Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

A Bayesian Approach to Evaluation of Land Warfare Systems

Lee S. Dewald, Sr., Ph.D.
Robert Holcomb, Ph.D.
Sam Parry, Ph.D.
Alyson G. Wilson, Ph.D.

24 October 2012

Approved for public release; distribution is unlimited.

Land Warfare Problem Space

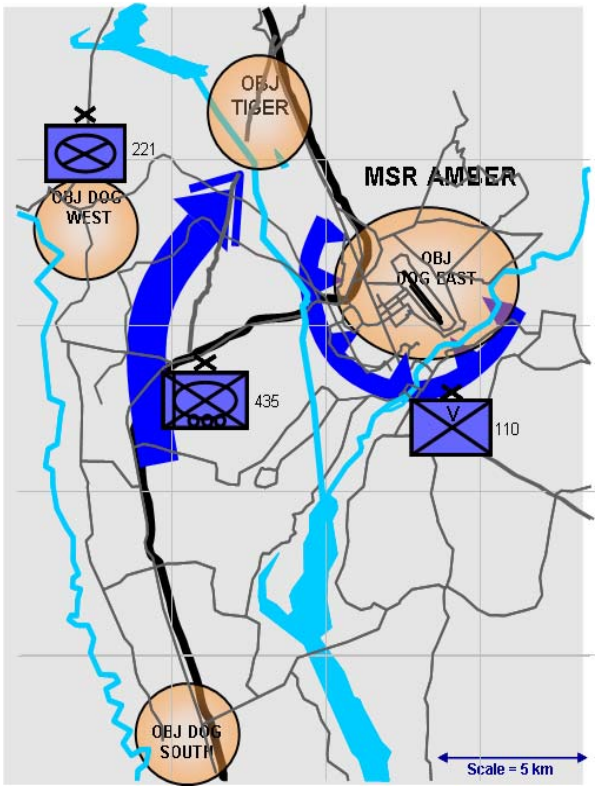
- The Land Warfare group in IDA evaluates the employment of Blue and Red combined arms brigades and battalions in realistic settings of mock combat for the purpose of evaluating weapons systems, tactics or operational concepts

- Several Measures of Effectiveness (MOEs) have been used to measure results in operational testing:
  - Number of Enemy (Red) casualties
  - Number of Friendly (Blue) casualties
  - Loss Exchange Ratio (LER)

- Historically, the evaluations considered each test event independently, for purposes of calculating sample sizes and establishing confidence intervals on relevant metrics

- By employing a Bayesian approach we hope to demonstrate that we can treat an overall test program in a cumulative manner.

- Basic objective is to use prior test events (or modeling, if testing has not yet begun) to provide insight into distribution of the parameter of interest



Approved for public release; distribution is unlimited.

Hypothesis, Objective and Data Sources

- Study Hypothesis: Use of the Bayesian approach provides cumulative information for design of the next field test which will increase the usability of data, provide more robust estimates of MOEs and reduce test costs

- Study Objective: Utilize data from previous field tests and simulations sequentially, to use prior information about MOE distributions to inform MOE distributions as new data becomes available

- Study Data Sources
  - Combat Simulations
    - CASTFOREM – a high resolution Brigade Level Constructive Simulation developed by TRADOC Analysis Center used for Army AoA's until replaced by COMBAT XXI around 2010.
    - JANUS – a high resolution Brigade Level Interactive Simulation developed by LLNL and the Army and used by TRAC for analysis and training since the early 1980's.
    - JCATS – a high resolution Brigade Level Constructive/Interactive Simulation developed by LLNL from JANUS in the mid-1990's to focus on Urban Combat. Currently used by TRADOC Schools and 23 Allied nations for air-land-sea analysis.
  - Field Tests/Exercises
    - A sequence of battles at the Army Warfighting Experiments (AWE) conducted at the National Training Center (NTC) in 1996-1997 to investigate the benefits of digitization to the Army
    - Division Capstone Exercise (DCX) conducted at the NTC after the AWE 1997 exercises were concluded.

