SIMULATION-BASED ESTIMATION OF QUANTILES

E. Jack Chen
W. David Kelton

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

ABSTRACT

This paper discusses implementation of a sequential quantile-estimation algorithm for highly correlated steady-state simulation output. Our primary focus is on issues related to computational and storage requirements of order statistics. The algorithm can compute exact sample quantiles and process sample sizes up to several billion without storing and sorting the whole sequence. The algorithm dynamically increases the sample size so that the quantile estimated satisfies a pre-specified precision requirement.

1 INTRODUCTION

A properly selected set of *quantiles* (percentiles) reveals all the essential distributional features of output random variables analyzed by simulation. For $0 < p < 1$, the p quantile of a distribution is the value at or below which $100p$ percent of the distribution lies. Quantiles are also more robust to outliers than are the mean and standard deviation. A few exceptional observations do not affect the quantiles as heavily as they affect the mean and standard deviation. However, quantiles have seldom been used in simulation studies. We believe that the reason for this lies in the complexity of quantile estimation.

Estimating quantiles is computationally more difficult than estimating the mean and variance. One of the basic problems in estimation of quantiles is that the whole output sequence must be stored and sorted if the estimates are based on order statistics. In the past, because of limited computational capability of computers, the quantile estimation procedures of Iglehart (1976) and Seila (1982a, 1982b) can only estimate quantiles of regenerative processes. For general processes, special algorithms, such as the maximum transformation of Heidelberger and Lewis (1984) and the P^2 algorithm of Jain and Chlamtac (1985), were needed to obtain estimates. But the advance of computer technology has alleviated some of the computational issues. Now, a 200MHz Pentium computer with only 32 megabytes of RAM can store and sort one million observations ten times in less than three minutes and has the capacity to store and sort up to four million observations. However, depending on the underlying distribution, the quantile to be estimated, and the precision required, the required sample size can be hundreds of millions. Therefore, using "brute force" to store and sort all the observations can be applied in only very limited cases. Algorithms that can be used to estimate quantiles without storing and sorting all observations are still required.

The output processes of virtually all dynamic simulations are nonstationary and autocorrelated. Thus, classical statistical techniques based on i.i.d. (independent and identically distributed) observations are not directly applicable. By contrast, order statistics 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 (a relatively mild assumption; see Section 2) of continuous random variables. Yet when the output processes of simulations are stationary, most of them satisfy the ϕ-mixing conditions. Therefore, classical statistical techniques can be used on the order-statistics quantile estimators of those processes.

The problem with which we are concerned is the estimation of a quantile for a discrete-time, covariance-stationary stochastic process. We discuss a procedure for estimating quantiles from simulation output. The proposed procedure will control the length of a simulation run so that the quantile estimated satisfies a pre-specified precision requirement.

In Section 2, we discuss some theoretical background of simulation output analysis. In Section 3, we present our methodologies and proposed procedure for quantile estimation. In Section 4, we give concluding remarks.

2 THEORETICAL BACKGROUND

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 $100p^{th}$ percentiles as the p quantile, such that $F(x_p) =$

$Pr(X \leq x_p) = p$. Thus, $x_p = inf\{x : F(x) \leq p\}$. 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 $X_1, X_2, \ldots, X_n$) then a point estimator for x_p based on the order statistics is the sample p quantile $\hat{x}_p$,

$$\hat{x}_p = y_{\lceil np \rceil} \qquad (1)$$

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

To define ϕ-mixing, let $\{X_i; -\infty < i < \infty\}$ be a stationary sequence of random variables defined on a probability space (Ω, $\mathcal{A}$, P). Thus, if $\mathcal{M}_{-\infty}^{k}$ and $\mathcal{M}_{k+j}^{\infty}$ are respectively the sequences generated by $\{X_i; i \leq k\}$ and $\{X_i; i \geq k+j\}$, and if $E_1 \in \mathcal{M}_{-\infty}^{k}$ and $E_2 \in \mathcal{M}_{k+j}^{\infty}$, then for all k $(-\infty < k < \infty)$ and j $(j \geq 1)$, if

