Division of Research
Graduate School of Business Administration
The University of Michigan

April 1982
Revised May 1982

TWO ARMA-BASED CONFIDENCE-INTERVAL PROCEDURES FOR THE ANALYSIS OF SIMULATION OUTPUT

Working Paper No. 304

Richard W. Andrews

Thomas J. Schriber

The University of Michigan

FOR DISCUSSION PURPOSES ONLY

ABSTRACT

Two methods are presented for building interval estimates on the mean of a stationary stochastic process. Both methods fit an autoregressive moving-average (ARMA) model to observations on the process. The model is used to estimate the variance of the sample mean and the applicable degrees of freedom of the t statistic. Fitting of the ARMA model is totally automated. The ARMA-based confidence intervals perform well with data generated from ARMA processes. With data generated from queuing-system simulations, the coverage of the confidence intervals is less than satisfactory. It is shown that with queuing-system data, the sample mean and its estimated standard deviation are strongly positively correlated, and that the residuals of the fitted models are not normally distributed. These factors contribute adversely to the coverage of the confidence-interval procedures with queuing data.

We introduce and test two confidence interval procedures (CIPs) for the mean of a univariate output random variable from a simulation model operating at steady state. The CIPs are based on an autoregressive moving-average (ARMA) model and are fixed-sample-size procedures. The threefold purpose of this research has been:

1. to develop two versions of an ARMA-based confidence-interval procedure;
2. to measure the effectiveness of both versions by subjecting them to comprehensive testing; and
3. to develop and report guidelines for using this procedure with output data from simulation models.

As used in this report, a confidence-interval procedure consists of four steps:

a. computation of the sample mean;
b. estimation of the variance of the sample mean;
c. determination of the number of degrees of freedom; and
d. computation of an interval estimate for the process mean, using the t distribution with the aforementioned quantities.

These four steps correspond to the first four steps given in Fishman [1978, p. 236] for forming interval estimates. After reviewing the pertinent fundamentals of ARMA processes in Section 1, the steps making up the ARMA-based confidence-interval procedure are explained in detail in Section 2.

The proposed CIPs have been subjected to comprehensive empirical testing, using the research framework suggested for this purpose in Schriber and Andrews [1981]. Empirical testing of a CIP involves the generation of data

from a series of theoretical output processes (TOPs) with known means. In Section 3, we discuss the eight TOPs used to evaluate the CIPs proposed here, and indicate the reasons these TOPs were chosen for testing purposes.

After a brief discussion of the testing environment in Section 4, the empirical results of the testing are given in Section 5. For each TOP used, the resulting measures of effectiveness (MOEs) are presented in a corresponding table of the form introduced in Schriber and Andrews [1981]. The tables display the performance characteristics of the CIP when used to process data generated by the associated TOPs.

Both CIPs perform well with data generated by ARMA TOPs, which in this research are tailor-made (Schriber and Andrews [1981]). However, when used to process observations produced by models of two queuing systems, the ARMA-based CIPs did not perform satisfactorily for all the measures of effectiveness. Nevertheless, they did as well as or better than the pure autoregressive (AR) confidence-interval procedure presented by Fishman [1971] and further investigated by Andrews and Schriber [1978].

In Section 5, we also investigate possible reasons for the failure of these CIPs to perform in better fashion on data generated by the queuing-system models. This investigation takes the form of empirically determining the extent to which the underlying assumptions were satisfied by the queuing system data. This discussion concludes with the recommendation that the ARMA-based CIP be used with queuing data only after the data have been tested appropriately. In particular, a test should be performed to see if there is a significant correlation between the sample mean and the estimated standard deviation of the sample mean. Furthermore, the distribution of the residuals should be tested for normality.

1. AUTOREGRESSIVE MOVING-AVERAGE MODEL

The autoregressive moving-average model on which the confidence-interval procedure is based is given by (1):

$$X_t = \phi_1 X_{t-1} + \cdots\cdots + \phi_p X_{t-p} + \theta_0 + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots\cdots - \theta_q \varepsilon_{t-q}$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$$E[\varepsilon_i, \varepsilon_j] = \begin{cases} \sigma_\varepsilon^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad (1)$$

$$Cov(\varepsilon_t, X_s) = 0 \text{ if } t > s.$$

This is the familiar model as given in Box and Jenkins [1976]. It is referred to as ARMA(p,q). When q = 0, the model is the pure autoregressive (AR) process which Fishman [1971] used as the basis for a confidence-interval methodology. Here, we allow for the presence of moving-average (MA) terms, thereby extending the pure AR model to a mixed AR-MA form with the objective of achieving an improved confidence-interval methodology for those cases in which MA terms are of importance.

Further motivation for developing an ARMA-based CIP is provided by Steudel and Wu [1977, p. 748], who state that "...any uniformly sampled wide-sense stationary stochastic process can be adequately described by a discrete autoregressive moving-average (ARMA) model of order n and n-1." On the basis of their limited empirical results, Steudel and Wu tentatively conclude, for example, that the "current system content" output variable for an M/M/1 queuing system is adequately modeled by an ARMA(1,0) model. In a companion paper, Steudel et al. [1978, p. 292] conclude that "Queue behavior is shown to be adequately described by a first order autoregressive AR(1) model if the

job selection discipline does not depend on operation processing time. In those cases where processing time was used in job selection, a second order autoregressive AR(2) model is adequate to characterize the queues." And Schmeiser and Kang [1981] have shown analytically that when batch means for any batch size are formed for an AR(1) process, the resulting process is ARMA(1,1).

The confidence-interval procedures we propose are for output processes which have reached steady state or, equivalently, for output processes which are stationary. Because the random disturbances, ε, in (1) are assumed to be normally distributed, we can equivalently consider the output process to be second-order stationary. In theory, discrete-time-parameter stationary processes can be adequately described by an ARMA(p,q) model if the p and q values are allowed to be appropriate nonnegative integers (Cox and Miller [1965, p. 288]). As emphasized by the concept of parsimony in the time series literature (Box and Jenkins [1976, p. 17]), small values for p and q can adequately fit a given data set in most situations.

The ARMA model in (1) has the following properties. The mean of the process is given by

$$\mu_X = \theta_0(1 - \sum_{i=1}^{p} \phi_i)^{-1} .$$

The variance of the process is given by

$$\sigma_X^2 = \sigma_\varepsilon^2 \, R(\tilde{\phi},\tilde{\theta}).$$

The specific form of the function R depends on the order (p,q). For example, the Yule-Walker equations (Box and Jenkins [1976, p. 75]) can be used to show that for an ARMA(1,1) model,

$$\sigma_X^2 = \sigma_\varepsilon^2 \, (1 + \theta_1^2 - 2\phi_1\theta_1)/(1 - \phi_1^2).$$

The spectral density function for the general ARMA(p,q) model (Fuller [1976, p. 146]) is given by (2):

$$f(\omega) = \sigma_\varepsilon^2(2\pi)^{-1}\left(1 - \sum_{j=1}^{q} \theta_j e^{-i\omega j}\right)^2\left(1 - \sum_{j=1}^{p} \phi_j e^{-i\omega j}\right)^{-2} \qquad (2)$$

$$-\pi \leq \omega \leq \pi.$$

As will be seen in Section 2, the spectral density plays an integral role in estimating the variance of the sample mean.

In most ARMA modeling applications, the objective is to find an adequate representation of the data under investigation. The procedures suggested in Box and Jenkins [1976] for finding an adequate ARMA model involve the well-known steps of identification, estimation, and diagnostic checking. The order of the resulting model and the estimated parameters are of central importance. The fitted model is then used as a surrogate for the actual process, and provides a basis for forecasting.

This contrasts with our situation, in which fitting an ARMA model is a means to the end of forming an interval estimate on the mean of a stationary simulation output process. Of course, the important steps of identification, estimation, and diagnostic checking must be carried out, but the resulting model order and parameter values are not the end result; the success of our overall procedure will be judged principally by characteristics of the confidence intervals which are ultimately produced.

