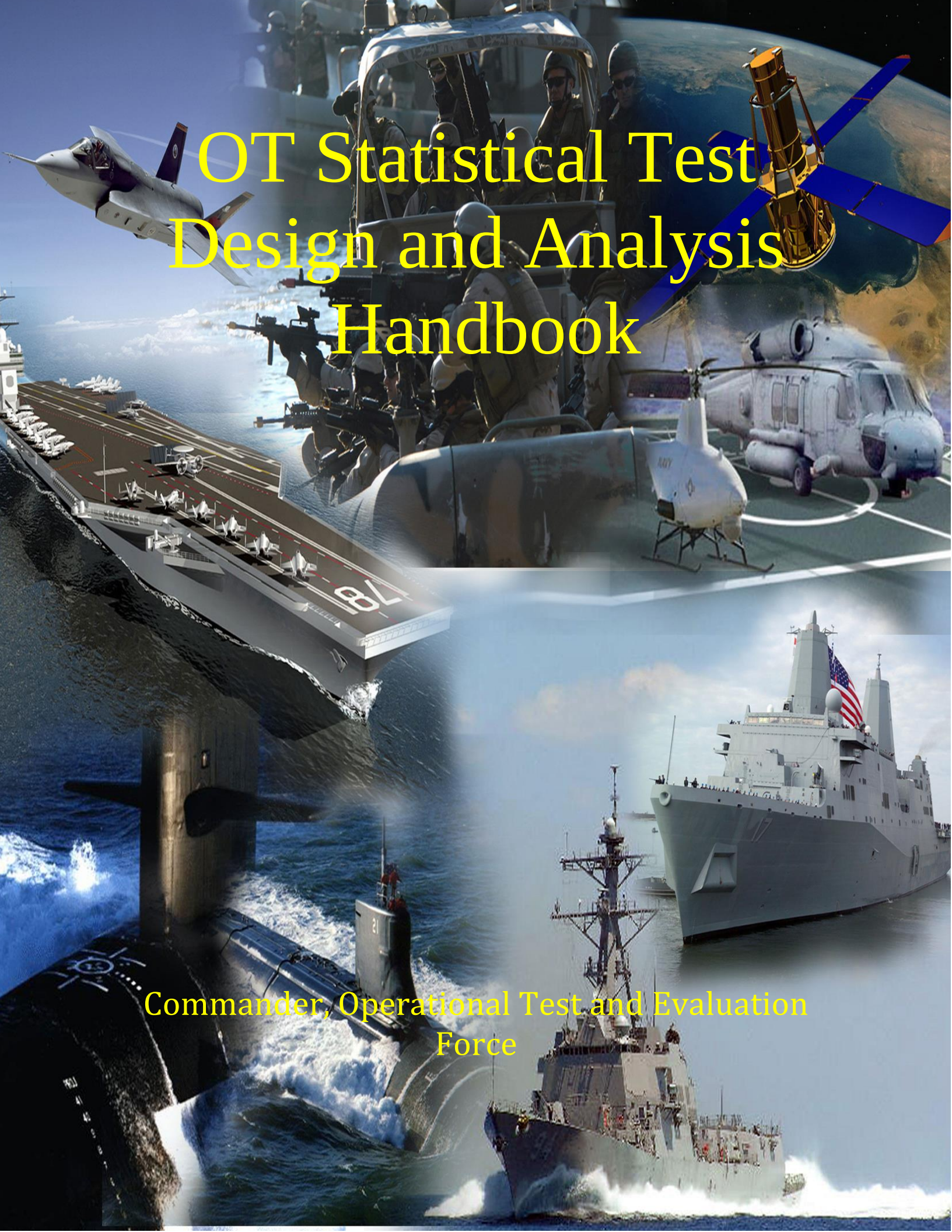
OT Statistical Test Design and Analysis Handbook



Commander, Operational Test and Evaluation Force

Version 1.0

11 May 20

RECORD OF REVISIONS

Number of Change	Summary of Changes	Updated
1	This is the initial OT Analysis Handbook	07 MAY 20
2	Signing process change	11 May 20

BEYRODT.DAVID.CHARLES.1044409379

Digitally signed by BEYRODT.DAVID.CHARLES.1044409379
Date: 2020.05.11 09:18:59 -04'00'

THIS PAGE INTENTIONALLY LEFT BLANK.

OT Analysis Handbook

TABLE OF CONTENTS

TABLES

FIGURES

SECTION 1 - INFERENTIAL STATISTICS IN OPERATIONAL TEST OVERVIEW

1.1 INTRODUCTION

This document provides an overview of the statistical test design and analysis techniques used at COMOPTEVFOR and a quick reference guide for warfare divisions Operational Test Directors (OTDs), Lead Test Engineers (LTEs), and contractors responsible for quantitative-based test design, planning and reporting. This handbook is not intended to be a statistical theory textbook, but rather a guide for warfare divisions on which techniques are most commonly used and what tools are available to assist OTDs, LTEs, and contract support in calculating the results. For more advanced STAT techniques considered beyond the scope of the warfare divisions, this handbook will detail the process for integrating COMOPTEVFOR 01B Analyst support into the existing test design, planning, and reporting processes and how to interpret the results produced by statisticians for evaluation. STAT techniques are used throughout the entire test life cycle, from test design where required resources are determined, to test reporting on relevant results to the Fleet. This handbook is intended implement COMOPTEVFOR OT&E policies and complement COMOPTEVFOR's primary OT&E handbooks. Those primary handbooks are:

- Integrated Evaluation Framework (IEF) Checklist
- Suitability Handbook
- Test Planning Handbook
- Test Execution Handbook
- Test Reporting Handbook
- Cyber Survivability Handbook

1.2 DISCUSSION: WHY USE INFERENTIAL STATISTICS?

The purpose of Integrated Test/Operational Test (IT/OT) is to conduct a test that informs the Fleet of the capability being delivered by the System Under Test (SUT). Inferential statistics are quantitative methods that provide testers the ability to use the results of data collected from a sample (IT/OT test period) in order to reach meaningful conclusions about how the SUT will perform in the Fleet. Conversely, descriptive statistics (mean, median, range), only apply to the sample collected. In other words, descriptive statistics only describe the test period itself and cannot be generalized to expected Fleet performance. Inferential statistics are therefore important tools for IT/OT.

Inferential statistical test designs are only created for critical measures (as defined by the Mission Based Test Design (MBTD) process). The designed test must be executable, and defendable as the minimum amount of data needed based on what has been defined as adequate testing. These designs provide the primary basis for resources/funding that are agreed upon in the TEMP.

COMOPTEVFOR follows traditional empirical scientific methods to design, plan, execute, and analyze IT/OT critical measures by: (1) identifying the question of interest, (2) defining the population to which the question applies, (3) collecting a sample of data, (4) carrying out data analyses, and (5) formulating a probabilistic answer to the question. A "probabilistic" answer may be the best the tester can offer from a finite sample of data because the latter is an imperfect representation of the population. In other words, samples are subject to random sampling error and are never a perfect representation of the "real world". Inferential statistics is a set of quantitative tools that allow the tester to make a reasonable evaluation in the presence of imperfect information.

1.3 TYPES OF IT/OT QUESTIONS ANSWERED BY STATISTICAL METHODS

There are different types of inferential statistical tests or designs, each intended to address different types of questions in IT/OT. Table 1-1 provides a high-level overview of the most common IT/OT test questions and the associated statistical methods that help answer the question.

Table 1-1. Types of IT/OT Test Questions and Associated Statistical Methods
UNCLASSIFIED

Type of IT/OT Question to Be Answered	Type of Statistical Method
Does SUT perform differently in different combinations of conditions (factors)? What is the predicted performance based on a given combination of conditions (factors)?	Response Variables (DOE and Regression Analysis
What is the range of plausible "real world" (Fleet) predicted values based on a given combination of conditions (factors)? How sure are we in the predictions?	Prediction Intervals
What is the range of plausible "real world" (Fleet) values for a single SUT parameter based on the uncertainty of the test results? How confident are we in the test results?	Confidence Intervals
Does SUT meet threshold? (Assumes no factors effect from different conditions)	Hypothesis testing: One-sample test of sample means or proportions.
Does SUT improve upon Legacy system? (Assumes conditions (factors) other than version of SUT are ignored.)	Hypothesis testing: Two-sample test of sample means or proportions.
Are two versions of SUT functionally equivalent (this is used trying to show that two things are the same, whereas previous tests aim at showing difference)? Do M&S results for a SUT perform the same as live results?	Hypothesis testing: Two One-Sided Test (TOST) of sample means or proportions.
Do two or more versions of SUT perform differently?	Single-factor ANOVA (one-way ANOVA)
What proportion of all SUT measurements fall above or below a threshold value? How confident are we in the test results? (Example: Can a radar maintain track 90% of the time for an inbound target? Is network throughput >2mbps more than 80% of time throughout a mission?)	Tolerance Intervals

SECTION 2 - INFERENTIAL METHOD: RESPONSE VARIABLES

2.1 DISCUSSION: WHY DO WE DO IT?

When the outcome of the critical measure is dependent on the conditions (factors) it will operate in and we have the means to systematically control those conditions in a test, then that critical measure shall be designated as a response variable (RV). The test design and analysis process associated with a RV facilitate answering the following questions for the warfighter:

- Which conditions (factors) effect the critical measure? i.e. What matters?
- How much effect do the conditions have on the critical measure? Just because a condition (factor) might have a "statistical" effect, that is not enough. The magnitude of the effect, in the presence of the other conditions (factors), must be considered to determine if it has practical effect on the warfighter.
- What is the predicted performance of the SUT (as evaluated by the outcome of the critical measure) given a specific combination of conditions (factors) that the warfighter expects to see on a given operational day?

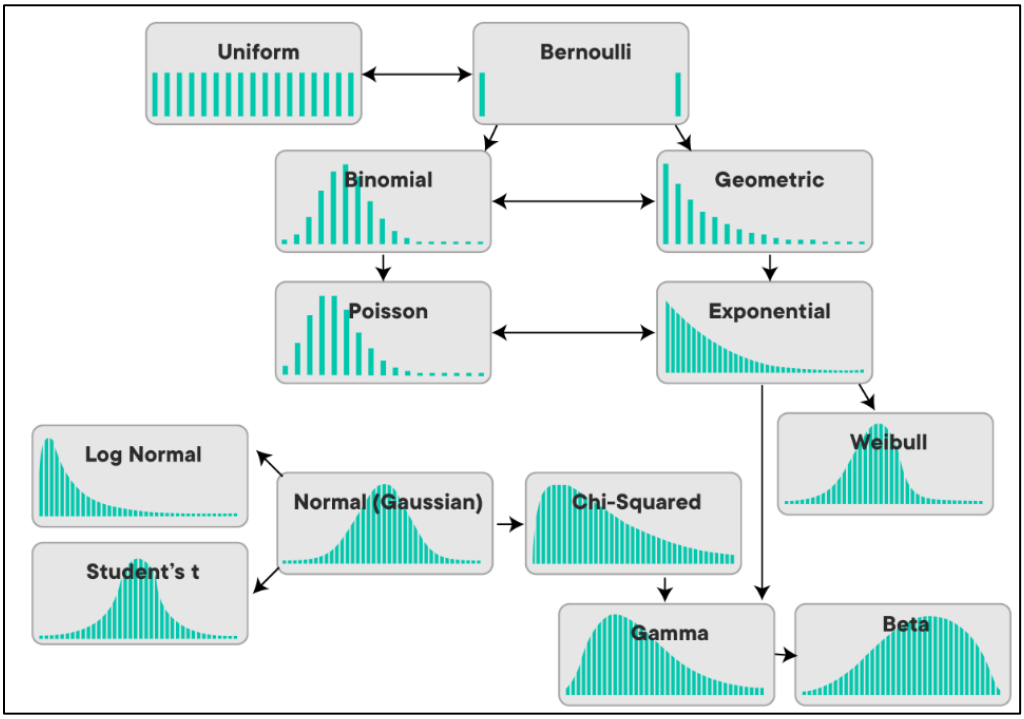
Response variable analysis is the inferential method which allows testers to answer those questions. Response variable analysis can be decomposed into four phases: Plan, Design, Test, and Analyze. Within these four phases, the following seven steps describe the process throughout the operational test and evaluation lifecycle. This process is captured within COMOPTEVFOR's MBTD, test execution, and Post-Test Iterative Process (PTIP) as shown in Table 2-1:

Table 2-1. Response Variable Analysis Traceability to IT/OT Processes UNCLASSIFIED				
Response Variable Analysis Process[1]		**Associated OPTEVFOR Process**		
Phase	**Step**	**Process Step**	**Associated Product**	**Associated Decisional Milestone/Meeting**
Plan	1. Recognition and statement of the problem	MBTD Steps 1-4 (IEF Checklist)	IEF Section 1	IPR-1
Plan	2. Selection of the response variable(s)	MBTD Steps 5-8 (IEF Checklist)	IEF Section 1	IPR-1
Plan	3. Choice of factors, levels, and ranges	MBTD Step 9 (IEF Checklist, OT Analysis Handbook)	IEF Section 2	DWG, IPR-2
Design	4. Choice of experimental design	MBTD Step 9 (Test Handbook, OT Analysis Handbook)	IEF Section 2	DWG, IPR-2
Test	5. Performing the experiment	Test execution (Test Execution Handbook)	Test data	Post-test Brief
Analyze	6. Statistical analysis of data	PTIP (Test Reporting Handbook, OT Analysis Handbook)	RV Analysis Outbrief	CEWG
Analyze	7. Conclusion and recommendation	Test Reporting (Test Reporting Handbook, OT Analysis Handbook)	DAS and Test Report	E-SERB

Selection of response variables should be done jointly with the 01B team. Examples of response variables include miss distance for a new air-to-ground weapon, detection ranges for sensors, or message throughput/error rates for communications systems where the associated controlled conditions can be systemically varied during testing. Implications of the type of response variable and distribution should be considered. Figure 2-1 visually depicts different types of distributions. The test team should inform the 01B Analyst what the expected distribution of the response based on historic data for a previous SUT version or a similar SUT or based on subject matter expertise (operator and/or engineering).

[1] Montgomery, D.C. (2008). Design and Analysis of Experiments. John Wiley and Sons

Figure 2-1. Visual Representation of Distribution Types[2]
UNCLASSIFIED



2.2 TEST DESIGN: DESIGN OF EXPERIMENTS (DOE)

The validity of the data to be used for response variable analysis is largely determined by the adequacy of the test design. DOE refers to the design process of planning an experiment so that appropriate data will be collected and analyzed, resulting in valid and objective conclusions from which Fleet expected operational performance can be inferred. The use of DOE ensures testers identify the variations in conditions and required sample size needed to evaluate critical measures chosen as response variables. The end goal is to ensure that statistical analysis of test results can detect whether the SUT's performance is impacted by the operational environment and how the conditions affect any variation in the results. Proper use of DOE will yield data that produces defendable results, identifies and mitigates the risks of making inaccurate conclusions and reduces uncontrolled experimental error[3]. DOE uses mathematical techniques to create the most efficient design for the desired test objectives, in keeping with responsible use of resources for operational test. DOE is recommended for use by the test community in DODI 5000.02 and the Defense Acquisition Guidebook (DAG).

2.2.1 How do we do it?

Close consultation with the 01B team is essential. Generation of the DOE draft run matrix is predicated on the following required inputs from the test team and will determine the size of test and number data points that need to be collected:

[2] Sharma, A. (2019). Understanding Different Types of Distributions You will Encounter as a Data Scientist. From https://medium.com/mytake/understanding-different-types-of-distributions-you-will-encounter-as-a-data-scientist-27ea4c375eec
[3] DAU "CLE 085 Scientific Test and Analysis Techniques (STAT) in Test and Evaluation (T&E)", Lesson 2, pg 4

- Type of DOE needed
- Response Variable Type
- Identification of Factors and Levels
- Design Constraints
- Effect Size and Statistical Signal-to-Noise

Once the inputs have been communicated to the 01B team, the 01B Analyst will use statistical software to generate the DOE draft run matrix for the test team. The typical "goal" the analyst will try to achieve is to create a DOE where each effect of interest has 80% power with 80% confidence. Power is defined as the probability of accurately identifying an effect (or difference) on the response variable when one exists. Confidence is defined as the probability of accurately concluding that there is not a significant effect (or difference) when one does not exist. Further discussion on the required test team's inputs is below.

2.2.1.1 Type of DOE

There are different types of DOEs that provide different levels of quantifiable information. The first input the tester must define is what question the DOE is trying to answer. The goal is to select the right tool for the task. The more information required the more complex the DOE will be and the more resources (in terms of number of runs) will be required. Testers shall work with 01B CTFs and Analysts to select the right type of DOE from Table 2-2 based on the test objective.

