Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

USE OF THE INTERVAL STATISTICAL PROCEDURE FOR SIMULATION MODEL VALIDATION

Robert G. Sargent

Department of Electrical Engineering
and Computer Science
L. C. Smith 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

In this tutorial we discuss the use of a recently published statistical procedure for the validation of models that have their required model accuracy specified as a range, often called the acceptable range of accuracy. This new statistical procedure uses a hypothesis test of an interval, considers both Type I and Type II errors through the use of the operating characteristic curve, and provides the model builder's risk curve and the model user's risk curve. A detailed procedure for validating simulation models using this interval hypothesis test is given, computer software developed for this procedure is briefly described, and examples of simulation model validation using the procedure and software are presented.

1 INTRODUCTION

This tutorial describes how to use a new statistical procedure recently published by Sargent (2014) for use in validating a model against an observable system (or another model). An observable system implies that data can be collected on the system to compare with the model's behavior. This new procedure uses a statistical interval-based hypothesis test to determine if the difference between a model's output and the corresponding system output is within some range (interval) for a set of experimental conditions. (A set of experimental conditions has a set of values for the set of variables that define the domain of applicability of a model.) This procedure would be used when conducting operational validity of a model, which consists of determining whether a model's output behavior has the required amount of accuracy for the model's intended purpose over the domain of the model's intended applicability. (For a discussion of operational validity, see, e.g., Sargent 2013.)

The amount of accuracy required of a 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. Current statistical procedures that use hypothesis tests in operational validity only test for a single point (see, e.g., Banks et al. 2009) or consider ranges indirectly (see Balci and Sargent 1981, 1982a, 1982b, and 1983). The new statistical procedure uses an interval in its hypothesis test to determine model validity.

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 traditional statistical applications, both quantities are often specified in advance of any sampling; but, unfortunately, this rarely occurs in model validation, so in simulation modeling applications, this desirable requirement may not be in effect. 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.

The new statistical procedure for hypothesis testing considers 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 $P_A(E)$. The probability of Type I error, $\alpha(E)$, is $1 - P_A(E)$ 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.

The reminder of this paper is organized into two major sections plus a summary. Section 2 discusses the statistical procedure, including a list of steps to be used in applying the interval hypothesis test in model validation. Section 3 contains detailed examples regarding the use of the interval hypothesis test in determining simulation model validity. Section 4 is the summary.

2 STATISTICAL PROCEDURE

2.1 Operational Validity of a Model Output

We are interested in determining if the mean of a model's output (performance measure) is satisfactory for its intended use as part of conducting operational validity of that model. We compare the difference between the mean μ_s of some output variable (performance measure) from the true system and the mean μ_m of the corresponding output variable from a model of the system; and we wish to determine whether the difference between μ_s and μ_m is within the model's acceptable range of accuracy for that variable's mean under the set of experimental conditions specified.

Specifically, we want to determine if $D = \mu_m - \mu_s$ is contained in the acceptable range of accuracy. The acceptable range of accuracy for D is given by L for the lower limit and U for the upper limit. This interval (L, U) will include the value of zero and often $U = -L$. The interval statistical procedure will be used to test the hypothesis:

H_0: $L \le D \le U$
vs.
H_1: $D < L$ or $D > U$.

The closed-interval form of the null hypothesis H_0 is nonstandard in the statistical literature, which typically presents H_0 as a simple equality (e.g., H_0: $D = d$ for some specified d) or an open one-sided interval (e.g., H_0: $D \le U$). The main contribution of the current paper is to illustrate how such interval-based null hypotheses can be used in model validation.

Test statistics for testing means commonly use the t-distribution when the variances are unknown. We assume that the variances of the model and system outputs are unknown but *equal* unless stated otherwise, and thus will use the t-distribution for our hypothesis testing. This procedure requires the data from both the simulation model and the system to be approximately Normal Independent and Identically Distributed (NIID). This can be accomplished in model validation for both the system data and the simulation model data by using 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_s the number of 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 $\alpha_U + \beta_U = 1$.

2.2 Statistical Foundation and Notation

As above, we denote the probability of Type I error by $\alpha_L = \alpha(D = L)$ or $\alpha_U = \alpha(D = U)$. In addition, for convenience of exposition, we assume that the simulation model and the system have equal population variances, i.e., $\sigma^2 = \sigma_m^2 = \sigma_s^2$. Then using straightforward manipulations, we find that the acceptance region A for the hypothesis test H_0 is given by $T \in A$, where

$$A \equiv \left[t\left(\alpha_L, n_m + n_s - 2, \frac{L/\sigma}{\sqrt{\frac{1}{n_m} + \frac{1}{n_s}}}\right), t\left(1 - \alpha_U, n_m + n_s - 2, \frac{U/\sigma}{\sqrt{\frac{1}{n_m} + \frac{1}{n_s}}}\right)\right],$$

