# Detecting Initialization Bias in Simulation Output

LEE W. SCHRUBEN
*Cornell University, Ithaca, New York*
(Received February 1980; accepted December 1981)

A general approach to testing for initialization bias in the mean of a simulation output series is presented. The output is transformed into a standardized test sequence that can be contrasted with a known limiting stochastic process. This transformation requires very little computation and the asymptotic theory is applicable to a wide variety of simulations. An initialization bias test is developed and several examples of its application are presented.

SIMULATION of a stochastic system with a computer program requires that starting conditions for each run be completely specified. These initial conditions should be randomly selected in accordance with the true joint probability of their occurrence. However, this is typically impractical or impossible. Often the selected starting conditions represent extremely unusual system states. Some examples: An industrial production simulation initially has no work in process and all machines are in good working order; a distribution system simulation is started with all items in stock; a simulated telephone exchange has no calls in progress. Each run of these simulation programs begins with a sequence of events that might have a low probability of occurring had the events not been artificially induced by the initial conditions. The result is that the output is contaminated by an *initialization bias.*

Initialization bias can be a major source of error in estimating the steady-state value of a simulated system performance measure. This problem is particularly troublesome when several independently seeded runs of the program are made and the results are used to construct confidence intervals. The frequency with which confidence intervals based on biased output include the true performance value generally decreases as more runs are made. This is caused by the intervals shrinking about an inaccurate point estimate. The influence of simulation initialization bias on interval coverage is demonstrated in Law [1977].

The usual method of controlling simulation initialization bias is to allow the program to run for a "warm-up" period before output data are collected. This is equivalent to discarding some early portion of the output. The procedure is referred to as output *truncation* and the

*Subject classification:* 563 application of diffusion processes, 767 simulation initialization bias tests.



**Operations Research**
Vol. 30, No. 3, May–June 1982
0030-364X/82/3003-0569 $01.25

observation after which data are retained for analysis is called the *truncation point.* There is a trade-off involved in selecting a truncation point. If too little of the initial output is discarded the remaining bias may adversely affect the results; ignoring too much of the output is wasteful (resulting in a higher than necessary variability in summary statistics).

The literature on simulation methodology contains many techniques for controlling initialization bias (see the survey by Wilson and Pritsker [1978]). New methods are frequently suggested (e.g. Kelton [1980]). These initialization bias control procedures are often elaborate and offer no assurance that initialization bias will be effectively controlled (Gafarian et al. [1978]). Of central interest in many simulation studies are two basic questions addressed in this article:

1. Is initialization bias in a simulation run a serious problem?
2. If so, is a particular initialization bias control procedure effective?

The first question is asked when deciding whether or not to expend effort in initializing a simulation program; the second question is asked once an initialization strategy is adopted.

In Sections 1 and 2 of this article, a general approach to detecting initialization bias in the mean of a simulation output series is presented. The basic concept involves standardizing the stochastic process being simulated so that it represents "noise" in which a "signal," due to initialization bias, may be detected. In Section 3, an initialization bias detection procedure is developed using this approach. Several applications are presented in Section 4. These examples illustrate the validity and power of the procedure for a variety of simulation programs.

## 1. BACKGROUND AND THEORY

A run of a simulation program results in a sequence of output data, $y_1, y_2, \cdots, y_n$. Here $n$ is called the *run length.* The $i$th observation, $y_i$, may represent the output from a single run, or

(i) An average over several runs that are not necessarily independently seeded, or
(ii) An average of equal sized, adjacent (not necessarily exclusive) groups or "batches" of observations from a single run, etc.

Data gathering may have begun at the start of the run(s) or may have been delayed to allow the simulation to "warm up."

The objective of the simulation run is to estimate the mean, $\mu$, of the process being simulated. This includes a variety of simulation studies. For example, $y_i$ can assume the values of zero or one indicating the absence or presence of a particular condition during the run; $\mu$ then is the steady-state probability that the condition is present.

The estimator of $\mu$ will be the average of the output series, $\bar{Y}_n$. The expected value of $\bar{Y}_n$ is denoted as $\bar{\mu}_n$. The bias, $b(\bar{Y}_n)$, in $\bar{Y}_n$, as an estimator of $\mu$, is given by $b(\bar{Y}_n) = \bar{\mu}_n - \mu$.

The simulation output may be conceptualized in a fairly general manner as a continuous time stochastic process, $\{Z_t;\ 0 < t < \infty\}$. $Z_t$ represents the "state" of the simulation at time $t$ and may take on values in a multiple dimensional or other suitable space. When the simulation is run the output consists of observations, $y_i$, of the stochastic process,

$$Y_i = \mu_i + X_i;\ i = 1, 2, \cdots, n. \tag{1}$$

The time that $Y_i$ is observed is $t_i$ and the run is started at $t_0 = 0$. The stochastic process $X_i$ is a real valued function of the process

$$\left\{ \int_{t_{i-1}}^{t_i} Z_s ds;\ i = 1, 2, \cdots . \right\}$$

The unknown deterministic function, $\mu_i$, represents the potential shift in the output mean that is introduced by initializing and running the simulation. Without loss in generality take $E[X_i] = 0$ so that $\mu_i = E[Y_i]$. Here the statement, "*initialization bias is present*" means that $\mu_i$ changes value for at least one observation in the output series. The statement "*no initialization bias is present*" means that the function $\mu_i$ does not change for the entire run. Transient initialization effects on higher order moments in the output process may also exist (Schruben [1981]). These second order effects of initializing a simulation program are not studied in this paper.

We will apply the theory for dependent stochastic processes presented in Chapters 20 and 21 of Billingsley [1968]. Certain technical assumptions are needed concerning the dependencies in the simulation output process. In particular, we assume that $\{X_i\}$ is stationary and $\phi$-*mixing* with a finite variance.

The property of $\phi$-mixing can be defined as follows. Let $E$ be an event (with a positive probability of occurring) that depends only on the behavior of a process up to time $t_k$. Let $F$ be an event that depends only on the behavior of the process after time $t_{k+n}$. That is, $n$ time units pass between possible occurrences of event $E$ and event $F$. We say that the process is $\phi$-mixing if the supremum (over $k$, $E$, and $F$) of $|\text{Prob}(F|E) - \text{Prob}(F)|$ is bounded by a real valued function of $n$, $\phi_n$, with $\lim_{n\to\infty} \phi_n = 0$. Intuitively, the distant future behavior of the process (i.e. event $F$) is almost independent of the present or past behavior of the process (i.e. event $E$).

If we also assume that $\sum \phi_n^{1/2}$ converges, a functional central limit theorem (Theorem 21.1 of Billingsley) applies. A slight modification (we

center the output around the sample mean) of this theorem forms the foundation for our approach. Readers comfortable with applications of the classical (i.i.d. scalar) central limit theorem will recognize that we are using a *process* version of a central limit theorem in a familiar manner: to obtain a limiting distribution for a test statistic.

The assumption that the output process is $\phi$-mixing is less restrictive than the ARMA time series models (Fishman [1973]) and $m$-dependent models (Mechanic and McKay [1966]) that have been postulated for simulation output. A proof that ARMA time series have the properties we assume for $\{X_i\}$ is given by Johnson and Bagshaw [1974]. In fact, a broad class of dependent stochastic processes has the property of $\phi$-mixing.

No specific assumptions about the function $\mu_i$ are made. Some examples of $\mu_i$ commonly encountered in simulation runs are illustrated in Figure 1. The dashed lines in these figures are at $\bar{\mu}_n$. Figure 1-a represents a transient mean function that monotonically approaches an asymptote (typical in simulations of queues). Figure 1-b is an example of the mean function for a simulation model with a feedback mechanism (see Schruben [1981] for an example of such a simulation). Figure 1-c illustrates a simulation (with a slowly dissipating initialization bias) that has not been run long enough to allow the program to warm up. An effective bias detection procedure would indicate that initialization bias is present in these and other situations in which $\mu_i$ is not constant throughout the run.

