Proceedings of the 2004 Winter Simulation Conference
R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

EXPERIMENTAL PERFORMANCE EVALUATION OF HISTOGRAM APPROXIMATION FOR SIMULATION OUTPUT ANALYSIS

E. Jack Chen

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

W. David Kelton

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

ABSTRACT

We summarize the results of an experimental performance evaluation of using an empirical histogram to approximate the steady-state distribution of the underlying stochastic process. We use a runs test to determine the required sample size for simulation output analysis and construct a histogram by computing sample quantiles at certain grid points. The algorithm dynamically increases the sample size so that histogram estimates are asymptotically unbiased. Characteristics of the steady-state distribution, such as the mean and variance, can then be estimated through the empirical histogram. The preliminary experimental results indicate that the natural estimators obtained based on the empirical distribution are fairly accurate.

1 INTRODUCTION

A concern of simulation output analysis is to estimate the sampling error of an estimator, i.e., the error caused by the estimator's randomness. This gives the experimenter an idea of the precision with which the estimator reflects the true but unknown parameter. For example, when estimating the steady-state performance, such as the mean μ, of some discrete-time stochastic output process $\{X_i : i \geq 1\}$ via simulation, we would like an algorithm to determine the simulation run length N so that the mean estimator (sample mean $\bar{X}(N) = \sum_{i=1}^{N} X_i/N$) is approximately unbiased (i.e. the asymptotic approximation is valid), the confidence interval (CI) is of a pre-specified width, and the actual coverage probability of the CI is close to the nominal coverage probability $1-\alpha$. Because we assume the underlying distribution is stationary, i.e., the joint distribution of the X_i's is insensitive to time shifts, the mean estimator will be unbiased. However, the usual method of CI construction from classical statistics, which assumes independent and identically distributed (i.i.d.) observations, is not directly applicable since simulation output data are generally correlated.

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 output random variables analyzed by simulation. Histograms can be constructed with a properly selected set of quantiles.

In this paper, we propose using the runs test to determine the simulation run length and construct an empirical histogram to approximate the underlying distribution. Distributional features, such as the variance, are then estimated through this empirical histogram. The only required condition is that the autocorrelations of the stochastic process output sequence die off as the lag between observations increases, in the sense of ϕ-mixing, see Billingsley (1999). This mild assumption is satisfied in virtually all practical settings. These weakly dependent processes typically obey a central limit theorem for dependent processes of the form

$$\sqrt{N}\frac{[\bar{X}(N)-\mu]}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \text{as} \quad N \to \infty,$$

where

$$\sigma^2 \equiv \lim_{N\to\infty} N\text{Var}[\bar{X}(N)] = \sum_{i=-\infty}^{\infty} \gamma_i$$

is the steady-state variance constant (SSVC), and $\gamma_i = \text{Cov}(X_k, X_{k+i})$ for any k is the lag-i covariance. If the sequence is independent, then the SSVC is equal to the process variance $\sigma_x^2 = \text{Var}(X_i)$. For a finite sample N, let

$$\sigma^2(N) = \gamma_0 + 2\sum_{i=1}^{N-1}(1-i/N)\gamma_i.$$

It follows that $\lim_{N\to\infty}\sigma^2(N)=\sigma^2$ so

$$\text{Var}[\bar{X}(N)]=\sigma^2(N)/N\approx\sigma^2/N, \tag{1}$$

provided that N is sufficiently large.

In the non overlapping batch-means method, the simulation output sequence $\{X_i : i=1,2,\ldots,N\}$ is divided into b adjacent non overlapping batches, each of size m. For simplicity, we assume that N is a multiple of m so that $N=bm$. The sample mean, $\bar{X}_j$, for the j^{th} batch is

$$\bar{X}_j=\frac{1}{m}\sum_{i=m(j-1)+1}^{mj}X_i \quad \text{for } j=1,2,\ldots,b.$$

Then the grand mean $\hat{\mu}$ of the individual batch means, given by

$$\hat{\mu}=\frac{1}{b}\sum_{j=1}^{b}\bar{X}_j, \tag{2}$$

is used as a point estimator for μ. Here $\hat{\mu}=\bar{X}(N)$, the sample mean of all N individual X_i's, and we seek to construct a CI based on the point estimator obtained by (2).

The rest of this paper is organized as follows. In Section 2, we present the methodologies of constructing an empirical histogram estimation. In Section 3, we present algorithms to estimate variance. In Section 4, we show some empirical-experimental results of mean and histogram estimation. In Section 5, we give concluding remarks.

