Technical Report CS82005-R*

VALIDATION OF MULTIVARIATE RESPONSE TRACE-DRIVEN SIMULATION MODELS

by

Osman Balci

and

Robert G. Sargent**

Department of Computer Science
Virginia Polytechnic Institute & State University
Blacksburg, Virginia 24061

April 1982

---

*Cross listed as Working Paper #82-002 in the Department of Industrial Engineering and Operations Research, Syracuse University, Syracuse, New York 13210

**Department of Industrial Engineering and Operations Research, Syracuse University, Syracuse, New York 13210

ABSTRACT

A procedure is developed by using Hotelling's one-sample T^2 test to test the validity of a multivariate response trace-driven simulation model that represents an observable system. The validity of the simulation model is tested with respect to the mean behavior under a given experimental frame.

A procedure for cost-risk analysis for the one-sample T^2 test is developed. By using this procedure, a trade-off analysis can be performed and judgement decisions can be made as to what data collection budget to allocate, what data collection method to use, how many paired observations to collect on the model and system response variables, and what model builder's risk to choose for testing the validity under a satisfactory model user's risk.

The procedure for validation and the cost-risk analysis are illustrated for a trace-driven simulation model that represents a time-sharing computer system with two performance measures of interest.

Index Terms: Validation, Trace-Driven Modelling, Simulation, Cost-Risk, Statistical Testing.

1. INTRODUCTION

A common problem encountered in computer system simulation is that of determining whether the representation of the computerized model is sufficiently accurate for the purpose for which the model is to be used [4]. "Substantiation that a computerized (simulation) model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model" is usually referred to as (simulation) model validation [20] and is the definition used in this paper.

A simulation model should be developed for a specific purpose or application and its adequacy or validity should be evaluated only in terms of that purpose with regard to experimental frame(s). As defined by Zeigler [24], an experimental frame, "... characterizes a limited set of circumstances under which the real system is to be observed or experimented with." A model may be valid in one experimental frame but invalid in another. Hence, the validity of the model should only be tested with respect to a set of experimental frames determined by the purpose for which the model is intended, and not for all possible experimental frames (or all sets of conditions) [18, 19].

The validity of a simulation model is tested under a given experimental frame and for an acceptable range of accuracy related to the purpose for which the model is intended. The acceptable range of accuracy is the amount of accuracy that is required for the simu-

lation model to be valid under a given experimental frame. The range of accuracy or the amount of agreement between the simulation model and the system is measured by a validity measure [3, 5]. The acceptable range of accuracy determines a range of the validity measure and this range is called an acceptable validity range [3, 5].

It is generally preferable to use some form of objective analysis to perform model validation. A common form of objective analysis for validating simulation models is statistical hypothesis testing [4]. In using a statistical test for validation, one should consider the type of the simulation model with regard to the way it is driven and with regard to the way its output is analyzed.

There are basically two types of simulation models with regard to the way they are driven: self- and trace-driven simulation models. Self-driven (distribution-driven or Monte Carlo) simulation [12] is a technique which uses random numbers in sampling from distributions or stochastic processes. Trace-driven (or retrospective [16]) simulation is a technique which combines measurement and simulation by using the actual data collected on the system as the model input [12, 23].

There are basically two types of simulation models with regard to analysis of the output: steady-state and terminating simulation models [10, 14]. A steady-state simulation "is one for which the quantity of interest is defined as a limit as the length of the simulation goes to infinity" [14]. A terminating simulation "is one for

which any quantities of interest are defined relative to the interval of simulated time $[0,T_E]$, where T_E, a possibly degenerate random variable, is the time that a specified event E occurs" [14].

The purpose of this paper is to give a procedure for validating a multivariate response trace-driven steady-state or terminating simulation model with respect to its mean behavior by using Hotelling's one-sample T^2 test [15] and the methodology for cost-risk analysis given in [5]. Cost-risk analysis for the one-sample T^2 test is presented in section 2, and the assumptions underlying the test together with some remedial measures are given in section 3. The procedure for validation is given in section 4 and is illustrated in section 5 for a simulation model of a time-shared computer system. Finally, conclusions are given in section 6.

2. COST-RISK ANALYSIS FOR THE ONE-SAMPLE T^2 TEST

In using Hotelling's one-sample T^2 test to test the validity of a trace-driven simulation model under a given experimental frame and for an acceptable range of accuracy consistent with the intended application of the model, we have the following hypotheses [5]:

H_0: Model is valid for the acceptable range of accuracy under the experimental frame.

H_1: Model is invalid for the acceptable range of accuracy under the experimental frame.

There are two possibilities for making a wrong decision in using the one-sample T^2 test for the purpose of validating a trace-driven simulation model. The first one, type I error, is rejecting the validity of the model when it is actually valid, and the second one, type II error, is accepting the validity of the model when it is actually invalid. The probability of making the first type of wrong decision is called model builder's risk (α) and the probability of making the second type of wrong decision is called model user's risk (β) [5].

These risks can be decreased at the expense of increasing the sample sizes of observations. However, increasing the sample sizes will increase the cost of data collection. Therefore, schedules and graphs can be constructed by following the methodology developed in [5] and the relationships among the model builder's risk, model user's risk, acceptable validity range, sample sizes, and cost of data collection can be determined. The model sponsor, model user, and model builder, individually or together, can examine the cost-risk trade-offs by using the schedules and graphs and can make judgement decisions as to what risks to take, what data collection budget to allocate, what method to use to measure the system and collect the trace data, and how many observations to collect on each of the model and system response variables.

The methodology in [5] will now be followed to develop a procedure for the one-sample T^2 test to construct the schedules. Step

1 of the methodology requires the determination of an appropriate statistical test for testing the validity of a multivariate response trace-driven simulation model.

A diagrammatic concept of a multivariate response trace-driven simulation model is given in Figure 1. A trace-driven simulation model is driven by using input sequences that are extracted from trace data rather than from a random number generator. The trace data is a stream of significant events that are observed in a real operational system and recorded with the time of their occurrences. One trace-data sequence is one observation (that is, one realization) from the ensemble (or the sample space) of all possible trace sequences [12]. The trace-data sequences must be collected in such a way that they are independent and identically distributed (iid) so that the system response variables and the model response variables will each have iid observations.

The trace-driven modeling technique which is usually used for computer performance evaluation has four distinct stages [12, 23]: (1) the real system is measured and the raw trace data is collected by a monitor in a real system, (2) the raw trace data is refined by a trace analysis program until it is suitable as input to the model and is analyzed to produce performance metrics, (3) the refined trace data is input to the trace-driven simulator which interprets the input data in a deterministic manner, and (4) the trace-driven model is validated. The four distinct components of trace-driven modeling are illustrated in Figure 1.

