V6.0

A Bayesian Approach to Evaluation of Operational Testing of Land Warfare Systems

By

Lee Dewald, Sr., Ph.D.
Virginia Military Institute
Lexington, VA 24450

Robert Holcomb, Ph.D.
Institute for Defense Analyses
Alexandria, VA 22311

Sam Parry, Ph.D.
Institute for Defense Analyses
Alexandria, VA 22311

Alyson Wilson, Ph.D.
North Carolina State University
Raleigh, NC 27695

Abstract

In this paper we critique perceived deficiencies in current practices in analyzing operational test metrics and demonstrate how a Bayesian approach could be applied to operational tests to improve confidence in the assessment of the test results. We develop an approach and apply it to the analysis of Percent Blue Survivors and Percent Red Killed as univariate measures of effectiveness in operational testing, using historical data.

CURRENT EVALUATION METHODS

Analyses used in the operational testing (OT) of land warfare systems typically rely on traditional statistical methods. This methodology is appropriate for cases where there are adequate samples and no important prior information. Experimental observations are obtained from randomly selected samples from a specified or known population, and the assessment relies on the ability to replicate the experiment an adequate number of times to achieve an appropriate level of confidence for operational effectiveness or suitability.

However, some metrics in operational testing involve force-on-force operations, which rarely meet these conditions. Instead, such operations are characterized by few samples, the availability of significant information from prior testing, and varying test conditions. Force-on-force battles are unique social events, not random samples drawn from some cosmic urn containing all possible battles. There are significant costs that make it very difficult to generate adequate replications of force-on-force battles to achieve meaningful levels of confidence in the metrics of interest.

In a typical operational test program, we are interested in examining force-level measures of merit that result when the land forces employ some new (often prototype) system in place of some older, potentially less capable, system. In the conventional analytical approach, the measure of effectiveness (MOE) is presumed to be a random variable representing true population parameters (e.g., mean and variance for Blue losses in a battle). The true values of these parameters are unknown, and data are collected during the operational test to obtain estimates of these parameters and to quantify the uncertainties in the estimates. Typically the results from one operational test are not used in conjunction with other test data; instead, the data are gathered, analyzed, and reported in a stand-alone manner. The acquisition executive who decides if the new system merits purchase must then deal with a multitude of stand-alone test results or merely consider the most recent.

This paradigm leads us to ask the wrong question. Typically, after an OT event, we ask "What do we know about Metric A, given the data we saw in the operational test?" This is an incomplete question, specifically because it fails to take into account whatever we might know about Metric A *before* the operational test data was analyzed. Typically systems have undergone developmental testing before they enter into the first operational test, and most systems have multiple operational tests before their final acquisition decision is made. Before the first operational test is conducted for a new system, there are constructive simulations conducted to examine and predict unit performance in combat when equipped with the new system. The results of the most recent operational test should inform and refine our state of knowledge about Metric A, not merely examine it as stand-alone event. This is particularly relevant for an acquisition decision that follows many years of developmental and operational testing such as occurs in any major defense acquisition program.

The report of the National Research Council on "Statistics, Testing, and Defense Acquisition" (Cohen et al., 1998) found that the current evaluation methods used during operational testing of combat systems

- Restrict application of statistical techniques
- Prevent integration of all available and relevant information for use in planning and conducting tests, and making production decisions
- Cannot achieve standard statistical measures of confidence for many relevant parameters using reasonable amounts of testing resources

In the next section we will describe aspects of a Bayesian approach to data analysis before we present a new paradigm for evaluation, one we believe supports the improvements the NRC is seeking.

ELEMENTS OF BAYESIAN DATA ANALYSIS

Lynch (2007) makes the following observation about Bayesian statistical methods:

> Put generally, the goal of Bayesian statistics is to represent prior uncertainty about model parameters with a probability distribution and to update this prior uncertainty with current data to produce a posterior distribution for the parameter that contains less uncertainty. This perspective implies a subjective view of probability—probability represents uncertainty—and it contrasts with the classical perspective. From the Bayesian perspective, any quantity for which the true value is uncertain, including model parameters, can be represented with probability distributions. From the classical perspective, however, it is unacceptable to place probability distributions on parameters, because parameters are assumed to be fixed quantities: Only the data are random, and thus, probability distributions can only be used to represent the data (p. 50).