Approved for public release; distribution is unlimited.

IDA

Evaluation Methods

- Operational tests involving land warfare combat systems are typically characterized by:
  - Many parameters and uncontrolled variables
  - Relatively few replications and observations available
  - Significant information from earlier operational testing or Analysis of Alternatives simulations usually available
  - Force-on-force battles are unique and dynamic events, not a random sample from a population of all battles
  - Cost and time frequently restrains the number of replications, impacting confidence and power
- The classical evaluation methodology is constrained by test resources and classical statistical methods
  - No important prior information normally carried forward from test event to test event
  - Number of replicates to gain high confidence and power are expensive to conduct
- We believe the Bayesian approach is more appropriate for land warfare system evaluations because it fits the circumstances of continuous evaluations, multiple tests, accumulation of evidence over years of a complex test program, and relationship to a prior based on reasonable estimation rather than dealing with single point estimates

Approved for public release; distribution is unlimited.

The Bayesian Approach

- **Hypothesis**: Design and analysis of a field test will be improved when all available relevant and credible information from previous simulations, developmental tests, and operational tests is considered

- **Current testing process**
  - Each test event is treated as a stand-alone item
  - Attempt to create enough trials to get an adequate sample size (generally not possible due to number of variables involved)
  - Done first with Limited User Tests, then with IOT&E, generally not incorporating previous AoA modeling results

- **Proposed Bayesian process**
  - Begin with AoA/simulation results done before first field testing event to form a distribution on the MOE of interest
  - Results of the first field test are combined with the simulation distribution which serves as the basis for a more refined estimate of the distribution of the MOE of interest
  - Each subsequent field test continues to refine the estimate of the distribution of the MOE

- **Potential Benefits**
  - When a probability interval for a specified MOE reaches an acceptable threshold, future tests can be limited to verification of the parameter rather than discovery
  - The test design for the next test may be significantly simplified (and less costly) based on the remaining MOE that must be tested

Approved for public release; distribution is unlimited.

IDA

Our Approach In the Examples That Follow

- The study approach was to take existing data from past events and demonstrate that, had we used Bayesian techniques at the time, our estimates for a selected MOE would have smaller variance than otherwise as we progressed from event to event
- We began with a simple univariate example, then moved to a more complex MOE and a hierarchical model
- Simple example was to examine Red Kills and Blue Losses from initial combat simulations, then progress to data from field tests done during the Advanced Warfighting Experiments of the late 1990's
- More complex example involved Loss Exchange Ratios, acknowledging that Red Kills and Blue Losses were not independent of one another

Approved for public release; distribution is unlimited.

IDA

Simple Univariate Model For Casualties

- GIVEN: Starting force of size n from AWE Field Test of Blue Attack using digitized forces. Examined Red kills and Blue losses separately

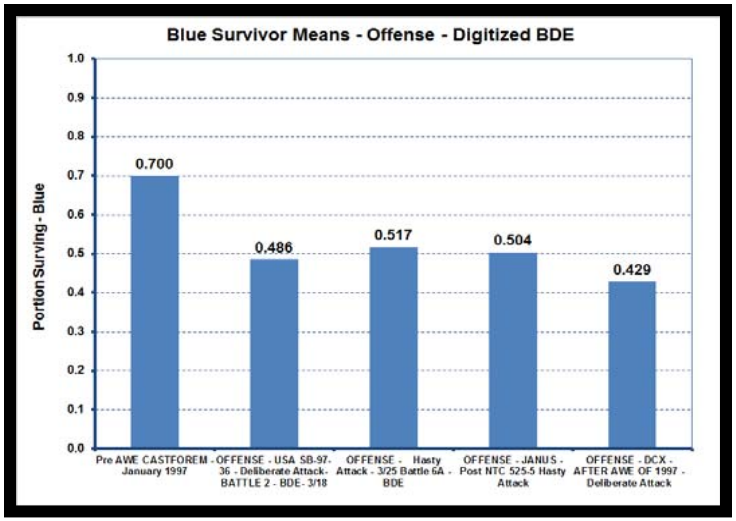
- ASSUMPTIONS:
  - Each kill (or loss) is an independent Bernoulli Trial with probability of kill = p
  - X, number of kills (or losses) in AWE, has a Binomial distribution (n,p) given a value for p.
  - Prior distribution on p is Beta (a,b) obtained from CASTOREM model run done before field trials began

