Technical Report CS81016-R*

VALIDATION OF MULTIVARIATE RESPONSE SIMULATION MODELS BY USING HOTELLING'S TWO-SAMPLE T^2 TEST

by

Osman Balci

and

Robert G. Sargent**

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

December 1981

---

*Cross listed as Working Paper #81-011 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 two-sample T^2 test to test the validity of a multivariate response 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 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 observations to collect on each of 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 is illustrated for a simulation model that represents an M/M/1 queueing system with two performance measures of interest.

1. INTRODUCTION

A common problem encountered in 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 [21] 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) [19,20].

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 simulation 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].

In using a statistical test for validation, one should consider the type of simulation model with regard to the way its output is analyzed. There are basically two types of simulation models with regard to analysis of the output: steady-state and terminating simulation models [10,16]. 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" [16]. 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" [16].

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" indicates that the simulation model input data and the system input data come from the same population but they are independent from each other. Hence, the simulated data are also expected to be independent from the actual system output data and this is subsequently considered in choosing or developing a statistical procedure 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]. Later, Shannon emphasized the importance of the multiple response problem by indicating that "unfortunately, a more serious problem exists with the validation of multivariate response models" [23, p. 229]. After illustrating the multiple response problem, Shannon suggested some statistical procedures for validating multivariate response simulation models among which is the Hotelling's T^2 Test [23, p. 231].

The purpose of this paper is to give a procedure for validating a multivariate response simulation model with respect to its mean behavior by using Hotelling's two-sample T^2 test and the methodology for cost-risk analysis given in [5]. In section 2, Hotelling's two-sample T^2 test is introduced and in section 3, the assumptions underlying the test together with some remedial measures are presented. The procedure for validation is given in section 4 and is illustrated in section 5 by an example for terminating simulation. Finally, conclusions are given in section 6.

2. HOTELLING'S TWO-SAMPLE T^2 TEST

Hotelling's two-sample T^2 Test [17] is a multivariate statistical test to test the equivalence of the means of two multivariate normal populations. It can be used to test the validity of a multivariate response simulation model with respect to its mean behavior by using the methodology given in [5].

In the first step of the methodology in [5] , Hotelling's two-sample T^2 test is used to test the following hypotheses:

H_0: Model is valid for the acceptable range of accuracy under the experimental frame. (1)

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 two-sample T^2 test for testing the hypotheses in (1). 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].

Assuming that the sample sizes of observations on each of the k model and system response variables are n and N, respectively, 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, and let Σ_m and Σ_s be the model and system variance-covariance matrices, respectively. Let us assume that for the purpose for which the simulation model is intended, the validity of the model can be determined with respect to its mean behavior and the acceptable range of accuracy can be stated as (2)

$$| \underline{\mu}^m - \underline{\mu}^s | \leq \underline{\delta}$$

where $\underline{\delta}$ is a vector of the largest acceptable differences.

Step 2 of the methodology in [5] requires the determination of the test statistic, the decision rule from the test statistic, the validity measure, and the power function of the test. Let x_{ij} and y_{ij} represent the ith independent observations of the jth steady-state or terminating model and system response variables, respectively. Let $\bar{x}_j$, $\bar{y}_j$ and S be the estimates of μ_j^m, μ_j^s and the common variance-covariance matrix Σ of Σ_m and Σ_s respectively, where

$$\bar{x}_j = (1/n) \sum_{i=1}^{n} x_{ij}, \qquad j = 1, \ldots, k \tag{3}$$

$$\bar{y}_j = (1/N) \sum_{i=1}^{N} y_{ij}, \qquad j = 1, \ldots, k \tag{4}$$

and

$$S = (A_1 + A_2)/(n+N-2) \tag{5}$$

where

$$A_1 = \sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})' \tag{6}$$

$$A_2 = \sum_{i=1}^{N} (\underline{y}_i - \bar{\underline{y}})(\underline{y}_i - \bar{\underline{y}})' \tag{7}$$

where

$$\underline{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{ik}], \quad i = 1, \ldots, n \tag{8}$$

$$\underline{y}_i' = [y_{i1}, y_{i2}, \ldots, y_{ik}], \quad i = 1, \ldots, N \tag{9}$$

$$\bar{\underline{x}}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k], \tag{10}$$

and

$$\bar{\underline{y}}' = [\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k]. \tag{11}$$

Then, the test statistic of the two-sample T^2 test is computed as [17]

$$T^2 = \frac{nN}{n+N} (\bar{\underline{x}} - \bar{\underline{y}})' S^{-1} (\bar{\underline{x}} - \bar{\underline{y}}). \tag{12}$$

The T^2 statistic has the central T^2 distribution when $\underline{\mu}^m = \underline{\mu}^s$ is true. For T^2 to have an F distribution, the expression of T^2 is weighted by the factor $(n+N-k-1)/k(n+N-2)$ so that

$$F = \frac{n + N - k - 1}{k(n + N - 2)} \quad T^2 \sim F_{k,n+N-k-1} \tag{13}$$

where "$\sim$" denotes "is distributed as" and $F_{k,n+N-k-1}$ is the F distribution with degrees of freedom k and $n+N-k-1$. Thus, the decision rule for testing the validity of the model with specified maximum model user's risk of β^* for a given acceptable range of accuracy 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

$$T^2 \leq \frac{k(n+N-2)}{n+N-k-1} F_{\alpha^*;k,n+N-k-1} \tag{14}$$

and reject otherwise, where $F_{\alpha^*;k,n+N-k-1}$ is the upper α^* percentage point of F distribution with degrees of freedom k and $n+N-k-1$.