2. CONFIDENCE-INTERVAL METHODOLOGY

The flowchart in Figure 1 displays the six key steps involved in applying the ARMA-based procedures for building a confidence interval. The overall objective of the first five steps is to estimate the variance of the sample mean. These five steps, taken together, correspond to Step b in the

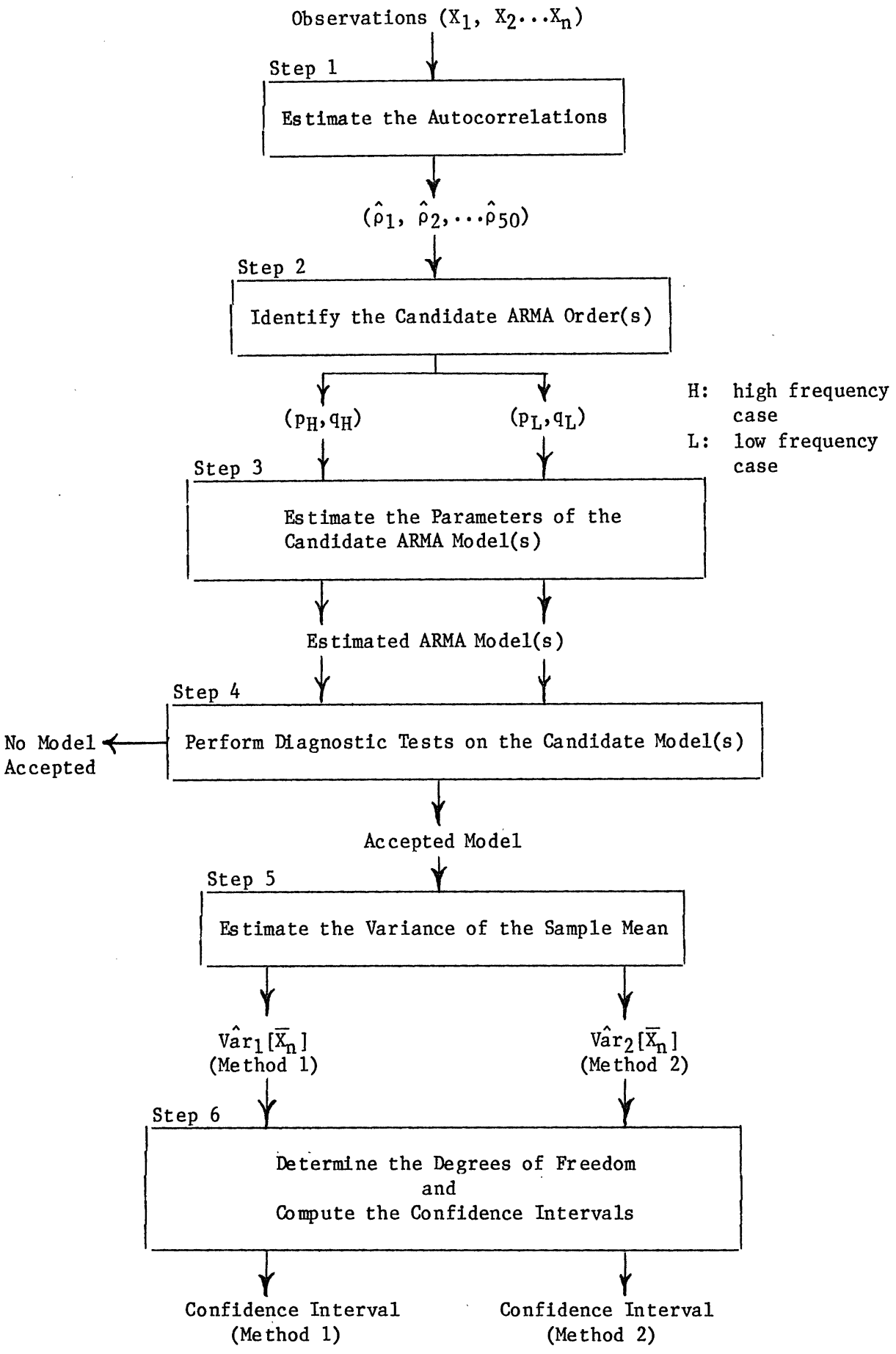

Figure 1: Overview of the Steps Involved in the ARMA-Based CIPs

introduction. The sixth step corresponds to Steps c and d in the introduction. Detailed commentary on these steps follows.

Step 1. Compute the Sample Autocorrelations

For a sequence of n observations $(X_1, X_2, \ldots, X_n)$, the first 50 sample autocorrelations are calculated. The sample autocorrelation of lag s is given by

$$\hat{\rho}_s = \frac{\sum_{i=1}^{n-j} (X_i - \bar{X})(X_{i+s} - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \; ; \; s = 1, 2, \ldots, 50.$$

Step 2. Identify the Candidate ARMA Order(s)

The methodology of Box and Jenkins [1976] for identifying the order of an ARMA process entails a visual inspection of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) estimated from the data. This inspection involves a subjective, time-consuming procedure which it would be desirable to automate. Recently, several algorithms for automating the identification step have been proposed (Gray et al. [1978]; Beguin et al. [1980]; Tiao and Tsay [1981]). We use the algorithm proposed by Gray et al. We apply the algorithm by using the 50 sample autocorrelations from Step 1 above to compute what Gray et al. term a D statistic. We compute this D statistic for each of 12 ARMA orders corresponding to all combinations of p and q for p = 1, 2, and 3 and q = 0, 1, 2, and 3. Two sets, each consisting of 12 D statistics, are computed: one for what is called the high-frequency case, and the other for what is called the low-frequency case. The (p,q) combination resulting in the largest D statistic for the high-frequency case is then a candidate model, as is the (p,q) combination corresponding to the largest D statistic for the low-frequency case.

Step 3. Estimate the Parameters of the Candidate ARMA Model(s)

In this step, the p autoregressive and the q moving-average coefficients along with the variance of the disturbance term, σ_ε^2, are estimated for the two candidate ARMA models. If both candidate models have identical orders, there is really only one candidate model, and the estimation process need be performed only once. The estimation procedure uses subroutines from the International Mathematical & Statistical Library (IMSL). The key subroutine, FTMXL, does the estimation by using the conditional likelihood method described in Box and Jenkins [1976, pp. 209-10]. (See the IMSL Library Reference Manual [1980] for documentation.)

Step 4. Perform Diagnostic Tests on the Candidate Model(s)

In this step, test statistics for the candidate model(s) are computed and evaluated to determine whether to accept a model as adequately fitting the data. Included among these statistics are the t statistic for each of the p autoregressive and q moving-average coefficients in the model(s). A model is judged unacceptable unless the t value for each AR and MA term is at least 1.96. This cutoff value of 1.96 was chosen to correspond to an $\alpha = .05$ test in the case of a large number of degrees of freedom.

In addition to the coefficient t statistics, the Ljung-Box [1978] portmanteau statistic, Q, was calculated for the candidate model(s). The value of Q is given by

$$Q = n(n+2) \sum_{k=1}^{10} (n-k)^{-1} \hat{r}_k^2 \quad ,$$

where $\hat{r}_k$ is the estimated autocorrelation of lag k of the residuals of the estimated model. We use 10 autocorrelations in the calculation of this statistic. Q therefore has an approximate χ^2 distribution with 10 - p - q

degrees of freedom. For a model to be acceptable, the achieved significance level of the Q statistic must be greater than .05.

In the case of two candidate ARMA models, it is possible that both models will pass the tests on the coefficient t statistics and on the Q statistic. If this occurs, we choose that model which has the largest minimum achieved significance level on the coefficients. The purpose of this procedure is to discard that model which has the least significant parameter estimate.

It is possible to come up empty in Step 4 if the model or models identified in Step 2 and estimated in Step 3 fail to pass the tests on the coefficient- and Q-statistics. In this case, and as shown in Figure 1, we proceed no further with the building of a confidence interval. We believe that for many simulation studies, the discarding of a sequence of data is not serious. In some data-collection environments, however, such an outcome would not be acceptable--for example, if the cost of generating a data set is unduly high.

Step 5. Estimate the Variance of the Sample Mean

Once a model has been accepted, it is used in this step to estimate the variance of the sample mean. Two alternative estimation methods are used for this purpose, both of which make use of an estimate of the spectral density function (2) as an intermediate step. The estimate of (2) is given by

$$\hat{f}(\omega) = \hat{\sigma}_\varepsilon^2 (2\pi)^{-1}\left(1 - \sum_{j=1}^{q} \hat{\theta}_j e^{-i\omega j}\right)^2 \left(1 - \sum_{j=1}^{p} \hat{\phi}_j e^{-i\omega j}\right)^{-2}, \qquad (3)$$

where $\hat{\phi}_j$, $\hat{\theta}_j$, and $\hat{\sigma}_\varepsilon^2$ are the respective estimates of the autoregressive and moving-average coefficients, and of the variance of the disturbances. We proceed to describe the alternative variance-estimation methods.

