Reducing Simulation Runs for Future Combat System Key Performance Parameter Analysis

Lieutenant Colonel Thomas M. Cioppa, Ph.D.
US Army Training and Doctrine Command Analysis Center
PO Box 8695
Monterey, CA 93943-0695

Abstract: The Future Combat System (FCS) Key Performance Parameter (KPP) simulation runs executed in the stochastic Combined Arms and Support Task Force Evaluation Model (CASTFOREM) simulation can exceed 40 hours per replication. Historically, each combination of input factor levels requires 21 replications. These replications, when coupled with the large number of combinations of different input factors, can exceed manpower and computer resources. A methodology to minimize the number of replications required is proposed. Applying normality theory to this issue can be misleading since the output measures may not be normally distributed. The proposed methodology incorporates two techniques to determine the minimum number of replications required and reduce the required number of simulation runs. The first technique examines the output measure using the coefficient of variation, which is defined for the output measure as its standard deviation divided by the mean. The second technique is to use bootstrapping for the executed simulation runs. A bootstrapping sample is obtained by randomly sampling from the original data points. Bootstrap confidence intervals can be constructed to compare various alternatives. These techniques were robustly applied to recent FCS KPP CASTFOREM simulation runs and showed substantial merit.

I. INTRODUCTION

The time to execute a simulation is always of concern and interest to the analyst. The valuable and limited resource of time is best applied to ensuring the simulation setting is accurate, data inputs are valid, and sufficient time is available to analyze the simulation output to provide results to decision makers (Law and Kelton, 2000). The time problem is compounded in military simulations where scenario establishment is time-consuming. Recently, the Future Combat System (FCS) Key Performance Parameter (KPP) analysis required extensive simulation support.

The simulation used at the US Army Training and Doctrine Command Analysis Center (TRAC) at White Sands is the Combined Arms and Support Task Force Evaluation Model (CASTFOREM). CASTFOREM is used to evaluate weapon systems and unit tactics, brigade and below by simulating intense battle conditions at battalion and brigade levels. It models a range of operations to include ammunition resupply, aviation, close combat; combat service support; C3, countermobility, logistics, engineering, mine warfare, fire support, intelligence and electronic warfare, mobility; survivability, and air defense. The time to execute one replication of one scenario exceeded 40 hours.

Previous experience with CASTFOREM showed that after 21 replications of most scenarios, the variance of the measure of effectiveness (MOE) had stabilized (Cherolis, 1992). The problem faced in the FCS KPP study was that even 21 replications of a single scenario could exceed 35 days. An alternative methodology for reducing the required number of replications was needed.

II. BACKGROUND AND NOTATION

Cherolis' (1992) thesis is based on the assumptions that the replications are independent and produce a sequence of independent, identically distributed random variables X_1, X_2, X_3, ..., X_n . The Central Limit Theorem is used to derive confidence intervals and hypothesis tests. When n is sufficiently large, the distribution of the random variable is

$$\frac{\overline{X}}{s_n / \sqrt{n}} \ ,$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size. This distribution is approximately normally distributed, where $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $s_n = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$. Thus, for sufficiently large n, an approximate $100 \times (1 - \alpha)\%$ confidence interval for the population mean, μ , is given by

$$\overline{X} \pm t_{n-1,1-\alpha/2}\sqrt{\frac{s_n^{\ 2}}{n}},$$

where $0 < \alpha < 1$ and $t_{n-1,1-\alpha/2}$ is the upper $1 - a/2$ critical point for the t distribution on $n - 1$ degrees of freedom.

Although the 21 replications is commonly used, it is quite possible that more than 21 replications (thus more than 35 days) may be required per scenario or alternative depending on the precision required. A justifiable methodology that could reduce the number of replications required was needed.

III. COEFFICIENT OF VARIATION AND BOOTSTRAP

The coefficient of variation (CV) is defined as the standard deviation divided by mean. It is a statistical measure of the deviation of a variable from its mean. There are no units associated with this measure. A smaller value is better and implies less variability. The data does not have to be normally distributed. A data set with a higher CV will have a larger confidence interval than a data set with a smaller CV. The one drawback of the CV is a measure is that MOE data must be quantitative and positive. This is not considered a significant problem since the majority of MOE data from constructive simulations satisfies these two requirements. From different fields and experience, a CV less than 0.20 indicate a reasonable amount of variability.

The bootstrap method (Efron and Tibshirani, 1994) is commonly used to resample from sparse data sets. A bootstrap sample $x^* = (x^*_1, x^*_2, x^*_3, \ldots, x^*_n)$ is obtained by randomly sampling n times, with replacement, from the original data points $x_1, x_2, x_3, \ldots, x_n$. The corresponding measure of interest (e.g., mean or median) is taken. For example, assume we have seven data points of (3, 9, 8, 5, 6, 1, 10) and its mean is 6. One bootstrap sample of these seven data points might be (6, 6, 1, 8, 1, 8, 10) and its mean is 5.714. A total of 1000 bootstrap samples are done. Above is only one example of the 1000 bootstrap samples. This procedure is done rapidly (within seconds) using a computer. A bias-corrected and accelerated bootstrap confidence interval (BCa) is calculated (via computer) from the 1000 samples and can be used to compare alternatives.

