$ only when estimating the 0.45 quantile. The sample sizes are larger than the independent cases because the lag l for the QI sequence that appears to be independent is larger.

If $\rho < 1$, the waiting-time distribution function of a stationary M/M/1 delay in queue is discontinuous at $F(x) \to 1 - \rho$, (i.e. $x = 0$); thus, the quantiles for M/M/1 delay-in-queue are applicable only when the estimated quantiles are larger than or equal to $1 - \rho$. Therefore, it is useful to know whether a desired quantile is attainable before conducting an informative experiment.

The summary of our experimental results of the M/M/1 delay-in-queue process is summarized in Table 5. We experienced some problems when estimating the 0.25 quantile of the M/M/1 queuing process with $\rho = 0.75$, because the distribution is not continuous at the true quantile value 0. Thus, the derivative does not exist at 0.25 quantile. Because the distribution function has a jump at this quantile point, the procedure often obtains ∞ as an estimate of the derivative since $\hat{x}_p = \hat{x}_{p+1/N}$ in this case. Therefore, we estimate the 0.30 quantile instead of the 0.25 quantile. However, the procedure can return the quantile estimate obtained in the first phase. Users should then investigate if the distribution is continuous at this particular quantile. Again, all quantile estimators satisfy the precision requirement of equation (4), and CI coverages are above or close to the specified 90%. The average CI half-width of the 0.90 quantiles of the M/M/1 delay in queue is much larger than the other quantiles since the quantile under estimation has a larger value.

The summary of our experimental results of the M/M/2 delay-in-queue process is listed in Table 6. If $\rho < 1$, the theoretical steady-state distribution of this M/M/2 queuing process is $F(x) \to 1 - 9e^{-x}/14$ (Hillier and Lieberman, 2001), where $x \geq 0$. Therefore, for this M/M/2 process quantiles less than 5/14 are not attainable, we estimated 0.40 quantile instead. All estimators satisfy the probability coverage requirements. Moreover, the percentages of the CI's that cover the true quantiles are close to the specified nominal value of 90%. The sample size determined by the

Table 3: Coverage of 90% Confidence Quantile Estimators for the $MA1(0.75)$ Process

Item	Quantile			
p	0.25	0.45	0.75	0.90
quantile	-0.842737	-0.156726	0.842737	1.60216
cover p	100%	100%	100%	100%
coverage	91%	93%	90%	88%
avg. rp	0.004897	0.010006	0.004437	0.002558
stdev rp	0.003227	0.007751	0.003252	0.001905
avg. hw	0.011336	0.005633	0.010974	0.013519
stdev hw	0.005510	0.002892	0.005519	0.006688
avg. sp	80805	326563	80805	80805
stdev sp	6977	42356	6977	6977

Table 4: Coverage of 90% Confidence Quantile Estimators for the $AR1(0.75)$ Process

Item	Quantile			
p	0.25	0.45	0.75	0.90
quantile	-1.01928	-0.189558	1.01928	1.93779
cover p	100%	100%	100%	100%
coverage	91%	92%	82%	91%
avg. rp	0.003340	0.009194	0.003583	0.001861
stdev rp	0.002279	0.007469	0.002411	0.001473
avg. hw	0.012107	0.005560	0.010085	0.012185
stdev hw	0.005923	0.002606	0.005668	0.006173
avg. sp	348067	1424348	348067	348067
stdev sp	54251	344237	54251	54251

Table 5: Coverage of 90% Confidence Quantile Estimators for the $M/M/1$ Delay-in-queue Process with $\rho = 0.75$

Item	Quantile			
p	0.30	0.50	0.75	0.90
quantile	0.275972	1.62186	4.39445	8.05961
cover p	100%	100%	100%	100%
coverage	90%	91%	92%	90%
avg. rp	0.005469	0.004432	0.003583	0.003919
stdev rp	0.004258	0.003228	0.002532	0.002947
avg. hw	0.005576	0.021964	0.051527	0.105334
stdev hw	0.003185	0.012968	0.029221	0.058994
avg. sp	5491879	1363070	1363070	1363070
stdev sp	1541480	297519	297519	297519