Method 1

For any stationary process it can be shown that

$$\mathrm{Var}[\bar{X}_n] = c_n \sigma_x^2/n \qquad \text{where} \tag{4}$$

$$c_n = 1 + 2 \sum_{i=1}^{n-1} (1 - i/n)\, \rho_i .$$

(See Schmeiser [1982].) Hence, if we can estimate c_n and σ_x^2, we can estimate $\mathrm{Var}[\bar{X}_n]$. Using (3), we construct the autocovariance function (5):

$$\hat{\gamma}(s) = 2\int_0^{\pi} \hat{f}(\omega) \cos s\omega \, d\omega; \qquad s = 0, 1, \ldots, q. \tag{5}$$

The value of the integral in (5) is calculated using Simpson's rule, with the integrand evaluated at 40 uniformly spaced intervals ranging from 0 to π. For $i \geq q + 1$, $\hat{\gamma}(i)$ is then computed from the recursive relationship:

$$\hat{\gamma}(i) = \sum_{j=1}^{p} \phi_j \, \hat{\gamma}(i-j). \tag{6}$$

(See Box and Jenkins [1976, p. 75].) Using (4), we then have

$$\hat{\mathrm{Var}}_1[\bar{X}_n] = \hat{c}_n \, \hat{\gamma}(0)/n, \qquad \text{where}$$

$$\hat{c}_n = 1 + 2 \sum_{i=1}^{n-1} (1 - i/n)\, \hat{\gamma}(i)/\hat{\gamma}(0).$$

Method 2

If the spectral density function of a stationary time series X_t is continuous, then

$$\lim_{n\to\infty} n \, \mathrm{Var}[\bar{X}_n] = 2\pi f(0),$$

where $f(0)$ is the spectral density of X_t evaluated at zero (see Fuller [1976], p. 232). As a second estimate of the variance of the sample mean, we therefore use

$$\hat{Var}_2(\bar{X}_n) = 2\pi\hat{f}(0)/n, \tag{7}$$

where from (3), $\hat{f}(0)$ is

$$\hat{f}(0) = \hat{\sigma}_\varepsilon^2 (2\pi)^{-1}(1 - \sum_{j=1}^{q} \hat{\theta}_j)^2 (1 - \sum_{j=1}^{p} \hat{\phi}_j)^{-2}.$$

(See Pritsker and Pegden [1979, p. 481].)

Step 6. Determine the Degrees of Freedom and Compute the Confidence Intervals

The $(1 - \alpha)100\%$ confidence interval for μ is given by

$$\bar{X}_n \pm t_{\alpha/2,k} \hat{SD}_i(\bar{X}_n),$$

with $t_{\alpha/2,k}$ denoting the $1 - (\alpha/2)$ percentile of the t distribution with k degrees of freedom, and

$$\hat{SD}_i[\bar{X}_n] = SQRT[\hat{Var}_i(\bar{X}_n)]; \qquad i = 1, 2 \text{ (the two methods)}.$$

We estimate the k degrees of freedom in a manner similar to that suggested by Fishman [1971]. If we have a sample of m independent observations from a distribution with mean μ and variance σ_x^2 identical to the mean and variance of the stationary process of interest, then

$$Var[\bar{X}_m] = \sigma_x^2/m. \tag{8}$$

Equating the right-hand sides from (8) and (4), we have

$$n = mc_n.$$

This can be interpreted to mean that in a degrees-of-freedom sense, each independent observation is equivalent to c_n correlated observations. (See Schmeiser [1982] for a further discussion.) We therefore specify the degrees of freedom to be

$$(n/\hat{c}_n) - p - q - 1, \tag{9}$$

where, in addition to adjusting the equivalent sample size by dividing n by $\hat{c}_n$, we also lose a degree of freedom for each estimated autoregressive and moving-average coefficient and for the estimated mean. If (9) is less than 1, we set the degrees of freedom equal to 1. Or, if $\hat{c}_n < 1$, meaning that $n/\hat{c}_n > n$, we set the degrees of freedom to $n - p - q - 1$.

3. TESTING PROCEDURE

Eight theoretical output processes were chosen to comprehensively test the ARMA-based confidence-interval procedures, using fixed retained sample sizes of n = 100, 200, 300, and 400. For each TOP/sample-size combination, enough replications were generated to build 100 confidence intervals. Each replication was produced under stationary conditions; in addition, the first 50 observations were deleted from each replication. Figure 2 shows the replication design for one TOP with the retained sample size set at 100. In Figure 2, X_i^j denotes the i-th observed value

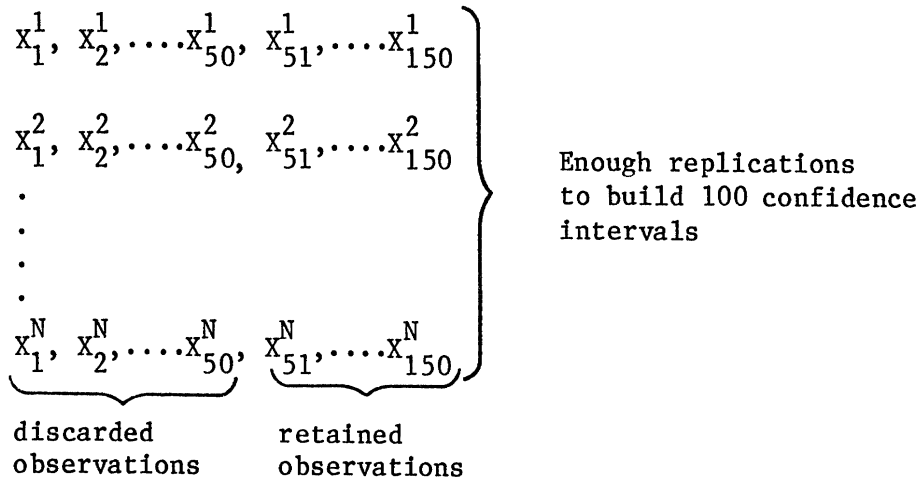

Figure 2: Replication Design

in the j-th replication. In general, more than 100 replications are needed to build 100 confidence intervals, because not every replication results in a

statistically acceptable ARMA model. The number of replications needed to obtain 100 confidence intervals, denoted by N in Figure 2, is reported in the measure-of-effectiveness tables in Section 4.

Six of the eight TOPs used in the testing were tailor-made; that is, they were autoregressive moving-average processes. This set of ARMA processes was carefully chosen to include two pure AR processes and one ARMA process, whose use for testing purposes has previously been reported in the literature, and to include ARMA processes providing a representative range of behavior in terms of their autocorrelation functions and their limiting value of c_n. The c_n value provides an important measure of correlation structure, and, as discussed above, indicates the number of dependent observations equivalent to one independent observation.

We proceed to discuss each of the eight TOPs individually, providing the rationale for their choice and describing their relevant properties.

TOP 1

The equation for this pure autoregressive TOP is

$$X_t = .5X_{t-1} + .5 + \varepsilon_t \qquad \varepsilon_t \sim N(0,1). \tag{10}$$

This is the first of two autoregressive TOPs used by Fishman [1971] to evaluate his proposed AR-based confidence-interval procedure. We include it here to support comparisons between the AR- and ARMA-based CIPs.

For (10), $E[X_t] = 1$, $SD[X_t] = \sqrt{4/3}$, and $SD[\overline{X}_n] \approx 2/\sqrt{n}$. Furthermore, for any stationary AR(1) process it can be shown that

$$\lim_{n\to\infty} c_n = (1 - \phi_1^2)/(1 - \phi_1)^2, \tag{11}$$

where ϕ_1 is the autoregressive coefficient. Substituting the $\phi_1 = 0.5$ from (10) into (11), $c_n \to 3$, which can be thought of as a global measure of the degree of dependence inherent in this model when estimating Var $(\bar{X}_n)$.

The ACF for an AR(1) process is given by

$$\rho_i = 0.5^i \qquad i \geq 1.$$

The ACF for the process in (10) is consequently always positive, and decays exponentially. In our experience, such ACF behavior is representative of outputs from queuing system simulations.

TOP 2

This TOP, also purely autoregressive and specified as

$$X_t = .5\, X_{t-1} - .25\, X_{t-2} + .5 + \varepsilon_t \qquad \varepsilon_t \sim N(0,1), \tag{12}$$

is the second of the two AR TOPs used by Fishman [1971]. For (12), $E[X_t] = 2/3$, $SD[X_t] = 1.13$, and $SD[\bar{X}_n] \approx 1.31/\sqrt{n}$. For a stationary AR(2) process, we have

$$\lim_{n\to\infty} c_n = (1 + \phi_1 + \phi_1\phi_2 - \phi_2^2)/(1 - \phi_2)(1 - \phi_1 - \phi_2), \tag{13}$$

where ϕ_1 and ϕ_2 are the autoregressive coefficients. Substituting the appropriate values from (12) into (13), $c_n \to 1.4$ for this process.