$$|P(E_2|E_1) - P(E_2)| \leq \phi(j), \quad \phi(j) \geq 0,$$

where $1 \geq \phi(1) \geq \phi(2) \geq \cdots$, and $lim_{j \to \infty}\phi(j) = 0$, then $\{X_i; -\infty < i < \infty\}$ is called ϕ-mixing. Roughly speaking $X_1, X_2, \cdots, X_n$ is ϕ-mixing if X_i and X_{i+j} become essentially independent as j becomes large. 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.

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 the following three conditions are satisfied:

1. The process $\{X_i\}$ satisfies the ϕ-mixing condition.
2. The cumulative distribution function, $F(x)$, is absolutely continuous.
3. The density function, $f(x)$, is finite, positive, and absolutely continuous for all $x = F^{-1}(t)$ and $0 < t < 1$.

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 $P_{x_p}(0)$. Here $P_{x_p}(0)$ is the initial point on the spectrum of the binary process $\{I_n(x_p)\}$, where

$$I_n(x) = \begin{cases} 1 & \text{if } X_n \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

$$P_{x_p}(0) = \sum_{k=-\infty}^{k=\infty} Cov[I_n(x_p), I_{n+k}(x_p)]$$

and

$$\sigma^2(\hat{x}_p) = P_{x_p}(0)/[nf^2(x_p)].$$

Therefore,

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

as $n \to \infty$.

The factor $P_{x_p}(0)$ measures not only the variance of each individual observation but also the correlation between observations. One can estimate $P_{x_p}(0)$ using the methods of Heidelberger and Welch (1981a, 1981b). The ratio of $P_{x_p}(0)$ to its value $p(1-p)$ for the i.i.d. process with identical marginal distributions gives us a measure of the inflation factor of the required sample size over the independence case.

3 METHODOLOGIES

This section presents the methodologies we will use for our quantile estimation. The results from Sen (1972) provided us with a strong theoretical basis for using classical statistical techniques to develop order-statistics quantile estimators. However, data are almost always expensive and scarce. Increasing the sample size is costly and time-consuming, both in the sampling procedure and in processing the data. 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. Therefore, both asymptotic theory and workable finite-sample approaches are needed by the practitioner.

3.1 Proportional Half-Width

Usually, the stopping rule of absolute-precision simulation procedures will ensure that

$$\hat{x}_p \in x_p \pm \epsilon \qquad (2)$$

with a certain confidence level, where $\hat{x}_p$ is the estimated quantile, x_p is the true (but unknown) quantile, and ϵ is the maximum allowed half-width of confidence.

Since we are estimating quantiles, a different precision requirement can be used. We can control the precision by ensuring that the p quantile estimator

$$\hat{x}_p \in x_{[p\pm\epsilon']_0^1} \tag{3}$$

where

$$[p\pm\epsilon']_0^1 = \begin{cases} p \pm \epsilon' & \text{if } 0 \le p-\epsilon' \text{ and } p+\epsilon' \le 1, \\ \left[0, p+\epsilon'\right] & \text{if } 0 > p-\epsilon' \text{ and } p+\epsilon' \le 1, \\ \left[p-\epsilon', 1\right] & \text{if } 0 \le p-\epsilon' \text{ and } p+\epsilon' > 1. \end{cases}$$

That is, if

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

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

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

where ϵ' is the maximum proportion half-width of the confidence. We would like to point out that the absolute half-width ϵ has the same *measurement unit* as the variate under investigation and can be any positive value. However, 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 second precision requirement (i.e. equation (3)), 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

$$n_p \ge \frac{z_{1-\alpha/2}^2 p(1-p)}{(\epsilon')^2} \tag{4}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, ϵ' is the maximum proportion half-width of the confidence interval, and $1-\alpha$ is the confidence level.