An example of the differences between classical and Bayesian approaches can be found in the approaches to hypothesis testing. In the classical approach, based on statistical sampling, the question examined is, "If the hypothesis is correct, what is the probability of seeing the data that I observed?" In the Bayesian approach the question is, "What is the probability, given the data I observed, that the hypothesis is correct?" One might suspect that acquisition decision-makers believe they are asking the second question, while they are usually being given the answer to the first. There is an extensive history and discussion available on the differences in the two approaches: good summaries are given in Jaynes (1984, 2003) and Sivia (2006). For a recent book on the subject see McGrayne (2011).

When decision-makers are making acquisition decisions, they are acting in the face of some uncertainty, and to reduce that uncertainty, they are seeking relevant, credible evidence that has some degree of inferential weight. Operational test results are one source of that evidence, and an argument for the Bayesian approach is that one can use prior information (such as combining results from several tests) to strengthen the inferential weight of the evidence provided. Establishing a prior probability distribution is key to the Bayesian approach. This is accomplished through the analysis of data sets from such sources as prior operational testing of similar systems or computer simulations.

The Bayesian approach requires the specification of a probability model: specifically, a joint probability density function (pdf) for all observable and unobservable quantities that impact the problem. (See Hoff (2009) for a detailed mathematical treatment.) This model will be consistent with our knowledge of the underlying problem and the data collection process. We designate the pdf as $p(\Theta, Y)$ where Θ is the vector of unknown parameters and Y is a vector of data. By the definition of conditional probability, $p(\Theta, Y) = p(\Theta)p(Y|\Theta)$. $p(\Theta)$ is our prior distribution on the vector Θ, and $p(Y|\Theta)$ represents the distribution of the data given the particular value of the vector Θ.

Next we condition on the observed data, Y, to calculate and interpret the appropriate posterior distribution, $p(\Theta|Y)$, which is the conditional pdf of the unobservables of interest given the observed data, Y. We can think of p(Θ|Y) as summarizing the information we have about Θ after observing Y—to include any information we had from the prior distribution p(Θ).

Using the laws of probability, we write p(Θ|Y) = p(Θ)p(Y|Θ)/p(Y), where

$$p(\mathrm{Y}) = \int_{-\infty}^{+\infty} p(\Theta)p(Y \mid \Theta)d\Theta$$

This expression for p(Θ|Y) is called *Bayes' Theorem.* This formula implies that to quantify our current uncertainty about the unknowns we are modeling, p(Θ|Y), we must specify the information we have before this experiment, p(Θ), and our data distribution, p(Y|Θ).

Finally, as we would in any modeling problem, we evaluate the fit of the model and the implications of the resulting posterior pdf. We ask, "Does the model fit the data? Are the conclusions reasonable? How sensitive are the results to the modeling assumptions?" If necessary we alter or expand the model and repeat the process above. The posterior pdf informs the next prior distribution for any subsequent modeling or testing of similar systems with the inclusion of new data.

A UNIVARIATE MODEL FOR PREDICTING BLUE SURVIORS AND RED CASUALTIES

As an example, we demonstrate the process described in the preceding section by constructing one-parameter models for Blue Survivors and Red Killed using information from a series of war games and operational testing of digitized and non-digitized forces from the 1997 Advanced Warfighting Experiments (AWEs). In these experiments, a Blue force was pitted in mock combat against a similar Red force of brigade-size. The combat took place at the National Training Center (NTC) in the Mohave Desert in California, and included a laser tag system to accurately record shots and kills by the opposing sides. The data used in the subsequent examples was taken from the results of several of these mock combats.

The Blue forces in these examples were equipped with a new digitized computer screen coupled with a Global Positioning System sensor that displayed to every Blue vehicle the location on a map background of all the other Blue systems. This system provided situational awareness to the Blue forces, and was expected to improve the outcome of the battles. The Red forces were not similarly equipped. In the actual operational assessment, the data collected from these experiments were also compared to a baseline non-digitized Blue force, but for illustration, we will only focus on understanding the performance parameters of the digitized force.