Table 2-2. Types of DOEs
UNCLASSIFIED

Type	Description		OPTEVFOR Minimum/Adequate Criteria
Full Factorial	Every combination of factors and levels of interest; cost prohibitive when many factors are involved. Appropriate for small designs.		Power $\geq$ 80% for all DOE terms, α=0.2[3]
Screening	Typically smaller designs with 2-level only factors; aimed at identifying the factors that effect a response. Good for sensitivity analysis.		
Optimal (D-, I-)	Good for multi-factor, multi-level experiments with both continuous and categorical factors when a full factorial design is not feasible (cost or disallowed combinations).	Main Effect (ME) Only: Evaluate factor effect.	Power $\geq$ 80% for all DOE terms, α=0.2[3] with exceptions[1]
		ME+2FI: Evaluate factor effect of main effects and interactions between factors. Can possibly be used to build a predictive metamodel[2].	
		ME+2FI+Quadratic: Evaluate factor effect of main effects, interactions, and curvature between continuous factor levels. Can possibly be used to build a predictive metamodel[2].	
Space Filling	Spreads combinations throughout the design region. Best for predictive metamodeling. These designs should be implemented when factor effect DOEs have been executed and analyzed. Since only factors that have effect should be used, Power is no longer the driving consideration for minimum/adequate.		Graphical displays can be used to ensure the design-space is adequately covered by test points.

Table 2-2. Types of DOEs
UNCLASSIFIED

Type	Description	OPTEVFOR Minimum/Adequate Criteria

Note 1. Exceptions:

a) 80% Power for all possible model terms is not realistic for large, complex designs with a high-order number of factors and levels. 01B Analysts will work with the test teams to present viable options to stakeholders and decision makers.

b) 80% Power for hard-to-change factors is not often achievable. Expect power for hard-to-change terms to be less than 50%. However, they are still critical to include in DOEs as any interaction terms that include hard-to-change factors should still be designed for adequate power.

c) 80% Power for quadratic terms is not often achievable. However, the information these terms provide is critical if there is an expectation of a non-linear effect curve between the levels of a continuous factor.

Note 2. A metamodel refers to a predictive statistical equation that is derived from the regression analysis of a response variable. This is described in more detail in Section 2.3 Post-Test: Regression Analysis.

Note 3. It is common for the test team at COMOPTEVFOR to set α to 0.20. Confidence $(1 - \alpha)$ is the probability of not making a Type I error. It is recommended that the analyst set α after weighing the costs of a Type I error. If the cost is estimated to be high, it may be appropriate to set α at levels smaller than 0.20. Target values of β (probability of Type II error) and $1 - \beta$ (statistical power): It is common at COMOPTEVFOR to aim for a power of 0.80. It is recommended that the analyst set β after weighing the costs of a Type II error.

2.2.1.2 Response Variable Type

In the IT/OT environment, critical measures are typically defined as continuous or discrete. "Continuous" refers to interval or ratio scales (e.g., detection range in nm, miss distances in meters, or gallons of fuel burned). "Discrete" refers to nominal or sometimes ordinal scales. The most common discrete variable used in T&E is the two-valued "Success" or "Failure" variable distributed according to the binomial distribution. Response variables should be continuous (vice binomial) if possible, as continuous variables provide more useful information about system performance across an operating environment. Binomial response variables also require significantly more resources (as much as 5-10 times) since more data is needed to obtain minimum/adequate confidence in response compared to continuous measures. In many cases, the KPPs or critical measures specified in the SUT program requirements documents (such as Capability Description Document (CDD), Capability Production Document (CPD), Operational Requirements Document (ORD), Urgent/Emergent Operational Needs Statement (U/EONS), Top Level Requirements Document (TLR)) are binomial metrics; the OTD may elect to create a continuous measure from those binomials to use for DOE. Ideally, response variables would be directly specified in the CDDs, but in some cases, SUT performance may be better described by derived or OTA created measures.

2.2.1.3 Identification of Factors and Levels

A focus of DOE is to reduce the large set of unconstrained conditions developed in the initial MBTD step to a manageable set of conditions based on what will significantly affect the response variable and the tasks to which it is associated. The OTD brings operational experience and judgment to help pare down the conditions to important ones. Test teams shall review the

conditions that have been associated with the critical task. All conditions fall into three categories: controlled, constant, or recordable. Conditions and their associated levels should be prioritized by expected impact on system performance and the likelihood operators will encounter them in the intended operating environment. Test teams should use Table 2-3 to prioritize the assigned controlled conditions that will be controlled and varied to assess the operational effectiveness of the SUT.

The reduced number of conditions selected to be controlled and varied are defined as DOE factors. Those that are not controlled or held constant should be identified as recordable. DOE levels refer to the specific variation within a factor that will be evaluated. For example, assume a tester wants to assess the effect of target size (controlled condition/factor) on a radar detection range (response variable). The tester has determined that there are three operationally meaningful target sizes: Small, medium, and large. These three target sizes define the three DOE levels for the "target size" factor.

The numbers of conditions and levels directly influence the design selected, the resulting run matrix, and the IT/OT resource requirements. The levels for each controlled condition will then be varied in a test design (run matrix) produced by the 01B Analyst. Again, Table 2-3 serves as a guide for prioritizing conditions and determining which will be defined as DOE factors to be controlled and varied in test.

Table 2-3. Factor Prioritization Matrix
UNCLASSIFIED

		Likelihood of Encountering Level During Operations		
		Multiple levels occur at balanced frequencies (e.g., 1/3, 1/3, 1/3)	Some levels arebalanced, others are infrequent (e.g., 5/10, 4/10, 1/10)	One level dominates (e.g., 8/10, 1/10, 1/10)
Effect of Changing Level on Performance		**Balanced**	**Mixed**	**Dominant**
Significant Effect on Performance	**High**	Vary all	Vary balanced levels, demonstrate infrequent levels	Fix dominant level, demonstrate others
Moderate Effect on Performance	**Medium**	Vary all	Vary balanced levels, demonstrate others	Fix dominant level, demonstrate others
Low Effect on Performance	**Low**	Fix levels or record level used	Fix levels or record level used	Fix dominant level

2.2.1.4 Design Constraints

Basic DOE assumes that all conditions are controllable and can be perfectly randomized in a run order. Real-world testing is not always that simple. The good news is that the 01B Analyst can take constraints into consideration when calculating a DOE test design. It is critical that these constraints are programmed in to the DOE in advance so that risks to post-test analysis are mitigated and unavoidable limitations to the test design are acknowledged by all stakeholders in advance. Failure to plan for test design constraints in advance places severe risk on the post-test evaluation of the critical measure. In other words, the tester would not be able to deliver on the level of information promised in the test plan. This is avoidable through early identification of

constraints by all stakeholders (e.g. OT Squadrons, DT testers, engineering SMEs) and communication with the 01B team.

2.2.1.4.1 Hard to Change (HTC) Conditions (Factors)

One way DOE mitigates analytic risk in post-test analysis is through randomization. This means that conditions (factors) can be varied randomly from run to run. However, sometimes there are real world constraints that make it difficult or impossible to completely randomize. In this case, the condition (factor) becomes classified as a hard to change (HTC) factor. An example would be a condition called "Time of Day" with the two levels being "Day" and "Night". If the tester had to flip flop between "Day" and "Night" every other run, the tester would only be able to get two runs completed per day! However, it might be feasible for the tester to do three runs in the late afternoon, wait an hour for twilight, and then do three runs in the early evening; thereby tripling the number of runs that can be executed in a day. These groupings of a level of a condition (factor) are called whole-plots.

With good communication from the test team on what is executable, the 01B Analyst can design the run matrix using whole plots to fix a level of a condition in a group of runs. The limitation of using whole plots is that the power of the HTC factor will likely never be at least 80%. A tester should expect to see power values less than 50% for a HTC factor. However, it is still important for the overall test objective because any interaction terms will still have sufficient power. For example, if the DOE included a second easy to change condition (factor) called "Mode" with two levels "Mode A" and "Mode B", and the DOE was properly designed with one HTC factor and one easy to change (unconstrained) factor, the tester would still be able to assess if the combination (interaction) between "Time of Day" and "Mode" had an effect on the response. It is essential that these constraints are implemented prior to test execution so that the statistician can reduce the limitations on test objectives and the tester can clearly communicate post-test expectations to all stakeholders. See Table 2-4 below for an example of a standard DOE run matrix versus one with a HTC factor.

Table 2-4. Un-Constrained vs. HTC DOE Run Matrix Example
UNCLASSIFIED

Un-Constrained DOE (Randomized) – Not Executable!			**DOE with One HTC Factor – Executable**			
Run Number	Factor 1: Time of Day	Factor 2: Mode	Nun Number	Whole Plot Number	Factor 1: Time of Day (HTC)	Factor 2: Mode
1	Day	Mode A	1	1	Day	Mode B
2	Night	Mode B	2	1	Day	Mode A
3	Day	Mode A	3	1	Day	Mode A
4	Night	Mode A	4	2	Night	Mode B
5	Day	Mode B	5	2	Night	Mode A
6	Night	Mode B	6	2	Night	Mode B
7	Day	Mode B	7	3	Day	Mode B
8	Night	Mode B	8	3	Day	Mode A
9	Day	Mode B	9	3	Day	Mode A
10	Day	Mode A	10	4	Night	Mode B
11	Night	Mode A	11	4	Night	Mode B
12	Night	Mode A	12	4	Night	Mode A

2.2.1.4.2 Very Hard to Change Conditions (Nuisance Factors)

There are also times when the levels of a condition (factor) cannot be varied at all, even within groupings (whole plots). Using the previous example, assume the factor was "Season" instead of "Time of Day" with the levels being "Winter" and "Summer". Even if it were defined as a HTC, it would still take 4 years to complete the test! In this case, "Season" is classified as a very hard to change, or a nuisance factor. A nuisance factor is one that we know has effect on the response variable, but cannot be controlled. Anything that "matters", or has effect on the response variable, must be dealt with in the DOE. The risk is that if not accounted for, the resultant data will be completely invalid for response variable analysis and no conclusions can be drawn from the test. This risk is easily mitigated through the test team's collaboration with the 01B Analyst throughout the test design process so that the nuisance factor can be properly dealt with.

There are a few different ways to handle a nuisance factor. First, the test team must prioritize the levels of the condition (factor) in accordance with Table 2-3. If the likelihood of encountering all levels except one are low, the strategy would be to include the dominant level as a constant condition in the DOE while exploring the other levels through demonstrations. However, if multiple levels are critical and are operationally likely, then a technique called blocking will be used. In laymen's terms, it means that a separate DOE test design and run matrix will be applied to each level of the blocked nuisance factor. Following the example, that means there would be two separate DOEs around the other conditions of interest: One DOE run matrix for "Summer" and one DOE run matrix for "Winter". The limitations of this approach are the increased resources needed to execute two DOE run matrices and "Summer" and "Winter" results cannot be statistically compared to one another using regression techniques (see the 01B Analyst for alternative comparison strategies).

2.2.1.4.3 Disallowed Combinations

A disallowed combination simply refers to unrealistic or operationally irrelevant combinations of levels between conditions (factors). Consider an example using two continuous conditions (factors): "Launch Altitude" defined between 5,000-20,000 ft and "Engagement Range" defined between 1-10nm. A disallowed combination might be a launch altitude of 20,000 ft with an engagement range of 1nm (which might make target intercept physically impossible). In this case, a disallowed combination might be defined where engagement ranges between 1-3nm are disallowed when the launch altitude is 20,000 ft. Disallowed combinations involving categorical factors simply omit that unique combination. However, disallowed combinations using continuous factors, as exemplified above, can implement linear constraint equations. The test team will work with the 01B Analyst to ensure an operationally relevant test design space by clearly defining any disallowed combinations between levels of conditions (factors).

2.2.1.5 Necessity of Replication

If the critical measure response variable is binomial, power calculations are not the only consideration for a minimum/adequate test design. Adequate test designs also rely on replication to mitigate additional risks specific to analysis of binomial responses. Replication refers to the repetition of the same input conditions. In other words, one would see a run matrix with the same

combination of controlled conditions repeated in subsequent runs. Replication demands significantly more resources; however, the likely alternative is data that is invalid for informing the desired test objective. If testers are using a binomial response, work with the 01B Analyst to determine if replication is necessary. The risk is easily mitigated through the test team's collaboration with the 01B Analyst and understanding the unique circumstances when replication is needed, even when power goals are achieved in fewer runs.

The following present the primary two reasons replication might be necessary for a specific test of a binary response:

- The SUT is expected to perform well (i.e. very few number of failures are observed in the data). For example, for a probability of detection response variable, assume a test is run 100 times under systemically controlled and varied factors/levels. If the SUT detects the object of interest 90 out of the 100 times, there are 90 "successes" and ten "failures". The opposite is true if the SUT is expected to perform poorly (i.e. very few number of successes).
- When the critical measure is being evaluated through Modeling and Simulation (M&S), the models (or federation of models) being used are often non-deterministic. Non-determinism means that even with identical inputs, different responses are observed. If non-deterministic M&S is brought forward for IT/OT, the magnitude of the non-determinism must be quantified for accreditation to even be considered.

If desired, see the 01B Analyst for the mathematical theory behind why these circumstances cause significant analysis risk. The main point is for the test team to be aware that the above special circumstances might result in a test design that requires replication and that the need for additional resources can be effectively communicated to all stakeholders.

2.2.1.6 Effect Size and Statistical Signal-to-Noise

The effect size is a key component when evaluating how believable and useful a statistical result is. The standard deviation of the effect size is critically important, as it quantifies how much uncertainty is included in the statistical result. Using historical data (if available) or subject matter expertise (operator or engineer), the test team will determine the anticipated distribution and variability of the response variable, as these are essential to the definition of effect size and the statistical calculations. See Figure 2-1 for a visual display of typical distributions.

In DOE, the effect size establishes the test team's required test sensitivity by answering, "How much difference does there need to be between a factor and "noise" in order to say that the factor affects the outcome of the response variable?" This question describes the statistical signal-to-noise ratio (SNR). The effect size chosen by the test team must be explained in terms of operational relevance. The 01B Analyst will assist in calculating the appropriate SNR for the DOE based on the test team's desired effect size and expected distribution and variability of the response variable.

For binomial responses, SNR is a less concrete concept. Instead, a mathematical approximation of SNR is used based on an accepted probability-based effect size. Academic literature supports three main methods for computing the mathematical approximation of SNR for binomial responses. COMOPTEVFOR uses the normal approximation method with variance adjustment

factor to estimate SNR because it is the most conservative of the three recognized methods. The 01B Analyst will compute the SNR approximation based on the test team's input on an operationally acceptable effect size. Appendix A identifies references where the three primary methods are discussed in further detail.

2.2.2 Verify the Design

This is an iterative process, which continues through test execution. The OTD must apply his or her operational expertise, knowledge of the system, and any previous test results when evaluating the practical meaning of a targeted effect size, power, and confidence. The test team must also ensure that every planned run, and the specified run order, is viable; not incorporating disallowed combinations of conditions (conditions that cannot occur in the real world). Likewise, the 01B Analyst must ensure that the choices made reflect a sound understanding of the relevant engineering and physical processes. The resulting run matrix provides the test team with a plan to assess the response variable (how many times a vignette will be run, under what conditions, etc.). It is acknowledged and understood that things can change between original test design and test execution as resources are adjusted or new, relevant information is learned about the SUT. Communication with the 01B CTF and Analyst is imperative in order to ensure that the appropriate statistical rigor is applied to any necessary run matrix changes so that test objectives are still met.

2.2.3 What do we do with it

Based on the effect size chosen by the test team, the 01B Analyst will calculate a decision table of varied sample sizes versus SNRs. The test team and stakeholders can use this table to weigh the priority of mitigating uncertainty in results versus resources. The 01B Analyst will also draft the "Test Design", "Sample Size for Statistical Significance", and "Post-test Analysis" of the associated critical measure in the IEF in accordance with the IEF Checklist. The test team will present the test design and associated run matrix at decisional meetings in accordance with the IEF Checklist and Test Planning Handbook.

2.3 POST-TEST: REGRESSION ANALYSIS

Regression Analysis is a powerful statistical method that allows you to examine the influence of one or more factors on the RV of interest. The OTD with LTE guidance is responsible for execution of the Post-Test Iterative Process (PTIP) in accordance with the Test Reporting Handbook to include completion of the RV analysis. The OTD is responsible for ensuring completion of the RV analysis steps detailed in Appendix B. This handbook augments the PTIP with more detailed guidance for understanding, interpreting, and reporting RV analysis findings.

2.3.1 How do we do it?

The 01B Analyst will typically complete the RV statistical analysis. 01B Analysts rely on the divisional OTDs, LTEs, analysts, and contract teams to scrub the RV data for typos, outliers, and obvious problems. The Division does this by preparing and generating a Data Analysis Package for the 01B Analyst. The same procedures and policies apply if the OTD is relying on an analyst from a different COMOPTEVFOR division or external organization.

A Data Analysis Package is a data package with properly validated and scored RV data (in accordance with the Test Reporting Handbook). The Data Analysis Package should organize the RV data into a properly classified and marked human readable file (e.g. Microsoft Excel) where each row represents a single run and columns represent the response variable, controlled conditions (factors), and recordable conditions. See Appendix B for full Data Analysis Package requirements.

