CASD 2016

Raytheon

Customer Success Is Our Mission

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

MISSION:

A WORLD OF INNOVATION

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

October 2016

Raytheon

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') <u>performance</u> across a factor space**
   a) Establish <u>summary statistic</u>(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 <u>design</u> space**
   a) Perform local sensitivity analysis and/or design optimization
   b) Create trustworthy surrogate models for well-defined purposes
3. **Support <u>tests</u> (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 (<u>SimV&V</u>)**
   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) δ**
4. **Define the experiment's Factor Space:**
   a) Control Factor set $\mathbf{X}_C$: type (numeric/categorical), units, and ranges/levels
   b) Uncertainty Factors set $\mathbf{X}_U$: type, units, distribution types & parameter values
   c) Constants: List critical simulation inputs, including any screened Control Factors
5. **Screen Control Factors $\mathbf{X}_C$ and/or inadmissible $\mathbf{X}_C$ treatments**
   a) Select experimental design – $\mathbf{X}_C$ treatments
   b) Set number of replicates – random $\mathbf{X}_U$ factor draws – for each $\mathbf{X}_C$ treatment
   c) Establish simulation run sequence, and execute & analyze the Screening runs
   d) Select $\mathbf{X}_C$ 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 – $\mathbf{X}_C$ treatments – e.g., Latin hypercube sampling
   c) Set number of replicates – random $\mathbf{X}_U$ factor draws – for each $\mathbf{X}_C$ 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 <u>populations</u> vs. <u>samples</u>**
  - **"Population" terms are denoted using Greek symbols—e.g.,** moments (μ, σ^2, …) and median $\tilde{\mu}$ of random variable X (note: binomial (pass/fail) parameter π often replaces μ 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 μ; sample variance $S^2 = \hat{\sigma}^2$
    - Summary statistic(s), e.g., $\hat{\mu}$, x_q ; sample-proportion <u>estimate</u> of π : $p = (\#\ successes) \div (\#\ attempts)$; (model coefficients $\hat{\beta}_{ij}$ not covered)
- **The relation $M = QN$ refers to the <u>QN Allocation Problem</u>: How best to allocate M runs between N $\mathbf{X}_C$ hypercube scenarios ("treatments"), and Q $\mathbf{X}_U$ replicates randomly drawn drawn per $\mathbf{X}_C$ scenario**
- **For Performance Verification, we must also distinguish between <u>bin</u>-level parameters or estimators (e.g., μ_π, $\bar{X}_{bin}$), vs. <u>scenario</u>-level parameters or estimators (e.g., π_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 π_i, i.e. $E\{\Pi\}$ or μ_π
- In this case, Theorem 1 applies when $Q = 1$ replicate per scenario:

Theorem 1: Let Π 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 Π. In equation form,*

