Proceedings of the 1995 Winter Simulation Conference
ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman

USING NONPARAMETRIC STATISTICS IN SIMULATION ANALYSIS: A REVIEW

Enver Yücesan

INSEAD
European Institute of Business Administration
Boulevard de Constance
77305 Fontainebleau Cedex, FRANCE

ABSTRACT

Techniques that make the minimum of assumptions about the underlying characteristics of the simulation output series are particularly useful for simulation analysis. This tutorial discusses robust non-parametric techniques with immediate applicability to such crucial steps in simulation analysis as sampling, experimental design, and output analysis. Algorithms are provided for various tasks. This is a revised version of the paper that appeared in the Proceedings of the 1994 Winter Simulation Conference.

1 MOTIVATION

Even though the construction and execution of simulation models have been largely facilitated by the impressive advances in simulation packages, correct and effective analysis of the results requires considerable care. Since a simulation is a statistical sampling experiment, appropriate statistical methods are essential to avoid erroneous conclusions, ultimately leading to poor decisions. In particular, special attention must be paid to the details of sampling, experimental design, and data analysis.

Considerable attention has been devoted to these problems yielding rigorous procedures for output analysis, mostly customized from classical statistical techniques. The latter are based on stringent assumptions about the properties of the output process. Unfortunately, a large number of simulation practitioners using high-level simulation languages have little knowledge or interest in verifying whether the underlying assumptions of the given technique are satisfied by the simulation output. Moreover, these packages do not always provide appropriate utilities for correct output analysis. Given this environment, methods that make the minimum of assumptions about the stochastic properties of the output sequence are particularly useful.

The objective of this tutorial is to revisit an old nonparametric technique, which addresses simulation analysis issues in a hypothesis testing framework. The randomization (permutation) tests were first proposed by Fisher (1925, 1935); the computational burden, however, had rendered such tests infeasible for practitioners. The approach is extremely flexible in that it enables the use of a wide variety of test statistics. Moreover, no assumptions are needed concerning the distribution of the sample of observations. It is, on the other hand, computationally intensive; however, the required computing power does not exceed the capabilities of a personal computer. A comprehensive treatment of permutation tests can be found in Good (1994).

This tutorial is organized as follows: Section 2 introduces the concept of randomization tests. Various applications in simulation analysis are presented in Section 3. Section 4 offers some concluding comments.

2 PRELIMINARIES

The problem of statistical inference within the Neyman-Pearson framework can be described in the following manner: some null hypothesis, H_0, concerning the nature of the probability law governing N observations, $x_1,x_2,...,x_N$, is to be tested. Some alternative hypothesis, H_1, concerning the nature of this law is also specified. To conduct the test, a region, w, is selected in the sample space, Ω, which is such that, if the sample point falls into ω, H_0 is rejected. This is the so-called *critical region.*

Adopting the notation of Box and Andersen (1955), let $\mathbf{X}_r$ be a vector of observations $(x_1,x_2,...,x_N)$, and let $p_0(\mathbf{X}_r)$ and $p_1(\mathbf{X}_r)$ represent the probability laws under H_0 and H_1, respectively. Then w is selected such that:

(1) $\int_\omega p_0(\mathbf{X}_r)\, d\mathbf{X}_r = \alpha$,

(2) $\int_\omega p_1(\mathbf{X}_r)\, d\mathbf{X}_r$ is maximized

The problem is to devise exact tests of significance when the form of the underlying distribution, $p_0(\mathbf{X}_r)$, is not known. The traditional practice is to build implicitly

more structure into the null hypothesis. For instance, the conventional t-test is a test of the hypothesis that two variables have common means; moreover, they are independently and normally distributed with common variances. The only part of this hypothesis of real interest is that the variables have common means; the condition that the variables are normally distributed with constant means and constant variances is added as a matter of convenience, simply to be able to specify $p_0(\mathbf{X}_r)$. As a consequence, if the null hypothesis is rejected, it may be either due to the dependence of variables or due to the fact that the variables are not distributed according to the normal law. The purpose of the randomization procedure is to construct valid tests without having to add extraneous, but analytically convenient, conditions to the null hypothesis.