Prior to the execution of the digitized Blue force exercises, a constructive computer simulation, using a model called CASTFOREM, was run to provide a prediction for how the digitized Blue force might perform in the field. CASTFOREM is a high-resolution brigade-level constructive simulation developed by TRADOC Analysis Center used for Army Analysis of Alternatives until replaced by COMBAT XXI around 2010. After the first two field experiments were conducted, additional model runs were made using JANUS, a high-resolution brigade-level interactive simulation developed by Lawrence Livermore National Laboratory (LLNL) and the Army and used by the TRADOC Analysis Center for analysis and training since the early 1980's.

We start our analysis by considering data from pre-AWE CASTFOREM runs made in January 1997. Hundreds of runs were made under a variety of conditions; fifty runs were made for each specific condition. Unfortunately, we do not have the complete data from these runs, but only summary statistics—specifically, the mean and standard deviation. To develop our prior distribution, we use the statistics from relevant runs: specifically, we consider summaries of

Percent Blue Survivors (PBS) and Percent Red Killed (PRK) from the 50 runs of the Deliberate Attack scenario for the digitized brigade. The mean PBS was 0.700, with a standard deviation of 0.06417; the mean PRK was 0.380, with a standard deviation of 0.06797.

From the previous section, to perform the Bayesian analysis, we need to specify Y, Θ, p(Y|Θ), the distribution of the data given the particular value of the vector Θ, and p(Θ), the prior distribution for the parameter vector. In our example, we are performing two parallel analyses, and we will use subscripts to index them. Consider first the analysis of PBS. The data that we will observe, Y_B, are the number of Blue Survivors in a battle. (With parallel notation, Y_R is the number of Red Killed in a battle.) The unknown parameter is θ_B, the probability that an individual Blue vehicle survives the battle. (θ_R is the probability that an individual Red vehicle is killed in the battle.)

For this analysis, we assume that each Blue vehicle has the same probability, θ_B, of surviving a battle, and similarly that each Red vehicle has the same probability, θ_R, of being killed. This is a simplifying assumption, but one that is frequently made. Given this assumption, the goal of our analysis is to estimate θ_B and θ_R (the proportion of Blue Survivors and Red Killed) and quantify our uncertainty about these estimates.

With these assumptions, $p(Y_B|\theta_B)$ is Binomial(n_B, θ_B), where n_B is the original number of vehicles in the battle. The mathematical expression is

$$p(Y_B|\theta_B) = \binom{n_B}{y_B} {\theta_B}^{y_B}(1-\theta_B)^{n_B-y_B}$$

For the Bayesian analysis, we also need to specify our prior information, $p(\theta_B)$, about the probability of a Blue vehicle surviving. In particular, $p(\theta_B)$ must be a probability density function. A common method for specifying a prior distribution is to select a flexible distribution family and then use prior knowledge to specify the parameters of the family.

In this case, we will use the Beta distribution. Beta distributions specify probabilities on the interval (0,1), which is what we need since we are capturing information about a proportion. The mathematical expression is

$$p(\theta_B) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} {\theta_B}^{\alpha-1}(1-\theta_B)^{\beta-1}$$

To specify the prior distribution, we need to choose α and β.

We use the information from the CASTFOREM runs to choose α and β. One method commonly used to choose parameters is called *moment matching*. The mean proportion of Blue Survivors from CASTFOREM was 0.700, and the standard deviation was 0.06417. Suppose that we choose α and β so that the mean of the beta distribution is 0.700 and the standard deviation is 0.06417, or as close as we can come given the mathematical expressions. In particular, the mean and standard deviation of a beta distribution are functions of α and β:

$$E[\theta_B] = \frac{\alpha}{\alpha + \beta}$$

$$SD[\theta_B] = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

Solving, we find $\alpha = 35$ and $\beta = 15$. This distribution is plotted in Figure 1.

We can use the Beta(35,15) distribution as the prior distribution for θ_B, or we could adjust it. For example, we may believe that combat simulations tend to underestimate the variability of Blue Survivors as compared to actual combat situations. In that case, we might choose to change the parameters of the prior distribution to have the same mean, but a larger standard deviation. Choosing a prior distribution is not an "analyst free" process. Careful consideration must be given to be sure that the data being used to develop the prior are relevant and valid, and they must be carefully modeled in light of the field test/operational test data.