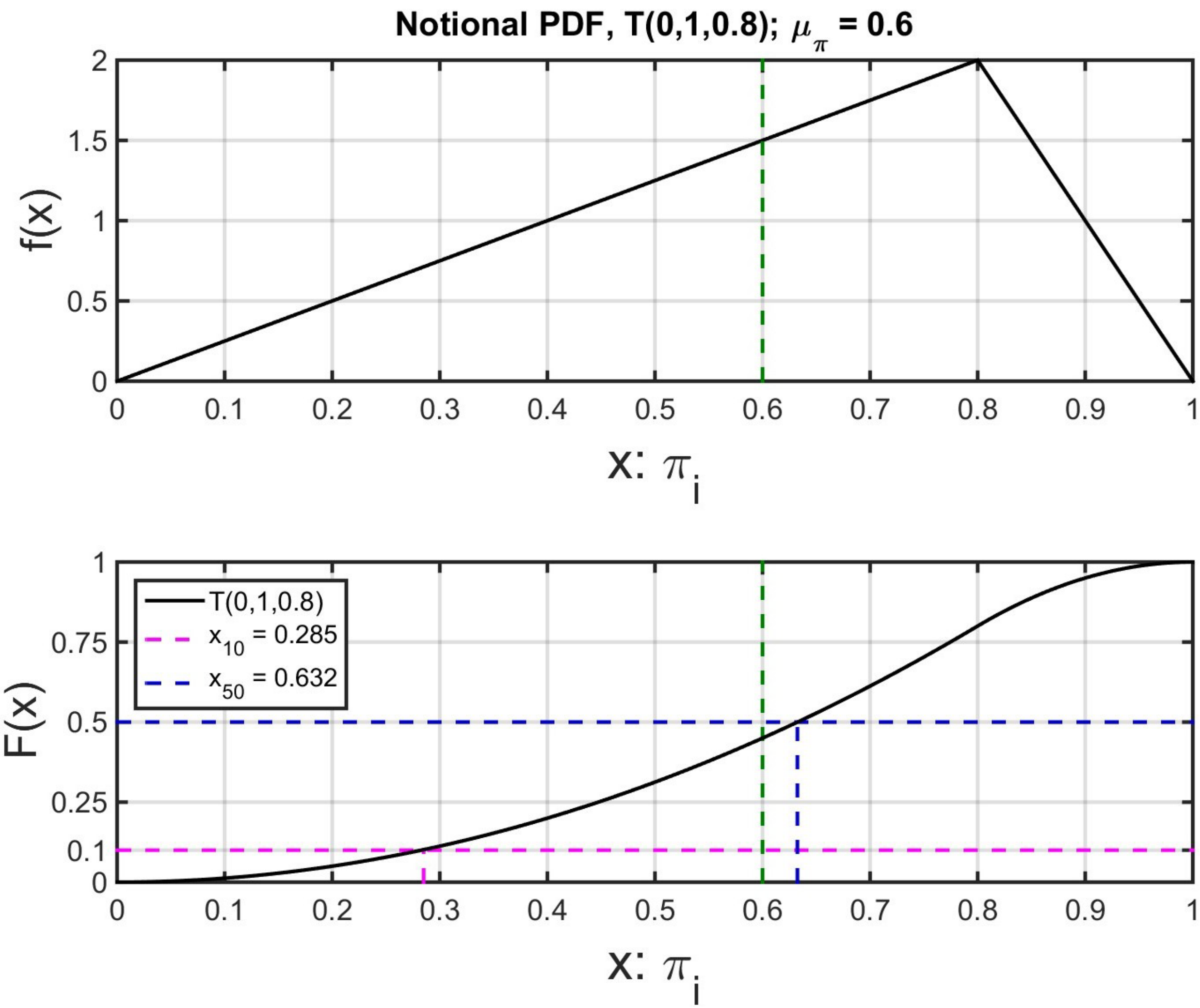
$$y = \sum_{i=1}^{N} x_i \tag{1}$$

where $x_i \sim \beta(1, \pi_i)$, and π_i is the i^{th} independent realization of Π. Then,

$$y \sim \beta(N, \mu_\pi) \tag{2}$$

This is true independent of the underlying distribution $f(\pi)$.