2 HISTOGRAM APPROXIMATION

This section presents an overview of constructing an empirical histogram to approximate the underlying distribution based entirely on data.

2.1 Natural Estimators

Let $F(\cdot)$ denote the unknown steady-state cumulative distribution function (c.d.f.) of the output random variable X of the process under study and Ψ a property of $F(\cdot)$, such as the mean, variance, or a quantile. The natural point estimator for Ψ, denoted by $\hat{\Psi}$, is typically the sample mean, the sample variance, the sample quantile, or a simple function of the relevant order statistics, chosen to imitate the performance measure Ψ. Furthermore, the natural estimators are appropriate for estimating any Ψ, regardless of dependency, which follows from the empirical distribution function $F_N(\cdot)$ converging to $F(\cdot)$.

The empirical discrete c.d.f. $F_N(x)$ based on samples $\{X_i : i=1,2,\cdots,N\}$ of X is constructed as follows:

$$F_N(x)=\begin{cases}0 & \text{if } x<X_{(1)},\\ i/N & \text{if } X_{(i)}\le x<X_{(i+1)},\ 1\le i\le N-1,\\ 1 & \text{if } X_{(N)}\le x,\end{cases}$$

where $X_{(1)}\le X_{(2)}\le\cdots\le X_{(N)}$ are the order statistics obtained by sorting the observations $\{X_i : i=1,2,\cdots,N\}$ in ascending order. Furthermore, let $I(\cdot)$ denote the indicator function. Then $F_N(x)=\sum_{i=1}^{N}I(X_i\le x)/N$. Recall that the p quantile x_p is the value such that $x_p=F^{-1}(x)=\inf\{x : F(x)\ge p\}$, $p\in(0,1)$. A natural estimator for x_p is the sample quantile $\hat{x}_p=F_N^{-1}(x)=\inf\{x : F_N(x)\ge p\}$.

2.2 Implementation

Chen and Kelton (2003) use grid points to construct a histogram (multiple quantiles) as an empirical distribution of the underlying steady-state distribution. The total number of grid points is $G=203$. The value of the grid points $g_0, g_1, \ldots, g_{202}$ will be constructed as follows: g_0 and g_{202} are set to $-\infty$ and ∞ (in practice, the minimum and maximum values on the underlying computer) respectively. If the analyst knows what may be the minimum or maximum values of the distribution, those values should be used instead. For example, the waiting-time of any queuing system cannot be negative, so the analyst should enter 0 as the minimum. Grid point g_{51} is set to the minimum of the initial n, $2n$, or $3n$ samples, depending on the degree of correlation of the sequence. The number of main grids, G_m, is set to 100. Grid points g_{i+51}, $i=1,2,\ldots,G_m$, are set to the i/G_m quantile of the initial n, $2n$, or $3n$ samples. We will set grid points g_1 through g_{50} and g_{152} through g_{201} to appropriate values so that g_1 through g_{52} will have the same segment length and g_{150} through g_{201} will have the same segment length. Therefore, the grids will be dense where the estimated probability density function is high and will be sparse where the estimated probability density function is low. Furthermore, each grid will contain no more than $1/G_m$ of the distribution. A corresponding array of $n_1, n_2, \ldots, n_{202}$ is used to store the number of observations between two consecutive grid points. The number of observations in the grid will be updated each time a new observation x_i is generated, i.e., the number stored in n_j is increased by one if $g_{j-1}<x_i\le g_j$.

Let the p quantile estimate satisfies the precision requirement

$$\Pr[x_p\in\hat{x}_{[p\pm\delta]_0^1}]\ge 1-\alpha \tag{3}$$

where

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

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 approximately the minimum n_p that satisfies

$$n_p \ge \frac{z_{1-\alpha/2}^2 p(1-p)}{\delta^2},$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, δ is the maximum proportional half-width of the confidence interval, and $1-\alpha$ is the confidence level. See Chen and Kelton (2004) for details.

2.3 Determining the Simulation Run Length

The natural estimators are computationally reasonable and perform well statistically (Goldsman and Schmeiser 1997). However, 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. Chen and Kelton (2003) propose using *runs* (runs-up and runs-down) tests to determine the simulation run length (i.e., sample size) for simulation output analysis. Briefly, a run up (run down) is a monotonically increasing (decreasing) subsequence and we consider the length of a run up (run down). A chi-square test is applied to the frequency of different run lengths to check whether a sequence appears to be independent. The runs test looks solely for independence and has been shown to be very powerful (Knuth 1998).

