Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

A TUTORIAL ON THE OPERATIONAL VALIDATION OF SIMULATION MODELS

Robert G. Sargent

Department of Electrical Engineering
and Computer Science
College of Engineering and Computer Science
Syracuse University
Syracuse, NY 13244, USA

David M. Goldsman
Tony Yaacoub

H. Milton Stewart School of Industrial
and Systems Engineering
Groseclose Building
Georgia Institute of Technology
Atlanta, GA 30332, USA

ABSTRACT

This tutorial discusses in depth the operational validation of simulation models after a brief overview of verification and validation of simulation models. The discussion of operational validation first covers the different approaches used for observable and non-observable systems. Next, various types of graphical displays of model output behavior are presented; this is followed by how these displays can be used in determining model validity by the model developers, subject matter experts, and others when no system data are available; and how these displays can be used as reference distributions for operational validation when system data are available. Lastly, the use of the "interval hypothesis test" is covered for operational validation when sufficient system data are available. Various examples are presented.

1 INTRODUCTION

This tutorial discusses the operational validation of simulation models. First, we give a brief overview of verification and validation of simulation models. Then, operational validity is discussed in depth.

To determine whether a model and its results are "correct" for a specific use or purpose, model verification and validation are performed. Model verification is defined as "ensuring that the computer program of the computerized model and its implementation are correct," and model validation is defined as the "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model." A model should be developed for a specific purpose and its validity should be determined with respect to that purpose. A developed model should usually be a parsimonious model, meaning that the model is as simple as possible yet meets its purpose. The accuracy of a model (sometimes referred to as model fidelity) should only be what is needed to satisfy the model's intended purpose or use and should usually be specified by the model's *acceptable range of accuracy*. A model should be valid over its domain of applicability, and this is usually determined by ensuring that the model is valid for numerous sets of experimental conditions that define the domain of a model's intended applicability. (A set of experimental conditions contains a set of values for the set of variables that define the domain of applicability.) If the purpose of a model is to answer a variety of questions, the validity of the model needs to be determined with respect to each question. (For a detailed discussion on simulation model verification and validation, see, e.g., Sargent 2013).

We present in Figure 1 a graphical paradigm developed by Sargent (1981, 1982, 1983, 2001b, 2013) called the "Simplified View of the Model Development Process" that shows how verification and validation are related to the model development process. The *problem entity* is the system (real or proposed), idea, situation, policy, or phenomenon to be modeled; the *conceptual model* is the mathematical/logical/graphical

978-1-5090-4486-3/16/$31.00 ©2016 IEEE

representation (mimic) of the problem entity developed for a particular study; and the *computerized model* is the conceptual model implemented on a computer. The conceptual model is developed through an *analysis and modeling phase*, the computerized model is developed through a *computer programming and implementation phase*, and inferences about the problem entity are obtained by conducting computer experiments on the computerized model in the *experimentation phase*.

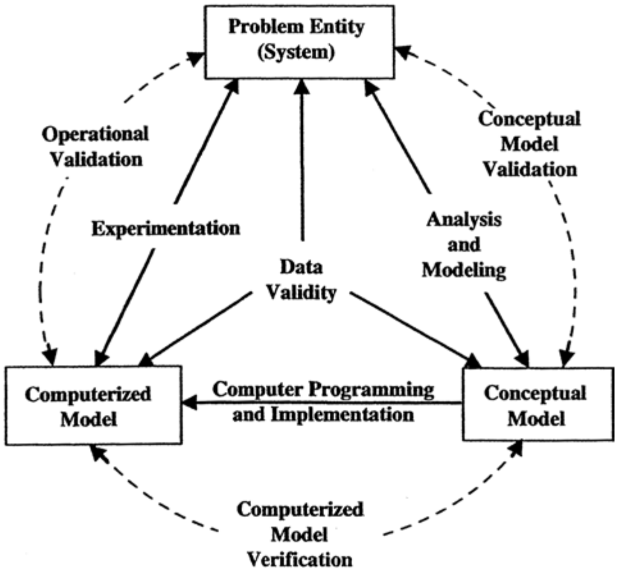


Figure 1: Simplified version of the model development process.

We now relate model verification and validation to this simplified version of the model development process. (See Figure 1.) *Conceptual model validation* is defined as ensuring the theories and assumptions underlying the conceptual model are correct and the model representation of the problem entity is "reasonable" for the intended purpose of the model. *Computerized model verification* is defined as assuring that the computer programming and implementation of the conceptual model are correct. *Operational validation* is defined as determining whether the model's output behavior has a satisfactory range of accuracy for the model's intended purpose over the domain of the model's intended applicability. *Data validity* is defined as ensuring that the data necessary for model building, model evaluation and testing, and conducting the model experiments to solve the problem are adequate and correct.

There are three basic decision-making approaches for deciding whether a simulation model is valid. Each of these three approaches uses a different decision-maker. All of the approaches require the model development team to conduct verification and validation as part of the model development process, which as shown in Figure 1 consists of three major model development steps and three verification and validation steps. (See Sargent 2013 and 2015a for further discussion on the model development process.)

One decision-making approach, and a frequently used one, is for the model development team itself to make the decision as to whether a simulation model is valid. The decision is based on the results of the various tests and evaluations conducted as part of the model development process.

A second decision-making approach is to have the user(s) of a simulation model decide the validity of the model. In this approach, the users of the simulation model are heavily involved with the model development team when the team is conducting verification and validation of the model and the users determine if the model is satisfactory in each phase of verification and validation. This approach is generally used with a model development team whose size is not large.

Another decision-making approach, usually called "independent verification and validation" (IV&V), uses a third party to decide whether the simulation model is valid. The third party (the IV&V team) is independent of both the simulation development team(s) and the model sponsor/user(s). The IV&V approach is generally used with the development of large-scale simulation models, whose development usually involves several teams. The IV&V team needs to have a *thorough* understanding of the intended purpose(s) of the simulation model in order to conduct IV&V.

The remainder of this paper discusses operational validation of simulation models in depth.

2 OPERATIONAL VALIDATION

Operational validation is determining whether the simulation model's output behavior has the accuracy required for the model's intended purpose over the domain of the model's intended applicability. This is where much of the validation testing and evaluation take place. Since the computerized (simulation) model as shown in Figure 1 is used in operational validation, any deficiencies found may be caused by what resulted from any of the earlier steps that are involved in developing the simulation model, including developing the system's theories or having invalid data.

The amount of accuracy required of a simulation model is usually specified by the range within which the difference between a model's output variable and the corresponding system output variable must be contained. This range is commonly known as the model's *acceptable range of accuracy*. If the variables of interest are random variables, then properties and functions of the random variables (such as means) are often what are of primary interest and are the quantities that are used in determining model validity. A model's acceptable range of accuracy should usually be specified prior to starting the development of the model or very early in the model development process. The accuracy of a model should be tested under numerous sets of experimental conditions that define the domain of a model's intended applicability. A model is considered valid for a set of experimental conditions if the model's accuracy is within its acceptable range of accuracy. A model may be valid for one set of experimental conditions and invalid for another.

Numerous validation techniques are applicable to operational validity (see, e.g., Sargent 2013). Which techniques and whether to use them objectively or subjectively must be decided by the model development team and the other interested parties. The major attribute affecting operational validity is whether the problem entity (or system) is observable, where observable means it is possible to collect data on the operational behavior of the problem entity. Table 1 gives a classification of the validation techniques used in operational validity based on the decision approach and system observability. "Comparison" means comparing the simulation model output behavior to either the system output behavior or another model output behavior using graphical displays and/or statistical tests and procedures. "Explore model behavior" means to examine the output behavior of the simulation model using appropriate validation techniques, including parameter variability-sensitivity analysis (Sargent 2013). Various sets of experimental conditions from the domain of the model's intended applicability should be used for both comparison and exploring model behavior.