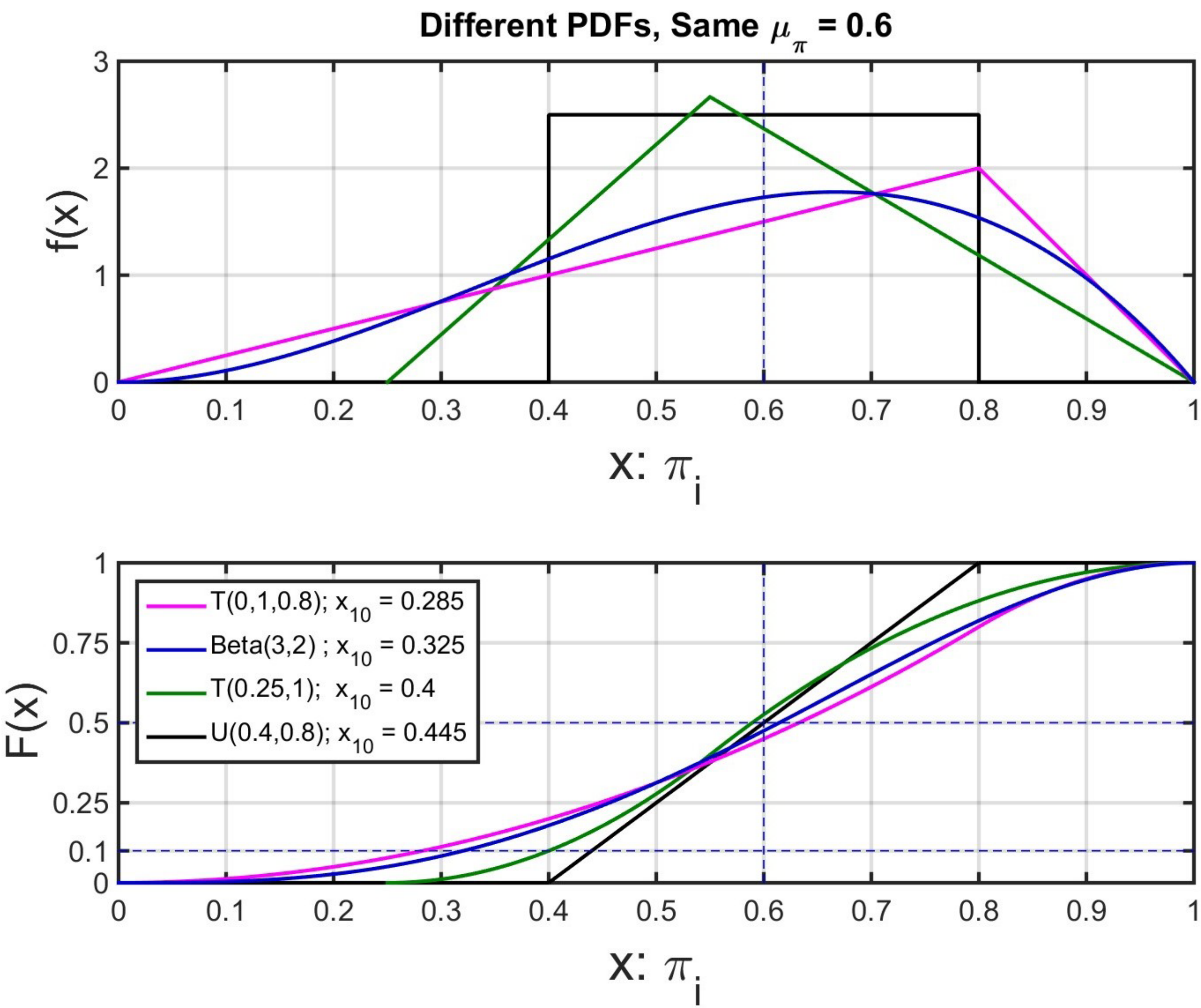
Example: Notional distribution of binomial parameters π_i



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 π_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 π_i all with $\mu_\pi = 0.6$



From a <u>scenario coverage</u> point of view, Theorem 1 is good news. But nothing is said or known regarding the <u>dispersion</u> of π_i around μ_π.

If this insight is desired, we must set $Q > 1$ and hence $N = M/Q$, reducing coverage of the scenario hypercube.

Raytheon

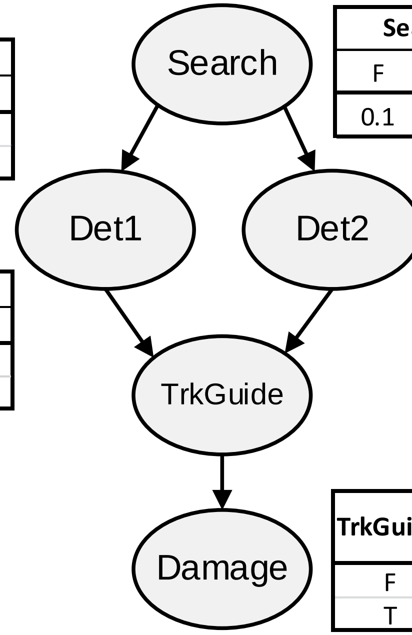
Tradeoffs for DASE Step 2 (Objective)

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

Search	Det1	
	F	T
F	1	0
T	0.1	0.9

Search	Det2	
	F	T
F	1	0
T	0.4	0.6



Search	
F	T
0.1	0.9

Det1	Det2	TrkGuide	
		F	T
F	F	0.55	0.45
F	T	0.12	0.88
T	F	0.4	0.6
T	T	0.1	0.9

TrkGuide	Damage		
	None	Disable	Kill
F	1	0	0
T	0.03	0.02	0.95

Hurst, T.N. J.J. Ballantyne, A.T. Mense, "Building Requirements-Flow Models using Bayesian Networks and Designed Simulation Experiments," *Proceedings*, Joint Statistical Meetings (2014).

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 δ)

Raytheon

<u>Axiom 6</u>: Given finite M&S fidelity and resources, the confidence half-interval ε and/or null/alternate difference Δ should be no smaller than M&S δ.