Within this procedure, $\mathbf{X}_r$ is regarded as a member of the set $\mathbf{X}$, which contains N! samples $\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_{N!}$ of observations $(x_1,x_2,...,x_N)$ in all possible permutations. Then the probability that $\mathbf{X}_r$ is observed given that it belongs to the set $\mathbf{X}$ is:

$$p_0(X_r|X) = \frac{p_0(X_r)}{\sum_{j=1}^{N!} p_0(X_j)} = \frac{p_0(X_r)}{p_0(X)}$$

Thus, $p_0(\mathbf{X}_r) = p_0(\mathbf{X}_r|\mathbf{X})p_0(\mathbf{X})$, and $\int_\omega p_0(\mathbf{X}_r|\mathbf{X}) p_0(\mathbf{X})\, d\mathbf{X}_r = \alpha$.

The last equality is satisfied only if $p_0(\mathbf{X}_r)$ is some symmetric function of the observations $x_1,x_2,...,x_N$. This is possible if the observations are independent and identically distributed; but it is not possible, if, for instance, they are not identically distributed or are serially correlated. If $p_0(\mathbf{X}_r)$ is a symmetric function, we have $p_0(\mathbf{X}_r) = p_0(\mathbf{X}_q)$ for r,q=1,2,....,N! and $p_0(\mathbf{X}_r|\mathbf{X}) = 1/N!$.

Hence, at significance level α, an exact test of hypothesis can be constructed by arranging that K out of the N! permutations are included in ω, and the remaining N!-K permutations are included in $\Omega-\omega$, where $K = \lfloor \alpha N! \rfloor$. Then we have

$$\int_\Omega p_0(x) \sum_{r=1}^{K} p_0(x_r|x)dx = \frac{K}{N!}\int_\Omega p_0(x)dx$$

where Ω represents the sample space. While no restrictive assumption is needed concerning the null distribution, $p_0(\cdot)$ a class of probability laws for the alternative hypothesis, $p_1(\cdot)$, has to be specified in order to obtain a most-powerful test. Trade-offs between the validity and power of randomization tests are discussed in Box and Andersen (1955).

It is also desirable to derive the permutation distribution and make it practical to carry out tests of significance. Except for very small samples, the calculations to determine whether the observed value of the sample point belongs to the critical region are extremely tedious. In fact, the computation of the permutation distribution or p-value results in exponential time bounds as one enumerates all possible permutations of the data (Spino and Pagano 1991). In such cases, an alternative approach is to use an approximation to the discrete distribution of the test statistic by means of some familiar continuous distribution for which tables are available (Scheffé 1959). Pagano and Tritchler (1983) and Spino and Pagano (1991) introduced polynomial-time algorithms to compute the permutation distribution in a matched-pairs design. Other such algorithms have been proposed by Edgington (1980), and Berry and Mielke (1985). However, even with polynomial-time bounds, the time complexity of these algorithms may be unacceptable when the sample sizes are large.

Another alternative is then to approximate the permutation distribution to any desired level of precision by sampling from the complete reference set, $\mathbf{X}$, of permutations in order to reduce the computational difficulties and approximate p-values for larger samples (Noreen 1989, Berry and Mielke 1985, and Edgington 1980). In other words, the distribution of the test statistic under the null hypothesis is approximated by shuffling the data and recomputing the test statistic. Each shuffle generates one permutation of the variables. One thousand shuffles can then be viewed as a sample of size 1000 from the population of all possible permutations.

The significance of the actual test statistic for the original unshuffled data is then assessed relative to this empirically generated distribution. The null hypothesis is rejected if the actual value of the test statistic for the original data is unusually large. Noreen (1989) calls this procedure an *approximate randomization test.*

Hoeffding (1952) demonstrates that randomization tests are asymptotically as powerful as analogous conventional parametric tests when the assumptions underlying the parametric test are true. Validity proofs are also provided by Foutz (1980), who also discusses the power of randomization tests.

A similar methodology, first suggested by Efron (1979) is *bootstrapping,* which proceeds as if the sample *is* the population for purposes of estimating the sampling distribution of the test statistic. That is, artificial samples are drawn *with replacement* from the sample itself.

The idea may first seem a bit odd. However, given a sample of observations $\{x_1,x_2,...,x_m\}$, the maximum likelihood nonparametric estimator of the population distribution is the one that assigns a probability mass of 1/m on each of the observations (Efron and Tibshirani 1984). The implication is that, when the sample

contains all of the available information about the population, it is almost natural to proceed as if the sample *is* the population.