The procedure will increase the simulation run length progressively until n samples (taken from the original output sequence) appear to be independent, as determined by the runs test. This is accomplished by *systematic sampling*, i.e., select a number l between 1 and L, choose that observation, and then every l^{th} observation thereafter. Here the chosen l will be sufficiently large so that systematic samples are approximately uncorrelated. A *quasi independent* (QI) subsequence is defined as the sequence of systematic samples that appear to be independent, as determined by the runs test. A QI subsequence always exists in simulation output sequences since we assume the autocorrelations of the underlying process die off as the lag between observations increases. Hence, the procedure will always terminate gracefully.

We estimate by the runs test the lag l at which the systematic samples appear to be independent. Chen and Kelton (2003) give some experimental results of using the runs test to check whether a sequence appears to be independent, and demonstrate that the lag l at which the systematic samples appear to be independent and the strength of the autocorrelation of the output sequence is highly correlated. Based on those experience, we set L, the limit of l, to 2^{10}. The minimum required sample size is then $N = nl$, i.e., the minimum required simulation run length computed based on the data, which is called the *computation run length*. We set $n = 4000$, the minimum recommended sample size for the runs tests, in our procedure. While the initial 4000 observations may provide a CI that exceeds the user's precision requirement, such an initial sample size is usually easy and inexpensive to generate. The procedure will iterate repeatedly to increase the lag l until it has obtained n systematic samples that appear to be independent, as determined by the runs test.

Optimally, we would like to search for the minimum l such that the systematic samples pass the runs test of independence. However, in order to process the data in one pass, we double the lag l every two iterations. Thus, the lag l determined dynamically as the simulation proceeds is likely to be much larger than the minimum lag l to pass the test of independence, and consequently, the empirical simulation run length will likely be longer than the computed run length.

3 ESTIMATING THE VARIANCE

In this section, we discuss methods of estimating the variance of sample means using the histogram approximation and batch means.

3.1 Histogram Variance Estimates

Once the QI algorithm has determined that the sample size is large enough for the asymptotic approximation to become valid, we can then compute the mean and variance based on the estimated histogram. The mean is estimated by $\bar{X}(N)$. The variance is conservatively estimated by

$$S_H^2 = \sum_{j=1}^{G-1} \max((g_{j-1} - \bar{X}(N))^2, (g_j - \bar{X}(N))^2) P_j.$$

Note that $N = nl = \sum_{j=1}^{G-1} n_j$ and $P_j = n_j/N$.

This would then lead to the $100(1-\alpha)\%$ CI for μ,

$$\bar{X}(N) \pm z_{1-\alpha/2} \frac{S_H}{\sqrt{n}}.$$

Here n is the sample size used for the runs test. Even though $\bar{X}(N)$ is the sample mean of N samples, we use n to compute the standard error of $\bar{X}(N)$ to adjust for the autocorrelation. Furthermore, even though we use the standard error S instead of the standard deviation σ to construct the CI, we use $z_{1-\alpha/2}$

instead of $t_{1-\alpha/2,n-1}$ to estimate the CI half-width since $n = 4000$, where $t_{1-\alpha/2,n-1}$ is the $1-\alpha/2$ quantile of the t distribution with $n-1$ degrees of freedom.

Let the half-width be

$$H_H = z_{1-\alpha/2}\frac{S_H}{\sqrt{n}}.$$

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 γ of the magnitude of the final point estimator $\bar{X}(N)$. If the relevant requirement

$$H_H \le \epsilon \text{ or } H_H \le \gamma|\bar{X}(N)|$$

for the precision of the confidence interval is satisfied, then the QI procedure terminates, and we return the point estimator $\bar{X}(N)$ and the CI with half-width H. If the precision requirement is not satisfied, then the procedure will increase the sample size to $n'l$, where

$$n' = (\frac{H_H}{\epsilon})^2 n \text{ or } n' = (\frac{H_H}{\gamma\bar{X}(N)})^2 n.$$

Furthermore, the half-width will be computed by

$$H_H = z_{1-\alpha/2}\frac{S_H}{\sqrt{n'}}.$$

The quasi-independent algorithm uses the runs test to determine the simulation run length, which has strong theoretical basis. The QI algorithm is easy to implement because we only need to tally the number of observations in each grid as the sample sizes increase once the grid points have been set up.

