Statistical Methods for Model Validation

January 26, 2017

Outline

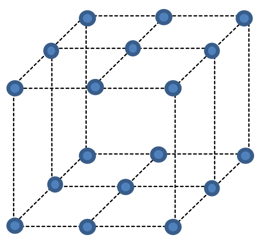
- **Overview of DOE**
- **M&S Validation: What, Why, and Who?**
- **Statistical Techniques for Validation**

What is Design of Experiments (DOE)?

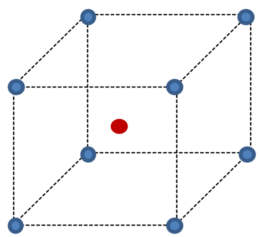
A Structured Approach to Picking Test Points

Tied to Test Objectives

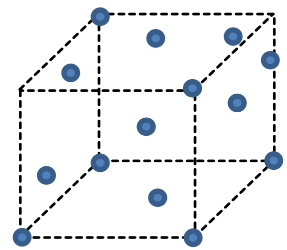
Connected to the Anticipated Analysis



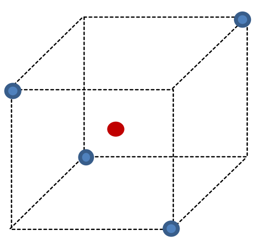
General Factorial
3x3x2 design



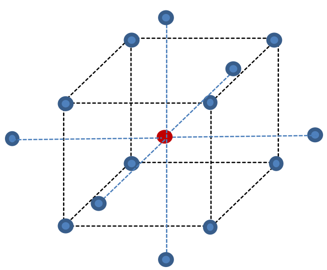
2-level Factorial
2^3 design



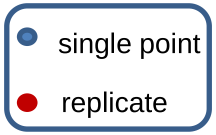
Optimal Design
IV-optimal



Fractional Factorial
2^{3-1} design



Response Surface
Central Composite design



"Just Enough" test points: most efficient!

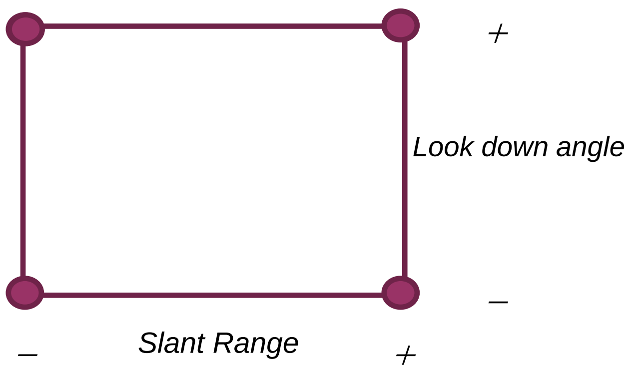
Test Design must support the Analysis we expect to perform!

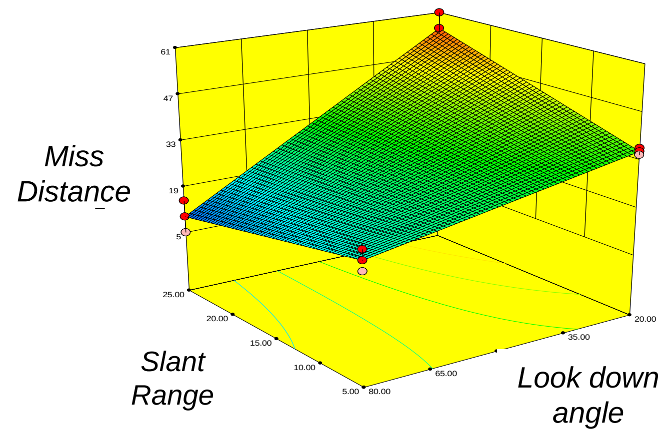
Which **factors** **in the operational space are essential to understand?**

Are **interactions** **between factors likely?**

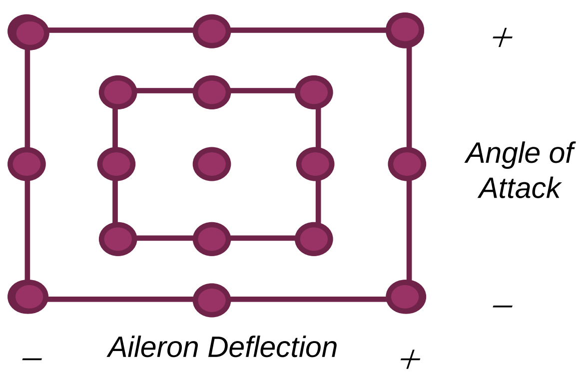
What about quadratic terms to explain **curvature****?**

