Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

ADAPT SELECTION PROCEDURES TO PROCESS CORRELATED AND NON-NORMAL DATA WITH BATCH MEANS

E. Jack Chen

BASF Corporation
333 Mount Hope Avenue
Rockaway, New Jersey 07866, U.S.A.

ABSTRACT

Many simulation output analysis procedures are derived based on the assumption that data are independent and identically distributed (i.i.d.) normal; examples include ranking and selection procedures and multiple-comparison procedures. The method of batch means is the technique of choice to "manufacture" data that are approximately i.i.d. normal when the raw samples are not. Batch means are sample means of subsets of consecutive subsamples from a simulation output sequence. We propose to incorporate the procedure of determining the batch size to obtain approximately i.i.d. normal batch means into the selection procedures of comparing the performance of alternative system designs. We performed an empirical study to evaluate the performance of the extended selection procedure.

1 INTRODUCTION

Many simulation output analysis procedures are derived based on the assumption that data are independent and identically distributed (i.i.d.) normal; examples include ranking and selection procedures and multiple-comparison procedures. The goal of selection is to find the design that has the smallest or largest mean among a finite number of alternatives. When the raw samples are not i.i.d. normal, the method of batch means (BM) is the technique of choice to "manufacture" data that are approximately i.i.d. normal. See Law and Kelton (2000) for more details on batch means. Nakayama (1995) and Goldsman et al. (2000) have adapted ranking and selection procedures for steady-state simulation using batch means. However, they did not address the issue of determining the batch size m. Chen and Kelton (2007) have developed a procedure to manufacture batch means that appear to be i.i.d. normal, as determined by the von Neumann (1941) test of independence and the chi-square test of normality. In the remainder of the paper, we will use "data that appear to be i.i.d. normal" to mean that "the data has passed both tests of independence and normality." In this paper, we incorporate the BM procedure with other simulation analysis procedures, namely, the indifference-zone selection procedures.

A concern of simulation output analysis is to estimate the sampling error (or variance) of an estimator of some unknown parameter, i.e., the error caused by the estimator's randomness. This estimate provides information regarding the precision with which the estimator reflects the true but unknown parameter. Because we assume the process is already in steady-state and the underlying process is stationary, i.e., the joint distribution of the X_ι's is insensitive to time shifts, the mean estimator will be unbiased. Here $\{X_\iota : \iota = 1,2,\ldots,n\}$ is the simulation output sequence. On the other hand, the confidence interval (c.i.) coverage's closeness to the specified nominal value is dependent on the accuracy of the mean estimator and the c.i. half width, which in turn is dependent on the variance estimator.

Chen (2004) shows that the precision of selection procedures is closely related to the c.i. construction. However, the usual method of c.i. construction from classical statistics, which requires i.i.d. normal observations, is not directly applicable since simulation output data are generally correlated and non-normal. The BM procedure of Chen and Kelton (2007) determines batch sizes based entirely on data and does not require user intervention. 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). Let μ be the expected value of the process and let $\bar{X}(n) = \sum_{\iota=1}^{n} X_\iota / n$ be the sample mean

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

of n samples. These weakly dependent, stationary processes typically obey a central limit theorem of the form

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

where Ω^2 is the steady-state variance constant (SSVC), $N(\mu,\sigma^2)$ denotes the normal distribution with mean μ and variance σ^2, and $\xrightarrow{D}$ denotes convergence in distribution.

For correlated sequences, the SSVC

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

where $\gamma_i = \text{Cov}(X_k, X_{k+i})$ for any k is the lag-i covariance. For stationary processes, a sufficient condition for the SSVC to exist is $\sum_{-\infty}^{\infty} |\gamma_i| < \infty$. If the sequence is independent, then the SSVC is equal to the process variance $\sigma_x^2 = \text{Var}[X_l]$. Our procedures may fail when the underlying stochastic process does not satisfy the limiting result of (1).

The rest of this paper is organized as follows. In Section 2, we present some background on ranking and selection, and the batch-means method of estimating the SSVC. In Section 3, we present the algorithm to estimate batch sizes such that the batch means appear to be i.i.d. normal, and selection using batch means. In Section 4, we show some empirical-experimental results. In Section 5, we give concluding remarks.

2 BACKGROUND

In this section, we review the basis of indifference-zone selection procedures, non-overlapping batch means, and the Von Neumann test of independence.

2.1 Basis of Indifference-Zone Selection

Selection procedures are used to select the "best" design from among k competing designs, and these procedures often are derived based on the assumption that the input data are i.i.d. normal. Let μ_{i_l} be the l^{th} smallest of the μ_i's, so that $\mu_{i_1} \leq \mu_{i_2} \leq \ldots \leq \mu_{i_k}$. The goal is to find a design with the smallest expected response μ_{i_1}. In practice, however, if μ_{i_1} and μ_{i_2} are very close together, we are indifferent if we mistakenly choose design i_2, whose expected response is μ_{i_2}. The "practically significant" difference d^* (a positive real number) between the best and next-best design is called the *indifference zone* in the statistical literature and represents the smallest difference about which we care. Therefore, we want a procedure that avoids making a large number of replications or batches to resolve differences less than d^*. Let P(CS) denote the probability of correct selection, i.e., the best design is selected. Then we want $P(\text{CS}) \geq P^*$ provided that $\mu_{i_2} - \mu_{i_1} \geq d^*$, where the minimal correct selection probability P^* and the "indifference" amount d^* are both specified by the user. For a comprehensive review of indifference-zone selection, please see Bechhofer et al. (1995).