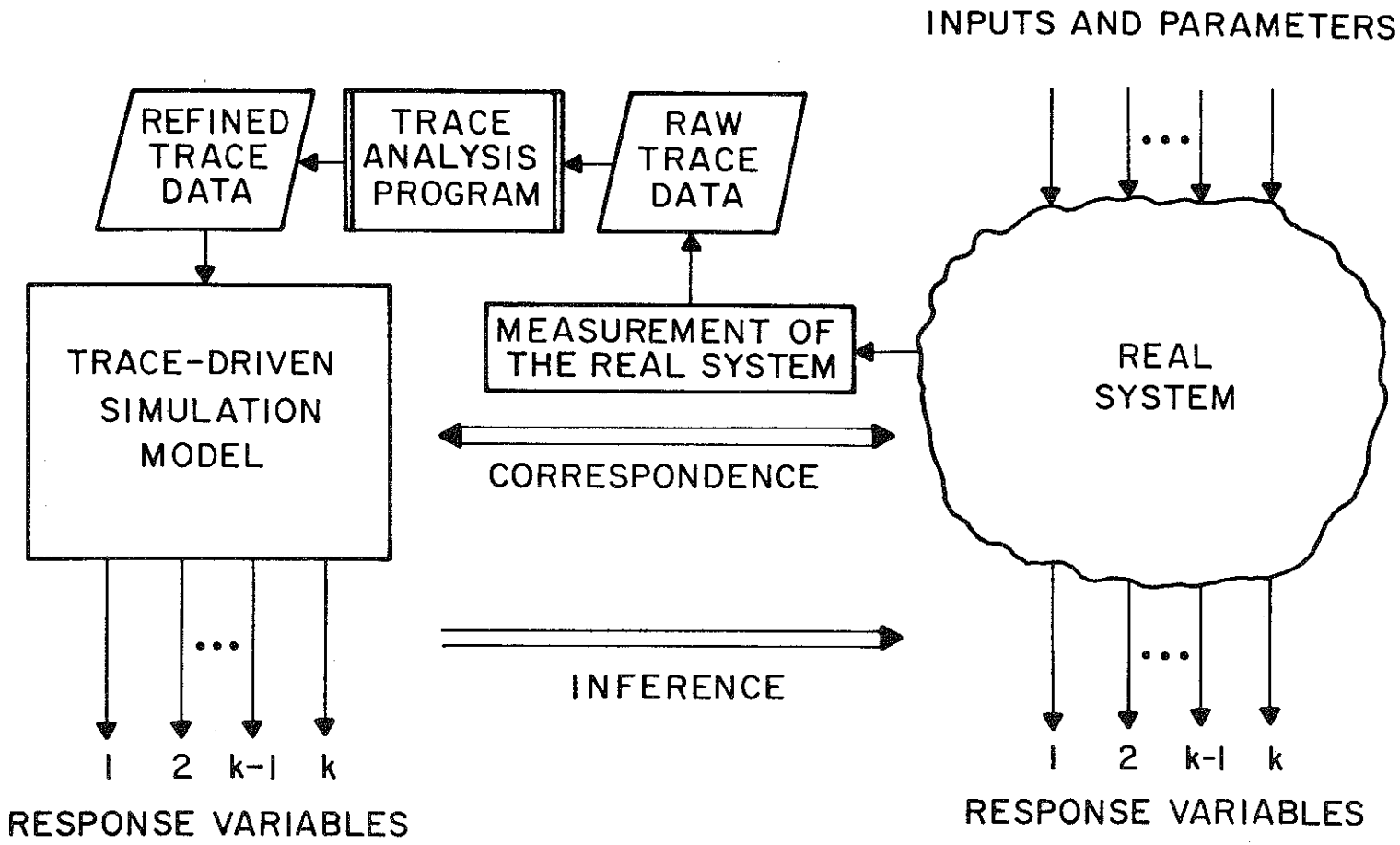


FIGURE 1. A Diagrammatic Concept of a Multivariate Response Trace-Driven Simulation Model.

The validity of a multivariate response simulation model is tested by comparing the model response variables with the corresponding system response variables when the simulation model is run with the "same" input data that drive the real system. Here, the "same" has different implications depending upon the way the simulation model is driven. In validating a trace-driven simulation model, the model input data is exactly the same as the system input data and therefore, the simulated data is correlated with the actual system output data. Due to this correlation, a one-sample significance test should be used for validation.

The validity of a multivariate response simulation model should be tested by using a multivariate statistical procedure. It would not be proper to test the validity of a multivariate response simulation model by testing the validity separately for each of the response variables because of the multiple response problem mentioned by Burdick and Naylor [7] and emphasized by Shannon [22].

Hotelling's one-sample T^2 test [15] can then be used for validating multivariate response trace-driven simulation models since it is a one-sample multivariate significance test.

Step 2 of the methodology requires the determination of the test statistic, the decision rule from the test statistic, the validity measure, and the power function of the test.

Assuming k response variables from the model and from the system, let $(\underline{\mu}^m)' = [\mu_1^m, \mu_2^m, \ldots, \mu_k^m]$ and $(\underline{\mu}^s)' = [\mu_1^s, \mu_2^s, \ldots, \mu_k^s]$ be the k dimensional vectors containing the population means of the model and system response variables. Furthermore, let μ_j^d $(= \mu_j^m - \mu_j^s)$ represent the population mean of the differences between the paired observations collected from the jth (j = 1,2,...,k) model and system response variables.

The acceptable range of accuracy, which is determined with respect to the purpose for which the model is intended, can be expressed in terms of the allowable differences between the population means and can be stated as

$$|\underline{\mu}^d| \leq \underline{\delta} \tag{1}$$

where $\underline{\mu}^d$ is a k dimensional vector containing the elements μ_j^d and $\underline{\delta}$ is a vector of the largest acceptable differences.

Independent observations can be collected from a steady-state [9] trace-driven simulation model by using the method of replications or the method of batch means with sufficient batch size. The method of replications can be used for collecting independent observations from a terminating [14] trace-driven simulation model. Let x_{ij} and y_{ij} represent the ith independent paired observations of the jth steady-state or terminating model and system response variables. The paired observations x_{ij} and y_{ij} represent the ith

independent paired replication values when the method of replications is used, and the ith independent paired batch mean values when the method of batch means is used. The one-sample T^2 test requires that

$$n_j = N \quad \text{and} \quad N_j = N \quad \text{for} \quad j = 1,\ldots,k \ . \tag{2}$$

Thus, we can obtain the following paired data matrices which are correlated with each other.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix} \qquad \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nk} \end{bmatrix} \tag{3}$$

MODEL DATA MATRIX SYSTEM DATA MATRIX

Let the difference between the paired observations x_{ij} and y_{ij} be denoted as d_{ij}, that is, let

$$d_{ij} = x_{ij} - y_{ij}, \quad i = 1,\ldots,N; \ j = 1,\ldots,k \ . \tag{4}$$

Then, the matrix of differences between the paired observations collected from the model and system response variables is given as

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1k} \\ d_{21} & d_{22} & \cdots & d_{2k} \\ \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nk} \end{bmatrix} \tag{5}$$

MATRIX OF DIFFERENCES

Let $\bar{d}_j$ and S_d be the estimates of μ_j^d and the variance-covariance matrix of the differences between the paired observations on the model and system response variables $\$_d$, respectively, where

$$\bar{d}_j = \frac{1}{N} \sum_{i=1}^{N} d_{ij}, \quad j = 1,\ldots,k \tag{6}$$

$$S_d = \frac{1}{N-1} \sum_{i=1}^{N} (\underline{d}_i - \underline{\bar{d}})(\underline{d}_i - \underline{\bar{d}})' \tag{7}$$

where

$$\underline{d}_i' = [d_{i1}, d_{i2}, \ldots, d_{ik}], \quad i = 1,\ldots,N \tag{8}$$

and

$$\underline{\bar{d}}' = [\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_k] \ . \tag{9}$$

