

Simulation Validation Using a Non-Parametric Statistical Method

Capt Brian D. Smith

Air Force Operational Test and Evaluation Center
Kirtland AFB, NM

“This briefing is approved for public release; distribution is unlimited.”
AFOTEC Public Affairs Release Number: 2006-10-09

1.0 Background

A primary advantage to using modeling and simulation (M&S) in a test program is it can often answer test measures that, if answered using real-world data, would require unrealistically expensive, time-consuming, or complex test events. Simulation outputs, though, are only as good as the underlying assumptions and models built into the simulation. For even moderately complex simulations, it is not easy to predict the quality of simulation results based purely on the logic that if component models¹, theoretical component interactions, and simulation inputs are valid, then the simulation results will be as valid as the results of a real-world test event under the same conditions.

1.1. The Problem of Validating Simulation Outputs

Direct comparison of simulation results to real-world test data is often conducted as part of a simulation validation effort. Assuming simulation input parameters are adequately matched to known real-world parameters, differences between simulation outputs and real-world test data could imply: 1) the test data is not statistically representative (contains multiple outliers), or 2) the simulation is not an adequate representation of the real process or system given the chosen input parameters. The analyst charged with simulation validation needs to eliminate the first possibility before focusing on the second possibility.

The analyst is faced with a difficult question: has enough data been collected under these conditions to be considered a statistically significant sample for comparison with the simulation? Consider a validation effort for a Monte Carlo simulation that produces one or more static parameters of interest to a test program. Such static parameters might include target miss distance in a weapon system simulation, breakage rate of a physical component in a reliability simulation, or supply depletion in a campaign simulation. Proponents of M&S in testing sometimes argue that a simulation will predict these static parameters in operating regimes that were never exercised in the real-world. For example, the simulation will predict weapon probability-of-kill at a launch altitude of 200 ft when, due to safety restrictions, the real-world launch tests never took place below 1000 ft. This is a paradox of sorts: the simulation is supposed to alleviate the need to conduct costly live tests, but the live tests are the best indication that the simulation is producing realistic results.

1.2. Population Characterization of the Simulation Outputs

A Monte Carlo simulation induces random noise at various points in the simulation process to replicate random variation that will occur in the real-world process or system. These random inputs undergo linear and/or non-linear operations in the simulation process, and the outputs of these processes will form statistical distributions that may be difficult to characterize analytically. For example, in a complex, mixed linear/non-linear simulation, Gaussian (normal)-distributed random inputs do not guarantee Gaussian-distributed outputs. But, because simulations can be run many times over

¹ “Component model” is used in the broad sense to describe individual pieces within an overarching simulation rather than the more narrow sense to describe physical “components” being modeled in the simulation.

(producing thousands of data points), it is fairly easy to numerically approximate the population parameters².

One might consider a real-world test event to be a “simulation” in which each event outcome is essentially a random sample from some hypothetical statistical population. In general, what the simulation validation analyst would like to show is the hypothetical test event population is statistically similar to the hypothetical (or computed) simulation population. Like the simulation output populations, the live event data populations may be difficult or impossible to predict analytically. Since test events may be duplicated only a few times (or not at all), few data points are generated, and sample statistics may produce very skewed estimates of the population statistics. Furthermore, it is likely that the populations for unique experiments are statistically different, because the system performance may differ for each set of test parameters³.

If the analyst could rely on data across experiments, the sample size is effectively increased, and the validation data becomes more statistically significant. Also, if the technique used to aggregate the test data from different experiments could account for the *expected performance differences* between experiments, the analysis of aggregate results might be more justifiable. The technique described below will attempt to combine data across experiments so the aggregate result is useful in determining how faithful a simulation models real system performance.

2.0 Technique

In section 2.1, the mathematical foundations of the technique will be presented. In section 2.2, previous work using related techniques will be acknowledged.

2.1. Mathematical Foundations

Consider Equation 1 relating the cumulative probability distribution function (cdf) with its corresponding probability density function (pdf). The random variable X is continuous over the interval $[-\infty, +\infty]$. The upper limit of integration, x , is a constant.

$$F_X(x) = \int_{-\infty}^x f_X(t)dt \quad (1)$$

By definition, $F_X(x)$ is always a non-decreasing function.

Consider a derivation shown in Leon-Garcia [3] that assumes a random variable, Z , such that

$$Z = F_X^{-1}(U) \quad (2)$$

where U is a uniformly distributed set of values between 0.0 and 1.0. The random variable Z is the *inverse cdf* for the random variable X with cumulative probabilities expressed as

² Population parameters refer to all parameters needed to describe a particular population of interest. For example, a Gaussian population’s sufficient parameters (statistics) would be mean and standard deviation. Throughout this paper, the term “population” will be used interchangeably with “population parameters” since the sufficient parameters uniquely define their associated population.

³ In fact, if performance was not expected to differ across dissimilar experiments, one might question entirely the need for conducting multiple experiments.

$$P[Z \leq x] = P[F_X^{-1}(U) \leq x] \quad (3)$$

By applying the cdf, $F_X(\bullet)$, to both sides of the rightmost bracketed inequality in (3), the result becomes

$$P[Z \leq x] = P[F_X(F_X^{-1}(U)) \leq F_X(x)] = P[U \leq F_X(x)] \quad (4)$$

Because U is uniformly distributed in the range $[0,1]$, for a constant h , $P[U \leq F_X(x) = h] = h$, and

$$P[Z \leq x] = h = F_X(x). \quad (5)$$

Equation 2 implies that a uniformly distributed set of numbers between 0 and 1 can produce samples in x that obey an arbitrary probability distribution (see Figure 1). This inverse cdf theorem is often used in computer algorithms to produce random numbers obeying a desired distribution using a simple *uniform* random number generator. Note that this theorem is completely general. The principle applies to *any and all probability distributions* (although not all continuous distributions have a closed-form inverse cdf).

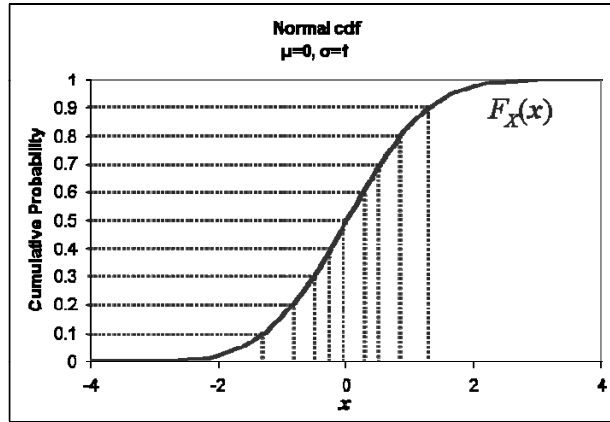


Figure 1: Inverse cdf example. Using a normal cdf as an example, one can see how uniformly spaced cumulative probabilities, when mapped through the cdf, produce normally distributed data in x .