In testing the null hypothesis $\text{H}_0 : \mu_i \leq \mu_{i_1}$, we reject the null hypothesis and conclude with confidence level $1-\alpha$ that $\mu_i > \mu_{i_1}$ is the same as checking that the lower endpoint of the one-tailed $1-\alpha$ c.i. is positive, i.e., $\mu_i - \mu_{i_1} > \hat{\mu}_i - \hat{\mu}_{i_1} - w_i > 0$, where $\hat{\mu}_i = \bar{X}_i(n_i)$ is the sample means of n_i samples from process i, and w_i denotes the one-tailed $1-\alpha$ c.i. half width. With independent sampling and $n_i = n_{i_1}$, $w_i = t_{1-\alpha,n_i-1}\sqrt{(S_i^2(n_i)+S_{i_1}^2(n_{i_1}))/n_i}$. Let $X_{i\iota}$ denote the ι^{th} sample from process i. Here

$$S_i^2(n_i) = \frac{1}{n_i - 1}\sum_{\iota=1}^{n_i}(X_{i\iota} - \hat{\mu}_i)^2,$$

is the unbiased estimator of σ_i^2 and $t_{1-\alpha,f_i}$ is the $1-\alpha$ quantile of the t distribution with f_i degrees of freedom (df). For additional details on duality of hypothesis test and c.i., please see Rice (1995). The half width w_i depends on the sample sizes and becomes smaller as the sample sizes become larger. This implies that the sample sizes (n_i and n_{i_1}) should be large enough so that $w_i < \hat{\mu}_i - \hat{\mu}_{i_1}$. By symmetry of the normal distribution $\text{P}\{\hat{\mu}_i - \hat{\mu}_{i_1} > (\mu_i - \mu_{i_1}) - w_i\} \geq 1-\alpha$. To obtain $\text{P}\{\hat{\mu}_i - \hat{\mu}_{i_1} > 0\} \geq 1-\alpha$, the sample size should be large enough so that $w_i < \mu_i - \mu_{i_1}$.

Let $d_{i_l} = \mu_{i_l} - \mu_{i_1}$ and $\hat{d}_{i_l} = \hat{\mu}_{i_l} - \hat{\mu}_{i_1}$. Let $CI1$ denote the the one-tailed $1-\alpha$ c.i. of d_{i_l} obtained by procedures that are developed based on the least favorable configuration (LFC) (i.e., assuming $\mu_{i_1} + d^* = \mu_{i_2} = \cdots = \mu_{i_k}$). Let $CI2$ denote the one-tailed $1-\alpha$ c.i. of d_{i_l} obtained by procedures that take into account sample means. The confidence interval $CI1 = (\hat{d}_{i_l} - d^*, \infty]$

since the procedures achieve $w_{i_l} < d^*$. On the other hand, $CI2 = (\hat{d}_{i_l} - d_{i_l}, \infty] \approx (0, \infty]$ since the procedures attempt to achieve $w_{i_l} < d_{i_l}$. Hence, the allocated sample sizes are just large enough to conclude $\mu_{i_1} < \mu_i$ (provided $\mu_{i_1} + d^* \leq \mu_i$) with a desired confidence but not more than necessary. Note that the smaller the indifference amount d^*, the smaller the c.i. half width needs to be and the larger the required sample sizes need to be.

2.2 Non-Overlapping Batch Means

The batch means method is a relatively simple idea that can be used to manufacture almost i.i.d. normal "samples" by breaking the output sequences into a few large batches of many individual observations. In the non-overlapping batch-means method, the simulation output sequence $\{X_\imath : \imath = 1, 2, \ldots, n\}$ 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

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

Under the null hypothesis that the batch size m is large enough such that the batch means are independent, we have the SSVC estimator $\Omega^2(m) \equiv m\mathrm{Var}[\bar{X}_j] = \gamma_0 + 2\sum_{i=1}^{m-1}(1 - i/m)\gamma_i$. The grand mean $\hat{\mu}$ of the individual BM, given by

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

is used as a point estimator for μ. Here $\hat{\mu} = \bar{X}(n)$, the sample mean of all n individual $X_\imath$'s.

The method of BM is a well-known technique for estimating the variance of point estimators computed from simulation experiments. The BM method tries to reduce autocorrelation by batching observations. The BM variance estimator (of estimating the variance of sample means, i.e., $\mathrm{Var}[\bar{X}_j]$) is simply the sample variance of the mean estimator $\bar{X}_j$ computed from the sample (batch) means of subsets of consecutive subsamples, i.e.,

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

The BM estimator for Ω^2 is

$$\hat{\Omega}^2 \equiv mS^2(b). \tag{3}$$

The $1 - \alpha$ c.i. constructed by the BM method is

$$\hat{\mu} \pm t_{1-\alpha/2,b-1} S(b)/\sqrt{b} = \hat{\mu} \pm t_{1-\alpha/2,b-1} \sqrt{mS^2(b)}/\sqrt{mb} = \hat{\mu} \pm t_{1-\alpha/2,b-1} \hat{\Omega}/\sqrt{n}.$$

Asymptotic validity of this c.i., i.e., that the coverage of the c.i. is close to the nominal coverage probability, often depends on the assumption that the batch means are approximately i.i.d. normal. That is, for large batch size m the batch means are approximately i.i.d. normal with unknown mean μ and unknown variance $\Omega^2(m)/m$. However, some BM methods (for example, overlapping BM by Meketon and Schmeiser 1984 and ASAP3 by Steiger et al. 2005) construct a c.i. of the mean by adjusting the c.i. half width based on the strength of the autocorrelations between BM. However, (3) may not hold since the BM may be correlated. The key idea for ASAP3 is that the batch means may become approximately jointly normally distributed before becoming approximately independent; thus, it adjusts the half width to counter any remaining dependence between the batch means. Consequently, these procedures are able to deliver valid c.i. with relatively small sample sizes (and relatively large half widths). Moreover, the BM manufactured by these procedures are likely to be correlated and cannot be used as input data for procedures that require i.i.d. normal data.