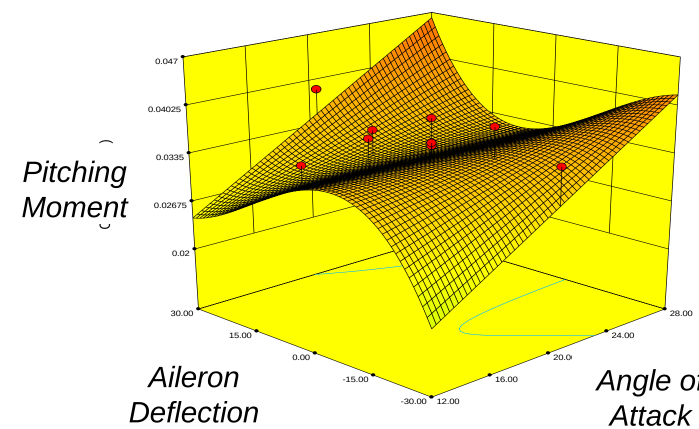
a)





b)



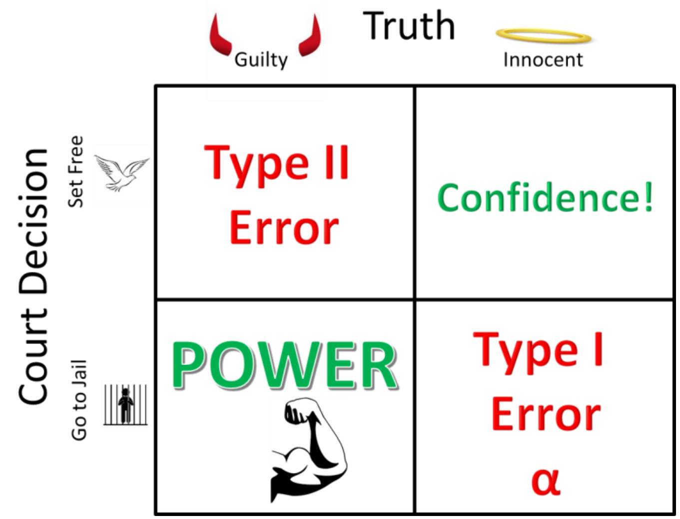


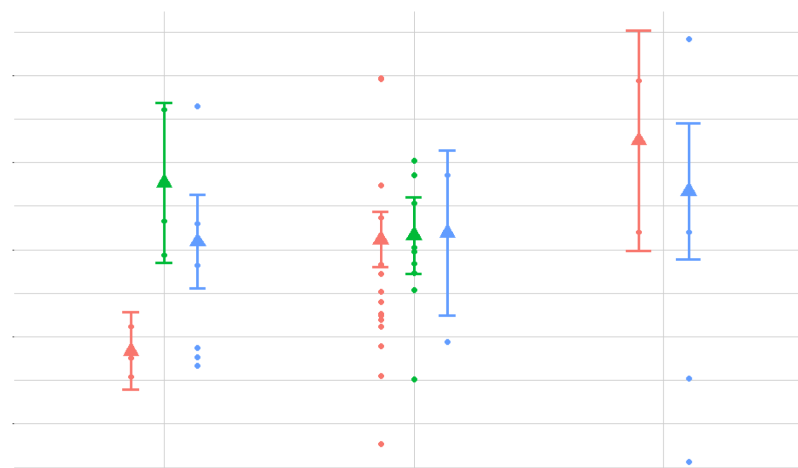
How Much Testing is Enough?

- **[Confidence](#) describes the risk of "False Positive" (Type I Error)**
  - Associated with the null hypothesis
  - What risk are we willing to accept of falsely rejecting the null hypothesis?

- **[Power](#) describes the risk of a "False Negative" (Type II Error)**
  - Associated with the alternative hypothesis
  - What risk are we willing to accept of falsely failing to reject the null hypothesis?

- **Power provides a strong indication of how wide the confidence intervals will be when reporting results**





Outline

- Overview of DOE
- **M&S Validation: What, Why, and Who?**
- Statistical Techniques for Validation

Uses for Modeling & Simulation (M&S) in Operational Testing

- **Supplement or augment live test data when experiments are cost and/or safety prohibitive**
- **Examine threats incapable of being reproduced for testing**
- **Characterize rare events or threats**
- **Allow for end-to-end mission evaluation**
- **Inform experimental design decisions**

M&S can never fully replace testing in the true operational environment (open air, at sea, etc.)