To obtain a *high* degree of confidence in a simulation model and its results, comparisons of the model's and system's output behaviors for several different sets of experimental conditions are usually required. Thus if a system is not observable, which is often the case, it is usually not possible to obtain a high degree of confidence in the model. In this situation, the model output behavior(s) should be explored as thoroughly as possible and comparisons should be made to other valid models whenever possible.

2.1 Explore Model Behavior

The simulation model output behavior can be explored either qualitatively or quantitatively. In qualitative analysis the directions of the output behaviors are examined and also possibly whether the magnitudes are "reasonable." In quantitative analysis both the directions and the precise magnitudes of the output behaviors are examined. Experts (e.g., subject matter experts) on the system often know the directions and frequently know the "general values" of the magnitudes of the output behaviors. Many of the validation techniques can be used for model exploration; and specifically, parameter variability-sensitivity analysis should routinely be used. Graphs of the output data discussed in Subsection 2.2 below can be used to display the simulation model output behavior. Furthermore, a variety of statistical approaches can be used in performing model exploration including metamodeling and design of experiments. (See, e.g., Kleijnen 1999 and 2015 for discussions on the use of these statistical approaches.) Additionally, numerous sets of experimental frames should be used in performing model exploration. The results of exploring model behavior can be used when no system data are available to aid in making a subjective decision regarding operational validation of a simulation model by the model development team, subject matter experts, users, and others.

Table 1: Operational Validation Classification.

Decision Approach	Observable System	Non-observable System
Subjective Approach	• Comparison Using Graphical Displays • Explore Model Behavior	• Explore Model Behavior • Comparison to Other Models
Objective Approach	• Comparison Using Statistical Tests and Procedures	• Comparison to Other Models Using Statistical Tests

2.2 Comparisons of Output Behaviors

There are three basic approaches used in comparing the simulation model output behavior to either the system output behavior or another (validated) model output behavior: (1) the use of graphs to make a subjective decision, (2) the use of confidence intervals to make an objective decision, and (3) the use of hypothesis tests to make an objective decision. It is preferable to use one of the statistical comparison approaches (2) or (3) because they allow for objective decisions. However, it is often not possible in practice to use either of these approaches because (a) the statistical assumptions required cannot be satisfied or only with great difficulty (required assumptions are usually data independence and normality), (b) there is an insufficient quantity of system data available, which causes the statistical results to be "meaningless" (e.g., the length of a confidence interval developed in the comparison of the system and simulation model means is too long for any practical usefulness), and/or (c) the behavior of the problem entity is changing, e.g., is highly nonstationary. As a result, the use of graphs is the most-commonly used approach for operational validation, but extreme care must be used for this approach. Each of these three approaches is discussed below using system output data. (Note: these same approaches can also be used with output data from a validated model instead of system output data when appropriate.)

2.2.1 Comparisons Using Graphical Displays

The output behavior data of the simulation model and the system can be graphed for various sets of experimental conditions to aid in determining if the model's output behavior has sufficient accuracy for the model's intended purpose. Three types of graphs are used: histograms, box (and whisker) plots, and behavior graphs using scatter plots (see, e.g., Hines, et al. 2003 for a discussion of graphs in general). These three types of displays (or graphs) allow data to be statistically dependent (i.e., correlated) and non-normal, which often occurs in behavior data of systems and simulation models. The use of graphical displays of (a) data for operational validation is discussed in Sargent (1996), and (b) simulation data for statistical references is developed in depth in Sargent (2001a).

Let us first illustrate histograms and box plots through an example. Consider the system in Figure 2 and a model of it in Figure 3 where transit times from nodes 1 to 9 are the outputs of interest. The label on each arc (edge) of the graphs is the associated transient time and probability of being selected for transit. The mean transit time through the system and model are respectively 7.760 and 6.890. Figures 4 and 5 show the system and model histograms each containing 5000 observations obtained by simulation, and Figure 6 depicts the box plots of each using the same sets of 5000 observations. A large number of observations were used here to illustrate that one can detect differences when differences occur through the use of histograms and box plots. We can observe from the histograms that there is slightly more variability in the system than in the model and furthermore the mode of the system is about seven and is about six for the model. From the box plots one can also see that the system is more variable than the model and that both have about the same median values.

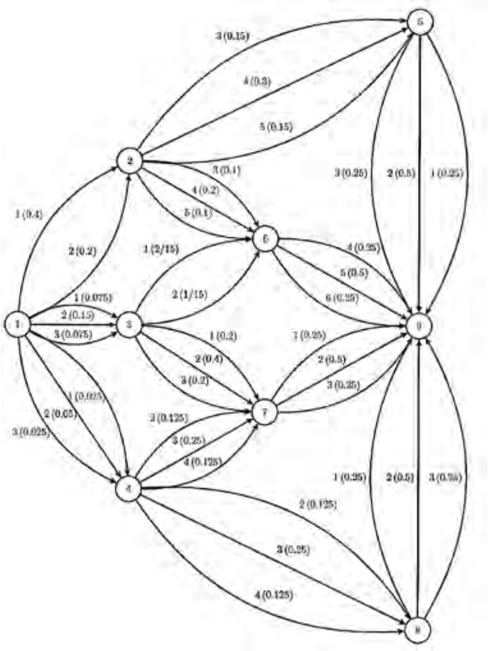


Figure 2: System.

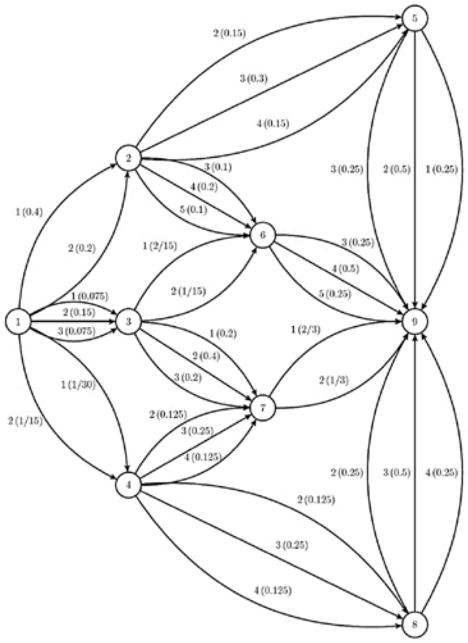


Figure 3: Model.

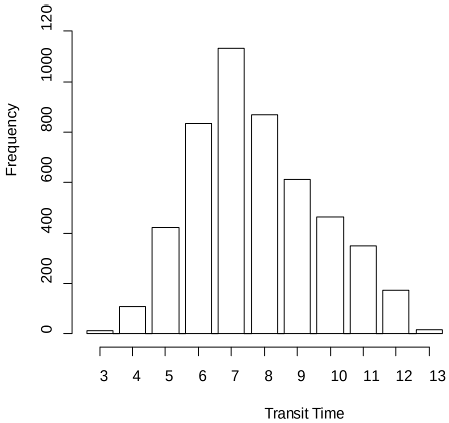


Figure 4: System histogram.

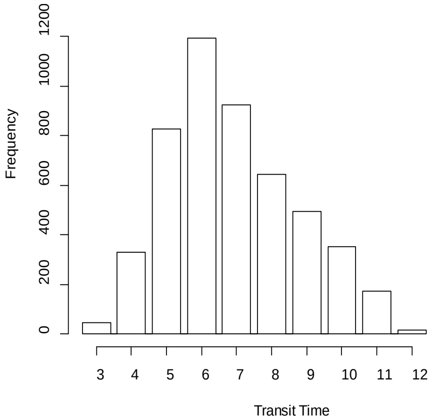


Figure 5: Model histogram.