Then, the test statistic of the one-sample T^2 test is computed as [15]

$$T^2 = N\,\bar{\underline{d}}'S_d^{-1}\bar{\underline{d}} \ . \tag{10}$$

The T^2 statistic has the central T^2 distribution when $\underline{\mu}^d = \underline{0}$ is true. For T^2 to have an F distribution, the expression for T^2 must be weighted by the factor $(N-k)/k(N-1)$ so that

$$F = \frac{N-k}{k(N-1)}\,T^2 \sim F_{k,N-k} \ , \tag{11}$$

where $F_{k,N-k}$ is the F distribution with degrees of freedom k and N-k. Thus, the decision rule for testing the validity of the model with specified maximum model user's risk of β^* for the acceptable range of accuracy (1) and with the minimum model builder's risk of α^* is the following: Accept the validity of the model with respect to the validity measure under the given experimental frame if

$$T^2 \leq \frac{k(N-1)}{N-k}\,F_{\alpha^*;k,N-k} \tag{12}$$

and reject otherwise, where $F_{\alpha^*;k,N-k}$ is the upper α^* percentage point of F distribution with degrees of freedom k and N-k.

When $\underline{\mu}^d = \underline{0}$ is not true, the quantity F in (11) has the noncentral F distribution with noncentrality parameter

$$\lambda = N(\underline{\mu}^d)'\Sigma_d^{-1}(\underline{\mu}^d) \tag{13}$$

and degrees of freedom k and N-k [15]. When the noncentrality parameter λ is equal to zero, $\underline{\mu}^d = \underline{0}$ holds perfectly and it implies that the model is a perfect representation of the system with respect to its mean behavior. Any difference between $\underline{\mu}^d$ and $\underline{0}$ will result in a value for λ which is greater than zero. As the difference between $\underline{\mu}^d$ and $\underline{0}$ increases, the value of λ will also increase. Hence, the noncentrality parameter λ is a validity measure for the one-sample T^2 test.

Substituting the acceptable range of accuracy (1) into (13), we obtain the upper bound of the acceptable validity range as

$$\lambda^* = N\underline{\delta}'\ddagger_d^{-1}\underline{\delta} \ . \tag{14}$$

The variance-covariance matrix $\ddagger_d$ should be estimated from a pilot run and/or earlier data.

The power function of the one-sample T^2 test is given by

$$1 - \beta(\lambda) = \Pr(F' > F_{\alpha^*;k,N-k}) \tag{15}$$

where F' has the noncentral F distribution with the prescribed parameters. From (15) we obtain the model user's risk β as a function of λ as

$$\beta(\lambda) = 1 - \Pr(F' > F_{\alpha^*;k,N-k}) \ . \tag{16}$$

$\beta(\lambda)$ is tabulated in [11] for several values of α^*,k,N-k, and various values of ϕ, where

$$\phi = \sqrt{\frac{2\lambda}{k+1}} \quad . \tag{17}$$

The maximum model user's risk β^* is given by $\beta(\lambda^*)$.

We now proceed to Step 3 of the methodology. We must first obtain an objective function in terms of the sample size N for the optimization problem in [5]. Recognizing that the power of the one-sample T^2 test is a monotonically increasing function of the noncentrality parameter λ [15], we can maximize the power or minimize the model user's risk β by maximizing the noncentrality parameter λ. The noncentrality parameter λ can be maximized by maximizing N since in its expression (13), $(\underline{\mu}^d)'\underline{\Sigma}_d^{-1}(\underline{\mu}^d)$ is a constant. Therefore, we have

$$\text{minimize(model user's risk } \beta) \equiv \text{maximize}(\lambda) \equiv \text{maximize}(N) \quad .$$

Noting that the degree of freedom N-k must be greater than or equal to 1, the optimization problem can be stated as

$$\begin{aligned} &\text{Maximize:} && N \\ &\text{Subject to:} && C_d N \leq B - c_0 - C_0 \\ & && N \geq k+1 \\ & && N \text{ integer} \end{aligned} \tag{18}$$

where $C_d = \sum_{j=1}^{k} (c_j + C_j)$; c_j and C_j are the unit costs of collecting one observation from the jth model and system response variables, respectively, B is the budget for data collection from the model and system, and c_0 and C_0 are the overhead data collection costs on the model and on the system.

Let the largest integer less than or equal to x be denoted by $\lfloor x \rfloor$. If $\lfloor (B-c_0-C_0)/C_d \rfloor \geq k+1$, then the optimal solution to problem (18) is given as $N^* = \lfloor (B-c_0-C_0)/C_d \rfloor$; otherwise (18) is infeasible.

We now have completed Step 3 of the methodology. Using the first three steps and following the remaining Steps 4 through 9, we present the following procedure for the one-sample T^2 test to construct the schedules in Table 1 for "a" values of ($\underline{c}$, c_0, $\underline{C}$, C_0), for "b" values of (B), for "c" values of (λ^*), and for "d" values of (α^*).

Procedure for the One-Sample T^2 Test to Construct the Schedules in Table 1.

Step 1: Initialize the counts, $u_1 = u_2 = u_3 = u_4 = 0$ and input k, $\hat{\xi}_d$, a, b, c, and d. Go to Step 2.

Step 2: Input $\underline{c}$, $\underline{C}$, c_0, and C_0. Compute $C_d = \underline{c}'\underline{1} + \underline{C}'\underline{1}$. Go to Step 3.

Step 3: Input B and compute $B' = B - c_0 - C_0$. Go to Step 4.

Step 4: If $\lfloor B'/C_d \rfloor < k+1$, go to Step 11; otherwise set $N^* = \lfloor B'/C_d \rfloor$ and go to Step 5.

TABLE 1. The Schedules.

$\underline{c}$, c_0	$\underline{C}$, C_0	B	N^*	CDC	λ^*	α^*	β^*
$\underline{c}_1, c_{01}$	$\underline{C}_1, C_{01}$	B_{11}	N^*_{11}	CDC_{11}	λ^*_{111}	α^*_{1111} ⋮ α^*_{111d}	β^*_{1111} ⋮ β^*_{111d}
					⋮	⋮	⋮
					λ^*_{11c}	α^*_{11c1} ⋮ α^*_{11cd}	β^*_{11c1} ⋮ β^*_{11cd}
		⋮	⋮	⋮	⋮	⋮	⋮
		B_{1b}	N^*_{1b}	CDC_{1b}	λ^*_{1b1}	α^*_{1b11} ⋮ α^*_{1b1d}	β^*_{1b11} ⋮ β^*_{1b1d}
					⋮	⋮	⋮
					λ^*_{1bc}	α^*_{1bc1} ⋮ α^*_{1bcd}	β^*_{1bc1} ⋮ β^*_{1bcd}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\underline{c}_a, c_{0a}$	$\underline{C}_a, C_{0a}$	B_{a1}	N^*_{a1}	CDC_{a1}	λ^*_{a11}	α^*_{a111} ⋮ α^*_{a11d}	β^*_{a111} ⋮ β^*_{a11d}
					⋮	⋮	⋮
					λ^*_{a1c}	α^*_{a1c1} ⋮ α^*_{a1cd}	β^*_{a1c1} ⋮ β^*_{a1cd}
		⋮	⋮	⋮	⋮	⋮	⋮
		B_{ab}	N^*_{ab}	CDC_{ab}	λ^*_{ab1}	α^*_{ab11} ⋮ α^*_{ab1d}	β^*_{ab11} ⋮ β^*_{ab1d}
					⋮	⋮	⋮
					λ^*_{abc}	α^*_{abc1} ⋮ α^*_{abcd}	β^*_{abc1} ⋮ β^*_{abcd}

Step 5: Compute CDC = $c_0 + C_0 + C_d N^*$, $\nu_1 = k$ and $\nu_2 = N^* - k$. Go to Step 6.

