Proceedings of the 1988 Winter Simulation Conference
M. Abrams, P. Haigh, and J. Comfort (eds.)

Tutorial: Analysis of simulation output to compare alternatives

Gordon M. Clark
Department of Industrial and Systems Engineering
The Ohio State University
Columbus, Ohio 43210
currently on sabbatical with
IBM Application Systems Division
P.O. Box 28331
Atlanta, GA 30358-0311

ABSTRACT

A fundamental use for system simulation is to predict system performance in new situations or with new system designs. In other words, the analyst uses simulation to aid in answering "what if" questions. We say that simulation is used to compare alternatives. This tutorial assumes that the simulation output is stochastic and has a distribution that is unknown to the analyst. The tutorial presents methods for making comparisons by either ranking alternatives or selecting the best alternative with respect to a single output performance measure. To increase the effectiveness of the comparisons, these methods consider the use of the variance reduction technique of common random numbers.

1.0 INTRODUCTION

This tutorial addresses the problem of analyzing a simulation which represents system performance by moving the system from state to state as simulated time elapses and calculates output performance measures as the simulation progresses. The time the system is in a state and/or the sequence of state changes may be random variables so the output performance measures are subject to random or stochastic variation if we repeat the simulation with the same inputs. The analyst uses the simulation to compare system performance in different situations or with different system designs. For example, an industrial engineer may use a simulation of a manufacturing shop to estimate the effect on lead time and on shop throughput realized by reducing the setup time for a particular operation. We define lead time as the time to produce an assembly and shop throughput as the mean number of assemblies produced per shift. The emphasis is on comparing the simulation output obtained from two or more sets of inputs. Another way of stating the output analysis problem is that the analyst wants to answer "what if" questions. We are assuming that the analyst will rarely use a simulation to estimate performance for a single set of inputs. We will call each set of inputs defined for comparison purposes an *alternative.*

Output analysis is complicated by the stochastic nature of output performance measures. In addition, this tutorial will make the following assumptions concerning the probability distribution of output performance measures:

- The form of the true underlying distribution is unknown.
- The variance can and will change with the alternative.

The above assumptions are common occurrences in system simulation. Even so, we will measure the stochastic or statistical variation in our estimates by interval estimates called confidence intervals which are random intervals that are purported to contain the true performance measure value with a stated probability. We use the term *replication* to specify repetition of a simulation with fixed inputs but different outputs due to different random numbers. Wide confidence intervals may mean that we need to add replications to reduce the statistical variation in our estimates.

1.1 Example Analysis Problem

To motivate the tutorial and show typical problems, the tutorial will illustrate the output analysis methods using the following production control problem. We have a network of machines representing several manufacturing operations.

The Production Control Manager must select one of several different priority rules for sequencing jobs to be processed by each machine. Candidate sequencing rules include shortest processing time, earliest job due date, and earliest operation due date. We will call each priority rule an alternative, and the analysis must compare the alternatives based on output performance measures. Important performance measures are job tardiness and throughput.

1.2 Output Analysis Issues

We will find that this problem illustrates many challenging output analysis issues. The selection of a sequencing rule may be based on the rule which is expected to be best during the next shift given the state of jobs at the start of the shift. For output analysis purposes, we call this a *finite-horizon* analysis. Some authors call simulations conducted for finite-horizon analyses as *terminating* simulations. On the other hand, the Production Control Manager may want to compare these sequencing rules based on their long-run or *steady-state* performance. We call analyses of steady-state performance *infinite-horizon* analyses. After we obtain some output data from a simulation producing tardiness values we will find that the data clearly come from a distribution that is far from normal. A significant issue in this tutorial will be analysis procedures we can use to construct approximate confidence intervals using statistics developed for normal distributions. Another issue is that we potentially have more than one performance measure, viz., tardiness and throughput. Decision analysis methods involving multiple measures such as presented by Keeney and Raiffa (1976) are beyond the scope of this tutorial, but it will discuss the effect on confidence intervals of considering multiple performance measures. The analyst may want to simultaneously compare all combinations of the alternatives, e.g., generate a ranking with respect to a single output measure, or the analyst may want to select the best alternative based on each output measure. An option potentially leading to much narrower confidence intervals when comparing alternatives accrues from being able to use the same random numbers for each alternative. This is an example of a variance reduction technique and is called *common random numbers.*

The tutorial will first examine finite-horizon analyses and later extend the methods examined for infinite-horizon analyses. The tutorial will discuss the above issues in the context of the finite-horizon analysis in the following order:

1. Batching to Approach Normality
2. Bounds on Simultaneous Confidence Intervals
3. Common Random Numbers
4. All Pairwise Multiple Comparisons
5. Selection of the Best

1.3 Notation

Let X_i represent the ith value observed of the random variable X and n be the total number of observations for X. Assume that all values of X are independent and identically distributed. Define the following quantities:

E(X)	expected value of X
V(X)	variance of X
Y	another random variable
$\rho(X,Y)$	correlation between X and Y
$\overline{X}$	$\sum_{i=1}^{n} X_i/n$
$s^2(X)$	$\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$, the unbiased estimator of the variance of X
$t_{d,\beta}$	quantile of the t distribution with d degrees of freedom which is exceeded with probability β
$H(\alpha)$	$t_{n-1,1-\alpha/2}s(X)/\sqrt{n}$, half width of the $(1-\alpha)\times 100$ % confidence interval
$L(\alpha)$	$\overline{X} - H$, lower $(1-\alpha)\times 100\%$ confidence boundary for E(X) when X is normally distributed
$U(\alpha)$	$\overline{X} + H$, upper confidence boundary analogous to $L(\alpha)$

In the following discussion for finite-horizon simulation, let n be the total number of replications, and γ be the number of alternatives compared. The performance measure observed on replication i for alternative j is $X_i^{(j)}$. Denote alternative j by Π_j, and the mean of $X_i^{(j)}$ by θ_j.

2.0 BATCHING TO APPROACH NORMALITY

We can operate a finite-horizon simulation so that each replication is independent and identically distributed. If b is a factor of n, the total number of replications, we can regard the n observations as m batches of size b, where n/b = m. Let $Y_s(b)$ be the average value of a performance measure for batch s, where s = 1,m. Under weak assumptions as b becomes large, the central limit theorem shows that the values of $Y_s(b)$ approach a normal distribution even when X is not normally distributed. That is, when the data is not normal, we can batch to improve the accuracy of confidence intervals calculated under the assumption of normal data. That is, we can compute $L(\alpha)$ and $U(\alpha)$ using the batch means $Y_s(b)$, for s = 1,m. However, large batches mean fewer batches which will increase the width of the confidence interval. Schmeiser (1982) shows that the effect on confidence interval width is negligible even when the data is truly normal so long as the number of batches is not smaller than 30. For the simulations discussed in this paper, we are never sure the output data is normal. The conclusion is that we should always use batches and limit the number of batches to 30 even with very large samples. In the remainder of this paper, the output performance measure $X_i^{(j)}$ represents the average of batch i for alternative j.

Kleijnen (1987) discusses the effect of non-normal data and describes methods for empirically checking for normal data. He recommends the use of a modified t statistic developed by Johnson (1978) when the data is clearly not normal.

3.0 BOUNDS ON SIMULTANEOUS CONFIDENCE INTERVALS

Consider the single alternative case with two performance measures, e.g., tardiness and throughput. Assume we calculated an average tardiness value from a set of n replications and an average throughput from another independent set of n' replications. If we estimate a 95% confidence interval for tardiness and a 95% confidence interval for throughput, then the probability that one or both of these confidence intervals do not cover the true value of their respective means is 1 - $.95^2$ or .0975. In this case when the confidence intervals are independent, we would need $\sqrt{.95} \times 100$ = 97.47% confidence intervals for each performance measure to give us 95% confidence that both intervals cover their respective means.

In actual practice, we use data from the same replications to calculate confidence intervals on both performance measures. This practice means that the confidence intervals are correlated. Kleijnen (1975, 1987) describes how one can use the Bonferroni inequality to calculate a conservative joint confidence interval when the amount of correlation is unknown. Assume that we want simultaneous confidence intervals for two unknown mean performance measure values, i.e, μ_1 and μ_2. Let $L_k(\alpha_k)$ be the lower $(1 - \alpha_k) \times 100$ % confidence boundary for μ_k , and let $U_k(\alpha_k)$ be its corresponding upper boundary. If S_k implies the statement that $L_k(\alpha_k) \leq \mu_k \leq U_k(\alpha_k)$ and $P_{1-\alpha}$ is the probability that both S_1 and S_2 are true, then by the Bonferroni inequality

$$P_{1-\alpha} \geq 1 - \alpha_1 - \alpha_2 .$$

4.0 COMMON RANDOM NUMBERS

Because we construct the simulation model, we have the ability to design and control simulation experiments to increase their effectiveness. Another tutorial by David Kelton in this conference will discuss this issue further. Some of the approaches for doing this are called variance reduction methods. Nelson (1987) reviews the variance reduction method of common random numbers as well as two other commonly used methods. Bratley, Fox, and Schrage (1987) give a comprehensive description of variance reduction in simulation.