The ACF for an AR(2) process is given by

$$\rho_1 = 0.4$$

$$\rho_i = 0.5\, \rho_{i-1} - 0.25\, \rho_{i-2} \qquad i = 2,\ 3,\ 4, \ldots.$$

This ACF is of a damped sinusoidal form. Although in our experience this form is not typical of outputs from queuing system simulations, we include this process to support comparison to the greatest extent possible with Fishman's [1971] work.

TOP 3

This AR(1) TOP is given by

$$X_t = .8\, X_{t-1} + 200 + \varepsilon_t \qquad \varepsilon_t \sim N(0,3600). \tag{14}$$

This TOP was selected for test purposes because Steudel and Wu [1977] indicate that the behavior of an M/M/1 queuing system having a high server utilization can be modeled by an AR(1) process for which ϕ_1 approaches 1. For the process in (14), $E[X_t] = 1{,}000$, $SD[X_t] = 100$, and $SD[\overline{X}_n] \approx 300/\sqrt{n}$. Substituting 0.8 from (14) into (11) indicates that $c_n \to 9$ for this process. And, as for TOP 1, the ACF decays exponentially.

TOP 4

This ARMA TOP, which is the first of three mixed ARMA models used for testing purposes here, is given by

$$X_t = .7X_{t-1} + 300 + \varepsilon_t + .4\varepsilon_{t-1} \qquad \varepsilon_t \sim N(0,2965.1). \tag{15}$$

For this process, $E[X_t] = 1{,}000$, $SD[X_t] = 100$, and $SD[\overline{X}_n] \approx 254/\sqrt{n}$. For any stationary ARMA(1,1) process,

$$\lim_{n\to\infty} c_n = 1 + 2\,\frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{(1 - \phi_1)(1 + \theta_1^2 - 2\phi_1\theta_1)}. \tag{16}$$

Using the coefficients from (15) in (16), $c_n \to 6.46$ for this process.

The ACF for this ARMA(1,1) process has

$\rho_1 = 0.8186$, and

$\rho_i = 0.7\,\rho_{i-1}$, for $i = 2, 3, 4, \ldots$.

This ACF is always positive, and decays exponentially. As mentioned above, our experience indicates that output from queuing system simulations often exhibits such behavior.

TOP 5

This ARMA(2,1) model, given by

$$X_t = 1.32X_{t-1} - .68X_{t-2} + 360 + \varepsilon_t - .8\varepsilon_{t-1}$$

$$\varepsilon_t \sim N(0,\ 5373.1), \tag{17}$$

was used by Gray et al. [1978] to test their algorithm for automatic identification of an ARMA process. We use it here to support comparison with their results in terms of the ability of their algorithm to correctly identify an ARMA process.

For the process in (17), $E[X_t] = 1{,}000$, $SD[X_t] = 100$, and $SD[\bar{X}_n] \approx 40/\sqrt{n}$.

For a stationary ARMA(2,1) process we have

$$\lim_{n\to\infty} c_n = \frac{\phi_2^2\theta_1^2 - 2\phi_2^2\theta_1 + \phi_2^2 - \phi_1\theta_1^2 + 2\phi_1\phi_2\theta_1 - \phi_1\phi_2\theta_1^2 + 2\phi_1\theta_1 - \phi_1\phi_2 - \theta_1^2 + 2\theta_1 - \phi_1 - 1}{(1-\phi_1-\phi_2)(2\phi_1\theta_1 + \phi_2 + \phi_2\theta_1^2 - \theta_1^2 - 1)}. \tag{18}$$

Using values from (17) in (18), $c_n \to 0.16$ for this process. This $c_n < 1$ indicates that the variance of the sample mean for this correlated process will be smaller than the variance of the sample mean for an independent process with the same underlying variance. This might contribute to the notable ability of Gray et al.'s algorithm to correctly identify data generated by (17) as coming from an ARMA(2,1) process (as reported by Gray et al., and as substantiated here in Section 5).

The ACF for this ARMA(2,1) process has

$\rho_1 = 0.5299$, and

$\rho_i = 1.32\ \rho_{i-1} - 0.68\ \rho_{i-2}$ for $i = 2,\ 3,\ 4, \ldots$.

The ACF consequently shows damped sinusoidal behavior.

TOP 6

This ARMA(2,1) model is given by

$$X_t = .9X_{t-1} - .18X_{t-2} + 280 + \varepsilon_t + .9\varepsilon_{t-1}$$

$$\varepsilon_t \sim N(0,\ 1271.5). \qquad (19)$$

This process has $E[X_t] = 1,000$, $SD[X_t] = 100$, and $SD[\bar{X}_n] \approx 242/\sqrt{n}$. Using values from (19) in (18), $c_n \to 5.85$ for this process.

The ACF is given by

$\rho_1 = 0.8597$, and

$\rho_i = 0.9\ \rho_{i-1} - 0.18\ \rho_{i-2}$ for $i = 2,\ 3,\ 4, \ldots$.

This ACF is always positive and decreases exponentially.

With its $c_n > 1$ and its ACF properties, TOP 6 was designed to contrast with TOP 5. We believe that in these measures TOP 6 corresponds more closely to typical queuing system simulation output than does TOP 5.

TOP 7

TOPs 7 and 8 are based on output processes associated with queuing system simulations. These two TOPs were chosen to investigate the potential applicability of the ARMA-based CIP to the queuing system class of non-ARMA TOPs.

TOP 7 is based on a materials-handling problem described in Hillier and Lieberman [1974, p. 465], in which the output random variable is cost per unit time. Here is a paraphrased statement of the problem:

> "A certain materials-handling unit is used to transport goods between producing centers in a job shop. Calls for the materials-handling unit to move a load come essentially at random (i.e., according to a Poisson input process)

> at a mean rate of two per hour. The total time required to move a load has an exponential distribution with an expected time of d minutes. The total equivalent uniform hourly cost (capital recovery cost, plus operating cost) for the materials-handling unit is FC. The estimated cost of idle goods (waiting to be moved, or in transit) because of increased in-process inventory is $10 per load per hour. Furthermore, the scheduling of the work at the producing center allows for just one hour from the completion of a load at one center to the arrival of that load at the next center. Therefore, an additional $20 per load per hour of delay (including transit time) after the first hour is to be charged for lost production. What is the expected cost of this system (defined as the sum of delay cost and equipment cost) as a function of d and FC?"

We simulate the behavior of this system, setting specific values for d and FC and taking periodic observations on the cost accumulated by the system over time. Letting TC be the cost per hour, it can be shown that for $d < 30$,

$$E[TC] = \{2d/(60 - 2d)\}\{20 \exp((2d - 60)/d) + 10\} + FC.$$

Hence, the performance of the ARMA-based confidence-interval procedure can be measured in terms of the known mean of the cost variable.

One key decision in using this system as a test case involves setting the interobservation time. We set this value at eight simulated hours (one shift), which is 16 times the expected job interarrival time. By way of comparison, Fishman [1971] chose an interobservation time 4 times the interarrival time to observe current queue content in an M/M/1 system used to test his AR-based confidence-interval procedure. In contrast, Steudel and Wu [1977] recommend that an interobservation time 10 times greater than the service time be used in observing current queue content in job-shop simulations. In general, no comprehensive guidelines have been reported for choosing the interobservation time in experiments of this type.

Another key decision in this system involves selecting a method to follow in accumulating the system cost. In one method, the cost could be based only on those jobs which have left the system during the current observation period. In another method, the cost could be based on all jobs which have been, and perhaps still are, in the system at any time during the course of the current observation period. The expected total cost per unit time will be the same for either method, but the variability of the cost will not. We use the second of these two cost-accumulation methods because the variability associated with it is smaller.

In setting parameter values for this TOP, we chose d = 24 minutes and FC = \$10. This results in a server utilization of 0.8 and leads to an expected daily system cost of \$788.18.

TOP 8

For TOP 8 we worked with an M/D/3 queuing system, choosing current system content as the output random variable of interest. Analytic solutions for this system have been evaluated numerically by Hillier and Yu [1981], making it easy to assess performance of the ARMA-based confidence-interval procedure in terms of the known system properties. By way of contrast with TOP 7, the output random variable chosen here takes on a noncumulative value; that is, the value observed is a point value, not a value accumulated during the course of the observation period.

In the M/D/3 system, interarrival time was set to 5 minutes, service time to 13.5 minutes, and interobservation time to 20 minutes. The interarrival and service-time settings result in an 0.9 server utilization, and give an expected system content of 6.42.

4. TESTING ENVIRONMENT

The software used in this research consisted of custom-built modules combined with proven existing routines. The existing routines included certain IMSL [1980] subroutines and the Michigan Interactive Data Analysis System (Fox and Guire [1976]). With the one exception noted below, the custom-built modules were written in FORTRAN, were checked out under an interactive FORTRAN interpreter, and then were translated under an optimizing FORTRAN compiler prior to their use in making the production runs.