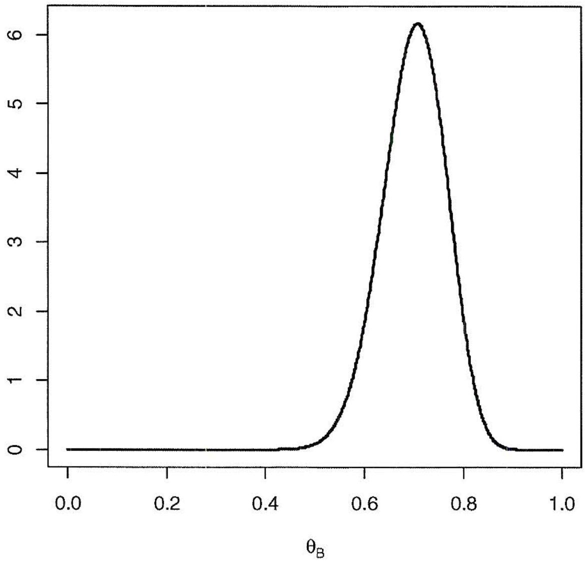

Figure 1. Beta(35,15) distribution.

For the purposes of our analysis, we will summarize our prior information from the CASTFOREM simulations with the Beta(35,15) distribution. Similar arguments lead to a prior distribution for θ_R of Beta(19,31).

Now we want to update our knowledge about θ_B and θ_R using data from our first field test. The first AWE field test was a Blue deliberate attack conducted on 18 March 1997 as part of an AWE Heavy Brigade rotation at the National Training Center. In this battle, we observed 70 of 166 Blue vehicles survive and 60 of 112 Red vehicles killed.

We specified the initial prior distributions on our probability of Blue surviving (Red being killed) using beta distributions in order to take advantage of the conjugate prior relationship and mathematical convenience of the beta distribution to the binomial distribution. Lynch (2007) or Hoff (2009) discuss this property in detail. The simple updating rule when there is a beta prior and a binomial data model is

$$\text{Prior} = \text{Beta}(\alpha,\beta)$$

$$\text{Observed count } (y) = \text{Binomial}(n,\theta)$$

$$\text{Posterior} = \text{Beta}(\alpha + y, \beta + n - y)$$

Given our data and our prior, the posterior distribution for PBS is Beta(105,111) and for PRK is (79,83). The posterior mean for PBS, which is used as an estimate of θ_B is 0.486 with posterior standard deviation 0.03393. These values are calculated from our earlier expressions for $E[\theta_B]$ and $SD[\theta_B]$. (For θ_R, the posterior mean is 0.488 with posterior standard deviation 0.03915.) The mean for PBS changed from 0.7000 in the CASTFOREM runs to 0.486 in the posterior distribution after incorporating the initial field test data, and the new standard deviation for PBS is 0.03393, whereas the original value in the prior distribution was 0.06417. The prior and posterior distributions for PBS are plotted in Figure 2, with the prior given as the dotted line and the posterior as the solid line.

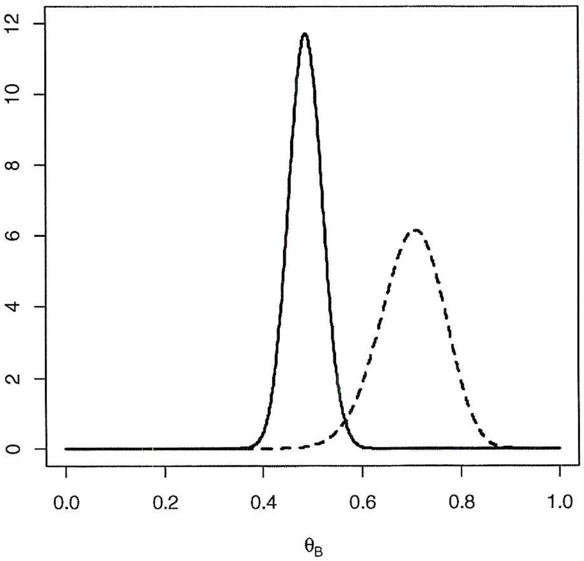

Figure 2. Prior (dotted) and posterior distributions (solid) using CASTFOREM and first AWE.