When $\underline{\mu}^m = \underline{\mu}^s$ is not true, the quantity F in (13) has the noncentral

F distribution with noncentrality parameter

$$\lambda = (nN/(n+N))(\underline{\mu}^m - \underline{\mu}^s)'\underline{\Sigma}^{-1}(\underline{\mu}^m - \underline{\mu}^s) \tag{15}$$

which is the validity measure for the two-sample T^2 test [5]. Substituting the acceptable range of accuracy (2) into (15), the upper bound of the acceptable validity range is obtained as

$$\lambda^* = (nN/(n+N))\underline{\delta}'\underline{\Sigma}^{-1}\underline{\delta} \tag{16}$$

The model user's risk β which is one minus the power of the two-sample T^2 test is given as a function of the validity measure λ as

$$\beta(\lambda) = 1 - \Pr(F' > F_\alpha^*;k,n+N-k-1) \tag{17}$$

where F' has the noncentral F distribution with the noncentrality parameter λ and degrees of freedom k and n+N-k-1. The maximum model user's risk β^* is given by $\beta(\lambda^*)$.

The model user's risk $\beta(\lambda)$ which is extremely important in model

validation and the model builder's risk α 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 to show the relationships among the risks, acceptable validity range ($0 \leq \lambda \leq \lambda^*$), sample sizes, and cost of data collection. 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 budget to allocate, what data collection method to use, and how many observations to collect on each of the model and system response variables for the validation of the model. The reader is referred to [5] for the construction of the schedules and graphs for using Hotelling's two-sample T^2 test to test the validity of a simulation model.

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 and graphs are constructed with no cost parameters for several enumerated values of the sample sizes. 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 observations to collect. Notice that the model user's risk $\beta(\lambda)$ is minimized (since the term $nN/(n+N)$ in (15) and λ are maximized) when $n=N$ for a given value of $n+N$ in the case where the data collection cost is not considered.

The data collection cost is dependent upon the method of data collection that is selected in the validation procedure. Basically, there are three methods of data collection from a simulation model, namely, method of replications [9], method of batch means (9), and regenerative method with batched cycles [12]. All of these three methods can be used for steady-state simulation models and the method of replications can be used for terminating simulation models. These methods can also be employed, in a similar manner, for collecting data from the real system.

3. ASSUMPTIONS OF THE TEST AND REMEDIAL MEASURES

Three assumptions are fundamental to the statistical theory underlying the two-sample T^2 test: (1) independence, (2) multivariate normality, and (3) equality of variance-covariance matrices.

Independence

There are two independence assumptions underlying the two-sample T^2 test which must be satisfied. The first one is the independence between the model and the system output data matrices. The second one is the independence among the observation vectors in the model output data matrix (8) and in the system output data matrix (9). The first independence assumption is satisfied by the use of random numbers in sampling from the distributions of system input variables to obtain the values that drive the simulation model. The second one can be satisfied for steady-state simulation models by using one of the three

major methods of data collection, namely, method of replications, method of batch means (with sufficient batch size), and regenerative method with batched cycles. For terminating simulation models, the second independence assumption can be satisfied by using the method of replications.

Multivariate Normality

The model response variables and the system response variables are each assumed to have a multivariate normal distribution.

Everitt [8] investigated the effects of departures from normality on the two-sample T^2 test. For situations involving from two to ten variables and with respect to the significance level, he reported that "it may perhaps be concluded that Hotelling's two-sample T^2 test is fairly robust against departures from normality; certainly the test appears less affected by departures from this assumption than from that of the equality of variance-covariance matrices." The multivariate normality must be tested since the power of the test is a matter of concern in model validation and the equality of the variance-covariance matrices must be justified, at least approximately.

An excellent broad review of the assessment of multivariate normality has been given by Gnanadesikan [11, pp. 137-195]. Considering the computational difficulty involved in assessing joint multivariate normality, we will follow the natural, simple, and preliminary step suggested by Andrews et al. [2] and evaluate the normality of multiresponse observations by studying the reasonableness

of marginal normality for the data on each of the response variables before we attempt to assess joint multivariate normality. For this purpose, one can utilize various ways of assessing the hypothesis of univariate normality. Some of the most common methods are [2]: (i) likelihood ratio tests associated with transformations for enhancing univariate normality, (ii) skewness and kurtosis tests, (iii) omnibus tests for normality such as Shapiro and Wilk's W-test, (iv) goodness of fit tests such as the χ^2-test and the Kolmogorov-Smirnov test, and (v) graphical methods including normal probability plots.

Univariate normality of the model response variables can be achieved for steady-state simulations by increasing the run length when the method of replications is used, the batch size when the method of batch means is used, and the number of cycles in a batch when the regenerative method with batched cycles [12] is used. The effects of the size and the number of batches on the normality in the methods of batch means, are discussed in [15,22]. In a similar manner, the system response variables can be observed by using one of the aforementioned three methods to try to achieve univariate normality.

In case the univariate normality of a response variable cannot be achieved by using one of the above mentioned three methods, data-based transformations of univariate observations can be tried to enhance the normality of its distribution. Hoyle [13] reviews the ways of developing a suitable transformation and the various transformations that are to be found in the literature with an extensive bibliography. The transformations which help to correct the lack of normality usually are also effective in making the variances of the observations

more equal since lack of normality and unequal variances usually tend to go hand in hand [18]. Although the transformation of data can be used to enhance the normality, one must be extremely careful in using it since the accuracy of the representation of the original data may be badly affected by the transformation [18].