3.2 A Hybrid Approach

The use of batch means is a well-known technique for estimating the variance of point estimators computed from simulation experiments. The batch-means variance estimator is fundamentally different than the histogram-sample-mean variance estimator. While the histogram-sample-mean variance estimator is computed indirectly from the sample variance, i.e., (1), the batch-means variance estimator is simply the sample variance of the mean estimator computed on means of subsets of consecutive subsamples, i.e.,

$$S_B^2 = \frac{1}{b-1}\sum_{j=1}^{b}(\bar{X}_j - \hat{\mu})^2. \quad (4)$$

The asymptotic validity of batch means depends on both the assumption of approximate independence of the batch means and the assumption of the batch means being approximately normally distributed.

We divide the entire output sequence into 100 batches. For details on batch-size effects in the analysis of simulation output once the sample size is fixed, see Schmeiser (1982). Let l denote the smallest lag at which the output sequence appears to be independent, as determined by the runs test. Then the batch size will be $40l$. To reduce the storage requirement, we allocate a buffer with size $3s$ ($t = 10$ and $s = n/t = 400$) to keep sample means. Initially, the sample means are obtained every $t = n/s$ observations and will be doubled every two iterations. The following shows the total number of observations and the number of observations used to compute sample mean in each cell at each iteration:

It	0	1_A	1_B	2_A	2_B	...	k_A	k_B
TO	n	$2n$	$3n$	$4n$	$6n$	...	$2^k n$	$2^{k-1}3n$
BU	s	$2s$	$3s$	$2s$	$3s$	...	$2s$	$3s$
OU	t	t	t	$2t$	$2t$	...	$2^{k-1}t$	$2^{k-1}t$

The *It* row shows the index of the iteration. The *TO* row shows the total number of observations at a certain iteration. The *BU* row shows the number of sample means obtained. The *OU* row shows the number of observations used to obtain the sample means stored in the buffer. For example, at the end of the 1_B^{th} iteration, the total number of observations is $3n$, there are $3s$ sample means in the buffer, and each sample mean is the average of t consecutive observations. At the beginning of the 2_A^{th} iteration, we reduce the number of sample means in the buffer from $3s$ to $3s/2$ by taking the average of every two consecutive sample means; consequently, each sample mean is the average of $2t$ consecutive observations. We will generate $s/2$ sample means that are the average of $2t$ consecutive observations at the 2_A^{th} iteration; thus, we will have $2s$ sample means at the end of the iteration. Depending on the number of buffers used, the batch means will be the average of 8 or 12 sample consecutive means.

Because observations x_i and x_{i+l} appear to be independent, we *assume* the batch size $40l$ is large enough such that the batch means will appear to be independent. Moreover, based on the common rule of thumb that the average of $n \ge 30$ i.i.d. observations is approximately normally distributed, the batch means with batch size $40l$ should be approximately normally distributed. Our experimental results indicate that the batch means constructed from this algorithm are generally normally distributed. The variance is then estimated as the sample variance of batch means, i.e., (4).

Both the histogram-approximation and batch-means algorithm of variance estimation use the same stopping rule and can be easily performed in the same simulation run. We propose a new approach to construct a CI. Let H_B, H_H, and H_A denote the $1-\alpha$ CI half-widths obtained from

the batch-means, histogram-approximation, and hybrid approaches. Note that

$$H_B = t_{1-\alpha/2,b-1}\frac{S_B}{\sqrt{b}}.$$

The hybrid approach sets the half-width

$$H_A = (H_B + H_H)/2.$$

That is, it uses the average of these two CI half-widths.

The ASAP2 procedure (Steiger et al. 2002) starts with an initial sample size of 4096 and increases the sample size by a factor of $\sqrt{2}$ at each iteration, thus, doubles the sample size every two iterations. ASAP2 iterates repeatedly until the batch means pass the Shapiro-Wilk test for multivariate normality and then delivers a correlation-adjusted confidence interval that accounts for dependency between the batch means. The WASSP procedure (Lada, Wilson, and Steiger 2003) extends the ASAP2 and use wavelet-based spectral method on the batch means to estimate the steady-state variance constant. The proposed QI procedure starts with an initial sample size of 4000 and doubles the sample sizes every two iterations. The QI procedure iterates repeatedly until the systematic samples pass the runs test for independence and then delivers a classical confidence interval without any adjustment since we assume the simulation run length determined by the QI procedure is long enough to ensure independence between batches. Hence, the QI procedure generally requires a larger sample size and delivers a tighter CI when no user precision is specified. Furthermore, ASAP and ASAP2 require storing and repeatedly using the entire output sequence. On the other hand, the QI-based procedures need only process each observation once and do not require storing the entire output sequence.