The output series is standardized and analyzed as a "signal" due to a changing mean function $\mu_i$ in the presence of "noise" due to $X_i$. Standardization is similar to that used in common applications of the classical central limit theorem; the sequence is centered at zero and its magnitude scaled. In addition, the run duration is standardized to the unit interval. Instead of a single limiting normal random variable, a limiting stochastic process is employed in the analysis.

The limiting stochastic process used is a standard Brownian bridge $\{\mathscr{B}_t;\ 0 \le t \le 1\}$, i.e., Brownian motion on the unit interval conditioned to start and return to zero. $\mathscr{B}_t$ here is the process analogue to the standard normal random variable used in applications of the classical central limit theorem. The Brownian bridge process has continuous sample paths and four notable properties:

1. $\mathscr{B}_0 = \mathscr{B}_1 = 0$,
2. $E[\mathscr{B}_t] = 0;\ 0 \le t \le 1$,
3. $\mathrm{Cov}(\mathscr{B}_{t_1}, \mathscr{B}_{t_2}) = \min(t_1, t_2)(1 - \max(t_1, t_2))$, and
4. Sets of $\mathscr{B}_t$ have a jointly normal distribution.

## 2. STANDARDIZING THE SIMULATION OUTPUT

Tests for the presence of initialization bias in simulation output can be

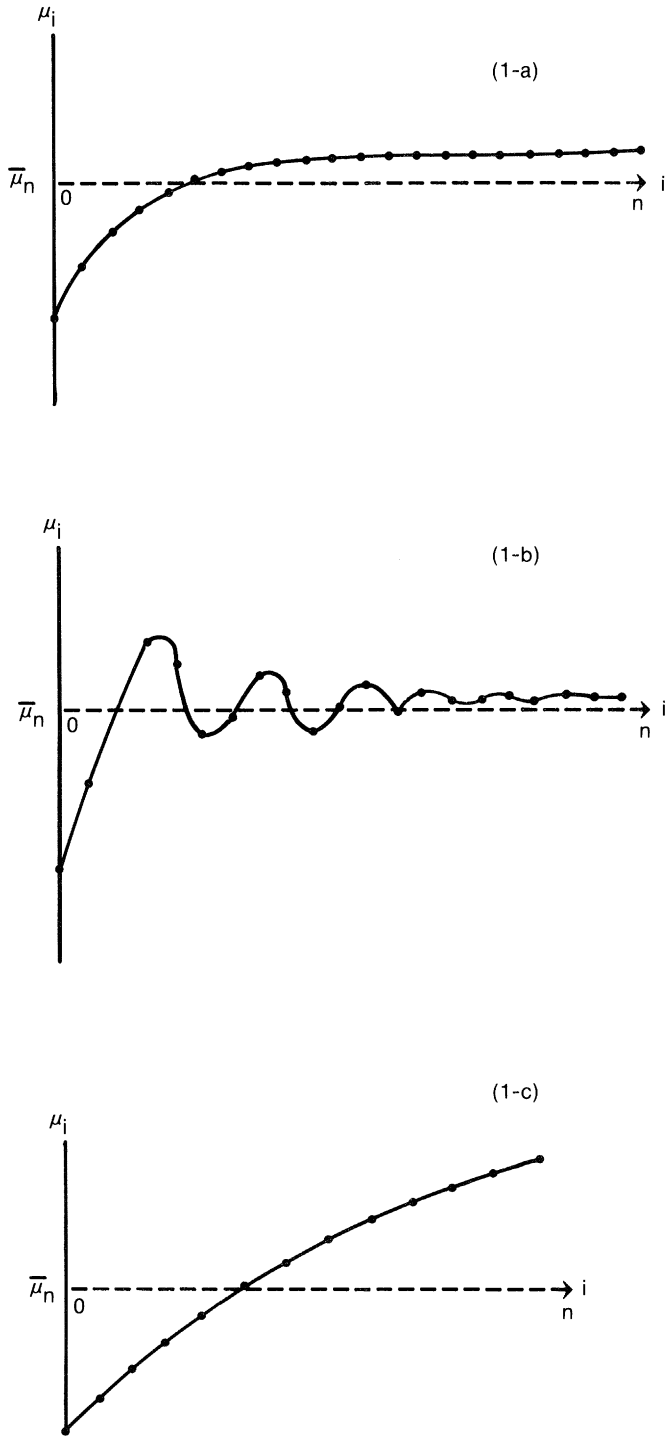


**Figure 1.** Some examples of transient mean functions, $\mu_i$, due to initialization of a simulation program.

based on the sequence of partial sums,

$$S_n(k) = \bar{Y}_n - \bar{Y}_k;\ k = 1, 2, \cdots, n, \tag{2}$$

with $S_n(0)$ defined to be zero. The sequence $S_n(k)$ contains the differences between the average of the entire output series, $\bar{Y}_n$, and the average of the first $k$ observations, $\bar{Y}_k$.

By substituting (1) into (2), $S_n(k)$ can be expressed as a deterministic (but unknown) "signal" component and a stochastic "noise" component, i.e., $S_n(k) = M_n(k) + X_n(k)$, where $M_n(k) = \bar{\mu}_n - \bar{\mu}_k$ represents the deterministic signal and

$$X_n(k) = n^{-1} \sum_{i=1}^{n} X_i - k^{-1} \sum_{i=1}^{k} X_i$$

represents the stochastic noise in $S_n(k)$.

## 2.1 The Noise Process

The noise process $X_n(k)$ is scaled and a functional central limit theorem is employed to show convergence to $\mathscr{B}_t$ as the run length becomes large. To illustrate the scaling arguments, consider the simple situation in which $\{X_i\}$ is a sequence of independent identically distributed random variables with finite variance $\sigma^2 = \mathrm{Var}(X_i)$. The random variables $W_i = \bar{X}_n - X_i$ $(i = 1, 2, \cdots, n)$ have the following properties

(1) $E[W_i] = 0$
(2) $\mathrm{Var}(W_i) = ((n-1)/n)\sigma^2$
(3) $\mathrm{Cov}(W_i, W_j) = -(\sigma^2/n) i \neq j.$

The second moments of $W_i$ can be easily found by noting that $\mathrm{Cov}(\bar{X}_n, X_i) = \mathrm{Var}(\bar{X}_n)$. The values of the process $kX_n(k)$ are equal to $\sum_{i=1}^{k} W_i$.

The covariance structure of $\{kX_n(k);\ k = 1, 2, \cdots, n\}$ is

$$\mathrm{Cov}(lX_n(l), kX_n(k)) = \sum_{s=1}^{k} \sum_{r=1}^{l} \mathrm{Cov}(W_s, W_r)$$
$$= (1/\mathrm{n})\min(l, k)(n - \max(l, k))\sigma^2.$$

If the sequence $\{kX(k);\ k = 1, 2, \cdots, n\}$ is now divided by $\sqrt{n}\,\sigma$ then

$$\mathrm{Cov}(lX_n(l)/(\sqrt{n}\sigma), kX_n(k)/(\sqrt{n}\sigma)) = (\min(l, k)/n)(1 - (\max(l, k)/n)).$$

Next the run duration is scaled to the unit interval by defining $t_k = k/n$. For values of $t_l$ and $t_k$ equal to $1/n$ or $2/n$ or $\cdots$ or 1,

$$\mathrm{Cov}((t_l n X_n(t_l n)/(\sqrt{n}\sigma)), (t_k n X_n(t_k n)/(\sqrt{n}\sigma)))$$
$$= \min(t_l, t_k)(1 - \max(t_l, t_k)).$$

Define

$$B_t = tnX_n(tn)/(\sqrt{n}\sigma);\ t = 1/n, 2/n, \cdots, 1$$

with $B_0 = 0$ (note that $X_n(n) = 0$ so $B_1 = 0$). Then for integer values of $tn$, $B_t$ has the following properties

(1) $B_0 = B_1 = 0$
(2) $E[B_t] = 0$
(3) $\text{Cov}(B_{t_l}, B_{t_k}) = \min(t_l, t_k)(1 - \max(t_l, t_k))$.