One of the methods proposed by Box and Cox [6] for obtaining a power transformation of a single variable x so as to improve the normality of its distribution is presented in Table 1. The value of $\hat{\theta}$ which maximizes $L_{max}(\theta)$ is found by enumeration in this paper.

Andrews, et al. [1,2] proposed methods, which are extensions of the techniques of Box and Cox, for obtaining transformations of multivariate observations to enhance the normality of their distribution. Table 2 presents the steps in applying their test using the power class of transformations. The values of $\underline{\hat{\theta}}$ which maximize $L_{max}(\underline{\theta})$ are found by enumeration in this paper. The procedures presented in Tables 1 and 2 not only indicate when the data are non-normal but also they suggest data transformations which may be used to enhance normality.

Equality of Variance-Covariance Matrices

The third assumption underlying the two-sample T^2 test is that the model response variables and the system response variables have the same variance-covariance matrix Σ of full rank k (number of response variables). We further assume that Σ is unknown.

TABLE 1. 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

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

where $\chi^2_{\gamma;1}$ denotes the upper $100\gamma\%$ point of the chi-square distribution with one degree of freedom.

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

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

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

(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

$$2\{L_{max}(\hat{\underline{\theta}}) - L_{max}(\underline{1})\} \leq \chi^2_{\gamma;k}$$

where $\chi^2_{\gamma;k}$ denotes the upper $100\gamma\%$ point of the chi-square distribution with k degrees of freedom.

Ito and Schull [14] have shown that the true significance level and the power function of the T^2 test is unaffected by discrepancies between Σ_m and Σ_s so long as the sample sizes (n,N) are equal to each other and fairly large. For small and unequal sample sizes, a separate test of the equality of the variance-covariance matrices should be performed.

A multivariate analog of Bartlett's test presented in Table 3 can be used to test the equality of the variance-covariance matrices [17]. The tests of Table 3 are very sensitive to nonnormality, and therefore multivariate normality must be assessed before using these tests.

If the assumption of equality of variance-covariance matrices is not satisfied, then the sample sizes (n,N) must be made equal and fairly large. In this case, unequal variance-covariance matrices have no effect upon the size of the Type I error probability or the power function of the two-sample T^2 test [14].

4. PROCEDURE FOR VALIDATION

The steps in using the two-sample T^2 test, for validating a multivariate response steady-state or terminating simulation model 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 two-sample t test.

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

TABLE 3. Testing the Equality of Variance-Covariance Matrices.

(a) Compute

(i) $M = (n+N-2)\ln|S| - (n-1)\ln|A_1/(n-1)| - (N-1)\ln|A_2/(N-1)|$

(ii) $U = 1 - \frac{2k^2+3k-1}{6(k+1)}\left(\frac{1}{n-1} + \frac{1}{N-1} - \frac{1}{n+N-2}\right)$

(iii) $\nu_1 = k(k+1)/2$

(b) If $k < 6$, $n > 20$, and $N > 20$ go to (d); otherwise compute

(i) $G_1 = 1-U$

(ii) $G_2 = \frac{1}{6}(k-1)(k+2)\left[\frac{1}{(n-1)^2} + \frac{1}{(N-1)^2} - \frac{1}{(n+N-2)^2}\right]$

(iii) $\nu_2 = (\nu_1+2)/(G_2-G_1^2)$

(iv) $V = [1-G_1-(\nu_1/\nu_2)]/\nu_1$

(c) The equality is accepted at the significance level γ if

$$MV \leq F_{\gamma;\nu_1,\nu_2}$$

where $F_{\gamma;\nu_1,\nu_2}$ denotes the upper γ percentage point of F distribution with ν_1 and ν_2 degrees of freedom. Terminate.

(d) The equality is accepted at the significance level γ if

$$MU \leq \chi^2_{\gamma;\nu_1}$$

where $\chi^2_{\gamma;\nu_1}$ denotes the upper γ percentage point of chi-square distribution with ν_1 degrees of freedom. Terminate.

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

$$| \mu_j^m - \mu_j^s | \leq \delta_j, \quad j = 1, \ldots, k$$

where δ_j is the largest acceptable difference. Go to 3.

3. If a trade-off analysis among the model builder's risk, model user's risk, cost of data collection, sample sizes, and validity measure is desired, go to 4; otherwise select a data collection method and go to 5.

4. Perform the procedure given in [5] to construct the schedules and graphs. Select a data collection method and choose appropriate values for the model builder's risk, model user's risk, sample sizes of observations, and data collection budget by examining the schedules and graphs. Go to 6.

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

6. Collect n and N independent observations from each model and system response variable, respectively, by running the simulation model with the same input data that drive the real system. Set j = 1 and go to 7.

7. Apply the univariate normality test in Table 1 to system response variable j. If system response variable j is found reasonably normal, go to 8; otherwise go to 9.

8. Compute $j = j+1$. If $j \leq k$, go to 7; otherwise set $j = 1$ and go to 11.

9. If all possible data-based transformations are tried, go to 10; otherwise apply a data-based transformation to system response variable j to enhance the normality of its distribution. Go to 7.

10. If the normality can be enhanced by increasing the batch size when the method of batch means or the regenerative method with batched cycles is used or by increasing the run length when the method of replications is used then go back to 3 to do so; otherwise, search for another statistical test or validation technique and terminate.

