Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

THE MORE PLOT: DISPLAYING MEASURES OF RISK & ERROR FROM SIMULATION OUTPUT

Barry L. Nelson

Department of Industrial Engineering & Management Sciences
Northwestern University
Evanston, IL, U.S.A.

ABSTRACT

The focus on mean (long-run average) performance as the primary output measure produced by simulation experiments diminishes the usefulness of simulation for characterizing risk. Confidence intervals on means are often misinterpreted as measures of future risk, when in fact they are measures of error. We introduce the Measure of Risk & Error (MORE) plot as a way to display and make intuitive the concepts of risk and error and thus support sound experiment design and correct decision making.

1 INTRODUCTION

Confidence intervals (CIs) for the mean are one of the first topics on statistical inference in introductory statistics courses, and stochastic simulation classes build on this foundation by emphasizing the need for CIs on estimated measures of system performance. A CI is a measure of error that attempts to cover, with high probability, the unknown value of a performance parameter implied by the simulation model. More relevant for decision making, however, is some sense of what the future might bring, say if the system design described by the simulation is implemented. The mean might be a good single-point guess of the future, but neither the mean nor a confidence interval on it capture any measure of the future risk.

Confusion about the meaning and relationship of risk and error is pervasive, and the presentation of output results by most simulation tools does nothing to alleviate it. Here we introduce the *Measure of Risk & Error* (MORE) plot as a possible default output display to make their relationship clear and facilitate the kind of sequential experimentation that simulation should readily support. No background in inferential statistics is required to interpret the MORE plot, and only the descriptive statistics concepts of sample average and histogram are essential. To illustrate this point, we first describe the MORE plot without reference to the statistical methods used to construct it, which are described later in the paper.

2 MORE PLOT

Suppose we have run a simulation experiment and one of the outputs is the number of barrels, in thousands, of a particular chemical that we need annually. This number depends on a complex host of things: demand for our product, yield loss, etc. We might be interested in how much to stock or on whether we should pay for an option to get more at a fixed price later in the year. After constructing a simulation to generate the annual need for barrels of the chemical, and simulating a number of years of consumption, we get the histogram shown in Figure 1 (top).

There are at least two questions we need to answer: How many barrels should we purchase or have an option on, and have we done enough simulation to really answer that question? Since humans love to average, we add the sample average to the histogram. And since it is clear that we could need much more or much less than the average, we also box a big chunk of the possible need and label it in an easy to understand way; see Figure 1 (middle). We immediately obtain an important insight by looking at the simulation output data in this way: The future is uncertain and our needs can be within a wide range around the average.

In baseball a players batting average last year is a meaningful historical statistic. However, a simulation is not trying to create history; instead it is trying to say something about what will happen in the future and whether we can live with it; the average does not always tell us, while the "risk box" in a MORE plot often does.

Have we done enough simulation to be confident in making any decision yet? As a final embellishment, we add a measure of error on each of the arrow heads, as shown in Figure 1 (bottom). These intervals imply that we are highly confident that the arrow head belongs *somewhere* in the interval, we just are not sure where. Now it becomes obvious that we have not done nearly enough simulation

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

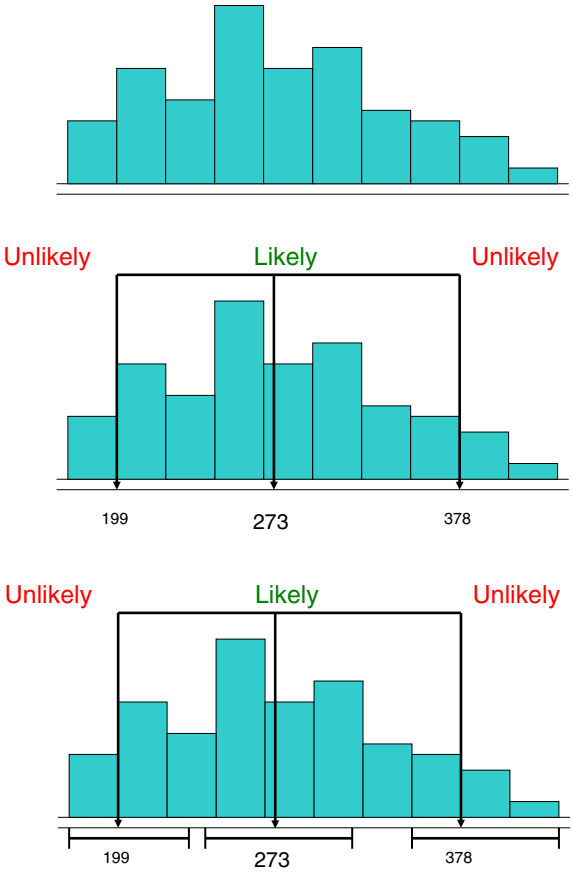

Figure 1: Construction of a MORE plot.