These resampling techniques yield the following kinds of nonparametric tests:

Permutation Test:
1. Choose a test statistic, T(X).
2. Compute T for the original set of observations.
3. Obtain the permutation distribution of T by repeatedly rearranging the observations.
4. Obtain the upper a-percentage point of the permutation distribution and accept (or reject) the null hypothesis according to whether T for the original observations is smaller (or larger) than this value.

Rank Test:
1. Choose a test statistic, T.
2. Replace the original observations by their ranks. Compute T for the original set of ranks.
3. Obtain the permutation distribution of T by repeatedly rearranging the ranks and recomputing the test statistic.
4. Accept or reject the hypothesis in accordance with the upper α-percentage point of this permutation distribution.

Bootstrap Test:
1. Choose a test statistic, T(X).
2. Compute T for the original set of observations.
3. Obtain the bootstrap distribution of T by repeatedly resampling from the observations with replacement.
4. Obtain the upper a-percentage point of the bootstrap distribution and accept (or reject) the null hypothesis according to whether T for the original observations is smaller (or larger) than this value.

3 APPLICATIONS

3.1 Testing Random Number Generators

Random variate generation constitutes an active area of research. Following L'Ecuyer (1990), we define a random number generator as a structure $G=(S,\mu,f,U,g)$ where S is a finite set of states, μ is a probability distribution on S, called the initial distribution, U is a finite set of output symbols, $f:S\rightarrow S$ is a transition function, and $g:S\rightarrow U$ is the output function. A generator operates as follows:
1. Select the initial state $s_0 \in S$ according to μ. Let $u_0 \leftarrow g(s_0)$.
2. For, i=1,2,..., let $s_i \leftarrow f(s_{i-1})$ and $u_i \leftarrow g(s_i)$.
The sequence of observations $\{u_0,u_1,...\}$ is the output of the generator. The initial state s_0 is called the *seed*. The output sequence should look as if the u_i's were the values of IID random variables, uniformly distributed over U. An ideal generator would be such that, using reasonable computing resources and time, it is impossible to distinguish between the generator's output and a sequence of truly IID uniform variates over U.

In practice, however, this is either supported by a theoretical basis or verified by statistical analysis. Theoretical tests use the numerical parameters of a generator to assess its global characteristics. Statistical tests, on the other hand, are based on the actual output of the generator to examine how closely they resemble IID uniform random variates.

Approximate randomization tests can be used to this end. Such tests are rather easy to design: any function of a finite set of IID uniform random variables can be used as a test statistic to define a test of hypothesis. To gain power, the test can be repeated N times, and the empirical distribution of the values of the test statistic can be compared to its theoretical distribution, for instance. A specific algorithm can be devised as follows: (Yücesan 1992)
1. Generate N sets of random numbers, each set containing k numbers.
2. Sort each set. This produces a permutation of indices in each set. There are k! possible permutations for each set. If the random numbers are independent, then all permutations are equally likely.
3. Count the number of each permutation occurring for these N sets and apply a chi-square test.

3.2 Resampling of Empirical Distributions

Stochastic discrete-event simulations are driven by parametric or empirical distributions, typically fitted to the observed (finite) samples. Any finite sample yields a distribution estimate with some error. The nature of the error in the empirical distribution's approximation to the true distribution function is well understood. Yet, this error is typically ignored in the analysis of simulation output.

A correct statistical analysis should include an assessment of the errors that result from using finite-sample estimates for probability distributions used in the simulation. Ideally, several empirical samples should be taken and a mixed-effects analysis of variance (ANOVA) conducted to estimate the magnitude and the significance of the empirical distribution random effect. Such an approach may however be costly.

A bootstrap resampling method provides an inexpensive resampling approach to assess the true estimation error (Barton and Schruben 1993). In a Monte Carlo simulation experiment, the bootstrap resampling method can be implemented as follows:
1. Sample, with replacement, n values from $\{x_{(1)}, x_{(2)},...,x_{(n)}\}$, the observed data ordered from the smallest to the largest value. Call these values $v_{(1)}, v_{(2)},...,v_{(n)}$.
2. Compute an empirical distribution based on the v sample.