the test statistic

$$T \equiv \frac{\bar{M} - \bar{S}}{S_P\sqrt{\frac{1}{n_m} + \frac{1}{n_s}}},$$

the sample mean of the model values is denoted by $\bar{M}$, the sample mean of the system values is denoted by $\bar{S}$, the pooled variance estimator from the $n_m + n_s$ NIID observations is denoted by

$$S_P^2 \equiv \frac{(n_m - 1)S_m^2 + (n_s - 1)S_s^2}{n_m + n_s - 2},$$

and S_m^2 and S_s^2 respectively denote the sample variances of the model and system NIID observations. Further, the notation $t(\gamma, k, \delta)$ denotes the $100\,\gamma\%$ quantile of the noncentral *t*-distribution with *k* degrees of freedom and noncentrality parameter δ (see, for instance, Hines et al. 2003), and $\alpha = \max(\alpha_L, \alpha_U)$.

We denote the probability of Type II error by $\beta_L = \beta(D = L)$ or $\beta_U = \beta(D = U)$. Then by definition of the test statistic T,

$$\beta_L = P(\text{Accept } H_0 \mid H_0 \text{ false with } D = L) = P(T \in A \mid D = L)$$

and

$$\beta_U = P(\text{Accept } H_0 \mid H_0 \text{ false with } D = U) = \text{P}(T \in A \mid D = U).$$

2.3 Interval Statistical Procedure

We now outline a new statistical procedure for model validation that incorporates the augmented interval-based null hypothesis. Table 1 presents the explicit steps needed to carry out the interval statistical procedure to determine if the mean of a specific simulation model output variable has sufficient accuracy to satisfy its acceptable range of accuracy for a specific set of experimental conditions. Step 0 gives what should be determined in the validation process prior to the start of the interval statistical procedure. Subsection 2.1 contains information regarding both Steps 0 and 1. In Step 2, the Initial Sample Sizes for

the model and system are selected. The sample data are collected from both the model and system and analyzed to obtain estimates of the means and variances of the model and system output of interest. The pooled variance estimate is calculated using the two output variance estimates and sample sizes.

Step 3 involves investigating the trade-off for different values of α and β at both L and U regarding the model builder's and model user's risk curves; the purpose of the investigation is to select the β values for L and U to be used in the hypothesis test. Step 4 concerns the evaluation of sample sizes larger than the Initial Sample Sizes with respect to their effects on the model builder's and model user's risk curves to select the Final Sample Sizes to be used in the hypothesis test.

Step 5 involves collecting additional data as determined in Step 4 and analyzing all of the collected data to obtain overall estimates of the means and variances of the model and system outputs, as well as an updated pooled variance estimate. Then, using these new estimates, the hypothesis test acceptance region and test statistic are calculated, and the risk curves are re-calculated. Next we determine if the test statistic falls within or outside of the acceptance region to ascertain whether or not the model is invalid or accepted as valid for this specific test with the risks as shown by the final risk curves.

Table 1: 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 that is to be 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 Sample and Analysis • Select Initial Sample Sizes for the model and system. • Collect the sample data. • Analyze the sample data to obtain model and system mean and variance estimates for the hypothesis test. • Calculate the pooled variance estimate.
Step 3: Investigate Alpha-Beta Trade-off • Select the beta values for L and 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 and 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 U to use in the hypothesis test.
Step 4: Investigate Sample Sizes • Evaluate the model builder's and the model user's risk curves using different (larger) sample sizes than the Initial Sample Sizes. Sample sizes should be selected from a set of feasible choices. (Note: The model and system variance estimates calculated in Step 2 are used to calculate each new pooled variance estimate for the increased sample sizes used for developing each new set of risk curves for the increased sample sizes.) • Determine Final Sample Sizes.
Step 5: Conduct Hypothesis Test • Collect additional samples if sample sizes were increased in Step 4. • If new samples have been collected, calculate the new sample mean and variance estimates as appropriate, and a new pooled variance estimate. • Calculate the acceptance region, the test statistic, and the final risk curves using the selected β values from Step 3, all sample values, 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.

2.4 Computer Software

Appendix A gives a snippet of R code (R Development Core Team 2008) to illustrate how one calculates the test statistic T, the pooled variance S_P^2 (if σ^2 is not specified), the acceptance region A of the test statistic T, and the value of $\beta(\delta)$ when $\delta = D - U$ is specified for the upper case, or when $\delta = D - L$ is specified for the lower case. The inputs are $\beta(L)$, $\beta(U)$, S_m^2, S_s^2, n_m, n_s, L, and U for the function $\beta(\delta)$, and $\bar{M}$ and $\bar{S}$ are additionally specified for the calculation of the test statistic.

