The DASE Axioms: Designing Simulation Experiments for Verifying Performance of Software-Intensive Systems

Terril N Hurst
Isaac M Goodrich
Carin E Leigh
Colin F Pouchet
Matthew A Tolman
Marc A Wynn
Jonathan T Zink

October 2016
Background

- During the past decade, Raytheon engineers have collaborated deeply with DoD to establish rigorous methods for applying and expanding Design-of-Experiment (DOE) principles when the sample source is modeling & simulation (M&S)
- The resultant protocol is called DASE: Design & Analysis of Simulation Experiments

Significant innovation was required to apply DASE to System Performance Verification, a Category-1 DASE objective that requires inference spanning the system’s full operational space
Why we conduct simulation experiments: Four categories, each with a different focus*

1. **Evaluate/compare system(s’) performance across a factor space**
   a) Establish **summary statistic(s)** across scenarios and/or alternative systems
   b) Isolate outliers to be diagnosed using experiments in the other categories

2. **Explore a specific system’s design space**
   a) Perform local sensitivity analysis and/or design optimization
   b) Create trustworthy surrogate models for well-defined purposes

3. **Support tests (e.g., Bench Top, HWIL, Captive Carry, Flight)**
   a) Assist test scenario allocation (i.e., which cases to test)
   b) Support pre-test activities (e.g., Shot-Box, Range Safety Review)
   c) Conduct post-test re-construction & data analysis (e.g., Failure Review Board)

4. **Verify & validate the simulation (SimV&V)**
   a) Check assumptions and implementation of models & simulations
   b) Compare simulation results with real-world data from tests

*It is critical **not** to design one experiment spanning categories; otherwise, the inevitable result is confusion & frustration.
Steps in the DASE process

*Plan, execute, and report results accordingly*

1. **Establish Basis** (sponsor, req’ts, SMEs, credible sim/tools, …, *time!* )
2. **State this experiment’s quantifiably specific Objective & Category** (1 - 4)
3. **Define measured Response(s) & practically Discernible Difference(s) \( \delta \)**
4. **Define the experiment’s Factor Space:**
   a) Control Factor set \( X_C \): type (numeric/categorical), units, and ranges/levels
   b) Uncertainty Factors set \( X_U \): type, units, distribution types & parameter values
   c) Constants: List critical simulation inputs, including any screened Control Factors
5. **Screen Control Factors** \( X_C \) and/or inadmissible \( X_C \) treatments
   a) Select experimental design – \( X_C \) treatments
   b) Set number of replicates – random \( X_U \) factor draws – for each \( X_C \) treatment
   c) Establish simulation run sequence, and execute & analyze the Screening runs
   d) Select \( X_C \) factors / treatments to be held fixed (eliminated—i.e., moved to Table 4c)
6. **Sample for empirical modeling** (“the main DOE”)
   a) Select model type & form—e.g., summary statistic(s), (non)linear regression, logistic, tree
   b) Select experimental design – \( X_C \) treatments – e.g., Latin hypercube sampling
   c) Set number of replicates – random \( X_U \) factor draws – for each \( X_C \) treatment
   d) Establish simulation run sequence, and execute/analyze the Modeling runs
7. **Analyze & present Results**—in the following order:
   a) Look at the data (scatterplots, time series, etc.)
   b) Aggregate the data (e.g., histograms, box plots, etc.)
   c) Only after 7a & 7b, compute & test summary statistics and/or model coefficients & residuals
   d) Decide action, including whether follow-on experiments will be required for decision-making
Presentation Contents

1. Background / Introduction (just completed)

2. Report DASE Lessons-Learned, presented as 11 axioms and one theorem

3. Demonstrate implications & consequences of the DASE axioms for 3 levels of demanded statistical rigor

4. Offer pragmatic recommendations for applying DASE to verify performance of software-intensive systems

More detail is found in the white paper and in the references on the final 2 slides
Language & terminology