11. Apply the multivariate normality test in Table 2 to system response variables. If multivariate normality is achieved, set I $= \emptyset$ and go to 12; otherwise go to 7 to improve marginal normalities.

12. If a data-based transformation is applied to system response variable j to enhance the normality of its distribution, apply the same transformation to model response variable j. Go to 13.

13. Apply the univariate normality test in Table 1 to model response variable j. If model response variable j is found reasonably normal, go to 14; otherwise store j in I and go to 14.

14. Compute $j = j+1$. If $j \leq k$, to to 12; otherwise set $j = 1$ and go to 15.

15. If $I = \emptyset$, go to 17; otherwise go to 16.

16. If the incompatibility between the distributions of the jth ($j \in I$) model and system response variables is believed to be created because of the run length (when the method of replications is

used), the batch size (when the method of batch means is used), or the number of cycles in a batch (when the regenerative method with batched cycles is used) and/or because of the values chosen for the sample sizes, risks and/or the estimate of the common variance-covariance matrix, then go back to 3 to choose new values; otherwise revise the model and go to 3.

17. Apply the multivariate normality test in Table 2 to model response variables. If multivariate normality is achieved, go to 18; otherwise go to 7 to improve marginal normalities.

18. If $n \stackrel{\sim}{=} N$ and is sufficiently large, go to 20; otherwise go to 19.

19. Test the equality of the variance-covariance matrices. If they are found to be at least approximately equal to each other, then go to 20; otherwise make the sample sizes (n,N) approximately equal and fairly large by taking the cost-risk or sample size-risk trade-offs into consideration and go to 6.

20. Apply the two-sample T^2 test to test the validity of the model with respect to the validity measure for the given acceptable range of accuracy under the given experimental frame. If the model is found valid, go to 22; otherwise go to 21.

21. Determine due to which variable(s) the invalidity occurs by testing the equality of the population means for the given acceptable range of accuracy, separately, for each of the response variables [17] or by constructing simultaneous confidence intervals for the differences in means [17]. If the invalidity for the given acceptable range of accuracy is believed to be created because of the values chosen for the sample sizes, risks, and/or the estimate of the common variance-covariance matrix, then

go back to 3 to choose new values; otherwise revise the model and go to 3.

22. 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 graphs for cost-risk analysis, assessment of multivariate normality, and the two-sample T^2 test together with the validation procedure are illustrated.

A multivariate response simulation model representing an M/M/1 queueing system is considered and the arrival process is assumed to be part of the model. The simulation model is represented by a computerized model of M/M/1 with an arrival rate (a_r) of 0.8 and a service rate (s_r) of 1. Similarly, the real system is represented by a computerized model of M/M/1 with $a_r = 0.79$ and $s_r = 1$.

Terminating simulation is considered and the data are collected by using the method of replications. The data collection from the model and the system is done on an IBM 370 by using the multiplicative congruential random number generator $W_n = 7^5 W_{n-1} \pmod{2^{31} - 1}$. The initial (starting) conditions are assumed to be an empty system and the first arrival takes place at time zero.

It is assumed that there are two response variables (performance measures) of interest, namely, the average queue length for the first 300 customers (response variable 1) and the average waiting time in the system for the first 300 customers (response variable 2). The

steps of the procedure given in section 4 will be followed for validating the simulation model.

The experimental frame under which the validity of the simulation model is going to be tested with respect to its mean behavior is determined by the exponential service times with service rate s_r and the first-come first-served queue discipline. Assuming that the intended application of the model is to analyze the mean behavior of the system with respect to the performance measures chosen, the acceptable range of accuracy for the population means is specified as

$$\begin{aligned} | \mu_1^m - \mu_1^s | &\leq 0.38 \\ | \mu_2^m - \mu_2^s | &\leq 0.28 . \end{aligned} \tag{18}$$

Assuming that a trade-off analysis is desired, we go to step 4 to construct the schedules. An estimate of the common variance-covariance matrix is needed to construct the schedules. In a pilot run of the simulation model, ten independent observations are obtained by way of replication, each replication being for the first 300 customers, and the estimate of the common variance-covariance matrix is found as

$$\hat{\Sigma} = \begin{bmatrix} 2.9667 & 3.5584 \\ 3.5584 & 4.3010 \end{bmatrix} . \tag{19}$$