Once the division has organized and delivered the Data Analysis Package, the 01B Analyst will conduct the RV analysis with two major parts: Exploratory Data Analysis (EDA) and System Characterization.

2.3.1.1 EDA

This step may seem obvious, but is critical. EDA is the means to identify and document any obvious trends, discrepancies, or other interesting findings. The right visual display of a dataset can uncover anomalies and provide insights that go beyond what most quantitative methods are capable of discovering. EDA complements the model-based approaches that will be discussed in follow-on sections.

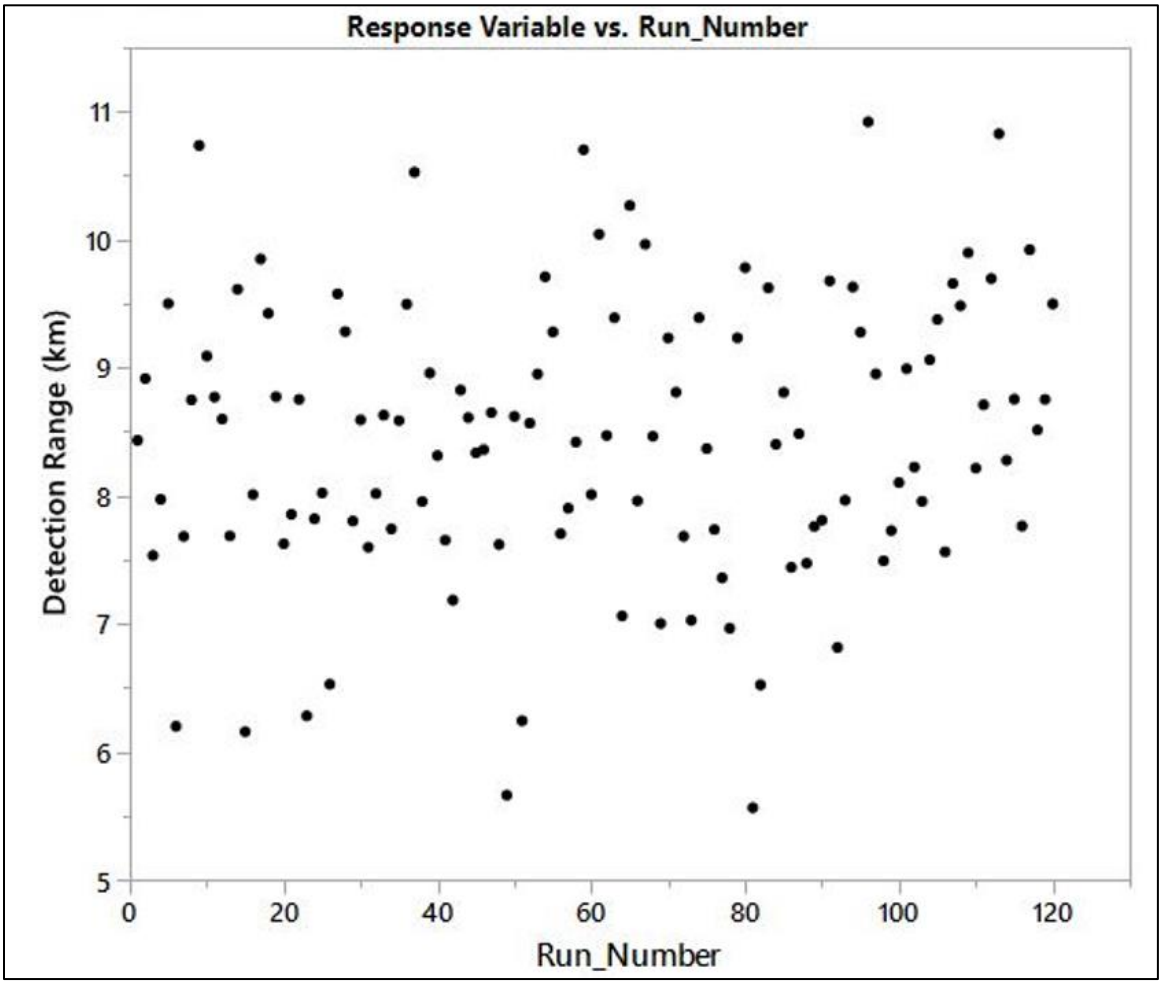
The primary objective of EDA is to maximize the insight into a data set and into the underlying structure of a data set, while extracting specific items about a data set like a good-fitting model, a list of outliers, a sense of robustness of conclusions, a ranked list of important factors, and conclusions as to whether individual factors are statistically significant. If test data was not collected in accordance with the planned DOE, EDA also assesses the validity of the data for the planned analysis objectives. EDA uses multiple techniques rather than depending on any single technique. Different plots have a different basis, focus, and sensitivities, and therefore bring out different aspects of the data. These techniques increase our reassurance that our conclusions are valid. Such visualizations are the shortest path to gaining insight into a data set in terms of testing assumptions, outlier detection, relationship identification, factor effect determination, model selection, and model validation.

Some EDA techniques are broad-brushed and apply almost universally, but many EDA techniques are situationally specific. Below are a few common graphical displays produced during EDA, how a test team should interpret these, and why these are important in reporting results.

2.3.1.1.1 Run-Sequence Plots

Run-Sequence plots are an easy way to graphically summarize a data set. Figure 2-2 shows an example of a run-sequence plot where the outcome of the Response Variable is plotted against the run order. When examining run-sequence plots, 01B analysts are looking for potential outliers, any non-randomness, or significant shifts to the mean and variance in the data over time. These run-sequence plots are for investigation and diagnostics purposes and are not typically included in the DAS for the final report unless something significant occurred such as outliers; but will be retained within 01B and made available upon request. Figure 2-2 exemplifies a case when the DOE was executed properly and there are no concerning outliers. This is shown through the randomization of responses across time and no responses at extreme values. Conversely, there would be significant risk to meeting analysis test objectives if the plotted responses showed significant trending over time.

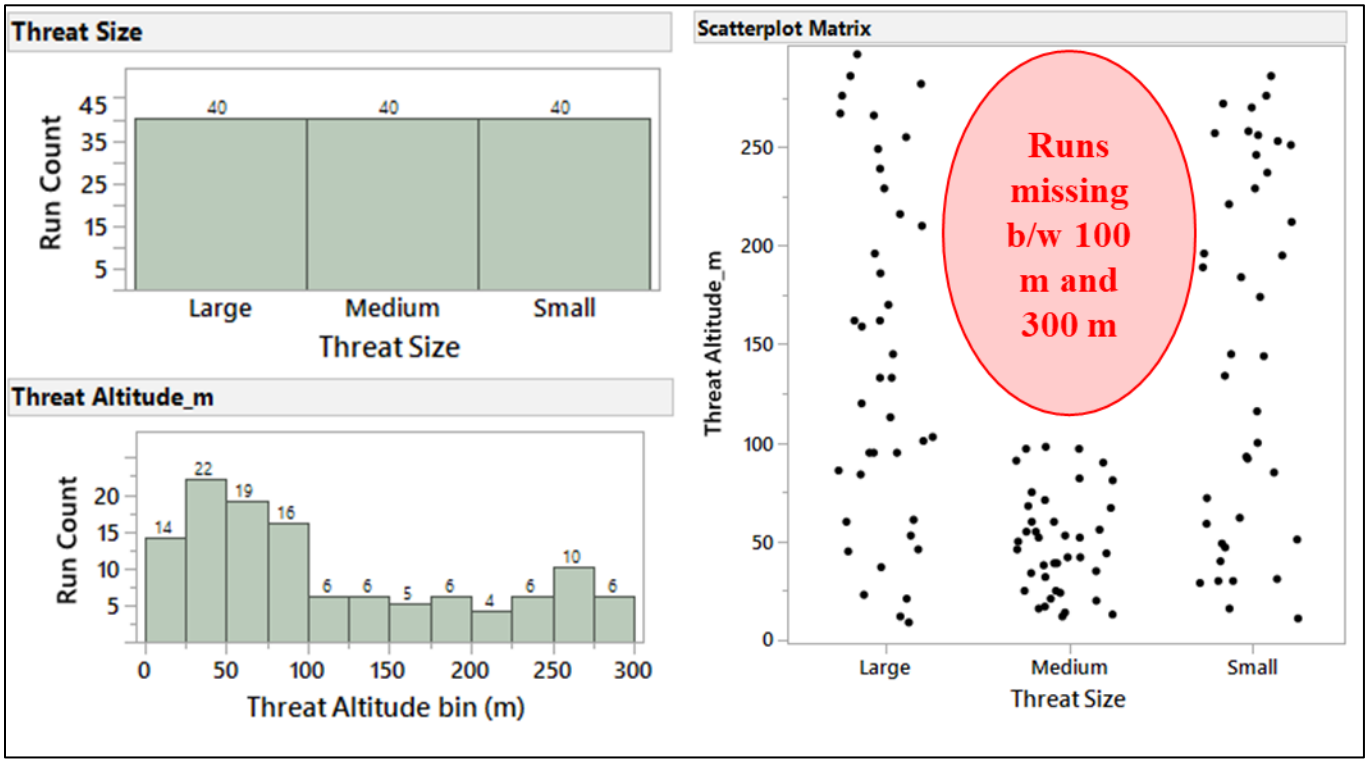
Figure 2-2. Response Variable by Run Order
UNCLASSIFIED



2.3.1.1.2 Missing or Unbalanced Runs

An important aspect of EDA is to check for missing runs whether it is fewer total executed runs than planned or a different mixture of the controlled conditions. Missing data or executing runs using different controlled conditions than planned can negatively affect the post-test RV analysis. 01B analysts have several ways to check for missing or unbalanced runs. Figure 2-3 shows two styles of plots that 01B Analyst use to examine for missing or unbalanced runs. On the left side of Figure 2-3, there are two distribution plots for each controlled condition: Threat Size (top left) and Threat Altitude (bottom left). On the right side of Figure 2-3, there is a scatterplot matrix. The distribution plots (left) in this example do not necessarily show an imbalance in the execution of runs, but the scatterplot matrix (right) shows a significant gap for the Threat Size [Medium] level where no runs were conducted from a Threat Altitude greater than 100 meters. This gap in Threat Altitude runs greater than 100 meters for the Threat Size [Medium] means that nothing is known about the SUT's performance at those conditions. In cases like these, the OTD should work with the 01B analyst to properly document the limitation in the DAS and assess the limitation's severity and impact for the Test Limitation section of the test report (in accordance with the Test Reporting Handbook).

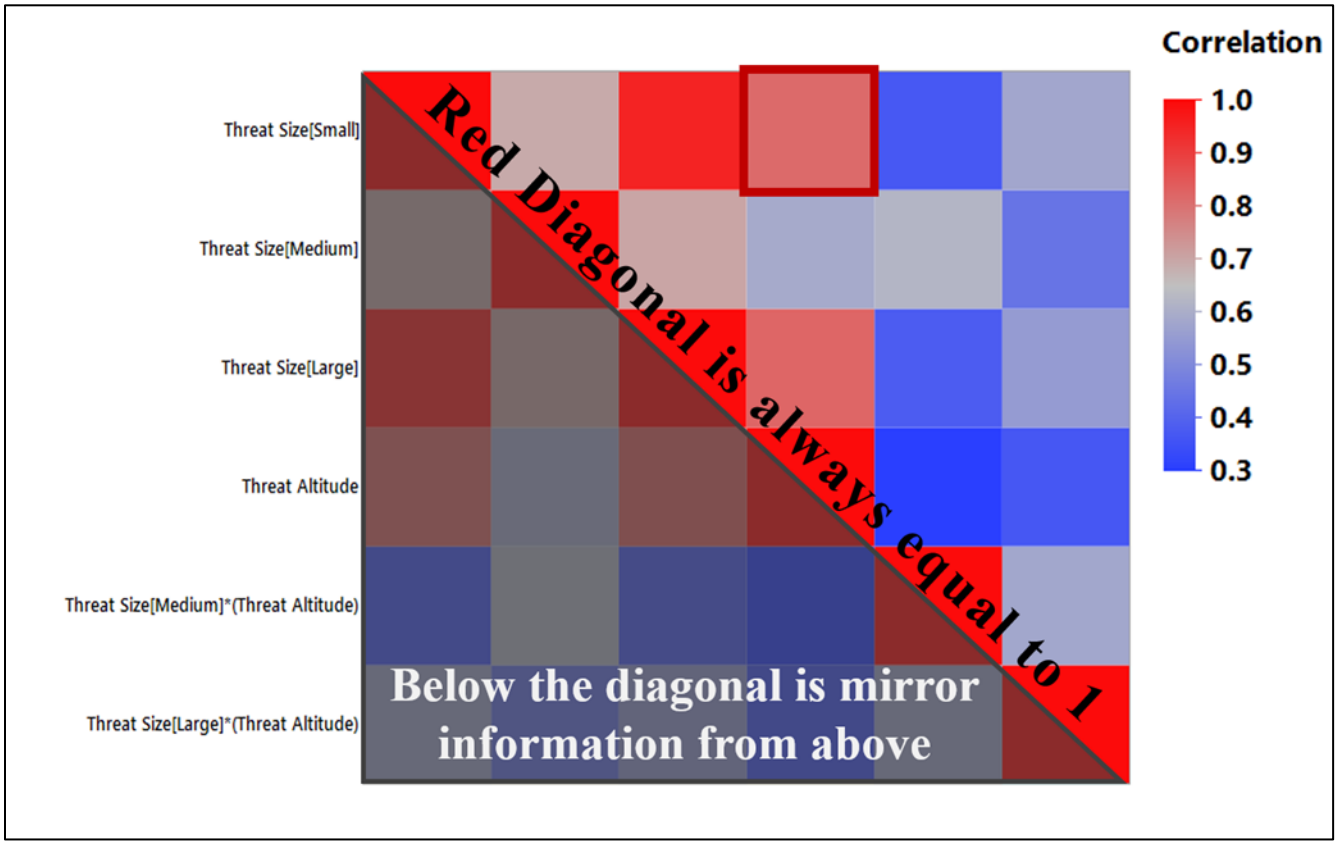
Figure 2-3. Distribution plots (left) and a Scatterplot Matrix (right) of the Controlled Conditions
UNCLASSIFIED



2.3.1.1.3 Color Correlation Map

Regression analysis is based on the assumption that controlled conditions (factors) are independent from one another. High correlation values indicate dependency among controlled conditions (factors). It is important to understand that highly correlated factors might result in limitations to the final regression analysis. A color correlation map shows the correlation between different conditions. Each colored cell shows the correlation between two conditions. A correlation between two conditions indicates that as one condition's value changes, the other condition tends to change in a specific direction (either increasing or decreasing). Figure 2-4 shows a color correlation map for two controlled conditions (Threat Altitude and Threat Size) and their interaction produced in JMP® statistical software. A color correlation map provides an easy visual where high correlations might exist, however there is a numeric value associated with each combination. In Figure 2-4, the line of red cells going from the top left to the bottom right is the main diagonal and shows that each variable always perfectly correlates with itself (correlation value = 1). Typically, a correlation map is "square" with the same variables shown in the rows and columns, and are symmetrical with the same correlations shown above the diagonal being a mirror image of those below the diagonal. Correlation map can help identify anomalies from test execution. For example in Figure 2-4, the correlation between Threat Size [Small] and Threat Altitude is high (red in a cell means high correlation). A general rule of thumb is that correlation values greater than 0.7 are considered highly correlated. High correlations should be identified and further evaluated before system characterization begins.

Figure 2-4. Color Correlation Map
UNCLASSIFIED



These EDA examples are common graphical displays, but are not all-inclusive of the EDA plots, graphs, and techniques. Additional examples can be seen in the "RV Analysis Outbrief Template.ppt" found at Y:\OT&E Reference Library\OT Analysis Handbook.

2.3.1.2 System Characterization

The primary purpose of modeling the RV data is to identify the statistically significant conditions and quantify their influence on the output response. This is known as system characterization.

2.3.1.2.1 Create an Empirical Model from the Data

There are two broad categories of models: physical and statistical. Physical models are mathematical representations found in nature such a Newton's Laws of Motion or Maxwell's equations of electricity and magnetism. A statistical model (also known as a metamodel or a regression) is a mathematical relationship between the response variable and the conditions. Due to the complexity of a SUT, physical models may not exist, but a statistical regression is still effective to identify which conditions impact the SUT's performance. The anticipated statistical model expected to be used during PTIP was documented and detailed in the test design section of the IEF and the Data Analysis Plan within the Post-Test Guidance enclosure of the Test Plan.

During test design and planning, decisions such as the type of RV (a continuous or binary (Yes/No) response variable) and which interactions between conditions would likely impact the RV were discussed and made. During PTIP, the 01B analyst verifies the validity of those planning decisions as related to the actual RV data collected in test. The 01B analyst explores numerous potential

modeling methods such as Standard Least Squares, Generalized Liner Model (GLM), Penalized regression, Chi-Square (χ^2) test, or logistical regression and determines the best model. If desired, see Appendix A for references on the different modeling methods.