- When discussing DASE / DOE, it is critical to distinguish between terms regarding populations vs. samples
  - “Population” terms are denoted using Greek symbols—e.g., moments ($\mu$, $\sigma^2$, ...) and median $\tilde{\mu}$ of random variable $X$ (note: binomial (pass/fail) parameter $\pi$ often replaces $\mu$ in what follows)
  - “Sample” terms are denoted using Latin symbols or Greek symbols under a bar or caret—e.g.,
    - Unbiased estimator $\bar{X}$ of population mean $\mu$; sample variance $S^2 = \hat{\sigma}^2$
    - Summary statistic(s), e.g., $\hat{\mu}$, $x_q$; sample-proportion estimate of $\pi$: $p = (#\text{ successes}) \div (#\text{ attempts})$; (model coefficients $\hat{\beta}_{ij}$ not covered)

- The relation $M = QN$ refers to the QN Allocation Problem: How best to allocate $M$ runs between $N X_C$ hypercube scenarios (“treatments”), and $Q X_U$ replicates randomly drawn per $X_C$ scenario

- For Performance Verification, we must also distinguish between bin-level parameters or estimators (e.g., $\mu_\pi$, $\bar{X}_{bin}$), vs. scenario-level parameters or estimators (e.g., $\pi_i$, $p_i$, $i = 1$ to $n$ scenarios)
Axioms related to DASE Step 1 (Basis)

**Axiom 1:** The *Performance Specification* consists of requirements that are stated in terms of verifiable population parameters, and the *Performance Verification Plan* spells out in detail how sampling will occur in order to collect data for estimating the population parameters.

**Axiom 2:** No sample of simulation runs should be regarded as perfectly representing actual performance of the system being simulated.

**Axiom 3:** Two weeks is sufficiently short for executing a full set of performance-specification runs.

This axiom sets allotted run-size $M$. Although this length of time may vary in other contexts, it has proven to be acceptable for the execute/analyze cycle on most programs.

- Modern computing facilities consist of scores or hundreds of nodes
- Scripting is vital for eliminating human errors (e.g., copy/paste/edit) within the tens of thousands of M&S input/output files
Axioms related to DASE Step 2 (Objective)

Axiom 4: The objective of a performance-verification simulation experiment involves either constructing a confidence interval or performing a hypothesis test, including confidence and power values, regarding one or more population parameters.

- The most common parameter stated in a requirement is the expected value of a distribution of pass/fail binomial parameters \( \pi_i \), i.e. \( E\{\Pi\} \) or \( \mu_\pi \)
- In this case, Theorem 1 applies when \( Q = 1 \) replicate per scenario:

Theorem 1: Let \( \Pi \) be a random variable which represents the population of possible binomial parameters, and let \( f(\pi) \) denote the associated probability density function (zero outside of the interval \([0,1]\)) with mean \( \mu_\pi = E\{\Pi\} \). Let \( Y \) be a new random variable which is the sum of \( N \) binomially distributed random variables of sample size 1, each with a probability of success which comes from an independent realization of \( \Pi \). In equation form,

\[
y = \sum_{i=1}^N x_i
\]

where \( x_i \sim \beta(1, \pi_i) \), and \( \pi_i \) is the \( i \)th independent realization of \( \Pi \). Then,

\[
y \sim \beta(N, \mu_\pi)
\]

This is true independent of the underlying distribution \( f(\pi) \).
Population-based requirements enable a conceptually simple representation of all plausible scenarios, regardless of the complexity of the factor space being sampled. If each of all admissible scenarios were simulated with full replication, the actual distribution of binomial parameters $\pi_i$ would be known, along with all moments, quantiles, etc. Theorem 1 allows maximal scenario coverage without knowing the actual distribution.
Distributions of binomial parameters $\pi_i$ all with $\mu_\pi = 0.6$

From a scenario coverage point of view, Theorem 1 is good news. But nothing is said or known regarding the dispersion of $\pi_i$ around $\mu_\pi$.

If this insight is desired, we must set $Q > 1$ and hence $N = \frac{M}{Q}$, reducing coverage of the scenario hypercube.
Tradeoffs for DASE Step 2 (Objective)

Mandating a second population parameter eliminates the $Q = 1$ option. Results:

- More statistical precision regarding $F(\pi)$, but
- Reduction in scenario hypercube coverage, as well as
- Fewer scenario data points for constructing Bayesian networks to construct probability models of derived requirements for algorithm performance

Axiom 5: Performance Assessment Working Group (PAWG) agrees upon sampling tradeoffs and documents these tradeoffs within the Performance Verification Plan.