4 EMPIRICAL EXPERIMENTS

In this section, we evaluate the performance of using the hybrid approach to construct a CI. We use $n = 4000$ samples for the runs test and test the procedure with four stochastic processes:

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

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

  where ϵ_i is i.i.d. $\mathcal{N}(0, 1)$ and $0 < \theta < 1$. μ is set to 2 and θ is set to 0.9 in our experiments. This process is denoted MA1. It can be shown that X has asymptotically a $\mathcal{N}(\mu, 1 + \theta^2)$ distribution.
- Steady-state of the *first-order auto-regressive* process, generated by the recurrence

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

  where ϵ_i is i.i.d. $\mathcal{N}(0, 1)$, and $0 < \varphi < 1$. μ is set to 2, φ is set to 0.9, and X_0 is set to a random variate drawn from the steady-state distribution in our experiments. This process is denoted AR1 It can be shown that X has asymptotically a $\mathcal{N}(\mu, \frac{1}{1-\varphi^2})$ distribution.
- Steady-state of the M/M/1 delay-in-queue process with the arrival rate ($\lambda = 0.9$) and the service rate ($\nu = 1$). This process is denoted MM1. The traffic intensity of this process is $\rho = \lambda/\nu = 0.90$. The steady-state mean waiting time for this M/M/1 queuing system is 9.0; see Hillier and Lieberman (2001).
- Steady-state of the M/M/2 delay-in-queue process with the arrival rate ($\lambda = 9$) and the service rate ($\nu = 5$). This process is denoted MM2. The traffic intensity of this process is $\rho = \lambda/(s\nu) = 0.90$. The steady-state mean waiting time for this M/M/2 queuing system is 81/95; see Hillier and Lieberman (2001).

The confidence level of the runs test of independence is set to 90.25%, i.e., independent sequences will fail the test of independence approximately 9.75% of the time. Note a lower confidence level of the runs test of independence will increase the simulation run length.

4.1 Goodness-of-Fit Test

In this experiment, we aim to check the goodness of fit of the empirical distribution. We use the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests comparing the empirical distribution with the true steady-state distribution. The AD test statistic is approximated by

$$A^2_\eta = \eta\sum_{i=k}^{K}(F_\eta(x) - F(x))^2\Psi(x)f(x)(g_i - g_{i-1}).$$

Here $k = \min\{j : P_j > 0\}$, $K = \min\{j : \sum_{i=k}^{j} P_i = 1\}$, $\eta = K - k + 1$; $x = (g_{i-1} + g_i)/2$, $\Psi(x) = 1/(F(x)(1 - F(x))$, and $f(x)$ is the steady-state probability density function.

Table 1 lists the the average and standard deviation of the test statistics of these four stochastic processes, obtained with the simulation run length and the computation run length, respectively. We obtain the computation run length by generating the same random number stream repeatedly, so we can increase the lag l by one at each iteration. Each

Table 1: Goodness-of-Fit Test Statistics

Process	MA1	AR1	MM1	MM2
avg. KS	0.192358	0.147152	0.175046	0.179356
sdev KS	0.054680	0.047038	0.057412	0.084024
avg. AD	0.056773	0.056555	0.073777	0.372421
sdev AD	0.048956	0.056657	0.041187	0.352225
avg. KS	0.169879	0.158931	0.215457	0.185488
sdev KS	0.040643	0.040346	0.082329	0.069682
avg. AD	0.045038	0.057798	0.069351	0.490271
sdev AD	0.037782	0.045869	0.042467	0.875112

design point is based on 100 independent simulation runs. The 90% confidence critical value of the KS and AD tests are 1.1224 and 1.933, respectively. These results indicate the empirical distributions constructed with the proposed procedure are excellent approximations to the true steady-state distributions. Since the computation run length is much shorter than the simulation run length, the test statistics are not as good, but they are well within the acceptable region.

In some sense, a simulation is just a *function*, which may be vector-valued or stochastic. The explicit form of this function is unknown and probably very complicated even if it were known. These empirical distributions (histograms) *numerically* characterize the response function over the parameter range, even though we do not have an *algebraic* formula with which to characterize it. In this sense, we have generated a response surface or metamodel.