If we use equation (2) as the precision requirement, then the sample size n_p needs to be inflated by a factor of $1/f^2(x_p)$. That is

$$n'_p = n_p/f^2(x_p).$$

Furthermore, if the sequences are ϕ-mixing instead of i.i.d., then the sample size needs to be further inflated by a factor of $P_{x_p}(0)/p(1-p)$.

3.2 Test of Independence

Because the required sample sizes are drastically different between i.i.d. and correlated sequences, it is beneficial to check whether the input data appear to be independent. We use a *runs-up* test for this purpose, see Knuth(1981, pp. 65-68). The observation immediately following a run is discarded so that subsequent runs are independent. Therefore, a straightforward chi-square test can be used. The runs-up tests looks solely for independence and has been shown to be very powerful. If the input data sequence appears to be dependent, then a sequential procedure will be used.

The test statistic of this simple runs-up test is sensitive to high correlation with the correlation coefficient of first-order autoregression data sequence. The runs-up test statistic becomes larger as the lag 1 (positive) correlation of the input data sequence becomes stronger. The variance of the input data sequence that are positively correlated at several different lags will be at least as large as the variance of the input data sequence that are correlated only at lag 1 with the same correlation. Therefore, we can use the runs-up test statistic to compute the lower bound of the required sample size for a sequential procedure.

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

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

where

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

$$0 < \rho < 1,$$

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

The AR(1) process, as defined above, has been used as a model for simulation output processes by numerous authors, for example, Law and Kelton (1984); it shares many characteristics observed in simulation output processes, including first- and second-order non-stationarity and autocorrelations that decline exponentially with increasing lag (so AR(1) sequences are ϕ-mixing sequences). 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. $\mathcal{N}(0, 1)$. Therefore, $x_i \xrightarrow{\mathcal{D}} \mathcal{N}(0, \frac{1}{1-\rho^2})$. The spectrum of the AR(1) process is $P_{x_p}(0) = 1/(1-\rho)^2$.

We ran ten replications of the runs-up test for each AR(1) process with ϵ' set to 0.005. The average of the test statistics with corresponding spectrum and required sample size are listed in Table 1. We could plot a graph of the runs-up test statistic vs. the required sample size to get an idea of the relationship between these two variables. Because of the convex nature of their relationship, we can use linear interpolation to get a conservative estimate of the lower bound of the required sample size for any given runs-up test statistic.

Table 1: The Required Sample Size for AR(1) with the Maximum Proportional Half-Width Set to 0.005

ρ	$P_{x_p}(0)$	Test Statistic	Sample Size
0.5	4	2659	4330
0.75	16	8782	17319
0.80	25	10954	27061
0.85	45	14009	48709
0.90	100	16877	108241
0.93	200	19804	216482
0.95	400	21052	432964

3.3 Sequential Procedure

One of the most important aspects of the design of experiments is the determination of the appropriate sample size of the basic experiment. Because almost all simulation output processes are autocorrelated and nonstationary, no fixed-sample-size procedures can be relied upon to produce a c.i. that covers x_p with the desired probability level, if the fixed sample size is too small for the system being simulated. In addition to this problem of coverage, a simulator might want to determine a sample size large enough to produce a c.i. with a small absolute precision ϵ or a small proportional precision ϵ'. It will seldom be possible to know in advance even the order of the magnitude of the sample size needed to meet these goals in a given simulation problem, so some sort of procedure to increase iteratively this sample size would be needed. Consequently, sequential procedures have been developed (Law and Kelton 1991).

We propose a simple sequential algorithm with combined precision stopping rules and use order statistics to estimate quantiles. The *zoom-in* algorithm, which is an *enclosure method*, gives upper and lower bounds on the quantile x_p at every step. We assume that the p quantile of F_x is finite, i.e. $-\infty < x_p < \infty$. That implies that the initial lower and upper bounds of x_p are greater than $-\infty$ and less than ∞ respectively. In our algorithm, the default lower and upper bounds are set to -10^{308} and 10^{308} respectively, which are roughly the minimum and maximum of floating-point numbers on a 32-bit computer. However, any *a priori* knowledge regarding the value of a quantile can be used for the lower and upper bound to reduce the needed iterations.