As the run duration, $n$, goes to infinity, $t$ may take on essentially all values in the unit interval. A central limit argument can be used to give the asymptotic normality of $X_n(k)$. Thus it may be argued that $\lim_{n\to\infty} B_t = \mathscr{B}_t$.

The convergence of $[tn]X_n([tn])/(\sqrt{n}\sigma)$ (where $[\cdot]$ is the greatest integer function) to a Brownian bridge process is demonstrated for more generally dependent stochastic process using Theorem 21.1 in Billingsley. For dependent processes $\sigma^2 = \lim_{n\to\infty} n \text{ Var}(\bar{X}_n)$ is used to scale the magnitude of the process. In the case of independent identically distributed $X_i$, $\sigma^2 = \text{Var}(X_i) = n \text{ Var}(\bar{X}_n)$. Since simulation runs are typically rather long (especially when initialization bias is suspected) it is reasonable to expect the functional central limit theorem employed here to produce good results. Estimators of $\sigma^2$ are discussed in Section 3.2.

### 2.2. The Signal

The objective is to detect the possible presence of an unknown nonzero deterministic signal, $M_n(k) = \bar{\mu}_n - \bar{\mu}_k$, due to the presence of initialization bias. Scaling the magnitude of $M_n(k)$ and the run duration to the unit interval in the same manner as done to standardize the noise process results in a standard deterministic signal $D_t$ given by

$$D_t = tnM_n(tn)/(\sqrt{n}\sigma);\ t = 1/n, 2/n, \cdots, 1.$$

Note that if $\mu_i$ is constant throughout the run then $D_t$ is zero for all values of $t$.

Figures 2-a, 2-b, and 2-c illustrate the signals, $D_t$, due to the corresponding mean functions in Figures 1-a, 1-b, and 1-c. There are several noteworthy features of the signals in Figure 2. First, they are not zero. Second, there is a prominent peak in the signal. Third, the location of this prominent peak occurs relatively near the beginning of the run (small $t$). Tests for initialization bias in a simulation output series can be based on these and other characteristics of the signal $D_t$. In the next section a test is developed based on the magnitude of the peak in the signal and its location.

## 3. TESTS FOR INITIALIZATION BIAS

Testing for the presence of initialization bias in simulation output

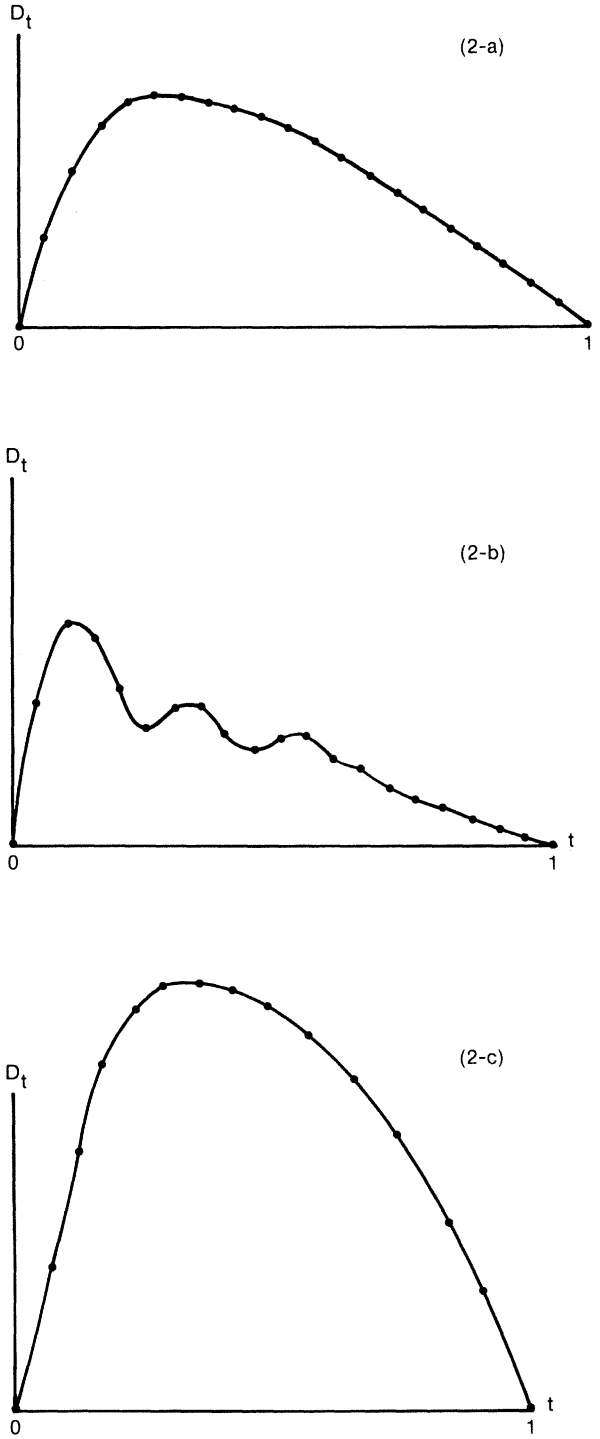


**Figure 2.** Signals, $D_t$, due to the corresponding transient mean functions in Figure 1.

involves analysis of the standardized test sequence,

$$T_n(t) = [tn]S_n([tn])/(\sqrt{n}\sigma);\ t \in (0, 1], \tag{3}$$

with $T_n(0) = 0$. The test presented here focuses on the maximum deviation of the test sequence from zero (if ties occur only the first maximum is considered). The basis for this test is the prominent peak in the "signal" that occurs at a relatively small value of $t$ (characteristic of the presence of initialization bias).

It is presumed here that the experimenter wishes to guard against bias of a particular sign. For example: the simulation of a queueing system may be most conveniently started in an undercongested state, inducing a negative bias in expected waiting times. In what follows, it is presumed that the experimenter suspects negative bias (typical of a simulation model that is started "empty and idle"). If positive bias is suspected, the output series is multiplied by $-1$ before testing. Knowledge of the sign of potential initialization bias allows a more powerful "one-sided" test than is possible otherwise (see the Appendix in Schruben [1980] for a two-sided test). The sign of the suspected bias is information often available to the simulation user. The observed maximum deviation in the test sequence is going to be squared before testing for bias. In Section 4.4 the consequences of testing for bias of the wrong sign are discussed.

### 3.1. The Situation Where No Initialization Bias Is Present

Consider the case in which $\mu_i$ is a constant throughout the run of a simulation program (we would hope that $\mu_i = \mu$). As shown in Section 2.1, the sequence $T_n(t)$ can then be modeled as a standard Brownian bridge process, $\{\mathscr{B}_t;\ t \in [0, 1]\}$. Let $t^*$ denote the location of the maximum of $\mathscr{B}_t$ and $s^* = \mathscr{B}_{t^*}$. The joint density of $t^*$ and $s^*$ is given by

$$f_{t^*,s^*}(t, s) = 2s^2/(t(1 - t))\phi(s, 0, t(1 - t));\ s \geq 0,\ t \in [0, 1], \tag{4}$$

where $\phi(s, \mu, v^2)$ is the normal density function at $s$ with mean $\mu$ and variance $v^2$. The joint density (4) of $t^*$ and $s^*$ can be found by conditioning the result in Shepp [1979] on Brownian motion being zero at time equal to one.