Now consider a sample of values, S_X , composed of individual data elements s_x . Assume some unspecified continuous cdf, $F_X(x)$, exists, and one wishes to determine whether or not the sample S_X is distributed according to X . Equation 2 states that, when the inverse cdf $F_X^{-1}(U)$ operates on the uniformly distributed set U , the result is the random variable Z . Conversely, if the data in sample S_X is truly distributed as X , then, for large sample sizes, the cumulative probabilities that result from $F_X(S_X)$ will be distributed uniformly.

This can be restated in terms of test events and populations. Consider the sample S_X to be composed of individual data elements, s_x , that represent any measure of interest acquired from real-world experiments. Each data element is, in effect, a random sample from a hypothetical (and perhaps uncharacterized) statistical population. Assume that a simulation predicts a particular statistical population⁴, X_{sim} , and the analyst wishes to decide if the sample S_X is distributed according to X_{sim} . If S_X is sufficiently large, then $F_X(S_X)$ should result in a uniformly distributed set of cumulative probabilities.

⁴ “Predicts” merely implies the population characteristics could be computed numerically from multiple simulation trials.

If S_X is not distributed according to X , then uniformity should not be expected for large samples. One might call a simulation “ideal” when its output statistical populations precisely match those of the real-world process or system.

Note this development makes *no assumptions about underlying population distributions*. A conventional “goodness-of-fit” test applied directly to the sampled data requires assumptions about statistical populations and requires multiple data points from each hypothesized population. By testing for cumulative probability uniformity, even if each experiment represents a different population pair (simulation and test), the cumulative probabilities can be analyzed in aggregate⁵. In essence, the analyst can examine a wider set of test factors (valuable for demonstrating *overall system performance* across many regimes) while producing a statistically significant data set used to validate a simulation prediction of a *specific parameter of interest*. This technique trades statistical significance at the individual experiment level for statistical significance in the aggregate. Fewer repetitions are conducted for each experiment, but one can exercise a broader sampling of the parameter space (i.e., larger number of unique experiments) and still keep the total number of trials fixed.

2.2. Previous Work

The technique described herein benefits greatly from an assortment of internal government working papers and correspondence produced in support of the AIM-9X missile program. That simulation validation effort applied R.A. Fisher’s combined probability test (an already well established technique in biological research) to the unique problems of operational testing. As part of a related effort, Arthur Fries of the Institute for Defense Analyses (IDA) published a paper [2] for the 6th Annual U.S. Army Conference on Applied Statistics. Relying on some discussions with A. Rex Rivolo (also of IDA), Fries’ paper investigated the theoretical underpinnings for a Fisher combined probability test applied to simulation validation. This paper is intended to demonstrate the utility of Fries’ techniques through “simulated” validation problems involving artificial data. The intent is to move beyond the theoretical statistics and into the realm of widespread application. Also, unlike previous works, this paper does not rely on Fisher’s logarithmic transformation of tail probabilities, but instead uses the raw cumulative probabilities in a Kolmogorov-Smirnov goodness-of-fit test. Nonetheless, the contributions of Fries, Rivolo, and many “behind the scenes” analysts cannot be overemphasized.

3.0 Demonstration of Theory

The techniques in Section 2.1 are entirely general in the sense that the only assumption is simulation and test results can be represented as *ratio data*⁶. To demonstrate an application of the described technique, a set of software tools was developed to produce “simulation data” and “test data”. In an actual simulation validation problem, simulation data would be produced by software that simulates a process or system, and test data would be produced in a real-world experiment involving the real process or system. By creating simulation and test data from a set of software tools (as is done here), it is possible to verify the statistical theory described in Section 2.1. Theory verification would not

⁵ The basis for this fortunate fact is cumulative probabilities (regardless of which experiment produced them) are hypothesized to belong to a uniform distribution over the range 0.0 to 1.0 (regardless of the underlying population distributions). Uniform distributions are completely described by their bounds, so *all* cumulative probabilities from *all* experiments are hypothesized to be random draws from a uniform distribution between 0.0 and 1.0.

⁶ Ratio data, in statistical parlance, must have a natural zero starting point. Also, ratios of the data, as well as differences, are meaningful comparisons. A large majority of real-world test data is ratio data.

be possible using an actual validation problem with real data, but simulation validation *is* the intended application for this technique.

The software tools used here also perform the uniformity tests on the cumulative probabilities. This function is identical to what would be done in an actual validation problem, so these tools have applicability beyond verification of the statistical theory. A brief discussion of the methodology in the tools follows.

3.1. Creating Simulation and Test Cases

In this analysis, a simulation/test pair will involve random variables X_{sim} and X_{test} . Each pair will produce a single realization (random draw), x_{test} , from X_{test} , and this should be interpreted as the measure or statistic computed from a real-world experiment. Each pair will also produce a vector of realizations, \mathbf{x}_{sim} , from X_{sim} , and this should be interpreted as the set of predicted values of the measure or statistic produced from a Monte Carlo simulation. Each set \mathbf{x}_{sim} will contain results from 100 Monte Carlo trials. The mean and standard deviation of the vector \mathbf{x}_{sim} could function as an estimator of x_{test} . Note that an individual pair represents the results of a real-world test conditioned upon a given set of test conditions and a given system configuration. This test would be subsequently reproduced in a simulation exercise.

Multiple pairs will comprise a set of cumulative probabilities for uniformity testing. This is analogous to completing several test events (or experiments), each with a different set of test factors. Notationally, the i -th individual case, and the realizations for that case, would be denoted as

$$\begin{aligned} X_{sim}^i &\rightarrow \mathbf{x}_{sim}^i = \{x_{sim,1}^i, x_{sim,2}^i, \dots, x_{sim,n}^i\} \quad \text{for } i = 1, \dots, I \quad \text{and for } n = 1, \dots, N \\ X_{test}^i &\rightarrow x_{test}^i \quad \text{for } i = 1, \dots, I \end{aligned} \quad (6)$$

where I is the number of pairs in the series of experiments and N is the number of Monte Carlo trials in each case. The i -th pair produces a single cumulative probability p_i . For this analysis, $I = 10$.