3 SIMULATION EXAMPLES

In this section, we illustrate our new interval statistical procedure for simulation model validation via several simulation model examples. The interval statistical procedure given in Table 1 will be followed, and we will use the specially developed computer software briefly discussed in Subsection 2.4. We choose a very simple system to model in our examples as the purpose of the examples is to demonstrate how the new interval statistical procedure works.

3.1 System and Model Descriptions

The system to be modeled is a single-server queueing system with an infinite allowable queue with customers served in order of arrival. The performance measure of interest is the *mean time* for the costumers to traverse this queueing system in steady state. (For a discussion on queueing systems and the queueing results used in this section, see, e.g., Gross et al. 2008.) The true *system* under study is an M/G/1 queueing process whose service time distribution is log normal. The Poisson arrival process has an arrival rate of 0.1 per minute and the service time has a mean of 5 minutes, which gives a system utilization, also called traffic intensity, of 0.5. The *simulation model*, hereafter referred to as the 'model', developed for this system is an M/M/1 queueing model. Let μ_m be the mean time of the model, μ_s be the mean time of the system, and $D = \mu_m - \mu_s$ be the difference between the two mean times. (Note: In all of the calculations, the system mean will be subtracted from the model mean for purposes of consistency.) The accuracy required of the 'mean time' of the model is specified by D having an acceptable range of accuracy of $L = -1.0$ and $U = 1.0$, unless otherwise specified. We wish to test the validity of this model using the interval statistical test for the experimental condition where the customer arrival rate is 0.1 per minute and the service time has a mean of 5 minutes. In our model the arrival process is Poisson and the service time distribution is exponential as our model is an M/M/1 queueing model. We now have specified the requirements of Step 0 of the Interval Statistical Procedure in Table 1.

Thus, the arrival processes for the model and the system are identical as are the means of the service times. The only differences between the model and the system are the service time distributions which in any case have identical means. (We note that the log normal distribution used for the system service time is a two-parameter distribution whereas the exponential distribution used for the model is a one-parameter distribution.) Alternatively, if we view the true M/G/1 system as another simulation model, then the examples can be viewed as validating a simulation model against another simulation model instead of against a system.

In Step 1 of our Procedure given in Table 1, we first specify the statistical hypothesis to be tested. This is the hypothesis given in Subsection 2.1. Next the statistical test to be used is selected — the two-sample *t*-test discussed in Subsections 2.1 and 2.2. This test requires NIID observations. We obtain approximately NIID observations from both the model and system using the method of batch means as discussed in Section 2. For both the model and system we will use a truncation point of 750 customers

and batch sizes of 750 customers. Sample sizes refer to the number of batch means (which we use as approximately NIID observations).

3.2 Example One

We next carry out Step 2 of the Procedure described in Table 1, as we have completed Step 1. We select the Initial Sample Sizes to be n_s = 10 and n_m = 15. We collect the NIID observations (batch means) of these sample sizes for the system and model. Analyzing the collected data, we obtain $\bar{S}$ = 11.645, S_s^2 = 1.386, $\bar{M}$ = 9.888, and $S_m^2 = 1.015$. Next the pooled variance estimate using the formula in Section 2 is $S_P^2 = 1.160$.

Moving to Step 3, 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 1 and 2 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 1 is ($D - L$) and in Figure 2 is ($D - U$). (Note, e.g., that the value of D is –1 at the x-axis location of 0 in Figure 1.) The model builder's risk curves are the curves on the right side of zero in Figure 1 (Lower Case Curves) and to the left of zero in Figure 2 (Upper Case Curves). The model user's risk curves are the curves on the left side of zero in Figure 1 and on the right side of zero in Figure 2. The risk curves in Figures 1 and 2 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 value 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 $\alpha_U + \beta_U = 1$.)

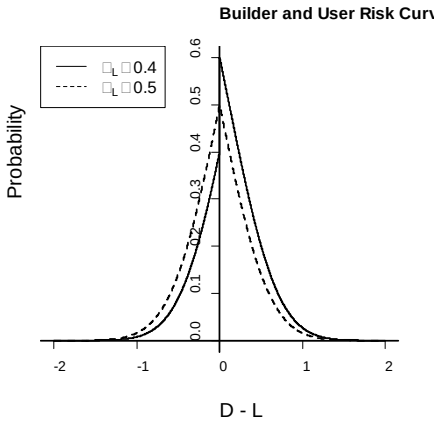

Figure 1: Example One β risk curves at L.

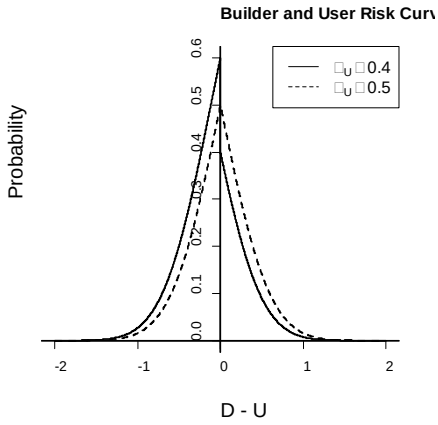

Figure 2: Example One β risk curves at U.