Rather than use $s^*$ directly, consider another random variable, $x^* = s^{*2}/(t^*(1 - t^*))$. The joint density of $t^*$ and $x^*$ is given by

$$g_{t^*,x^*}(t, x) = (x)^{1/2}e^{-x/2}/\sqrt{2\pi};\ x \geq 0,\ 0 \leq t \leq 1. \tag{5}$$

From this joint density of $t^*$ and $x^*$ note the following:

(1) $t^*$ and $x^*$ are independent,
(2) $t^*$ is uniformly distributed on the unit interval, and
(3) $x^*$ has a $\chi^2$ distribution with 3 degrees of freedom.

Let $\hat{t}$ denote the observed location of the (first) maximum in $\{T_n(t)\}$ and $\hat{s} = \sigma T_n(\hat{t})$. The initialization bias testing procedure involves assessing whether or not the observed value of $\hat{x} = \hat{s}^2/(\sigma^2(\hat{t}(1 - \hat{t}))$ would be *unusual* if the output series contained *no* initialization bias. If $|\bar{\mu}_n - \bar{\mu}_k|$ is increased for $k < n$ then $\hat{x}$ is likely to increase. Therefore, large values of $\hat{x}$ are regarded as unusual.

### 3.2. Estimation of $\sigma^2$

The quantity $\sigma^2$ in (3) is the limiting variance (as $n \to \infty$) of $(n)^{-1/2} \sum_{i=1}^{n} X_i = n^{1/2}\bar{X}_n$ and is used to scale the magnitude of the test sequence presented in this paper. Since $\sigma^2$ is unknown it must be estimated from the simulation output. Several possible estimators, $\hat{\sigma}^2$, for $\sigma^2$ have been suggested in the simulation methodology literature. Often it is appropriate to regard $\nu\hat{\sigma}^2/\sigma^2$ as having an approximate $\chi^2$ distribution with $\nu$ degrees of freedom. Some alternative estimators for $\sigma^2$ are presented in Table 32 of Fishman along with estimated degrees of freedom, $\nu$ (note that $\sigma^2 = \lim_{n\to\infty} n \operatorname{var}(\bar{X}_n)$ is denoted as $m$ in this table).

The experiments discussed in Section 4 use an estimator of $\sigma^2$ that is based on fitting the latter portion of the simulation output to a $p$-order autoregressive model, i.e.,

$$Y_i = \phi_1 Y_{i-1} + \phi_2 Y_{i-2} + \cdots + \phi_p Y_{i-p} + \epsilon_i. \tag{6}$$

Here $\{\epsilon_i\}$ is a sequence of independent random variables that are normally distributed with zero mean and variance $\sigma_\epsilon^2$. There are many computer programs available for fitting the above autoregression (e.g., Appendix B of Fishman). Once estimates are made for $\phi_1, \phi_2, \cdots, \phi_p, \sigma_\epsilon^2$ (denoted as $\hat{\phi}_i$ for $i = 1, \cdots p$ and $\hat{\sigma}_\epsilon^2$) they may be used to estimate $\sigma^2$ by

$$\hat{\sigma}^2 = \hat{\sigma}_\epsilon^2(1 - \hat{\phi}_1 - \hat{\phi}_2 - \cdots - \hat{\phi}_p)^{-2}. \tag{7}$$

The degrees of freedom, $\nu$, used with this estimator is given by

$$\nu = (n\hat{\sigma}_\epsilon)/(2\hat{\sigma}(p - \textstyle\sum_{i=1}^{p} (p - 2i)\hat{\phi}_i)). \tag{8}$$

(See the footnote to Fishman in the references.)

The reader probably noted that it was suggested that only some "latter portion" of the simulation output should be used to estimate $\sigma^2$. In the experiments presented in Section 4, only the last half of each output sequence is used to fit the autoregressive model. More of the output should be used if there is a reasonable belief that initialization bias is small. Some crude truncation of the output is suggested prior to estimating $\sigma^2$ (truncating half of the output is certainly crude).

The philosophy here is that it is more beneficial to detect initialization bias when it is present than it is detrimental to falsely conclude bias is present when it is not. Therefore, it is considered better to underestimate

$\sigma^2$ (and have the magnitude of the scaled test sequence be too large) than to overestimate $\sigma^2$. Most common estimators of $\sigma^2$ have a tendency to be too small in the presence of the positive serial correlation that is typical of simulation output. This is fortunate in this particular application of these estimators (but problematic in other applications such as confidence interval estimation). As demonstrated in Section 4 of this paper, using the last half of the output to estimate $\sigma^2$ produced good results. In Section 4.6, a modification is suggested that eliminates the need to estimate $\sigma^2$.

### 3.3. A Test Procedure

As discussed earlier, a large positive maximum value of the scaled test sequence $T_n(t)$ is unusual if no negative initialization bias is present. To quantitatively measure bias in a particular simulation run, a procedure is presented in this section to estimate the probability that a test statistic *more unusual* than that observed will occur if there is no initialization bias present. This probability is denoted by $\hat{\alpha}$. Large values of $\hat{\alpha}$ indicate that the output does not contain a significant negative initialization bias (*more unusual* test statistics are likely). Small values of $\hat{\alpha}$ indicate that the results are highly unusual for unbiased runs (suggesting the presence of negative initialization bias). The quantity $\hat{\alpha}$ is the observed level of significance for this test of initialization bias in the mean.

In a hypothesis testing framework the null hypothesis is that the output process has a constant mean. A "no negative bias" hypothesis is rejected if the observed value of $\hat{\alpha}$ is less than the specified probability, $\alpha$, of rejecting a true hypothesis ($\alpha$ is the type I error level).

The procedure for estimating $\hat{\alpha}$ is based on results developed so far:

(i) $\hat{x} = \hat{s}^2/(\sigma^2(\hat{t}(1-\hat{t}))$ has approximately a ${\chi_3}^2$ distribution when no initialization bias is present, and
(ii) $\nu\hat{\sigma}^2/\sigma^2$ has approximately a ${\chi_\nu}^2$ distribution.

It follows that if $\hat{\sigma}^2$ and $\hat{x}$ were independent (see Recommendation 3 of Section 4.6) when no bias is present, then

$$\hat{h} = \hat{s}^2/(3\hat{\sigma}^2\hat{t}(1-\hat{t})) \tag{9}$$

will have approximately an $F$ distribution with 3 and $\nu$ degrees of freedom. The above results lead to a procedure:

*Step 1.* Find $\hat{s}$, the (first if ties occur) global maximum of $\{kS_n(k)/\sqrt{n}$ for $k = 1, \cdots, n\}$, and its location $\hat{k}$.
*Step 2.* Compute $\hat{\sigma}^2$ and $\nu$ (Equations 7 and 8).
*Step 3.* Set $\hat{t} = \hat{k}/n$ and compute $\hat{h}$ using Equation 9.
*Step 4.* Compute $\hat{\alpha} = \bar{F}_{3,\nu}(\hat{h})$ where $\bar{F}_{3,\nu}(\cdot)$ is the upper tail of the distribution function for an $F$ variate with 3 and $\nu$ degrees of freedom.

*Step 5* (optional). Reject a hypothesis of no negative bias if $\hat{\alpha} < \alpha$.