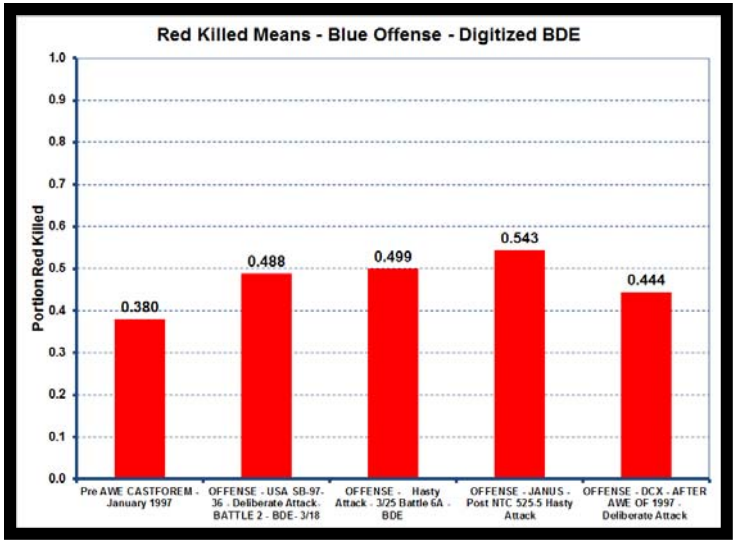
- RESULTS:
  - Posterior distribution on p is Beta (A,B) where A = a + X and B = b + n – X
  - E(p) = A/(A+B) and Var(p) = AB/[(A+B)^2(A+B+1)]
  - E(p) is weighted average of mean of prior and X/n

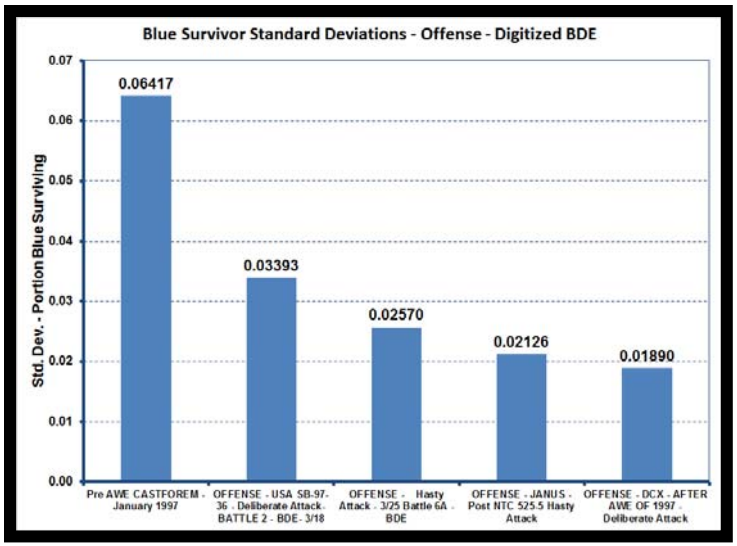
Approved for public release; distribution is unlimited.