The general algorithm is described. A minimum of five replications is conducted on the simulation. The CV is calculated. If the CV is less than or equal to 0.20, the five replications are bootstrapped and the BCa obtained. If after five replications, the CV is greater than 0.20, another replication is done and a new CV calculated. This procedure is terminated when the CV is less than or equal to 0.20. Note that the CV makes no assumption of normality.

IV. APPLICATION TO DATA SETS

The TRAC element at Monterey previously gained insights on MOE data characteristics from TRAC-White Sands Night Vision and Electronic Surveillance Directorate Search and Targeting Acquisition Modeling Project. In most of the MOE examined, the CV was under 0.20 in most all cases by the time the tenth replication was analyzed.

There were 36 MOE's initially identified for the FCS KPP analysis conducted by TRAC-White Sands. Each of the MOE had four alternative force structures. Thus, a total of 144 data sets existed to determine the soundness of using the CV and bootstrap. There were 11 replications per alternative.

The mean was calculated for the 11 replications. The CV, test for normality, and mean and median 90% BCa were calculated for the first five replications, then for the first six replications, then for the first seven replications, then for the first eight replications, then for the first nine replications, then for the first ten replications, and finally for all 11 replications.

As an example, the first data set included the 11 data points of (279, 287, 356, 297, 302, 291, 294, 288, 286, 352, 306). The sample mean of these 11 replications is 303.5. The sample 90% confidence interval (using parametric statistics) of the 11 replications is (289.2, 317.7), but note the data is non-normal. The true population mean is unknown which is typical in military analysis. Bootstrap samples of 1000 were taken for each of the number of replications. Note that a different 1000 bootstrap samples can yield slightly different numbers, but these differences are negligible. Furthermore, the BCa has a slightly wider confidence interval, but it does not require normality assumptions.

Table 1 shows the results of the CV's and BCa's for this data set. For five replications, it is found that the data is not normally distributed (Kolmogorov-Smirnov test for normality). The CV was 0.087 which indicates little variability in the MOE. The mean 90% BCa is (287.8, 330.4) compared to the 90% confidence interval of the 11 replications of (289.2, 317.7). This indicates that the BCa is a reasonable approximation even with over a 50% decrement in replications required. As the number of replications examined increases, the CV remains relatively constant. Furthermore, the mean 90% BCa experiences some fluctuations, but also is relatively consistent.

Number of Replications	Normal	CV	Mean 90% BCa	Median 90% BCa
5	No	0.087	(287.8, 330.4)	(279, 302)
6	No	0.087	(289.2, 327.2)	(283, 302)
7	No	0.087	(290.7, 324.5)	(279, 297)
8	No	0.087	(290, 318.1)	(287, 297)
9	No	0.087	(289.2, 316.1)	(286, 294)
10	No	0.086	(292.7, 322.8)	(287, 299.5)
11	No	0.086	(293.1, 319.8)	(287, 302)

Table 1. CV's and Confidence Intervals for Representative MOE Data.

Table 1 was for one MOE for one of the four alternatives. For this specific MOE, the other three alternatives were examined. When a CV < .20 was achieved, the procedure was terminated at that number of replications. For this particular MOE, after five replications, each alternative had a CV < .20. Table 2 provides the CV's and confidence intervals for the remaining three alternatives. Note that each of these alternatives also had 11 replications.

	True Mean	True Median	Replications	Normal	CV	Mean 90% BCa	Median 90% BCa
Alternative 2	316.3	314	5	Yes	0.131	(276.6, 318.6)	(258, 316)
Alternative 3	349.9	361	5	Yes	0.066	(337.2, 361.6)	(326, 364)
Alternative 4	235.4	241	5	Yes	0.109	(218.2, 264)	(200, 261)

Table 2. CV and Confidence Intervals for Remaining Three Alternatives for Representative MOE Data.

An important consideration for the analyst is how to present the information to senior decision makers. TRAC-White Sands commonly employs box plots to compare alternatives using the Sheffe, Tukey, Bonferroni, or Fisher's least significant differences approach. One drawback of these approaches is that each assumes normality and equal variances among the alternatives. If the normality assumption is not valid, the non-parametric Kruskal-Wallis test can be used to compare all alternatives and then be followed with the Wilcoxan test to compare a particular pair of treatments. If equal variances are not assumed, then the Games-Howell test and one of Tamhane's tests are recommended.

