SOME EXAMPLES OF SIMULATION MODEL VALIDATION USING HYPOTHESIS TESTING

Osman Balci

Virginia Polytechnic Institute & State University
Blacksburg, VA 24061

Robert G. Sargent

Syracuse University
Syracuse, NY 13210

ABSTRACT

The use of hypothesis testing with cost-risk analysis is illustrated for simulation model validation by two examples. In the first example, Hotelling's two-sample T^2 test with cost-risk analysis is used for illustrating the validation of a multivariate response self-driven steady-state simulation model representing a single server M/M/1 queueing system. In the second example, the validation of a multivariate response trace-driven terminating simulation model representing an M/M/1 system is illustrated by the use of Hotelling's one-sample T^2 test with cost-risk analysis.

1. INTRODUCTION

Validation, being one of the most important steps in the development of a computerized simulation model, is usually referred to as "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" [37].

The validity of a simulation model is usually tested for different sets of experimental conditions 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.

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 [3]. 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 [24] is a technique which uses random numbers in sampling from distributions or stochastic processes. Trace-driven (or retrospective [32]) simulation is a technique which combines measurement and simulation by using the actual data collected on the system as the model input [24, 40]. There are basically two types of simulation models with regard to analysis of the output: steady-state and terminating simulation models [15, 25]. 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" [25]. 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 dengenerate random variable, is the time that a specified event E occurs" [25].

The purpose of this paper is to give some examples to illustrate the use of statistical hypothesis testing for simulation model validation. In the next section, validation techniques and statistical techniques proposed for validation will be tabulated and the use of hypothesis testing with cost-risk or sample size-risk analysis will be introduced for model validation. Examples will be given in section 3 and conclusions will be stated in section 4.

2. VALIDATION TECHNIQUES AND HYPOTHESIS TESTING

The existing literature on simulation model validation [3] generally falls into two broad areas, namely, validation techniques and statistical techniques proposed for validation as shown in Tables 1 and 2 that also contain the related reference number(s) for each technique. Some of the validation techniques in Table 1 can also be used statistically by introducing a statistical test.

In using statistical hypothesis testing to test the validity of a simulation model under a given set of experimental conditions and for an acceptable range

Proceedings of the 1982
Winter Simulation Conference
Highland * Chao * Madrigal, Editors

82CH1844-0/82/0000-0621 $00.75

TABLE 1. Validation Techniques.

Event Validity	[20]
Face Validity	[20]
Field Tests	[39, 45]
Graphical Comparisons	[9, 14, 26, 47]
Historical Methods	[30]
Hypothesis Validity	[20]
Internal Validity	[20]
Multistage Validation	[30]
Predictive Validation	[11]
Schellenberger's Criteria	[19, 36]
Sensitivity Analysis	[20, 27, 28, 45]
Submodel Testing	[34, 35]
Traces	[34, 35]
Turing Test	[29, 38, 43, 45]

TABLE 2. Statistcal Techniques Proposed for Validation.

Analysis of Variance	[30]
Confidence Intervals	[5, 39]
Factor Analysis	[8]
Hotelling's T^2 Tests	[4, 6, 7, 39]
Multivariate Analysis of Variance (MANOVA)	[18]
- Standard MANOVA	
- Permutation Methods	
- Nonparametric Ranking Methods	
Nonparametric Goodness-of-fit Tests	[16, 30]
- Kolmogorov-Smirnov Test	
- Cramer-Von Mises Test	
- Chi-square Test	
Nonparametric Tests of Means	[39]
- Mann-Whitney-Wilcoxon Test	
- Analysis of Paired Observations	
Regression Analysis	[1, 8, 21]
Theil's Inequality Coefficient	[23, 31, 42]
Time Series Analysis	
- Spectrial Analysis	[13, 17, 21, 22, 45, 46]
- Correlation Analysis	[46]
- Error Analysis	[10, 44]
t-Test	[39, 41]

of accuracy consistent with the intended application of the model, we have the following hypotheses:

H_0: Model is valid for the acceptable range of accuracy under the given set of experimental conditions.

