INSTITUTE FOR DEFENSE ANALYSES

Statistical Methods for M&S V&V: An Intro for Non-Statisticians

Keyla Pagan-Rivera, Project Leader

John T. Haman
Kelly M. Avery
Curtis G. Miller

OED Draft

March 2024

IDA Product ID-3000770

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task 229998, "Statistical Methods for M&S V&V: An Intro for Non-Statisticians Task," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Dr. Megan L. Gelsinger, Mr. Phillip W. Meade, Dr. Rebecca M. Medlin, and Dr. Russel D. Miller from the Operational Evaluation Division, and Dr. John W. Dennis III from the Strategy, Forces and Resources Division.

For more information:
Dr. Keyla Pagan-Rivera, Project Leader
kpaganri@ida.org • 703-845-6936

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Project ID 3000770

Statistical Methods for M&S V&V: An Intro for Non-Statisticians

Keyla Pagan-Rivera, Project Leader

John T. Haman
Kelly M. Avery
Curtis G. Miller

Executive Summary

IDA was invited to provide a short training to the Navy Modeling & Simulation (M&S) Verification Validation & Accreditation (VV&A) Working Group's Subgroup on Validation Statistical Method Selection. DOT&E reviewed and approved the development of this material. The briefing is intended to motivate and explain the basic concepts of applying statistics to verification and validation for an audience that is technical but not necessarily trained in statistics.

After a brief introduction to IDA and Test Science, the training starts off by motivating why statistics is important when conducting a validation analysis. Analysts should not simply rely on visual inspection of data when comparing live test results to model output. This introduction also introduces key terminology and discusses relevant DoD and DOT&E policy on M&S VV&A.

The next section of the training focuses on statistical techniques that can be used to either compare live data to M&S output or explore the behavior of the model on its own. The slides introduce the various analytical goals of VV&A and outline which methods are recommended in certain conditions. They then go into some more technical detail on the specific techniques that frequently perform the best. The training also emphasizes that rigorous analysis requires an appropriate data collection method, and introduces methods such as classical design of experiments and space-filling designs.

The training concludes with an example that walks through an application of multiple statistical methods to a realistic dataset. It highlights the inadequacies of techniques that do not account for factor effects and provides snippets of R code that can be used to perform each analysis technique presented.

Statistical Methods for M&S V&V: An Intro for Non-Statisticians

John Haman

Kelly Avery

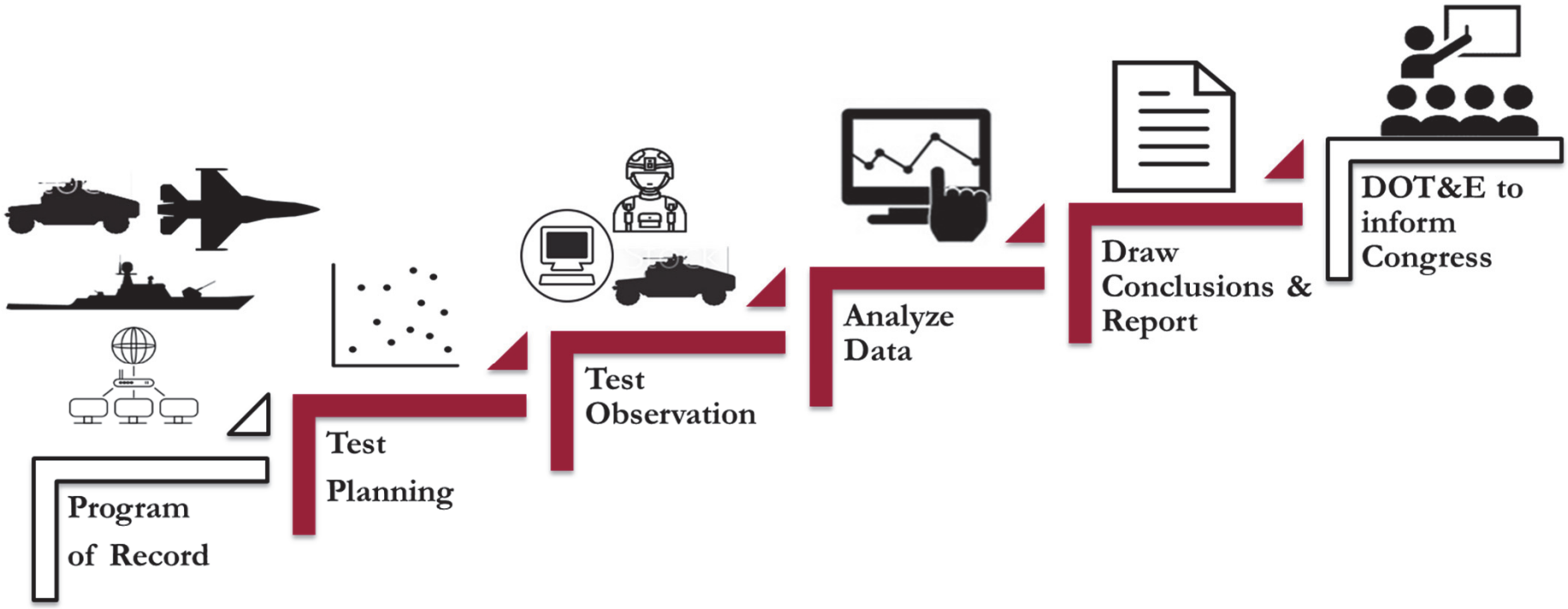
Curtis Miller

December 2023

Institute for Defense Analyses

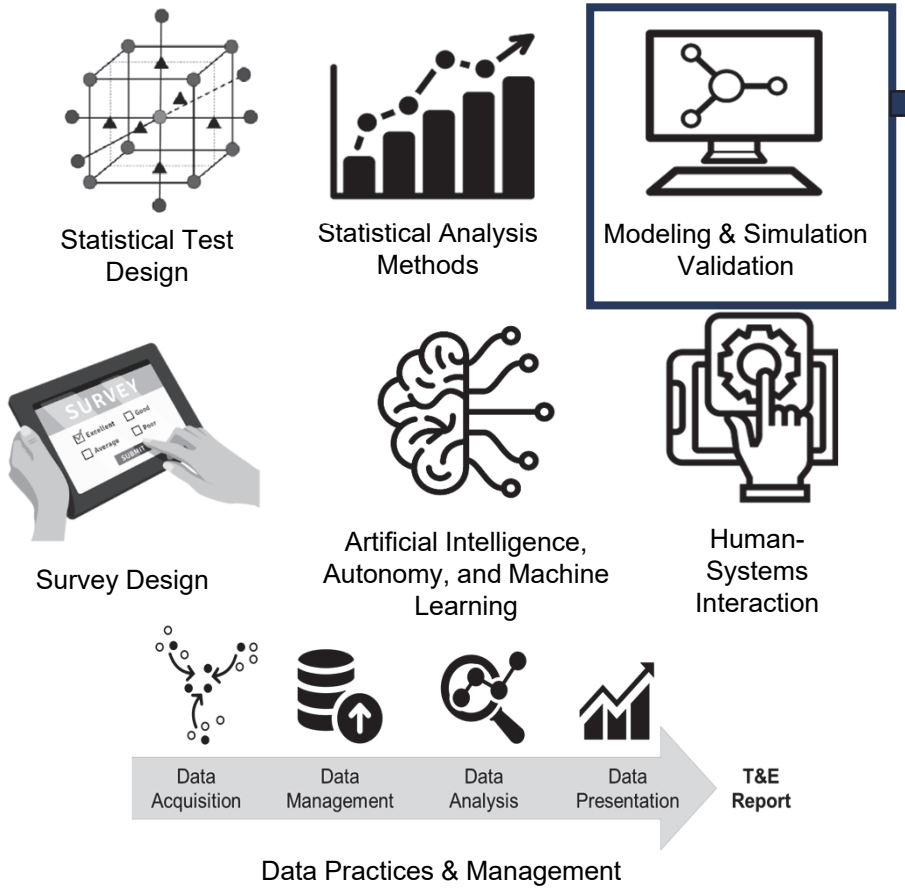
730 East Glebe Road ● Alexandria, Virginia 22305

IDA's Operational Evaluation Division provides technical and analytical support to DOT&E



The Test Science Team within OED develops, applies, and disseminates statistical, psychological, and data science methodologies

Core Areas of Expertise



Available resources:

- *Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation*
- *Space-Filling Designs for Modeling & Simulation*
- *Metamodeling Techniques for Verification and Validation of Modeling and Simulation Data*
- Multiple seminars and trainings ranging from 1 hour to 1 day
- Testscience.org