The two queuing-system TOPs were built in GPSS. The GPSS models were run under GPSS/H (Henriksen and Crain [1982]), a state-of-the-art GPSS implementation. GPSS/H uses the Tausworthe [1965] algorithm to generate uniform 0-1 random numbers. The exponentially distributed interarrival and service times in the GPSS models were sampled using the standard natural logarithm function from the FORTRAN library. This is superior to the more conventional GPSS approach of using a piecewise linear approximation to the inverse cdf of the exponential distribution (see Schriber [1974, p. 163]).

The computing work was accomplished on an Amdahl 470/V8 operating under the Michigan Terminal System at The University of Michigan.

5. TEST RESULTS

Results of using the two versions of the ARMA-based confidence-interval procedure with the eight theoretical output processes are presented here in a set of eight identically formatted tables, following the suggestion of Schriber and Andrews [1981]. The four table rows correspond to 100 replications consisting of 100, 200, 300, and 400 observations, respectively. Each of the five table columns corresponds to a particular measure of effectiveness of the confidence-interval procedure. These five MOEs will be described before the tables themselves are presented and discussed.

For the six ARMA TOPs, table column 1 (MOE 1) reports the percentage of accepted ARMA models whose (p,q) order matched that of the known underlying ARMA process. Recall that a candidate ARMA model was accepted only if various test statistics had satisfactory values. Also note that because each MOE 1 percentage is based on 100 replications, the percent can alternatively be thought of as an actual count. This MOE measures the ability of the Gray et al. algorithm to correctly identify the order of the ARMA process used to generate the time series being analyzed. MOE 1 does not have an interpretation for queuing-system TOPs 7 and 8, and so is marked NA (not applicable) in Tables VII and VIII. The orders of the ARMA models accepted for the queuing-system TOPs are of interest, however, and will be presented separately when test results for those TOPs are discussed.

Table column 2 (MOE 2) provides a measure of the coverage properties of the confidence intervals built for the accepted ARMA models. In particular, the number reported in column 2 is the achieved significance level of a χ^2 test for uniformity in the distribution of the random variable

$$\eta^* = \inf\{\eta:\theta \varepsilon C(X_1, X_2, \ldots, X_n; \eta)\},$$

where θ = the process parameter of interest,

η = a confidence level, and

$C(X_1, X_2 \ldots X_n;\eta)$ = a confidence interval based on the sequence $X_1, X_2, \ldots, X_n$ at confidence level η.

The random variable η^* is the confidence level that just succeeds in covering the parameter of interest, which in our case is the process mean. The distribution of η^* is referred to as the coverage function. For a theoretically perfect confidence-interval procedure, η^* follows a uniform (0,1) distribution. (See Schruben [1980] for details.)

We conducted the χ^2 goodness-of-fit test by dividing the (0,1) interval into 10 cells of equal width and then computing the corresponding test statistic. Low values of the achieved significance level of this statistic would indicate that the observed η^*'s do not conform to the theoretically correct uniform (0,1) distribution. In particular, any value lower than, say, 0.05 suggests that the CIP is suspect in terms of its ability to produce meaningful confidence intervals for the TOP at hand.

Table column 3 (MOE 3) provides a measure of the degree of variability in the halfwidths of the confidence intervals. In particular, it reports the estimated coefficient of variation ($\hat{CV}$) of the standard error of the mean. This estimate is computed as follows:

$$\hat{CV} = \hat{SD}\,[\hat{SD}(\bar{X}_n)]/\overline{\hat{SD}(\bar{X}_n)},$$

where

$$\overline{\hat{SD}(\bar{X}_n)} = 100^{-1} \sum_{j=1}^{100} \hat{SD}_j(\bar{X}_n);$$

$$\hat{SD}[\hat{SD}(\bar{X}_n)] = 99^{-1} \sum_{j=1}^{100} (\hat{SD}_j(\bar{X}_n) - \overline{\hat{SD}(\bar{X}_n)})^2;$$

and $\hat{SD}_j(\bar{X}_n)$ is the standard error on the jth replication. For iid observations taken from a normal distribution, Schmeiser [1982] derived CV analytically, one form of which is

$$CV = \{[\Gamma(\tfrac{n+1}{2})]^2 - [\Gamma(\tfrac{n}{2})]^2\}^{1/2}/\Gamma(\tfrac{n}{2}).$$

Note that CV in this case depends only on the sample size, n. For iid normal samples of size 100, 200, 300, and 400, the corresponding CV values would be 0.071, 0.050, 0.041, and 0.035. These numbers provide a benchmark for MOE 3.

Table column 4 (MOE 4) provides the two conventional measures for reporting the properties of a confidence-interval procedure. This column indicates the average relative halfwidth of the confidence intervals built at a 95% confidence level, and the percentage of these intervals which cover the process mean.

Table column 5 (MOE 5) indicates how many replications had to be generated to obtain 100 usable replications. A usable replication is one to which an ARMA model can be fitted acceptably in the statistical sense described in Section 3. This measure, which involves the ability of the ARMA-based CIP to produce a confidence interval for the TOP at hand, is reported as the ratio of replications generated to replications used.

In examining the MOE tables, the following points should be kept in mind:

1. How do the two versions of the ARMA-based CIP compare?
   (Only MOEs 2, 3, and 4 will depend on the version in question.)
2. How adequately do the two CIPs perform for the TOP used?
3. Do the properties of the CIP as measured by the MOEs improve as sample size increases?
4. If one or more table values are not what we would expect, can we identify the underlying cause and their implications?

Tables I and II give the results of analyzing the AR(1) and AR(2) models used by Fishman [1971]. In the AR(1) case, 83, 88, 86, and 98 percent of the accepted ARMA models were correctly identified to be of (1,0) order for replications consisting of 100, 200, 300, and 400 observations, respectively (column 1, Table I). This indicates good performance on the part of the identification algorithm for this AR(1) TOP. For the AR(2) case, the percentages of accepted ARMA models which matched the underlying (2,0) order were

also in the 80% range except for sample size 100 (column 1, Table II). To obtain 100 acceptable ARMA models, we needed at most 147 replications for the AR(1) TOP (column 5, Table I), but as many as 202 replications from the AR(2) TOP were needed to obtain 100 acceptable models (column 5, Table II).

The achieved significance levels of the coverage function (column 2) are consistently excellent in Table I, and are completely satisfatory in Table II for samples of size 200, 300, and 400. For samples of size 100 in Table II, however, the hypothesis regarding uniformity of η^* would be rejected at a 0.05 significance level. Note that this is also the sample size in Table II for which the relatively small percent of correct identification and relatively large replication ratio were experienced.

The entries in column 3 (MOE 3) in Tables I and II provide a measure of the variability in the width of the confidence intervals. The column 3 values are larger than the iid-normal benchmark values reported above, which might reflect the dependency in the data. Like the benchmark values, the column 3 values decrease as sample size increases. For TOPs 1 and 2, and for the other six TOPs as well, the variability of the confidence interval width is larger for Method 2 than for Method 1. In all cases, however, the differences are small.

Column 4(b) in Tables I and II indicates that the coverage rate of confidence intervals at a 95% confidence level was close to 0.95 in most cases. The entries in column 4(a) report the average relative halfwidth. This measure is useful when it is of interest to estimate the sample size required to achieve a specified relative halfwidth in a confidence interval. These values range from .19 in Table II for a sample size of 400, to .409 in Table I for a sample size of 100.

TABLE I

Measures of Effectiveness for Analysis of an ARMA (1,0) Model

Number of Observations per 100 Accepted Replications		1 % Correct Identification	2 Achieved Significance for Coverage Function	3 Coefficient of Variation of the Standard Error	4(a) Average Relative Half-Width	4(b) % Coverage (95% Confidence Level)	5 Replication Ratio
100	Method 1	83	0.554	0.228	0.408	93	1.47
	Method 2		0.437	0.247	0.409	93	
200	Method 1	88	0.384	0.207	0.284	92	1.21
	Method 2		0.401	0.219	0.284	90	
300	Method 1	86	0.699	0.155	0.223	91	1.21
	Method 2		0.534	0.174	0.223	90	
400	Method 1	98	0.964	0.097	0.199	95	1.12
	Method 2		0.964	0.101	0.199	95	

TABLE II

Measures of Effectiveness for Analysis of an ARMA (2,0) Model

Number of Observations per 100 Accepted Replications		1 % Correct Identification	2 Achieved Significance for Coverage Function	3 Coefficient of Variation of the Standard Error	4(a) Average Relative Half-Width	4(b) % Coverage (95% Confidence Level)	5 Replication Ratio
100	Method 1	64	0.024	0.352	0.385	88	2.02
	Method 2		0.005	0.403	0.376	85	
200	Method 1	85	0.225	0.167	0.273	89	1.52
	Method 2		0.225	0.179	0.272	89	
300	Method 1	86	0.154	0.129	0.227	93	1.40
	Method 2		0.154	0.128	0.227	93	
400	Method 1	89	0.122	0.127	0.190	86	1.46
	Method 2		0.122	0.131	0.190	86	

The only measure which can be compared with the results reported by Fishman [1971] is MOE 1. For these two models, Fishman reported correct identification in 77% to 86% of the cases. These percentages are not directly comparable with MOE 1 here because our percentages do not include those replications for which no statistically acceptable ARMA model could be fitted. Furthermore, Fishman used a variable sample size scheme, extending the size as necessary to achieve a stated relative halfwidth. His sample sizes ranged from about 250 to 350. Fishman did not report coverage measures for the AR(1) and AR(2) models, so coverage comparisons cannot be made.