4.2 Evaluation of Independence of the Batch Means

In this experiment, we evaluate the asymptotic validity of independence of the batch means. We generate 4000 batch means with batch size $40l$, where l is determined by the procedure described in Section 3.2. We then apply the runs test of independence of these 4000 batch means to determine whether they appear to be independent. Recall that the confidence level of the runs test of independence is set to 90.25%. Consequently, we will encounter Type I error, i.e., reject the null hypothesis that the underlying distribution is normal when it is true, approximately 9.75% of the time.

Table 2 lists the the proportion of the batch means obtained with the simulation run length and the computation run length, respectively, that appear to be independent, as determined by the runs test. Each design point is based on 100 independent simulation runs. The observed proportions are smaller than the nominal value of 0.9025, indicating that approximately 10% and 15% of the batch means obtained by the simulation run length and the computation run length, respectively, are not independent. Since the runs test of independence has great power, we believe those batch means that appear to be dependent are only slightly correlated. Our experimental results reflect that, i.e., the CI coverage for the mean constructed with these batch means are close to the nominal value; see Chen and Kelton (2003).

Table 2: Proportion of Batch Means that Appear to be Independent

Process	MA1	AR1	MM1	MM2
proportion	0.77	0.81	0.81	0.83
proportion	0.83	0.73	0.72	0.74

In general, the proportion of batch means that appear to be independent should be higher with the simulation run length than with the computation run length since the simulation run length is at least the computation run length. The proportion is lower with the simulation run length than with the computation run length when MA1 is the underlying process. We believe this is because of the stochastic nature of the experiment. Note that the MA1 process is only weakly correlated, so the simulation run length and the computation run length are approximately the same; see Chen and Kelton (2003).

4.3 Evaluation of Normality of the Batch Means

In this experiment, we evaluate the asymptotic validity of normality of the batch means. We generate 100 batch means with batch size $40l$, where l is determined by the procedure described in Section 3.2. We then apply the chi-square test to these 100 batch means to determine whether they appear to be normal. We check the proportions of the value of batch means in each interval bounded by ($-\infty$, $\hat{\mu} - 0.96S$, $\hat{\mu} - 0.43S$, $\hat{\mu}$, $\hat{\mu} + 0.43S$, $\hat{\mu} + 0.96S$, ∞), where $\hat{\mu}$ and S are, respectively, the grand sample mean of these N observations and standard error of these b batch means. These intervals are strategically chosen so that the proportions of batch means in each interval are approximately equal. Under the null hypothesis that the batch means are normally distributed, the proportions of batch means in each interval are approximately (0.1685, 0.1651, 0.1664, 0.1664, 0.1651, 0.1685). We use a 0.9 confidence level for the chi-square test.

Table 3 lists the the proportion of the batch means obtained with the simulation run length and the computation run length, respectively, that appear to be normal, as determined by the chi-square test. Each design point is based on 100 independent simulation runs. The observed proportions are greater than the nominal value of 0.90, when the underlying distributions are normally distributed.

Table 3: Proportion of Batch Means that Appear to be Normal

Process	MA1	AR1	MM1	MM2
proportion	0.97	0.95	0.68	0.65
proportion	0.98	0.94	0.46	0.42

Furthermore, our experimental results indicate that the autocorrelations among observations have little impact on this test of normality, i.e., if the underlying steady-state distribution is normal, it is likely to pass this test of normality even the samples are correlated. On the other hand, when the underlying distributions are far from normal the proportion of batch means that appear to be normal is smaller than the nominal value. We can enhance the procedure by increasing the batch size progressively until the obtained batch means pass the normality check. Moreover, our experimental results indicate that CIs constructed with the assumption that samples are i.i.d. normal generally have coverages close to the nominal value when samples are independent but not normal.

4.4 Performance Evaluation of Confidence Intervals

Since the empirical histograms provide good approximations of the underlying distribution, we evaluate the performance of the CI half-width estimated based on the hybrid approach. In these experiments, no relative precision or absolute precision was specified, so the half-width of the CI is the result of the default precision.

Table 4 lists the experimental results of these four processes with the simulation run length and the computation run length. Each design point is based on 1000 replications. The μ row lists the true mean. The $\bar{X}$ row lists the grand sample mean. The *avg. rp* row lists the average of the relative precision of the estimators. Here, the relative precision is defined as $rp = |\bar{X} - \mu|/\bar{X}$, where $|x|$ is the absolute value of x. The *sdev rp* row lists the standard deviation of the relative precision of the estimators. The *avg. srl* row lists the average of the simulation run length. The *avg. crl* row lists the average of the computation run length. The *sdev samp* row lists the standard deviation of the simulation or computation run length. The *avg. hw* row lists the average of the CI half-width. The *sdev hw* row lists the standard deviation of the CI half-width. The *coverage* row lists the percentage of the CIs that cover the true mean value.