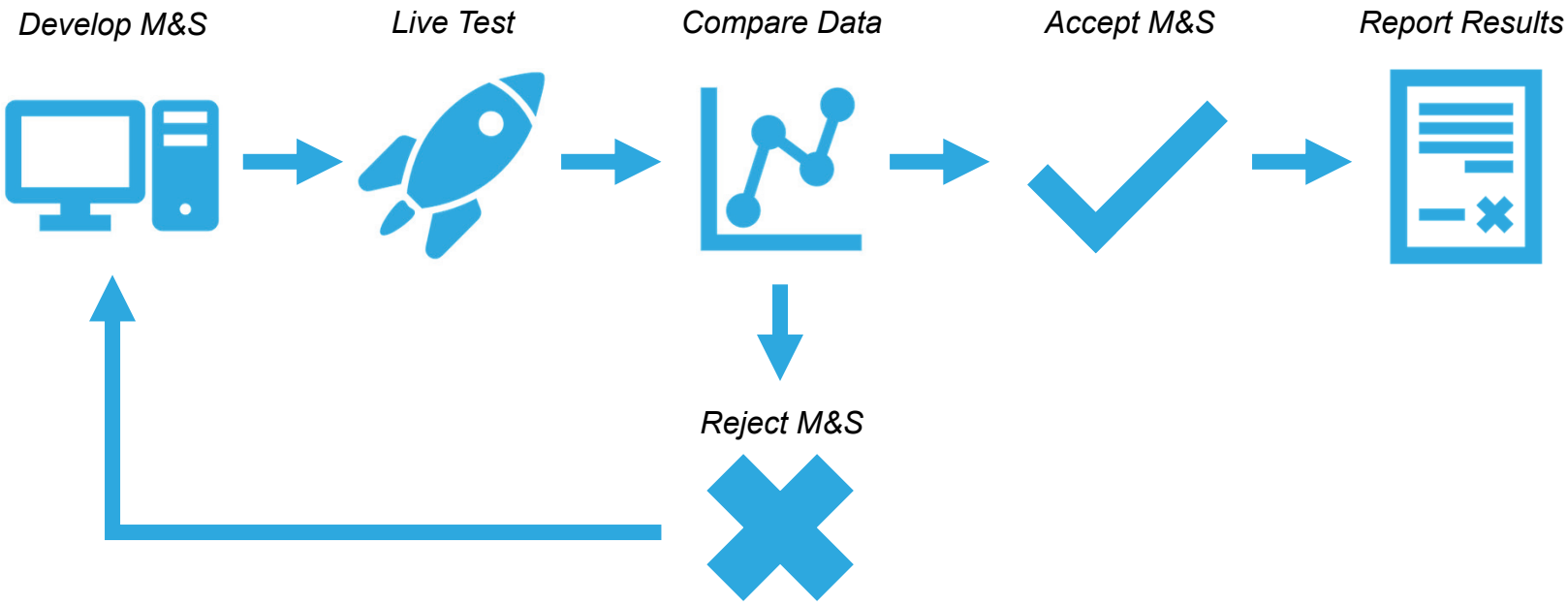
Outline

1. Motivation and Statistical Concepts
2. Statistical Methods for V&V
3. Example with Code

Outline

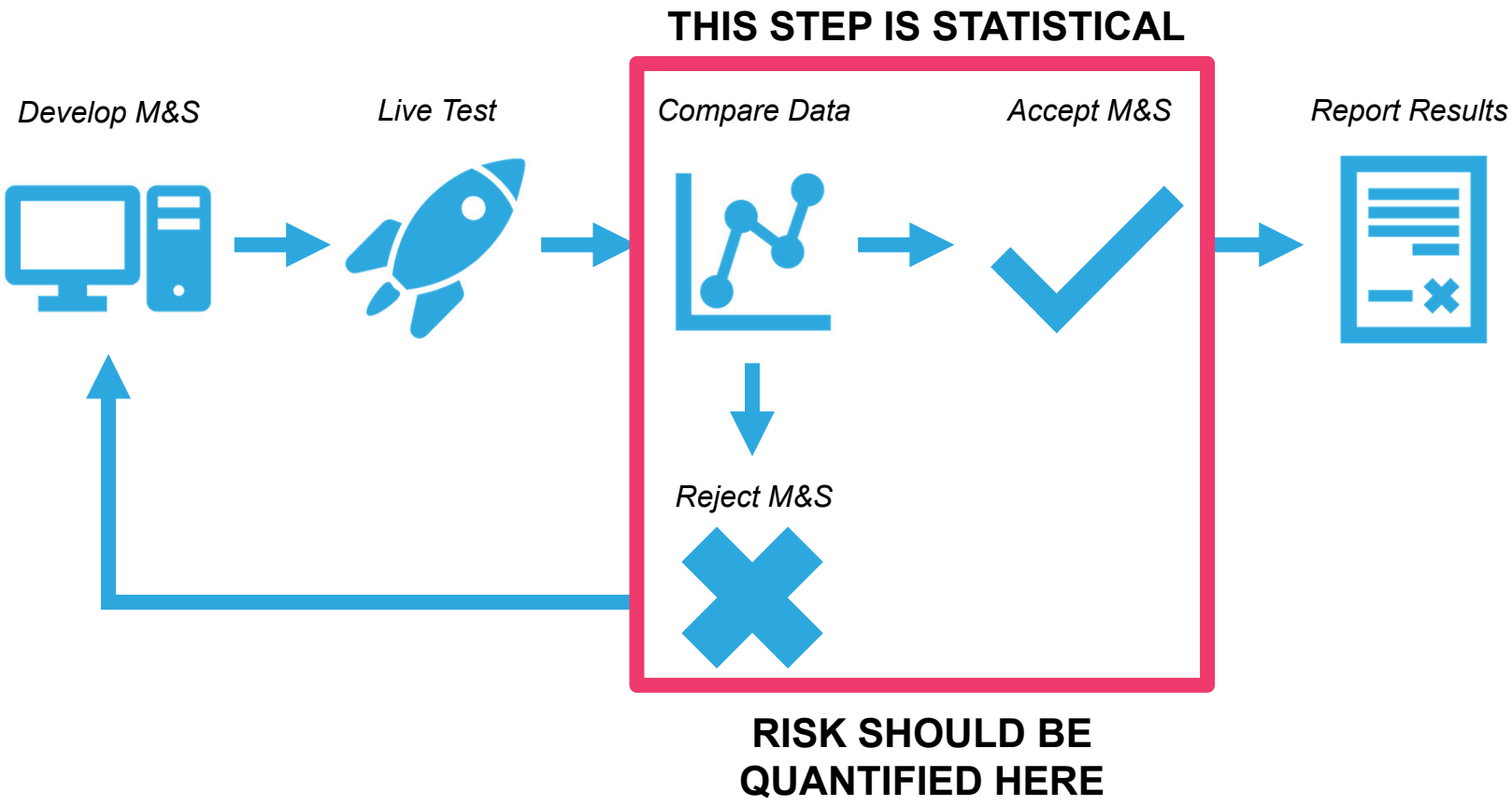
1. Motivation and Statistical Concepts
2. Statistical Methods for V&V
3. Example with Code

If M&S output will be used as part of OT evaluation, we must understand the capabilities and limitations of the M&S and compare the outputs to available live data



M&S: Modeling and Simulation; OT: Operational Testing

If M&S output will be used as part of OT evaluation, we must understand the capabilities and limitations of the M&S and compare the outputs to available live data



M&S: Modeling and Simulation; OT: Operational Testing

Why statistics?

- SMEs can tell you whether the M&S includes all of the important factors and outputs.
- But eyeballs are not the best measurement tools for comparing numbers.

Subject Matter Expert (SME) review is important, but it is not sufficient.

M&S: Modeling and Simulation

Humans are programmed to find patterns, and so they do, even when patterns aren't there

This bread looks like a rabbit

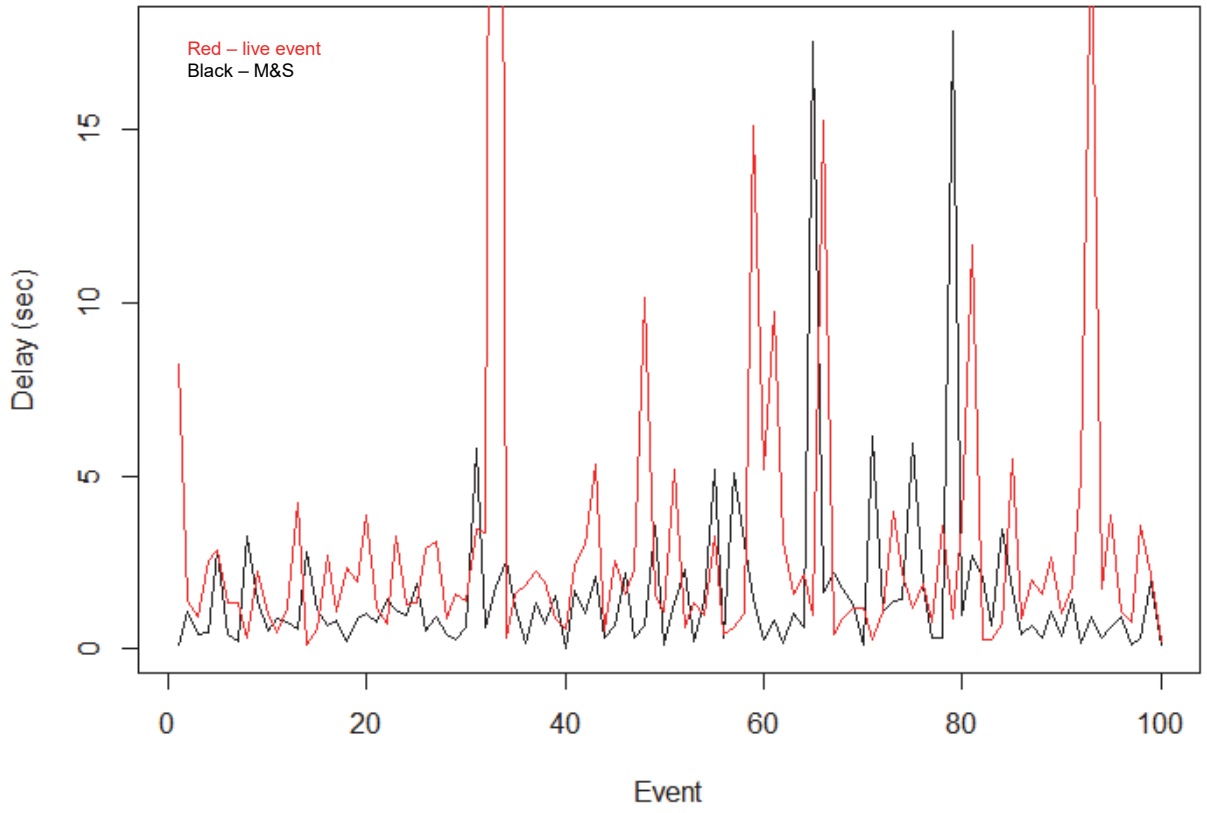
Since the 1930s (and probably before), people have been drawing pictures of noise and overinterpreting them

Chart courtesy of StockCharts.com

There's even a word for this

Apophenia: the tendency to mistakenly perceive connections and meaning between unrelated things.

Are these live and simulated delays equivalent?

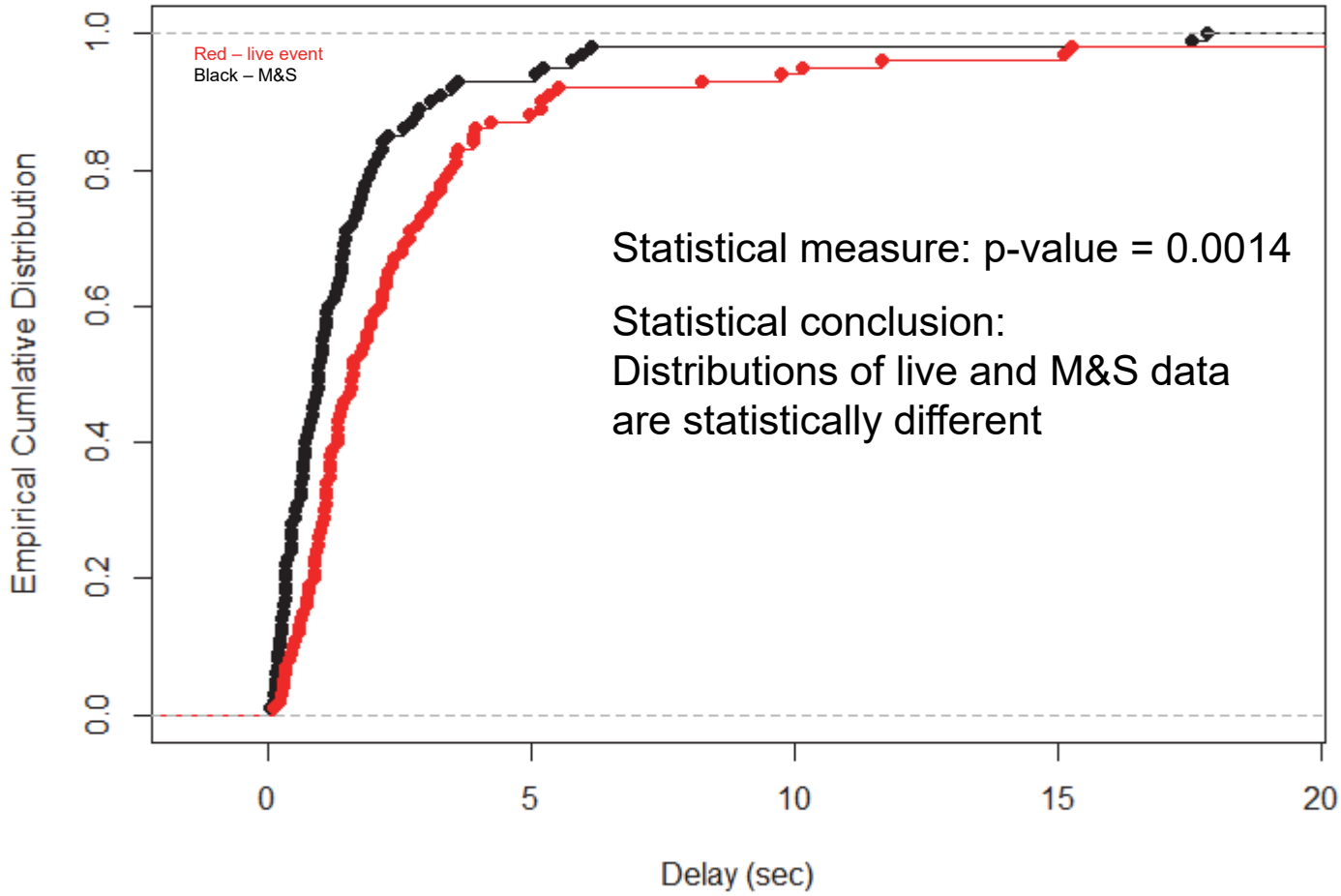


To the eye, the spread in the live and M&S data look pretty similar

M&S: Modeling and Simulation

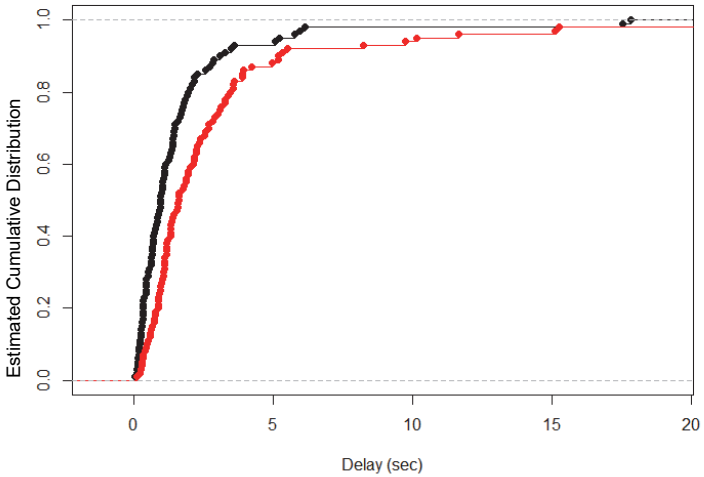
A statistical test can examine whether the distributions are equivalent.

Example of a test for statistical validation (Kolmogorov-Smirnov Test)



Null Hypothesis: M&S outputs match real-world outcomes
Alternative Hypothesis: M&S outputs typically differ from real-world outcomes

Is a statistically significant difference meaningful?