In Step 4 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 performance measure variance estimates from Step 2 and the increased sample sizes. Our software was used to produce Figure 3, which shows the Initial Sample Size risk curves along with the risk curves for sample sizes of n_s = 15 and n_m = 40, where $S_P^2 = 1.113$. (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_s = 15 and n_m = 40 as the sample sizes for the hypothesis test.

In Step 5 the first action required is to collect the additional samples that are called for from Step 4. We obtain 5 more system NIID observations and 25 additional model NIID observations (batch means).

Next we analyze the total samples of 15 system observations and 40 model observations. We obtained $\bar{S}$ = 11.734, S_S^2 = 2.240, $\bar{M}$ = 9.957, and $S_m^2 = 1.254$. Next we calculate the pooled variance estimate using the formula in Section 2 to obtain $S_P^2 = 1.514$. We use the software program with the newly calculated pooled variance estimate of 1.514 to produce the final risk curves shown in Figure 4, along with the acceptance region for the test statistic *T*, (–2.436, 2.436), and the *T* value of –4.771. 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 (–0.907, 0.907) and the difference between the two sample means is 9.957 – 11.734 = –1.777, 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 curves plotted in Figure 4.

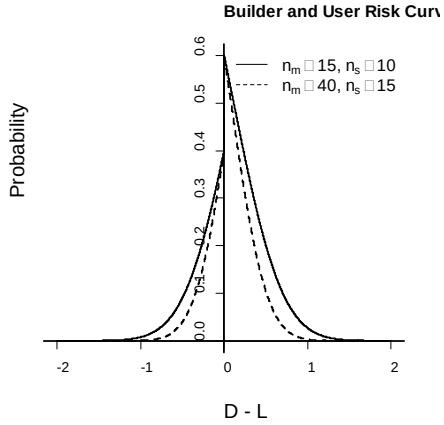

Figure 3: Example One sample size risk curves at *L*.

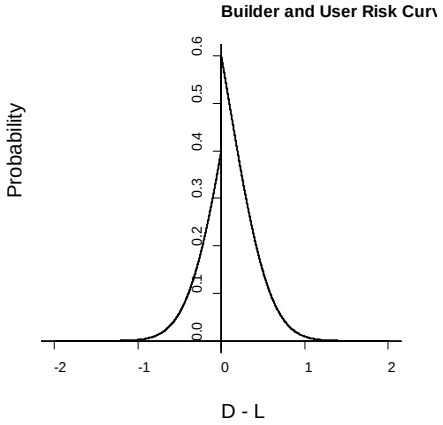

Figure 4: Example One final risk curves at *L*.

The reason that this (simulation) model mean time output is not acceptable is that the service times of the system server have a larger variance than the service times of the model server, while both have the same mean of 5 minutes. The variance of the service time of the log normal used for the system was 42.96. (Note: the parameters of the log normal were (1.11, 1).) The true excepted mean time for both the M/M/1 model and the M/G/1 system can be calculated, and they turn out to be 10 minutes for the M/M/1 and 11.8 minutes for the M/G/1, resulting in the mean time for the system to be larger by 1.8 minutes. We note that the difference *D* is –1.8 minutes, and thus it falls outside the acceptable range of accuracy for *D* of *L* = –1 and *U* = +1. Thus, this example illustrates the effect that the variability of service times has on the behavior of a queueing system; and therefore variances of service times are extremely important when investigating queueing systems.

3.3 Example Two

Example Two will illustrate the use of the interval statistical procedure when *L* is not equal to –*U*. Let Example Two be identical to Example One except that *L* = –2. For Example Two, this allows the model mean time to be acceptable from 2 minutes less than the system mean time up to 1 minute larger than the system mean time; this gives *D* a range of (–2, 1) as opposed to (–1,1) in Example One. Using the same sample sizes and data observations from Step 2 of Example One for Example Two causes $\bar{S}$, S_S^2, $\bar{M}$, S_m^2, and S_P^2 to have the same values for both examples. Using the same β values in Step 3 for Example Two as were used in Example One results in the same risk curves for both Examples except that the location of risk curves for *L* is shifted because of the use of a different value for *L*. Thus, the risk curves shown in Figures 1 and 2 apply to both Examples except that the value of *D* on the *x*-axis of Figure 1 at location 0 for Example Two is –2, whereas for Example One it was –1 because the *x*-axis of Figure 1 is *D* – *L*. The

value of *D* on the *x*-axis in Figure 2 is identical for both Examples One and Two because the same value for *U* is used in both Examples.

