Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

HOW SIMULATION LANGUAGES SHOULD REPORT RESULTS: A MODEST PROPOSAL

Jamie R. Wieland

Dept. of Management & Quantitative Methods
Illinois State University
Normal, IL 61790, USA

Barry L. Nelson

Dept. of Industrial Engineering & Management Sciences
Northwestern University
Evanston, IL 60208, USA

ABSTRACT

The focus of simulation software environments is on developing simulation models; much less consideration is placed on reporting results. However, the quality of the simulation model is irrelevant if the results are not interpreted correctly. The manner in which results are reported, along with a lack of standardized guidelines for reports, could contribute to the misinterpretation of results. We propose a hierarchical report structure where each reporting level provides additional detail about the simulated performance. Our approach utilizes two recent developments in output analysis: a procedure for omitting statistically meaningless digits in point estimates, and a graphical display called a MORE Plot, which conveys operational risk and statistical error in an intuitive manner on a single graph. Our motivation for developing this approach is to prevent or reduce misinterpretation of simulation results and to provide a foundation for standardized guidelines for reporting.

1 INTRODUCTION

Modern simulation software environments provide substantial support for developing simulation models, both the stochastic input models that drive the simulation, and the logical model that describes how the system reacts to the stochastic inputs. Modeling is clearly the critical starting point for a successful simulation study, so this emphasis is well placed. At the end of the day, however, the primary reason for undertaking a simulation study is to generate system performance estimates. If these performance estimates are not interpreted correctly, then the quality of the simulation model is irrelevant.

There are at least two ways that performance estimates can be misinterpreted. One is that the meaning of the estimate, as it relates to the system design objectives that motivated the study or the operational decisions that have to be made, is misunderstood. We believe that this often occurs when decisions are, incorrectly, based on long-run averages when the real focus should be on risk, which typically has little to do with average performance. The other common misinterpretation occurs when system performance has not been estimated precisely enough to support the decision for which it was intended. This is a statistical problem related to not running the simulation long enough (in terms of run length or number of replications or sometimes both). *We believe that it is possible to reduce the incidence of both types of misinterpretations.*

One factor that likely contributes to both of these issues is the manner in which the observed simulation performance is reported. The types of information reported varies among simulation software environments. Some reports create more opportunity for misinterpretation than others, but lack of standardized guidelines that are widely used for reporting results could be responsible for user errors. Law and Kelton (2000) provide some high-level guidelines for reporting results which we have not observed to be widely used in practice.

In this paper we propose a hierarchical approach for reporting simulation results, where each level provides a summary of the results, but also allows the analyst to drill deeper depending on what is required by their particular problem and their own statistical expertise. The objective in developing this approach is to prevent or reduce the incidence of misinterpretation.

Our approach combines recent work on output analysis by Song and Schmeiser (2009) and Nelson (2008). Song and Schmeiser introduced rules for displaying the "significant digits" of performance estimates and their respective standard errors. Nelson described a graphical method, called a MORE (measure of risk and error) plot, as a way to display risk—that is, uncertainty about future outcomes—and a measure of statistical error—which addresses whether or not the simulation has been run long enough—in an intuitive way on a single plot.

2 APPROACHES FOR REPORTING RESULTS

Simulation environments usually provide summary measures of observed performance, as they should, along with the ability to store or access the raw data. However, they typically display only one type of summary. Law and Kelton (2000) note that for each performance measure of interest, the average observed value, the minimum observed value, and the maximum observed value are usually provided. We have seen the following choices implemented in various simulation environments (sometimes supported by a graphical presentation):

1. Report the sample mean of each output.
2. Report the sample mean of each output and a confidence interval (CI) on the mean.
3. Report the sample mean of each output, a CI on the mean, and the maximum and minimum observed value.
4. Report the sample mean of each output, a CI on the mean, and a number of other summary measures such as the sample standard deviation and every 5th percentile.
5. Provide access to, or at least a way to record and export, all of the raw data generated by the simulation for analysis via other software.