VV&A

- **All M&S used in T&E must be accredited by the intended user (PM or OTA). DOT&E determines if a model has been adequately VV&A'd to use in Operational Testing.**

- **"Verification is the process of determining if the M&S accurately represents the developer's conceptual description and specifications and meets the needs stated in the requirements document."**

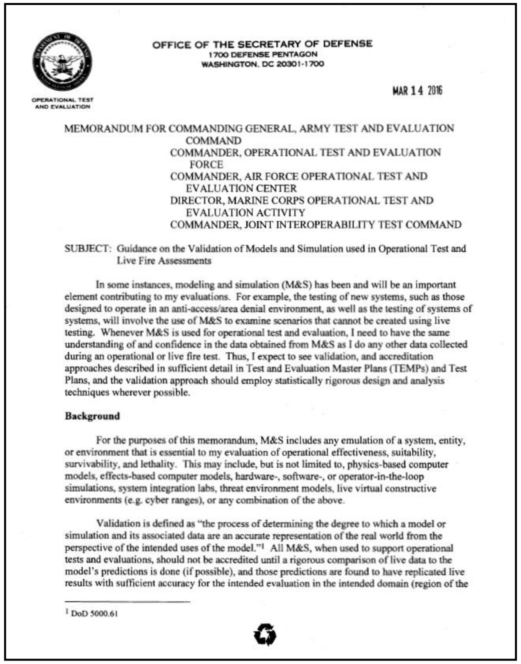
- **"Validation is the process of determining the extent to which the M&S adequately represents the real-world from the perspectives of its intended use."**

- **"Accreditation is the official determination that the M&S is acceptable for its intended purpose."**

"A model should be developed for a specific purpose (or application) and its validity determined with respect to that purpose" (Sargent 2003)

DOT&E Guidance Memo (Mar 14 2016):

- **Provides guidance on the validation of models and simulations used in operational test and live fire assessments**
- **TEMPs and Test Plans must describe the validation and accreditation process in sufficient detail**
- **Rigorous statistical design and analysis techniques should be used wherever possible**
  - Apply design of experiments principles when planning data collection for the M&S and the live test (if applicable)
  - Employ formal statistical analysis techniques to compare live and M&S data



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST AND EVALUATION

MAR 14 2016

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND

SUBJECT: Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments

In some instances, modeling and simulation (M&S) has been and will be an important element contributing to my evaluations. For example, the testing of new systems, such as those designed to operate in an anti-access/area denial environment, as well as the testing of systems of systems, will involve the use of M&S to examine scenarios that cannot be created using live testing. Whenever M&S is used for operational test and evaluation, I need to have the same understanding of and confidence in the data obtained from M&S as I do any other data collected during an operational or live fire test. Thus, I expect to see validation, and accreditation approaches described in sufficient detail in Test and Evaluation Master Plans (TEMPs) and Test Plans, and the validation approach should employ statistically rigorous design and analysis techniques wherever possible.

Background

For the purposes of this memorandum, M&S includes any emulation of a system, entity, or environment that is essential to my evaluation of operational effectiveness, suitability, survivability, and lethality. This may include, but is not limited to, physics-based computer models, effects-based computer models, hardware-, software-, or operator-in-the-loop simulations, system integration labs, threat environment models, live virtual constructive environments (e.g. cyber ranges), or any combination of the above.

Validation is defined as "the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model."[1] All M&S, when used to support operational tests and evaluations, should not be accredited until a rigorous comparison of live data to the model's predictions is done (if possible), and those predictions are found to have replicated live results with sufficient accuracy for the intended evaluation in the intended domain (region of the

[1] DoD 5000.61

When M&S is used as part of OT evaluations of effectiveness, suitability, survivability, or lethality, we should ensure we understand and characterize the usefulness and limitations of the M&S!

Validation Strategies

- **A statistical comparison of the model output to live data should be a *portion* of a larger validation plan**
  - Both quantitative and qualitative evaluations are necessary to understand the strengths and weaknesses of the model across the operational envelope
    - » Face validity, SME evaluation, comparison to other models, comparison to historical data, etc. are all acceptable methods, but should not be the ONLY validation methods used
    - » Must consider sensitivity analysis (do changes to inputs produce reasonable changes to outputs?) and predictive validation (can the model predict live test outcomes?)

- **Developing a validation strategy and designing associated experiments takes a lot of coordination among several groups of people!**
  - Testers, including statisticians
  - Model developers
  - Users
  - Subject Matter Experts / Independent evaluators
  - Etc.