Selecting the same β values for the hypothesis test in Step 3 that were selected in Example One, namely, $\beta_L = \beta_U = 0.40$, results in identical risk curves for Examples One and Two when both have the same sample sizes. The only difference in Figure 3 between the risk curves for Examples One and Two is the value of *D* on the *x*-axis, recalling that the *x*-axis in Figure 3 is D – L. Let us select in Step 4 the same sample sizes for the hypothesis test that were selected in Example One, i.e., $n_s = 15$ and $n_m = 40$.

Proceeding to Step 5, the first action is to collect the additional samples as specified in Step 4. Let us use the same additional observations that were used in Example One. Thus, we obtain the same values for $\bar{S}$, S_s^2, $\bar{M}$, S_m^2, and S_P^2 that were obtained for Example One. We use the software program with the newly calculated pooled variance estimate of 1.514 to produce the final risk curves shown in Figure 4, along with the acceptance region for the test statistic *T*, (–5.111, 2.436), and the *T* value of –4.771. Since the value of *T* falls within the acceptance region, the null hypothesis is not rejected, meaning that the simulation model has a mean time output that is acceptable. (The acceptance region can be calculated for *D*, which is (–1.907, 0.907), and the difference of the two sample means is 9.957 – 11.734 = –1.777, which, of course, falls within the acceptance region; this is consistent with the fact that the *T* value falls within its acceptance region.) Therefore, the model has been determined to be valid with the risks as depicted by the risk curves plotted in Figure 4. Note that the accuracy required of the simulation model in this example was not as stringent as Example 1**,** which allowed the model to be acceptable.

3.4 Example Three

Example Three differs from Example One by using a different variance for the system service times than what was used in Example One. We proceed as in Example One. We start with Step 2 of the procedure given in Table 1 since Step 1 has been completed as discussed in Section 3.1. We select Initial Sample Sizes of $n_s = 10$ and $n_m = 20$ and proceed to collect the NIID observations (batch means) of these sample sizes on the system and model. Analyzing the collected data we obtain $\bar{S} = 9.337$, $S_s^2 = 1.558$, $\bar{M} = 10.202$, and $S_m^2 = 1.769$. Next the pooled variance estimate using the formula in Section 2 is calculated to obtain $S_P^2 = 1.701$.

Proceed to Step 3 where we evaluate the risk curves using different β values to select β values for *L* and *U* to use in the hypothesis test. In our example the same β values will be used at *L* and *U*. We use our software for this statistical procedure to obtain the risk curves shown in Figure 5 for β values of 0.3 and 0.5. Note that the pooled variance estimate obtained in Step 2 is used in calculating these risk curves. The risk curves are examined to evaluate the trade-offs between the model builder's risk and the model user's risk for different specified β values at *L* and *U*. Based on our evaluation of the risk curves, we select $\beta_L = \beta_U = 0.30$ as our values to use in the hypothesis test for determining validity. (This gives $\alpha_L = \alpha_U = 0.70$ since $\alpha_L + \beta_L = 1$ and $\alpha_U + \beta_U = 1$.)

In Step 4 we explore increased feasible sample sizes to determine their effects on the risk curves. The pooled variance estimate must be calculated for each new pair of increased sample sizes using the variance estimates from Step 2 and the increased sample sizes. Our software was used to produce Figure 6 which displays the Initial Sample Size risk curves along with the risk curves for sample sizes of $n_s = 15$ and $n_m = 35$ using $S_P^2 = 1.708$. (These sets of risk curves are for $\beta_L = \beta_U = 0.30$ as these values were the ones selected in Step 3.) After exploring different feasible sample sizes, we decided to use $n_s = 15$ and $n_m = 35$ as the sample sizes for the hypothesis test.

In Step 5 the first action required is to collect the addition samples that were determined in Step 4. We need to obtain 15 more model NIID observations (batch means) and 5 more system NIID observations. Next we analyze the total samples of 35 model observations and 15 system observations. We obtained $\bar{S} = 9.472$, $S_s^2 = 0.844$, $\bar{M} = 9.981$, and $S_m^2 = 1.058$. Next we calculate the pooled variance estimate using the formula in Section 2 to obtain $S_P^2 = 0.996$. We obtain from the software program using the newly calculated pooled variance estimate of 0.996 the final risk curves shown in Figure 7, the

acceptance region for the test statistic *T*, (–2.717, 2.717), and the *T* value of 1.651. Since the value of the *T* statistic falls within the acceptance region, the null hypothesis is not rejected, meaning that the simulation model has a mean time that is acceptable. (The acceptance region can be calculated for *D* which is (–0.837, 0.837), and *D* is 9.981 – 9.472 = 0.509, which falls inside the acceptance region; this is consistent with the fact that the *T* value falls inside its acceptance region.) Thus, the model has been determined to be valid with respect to the mean system time with the risks as shown by the final risk curves plotted in Figure 7. (Of course, validity of the mean behavior does not imply other measures such as variances or distributions are also valid. Each property or function of a random output of interest must be tested individually for validity.)