These choices for displaying summary measures of observed performance illustrate the lack of standardized guidelines for reporting simulation results. We contend that no single summary of the data is adequate, nor is it sufficient to simply retain all of the raw data. We propose, instead, a four-level approach, starting at Level Three and moving down to Level Zero, which is simply the raw data. Level Three provides a very concise, but still statistically meaningful, summary, while each lower level drills deeper. The levels should be connected by hyperlinks, as illustrated later in the paper, to facilitate this deeper exploration.

The advantage of a multi-level report, compared to displaying only one type of summary, is that detailed information is easily accessible for skilled users, but inexperienced users are not overwhelmed with too much information. A multi-level approach allows all users to access information on an as-needed basis. For the inexperienced user it is important that the top level of results be concise, yet not misleading. Even reporting CIs along with the sample means at the highest level provides an opportunity for (incorrectly) interpreting the CI as a measure of future risk rather than a measure of statistical error.

Guidelines for output reports have been suggested by Law and Kelton (2000). They state that output reports should include the following:

1. Summary statistics including the average observed value, the minimum observed value, and the maximum observed value. If a standard deviation estimate is also provided, then the user should be sure that it is based on a statistically acceptable method, such as independent replications or batch means.
2. A variety of static graphs, including a histogram, a time plot, and a correlation plot.
3. Access to the raw data.

Consistent with Law and Kelton's guidelines, our approach includes the average observed value, the minimum observed value, the maximum observed value, a histogram, and access to the raw data. We also include a standard deviation estimate and agree that when such an estimate is provided that it should be based on a statistically acceptable method, such as independent replications or batch means. However, we do not think that it is the responsibility of the user to assure that valid methods were used. Such features should be embedded into simulation software environments. By default, our proposed approach does not generate time-series plots or correlation plots. However, we do include a list of recommended plots in Section 7.

The novelty of our approach lies not only in utilizing significant digits and MORE Plots, as previously discussed, but also in the proposed hierarchy through which we recommend information be linked. Law and Kelton do not provide concrete guidelines for how and in what order the suggested information should be displayed. We argue that providing all users with all information simultaneously could result in misinterpretation.

Law and Kelton also recommend that when simulating multiple scenarios a database be constructed to store the observations from each scenario, with the option of plotting results across scenarios on a single graph. In this paper we focus on reporting output for a single scenario, but our approach could be extended to multiple scenarios.

In the next five sections we describe each level of reporting and illustrate them with an example. In Section 8 we also describe a prototype implementation of this four-level approach in VBA for Excel that is available (free) for download.

Table 1: Sample means for both performance measures, along with estimated standard errors, as would be displayed by current software environments assuming that 5 digits are used for reporting output

Performance Measure	Sample Mean	Estimated Standard Error
Expected Patient Wait Time	122.72	10.331
Doctor Utilization	0.7637	0.1383

3 AN EXAMPLE

For describing each of the four levels of the output report, we consider the following example problem: Suppose a simulation model has been constructed to analyze potential staffing changes in a hospital emergency department during the peak-arrival time window on the weekend. Hospital administration is concerned with analyzing not only the effects that staffing changes will have on expected patient wait times (measured in minutes), but also the effects on the distribution of individual patient wait times and the number of patients waiting. Utilization of doctor's time may also be of interest when considering whether or not to adjust staffing levels.

4 LEVEL THREE: SAMPLE MEANS TO SIGNIFICANT DIGITS

Clearly there is value in having a comprehensive, easy-to-digest summary of all of the performance measures generated by the simulation. For this purpose an appropriately organized and uncluttered table of sample means is hard to beat. However, even at this level it is important to avoid implying a level of statistical precision that is not supported by the data, which is almost certain to happen when a predetermined number of digits are displayed. To avoid this pitfall, we adopt the "significant digits" procedure of Song and Schmeiser (2009).