(1)

H_1: Model is invalid for the acceptable range of accuracy under the given set of experimental conditions.

There are two possibilities for making a wrong decision in statistical hypothesis testing. The first one, type I error, is accepting the alternative hypothesis (H_1) when the null hypothesis (H_0) is actually true, and the second one, type II error, is accepting the null hypothesis when the alternative hypothesis is actually true. The probability of making the first type of wrong decision is called <u>model builder's risk</u> (α) and the probability of making the second type of wrong decision is called <u>model user's risk</u> (β) [2, 4].

In validation, the model user's risk is extremely important and must be kept at a small value. Model user's risk can be decreased by increasing the sample sizes of observations and/or the model builder's risk. However, increasing the sample sizes will increase the cost of data collection. In those cases where the data collection cost is a major factor to consider, a cost-risk trade-off analysis becomes necessary. Otherwise, if the cost of data collection is not a relatively important factor, then a sample size-risk analysis can be performed without considering the data collection cost. In any case, model user's risk must be kept at a small value by choosing appropriate values for the sample sizes and model builder's risk.

In the next section, two examples will be presented by using the validation and cost-risk analysis procedures that are given in [4, 6, 7].

3. EXAMPLES

Two examples are given, in this section, to illustrate the use of hypothesis testing with cost-risk analysis for simulation model validation [4, 6, 7]. In the first example, the validation of a multivariate response self-driven steady-state simulation model is illustrated by using Hotelling's two-sample T^2 test. In the second example, we illustrate the validation of a multivariate response trace-driven terminating simulation model by using Hotelling's one-sample T^2 test.

In each of the two examples given in this section, the random variate generation is done on an IBM 370 by using the Inverse Transform Method [12] and the multiplicative congruential random number generator $W_n = 7^5 W_{n-1} (\text{mod } 2^{31}-1)$. The simulation programs are coded in FORTRAN and the initial (starting) conditions are assumed to be an empty system and the first arrival takes place at time zero.

3.1 Self-Driven Steady-State Simulation

A computerized self-driven steady-state simulation model of M/M/1 queueing system with the arrival rate of customers per unit of time (a_r) = 0.79 and the service rate of the server per unit of time (s_r) = 1 is treated as being the real system for illustrating the validation of a self-driven steady-state simulation model with a_r = 0.8 and s_r = 1. For the purpose of study, it is assumed that the arrival process is part of the model and there are two response variables (performance measures) of interest, namely, the utilization of the server (response variable 1) and the average waiting time in the system (response variable 2). The steps of the procedure given in [6] will be followed for validating the model by using the two-sample T^2 test.

The set of experimental conditions 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

$$|\mu_1^m - \mu_1^s| \leq 0.048$$
$$|\mu_2^m - \mu_2^s| \leq 0.28 \tag{2}$$

where (μ_1^m, μ_2^m) and (μ_1^s, μ_2^s) are the population means of the first and second model and system response variables, respectively.

Assuming that a cost-risk trade-off analysis is desired, we need to construct the schedules for which an estimate of the common variance-covariance matrix is required. Therefore, ten independent observations are obtained in a pilot run from each of the model response variables by using the method of batch means [12, 33]. The data collected for the first 600 customers are deleted and each batch is composed of 300 customers representing one independent observation. The estimate of the common variance-covariance matrix is found as

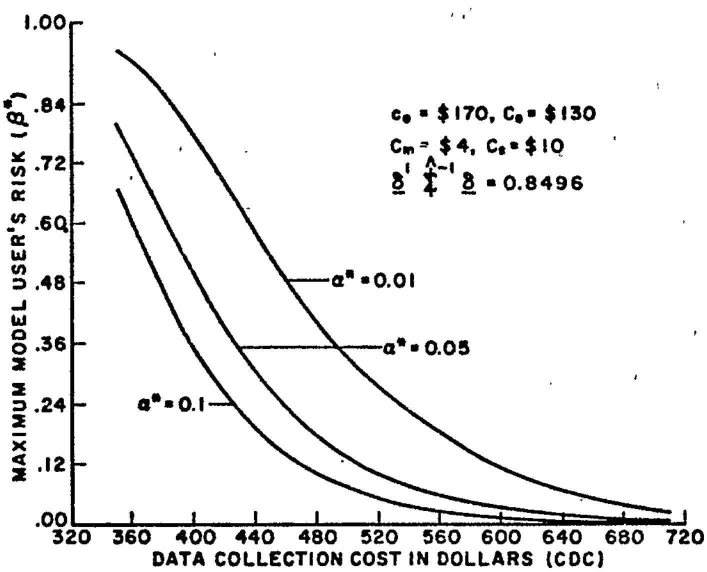
$$\hat{\Sigma} = \begin{bmatrix} 0.0052 & 0.2479 \\ 0.2479 & 21.7395 \end{bmatrix}. \tag{3}$$