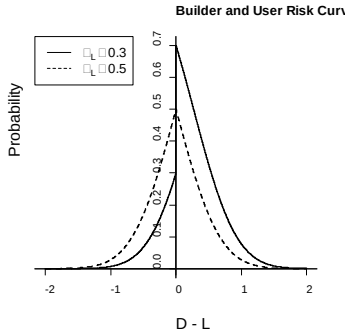

Figure 5: Example Three β risk curves at *L*.

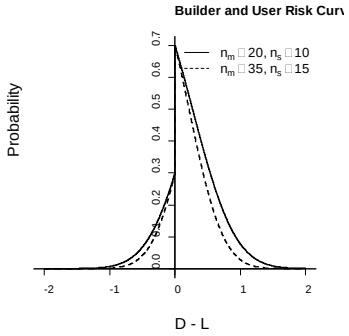

Figure 6: Example Three sample size risk curves at *L*.

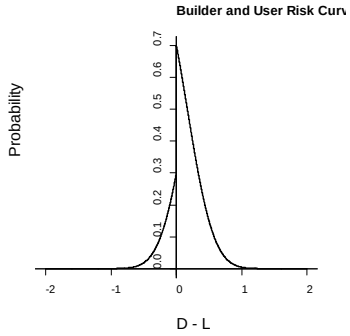

Figure 7: Example Three final risk curves at *L*.

The reason that this (simulation) model mean time output is acceptable is because the service times of the system server have the same mean as the model and a variance of 18.88 that is close to that of the model server's variance of 25.0. (Note that the parameters of the log normal used as the distribution for the system server were (1.33, 0.75).) The true expected mean time for the M/M/1 model is 10 minutes and for the M/G/1 is 9.39 minutes. The difference *D* between these two means is 0.61 minutes and falls within the acceptable range of accuracy for *D* of $L = -1$ and $U = +1$. If the acceptance range of accuracy was smaller, then the simulation model may not have the accuracy required.

3.5 Example Four

Example Four will illustrate the use of the interval statistical procedure when the β values at *L* and *U* are different. Let Example Four be identical to Example Three except that values for β at *U* will be different. Using the same sample sizes and data observations from Step 2 of Example Three for Example Four results in $\bar{S}$, S_s^2, $\bar{M}$, S_m^2, and S_P^2 having the same values for both examples. In Step 3 we investigate the risk curves for different β values at *L* and *U* separately and also decide the β values to use for hypothesis testing at *L* and *U* separately. Let us consider *L* first. The same β values at *L* for Examples Three and Four will produce the same risk curves for both Examples of which two sets of risk curves are shown in Figure 5. Let us select for the hypothesis test in Example Four the same β value of 0.30 for *L* that was selected in Example Three and whose risk curves are shown in Figure 5. Now we evaluate the risk curves at *U* for different values of β. Two sets of risk curves are shown in Figure 8 for β values of 0.2 and 0.4 at *U* obtained from our software which used the pooled variance estimate obtained in Step 2. After evaluating the trade-offs between the model builder's risk and the model user's risk for different specified β values it was determined to use a β value of 0.2 at *U* for the hypothesis test.

Moving to Step 4 where we decide on the sample sizes to use for the hypothesis test by exploring increased feasible sample sizes to determine their effects on the risk curves. The pooled variance estimate must be calculated for each new pair of increased sample sizes using the variance estimates from Step 2

and the increased sample sizes. The effect of sample sizes on the risk curves must be evaluated at *L* and *U* together as only one set of sample sizes are used for our hypothesis test. Our software was used to produced Figure 9 which displays for *U* the Initial Sample Size risk curves along with the risk curves for sample sizes of n_s = 15 and n_m = 35 using $S_P^2 = 1.708$. (These sets of risk curves are for β_U = 0.20 as these values were the ones selected in Step 3.) In Example 3 we produced Figure 6 which contains the risk curves at *L*, which would have been produced here if it had not already been displayed. After exploring different feasible sample sizes for both *L* and *U*, we decided to use n_s = 15 and n_m = 35 as the sample sizes for the hypothesis test.

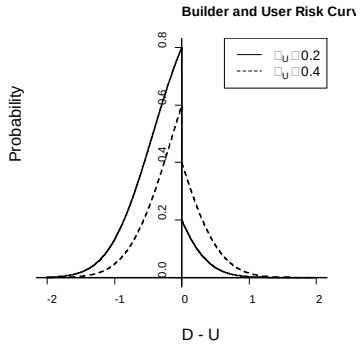

Figure 8: Example Four β risk curves at *U*.

