Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Tests for Initialization Bias in Simulation Output

L. Schruben, H. Singh, L. Tierney,

Optimal Tests for Initialization Bias in Simulation Output

L. SCHRUBEN and H. SINGH
Cornell University, Ithaca, New York

L. TIERNEY
Carnegie-Mellon University, Pittsburgh, Pennsylvania

(Received August 1981; accepted July 1983)

We present a family of tests for detecting initialization bias in the mean of a simulation output series using a hypothesis testing framework. The null hypothesis is that the output mean does not change throughout the simulation run. The alternative hypothesis specifies a general transient mean function. The tests are asymptotically optimal based on cumulative sums of deviations about the sample mean. A particular test in this family is applied to a variety of simulation models. The test requires very modest computation and appears to be both robust and powerful.

THE PROBLEM of testing for a change in the mean of a stochastic process is particularly important in simulation studies. The initial state of the simulated system must be specified each time the program is run. It is often inconvenient or impossible to insure that these initial conditions are typical. An unusual sequence of events might be induced by initializing a simulation in a particular way. As a result, estimation of the equilibrium simulation response is complicated by the possible presence of *initialization bias.* In studies of the steady state characteristics of a simulation model, it is important to identify initialization bias and to evaluate efforts to control this problem.

The usual approach to initializing a simulation is to start the program in some convenient fashion and allow an arbitrary "warm up" period to pass before retaining data for analysis. This procedure is known as output truncation. Researchers have proposed many methods for selecting a warm up period (see the survey in Wilson and Pritsker [1978]). These techniques are often elaborate and give no assurance that initialization bias will be effectively controlled. When the simulation is expensive to run, some measure of the initialization effect might be desired before using data truncation. Throwing out a large amount of the data collected

Subject classification. 767 statistical analysis.

Operations Research
Vol. 31, No. 6, November–December 1983
0030-364X/83/3106-1167 $01.25

at the beginning of each run may be wasteful. Perhaps more important, retaining biased data can be misleading. For example, the frequency that confidence intervals using biased data will cover an unknown parameter generally *decreases* as more replications of the simulation are run. This phenomenon arises because the intervals shrink about a biased point estimate.

This paper formulates the problem of detecting the presence of initialization bias using statistical hypotheses concerning the mean of the output process. Classical statistical theory (likelihood ratios) is used to derive the form of the optimal test. We test whether the observed output is consistent with a *null hypothesis* that the mean of the process did not change throughout a run. The optimal test against a specified alternative transient mean function is presented. The paper does not consider second order effects of initialization, such as a change in the process variance.

The tests presented here are based on theory presented in Schruben [1982, 1983]. The test in Schruben [1982] is appropriate if the experimenter is unwilling to say anything about possible forms of initial transient behavior in the simulation. When particular initial transient functions are specified as alternatives to the null hypothesis, more powerful tests than the test in Schruben [1982] are possible.

We suggest a particular test that is most powerful against a simple quadratic transient mean function, and examine the performance of this test using some artificial test data and some real simulations.

Often the presence of initialization bias is obvious from a visual inspection of the simulation output. With this in mind, we anticipate that the primary utility of the test proposed here will be in testing for remaining bias in simulation output that has been truncated. Not rejecting the null hypothesis supports the assertion that initialization bias in the output process mean has been effectively controlled. It is important to recognize that we are testing for consistency of the null hypothesis with the actual output. Like all statistical tests, rejection of a false hypothesis is unlikely if the sample (here the run length) is very small. The applications presented in Section 2 all involve relatively short runs. These examples indicate "small sample" robustness of the asymptotic theory.

1. OPTIMAL TESTS FOR INITIALIZATION BIAS

Consider a simulation output sequence, $Y_1, \cdots, Y_n$. The expected value of Y_i is given by

$$\mu_i = \mu(1 - a_i). \tag{1}$$

The function μ_i is called the transient mean function. The a_i's can be chosen to reflect more or less arbitrary behavior in the mean of the simulated process. When the simulated process is asymptotically stationary then $\lim_{i\to\infty} a_i = 0$, and the a_i's represent changes in the output mean due to initializing the simulation program. When there is no initialization effect in the output mean, then $a_i = 0$ for all i. The value of the steady-state mean, μ, is unknown. Formally, the null hypothesis is that there is no initialization bias in the output mean, i.e.