The overhead costs for statistical data collection by way of batch means method for the model and for the system are assumed to be \$170 and \$130, respectively. It is assumed that the unit cost of collecting one independent observation (one batch) 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 [4], is performed and the schedules are constructed.

In determining the sample sizes and the risks, the following two questions of particular interest will be considered: (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_0), (2) overhead data collection cost of the system (C_0), (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 (δ_j, j = 1,2), (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_0, C_0, C_m, C_s, B, α^*, and $\underline{\delta}$?

In order to answer the first question, assuming that c_0 = \$170, C_0 = \$130, C_m = \$4, C_s = \$10, $\alpha^* = 0.1$, $\beta^* = 0.0438$, and $\underline{\delta}' \doteq [0.048, 0.28]$ which give $\underline{\delta}'\hat{\Sigma}^{-1}\underline{\delta} \doteq 0.8469$, Figure 1 is constructed by using the data contained in the schedules. The data collection cost is read from Figure 1 (or from the schedules) as \$530 for $\alpha^* = 0.1$ and $\beta^* = 0.0438$. Thus, the necessary data collection budget B is \$530, and the sample sizes corresponding to c_0, C_0, C_m, C_s, and B are read from the schedules as $n^* = 20$, $N^* = 15$. The acceptable validity range corresponding to these sample sizes is read from the operating characteristic curves in Figure 2 (or from the schedules) as $0 \leq \lambda \leq 7.282$. Notice that for these sample sizes, β^* would be 0.0876 and 0.2546 for $\alpha^* = 0.05$ and 0.01, respectively.

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



In order to answer the second question, assuming that c_0 = \$170, C_0 = \$130, C_m = \$4, C_s = \$10, B = \$710, $\alpha^* = 0.1$, and $\underline{\delta}'$ = [0.048,0.28], Figure 3 is constructed by using the data in the schedules. The optimal sample sizes corresponding to c_0, C_0, C_m, C_s, and B are read from the schedules as $n^* = 40$ and $N^* = 25$. The corresponding λ^* is calculated as $n^*N^*\underline{\delta}'\hat{\Sigma}^{-1}\underline{\delta}/(n^*+N^*) = 13.071$. Then, the value of the maximum model user's risk β^* is read from Figure 3 (or from the schedules) for $\alpha^* = 0.1$ as 0.0018. Thus, we get $0 \leq \beta \leq 0.0018$, $0.1 \leq \alpha \leq 0.9982$, and $0 \leq \lambda \leq 13.071$. Notice that for $n^* = 40$ and $N^* = 25$, we could also get $\beta^* = 0.0047$ and 0.025 for

$\alpha^* = 0.05$ and 0.01, respectively.

FIGURE 2. Operating Characteristic Curve for $n^* = 20$ and $N^* = 15$.