Step 6: Input $\underline{\delta}$ and compute $\lambda^* = N^* \underline{\delta}' \hat{\Phi}_d^{-1} \underline{\delta}$ and $\phi^* = \sqrt{2\lambda^*/(k+1)}$. Go to Step 7.

Step 7: Input α^* and retrieve β^* from the tables [11] for the values of α^*, ν_1, ν_2, and ϕ^*. Go to Step 8.

Step 8: Output $(\underline{c}, c_0)$, $(\underline{C}, C_0)$, B, N^*, CDC, $0 \leq \lambda \leq \lambda^*$, $\alpha^* \leq \alpha \leq 1-\beta^*$, and $0 \leq \beta < \beta^*$. Go to Step 9.

Step 9: Compute $u_4 = u_4+1$. If $u_4 < d$, go to Step 7; otherwise set $u_4 = 0$ and go to Step 10.

Step 10: Compute $u_3 = u_3+1$. If $u_3 < c$, go to Step 6; otherwise set $u_3 = 0$ and go to Step 11.

Step 11: Compute $u_2 = u_2+1$. If $u_2 < b$, go to Step 3; otherwise set $u_2 = 0$ and go to Step 12.

Step 12: Compute $u_1 = u_1+1$. If $u_1 < a$, go to Step 2; otherwise terminate.

The model sponsor, model user, and model builder, individually or together, can perform a cost-risk analysis for the one-sample T^2 test by using the schedules in Table 1 constructed by following the above procedure. By examining the schedules and/or the graphs of the data contained in the schedules, cost-risk trade-offs can be determined and judgement decisions can be made as to what risks to take, what budget to allocate, what method to use to measure the system and collect the trace data, and how many paired observations to

collect for testing the validity of a steady-state or terminating trace-driven simulation model by using the one-sample T^2 test. Construction of the schedules and graphs will be illustrated by the example of section 5.

In those cases where the data collection cost is not a relatively important factor to consider, a sample size-risk analysis can be performed without considering the data collection cost. In this case, the schedules in Table 1 are constructed with no cost parameters for several enumerated values of the sample size N. Then, by examining the schedules and/or the graphs of the data contained in the schedules, sample size-risk trade-offs can be determined and judgement decisions can be made as to what risks to take with respect to how many paired observations to collect.

3. ASSUMPTIONS OF THE TEST AND REMEDIAL MEASURES

Two assumptions are fundamental to the statistical theory underlying the one-sample T^2 test: (1) independence, and (2) multivariate normality.

3.1 Independence

The observation vectors ($\underline{d}_i$, i = 1,...,N) in the matrix of differences (5) must be independent from each other in using the one-

sample T^2 test. This implies the independence among the observation vectors in the model data matrix (3) and the independence among the observation vectors in the system data matrix (3).

This assumption can be satisfied by using one of the two major approaches for obtaining independent observations, namely, the method of replications and the method of batch means with sufficient batch size. In using the method of batch means, the batch size must be chosen appropriately to obtain independence. Both of the two methods can be used for steady-state trace-driven simulations and the method of replications can be used for terminating trace-driven simulations.

3.2 Multivariate Normality

Everitt [8] investigated the effects of departures from normality on the one-sample T^2 test. For situations involving from two to ten variables and with respect to the significance level, he reported that "Hotelling's one-sample T^2 test is badly affected by departures from multivariate normality due to skewness, but is fairly robust against departures due to kurtosis." The multivariate normality must be tested since the power of the test is a matter of concern in model validation and the test is sensitive to the departures from multivariate normality due to skewness.

Univariate normality of the response variables of a steady-state simulation model may be achieved by increasing the batch size when the method of batch means is used and by increasing the run

length when the method of replications is used. The effects of the size and the number of batches on the normality, in the method of batch means, are discussed in [13, 21]. In a similar manner, the system response variables can be observed by using one of the aforementioned two methods to try to achieve univariate normality.

The paired observations x_{ij} and y_{ij} represent the averages of the jth response variables in the ith replication when the method of replications is used and the averages of the jth response variables in the ith paired batch when the method of batch means is used. As Box, Hunter and Hunter [6] point out, the distribution of the differences between sample averages ($\underline{d}_j$, j = 1,...,k) would be expected to be nearly normal because of the central limit effect even if the distributions of the original observations had been moderately nonnormal. Univariate normality of the differences between sample averages may then be achieved for steady-state and terminating simulations by increasing the sample size of differences.

Andrews et al. [2] indicate that although univariate normality of each variable does not imply multivariate normality of all the variables, the presence of many types of nonnormality is often reflected in the distribution of each variable as well. Hence, univariate normality of each variable should first be tested and then, upon the achievement of univariate normality, the multivariate normality should be tested. Univariate normality can be tested by using the Box-Cox transformation test [2] given in Table 2 or

TABLE 2. Box-Cox Transformation Test for Univariate Normality.

(a) Numerically find $\hat{\theta}$ to maximize $L_{max}(\theta)$ where

(i) $$L_{max}(\theta) = -\frac{n}{2}\ell n\hat{\sigma}^2 + (\theta-1)\sum_{i=1}^{n}\ell n(x_i)$$

(ii) $$\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}\{x_i^{(\theta)} - \bar{X}^{(\theta)}\}^2$$

(iii) $$\bar{X}^{(\theta)} = (1/n)\sum_{i=1}^{n} x_i^{(\theta)}$$

(iv) $$x_i^{(\theta)} = \begin{cases} (x_i^{\theta} - 1)/\theta & \text{for } \theta \neq 0, \\ \ell n(x_i) & \text{for } \theta = 0. \end{cases}$$

(v) x_i, $i = 1,\ldots,n$; $x_i > 0$

(b) Obtain the significance level γ from

$$2\{L_{max}(\hat{\theta}) - L_{max}(1)\} \leq \chi_1^2(\gamma)$$

where $\chi_1^2(\gamma)$ denotes the upper $100\gamma\%$ point of the chi-squared distribution with one degree of freedom.

TABLE 3. The Transformaton Test for Multivariate Normality.