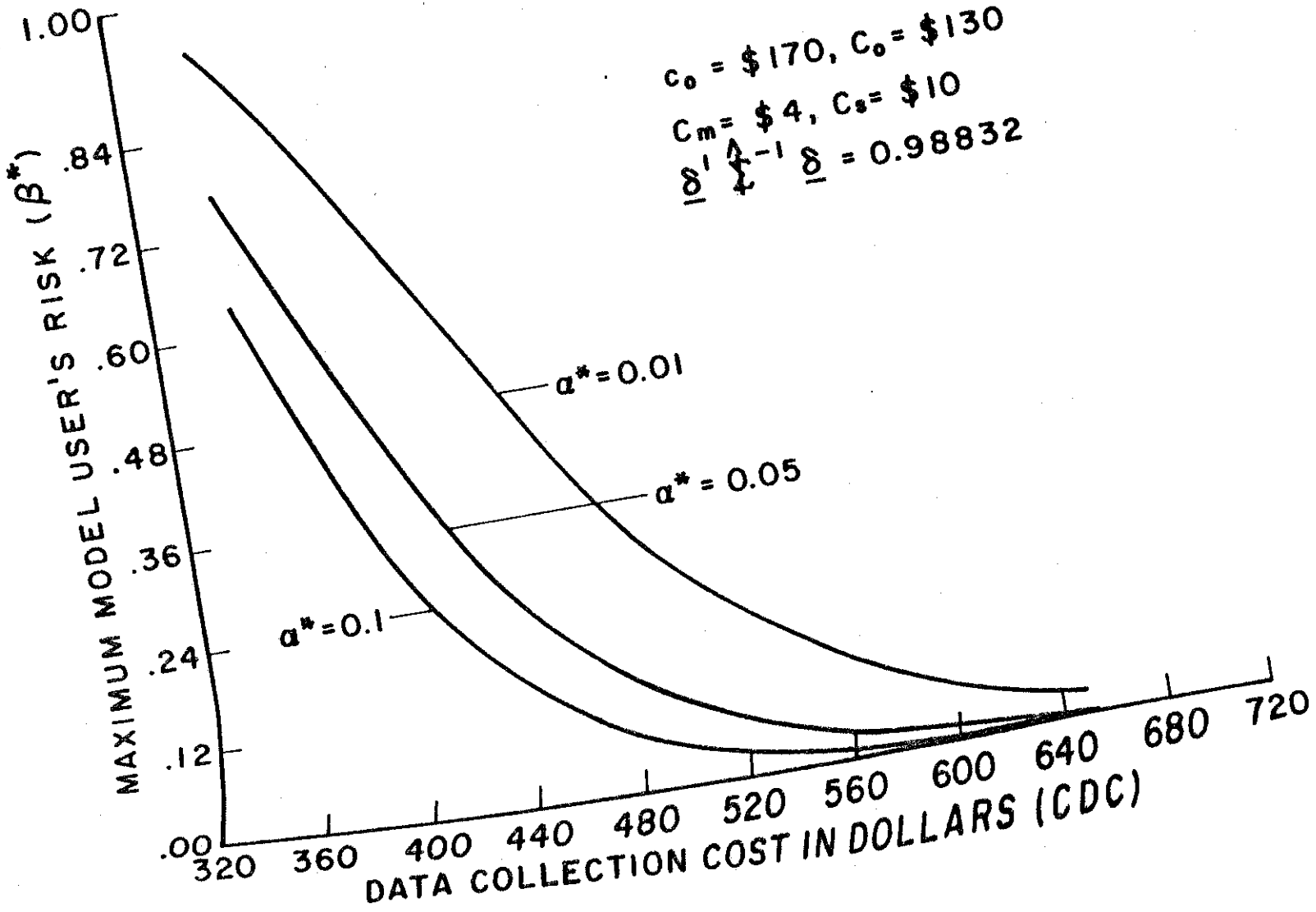
The overhead costs for statistical data collection by way of

replication for the model and for the system are estimated to be \$170 and \$130, respectively. It is estimated that the unit cost of collecting one independent observation (one replication) from each model response variable is \$2 and from the first and second system response variables it is \$4 and \$6, respectively. The procedure for constructing the schedules for the two-sample T^2 test, given in [5], is performed and the schedules are constructed.

Two questions of particular interest are: (i) what budget (B) and sample sizes (n,N) would be required for the given values of the following: (1) overhead data collection cost of the model (c_o), (2) overhead data collection cost of the system (C_o), (3) sum of the unit costs of data collection from the model (C_m), (4) sum of the unit costs of data collection from the system (C_s), (5) minimum model builder's risk (α^*), (6) maximum model user's risk (β^*), and (7) the acceptable range of accuracy ($\delta_j, j=1,\ldots,k$); (ii) what would be the maximum model user's risk, maximum model builder's risk, and the acceptable validity range for the given values of c_o, C_o, C_m, C_s, B, α^*, and $\underline{\delta}$?

In order to answer the first question, assuming that c_o = \$170, C_o = \$130, C_m = \$4, C_s = \$10, α^* = 0.05, β^* = 0.04, and $\underline{\delta}'$ = [0.38, 0.28] which give $\underline{\delta}'\hat{\Psi}^{-1}\underline{\delta}$=0.98832, Figure 1 is constructed by using the data contained in the schedules. In Figure 1, the relationships among the maximum model user's risk (β^*), minimum model builder's risk (α^*), and data collection cost (CDC = $c_o + C_o + n\,C_m + N\,C_s$) are shown for the given values of the parameters. The data collection cost is read from Figure 1 (or from the schedules) as \$550 for α^* = 0.05 and β^* = 0.04

FIGURE 1. Cost Versus Maximum Model User's Risk.



Thus, the necessary data collection budget B is \$550, and the sample sizes corresponding to c_o, C_o, C_m, C_s, and B are read from the schedules as $n^* = 25$ and $N^* = 15$. The acceptable validity range corresponding to these sample sizes is read from the operating characteristic curves in Figure 2 (or from the schedules) as $0 \leq \lambda \leq$ 9.265. Notice that for these sample sizes, β^* would be 0.018 and 0.136 for $\alpha^* = 0.1$ and 0.01, respectively.

In order to answer the second question, assuming that c_o = \$170, C_o = \$130, C_m = \$4, C_s = \$10, B = \$650, $\alpha^* = 0.1$, and $\underline{\delta}' = [0.38, 0.28]$, Figure 3 is constructed by using the data contained in the schedules. In Figure 3, 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 optimal sample sizes corresponding to c_o, C_o, C_m, C_s, and B are read from the schedules as $n^* = 35$ and $N^* = 21$. The corresponding λ^* is calculated as $n^* N^* \underline{\delta}' \hat{\Psi}^{-1} \underline{\delta} / (n^* + N^*) = 12.972$. Then, the value of the maximum model user's risk β^* is read from Figure 3 (or from the schedules) for $\alpha^* = 0.01$ as 0.029. Thus, we get $0 \leq \beta < 0.029$, 0.1 $\leq \alpha \leq$ 0.971 and $0 \leq \lambda \leq 12.972$. Notice that for $n^* = 35$ and $N^* = 21$, we could also get $\beta^* = 0.005$ and 0.002 for $\alpha^* = 0.05$ and 0.1, respectively.