The techniques here can be extended to multi-dimensional data, but this paper will focus on one-dimensional data for simplicity of discussion. For thoroughness of analysis, multiple sets of experiments, or batches, will be run. Each batch will contain a different set of simulation/test pairs, so the results produced by the theory in Section 2.1 can be viewed in an aggregate sense⁷. Batch analysis will involve 100 sets of I cases.

3.2. Replicating Simulation and Test Data

The simulation data and test data will be formed as sets of random draws from the random variables X_{sim} and X_{test} and implemented in the software tools. To demonstrate that the technique does not rely on assumptions about underlying probability distributions, the analysis will include data derived from both Gaussian and Rayleigh random variables⁸. Remember from Section 2.1 that an “ideal” simulation would produce data from the identical statistical population that produces test data. In these tools, an “ideal” simulation would use X_{sim}^i having identical means and covariances as the test variables

⁷ I cases produce x_{test}^i and p_i for $i=1, \dots, I$. A single uniformity test requires all I cases in the set, so a batch process allows analysis of many uniformity tests on widely varied data.

⁸ It can be shown that W , the root sum square ($W = \sqrt{X^2 + Y^2}$) of two jointly Gaussian random variables, X and Y , is Rayleigh distributed [3]. Since Gaussian random number generators are readily available in many software packages, Rayleigh data provides an easily implemented alternative probability distribution.

X_{test}^i . Conversely, non-ideal simulations imply X_{sim}^i has different statistics than X_{test}^i . Note that test data and simulation data will always be the result of independent random draws from the corresponding random variables.

3.3. Cumulative probabilities for each test

Per the development of Section 2.1, a cumulative probability, p_i , is computed for each test realization x_{test}^i based on the approximated cumulative distribution function belonging to X_{sim}^i . For this analysis, the simulation data will be used as a discrete approximation to the cdf of X_{sim} , and interpolation between points $x_{sim,n}$ and $x_{sim,n+1}$ will be performed to smooth the approximation. Remember that in real-world validation problems, X_{test}^i is totally unknown, so the theory in Section 2.1 only exploits knowledge of X_{sim}^i (which is always known). In theory, one could compute cumulative probabilities based on some *a priori* knowledge of the statistical distributions for X_{sim} . Numerical approximation has the added benefit of showing that the theory requires no assumptions about distributions of data. Cumulative probabilities will be tested for uniformity using a Kolmogorov-Smirnov test, but other “goodness-of-fit” tests could be equally valid.

4.0 Results

Six cases were run using the previously described software tools, where each case contains 100 sample sets (or batches) of cumulative probabilities. Each batch includes $I = 10$ simulation/test event pairs (yielding ten cumulative probabilities). The i -th pair represents 100 random draws from the simulation random variable X_{sim}^i and a single “test event” draw from the random variable X_{test}^i .

A K-S test for uniformity is performed on every batch of I cumulative probabilities, producing 100 K-S test statistics. Other goodness-of-fit tests could be applied here, but the K-S test is mathematically straightforward and lends itself to visualization. A Kolmogorov-Smirnov test, as applied to a single batch of I probabilities hypothesized to be uniform, compares the maximum y-axis deviation of the expected cumulative frequencies from the ideal uniform cdf (denoted in Figure 2 by the black line of slope 1.0). *Expected cumulative probability* (equivalent to frequency) of the i -th *empirical cumulative probability* is computed as $p_{exp,i} = i / I$ for $i=1, \dots, I$, where I represents the total number of cumulative probabilities in the set. The gray-shaded region represents the region of acceptance for the K-S test as determined by the K-S test critical value⁹. Any single probability falling outside this region will cause its batch to fail the K-S test for uniformity, and some batches may have more than one probability outside the shaded region [1].

In the first three cases, the I Gaussian or Rayleigh random variables have independently and randomly chosen means and standard deviations (within predefined bounds)¹⁰, and all 100 I pairs are distinct (i.e., there is no repetition of random variable parameters within any sample set or across sample sets). The randomly chosen random variable parameters are intended to reduce the possibility that a chance combination of parameters makes the theory appear valid when it might be generally invalid.

⁹ The K-S test critical value is a function of the chosen α (0.1 in these examples) and the sample size ($N = 10$).

¹⁰ The Rayleigh data is created from two, independent Gaussian random variables with randomly chosen means and standard deviations.

However, this technique confounds two sources of uncertainty: 1) the randomly chosen population parameters, and 2) the randomly drawn test event and its associated cumulative probability.

Cases 3 through 6 eliminate the first source of uncertainty inherent to Cases 1 through 3 by fixing population parameters for all batches. In other words, the parameters for the i -th simulation population are identical for all 100 batches, and the parameters for the i -th test event population are identical for all 100 batches. Within a batch, however, the parameters for the $I=10$ populations (both simulation and test event) will differ. The remaining uncertainty factor is the test event random draw that occurs for each pair in a batch, but this uncertainty will always exist in a real-world validation problem. The i -th test event random draw is independent for each of the 100 batches.

4.1. A Visual Inspection

Before analyzing the K-S test results, it is worth looking at the raw cumulative probabilities. Figure 2 shows expected cumulative frequencies (reduced to probabilities) versus the empirical cumulative probabilities computed from the approximated cdfs for all 100 runs in a case. Probabilities shown in Figure 2 were derived from the Gaussian-distributed data, and the Rayleigh data (not shown) was virtually identical in appearance.

Note that the probabilities in Figure 2(a) have a structure that mimics that of the shaded region. These probabilities were produced using an “ideal” case in which the simulation random variables were statistically identical to the test random variables. Incidentally, as the alpha value on the K-S test changes, the shaded region would widen (for smaller alpha) or narrow (for larger alpha), thereby changing the number of points falling inside the region of acceptance. The probabilities in Figure 2(b) came from the significantly skewed Case 3 (to be described in more detail below), and the structure here is noticeably different than in 2(a). Case 2 (not shown) yields probabilities with a structure somewhat between Figures 2(a) and (b). Most importantly, because visual analysis of probability data is rather subjective, a goodness-of-fit test, like K-S, generally makes a more powerful statement about uniformity.