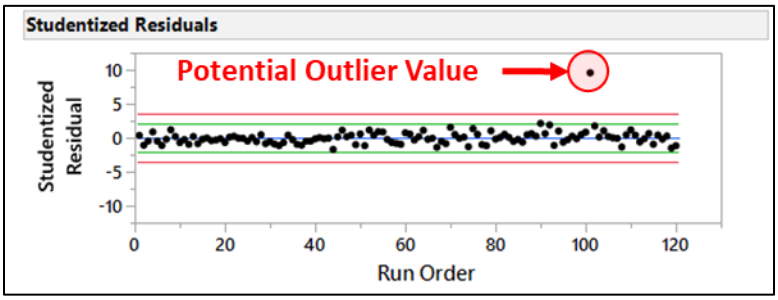
2.3.1.2.2 Statistical Meta-model Validation

The validity of the results reflected in the statistical meta-model are contingent upon certain mathematical assumptions being met. In other words, these assumptions being met means that "math" is correct and the statistical meta-model output is valid. Therefore, statistical meta-model validation is possibly the most important step in building a model. It is also one of the most overlooked. There are many statistical tools for model validation, but the primary tool for most applications is graphical residual analysis. In regression analysis, the residual is the difference between the actual RV value observed and the predicted value of the RV from the model. Different types of plots of the residuals from a fitted model provide information on the adequacy of different aspects of the model. Graphical methods have the advantage for model validation because they readily illustrate a broad range of complex aspects of the relationship between the model and the data. If the model fit perfectly, the residuals would be small and would approximate any random error from testing.

A common graphical residual analysis plot is the Studentized Residual plot as shown in Figure 2-5. In Figure 2-5, the x-axis is the run order executed during test and the y-axis is the Studentized Residual values. Regression analysis is predicated on the validity of a few entering assumptions. One is that there is no bias amongst the residuals. In laymen's terms, the resulting predictive model underestimates as often as it overestimates. This is shown in figure 2-5 visually when there are roughly as many points below the 0 x-axis as there are above. When modeling multiple conditions with different units such as Threat Altitude in meters and Threat Size as unit-less categories (Small, Medium, Large), a Studentized Residual normalizes these conditions so that the residual value associated with each combination of conditions can be compared to each other. Graphical analysis of studentized residuals are an important technique in the detection of outliers as shown highlighted in the red circle of Figure 2-5.

Figure 2-5. Studentized Residual Plot

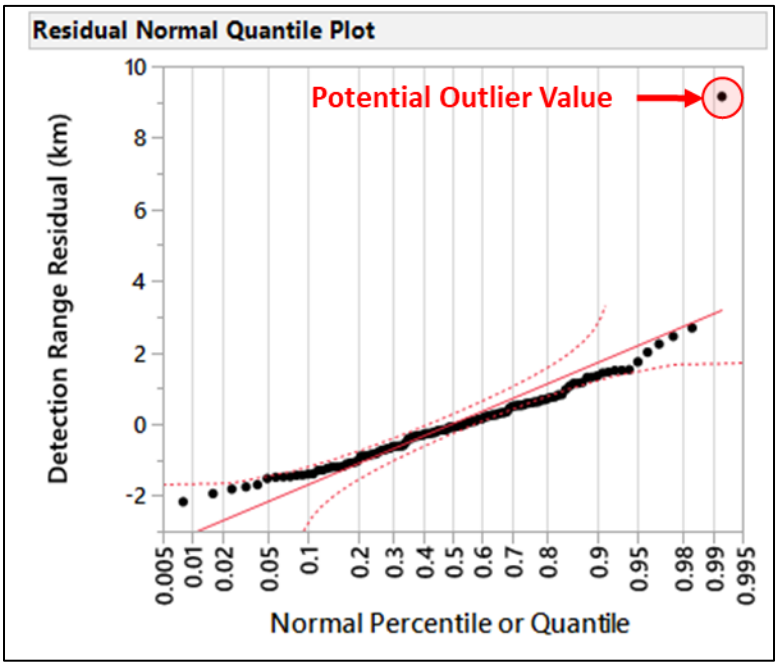
UNCLASSIFIED



Another common residual analysis plot is the Normal Quantile by Residual plot. Quantiles, often referred to as "percentiles" are data points below which a certain proportion of the data falls. For example, imagine the classic bell-curve, or standard Normal distribution, with a mean of zero. The 0.5 quantile or 50th percentile is where half the data lies below the mean of zero. Figure 2-6 is an example of a Residual Normal Quantile Plot where the x-axis is the Quantile or percentile value and the y-axis represents the residual values. The points in Figure 2-6 should form a line (red) that

is roughly straight. If the points do not follow the red line, the assumption of normality amongst residuals is violated. The 01B Analyst will likely need to perform a transformation on the responses in order to meet the normality criteria. When one or two residual values fall outside this line, those data points should be investigated as potential outliers that may skew the RV analysis. The OTD should work their 01B analyst to determine if outliers should be included or excluded from the final model.

Figure 2-6. Residual Normal Quantile Plot
UNCLASSIFIED



These model validation examples are a few common graphical displays used, but are not all-inclusive.

2.3.1.2.3 Importance of main effects and interactions

Once the model is validated, the model can determine which conditions impact the SUT's RV performance. Figure 2-7 exemplifies the results of the model's effect test. This test shows a significance test for each condition in the model. The test for a given condition tests with a given condition against when all parameters associated with that condition are zero. In Figure 2-7, the L-R ChiSquare and *Prob>Chisq* or p-value provide information about whether each individual condition or interaction is related to the response. The p-value is the probability of finding the observed results when removing the condition from the model. 01B analysts are looking for small p-values as these are indications that the conditions have an effect on the RV. It is good practice to decide in advance of the test how small a p-value is required to note an effect. This is exactly analogous to choosing a significance level (generally the significance level for operational test below 0.2).

Figure 2-7. Importance of Main Effects and Interactions
UNCLASSIFIED

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
Threat Size	2	168.87623	<.0001*
Threat Altitude_m	1	0.3206247	0.5712
Threat Size*Threat Altitude_m	2	0.5505252	0.7594

These tests are known as *partial tests*, because each test is adjusted for the other conditions in the model. This highlights this type of analysis: The ability to determine if a condition has effect, not in isolation, but in lieu of all other conditions with potential effect. If conditions are correlated, the *p*-values can change a great deal as other variables are added to or removed from the model. Note that there are other types of tests for individual conditions that are available, but this discussion is beyond the scope of this handbook and we limit our discussion to partial tests.

2.3.2 What do we do with it?

The following general policy applies once the 01B Analyst has completed the RV analysis (See Appendix B for the full checklist of requirements):

- All RV analyses findings will be documented in a RV Analysis Outbrief in accordance with the approved PowerPoint template (File: "RV Analysis Outbrief Template.ppt," found at Y:\OT&E Reference Library\OT Analysis Handbook).
- All RV Analysis Outbrief PowerPoints will be peer reviewed by a second 01B Analyst.
- If the OTD is relying on an analyst from a different COMOPTEVFOR division or external organization to perform the RV analysis, the analyst will forward the completed RV Analysis Outbrief PowerPoint to the 01B Lead Analyst for peer-review assignment.
- The analyst completing the RV analysis will brief the RV Analysis Outbrief PowerPoint to the 01B Test Design Director with the test team members and other stakeholders as applicable, prior to the COI Evaluation Working Group (CEWG).
- Once analysis results are approved by the 01B Test Design Director, the RV Analysis Outbrief will be provided to the test teams prior to the associated CEWG.

The test team shall embed the final RV Analysis Outbrief PowerPoint in the DAS. In the DAS, each RV should have a dedicated section or tab within the document. Within this section, the most important findings of the final RV Analysis Outbrief PowerPoint should be included such as limitations discovered during EDA, the details of the final model, model validations graphs, and effects test results, to include the identification of all factors as; significant, not significant, or undetermined. Factors identified as significant may drive tactics development/operational employment guidance updates and should be included in any future FOT&E test design development. During the draft DAS review with the COMOPTEVFOR Technical Director, the OTD will briefly discuss how the findings/results of any RV analysis have been incorporated into measures evaluation in the DAS.

THIS PAGE INTENTIONALLY LEFT BLANK.

SECTION 3 - INFERENTIAL METHOD: CONFIDENCE INTERVALS

3.1 DISCUSSION: WHY DO WE DO IT?

Just because something is observed in test does not mean it perfectly represents the "real world". So what is the difference between what is learned from test and what the "real world" answer is? Inferential statistics is the means by which one can quantify the uncertainty of test results and provide a level of confidence that the test reflects actual Fleet performance. When the critical measure is not subject to variation due to different conditions and the test objective is to measure a population parameter, like a mean, a confidence interval provides the tester quantification of how accurate the test results might reflect the "actual" SUT performance. A confidence interval for a population mean is probably the most common type, but you can also use confidence intervals for the standard deviation (or sigma), proportions, rates of occurrence, regression coefficients, and the differences between populations. They are not an inferior statistical method to response variable analysis. They simply satisfy a different test objective: examination of stochastic results when there are no condition/factor effects.

3.2 TEST DESIGN: SAMPLE SIZE CALCULATION

3.2.1 How do we do it?

The test team will work with the 01B team to determine an acceptable range of uncertainty in the response and choose the correct type of confidence interval based on the objectives for the measure. Desired sensitivity of test must be operationally relevant and consistent with expected test results. Based on the effect size chosen by the test team, the 01B Analyst will calculate a decision table of varied sample sizes versus effect size.

3.2.1.1 Two-side Confidence Intervals

The two-sided confidence interval is used when the test team desires a description of the likely "real world" range of performance of the SUT based on the sample parameter observed in test. A two-sided confidence interval for a population parameter is defined as an interval with margins of error above and below the sample parameter. The ends of the two-sided confidence interval are respectively called the lower confidence limit (LCL) and upper confidence limit (UCL).

The width of the confidence interval, which represents the precision of the performance estimate, is affected by the following:

- Sample size or N: the larger the sample size, the smaller the interval, all other quantities held constant.
- Confidence level: the higher the confidence level, the larger the interval, all other quantities held constant.

- Standard deviation. The larger the standard deviation, which estimates run-to-run variability, the larger the interval, all other quantities held constant

3.2.1.2 One-sided Confidence Intervals

One-sided confidence intervals are used to test confidence of a measure meeting or exceeding a threshold requirement. In other words, performance of the SUT is expected to fall on the "passing side" of the given threshold. A margin of error is attached to the "non-passing" side and reveals the range of performance values that might occur with the chosen level of confidence.

A one-sided confidence interval for a population parameter requires the construction of a margin of error on one side of the parameter—lower or upper. If the margin of error is placed on the low side, the LCL is the left-most value of this margin. The one-sided confidence interval, therefore, consists of all values AT and ABOVE the LCL. Similarly, for a margin of error on the high side, the UCL is the right-most value of the margin and the one-sided confidence interval consists of all values AT and BELOW the UCL.

3.2.2 What do we do with it?

OPTEVFOR typically designs tests for an 80% confidence level. Just as with response variable critical measures, the test team shall use historical data (if available) or subject matter expertise to determine the anticipated distribution and standard deviation (variability) of the critical measure. Given these two fixed inputs, and threshold values if applicable, sample size becomes the primary decisional variable when evaluating which confidence interval (i.e. level of uncertainty between the test and the "real world") proposed by the 01B Analyst is acceptable as minimum/adequate. When the minimum/adequate sample size has been confirmed by the test team, the 01B Analyst will draft the "Sample Size for Statistical Significance" and "Post-test Analysis" of the associated critical measure in the IEF in accordance with the IEF Checklist. The test team will present the test design at decisional meetings in accordance with the IEF Checklist and Test Planning Handbook.

3.3 TWO-SIDED CONFIDENCE INTERVAL CALCULATIONS ON A MEAN FOR CONTINUOUS DATA

3.3.1 How do we do it?

3.3.1.1 Using JMP®

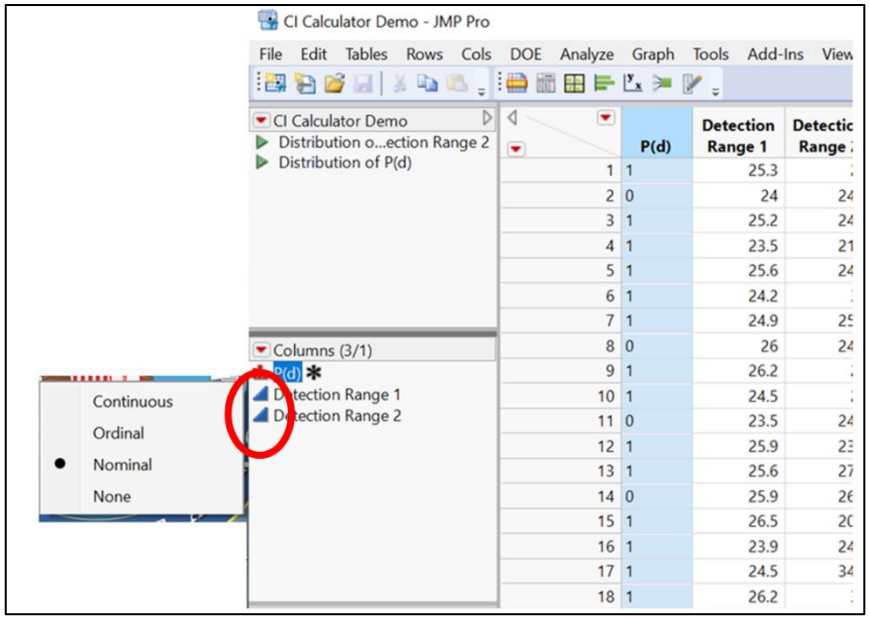
JMP® is a statistical and data visualization software package that is available for all OPTEVFOR personnel. The following steps guide test teams in calculating a two-sided continuous confidence interval using JMP®:

1. Open your data in a JMP® data table. If your data is saved in a column of an Excel spreadsheet and has been checked for errors (all numeric values for a continuous measure or only two responses for a binomial, i.e. no typos), simply open JMP® and then from the top bar, select Files > Open. Navigate to your Excel file and select your file. An "Excel Import Wizard" will open. Select the sheet in Excel where your data is stored. Select import. For other methods of importing data, see Appendix A for links to tutorials on basic JMP® functionality.

2. From an open JMP® data table, ensure your data is coded correctly prior to generating Confidence Interval calculations. Proceed to Step 3 for coding verification.
3. Continuous measures should have a blue ramp/triangle icon next to the Column title in the "Columns" box on the far middle-left portion of the screen. If it is not correct, left click on the column name within the "Columns" box and select "Continuous". The icon should change (See Figure 3-1).

Figure 3-1. Verifying JMP® Data Type

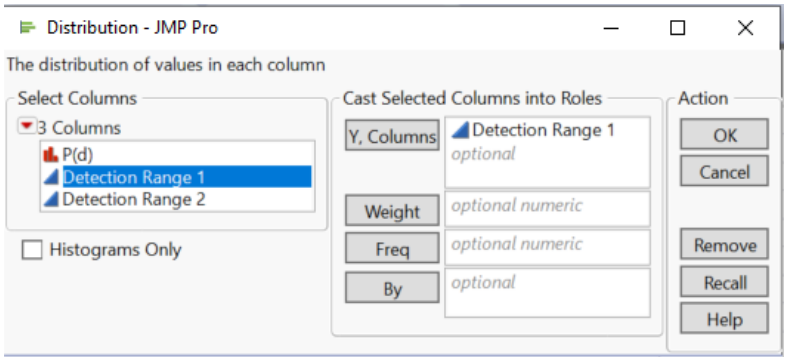
UNCLASSIFIED



4. From an open JMP® data table, select Analyze > Distribution. (See Figure 3-2)
5. Select the variables from "Select Columns"
6. Click "Y, Columns"
7. Click "Ok"

Figure 3-2. JMP® Distribution Window

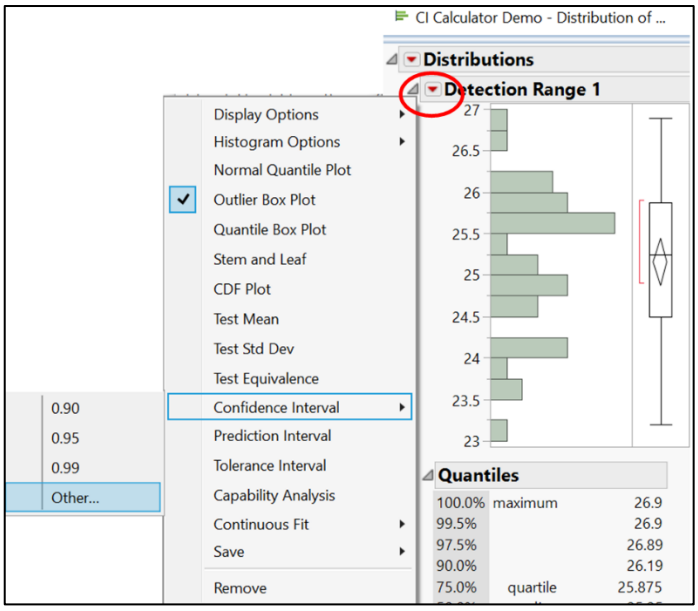
UNCLASSIFIED



8. In the resulting window, click on the red triangle drop down for the variable and select Confidence Interval and then “Other…” (See Figure 3-3)

Figure 3-3. JMP® Distribution Output Window

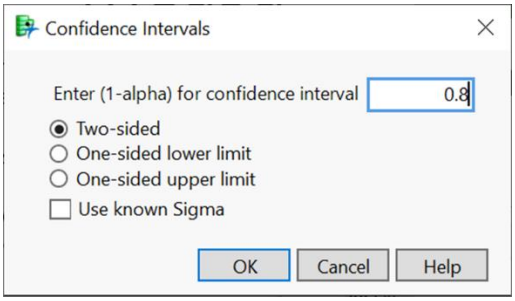
UNCLASSIFIED



9. In the resulting window, simply enter the desired confidence level and which type of CI calculation is needed. Type in 0.8 and select Two-sided for an 80% two-sided confidence interval. (See Figure 3-4)

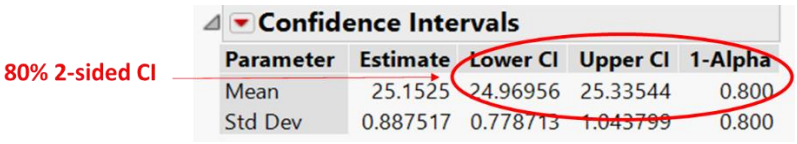
Figure 3-4. JMP® Continuous Measure Confidence Interval Input Window

UNCLASSIFIED



10. The solution is added to the bottom of the Distribution Output Window (See Figure 3-3). The solution will look similar to that in Figure 3-5 below.