FIGURE 2. Operating Characteristic Curves for $n^* = 25$ and $N^* = 15$.

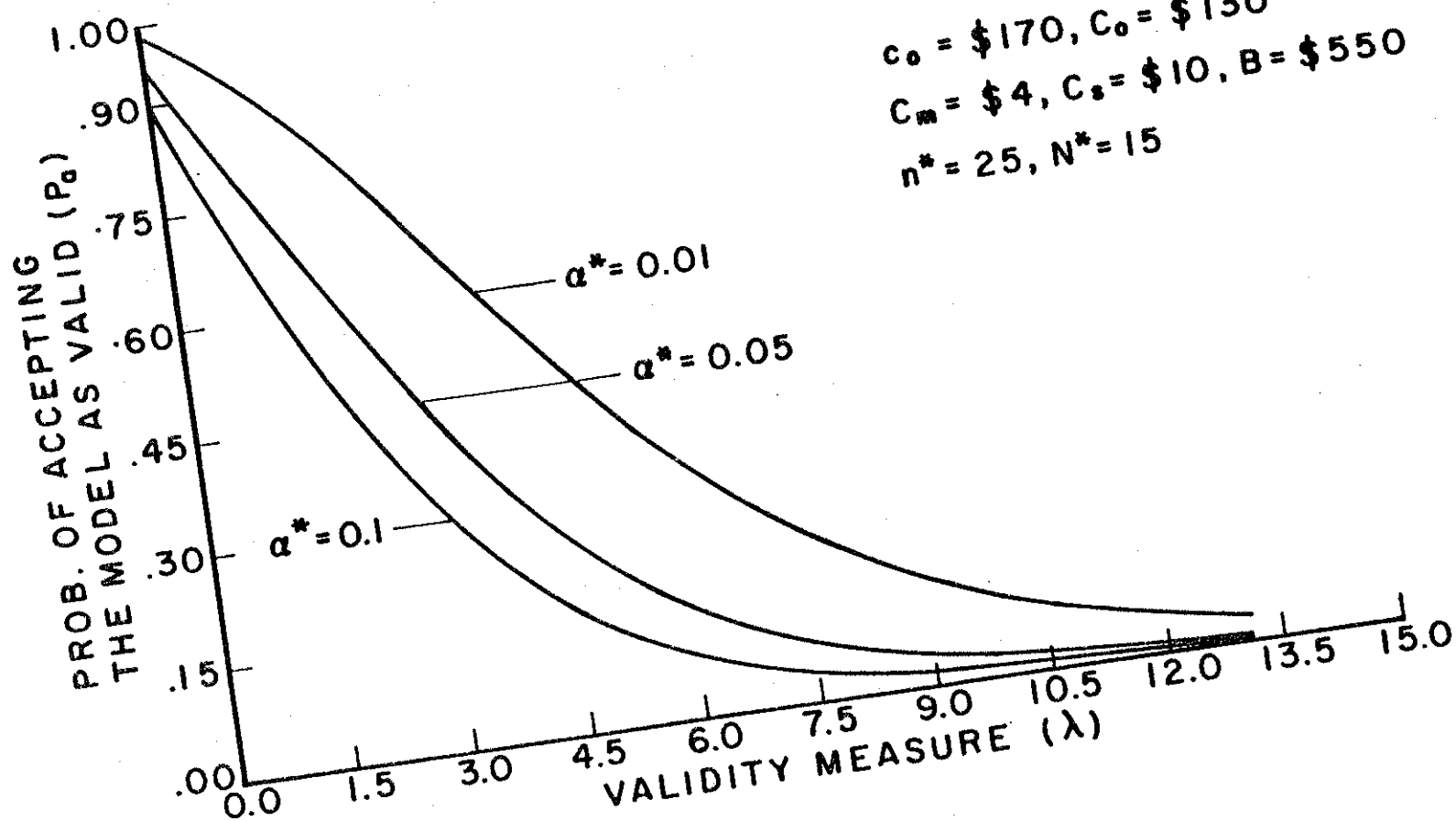
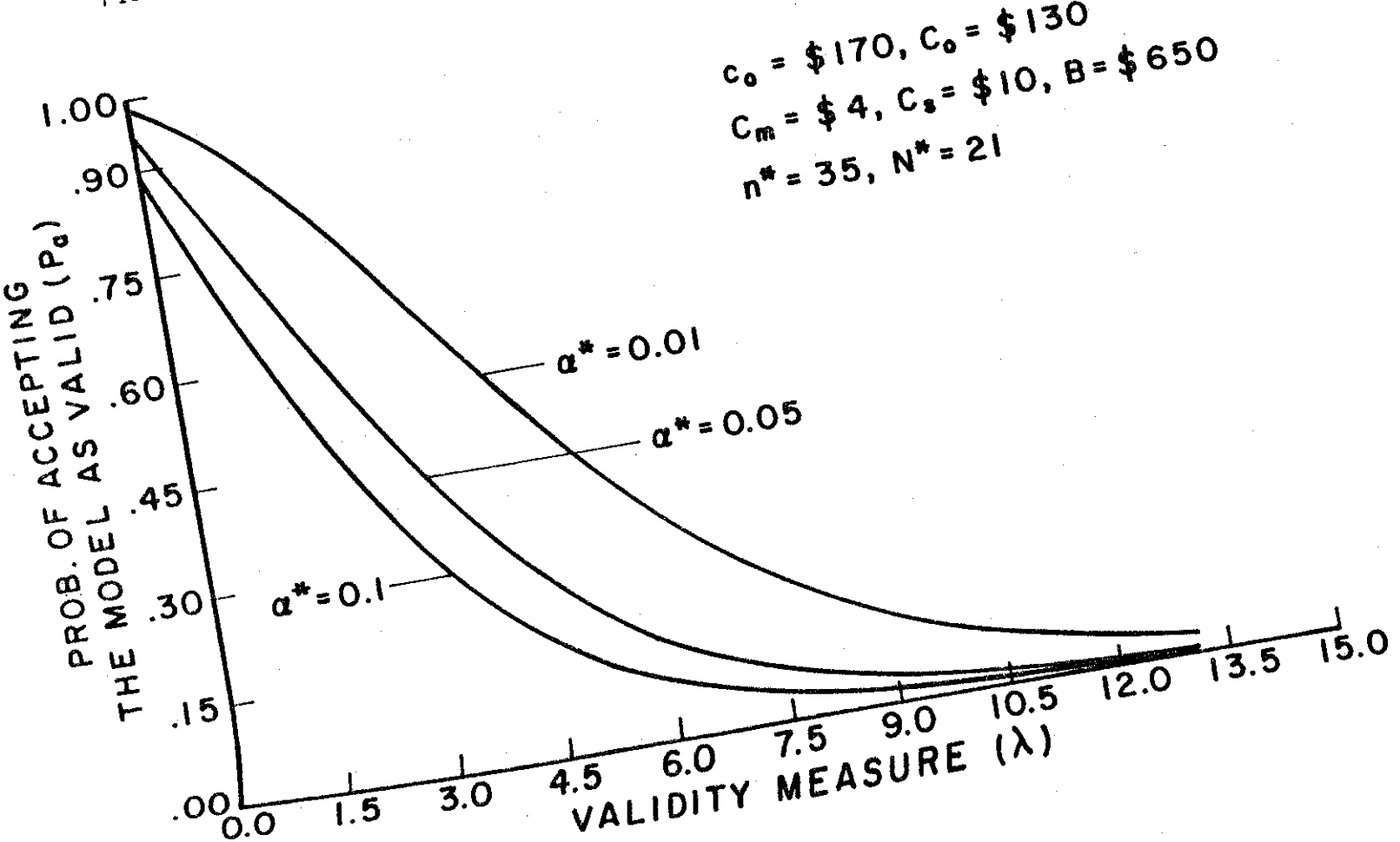