$$H_0\colon a_i = 0 \quad \text{for all} \quad i.$$

The alternative hypothesis is,

$$H_1\colon a_i = \text{an arbitrary (specified) function of } i.$$

As pointed out in Schruben [1982] the behavior of the cumulative sum process, $S_k = \bar{Y}_n - \bar{Y}_k$, with $\bar{Y}_k = 1/k \sum_{i=1}^{k} Y_i$, and $S_0 = S_n = 0$, is highly sensitive to the presence of nonstationarity in the mean of the output. We will design a test for initialization bias using the $\{S_k\}$ process. In this section we show how to obtain a test statistic that is optimal in a particular setting and can thus be expected to perform reasonably well in similar settings.

We assume under the null hypothesis that the output series is stationary and satisfies some regularity conditions requiring, essentially, that two observations that are far apart in time are approximately independent. Theorem 21.1 of Billingsley [1968] then implies that under H_0, the process

$$B_t = ([nt]S_{[nt]})/\sqrt{n}; \qquad 0 \le t \le 1, \tag{2}$$

(with [.] denoting the greatest integer function) converges weakly as $n \to \infty$ to a Brownian bridge process, $\{\mathscr{B}_t\}$, with variance $\sigma^2 = \lim_{n\to\infty} n \operatorname{Var}(\bar{Y}_n)$. A Brownian bridge process is a zero mean Gaussian process on the unit interval with continuous paths and covariance function, $\operatorname{Cov}(\mathscr{B}_r, \mathscr{B}_t) = \sigma^2(\min(r, t) - rt)$. Further theoretical details are presented in Schruben [1983].

If the series $\{Y_i\}$ is not stationary but can be written as

$$Y_i = X_i + \mu_i$$

for some zero mean stationary series $\{X_i\}$ that satisfies the regularity conditions mentioned above, then $\{B_t\}$ converges weakly to $\{\mathscr{B}_t\}$ plus an unknown deterministic mean function,

$$(1/n) \textstyle\sum_{i=1}^{n} \mu_i - (1/k) \sum_{i=1}^{k} \mu_i.$$

Instead of working directly with the series $\{Y_i\}$, consider the series, $\{\hat{Y}_i, i = 1, \cdots, n\}$ of uncorrelated normal random variables with means equal to μ_i and variance, σ^2, as defined earlier.

The series $\{\hat{Y}_i\}$ for our purposes is asymptotically equivalent to the series $\{Y_i\}$ and is easier to work with. That is, $\{S_k\}$ has the same limiting distribution when defined for the series $\{\hat{Y}_i\}$ as for the series $\{Y_i\}$. A test that performs optimally for the series $\{\hat{Y}_i\}$ should perform well for the original series if the run duration is large enough.

Fairly long runs are typical in simulations, particularly if initialization bias is suspected. One might reasonably expect that theory based on *long run* lengths would produce results comparable to theory based on, say, a large number of *replications*. Simulation practitioners are familiar with asymptotic theory requiring a large number of replications as the basis for many successful applications of the classical (scalar) central limit theorem in the analysis of simulation output. The (process) central limit theorem we use here appears to have "small sample" (short run) robustness similar to the more familiar (scalar) central limit theorem.

To derive an optimal test for bias based on $\{S_k\}$ (based on the series $\{\hat{Y}_i\}$), assume for the moment that we know that the *sign* of the steady state mean, μ, is not negative, but that its magnitude remains unknown. Also assume, for the moment, that σ^2 is known. Under these assumptions, we can derive the most powerful test of the null hypothesis of no bias in the output mean against any specified alternative hypothesis (the a_i's) using the Neyman-Pearson Lemma.

The joint density of $\{S_k, k = 1, \cdots, n-1\}$ is equal to

$$f(S_1, \cdots, S_{n-1}; a_1, \cdots, a_n) = n!/((2\Pi\sigma^2)^{(n-1)/2}\sqrt{n})\cdot$$
$$\exp\{(\textstyle\sum_{i=1}^n \mu_i)^2/(2n\sigma^2) - (1/2\sigma^2)\,[\sum_{i=1}^n ((i-1)S_{i-1} - iS_i - \mu_i)^2]\}.$$