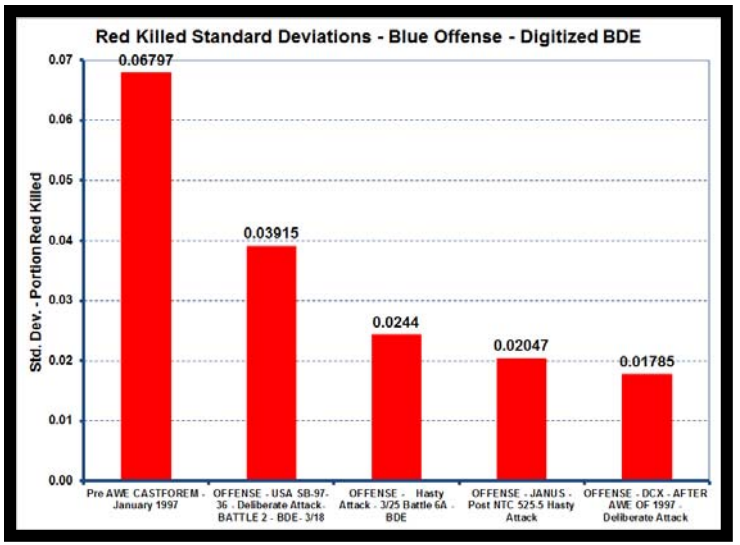
IDA

Summary of Posterior Distributions from Univariate Cases





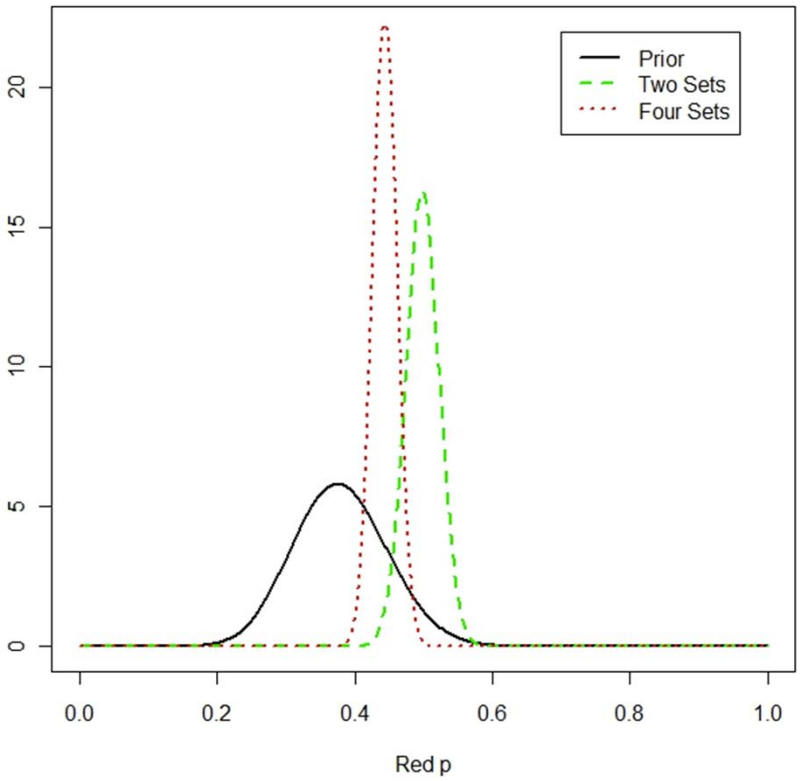


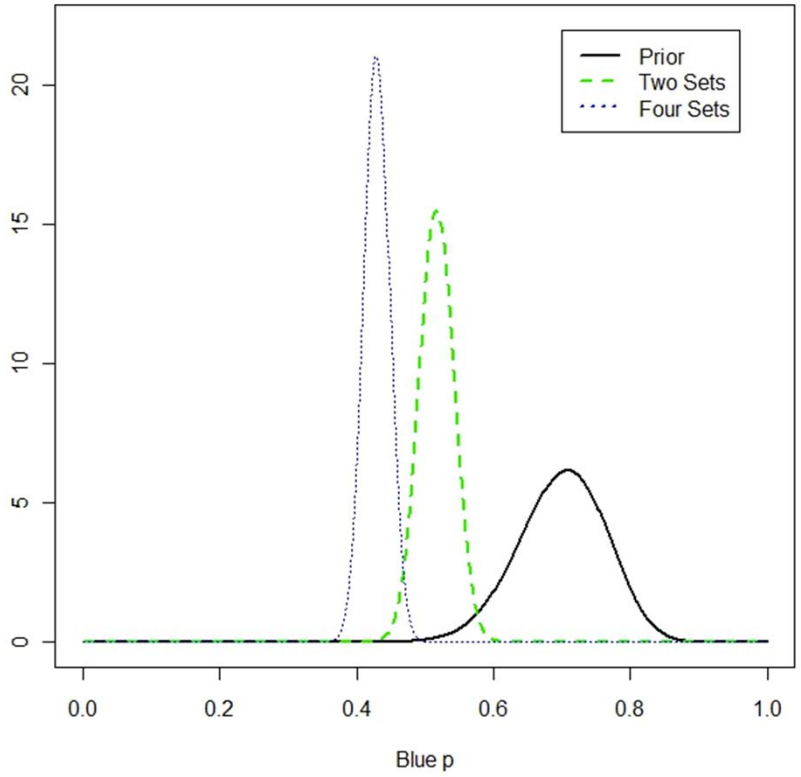