In Step 1, order $n$ storage locations are required and the computation is insignificant (about that of summing the output twice). Alternative estimators for $\hat{\sigma}^2$ and $\nu$ in Step 2 are given in Table 32 of Fishman; Equations 7 and 8 are efficiently obtained using a modification of the subroutine in Appendix B of Fishman. In Step 4,

$$\bar{F}_{3,\nu}(\hat{h}) = \text{Prob}\{\text{an } F_{3,\nu} \text{ variate } \textit{exceeds } \hat{h}\}$$

can be easily obtained.

## 4. SOME APPLICATIONS

The significance test for initialization bias is applied to five different simulation programs. Two questions are of particular concern: (1) Is the testing procedure valid? and (2) Is the procedure effective?

If the procedure is valid, the observed significance level, $\hat{\alpha}$, should be uniformly distributed between zero and one when no initialization bias is present. The observed significance level is the complementary distribution function for the test statistic, $\hat{h}$, evaluated at the observed value of the test statistic. If the statistic, $\hat{h}$, has the assumed distribution, then the fact that $F_{\hat{h}}(\hat{h})$ is uniform is familiar to many simulation users. (It is the basis for the "inversion method" of stochastic variate generation.)

For the procedure to be effective in detecting initialization bias the significance level should be small when initialization bias is present. The distribution function $F_{\hat{\alpha}}(\alpha)$ is called the "power function" of the test. The more powerful the test is in detecting negative bias the more rapidly $F_{\hat{\alpha}}(\alpha)$ will go to one as $\alpha$ increases from zero when negative bias is present.

In the examples that follow, several independent runs are made both with and without initialization bias. The results are presented in the form of the empirical distribution functions ("power functions") of the observed significance level.

### 4.1. Model 1: A Telephone Exchange

This model was used in a study of a data collection system for a telephone exchange (McDaniel [1979]). If the interarrival times between calls and the call durations are independent and exponentially distributed, the steady-state distribution of the number of calls in the exchange can be computed. Twenty-five independently seeded runs of the simulation were made with the system initially empty and 25 runs were made with the initial state randomly selected from the steady state distribution. Each run consisted of 200 call completions (a run length that is considerably shorter than that used in the actual study). The output, $Y_i$, is the number of calls in the exchange when the $i$th call is completed. The

transient mean behavior when the system was initially empty is like that in Figure 1-c. Empirical distributions of the significance level, $\hat{\alpha}$, appear in Figure 3. Figure 3-a, corresponding to the runs without initialization bias, shows that the test procedure is apparently valid ($\hat{\alpha}$ is approximately

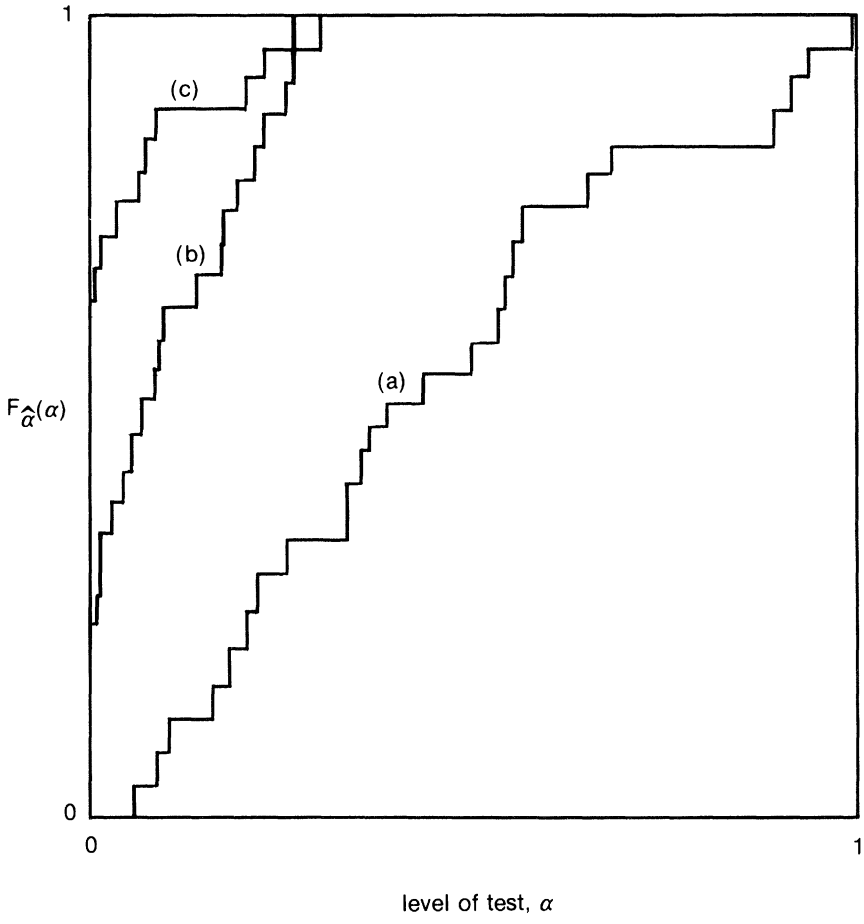


**Figure 3.** Empirical distribution functions of significance levels for a telephone exchange simulation. (*a*) System initially in steady state ($n = 200$), (*b*) system initially empty ($n = 200$), and (*c*) system initially empty ($n = 100$).

uniform). Figure 3-b, corresponding to the runs with initialization bias, indicates that the test was rather powerful in detecting the bias. For Figure 3-c, run lengths were halved for the biased model. The significance of the initialization bias increased (smaller $\hat{\alpha}$'s) for these shorter runs (as expected?).

Readers comfortable with a hypothesis testing framework may interpret Figure 3 (and Figures 4, 5, 6, and 8) as a plot of the probability of rejecting a hypothesis of no initialization bias as a function of the probability of rejecting the hypothesis when it is true, i.e., a plot of power as a function of the type I error level. From Figure 3-c, a test for a hypothetical constant mean at a 0.1 level would have rejected the hypothesis in 22 of the 25 runs or for 88% of the sample.

### 4.2. Model 2: An Inventory System

This simulation program models the inventory level in a warehouse area of fixed capacity used to store output from a simplified production facility. When the production line is working the rate of input to the storage area is twice the demand or outflow rate. When the warehouse is filled production stops. The production process is subject to random shutdowns according to a Markov process. Again the steady state distribution of the storage contents can be calculated for this model.

Fifty independently seeded runs were made starting with both the storage empty and the storage level randomly chosen according to its steady state distribution. A run consisted of 200 observations (here an observation was the daily average storage contents). The program "warmed up" very rapidly. The transient mean function is like that in Figure 1-a when the storage was initially empty. The empirical distribution functions of $\hat{\alpha}$ (presented in Figure 4) indicate that the test is apparently valid and quite effective in detecting initialization bias.

### 4.3. Model 3: A Computer Time Sharing System

The computer time-sharing simulated here is discussed in Adiri and Avi-Itzhak [1969]. A thorough simulation study of this system is present in Sargent [1979].

With this model two experiments were conducted each involving 50 independently seeded runs. Each run consisted of 200 successive job completions. Figure 5-a shows the empirical significance level distribution function for runs that included a warm-up period of 100 job completions before the 200 response time observations were collected. In this case a small remaining initialization bias is indicated. Figure 5-b shows the significance level distribution function corresponding to runs with no warm-up period. The measured initialization bias is more significant when no warm-up period is used (as expected?).

The indication of initialization bias in runs with a warm-up period of 100 jobs (Figure 5-a) demonstrates the power of this test; bias after the warm-up period was not *visually* apparent in the original output.