3. Repeat this process t times and run the simulation with each of the resulting distributions to collect t bootstrap samples of the simulation process. Perform a mixed-effect ANOVA.

The example of an M/M/1 queue demonstrates that the variation in simulation output due to empirical sampling can be significant even for relatively large empirical distributions. Resampling schemes provide an improvement over the standard approach.

3.3 Tests for Initialization Bias

The objective here is to detect any significant change in the mean of the output process. The null hypothesis is that there is no initialization bias in the output mean. To test this hypothesis, a preliminary step is taken where the output series of N observations is partitioned into b non-overlapping batches, each of size m (N = bm). This is done in order to control the serial correlation in simulation output. Working with the batched process, however, is conceptually identical to working with the original process.

The test is applied to b batch means, rather than the original output sequence. The batch means are further partitioned into two groups. Initially, the first group contains only the first batch mean, while the second group contains the remaining b-1 batch means.

The test statistic could be the *absolute value of the difference between the grand means of the two groups.* The randomization test is applied to the batch means.

The procedure is repeated by redefining the groups. In the second iteration, the first group contains the first two batch means while the second group contains the remaining b-2 batch means. The randomization test is applied to the new groups. In the third iteration, the first group contains the first three batch means while the second group contains the remaining b-3 batch means, and so on. The procedure can be continued until the first group contains the first b-1 batch means and the second group contains the last batch mean.

The truncation point is then determined by the earliest iteration where the null hypothesis is not rejected. Note that this may be the very first iteration (implying no significant initialization bias in the output series) or may never happen (implying that the system has not yet settled into a steady state). This assumes that the significance level of the test is a more or less monotonically increasing function of the number of batches in the first group. .This is true when there *is* initialization bias in the sequence; otherwise, it is basically a random variable *uniformly* distributed over the interval [0,1].

Example: M/M/1 Queue

This is a single-server queue with Poisson arrivals at rate λ and exponential service times at rate μ. The system is analyzed at a traffic intensity, ρ, of 0.7 to estimate the average customer delay in queue. Table 1 shows two scenarios: one where the system starts empty and idle ($W_0 = 0$) and the other where the initial delay is sampled from the (known) steady-state distribution ($W_0 \sim SS$). The total number of batches is 30 with a batch size of 500. The asterisk (*) denotes the truncation point (Yücesan 1993).

$W_0 = 0$		$W_0 \sim SS$	
Batch in G1	Sign. Level	Batch in G1	Sign. Level
1 (*)	0.061	1	0.515
2	0.570	2	0.765
3	0.890	3	0.880
4	0.960	4	0.635
5	0.320	5	0.905
6	0.260	6	0.135
7	0.260	7	0.425
8	0.570	8	0.150
9	0.315	9	0.270
10	0.405	10	0.270

Table 1: Significance Levels

3.4 Comparing Alternative System Configurations

The real utility of simulation lies in comparing different alternatives that might represent competing system designs. Conventional statistical techniques are not directly applicable to the analysis of simulation output data in the of some performance measure. Such a situation arises when a new policy is proposed to replace the existing one. For i=1,2, let X_{ij} be the output of the j^{th} independent simulation run with the i^{th} system (or alternative policy), and let $\mu_i = E[X_{ij}]$ be the expected response of interest under this system. It is desired to assess the significance of the difference between the two expected responses, namely $\delta=\mu_1-\mu_2$. Law and Kelton (1991, §10.2) describe a parametric approach for comparing the two systems. The traditional one-way model is $X_{ij} = \mu+\alpha_i+\varepsilon_{ij}$ for i=1,2, and j=1,2,...,n_i where μ is the common mean, α_i is the so-called fixed treatment (i.e., policy) effect, the ε_{ij}'s are IID variables normally distributed with mean zero and unknown variance σ_e^2 representing the random error for the j^{th} replication under policy i, and n_i is the total number of replications conducted under policy i. Sometimes, it is required that $\sum_i\alpha_i = 0$. The underlying assumption of the paired-t test (or confidence interval) is that the differences between pairs of observations, $Z_j = X_{1j} - X_{2j}$, are normally distributed. It is then necessary that $n_1 = n_2 = n$. It is also possible to apply the classical

two-sample t-test without pairing up observations; to obtain a valid test, however, it is necessary to assume that $Var(X_{1j}) = Var(X_{2j})$. This assumption is harder to justify as the variances usually depend upon the specific alternatives under consideration. Note that heuristic approaches exist to handle the cases where variances are unequal (Welch 1938). Randomization tests, on the other hand, eliminate the need to make such an assumption altogether (Yücesan 1994).