There are a number of batch-size-determination procedures that aim to manufacture independent batch means. However, these methods have focused on selecting a batch size large enough to achieve near independence of the batch means and have ignored the question of normality based on one assumption. If the batch size is large enough for the batch means to be approximately independent, then the batch size should be large enough for the batch means to be approximately normally distributed as well. Evidently, this assumption is not true in general (see Chen and Kelton 2007). Our experimental results indicate that it is not as critical to ensure that the batch means are normally distributed as it is to ensure that the batch means are independent in terms of c.i. coverage. However, some other procedures (e.g., ranking and selection) will require the data to be i.i.d. normal.

2.3 The von Neumann Test of Independence

We briefly review the von Neumann test (Fishman 2001, von Neumann 1941) for the hypothesis H_0: the batch means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_b$ are uncorrelated. The von Neumann ratio is

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

Note that C_b is an estimator of the lag-1 autocorrelation $\omega_1 \equiv \mathrm{Corr}[\bar{X}_j, \bar{X}_{j+1}]$, adjusted for end effects that diminish in importance as the number of batches b increases. If $\{X_\iota : \iota = 1, 2, \ldots, n\}$ has a monotone non-increasing autocorrelation function, then ω_1 is positive and decreases monotonically to zero as the batch size m increases. If for the given b and m, H_0 is true for $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_b$, then $\omega_1 = 0$.

The von Neumann test statistic for H_0 is

$$Z = \sqrt{\frac{b^2-1}{b-2}}C_b.$$

Let α_{ind} denote the acceptable level of committing a Type I error (i.e., rejecting the null hypothesis when it is true). Under H_0, $Z \sim N(0,1)$, so one rejects H_0 at level $1-\alpha_{\mathrm{ind}}$ if $Z > z_{1-\alpha_{\mathrm{ind}}}$, where $z_{1-\alpha_{\mathrm{ind}}}$ is the $1-\alpha_{\mathrm{ind}}$ quantile of the standard normal distribution.

3 SELECTION USING BATCH MEANS

In this section, we discuss the strategy of estimating the required batch size such that the batch means appear to be i.i.d. normal. While there are many BM procedures, most of them are not suitable to be used as a pre-processor of procedures that require i.i.d. normal data; see Section 2.2. One of the BM procedures that is a good candidate for this purpose is the quasi-independent-and-normal (QIN) procedure of Chen and Kelton (2007). The procedure is a data-based algorithm, i.e., the procedure can be embodied in a software package whose input is the data $\{X_\iota : \iota = 1, 2, \ldots, n\}$ and whose output are i.i.d. normal batch means. We briefly review the QIN procedure and subsequently incorporate the QIN procedure into a selection procedure.

3.1 Validation of Normality

To determine whether the batch means appear to be normally distributed, we check the proportions of the values of batch means in each interval bounded by -∞, $\hat{\mu} - 0.9674S$, $\hat{\mu} - 0.4303S$, $\hat{\mu}$, $\hat{\mu} + 0.4303S$, $\hat{\mu} + 0.9674S$, ∞, where $\hat{\mu}$ and S are, respectively, the grand sample mean of the n observations and the standard deviation of these b batch means. The constants involved in these intervals are strategically chosen so that the proportions of the batch means in each interval are equal under the null hypothesis that the batch means are normal. We apply the chi-square test of normality (see Law and Kelton 2000, Rice 1995) to these batch means, using a confidence level of 0.9.

The probability of this chi-square test of normality to commit either Type I or Type II errors is high when the data are highly correlated. Recall that a Type II error is accepting the null hypothesis when it is false. Since the QIN procedure performs the normality test only on BM that passed the test of independence, this deficiency of the chi-square test of normality will not cause difficulty for the QIN procedure. Our experimental results indicate that the residual correlation of BM that passed the test of independence is weak, if any.

The powers of the independence and the normality tests increase as the number of batches used to perform the tests increases. Based on our experimental results, we recommend that the sample size used for this normality test and the von Neumann test be at least 180. There are other more sophisticated normality tests, for example, the Shapiro-Wilk test (Bratley et al. 1987). We chose the chi-square test because it is easy to apply and serves our purpose well. Nonetheless, other normality tests could be used in place of the chi-square test of normality if desired.

For instance, the average batch size for $U(0,1)$ (uniformly distributed between 0 and 1) observations to pass this chi-square normality test is 1.85. It is known that the average of two $U(0,1)$ random variates is a random variate having a triang$(0,1,0.5)$ distribution, where triang(a,b,c) denotes a triangular distribution on $[a,b]$ with mode c. Even though the triangular distribution is symmetric, it is certainly not nearly normal. This result indicates that the power of the chi-square normality test with only 5 degrees of freedom (df) is weak. Nevertheless, if we mistakenly treat this kind of non-normal data as being approximately normal, the c.i. coverage remains fairly accurate.