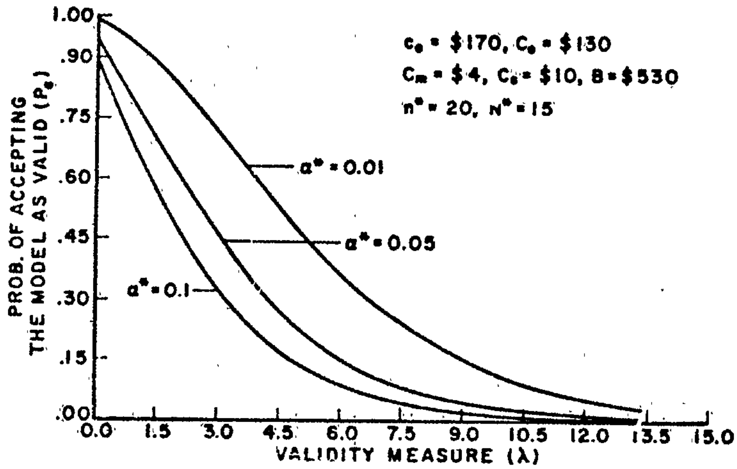
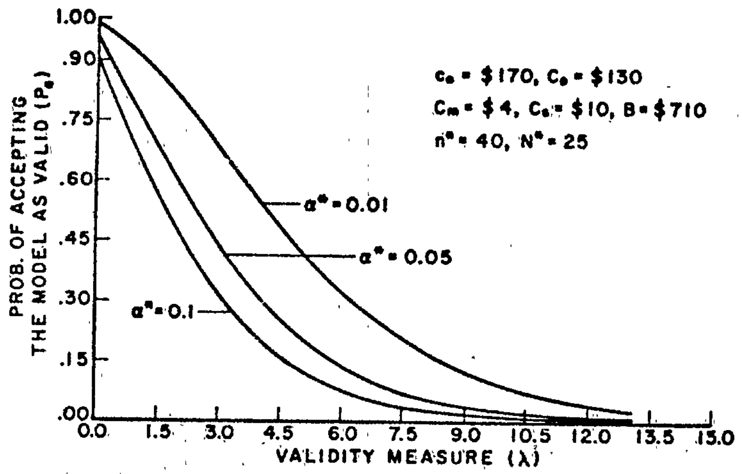


FIGURE 3. Operating Characteristic Curves for $n^* = 40$ and $N^* = 25$.



For illustrative purposes, the remaining steps of the validation procedure will be carried out for $n^* = 25$ and $N^* = 15$.

The simulation model and the system are run for 20 and 15 independent batches, respectively, in steady-state after deleting the data collected during the transient period of 600 customers. Each batch is composed of 300 customers. The data obtained are given in Table 3.

The Steps 7 through 17 of the validation procedure are followed and the results of the univariate and multivariate normality tests and the transformations are given in Table 4. The original system and model response variable 1 are found univariate normal with approximate significance level of 0.6075 and 0.8931,

TABLE 3. Data Collected for Validation.

SYSTEM		MODEL	
Var. 1	Var. 2	Var. 1	Var. 2
0.78	13.84	0.83	27.90
0.80	5.59	0.78	5.64
0.75	3.25	0.75	3.78
0.87	6.10	0.91	6.68
0.73	2.64	0.78	3.25
0.72	2.88	0.80	2.89
0.91	7.42	0.76	3.98
0.75	3.18	0.79	4.08
0.77	4.44	0.67	2.01
0.83	5.67	0.87	3.71
0.81	7.16	0.82	3.90
0.81	3.12	0.75	4.26
0.81	4.09	0.82	6.37
0.80	4.66	0.79	3.50
0.84	6.56	0.77	3.28
		0.88	12.29
		0.75	5.39
		0.92	4.97
		0.66	2.88
		0.75	3.32

respectively. The original system response variable 1 and the transformed system response variable 2 are found multivariate normal with approximate significance level of 0.2753. Similarly, the original model response variable 1 and the transformed model response variable 2 are also found multivariate normal with approximate significance level of 0.2382.

Following the procedure, the equality of the variance-covariance matrices of the model and system response variables is tested. The test statistic F is found to be 0.4829 which is less than $F_{0.1;3,\infty} = 2.08$ and the equality is accepted at the significance level of 0.1. Then, 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 0.276 which is less than 5.122 at $\alpha^* = 0.1$ and the equality of the population means is accepted at $\alpha^* = 0.1$. Finally, it is concluded that the model is valid with respect to the validity measures for the acceptable range of accuracy under the given set of experimental conditions.