Sim 90% of delays <3.1 sec
Live 90% of delays <5.2 sec

SMEs can help you decide whether the difference affects conclusions

SME: Subject Matter Expert

We should use quantitative comparison methods whenever possible

Work with SMEs to identify:

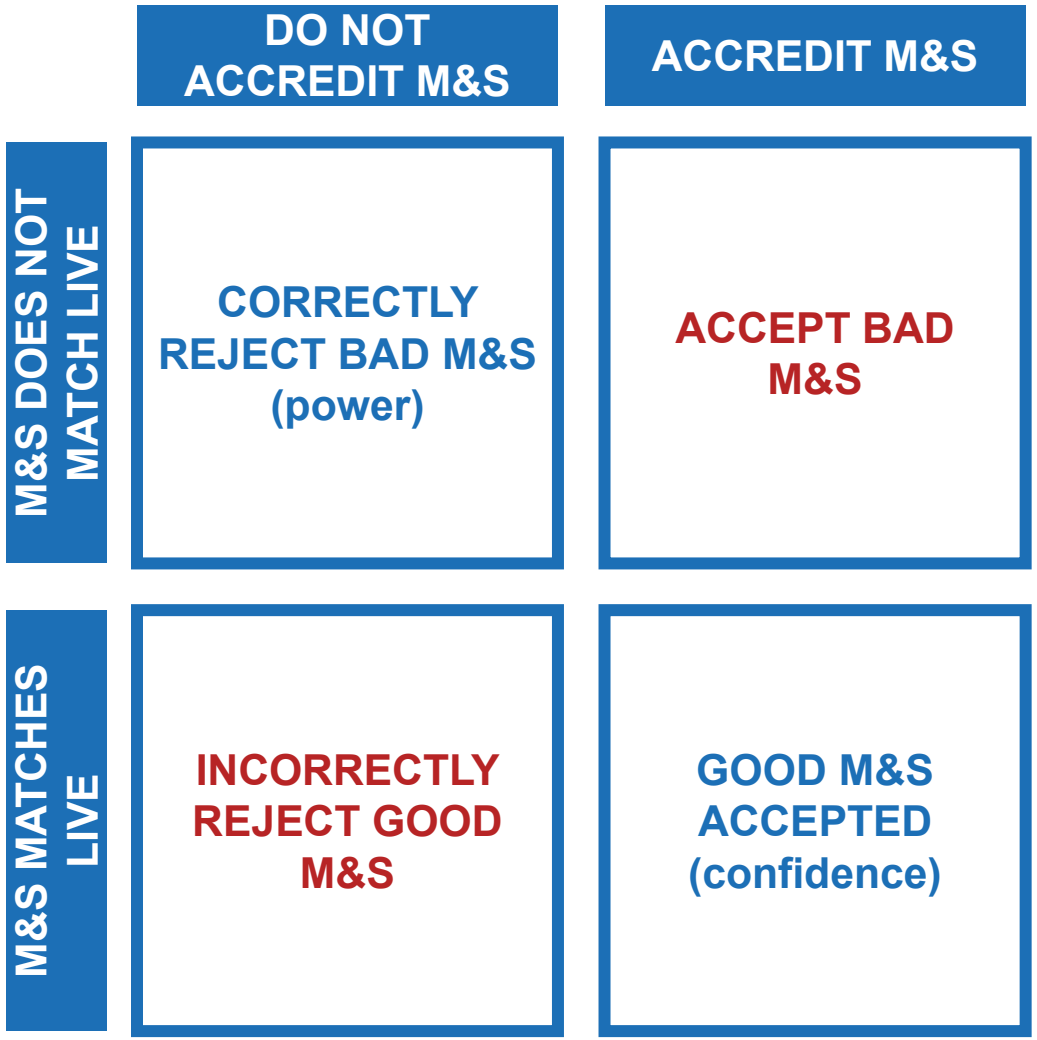
- Key M&S outputs and live data
- Criteria for comparing M&S to live data
- Quantitative methods of comparing data
- Differences that are meaningful

M&S: Modeling and Simulation; SME: Subject Matter Expert

As in live testing, power and confidence apply to M&S validation

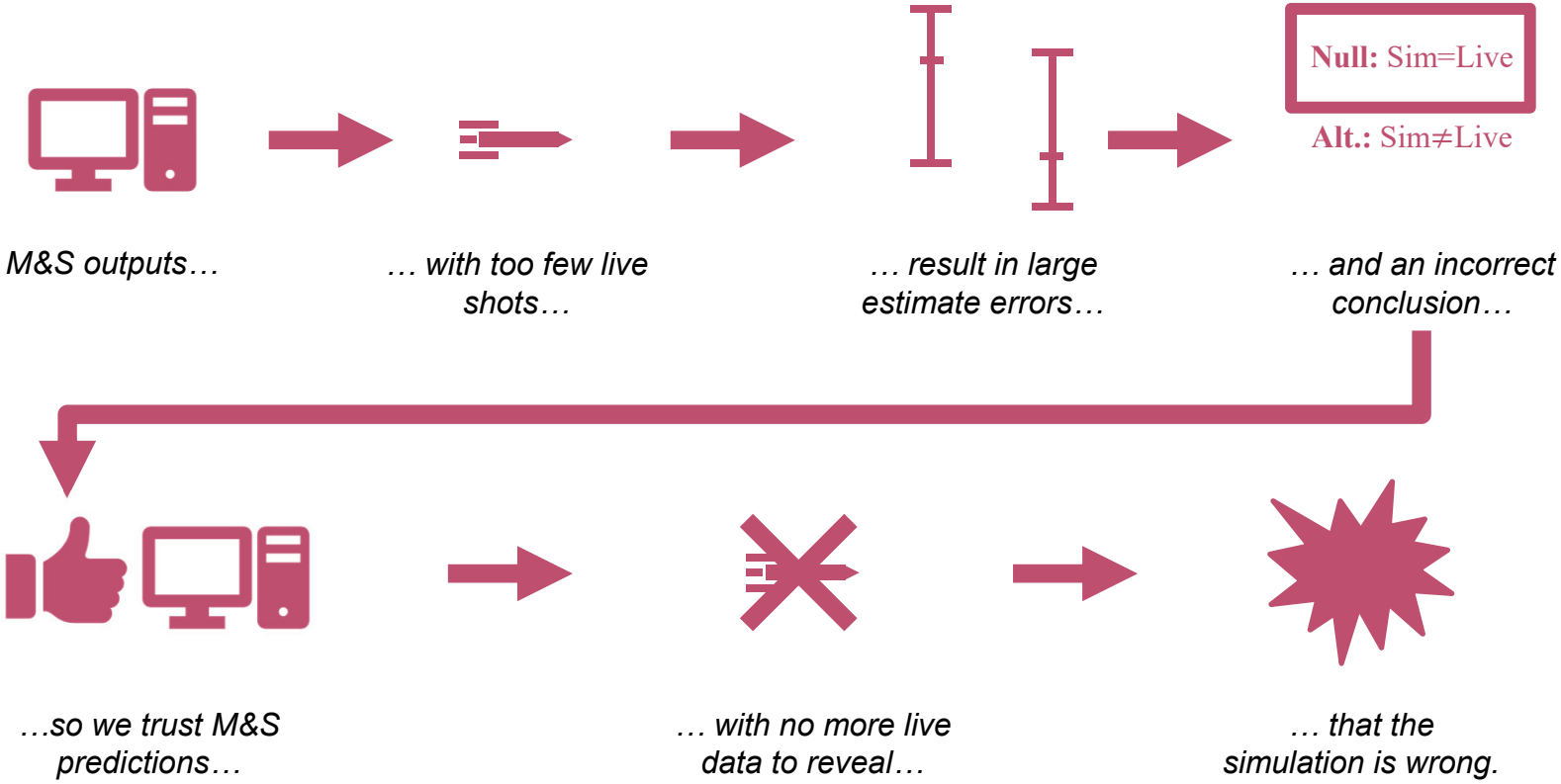
Null Hypothesis: M&S outputs match real-world outcomes

Alternative Hypothesis: M&S outputs typically differ from real-world outcomes



M&S: Modeling and Simulation

Low power can result in unsuitable M&S systems being used for OT assessments

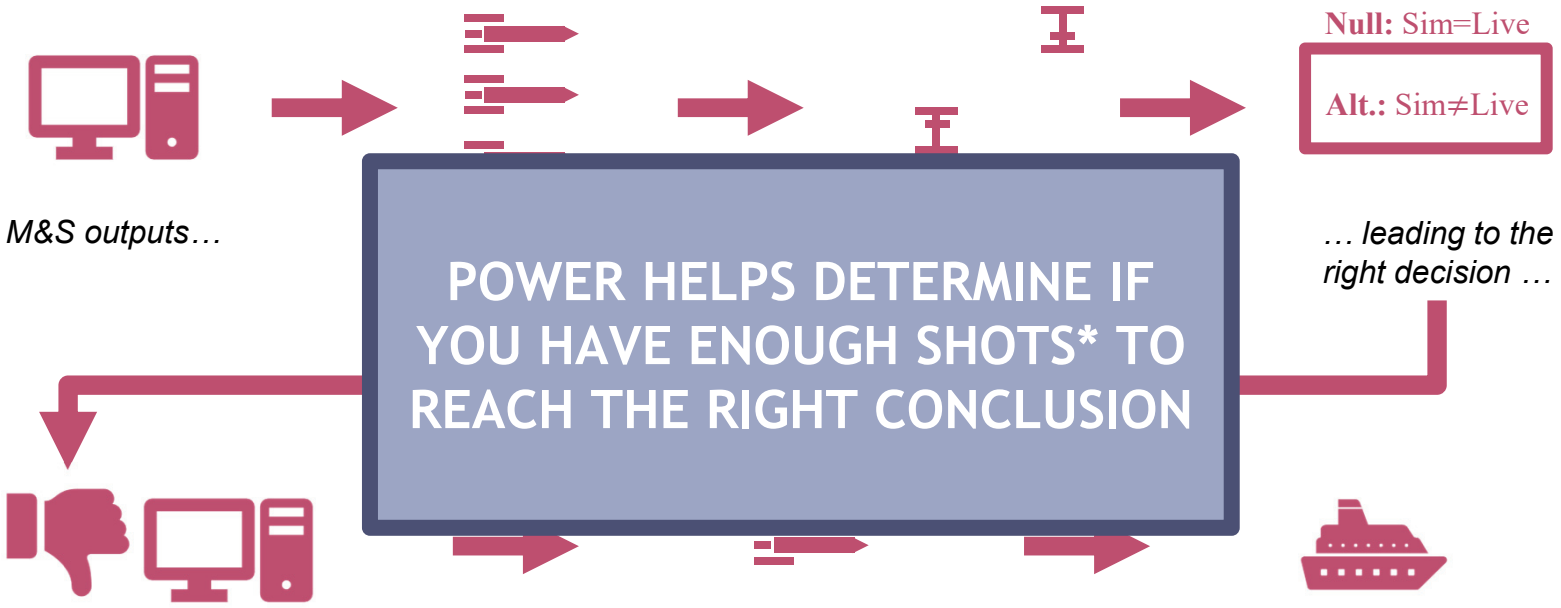


Alt.: Alternative; M&S: Modeling and Simulation; OT: Operational Test

High power reduces the risk of using unsuitable M&S systems for OT assessments

Alt.: Alternative; M&S: Modeling and Simulation; OT: Operational Test

High power reduces the risk of using unsuitable M&S systems for OT assessments



* Among other things; power is not just the number of shots. Also, power is not all we should care about.

Alt.: Alternative; M&S: Modeling and Simulation; OT: Operational Test

Validation and accreditation are not binary!

- The validation process should uncover trends, capabilities, and limitations across the M&S space

- Models are accredited for a specific intended use

Statistics can help facilitate holistic and nuanced analyses and decisions

M&S: Modeling and Simulation

There is inherent uncertainty with any model even after VV&A

M&S uncertainty must be quantified!

Uncertainty Quantification is the rigorous analysis and measurement of variations in mathematical models and simulations, aiding in understanding the consistency of the predictions with the phenomena they represent, and decision-making in complex systems.

