INSTITUTE FOR DEFENSE ANALYSES

Scientific Test and Analysis Techniques: Continuous Learning Module

Yevgeniya Pinelis, *Project Leader*

Laura J. Freeman
Heather M. Wojton
Denise J. Edwards
Stephanie T. Lane
James R. Simpson

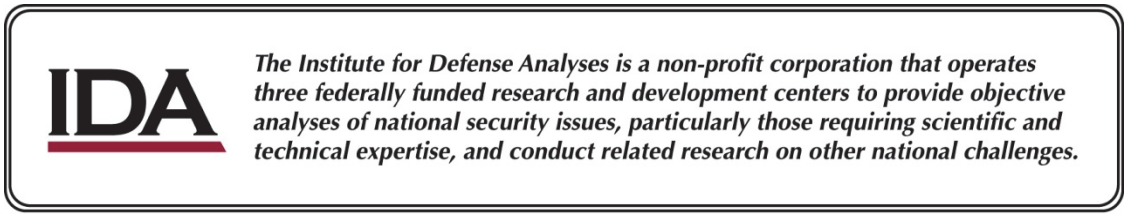
August 2018

Approved for public release.
Distribution is unlimited.

IDA Non-Standard Document
NS D-8920

Log: H 2018 -000016

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001 Task 2299(90), "Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For more information:
Yevgeniya Pinelis, Project Leader
ypinelis@ida.org • 703-845-6899

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS D-8920

Scientific Test and Analysis Techniques: Continuous Learning Module

Yevgeniya Pinelis, *Project Leader*

Laura J. Freeman
Heather M. Wojton
Denise J. Edwards
Stephanie T. Lane
James R. Simpson

Executive Summary

A. Introduction

In recent years the test and evaluation (T&E) community has taken steps to incorporate Scientific Test and Analysis Techniques (STAT) into test planning and analysis. This document provides course material for a new Defense Acquisition University (DAU) continuous learning module. IDA developed the technical content for the course, which the DAU is turning into an online module. The four main course modules are design of experiments, observational studies, survey design, and statistical analysis. It is designed as a four-hour online course and will be included in the required course modules for future test and evaluation certificates issues by the DAU.

B. Lesson 1: Scientific Test and Analysis Techniques (STAT) in Test and Evaluation (T&E) Overview

STAT is defined as the scientific and statistical methods, with associated processes, used to enable the development of efficient, rigorous test strategies so as to yield defensible test results. It encompasses many techniques including, but not limited to, design of experiments, observational studies, and survey design. Lesson 1 provides a summary of the benefits of STAT. It summarizes guidance from the Office of the Secretary of Defense and Service on the use of STAT in test planning and analysis.

C. Lesson 2: Design of Experiments

The second module introduces design of experiments, a process to plan, design, execute, and analyze tests. It summarizes common test design objectives and how those objectives link to evaluation frameworks captured in the Test and Evaluation Master Plan. The lessons provide best practices to select response variables, factors, and design strategies. Numerous examples illustrate the application of experimental design in Defense testing. The lesson provides tools for evaluating test designs including statistical power and correlation of variables.

D. Lesson 3: Observational Studies

Lesson 3 covers observational studies, which are methods to determine how much data is enough when testers cannot directly control the factors of interest in a test. Observational studies often occur during large-scale training exercises or when data is collected during actual operations. The lesson discusses the difference between correlation and causation. Examples illustrate different types of observational studies and how they differ from designed experiments.

E. Lesson 4: Survey Design

Lesson 4 provides the scientific methodology to design and use surveys in Defense testing. Surveys are instruments that measure people's thoughts and feelings on subjects like usability of a system or the workload experienced by the operator. Lesson 4 highlights that surveys should be designed to answer specific evaluation goals. It breaks down the structure of a survey's questions and summarizes best practices for using surveys in Defense testing. Examples of good and bad questions reinforce the best practices provided.

F. Lesson 5: Statistical Analysis

Lesson 5 introduces the concept of a statistical model and its importance in summarizing data from testing. The lesson covers best practices to review data, fit models, and check the assumptions of those models. Topics covered include descriptive statistics, simple-linear regression, multiple-linear regression, model selection, and model checking. An example illustrates the process of fitting a model to data.

Contents

Scientific Test and Analysis Techniques (STAT) in T&E Module Introduction

Welcome

Welcome to the Scientific Test and Analysis Techniques (STAT) in T&E module.

This module provides you with a basic understanding of STAT in T&E and describes several scientific test and analysis techniques in detail.

