Statistical Techniques for Modeling and Simulation Validation

Kelly McGinnity, Institute for Defense Analyses

Amanda Muyskens, Summer Associate, North Carolina State University

April 4, 2017
Outline

- DOE: A quick review
- M&S Validation
- Statistical Techniques
- Simulation Study
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
- Fractional Factorial
  - $2^{3-1}$ design
- Response Surface
  - Central Composite design

**Optimal Design**
- IV-optimal

**“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)**

- **Slant Range**
- **Look down angle**

---

**b)**

- **Aileron Deflection**
- **Angle of Attack**

---
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

- DOE: A quick review
- M&S Validation
- Statistical Techniques
- Simulation Study
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.)
Validation

- **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*

- Both **quantitative** and **qualitative** evaluations are necessary to understand the strengths and weaknesses of the model across the operational envelope
  - A statistical comparison of the model output to live data should be a *portion* of a larger validation plan

*DoD 5000.61
DOT&E Guidance Memos (March 14, 2016 / January 17, 2017):

- 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
  - Employ formal statistical analysis techniques to compare live and M&S data

- In addition to direct live vs. sim comparisons, assess M&S output across the entire operational domain
  - Empirical models (statistical emulators) can be useful for quantifying uncertainty
Outline

- DOE: A quick review
- M&S Validation
- Statistical Techniques
- Simulation Study
Goal of the Statistical Comparison (in English)

Do the simulation data and the live 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?
Goal of the Statistical Comparison (in stats speak)

- $H_0$: 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
- Can assess the statistical performance of various techniques using Monte Carlo simulation
Statistical Test Options (not exhaustive!)

- **Parametric Tests**
  - t-test (or log t-test)
  - Kolmogorov-Smirnov Test

- **Non-parametric Tests**
  - Kolmogorov-Smirnov Test
  - Fisher’s Combined Probability Test
  - Wilcoxon Rank Sum 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?
Preview of Recommendations

- The properties of your observed data and the structure of your factors dictates which method is best
  - 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*

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

  \[
  \text{Scaled data} = \frac{\text{each individual data point} - \text{mean (all data in that condition)}}{\text{std dev (all data in that condition)}}
  \]

**Works better for our problem**

*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: $X = -2 \sum \ln(p)$ 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,
    
    \[
    \text{Detection Range} = \beta_0 + \beta_1 \text{TestType} + \beta_2 \text{Threat} + \beta_3 (\text{TestType} \times \text{Threat}) + \epsilon
    \]
  - 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

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

- DOE: A quick review
- M&S Validation
- Statistical Techniques
- Simulation Study
Monte Carlo Power Simulation

- **Risk analysis technique**
  - Model possible outcomes by randomly drawing from a probability distribution and calculating results over and over, each time using a difference set of random values

- **Result of interest in this case is whether or not a statistical test rejects the null hypothesis, given the alternative is true**
  - Doing this many many times and calculating the proportion of times the test succeeds is an estimate of statistical power!

- **Simulation conditions:**
  - Distribution of response variable (symmetric, skewed, binary)
  - Sample size of live test (small, medium, large*)
  - Structure of factors (univariate, distributed level effects, designed experiment)
  - Effect size of interest (range of differences in means and variance ratios)

- **Running the simulation for each statistical test over all conditions of interest and comparing their power allows us to make informed recommendations**

* For symmetric and skewed data: Small = 2-5, Moderate = 5-10, Large = 11-20; For binary data: Small = 20, Moderate = 40, Large = 100
Univariate Mean Change Results

Null Rejection Rate for $\alpha = 0.2$

$\mu$ Actual - $\mu$ Modeled

Symmetric
Univariate Variance Change Results

Fisher
Non-Par KS

Symmetric
T-test Power

Symmetric
Fisher’s Test Power

Symmetric
Kolmogorov-Smirnov Test Power

Symmetric
Designed Experiment Mean Change Results

Regression

Symmetric
Designed Experiment Variance Change Results

Sc Non-Par Ks
Em & Pred

Symmetric
Emulation & Prediction Power

Symmetric
Non-Parametric KS Test Power

Symmetric
## Detailed Recommendations

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Structure Of Factors</th>
<th>Small Sample Sizes</th>
<th>Moderate Samples Sizes</th>
<th>Large Sample Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Symmetric</strong></td>
<td>Univariate</td>
<td>Fisher's Combined</td>
<td>T-test Fisher's Combined Non-Par KS</td>
<td>T-test Fisher's Combined Non-Par KS</td>
</tr>
<tr>
<td></td>
<td>Distributed Level Effects</td>
<td>Combo Test</td>
<td>Sc Non-Par KS</td>
<td>Sc Non-Par KS</td>
</tr>
<tr>
<td></td>
<td>Designed Experiment</td>
<td>Linear Regression</td>
<td>Linear Regression Sc Non-Par KS Emulation &amp; Pred</td>
<td>Sc Non-Par KS</td>
</tr>
<tr>
<td><strong>Skewed</strong></td>
<td>Univariate</td>
<td>Fisher's Combined</td>
<td>Log T-test Fisher's Combined Non-Par KS</td>
<td>Log T-test Fisher's Combined Non-Par KS</td>
</tr>
<tr>
<td></td>
<td>Distributed Level Effects</td>
<td>Combo Test</td>
<td>Sc Non-Par KS</td>
<td>Sc Non-Par KS</td>
</tr>
<tr>
<td></td>
<td>Designed Experiment</td>
<td>Lognormal Regression Sc Non-Par KS Emulation &amp; Pred</td>
<td>Lognormal Regression Sc Non-Par KS Emulation &amp; Pred</td>
<td>Sc Non-Par KS</td>
</tr>
<tr>
<td><strong>Binary</strong></td>
<td>Univariate</td>
<td>Fisher's Exact</td>
<td>Fisher's Exact</td>
<td>Fisher's Exact</td>
</tr>
<tr>
<td></td>
<td>Distributed Level Effects</td>
<td>Logistic Regression</td>
<td>Logistic Regression</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td></td>
<td>Designed Experiment</td>
<td>Logistic Regression</td>
<td>Logistic Regression</td>
<td>Logistic Regression</td>
</tr>
</tbody>
</table>

**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
Conclusions

• **Design of experiments** techniques should be used to efficiently cover the simulation domain and inform live testing

• **Regression analysis** is the most powerful comparison when matched experimental designs are used

• More robust nonparametric techniques, such as the Kolmogorov-Smirnov test, provide widely applicable solutions for the comparison

• How sensitive / powerful a test needs to be depends on the validation goals and the defined acceptability criteria

Use the method(s) that makes most sense for your observed data!!!
BACKUP
• 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)
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:

<table>
<thead>
<tr>
<th></th>
<th>Pass</th>
<th>Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Sim</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

  - 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

    \[
    \frac{(a + b)! \times (c + d)! \times (a + c)! \times (b + d)!}{n! \times a! \times b! \times c! \times d!}
    \]

  - 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
Empirical Type I Correction

- Be cautious of unexpected Type I error rates
  - Particularly for small sample sizes or when using a statistical test outside the context it was designed for

- We chose to empirically correct the type I error for all methods via simulation before comparing power performance
Example of Type I Error Correction

Before

After

Rejection Rate vs. Effect Size

α vs. Effect Size