M&S: Modeling and Simulation; QOI: Quantity of Interest; VV&A: Verification, Validation, and Accreditation

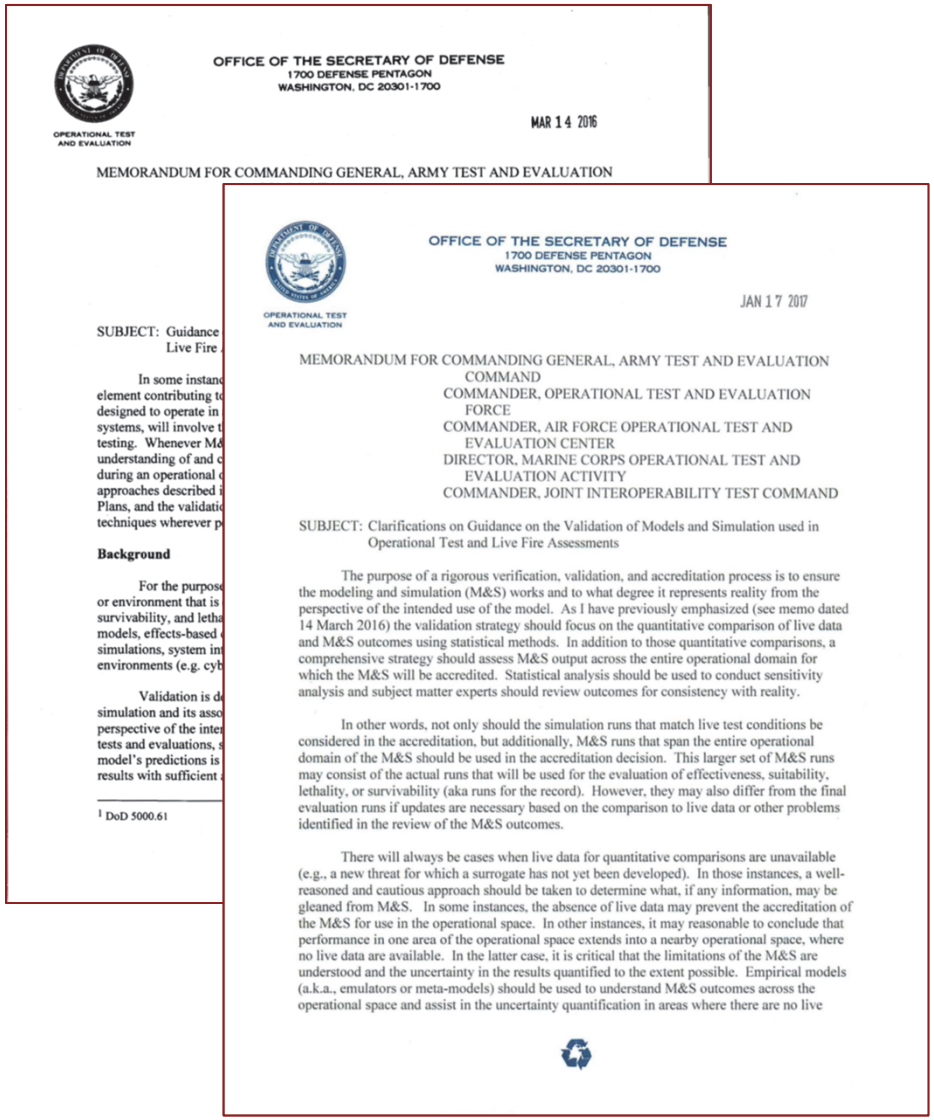
Current DoD/DOT&E Policy & Guidance on M&S V&V

DoD 5000.61

MIL STD 3022

2016 DOT&E memo

2017 DOT&E memo



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

SUBJECT: Guidance
Live Fire

In some instanc
element contributing t
designed to operate in
systems, will involve t
testing. Whenever M&
understanding of and c
during an operational
approaches described i
Plans, and the validati
techniques wherever p

Background

For the purpos
or environment that is
survivability, and letha
models, effects-based
simulations, system in
environments (e.g. cyb

Validation is d
simulation and its asso
perspective of the inte
tests and evaluations,
model's predictions is
results with sufficient

[1] DoD 5000.61

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

OPERATIONAL TEST AND EVALUATION

JAN 17 2017

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

The purpose of a rigorous verification, validation, and accreditation process is to ensure the modeling and simulation (M&S) works and to what degree it represents reality from the perspective of the intended use of the model. As I have previously emphasized (see memo dated 14 March 2016) the validation strategy should focus on the quantitative comparison of live data and M&S outcomes using statistical methods. In addition to those quantitative comparisons, a comprehensive strategy should assess M&S output across the entire operational domain for which the M&S will be accredited. Statistical analysis should be used to conduct sensitivity analysis and subject matter experts should review outcomes for consistency with reality.

In other words, not only should the simulation runs that match live test conditions be considered in the accreditation, but additionally, M&S runs that span the entire operational domain of the M&S should be used in the accreditation decision. This larger set of M&S runs may consist of the actual runs that will be used for the evaluation of effectiveness, suitability, lethality, or survivability (aka runs for the record). However, they may also differ from the final evaluation runs if updates are necessary based on the comparison to live data or other problems identified in the review of the M&S outcomes.

There will always be cases when live data for quantitative comparisons are unavailable (e.g., a new threat for which a surrogate has not yet been developed). In those instances, a well-reasoned and cautious approach should be taken to determine what, if any information, may be gleaned from M&S. In some instances, the absence of live data may prevent the accreditation of the M&S for use in the operational space. In other instances, it may reasonable to conclude that performance in one area of the operational space extends into a nearby operational space, where no live data are available. In the latter case, it is critical that the limitations of the M&S are understood and the uncertainty in the results quantified to the extent possible. Empirical models (a.k.a., emulators or meta-models) should be used to understand M&S outcomes across the operational space and assist in the uncertainty quantification in areas where there are no live

M&S: Modeling and Simulation; V&V: Verification and Validation

Update to policy is in progress via the DoDM on M&S VV&A in OT&E and LFT&E

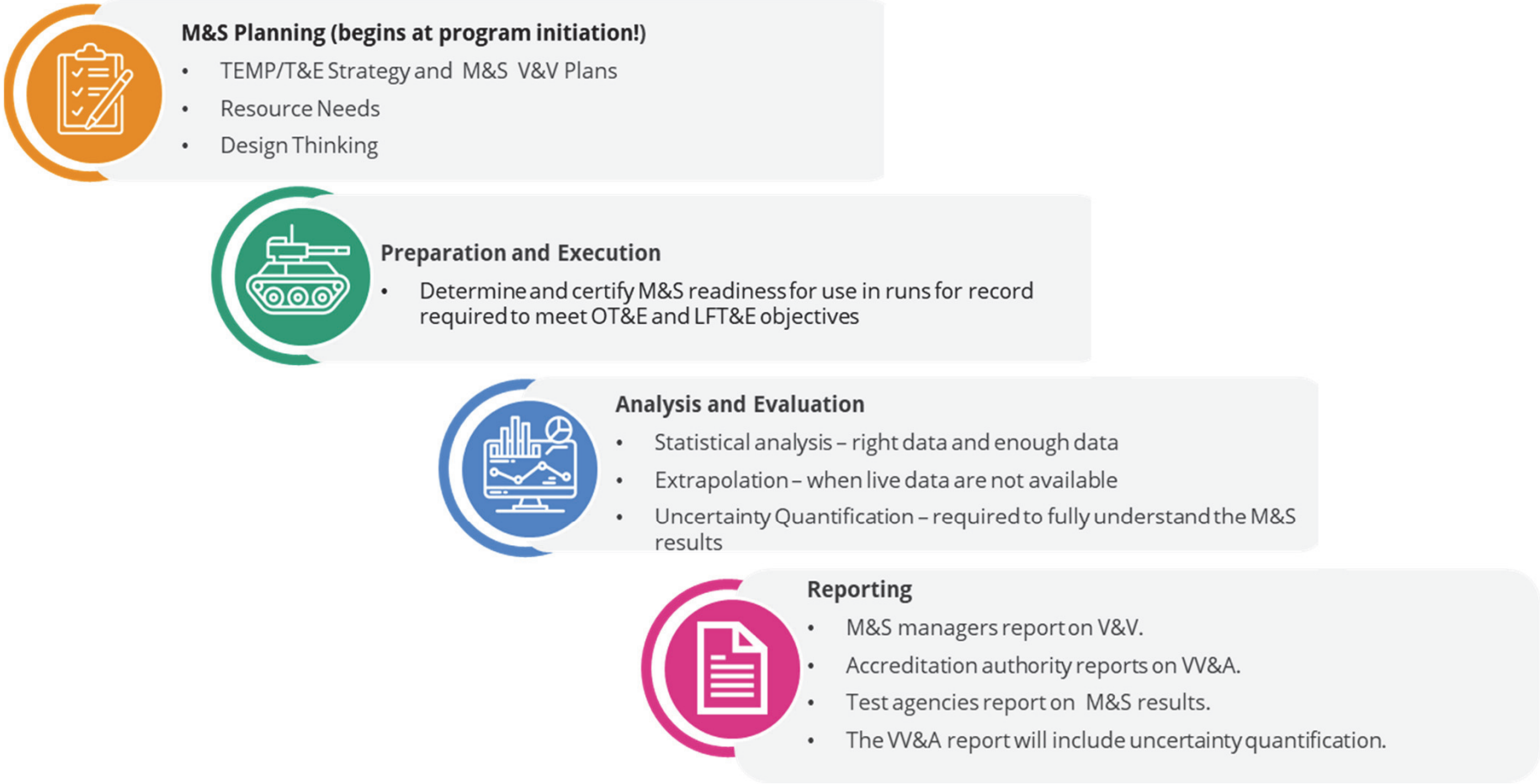
This DoDM assigns responsibilities, and provides procedures for verification, validation, and accreditation (VV&A) of modeling and simulation (M&S) tools critical to meeting the operational test and evaluation (OT&E) and live fire test and evaluation (LFT&E) objectives of DoD systems and services acquired via the Defense Acquisition System.

Major change:

- DOT&E is the approval authority for the M&S and V&V strategy, outlined in the TEMP/T&E strategy, and the M&S V&V plans supporting the M&S VV&A for OT&E and LFT&E.

TEMP: Test and Evaluation Master Plan

The M&S VV&A process outlined in the DoDM continues to emphasize scientific and statistical best practices



LFT&E: Live Fire Test and Evaluation; M&S: Modeling and Simulation; OT&E: Operational Test and Evaluation; TEMP: Test and Evaluation Master Plan; VV&A: Verification, Validation, and Accreditation

Outline

1. Motivation and Statistical Concepts
2. Statistical Methods for V&V
3. Example with Code

Statistical validation analyses fall into two broad categories

1. Exploration of the M&S space on its own
   - Variation analysis
   - Sensitivity analysis
   - Emulators/Meta-models

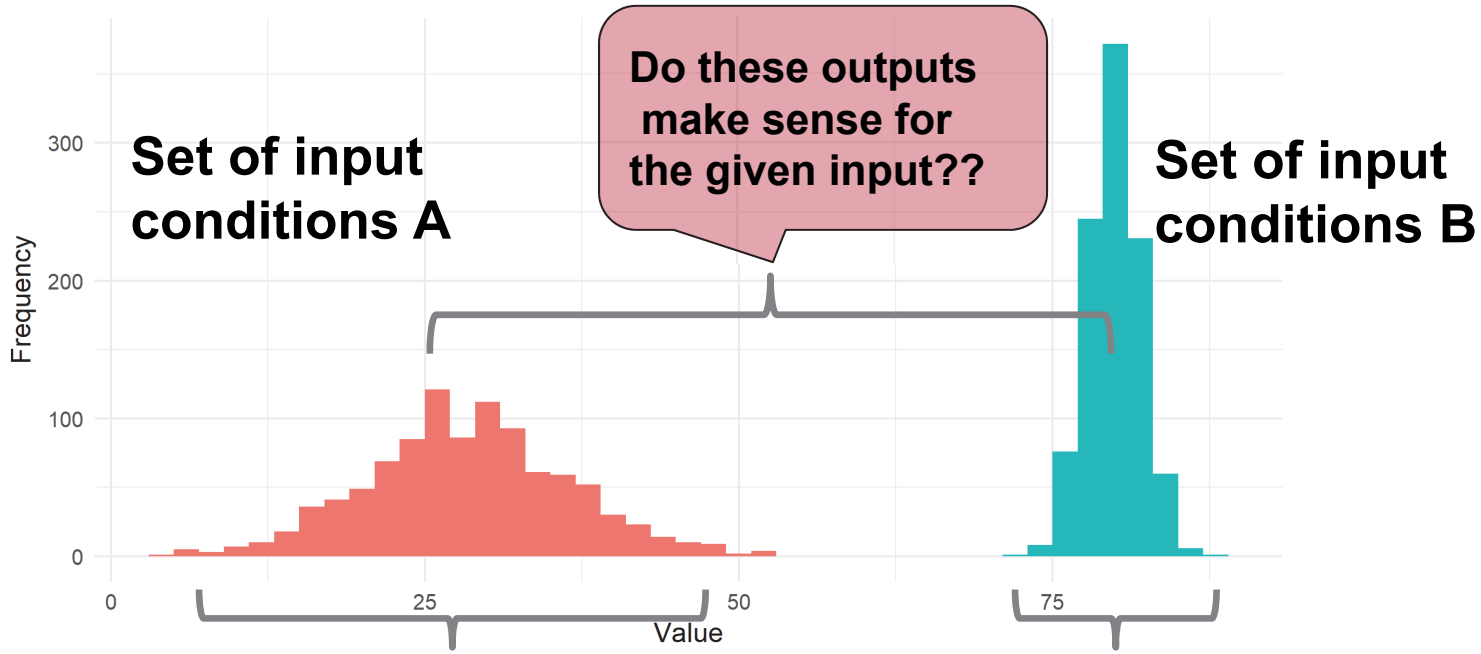
2. A comparison of the M&S to live-test data (or the highest fidelity data available at the time)
   - Distribution tests
   - Regression-based approaches

The right data collection strategy (test design) depends on the analysis goal above

Exploring the M&S Space

We should fully understand and evaluate the behavior of the model itself.