From A.Law, *Simulation Modeling and Analysis* (Ch. 5, "Validation"):

Given "true" (unknowable) system model means μ_S and μ_M, the error in estimator $\hat{\mu}$ is given by

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

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

The first error term ε is <u>statistical</u>; the second, δ is <u>practical</u> (M&S)

Typical declared M&S $\delta_\pi = 0.05$ (probability points). A precise value of δ_π 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.**[26]
- **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 $\mathbf{X}_C$, and
  - The remaining, vast majority of factors constitute the set of "uncertainty" factors $\mathbf{X}_U$ (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 $\{\mathbf{X}_C, \mathbf{X}_U\}$ is documented within the *Performance Verification Plan.*

Axioms related to DASE Steps 5 & 6

(Control factor <u>and/or</u> 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 <u>may</u> be both appropriate and feasible to screen factors having relatively mild <u>and</u> constant main effects and interactions
- In Category 1, all factors must be explored, within tactically relevant scenarios. Therefore:

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

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

<u>Axiom 10</u>: 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

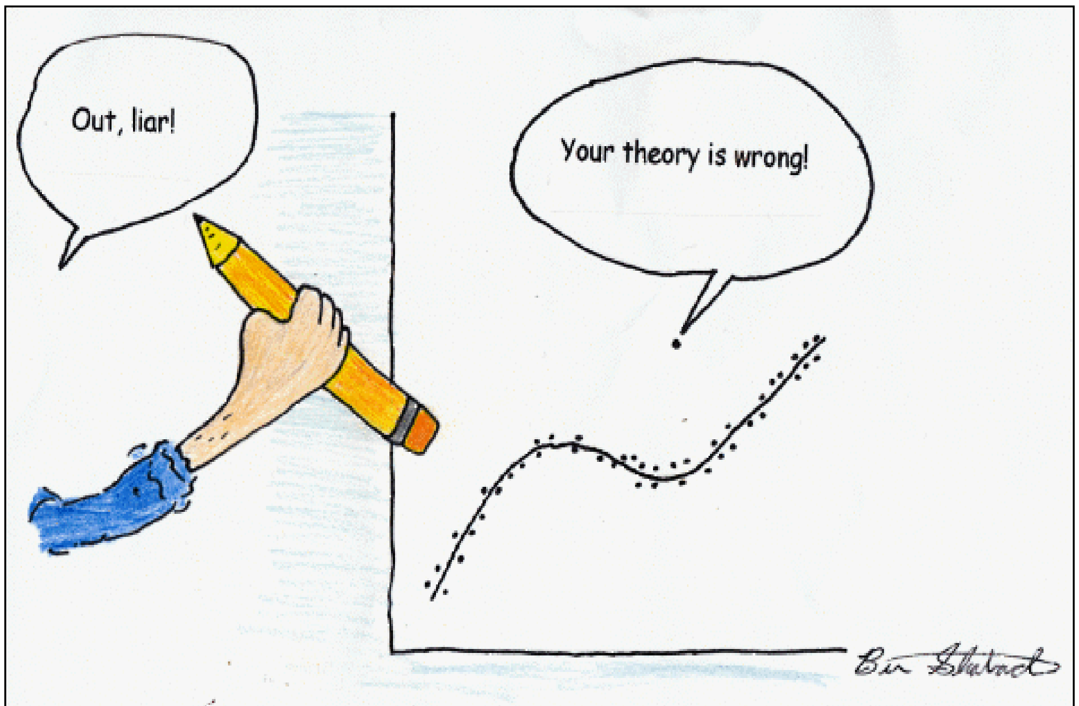
After space-filling sampling of kinematic treatments (e.g., range to target, Mach, target aspect, etc.)**, scenarios involving the kinematic factors are filtered according to agreed-upon criteria** (e.g., Pr(Guide-to-Target), Time-of-Flight, etc.)

- 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 ($\mathbf{X}_C$) space and the uncertainty ($\mathbf{X}_U$) 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, <u>especially marginal means</u>, are very fragile in the presence of outliers
- It is often the <u>outliers</u> that hold keys to improving system performance



<u>Axiom 11</u>: Fully automatic generation of statistical estimators before reviewing raw data is to be avoided.