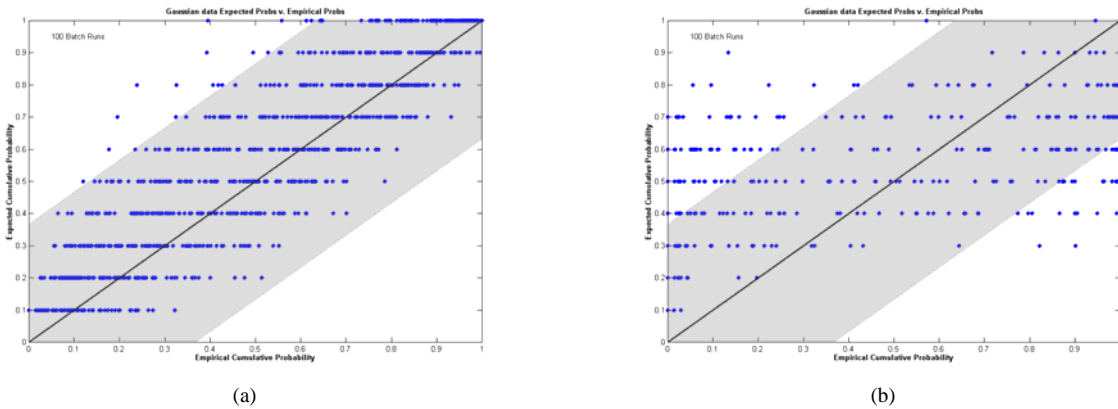


Figure 2. Expected cumulative frequencies (probabilities) are plotted versus the cumulative probabilities computed from the approximated pdfs of Gaussian distributed data: (a) the “ideal” simulation of Case 1, (b) the significantly skewed sim of Case 3. Each figure contains 1000 cumulative probabilities (100 batches of 10 probabilities).

The first case assumed an ideal simulation in which X_{sim}^i is statistically identical to X_{test}^i for all I pairs and all sample sets. Section 2.0 theory would predict that cumulative probabilities within each batch of size $I=10$ should appear uniformly distributed.

4.2. Testing for Uniformity

Case 1

Histograms for the K-S test results were created for each case, and the resulting cumulative distributions are presented below¹¹. The two-sided K-S test critical value for $\alpha=0.1$, $I=10$ is 0.368 and is denoted by a red vertical line in the cdf plots. Figure 3 shows the critical value mapping near the 95-th percentile for both the Gaussian and Rayleigh data, implying about 95% of all sample sets passed the K-S test for uniformity. Passing the K-S test implies that the simulation is faithfully predicting the test distribution, and the simulation can be considered a good predictor of system/process outcome.

The 5% of batches that failed the K-S test represent a Type I error region. Type I error, in this context, is committed by declaring a simulation to be a poor predictor of the real-world process when it is, in fact, a faithful predictor. Figure 3 demonstrates that, about 5% of the time, a single sample set of size $I=10$ would fail a K-S test even though the simulation and test populations are statistically identical. As with all statistical tests, the random variation in the sample leads to this Type I error.

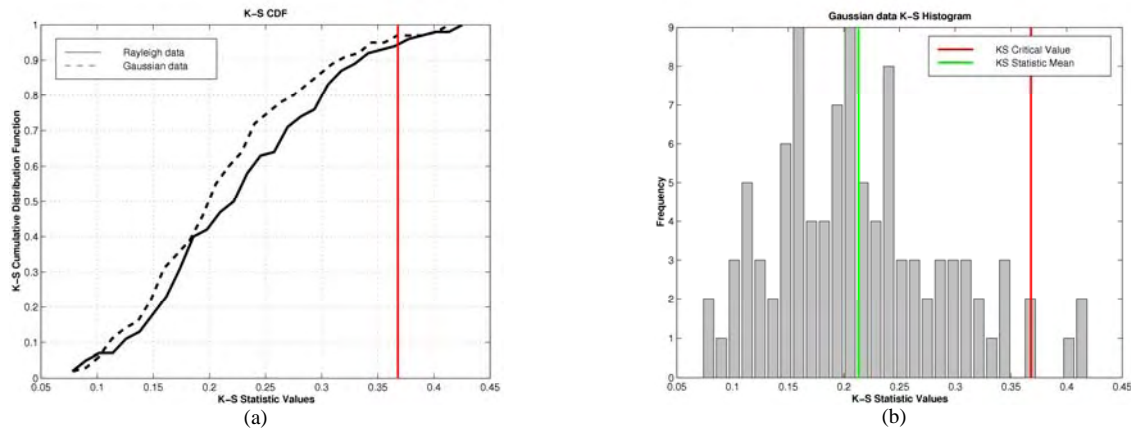


Figure 3: Distribution of K-S test results for Case 1, the “ideal” simulation:

- the cdfs for the Gaussian data results (dashed line) and the Rayleigh data results (solid line). The K-S test critical value (0.368) is denoted by the red vertical line. The intersection of the cdf and critical value indicates the percentage of batches that passed the K-S test for uniformity.
- the histogram of the Gaussian data results, with red vertical line showing the K-S critical value and green vertical line denoting the mean of the histogram.

Case 2

The second case assumed the simulation does not faithfully reproduce the test populations. Results appear in Figure 4. The test distributions for this batch were randomly chosen such that their means are between 50% and 150% of the corresponding simulation means (i.e., skewed 50% in both directions) and their standard deviations are between 100% and 110% of the corresponding simulation standard deviations (i.e., as much as 10% greater variation about the mean)¹². The K-S cumulative

¹¹ The K-S test histograms presented here are used only for visualization of the batch results, and they are not related to the pdf/cdf of the underlying data. In a real validation problem, only one K-S test is performed, so K-S histograms and K-S cumulative distributions would be unnecessary.

¹² When producing skewed Rayleigh data, the means and standard deviations of the component Gaussian random variables are adjusted prior to construction of the Rayleigh data.

distributions show the critical value maps near the 65-th percentile for both Gaussian and Rayleigh data. This implies 35% of all sample sets failed the K-S test for uniformity.

Depending on the specific combinations of simulation/test event population pairs in Case 2, some batches might represent fairly “good” simulations while other batches might represent fairly “bad” simulations. Therefore, Case 2 should be viewed as a plausible real-world validation scenario in which simulation predictions are presumed to be fairly close to real-world results, but the analyst cannot easily determine how close they truly are. Lastly, in Case 2 batches, a few population pairs might be statistically similar while other pairs in the same batch might be statistically dissimilar¹³. Cases 5 and 6 will address this problem more thoroughly. For the variations in simulation/test populations chosen in Case 2, it is likely the analyst would decide the simulation is “not a bad representation” of reality. The null hypothesis is, therefore, not rejected.