Axiom related to DASE Step 3
(Response and M&S Discernible Difference $\delta$)

**Axiom 6:** Given finite M&S fidelity and resources, the confidence half-interval $\varepsilon$ and/or null/alternate difference $\Delta$ should be no smaller than M&S $\delta$.

From A.Law, *Simulation Modeling and Analysis* (Ch. 5, “Validation”): Given “true” (unknowable) system model means $\mu_S$ and $\mu_M$, the error in estimator $\hat{\mu}$ is given by

$$\text{error in } \hat{\mu}_M = |\hat{\mu}_M - \mu_S| = |\hat{\mu}_M - \mu_M + \mu_M - \mu_S|$$

$$\therefore \text{error in } \hat{\mu}_M \leq |\hat{\mu}_M - \mu_M| + |\mu_M - \mu_S| \text{ (triangle inequality)}$$

The first error term $\varepsilon$ is statistical; the second, $\delta$ is practical (M&S)

Typical declared M&S $\delta_\pi = 0.05$ (probability points). A precise value of $\delta_\pi$ is difficult to decide with any confidence, but it is important for setting a statistical-precision threshold.

Following Axiom 6 minimizes wasteful loss of scenario hypercube coverage mentioned in connection with Axiom 5.
Axiom related to DASE Step 4  (Factor Space \( \{X_C, X_U\} \))

- Simulating a software-intensive, closed-loop system to verify performance over the entire operational envelope involves hundreds of correlated variables, which, strictly speaking, should each be regarded as a random (not fixed) effect—i.e. inference should be done regarding its population of levels. But this is not currently feasible.  

- In M&S, the degree of control is entire (unlike real-world experiments): all variables are controllable & repeatable, so where’s the uncertainty, and thus need for statistics?
  - Factors having “known” values for a given scenario (e.g., initial range, altitude, target type, etc.) are designated as “control” factors \( X_C \), and
  - The remaining, vast majority of factors constitute the set of “uncertainty” factors \( X_U \) (e.g., rocket motor variations, sensor imperfections, target countermeasures, winds), each modeled with a probability distribution

**Axiom 7:** Assignment of each factor to the sets \( \{X_C, X_U\} \) is documented within the Performance Verification Plan.
Axioms related to DASE Steps 5 & 6
(Control factor and/or treatment screening; sampling-for-score)

The role of DASE Step 5 differs for Category-1 objectives vs. the other three categories of objectives, which may involve surrogate model construction for answering questions regarding a tightly restricted subspace

- In Categories 2-4, it may be both appropriate and feasible to screen factors having relatively mild and constant main effects and interactions
- In Category 1, all factors must be explored, within tactically relevant scenarios. Therefore:

**Axiom 8:** Nonsensical control-factor treatments should be identified & screened prior to drawing from the full set of uncertainty factors.

**Axiom 9:** The Performance Assessment Working Group (PAWG) works together to assure that sampling reflects scenarios that are tactically relevant.

**Axiom 10:** The DASE Category-1 experimental design for constructing summary-statistics and Bayes nets is space-filling, i.e. Latin hypercube sampling, with maxi-min spacing.
Example: “Green-pointing” to identify kinematically feasible scenarios

After space-filling sampling of kinematic treatments (e.g., range to target, Mach, target aspect, etc.), scenarios involving the kinematic factors are filtered according to agreed-upon criteria (e.g., Pr(Guide-to-Target), Time-of-Flight, etc.)

- The surviving kinematic scenarios collapse into a single, categorical factor, “kinematics,” akin to “subjects” in a biostatistics study. Each subject is a legitimate (“green point”) treatment for use in performance-scoring in the presence of uncertainty

- This categorical factor must have sufficient levels (“subjects”), both to represent the basic scenario ($X_C$) space and the uncertainty ($X_U$) space with as much power and confidence as is affordable given allotted run-size $M$

- Each “kinematics” level (“subject”) is then mapped to randomly drawn values from the $X_U$ (uncertainty) factors
Axiom related to DASE Step 7
(results review, analysis, conclusions, and next steps)

Just as crucial as starting with a well-defined objective is “letting the data speak for itself” before imposing simplifying statistical assumptions, logic, and math models