Let us obtain 500 sample means of size 20 from the model (which has an overall sample mean of 6.901) and put these 500 observations into a histogram as shown in Figure 7 as a statistical reference for sample means of size 20. Similarly, let us obtain a sample of size 20 from the system, which has a sample mean of 7.80; and put this system sample mean on the histogram in Figure 7 with an "X". One can see from the Figure 7 that the system sample mean is in the tail of the histogram and thus the system and model means are most likely not equal but are reasonably close to each other.

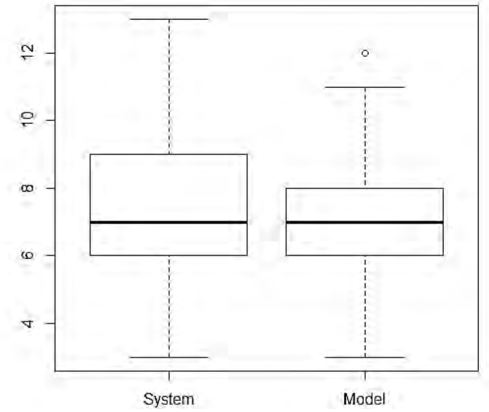


Figure 6: Box plots of the system and model.

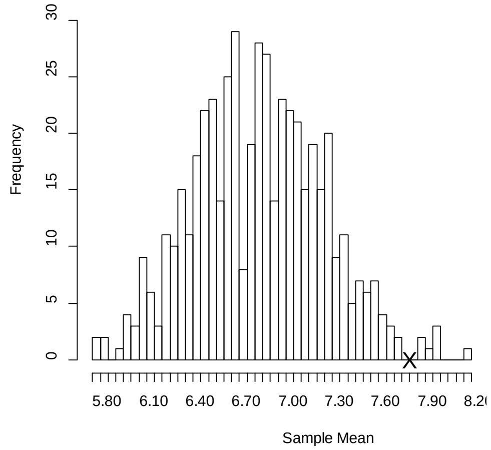
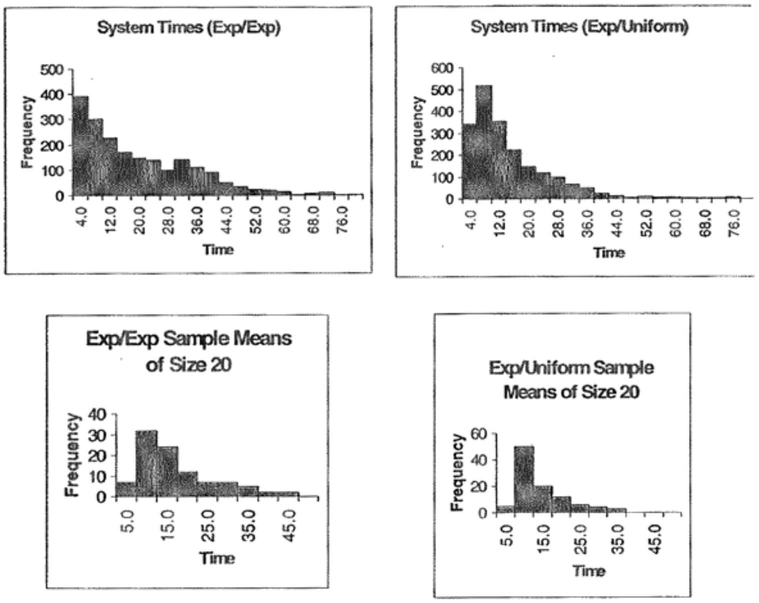


Figure 7: Model mean histogram of size 500.

Let us look at another example: a single-server queueing system with an infinite allowable queue with exponential inter-arrival times and two different service time distributions. One service time distribution will be an exponential and the other one will be a uniform with both having the same mean. One can readily see from the histograms in Figure 8 and the box plots in Figure 9 that the system times are different between these two systems with different service time distributions.


System Times (Exp/Exp)

System Times (Exp/Uniform)

Exp/Exp Sample Means of Size 20

Exp/Uniform Sample Means of Size 20

Figure 8: Histograms of single-server queueing system.

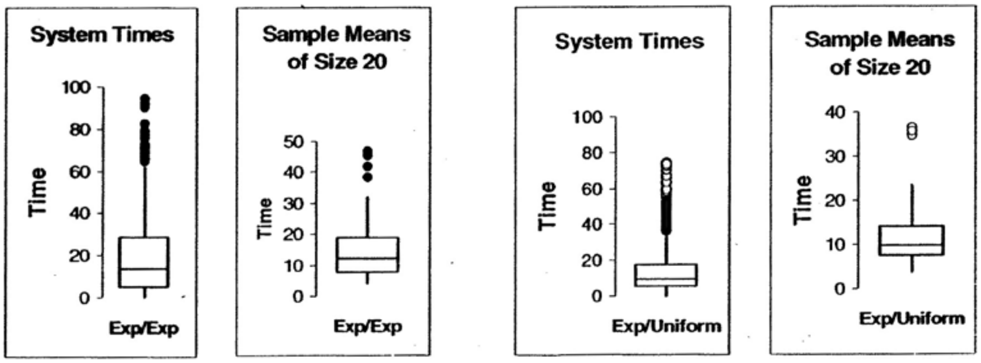


Figure 9: Box plots of single-server queueing system.

Next, let us look at a hospital model by Lowery (1996) that was validated by using histograms and box plots. Figures 10 and 11 give two of the graphical displays that were used. We note in Figure 10 that the system observation for 24 weeks lay within the histogram of the model observations of 24 weeks. In Figure 11 we note that the model has a slightly larger variation than the system and slightly lower median.

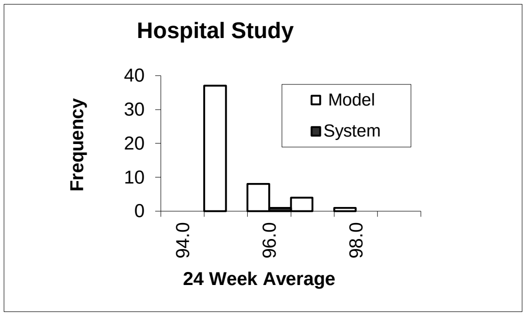


Figure 10: Histogram of hospital data.

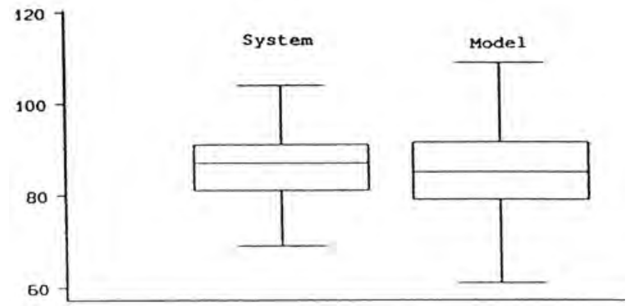


Figure 11: Box plots of hospital data.

Let us next look at the use of behavior graphs for validation of simulation models. Recall that behavior graphs use scatter plots. A variety of graphs should be used with different types of (1) measures such as the mean, variance, maximum, distribution, and times series of a variable, and (2) relationships such as those between (a) two measures of a single variable and (b) measures of two variables. It is important that the measures and relationships selected for validating a simulation model be determined with respect to the model's intended purpose. We will look at the work of Anderson and Sargent (1974) that used behavior graphs in validating a model of a computer system. Figure 12 gives a queueing network of the computer system. Figures 13–15 give three of the behavior graphs that were used. Figure 14 illustrates the use of two measures of the same variable. The other two figures use different variables for the relationships. Studying these graphical displays, one notes that there is a difference between the simulation model and computer system whose cause was determined to be that the actual job scheduling rule used in the computer system was different than what was described for the computer system and was used in the simulation model.