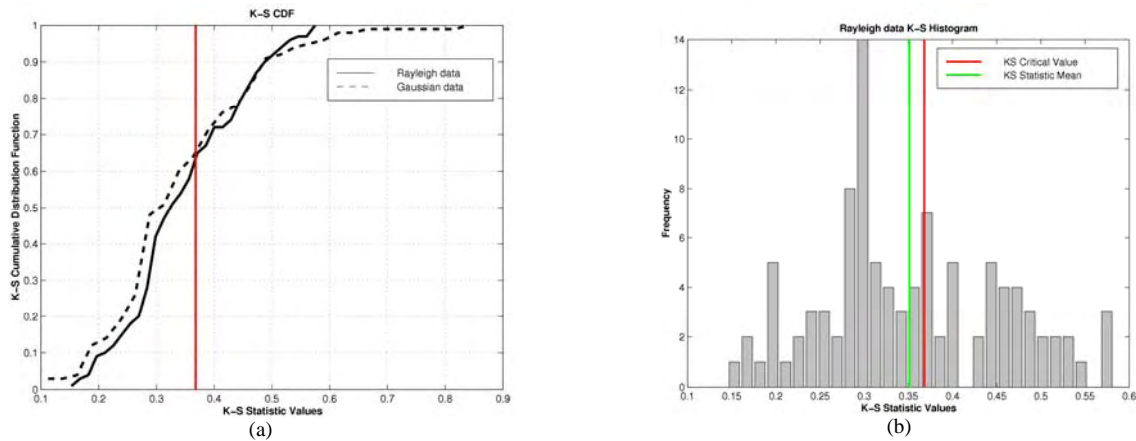


Figure 4: Distribution of K-S test results for batch set two, the moderately skewed simulation:
a) cumulative distributions for the Gaussian and Rayleigh data results
b) histogram of the Rayleigh data results

Case 3

Case 3 (see Figure 5) shows the effect of having significant differences between simulation outputs and real-world performance. Case 3 has a larger window of population pair variability than did Case 2, and it is very unlikely that any simulation/test population pairs will be statistically similar. The test distributions for this batch were randomly chosen such that their means are between -100% and 300% of the corresponding simulation means¹⁴ (i.e., skewed 200% in both directions) and their standard deviations are between 50% and 200% of the corresponding simulation standard deviations (i.e., as much as 100% greater or 50% less variation about the mean). The K-S cumulative distributions show the critical value maps near the 20-th percentile for the Gaussian data and the 10-th percentile for the Rayleigh data. This implies 80% (Gaussian) and 90% (Rayleigh) of all sample sets failed the K-S test for uniformity.

¹³ Because of the way population parameters were randomly assigned, it is *extremely unlikely* that any simulation/test population pairs are statistically identical in Cases 2 or 3.

¹⁴ Using the developed tools, the mean of X_{test}^i can be on either side of the simulation mean and take on values of opposite sign. In the context of Case 3, -100% and 300% imply μ_X varies between -1.0 and +3.0 times the simulation mean.

Figure 5 also illustrates the likelihood of a hypothesis test Type II error. Type II error, in this application, is committed by declaring a poor simulation to be a faithful predictor of the real-world process population. Keep in mind that “good” (or faithful) and “poor” are subjective labels applied to the differences in statistical populations (a rather strict test for validation purposes). The likelihood of a Type II error is about 10-20% for this case, because 10-20% of the samples *pass* the K-S test for uniformity when theory would suggest the underlying populations should yield non-uniform cumulative probabilities. Also keep in mind that, in a real validation problem, only the Type I error can be controlled in this technique (by changing α). Type II error probability (β) is essentially fixed once a sample size (I) and confidence level (α) have been chosen.

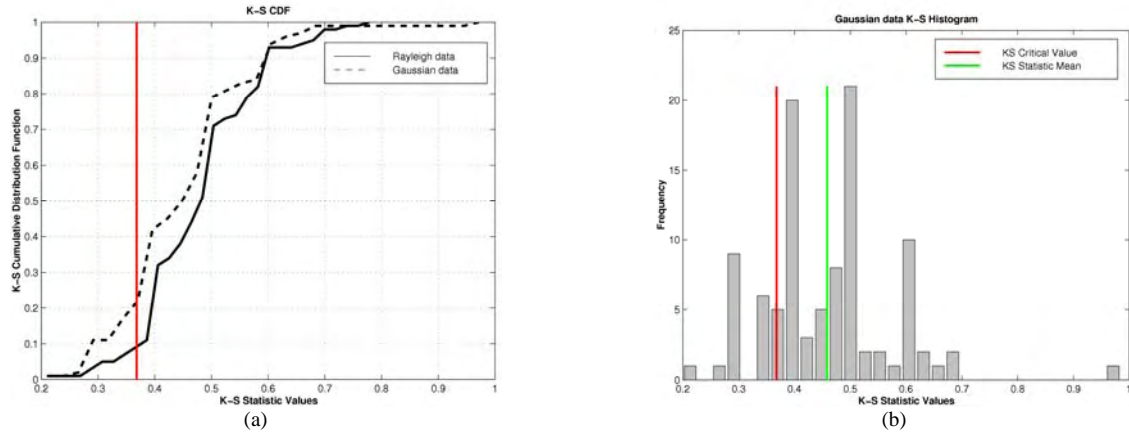


Figure 5: Distribution of K-S test results for batch set three, the significantly skewed simulation:
a) cumulative distributions for the Gaussian and Rayleigh data results
b) histogram of the Gaussian data results

This third case demonstrates that a simulation with significant variation from real-world performance will generally fail the K-S test for uniformity. If α were increased, thereby lowering the K-S critical value, even fewer samples would pass the uniformity test. This would lower the likelihood of declaring a bad simulation good, but it increases the likelihood of declaring a good simulation bad. The analyst must choose α based on these risks and based on knowledge that smaller α 's will make small discrepancies between simulation and real-world more difficult to detect.

Case 4

This case is very similar to Case 1 in the sense that the simulation and test populations for all ten pairs are statistically identical. As previously stated, the pairs are dissimilar within the batch (i.e., each population pair is unique). As shown in Figure 6, removing the added source of uncertainty from Case 1 does not change the outcome of the K-S test results. The K-S critical value maps into the 95-th percentile of the K-S cdf, implying 95% of all batches passed the test for uniformity at the chosen alpha.

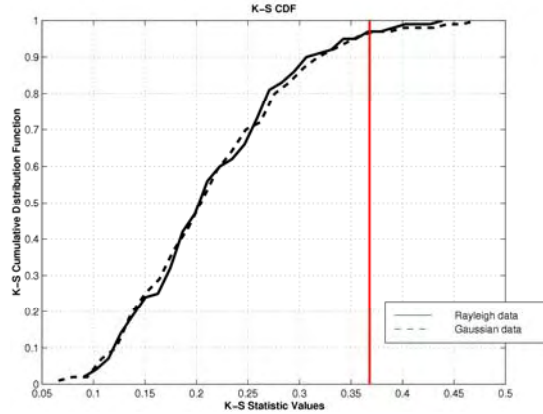


Figure 6: Results for Case 4, an “ideal” simulation. This case is similar to Case 1, except the i -th population pair is statistically identical across all 100 batches (see text for further explanation).