The randomization procedure used here assumes that, under the null hypothesis, H_0: $\delta = \mu_1 - \mu_2 = 0$, the distribution of the $m = n_1 + n_2$ observations remains invariant under all permutations. This assumption is satisfied if the data, i.e., the simulation output, are IID. This, in turn, is easily achieved by running *independent replications* of the model.

The procedure starts with the selection of an appropriate test statistic (e.g., the *absolute value of the difference between the mean responses* under the two policies). The value of the test statistic is computed for the original data. Let Group i include the responses of simulation runs under Policy i, i=1,2. Let $d = |\bar{X}_1 - \bar{X}_2|$ be the test statistic, where $\bar{X}_i = (1/n_i)\Sigma_j X_{ij}$ for i=1, 2. The significance of the observed value of the test statistic is then assessed through a permutation test. More specifically, the data set is permuted and the first n_1 data points are arbitrarily included in Group 1 and the remaining $n_2 = m - n_1$ data points are included in Group 2. The test statistic, d, is then recomputed for the newly formed groups. The objective is to determine, through a large number of permutations, how unusual the original value of the test statistic is with respect to the permutation distribution.

Example: (s,S) Inventory Model

The example is an (s,S) inventory system from Law and Kelton (1991, §1.5). An (s,S) inventory system is one in which the inventory position of a single item is reviewed periodically. Different (s,S) combinations correspond to different "policies" or "system configurations." The total operating cost for this inventory system includes ordering, holding, and shortage costs. Suppose that it is desired to compare the current policy proposed policy of (20,80) in order to improve the average total operating cost per month.

The system is simulated for 120 months under each policy. The results of the five independent replications for each policy, X_{ij}, i=1,2; j=1,...,5, are listed in Table 2.

Law and Kelton (1991) construct confidence intervals on the difference of the average total operating costs per month under the two policies. They conclude that the (20,80) policy results in lower costs at the 90% confidence level.

Run	(20,40)	(20,80)
1	126.97	118.21
2	124.31	120.22
3	126.68	122.45
4	122.66	122.68
5	127.23	119.40

Table 2: Average Operating Cost/Month

One can also cast the problem as a hypothesis test and assess the significance of the test statistic through an approximate randomization test. To this end, the null hypothesis is that *there are no differences between the two policies in terms of the average total operating costs per month.* The test statistic is defined to be d = |average_cost1 - average_cost2|. For the above data, the value of the original test statistic is given by d = 4.98. The next task is to determine whether this difference is significant. Since there are only 10 data points resulting in 252 different permutations, a complete enumeration is possible. For illustration purposes, we apply the approximate randomization test by randomly generating 199 permutations from the set of all possible permutations. The estimated significance level is computed to be 0.030, also leading to the rejection of the null hypothesis of no difference at 90% confidence level.

3.5 Threshold Bootstrap

Kim et al. (1993) discuss bootstrapping techniques applicable to autocorrelated data. They identify three such approaches in the literature. In the first approach, an ARIMA model is fit to the data and pseudo-series are created by resampling residuals and adding them to the fitted model. In the second approach, the data series is divided into adjacent blocks of fixed length and pseudo-data are created by concatenating blocks chosen by resampling without replacement. This is referred to as the *moving block bootstrap.* In the third approach, the data series are resampled by concatenating blocks whose starting point is chosen at random and whose length is geometrically distributed with some mean p. This is called the *stationary bootstrap*.

Kim et al. (1993) develop the *binary bootstrap,* which resamples from the runs of zeros and ones that comprise any binary series. By using runs rather than blocks, they avoid the problem of selecting the block size. Encouraged by the success of the approach, they generalize the method to non-binary data. The technique is as follows:

0. Generate a time series with N values.
1. Select a threshold value (eg, the sample mean).
2. Divide the series into runs that are either above or below the threshold.
3. Create a bootstrap replication by concatenating runs resampled with replacement. Truncate if the total length exceeds N.