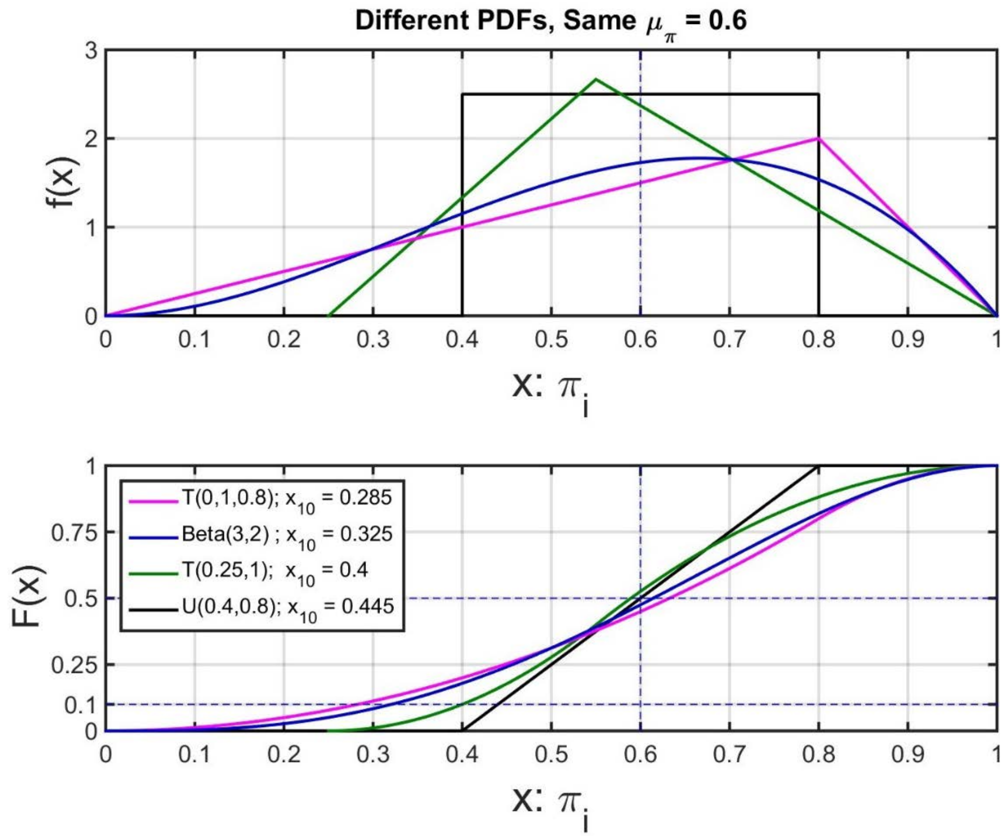
Disregarding Axiom 11 is tempting, given the volume of M&S data and ease of scripting. <u>Just Say No</u>.

Implications & Application Examples

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

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

3: Full precision at both the bin <u>and</u> the individual scenario level



- $\delta_\pi = 5$ points = ε_π; 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

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

Difference in sample sizes grows greater when demanding more precision

Recommendations for allocating M runs

<u>Although seeking more statistical precision is understandable</u>,

a) keeping confidence half-interval ε_π close to M&S δ (DASE Axiom 6),

b) using confidence intervals rather than hypothesis tests, and

c) setting $Q = 1/\delta_\pi$ when seeking individual estimates of π_i in order to estimate quantile(s) x_q, <u>will all</u>

- 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 <u>any</u> type (DASE Axiom 11).

Summary of DASE axioms & sampling theorem

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 ε and/or null/alternate difference Δ should be no smaller than the M&S discernible difference δ.
7. Factor assignments to $\{\mathbf{X}_C , \mathbf{X}_U\}$ is documented within the *Performance Verification Plan*.
8. Nonsensical control-factor treatments are identified & screened prior to drawing from $\mathbf{X}_U$.
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 $\mathbf{X}_U$ replicate per $\mathbf{X}_C$ scenario.

Bibliography (p. 1 of 2)