- **This integrated V&V team must also decide on appropriate accreditation criteria in accordance with the intended use of the model**

Outline

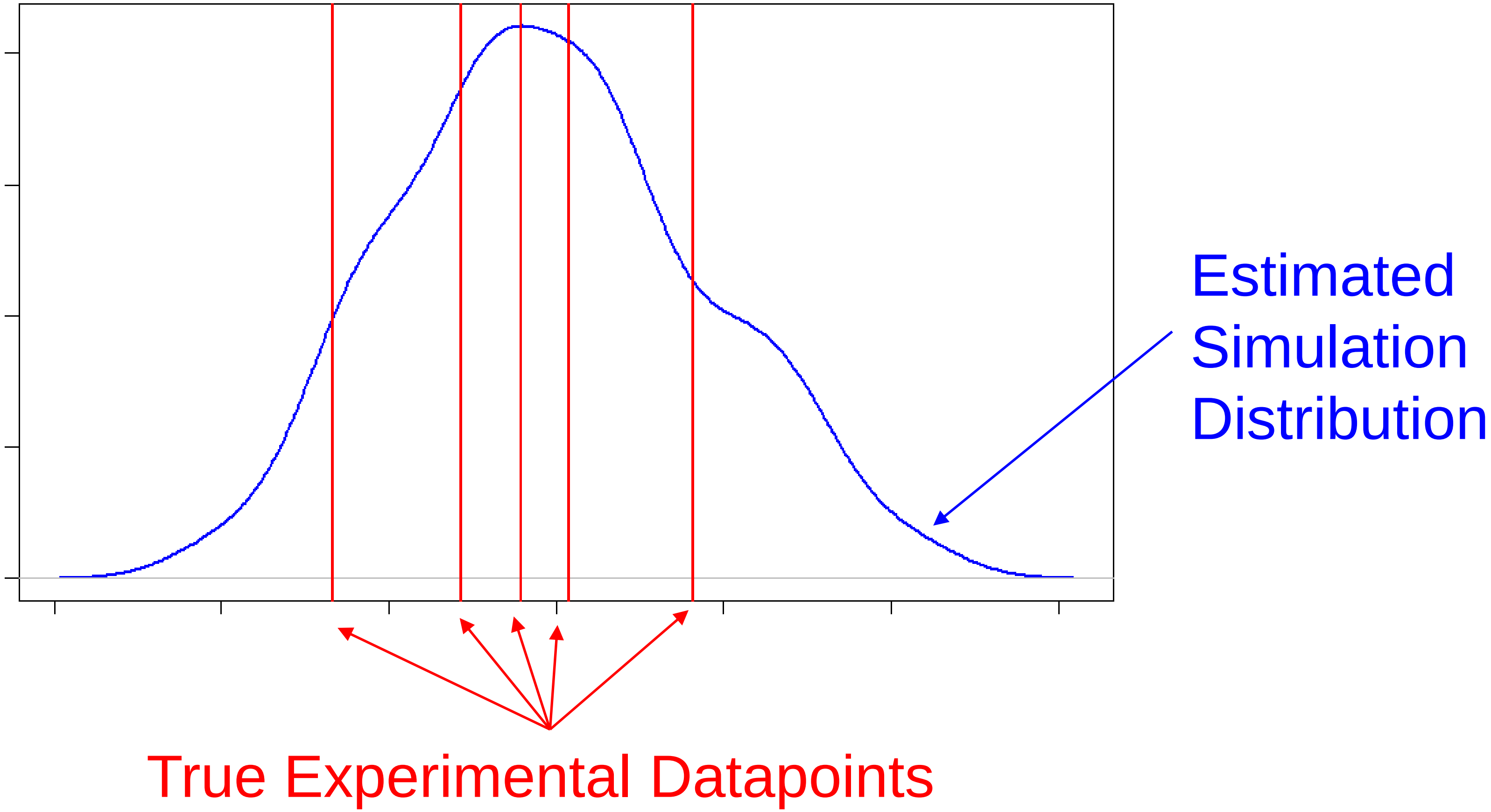
- Overview of DOE
- M&S Validation: What, Why, and Who?
- **Statistical Techniques for Validation**

Do the simulation data and the experimental data agree?