Approved for public release; distribution is unlimited.

IDA

Shrinkage of Variability of Estimates

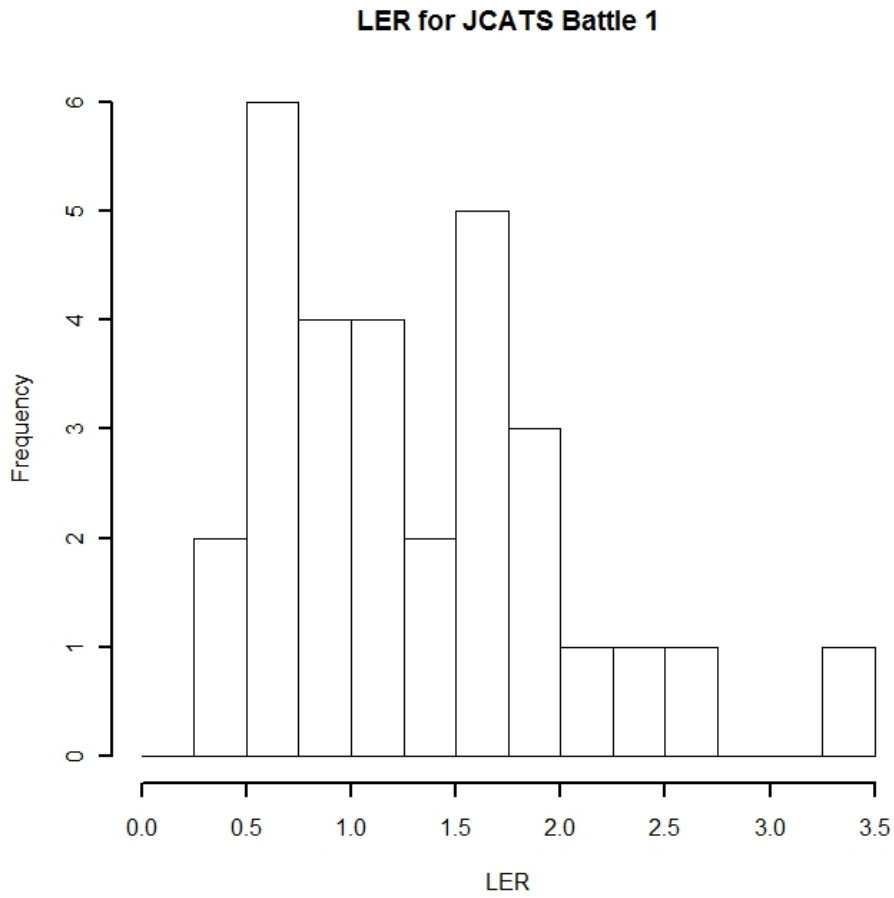




Approved for public release; distribution is unlimited.