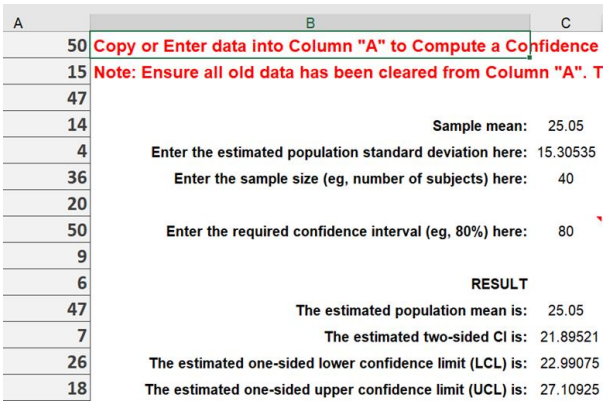
Figure 3-5. JMP® Continuous Measure Confidence Interval Output



3.3.1.2 Using Microsoft Excel

For those test teams that do not have access to JMP® software, an Excel tool has been developed for OPTEVFOR use titled "COTF CI Calculator.xls" and is available at Y:\OT&E Reference Library\OT Analysis Handbook. Select the worksheet "Continuous Measure" and follow the on-screen directions to copy the test data results into Column A. Once the data is properly copied/entered, select the desired confidence interval (typically 80%) and select enter. The results will populate at the bottom of the worksheet.

Figure 3-6. Microsoft Excel COTF CI Calculator Screen Shot



3.3.1.3 Using Online Calculators

There are many calculators available online. See Appendix A for a list of URLs for available online tools.

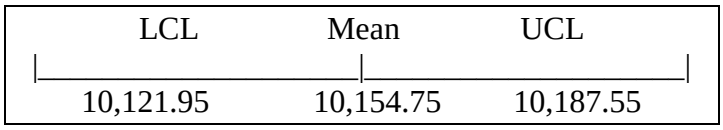
3.3.2 What do we do with it?

Confidence Intervals shall be calculated prior to the CEWG. Confidence Intervals for critical measures are addressed, COI-by-COI, in the data section of the DAS in accordance with the Test Reporting Handbook.

3.3.3 Example

A test team ran a test to assess the maximum range of a gun weapon system (GWS). Sample size was 30 gun firings. The mean range of test results was 10,154.75 yards. The team calculated the lower confidence limits (LCL) and upper confidence limits (UCL) of the confidence interval as 10,121.95 and 10,187.55 yards. The test team concluded that there is 80 percent confidence that this interval contains the real performance of the GWS. Figure 3-7 depicts the two-sided confidence interval.

Figure 3-7. Two-sided confidence interval on a test sample mean



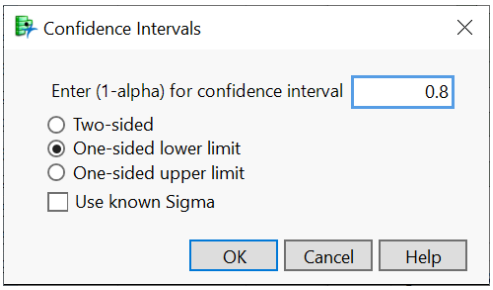
3.4 ONE-SIDED CONFIDENCE INTERVAL CALCULATIONS ON A MEAN FOR CONTINUOUS DATA

3.4.1.1 Using JMP®

JMP® is a statistical and data visualization software package that is available for all OPTEVFOR personnel. The following steps guide test teams in calculating a one-sided continuous confidence interval using JMP®:

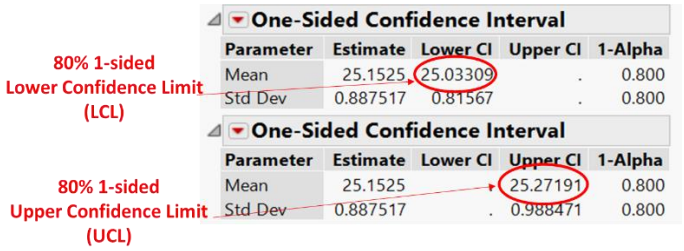
1. Open your data in a JMP® data table. If your data is saved in a column of an Excel spreadsheet and has been checked for errors (all numeric values for a continuous measure or only two responses for a binomial, i.e. no typos), simply open JMP® and then from the top bar, select Files > Open. Navigate to your Excel file and select your file. An "Excel Import Wizard" will open. Select the sheet in Excel where your data is stored. Select import. For other methods of importing data, see Appendix A for links to tutorials on basic JMP® functionality.
2. From an open JMP® data table, ensure your data is coded correctly prior to generating Confidence Interval calculations. Proceed to Step 3 for coding verification.
3. Continuous measures should have a blue ramp/triangle icon next to the Column title in the "Columns" box on the far middle-left portion of the screen. If it is not correct, left click on the column name within the "Columns" box and select "Continuous". The icon should change (See Figure 3-1).
4. From an open JMP® data table, select Analyze > Distribution.
5. Select the variables from "Select Columns"
6. Click "Y, Columns"
7. Click "Ok" (See Figure 3-2)
8. In the resulting window, click on the red triangle for the variable and select Confidence Interval and then "Other…" (See Figure 3-3)
9. In the resulting window, simply enter the desired confidence level and which type of CI calculation is needed. Type in 0.8 and select One-sided lower or upper for an 80% one-sided confidence interval. (See Figure 3-8)

Figure 3-8. JMP® Continuous Measure Confidence Interval Input Window



10. The solution is added to the bottom of the Distribution Output Window (See Figure 3-3). The solution will look similar to that in Figure 3-9 below.

Figure 3-9. JMP® Continuous Measure Confidence Interval Output



3.4.1.2 Using Microsoft Excel

For those test teams that do not have access to JMP® software, an Excel tool has been developed for OPTEVFOR use titled "COTF CI Calculator.xls" and is available at Y:\OT&E Reference Library\OT Analysis Handbook. Select the worksheet "Continuous Measure" and follow the on-screen directions to copy the test data results into Column A. Once the data is properly copies/entered, select the desired confidence interval (typically 80%) and select enter. The results will populate at the bottom of the worksheet. This calculator already takes into account the different calculations for a two-sided or one-sided confidence interval. Simply enter the confidence interval desired and both two-sided and one-sided results will be correctly calculated. See Figure 3-6.

3.4.1.3 Using Online Calculators

There are many calculators available online. See Appendix A for a list of URLs for available online tools. Unless the calculator specifically states it will compute one-sided CIs, double the alpha for one-sided calculations.

3.4.2 What do we do with it?

Confidence Intervals shall be calculated prior to the CEWG. Confidence Intervals for critical measures are addressed, COI-by-COI, in the data section of the DAS in accordance with the Test Reporting Handbook.

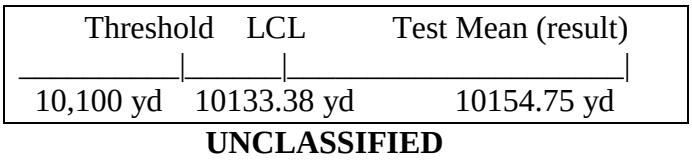
3.4.3 Example

A test team ran a test to assess the maximum range of a gun weapon system (GWS). Sample size was 30 gun firings. There is a threshold value that maximum range exceeds 10,100 yards. The mean range during the test was 10,154.75 yards. The one-sided lower confidence limit (LCL) was 10,133.38 yards. The test team concludes that there is 80 percent confidence that the real performance of the NGFS weapon lies at or above 10,133.38 yards.

The graphical display in figure 3-10 shows how the one-sided confidence interval is used to support the hypothesis that the stated threshold is met. Note that the lower bound of confidence interval (10,133.38 yards) is above the threshold; in other words, the interval excludes that threshold. This provides statistical evidence (at the 80 percent confidence level) that the threshold has been met.

If, however, the LCL was below the threshold value (in this example less than 10,100 yards), the test team cannot conclude that threshold has been met because the LCL value is less than threshold. Although the test result average is 10,154.75 yards, which is above threshold, there is too much uncertainty to conclude that the actual fleet results are greater than threshold. In other words, although the test results qualitatively indicate positive results, the test team cannot conclude quantitatively, with statistical confidence, that threshold has been met.

Figure 3-10. One-sided Confidence intervals on GWS Range



3.5 TWO-SIDED CONFIDENCE INTERVAL ON A BINOMIAL PROPORTION

3.5.1 How do we do it?

3.5.1.1 Using JMP®

JMP® is a statistical and data visualization software package that is available for all OPTEVFOR personnel. The following steps guide test teams in calculating a two-sided continuous confidence interval using JMP®:

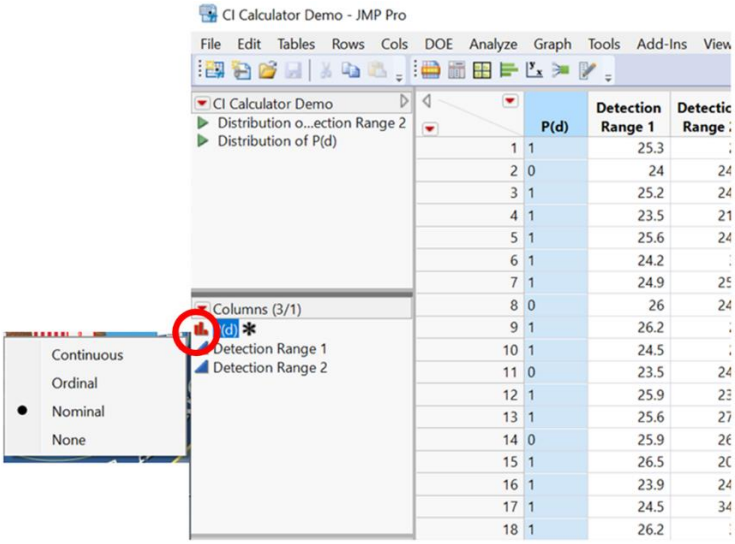
1. Open your data in a JMP® data table. If your data is saved in a column of an Excel spreadsheet and has been checked for errors (all numeric values for a continuous measure or only two responses for a binomial, i.e. no typos), simply open JMP® and then from the top bar, select Files > Open. Navigate to your Excel file and select your file. An “Excel Import Wizard” will open. Select the sheet in Excel where your data is

stored. Select import. For other methods of importing data, see Appendix A for links to tutorials on basic JMP® functionality.

2. From an open JMP® data table, ensure your data is coded correctly prior to generating Confidence Interval calculations. Proceed to Step 3 for coding verification.
3. Categorical and Binary measures should have a red bar chart icon next to the Column title in the "Columns" box on the far middle-left portion of the screen. If it is not correct, left click on the column name within the "Columns" box and select "Nominal". The icon should change (See Figure 3-11).

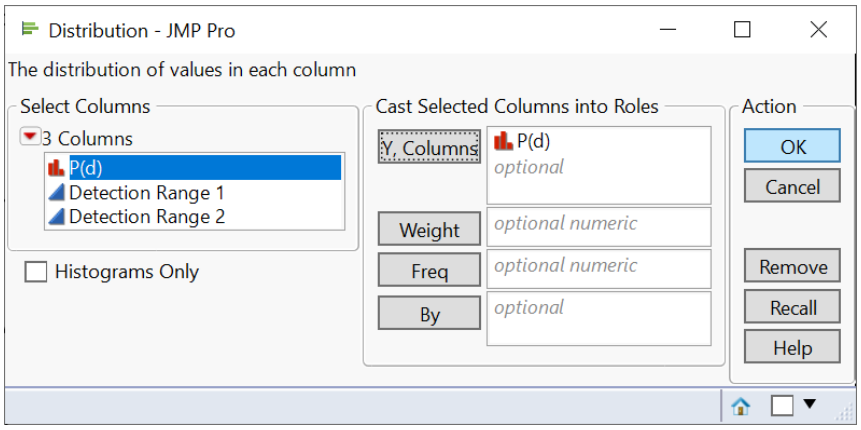
Figure 3-11. Verifying JMP® Data Type

UNCLASSIFIED



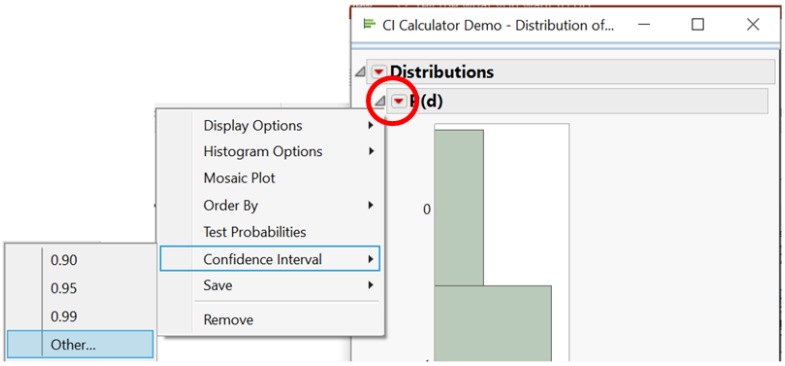
4. From an open JMP® data table, select Analyze > Distribution. (See Figure 3-12)
5. Select the variables from "Select Columns"
6. Click "Y, Columns"
7. Click "Ok"

Figure 3-12. JMP® Distribution Window



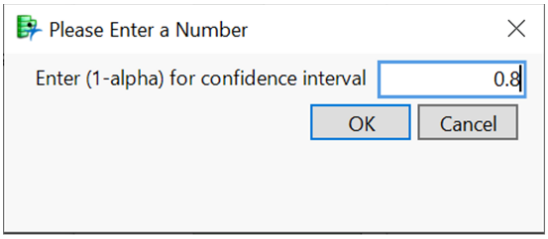
8. In the resulting window, click on the red triangle drop down for the variable and select Confidence Interval and then Other… (See Figure 3-13)

Figure 3-13. JMP® Distribution Output Window



9. In the resulting window, simply enter the desired confidence level and which type of CI calculation is needed. Type in 0.8 and select Two-sided for an 80% two-sided confidence interval. (See Figure 3-14)

Figure 3-14. JMP® Binomial Measure Two-Sided Confidence Interval Input Window

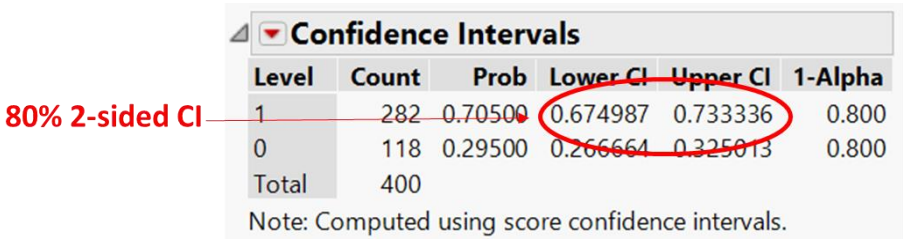


For 80% 2-sided CI

10. The solution is added to the bottom of the Distribution Output Window (See Figure 3-13). The solution will look similar to that in Figure 3-15 below.