3.2 Trace-Driven Terminating Simulation

A multivariate response trace-driven terminating simulation model (M/M/1, $a_r = 0.6$, $s_r = 0.99$) representing a single server M/M/1 queueing system ($a_r = 0.6$, $s_r = 1$) has two response variables (performance measures) of interest, namely, the average queue length for the first 500 customers (response variable 1) and the average waiting time in the system for the first 500 customers (response variable 2). The steps of the procedure given in [7] will be fol-

TABLE 4. Normality Tests and Transformations for $n^* = 20$ and $N^* = 5$.

Univariate Power Transformation Tests

	Response Variable	Transformed by	$\hat{\theta}$	$2\{L_{max}(\hat{\theta})-L_{max}(1)\}$	Approximate γ	Univariate Normal?
System	1	--	-1.048	0.3032	0.6075	Yes
System	2	--	-0.594	8.9389	<0.005	No
System	2	$(y_2^{-0.594}-1)/(-0.594)+2$	0.100	0.0315	0.8727	Yes
Model	1	--	0.678	0.0197	0.8931	Yes
Model	2	$(x_2^{-0.594}-1)/(-0.594)+2$	-1.431	0.8702	0.3800	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, 2	-4.100	-2.700	2.6304	0.2753	Yes
Model	1, 2	-0.801	-2.457	2.9148	0.2382	Yes

lowed for validating the simulation model by using the one-sample T^2 test.

The experimental conditions 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 is specified as

$$|\mu_1^d| \leq 0.154$$
$$|\mu_2^d| \leq 0.28 \qquad (4)$$

where μ_1^d is the population mean of the differences between the paired observations on the first model and system response variables, average queue length for the first 500 customers; μ_2^d is the population mean of the differences between the paired observations on the second model and system response variables, average waiting time in the system for the first 500 customers.

Assuming that a cost-risk trade-off analysis is desired, we need to construct the schedules for which an estimate of the variance-covariance matrix is required. Therefore, five independent paired observations are obtained in pilot runs on the model and system response variables by way of replicating the trace-driven simulation model with the same trace-data that drive the real system. The trace driven simulation is obtained by using the same sequence of random numbers to generate the same arrival pattern to the model and to the system and by using another sequence of random numbers to generate the same pattern of service times in the model and in the system. The estimate of the variance-covariance matrix of differences between the paired observations on the model and system response variables is found as

$$\hat{\Sigma}_d = \begin{bmatrix} 0.2162 & 0.4147 \\ 0.4147 & 0.7959 \end{bmatrix}. \qquad (5)$$

The overhead costs for statistical data collection by way of replication for the model and for the system are estimated to be \$200 and \$1400, respectively. It is estimated that the unit cost of collecting one independent observation (one replication) from each model response variable is \$5 and from the first and second system response variables it is \$15 and \$20, respectively. The procedure for constructing the schedules for the one-sample T^2 test, given in [7], is performed and the schedules are constructed.

A question of particular interest is "what would be the maximum model user's risk, maximum 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 that c_0 = \$200, C_0 = \$1400, C_m = \$10, C_s = \$35, B = \$2300, $\alpha^* = 0.05$, and $\underline{\delta}' = [0.154, 0.28]$ which give $\underline{\delta}'\hat{\Sigma}_d^{-1}\underline{\delta} = 0.80137$, first the optimal sample size N is read from the schedules corresponding to B = \$2300 as 15 and then Figures 4 and 5 are constructed by using the data contained in the schedules. In Figure 4, the relationships among maximum model user's risk (β^*), mini-

FIGURE 4. Cost Versus Maximum Model User's Risk.