Common random numbers is frequently used in simulation to reduce variance when one is comparing alternatives. The objective is to use the same random numbers for each alternative to obtain a large positive correlation among their performance measure values on each replication and thus on each batch. This reduces the variance of the difference between their respective performance measure values. For alternatives 1 and 2,

$$V(X_i^{(1)} - X_i^{(2)}) = V(X_i^{(1)}) + V(X_i^{(2)}) - 2\rho(X_i^{(1)}, X_i^{(2)})\sqrt{V(X_i^{(1)})}\sqrt{V(X_i^{(2)})} \quad (1)$$

For example, if we want to estimate the difference in mean waiting times for jobs entering a queue when two alternatives have unequal mean service times, we could use the same random number(s) for the kth interarrival time (as well as for all other interarrival times) for each alternative. This may increase the correlation between $X_i^{(1)}$ and $X_i^{(2)}$, where these quantities are mean waiting times. The problem of coordinating random numbers so that identical random numbers determine analogous events in each alternative is called the *synchronization* problem. See Bratley, Fox, and Schrage (1987) for suggestions on how to obtain synchronization.

The usual way to estimate $\theta_j - \theta_k$, for two different alternatives, π_j and π_k, when using common random numbers is to form pairs for each batch, i.e.,

$$D_i^{(jk)} = X_i^{(j)} - X_i^{(k)}$$

and calculate $\overline{D}^{(jk)}$ and $s^2(D_i^{(jk)})$. Then calculate a confidence interval for $\theta_j - \theta_k$ using the equations for $L(\alpha)$ and $U(\alpha)$ with the random variable X replaced by $D^{(jk)}$. Nelson (1986) has shown that calculation of the confidence interval in this way is computationally equivalent to using (1) and replacing values in (1) with their respective estimates.

5.0 ALL PAIRWISE MULTIPLE COMPARISONS

To enhance the power of the comparisons, we want to simulate γ different alternatives using synchronized random numbers. To rank the alternatives, we will in essence be making

$$\psi = \binom{\gamma}{2}$$

pairwise comparisons. For example, with four alternatives we would make six pairwise comparisons. The statistics $X_i^{(j)} - X_i^{(k)}$ and $X_i^{(j)} - X_i^{(u)}$ are correlated because of the common random numbers and the common quantity $X_i^{(j)}$.

We can estimate ψ simultaneous confidence intervals that jointly hold with probability of at least $1 - \alpha$ by using the Bonferroni inequality. This may allow us to rank the alternatives if we allow three possibilities:

1. Regard $\theta_j = \theta_k$ when the confidence interval for $\theta_j - \theta_k$ includes zero
2. Regard $\theta_j > \theta_k$ when the lower bound on the confidence interval for $\theta_j - \theta_k$ is positive
3. Permit contradictory relations, e.g., $\theta_1 = \theta_2$, $\theta_2 = \theta_3$, and $\theta_1 > \theta_3$

In practice, the third possibility can usually be resolved by adding replications to reduce the confidence interval widths.

6.0 SELECTION OF THE BEST

We assume here that more is better so that we want to identify the alternative having the largest mean. We attempt to do this by running a number of replications on each alternative and declaring the alternative with the largest average performance measure value as the "best." The challenge is to determine the number of replications which give a specified probability of actually choosing the best alternative. Call that probability P^*. When other competing alternatives have means almost as large as the best alternative, the required number of replications can be quite large. That is why we resort to the indifference zone approach. The length of the indifference zone is δ which is specified by the analyst before conducting the simulation experiments. Any alternative with a mean within δ of the truly best alternative is a satisfactory choice or a correct selection.

When we assume normally distributed outputs, unknown variances, and unequal variances, we consider two possible procedures for implementing the indifference zone approach. Dudewicz and Dalal (1975) present an exact procedure for these assumptions when the outputs for each alternative are independent. We should use common random numbers so the specified number of observations by the Dudewicz and Dalal procedure are probably in excess of the minimum requirement. Clark and Yang (1986) present a procedure that does consider the correlations induced by common random numbers, but they use the Bonferroni inequality and another bounding assumption which renders their procedure conservative. Clark and Yang's procedure is preferred when the correlations are large and the number of alternatives is not large.