The results of using the histogram approximation and batch-means approach separately are not listed here, but we make some observations. For the MA1 and AR1 processes, the CI coverages are above the specified 90% confidence level under both simulation and computation run length. Since the steady-state distribution of the MA1 and AR1 process extends to $-\infty$ and ∞ in each tail, the histogram variance estimator seems conservative, i.e., achieves high coverage with larger CI half-width. The standard deviation of the CI half-width from the histogram approach is significantly smaller than the batch-means approach. The observed coverage of the hybrid approach is between the histogram-approximation and batch-means approaches. The simulation run length is significantly longer than the computation run length and achieved significantly narrower CI half-width under the batch-means approach, but the improvement of CI half-width is small under histogram-approximation approach.

Table 4: Coverage of 90% confidence intervals from AHB

Process	MA1	AR1	MM1	MM2
μ	2.00	2.00	9.00	0.852632
$\bar{X}$	2.000021	1.999499	8.997934	0.852395
avg. rp	0.008043	0.011576	0.015386	0.016181
sdev rp	0.006174	0.009084	0.011714	0.012606
avg. srl	8968	122752	1377024	1307136
sdev samp	2759	40663	537098	486560
avg. hw	0.036058	0.056863	0.227210	0.027882
sdev hw	0.001936	0.003537	0.027182	0.002893
coverage	92.28%	95.0%	90.3%	90.4%
$\bar{X}$	2.000714	2.002307	8.992556	0.848514
avg. rp	0.008207	0.014312	0.021803	0.022567
sdev rp	0.006394	0.010306	0.016961	0.017255
avg. crl	8652	83812	656016	624896
sdev samp	1489	10280	86451	79593
avg. hw	0.036188	0.061223	0.329302	0.032875
sdev hw	0.001686	0.002831	0.030914	0.003225
coverage	91.2%	93.1%	81.9%	83.3%

For the M/M/1 delay-in-queue process, the CI coverages are near the specified 90% with simulation run length, but the coverages are below the nominal values with the computation run length, especially under the histogram approximation. Since the steady-state distribution of the M/M/1 delay-in-queue process is bounded below by zero, the histogram variance estimator of sample means is biased low. This is caused by estimating the variance of sample means indirectly from the sample variance and the asymptotic approximation may not be valid yet with computation run length. The extreme values may not have occured with the necessary frequency to obtain an unbiased variance estimate. Several extreme values do not affect the quantile much, but those extreme values can significantly increase variance. Note that the batch-means variance estimator is computed directly from several batch means. The relative precision is significantly lower while the CI half-width is only slightly wider with the computation run length, resulting in low CI coverage. Again, the standard deviation of the half-width obtained by the histogram approach is smaller than in the batch-means approach. It is interesting that the hybrid approach obtains the best CI coverage while its half-width is smaller than the half-width of the histogram approximation with simulation run length. When estimating the M/M/1 delay-in-queue process with $\rho = 0.9$, there were two independent simulation runs that terminated by reaching the iteration limit, i.e., $l = 2^{10}$.

For the M/M/2 delay-in-queue process, the CI coverages of the histogram-approximation and batch-means approaches are just below the specified 90% with simulation run length. However, the hybrid approach obtains

CI coverage higher than the nominal value while its half-width is smaller than the half-width of the batch-means approach with simulation run length. In these simulation runs, there was one independent simulation run that reached the iteration limit L. Other results are similar to the M/M/1 delay-in-queue process. These experimental results indicate that the histogram variance estimates are often too small to achieve the desired coverage with the computation run length when the underlying distribution is bounded in one tail and the probability density function has relatively high values around the bound. One remedy for the histogram variance estimator being biased low with computation run length when the underlying distribution is bounded in one tail is heuristically to adjust the variance estimators when the output sequence is highly correlated.