FIGURE 3. Operating Characteristic Curves for $n^* = 35$ and $N^* = 21$.



If the answer to the first question were considered to be satisfactory, the optimal sample sizes $n^* = 25$ and $N^* = 15$ would be chosen. If the answer to the second question were considered to be satisfactory, the optimal sample sizes $n^* = 35$ and $N^* = 21$ would be chosen. For illustrative purposes, the remaining steps of the validation procedure will be carried out for $n^* = 25$ and $N^* = 15$.

In step 6 of the procedure, the simulation model and the system are replicated 25 and 15 times, respectively, for 300 customers in each replication. The data obtained are presented in Table 4. Setting $j = 1$, we go to Step 7 and apply the univariate normality test in Table 1 to system response variable 1. The results of this and the other tests to be conducted are presented in Table 5. The test for system response variable 1 indicates some normality with an approximate significance level of 0.0579. However, the normality of its distribution should be improved to enhance better multivariate normality. Therefore, the power transformation with parameter -0.126, as suggested by the test, will be used to transform the original observations to improve the normality of its distribution. The location parameter is shifted by adding 2 to avoid negative values. Thus, we go back to Step 7 and apply the univariate normality test in Table 1 to the transformed system response variable 1. The results as shown in Table 5 indicate univariate normality with an approximate significance level of 0.9396. So, we come back to Step 7 again after setting $j = 2$ in Step 8. This time, the univariate normality test in Table 1 is applied to system response variable 2. The results given in Table 5 indicate some normality with an approximate significance level of 0.0721. However, the normality of its distribution should be

TABLE 4. Data Collected for Validation ($N^* = 15$ and $n^* = 25$)

SYSTEM		MODEL	
Var. 1	Var. 2	Var. 1	Var. 2
2.0273	3.6520	3.5400	5.5352
1.9247	3.4844	1.6440	3.0283
3.3547	4.9845	1.2820	2.4976
3.0529	4.9914	3.3268	5.0149
1.6762	3.0550	1.4878	2.8773
2.4553	4.0731	4.8299	7.3234
4.8960	6.9151	2.4047	4.2181
1.3033	2.7547	2.1970	3.6049
2.9373	4.9382	1.3670	2.7532
2.7483	4.1511	1.1619	2.3480
2.8061	4.5167	3.6845	5.8066
3.8587	5.7909	3.0052	5.2354
1.6766	3.0931	1.8239	3.2580
6.3030	8.6714	4.4252	6.5992
4.3003	6.7946	3.1672	4.7395
		2.6665	4.1884
		2.0531	3.6452
		2.9899	4.6400
		2.8934	4.4813
		2.9391	4.8431
		2.7012	4.2025
		1.6411	3.2391
		7.4164	10.0199
		1.1885	2.3009
		5.1204	7.0183

TABLE 5. Normality Tests and Transformations for $n^* = 25$ and $N^* = 15$.