Table III reports results for the AR(1) model used as TOP 3. The table contains superb values for all measures. In terms of identification, the procedure did better with this process than it did with the AR(1) process reported in Table I. TOP 3 had $\phi_1 = 0.8$, as compared with $\phi_1 = 0.5$ for TOP 1. The corresponding limiting values of c_n were 9 and 3. It would therefore seem that the Gray et al. algorithm correctly identifies the underlying order a higher percent of the time when the correlation structure as measured by the limiting value of c_n is stronger.

Table IV corresponds to TOP 4, which is a mixed ARMA(1,1) model. The Table IV results are excellent except for the small percent of correctly identified models and the low achieved significance level for η^* at sample size 100. It is worthwhile to attempt to explain why the coverage at a 95% confidence level is satisfactory when n = 100, whereas the small achieved significance level for MOE 2 at the same sample size indicates that the coverage was not satisfactory for all confidence levels. One possible cause is that there may be a significant correlation between the estimate of the mean, $\overline{X}_n$, and the estimated standard deviation of $\overline{X}_n$. If this correlation exists, then the numerator and denominator of a statistic assumed to follow

TABLE III

Measures of Effectiveness for Analysis of an ARMA (1,0) Model

Number of Observations per 100 Accepted Replications		1 % Correct Identification	2 Achieved Significance for Coverage Function	3 Coefficient of Variation of the Standard Error	4(a) Average Relative Half-Width	4(b) % Coverage (95% Confidence Level)	5 Replication Ratio
100	Method 1	98	0.437	0.273	0.058	87	1.08
	Method 2		0.514	0.281	0.059	87	
200	Method 1	97	0.419	0.246	0.042	93	1.15
	Method 2		0.494	0.254	0.042	93	
300	Method 1	98	0.384	0.173	0.033	91	1.08
	Method 2		0.419	0.176	0.033	89	
400	Method 1	100	0.494	0.164	0.029	98	1.13
	Method 2		0.534	0.166	0.030	98	

the t distribution violate the assumption that they are independent. This is one of the potential problems which Schruben [1980] indicates can lead to poor performance of the empirical coverage function. However, this is not a problem with this ARMA(1,1) process. For the 100 usable replications at sample size 100, the achieved correlation between $\overline{X}_n$ and the estimated standard deviation of $\overline{X}_n$ was 0.15 for both methods of estimating the variance of the sample mean. At a significance level of .05, the critical value of the correlation coefficient is .20, so the hypothesis that these statistics are uncorrelated would be accepted. The poor realization of the coverage function is traceable to the fact that 21 (20 in the case of Method 2) of the 100 confidence intervals had an achieved η^* between 0.60 and 0.70. This may be attributable to unexplained randomness.

TABLE IV

Measures of Effectiveness for Analysis of an ARMA (1,1) Model

Number of Observations per 100 Accepted Replications		1 % Correct Identification	2 Achieved Significance for Coverage Function	3 Coefficient of Variation of the Standard Error	4(a) Average Relative Half-Width	4(b) % Coverage (95% Confidence Level)	5 Replication Ratio
100	Method 1	44	0.007	0.286	0.052	93	1.26
	Method 2		0.029	0.296	0.053	93	
200	Method 1	76	0.720	0.205	0.036	93	1.41
	Method 2		0.851	0.208	0.036	93	
300	Method 1	84	0.154	0.170	0.028	94	1.18
	Method 2		0.154	0.174	0.029	95	
400	Method 1	88	0.699	0.128	0.025	97	1.19
	Method 2		0.779	0.129	0.025	97	

Tables V and VI correspond to TOPs 5 and 6, both of which are ARMA(2,1) models. Three points should be kept in mind when examining these tables:

1. In Table V, the alternative methods for estimating $\mathrm{Var}[\bar{X}_n]$ give discernibly different results for the first time;
2. The percent of correct identifications in Table VI is low for the first time; and
3. The values of the average relative halfwidth are extremely small, which merits comment.

In Table V, at sample sizes of 100 and 200, MOE 2 reports low values for Method 2 but acceptable values for Method 1. It should be noted that Method 1 uses equations (5) and (6) to smooth out the autocorrelation function. This results in better empirical autocorrelation properties for this particular theoretical output process.

The inferior results for Method 2 at sample sizes of 100 and 200 can be explained in terms of the values calculated for $\hat{\mathrm{Var}}_2[\bar{X}_n]$. From (7),

$$\hat{\mathrm{SD}}_2[\bar{X}_n] = \left[\frac{\hat{\sigma}_\varepsilon^2 \left[1 - \sum \hat{\theta}\right]^2}{n \left[1 - \sum \hat{\phi}\right]^2}\right]^{1/2} .$$

The resulting estimated standard deviation may be very large if $\sum \hat{\phi} \approx 1$ or it may be very small if $\sum \hat{\theta} \approx 1$. For the replications with sample sizes of 100 and 200, small values for $\hat{\mathrm{SD}}_2[\bar{X}_n]$ occurred quite often. The following list gives some of the cases for which $\hat{\mathrm{SD}}_2[\bar{X}_n]$ was very small compared to $\hat{\mathrm{SD}}_1[\bar{X}_n]$.

TABLE V

Measures of Effectiveness for Analysis of ARMA (2,1) Model No. 1

Number of Observations per 100 Accepted Replications		1 % Correct Identification	2 Achieved Significance for Coverage Function	3 Coefficient of Variation of the Standard Error	4(a) Average Relative Half-Width	4(b) % Coverage (95% Confidence Level)	5 Replication Ratio
100	Method 1	81	0.304	0.479	0.010	96	1.66
	Method 2		0.046	0.613	0.009	86	
200	Method 1	92	0.225	0.277	0.006	88	1.40
	Method 2		0.024	0.353	0.005	82	
300	Method 1	95	0.658	0.224	0.005	95	1.39
	Method 2		0.086	0.249	0.005	89	
400	Method 1	96	0.883	0.172	0.004	91	1.26
	Method 2		0.494	0.188	0.004	91	

TABLE VI

Measures of Effectiveness for Analysis of ARMA (2,1) Model No. 2

Number of Observations per 100 Accepted Replications		1 % Correct Identification	2 Achieved Significance for Coverage Function	3 Coefficient of Variation of the Standard Error	4(a) Average Relative Half-Width	4(b) % Coverage (95% Confidence Level)	5 Replication Ratio
100	Method 1	11	0.456	0.328	0.057	94	1.47
	Method 2		0.514	0.338	0.058	94	
200	Method 1	10	0.575	0.248	0.038	97	1.77
	Method 2		0.419	0.252	0.038	97	
300	Method 1	5	0.225	0.167	0.031	99	1.57
	Method 2		0.225	0.169	0.031	99	
400	Method 1	12	0.679	0.202	0.005	95	2.03
	Method 2		0.616	0.204	0.005	95	

		η^*	
$\hat{SD}_1[\bar{X}_n]$	$\hat{SD}_2[\bar{X}_n]$	Method 1	Method 2
2.23	0.91	1.0	1.0
2.58	1.31	0.92	1.0
2.83	0.72	0.27	0.83
2.03	0.72	0.81	1.0
2.51	0.94	0.26	0.63
1.29	0.22	0.20	0.85

As can be seen in these examples, the small values of $\hat{SD}_2[\bar{X}_n]$ yield large values of η^* which, in turn, result in an empirical coverage function not conforming to uniform (0,1). As further evidence of this point, the average value of $\hat{SD}_1[\bar{X}_n]$ for the 100 replications was 5.16 for sample size 100 (2.97 for sample size 200), whereas for $\hat{SD}_2[\bar{X}_n]$ it was 4.56 for sample size 100 (2.73 for sample size 200). This suggests that $\hat{Var}_2[\bar{X}_n]$ is underestimating $Var[\bar{X}_n]$. At sample sizes of 300 and 400 the average $\hat{SD}_2[\bar{X}_n]$ was again smaller than the average $\hat{SD}_1[\bar{X}_n]$; however, the difference was not sufficient to degrade the coverage function because the extreme lower tail values were not persistent. For example, of the 100 replications at sample size 100, no $\hat{SD}_1[\bar{X}_n]$ values were smaller than 2.0; however, 12 of the $\hat{SD}_2[\bar{X}_n]$ were smaller than 2.0. For sample size 200, no $\hat{SD}_1[\bar{X}_n]$ values were smaller than 1.2, but 7 of the $\hat{SD}_2[\bar{X}_n]$ values were. This explains the poor performance of the coverage functions for Method 2 at sample sizes of 100 and 200.