2.2.2 Confidence Intervals

Confidence intervals (c.i.'s) and simultaneous confidence intervals (s.c.i.'s) can be obtained for the differences between means, variances, and distributions of different output variables of a simulation

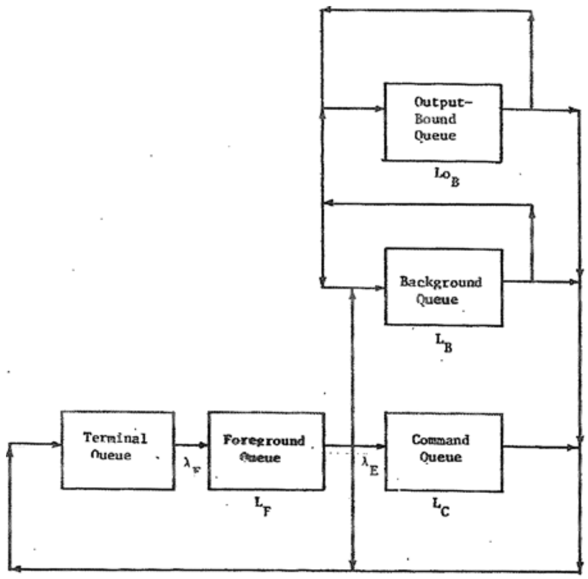


Figure 12: Queue network of computer system.

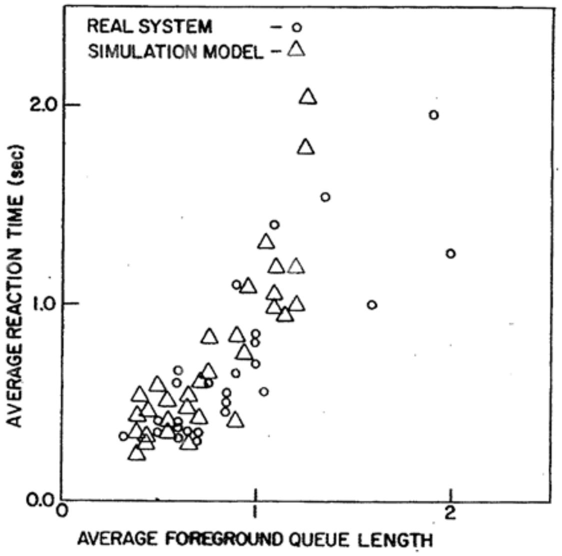


Figure 13: Behavior of foreground queue.

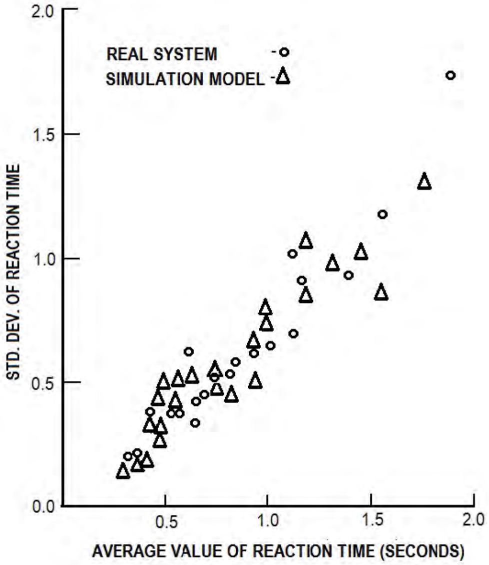


Figure 14: Computer reaction time.

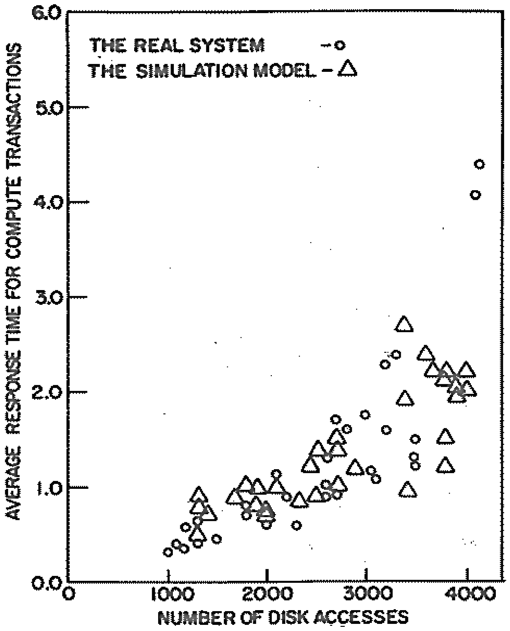


Figure 15: Response time versus disk access.

model and a system for each set of experimental conditions. These c.i.'s and s.c.i.'s can be used as the model range of accuracy for model validation, where the model range of accuracy is the confidence interval or region (for the s.c.i.'s) around the estimated difference between some function (e.g., the mean) of the model and system output variable being evaluated. Balci and Sargent (1984) give details on the use of c.i.'s and s.c.i.'s for operational validity, including a general methodology.

To construct the model range of accuracy, a procedure containing a statistical technique and a method of data collection must be developed for each set of experimental conditions and for each variable of interest for both the simulation model and the system. The statistical techniques used can be divided into two groups: (1) univariate statistical techniques and (2) multivariate statistical techniques. The univariate techniques can be used to develop c.i.'s and with the use of Bonferroni's inequality (Law 2014) s.c.i.'s. The multivariate techniques can be used to develop a s.c.i. Both parametric and nonparametric statistical techniques can be used. The method of data collection must satisfy the underlying assumptions of the statistical technique that is being used. One approach to developing a model range of accuracy is to use the standard statistical techniques and data collection methods common in simulation output analysis (Banks et al. 2010, Law 2014) for both the simulation model and the system, e.g., using the methods of replication or (nonoverlapping) batch means. We note that these c.i.'s and s.c.i.'s will be around a specific value for the difference between the simulation model and the system.

2.2.3 Interval Hypothesis Test

In conducting operational validation, one is interested in determining if the accuracy of a model is within its acceptable range of accuracy specified by L and U; where L and U are the lower and upper values of the range of the model's acceptable range of accuracy, respectively. Thus one desires a hypothesis test that determines whether a value is within a range. Classical hypothesis tests determinations are typically about a specific point. Sargent (2015b) recently developed an interval statistical procedure for model validation that includes an interval hypothesis test that is based on the use of the operating characteristic curve. This interval statistical procedure is discussed by Sargent, Goldsman, and Yaacoub (2015) using examples involving the validation of a simple queueing simulation model to illustrate various usages of this interval validation procedure.

Table 2 gives different ways this interval hypothesis test can be used for validation of models. In using Table 2, one determines the row to be used based on the type of MODEL being validated and the type of REFERENCE the type of MODEL is being compared (tested) against to find the STATISTICAL PROCEDURE to be used. Appendices A and B contain the statistical procedures for using this new interval hypothesis test for validation of models for the one-sample and two-sample tests, respectively.

Table 2: Interval Hypothesis Tests for the Interval Statistical Procedure.