3.2 Manufacturing i.i.d. Normal Batch Means

The QIN procedure progressively increases the batch size m until the batch means appear to be i.i.d. normal, as determined by the von Neumann and the chi-square tests. This procedure can be viewed as the initial step of the selection procedures or other procedures that require i.i.d. normal data. We divide the entire output sequence into $b = 180$ batches. We allocate a buffer with size $3b$ to keep *primitive batches* (PB), i.e., the average of l (a positive integer) observations. Batch means are then computed from PB in the buffer. Initially each observation is treated as a PB; as the procedure proceeds, l (the number of observations used to compute PB) will be doubled in every other iteration. The steps between collecting more observations is considered an iteration.

To facilitate the description of the algorithm, we break each iteration into A and B sub-iterations so that the most relevant quantities can be described in simpler forms. The procedure performs both independence and normality tests in both sub-iterations and terminates when BM appear to be i.i.d. normal. Table 1 illustrates how sampling progresses from iteration to iteration. The *Iter* row shows the index of the iteration. The n row lists the total number of observations to be taken at a certain iteration. The p_b row lists the total number of PB in the buffer at a certain iteration. The l row lists the number of observations used to obtain the PB in the buffer. The m lists the batch size. There may be b, $2b$, or $3b$ PB in the buffer. We aggregate the available PB into b BM by averaging the adjacent PB.

Table 1: Properties of QIN at each iteration

Iter	0	1_A	1_B	2_A	2_B	...	q_A	q_B
n	b	$2b$	$3b$	$4b$	$6b$	...	$2^q b$	$2^{q-1}3b$
p_b	b	$2b$	$3b$	$2b$	$3b$	...	$2b$	$3b$
l	1	1	1	2	2	...	2^{q-1}	2^{q-1}
m	1	2	3	4	6	...	2^q	$2^{q-1}3$

The quasi-independent-and-normal algorithm:

The size of the buffer used to store the PB is $3b$. l is the number of observations used to compute PB. δ is the incremental sample size. q is the index of iterations. Each iteration q contains two sub-iterations q_A and q_B.

1. Initialization: Set $b = 180$, $l = 1$, $\delta = b$, and $q = 0$.
2. Generate δ PB, where each PB is the average of l observations. Store these PB in the buffer.
3. If this is the initial iteration, set $c = 1$. If this is a q_A iteration, set $c = 2$. If this is a q_B iteration, set $c = 3$. Set the value of the batch means to the average of c consecutive PB in the buffer (i.e., $\bar{X}_j = \sum_{c(j-1)+1}^{cj} V_j/c$ for $j = 1,2,\ldots,b$, where V_j is the value of the j^{th} PB).
4. Carry out the von Neumann and the chi-square tests to determine whether these b batch means appear to be i.i.d. normal.
5. If the batch means appear to be i.i.d. normal, go to step 10.
6. If this is the initial or a q_B iteration, set $q = q+1$ and start a q_A iteration. If this is a q_A iteration, start a q_B iteration.
7. If this is a q_A iteration ($q > 1$), then re-calculate the PB in the buffer by taking the average of two consecutive PB, and reindex the rest of the $3b/2$ PB in the first half of the buffer. Set $l = 2^{q-1}$ and $\delta = b/2$.
8. If this is a q_B iteration ($q > 1$), set $\delta = b$.
9. Go to step 2.
10. Deliver the batch size m and the i.i.d. normal batch means.

Since the BM manufactured by the QIN procedure appear to be i.i.d. normal, we can use these BM to construct a classical c.i. without any adjustment and use these BM as input for other simulation procedures that require i.i.d. normal data. The QIN procedure needs to process each observation only once and does not require storing the entire output sequence. Furthermore, the derivation of the procedure is rather straightforward, which makes the QIN procedure easy to understand and simple to implement as a practical procedure for manufacturing approximately i.i.d. normal data.

3.3 Selection via All Pairwise Comparisons

Selection procedures that are derived based on the LFC are conservative, i.e., they often allocate large sample sizes and achieve higher than the specified precision. Chen and Kelton (2005) propose a sequential selection procedure that takes into account the difference of sample means and is effective in terms of sample sizes. This selection procedure achieves P(CS) $\approx P^*$, provided that the difference between the best and the second-best designs is at least d^*, and d^* is at least 10% of the standard error of the difference between the sample means of the best and the second best. It is our intention for the batch means to play the role of the i.i.d. normal observations that the original version of the procedure requires. The batch sizes are determined dynamically and are random variables. Moreover, the batch size is likely to be one when the raw data are i.i.d. normal. The extension of the selection procedure is as follows:

The Sequential Selection Procedure via All Pairwise Comparisons:

1. Initialize the set I to include all k designs. Let $\bar{X}_{i\iota}$ be the ι^{th} batch mean of design i. Let $b_{i,q}$ be the number of batches allocated for design i, and $\hat{\mu}_{i,q}$ be the sample mean (i.e., the average of batch means) of design i at the q^{th} iteration.
2. Run the QIN procedure to determine the batch size m_i for each design i. Specify the initial number of batches b_0, the value of the indifference amount d^*, and the minimal required precision P^*. Set $P = 1-(1-P^*)/(k-1)$, the iteration index $q = 0$, and the current number of batches $r = b_{1,q} = b_{2,q} = \cdots = b_{k,q} = b_0$. Simulate b_0 batches for each design $i \in I$.
3. Perform all pairwise comparisons and delete inferior design i from I; i.e., $\hat{\mu}_{i,q} > \hat{\mu}_{j,q} + w_{ij}$, $i, j \in I$. Note that w_{ij} is the one-tailed P c.i. half width.
4. If $w_{ij} < d^*$ and $\hat{\mu}_{i,q} > \hat{\mu}_{j,q}$, remove design i from I.
5. If there is only one element (or the pre-determined number of best designs) in I, go to step 9.
6. Compute the critical value $h_t = \sqrt{2}t_{P,r-1}$, where $t_{P,f}$ denotes the P quantile of the t distribution with f df.
7. Let $\hat{\mu}_{B,q} = \min_{i\in I}\hat{\mu}_{i,q}$. For $\forall i \in I$, compute $\hat{d}_{i,q} = \max(d^*, \hat{\mu}_{i,q} - U(\hat{\mu}_{B,q}))$, where $U(\hat{\mu}_{B,q})$ is the upper one-tailed P^* confidence limit of the unknown true mean μ_B at the q^{th} iteration, and compute