(a) Numerically find $\hat{\underline{\theta}}$ to maximize $L_{max}(\underline{\theta})$ where

(i) $\hat{\underline{\theta}}' = [\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k]$ and $\underline{\theta}' = [\theta_1, \theta_2, \ldots, \theta_k]$

(ii) $$L_{max}(\underline{\theta}) = -\frac{n}{2}\ell n|\hat{\Phi}| + \sum_{j=1}^{k}[(\theta_j - 1)\sum_{i=1}^{n}\ell n y_{ij}]$$

(iii) $$\hat{\Phi} = (1/n)(Y^{(\underline{\theta})} - \underline{1}\cdot\hat{\underline{\mu}}')'(Y^{(\underline{\theta})} - \underline{1}\cdot\hat{\underline{\mu}}')$$

(iv) $$\hat{\underline{\mu}} = (1/n)Y^{(\theta)'} \cdot \underline{1}$$

(v) $$Y^{(\underline{\theta})} = (y_{ij}^{(\underline{\theta})})$$

(vi) $$y_{ij}^{(\underline{\theta})} = \begin{cases} (y_{ij}^{\theta_j} - 1)/\theta_j & \text{for } \theta_j \neq 0, \\ \ell n\, y_{ij} & \text{for } \theta_j = 0. \end{cases}$$

(vii) $Y = (y_{ij})$, $i = 1,\ldots,n$; $j = 1,\ldots,k$; $y_{ij} > 0$

(b) Obtain the significance level γ from

$$2\{L_{max}(\hat{\underline{\theta}}) - L_{max}(\underline{1})\} \leq \chi_k^2(\gamma)$$

where $\chi_k^2(\gamma)$ denotes the upper $100\gamma\%$ point of the chi-squared distribution with k degrees of freedom.

by using the Shapiro-Wilk test described in [9]. Multivariate normality can be tested by using the power transformation test [2] presented in Table 3.

4. PROCEDURE FOR VALIDATION

The steps in using the one-sample T^2 test, for validating multivariate response steady-state or terminating trace-driven simulation models with respect to the validity measure for an acceptable range of accuracy under a given experimental frame, are presented below. For the univariate case, the procedure can easily be modified for the use of the one-sample t test.

1: Determine the experimental frame under which the validity of the simulation model is going to be tested. Go to 2.

2: Specify the acceptable range of accuracy for the population means with respect to the intended application of the model as

$$|\mu_j^d| \leq \delta_j, \quad j = 1,\ldots,k$$

where δ_j is the largest acceptable difference between the population means of the jth model and system response variables. Go to 3.

3: If a trade-off analysis among the model builder's risk,

model user's risk, cost of data collection, sample size, and validity measure is desired, go to 4; otherwise select a data collection method and go to 5.

4: Perform the procedure given in section 2 to construct the schedules in Table 1. Select a data collection method and choose appropriate values for the model builder's risk, model user's risk, sample size of paired observations, and data collection budget by examining the schedules and/or the graphs of the data contained in the schedules. Go to 6.

5: Determine the sample size of paired observations, model builder's risk, and model user's risk. Go to 6.

6: Collect N independent paired observations on the model and system response variables. Set $j = 1$, $I = \emptyset$, and go to 7.

7: Test the univariate normality of the differences between the paired observations on the jth model and system response variables. If it is found reasonably normal, go to 8; otherwise store j in I and go to 8.

8: Compute $j = j+1$. If $j \leq k$, go to 7; otherwise go to 9.

9: If $I = \emptyset$, go to 12; otherwise go to 10.

10: If the lack of reasonable normality in the distribution of the differences between the paired observations on the jth ($j \in I$) model and system response variables is believed

to be created because of the values chosen for the sample size of paired observations, model builder's risk, and/or model user's risk, then go back to 3 to choose new values; otherwise go to 11.

11: If there is a remedial measure to correct the nonnormality, such as increasing the batch size or the run length, go to 3; otherwise choose another statistical test or validation technique. Terminate.

12: Test the multivariate normality of the differences between the paired observations on the model and system response variables. If multivariate normality is achieved, go to 13; otherwise go to 11.

13: Apply the one-sample T^2 test to test the validity of the model with respect to the validity measure for the acceptable range of accuracy under the given experimental frame. If the model is found valid, go to 15; otherwise go to 14.

14: Determine, due to which variable(s) the invalidity occurs by testing $\mu_j^d = 0$ for the given acceptable range of accuracy, separately, for each of the response variables [15] or by constructing simultaneous confidence intervals for μ_j^d, $j = 1,\ldots,k$. If the invalidity is believed to be created because of the values chosen for the sample size, risks, and/or the estimate of the variance-covariance matrix, then go back to 3 to choose new values; otherwise

conclude that the model is invalid under the experimental frame considered, revise the model, and go to 3.

15: Conclude that the model is valid with respect to the validity measure for the acceptable range of accuracy under the given experimental frame. Terminate.

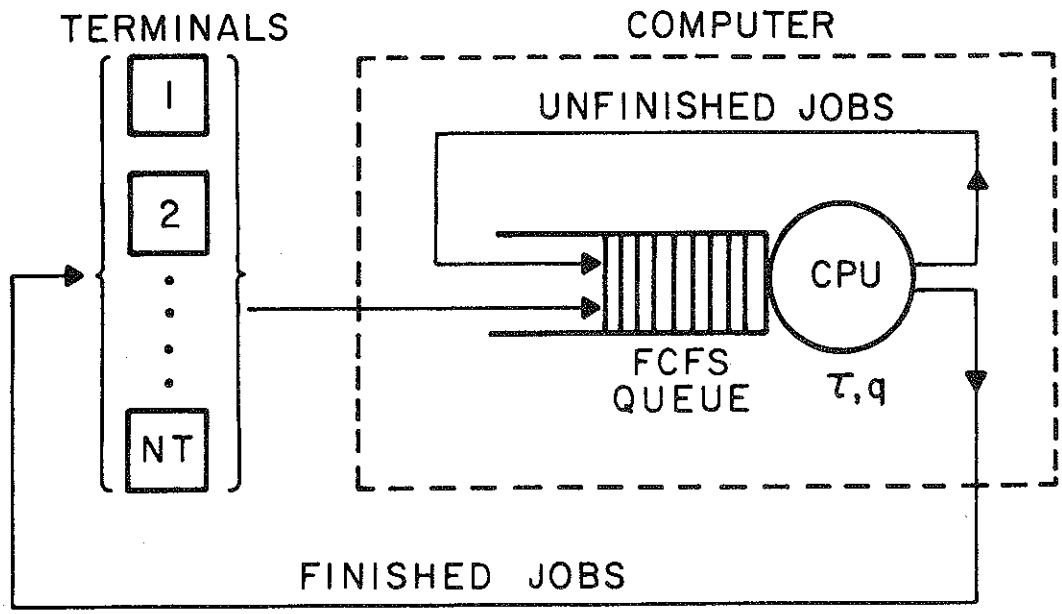
5. EXAMPLE

In this section, construction of the schedules and graphs for cost-risk analysis, assessment of multivariate normality, and the one-sample T^2 test together with the validation procedure are illustrated for the validation of a time-shared computer model [1, 17] in steady-state.

The model, as shown in Figure 2, consists of NT = 25 terminals and a single central processing unit (CPU). The user of each terminal "thinks" for an amount of time which is an independent and exponentially distributed random variable with a mean of 25 seconds. Then, the job produced at the terminal after the think time is sent to the CPU with a service time which is an independent and exponentially distributed random variable with a mean of 0.8 second. The arriving jobs join a single queue, with first come first serve (FCFS) discipline, in front of the CPU and are served in a round-robin manner. That is, the CPU allocates to each job a maximum service

quantum of length q = 0.1 second (not including overhead). If the job is finished within this service time period, it is returned back to its terminal after spending a fixed overhead time of τ = 0.015 second at the CPU. If the job is not finished within q = 0.1 second of service time, its remaining service time is decremented by q seconds and it is placed at the end of the queue after spending a fixed overhead time of τ = 0.015 second at the CPU.