- Variation Analysis of Monte Carlo variables
  - Run the simulation multiple times under the same conditions
- Sensitivity Analysis
  - Compare simulation outcomes under different conditions

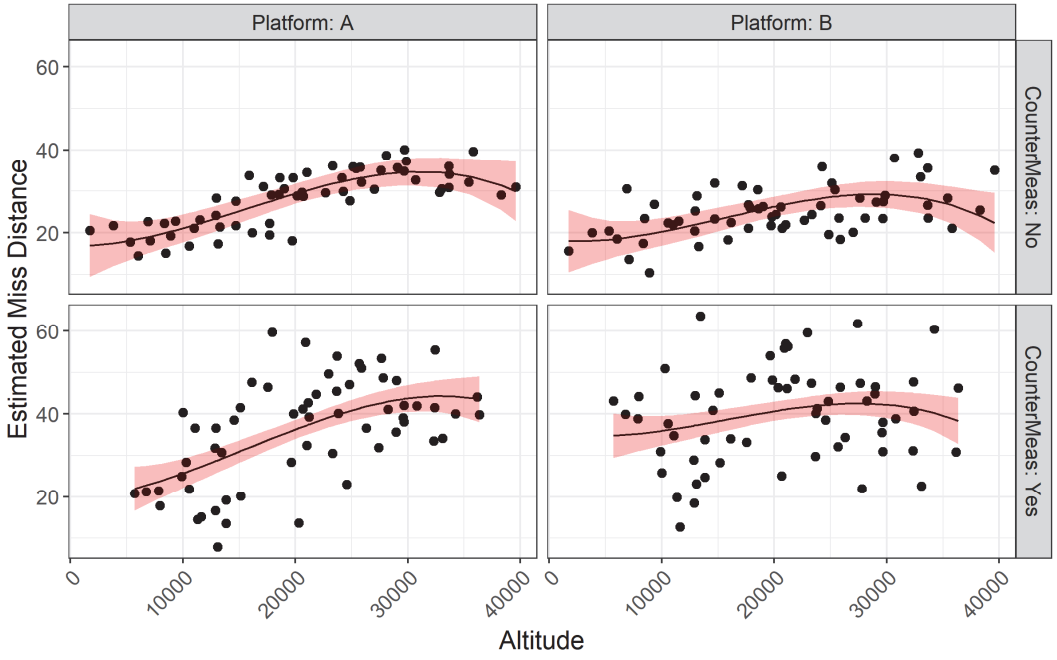


Variation analyses can be used to test for model robustness, scope test designs, and identify risk areas.

Meta-models (emulators) can be used to predict the output of the simulation at both tested and untested conditions

- Any appropriate statistical model can be an emulator (depends on the type and fidelity of the simulation)
- Supports prediction and uncertainty quantification across the M&S space

- Informs the choice of live test points
- In some cases may also serve as a surrogate for the M&S itself, which can save time and cut costs by avoiding the need to re-run the simulation over and over



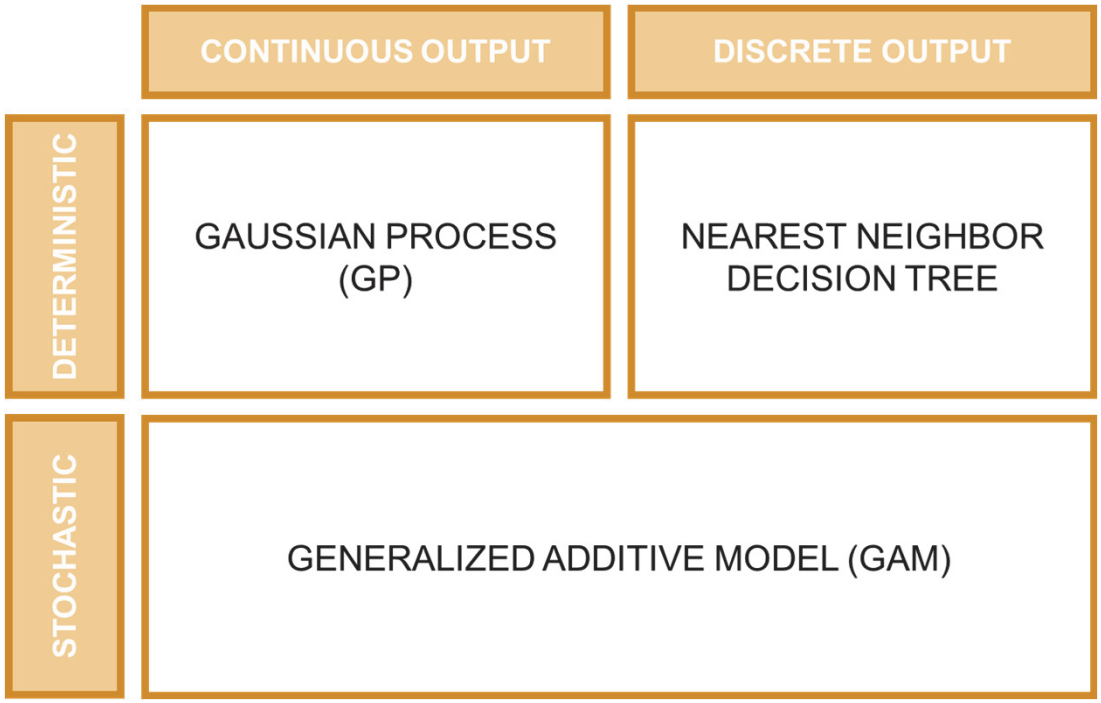
Note: All data are notional

M&S: Modeling and Simulation

We recommend different analysis procedures based on the M&S output

Deterministic models produce the same exact results for a particular set of inputs

Stochastic models possess some level of inherent randomness; the same inputs do not necessarily produce the same result



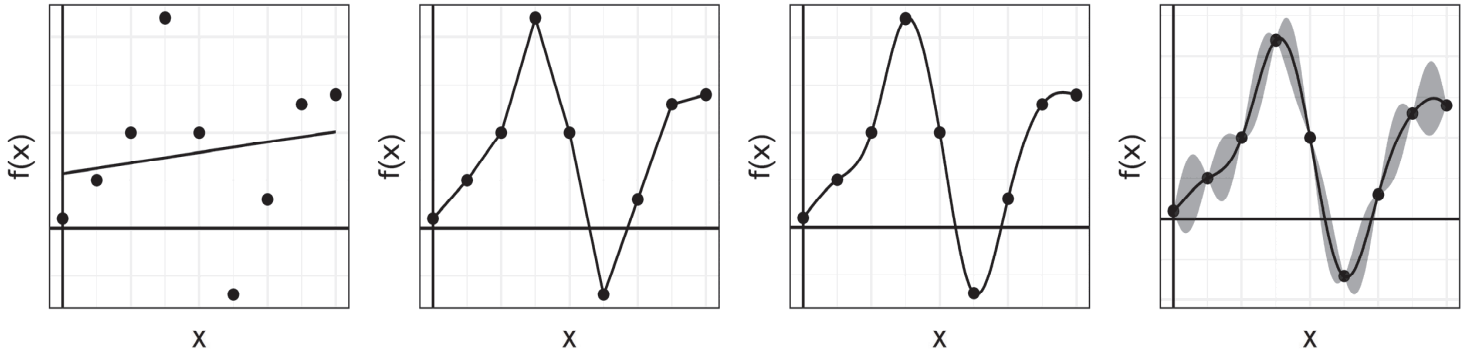
Source: *Metamodeling Techniques for Verification and Validation of Modeling and Simulation Data*, C.Miller and J. Haman, testscience.org

M&S: Modeling and Simulation

Statistical methods for analyzing most M&S outputs should interpolate and quantify uncertainty

Goals: Interpolate across a complex space

Quantify uncertainty at unobserved points



Linear Regression

Basic Interpolators and Splines

GPs/GAMs

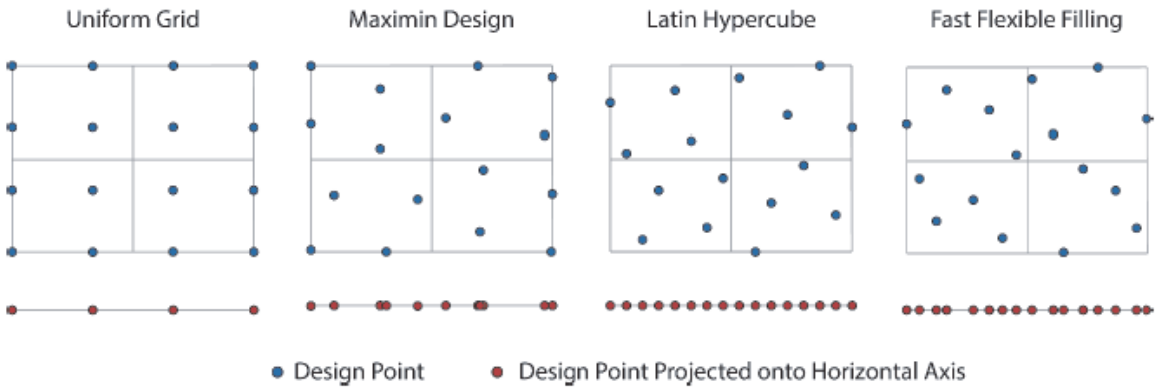
M&S: Modeling and Simulation

Test design(s) for M&S should facilitate this thorough understanding of the model itself

Design Properties

- Fill the M&S space
- Consider all input parameters (even those not able to be varied in live test)
- Consider the type of output expected (e.g. nonlinearities, curvature) when deciding on the number of points

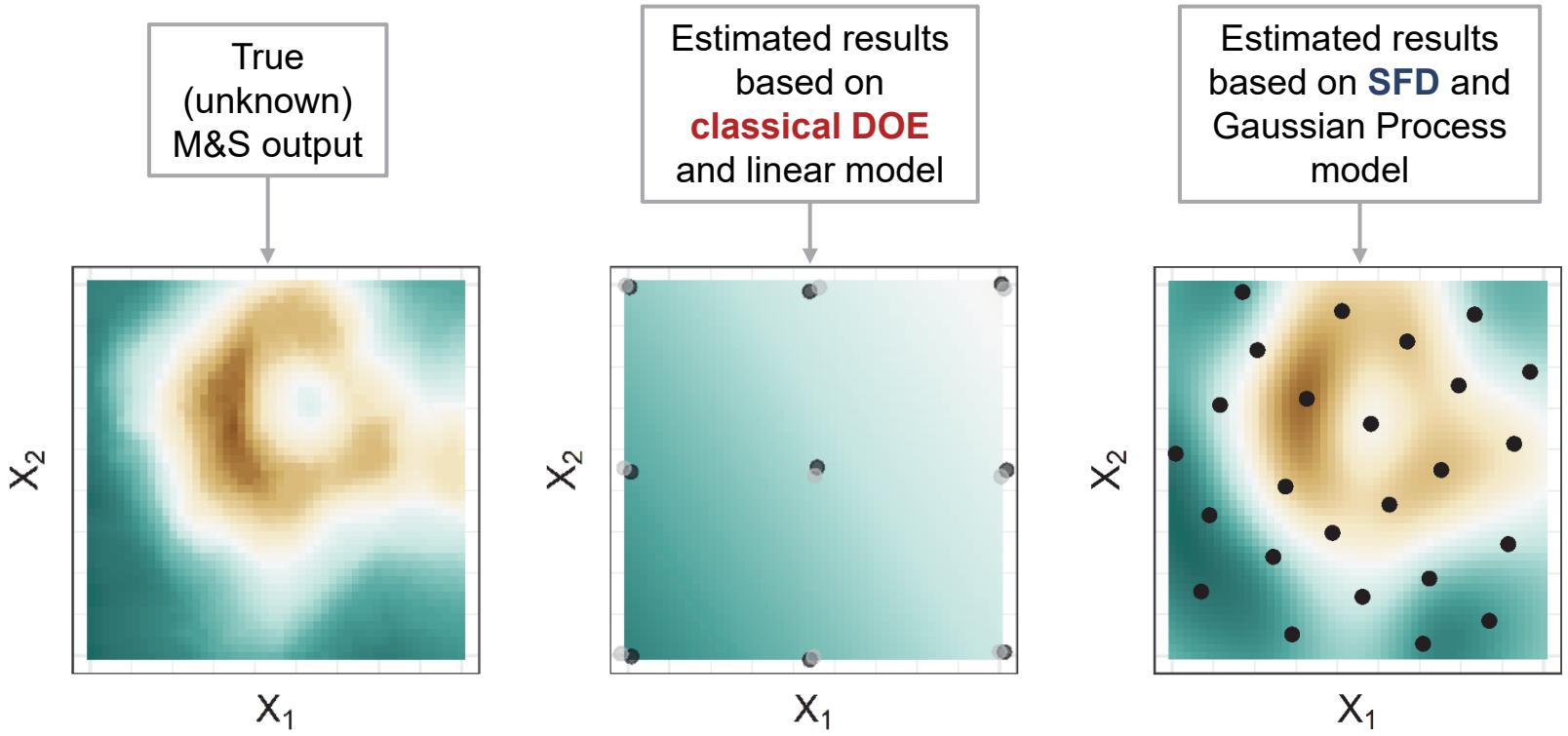
Space-Filling Designs are recommended



Space-Filling Designs and associated analyses are lesser known and infrequently used in the T&E community

M&S: Modeling and Simulation; T&E: Test and Evaluation

Space-Filling Designs capture M&S behavior more effectively than Classical DOE, without requiring additional test resources