The Song and Schmeiser procedure determines the number of digits of a point estimate that should be displayed when reporting results and discards all "meaningless" digits. Meaningless digits are determined by the point estimate's standard error and are essentially noise because they can be discarded with very little loss of statistical information. Song and Schmeiser's significant digits procedure provides a probability guarantee on the loss of statistical information, which, practically interpreted, states that all meaningless digits could just as well be replaced with randomly selected digits between 0 and 9, without loss of information. In other words, the chance that a meaningless digit is correct is only about one-in-ten.

The number of meaningless digits of a point estimator is inversely proportional to its standard error. When the standard error is of the same order of magnitude (or higher) as the point estimate, then all point-estimate digits are meaningless. In this case we display the point estimate as a single "X", along with an error message indicating that more data are needed to provide a meaningful estimate. When the standard error of the point estimate is orders of magnitude smaller than the point estimate itself, then few digits (if any) are omitted. When a digit of a point estimate is omitted, we replace it with an "X" which serves merely as a placeholder. Replacing all meaningless digits with X's could prevent users from inadvertently making decisions based on random noise because the system performance measure has not been estimated precisely enough to support the decision.

An alternative to using X's as placeholders for meaningless digits is to use scientific notation. The advantage of using X's in the display is that it immediately draws attention to itself. For example, consider a case where a point estimate is 20,032 with a standard error of 50. Using scientific notation, this estimate would be displayed as 2.00E4. Using X's as place holders, this estimate would be displayed as 20,0XX. Both of these notations have the same interpretation in that the last two digits, 3 and 2, are meaningless. However, we argue that 2.00E4 is more likely to incorrectly be interpreted as 20,000, whereas, the X notation is likely to spark the attention of users, prompting them to inquire further about how such a result should be interpreted. This issue may be of particular concern for simulation software environments targeted for applications where scientific notation is not widely used in practice.

Continuing the example of analyzing the effects of staffing changes in a hospital emergency department during the weekend peak-arrival time window, assume that a small number of replications were performed to estimate patient wait time and doctor utilization. Current software environments do not take significant digits into account when displaying estimates. Instead, they dump a pre-determined number of digits for all items displayed in the report. For example and without loss of generality, assume that the software pre-determined that 5 digits would be used for displaying all items in the report. Table 1 displays the sample means and standard errors for both patient wait times and doctor utilization as would be displayed by current simulation software environments.

Table 2: Level Three report output for the example problem

Performance Measure	Sample Mean
Expected Patient Wait Time	1XX
Doctor Utilization	X

Table 3: Level Three report output for the example problem after increasing the number of replications

Performance Measure	Sample Mean
Expected Patient Wait Time	11X
Doctor Utilization	0.8

Rather than using a pre-determined number of digits for each item in the report, the proposed Level Three report displays only the significant digits for each of the sample means. Table 2 displays the Level Three output report based on a small number of independent replications. The meaningless digits of the sample means, as determined by the standard error of each estimate, were omitted and replaced with X's.

In this example, only the first digit of the average patient wait time is significant; hence, the remaining digits are omitted. Due to the standard error being of the same order of magnitude as the average utilization, there are no significant digits. In this case only an X is reported, indicating that more simulation would be required to obtain a meaningful estimate.

For discussion purposes, we have increased the number of replications and displayed the Level Three report in Table 3. In this example, the average patient wait time was found to be 110.57, with a standard error of 2.3. The average doctor utilization was found to be 0.80, with a standard error of 0.04. Increasing the number of replications provided an additional digit of precision for both estimates.

5 LEVEL TWO: MORE PLOT

Each sample mean at Level Three should be linked to a MORE Plot. The MORE Plot not only displays both risk and a measure of statistical error (CI) in an intuitive way on a single plot, but also allows users to see all the results for any single output graphically summarized.