MODEL (Entity being validated)	REFERENCE (Model being compared against)	STATISTICAL PROCEDURE
Simulation Model	Analytical Model	-Perform one-sample test -Calculate value of Analytical Model -Collect data sample from MODEL or REFERENCE, as appropriate
Analytical Model	Simulation Model or System	
Simulation Model	Simulation Model or System	-Perform two-sample test -Collect one data sample from MODEL -Collect second data sample from REFERENCE

Two types of errors are possible in ascertaining model validity via hypothesis testing. Type I error is that of rejecting the validity of a valid model, and Type II error is that of accepting the validity of an invalid model. The probability of a Type I error, α, is called the model builder's risk, and the probability of Type II error, β, is called the model user's risk (Balci and Sargent 1981). In model validation, the model user's risk is especially important and must be kept small to reduce the probability that an invalid model is accepted as being valid. Thus both Type I and Type II errors must be carefully considered when using hypothesis testing for model validation.

This new interval hypothesis test allows consideration of both Type I and Type II errors through the use of the operating characteristic (OC) curve. The OC curve is defined as the probability of accepting the null hypothesis when event E prevails, denoted $P_{\mathrm{A}}(E)$. The probability of Type I error, $\alpha(E)$, is $1 - P_{\mathrm{A}}(E)$

when E has a value where the null hypothesis is true; and the probability of Type II error, $\beta(E)$, is $P_A(E)$ when E has a value where the alternative hypothesis is true. Note that the probabilities of Type I error, $\alpha(E)$, and Type II error, $\beta(E)$, are both functions of the event E. Moreover, it is common practice to specify α, which is called the significance level of the test, as the maximum allowable value of the probability of Type I error. (For a detailed discussion on hypothesis testing, Type I and Type II errors, and OC curves, see, e.g., Hines et al. 2003 or Johnson et al. 2010.) Furthermore, the model builder's and the model user's risk curves can be obtained from the OC curve.

We are going to limit our discussion to the testing of the means. (The interval hypothesis test can be used for testing other statistical properties provided that the appropriate OC curve can be calculated.) Let μ_M and μ_R be the means of the simulation model and reference (in our case, the system), respectively, and let $D = \mu_M - \mu_R$. Test statistics for testing means commonly use the t-distribution when the variances are unknown. We assume that the variances of the model and reference outputs are unknown and equal, and thus will use the t-distribution for our hypothesis testing. This procedure requires the data from both the simulation model and the reference to be approximately Normal Independent and Identically Distributed (NIID). This can be accomplished in model validation for both the simulation model data and the reference (system) data by using the standard methods that are carried out in simulation output analysis to obtain approximately NIID data; namely, for terminating simulations, one could use the method of replications and for nonterminating (steady-state) simulations, either the method of replications or the method of batch means (see, e.g., Law 2014). Let n_M indicate the number of model NIID data values (observations) and n_R the number of reference (e.g., system) NIID data values. The t-distribution with the appropriate test statistic will be used for testing the means of our NIID data. As mentioned above, both Type I and Type II errors are important in model validation and they are considered through the use of OC curves. The significance level of the test, α, for our situation is the maximum of $\alpha_L = \alpha(D = L)$ and $\alpha_U = \alpha(D = U)$. Also, letting $\beta_L = \beta(D = L)$ and $\beta_U = \beta(D = U)$, we note that $\alpha_L + \beta_L = 1$ and $\alpha_U + \beta_U = 1$.

Computer software using R code (R Development Core Team 2008) has been developed to use the interval statistical procedures given in Appendices A and B for testing of means. This software is briefly discussed in Sargent, Goldsman, and Yaacoub (2015), and various snippets of R code are at the website http://www2.isye.gatech.edu/~sman/validation/.

Let us use this interval hypothesis test to determine if the simulation model given in Figure 3 is a valid model of the system given in Figure 2, which is the reference for this model, with respect to the mean transit time. Let μ_M and μ_R be the mean transit time of the simulation model and reference (system), respectively. We want to determine if D is within (L, U). The interval hypothesis test will be used to test the hypothesis: H_0: $L \le D \le U$ vs. H_1: $D < L$ or $D > U$. Using Table 2, we find that we should apply the two-sample test and thus the Interval Statistical Procedure given in Appendix B. Since the model and system are terminating, we will use the method of replications to obtain our NIID observations. (The sample mean should be approximately normal due to the fact that each observation is a sum of random values (the arcs' transit times); and then we sum the observations to obtain the sample mean.) Let the model acceptable range of accuracy be ($L = -0.75$, $U = 0.75$), and let the set of experimental conditions be the values given in Figures 2 and 3. We have now specified the model validation formulation of Step 0 of the Statistical Procedure in Appendix B.

Proceeding to Step 1 of this Procedure, we have for the statistical hypothesis H_0: $-0.75 \le D \le 0.75$ vs. H_1: $D < -0.75$ or $D > 0.75$, and we will use the two-sample t-test for unknown and equal variances.

We next carry out Step 2 of the Procedure. We select the Initial Sample Sizes to be $n_R = 10$ and $n_M = 15$. We collect the NIID observations of these sample sizes for the reference (system) and model. Analyzing the collected data we obtain $\overline{M} = 7.067$ for the sample mean of the model, $\bar{R} = 7.300$ for the sample mean of the reference, $S_M^2 = 2.638$ for the sample variance of the model, $S_R^2 = 2.233$ for the sample variance of the reference, and $S_P^2 = 2.480$ for the pooled variance estimate.

Moving to Step 3 of the Procedure, we evaluate the risk curves using different β values in order to select β values for L and for U to use in the hypothesis test. In this example, the same β values will be used at L

and U. Our software for this interval statistical procedure is used to obtain the risk curves shown in Figures 16 and 17 for β values of 0.4 and 0.5. The pooled variance estimate obtained in Step 2 is used in calculating these risk curves. The horizontal (x) axis in Figure 16 is ($D – L$) and in Figure 17 is ($D – U$). (Note, e.g., that the value of D is –0.75 at the x-axis location of 0 in Figure 16.) The model builder's risk curves are the curves on the right side of zero in Figure 16 (Lower Case Curves) and to the left of zero in Figure 17 (Upper Case Curves). The model user's risk curves are the curves on the left side of zero in Figure 16 and on the right side of zero in Figure 17. The risk curves in Figures 16 and 17 are identical except for being "reversed," the reason being that the same values were used for β_L and β_U. (Our examples, unless stated otherwise, will henceforth use identical values for β_L and β_U, and we will present only the Lower Case Curves because the Upper Case Curves are just reversed images of the former.) The risk curves are examined to evaluate the trade-offs between the model builder's risk and the model user's risk for different β values at L and U. Based on our evaluation of the risk curves, we select $\beta_L = \beta_U = 0.40$ as our values to use in the hypothesis test for determining validity. (Note that this gives $\alpha_L = \alpha_U = 0.60$ since $\alpha_L + \beta_L = 1$ and $\alpha_U + \beta_U = 1$.)

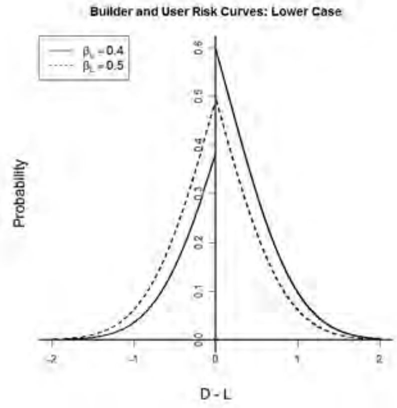


Figure 16: β risk curves at L.

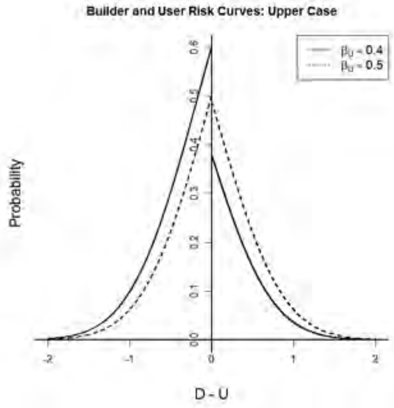


Figure 17: β risk curves at U.