### 4.4. Model 4: An *M/M/1* Queue

For this set of experiments an $M/M/1$ queue simulation with a traffic intensity of 0.9 was used. Each run involved averaging 10 independently seeded replications of 1,000 observations of customer waiting times in the queue. The run length here is entirely too short for the transient effects

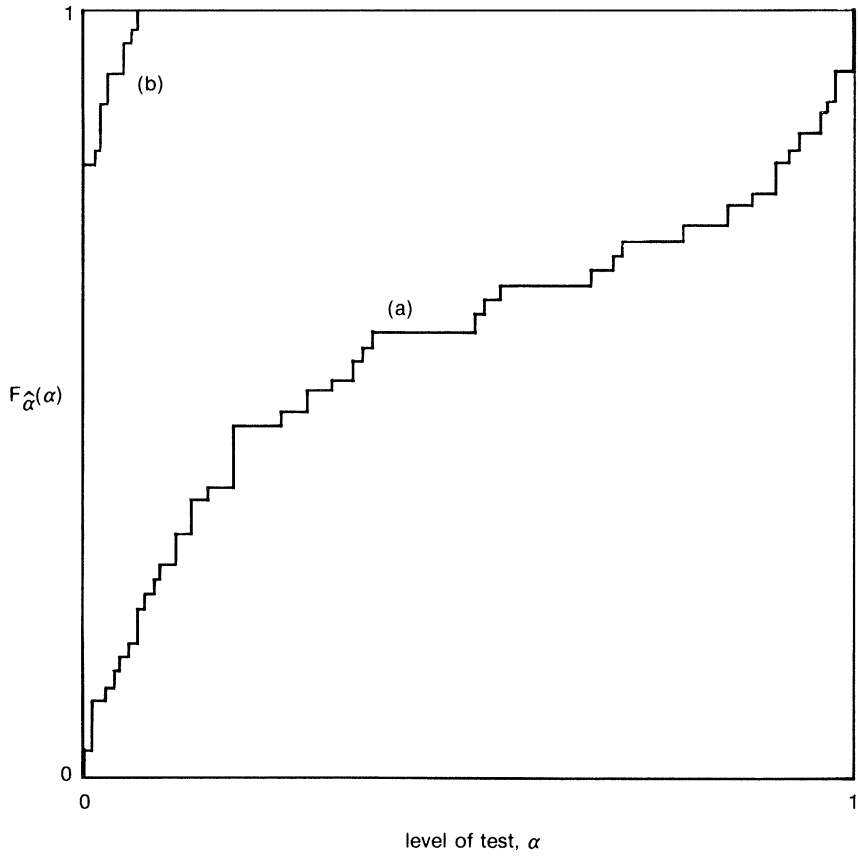


**Figure 4.** Empirical distribution functions of significance levels for a simulated inventory system. (*a*) System initially in a steady state and (*b*) system initially empty.

of initialization to dissipate (see Gafarian et al.). The transient mean function for the initially empty system is characterized by Figure 1-c.

A few comments about this example are in order. The average location of the maximum value in the test sequence in this experiment was near one-half of the run length (i.e. at $\hat{t} \doteq \frac{1}{2}$). There was also a tendency to severely overestimate $\sigma^2$ with the autoregressive estimator when the

output series contained a large change in the mean. Since the denominator of the test statistic is proportional to $(1 - \hat{t})\hat{t}\hat{\sigma}^2$, an overestimated $\sigma^2$ with $\hat{t}$ near ½ represents a "worst case" for the test presented here.

Figures 6-a and 6-b show the empirical distribution functions for the observed significance level when the initial state is sampled from the

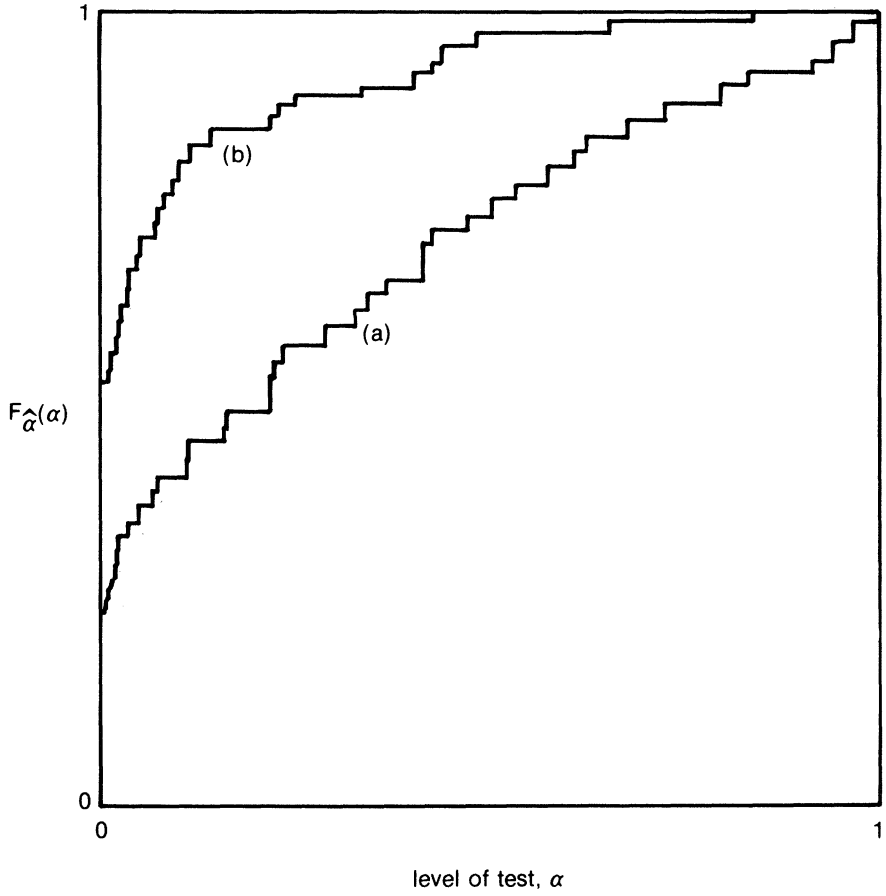


**Figure 5.** Empirical distribution functions of significance levels for a simulation of a computer time sharing system: (*a*) 100-job "warm-up" before data collection and (*b*) no "warm-up" before data collection.

steady state distribution and when the system is initially empty. Figure 6-a indicates that the test is applicable to this system. Figure 6-b shows the power of the test in detecting the initialization bias when the system is initially empty. Of hundreds of sets of experiments the result in Figure 6-b represents the worst performance for the test.

A third set of 50 experiments were run to produce Figure 6-c. The

initial state for these runs was twice the steady state mean. This induced a *positive* bias in the output. As before, the "one-sided" test for initialization bias was conducted to guard against potential *negative* bias. Figure 6-c gives a strong indication that bias of the "other sign" is present. Consistently large significance levels should prompt suspicion that ini-