Table 5 lists the experimental results from ASAP, ASAP2, and WASSP for AR1 process with correlation coefficient 0.9 and the M/M/1 process with traffic intensity 0.9 with the required relative precision set to 7.5%. These results are extracted directly from Steiger and Wilson 1999; Steiger et al. 2002; Lada, Wilson, and Steiger 2003. Hence, the underlying random number streams are different. Since these procedures terminate when they detect normality among batch means and deliver a correlation-adjusted CI half-width, they are able to provide valid CIs with relatively small sample sizes. On the other hand, the QI procedure generally requires larger sample sizes and delivers tighter CIs by default because it terminates only after it has obtained large enough samples to approximate the underlying distribution. We don't think this is a major drawback since wide half-widths provide little useful information. Furthermore, the empirical histogram can provide insight regarding the underlying distribution, such as quantiles.

Table 5: Coverage of 90% Confidence Intervals from ASAP and ASAP2

Procedure	ASAP	ASAP	ASAP2	WASSP
Process	AR1	MM1	MM1	MM1
avg. samp	24860	321468	281022	371380
coverage	91.0%	93.0%	92.0%	90.8%
avg. rp	0.059	0.069	0.070	-
avg. hw	0.118	0.620	0.628	0.5914
sdev hw	0.0003	0.003	0.002	0.006

The estimated required sample size for WASSP to obtain average M/M/1 CI half-width to be within 0.227210 is approximately 2516312 ($(0.5914/0.227210)^2 371380$), which is much greater than the average sample size 1377024 used by the QI procedure to obtain average CI half-width of 0.227210. In this regard, the QI procedure is very efficient in terms of sample size.

5 CONCLUSIONS

We have presented an algorithm for estimating the histogram of a stationary process. Some histogram estimates require more observations than others before the asymptotic approximation becomes valid. The proposed quasi-independent algorithm works well in determining the required simulation run length for the asymptotic approximation to become valid. The QI procedure estimates the required sample size based entirely on data and does not require any user intervention. Moreover, the QI procedure processes each observation only once and does not require storing the entire output sequence. Since the procedure stops when the QI subsequence appears to be independent, the procedure obtains high precision and small half-width with long simulation run length by default.

The histogram-approximation algorithm computes quantiles only at certain grid points and generates an empirical distribution (histogram) of the output sequence, which can provide valuable insights of the underlying stochastic process. The hybrid approach takes into consideration two different half-width estimators and may achieve better performance in terms of CI coverage and the average half-width. Since the QI procedure does not need to read the output sequence repeatedly, the storage requirement is minimal. The main advantage of the approach is that by using a straightforward runs test to determine the simulation run length and using natural estimators to construct the CI, we can apply classical statistical techniques directly and do not require more advanced statistical theory, thus making it easy to understand and simple to implement.

Because a histogram is constructed as an empirical distribution of the underlying process, it is possible to estimate other characteristics of the distribution, such as a proportion, quantile, or derivative (Chen 2003), under the same framework. Preliminary experimental results indicate that the natural estimators obtained based on the empirical distribution are fairly accurate.

REFERENCES

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

Chen, E. J. 2003. Derivative Estimation with Finite Differences. *Simulation: Transactions of The Society for Modeling and Simulation International* 79 (10): 598-609.

Chen, E. J. and W. D. Kelton. 2003. Determining Simulation Run Length with the Runs Test. *Simulation Modelling Practice and Theory* 11(3-4): 237-250.

Chen, E. J. and W. D. Kelton. 2004. Quantile and Tolerance-Interval Estimation in Simulation. *European Journal of Operational Research*. To Appear.

Goldsman D. and B. W. Schmeiser. 1997. Computational efficiency of batching methods. In *Proceedings of the*

1997 Winter Simulation Conference, ed. S. Andradóttir, K.J. Healy, D.H. Withers, and B.L. Nelson, 202–207. Piscataway, New Jersey:Institute of Electrical and Electronics Engineers.

Hillier, F. S. and G. J. Lieberman. 2001. *Introduction to Operations Research*. 7th ed. Boston: McGraw-Hill.

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

Lada, E. K., J. R. Wilson, and N. M. Steiger. 2003. A Wavelet-Based Spectral Method for Steady-State Simulation Analysis. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. E. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 422–430. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Schmeiser, B. 1982. Batch-size effects in the analysis of simulation output. *Operations Research* 30: 556–568.

Steiger, N. M. and J. R. Wilson. 1999. Improved batching for confidence interval construction in steady-state simulation. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, 442–451. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Steiger, N. M., E. K. Lada, J. R. Wilson, C. Alexopoulos, D. Goldsman, and F. Zouaoui. 2002. ASAP2: An improved batch means procedure for simulation output analysis. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 336–344. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

AUTHOR BIOGRAPHIES

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

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. 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.cba.uc.edu/faculty/keltonwd>.