FIGURE 2. A Time-Shared Computer Model with Round-Robin Service Discipline.



The time-shared computer model has two response variables (performance measures) of interest. The first response variable is the utilization of the CPU and the second one is the average response time which is the average time elapsed between the time the job leaves its terminal and the time it is finished being processed at the CPU.

For the purpose of illustration, the real system is represented by the same time-shared computer model with the same values stated above except that the fixed overhead time for the system is considered to be an independent random variable having an Erlang distribution with parameter 5 and a mean value of 0.015 second. Two simulation programs are coded in GPSS/H; one to represent the model and the other one to represent the real system. The GPSS/H program representing the real system is run and during the course of simulation the interarrival and service times of jobs with respect to each terminal are stored in a data-base with the help of a FORTRAN subroutine called in the GPSS/H program. The initial (starting) conditions are assumed to be all terminals being in the think state at time zero. Then, the GPSS/H program representing the model is driven by the interarrival and service times of jobs in the data-base and the simulation is carried out until running out of jobs from a terminal. The method of replications is used for data collection. The model and the system response vari-

ables are observed in pairs by running the model with the same trace-data that drive the real system.

The steps of the procedure given in section 4 will be followed for validating the time-shared computer model. The experimental frame under which the validity of the model is going to be tested with respect to its mean behavior is determined by the exponentially distributed think times and service time requests, one CPU with round-robin service discipline, and a single queue with first come first serve queue discipline. Assuming that the intended application of the model is to analyze the mean steady state behavior of the system with respect to the performance measures chosen, the acceptable range of accuracy is specified as

$$|\mu_1^d| \leq 0.02$$
$$|\mu_2^d| \leq 0.03 \qquad (19)$$

where μ_1^d and μ_2^d are the population means of the differences between the paired observations on the first and second model and system response variables, respectively.

The overhead costs for statistical data collection by way of replication for the model and for the system are estimated to be \$200 and \$1,400, respectively. It is estimated that the unit cost of collecting one independent observation (one replication) from

each model response variable is \$50 and from the first and second system response variables it is \$150 and \$200, respectively.

The model sponsor, model user, and model builder are willing to examine the trade-offs among the model user's risk, model builder's risk, acceptable validity range, sample size of observations, and data collection budget to determine appropriate values for these parameters. This will be done in two stages. In the first stage, schedules and graphs showing the relationships among the parameters will be constructed using the procedure of section 2. In the second stage, the trade-offs among the parameters will be examined by studying the relationships constructed and appropriate values for the parameters will be determined with respect to the intended application of the model.

In order to construct the schedules in Table 1, the variance-covariance matrix must be estimated first. According to a study performed by Sargent [17], it has been found that the time-shared computer model considered here reaches the steady-state conditions after the first 200 to 300 observations. Hence, after deleting the first 250 observations and each system run being composed of 500 observations in steady-state, five independent paired observations are obtained in pilot runs. The observations are given in Table 4 and the variance-covariance matrix is estimated to be

$$\hat{\Sigma} = \begin{bmatrix} 0.000659 & 0.000272 \\ 0.000272 & 0.265486 \end{bmatrix}. \qquad (20)$$

TABLE 4. Pilot Data.

MODEL		SYSTEM		DIFFERENCE	
Variable 1	Variable 2	Variable 1	Variable 2	Variable 1	Variable 2
0.81	3.48	0.82	2.80	-0.01	0.68
0.80	3.06	0.78	3.25	0.02	-0.19
0.82	3.61	0.83	3.11	-0.01	0.50
0.72	2.68	0.76	3.21	-0.04	-0.53
0.86	3.84	0.84	4.06	0.02	-0.22

The schedules in Table 1 are constructed by using the procedure of section 2 for c_0 = \$200; C_0 = \$1,400; C_m = \$100; C_s = \$350; B_i = \$2,950 + 450i, i = 1,2,...,12; $\underline{\delta}'\hat{\Sigma}_d^{-1}\underline{\delta} = 0.609$ (as calculated from the values given in (19) and (20)); $\underline{\delta}'\Sigma_d^{-1}\underline{\delta} = 0.03 + 0.05i$, i = 0,1, ...,18; and $\alpha^* = 0.01, 0.05, 0.10$.

The trade-offs among the parameters can be examined by studying the schedules and/or the graphs of the data contained in the schedules, and judgement decisions can be made to determine appropriate values for the maximum model user's risk (β^*), minimum and maximum model builder's risks ($\alpha^* \leq \alpha \leq 1-\beta^*$), acceptable validity range ($0 \leq \lambda \leq \lambda^*$), sample size of observations (N^*), and data collection budget (B).

A question of particular interest is "what would be the maximum model user's risk, maximum model builder's risk, and acceptable validity range for the given values of c_0, C_0, C_m, C_s, B, α^*, and

$\underline{\delta}$?" In order to answer this question, assuming c_0 = \$200, C_0 = \$1,400, C_m = \$100, C_s = \$350, B = \$8,350, α^* = 0.05, and $\underline{\delta}'$ = [0.02, 0.03], first the optimal sample size N^* is read from the schedules corresponding to B = \$8,350 as 15 and then Figures 3 and 4 are constructed by using the data contained in the schedules. In Figure 3, the relationships among maximum model user's risk (β^*), minimum model builder's risk (α^*), and data collection cost (CDC) are shown for the given values of the parameters. In Figure 4, operating characteristic curves are given for the specified values of the parameters to determine the probability of accepting the simulation model as valid for various values of the validity measure λ and to allow the determination of β^* for a given value of the upper bound of the acceptable validity range λ^*.

The upper bound of the acceptable validity range (λ^*) is calculated as $N^*\underline{\delta}'\hat{\underline{\Sigma}}_d^{-1}\underline{\delta}$ = 9.135. Then, the value of the maximum model user's risk β^* is read from Figure 4 (or from the schedules) for α^* = 0.05 and λ^* = 9.135 as 0.075. Thus, we get $0 \leq \beta \leq 0.075$, $0.05 \leq \alpha \leq 0.925$, and $0 \leq \lambda \leq 9.135$. Assuming that these values are satisfactory, we choose N^* = 15.

After choosing N^* = 15 as a result of the cost-risk analysis, 15 independent and identically distributed paired observations are obtained by deleting the first 250 observations and running the system for 500 observations in steady-state. The paired observations obtained and the differences between them are presented in Table 5.