There are two phases in the zoom-in algorithm. First, the zoom-in process determines the required simulation run length with pre-specified precision requirements. Second, replications will be executed iteratively to obtain a confidence interval. A *memory buffer*, which is an array of variables that are defined as *double* (in the C language), is allocated at the beginning of the computer program. The memory buffer is used to store and sort the observations that are deemed most likely to be the true quantile value. Let n_0 be the initial sample size, which is also the initial buffer size. We compute the sample quantile according to equation (1). The sample quantile lower and upper bound are $\hat{x}_{pl}$ and $\hat{x}_{pu}$, which can be estimated by

$$\hat{x}_{pl} = Y_{\lfloor n_{pl} \rfloor} \tag{5}$$

and

$$\hat{x}_{pu} = Y_{\lceil n_{pu} \rceil}, \tag{6}$$

where $n_{pl} = n_0(p - \delta_p)$, $n_{pu} = n_0(p + \delta_p)$, and $0 < \delta_p < 0.5$. If $n_{pl} < 0$, then the lower bound is not changed. If $n_{pu} > n_0$, then the upper bound is not changed.

Once we obtain the lower and upper bounds of the p quantile, the sample values X_i for $i = 1, \ldots, n_{pl} - 1$ and $i = n_{pu}+1, \ldots, n_0$ are no longer needed. More samples can then be generated to fill in the available slots in the memory buffer. We will store the newly generated sample values in the buffer only when the value is between the lower and upper bounds inclusively. If the sample value is less than the lower bound, we increase the counter n_{pl} by one. If the sample value is larger than the upper bound, we increase the counter n_{pu} by one. The total sample size n in the second iteration is then equal to $n_0 + n_{pl} + n_{pu}$. That is, the sample size is increased by $n_{pl} + n_{pu}$, and $n_0/2\delta_p < E(n) < n_0/\delta_p$. When the buffer is filled completely again, new sample-quantile lower and upper bounds can be recomputed. The gap δ_p is multiplied by DFACTOR (< 1, we use 0.90) in subsequent iterations. A test is also used to see whether the number of available slots is more than R% (we use 10%) of the buffer size, in which case DSIZE (we use 1,000) slots will be added to the buffer when the number of available slots is less than R% of the buffer size. The process can be repeated iteratively until the pre-specified stopping criteria are satisfied. In cases that the sample quantile is outside the range of the buffer, i.e., $np < n_{pl}$ or $np > n_{pu}$, then the simulation needs to be restarted with a larger memory buffer or larger DFACTOR. The lower and upper bounds estimated in the current run can be used for later runs so that the program can be re-started from where the program terminated without starting from the very beginning. The estimator can also be estimated by linear extrapolation, but the estimator will not be the exact order-statistics quantile in this case.

It is also possible that while the quantile estimator is inside the lower and upper bound during the sample-size-estimating phase, some of the quantile estimators are outside the range in subsequent replications. In these cases, the estimators will be treated as outliers and the lower or upper bounds of the sample quantiles will be used as estimators.

The proposed stopping criterion includes six parts:

1. The difference of consecutive quantile estimators have changed sign at least K times, and the last L consecutive quantile estimators do not increase or decrease monotonically. Based on the results of our empirical studies, we set both K and L to 4. If we assume the consecutive sample quantiles are independent, the possibility of having four monotonically increasing or decreasing sample quantiles is 1/24.
2. There is no new maximum or minimum in subsequent simulation runs.
3. The range covered by $\hat{x}_{pl}$ and $\hat{x}_{pu}$, i.e. $F(\hat{x}_{pu}) - F(\hat{x}_{pl})$, should be roughly the same as p_{lu}, which is the proportion of observations between $\hat{x}_{pl}$ and $\hat{x}_{pu}$. That is, $|(F(\hat{x}_{pu}) - F(\hat{x}_{pl})) - p_{lu}| < \epsilon'$.
4. The range covered by $(-\infty, \hat{x}_{pl})$ and $(\hat{x}_{pu}, \infty)$, should be roughly the same as p_l and p_u, which are the proportion of observations smaller than $\hat{x}_{pl}$ and larger than $\hat{x}_{pu}$, respectively. That is, $|F(\hat{x}_{pl}) - p_l| < \epsilon'$ and $|F(\hat{x}_{pu}) - (1 - p_u)| < \epsilon'$.
5. The absolute and relative difference of the consecutive quantile estimators $\hat{x}_p$ of the sequential procedure is within tolerance, i.e. less than ϵ'.
6. The coverage of $\hat{x}_{pl}$ and $\hat{x}_{pu}$ is less than ϵ'. That is, $F(\hat{x}_{pu}) - F(\hat{x}_{pl}) < \epsilon'$, i.e. $\delta_p < \epsilon'/2$.

The zoom-in algorithm:

1. Remark: bfmax is the maximum size of the memory buffer to be allocated, and ier is an error flag to the user. bfsize is the size of the buffer used to store and sort observations. x_{pl} and x_{pu} are the lower and upper bounds of the quantile to be estimated; they are initially set to $-\infty$ and, ∞ respectively. Dsize is the incremental size used to expand the buffer.
2. Simulate bfsize observations and use order statistics to estimate the quantile x_p, the lower bound x_{pl}, and the upper bound x_{pu}.
3. If the observations are determined to be i.i.d. then go to step 9.
4. Compute n_{pl}, the number of observations that are smaller than x_{pl}, and n_{pu}, the number of observations that are larger than x_{pu}.
5. Simulate more observations until all the available slots in the buffer are filled. Use order statistics to estimate the quantile x_p, the lower bound x_{pl}, and the upper bound x_{pu}.
6. If the stopping criteria are satisfied, go to step 9.
7. Otherwise, if the number of available slots in the buffer is less than R% of the current buffer size, increase the buffer size by Dsize. If $bfsize > bfmax$, set ier = 1 and exit.
8. Go to step 4.
9. Make another K replications, and use the values of bfsize, x_{pl} and x_{pu} computed above.
10. While the standard deviation of the sample quantiles is greater than tolerance, make 3 more replications, and set $K = K + 3$.
11. Set the quantile point estimator to the upper bound of the confidence interval of the observed $K + 1$ quantile estimators.

The first five stopping rules are connected by "and," the last stopping rule is connected by "or." That is, the procedure will terminate if either stopping rule 6 alone is satisfied or stopping rules 1 through 5 are satisfied. To avoid premature stopping of the sequential procedure, we will stop it only when stopping rules 2 through 5 have been satisfied in three consecutive iterations. Of course, the number of iterations that stopping rules 2 through 5 must be satisfied can be increased for conservative users. On the other hand, the sequential procedure will stop immediately whenever stopping rule 6 is satisfied, therefore, the zoom-in algorithm will always terminate gracefully. This is because once the value of ϵ' has been decided, the maximum number of iterations n_i can be computed by $n_i = \lceil \ln(\epsilon'/2\delta_p)$ / $\ln(\text{DFACTOR})\rceil$ (because $\text{DFACTOR}^n \delta_p < \epsilon'/2$). Therefore, if the number of iterations that stopping rules 2 through 5 must be satisfied was set too large, the sequential procedure will always be terminated by stopping rule 6. Then the sequential procedure will behave like a fixed-sample-size procedure, and the sample size determined by those stopping rules will be still large enough for highly correlated data but will be larger than required for data that are only slightly correlated.