Univariate Power Transformation Tests

Response Variable		Transformed by	$\hat{\theta}$	$2\{L_{max}(\hat{\theta})-L_{max}(1)\}$	Approximate γ	Univariate Normal?
System	1	--	-0.126	3.661348	0.0579	Improve
	2	--	-0.472	3.341516	0.0721	Improve
Model	1	$(y_1^{-0.126}-1)/(-0.126)+2$	0.840	0.007044	0.9369	Yes
	2	$(y_2^{-0.472}-1)/(-0.472)+2$	0.370	0.014038	0.9074	Yes
	1	$(x_1^{-0.126}-1)/(-0.126)+2$	0.816	0.020935	0.8911	Yes
	2	$(x_2^{-0.472}-1)/(-0.472)+2$	1.783	0.065094	0.8142	Yes

Multivariate Power Transformation Tests

Response Variable		$\hat{\theta}_1$	$\hat{\theta}_2$	$2\{L_{max}(\hat{\theta}_1,\hat{\theta}_2)-L_{max}(1,1)\}$	Approximate γ	Multivariate Normal?
System	1	0.760	-0.120	1.392487	0.4995	Yes
	2					
Model	1	1.010	2.210	5.280090	0.0757	Yes
	2					

improved to enhance better multivariate normality. Therefore, the power transformation with parameter -0.472, as suggested by the test, will be used to transform the original observations to improve the normality of its distribution. The location parameter is shifted by adding 2 to avoid negative values. Thus, we go back to Step 7 and test the univariate normality of system response variable 2. The results given in Table 5 indicate univariate normality with an approximate significance level of 0.9074. Thus, we go to Step 11 after setting j = 1 in Step 8, and apply the multivariate normality test in Table 2 to the transformed system response variables. The results as shown in Table 5 indicate multivariate normality with an approximate significance level of 0.4995.

Now, we go to Step 12 and transform the observations of model response variable 1 with the same transformation applied to system response variable 1. Then, in Step 13, apply the univariate normality test to the transformed model response variable 1. The results given in Table 5 indicate univariate normality with an approximate significance level of 0.8911. Thus, we come back to Step 12 again after setting j = 2 in Step 14. This time, model response variable 2 is transformed by the same transformation applied to system response variable 2, and the normality of its distribution is tested. The results of the test given in Table 5 indicate univariate normality with an approximate signficance level of 0.8142. In step 17, the multivariate normality test in Table 2 is applied to the transformed model response variables. The results of the test given in Table 5 indicate multivariate normality with an approximate significance level of 0.0757.

In Step 19, the equality of the variance-covariance matrices of the transformed model and system response variables is tested by using Table 3. The following values are obtained in the test: M=2.1885, U=0.9373, G_2 = 0.0041, and V = 0.3124. The test statistic F is found to be 0.6837 which is less than $F_{0.1;3,\infty}$ = 2.08 and the equality of the variance-covariance matrices is accepted at the significance level of 0.1.

In step 20, the two-sample T^2 test is applied to test the equality of the population means. As a result of the two-sample T^2 test, the test statistic T^2 is found to be 0.9773 which is less than 6.738 at $\alpha^* = 0.05$ and the equality of the population means is accepted at $\alpha^* = 0.05$. Finally, in Step 22, 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.

6. CONCLUSIONS

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

Some remedial measures are given to satisfy the assumptions underlying the two-sample T^2 test, namely, independence, multivariate normality, and equality of variance-covariance matrices. Construction of the graphs for cost-risk analysis, assessment of multivariate normality, and the two-sample T^2 test together with the validation procedure are illustrated by an example. In this example, a terminating simulation model representing an M/M/1 queueing system with two performance measures is considered.

REFERENCES

[1] Andrews, D.F., Gnanadesikan, R., and Warner, J.L., "Transformations of Multivariate Data," Biometrics, Vol. 27, (Dec. 1971), pp. 825-840.

[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., and Cox, D.R., "An Analysis of Transformatons," J.R. Statist. Soc. B, Vol. 26, (1964), pp. 211-252.

[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] Gnanadesikan, R., Methods for Statistical Data Analysis of Multivariate Observations, John Wiley & Sons, New York, 1977.

[12] Heidelberger, P., and Lewis, P.A.W., "Regression-Adjusted Estimates for Regenerative Simulations, with Graphics," Communications of ACM, Vol. 24, No. 4 (April 1981), pp. 260-273.

[13] Hoyle, M.H., "Transformations - An Introduction and a Bibliography," Int. Stat. Review, Vol. 41, No. 2 (1973), pp. 203-223.

[14] Ito, K., and Schull, W.J., "On the Robustness of the T_o^2 Test in Multivariate Analysis of Variance When Variance-Covariance Matrices are not Equal," Biometrika, Vol. 51, Parts 1 and 2, (June 1964), pp. 71-82.

[15] 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.

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

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

[18] Neter, J., and Wasserman, W., Applied Linear Statistical Models, Richard D. Irwin, Inc., Homewood, IL, 1974.

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

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

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

[22] Schmeiser, B., "Batch Size Effects in the Analysis of Simulation Output," Working Paper, School of Industrial Engineering, Purdue University, June 1980.

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

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