4. Compute the desired statistic.
5. Repeat steps 3 and 4 B times.
6. Analyze the statistics as if they were coming from independent replications.

Limited experimentation with the technique reveals a promising performance.

4 CONCLUSIONS

Techniques that make the minimum of assumptions about the underlying characteristics of the output series are particularly useful for simulation analysis. This tutorial discussed robust non-parametric techniques with immediate applicability to such crucial steps in simulation as sampling, experimental design, and output analysis.

The principal advantage of randomization tests, or resampling techniques in general, is the flexibility they provide in selecting the most appropriate test statistic for the case under study. Only mild assumptions are needed concerning the distribution of the sample of observations. Moreover, the present computing technology provides ample power to widely use these computationally intensive methods.

REFERENCES

Barton, R.R. and L.W. Schruben. 1993. Uniform and Bootstrap Resampling of Empirical Distributions. In *Proceedings of the 1993 Winter Simulation Conference*, 503-508.

Berry, K.J. and P.W. Jr. Mielke. 1985. Computation of Exact and Approximate Probability Values for a Matched Pairs Permutation Test. *Communications in Statistics - Part B: Simulation and Computation,* 14: 229-248.

Box, G.E.P. and S.L. Andersen. 1955. Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumptions. *Journal of the Royal Statistical Society, Series B,* 17: 1-34 .

Edgington, E.S. 1980. *Randomization Tests.* New York: Marcel Dekker.

Efron, B. 1979. Bootstrap Methods: Another Look at Jackknife. *Annals of Statistics,* 7: 1-26.

Efron, B. and R. Tibshirani. 1984. Bootstrap Methods for Standard Errors, Confidence Intervals and Other Measures of Statistical Accuracy. *Statistical Science,* 1: 54-77.

Fisher, R.A. 1925. *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd.

Fisher, R.A. 1935. *The Design of Experiments.* Edinburgh: Oliver & Boyd.

Foutz, R.V. 1980. A Method for Constructing Exact Tests from Test Statistics that Have Unknown Null Distributions. *J. Statist. Comput. Simul.*,10: 187-193.

Good, P. 1994. *Permutation Tests.* New York: Springer-Verlag.

Hoeffding, W. 1952. The Large Sample Power of Tests Based on Permutations of Observations. *Annals of Mathematical Statistics,* 23: 169-192.

Kim, Y.B., T.R. Willemain, J. Haddock, and G.C. Runger. 1993. The Threshold Bootstrap: A New Approach to Simulation Output Analysis. In *Proceedings of the 1993 Winter Simulation Conference,* 498-502.

L'Ecuyer, P. 1990. Random Numbers for Simulation. *Communications of the ACM,* 33: 85-97.

Noreen, E. 1989. *Computer Intensive Methods for Testing Hypotheses: An Introduction.* New York: Wiley.

Pagano, M. and D. Tritchler. 1983. On Obtaining Permutation Distributions in Polynomial Time. *Journal of the American Statistical Association,* 78: 435-440.

Scheffé, H. 1959. *The Analysis of Variance.* New York: Wiley.

Spino, C. and M. Pagano. 1991. Efficient Calculation of the Permutation Distribution of Trimmed Means. *Journal of the American Statistical Association,* 86: 729-737.

Welch, B.L. 1938. The Significance of the Difference between two Means when the Population Variances are Unequal. *Biometrica,* 25: 350-362.

Yücesan, E. 1992. Evaluating Random Number Generators Using Permutation Tests. In *Proceedings of the 1992 European Simulation Multiconference,* ed. Stephenson, 94-98.

Yücesan, E. 1993. Randomization Tests for Initial Bias in Simulation Output. *Naval Research Logistics,* 40: 643-663.

Yücesan, E. 1994. Comparing Alternative System Configurations: A Nonparametric Approach. *Annals of Operations Research,* 53: 471-484.

AUTHOR BIOGRAPHY

ENVER YUCESAN is an Associate Professor of Operations Research at INSEAD, in Fontainebleau, France. He holds an undergraduate degree in Industrial Engineering from Purdue University, and an MS and a PhD in Operations Research from Cornell University.