since the positions of the arrow heads are quite uncertain relative to the amount of the chemical we might use. We call the bottom figure a MORE plot.

Figure 2 shows the MORE plot we get if we run the simulation for many more years. *Notice that uncertainty about the future does not disappear; we cannot simulate away risk.* But we do improve our estimate of future uncertainty by running the simulation longer, as indicated by the shorter error intervals. With this information we can balance the various costs associated with the decision in light of the likely outcomes and do something rational.

The MORE plot is based on the catch phrase "use MOE to get MOR," which means *use measures of error to get measures of risk*. The box in the MORE plot is a measure of future risk, and that is what we often need to support our decision. The intervals are measures of error; they tell us if we have done enough simulation to reach a conclusion.

As a second example to illustrate how error diminishes with simulation effort, but risk does not, Figure 3 shows a sequence of MORE plots for a simulation of product cycle time as we increase the number of replications simulated. If our interest is in setting a promise date so that we are

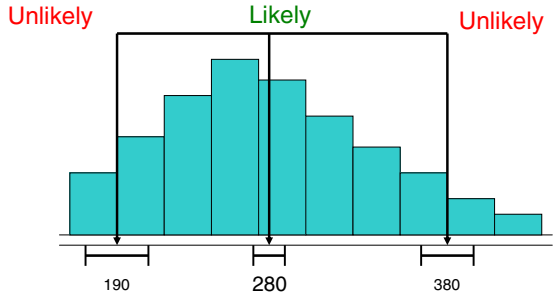

Figure 2: MORE plot with more data.

unlikely to be late, then the right-hand arrow head might be of most interest. Notice that the error in the sample average is smaller than the error in the risk box, which illustrates why we cannot always design the simulation based solely on the average.

The MORE plot is also relevant even when the mean is the quantity of primary interest. Figure 4 shows MORE plots for the aggregate (all-store) weekly profits generated by two different magazines titles; the simulation was performed to find the optimal stocking levels to maximize long-run average profit. While the weekly demand distributions of these two titles store by store are the same, the sales of one of them (bottom MORE plot) are very much driven by who is on the magazine cover. Since the same cover appears in all stores, larger swings in aggregate weekly sales occur for the title with the cover effect.

Under the optimal stocking formula both of these titles have the same long-run average profit, and will realize that profit over many weeks. However, the MORE plot shows that there is substantially more cash flow risk when there is a common cover effect. If the magazine distributor was unprepared for these big swings, then a few bad weeks of profits (low achieving covers) might cause them to abandon their "optimal" stocking policy thinking it must be wrong. That could be a big mistake, particularly if they use an ad hoc fix that results in an unknown loss of potential profit. The MORE plot shows them what to expect, not just over the long run, but also week to week.

3 DETAILS FOR I.I.D. DATA

The MORE plot adds to a histogram estimates of the mean and two percentiles (the arrow heads), and confidence intervals on each of them. Let the simulation output data be denoted by $Y_1, Y_2, \ldots, Y_n$, and suppose that they are performance measures extracted from different replications so that they are independent and identically distributed (i.i.d.). Without loss of generality assume that they have been sorted so that $Y_1 \leq Y_2 \leq \cdots \leq Y_n$.