This result can be derived using the equality, $\hat{Y}_i = (i-1)S_{i-1} - iS_i + \bar{\hat{Y}}_n$ to transform the density of $(\hat{Y}_1, \cdots, \hat{Y}_n)$ into the joint density of $(S_1, \cdots, S_{n-1}, \bar{\hat{Y}}_n)$. The density of the series average, $\bar{\hat{Y}}_n$, which is independent of $(S_1, \cdots, S_{n-1})$, is then factored out of this joint density. The likelihood ratio for these two hypotheses is therefore equal to

$$\begin{aligned} f(S_1, \cdots, S_{n-1}; a_1, \cdots, a_n)&/f(S_1, \cdots, S_{n-1}; 0, \cdots, 0)\\ &= \exp\{1/\sigma^2 \textstyle\sum_{i=1}^n ((i-1)S_{i-1} - iS_i)\mu_i\\ &\qquad - 1/2\sigma^2\,(\textstyle\sum_{i=1}^n \mu_i^2 - (1/n)\,(\sum_{i=1}^n \mu_i)^2)\}\\ &= \exp\{\mu/\sigma^2 \textstyle\sum_{i=1}^n ((i-1)S_{i-1} - iS_i)(1 - a_i)\\ &\qquad - \mu^2/2\sigma^2\,(\textstyle\sum_{i=1}^n a_i^2 - (1/n)\,(\sum_{i=1}^n \alpha_i)^2)\}.\end{aligned}$$

For any $\mu > 0$, this expression is an increasing function of

$$\begin{aligned} T &= \textstyle\sum_{i=1}^n ((i-1)S_{i-1} - iS_i)(1 - a_i)\\ &= \textstyle\sum_{i=1}^n iS_i(a_i - a_{i+1}).\end{aligned}$$

The second identity above is found by summation by parts. Thus (see, e.g. Bickel and Doksum [1977]), the most powerful test of H_0 versus H_1 based on the sequence $\{S_k\}$ is to reject H_0 when the statistic T is too large.

The critical rejection region for the test is determined by using the fact that (under H_0) the iS_i are jointly normal with zero mean and covariance function,

$$\mathrm{Cov}(iS_i, jS_j) = \sigma^2 n(\min(i/n, j/n) - i/n \cdot j/n).$$

Therefore, T being the sum of normal random variables, is also normally distributed with zero mean and variance

$$\sigma^2 \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} n(\min(i/n, j/n) - i/n \cdot j/n)(a_i - a_{i+1})(a_j - a_{j+1}).$$

To summarize: The optimal test statistic for

$$H_0\colon \mu_i = \mu \quad \text{for all} \quad i \quad \text{against} \quad H_1\colon \mu_i = \mu(1 - a_i)$$

is a weighted average of the sequence $\{kS_k\}$ with weights equal to $a_k - a_{k+1}$. That is,

$$T = \textstyle\sum_{k=1}^{n} c_k k S_k, \quad \text{with} \quad c_k = a_k - a_{k+1}$$

is the most powerful test statistic based on $\{S_k\}$.

We have derived the form of the asymptotically most powerful test based on partial sums of deviations for any suspected transient mean function. In practice (a_i's unknown), we can not use the optimal weights, c_k. We do know that in simulations of stationary processes $c_k \to 0$ since $\mu_i \to \mu$. In addition, if bias of a particular sign is suspected, then the signs of the weights can be appropriately chosen.

If the initial conditions might induce a negative bias ($E[\bar{Y}_n] < \mu$) in the output, then we suspect that,

$$\mu_i - \mu < 0 \Rightarrow \mu(1 - a_i) - \mu < 0 \Rightarrow a_i > 0$$

and the weights will be positive. If initialization of the simulation can potentially introduce a positive bias, then the weights should be negative. The sign of potential initialization bias is often known in practice, e.g. some queueing simulations may most conveniently be initialized in an undercongested state.

Any weighting with decreasing magnitude and the appropriate sign is optimal against some initial transient mean function and should perform reasonably well against similar transients. In the simulation examples presented in Section 2, we use the weighting,

$$c_k = 1 - (k/n). \tag{3}$$

This arbitrary weighting is optimal against a simple quadratic transient mean function with $a_k = 1/(2n)(k^2 - k(2n + 1))$ plus a constant ($a_k = 0$ if $k \gg n$). The test performed well in all the examples we studied.