In Step 4 of the Procedure we explore increased sample sizes to determine their effects on the risk curves. The pooled variance estimate must be calculated for each new set of increased sample sizes using the variance estimates from Step 2 and the increased sample sizes. One can observe from the risk curves in Figures 16 and 17 that the sample sizes need to be increased significantly if the hypothesis test is going to have "meaningful results" regarding the model acceptable range of accuracy given by (–0.75, 0.75). Our software was used to produce Figure 18, which shows the Initial Sample Size risk curve along with the risk curves for sample sizes of $n_R = 20$ and $n_M = 25$ and 40, where $S_P^2 = 2.459$ and 2.506, respectively. (These sets of risk curves are for $\beta_L = \beta_U = 0.40$ as these β values were selected in Step 3.) After evaluating different feasible sample sizes, it was decided to use $n_R = 20$ and $n_M = 25$ as the sample sizes for the hypothesis test.

In Step 5 of the Procedure, the first action required is to collect the additional samples that are called for from Step 4. We obtain 10 more reference (system) NIID observations and 10 additional model NIID observations (replications). Next we analyze the total samples of 20 reference observations and 25 model observations. We obtain $\bar{M} = 7.120$, $\bar{R} = 7.800$, $S_M^2 = 2.777$, $S_R^2 = 2.695$, and $S_P^2 = 2.740$. We use the software program with the newly calculated pooled variance estimate of 2.740 to produce the final risk curve shown in Figure 19, along with the acceptance region for the test statistic T, (–1.262, 1.262), and the T critical value of –1.369. Since the value of the T does not fall within the acceptance region, the null hypothesis is rejected, meaning that the simulation model does not have a mean time output that is acceptable. (The acceptance region can be calculated for D, which is (–0.627, 0.627); and the difference between the two sample means is 7.120 – 7.800 = –0.680, which, of course, falls outside the acceptance region. This is consistent with the fact that the T value falls outside of its acceptance region.) Thus, the model has been determined to be invalid with the risks as depicted by the final risk curve plotted in Figure 19.

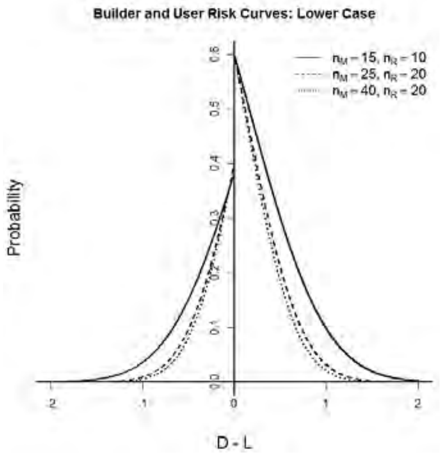


Figure 18: Sample size risk curves at *L*.

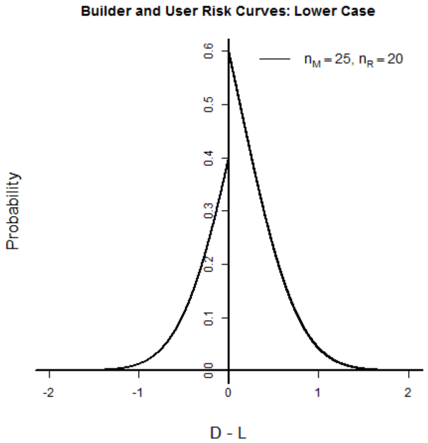


Figure 19: Final risk curve at *L*.

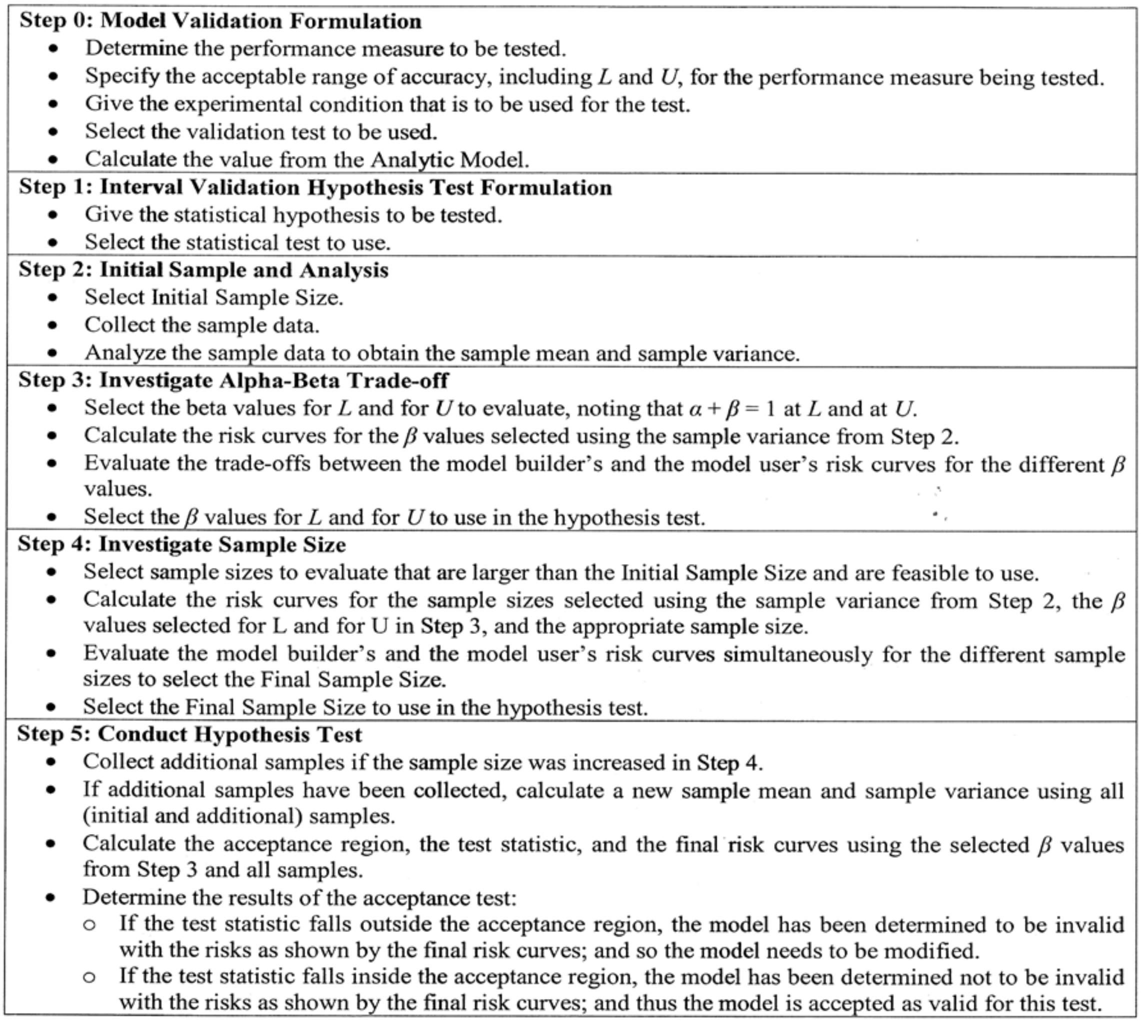
Let us now modify this example to illustrate the use of the interval statistical procedure when L is not equal to $–U$. We will keep everything the same except let $L = –1.00$. If we use the same seeds for generating the observations and make the same decisions in all of the steps of the Procedure, then everything remains the same, except in Step 5 the acceptance range of T changes to (–1.766, 1.262) and now $T = –1.369$. Since the value of $T = –1.369$ falls within the acceptance range for T, the null hypothesis now fails to be rejected and therefore the model has been determined to be valid with the risks as depicted by the risk curves plotted in Figure 19. (The acceptance region can be calculated for D, which is (–0.877, 0.627), and the difference of the two sample means is –0.68, which, of course, falls within the acceptance region.) Note that the accuracy required of the simulation model was not as stringent as previously required**,** which allowed the model to now be acceptable. We also note that $D = \mu_M – \mu_R = 6.890 – 7.760 = –0.870$, and thus the statistical results for both $L = –0.75$ and $L = –1.00$ agree with the theoretical results.