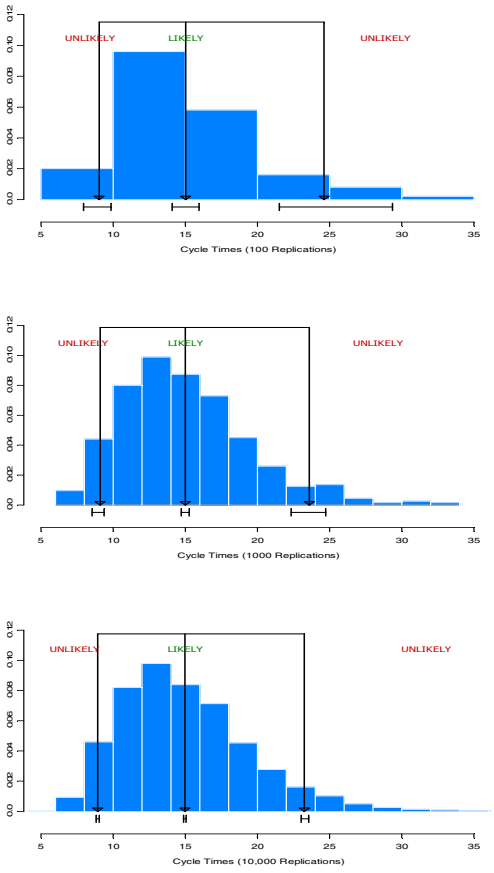

Figure 3: MORE plot of cycle time data.

The sample mean (middle arrow head) is placed at $\bar{Y} = \sum_{i=1}^{n} Y_i/n$. If n is large enough so that a histogram can be formed, then a reasonable choice for the interval under it is

$$\left[\bar{Y} - z_{1-\alpha/2}S/\sqrt{n},\ \bar{Y} + z_{1-\alpha/2}S/\sqrt{n}\right]$$

where S^2 is the sample variance of the data and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution (e.g., 1.96 if $\alpha = 0.05$ for a 95% CI).

For the percentile arrow heads that form the risk box, we chose the 5th and 95th percentiles of the data in this paper (meaning that the box contains 90% of the outcomes), but any percentiles (not necessarily symmetric) could be chosen. Generically, suppose we choose the β_L100th and β_U100th percentiles, with $\beta_L < 1/2$ and $\beta_U > 1/2$. Then we place the left arrow head at $Y_{\lfloor n\beta_L \rfloor}$ and the right arrow head at $Y_{\lfloor n\beta_U \rfloor}$ where $\lfloor \cdot \rfloor$ means to round down. This is an easy choice; more refined percentile estimators that interpolate two or more of the data points could also be used.

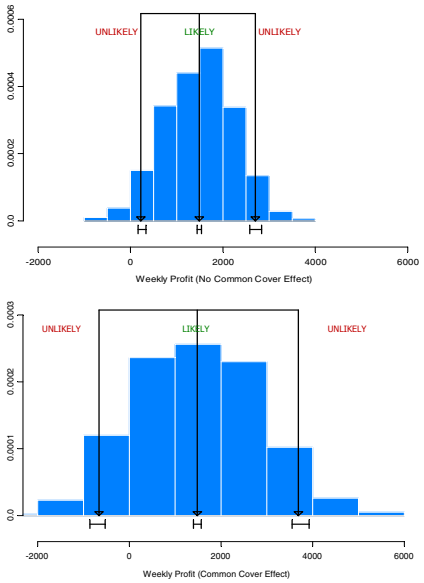

Figure 4: MORE plot of magazine profit data.

A large n approximate CI for the β100th percentile can be obtained as follows (Banks et al. 2005): Let

$$\begin{aligned} \beta_1 &= \beta - z_{1-\alpha/2}\sqrt{\frac{\beta(1-\beta)}{n-1}} \\ \beta_2 &= \beta + z_{1-\alpha/2}\sqrt{\frac{\beta(1-\beta)}{n-1}}. \end{aligned}$$

Then take as the interval $\left[Y_{\lfloor n\beta_1 \rfloor}, Y_{\lfloor n\beta_2 \rfloor}\right]$. An interval like this is needed for each percentile. Again, more refined methods exist, including nonparametric methods that do not depend on a normal approximation.

A few additional refinements are worth mentioning:

- Although less frequently used in engineering practice, the sample median could replace the sample mean as the center of the MORE plot.
- If it is desirable to have all three of the confidence intervals cover their respective parameters with probability $\geq 1-\alpha$, then the Bonferroni inequality implies that we form each interval at the $1-\alpha/3$ level (that is, use $z_{1-\alpha/6}$).
- To avoid any implication of statistical inference, the CIs in the MORE plot could be replaced, for instance, with ± 2 `standard error` (although it the case of the percentiles it is actually easier to form the CI than estimate the standard error).