All of the reported measures of effectiveness in Table VI are adequate except for MOE 1, where the percent of correctly identified ARMA models was very low at all sample sizes. An explanation for this misidentification lies with the choice of ϕ_2, which for this process has a value of -0.18. This

relatively small value results in data adequately fitted by ARMA(1,1) models. In fact, of the 100 fitted ARMA models for each of the four sample sizes, 62, 66, 62, and 63, respectively, were of (1,1) order. This largely explains why there were so many misidentifications. In spite of these misidentifications, however, the coverage properties were still satisfactory for this theoretical output process.

The values of the average relative halfwidth listed in column 4(a), Table VI, demonstrate how this measure decreases with increasing sample size. The very small values for this measure are misleading, however, in that there is no standard against which to compare them. Simply by changing the process mean for which the data were generated, the relative halfwidths can be changed without either improving or degrading the confidence-interval procedure. Average relative halfwidths consequently are not comparable across TOPs having nonidentical means. For example, the seemingly large values of this measure in Table II were for a process mean of 0.6667, whereas the process mean for the Table V TOP is 1000. It is important to be aware that the average relative-halfwidth measure should only be used for comparison purposes as sample size changes for a given TOP.

Tables VII and VIII report the results for queuing-system TOPs 7 and 8, respectively. The small values for MOE 2 in these tables indicate that the coverage function for these TOPs does not conform to a uniform distribution. The coverage percentages given as MOE 4(b) also fall short of the 95% level. We will now discuss three sources of possible error which could account for the poor performance of the coverage properties with respect to these queuing models.

TABLE VII

Measures of Effectiveness for Analysis of an M/M/1 Model

Number of Observations per 100 Accepted Replications		1 % Correct Identification	2 Achieved Significance for Coverage Function	3 Coefficient of Variation of the Standard Error	4(a) Average Relative Half-Width	4(b) % Coverage (95% Confidence Level)	5 Replication Ratio
100	Method 1	NA	0.005	0.650	0.483	83	1.64
	Method 2		0.003	0.675	0.490	83	
200	Method 1	NA	0.0	0.533	0.328	81	1.33
	Method 2		0.0	0.540	0.330	81	
300	Method 1	NA	0.0	0.590	0.279	87	1.30
	Method 2		0.0	0.595	0.280	87	
400	Method 1	NA	0.0	0.358	0.231	79	1.31
	Method 2		0.0	0.361	0.232	79	

TABLE VIII

Measures of Effectiveness for Analysis of an M/D/3 Model

Number of Observations per 100 Accepted Replications		1 % Correct Identification	2 Achieved Significance for Coverage Function	3 Coefficient of Variation of the Standard Error	4(a) Average Relative Half-Width	4(b) % Coverage (95% Confidence Level)	5 Replication Ratio
100	Method 1	NA	0.0	1.08	1.13	67	1.33
	Method 2		0.0	1.36	1.42	67	
200	Method 1	NA	0.0	0.689	0.345	62	1.27
	Method 2		0.0	0.725	0.354	62	
300	Method 1	NA	0.0	1.21	0.432	68	1.38
	Method 2		0.0	1.31	0.455	68	
400	Method 1	NA	0.0	0.841	0.392	71	1.61
	Method 2		0.0	0.888	0.409	71	

1. Correlation between sample mean and its standard error

The use of the t statistic in building a confidence interval assumes that $\overline{X}_n$ and $\hat{SD}[\overline{X}_n]$ are independent. The extent to which this assumption is satisfied by the data can be checked by using the 100 replications at a given sample size to estimate the correlation between these two statistics. We computed the correlation between the sample means and their standard deviations as estimated by Methods 1 and 2 for all eight TOPs. The resulting correlation coefficients appear in Table IX.

The critical value at $\alpha = .01$ for the Table IX correlations under the assumption of bivariate normality is 0.26. Inspecting the table, we see that the ARMA TOPs 1 through 6 had insignificant correlations for both versions of the estimators of $Var[\overline{X}_n]$. However, there were significant correlations in all cases involving queuing-system TOPs 7 and 8. This correlation between the sample means and their estimated standard deviations contributes adversely to the behavior of the coverage function.

2. Distribution of the disturbance terms

The ARMA model in (1) assumes that the disturbance terms are normally distributed. This assumption, in turn, forms part of the basis for testing the statistical acceptability of the fitted ARMA models, and for the subsequent confidence-interval methodology. The validity of this assumption can be tested by investigating the distribution of the residuals of the fitted models.

We chose four replications, each of sample size 400, on which to check the normality assumption for the residuals. The first two replications were for TOP 5, the ARMA(2,1) process for which these CIP procedures work well. The first replication had a reported η^* of 0.09, and the second had a

Table IX

Correlation of Sample Means and Their Standard Errors

Number of Observations per 100 Accepted Replications		Theoretical Output Processes 1	2	3	4	5	6	7	8
100	Method 1	-.11	-.01	-.20	.15	-.25	.22	.74	.57
	Method 2	-.11	.00	-.20	.15	-.25	.22	.74	.49
200	Method 1	.06	.01	-.05	.03	.09	-.01	.88	.90
	Method 2	.06	.00	-.05	.03	.10	-.01	.88	.90
300	Method 1	-.13	.01	.01	.09	-.07	.19	.85	.84
	Method 2	-.10	.01	.01	.09	-.07	.18	.85	.84
400	Method 1	.13	-.09	-.04	.04	.07	.00	.76	.88
	Method 2	.13	-.09	-.04	.04	.07	.01	.76	.88

reported η^* of 1.0. This means that the confidence interval from the first replication covered the true mean for all confidence levels at or above 9%. The confidence interval associated with the second replication covered only at confidence levels approaching 100%.

The other two replications were chosen from TOP 8. Their respective η^*'s were 0.37 and 1.0. Hence, these replications had specifications similar to those of the two replications chosen from TOP 5.

Table X summarizes the results of checking the residuals from the chosen four replications for normality with a mean of zero. The table indicates the mean, standard deviation, skewness, kurtosis, and the achieved significance level of the χ^2 goodness-of-fit test for normality.

As expected, the distribution of the residuals from the tailor-made TOP 5 is consistent with the underlying assumption of normality. Their kurtosis (which can be compared to a kurtosis of zero for the normal distribution) and skewness (which can be compared to a skewness of zero for the normal) substantiate normality, as does the achieved significance level of the associated χ^2 statistics.

However, the measures of skewness, kurtosis, and goodness-of-fit do not support the assumption of normality for the queuing-TOP residuals. The skewness and kurtosis are not close to zero, and the small values of the achieved significance levels of the goodness-of-fit tests indicate a rejection of the normality hypothesis. The lack of normality in the residuals is a factor contributing to the poor coverage encountered when working with observations produced by the queuing-system simulations.

Table X

Residual Analysis

	Replication 1 (TOP 5, η^*=.09)	Replication 2 (TOP 5, η^*=1.00)	Replication 3 (TOP 8, η^*=.37)	Replication 4 (TOP 8, η^*=1.00)
Mean	-.641	.516	-.007	-.022
Standard Deviation	76.12	72.64	1.98	1.98
Skewness	.061	.095	.430	.312
Kurtosis	-.220	-.009	.372	-.261
Achieved Significance Level of χ^2-Test	.22	.68	.00	.09

3. Aberrant behavior of the replicated means

A third possible contributor to poor coverage involves situations in which achieved sample means may persistently be far removed from the process mean. If the replicated means consistently tend to be far removed from the process mean, then this can result in poor performance characteristics for any confidence-interval procedure. The halfwidths needed to cover the process mean tend to be persistently large in such a situation, resulting in large values of η^* and strongly nonuniform behavior for the coverage function.