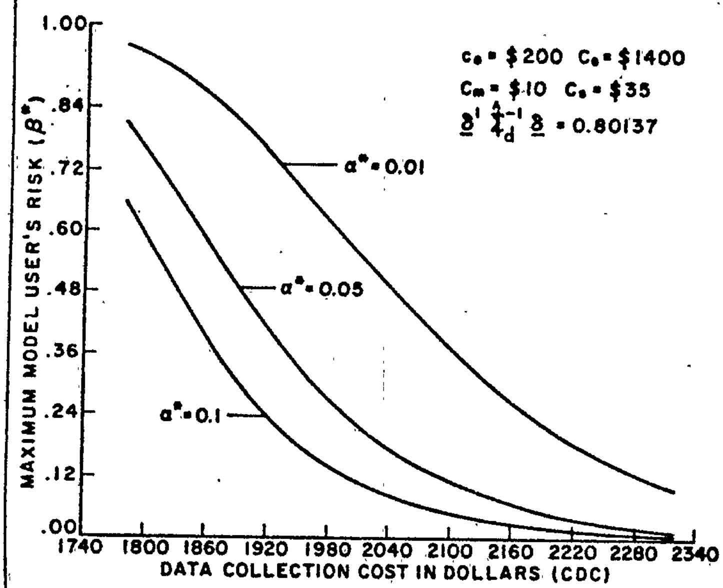
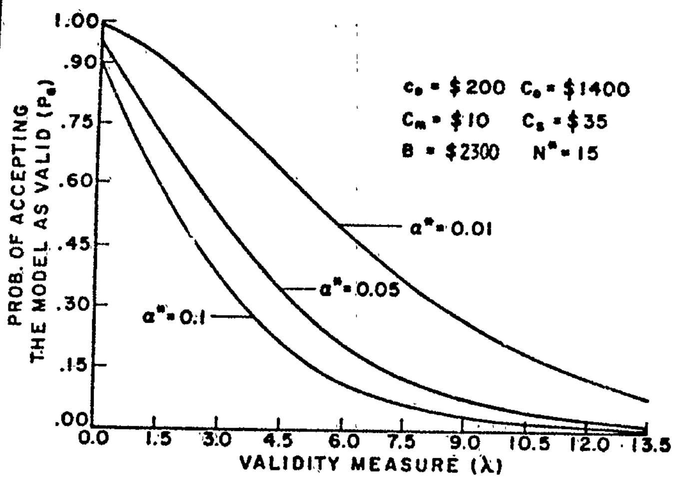


FIGURE 5. Operating Characteristic Curves for N^* = 15.



mum model builder's risk (α^*), and data collection cost (CDD = $c_0 + C_0 + (C_m + C_s)N^*$) are shown for the given values of the parameters. In Figure 5, 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 $N^* \underline{\delta}' \hat{\Sigma}_d^{-1} \underline{\delta}$ = 12.021. Then, the value of the maximum model user's risk β^* is read from Figure 5 (or from the schedules) for $\alpha^* = 0.05$ and $\lambda^* = 12.021$ as 0.0256. Thus, we get $0 \le \beta \le 0.0256$, $0.05 \le \alpha \le 0.9744$, and $0 \le \lambda \le 12.021$. Assuming that these values are satisfactory, we choose $N^* = 15$.

The simulation model and the system are replicated 15 times for 500 customers in each replication, by using the same sequence of random numbers for the model and for the system. The paired observations obtained and the differences between them are presented in Table 5.

Continuing with the validation procedure in [7], the Box-Cox transformation test for univariate normality [7] is applied to the differences between the paired observations on each of the two model and system response variables. The results of the tests are presented in Table 6. After achieving reasonable univariate normality, the multivariate normality of the differences is tested by using the transformation test for multivariate normality [7]. The results of this test are also given in Table 6. As shown in the table, multivariate normality is achieved at an approximate significance level of 0.8724.

TABLE 5. Data Collected for Validation.

MODEL		SYSTEM		DIFFERENCE	
Var. 1	Var. 2	Var. 1	Var. 2	Var. 1	Var. 2
0.691	1.996	0.946	2.627	-0.255	-0.631
0.945	2.561	0.744	2.189	0.201	0.372
0.909	2.426	0.901	2.554	0.008	-0.128
0.788	2.373	0.774	2.338	0.014	0.035
1.003	2.544	1.149	2.934	-0.146	-0.390
1.271	3.093	0.950	2.454	0.321	0.639
1.025	2.885	0.928	2.582	0.097	0.303
1.308	3.257	0.629	2.017	0.679	1.240
1.373	3.083	1.012	2.685	0.361	0.398
1.126	2.844	0.857	2.339	0.269	0.505
0.964	2.531	0.811	2.324	0.153	0.207
1.128	2.603	0.799	2.138	0.329	0.465
0.793	2.347	0.560	1.909	0.283	0.438
1.301	2.987	0.644	2.082	0.657	0.905
0.835	2.390	1.149	2.848	-0.314	-0.458