7.0 INFINITE-HORIZON ANALYSES

When we are looking for a comparison under steady-state conditions, we are faced with a more complex analytical task. Dave Kelton discusses experimental design problems with this type of experimentation in his tutorial. Since a steady-state comparison should be unaffected by initial conditions, we have a problem of eliminating the effect of initial transients. Also, when simulating under steady-state conditions, the successive observations can be very correlated, e.g. waiting times in a queue.

Law and Kelton (1982) and Bratley, Fox, and Schrage (1987) review the numerous approaches to this statistical analysis problem. Sargent, Kang, and Goldsman (1987) present empirical comparisons of several of these procedures. This tutorial will discuss the nonoverlapping batch means approach to estimating confidence intervals.

8.0 CONCLUSIONS

The procedures presented and reviewed in this tutorial provide a basis for measuring the risk the analyst assumes in making incorrect conclusions when comparing alternatives by stochastic simulation. In addition, they provide guidance in determining the number of replications required to reduce that risk to an acceptable level. They are easily programmed so they can be automated and used by individuals unfamiliar with mathematical statistics. However, any analyst should be aware of the possibility of stochastic variation in simulation output giving misleading information.

REFERENCES

Bratley, P., Fox, B. L., and Schrage, L. E.(1987). *A Guide to Simulation.* Second Edition, Springer-Verlag, New York.

Clark, G. M. and Yang, W.(1986). A Bonferroni selection procedure when using common random numbers with unknown variances. In: *Proceedings of the 1986 Winter Simulation Conference*(J. R. Wilson, J. O. Henriksen, and S. D. Roberts eds.) Institute of Electrical and Electronic Engineers, San Francisco, California, 313 - 315.

Dudewicz, E. J. and Dalal, S. R.(1975). Allocation of observations in ranking and selection with unequal variances. *Sankhya* 37, 28 - 78.

Johnson, N. J.(1978). Modified t tests and confidence intervals for asymmetric populations. *Journal American Statistical Association* 73, 536 - 544.

Keeney, R. L. and Raiffa, H.(1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley & Sons, New York.

Kleijnen, J. P. C.(1975). *Statistical Techniques in Simulation Part II.* Marcel Dekker, Inc., New York.

Kleijnen, J. P. C.(1987). *Statistical Tools for Simulation Practitioners.* Marcel Dekker, Inc., New York.

Law, A. M. and Kelton, W. D.(1982). *Simulation Modeling and Analysis*,McGraw-Hill Book Company, New York.

Nelson, B. L.(1986). Personal conversation. Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio.

Nelson, B. L.(1987). Variance reduction for simulation practitioners. In: *Proceedings of the 1987 Winter Simulation Conference*(A. Thesen, H. Grant, W. D. Kelton, eds.) Institute of Electrical and Electronic Engineers, San Francisco, California, 43 - 51.

Sargent, R. G., Kang, K., and Goldsman, D.(1987). An investigation of small sample size behavior of confidence interval estimation procedures. Working Paper #87-005, Department of Industrial Engineering and Operations Research, Syracuse University, Syracuse, New York.

Schmeiser, B.(1982). Batch size effects in the analysis of simulation output. *Operations Research*. 30, 556 - 568

BIOGRAPHY

GORDON M. CLARK is a professor of Industrial and Systems Engineering at The Ohio State University, and he is currently on sabbatical with the IBM Corporation at their Application Systems Division. He received the B.I.E. degree in Industrial Engineering from The Ohio State University in 1957, a M. Sc. in Industrial Engineering from The University of Southern California, and a Ph.D. from The Ohio State University in 1969. He was responsible for the development of several combat simulations at The Ohio State University while under contract with the U.S. Army (1965 - 1974). He was an associate editor of *Operations Research* (1975 - 1978) and a member of the board of directors for the Military Operations Research Society (1972 - 1976). He has consulted in the areas of simulation modeling and analysis and manufacturing systems for industrial companies and Battelle Memorial Institute. His research interests include simulation optimization, applications in manufacturing systems, and approximations to nonstationary queueing systems. He is a member of ORSA, TIMS, IIE, SCS, and SIGMA XI.

Until June, 1989
Gordon M. Clark
R17A
IBM Corporation
P.O. Box 28331
Atlanta, GA 30358-0311
(404) 956-6393

After June, 1989
Gordon M. Clark
Department of Industrial
and Systems Engineering
The Ohio State University
Columbus, OH 43210
(614) 888-1746