The system (reference) sample of size 20 having a sample mean of 7.800 used in the interval statistical tests discussed above was the same system sample whose sample mean was used in Figure 7. Our subjective analysis using the graphical approach and Figure 7 led us to conclude that the model mean transit time was probably close to but not equal to the system mean transit time. Using the interval statistical procedure we were able to quantify the results and obtain an objective decision of whether the model satisfied the model's acceptable range of accuracy. This example clearly shows the advantages of using the interval statistical procedure for model validation whenever it is possible.

3 SUMMARY

After a brief overview of verification and validation of simulation models, operational validity was discussed in depth. Validation approaches were discussed that use subjective and objective decision making for both observable and non-observable systems. Various types of graphical displays for exploring and comparing output behaviors were covered in detail and illustrated through examples. Statistical methods were discussed and a detailed example using a new interval hypothesis test was presented that illustrated why this approach should be used when possible. Detailed procedures for using this new interval statistical approach for model validation for both one-sample and two-sample interval hypothesis tests are given in two appendices.

A ONE-SAMPLE INTERVAL STATISTICAL PROCEDURE



Step 0: Model Validation Formulation

- Determine the performance measure to be tested.
- Specify the acceptable range of accuracy, including *L* and *U*, for the performance measure being tested.
- Give the experimental condition that is to be used for the test.
- Select the validation test to be used.
- Calculate the value from the Analytic Model.

Step 1: Interval Validation Hypothesis Test Formulation

- Give the statistical hypothesis to be tested.
- Select the statistical test to use.

Step 2: Initial Sample and Analysis

- Select Initial Sample Size.
- Collect the sample data.
- Analyze the sample data to obtain the sample mean and sample variance.

Step 3: Investigate Alpha-Beta Trade-off

- Select the beta values for *L* and for *U* to evaluate, noting that $\alpha + \beta = 1$ at *L* and at *U*.
- Calculate the risk curves for the β values selected using the sample variance from Step 2.
- Evaluate the trade-offs between the model builder's and the model user's risk curves for the different β values.
- Select the β values for *L* and for *U* to use in the hypothesis test.

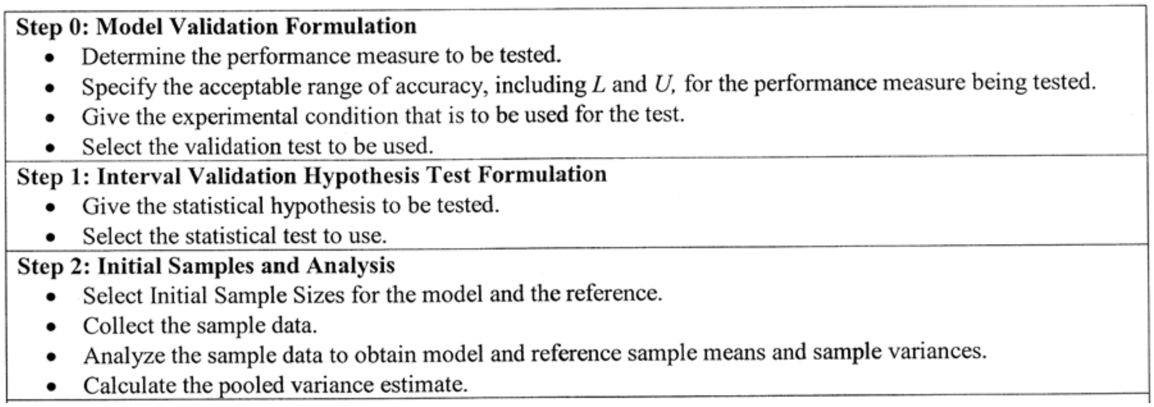
Step 4: Investigate Sample Size

- Select sample sizes to evaluate that are larger than the Initial Sample Size and are feasible to use.
- Calculate the risk curves for the sample sizes selected using the sample variance from Step 2, the β values selected for L and for U in Step 3, and the appropriate sample size.
- Evaluate the model builder's and the model user's risk curves simultaneously for the different sample sizes to select the Final Sample Size.
- Select the Final Sample Size to use in the hypothesis test.

Step 5: Conduct Hypothesis Test

- Collect additional samples if the sample size was increased in Step 4.
- If additional samples have been collected, calculate a new sample mean and sample variance using all (initial and additional) samples.
- Calculate the acceptance region, the test statistic, and the final risk curves using the selected β values from Step 3 and all samples.
- Determine the results of the acceptance test:
  - If the test statistic falls outside the acceptance region, the model has been determined to be invalid with the risks as shown by the final risk curves; and so the model needs to be modified.
  - If the test statistic falls inside the acceptance region, the model has been determined not to be invalid with the risks as shown by the final risk curves; and thus the model is accepted as valid for this test.

B TWO-SAMPLE INTERVAL STATISTICAL PROCEDURE



Step 0: Model Validation Formulation

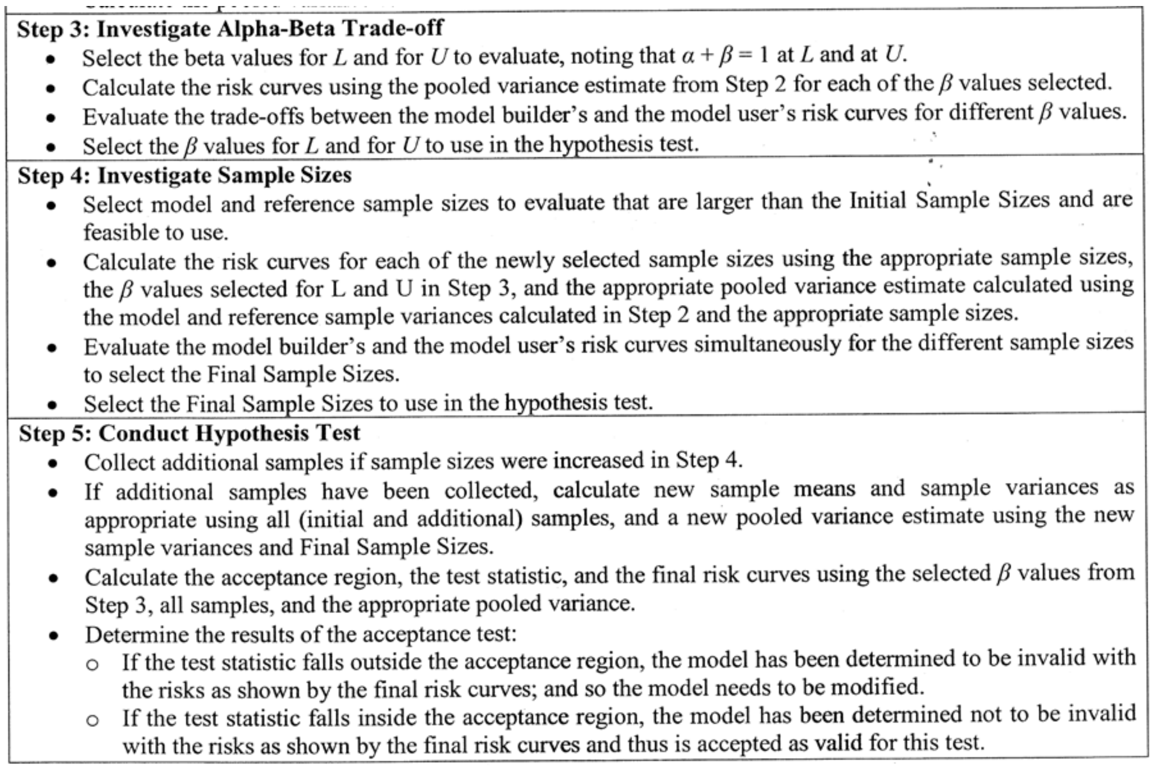
- Determine the performance measure to be tested.
- Specify the acceptable range of accuracy, including *L* and *U*, for the performance measure being tested.
- Give the experimental condition that is to be used for the test.
- Select the validation test to be used.