A More Complex Example

- We recognized that some MOEs are not simple Bernoulli trials, and so examined the more complex MOE of Loss Exchange Ratio, to capture the inter-relationship between Red kills and Blue losses
- These data represent the LER of one Blue offensive battle in JCATS replicated 30 times
- Note skewed nature of distribution and high degree of variability in LER obtained from each replication
- A hierarchical model was developed for the MOE of LER



Approved for public release; distribution is unlimited.

Data for Hierarchical Model

Non-Digitized Brigade

- Blue movement to contact
  - 2 CASTFOREM runs, 40 replications each, mean and standard deviation recorded
  - 3 field trials, LERs recorded
- Blue attack
  - 8 JCATS simulations, 30 replications each, LER recorded for each
  - 1 CASTFOREM run, 40 replications, mean and standard deviation recorded
  - 1 JANUS run, 40 replications, mean and standard deviation recorded
  - 2 field trials, LERs recorded
- Blue defense
  - 8 JCATS simulations, 30 replications each, LER recorded for each
  - 3 field trials, LER recorded

Approved for public release; distribution is unlimited.

IDA

Hierarchical Model

Assumptions

• For the purposes of this analysis, we assume that either the actual LERs or appropriate sufficient statistics were recorded for the CASTFOREM and JANUS runs. (We simulated data for this analysis.)

• We ignore the battle types. (This assumption can be relaxed by considering a hierarchical model using a Dirichlet process prior.)

Note

• The JCATS simulations have more variability than the JANUS/CASTFOREM simulations. JCATS uses the latest sensor algorithms (new AMSAA ACQUIRE Targeting Task Performance Metric) which are more sensitive to small differences in range, target exposed, area, and sensor FOV.

• We will model the simulation data, generate a predictive distribution for the field trials, and compare the observed field trials to this predictive distribution.

Approved for public release; distribution is unlimited.

IDA

Hierarchical Model Specification

Model

- For JCATS and JANUS/CASTFOREM, we model the LERs from each set of simulation runs as i.i.d. Gamma(α_i,β_i)
  - We think that each set of simulations is different, but that their results are related

- Let $\mu_i = \alpha_i/\beta_i$ and $\nu_i = \alpha_i/\beta_i^2$
  - μ and ν are the mean and variance of the gamma distribution

- For all of the sets of JCATS runs, we assume that the means and variances come from common distributions

$$\mu_i \sim \text{Gamma}(\gamma_\mu,\delta_\mu) \text{ and}$$
$$\nu_i \sim \text{Gamma}(\gamma_\nu,\delta_\nu)$$

Approved for public release; distribution is unlimited.

IDA

Hierarchical Model Specification

Model

• For all of the JANUS/CASTFOREM runs, we will assume that the means come from the same common distribution as the JCATS runs