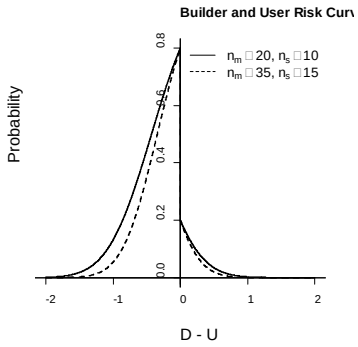

Figure 9: Example Four sample size risk curves at *U*.

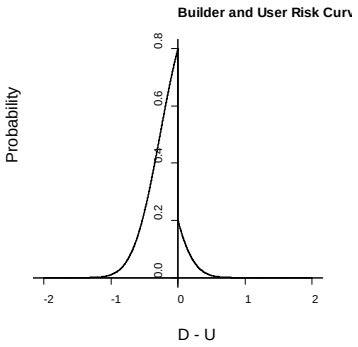

Figure 10: Example Four final risk curves at *U*.

Proceeding to Step 5, the first action is to collect the additional samples as specified in Step 4. Let us use the same additional observations that were used in Example Three. Thus, we obtain the same values for $\bar{S}$, S_s^2, $\bar{M}$, S_m^2, and S_P^2 that were obtained for Example Three. We use the software program with the newly calculated pooled variance estimate of 0.996 to produce the final risk curves shown in Figure 7 (also produced in Example Three) and Figure 10, along with the acceptance region for the test statistic *T*, (–2.717, 2.393), and the *T* value of 1.651. Since the value of *T* falls within the acceptance region, the null hypothesis is not rejected, meaning that the simulation model has a mean time output that is acceptable. (The acceptance region can be calculated for *D*, which is (–0.838, 0.737), and the difference between the two sample means is 9.981 – 9.472 = 0.509, which, of course, falls within the acceptance region; this is consistent with the fact that the *T* value falls within its acceptance region.) Thus, the model has been determined to be valid with the risks as depicted by the final risk curves plotted in Figures 7 and 10.

Note that the acceptance region for Example Four is smaller than the acceptance region for Example Three. The reason is because the model user's risk was reduced by using a smaller value for β at *U* in Example Four compared to the value used in Example Three. Having a smaller acceptance region reduces the probability of accepting a model as valid and hence reduces the probability of accepting an invalid model as valid. However, there is another side to this issue. Reducing the model user's risk increases the model builder's risk and thus the probability of a valid model being determined to be invalid is increased.

3.6 Some Comments on the Examples

The performance measure considered for model validity was the mean time customers spend in the queueing system. The mean system time of customers, *W*, in the M/G/1 can be calculated by using the Pollaczek-Khintchine (PK) formula (Gross et al. 2008). For the mean arrival rate and mean service time used in the Examples, the PK formula gives $W = 7.5 + 0.1\sigma_s^2$. Using this equation, we have in Table 2 different values for σ_s^2, W, and $D = \mu_m - \mu_s = 10.0 - \mu_s$. The values of *D* and *W* for the M/M/1 model are calculated using σ_s^2 = 25.0, which is the square of the mean service time. Notice that the values selected

for L and U were 10% of W of the M/M/1 except for Example Two where L had a larger negative value. The values for L and U should be determined by the accuracy required of a model which depends on its use. Unfortunately, setting these values is usually difficult and there has been little scholarly research on how to do so. The values of $L = -1$ and $U = 1$ allow service time variances of 15–35. One can readily see from Table 2 that the variance of the service time of Example One (42.96) lies outside of the acceptable range of accuracy, and the variance of the service time of Examples Three and Four (18.88) lies inside the acceptable range of accuracy.

Table 2: Expected values of mean system times for M/G/1.

σ_S^2	0	15.0	18.9	25.0	35.0	43.0
W	7.5	9.0	9.4	10.0	11.0	11.0
D	2.5	1.0	0.6	0.0	–1.0	–1.8

4 SUMMARY

This tutorial paper demonstrated a new hypothesis test for use in operational validity of simulation models. The presentation contained a detailed procedure for the use of the interval hypothesis test in simulation model validation, a short description of the statistics used for this procedure, a brief description of the software that has been developed to use this procedure, and examples of simulation model validation using the procedure and software. Examples involved the testing of a single-server queueing model regarding validity using batch means that are approximately NIID observations. These examples covered the use of the model builder's and model user's risk curves in the selection of the beta-alpha values and the sample sizes. The use of risk curves should be valuable in communicating about the risks involved in Type I and Type II errors for different sample sizes. Such information may help to obtain financial support in order to use larger sample sizes.