The second Blue offense AWE field test was a Blue Hasty Attack conducted on 25 March 1997 as part of an AWE Heavy Brigade rotation at the National Training Center. In this battle, we observed 90 of 161 Blue vehicles survive and 130 of 257 Red vehicles killed. If we assume that the data from the second field test have the same probabilities of survival for Blue and Red forces as the first field test, then we can simply use our posterior distribution from the previous analysis, add the additional data, and get our new posterior distribution—essentially, we are simply using the second battle to increase our sample size. Making these assumptions, the posterior distribution for PBS is Beta(195,182), with mean 0.517 and standard deviation 0.02570. The posterior distribution for PRK is Beta(209,210), with mean 0.499 and standard deviation 0.02440.

We recognize that these assumptions are almost certainly invalid, and we make them only to illustrate the simplest possible Bayesian analysis. However, as discussed when specifying prior distributions, combining information to improve prediction and confidence in a sensible way is not an "analyst-free" process. If these assumptions are invalid in practice, then we specify more mathematically complex data models and prior distributions. The sequence of specify prior, update using data to posterior distribution, consider more information, update again remains the same, although the mathematical details will differ.

At the conclusion of the 1997 AWE field tests, additional model runs were conducted using the JANUS model for a Blue Brigade Hasty Attack. We have summary statistics for 40 replicated JANUS runs. For the 40 replications, the starting Blue force size was 175, and the starting Red force size was 172. There were, on average, 92 Blue casualties (83 survivors) and 112 Red casualties. The standard deviations for Blue Survivors was 12.61; for Red casualties, 15.35.

Serious consideration must be given to what we want to do with this data. Do we want to treat each simulated vehicle as "the same" as a vehicle in the AWE field exercise? If they are not the same, what information do we have about how they differ? Are the simulation runs relevant to our assessment? How? The answers to these questions determine the specifics of the Bayesian analysis. Making the assumption that combatants in the simulation are the same as those in the field tests, we update our posterior distribution again Beta(278,274), with mean 0.504 and standard deviation 0.02126 for PBS. The posterior distribution for PRK is Beta(321,270), with mean 0.543 and standard deviation 0.02440.

The final field test data used for this example are from a Blue Deliberate Attack in the Division Capstone Exercise (DCX) after the AWE 1997 tests at the NTC. We observed 16 of 133 Blue vehicles survived and 23 of 183 Red vehicles killed. Given these observations, this battle is clearly relevant to our assessment, but we choose not to give each observation the same weight as in previous experiments. The reason for this is that the digitized systems experienced problems during the exercise reducing the potential credibility of the field test data. Thus, we choose to multiply both the sample size and the results by 0.2. This gives us the same mean, but decreases how much we learn from this experiment. In this case, we have chosen a specific weighting; however, using more complex statistical models, the weight can be chosen based on data (Reese et al. (2004); Ibrahim and Chen (2000); Anderson-Cook et al. (2007)).

The prior distribution (posterior from our previous analyses) for PBS is Beta(278,274). We update with 16*0.2 Blue vehicles survived of 133*0.2 total to find a posterior distribution of Beta(278 + 3.2,274 + 23.4) = Beta(281.2, 297.4), which has mean 0.486 and standard deviation 0.0208. A similar computation shows a posterior distribution of Beta(325.6, 302) for PRK, with mean 0.519 and standard deviation 0.0199. In Figure 3, we plot the prior and four updates leading to four posterior distributions.

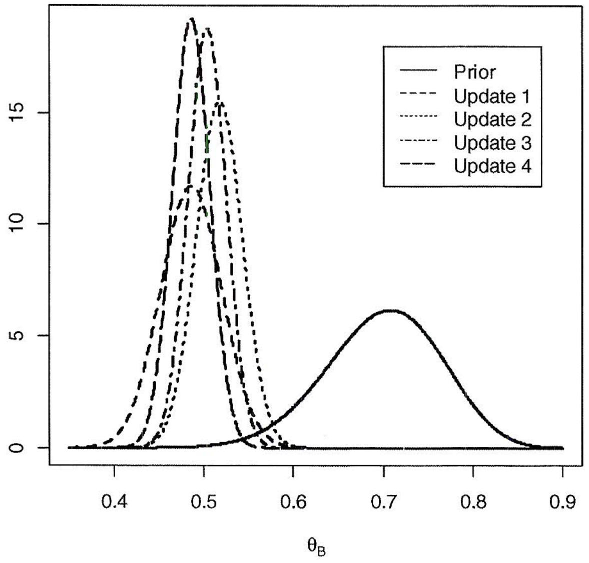

Figure 3. Prior and four updates for posterior distributions.