Figure 1 gives an example of the 0.75 quantile estimator for the waiting-time of an M/M/1 delay-in-queue process with arrival rate $\lambda = 0.75$, service rate $\mu = 1.0$, and $\epsilon' = 0.005$ at each iteration. Figure 2 shows the corresponding sample size at each iteration. The process iterates 26 times before it stops. During the sample-size-determination iterations, the quantile estimators are not independent, but we assume that those quantile estimators satisfy the ϕ-mixing conditions. The sample size grows exponentially during this

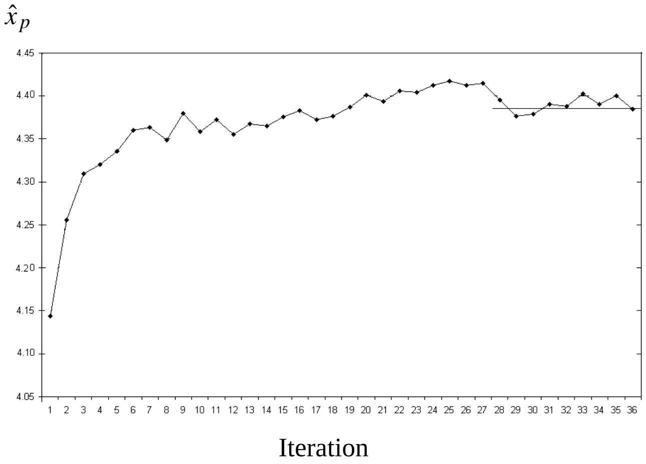

Figure 1: 0.75 Quantile Estimator for M/M/1 Queue with $\lambda/\mu = 0.75$, $\epsilon' = 0.005$

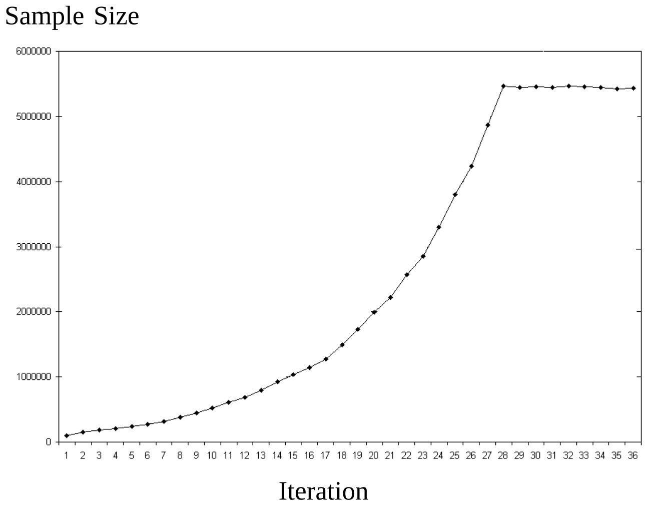

Figure 2: Sample Size for M/M/1 Queue with $\lambda/\mu = 0.75$, $\epsilon' = 0.005$ at Each Iteration

phase and the incremental sample size seems large enough to influence the distribution of the output sequence. Thus, the correlation between those estimators is very weak if they are correlated at all. The sequential sample quantiles converge to the true quantile value of 4.39445.

The process executes 9 more replications before it stops, and the average sample size of these 9 replications is 5,450,000. The sample quantile of each replication fluctuates around the true value. The average of these 10 sample quantiles is 4.3929, or approximately the 0.749903 quantile. Even though the sample size used in the simulation run is much smaller than the theoretical value, the precision of the estimator is still very good. We believe that this is because the theoretical value is for the worst-case scenario. For most cases, we are able to get good estimates with sample sizes considerably less than the theoretical value.