Following the procedure, the one-sample T^2 test is applied to test the validity. As a result, the test statistic T^2 is found to be 7.52 which is less than 8.23 at $\alpha^* = 0.05$ and the validity is accepted at $\alpha^* = 0.05$. Finally, it is concluded that the model is valid with respect to the validity measure for the acceptable range of accuracy under the given set of experimental conditions.

TABLE 6. Normality Tests.

M/M/1 Terminating Trace-Driven Model ($a_r = 0.6$, $s_r = 0.99$, $N^* = 15$) M/M/1 System ($a_r = 0.6$, $s_r = 1$, $N^* = 15$)					
Univariate Power Transformation Tests					
Difference on Response	$\hat{\theta}$		$2\{L_{max}(\hat{\theta})-L_{max}(1)\}$	Approximate γ	Univariate Normal?
1	0.856		0.0206	0.8916	Yes
2	0.951		0.0077	0.9341	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 2	0.772	0.981	0.2779	0.8724	Yes

4. SUMMARY

Two examples are presented to illustrate the use of hypothesis testing with cost-risk analysis for the validation of two types of simulation models. As the first type, a self-driven steady-state simulation model with two performance measures is considered and Hotelling's two-sample T^2 test with cost-risk analysis is used for the illustration. As the second type, a trace-driven terminating simulation model with two performance measures is considered and Hotelling's one-sample T^2 test with cost-risk analysis is used for the illustration.

BIBLIOGRAPHY

[1] Aigner, D.J. (1972), "A Note on Verification of Computer Simulation Models," Management Science, Vol. 18, No. 11, pp. 615-619.

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

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

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

[5] Balci, O. and R.G. Sargent (1981b), "A Methodology for Validating Multivariate Response Simulation Models by Using Simultaneous Confidence Intervals," Working Paper #81-009, Dept. of I.E.&O.R., Syracuse University.

[6] Balci, O. and R.G. Sargent (1982a), "Validation of Multivariate Response Simulation Models Using Hotelling's Two-Sample T^2 Test," to appear in Simulation.

[7] Balci, O. and R.G. Sargent (1982b), "Validation of Multivariate Response Trace-Driven Simulation Models," Working Paper #82-002, Dept. of I.E.&O.R., Syracuse University.

[8] Cohen, K.J. and R.M. Cyert (1961), "Computer Models in Dynamic Economics," Economics Quarterly, Vol. 75, pp. 112-127.

[9] Cyert, R.M. (1966), "A Description and Evaluation of Some Firm Simulation," Proc. IBM Scientific Computing Symposium on Simulation Models and Gaming, IBM, White Plains, NY.

[10] Damborg, M.J. and L.F. Fuller (1976), "Model Validation Using Time and Frequency Domain Error Measures," ERDA Report 76-152.

[11] Emshoff, J.R. and R.L. Sisson (1970), Design and Use of Computer Simulation Models, MacMillan, New York.

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

[13] Fishman, G.S. and P.J. Kiviat (1967), "The Analysis of Simulation Generated Time Series," Management Science, Vol. 13, No. 7, pp. 525-557.

[14] Forrester, J.W. (1961), Industrial Dynamics, MIT Press, Cambridge.

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

[16] Gafarian, A.V. and J.E. Walsh (1969), "Statistical Approach for Validating Simulation Models by Comparison with Operational Systems," Proc. Fourth International Conference on Operations Research, John Wiley & Sons, New York, pp. 702-705.

[17] Gallant, A.R., T.M. Gerig, and J.W. Evans (1974), "Time Series Realizations Obtained According to an Experimental Design," J. American Statistical Association, Vol. 69, No. 347, Applications Section.