The MORE Plot starts with a histogram of the output data, a familiar plot that displays the variability in actual performance that the analyst should expect to experience if the system design is implemented. The sample mean, from Level Three, is displayed on the histogram, along with a "risk box" that captures the likely future outcomes. Both the sample mean and the risk box are displayed by use of arrows pointing to their respective values on the horizontal axis of the histogram. Nelson (2008) suggested using the sample 5th and 95th percentiles to define the risk box, but other choices are possible. The key message that the risk box portrays is that actual performance in reality could differ, perhaps substantially, from the long-run average performance.

Besides conveying a sense of risk, the MORE Plot also answers the question, have we done enough simulation to be confident in making any decision yet? This is accomplished by adding CIs below the arrow heads that indicate the sample mean and risk box. These intervals portray how confident we are about where the three arrow heads belong. If these intervals are relatively wide, then it implies that we have not done enough simulation since their positions are quite uncertain. Although we cannot simulate away risk (that is, we cannot drive the width of the risk box toward zero), we can shrink the measures of error (CIs) by doing more simulation. Even without a sophisticated knowledge of statistics, an analyst can continue to increase the simulation effort until these confidence intervals are acceptably narrow on the scale of the output.

Output reports for many simulation environments include CIs when sample means are reported. In the proposed approach, we deliberately hold off providing confidence intervals by choosing to include them in Level Two rather than Level Three. Further, CIs are first displayed graphically, within a MORE Plot, rather than being listed in a table along with the sample means. The advantage of first displaying CIs within a MORE Plot is that it reduces the chance of interpreting the interval as capturing future risk, which is (correctly) captured by the risk box.

For the example problem, a MORE Plot for the patient wait time is displayed in Figure 1. The CIs on the horizontal axis of the MORE Plot indicate that we are still uncertain about the locations of the arrow for the average wait time and lower

and upper arrows for this risk box. If the hospital administration were interested in excessive wait times, then the upper arrow of the risk box may be of particular interest. The measure of error for the upper arrow of the risk box is relatively large, which indicates that much more simulation is required. For illustration, a MORE Plot based on a larger number of replications is shown in Figure 2.

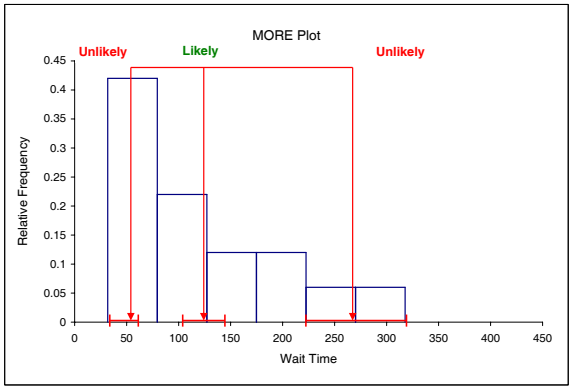

Figure 1: Level Two report for patient wait time

The risk box displayed in Figure 2 is quite wide, which indicates that observed waiting times can be quite different from the average waiting time. Even after increasing the number of replications, the confidence interval for the upper arrow of the risk box is still much larger than that for the sample mean and the lower arrow of the risk box. This indicates that basing the number of replications only on the sample average may not yield the same level of precision for other performance measures.

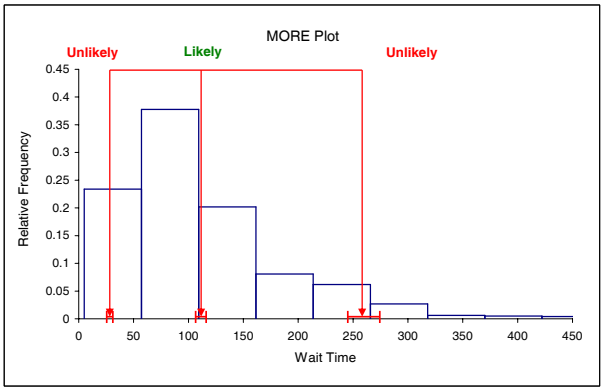

Figure 2: Level Two report for patient wait time based on a larger number of replications