4 CONCLUDING REMARKS

The results from our empirical experiments show that the procedure is excellent in achieving the pre-specified accuracy. However, the variance of the run length from our sequential procedure is large. We ran ten replications of quantile estimation and set the point estimator to the upper bound of the confidence interval. Figure 3 gives an example of the results of our 0.95 quantile estimator for an AR(1) processes with $\rho = 0.95$, $\alpha = 0.10\%$, and $\epsilon' = 0.0025$. The horizontal axis is the run number and the vertical axis is the cumulative distribution function value, i.e. $F(\bar{\hat{x}}_p)$. In the 100 runs we did, 97 of those 0.95 quantile estimators cover at least 95% of the distribution, and there are only three estimators that cover more than 95.05%. In those three cases that the coverages are less than 95%, the deviations are less than 0.001%, i.e. they cover more than 94.99% of the distribution. The maximum deviation of these 100 estimators is less than 0.06%, which is much smaller than the required precision $\epsilon' = 0.0025$.

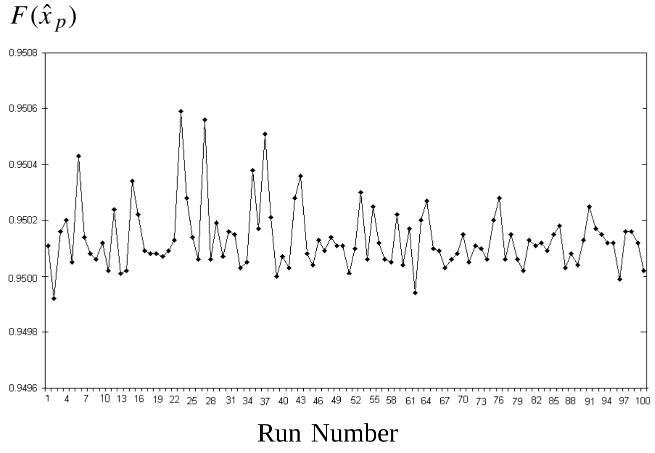

Figure 3: Sequential Estimators of 90% Confidence 0.95 Quantile for AR(1) with $\rho = 0.95$ and $\epsilon' = 0.0025$

Our proposed zoom-in algorithm requires storing and sorting only a sequence of the most likely *p* quantile values. Savings in storage and sorting are substantial for our method. The proposed procedure for estimating quantiles can process sample sizes up to several billion. Our approach has the desirable properties that it is a sequential procedure and it does not require the user to have *a priori* knowledge of values that the data might assume. This allows the user to apply this method without having to run a 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 simplicity of this method should make it attractive to simulation practitioners.

ACKNOWLEDGMENTS

We would like to thank the Ohio Supercomputer Center for the Cray T-90 time grant.

REFERENCES

Heidelberger, P., and P. A. W. Lewis. 1984. Quantile Estimation in Dependent Sequences. *Operations Research* 32:185–209.

Heidelberger, P., and P. D. Welch. 1981a. A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations. *Communications of the ACM*. 233–245.

Heidelberger, P., and P. D. Welch. 1981b. Adaptive Spectral Methods for Simulation Output Analysis. *IBM Journal of Research and Development* 25. No. 6:860–876.

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

Jain, R., and I. Chlamtac. 1985. The P^2 Algorithm for Dynamic Calculation of Quantiles and Histograms without Storing Observations. *Commun. Assoc. Comput. Mach.* 28:1076–1085.

Knuth, D. E. 1981. *The Art of Computer Programming*. Vol. 2. 2nd ed. Reading, Mass.:Addison-Wesley.

Law, A. M., and W. D. Kelton. 1984. Confidence Intervals for Steady-State Simulations: I. A Survey of Fixed Sample Size Procedures. *Operations Research* 32:1221–1239.

Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis*. 2nd 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 Ph.D. Candidate in the Department of Quantitative Analysis and Operations Management at the University of Cincinnati. He received a B.S. in engineering from National Taiwan University, an M.S. in computer science from Syracuse University, and an M.B.A. from Northern Kentucky University. His research interests are in the area of computer simulation.

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, statistical quality control, and stochastic models. He serves as Simulation Area Editor for *Operations Research*; he has also been Simulation Area Editor for the *INFORMS Journal on Computing* and *IIE Transactions*, and Associate Editor of *Operations Research*, the *Journal of Manufacturing Systems*, and *Simulation*. He is 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.