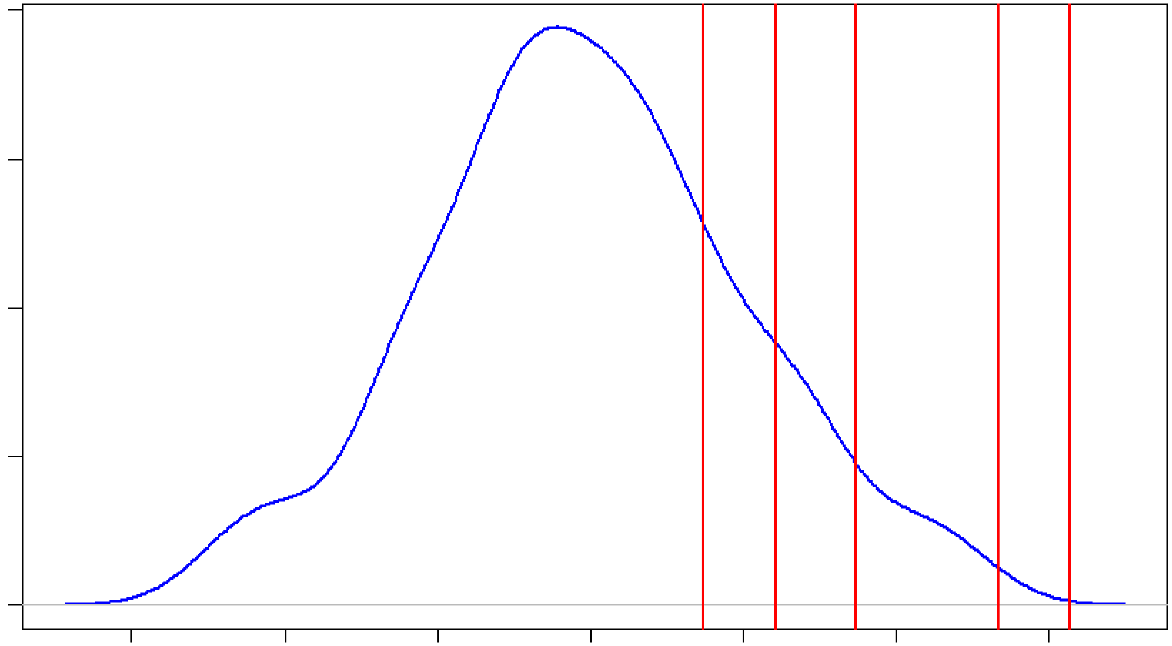
What uncertainty is there in the simulation data?

If they don't agree, can we identify the specific conditions where they disagree?