Cases 5 and 6 and “Meta-Analysis”

Cases 5 and 6 evaluate the effects of a mixed population set. A mixed set is, perhaps, a good way to describe a scenario in which the simulation is a good statistical predictor in certain regimes and a not-so-good predictor in other regimes. In Case 5, three of the simulation/test population pairs include simulation populations that are not statistically identical to the corresponding test population. The remaining seven pairs are statistically ideal in the sense that simulation and test populations are identical. In Case 6, six pairs are statistically different sets and the remaining four pairs are statistically ideal. Only the Gaussian data is shown for these cases. The K-S cdfs appear in Figure 7, and the cumulative probabilities appear in Figure 8.

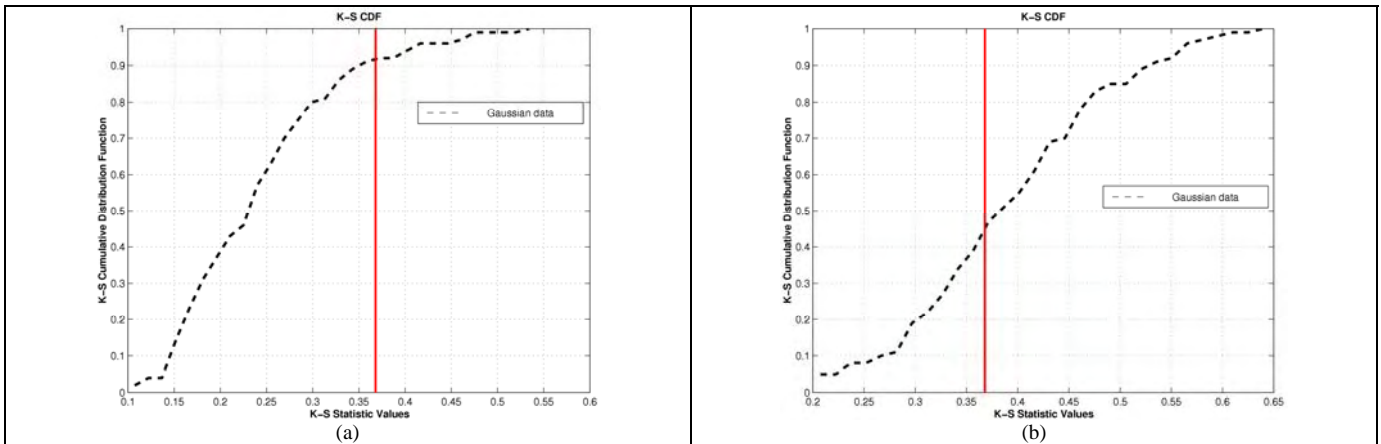


Figure 7: K-S test statistic cdfs for Cases 5 and 6. Case 5 (a) included three pairs of statistically different sim/test populations and seven pairs of statistically identical sim/test populations. Case 6 (b) included six different pairs and four identical pairs (see text for further explanation).

Note in Figure 8(a) the K-S test statistic mapped to slightly above 90% for Case 5, so 90% of all batches passed the K-S test for uniformity even though three pairs included non-ideal simulation populations. Conversely, in Figure 8(b), only about 45% of the batches passed the K-S test. The implication here is that a test for uniformity is flexible enough for situations in which a simulation works well in some regimes and not as well in others. Cases 5 and 6 also illustrate the value of so-called “meta-analysis”. Fries used this term to describe the multi-faceted process by which the analyst would decide

whether a simulation is good enough for its intended purpose based partially on the results of the uniformity test.

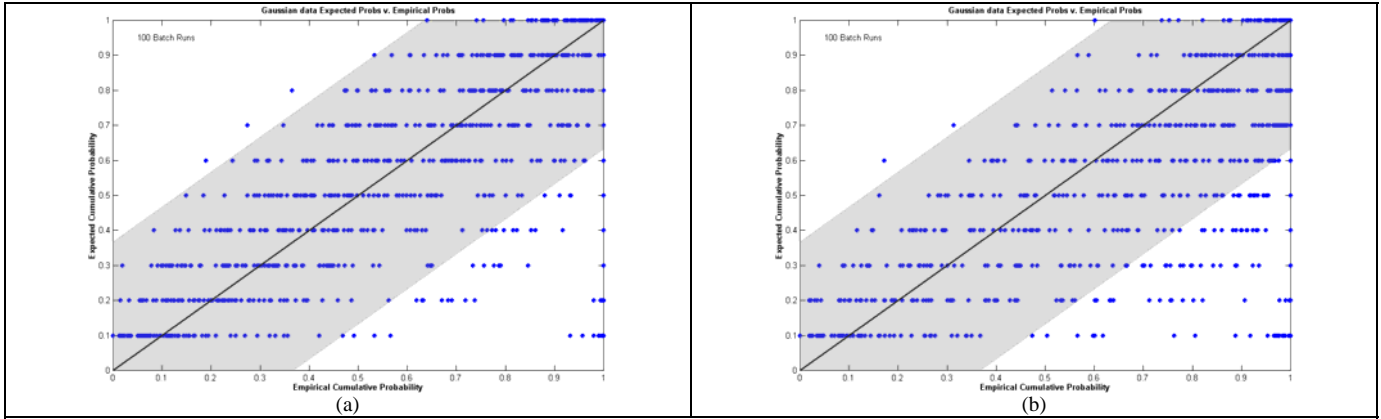


Figure 8: Cumulative probabilities for Cases 5, shown in (a), and 6, shown in (b). As expected, Case 6 has more probabilities falling outside the K-S test region of acceptance than does Case 5.

Cases 5 and 6 are only snapshots of what can occur in a real-world validation problem, because these problems have many variables. For example, the cases in this paper have $I=10$ cumulative probabilities produced from 10 sim/test population pairs. Each pair of Gaussian populations would have two means (simulation and test) and two standard deviations, all of which are independent. Therefore each batch of 10 probabilities incurs $(2 \times 2 \times 10) = 40$ degrees-of-freedom, all of which have some impact on the nature of the K-S test results. These degrees-of-freedom are confounded further by the randomness of real-world sampled test data. For the Rayleigh data case, the degrees-of-freedom would grow to 40^2 since each Rayleigh random variable is created from two, independent Gaussian distributions. In this author’s opinion, the meta-analysis of cumulative probabilities is a daunting exercise; a valuable exercise so long as the analyst remains cognizant of the “big picture”.