$$\delta_{i,q+1} = \lceil ((h_t S_i(r)/\hat{d}_{i,q})^2 - r)^+ \rceil.$$

   Note that $S_i^2(r)$ is the variance of the batch means of design i.
8. Set $q = q+1$. If $\delta_{i,q} = 0$, set $\delta_{i,q} = 1$. Set the incremental sample size at the q^{th} iteration $\delta_q = \min_{i\in I}\delta_{i,q}$. For $\forall i \in I$, simulate additional δ_q batches, set $r = r + \delta_q$. Go to step 3.
9. Let $\hat{\mu}_i = \sum_{\iota=1}^{b_i}\bar{X}_{i\iota}/b_i$ denote the final sample mean of the i^{th} design. Return the values B and $\hat{\mu}_B$, where $\hat{\mu}_B = \min\hat{\mu}_i$, $1 \le i \le k$, and i was not eliminated by all pairwise comparisons.

The batch size of each design m_i is determined independently, i.e., the batch sizes may be different among these designs and can be different at different simulation runs. Hence, the procedure can perform selection among designs having different families of distributions. A small initial number of batches b_0 can reduce the P(CS) of the selection procedure because a small b_0 increases the frequency of incorrect eliminations of the best design at the initial comparisons. Since the QIN procedure uses 180 batch means for the tests of independence and normality, we use all the samples that are available and set $b_0 = 180$. In some cases, doing so may result in allocating larger than necessary sample sizes and achieving higher than required precision.

The required number of batches for design i is calculated by

$$b_i = \lceil (h_t S_i(r)/\hat{d}_{i,q})^2 \rceil.$$

Let μ_i be the expected performance measure of design i and $\mu_B = \min \mu_i$. The basis for this sample size allocation strategy is to ensure that the one-tailed P c.i. half width of $\mu_i - \mu_B$ is less than $d_i = \max(d^*, \mu_i - \mu_B)$ (see Chen 2004). From (1), if the sample size n is large enough, then the c.i. half width of the mean of correlated sequences can be computed by $z_P \Omega/\sqrt{n}$, where z_P is the P quantile of the standard normal distribution. Consequently, to achieve the same precision as an i.i.d. normal sequence, the required number of unbatched raw samples of correlated sequence should be

$$n_i = \lceil (h_t \hat{\Omega}_i / \hat{d}_{i,q})^2 \rceil,$$

where $\hat{\Omega}_i$ is the estimator of Ω_i. Similarly, the SSVC estimate of design i $\hat{\Omega}_i^2 = m_i S_i^2(b_i)$, where m_i is the batch size, and b_i is the number of batches. Hence, the unbatched sample size is

$$\begin{aligned} n_i &= m_i b_i \\ &= m_i \lceil (h_t S_i(b_i)/\hat{d}_{i,q})^2 \rceil \\ &\approx \lceil (h_t \hat{\Omega}_i / \hat{d}_{i,q})^2 \rceil. \end{aligned}$$

Consequently, any technique of estimating SSVC Ω^2 can be used to estimate the required unbatched sample sizes, e.g., various standardized time series analysis (Healy et al. 2007) and spectrum analysis (Sullivan and Wilson 1989).

4 EMPIRICAL EXPERIMENTS

In this section, we present some empirical results from simulation experiments using the proposed procedure. We use 180 batch means for the von Neumann test of independence and the chi-square test of normality. We evaluate the performance of using the approximately i.i.d. normal batch means manufactured by the QIN procedure as input for the selection procedure. The basic characteristics of the BM manufactured by the QIN procedure are available in Chen and Kelton (2007). The confidence level of the von Neumann tests and the chi-square tests is set to 0.9 in our experiments. Note that a lower confidence level of these tests will increase the chance of committing a Type I error and will also increase the batch size and the simulation run length.

There are k alternative designs in the selection subset. The least favorable configuration is used, in which μ_1 has the smallest value, while $\mu_1 + d^* = \mu_2 = \mu_3 = \cdots = \mu_k$. We did not perform any experiments with the non LFC, since it has been shown that procedures that take into account the difference of sample means can significantly reduce the required sample sizes. The number of systems (k) is set to either 2 or 5 and the number of initial samples or batch means (b_0) is set to either 20 or 180. Recall that the procedure has generated 180 batches for the test of normality and test of independence. We want to select a design with the minimum mean: design 1. Furthermore, 1,000 independent experiments are performed to estimate the actual P(CS) by $\hat{P}$(CS): the proportion of the 1,000 experiments in which we obtained the correct selection.