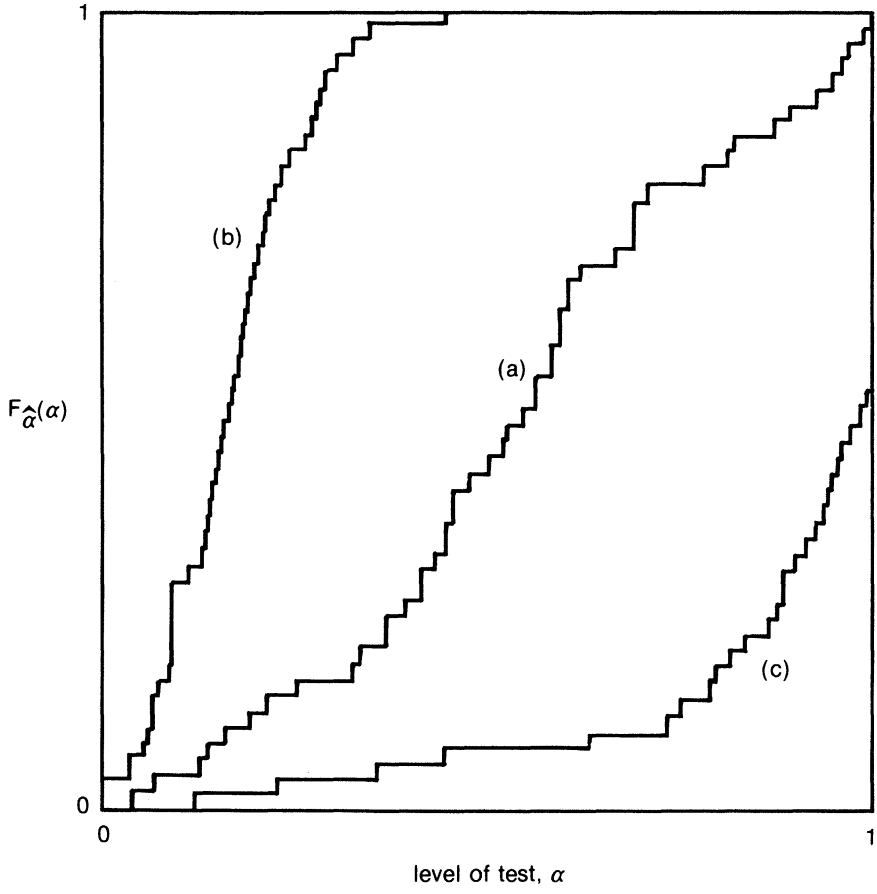


**Figure 6.** Empirical distribution functions of significance levels for a simulated $M/M/1$ queue ($\rho = 0.9$). (*a*) System initially in steady state, (*b*) system initially idle (negative bias), and (*c*) system initially highly congested (positive bias).

tialization bias (of the opposite sign anticipated) is present. An interesting situation where the wrong bias sign is tested is presented next.

### 4.5. Model 5: A Network of Queues

The system simulated in this set of experiments is a network of three

capacitated $M/M/s$ queues with feedback (blocked customers must re-enter the service queue just completed). Figure 7 is a schematic drawing of this system. Fifty runs were made with the system initially empty and 50 runs were made with the initial state randomly selected from a proposed steady-state distribution for the number of customers in the system. The output was the average number of customers in the system over 10 replications for 1,000 service completions.

The empirical distribution functions of the significance level for both these sets of experiments are presented in Figure 8. The test has almost perfect performance in detecting initialization bias (Figure 8-b); all values

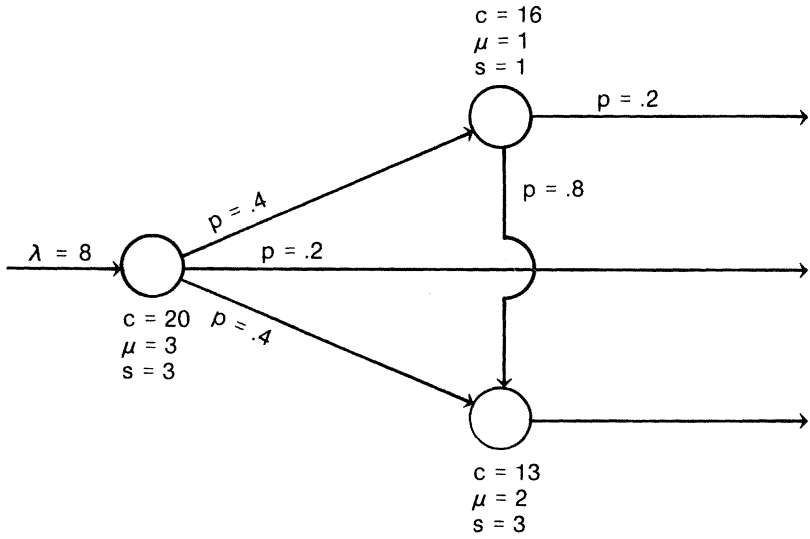


**Figure 7.** Queueing network for simulation experiment 5: $c$ = capacity, $s$ = number of servers, $\mu$ = individual server rate, $\lambda$ = arrival rate, and $p$ = probability that a departing customer follows a particular path.

of $\hat{\alpha}$ for the biased runs were zero to three decimal places. The significance level distribution for the "steady-state" runs (Figure 8-a) shows that large values of $\hat{\alpha}$ are likely. The initial state for these runs was selected according to an approximate steady-state distribution that overstates system congestion. The error in the proposed steady-state distribution for this queueing network was not recognized until after the simulation experiments were run. The results of the initialization bias test led to a more careful examination of the analytical solution technique and the error was discovered. This example illustrates another potentially useful application of the initialization bias test presented in this article: investigating proposed analytical solutions to stochastic models.

### 4.6. Summary of the Empirical Studies

The test for initialization bias presented in Section 3 was applied to many simulation models. Some of the results have been presented here. In all cases the test appeared to be valid and showed reasonable power in

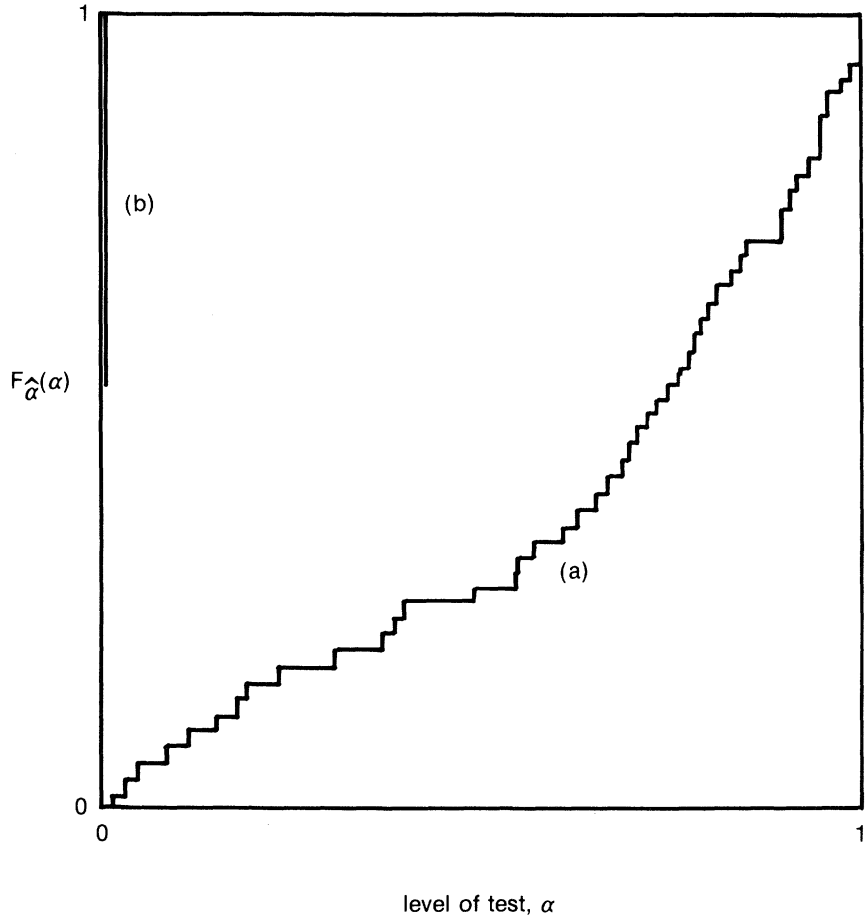


**Figure 8.** Empirical distribution functions of significance levels for a simulation of the queueing network in Figure 7. (*a*) System initially in "steady state" (see text) and (*b*) system initially empty.

detecting the presence of initialization bias. These experiments indicate that the asymptotic theory on which the test is based produces good results. Many of the simulation runs were unusually short. The results improve for longer runs, as expected from the theory.