A APPENDIX: EXAMPLE OF R CODE

We provide R code to calculate the model user's risk and the acceptance region of the test statistic T when the Type II errors at L and at U are specified.

```
beta = function(betaL,betaU,varm,vars,nm,ns,L,U,delta,sigma2,case){
  #case='upper' or 'lower', if sigma2='NA', pooled variance is used
  alphaL=1-betaL
  alphaU=1-betaU
  df=nm+ns-2                          #degrees of freedom
  Sp2=((nm-1)*varm+(ns-1)*vars)/df    #pooled variance Sp2
  if(case=='upper'){                  #if upper case is specified
    if(sigma2=='NA'){                      #if population variance is unknown
      ncp=(U+delta)/sqrt(Sp2*(1/nm+1/ns))   #non-centrality parameter with delta>0
      lower=qt(alphaL,df,L/sqrt(Sp2*(1/nm+1/ns))) #lower limit of acceptance region,
      #qt:quantile of t distribution
      upper=qt(1-alphaU,df,U/sqrt(Sp2*(1/nm+1/ns)))  #upper limit
    }else{           #if population variance is known (usually not the case)
      ncp=(U+delta)/sqrt(sigma2*(1/nm+1/ns))
      upper=qt(1-alphaU,df,U/sqrt(sigma2*(1/nm+1/ns)))
      lower=qt(alphaL,df,L/sqrt(sigma2*(1/nm+1/ns)))
    }
    beta=pt(upper,df,ncp)-pt(lower,df,ncp)} #calculation of user's risk.
    #pt(upper,df,ncp)=cdf evaluated at upper limit
```

```
  if(case=='lower'){        #if the case is lower
    if(sigma2=='NA'){
      ncp=(L-delta)/sqrt(Sp2*(1/nm+1/ns))      #delta is negative in this case
      lower=qt(alphaL,df,L/sqrt(Sp2*(1/nm+1/ns)))
      upper=qt(1-alphaU,df,U/sqrt(Sp2*(1/nm+1/ns)))
    }else{
      ncp=(L-delta)/sqrt(sigma2*(1/nm+1/ns))
      upper=qt(1-alphaU,df,U/sqrt(sigma2*(1/nm+1/ns)))
      lower=qt(alphaL,df,L/sqrt(sigma2*(1/nm+1/ns)))
    }
    beta=pt(upper,df,ncp)-pt(lower,df,ncp)}
  output=list(lower=lower,upper=upper,beta=beta)
  return(output) #returns lower and upper limits of acceptance region of T,
  #and the model user's risk at a specified delta
}
```

REFERENCES

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

Balci, O. and R. G. Sargent. 1982a. "Validation of Multivariate Response Simulation Models by Using Hotelling's Two-sample T^2 Test." *Simulation* 39(6): 185–192.

Balci, O. and R. G. Sargent. 1982b. "Some Examples of Simulation Model Validation Using Hypothesis Testing." In *Proceedings of 1982 Winter Simulation Conference*, edited byH. J. Highland, Y. W. Chao, and O. S. Madrigal, 620–629. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Balci, O. and R. G. Sargent. 1983. "Validation of Multivariate Response Trace-driven Simulation Models." In *Performance* 83, ed. A. K. Agrawada and S. K. Tripathi, 309–323. North Holland.

Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2010. *Discrete-Event System Simulation*. 5th ed. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Gross, D., J. M. Shortle, J. M. Thompson, and C. M. Harris. 2008. *Fundamentals of Queueing Theory*. 4th ed. New York: John Wiley.

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.

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

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

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

Sargent, R. G. 2104. "An Interval Statistical Procedure for Use in Validation of Simulation Models." *Journal of Simulation*, advance online publication, November 21, 2014; doi:10.1057/jos.2014.30.

AUTHOR BIOGRAPHIES

ROBERT G. SARGENT is a Professor Emeritus of Syracuse University. He received his education at The University of Michigan. Dr. Sargent has published extensively and has served his profession in numerous ways including being the General Chair of the 1977 Winter Simulation Conference (WSC), serving on the WSC Board of Directors for ten years and chairing the Board for two years, being a Department Editor for the *Communications of the ACM*, holding the Presidency and other offices of what is now the INFORMS Simulation Society, serving as the Founding President of the WSC Foundation, and

initiating the Simulation Archive. He has received several awards and honors for his professional contributions including the INFORMS Simulation Society's Lifetime Professional Achievement Award and their Distinguished Service Award, a WSC 40th anniversary landmark paper award, the WSC Wilson Award, the ACM SIGSIM Distinguished Contributions Award, service awards from ACM and IIE, and being elected a Fellow of INFORMS. His current research interests include the methodology areas of modeling, discrete-event simulation, model validation, and performance evaluation. Professor Sargent is listed in *Who's Who in America* and in *Who's Who in the World*. 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 graduate student at the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology, currently working towards a Ph.D. in Industrial Engineering. He received a B.S. in Mathematics from Clayton State University and an M.S. in Mathematics from Georgia Southern University. His research interests include statistical quality control and improvement, high-dimensional data and machine learning, and discrete-event simulation. His email is tryaacoub@gmail.com.