Many readers will recognize that there is a close connection between the risk box in a MORE plot and a *prediction interval (PI)* (Banks, et al. 2005). Loosely speaking, a PI is

what you get if you move the risk box arrow heads out to the far extremes of the error intervals because a PI simultaneously accounts for risk and error in a single interval. A PI is therefore conservative and does not provide a sense of what could happen—how much the risk box might shrink—with additional replications.

4 STEADY-STATE SIMULATION OUTPUT

When the performance measure of interest is a summary measure of a random variable Y that is defined by limit as the simulation run length (conceptually) goes to infinity, then care must be taken to construct a valid MORE plot.

Suppose that $Y_1, Y_2, \ldots, Y_n$ is the output of a single replication of a "steady-state simulation" after the impact of initial conditions have been mitigated. For instance, Y_i might be the cost incurred in the ith month of a supply chain simulation and we are interested in characterizing the fluctuations of monthly cost. If we assume that this output comes from a stationary stochastic process, then the histogram, sample mean and sample percentiles for the MORE plot can be obtained as described above. The difficulty arises when trying to construct the error intervals because the outputs are typically dependent, invalidating the CI's from the previous section.

Perhaps the simplest approach is to use the methods of batch means and batch percentiles (e.g., Wood and Schmeiser 1995) to form the CIs. This is hard to do in an entirely automated way, but for a MORE plot even a rough measure of the error in the sample means and percentiles is adequate.

Care must be taken if the replication-deletion approach—which is widely used for estimating the mean—is applied. In particular, the histogram must be formed from all of the raw data retained after deletion from all replications. Further, from each replication an estimate of the mean and each of the percentiles should be obtained; these are used to compute point estimates and CIs by averaging. In other words, we apply the methods of batch means and batch quantiles with each replication playing the role of a batch. This is critical since the run length of a steady-state simulation replication is arbitrary and not related to any property of the system we are trying to capture.

5 MISTAKES

Suppose we are simulating a call center during its operating hours from 8 AM to 7 PM, and we are interested in caller delay before talking to an operator. Caller load on the call center varies systematically throughout the day with peak loads around noon. The natural experiment design here is to make multiple replications, each representing a day of service, and a typical output is the average caller delay for the day.

A MORE plot is easily constructed from the output of this simulation, but it is just as easily misinterpreted. In particular, the risk box can be incorrectly interpreted as the likely delay that an individual caller will experience. This is incorrect because the MORE plot was constructed from *daily average delays*, not individual caller delays. Therefore, the risk box characterizes the variability of the daily average, not the individual callers' delays. In fact, the distribution of individual caller delay is not well defined in this problem since there is a strong dependence on the time of day the call was placed.

A second mistake is trying to squeeze too much information out of a MORE plot when too little data have been obtained. Typically the mean is more easily estimated than the percentiles; for instance, we might be comfortable estimating the mean from $n = 10$ replications/observations, but we cannot hope to estimate, say, the 5th percentile from such a small sample. The MORE plot is designed for situations in which enough data have been generated to form a reasonable histogram.

Finally, the MORE plot captures nothing about model risk, which relates to how faithfully the simulation model represents the real system of interest. For instance, model risk is present any time an input distribution that drives the simulation is based on a sample of data.

ACKNOWLEDGEMENTS

The author acknowledges the helpful comments of Bruce Ankenman, Ira Gerhardt, Wally Hopp, John Fowler and Jeremy Staum on the development of the MORE plot.

REFERENCES

Banks, J., J. S. Carson, B. L. Nelson and D. M. Nicol. 2005. *Discrete-event system simulation*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Wood, D. C. and B. W. Schmeiser. 1995. Overlapping batch quantiles. *Proceedings of the 1995 Winter Simulation Conference*, ed. C. Alexopoulos, K. Kang, W. R. Lilegdon and D. Goldsman, 303–308. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

AUTHOR BIOGRAPHY

BARRY L. NELSON is the Charles Deering McCormick Professor of Industrial Engineering and Management Sciences at Northwestern University and is Editor in Chief of *Naval Research Logistics*. 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/>.