1. Law, A.M., *Simulation Modeling and Analysis*, 5th ed., McGraw-Hill (2015).
2. Gilmore, J.M., "Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation," memorandum from the Office of the Secretary of Defense, Washington, D.C., Oct. 19, 2010.
3. Gilmore, J.M. (Director, Department of Defense Operational Test and Evaluation), "Memorandum for Users of the DOT&E TEMP Guidebook," 27 Feb. 2012.
4. Gawande, A., *The Checklist Manifesto: How to Get Things Right*, Metropolitan Books (2009).
5. Kleijnen, Jack P.C. *et al*, "State-of-the-Art Review: A User's Guide to the Brave New World of Designing Simulation Experiments," *Proc. 2005 Winter Simulation Conference*.
6. Sanchez, S.M., "Work Smarter, Not Harder: Guidelines for Designing Simulation Experiments," *Proc. 2005 Winter Simulation Conference.*
7. Collins, B.D., T.N. Hurst, and J. M Ard, "Designed Simulation Experiments, Part 1: Roots, Myths, and Limitations of Conventional DOE," AIAA Conference on Modeling & Simulation Technologies, 2011.
8. Hurst, T.N., C.S. Joseph, C.F. Pouchet, and B.D. Collins, "Designed Simulation Experiments, Part 2: DOE for the Digital Age," AIAA Conference on Modeling & Simulation Technologies, 2011.
9. Hurst, T.N., A.S. Cadenhead, S.H. Cole, and A.D. Post, "Applying Experimental Design Techniques to Missile Performance Simulation Experiments," AIAA National Forum on Weapon System Effectiveness (Tucson), 2009.
10. Hurst, T.N., M.T. Pittard, and K.Vander Putten, "Alternatives for Optimizing Algorithms using Designed Simulation Experiments," AIAA Conference on Modeling & Simulation Technologies, 2010.
11. Hurst, T.N., C.S. Joseph, and J.S. Rhodes, "Novel Experimental Design & Analysis Methods for Simulation Experiments Involving Algorithms," U.S. Army Conference on Applied Statistics (Cary, N.C.), 2010.
12. Hurst, T.N. and B.D. Collins, "Simulation Validation Alternatives when Flight Test Data Are Severely Limited," AIAA Conference on Modeling & Simulation Technologies, 2008.
13. Whelan, A. and P. Stevens, "Design & Analysis of Simulation Experiments (DASE) Approach to Circuit Card Assembly Thermal Analysis," *AIAA Thermophysics Conference*, 2011.
14. Hurst, T.N. C.F. Pouchet, A.T. Mense, "Verifying System Performance Using Designed Simulation Experiments," *Proceedings*, Army Conference on Applied Statistics (Monterey, CA), 2012.

Raytheon

Bibliography (p. 2 of 2)

15. Ard, J.M., K.I. Davidsen, T.N. Hurst, "Simulation-Based Agile Development," *IEEE Software* (0740-7459), 2014.
16. Mense, A.T., T.N. Hurst, J.J. Balantyne, "*QN* Allocation: Balancing the Number of Replicates vs. the Number of Treatments in a Designed Simulation Experiment," *Proceedings*, Joint Statistical Meetings (Montreal), 2013.
17. Box, G.E.P. Box, J.S. Hunter, W.G. Hunter, *Statistics for Experimenters*, 2nd ed., Wiley (2005).
18. Agresti, A., *Categorical Data Analysis*, Wiley (2013).
19. Conover, W.J. *Practical Nonparametric Statistics*, 3rd ed., Wiley (1999).
20. Hurst, T.N., J.J. Ballantyne, A.T. Mense, "Building Requirements-Flow Models using Bayesian Networks and Designed Simulation Experiments," *Proceedings*, Joint Statistical Meetings (CASD2014, Washington, D.C.), 2014.
21. Ballantyne, J.J., T.N. Hurst, A.T. Mense, "Learning Bayesian Network Structure using Data from Designed Simulation Experiments," *Proceedings*, Conference on Applied Statistics in Defense (Fairfax, VA), 2015.
22. Montgomery, D.C., *Design & Analysis of Experiments*, 8th ed., Wiley (2012).
23. McKay, M.D., Beckman, R.J., and Conover, W.J. (May 1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics* (American Statistical Association) **21** (2): 239–245.
24. Cioppa, T.W., "Efficient Nearly Orthogonal and Space-Filling Latin Hypercubes," *Technometrics* (vol.41 #1), 2007.
25. Tukey, J., *Exploratory Data Analysis* (Addison-Wesley), 1977.
26. DARPA Broad Agency Announcement: "Minimizing Uncertainty in Designing Complex Military Systems," http://www.darpa.mil/news-events/2015-01-08
27. Lunquist, E., "Technical Brief: Data Farming," *Defense News* (3 January 2013), http://archive.defensenews.com/article/20130103/TSJ01/301030005/Technical-Briefing-Data-Farming
28. Sanchez, S.M., "Better Data, Not Just Big Data," *Proceedings*, 2015 Winter Simulation Conference.