Additional summaries of each stage of the Bayesian process are shown in Figures 4 and 5. The means of the distributions of Blue Survivors and Red Killed are shown in Figure 4 after each subsequent data set is applied to the Bayesian process. The standard deviations of the distributions of Blue Survivors and Red Killed are shown in Figure 5 after each subsequent data set is analyzed using the Bayesian process.

DISCUSSION

This example illustrates a very simple Bayesian analysis for combining information from field tests and simulations. The example is not intended to illustrate how to implement the various models that could be developed, but instead to illustrate that mathematical techniques exist for combining information and that careful analyst input must be applied throughout the process. Adding additional relevant information in a principled way often improves the precision of the estimates that we can make about quantities of interest.

As was discussed, the assumptions about Blue Survivors and Red Killed are questionable. Likewise there has been no accounting for the obvious dependency between these two univariate measures: the number of Blue Survivors and the number of Red Killed are not independent random variables. One commonly used univariate measure that by its definition accounts for this dependency is the Loss Exchange Ratio (LER) which is the ratio of Red to Blue losses. In future work, we will consider LER, Number of Red Lost/ Number of Blue Lost. Early work by Olwell (1997) showed that LER can be modeled very well by the Inverse Gaussian distribution which is well described in a monograph by Chikara and Folks (1989).