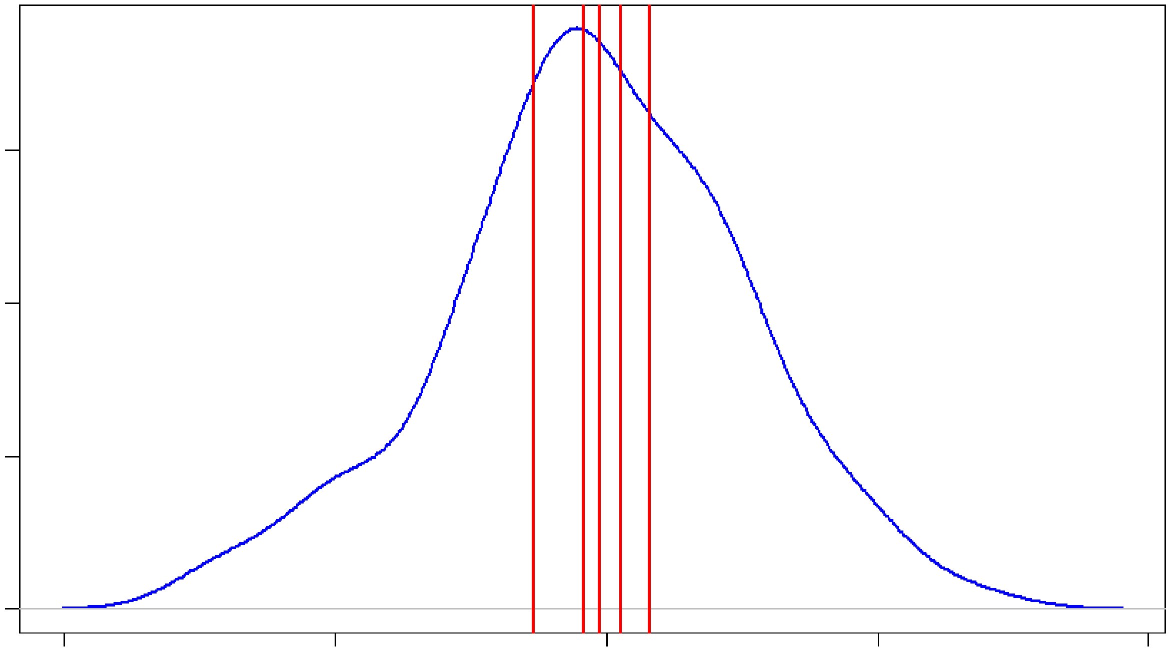
Data Example



Change in Mean



Change in Variance

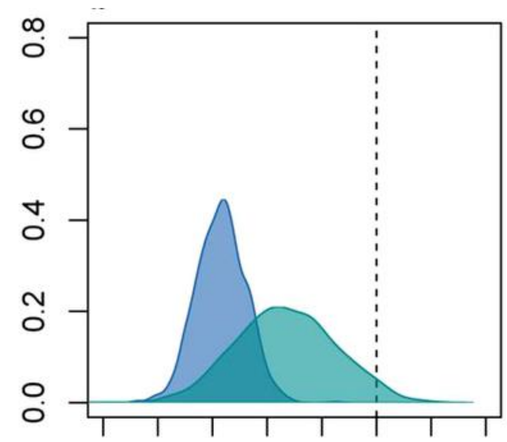


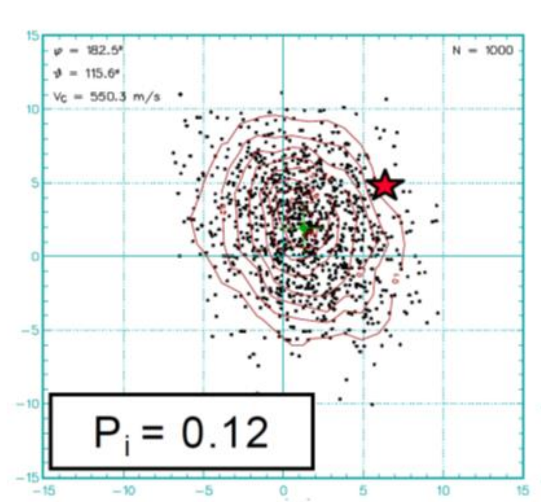
Validation Testing

- H_0: Simulation output matches the live data
- H_1: Simulation output does not match the live data
- "Matching" can be in terms of a variety of parameters, including the means, the variances, and the distributions
- Goal is to maximize power given a specified confidence level
- Higher power and confidence translates into less uncertainty about the difference between live and sim

Statistical Test Options

- **Parametric Tests**
  - t-test (or log t-test)
  - Kolmogorov-Smirnov Test
- **Non-parametric Tests**
  - Kolmogorov-Smirnov Test
  - Fisher's Combined Probability Test
  - Fisher's Exact Test
- **Regression Testing**
  - Multiple Regression (Linear, lognormal, or logistic)
  - Emulation and Prediction
- **Considerations**
  - Ignore factors or take factors into account?
  - Are a combination of techniques necessary in some cases?





Overall Recommendations

- **Use the method(s) that makes most sense for your observed data!!!**

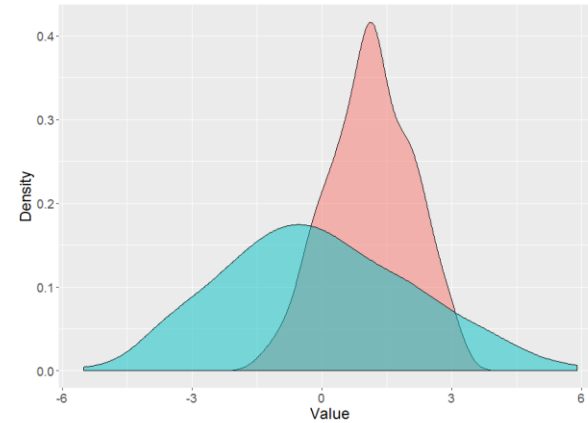
- **Multiple statistical techniques can be used to check for various types of differences between live and sim**

- **General classes of comparison methods that tend to work well:**
  - Non-parametric Kolmogorov-Smirnov test and Fisher's Combined Probability Test
    - » Work well for distribution comparisons
  - Regression analysis (to include variations like logistic and lognormal) with indicator variable for live/sim
    - » Works best for matched designed experiments
  - Statistical emulation and prediction
    - » Works well for lots of M&S data and limited live data

Recommendations determined via Monte Carlo power simulations

Kolmogorov-Smirnov (K-S) Test

- **Compare the distribution of live data to the distribution of M&S data**
  - The K-S test calculates the maximum distance between two CDFs



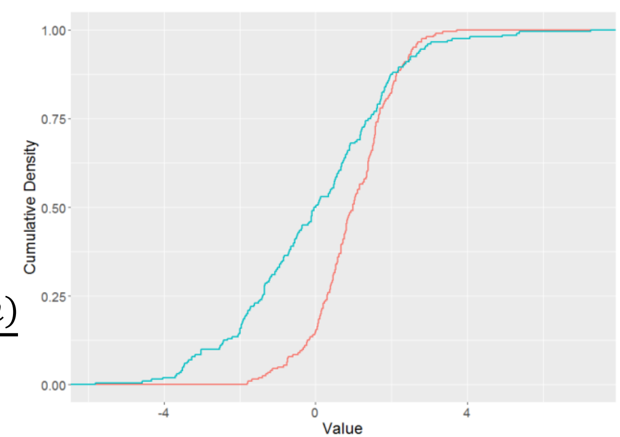
- **Parametric: Compare each of the data sets (live and sim) to a *reference distribution* (e.g. normal)**