The test presented here appears to be most useful when the transient

mean function is nearly constant or dissipates rapidly. The test is not particularly powerful when a large initial transient effect is present that remains throughout the run. A test based on the area (actually an optimally weighted sum) of the signal, $D_l$, is presented in Schruben et al. [1980]. That test has excellent power in large bias situations. Many of these large bias cases can be adequately identified by visual inspection of the simulation output. Extensive experiments were conducted involving variations in run duration, number of replications, output filtering, and method of variance estimation. The following recommendations are the result:

1. Approximately five or more independently seeded replications of the simulation should be run using the same initialization strategy. The test for initialization bias can be conducted for each replication or the total output can be averaged across replications and the resulting series tested.
2. Prior to testing, the output should be grouped into *small* adjacent, exclusive batches (say five observations per batch). The sequence of batch averages should be used in the test. This does not mask significant initialization bias and makes variance estimation easier.
3. If several replications are available $\sigma^2$ should be estimated using all of the runs *except* that run currently being tested for initialization bias. This allows justification of the (incorrectly) assumed independence of $\hat{\sigma}^2$ and $\hat{x}$. However, in the empirical studies presented in this paper, the same runs were used to obtain both $\hat{\sigma}^2$ and $\hat{x}$.
4. Plots of the output and of the test sequence should be compared to the patterns in Figures 1 and 2.
5. The computed significance level for the test is a guide and should not be substituted for intuition and common sense.

Some final comments concerning estimation of the limiting variance ($\sigma^2$): the autoregressive estimator used in the empirical studies presented here did *not* perform better than other estimators. This variance estimator was selected since it is applicable to experiments involving a single run and it is computationally convenient. An endorsement of the autoregressive method of variance estimation is not implied here. Variance estimation from simulation output is an active area of research. When better techniques are developed they should be used with the initialization bias testing approach presented here.

A modification of the test presented in this paper that does not require estimation of the scale parameter $\sigma^2$ is as follows:

*Step 1.* Compute $\hat{h}$ using the output from the first half of a run with $\hat{\sigma}^2 = 1$, call this $\hat{h}_f$.

*Step 2.* Compute $\hat{h}$ using the output from the last half of a run with $\hat{\sigma}^2 = 1$, call this $\hat{h}_l$.

*Step 3.* Set $\hat{\alpha} = \bar{F}_{3,3}(\hat{h}_f/\hat{h}_l)$.

Here we use the result that $3h\hat{\sigma}^2/\sigma^2$ has an approximate $X_3^2$ distribution as the run duration increases.

Independently seeded runs should be used for Steps 1 and 2 so that the statistics $\hat{h}_f$ and $\hat{h}_l$ are independent. However, preliminary testing indicates that this version of the test can perform reasonably well when the same run is used for both Steps 1 and 2. Like the test procedure presented in Section 3.3 the above procedure does not appear to have good bias detection power for short runs with severe initialization effects (again, these situations are probably easy to identify by a visual inspection of the output).

The procedure presented in the above paragraphs can be extended to estimation of $\sigma^2$ as one-third of the average of a sample of $\hat{h}$ statistics. Properties of this and other estimators of $\sigma^2$ based on weak convergence of partial sums to Brownian motion are currently under investigation. The techniques of bias detection presented in this paper and in Schruben et al. are representative of other potential applications of functional central limit theorems to the analysis of simulation output.

## 5. SUMMARY

An approach to testing for stationarity in the mean of a simulation output sequence has been presented. The basis for the technique is the asymptotic convergence of partial sums of deviations about the average to a Brownian bridge process. Tests for initialization bias in a simulation output series can be developed using this approach. One such test is presented in Section 3. This test appears both valid and powerful as indicated by the experiments in Section 4. In these experiments short run durations were used to illustrate the performance of the asymptotic theory in relatively short runs of a simulation program.

The test sequence $\{T_n(t)\}$ presented here can be used to detect initialization bias or to support arguments that initialization bias has been adequately controlled. The question remains: what if a simulation output sequence *fails* a test for initialization bias (say, $\hat{\alpha}$ is less than 0.1)? The experimenter may then want to consider one or both of the following two actions: truncate the output series, or increase the run length. Decisions of how much data to truncate or how long to run a program must be made. These problems are under investigation using the approach developed in this paper.

## ACKNOWLEDGMENT

The author wishes to thank David Heath of Cornell for his encouragement and assistance and Averill Law for his thoughtful comments and suggestions.

## REFERENCES

ADIRI, I., AND B. AVI-ITZHAK. 1969. A Time Sharing Queue with a Finite Number of Customers. *J. Assoc. Comput. Mach.* **16,** 315–323.

BILLINGSLEY, P. 1968. *Convergence of Probability Measures*, Ch. 20 and 21. John Wiley & Sons, New York. [Theorem 21.1 should have $E[\eta_0]$ changed to $E[\eta_0^2]$.]

FISHMAN, G. S. 1973. *Concepts and Methods in Discrete Event Digital Simulation.* John Wiley & Sons, New York. [Equation 8 of this paper differs from that in Table 32, page 289. This is probably not critical; see G. S. FISHMAN, *Principles of Discrete Event Simulation*, p. 252. Wiley, 1978.]

GAFARIAN, A. V., C. J. ANCKER, JR., AND T. MORISAKU. 1978. Evaluation of Commonly Used Rules for Detecting "Steady State" in Computer Simulation. *Naval Res. Logist. Quart.* **25,** 511–529.

JOHNSON, R. A., AND M. BAGSHAW. 1974. The Effect of Serial Correlation on the Performance of CUSUM Tests. *Technometrics* **16,** 103–112. [The last line of the Appendix should have $a_j$'s replaced by $X_j$'s.]

KELTON, W. D. 1980. The Startup Problem in Discrete-Event Simulation, unpublished Ph.D. dissertation (Industrial Engineering), University of Wisconsin, Madison, Wisconsin.

LAW, A. M. 1977. Confidence Intervals in Discrete Event Digital Simulation: A Comparison of Replication and Batch Means. *Naval Res. Log. Quart.* **24,** 667–678.

MCDANIEL, M. 1979. Evaluation of CGS Measurements in ETN-Case 49427-42, Bell Laboratories Memorandum, Bell Telephone Laboratories Inc., Holmdel, N.J.

MECHANIC, H., AND W. MCKAY. 1966. Confidence Intervals for Averages of Dependent Data in Simulations II, IBM Technical Report, 17–202, Yorktown Heights, N.Y.

SARGENT, R. G. 1979. An Introduction to Statistical Analysis of Simulation Output Data. In *Proceedings of the AGARD Symposium*, Paris, France.

SCHRUBEN, L. 1980. Detecting Initialization Bias in Simulation Output, Tech. Report 444, School of O.R. & I.E., Cornell University, Ithaca, N.Y. 14853.

SCHRUBEN, L. 1981. Control of Initialization Bias in Multivariate Simulation Response. *Commun. ACM* **24,** 246–252.

SCHRUBEN, L., H. SINGH AND L. TIERNEY. 1980. A Test of Initialization Bias Hypotheses in Simulation Output, Tech. Report 471, School of O.R. & I.E., Cornell University, Ithaca, N.Y. 14853.

SHEPP, L. A. 1979. The Joint Density of the Maximum and Its Location for a Wiener Process with Drift. *J. Appl. Prob.* **16,** 423–427.

WILSON, J. R., AND A. A. B. PRITSKER. 1978. A Survey of Research on the Simulation Start-up Problem. *Simulation* **31,** 55–58.