4.3. Single Batch Analysis Techniques

The preceding analysis was primarily designed to test the validity of using a non-parametric goodness-of-fit test to determine cumulative probability uniformity. Ultimately though, the analyst would like to apply this technique as a simulation validation tool. The obvious difference between the analysis of section 4.2 and a real-world validation problem is the real-world problem results in a single batch of cumulative probabilities (vice multiple batches in 4.2). As is true with any sample-based statistical technique, type I and II errors will go unrecognized in the single batch case. The probability of these errors is controllable to some extent by understanding the effects of sample size (the number of cumulative probabilities in the batch), the statistical power of the test, and the chosen alpha value.

Fortunately, the single-batch case provides more than just a test statistic on which to base a decision. Recall that a simulation which faithfully predicts the real-world process or system will produce uniformly distributed cumulative probabilities according to the technique described in section 3.0. The simplicity of a uniform distribution makes for straightforward “what if” analysis. Such analysis can indicate how close a non-uniform batch (as declared by a goodness-of-fit test) was to passing or how close a uniform batch was to failing. This is one function of the meta-analysis discussed in the previous section.

Three examples from the cases of section 4.2 were chosen for single-batch analysis. In all three batches, the set of cumulative probabilities failed the K-S goodness-of-fit tests. The K-S test lends itself to visualization, and the three batches are shown in Figure 9(a), (b), and (c). The gray-shaded regions indicate the region of acceptance for the K-S test at $\alpha=0.1$ and sample size of $I=10$. Remember that the y-axis expected cumulative probabilities are intrinsic to the uniformity assumption and computed for the i -th empirical cumulative probability as i/I . The x-axis empirical cumulative probabilities are derived from the pairing of simulation and real-world test data. During “what if” analysis, the analyst can evaluate the effect of moving the empirical probabilities horizontally, but not vertically. This horizontal tweaking mimics the effect of an altered test event result paired with an unaltered simulation distribution¹⁵.

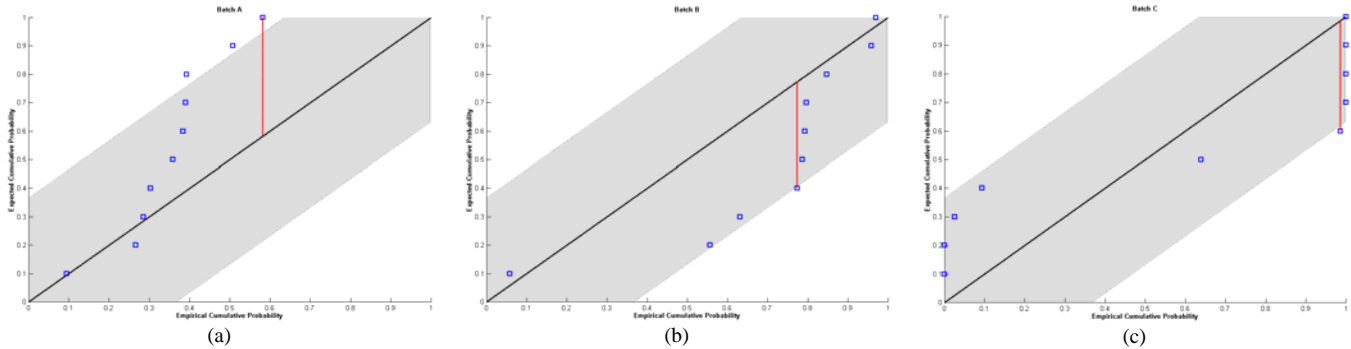


Figure 9: Single batches of 10 cumulative probabilities taken from the multiple batches analyzed in section 4.2. The gray-shaded region is the region of acceptance for the K-S test at $\alpha=0.1$. If any one of the cumulative probabilities in the batch lie outside this region, the K-S test will fail and the batch is declared non-uniform. Note that batches A and B were taken from Case 1, the ideal simulation, while batch C was taken from Case 3, the highly skewed simulation.

The batches in Figure 9(a) and (b) were both taken from Case 1, the “ideal” simulation. Recall that the simulation populations for Case 1 were statistically identical to the corresponding real-world populations. Therefore, batches A and B represent an unfortunate situation for the analyst: the pairing of the real-world data to a statistically ideal simulation resulted in a failed uniformity test (a type I error). The analyst, however, would have no idea that a type I error occurred and might conclude the simulation was a poor representation of the real process. Looking at batch A in Figure 9(a), three cumulative probabilities fell outside the region of acceptance. All three of these cumulative probabilities would need to be inside the shaded region for the batch to be declared uniform. Tweaking three probabilities in the same set might be unpalatable for most analysts, so this type I error would prevail and the analyst would declare the simulation to be suspect. On the other hand, batch B in Figure 9(b) was very close to passing. A slight decrease in the value of the fourth cumulative probability would allow the batch to pass the K-S test. In this situation, the analyst might conclude that the batch essentially passed, because a slightly different outcome in a single test event sways the results considerably.

Figure 9(c) shows a batch from Case 3, the highly skewed simulation. Most of the empirical cumulative probabilities are bunched near 0.0 and 1.0, and this implies the test results were very far from the simulation mean in every simulation/test pair. This is expected given the simulation of Case 3 was designed to be a particularly poor representation of the real-world process. At the same time, this batch illustrates a shortcoming of the K-S test with regards to statistical power. For the chosen alpha and sample size of 10, the K-S region of acceptance is relatively large, making a type II error relatively

¹⁵ Or similarly, an altered simulation distribution paired to an unaltered test event result.

likely (low statistical power, or high beta). If the analyst concluded that a slight decrease in the sixth cumulative probability were warranted, the batch would pass the K-S test. The null hypothesis that the simulation and test populations are statistically similar would remain true, according to this “what if” K-S result. Hopefully, the analyst would notice the bunching of probabilities at the tails and conclude that a type II error is very likely. Also note that increasing the alpha value (which increases the likelihood of a type I error) would decrease the likelihood of a type II error, and the region of acceptance would become narrower. Looking at Figure 9(c), as alpha increases, the shaded region encompasses fewer of the bunched probabilities and the K-S test fails.