- Means, especially marginal means, are very fragile in the presence of outliers
- It is often the outliers that hold keys to improving system performance

**Axiom 11:** Fully automatic generation of statistical estimators before reviewing raw data is to be avoided.

Disregarding Axiom 11 is tempting, given the volume of M&S data and ease of scripting. Just Say No.
Implications & Application Examples

1: A single, bin-level summary statistic (“grand mean”)

2: Two bin-level summary statistics (for dispersion estimate, 10^{th} percentile)

3: Full precision at both the bin and the individual scenario level

- \( \delta_\pi = 5 \) points = \( \varepsilon_\pi \); confidence level \( 1 - \alpha = 0.95 \); coverage fraction = 0.90
- See paper for other values and hypothesis-test sample size requirements
Comparison of required sample sizes

<table>
<thead>
<tr>
<th>Desired precision level</th>
<th>Sample size for 95% conf. interval*</th>
<th>Basis of sample size calculation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: single summary statistic $\mu_\pi$</td>
<td>386</td>
<td>Theorem 1 for $Q = 1$ replicate $\rightarrow N = 0.25 \left(\frac{1.96}{0.05}\right)^2$</td>
<td><strong>Maximizes</strong> $X_C$ hypercube coverage</td>
</tr>
<tr>
<td>2: second statistic $x_q$ to estimate dispersion</td>
<td>580</td>
<td>$M = QN = 20 \times 29$</td>
<td>Once $N$-size sample is available, compute 2-sided confidence interval on $x_q$ for $p = 0.10$: $F\left(\text{ceil}\left{Np - Z_{CL}\sqrt{Np(1-p)}\right}\right)^{-1} \leq x_q \leq F\left(\text{ceil}\left{Np - Z_{CL}\sqrt{Np(1-p)}\right}\right)^{-1}$ (see Conover)</td>
</tr>
<tr>
<td>3: full precision for all scenarios</td>
<td>77,200</td>
<td>$M = QN = 20 \times 386$</td>
<td><strong>Minimizes</strong> $X_C$ hypercube coverage</td>
</tr>
</tbody>
</table>

Difference in sample sizes grows greater when demanding more precision
Recommendations for allocating $M$ runs

Although seeking more statistical precision is understandable,
a) keeping confidence half-interval $\varepsilon_\pi$ close to M&S $\delta$ (DASE Axiom 6),
b) using confidence intervals rather than hypothesis tests, and
c) setting $Q = 1/\delta_\pi$ when seeking individual estimates of $\pi_i$ in order to estimate quantile(s) $x_q$, will all
• help deploy the allotted run size $M = QN$ most effectively,
• allow fuller coverage of algorithm/software paths, and
• provide a broader basis for constructing probability models of derived algorithm requirements (Bayes nets).

Regardless of the tradeoff decision made for precision vs. coverage (DASE Axioms 5-6), always display the raw data underlying estimators of any type (DASE Axiom 11).
1. The *Performance Specification* includes verifiable requirements, and the *Performance Verification Plan* spells out in detail how sampling will occur.

2. No sample of simulation runs should be regarded as perfectly representing actual performance of the system being simulated.

3. The computing resources and allowed time set the number $M = QN$ of runs for scoring bins of related scenarios.

4. A performance-verification experiment is done either to construct a confidence interval or to run a hypothesis test for summary statistic(s).

5. The Performance Assessment Working Group agrees upon sampling tradeoffs and documents these tradeoffs within the *Performance Verification Plan*.

6. Given finite M&S fidelity and resources, the confidence half-interval $\varepsilon$ and/or null/alternate difference $\Delta$ should be no smaller than the M&S discernible difference $\delta$.

7. Factor assignments to $\{X_C, X_U\}$ is documented within the *Performance Verification Plan*.

8. Nonsensical control-factor treatments are identified & screened prior to drawing from $X_U$.

9. The Performance Assessment Working Group assures that sampling reflects tactically relevant scenarios.

10. Latin Hypercube sampling is used to construct summary statistics and Bayesian networks.

11. Avoid fully automatic generation of statistical estimators before reviewing raw data.

12. Theorem 1 identifies the sampling distribution when drawing one $X_U$ replicate per $X_C$ scenario.
Bibliography (p. 1 of 2)