4.1 Experiment 1

We test the procedure with the following three independent stochastic processes:

- Observations are i.i.d. normal. The best design has a $N(0,1)$ distribution, while all other designs have a $N(0.2,1)$ distribution. The indifference amount d^* is set to 0.2.
- Observations are i.i.d. uniform between a and b, denoted $U(a,b)$. The best design has a $U(0,1)$ distribution, while all other designs have a $U(0.1,1.1)$ distribution. The indifference amount d^* is set to 0.1.
- Observations are i.i.d. exponential with mean μ, denoted expon(μ). The best design has an expon(1) distribution, while all other designs have an expon(1.2) distribution. The indifference amount d^* is set to 0.2.

Table 2 lists the experimental results. The $\hat{P}$(CS) column lists the proportion of correct selection. The T column lists the average of the number of total unbatched observations and batched observations (in parenthesis) used in each experimental design. The stdv(T) column lists the standard deviation of the number of total observations unbatched and batched (in parenthesis) at each independent simulation run. The normal(20) row lists the results when the underlying distributions are normally distributed and $b_0 = 20$. All other rows are defined similarly. The observed $\hat{P}$(CS)'s are around the specified nominal value of 0.95. Since the procedure is derived based on the conservative Bonferroni Inequality, the $\hat{P}$(CS)'s are generally higher than the specified P^* as the number of k increases. The obtained batch sizes are approximately 1.1, 2.0, and

5.1, respectively, for normal, uniform, and exponential distributions. The procedure correctly increases the batch size as the distributions depart farther away from the normal distribution. With $\alpha = 0.10$ in the normality test, data that are normally distributed will fail the test 10% of the time. The average batch sizes for normal distribution, 1.1, are close to the theoretical value, i.e., $\sum_{i=0}^{\infty} \alpha^i$.

Table 2: Results of Experiment 1 - $\hat{P}$(CS) of the independent processes with $P^* = 0.95$

	$k = 2$			$k = 5$		
Distribution	$\hat{P}$(CS)	T	stdv(T)	$\hat{P}$(CS)	T	stdv(T)
normal(20)	0.955	247 (220)	108 (101)	0.963	1148 (1016)	267 (260)
uniform(20)	0.946	96 (49)	30 (13)	0.975	405 (210)	86 (54)
expon(20)	0.954	323 (66)	149 (30)	0.971	1487 (308)	416 (98)
normal(180)	0.978	424 (360)	151 (0)	0.967	1252 (1074)	238 (127)
uniform(180)	1.000	732 (360)	257 (0)	1.000	1825 (900)	398 (0)
expon(180)	0.999	1911 (360)	944 (0)	0.995	4806 (900)	1566 (0)

In this setting, the required number of batches for each design is less than 180, except for $k = 5$ with normally distributed samples. Hence, the procedure results in allocating a larger than necessary number of batches and achieving high precision when the initial number of batches b_0 is set to 180. With such a large initial number of batches, the procedure has selected the best design at the initial iteration and no further simulation is required. Furthermore, many samples can be generated simultaneously when deploying ranking and selection in a parallel and distributed environment (see Chen 2005).

4.2 Experiment 2

Goldsman et al. (2000) investigated the P(CS) and sample size with a series of pre-determined batch sizes. In order to compare our results with theirs, we perform an experiment with a similar setting.

Suppose $X_{il} = \mu_i + \varphi(X_{il-1} - \mu_i) + \varepsilon_i$, $i = 1, 2, \ldots, k$, where $\varepsilon_i \sim N(0, 1 - \varphi^2)$. We denote this sequence the AR1(φ) processes, in which μ_1 was set to 0, while $\mu_2 = \mu_3 = \cdots = \mu_k = d^*$. Note that the marginal variance of each system σ^2 is 1. Let Ω_1^2 be the steady-state variance constant of the best system. The indifference amount d^* is set to $\sqrt{\Omega_1^2/70}$ and $\sqrt{\Omega_1^2/1000}$, respectively, for $\varphi = 0.3$ and $\varphi = 0.9$. Note that for the AR1(φ) processes, the SSVC $\Omega^2 = \sigma^2(1+\varphi)/(1-\varphi)$. Consequently, the indifference amounts are approximately 0.1629 and 0.1378, respectively, for $\varphi = 0.3$ and $\varphi = 0.9$.

Table 3 lists the experimental results. The b_0 column lists the initial number of batches. The φ column lists the value of the correlation coefficient. All other columns are as defined previously. The $\hat{P}$(CS)'s are around the nominal value when the auto-regressive sequences are pre-processed to become approximately i.i.d. normal batch means before used as input for the selection procedure. The residual autocorrelations between BM may have caused the observed P(CS) to be lower than desired when $b_0 = 20$ and $k = 2$. If this is a concern, a lower confidence level for the test of independence can be used to reduce the frequency of committing a Type II error.

Table 3: Results of Experiment 2 - $\hat{P}$(CS) of the AR1(φ) processes with $P^* = 0.95$

		$k = 2$			$k = 5$		
b_0	φ	$\hat{P}$(CS)	T	stdv(T)	$\hat{P}$(CS)	T	stdv(T)
20	0.3	0.939	629 (124)	319 (64)	0.953	2914 (581)	821 (171)
	0.9	0.942	8684 (121)	4044 (62)	0.950	41767 (561)	11785 (161)
180	0.3	0.989	2121 (360)	1740 (0)	0.965	5198 (911)	2636 (29)
	0.9	0.986	29016 (360)	17723 (0)	0.962	74892 (904)	27356 (14)

The observed average batch sizes are approximately 5 and 74, respectively, for $\varphi = 0.3$ and $\varphi = 0.9$. These results are consistent with the results of Goldsman et al. (2000) because the batch sizes determined by the QIN procedure are close to the pre-determined batch sizes with the allocated total unbatched observations and the observed P(CS).