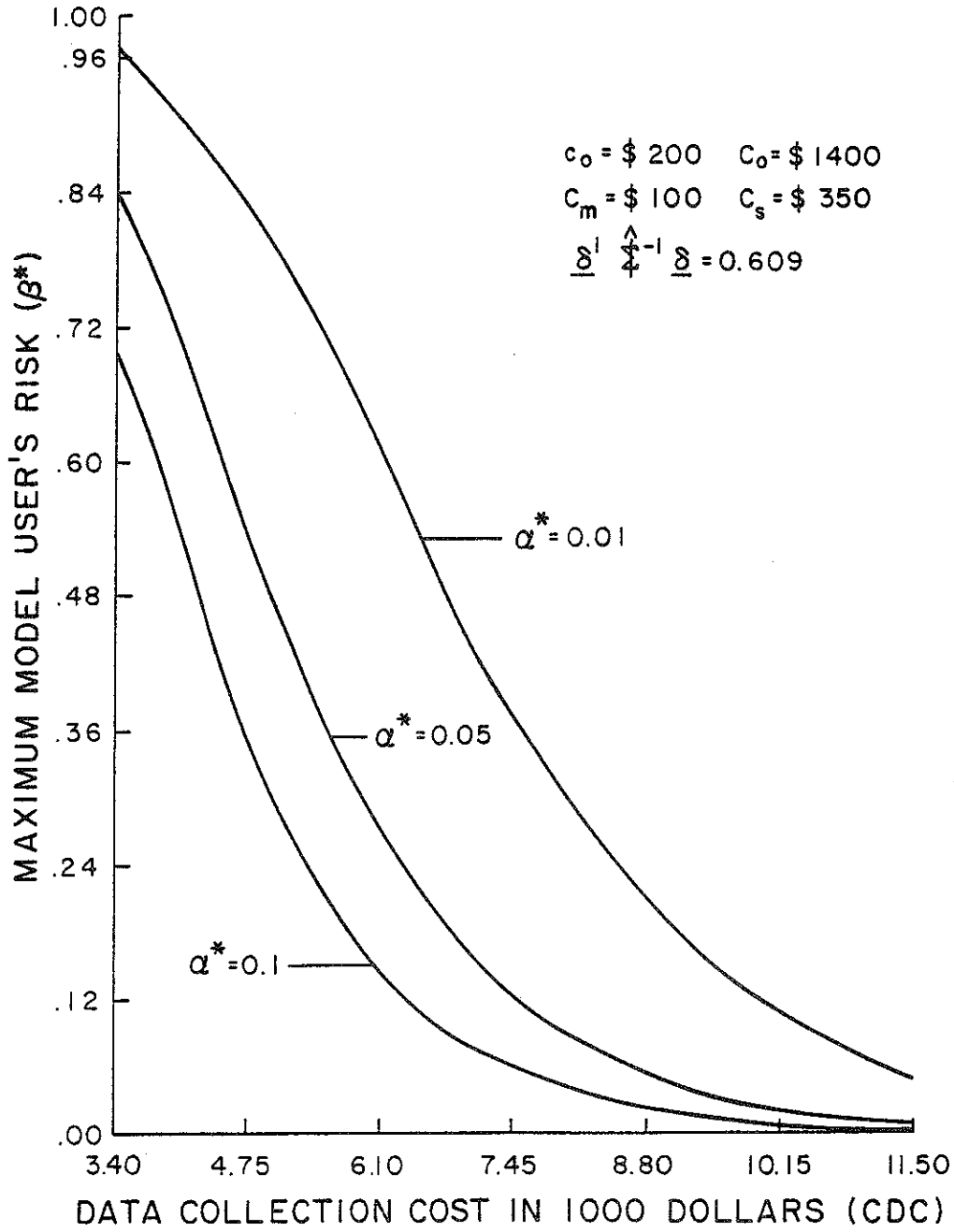


FIGURE 3. Cost Versus Maximum Model User's Risk.

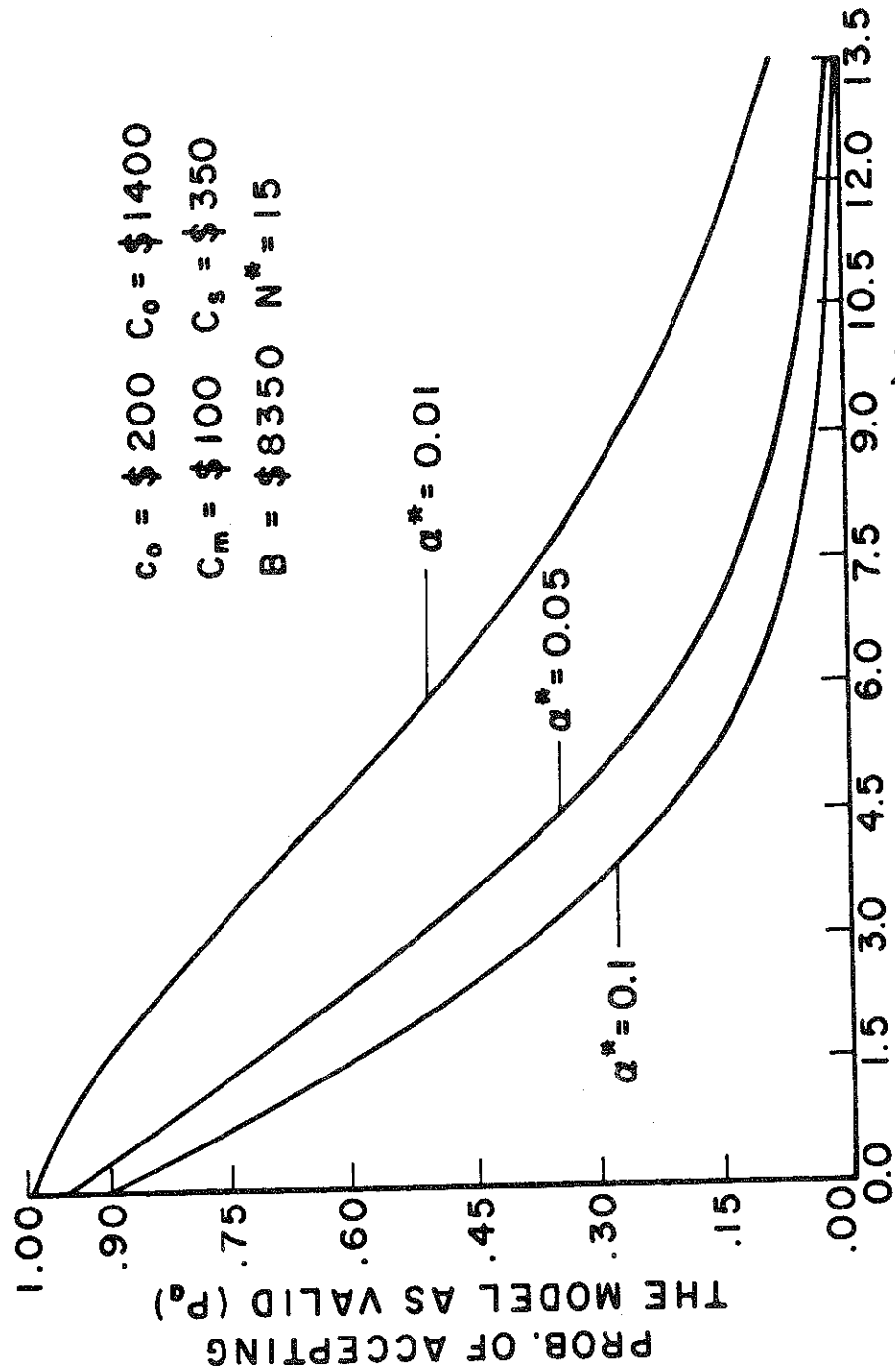


FIGURE 4. Operating Characteristic Curves.