Failing to understand M&S behavior means testers may include inaccurate predictions about system performance in their reports

Source: *Space-Filling Designs for Modeling & Simulation*, H. Yi, C. Miller, and K. Avery, testscience.org

DOE: Design of Experiments; M&S: Modeling and Simulation; SFD: Space-Filling Design

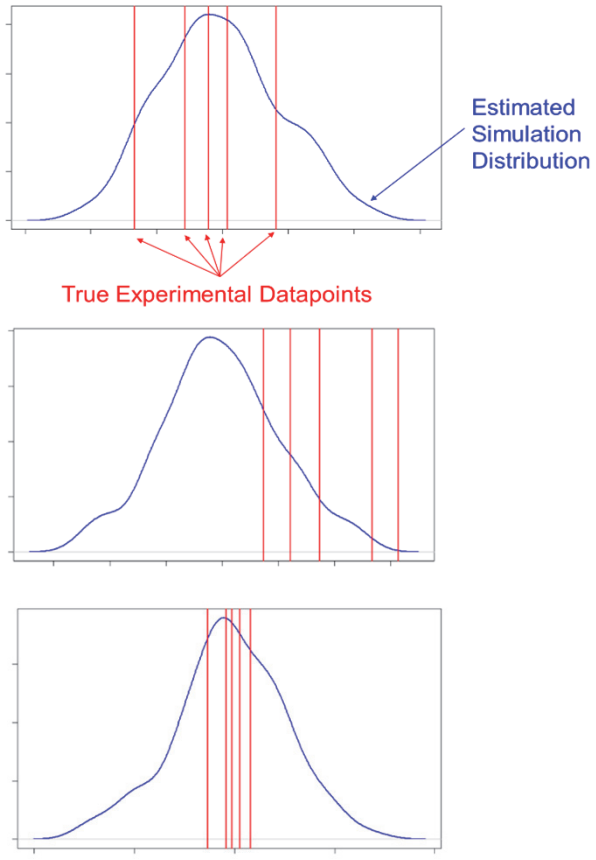
Comparing Live and M&S Data

There are dozens of techniques for comparing live data and simulation output.

H0: Simulation output matches the live data

HA: Simulation output does not match the live data

- Differences could be in terms of mean, variance, distribution, etc.
- Statistical techniques can be crucial for detecting differences in the presence of multiple factors and interactions or noisy data.
- We typically have more simulation data than live data.



Goal is to detect and quantify differences between live and sim, while accounting for other operational factors.

There is no single "best" method for this validation comparison.

The most effective method depends on:

- The distribution of the response variable
- The amount of live data (focus is on relatively small live-data sets)
- Presence or absence of factors/whether or not a statistically designed experiment was conducted

Distribution	Structure Of Factors	Small Sample Sizes	Moderate Samples Sizes	Large Sample Sizes
Skewed	Univariate			
	Distributed Level Effects			
	Designed Experiment			
Symmetric	Univariate			
	Distributed Level Effects			
	Designed Experiment			
Binary	Univariate			
	Distributed Level Effects			
	Designed Experiment			

*A small sample size is 2 to 4 samples for continuous data or 20 for binary data; medium size is 6 to 10 samples for continuous data or 50 for binary data; large size is 11 to 20 samples for continuous data or 100 for binary data.

Multiple techniques may be useful or necessary to fully understand the strengths and weaknesses of the M&S.

M&S: Modeling and Simulation

IDA conducted power simulation studies of relatively simple techniques to generate recommendations.

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, lognormal, and Poisson) 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

		Recommended Method by Sample Size		
Distribution	**Factors**	Small	Medium	Large
Skewed (Lognormal)	Univariate	Fisher's Combined	Log t-test Fisher's Combined Non-parametric K-S	Log t-test Fisher's Combined Non-parametric K-S
	Distributed	Log t-test Fisher's Combined Non-parametric K-S	Non-parametric K-S	Non-parametric K-S
	Designed Experiment	Log-Normal Regression Emulation & Prediction	Log-Normal Regression Emulation & Prediction	Log-Normal Regression Emulation & Prediction
Symmetric (Normal)	Univariate	Fisher's Combined	t-test Fisher's Combined Non-parametric K-S	t-test Fisher's Combined Non-parametric K-S
	Distributed	t-test Fisher's Combined Non-parametric K-S	Non-parametric K-S	Non-parametric K-S
	Designed Experiment	Regression Emulation & Prediction	Regression Emulation & Prediction	Regression Emulation & Prediction
Binary	Univariate	Fisher's Exact	Fisher's Exact	Fisher's Exact
	Distributed	Logistic Regression	Logistic Regression	Logistic Regression
	Designed Experiment	Logistic Regression	Logistic Regression	Logistic Regression

This is not final or exhaustive; different and more advanced techniques may be appropriate and perform even better!

M&S: Modeling and Simulation

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.

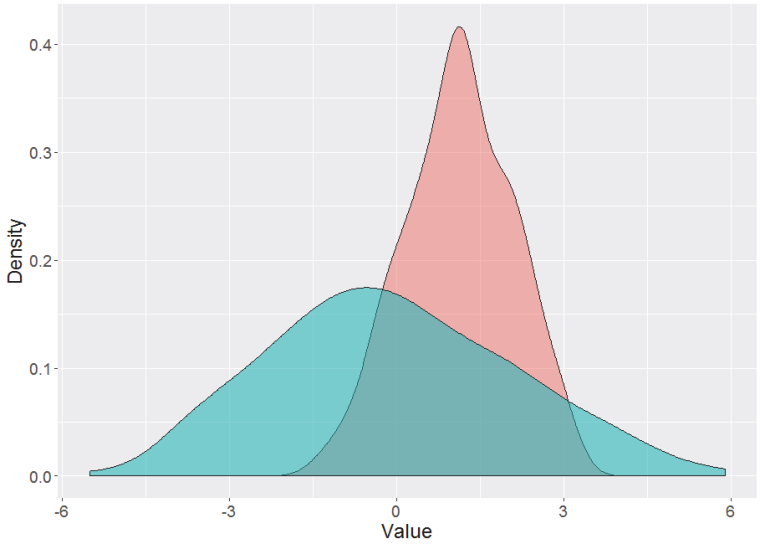
- Parametric: Compare each of the data sets (live and sim) to a *parametric reference distribution* (e.g., normal).

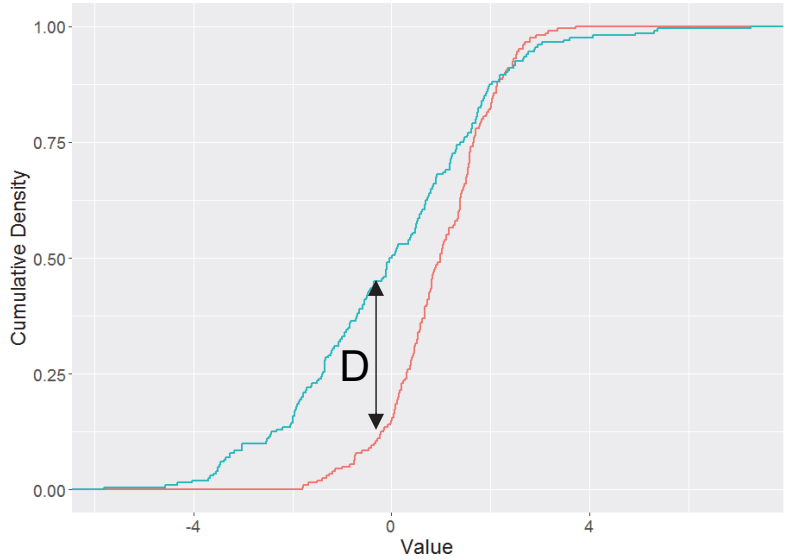
Works better for our problem

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

$$D_{m,n} = \max_x |F(x) - G(x)|$$

where F(x) is the observed CDF of a sample with size m and G(x) is the CDF of a sample of size n



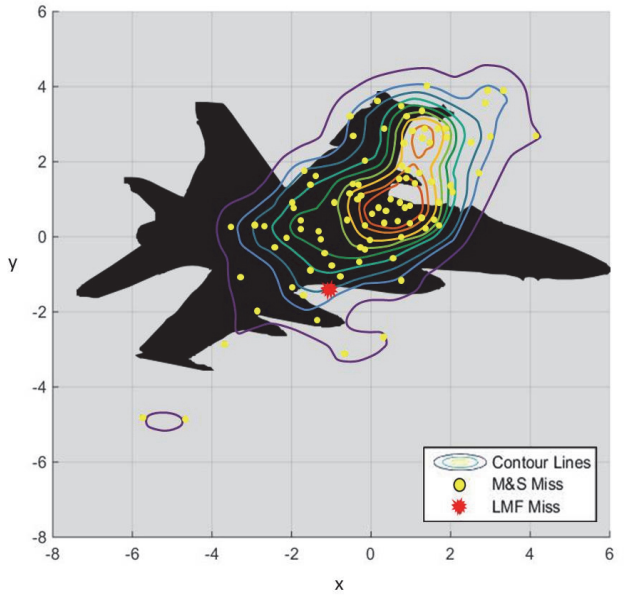


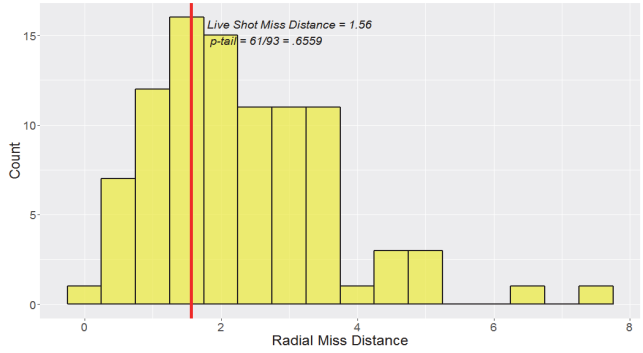
Note: All data are notional

CDF: Cumulative Density Function; M&S: Modeling and Simulation

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 may 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

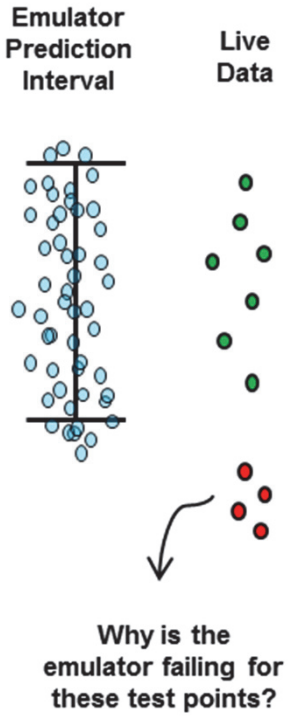
- Use bootstrapped regression if sample sizes are severely unbalanced to avoid correlation issues.

- Always keep in mind that practical significance (raw magnitude of the effect) matters in addition to statistical significance.

M&S: Modeling and Simulation

Emulation and Prediction

- Build a meta-model (i.e., statistical emulator) 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.
- The method works best if you used a designed experiment.
  - Strength is detecting differences in variance.
- Works well even when there is limited data.

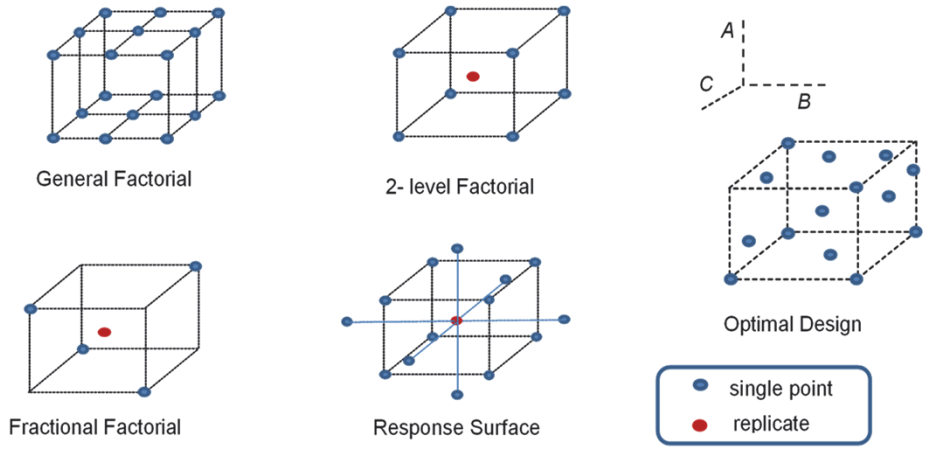


Test design(s) for M&S should facilitate this comparison with available live test data

Design Properties:

- Match the live test points (possibly with replicates)
- Support building a statistical model
- High power to detect differences between live and M&S

Classical DOE is recommended



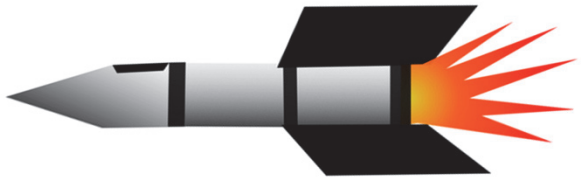
These design and analysis techniques are already widely understood and used

DOE: Design of Experiments; M&S: Modeling and Simulation

Outline

1. Motivation and Statistical Concepts
2. Statistical Methods for V&V
3. Example with Code

Consider a test of an air-to-ground missile.