4.3 Experiment 3

In this experiment, the output data are not only correlated but also non-normal. Suppose $X_{il} = \mu_i + \varphi(X_{il-1} - \mu_i) + \varepsilon_i - \sqrt{1-\varphi^2}$, $i = 1, 2, \ldots, k$, where $\varepsilon_i \sim \text{expon}(\sqrt{1-\varphi^2})$. We denote this sequence the EAR1(φ) (exponential AR1) processes, in which μ_1 was set to 0, while $\mu_2 = \mu_3 = \cdots = \mu_k = d^*$. The indifference amounts are set to approximately 0.1629 and 0.1378, respectively, for $\varphi = 0.3$ and $\varphi = 0.9$.

Table 4 lists the experimental results. All Columns are as defined previously. The $\hat{P}$(CS)'s are around the nominal value. The observed average batch sizes are approximately 7 and 76, respectively, for $\varphi = 0.3$ and $\varphi = 0.9$. For a slightly correlated sequence, i.e., $\varphi = 0.3$, the batch sizes for BM to appear to be i.i.d. normal are significantly different between the AR1 and EAR1 processes. On the other hand, for a highly correlated sequence, i.e., $\varphi = 0.9$, the batch sizes for BM to appear to be i.i.d. normal are approximately the same between the AR1 and EAR1 processes. It can be shown that the steady-state distribution of the EAR1 process is a gamma distribution (see Lewis et al. 1989). As the correlation coefficient φ increases, the shape parameter of the gamma distribution becomes larger, and the distribution has a higher degree of symmetry.

Table 4: Results of Experiment 3 - $\hat{P}$(CS) of the EAR1(φ) processes with $P^* = 0.95$

		$k=2$			$k=5$		
b_0	φ	$\hat{P}$(CS)	T	stdv(T)	$\hat{P}$(CS)	T	stdv(T)
20	0.3	0.947	671 (96)	305 (48)	0.956	3034 (433)	930 (132)
	0.9	0.938	8990 (120)	4399 (61)	0.947	41820 (549)	10147 (159)
180	0.3	0.999	2842 (360)	1582 (0)	0.979	7382 (900)	3253 (2)
	0.9	0.983	30824 (360)	17650 (0)	0.965	77072 (902)	27494 (9)

4.4 Experiment 4

Since batch means are often used to estimate the variance of sample means, we evaluate the accuracy of the c.i. constructed by the QIN procedure. In these experiments, neither the relative precision nor the absolute precision are specified, so the half width of the c.i. is the result of the default precision. In this experiment, we list the results of c.i. coverage when the entire sequences are divided into 180 batches.

We test the procedure with the following three stochastic processes:

- The AR1(φ) processes are as defined in Section 4.2. We set μ to 2 in our experiments. We set X_0 to a random variate drawn from $N(\mu, 1)$.
- The EAR1(φ) processes are as defined in Section 4.3. We set μ to 2 in our experiments. We set X_0 to a random variate drawn from gamma $((1+\varphi)/(1-\varphi), \sqrt{(1-\varphi)/(1+\varphi)})$ plus $\mu - \sqrt{(1+\varphi)/(1-\varphi)}$, where gamma$(a,b)$ denotes a gamma distribution with shape parameter a and scale parameter b.
- Steady-state of the M/M/1 delay-in-queue processes has the arrival rate (λ) and the service rate ($\mu = 1$). This process is denoted MM1(ρ), where $\rho = \lambda/\mu$ is the traffic intensity.

Table 5 lists the experimental results. The avg and stdv r rows list, respectively, the average relative precision and the standard deviation of the relative precision of the estimator $\hat{\mu}$. Here, the relative precision is defined as $r = |\hat{\mu} - \mu|/\hat{\mu}$. The T and stdv(T) rows list, respectively, the average sample size and the standard deviation of the sample size. The avg and stdv bsize rows list, respectively, the average batch size and the standard deviation of the batch size obtained by the procedure. The avg and stdv hw rows list, respectively, the average half width and the standard deviation of the half width obtained by the procedure. The coverage row lists the percentage of the c.i.'s that cover the true mean. For these three tested processes, the c.i. coverages are slightly below the specified 90% confidence level. Since the steady-state distribution of the AR1 process is normal, we believe that some of the batch means that passed the test of independence may be slightly correlated. For the EAR1 and M/M/1-queuing processes, samples are not only highly correlated but also non-normal. Furthermore, the steady-state distribution of the M/M/1-queuing processes is not only asymmetric but also discontinuous at $x = 0$. Hence, the batch size for the M/M/1-queuing-processes BM to appear i.i.d. normal is significantly larger than the batch size for the AR1 and EAR1 processes. In addition, some correlated batch means pass the test of independence and some non-normal batch means pass the test of normality; thus, the coverage of the EAR1 and M/M/1-queuing processes is slightly lower than that of the AR1 process.

Table 5: Results of Experiment 4 - Coverage of 90% confidence intervals of correlated samples

Process	AR1(0.9)	EAR1(0.9)	MM1(0.9)
μ	2.00	2.00	9.00
avg r	0.0160	0.0165	0.0121
stdv r	0.0134	0.0137	0.0103
T	14560	15250	3296931
stdv(T)	12481	13028	2992981
avg bsize	80.9	84.7	18316
stdv bsize	69.3	72.4	16628
avg hw	0.0613	0.0607	0.202
stdv hw	0.0117	0.0121	0.0641
coverage	88.1%	85.4%	87.0%

We would like to point out that the average relative precisions of the mean estimator are within 2% and the average half widths are around or within 3% of the mean. Hence, those c.i.'s that do not cover the true mean must have missed it by only a small amount. Interested readers can see Chen and Kelton (2007) for some performance comparisons of QIN against other batch-means procedures. Since the objective of the QIN procedure is to obtain approximately i.i.d. normal batch means, it generally requires larger sample sizes and delivers tighter c.i.'s by default.