$$\mu_i \sim \text{Gamma}(\gamma_\mu, \delta_\mu)$$

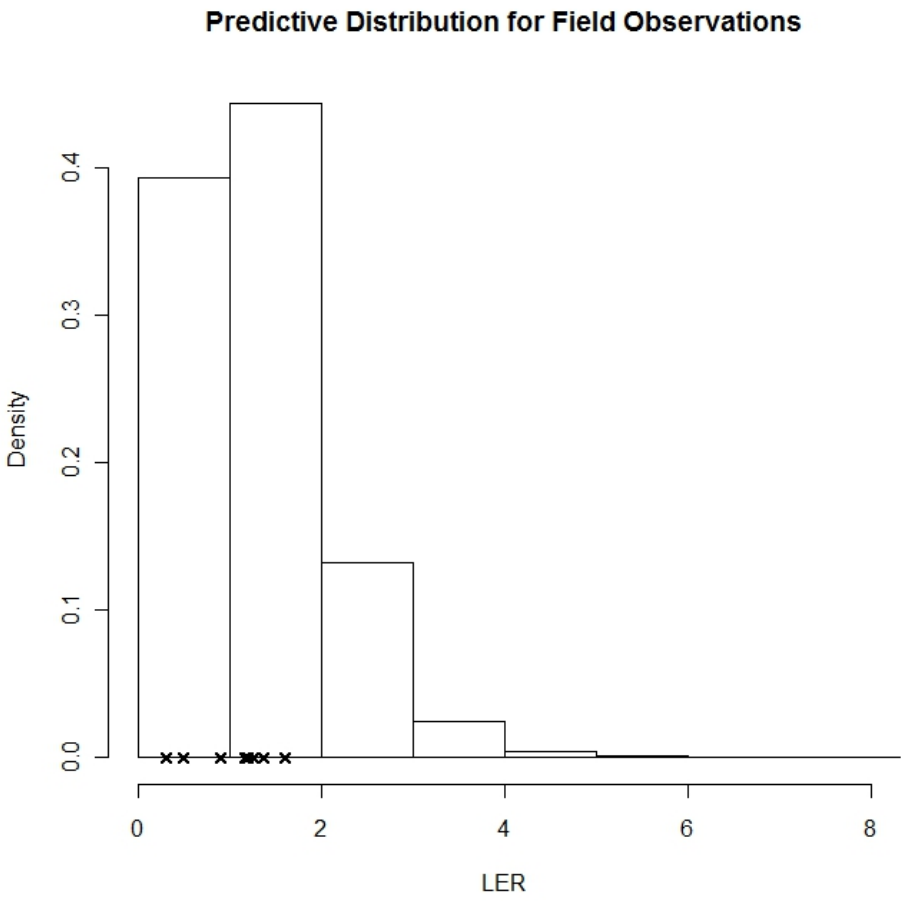
• However, we think that the JANUS/CASTFOREM runs have a different variance than JCATS, so we choose a different distribution to describe their relationship

$$\nu_i \sim \text{Gamma}(\upsilon^2\gamma_\nu, \upsilon\delta_\nu)$$

• We use this parameterization so that we can understand the relationship between the two variances

Approved for public release; distribution is unlimited.

Hierarchical Model Results



- A 90% credible interval for where we expect to see the field observations is (0.18, 2.7)
- The field trial data points are plotted as “X’s” on the predictive distribution. The actual min and max we observed in the field trials are 0.308 and 1.603
- The min and max from the JCATS simulations are 0.101 and 7.667

Approved for public release; distribution is unlimited.

Summary and Conclusions (1)

- By using Bayesian prior to refine the estimate for the MOEs chosen, we demonstrated that the variance around the mean of the metric of interest steadily decreased
- We observed this in both the Red and the Blue cases, and in both the simple and more complex MOEs
- The benefits from combining prior data are two:
  - Confidence in the value of the mean (likelihood) of the MOE of interest (e.g., LER) increases with each new data set when the variability of each data set is not large
    - Further testing and measurement of the MOE may not be required in subsequent tests
    - May significantly reduce the testing cost
  - Or, if there is large variability of the MOE in each data set which inhibits convergence, then
    - Cause of the variability in prior tests or model runs can be analyzed
    - These causes can be taken into account in the design of the next field test
    - Without this knowledge from Bayesian analyses, we would have no idea about the distribution of the MOE from the next field test

Approved for public release; distribution is unlimited.

IDA

Summary and Conclusions (2)

- In all cases, the use of the Bayesian approach provides vital information for design of the next field test which will increase the usability of data and potentially reduce the cost
- We don't get something for nothing: this requires careful analysis and identification of relevant information
- Conducting the analysis requires sufficient resolution in the data so that combining (and decisions about whether to combine) can be made
- This approach can create a linkage between the analysis community which performs AoAs and the test community, and help to insure common metrics, a common viewpoint of the benefits of a new system and cross-talk about requirements and performance that should prove beneficial to both communities

Approved for public release; distribution is unlimited.