Figure 3-15. JMP® Binomial Measure Two-sided Confidence Interval Output

UNCLASSIFIED



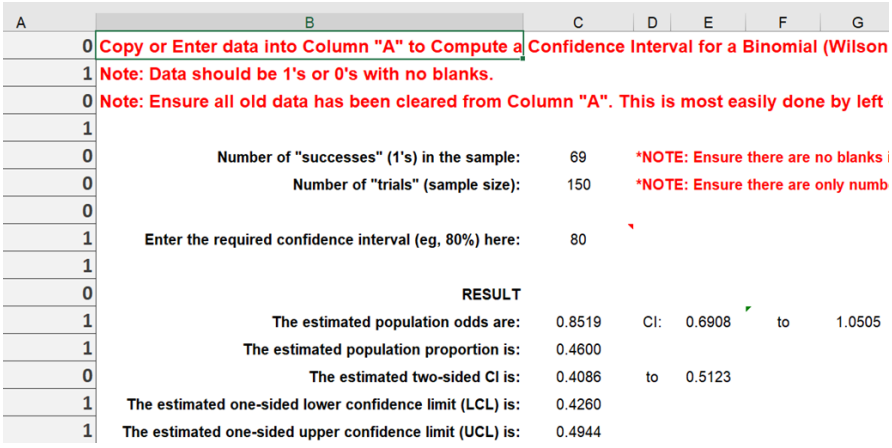
Note: Section 2-7 of the Suitability Handbook guides testers to calculate confidence intervals around two Mean Time Between Operational Mission Failure (MTBOMF) values to see if there is any overlap in values. The previous steps support those calculations. See the Suitability Handbook for further guidance.

3.5.1.2 Using Microsoft Excel

For those test teams that do not have access to JMP® software, an Excel tool has been developed for OPTEVFOR use titled "COTF CI Calculator.xls" and is available at Y:\OT&E Reference Library\OT Analysis Handbook. Select the worksheet "Binomial Measure" and follow the on-screen directions to copy the test data results into Column A. Once the data is properly copied/entered, select the desired confidence interval (typically 80%) and select enter. The results will populate at the bottom of the worksheet.

Figure 3-16. Microsoft Excel COTF CI Calculator Screen Shot

UNCLASSIFIED



3.5.1.3 Using Online Calculators

There are many calculators available online. See Appendix A for a list of URLs for available online tools.

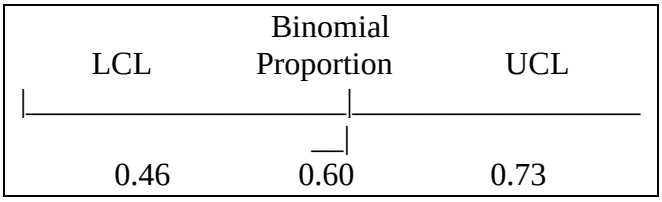
3.5.2 What do we do with it?

Confidence Intervals shall be calculated prior to the CEWG. Confidence Intervals for critical measures are addressed, COI-by-COI, in the data section of the DAS in accordance with the Test Reporting Handbook.

3.5.3 Example

A target detection system is the SUT. During test, twenty targets are deployed, with the goal of assessing probability of detection (P_{DETECT}). The overall P_{DETECT} test result (a binomial proportion) is 0.60 with the LCL and UCL values of 0.46 and 0.73 respectively. The test team concludes that there is 80 percent confidence that this interval contains the real performance of the target detection system.

Figure 3-17. Two-sided confidence interval on a binomial proportion



3.6 ONE-SIDED CONFIDENCE INTERVAL ON A BINOMIAL PROPORTION

3.6.1.1 Using JMP®

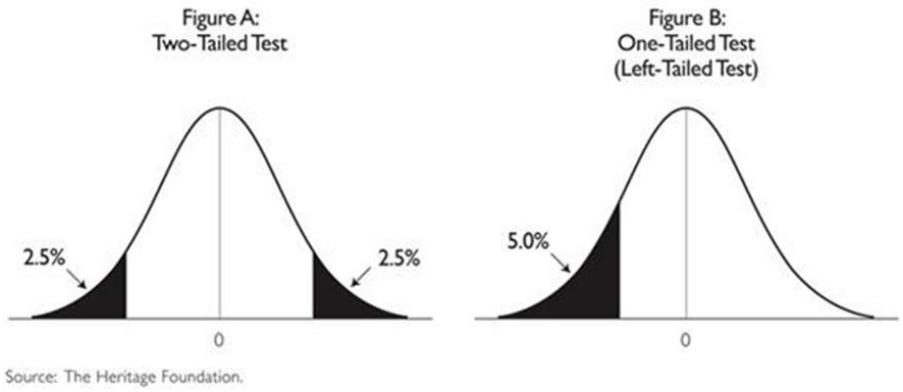
JMP® is a statistical and data visualization software package that is available for all OPTEVFOR personnel. The following steps guide test teams in calculating a one-sided continuous confidence interval using JMP®:

1. Open your data in a JMP® data table. If your data is saved in a column of an Excel spreadsheet and has been checked for errors (all numeric values for a continuous measure or only two responses for a binomial, i.e. no typos), simply open JMP® and then from the top bar, select Files > Open. Navigate to your Excel file and select your file. An "Excel Import Wizard" will open. Select the sheet in Excel where your data is stored. Select import. For other methods of importing data, see Appendix A for links to tutorials on basic JMP® functionality.
2. From an open JMP® data table, ensure your data is coded correctly prior to generating Confidence Interval calculations. Proceed to Step 3 for coding verification.
3. Categorical and Binary measures should have a red bar chart icon next to the Column title in the "Columns" box on the far middle-left portion of the screen. If it is not correct, left click on the column name within the "Columns" box and select "Nominal". The icon should change (See Figure 3-11).
4. From an open JMP® data table, select Analyze > Distribution.
5. Select the variables from "Select Columns"

6. Click “Y, Columns”
7. Click “Ok” (See Figure 3-12)
8. In the resulting window, click on the red triangle for the variable and select Confidence Interval and then Other… (See Figure 3-13)
9. For a one-sided interval in JMP, there is no option given between two-sided and one-sided intervals like there is for a continuous measure. Therefore, the user must modify the value entered for confidence. In a two-sided confidence interval, Confidence = 1-alpha. The one-sided confidence interval is mathematically equivalent to the two-sided confidence interval when the equation is adjusted such that Confidence = 1- (2*alpha). See Figure 3-18.

Figure 3-18. Two-sided versus One-sided Hypothesis Tests

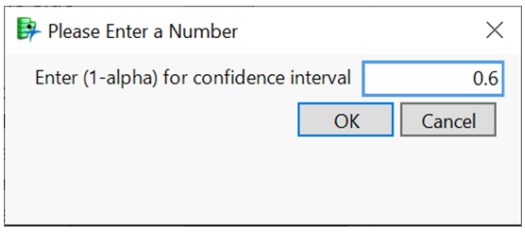
UNCLASSIFIED



10. So, for an 80% two-sided confidence interval: 0.8 Confidence = 1-alpha. Therefore alpha = 0.2. In order to use a two-sided calculator to compute a one-sided confidence interval, adjust the input as described above. In this example, Confidence = 1-(2*alpha) = 1-(2*0.2) = 1 – 0.4 = 0.6. Therefore, in the resulting window input 0.6 for confidence for a one-sided 80% confidence interval (as shown in Figure 3-19). Other desired confidence levels will follow the same adjustment to determine the correct input.

Figure 3-19. JMP® Binomial Measure One-Sided Confidence Interval Input Window

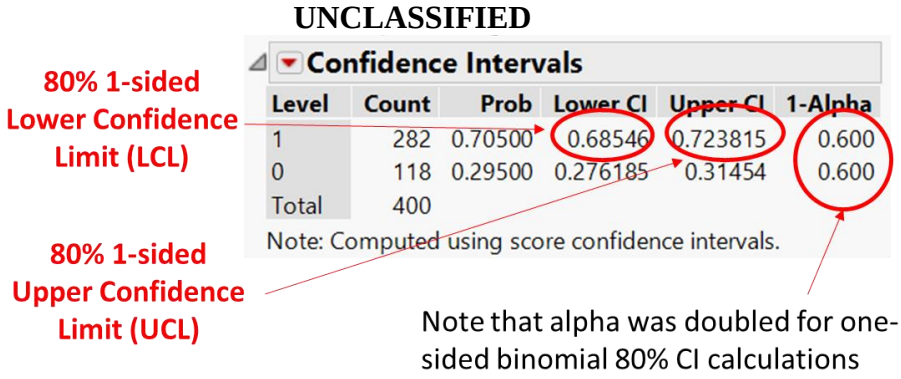
UNCLASSIFIED



For 80% 1-sided CI

11. The solution is added to the bottom of the Distribution Output Window (See Figure 3-13). The solution will look similar to that in Figure 3-20 below.

Figure 3-20. JMP® Continuous Measure Confidence Interval Output



3.6.1.2 Using Microsoft Excel

For those test teams that do not have access to JMP® software, an Excel tool has been developed for OPTEVFOR use titled "COTF CI Calculator.xls" and is available at Y:\OT&E Reference Library\OT Analysis Handbook. Select the worksheet "Binomial Measure" and follow the on-screen directions to copy the test data results into Column A. Once the data is properly copies/entered, select the desired confidence interval (typically 80%) and select enter. The results will populate at the bottom of the worksheet. This calculator already takes into account the different calculations for a two-sided or one-sided confidence interval. Simply enter the confidence interval desired and both two-sided and one-sided results will be correctly calculated. See Figure 3-16.

3.6.1.3 Using Online Calculators

There are many calculators available online. See Appendix A for a list of URLs for available online tools. Unless the calculator specifically states it will compute one-sided CIs, double the alpha for one-sided calculations.

3.6.2 What do we do with it?

Confidence Intervals shall be calculated prior to the CEWG. Confidence Intervals for critical measures are addressed, COI-by-COI, in the data section of the DAS in accordance with the Test Reporting Handbook.

3.6.3 Example

A target detection system is the SUT. During test, twenty targets are deployed, with the goal of assessing probability of detection (P_{DETECT}). Threshold for P_{DETECT} was established to be 0.50. The overall P_{DETECT} test result is 0.60 and the LCL value is 0.51. The test team concludes that there is 80 percent confidence that P_{DETECT} falls at or above 0.51. Because the interval excludes the threshold of 0.50, the test team may conclude that there is statistical evidence to support a conclusion that the threshold has been met. However, if the LCL value is 0.48 instead, the test team cannot conclude that threshold has been met because the LCL value is less than threshold. Although the test result is 0.60, and the threshold is 0.50, there is too much uncertainty to conclude that the actual fleet results are greater than 0.50. In other words, although the test

results qualitatively indicate positive results, the test team cannot conclude quantitatively, with statistical confidence, that threshold has been met. See figure 3-21.

Figure 3-21. One-Sided Confidence Interval on PDETECT

3.7 CONFIDENCE INTERVALS ON DATA WITH UNUSUAL OR UNKNOWN DISTRIBUTIONS

Because the sample mean and binomial proportion are the most commonly used statistics to summarize test results, the confidence interval discussion above focused on these two statistics. Both have known sampling distributions, with straightforward formulas for creating confidence intervals. There are times when there are no theoretical distributions available (non-parametric data) for confidence interval creation or when test data are shown to have unusual distribution properties (e.g., high skewness). For example, if the critical measure is a percentile (like the 70th percentile), there is no derived theoretical distribution, and consequently no analytic approach for creating confidence intervals. Further, consider a situation in which a sample mean is the test statistic of choice, but the sample test data are extremely skewed. In this situation, it is more reasonable to use the median instead of the mean because the median is less subject to the effects of skewness. These are just examples where alternative analytic strategies are needed to create confidence intervals (e.g. Wald method, Likelihood method, Empirical bootstrapping). "Empirical bootstrapping" refers to using the sample data themselves to characterize the shape and nature of the population distribution. If a test team has a critical measure that falls under this category, coordinate with the 01B Analyst for assistance in calculating the associated confidence intervals.

THIS PAGE INTENTIONALLY LEFT BLANK.

SECTION 4 - INFERENTIAL METHOD: HYPOTEHSIS TESTING

4.1 DISCUSSION: WHY DO WE DO IT?

There are many uses of hypothesis testing. The logic of hypothesis testing is fundamental to all inferential statistics. In fact, the means by which a factor is determined to be "statistically significant" in response variable analysis discussed in Section 2 is through a series of hypothesis tests calculated via statistical software packages. This section is not intended to discuss all applications of hypothesis testing throughout the statistics discipline, but discuss common, practical applications used in operational testing. Hypothesis testing is the appropriate quantitative tool to answer test team questions like, "Does this version of the SUT perform better than the last version?"

4.2 COMMON TYPES OF HYPOTHESIS TESTS IN OPERATIONAL TEST

In IT/OT, a hypothesis is a proposition regarding the operational effectiveness or suitability of the SUT. In the hypothesis discussed in this section, known as null hypothesis testing, there are two hypotheses of concern—the null and alternative. The alternative hypothesis is the proposition that the analyst is interested in addressing. The null hypothesis is the logical obverse of the null statement. Essentially, a proposition regarding the effectiveness of a SUT is supported when data indicates that the null hypothesis is unlikely and is rejected. For rejecting a null hypothesis, a test statistic is calculated. This test-statistic is then compared with a defined critical value and if it is found to be greater than the critical value, the hypothesis is rejected. Below are commonly used test statistics.

4.2.1 T-test

A t-test is used to compare the mean of two given samples from a continuous measure. A t-test assumes a normal distribution of the samples (See Figure 2-1) and is used when the population parameters (mean and standard deviation) are not known. There are three versions of a t-test:
1. Independent two-sample t-test, which compares mean for two groups. This test statistic is complimentary to the comparison of a two-sided confidence interval for each sample mean of the two groups being evaluated. If the confidence intervals between the two groups do not overlap, there is evidence that the groups are statistically different. The independent two-sample t-test provides an alternative method or additional confirmation in answering the same question. Sample applicable operational test questions:

- Does Version 2.0 of a missile engage land-based targets farther downrange than Version 1.0?
- Can a new field laptop download intelligence pictures faster than the existing laptop used in the fleet?

2. Paired sample t-test which compares means from the same group. Here, one SUT in the same configuration is compared at two different times. It is uncommon that this method is used in

operational test. Typically, if there is a distinguishable pause in time when testing a SUT, there are often changes or improvements made to the SUT which is more suitable for an independent two-sample t-test.
3. One sample t-test which tests the mean of a single group against a known mean or other set value (like a threshold). This test statistic is complimentary to a one-sided confidence interval for a sample mean of a continuous measure. It provides an alternative method or additional confirmation in answering the same question. In a one-sided confidence interval, the result simply reports if the SUT does or does not meet threshold given a fixed confidence level (typically 80%). A t-test can be used to evaluate what the actual confidence is. If a SUT fails to meet objective in a one-sided confidence interval, knowing that it meets threshold with 78% confidence versus 5% confidence is informative. Sample applicable operational test questions:

- Does a new radar (SUT) detect incoming targets greater than 10nm (threshold)?
- Can an operator don protective equipment in less than 5 minutes (threshold)?
- Can a gun weapon system engage an inbound target outside of 100 yards (threshold) from the ship?

4.2.2 Chi-Square Test

A chi-square test for independence examines whether distributions of two categorical variables differ from one another. In operational testing, the most common categorical response is the binomial "Success/Fail" or "Yes/No" response. The collection of these responses results in a probability of an event (e.g. detection, kill, engagement). In IT/OT, the test team might want to know if the binomial responses (probability) are related to a categorical condition such as the SUT version. In laymen's terms, is the "Success/Fail" output related to, or differ, based on which SUT version is used? This test statistic is complimentary to the comparison of a two-sided confidence interval for the binomial distribution of both groups being evaluated. If the confidence intervals between the two groups do not overlap, there is evidence that the groups are statistically different. The chi-square test provides an alternative method or additional confirmation in answering the same question. Sample applicable operational test questions:

- Is the probability of kill better for the new version of the weapon versus the old version?
- For M&S validation: Is there a difference in probability of detection between the M&S results and the live fire results?

Note: When dealing with small sample sizes, an alternative statistical test called the Fisher's exact test will be used. The 01B Analyst will advise the test team on the specific test to be used. The purpose of both is similar; it is simply a matter of selecting the right test based on the amount of data collected.

4.2.3 Two One-Sided Tests (TOST)

We often use hypothesis testing to assess if there is a difference between samples. But, there are times when the objective is to test for equivalence. For results found in the hypothesis tests described above (t-test and Chi-square test), it is not proper to assume that failure to reject the null hypothesis means one can conclude equivalence. Again, failing to detect a difference using

the previous tests does not mean that the samples are equivalent! A different method is required to test for equivalency.

This test is a functional test, meaning it is not a specific null hypothesis test. It draws on the conclusions of two one-sided null hypothesis tests to functionally conclude that the distributions from two samples are equivalent. For example, a test team wants to show that there has been no change in mean detection range between the old version and new version of a radar. The new capabilities of the new version were focused on the interface of the radar with the combat system and had nothing to do with the detection capabilities of the radar. Therefore, the test team intends to show that the new version still detects equivalently to the old version. Assume the test team has decided that a mean detection range within 3nm between the two versions is considered equal performance. The mean detection range for the old version is 20nm.