The test recommended in this paper is a member of a family of optimal tests. An experimenter with more knowledge of suspected transient mean behavior may choose a more appropriate test (i.e., select an optimal weighting c_k) from this family. Note that in some situations the form of a suspected initial transient is known. For example, in simple queueing simulations frequently the transient mean function is exponential. For these cases, experimenters should use optimally weighted test statistics. A situation, not directly related to simulation initialization, where the transient mean function might also be known would be in applying the test to industrial process control. Here the mechanism for the process going out of control (e.g. drift or wear) might be known from the physical machine characteristics. Application of the above test to industrial quality control problems appears to be straightforward.

When the weights (3) are used, the distribution of the test statistic under the null hypothesis (no bias in the mean) is normal with zero mean and variance,

$$\begin{aligned}\operatorname{Var}(T) &= \sigma^2 \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} n(\min(i/n,\, j/n) - i/n \cdot j/n)(1 - i/n)(1 - j/n) \\ &\approx n^3\sigma^2 \int_0^1 \int_0^1 (\min(s,\, t) - st)(1 - s)(1 - t)\,ds\,dt \\ &= n^3\sigma^2 2 \int_0^1 \int_0^t s(1 - s)(1 - t)^2\,ds\,dt \\ &= n^3\sigma^2/45.\end{aligned}$$

The above approximation uses the change of variables $i = ns$ and $j = nt$, so $didj = n^2dsdt$.

A remaining problem in applying the test developed above is that the scaling constant, σ^2, is generally unknown and must be estimated from simulated data. A number of estimators are available in the simulation methodology literature. Table 32 of Fishman [1973] presents formulas for some of the more established methods. Note that the table refers to σ^2 as m. The examples presented in the next section both used the "autoregressive" estimator and the "batched means" estimator. For the "batched means" estimator, our experiments used an arbitrary batch size of 50 observations. The test performance was essentially identical for these two estimators of σ^2. Both estimators are suitable when only the output from a single run is available.

We denote the estimator of σ^2 by $\hat{\sigma}^2$. The quantity $d\hat{\sigma}^2/\sigma^2$ can be assumed to have an asymptotic χ^2 distribution with d degrees of freedom (the appropriate degrees of freedom for each estimator of σ^2 is also presented in Table 32 of Fishman's book).

Using the weighting (3), we have the test statistic, T, which has an approximate normal distribution with zero mean and variance given by (4). If we divide T by the square root of the approximate $\chi_d{}^2$ statistic, $d\hat{\sigma}^2/\sigma^2$, over its degrees of freedom, d, we can partially justify treating,

$$\hat{T} = (\sqrt{45}/n^{3/2}\hat{\sigma}) \sum_{k=1}^{n} (1 - k/n)k(\bar{Y}_n - \bar{Y}_k) \tag{5}$$

as a t statistic with d degrees of freedom. To justify treating $\hat{T}$ as a t statistic, we would need to estimate σ^2 using *independently* seeded replicate runs of the simulation. However, the applications in Section 2 computed both $\hat{T}$ and $\hat{\sigma}^2$ from the same output. The results indicate that treating $\hat{T}$ as a t statistic appears to be reasonable.

Some final comments on estimation of σ^2: All the estimators mentioned earlier are designed for stationary processes. The presence of initialization bias will tend to inflate these estimators and thus reduce the power of the test for initialization bias because the test statistic will be scaled (divided) by too large a value. It is therefore advisable to use only the latter portion of an output series, say the last half, for estimating σ^2. When the initialization effects are moderate, the amount of bias in the latter portion of the output can be expected to be substantially smaller than the amount of bias in the earlier observations. The examples of the next section used the autoregressive estimator based on roughly the last half of each series to estimate σ^2. Estimation of σ^2 is an active area of simulation methodology research. When better estimators of σ^2 are developed these should be used with the test presented in this paper. A more detailed discussion of estimation of the scale parameter, σ^2, is presented in Schruben [1982].

Using the results developed so far, the test for the presence of negative initialization bias at a significance level of α is as follows:

Step 1. Compute $\hat{\sigma}^2$ and d using the autoregressive method or batched means method in Table 32 of Fishman.

Step 2. Compute the test statistic, $\hat{T}$, using (5).

Step 3. Reject the hypothesis of no *negative* initialization bias if $\hat{T} > t(d, \alpha)$. Here $t(d, \alpha)$ is the upper 100α-quantile of the t distribution with d degrees of freedom.