6 LEVEL ONE: DETAILED STATISTICAL SUMMARY

For many users Levels Three and Two will be entirely adequate, and Nelson (2008) argues that many fewer interpretation and statistical errors would occur if the MORE Plot was available. However, the statistically sophisticated analyst is likely to desire numerical information, and we suggest that Level One be a detailed statistical summary report that is hyperlinked to the MORE Plot. All of the measures that go into the MORE Plot should be displayed, along with a collection of standard measures such as the standard deviation, the range, and every 5th percentile of the data. See Figure 3.

7 LEVEL ZERO: RAW DATA

It is impossible to anticipate every type of analysis that might interest a user. Therefore, we recommend that there always be an option to retain all of the raw data (at least for selected performance measures) generated by the simulation. The raw

Summary Statistics

Number of Replications	50
Sample Average	122.7182792
Standard Error	10.33104015
Standard Deviation	73.05148543
Minimum Observation	31.84452982
Maximum Observation	317.8192237

MORE Plot

	Estimate	CI Lower Bound	CI Upper Bound
Sample Average	122.7182792	102.4698126	142.9667458
5th Percentile	52.44996801	31.84196909	59.67406594
95th Percentile	266.6287393	220.8534926	317.8062505

Percentiles

5th	52.44996801
10th	58.67973644
15th	62.23147408
20th	66.06386826
25th	69.88803525
30th	71.51901901
35th	75.91428402
40th	77.97396604
45th	80.58301067
50th	84.69621362
55th	108.4446675
60th	119.0345356
65th	133.7905837
70th	140.3398093
75th	163.6966434
80th	191.6330144
85th	211.2023178
90th	224.6949032
95th	266.6287397

Figure 3: Level One report for patient wait time

data are useful to further explore surprising results found in the summary measures, as well as supporting more specialized statistical analysis.

The sophisticated analyst may want to export the raw data to statistical analysis software, and this should be facilitated. If the simulation software developers want to include some additional analysis functionality to act on the raw data, then we recommend tools to plot the empirical cumulative distribution function, a time series plot and a correlation plot.

8 USING THE PROTOTYPE

A prototype is available for download at <users.iems.northwestern.edu/ nelsonb/prototype2003.xls> or at <users.iems.northwestern.edu/ nelsonb/prototype2007.xls> where 2003 or 2007 refers to the version of Excel in use. This prototype is an Excel VBA file that automatically generates all four levels of output reports for any given data sets. The prototype first requires users to import the raw data for each output measure in a separate column. It then asks for the total number of variables that should be included in the report, along with the column references for each variable. The prototype will then open new worksheets for each of the four report levels to display the relevant summary. Hyperlinks are included within each report to easily move between levels.

REFERENCES

Law, A. M. and W. D. Kelton. 2000. *Simulation modeling and analysis*, 3rd ed. New York: McGraw Hill.

Nelson, B.L. 2008. The MORE Plot: Displaying measures of risk and error from simulation output. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 413–416. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Song, W.T. and B.W. Schmeiser. 2008. Displaying statistical point estimators: the leading-digit procedure. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 407–412. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Song, W.T. and B.W. Schmeiser. 2009. Omitting meaningless digits in point estimates: the probability guarantee of leading-digit rules. *Operations Research* 57:109-117.

AUTHOR BIOGRAPHIES

JAMIE R. WIELAND is an Instructional Assistant Professor of Management and Quantitative Methods at Illinois State University. Her research interests are in stochastic operations research, with focus on simulation methodology and applied statistics. Her email address is <jamie.wieland@ilstu.edu>.

BARRY L. NELSON is the Charles Deering McCormick Professor and Chair of the Department of Industrial Engineering and Management Sciences at Northwestern University. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. His e-mail and web addresses are <nelsonb@northwestern.edu> and <www.iems.northwestern.edu/ nelsonb/