Test 1: Consider a one-sided t-test under the following hypotheses:

- Null: Mean detection range difference < 3nm (new version – old version mean detection range is 3nm or less)
- Alternative: Mean detection range difference > 3nm

If we fail to reject the null, one cannot conclude the new version's mean detection range is 3nm greater than the current version. It does not mean that one can then jump to the conclusion that the versions are equivalent! It simply means there was not enough evidence to conclude a difference.

Test 2: Consider a second one-sided t-test under the following hypotheses:

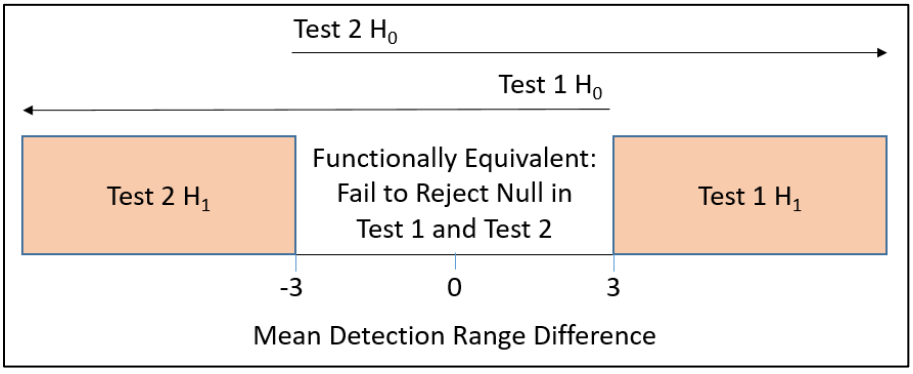
- Null: Mean detection range difference > -3nm (new version – old version mean detection range is -3nm or greater)
- Alternative: Mean detection range difference < -3nm

Again, failing to reject the null simply means that it cannot be concluded that that new version performs worse by 3nm.

By adding the knowledge of the two tests together, one can draw a functional conclusion based on the two inferential hypothesis tests that the two different versions have "equivalent" mean detection ranges (equivalent as defined by the test team as +/- 3nm). See figure 4-1.

Figure 4-1. TOST Equivalency Test Example

UNCLASSIFIED



4.3 HOW DO WE DO IT?

Hypothesis testing is accomplished using the following six steps:

1. Define the hypothesis. "H_0" ordinarily designates the null hypothesis and "H_1" or "H_A" the alternative hypothesis. A properly written null hypothesis contains one of the following: "=," "≥," or "≤." The following is an example pair of null and alternative hypotheses in which the critical measure is mean detection range. The test team intends to prove that the SUT has a mean detection range that is greater than 30nm. In hypothesis testing, what is intended to be proven is the alternative hypothesis and the counter is the null hypothesis. Given that, the hypothesis statement for this test objective would be:

   $H_{0:}$ mean detection range ≤ 30 nm
   H_1: mean detection range > 30 nm

2. Define the assumptions. Include expected distribution of the critical measure (See Figure 2-1), knowledge or lack of "real world" characteristics, and level of significance (confidence and power) selected for minimum/adequate testing (typically 80% for both for operational testing).

   a. It is common for the test team at COMOPTEVFOR to set α to 0.20. Confidence (1 - α) is the probability of not making a Type I error. It is recommended that the analyst set α after weighing the costs of a Type I error. If the cost is estimated to be high, it may be appropriate to set α at levels smaller than 0.20.

   b. Target values of β (probability of Type II error) and 1 – β (statistical power): It is common at COMOPTEVFOR to aim for a power of 0.80. It is recommended that the analyst set β after weighing the costs of a Type II error.

3. Define the test statistic and sample size. The 01B Analyst will select the right test statistic that is suitable for the test objective and calculate the required sample size based on the required confidence and power levels for the test. The 01B Analyst will draft the "Sample Size for Statistical Significance" and "Post-test Analysis" of the associated critical measure in the IEF in accordance with the IEF Checklist.

4. Collect the test data. Execute and collect data in accordance with the Test Planning and Test Execution Handbooks.

5. Calculate the test statistic. Once the data has been scored as valid, calculate the appropriate test statistic. Coordinate with the 01B Analyst for assistance if needed. Ensure the results are available prior to the CEWG in accordance with the Test Reporting Handbook.

6. Draw conclusions. Include the results in the DAS in accordance with the Test Reporting Handbook.

SECTION 5 - INFERENTIAL METHOD: TOLERANCE INTERVALS

5.1 DISCUSSION: WHY DO WE DO IT?

In the previous sections, the focus has been on quantifying the uncertainty around a single "real world" parameter, like an average (mean) or standard deviation. As we have seen, this is done via a *confidence interval,* which (at some level of significance) contains the chosen parameter value. A *tolerance interval,* on the other hand, contains (at some level of significance a chose proportion of the entire range of possible values. Sample applicable operational test questions:

- What range can we expect 90% of all radar detection ranges to fall within?
- Can we be confident that 95% of all transmissions will be less than 5 minutes?

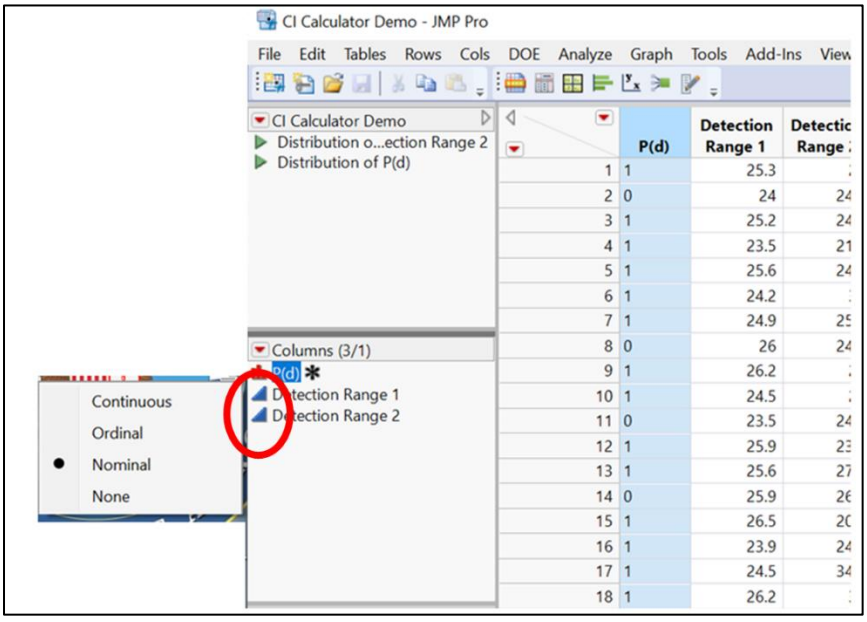
5.2 HOW DO WE DO IT: USING JMP®

JMP® is a statistical and data visualization software package that is available for all COMOPTEVFOR personnel. Although online calculators may exist for tolerance intervals, they are less frequently found than confidence interval calculators are. Unless one is experienced in creating one's own Excel tool or R-coding, JMP® is the recommended tool for calculating tolerance intervals at COMOPTEVFOR. The following steps guide test teams in calculating both two-sided and one-sided continuous tolerance intervals using JMP®:

1. Open your data in a JMP® data table. If your data is saved in a column of an Excel spreadsheet and has been checked for errors (all numeric values for a continuous measure or only two responses for a binomial, i.e. no typos), simply open JMP® and then from the top bar, select Files > Open. Navigate to your Excel file and select your file. An "Excel Import Wizard" will open. Select the sheet in Excel where your data is stored. Select import. For other methods of importing data, see Appendix A for links to tutorials on basic JMP® functionality.
2. From an open JMP® data table, ensure your data is coded correctly prior to generating Confidence Interval calculations. Proceed to Step 3 for coding verification.
3. Continuous measures should have a blue ramp/triangle icon next to the Column title in the "Columns" box on the far middle-left portion of the screen. If it is not correct, left click on the column name within the "Columns" box and select "Continuous". The icon should change (See Figure 5-1).

Figure 5-1. Verifying JMP® Data Type

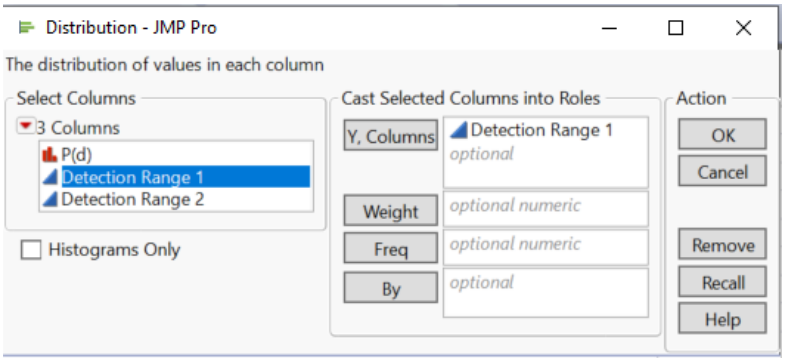
UNCLASSIFIED



4. From an open JMP® data table, select Analyze > Distribution.
5. Select the variables from "Select Columns"
6. Click "Y, Columns"
7. Click "Ok" (See Figure 5-2)

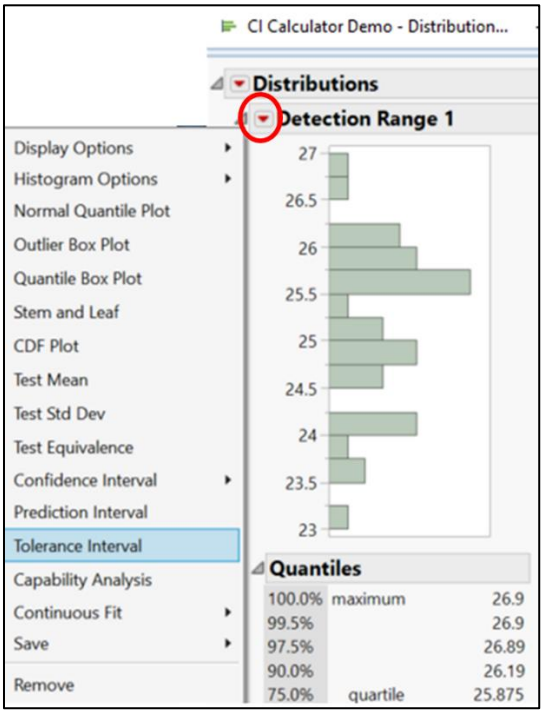
Figure 5-2. JMP® Distribution Window

UNCLASSIFIED



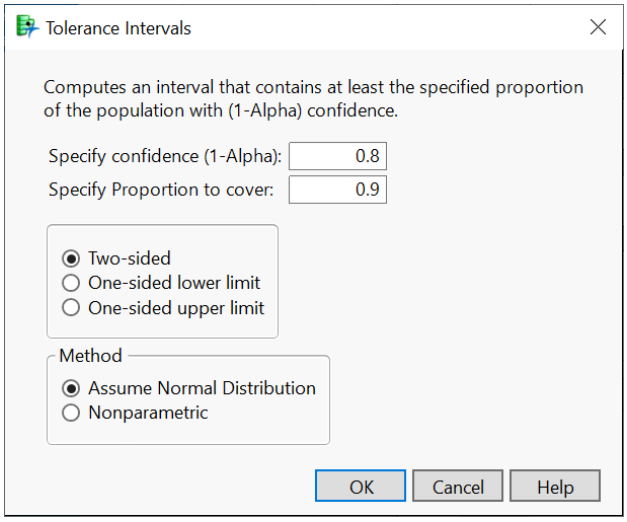
8. In the resulting window, click on the red triangle drop down for the variable and select Tolerance Interval (See Figure 5-3)

Figure 5-3. JMP® Distribution Output Window
UNCLASSIFIED



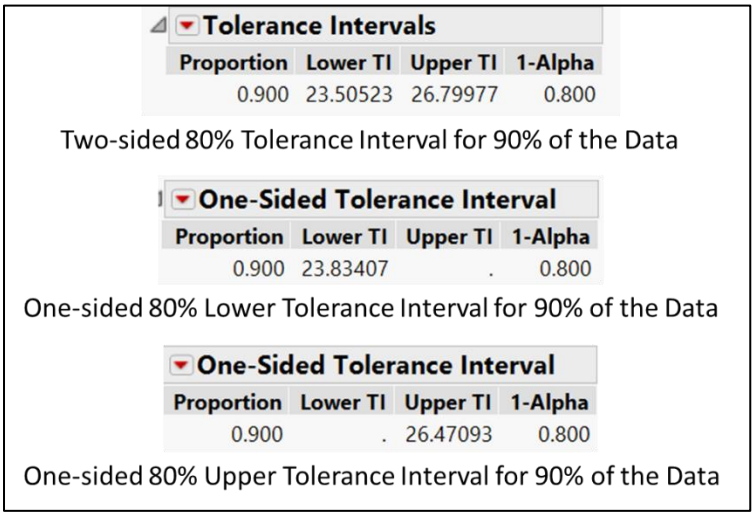
9. In the resulting window, simply enter the desired confidence level and what percentage of the data you wish to bound. Also select if your data is normally distributed or not (ask the 01B Analyst for assistance on this step if you are unclear). For example: Type in 0.8 for confidence and 0.9 for proportion and select Two-sided (See Figure 5-4). The results will produce a range that is interpreted as 80% confidence that 90% of "real world" results fall within that range. A one-sided result will produce a value that is interpreted as 80% confidence that 90% of "real world" results fall above (lower limit) or below (upper limit).

Figure 5-4. JMP® Continuous Measure Tolerance Interval Input Window
UNCLASSIFIED



10. The solution is added to the bottom of the Distribution Output Window (See Figure 5-3). The solution will look similar to that in Figure 5-5 below.

Figure 5-5. JMP® Continuous Measure Tolerance Interval Output
UNCLASSIFIED



5.3 WHAT DO WE DO WITH IT?

Tolerance Intervals shall be calculated prior to the CEWG. Tolerance Intervals for critical measures are addressed, COI-by-COI, in the data section of the DAS in accordance with the Test Reporting Handbook.

5.4 EXAMPLE

A test team ran a test to assess the maximum range of a gun weapon system (GWS). Sample size was 40 gun firings. The mean range of test results was 2,515.25 yards. The team calculated the two-sided 80% confidence interval for the mean (using Section 3 of this handbook) as

2,496.96 (LCL) and 2,533.54 (UCL) yards. The team also calculated the two-sided 80% tolerance intervals (using the steps above) to cover 95% of the data as 2,319.0 (lower tolerance interval) and 2,711.5 (upper tolerance interval) yards. What is the interpretation of these two test statistics?

- Confidence Interval: The test team is 80% confident that the AVERAGE gun engagement range in the real world is between 2,496.96-2,533.54 yards. This was based on an average of 2,515.25 yards observed in the test. Remember, the test average is not ground truth. The confidence interval quantifies the uncertainty of what the real world answer is likely to be.
- Tolerance Interval: The test team is 80% confident that 95% of ALL gun engagement ranges in the real world are between 2,319.0 and 2,711.5 yards. This was based on a minimum of 2,320 yards and a maximum of 2690 yards (and other characteristics of the sample distribution such as sample size and variability) observed in the test. Remember, the test minimum and maximum are not ground truth. The tolerance interval quantifies the uncertainty of what the real world span is likely to be.

After reading the test team's report, the Fleet asked OPTEVFOR if the test team could evaluate, with 95% confidence (instead of 80% used for the COI evaluation) if 95% of all gun engagements were greater than 2,300 yards, which for hypothetical tactical reasons, is a go-no go range for GWS engagements. Good news! The test team answered that they could provide insight to that question with the existing test data. Using the results from the same 40 gun firings, the test team calculated a one-sided 95% tolerance interval to cover 95% of the data as 2,325.61 yards (using the same steps above with the new inputs). Because the calculated lower tolerance level is greater than the goal of 2,300 yards, the test team informed the Fleet that they could be 95% confident that 95% of all gun engagements in the real world are greater than 2,325.61 yards, which means the goal is met.

THIS PAGE INTENTIONALLY LEFT BLANK.