If a test for *positive* initialization bias is desired, then the sign of the test statistic $\hat{T}$ should be changed in Step 2. A two-sided test for initialization bias can be computed in the usual manner for t tests by using the absolute value of the test statistic $\hat{T}$.

2. SOME EXAMPLES OF TEST APPLICATION

Our experiments applied the procedure suggested in Section 1 for testing for negative initialization bias to several simulation models. In each case, we ran two sets of experiments, one "biased" set with an initial transient mean and one "unbiased" set that incorporates an attempt to remove the initialization bias. These models are briefly described below. Schruben et al. [1980] present more complete descriptions of the experiments.

Model 1. This model is an artificial output series with an oscillating initial transient mean function. The series was generated using the second order autoregression,

$$e_i = Y_i - 0.75Y_{i-1} + 0.5Y_{i-2},$$

where $\{e_i\}$ is a sequence of uncorrelated standard normal random variables. To generate "initialization bias" for the biased runs, we added the linearly increasing function

$$b_i = \begin{cases} (i-30)/30 & i \leq 30 \\ 0 & i > 30 \end{cases}$$

to the series, $\{Y_i\}$. Each run consisted of 500 observations. Fifty runs each for the biased time series and the unbiased time series were made.

Model 2. This set of experiments simulated an $M/M/1$ queue with a traffic intensity of 0.9. The output series consisted of the waiting times for successive customers. Each "run" involved averaging across 10 replications of 500 service completions. In this situation, the biased runs were made with the system starting empty, and the unbiased runs were initiated by sampling from the known steady-state distribution for the number of customers in the system. The initial transient mean function here is an exponential which for these short run durations can be closely approximated by a quadratic function similar to the transient mean function for which the test is optimal.

Model 3. These experiments simulated a simple production-inventory model. This production facility is subject to random failures according to a Markov chain. When operational, the facility produces parts at a rate of 2 parts per time unit. The probability of production process failure is set at 0.2 and the probability of a failed production process resuming operation in a time unit was set at 0.7. The finished parts are placed in a storage area of fixed capacity. When the storage is full, production stops. Demand for parts from storage was at a rate of 1 part per time unit. The output series was the average daily contents of the storage area for runs of 200 days. The biased experiments were initialized with the storage empty and the unbiased runs were initiated with the storage contents chosen from the known steady-state distribution for this system. This simulation warmed up very rapidly and the short initial transient was often difficult to detect visually.

Model 4. This set of experiments simulated the time shared computer system described in the paper by Adiri and Avi-Itzhak [1969]. Each run produced 300 job response times. Fifty biased runs were started with no jobs being processed by the CPU. Fifty "unbiased" runs were made by allowing the simulation to warm up for 100 job completions before the 300 job response times were recorded. No initialization bias was visually apparent in output from the runs that included the 100 job warm up period.

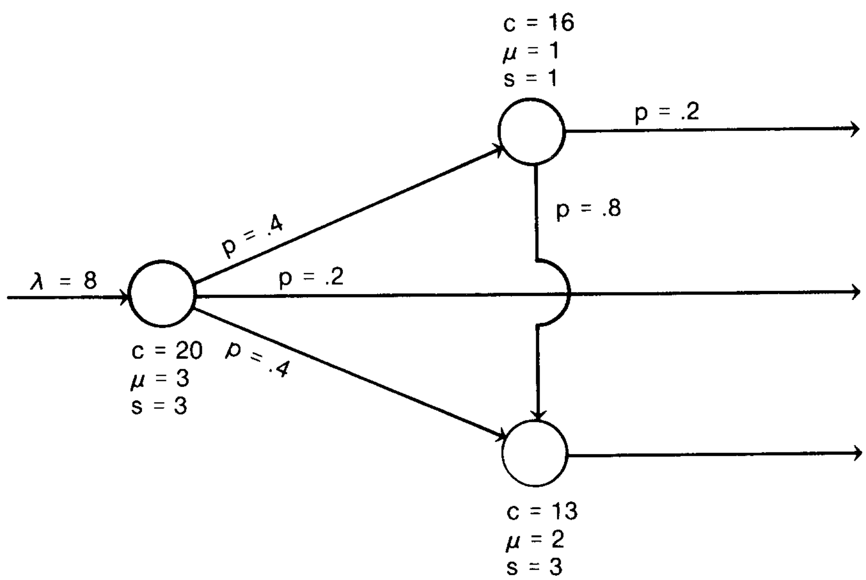

Figure 1. Queueing network for simulation Experiment 6. c = capacity, s = number of servers, μ = individual server rate, λ = arrival rate, and p = probability that a departing customer follows a particular path.