TABLE 5. Data Collected for Validation.

MODEL		SYSTEM		DIFFERENCE	
Variable 1	Variable 2	Variable 1	Variable 2	Variable 1	Variable 2
0.76	2.57	0.77	2.70	-0.01	-0.13
0.77	2.95	0.77	2.88	0.00	0.07
0.84	2.76	0.82	2.70	0.02	0.06
0.89	4.17	0.88	4.02	0.01	0.15
0.76	2.52	0.75	2.60	0.01	-0.08
0.79	3.62	0.81	3.88	-0.02	-0.26
0.86	3.37	0.86	3.17	0.00	0.20
0.79	3.37	0.78	3.38	0.01	-0.01
0.82	3.45	0.82	3.63	0.00	-0.18
0.77	2.54	0.77	2.76	0.00	-0.22
0.80	3.31	0.79	3.18	0.01	0.13
0.85	3.70	0.87	3.71	-0.02	-0.01
0.84	3.90	0.85	3.96	-0.01	-0.06
0.87	3.89	0.89	3.70	-0.02	0.19
0.73	2.68	0.75	2.67	-0.02	0.01

We are now in Step 6 of the validation procedure. Setting $j = 1$ and $I = \emptyset$, we go to Step 7 and apply the univariate normality test in Table 2 to the differences between the paired observations on the first model and system response variables and we repeat this for $j = 2$. The results of the tests are presented in Table 6.

After achieving reasonable univariate normality, we go to Step 12 and apply the multivariate normality test in Table 3. The results of this test are also given in Table 6. Multivariate normality is achieved at an approximate significance level of 0.953.

In Step 13, the one-sample T^2 test is applied to test the validity. As a result, the test statistic T^2 is found to be 0.877 which is less than $(28/13)F_{0.1;2,13}$ (see (12)). Thus, in Step 15, it is concluded that the model is valid with respect to the validity measure for the acceptable range of accuracy under the given experimental frame.

TABLE 6. Normality Tests.

Univariate Power Transformation Tests				
Difference on Response	$\hat{\theta}$	$2\{L_{max}(\hat{\theta})-L_{max}(1)\}$	Approximate γ	Univariate Normal?
1	5.37	0.039	0.860	Yes
2	1.52	0.079	0.790	Yes

Multivariate Power Transformation Test					
Difference on Response	$\hat{\theta}_1$	$\hat{\theta}_2$	$2\{L_{max}(\hat{\theta}_1,\hat{\theta}_2)-L_{max}(1,1)\}$	Approximate γ	Multivariate Normal?
1	5.6	1.4	0.096	0.953	Yes
2					

6. CONCLUSIONS

A procedure using Hotelling's one-sample T^2 test is presented for validating a multivariate response trace-driven simulation model of an observable system with respect to its mean behavior.

Construction of the schedules and graphs for cost-risk analysis, assessment of multivariate normality, and the one-sample T^2 test together with the validation procedure are illustrated by an example. In this example, a steady-state trace-driven simulation model representing a time-sharing computer system with 25 terminals and two performance measures is considered.

REFERENCES

[1] Adiri, I. and Avi-Itzhak, B., "A Time-Sharing Queue with a Finite Number of Customers," Journal of ACM, Vol. 16, No. 2, (April 1969), pp. 315-323.

[2] Andrews, D.F., Gnanadesikan, R. and Warner, J.L., "Methods for Assessing Multivariate Normality," in Multivariate Analysis III, P.R. Krishnaiah, Ed., Academic Press, New York, 1973, pp. 95-116.

[3] Balci, O., "Statistical Validation of Multivariate Response Simulation Models," Ph.D. Dissertation, Syracuse University, July 1981.

[4] Balci, O. and Sargent, R.G., "Bibliography on Validation of Simulation Models," Newsletter-TIMS College on Simulation and Gaming, Vol. 4, No. 2 (Spring 1980), pp. 11-15.

[5] Balci, O. and Sargent, R.G., "A Methodology for Cost-Risk Analysis in the Statistical Validation of Simulation Models," Communications of ACM, Vol. 24, No. 4, (April 1981), pp. 190-197.

[6] Box, G.E.P., Hunter, W.G., and Hunter, J.S., Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley & Sons, New York, 1978.

[7] Burdick, D.S. and Naylor, T.H., "Design of Computer Simulation Experiments for Industrial Systems," Communications of ACM, Vol. 9, No. 5, (May 1966), pp. 329-339.

[8] Everitt, B.S., "A Monte Carlo Investigation of the Robustness of Hotelling's One- and Two-Sample T^2 Tests," J. American Statistical Association, Vol. 74, No. 365, (March 1979), pp. 48-51.

[9] Fishman, G.S., Principles of Discrete Event Simulation, Wiley-Interscience, New York, 1978.

[10] Gafarian, A.V. and Ancker, C.J., "Mean Value Estimation from Digital Computer Simulation," Operations Research, Vol. 14, (1966), pp. 25-44.

[11] Graybill, F.A., Theory and Application of the Linear Model, Duxbury Press, North Scituate, 1976.

[12] Kobayashi, H., Modeling and Analysis: An Introduction to System Performance Evaluation Methodology, Addison-Wesley, Reading, MA, 1978.

[13] Law, A.M., "Confidence Intervals in Discrete Event Simulation: A Comparison of Replication and Batch Means," Naval Research Logistics Quarterly, Vol. 24, No. 4, (December 1977), pp. 667-678.

[14] Law, A.M., "Statistical Analysis of the Output Data from Terminating Simulations," Naval Research Logistics Quarterly, Vol. 27, (March 1980), pp. 131-143.

[15] Morrison, D.F., Multivariate Statistical Methods, McGraw-Hill, New York, 1976.

[16] Rubinstein, R.Y., Simulation and the Monte Carlo Method, John Wiley & Sons, New York, 1981.

[17] Sargent, R.G., "An Introduction to Statistical Analysis of Simulation Output Data," Proc. NATO AGARD Symposium on Modeling and Simulation of Avionics Systems and Command, Control, and Communication Systems, (No. 268), Paris, France, October 1979, pp. 3-1/3-13.

[18] Sargent, R.G., "Validation of Simulation Models," Proc. Winter Simulation Conference, San Diego, California, December 1979, pp. 497-503.

[19] Sargent, R.G., "Verification and Validation of Simulation Models," in Progress in Modeling and Simulation, Edited by F.E. Cellier, Academic, London, 1982.

[20] Schlesinger, S., et al., "Terminology for Model Credibility," Simulation, (March 1979), pp. 103-104.

[21] Schmeiser, B., "Batch Size Effects in the Analysis of Simulation Output," Operations Research, Vol. 30, No. 3, May-June 1982.

[22] Shannon, R.E., Systems Simulation: The Art and Science, Prentice-Hall, New Jersey, 1975.

[23] Sherman, S.W., "Trace Driven Modeling: An Update," Proc. Symposium on the Simulation of Computer Systems IV, (August 1976), pp. 87-91.

[24] Zeigler, B.P., Theory of Modelling and Simulation, John Wiley & Sons, New York, 1976.