5 CONCLUSIONS

We have presented an algorithm for estimating the required batch size so that the batch means appear to be i.i.d. normal and a strategy of incorporating the algorithm into selection procedures for comparing the means of several steady-state simulation responses. The QIN algorithm works well in determining the required batch size for the batch means to become approximately i.i.d. normal. Our approach has desirable properties; it is a sequential procedure and does not require users to have *a priori* knowledge of values that the data might assume. Thus, the user can apply the ranking and selection procedures without having to execute a separate pilot run to determine whether the raw samples appear to be i.i.d. normal. The experimental evaluation reveals that the QIN procedure determines batch sizes that are sufficiently large for achieving approximately i.i.d. normal batch means and adequate c.i. coverage.

The main advantage of the approach is its use of a straightforward test of independence and test of normality to obtain approximately i.i.d. normal batch means. We can apply classical statistical techniques directly without more advanced statistical theorems, thus making it easy to understand and simple to implement. The procedure estimates the required sample size based entirely on data and does not require user intervention. Furthermore, it needs to process each observation only once and does not require storing the entire output sequence. Clearly, we can increase the flexibility of selection procedures by using the QIN procedure to pre-process the observations.

Some extensions of ranking and selection include selecting a (restricted) subset (Chen 2008, Chen 2009), comparison-with-a-standard (control) and multiple comparisons (Chen 2006). Those procedures are also derived based on the assumption that the input data are i.i.d. normal. Hence, the discussion in this paper can also be applied to those procedures. Instead of selecting a single best design, the goal of subset selection is to select a subset of size m containing c of the v best from k alternatives. In comparison with a standard (control), we compare a finite number of designs with respect to a single standard (control). Multiple comparisons provide simultaneous c.i.'s on selected differences among the designs.

REFERENCES

Bechhofer, R. E., T. J. Santner, and D. M. Goldsman. 1995. *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*. New York: John Wiley & Sons, Inc.

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

Bratley, P., B. L. Fox, and L. E. Schrage. *A Guide to Simulation*. 2nd ed. New York: Springer-Verlag, 1987.

Chen, E. J. 2004. Using Ordinal Optimization Approach to Improve Efficiency of Selection Procedures. *Journal of Discrete Event Dynamic Systems* 14(2): 153-170.

Chen, E. J. 2005. Using Parallel and Distributed Computing to Increase the Capability of Selection Procedures. *Proceedings of the 2005 Winter Simulation Conference*, ed. Kuhl ME, Steiger NM, Armstrong FB, and Joines JA, 723-731. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Chen, E. J. 2006. Comparison With a Standard Via All-Pairwise Comparisons. *Journal of Discrete Event Dynamic Systems* 16(3):385-403.

Chen, E. J. 2008. Restricted Subset Selection. *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Moench, and O. Rose, 281-289. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Chen, E. J. 2009. Subset Selection Procedures *Journal of Simulation*. To Appear.

Chen, E. J. and W. D. Kelton. 2005. Sequential Selection Procedures: Using Sample Means to Improve Efficiency. *European Journal of Operational Research* 166(1): 133-153.

Chen, E. J. and W. D. Kelton. 2007. A Procedure for Generating Batch-Means Confidence Interval for Simulation: Checking Independence and Normality. *Simulation: Transactions of The Society for Modeling and Simulation International* 83: 683-694.

Fishman, G. S. 2001. *Discrete-Event Simulation: Modeling Programming and Analysis*. New York: Springer-Verlag.

Goldsman, D., W. S. Marshall, S.-H. Kim, and B. L. Nelson. 2000. Ranking and selection for steady-state simulation. *Proceedings of the 2000 Winter Simulation Conference*, eds. Joines JA, Barton RR, Kang K, and Fishwick PA, 544–553. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Healey, C., D. Goldsman, S.-H. Kim. 2007. Ranking and Selection Techniques with Overlapping Variance Estimators. *Proceedings of the 2007 Winter Simulation Conference*, ed. S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 522-529. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

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

Lewis, P. A. W., E. McKenzie, D. K. Hugus. 1989. Gamma Processes. *Comm. Statist. Stoch. Models* 5: 1-30.

Meketon, M. S. and B. Schmeiser. 1984. Overlapping batch means: Something for nothing? *Proceedings of the 1984 Winter Simulation Conference*, eds. Sheppard S, Pooch U, and Pedgen D, 227-230. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Nakayama, M. K. 1995. Selecting the best system in steady-state simulations using batch means. *Proceedings of the 1995 Winter Simulation Conference*, eds. Alexopoulos C, Kang K, Lilegdon WR, and Goldsman D, 362-366. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Rice, J. A. 1995. *Mathematical Statistics and Data Analysis,* 2nd ed., Duxbury Press, Belmont, California.

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2005. ASAP3: A batch means procedure for steady-state simulation output analysis. *ACM Transactions on Modeling and Computer Simulation* 15: 39–73.

Sullivan, D. W., J. R., Wilson. 1989. Restriced subset selection procedures for simulation. *Operations Research* 37:52-71.

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference and the variance. *Annals of Mathematical Statistics* 12: 367–395.

AUTHOR BIOGRAPHY

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