- **Non-parametric: Compare each of the data sets (live and sim) to *each other***

Works better for our problem

- **Scaling the data first can account for different conditions**
  - For each distinct condition:

$$\textit{Scaled data} = \frac{each\ individual\ data\ point - mean\ (all\ data\ in\ that\ condition)}{stan\ dev\ (all\ data\ in\ that\ condition)}$$



Note: All data are notional

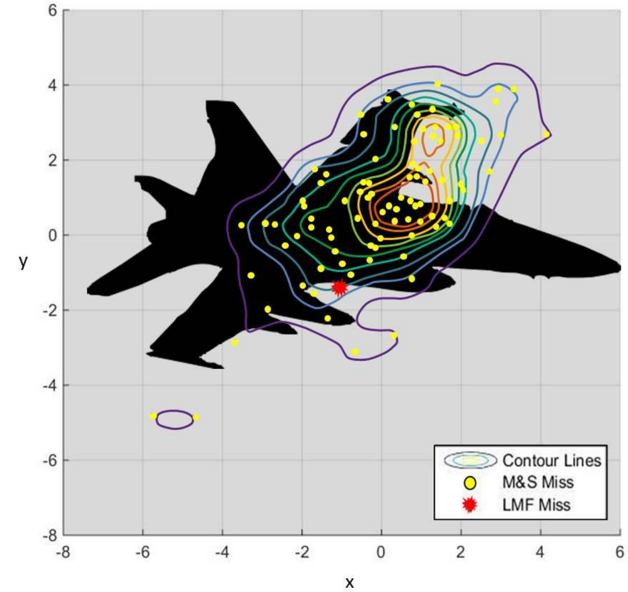
Fisher's Combined Probability Test

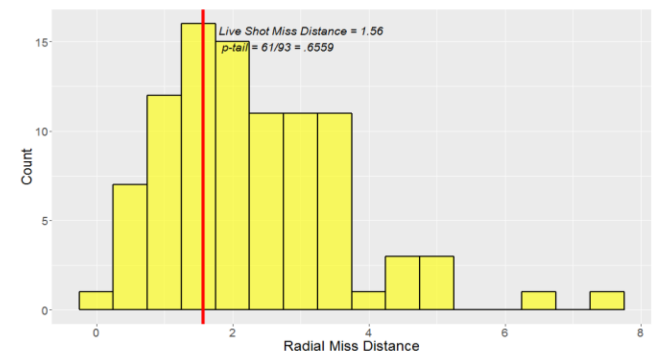
- **Compares distributions of continuous data**
  - Simulation "cloud" vs. 1 or more live shots per condition
  - Nonparametric

- **p-values can be calculated in a variety of ways**
  - 2 dimensionally using contours
  - 1 dimensionally using miss distance quantiles

- **Use a goodness-of-fit procedure to check for overall uniformity of the p-values**
  - Fisher's Combined probability test: $X = -2 \Sigma \ln(p)$ follows a chi-square distribution with 2N degrees of freedom
    - » Sensitive to one failed test condition
  - Kolmogorov-Smirnov test: compares observed p-values to a true uniform distribution

- **No formal test of factor effects**





Note: All data are notional

Regression Modeling: Parameterizing Live vs. Sim

- **Pool live and M&S data and build a statistical model**
  - Include an indicator term that indicates whether the data point comes from live or M&S (*test type*), as well as interaction terms between *test type* and other factors of interest
  - For example,

    $$\textit{Detection Range} = \beta_0 + \beta_1 TestType + \beta_2 Threat + \beta_3 (TestType * Threat) + \epsilon$$
  - If the *Test Type* effect is statistically significant, then the M&S runs are not providing data that are consistent with the live runs
  - If the interaction term is significant, there many be a problem with the simulation under some conditions but not others

- **The type of regression depends on the nature of the observed data**
  - Symmetric – use linear regression
  - Skewed – use lognormal regression
  - Binary – use logistic regression

- **Method works best if you used a designed experiment for both live and sim**
  - Must compute interaction terms to avoid rolling up results
  - Strength is detecting differences in means