- A key metric of interest is miss distance.
- Factors that might affect miss distance include:
  - The specific aircraft platform from which the missile is launched (A or B)
  - The altitude of the aircraft (Low, Medium, or High)
  - Whether or not countermeasures were in use (Yes or No)
- Only 12 missiles are available for live testing – one test was conducted in each of the 12 conditions (2 platforms x 3 altitudes x countermeasure on/off)
- For each condition executed during live test, 100 replicates were conducted in the M&S.
- The primary goal is to compare M&S results to live data.

M&S: Modeling and Simulation

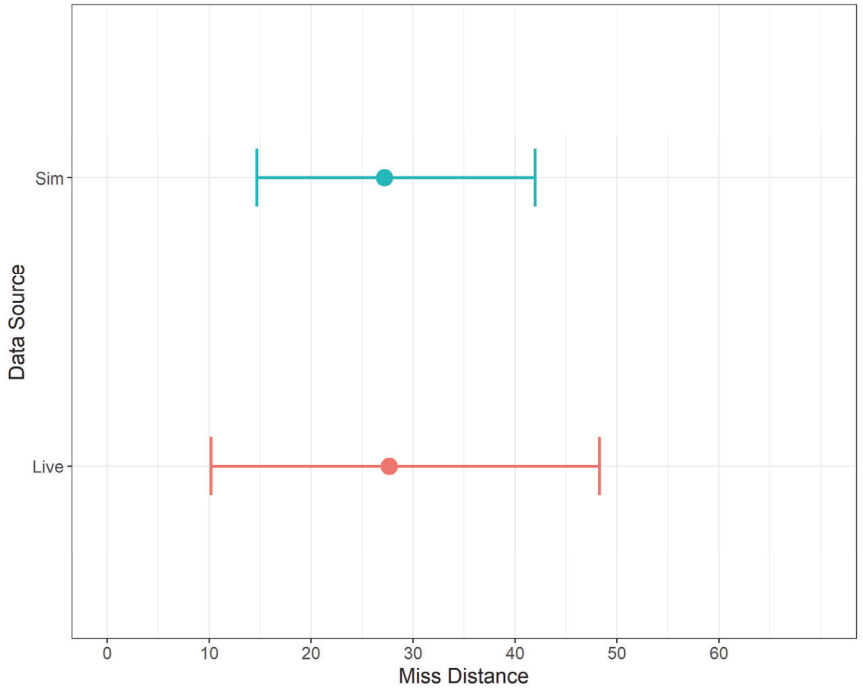
We will demonstrate multiple statistical techniques using this data set and show why some are better than others

		Recommended Method by Sample Size		
Distribution	**Factors**	Small	Medium	Large
Skewed (Lognormal)	Univariate	Fisher's Combined	Log t-test Fisher's Combined Non-parametric K-S	Log t-test Fisher's Combined Non-parametric K-S
	Distributed	Log t-test Fisher's Combined Non-parametric K-S	Non-parametric K-S	Non-parametric K-S
	Designed Experiment	Log-Normal Regression Emulation & Prediction	Log-Normal Regression Emulation & Prediction	Log-Normal Regression Emulation & Prediction
Symmetric (Normal)	Univariate	Fisher's Combined	t-test Fisher's Combined Non-parametric K-S	t-test Fisher's Combined Non-parametric K-S
	Distributed	t-test Fisher's Combined Non-parametric K-S	Non-parametric K-S	Non-parametric K-S
	Designed Experiment	Regression Emulation & Prediction	Regression Emulation & Prediction	Regression Emulation & Prediction
Binary	Univariate	Fisher's Exact	Fisher's Exact	Fisher's Exact
	Distributed	Logistic Regression	Logistic Regression	Logistic Regression
	Designed Experiment	Logistic Regression	Logistic Regression	Logistic Regression

The example shows snippets of R code, but many other software packages can perform these techniques. There are also web-based apps in many cases.

Comparing means

An initial statistical look at the data might involve testing whether the mean miss distance from the live data is equal to that of the simulation data.



The t-test does not reject, meaning that the mean miss distance of the live data is not statistically different than the mean miss distance of the simulation data.

However, this test ignores the variance of the data, and does not account for the test conditions in the DOE.

DOE: Design of Experiments

Comparing means

The t.test function in R is built in to base R

```
> t.test(x = filter(dat, Source=="Sim")$MissDist,
+        y = filter(dat, Source=="Live")$MissDist)

        Welch Two Sample t-test

data:  filter(dat, Source == "Sim")$MissDist and filter(dat, Source == "Live")$MissDist
t = -0.14392, df = 11.108, p-value = 0.8881
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.128540  6.252517
sample estimates:
mean of x mean of y
 27.22866  27.66667
```

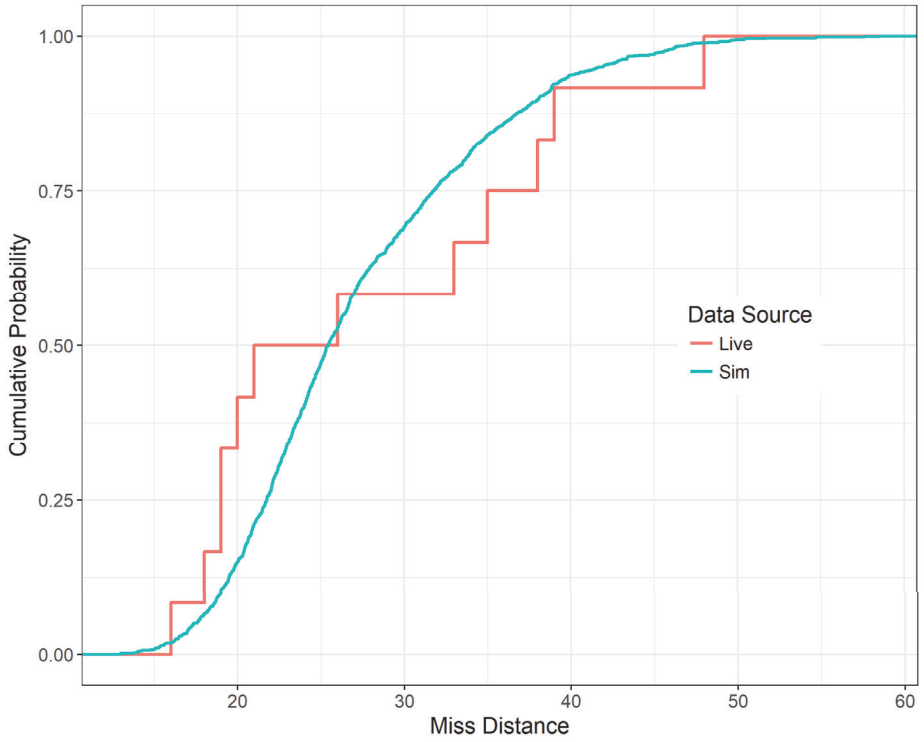
High p-value indicates the difference in means is NOT significant

Comparing distributions

A next statistical look at the data might be to compare the overall miss distance distribution of the live data to that of the simulation data.

Here the variance of the data is considered, but test conditions are still completely ignored.

The K-S test fails to reject the null hypothesis that the two data sets come from the same distribution.



Comparing distributions

The ks.test function in R is built in to base R

```
> ks.test(x = filter(dat, Source=="Live")$MissDist,
+         y = filter(dat, Source=="Sim")$MissDist)

        Two-sample Kolmogorov-Smirnov test

data:  filter(dat, Source == "Live")$MissDist and filter(dat, Source == "Sim")$MissDist
D = 0.31083, p-value = 0.2011
alternative hypothesis: two-sided
```

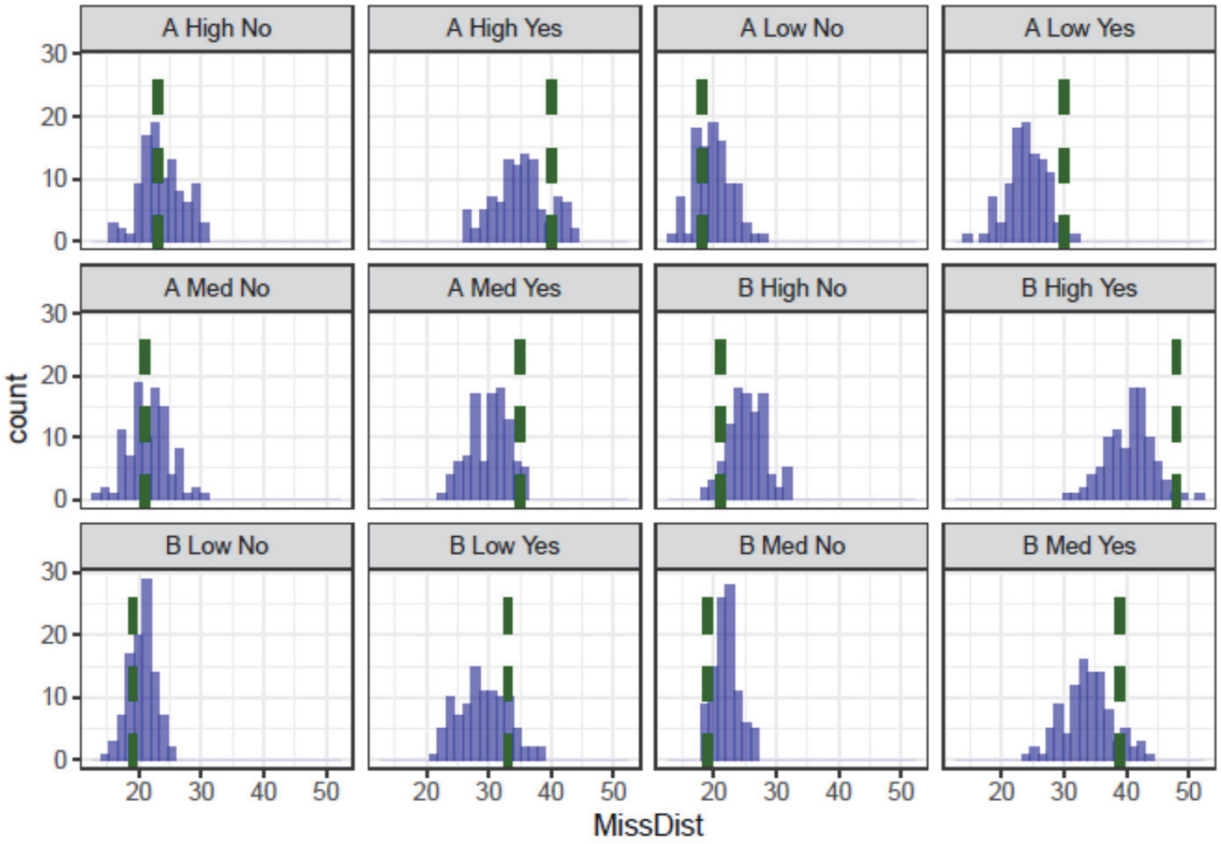
High p-value indicates the difference in distributions is NOT significant

K-S Test app available on testscience.org!

Comparing distributions across conditions

Fisher's test takes a stove-piped look at each of the 12 conditions and then attempts to summarize these individual findings using a global hypothesis test.

The results of this test are a bit more ambiguous.