Model 5. The simulation used for this set of experiments is that of a telephone exchange model (McDaniel [1979]). This model is basically an $M/M/S$ queueing system with no waiting line capacity. Each customer (a call) that enters the exchange occupies two servers (an input line and an output line). The simulation output was the number of lines busy, sampled at fixed intervals. Each run produced 100 observations, a run duration that was considerably shorter than that used in the actual study. The biased runs were initialized with the exchange empty and the unbiased runs were initialized with the number of calls in the exchange drawn from the known steady-state distribution.

Model 6. The system simulated in this set of experiments is a network of three capacitated $M/M/S$ queues with feedback (blocked customers must reenter the service queue just completed). Figure 1 is a schematic drawing of this system. Fifty biased runs were made with the system initially empty and fifty "unbiased" runs were made with the initial state randomly selected from an approximate steady-state distribution for the number of customers in each of the queues. This approximate steady-state distribution slightly overstates the congestion in the sytem (a positive bias). See Schruben [1982] for a discussion of testing for negative bias when a slight positive bias is present.

TABLE I
PERFORMANCE OF INITIALIZATION BIAS TESTS[a]

Simulation Model	"Unbiased" Runs[b]: Test presented in this paper	"Unbiased" Runs[b]: Test presented in Schruben [1982a]	Biased Runs: Test presented in this paper	Biased Runs: Test presented in Schruben [1982a]
1	12	10	70	60
2	4	5	95	30
3	10	11	95	98
4	11	12	50	72
5	9	4	72	88
6	7	6	100	100

[a] Percent of runs for which the null hypothesis was rejected at the 10% level.
[b] See text.

Table I presents the results of testing for initialization bias in these six simulation models at the 0.1 level. For comparison, the table includes the performance of the test in Schruben [1982]. The References report complete empirical power functions for all significance levels.

In all the simulations studied, the tests appear to be valid; observed significance levels had no serious departures from uniformity for the unbiased runs. The test presented in this paper had high power in detecting the presence of initialization bias in all of the test models. In particular, the performance of the test with the $M/M/1$ queue (Model 2) was good in light of the fact that it is difficult to detect initialization bias with this model. Model 2 was the only case in which the test in Schruben [1982] did poorly. The tests presented in this paper are optimal in detecting *specified* initial transients. The test used in the examples is optimal *only* against transients corresponding to the weighting in Expression 3. The computation involved in the two tests was minimal even on the smallest of computers (a Heathkit H89 home computer was used for some of the experiments).

3. SUMMARY

The test for initialization bias in simulations presented in this paper appears to be robust and powerful in widely differing situations. The forms of the initial transient mean function in the models used in the examples are quite different. Relatively short runs were used in the experiments to illustrate the robustness of the asymptotic theory on which the test is based.

We recommended that both the test presented here and the test presented in Schruben [1982] be conducted to support assertions that initialization bias in the mean of a simulation output process has been effectively controlled. If several replications of the simulation are run, we recommend that each output sequence generated be tested for bias prior to analysis. When the output fails to pass the tests, the experimenter is advised to either increase the run duration or increase the warm up period, or both. Methods for selecting a run duration and a warm up period are currently under investigation using the weak convergence theorem for the sequence $\{S_k\}$ presented in this paper.

ACKNOWLEDGMENT

The authors appreciate comments concerning this paper made by James R. Wilson. This research was partially supported by Office of Naval Research Grant N00014-81-K-0037.

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.

Bickel, P. J., and K. A. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*, pp. 209–210. Holden-Day, San Francisco.

Billingsley, P. 1968. *Convergence of Probability Measures*, Chap. 20 and 21. John Wiley & Sons, New York.

Fishman, G. S. 1973. *Concepts and Methods in Discrete Event Digital Simulation,* Table 32, p. 289. John Wiley & Sons, New York.

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

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.

SCHRUBEN, L. 1982. Detecting Initialization Bias in Simulation Output. *Opns. Res.* **30,** 569–590.

SCHRUBEN, L. 1983. Confidence Interval Estimate from Standardized Simulation Output. *Opns. Res.* **31,** 1090–1108.

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