- **Works well even when there is limited data**

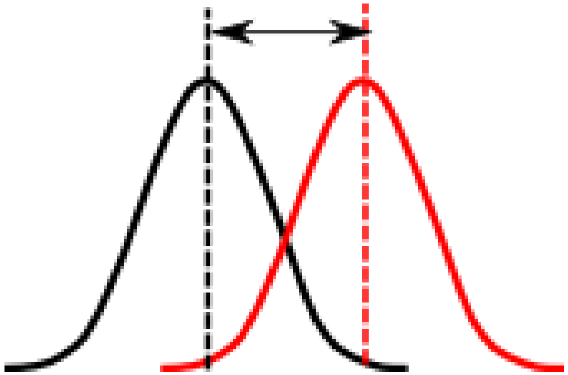
Conclusion

- **When used to support OT or live fire evaluations, M&S validation should include a rigorous comparison of live data with simulation output**
  - This means carefully considering how much data in what conditions should be collected during live testing to support the appropriate analyses
  - An integrated V&V team must coordinate to develop appropriate validation strategy and acceptability criteria
- **Testers should review the March 2016 memo and ensure TEMPs and Test Plans include the appropriate information**
  - If TEMP timelines are out of sync with VV&A planning, information can be presented to DOT&E in the form of an M&S concept briefing as soon as the details are available
- **M&S validation case studies and a detailed implementation handbook will be posted on the DOT&E webpage in the coming weeks and months**
  - Handbook will describe and recommend statistical methods such as those presented today

BACKUP

T-test

- **Parametric test to compare the means of 2 data sets (e.g. live and sim)**
- **Assumptions:**
  - Data is approximately normally distributed
  - Observations are independent of one another
- **If the data is skewed, a log transformation can be performed and a t-test conducted on the transformed data (we call this a log t-test for short)**
- **Powerful tool for detecting differences in means when assumptions are met**
- **Doesn't test for factor effects**
- **Cannot detect differences in variance**
- **Requires a moderate amount of live data**



Fisher's Exact Test

- Nonparametric test for binary or categorical data
- Consider the following contingency table:

	Pass	Fail	
Live	a	b	a+b
Sim	c	d	c+d
	a+c	b+d	n

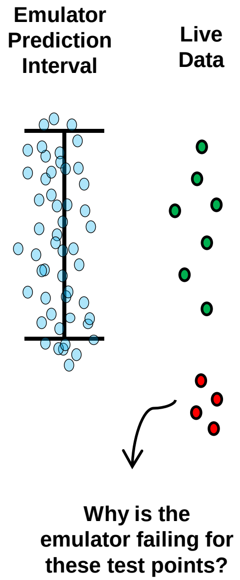
- Assuming the margins of the table are fixed, the exact probability of a table with cells a, b, c, d and marginal totals (a+b), (c+d), (a+c), and (b+d) equals

$$\frac{(a+b)! * (c+d)! * (a+c)! * (b+d)!}{n! * a! * b! * c! * d!}$$

- Works well when there are no factors and for small sample sizes

Emulation and Prediction

- **Build an empirical emulator (i.e. statistical model) from the simulation data**
- **As a new set of live data becomes available, compare each point with the prediction interval generated from the emulator under the same conditions**
  - If a live point falls within the prediction interval, that is evidence that the simulation is performing well under those conditions
- **Use the results to help inform future testing and/or fix the simulation**
  - Test for any systematic patterns to help explain where / why the simulation is failing in certain cases
  - Live data can then be used to update the simulation and continue to "train" the model
- **Method works best if you used a designed experiment**
  - Strength is detecting differences in variance
- **Works well even when there is limited data**



Detailed Recommendations

Distribution	Structure Of Factors	Small Sample Sizes	Moderate Samples Sizes	Large Sample Sizes
Symmetric	Univariate	Fisher's Combined	T-test Fisher's Combined Non-Par KS	T-test Fisher's Combined Non-Par KS
	Distributed Level Effects	Combo Test	Sc Non-Par KS	Sc Non-Par KS
	Designed Experiment	Linear Regression Sc Non-Par KS Emulation & Pred	Linear Regression Sc Non-Par KS Emulation & Pred	Sc Non-Par KS
Skewed	Univariate	Fisher's Combined	Log T-test Fisher's Combined Non-Par KS	Log T-test Fisher's Combined Non-Par KS
	Distributed Level Effects	Combo Test	Sc Non-Par KS	Sc Non-Par KS
	Designed Experiment	Lognormal Regression Sc Non-Par KS Emulation & Pred	Lognormal Regression Sc Non-Par KS Emulation & Pred	Sc Non-Par KS
Binary	Univariate	Fisher's Exact	Fisher's Exact	Fisher's Exact
	Distributed Level Effects	Logistic Regression	Logistic Regression	Logistic Regression
	Designed Experiment	Logistic Regression	Logistic Regression	Logistic Regression

Notes on sample sizes:
Simulation sample size = 100 in all cases; Live sample size (symmetric and skewed): Small = 2-5, Moderate = 5-10, Large = 11-20; Live sample size (binary): Small = 20, Moderate = 40, Large = 100