Step 1: Interval Validation Hypothesis Test Formulation

- Give the statistical hypothesis to be tested.
- Select the statistical test to use.

Step 2: Initial Samples and Analysis

- Select Initial Sample Sizes for the model and the reference.
- Collect the sample data.
- Analyze the sample data to obtain model and reference sample means and sample variances.
- Calculate the pooled variance estimate.



Step 3: Investigate Alpha-Beta Trade-off

- Select the beta values for L and for U to evaluate, noting that $\alpha + \beta = 1$ at L and at U.
- Calculate the risk curves using the pooled variance estimate from Step 2 for each of the β values selected.
- Evaluate the trade-offs between the model builder's and the model user's risk curves for different β values.
- Select the β values for L and for U to use in the hypothesis test.

Step 4: Investigate Sample Sizes

- Select model and reference sample sizes to evaluate that are larger than the Initial Sample Sizes and are feasible to use.
- Calculate the risk curves for each of the newly selected sample sizes using the appropriate sample sizes, the β values selected for L and U in Step 3, and the appropriate pooled variance estimate calculated using the model and reference sample variances calculated in Step 2 and the appropriate sample sizes.
- Evaluate the model builder's and the model user's risk curves simultaneously for the different sample sizes to select the Final Sample Sizes.
- Select the Final Sample Sizes to use in the hypothesis test.

Step 5: Conduct Hypothesis Test

- Collect additional samples if sample sizes were increased in Step 4.
- If additional samples have been collected, calculate new sample means and sample variances as appropriate using all (initial and additional) samples, and a new pooled variance estimate using the new sample variances and Final Sample Sizes.
- Calculate the acceptance region, the test statistic, and the final risk curves using the selected β values from Step 3, all samples, and the appropriate pooled variance.
- Determine the results of the acceptance test:
  - If the test statistic falls outside the acceptance region, the model has been determined to be invalid with the risks as shown by the final risk curves; and so the model needs to be modified.
  - If the test statistic falls inside the acceptance region, the model has been determined not to be invalid with the risks as shown by the final risk curves and thus is accepted as valid for this test.

REFERENCES

Anderson, H. A. and R. G. Sargent. 1974. "An Investigation into Scheduling for an Interactive Computer System." *IBM Journal of Research and Development* 18 (2): 125–137.

Balci, O. and R. G. Sargent. 1981. "A Methodology for Cost-risk Analysis in the Statistical Validation of Simulation Models." *Comm. of the ACM* 24 (6): 190–197.

Balci, O. and R. G. Sargent. 1984. "Validation of Simulation Models via Simultaneous Confidence Intervals." *American Journal of Mathematical and Management Science* 4 (3): 375–406.

Banks, J., J. S. Carson II, B. L. Nelson, and D. Nicol. 2010. *Discrete-event System Simulation.* 5th ed. Englewood Cliffs, NJ: Prentice-Hall.

Hines, W. W., D. C. Montgomery, D. M. Goldsman, and C. M. Borror. 2003. *Probability and Statistics in Engineering*. 4th ed. Hoboken, New Jersey: John Wiley.

Johnson, R. A., I. Miller, and J. E. Freund. 2010. *Miller & Freund's Probability and Statistics for Engineers*. 8th ed. New Jersey: Prentice Hall.

Kleijnen, J. P. C. 1999. "Validation of Models: Statistical Techniques and Data Availability." In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. Black Nembhard, D. T. Sturrock, and G. W. Evans, 647–654. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kelijnen, J. P. C. 2015. *Design and Analysis of Simulation Experiments*. 2nd ed. Heidelberg: Springer.

Law, A. M. 2014. *Simulation Modeling and Analysis*. 5th ed. New York: McGraw-Hill.

Lowery, J. 1996. "Design of Hospital Admissions Scheduling System Using Simulation." In *Proceedings of the 1996 Winter Simulation Conference*, edited by J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 1199–1204. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Sargent, R. G. 1981. "An Assessment Procedure and a Set of Criteria for Use in the Evaluation of Computerized Models and Computer-based Modeling Tools." Final Technical Report RADC-TR-80-409, U.S. Air Force.

Sargent, R. G. 1982. "Verification and Validation of Simulation Models." Chapter IX in *Progress in Modelling and Simulation*, edited by F. E. Cellier, 159–169. London: Academic Press.

Sargent, R. G. 1983. "Validating Simulation Models." In *Proceedings of the 1983 Winter Simulation Conference*, edited by S. Roberts, J. Banks, and B. Schmeiser, 333–337. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sargent, R. G. 1984. "Simulation Model Validation." Chapter 19 in *Simulation and Model-Based Methodologies: An Integrative View*, edited by T. I. Oren, B. P. Zeigler, and M. S. Elzas, 537–555. Heidelberg, Germany: Springer-Verlag.

Sargent, R. G. 1996. "Some Subjective Validation Methods Using Graphical Displays of Data." In *Proceedings of the 1996 Winter Simulation Conference*, edited by J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 345–351. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sargent, R. G. 2001a. "Graphical Displays of Simulation Model Data as Statistical References." In *Simulation 2001 (Proc. of the 4th St. Petersburg Workshop on Simulation)*, edited by S. M. Ermakor, Yu. N. Kashta-nov, and V. B. Melas, 109–118. Publisher: Chemistry Research Institute of St. Petersburg University.

Sargent, R. G. 2001b. "Some Approaches and Paradigms for Verifying and Validating Simulation Models." In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer,106–114. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sargent, R. G. 2013. "Verification and Validation of Simulation Models." *Journal of Simulation* 7:12–24.

Sargent, R. G. 2015a. "An Introductory Tutorial on Verifying and Validating Simulation Models," In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 1729–1740. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sargent, R. G. 2015b. "An Interval Statistical Procedure for Use in Validation of Simulation Models." *Journal of Simulation 9*: 232–237.

Sargent, R. G., D. M. Goldsman, and T. Yaacoub. 2015. "Use of the Interval Statistical Procedure for Simulation Model Validation." In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 60–72. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

ROBERT G. SARGENT is a Professor Emeritus of Syracuse University. He received his education at The University of Michigan. Dr. Sargent has received several awards and honors for his professional contributions. His email is rsargent@syr.edu.

DAVID GOLDSMAN is a Professor in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests include simulation output analysis, ranking and selection, and healthcare simulation. He is a Fellow of the Institute of Industrial Engineers. His email address is sman@gatech.edu and his webpage is www.isye.gatech.edu/~sman.

TONY YAACOUB is a Ph.D. student in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests include statistical quality control, machine learning, and discrete-event simulation. His email is tyaacoub@gatech.edu.