Comparing distributions across conditions

```
####### Fisher's Test

# Enumerate conditions
conditions <- expand.grid(Platform = c("A", "B"), Altitude = c("Low", "Med", "High"),
            CounterMeas = c("No", "Yes"))

# Calculate the percentage of simulation runs are greater than the single live run for each test condition
ptails <- vector(length = nrow(conditions))

for(i in 1:nrow(conditions)) {
ptails[i] <-
  dat %>%
  filter(Platform == conditions$Platform[i],
         Altitude == conditions$Altitude[i],
         CounterMeas == conditions$CounterMeas[i]) %>%
  select(MissDist, Source) %>%
  summarize(quant = length(MissDist[1:nsimu][MissDist > MissDist[nsimu+nlive]])) / (nsimu+nlive)
}
ptails <- as_tibble(do.call(rbind, ptails))

fish_dat <- cbind(conditions, ptails) %>%
  rename("pvalues"="V1")
```

Calculating each of the individual p-values (p-tails) takes a little bit of coding

```
> # Fisher's combined probability test
> fisher_test_stat <- -2*sum(log(fish_dat$pvalues))
> pchisq(fisher_test_stat, df = 2*length(fish_dat$pvalues), lower.tail = F)
[1] 0.2738278
```

Fisher's test statistic is fairly straightforward to code

```
> # Kolmogorov Smirnov test for uniformity of p-values
> ks.test(x = fish_dat$pvalues,
+         y = "punif", 0, 1)

        One-sample Kolmogorov-Smirnov test

data:  fish_dat$pvalues
D = 0.33168, p-value = 0.1119
alternative hypothesis: two-sided
```

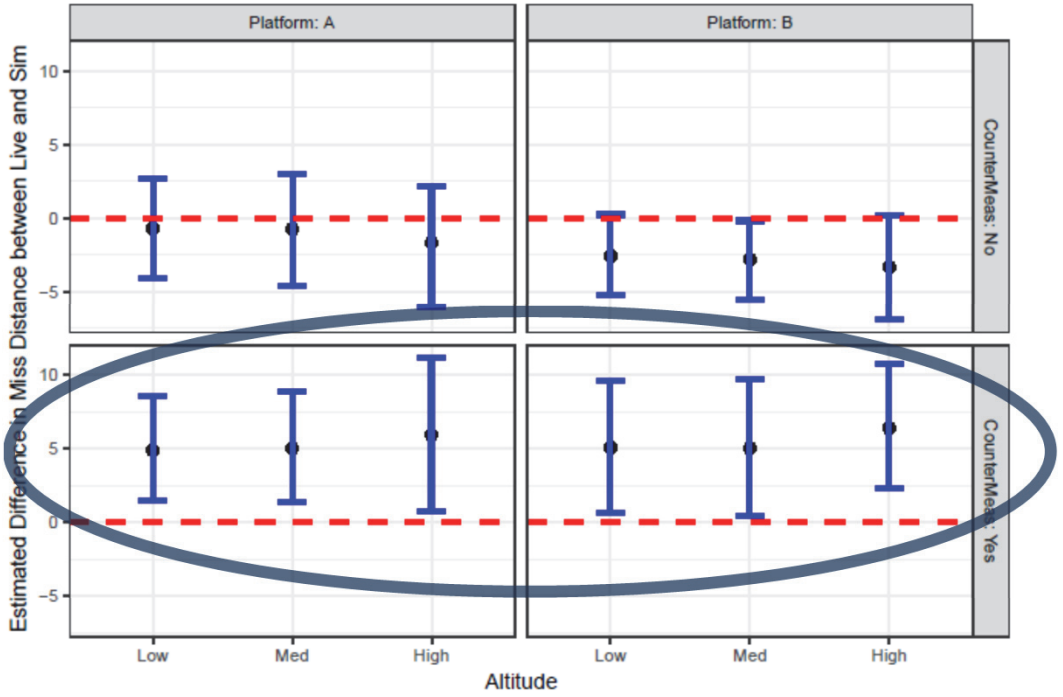
We can also use the KS test to test for uniformity of p-values

Regression techniques can uncover what simple techniques will miss.

Bootstrapped regression 1) controls for differences in sample size between live and sim and 2) mathematically accounts for two-way (or higher) interactions between factors in the test design.

Error bars on either side of zero indicate significant difference between live and sim!

The M&S tends to underestimate miss distance in the presence of countermeasures, but performs better for the no-countermeasures case.



M&S: Modeling and Simulation

Regression techniques

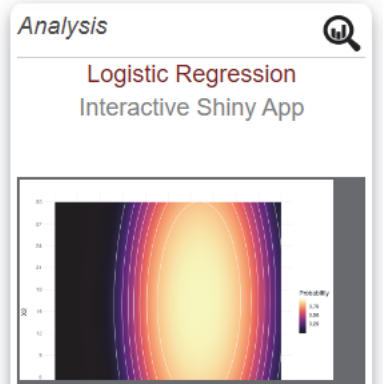
```r
# create list of nboot data sets, matching live and sim sample size
for (i in 1:length(dat_list)){
  dat_list[[i]] <- dat %>%
    split(list(.$Platform, .$Altitude, .$CounterMeas, .$Source)) %>%
    map(~ .x[sample(1:nrow(.x), 1, TRUE), ]) %>%
    do.call(rbind, .) %>%
    mutate(ID=i)
}
```

```r
# Run regression on each of nboot data sets
model <- dat_list %>%
  map(~lm(MissDist ~ (Platform + Altitude + CounterMeas + Source)^3, data = .x))
```

```r
# Save model terms and p-values
terms <- model %>%
  map(broom::tidy)  %>% do.call(rbind, .) %>%
  filter(term != "(Intercept)")
```

```r
# Compute median p-value and quantiles for each term
mu <- plyr::ddply(terms, "term", summarise,
                  p.median = median(p.value),
                  lb.estimate = quantile(estimate, 0.1),
                  ub.estimate = quantile(estimate, 0.9))
```

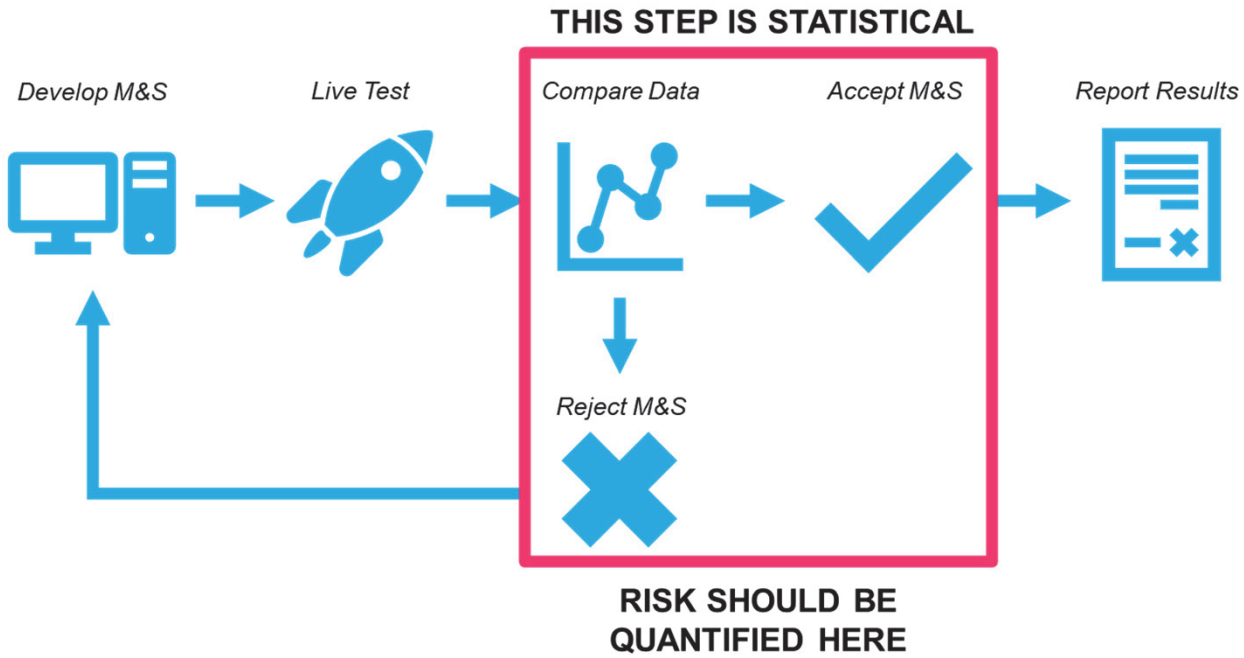
Tidyverse packages in R are useful for bootstrapping, model fitting, and results summaries



Logistic Regression app available on testscience.org!

Conclusions

- Statistical analysis techniques are essential pieces of a rigorous M&S validation strategy
  - Some analysis methods are better than others in identifying trends and uncovering limitations in M&S performance
  - Design of experiments techniques facilitate these analyses



M&S: Modeling and Simulation

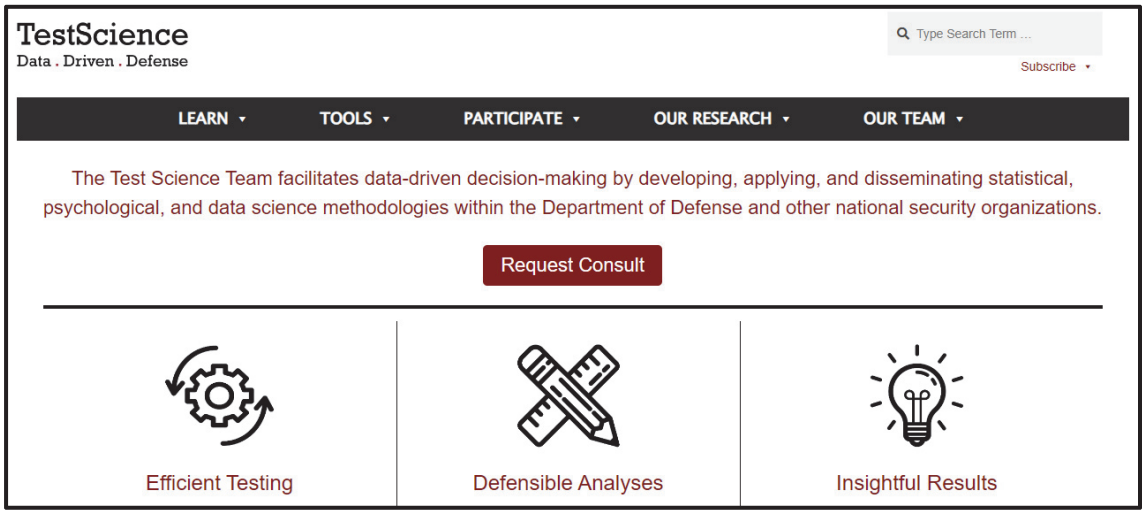
References on VV&A, DOE, and Statistical Analysis

- Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation. https://testscience.org/research-on-emerging-directions/
- National Academy of Sciences Report (ISBN 978-0-309-25634-6), "Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification (VVUQ)," 2012.
- https://vva.msco.mil
- Department Of Defense Standard Practice Documentation Of Verification, Validation, And Accreditation (VV&A); Mil-Std-3022 W/Change 1; 5 April 2012.
- Jack PC. Kleijnen, "Verification and validation of simulation models," European journal of operational research 82.1, pp. 145–162, 1995.
- S. Y. Harmon and Simone M. Youngblood, "A proposed model for simulation validation process maturity," The Journal of Defense Modeling and Simulation 2.4, pp. 179–190, 2005.
- T. Santner, B. Williams, W. Notz, The design and analysis of computer experiments springer-verlag, New York. (2003).
- Douglas C. Montgomery, "Design and analysis of experiments," John Wiley & Sons, 1990.
- R. T. Johnson, G. T. Hutto, J. R. Simpson, D. C. Montgomery, Designed experiments for the defense community, Quality Engineering 24 (2012) 60–79.
- R. H. Myers, Response surface methodology–current status and future directions, Journal of Quality Technology 31 (1999) 30.
- Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, Design and analysis of computer experiments, Statistical science (1989) 409–423.
- K.-T. Fang, R. Li, A. Sudjianto, Design and modeling for computer experiments, CRC Press, 2005.
- J. P. Kleijnen, Design and analysis of simulation experiments, volume 20, Springer, 2008.
- George E. P. Box, William G. Hunter and J. Stuart Hunter, "Statistics for experimenters," 2nd edition, John Wiley & Sons, 2005.
- M.A. Stephens, "EDF Statistics for Goodness of Fit and Some Comparisons," Journal of the American Statistical Association 69.347, pp. 730–737, 1974.
- Ronald A. Fisher, "Statistical Methods for Research Workers," Oliver and Boyd, 1925.
- Alan Agresti, "Categorical Data Analysis," 2nd edition, John Wiley & Sons, 2002.

The Test Science Website makes training materials, tools, best practices, and emerging research available to the T&E community

In order to make statistics and data science knowledge transparent to the community, we:

- Develop and facilitate training courses
- Publish research papers and articles in open literature
- Create user-friendly software to make analyses easier
- Maintain a repository of educational resources at testscience.org
- Host and co-sponsor the Defense and Aerospace Test & Analysis Workshop (DATAWorks)



T&E: Test and Evaluation

Every year DATAWorks offers relevant courses, tutorials, and talks to the T&E community

- Of particular in interest 2024:
  - Design of Experiments Short Course
  - Uncertainty Quantification Mini-Tutorial
  - Multiple M&S-related talks

M&S: Modeling and Simulation; T&E: Test and Evaluation

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-03-2024	OED Draft	

4. TITLE AND SUBTITLE

Statistical Methods for M&S V&V: An Intro for Non-Statisticians

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kelly Avery, John Haman, and Keyla Pagan-Rivera

5d. PROJECT NUMBER

3000770

5e. TASK NUMBER

229998

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER

3000770

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

This OED Draft has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

13. SUPPLEMENTARY NOTES

Project Leader: Keyla Pagan-Rivera

14. ABSTRACT

This is a briefing intended to motivate and explain the basic concepts of applying statistics to verification and validation. The briefing will be presented at the Navy M&S VV&A WG (Sub-WG on Validation Statistical Method Selection).

15. SUBJECT TERMS

Statistics; Modeling and Simulation (M&S); Verification, Validation, and Accreditation (VV&A)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Keyla Pagan-Rivera
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* 703-845-6936