APPENDIX A - REFERENCES ON STATISTICAL THEORY AND METHODS

A.1 ONLINE CALCULATORS

- Institute of Defense Analyses (IDA) Test Science interactive tools: https://testscience.org/interactive-tools/
- Statistics Kingdom: http://www.statskingdom.com/index.html
- Social Science Statistics: www.socscistatistics.com
- Vassar Stats: http://vassarstats.net/

A.2 OTHER T&E STAKEHOLDER REFERENCES

- IDA's Test Science website hosting a collection of T&E training, resources, and tools
- Air Force Institute of Technology STAT Center of Excellence Best Practices and Test Planning Guides

A.3 JMP® TUTORIALS

A.3.1 Exploratory Data Analysis

https://www.jmp.com/en_us/events/mastering/application-areas/data-visualization-and-exploratory-data-analysis.html

Within that link:

- Organizing and Getting the Most from JMP® Tables
- Preparing Data for Analysis
- Exploratory Data Analysis and Dynamic Graphs
- Basics for Using Graph Builder
- Creating, Using and Sharing Journals
- Using Formulas to Get the Most from Your Data

A.3.2 Design of Experiments

https://www.jmp.com/en_us/events/mastering/application-areas/design-of-experiments.html

Within that link:

11. Essentials of Designing Experiments using JMP®

- Specialized Custom DOE for Experienced Experimenters
- Split-Plot and Strip-Plot Design of Experiments
- Handling Constraints When Designing Experiments

- Using Blocking When Designing Experiments

A.3.3 Statistics, Predictive Modeling and Data Mining

https://www.jmp.com/en_us/events/mastering/application-areas/statistics-predictive-modeling-and-data-mining.html

Within that link:

- Producing and Interpreting Basic Statistics Using JMP®
- Data Mining and Predictive Modeling
- Specifying and Fitting Models
- Transforming Data to Make Better Predictions
- Building Better Predictive Models Part 1 and 2
- (Advanced) Using Generalized Regression in JMP® PRO to Create
- Robust Linear Models
- (Advanced) Building Linear Mixed Models Using JMP® PRO
- (Advanced) Fitting Repeated Measures Data using JMP® PRO
- (Advanced) Time Series Analysis and Forecasting

A.4 STATISTICS THEORY AND BACKGROUND

5.4.1 DOE

- Microsoft Word - Size of an Experiment
- Statistical Power Analysis for the Behavioral Sciences
- NIST D-Optimal designs
- Practical Statistical Power Analysis
- Power Calculations for Additive Interactions
- Powerjmp.pdf
- Monte Carlo Power calculations Binary

5.4.2 Statistical tests

- The $\chi 2$ Test of Goodness of Fit
- The chi-square test

5.4.3 Logistic Regression

12. Multiple Logistic Regression Analysis

13. FAQ: What are pseudo R-squareds?
14. Logistic Regression
15. Tests for the Interaction Odds Ratio in Logistic Regression with Two Binary X's (Wald Test)
16. Logistic Regression: Interaction Terms
17. Confidence Intervals for the Odds Ratio in Logistic Regression with Two Binary X's
18. PDF: Importance of Assessing the Model Adequacy of Binary Logistic Regression
19. Residuals from a logistic regression | Freakonometrics
20. Hosmer Lemeshow Applied-Logistic-Regression.pdf
21. Firth Bias Reduction of Maximum Likelihood Estimates on JSTOR
22. Firth Bias Rediction of MLE
23. MLR
24. Wald Statistic LogRegression
25. LogLikelihood Logistic Regression

A.4.1 Collinearity and Variance Inflation Factor (VIF

- Correlation Coefficients for Binary Data In Factor Analysis - Kaltenhauser - 1976 - Geographical Analysis - Wiley Online Library
- Multicollinearity | Introduction to Statistics | JMP
- collinear.pdf
- VIF Incorporating the Multinomial Logistic Regression in Vehicle Crash Severity Modeling: A Detailed Overview
- Variance Inflation Factors In Regression Models With Dummy Variables

A.4.2 LASSO Regression

- Standardization in LASSO | Freakonometrics
- Lasso

Miscellaneous

- Dummy Variable Interactions.pdf
- Poisson Confidence Interval (Incidence Rate) - StatsDirect
- What Is an ROC Curve? - The Analysis Factor
- IDA's Handbook on Statistical Design and Analysis Techniques for Modeling
- and Simulation dated February 2019

- IDA's Purpose of Mixed Effects Models in Test and Evaluation dated August 2019
- IDA Handbook on Power analysis Tutorial for Experimental Design Software dtd November 2014 (no link; inquire with 01B for PDF electronic copy)

APPENDIX B - CHECKLIST FOR RESPONSE VARIABLE ANALYSIS

This checklist leads test teams, supporting analyst, and any outside analysis organization through the preparation and analysis of response variable (RV) datasets. This supplements the PTIP Checklist in the Test Reporting Handbook to specify the parallel process associated with RV analysis prior to the CEWG.

1. Create Data Analysis Package (OTD)

- ☐ a) No later than the completion of the Scoring Board, the OTD will:
  - ☐ i) Create a data package with properly validated and scored data in accordance with the PTIP (see Test Reporting Handbook). As part of the analysis package, also provide the following information to aid in the analyst's initial review (executed immediately following the Scoring Board in accordance with the Test Reporting Handbook):
    - ☐ 1. Project title and name of point-of-contact
    - ☐ 2. Network location of applicable data files and other relevant project documents (Test Plan, IEF, etc.)
    - ☐ 3. Overall objective of the desired statistical analysis if different from what is detailed in the Test Plan
    - ☐ 4. Supporting data processing/wrangling/cleaning methods and analysis from other organizations (e.g. NSWC Corona, NUWC Newport)
    - ☐ 5. Any notes, logs, or other input from the data collectors relevant to measurement/calculation of response, results of a given run, recording of controlled or recordable conditions, etc.
  - ☐ ii) Ensure classified data, or data that will become classified after analysis, are built into data packages on SIPRNET
  - ☐ iii) Data must be organized in an excel spreadsheet by response variable with recordable and controlled conditions (factors) in columns with one row per associated Test Plan run (DOE design point/run),
  - ☐ iv. If not already verified as part of a Scoring Board, verify the appropriate data were collected according to IEF documentation or analysis package. This verification should include but is not limited to the following:
    - ☐ 1. Confirm consistency in units of measure
    - ☐ 2. Confirm the tested design space as set by the controlled factor levels (controlled conditions)
    - ☐ 3. Confirm runs were executed within factor level tolerance
  - ☐ v) OTD is responsible for ensuring this data package is sent to the 01B Analyst or statistician who will be conducting the regression analysis.

Completed by:_____
Date:_____

2. Test Design Review (01B Analyst or outside analysis organization under OTD/LTE management with 01B Analyst peer review)

- ☐ a) Upon receipt of the analysis package, the analyst will:
  - ☐ i. Review system description from the Test Plan and IEF
  - ☐ ii. Review the overall test objective(s) obtained from the Test Plan, IEF, notes, or analysis package documentation prepared by the test team
  - ☐ iii. Identify all response variables/measures, controlled conditions (factors) and levels, and recorded conditions (covariates) listed in the IEF documentation or included in the dataset that require analysis
  - ☐ v. Review any notes made by the OTD or other staff during data collection
- ☐ b) The analyst will document any questions, concerns, or other issues with the dataset in question, for anything not already identified as part of the Scoring Board (e.g. inconsistencies, gaps, etc.)
- ☐ c) Communicate discrepancies back to OTD and O1C AO.

Problem Resolution Required: □ yes / □ no

Completed by:_____
Date:_____

3. Exploratory Data Analysis (EDA) (01B Analyst or outside analysis organization under OTD/LTE management with 01B Analyst peer review)

- ☐ a) The analyst will perform an EDA to identify and document any obvious trends, discrepancies, or other interesting findings
  - ☐ i. Plot each controlled condition, recorded condition, and response variable executed run order and identify any noteworthy characteristics such as[1]:
    - ☐ Trends across actual run order
    - ☐ Design balance across factors
    - ☐ Missing data
    - ☐ Data entry errors
    - ☐ Lack of independence
    - ☐ Randomization
    - ☐ Blocking
    - ☐ Possible outliers

- ☐ ii. Compute relevant summary descriptive statistics for RVs, plus controlled conditions (factors) and recorded conditions of interest
- ☐ iv. Create multi-factor (factor vs. factor) plots across factor combinations of interest to identify missing data or regions of sparse data
- ☐ v. Plot response variable(s) vs. factors of interest and note the distribution that best describes the observed response trend[2]
- ☐ vi. Check for multicollinearity (dependencies among controlled and recordable contions/factors of interest) and compute Variance Inflation Factor (VIF) calculations

☐ b) Compile work from the EDA in Section 1 of the RV Analysis Outbrief PowerPoint format

☐ c) Review the EDA results with 01C AO, OTD, LTE, and other interested parties if there are any issues requiring resolution

Notes:
[1]Strict random order may not be maintained or possible during test execution and therefore the effects of randomization (or lack of) should be considered during the interpretation of run-sequence plots and other randomization related analyses. The analyst should also consider the effect of blocking and hard-to-change factors when analyzing data during an EDA or any subsequent statistical analysis. The as-executed design order/sequencing must be accounted for to ensure accurate characterization of performance.
[2]The general form of the model to be fit will be selected from this result

Problem Resolution Required: □ yes / □ no

Network Location of EDA report:_____

26. Completed by:_____
Date:_____

4. System Characterization and Analysis of Factor Effects (01B Analyst or outside analysis organization under OTD/LTE management with 01B Analyst peer review)

- ☐ a) Based on the results of the EDA in section 3, select the type of model or models that best describe (in general) the observed trends including the factors of interest
- ☐ b) Identify the alias structure of the selected/proposed model
- ☐ c) Using the various model selection techniques (forward, backward, and stepwise) reduce the model[3]
- ☐ d) Check model fit statistics (lack of fit) for the selected model

- ☐ e) Confirm the basic assumptions of the selected model fitting technique are being satisfied. Consider a transformation on the response variable if needed. For linear regression techniques, confirm (at a minimum) the following properties:
    - ☐ Independence
    - ☐ Constant Variance
    - ☐ Normality (distribution dependent) of residuals
    - ☐ Linearity
- ☐ f) Identify influential design points and response values using available diagnostic techniques
- ☐ g) If possible, cross-validate the selected model through validation techniques such as K-fold
- ☐ h) Compare results from the model predictions to the actual data and identify any significant disagreements
- ☐ i) If possible, compare model prediction results to off-design point validation runs and ensure validation results fall within the prediction intervals of the model
- ☐ j) Compile work from the System Characterization in Section 2 of the RV Analysis Outbrief PowerPoint format. Also state the specified significance level (probability of committing a type I error) of any test. If the test objective is to produce a predictive metamodel, include the prediction equation.
- ☐ k) If possible, determine the achieved signal-to-noise ratio (SNR) or effect size to aid in the development of future experimental designs.

Notes: 3The analyst should ensure that model hierarchy is preserve when reducing the model (e.g. factors involved in interactions should have their linear main effects in the model regardless of whether those terms themselves are statistically significant or not) .
Completed by:_____ Date:_____

5. Response Variable Analysis Outbrief (01B Analyst or outside analysis organization under OTD/LTE management with 01B Analyst peer review)

- ☐ a) Based on the outcome of the analysis and model fitting in section 4, draw statistical conclusions that directly address the objective statement(s) in section 1.
- ☐ b) Present results in graphical form (scatter plots, contour plots, etc.) if possible. Graphics should include any relevant confidence, prediction, or tolerance intervals if possible/required.
- ☐ c) All RV Analysis Outbrief PowerPoints will be peer reviewed by a second 01B Analyst. If the analyst completing the RV analysis is from a different COMOPTEVFOR division or external organization, the analyst will forward the

completed RV Analysis Outbrief PowerPoint to the 01B Lead Analyst for peer-review assignment.

- ☐ d) The analyst completing the RV analysis will brief the RV Analysis Outbrief PowerPoint to the 01B Test Design Director with the test team members and other stakeholders as applicable, prior to the COI Evaluation Working Group (CEWG).
- ☐ Once analysis results are approved by the 01B Test Design Director, the RV Analysis Outbrief should be provided to the test teams prior to the associated CEWG in accordance with the Test Reporting Handbook. The final RV Analysis Outbrief PowerPoint shall be embedded in the Data Analysis Summary (DAS).

Completed by:_____ Date:_____

Table B-1 summarizes useful statistical tools and techniques that can be applied throughout the response variable analysis process. This is by no means an exhaustive list of available techniques and tools. The analyst must use his or her judgement in presenting the requisite material pertinent to the specific SUT test objectives.

Table B-1: Summary of useful techniques and statistical tools

Analysis Step	Useful Techniques and Tools
3.a(i)	Run-Sequence Plot, Autocorrelation Plot, Lag Plot, Spectral Plot
3.a(ii)	Histogram, JMP Distribution Fitting Functions, Normal Probability Plot, QQ-plot
3.a(iii)	Mean, Median, Mode, Max, Min, Custom Percentiles/Quantiles, etc.
3.a(iv)	Factor vs. Factor Scatter Plot
3.a(v)	Response-Factor Scatter Plot
3.a(vi)	JMP Correlation Colormap, JMP Multivariate Scatterplot Matrix, Variance Inflation Factor (VIF)
4.a	JMP distribution fitting functions, JMP survival analysis function
4.c	Forward Elimination, Backward Elimination, or Stepwise Elimination, JMP Automated Stepwise Feature, Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), Pareto plot of effects, Normal Plot of Effects
4.d	Goodness of fit tests
4.e	i) Residuals vs. Time Plot, ii) Residuals vs. Fitted Values Plot, iii) Normal Probability Plot of Residuals, and iv) Scatterplot of Response vs. Factor
4.f	H matrix, Cook's D
4.g	Even-Odd, k-fold, and Single Omission Techniques
4.h	Scatter plots, Contour plots, Line plots, and Statistical Intervals

THIS PAGE INTENTIONALLY LEFT BLANK.

APPENDIX C - RELATIONSHIPS BETWEEN POWER, ALPHA, SIGMA, ACTUAL EFFECT SIZE, AND SAMPLE SIZE (N)

Power analysis is the process of determining sample size. "N" usually refers to the total sample size in a test. Power analysis primarily answers the question, "What sample size is needed?" given the values of a desired maximum levels of α and β, estimated σ, and estimated ES.

C.1 DEFINITIONS

- Type I error occurs when test results indicate that the system meets threshold when in fact it does not. In other words, even though test data tell the decision maker to "pass the system," the data are the result of random sampling error and misrepresent the long-run performance of the system.

- α is the probability of making a Type I error. If alpha is 0.20, 1 times in 5, data will indicate in error that the null hypothesis should be rejected.

- Type II error occurs when test results indicate that the system does not meet threshold when in fact it does. In other words, even though test data tell the decision maker not to pass the system, the data are a result of random sampling error and misrepresent the long-run performance of the system.

- β is the probability of making a Type II error. If β is 0.20, 1 time in 5, data will indicate in error that the null hypothesis should not be rejected.

- Confidence (1-α) is the probability of avoiding a Type I error. If α equals 0.20, then 4 times out of 5, the test results will not lead the decision maker to make a Type I error.

- Power (1-β) is the probability of avoiding a Type II error. If β equals 0.20, then 4 times out of 5, the test results will lead the decision maker to correctly reject the null hypothesis.

- Standard deviation (σ) is an index of the variability within a sample or population of data. In many statistics texts, "σ" refers to the population standard deviation, while "S" refers to the sample standard deviation value.

- ES is the difference between the null hypothesized value and the alternative hypothesized value. ES may be expressed in the measurement units of the critical or response variable, or may be expressed in terms of multipliers of the standard deviation. The latter is referred to as signal to noise ratio. Some statisticians refer to ES as delta (δ). For other statisticians, δ is the non-centrality parameter (discussed below).

C.2 IMPORTANCE OF SIGMA

In order to determine N, the analyst must first estimate σ in most statistical tests. (An example where σ is not specifically identified is the exact tests of binomial proportions.) In general, σ

represents mathematical noise that interferes with the statistical test's providing statistically reliable information.

C.3 IMPORTANCE OF ES

In order to determine N, the analyst must estimate the ES in all tests. When the real ES is large, the statistical test is more likely to lead the analyst to reject H_0 in favor of H_1. Further analysis indicates that when the ES is large, Ns may be comparatively small. When the ES is small, the N must be relatively large. Once ES values are estimated, details of power analysis differ depending on the type of the statistical test.

The interrelationship among α, 1-β, σ, ES, and N are summarized in Table C-1.

Table C-1. Relationships Between Power and α, σ, actual effect size, and N
UNCLASSIFIED

State of α, σ, or actual Effect Size	What Happens to Power? (Holding Constant on Other Parameters)	Comments
N increases	Increases	Best way of increasing power
N decreases	Decreases	--
α increases	Increases	When α is increased, Type I error increases
α decreases	Decreases	α may be decreased when the cost of Type I error is high
σ decreases	Increases	Efforts should be made to control variation due to poor test procedures; blocking may be used to reduce σ
σ increases	Decreases	--
real ES increases	Increases	"real ES" is the real difference between the null hypothesis value and the actual; Real ES cannot be modified prior to test
Real ES decreases	Decreases	--

The "takeaways" from this table are as follows: (1) The primary way to ensure adequate power is to modify N. (2) Another approach involves controlling σ with "blocking designs" discussed earlier in this document. Blocking decreases σ. (3) Table 1 indicates that the "real ES" is the real difference between the null hypothesis value and the actual value. "Real ES" is what it is; it cannot be manipulated by the analyst during test planning. However, in estimating the ES, the analyst may identify an ES that is of practical value—a value below which makes little operational difference. (4) Theoretically, α can be increased to increase power. However, best practice requires that α be kept at a level representing maximum acceptable Type I error risk.