To investigate this possibility here, we computed the grand mean for all 100 replications at each sample size for TOPs 7 and 8. Then we evaluated MOE 2 and 4(b) for each combination to determine the extent to which the ARMA-based confidence intervals covered these grand means. In essence, we were adjusting for the bias in the generating process. The MOE 2 and 4(b) coverage properties improved slightly but were still unsatisfactory except for sample size 400 with TOP 7. We conclude that the generation process was not the cause of poor coverage with the queuing TOPs.

The final aspect of the test results involves the order of the fitted ARMA models. The counts of the achieved orders for TOPs 7 and 8 are shown in Table XI. Table rows indicate the number of observations per replication; and table columns show the orders of the models, arranged by increasing sum of the autoregressive and moving-average orders. Each table cell shows the two applicable counts, with the TOP 7 and 8 counts in the upper and lower parts of each cell, respectively. For example, with 200 observations per replication, there were 61 and 79 ARMA(1,0) models fitted for respective TOPs 7 and 8. We conclude that low-order ARMA models produce statistically acceptable fits for queuing system data, at least for the case of the two queuing TOPs used for testing purposes in this work.

Number of Observations per 100 Accepted Replications	(p,q)											
	(1,0)	(1,1)	(2,0)	(2,1)	(1,2)	(3,0)	(3,1)	(2,2)	(1,3)	(3,2)	(2,3)	(3,3)
100	63 / 90	8 / 6	22 / 3	3 / 1	2 / 0	1 / 0	0 / 0	0 / 0	0 / 0	1 / 0	0 / 0	0 / 0
200	61 / 79	12 / 12	18 / 2	2 / 1	2 / 1	3 / 0	0 / 0	0 / 0	0 / 0	0 / 0	1 / 0	1 / 0
300	62 / 70	13 / 25	19 / 3	0 / 1	0 / 1	3 / 0	1 / 0	0 / 0	0 / 0	1 / 0	0 / 0	1 / 0
400	65 / 58	16 / 36	14 / 2	1 / 2	0 / 2	0 / 0	0 / 0	3 / 0	1 / 0	0 / 0	0 / 0	0 / 0

Table XI

Counts of the (p,q) Orders for TOP's 7 and 8

The results in Table XI can also be compared with the statement of Steudel and Wu [1977] that output from an M/M/1 queuing system can be fitted with an ARMA(1,0) model and that, in general, ARMA models of order (p,p - 1) provide acceptable fits for outputs from queuing-system simulations. In the TOP 7 M/M/1 system, 251 of the 400 total replications were fitted by ARMA(1,0) models. And in the TOP 8 M/D/3 queuing system, 302 of the 400 total replications were fitted by either ARMA(1,0) or ARMA(2,1) models.

6. CONCLUSIONS

Two ARMA-based confidence-interval procedures have been described, and results of subjecting these procedures to extensive testing have been reported. The differences in the performance characteristics of the alternative procedures are small. Both procedures work well when used to process ARMA-generated data for the ranges and combinations of autoregressive and moving-average orders and parameter values investigated. Although Method 1 produces somewhat more stable confidence intervals than Method 2 as measured by MOE 3, the differences in the performance characteristics of the two alternative procedures for tailor-made data are small on balance.

Both confidence-interval procedures perform in less-than-satisfactory fashion when used to process data produced by two queuing-system simulations chosen for testing purposes. The underlying cause for this poor performance may be the demonstrated correlation between the sample means and their standard errors in the queuing output, and/or the demonstrated nonnormality in the distribution of the residuals associated with the ARMA models fitted to the queuing-system output. We advise that practitioners conduct appropriate correlation and normality tests on simulation output prior to using these ARMA-based confidence-interval procedures.

ACKNOWLEDGEMENTS

This research was supported by the Office of Naval Research under Contract N00014-81-K-0120. The authors acknowledge the helpful comments of Clifford Ball, E. Philip Howrey, Joseph A. Machak, Robert G. Sargent, and Bruce Schmeiser.

REFERENCES

Andrews, R. W., and T. J. Schriber. 1978. Interactive Analysis of Output from GPSS-Based Simulations. Proceedings of the 1978 Winter Simulation Conference, 267-278. ACM, New York.

Beguin, J. M., C. Gourieroux, and A. Monfort. 1981. Identification of a Mixed Autoregressive-Moving Average Process: The Corner Method, in Time Series, ed. O. D. Anderson, 423-436. North-Holland, Amsterdam.

Box, G. E. P., and G. M. Jenkins. 1976. Time Series Analysis: Forecasting and Control, Rev. Ed. Holden-Day, San Francisco.

Cox, D. R., and H. D. Miller. 1965. The Theory of Stochastic Processes. Halstead Press, New York.

Fishman, G. S. 1971. Estimating Sample Size in Computing Simulation Experiments. Management Science 18, 21-38.

Fishman, G. S. 1978. Principles of Discrete Event Simulation. Wiley Interscience, New York.

Fox, D., and K. Guire. 1976. Documentation for MIDAS, 3rd ed. The University of Michigan, Ann Arbor, Mich.

Fuller, W. A. 1976. Introduction to Statistical Time Series. Wiley & Sons, New York.

Gray, H. L., G. D. Kelley, and D. D. McIntire. 1978. A New Approach to ARMA Modeling. Communications in Statistics B7, 1-77.

Henriksen, J. O., and R. Crain. 1982. GPSS/H User's Manual, 2nd Ed. Wolverine Software Corp. Annandale, VA.

Hillier, F. S., and G. J. Lieberman. 1974. Introduction to Operations Research. Holden-Day, San Francisco.

Hillier, F. S., and O. S. Yu. 1981. Queuing Tables and Graphs. ORSA Publications in Operations Research Series, Vol. 3. Elsevier North Holland, New York.

International Mathematical and Statistical Libraries Inc. Library Reference Manual, 8th Ed. 1980. Houston, Tex.

Ljung, G. M., and G. E. P. Box. 1978. On a Measure of Lack of Fit in Time Series Models. Biometrika 65, 297-303.

Pritsker, A. A. B., and C. D. Pegden. 1979. Introduction to Simulation and SLAM. Wiley & Sons, New York.

Schriber, T. J. 1974. Simulation Using GPSS. Wiley & Sons, New York.

Schriber, T. J., and R. W. Andrews. 1981. A Conceptual Framework for Research in the Analysis of Simulation Output. Communications of the ACM 24, 218-232.

Schmeiser, B. 1982. Batch Size Effects in the Analysis of Simulation Output. Operations Research 30, 4xx-4yy (forthcoming).

Schmeiser, B., and K. Kang. 1981. Properties of Batch Means from Stationary ARMA(1,1) Time Series. School of Industrial Engineering, Purdue University Research Memorandum No. 81-3. West Lafayette, Ind.

Schruben, L. W. 1980. Coverage Function for Interval Estimators of Simulation Responses. Management Science 26, 18-27.

Steudel, H. J., S. M. Pandit, and S. M. Wu. 1978. Interpretation of Dispatching Policies on Queue Behavior via Simulation and Time Series Analysis. AIIE Transactions 10, 292-298.

Steudel, H. J., and S. M. Wu. 1977. A Time Series Approach to Queuing Systems with Applications for Modeling Job-shop In-process Inventories. Management Science 23, 745-755.

Tausworthe, R. C. 1965. Random Numbers Generated by Linear Recurrence Modulo Two. Mathematics of Computation. 19, 201-209.

Tiao, G. C., and R. S. Tsay. 1981. Identification of Nonstationary and Stationary ARIMA Models. Department of Statistics, University of Wisconsin Technical Report No. 647. Madison, Wisc.

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Working Paper No. 304	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: Two ARMA-Based Confidence-Interval Procedures for the Analysis of Simulation Output	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Richard W. Andrews, Thomas J. Schriber	8. CONTRACT OR GRANT NUMBER(s): N00014-81-K-0120
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Michigan, Ann Arbor MI 48109	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, 800 North Quincy Street, Arlington VA 22217	12. REPORT DATE: April, 1982
	13. NUMBER OF PAGES: 47
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Simulation	Automatic identification of ARMA models
Confidence intervals	Queuing
Autoregressive	Output analysis
Moving average	Time series

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Two methods are presented for building interval estimates on the mean of a stationary stochastic process. Both methods fit an autoregressive moving-average (ARMA) model to observations on the process. The model is used to estimate the variance of the sample mean and the applicable degrees of freedom of the t statistic. Fitting of the ARMA model is totally automated. The ARMA-based confidence intervals perform well with data generated from ARMA processes. With data generated from queuing-system simulations, the coverage of the confidence intervals is less than satisfactory. It is shown that with queuing-system data,

DD FORM 1 JAN 73 1473

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

20 (continued)

the sample mean and its estimated standard deviation are strongly positively correlated, and that the residuals of the fitted models are not normally distributed. These factors contribute adversely to the coverage of the confidence-interval procedures with queuing data.