Table 6: Coverage of 90% Confidence Quantile Estimators for the $M/M/2$ Delay-in-queue Process with $\rho = 0.75$

Precision	Traffic Intensity ρ 0.75			
p	0.40	0.50	0.75	0.90
quantile	0.068993	0.251314	0.944462	1.86075
cover p	100%	100%	100%	100%
coverage	89%	90%	88%	84%
avg. rp	0.006573	0.002424	0.004137	0.004463
stdev rp	0.005017	0.001859	0.002950	0.003152
avg. hw	0.001627	0.002265	0.012300	0.024968
stdev hw	0.000828	0.001122	0.006728	0.014283
avg. sp	7925708	1326701	1326701	1326701
stdev sp	2921767	285481	285481	285481

QI procedure is roughly the same for the waiting-time of M/M/2 and the M/M/1 delay in queue with the same traffic intensity of $\rho = 0.75$. However, the CI coverage of M/M/2 delay in queue is not as good as M/M/1. Again, we believe this is caused by the half-width being too small.

5 CONCLUSIONS

We have presented an algorithm for estimating the histogram and quantile x_p of a stationary process. Some quantile estimates require more observations than others before the asymptotics necessary for quantile estimates become valid. Our proposed quasi-independent algorithm works well in determining the required simulation run length for a valid asymptotic approximation. 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 sequences. 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 needed to develop new algorithms that slows the rate of growth of simulation run lengths at later iterations.

The histogram approximation algorithm computes quantiles only at grid points and uses Lagrange interpolation to estimate the p quantile. The algorithm also generates an empirical distribution (histogram) of the output sequence, which can provide valuable insights to the underlying stochastic process. The first-phase of the procedure computes quantile estimates that satisfy a proportional half-width requirement. Based on the results determined in the first-phase, the procedure estimates the derivative at the p quantile and computes the required sample size for the second phase. The quantile estimates from the second phase satisfy absolute or relative precision requirements.

Our approach has the desirable properties of having a sequential procedure and not requiring users to have *a priori* knowledge of values that the data might assume. This allows the user to apply the two-phase quantile estimation procedure without having to execute a separate pilot run to determine the range of values to be expected or to guess and risk having to re-run the simulation. Both of these options represent 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 instead of advanced statistical theory, making it easy to understand, simple to implement, and fast to run.

In an effort to reduce the variance in simulation run length determined by the QI sequence, we have experimented with combining runs-up and runs-down tests together to check whether a sequence appears to be independent. Our preliminary results show that combining runs-up and runs-down tests has little effect on i.i.d. sequences. However, it generally results in a longer simulation run length for highly correlated sequences with no improvement in the variance of simulation run length.

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*. Vol. 76. No. 5: 273-276, 301-304.

Chen, E. J. 2002. Derivative Estimation with Finite Differences. Working paper.

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. 2000. 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.

Chen, E. J., and W. D. Kelton. 2001. Quantile and Histogram Estimation. *Proceedings of the 2001 Winter Simulation Conference*, ed. B.A. Peters, J.S. Smith, D.J. Medeiros, and M.W. Rohrer, 451–459. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Hogg, R.V., and A. T. Craig. 1995. *Introduction to Mathematical Statistics*. 5th ed. Englewood Cliffs, New Jersey:Prentice Hall,

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.

Wood, D. C., and B. W. Schmeiser. 1995. Overlapping Batch Quantiles. *Proceedings of the 1995 Winter Simulation Conference*, ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman. 303–308. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

AUTHOR BIOGRAPHY

E. JACK CHEN is a Senior Staff Specialist with BASF Corporation. He received a Ph.D. degree from University of Cincinnati. His research interests are in the area of computer simulation. His email address is <chenej@basf.com>.