[18] Garratt, M. (1974), "Statistical Validation of Simulation Models," Proc. Summer Computer Simulation Conference, Houston, TX, pp. 915-926.

[19] General Accounting Office (1979), "Guidelines for Model Evaluation," PAD-79-17, U.S.GAO, Washington, DC.

[20] Hermann, C.F. (1967), "Validation Problems in Games and Simulations with Special Reference to Models of International Politics," Behavioral Science, Vol. 12, pp. 216-231.

[21] Howrey, P. and H.H. Kelejian (1969), "Simulation Versus Analytical Solutions," in The Design of Computer Simulation Experiments, T.H. Naylor, Ed., Duke University Press, Durham, NC.

[22] Hunt, A.W. (1970), "Statistical Evaluation and Verification of Digital Simulation Models Through Spectral Analysis," Ph.D. Dissertation, The University of Texas at Austin.

[23] Kheir, N.A. and W.M. Holmes (1978), "On Validating Simulation Models of Missle Systems," Simulation, pp. 117-128.

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

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

[26] Miller, D.K. (1975), "Validation of Computer Simulations in the Social Sciences," Proc. Sixth Annual Conference on Modelling and Simulation, Pittsburg, PA, pp. 743-746.

[27] Miller, D.R. (1974a), "Sensitivity Analysis and Validation of Simulation Models," J. Theor. Biol., Vol. 48, pp. 345-360.

[28] Miller, D.R. (1974b), "Model Validation Through Sensitivity Analysis," Proc. Summer Computer Simulation Conference, Houston, TX, pp. 911-914.

[29] Mitroff, I.I. (1969), "Fundamental Issues in the Simulation of Human Behavior: A Case in Strategy of Behavioral Science," Management Science, Vol. 15, No. 12, pp. B635-B649.

[30] Naylor, T.H. and J.M. Finger (1967), "Verification of Computer Simulation Models," Management Science, Vol. 14, No. 2, pp. B-92-101.

[31] Rowland, J.R. and W.M. Holmes (1978), "Simulation Validation with Sparse Random Data," Computers and Electrical Engineering, Vol. 5, No. 1, pp. 37-49.

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

[33] Sargent, R.G. (1979), "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, pp. 3-1/3-13.

[34] Sargent, R.G. (1979), "Validation of Simulation Models," Proc. Winter Simulation Conference, San Diego, CA, pp. 497-503.

[35] Sargent, R.G. (1982), "Verification and Validation of Simulation Models," in Progress in Modeling and Simulation, F.E. Cellier, Ed., Academic, London.

[36] Schellenberger, R.E. (1974), "Criteria for Assessing Model Validity for Managerial Purposes," Decision Sciences, Vol. 5, No. 4, pp. 644-653.

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

[38] Schruben, L.W. (1980), "Establishing the Credibility of Simulations," Simulation, Vol. 34, No. 3, pp. 101-105.

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

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

[41] Teorey, T.J. (1975), "Validation Criteria for Computer System Simulations," Simuletter, Vol. 6, No. 4, pp. 9-20; also published in Proc. Symposium on the Simulation of Computer Systems III, August 1975, pp. 161-169.

[42] Theil, H. (1961), Economic Forecasts and Policy, North-Holland, Amsterdam.

[43] Turing, A.M. (1963), "Computing Machinery and Intelligence," in Computers and Thought, E.A. Feigenbaum and J. Feldman, Eds., McGraw-Hill, New York, pp. 11-15.

[44] Tytula, T.P. (1978), "A Method for Validating Missile System Simulation Models," U.S. Army Missile R&D Command, Technical Report E-78-11.

[45] Van Horn, R.L. (1971), "Validation of Simulation Results," Management Science, Vol. 17, No. 5, pp. 247-258.

[46] Watts, D. (1969), "Time Series Analysis," in The Design of Computer Simulation Experiments, T.H. Naylor, Ed., Duke University Press, Durham, NC, pp. 165-179.

[47] Wright, R.D. (1972), "Validating Dynamic Models: An Evaluation of Tests of Predictive Power," Proc. Summer Computer Simulation Conference, San Diego, CA, pp. 1286-1296.