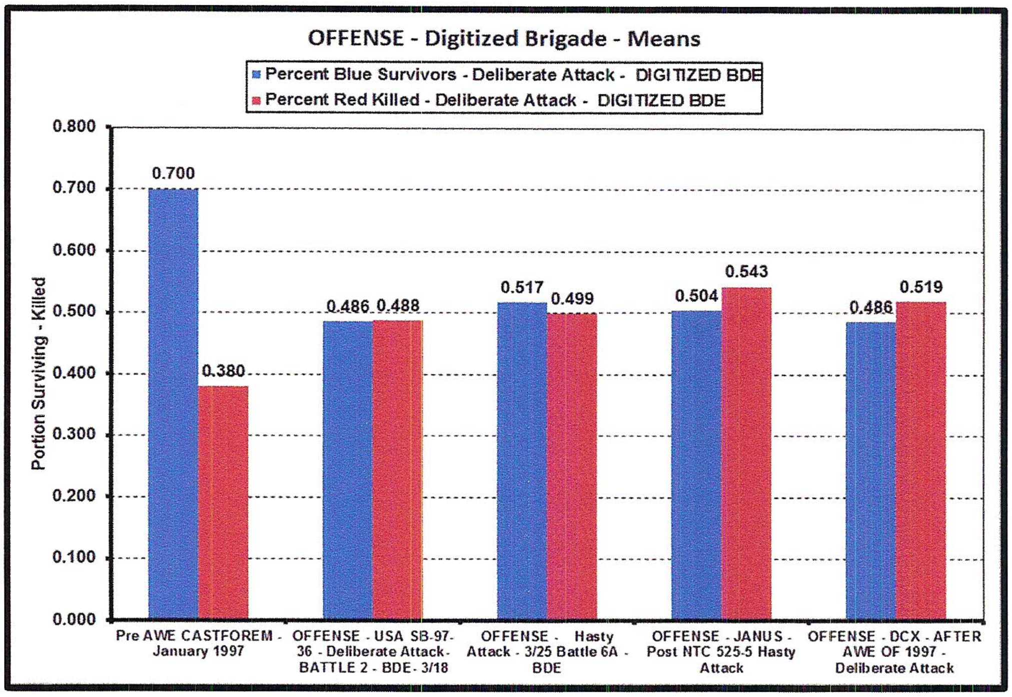

Figure 4: Means of Posterior Distributions

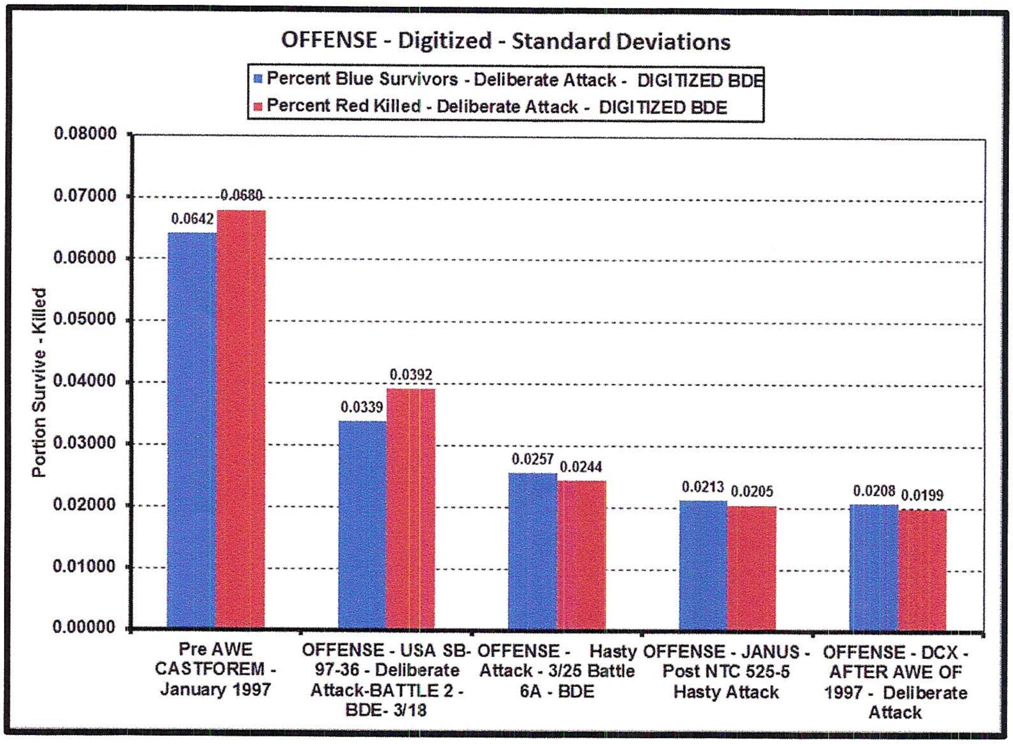

Figure 5: Standard Deviations of Posterior Distributions

REFERENCES

Anderson-Cook, C., T. Graves, M. Hamada, N. Hengartner, V. Johnson, C. S. Reese, A. Wilson (2007). Bayesian Stockpile Reliability Methodology for Complex Systems with Application to a Munitions Stockpile. *Journal of the Military Operations Research Society* 12(2): 25-38.

Chikara, R.S. and J.L. Folks. *The Inverse Gaussian Distribution.* Marcel Dekker: New York, 1989.

Cohen, Michael L., Rolph, John E., Steffey, Duane L., (eds), *Statistics, Testing and Defense Acquisition: New Approaches and Methodological Improvements*, National Academy Press: Washington, D.C., 1998.

Hoff, Peter, *A First Course in Bayesian Statistical Methods*, Springer: New York, 2009.

Ibrahim, J. G. and M-H. Chen (2000). Power Prior Distributions for Regression Models. *Statistical Science* 15(1): 46-60.

Jaynes, E.T., *Bayesian Methods: General Background*, an address presented at the Fourth Annual workshop on Bayesian/Maximum Entropy Methods, University of Calgary, August 1984.

Jaynes, E.T. *Probability Theory: The Logic of Science*, Cambridge University Press: Cambridge, UK, 2003.

Lynch, Scott M., *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, Springer: New York, 2007.

McGrayne, Sharon B., *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press: New Haven CT, 2011.

Olwell, D. H. (1997). Modeling Loss Exchange Ratios as Inverse Gaussian Variates: Implications. *Military Operations Research* 3(1): 51-67.

Reese, C. S., A. G. Wilson, M. S. Hamada, H. F. Martz, K. J. Ryan (2004). Integrated Analysis of Computer and Physical Experiments. *Technometrics* 46(2): 153-164.

Sivia, D.S. with Skilling, J. *Data Analysis: A Bayesian Tutorial, 2nd Edition,* Oxford University Press: Oxford, U.K., 2006.