4.4. Possible Pitfalls in Single Batch Analysis

In the previous section, the “what if” analysis was restricted to shifting empirical probabilities in the x -axis or changing the alpha value. It is very tempting to explore more elaborate “what if” techniques, and the obvious first step might be altering the number of probabilities in the batch. For instance, looking at batch B in Figure 9(b), one might be tempted to remove a probability and pretend a particular test event never occurred. One might be tempted to add a probability or two to batch C (Figure 9(c)) and make the claim that an extra test event or two might vindicate this simulation. Such proposals are not altogether different from analyzing the effects of statistical outliers in conventional parametric statistics.

However with the technique presented here, artificially altering the probabilities is somewhat treacherous. Remember that adding or removing probabilities from a batch changes both the K-S test region of acceptance as well as the y -axis expected probabilities. This author found that it was very easy to “game the system” by carefully choosing probabilities for removal to maximize the chances that the reduced set would pass the uniformity test. In fact, removing probabilities that appear to be outliers ignores a fundamental trait of this technique: a uniform distribution of probabilities implies that all values from 0.0 to 1.0 are equally likely to occur. In other words, so-called outliers do not exist by definition¹⁶. Removing a probability, akin to turning a blind eye to a test event, only reduces the ability to determine uniformity. As the theory in Section 3 shows, there are no “bad” cumulative probabilities, just bad sets of probabilities. An isolated cumulative probability is completely neutral – it says nothing about the simulation or test event that produced it. Likewise, adding a fictitious probability to a set merely obscures the true answer. If one were to add enough probabilities in the correct places, any set could be forced to appear uniform. The bottom line is that these “what if” techniques obscure any inferences made from the original data in hopes of achieving a desired outcome from the uniformity test.

One of the most commonly encountered questions will probably be, “Why wouldn’t my test event always fall near the simulation mean and produce cumulative probabilities near 0.5?” Figure 10 shows a normal cdf with $\mu = 0$, $\sigma = 1$. Recall that random samples from a normal distribution should fall within one standard deviation of the mean 68% of the time. The shaded box in Figure 10 spans ± 1 standard deviation in x , and this corresponds to cumulative probabilities between 0.16 and 0.84. Therefore, under normal distribution assumptions, a test event/simulation pair should produce a cumulative probability in this window 68% of the time and outside this window 32% of the time. Again, since cumulative probabilities are hypothesized to be uniform for an ideal simulation, any given

¹⁶ For distributions other than uniform, probability of occurrence usually decreases as the value moves farther from the mean. An “outlier” is a value considered to have an unusually low probability of occurrence. For uniform distributions, though, probability of occurrence is the same over the entire interval.

probability is equally likely and any equally sized region of cumulative probabilities is equally likely. Also, cumulative probabilities are sensitive to both the mean and standard deviation of the simulation distribution. Even when test events fall fairly close to the mean, sensitivity to standard deviation will cause the cumulative probability to be vary. Large numbers of cumulative probabilities near 0.5 might imply that simulation results have a much higher variance than does the real-world process, leading to a nearly flat pdf around the mean.

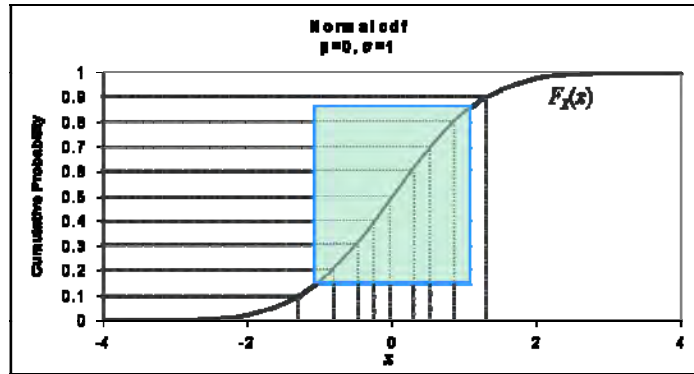


Figure 10: Using a normal cdf as an example, the shaded box spans ± 1 standard deviations from the mean in x and cumulative probability 0.16 to 0.84 in y . Assuming normally distributed data, cumulative probabilities derived from a test/simulation pair should be inside this box 68% of the time. Cumulative probabilities should be outside the box (<0.16 or >0.84) 32% of the time.

5.0 Conclusions

The analysis of Section 4.0 demonstrates the cumulative probabilities approach described in Section 2.0 is useful for determining whether simulation outputs are statistically similar to real-world test results. The cumulative probability technique aggregates independent experiments to provide an overall assessment of simulation output validity. Approximately 95% of the Gaussian and Rayleigh data example cases involving “ideal” simulations passed the uniformity test, and this passage rate was consistent with the chosen confidence level $\alpha=0.1$. When the simulation populations were skewed from the test event populations (either in mean or variance), the cumulative probabilities appeared substantially less uniform. Case 3 three analyzed the effect of a particularly poor simulation, and the data yielded very non-uniform cumulative probabilities. Approximately 80% (Gaussian) and 90% (Rayleigh) of these batch cases failed the K-S test for uniformity.

For the purposes of this paper, the important trend to note is that fewer batches pass the K-S test when simulation population parameters are permitted to deviate from the test populations. Cases 1 through 3 confound two sources of variability (random sampling of probabilities in addition to random assignment of population parameters), yet the uniformity tests answer the question, “Were the simulation and test populations statistically similar?” Cases 4 through 6 eliminate the random population assignment across batches, and the technique remains equally valid. Cases 5 and 6 showed that the technique is sufficiently flexible to handle a simulation that performs well in some regimes and not as well in others.

Although beyond the scope of this paper, the next logical question is, “How much simulation/test population variability will this technique tolerate before it indicates a simulation is not a good statistical

predictor of the real-world process?” That answer would require a carefully structured sensitivity analysis with a number of fixed assumptions regarding simulation and test population characteristics.

Finally, the single batch analysis most closely illustrates the problems inherent to an actual validation effort. As with any statistical analysis from sampled data, any answer gleaned from only one set of probabilities entails a degree of risk (Type I error), and the analyst must assess this risk. Fortunately, if real-world experiments are replicated, the certainty of this technique improves, because each replication produces another cumulative probability. With a careful regard for Type I error and a realistic view of meta-analysis approaches, the cumulative probability technique provides the analyst with a quantitative tool to supplement the often subjective techniques used in simulation validation.

6.0 References

1. Conover, W.J., *Practical Nonparametric Statistics*, John Wiley & Sons, Inc., 1999.
2. Fries, Arthur, *Another “New” Approach For “Validating” Simulation Models*, 6th US Army Conference on Applied Statistics, 18-20 October, 2000.
3. Leon-Garcia, Alberto, *Probability and Random Processes for Electrical Engineering*, 2nd ed, Addison Wesley Longman, 1994.