The proposed alternative emphasizes the BCa and does not rely on normality or equal variance assumptions. The mean 90% BCa's can be displayed and indicate the magnitude and range of the alternatives. Figure 1 illustrates this approach and uses the data from Tables 1 and 2. Thus, if "bigger is better," then Alt 3 > Alt 1 = Alt 2 > Alt 4. This is obtained by examining if significant overlap of the green bands occurs between alternatives. Since there is no overlap between Alternative 3 and the remaining alternatives, Alternative 3, having the highest band, is determined to be best. Although Alternative 1 and Alternative 2 do not coincide exactly (although substantial overlap exists in their BCa's), there is sufficient visual evidence for the senior decision maker to suggest that there is not a remarkable difference between the two. Furthermore, insight is gained that Alt 3 has the least variability.

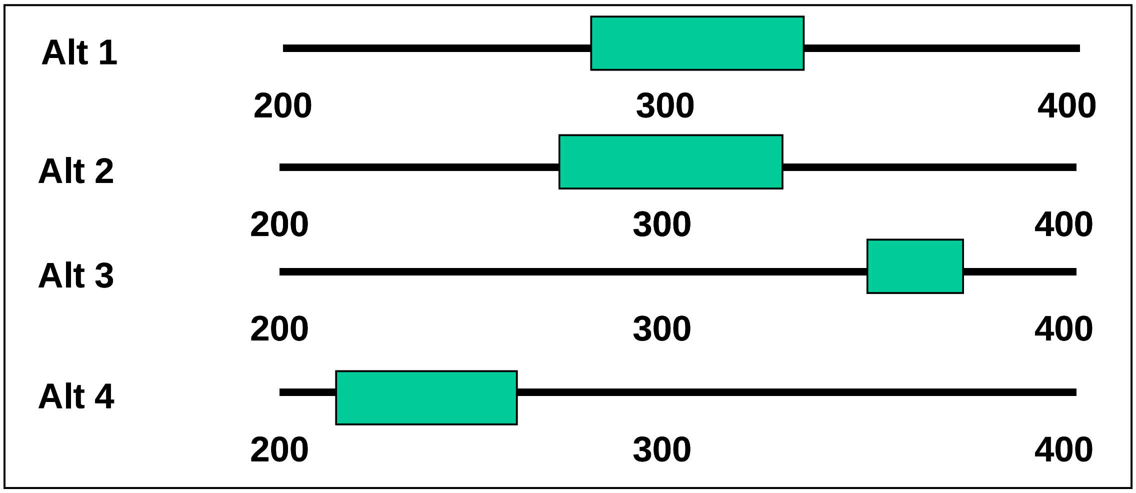

Figure 1. Comparing the Four Alternatives for the Representative MOE Data.

This methodology was executed for each of the remaining 35 MOE's. These results can be presented upon request. The purpose of this work was not to check the work of TRAC-White Sands, but to determine the merits and implementation insights of the CV and bootstrap methods.

V. SUMMARY

Although both the mean and median can be used, the mean appears to be sufficient. There was not a significant difference when comparing alternatives on whether the data was normally distributed or not (assessed using the Kolmogorov-Smirnov Test). The CV was less than .10 in almost 80% of the 144 data sets after five replications. When the bootstrap procedure was done on these five replications, the resulting mean 90% BCa included the 11-replication sample mean in all cases. The CV was less than .20 in over 86% of the 144 data sets after eight replications. When the bootstrap procedure was done on these eight replications, the resulting mean 90% BCa included the 11-replication sample mean in all cases. Approximately 14% of the CV values were greater than .20 (with some greater than .70), but after eight replication, the resulting mean 90% BCa included the 11-replication sample mean in all cases.

If the CV is high for a particular MOE in one alternative, it was found that it is high for all of the alternatives for that MOE. The CV does not significantly change from 5-11 replications. For example, from our MOE example for alternative 1, the CV after five replications was .087 and after 11 replications, the CV was .086. The magnitude of the MOE value does not effect the CV (unitless). For example, the MOE example for Alternative 1 had values ranging from 279 to 352 and had a CV of .087. Another MOE we examined had values ranging from .79 to .831 and had a CV of .017.

If there are available resources, then there is nothing that substitutes for the actual data obtained from executing the simulation. The CV value (especially when paired with a "picture" of the data) appears to be a good measure to determine how many replications are required and does not require normality assumptions. If the FCS KPP simulation runs do require significant resources (mainly time), the bootstrap appears to offer good results after five replications when compared to the 11 replications. Finally, the 90% mean BCa is an excellent analytical and visual tool to show where differences between alternatives exist.

REFERENCES

Cherolis, George. *A Sequential Stopping Rule for Reducing Production Times During CASTFOREM Studies*. Master's Thesis, New Mexico State University, 1992.

Efron, Bradley and Tibshirani, Robert J. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1994.

Law, Averil M. and Kelton, W. David. *Simulation Modeling and Analysis, 3rd Edition*. New York: McGraw Hill Publishers, 2000.