Topics covered in this module will include:

- Overview of STAT in T&E
- Design of Experiments (DOE)
- Observational Studies
- Survey Design and Analysis
- Statistical Analysis

Module Objectives

The Scientific Test and Analysis Techniques (STAT) in Test and Evaluation (T&E) module is composed of five lessons (Lesson 1 – Lesson 5).

Lesson 1

Overview

- Know the definition of STAT
- Understand the value of STAT in T&E
- Identify the policy and guidance for STAT in T&E

Lesson 2

Design of Experiments (DOE)

- Know the definition of DOE
- Understand the phases of DOE
- Understand the 4 steps in the Plan and Design phases
- Understand design options (classic experimental designs, optimal designs)
- Understand Statistical Measures of Merit (statistical model supported, power, confidence, correlation coefficients)
- Understand design evaluation
- Understand common design mistakes

Lesson 3

Observational Studies

- Define Observational Study
- Distinguish between DOE and an observational study
- Understand why observational studies are used for data collection
- Identify the various types of observational studies

Lesson 4

Survey Design and Analysis

- Understand the purpose of a survey and how it can be used in T&E
- Understand the parts of a survey, structured and unstructured questions, and survey quality
- Understand empirically-vetted and custom-made surveys, and their use
- Understand that surveys are part of a detailed test plan

Lesson 5

Statistical Analysis

- Define statistical analysis and its key elements
- Understand the analysis checklist
- Understand pre-modeling and exploratory data analysis
- Understand several types of statistical models
  - ANOVA
  - Regression
  - General Linear Models
- Understand model selection and checking assumptions

The introduction also will include about six slides that provide generic information about taking a CLM.

DAU has a stock set.

Lesson 1
Scientific Test and Analysis Techniques (STAT) in T&E Overview

= hyper link to a call out

Introduction and Objectives

This lesson introduces an overview of Scientific Test and Analysis Techniques (STAT) in Test and Evaluation (T&E).

Upon completion of this lesson, you will be able to:

- Know the definition of STAT
- Understand the value of STAT in T&E
- Identify the policy and guidance for STAT in T&E

Scientific Test and Analysis Techniques (STAT)

- Scientific Test and Analysis Techniques (STAT): The scientific and statistical methods, with associated processes, used to enable the development of efficient, rigorous test strategies so as to yield defensible test results. (As defined in the Defense Acquisition Guidebook, 15 May 2013)

- STAT encompasses techniques such as design of experiments (DOE), observational studies, and survey design.

- The specific objective(s) of the test determines the suitability and specific application of each method.

Scientific Test and Analysis Techniques (STAT)

- STAT is all encompassing of scientifically defensible methods and processes. In addition to DOE, observational studies, and survey design, there are other methods that are employed on a regular basis.

  They include:

  - Reliability Analysis
  - Reliability Growth
  - Operating Characteristic (OC) curves
  - Natural Experiments
  - Hypothesis Testing
  - Regression Analysis
  - Data Analysis Techniques
  - Optimization
  - Distribution fitting (especially of historical data to inform modifications/ increments)

Value of STAT in T&E

- The proper and early use of STAT produces tests yielding defensible results as well as answering the test objectives, identifying risks of making inaccurate conclusions, and reducing uncontrolled experimental error.

- STAT is applied to test design and analysis throughout all phases of the acquisition life cycle. Various types of test events (e.g., contractor, developmental, live fire, operational, cybersecurity, interoperability, reliability, and M&S) can utilize STAT to achieve defensible results.

Value of STAT in T&E

- A statistical, scientifically based approach to testing also informs the systems engineering process, and enables a better understanding of the true state of technology and system performance throughout the acquisition life cycle.

- STAT enables estimation of technical performance requirements as well as the mission-oriented metrics of operational effectiveness, operational suitability and survivability (including cybersecurity), or lethality over the entire operational envelope.

Policy and Guidance for STAT in T&E

There are policy and guidance on the use of STAT in T&E. Some key ones are listed below:

- DoD
  - DoDI 5000.02, January 2015
  - Defense Acquisition Guide, May 2013
  - DOT&E Guidance on the Use of DOE, October 2010
  - DOT&E TEMP Guidebook 3.0, November 2015

- DoD Components
  - Army
    - AR 73-1, November 2016
    - DA Pam 73-1, June 2017
  - Navy
    - SECNAV 5000.2E / Navy Guidebook
  - Air Force
    - AFI 99-103,
    - AF Memorandum 63119, Certification for Dedicated OT&E

- You should go to the DoD/DoD Component home pages to obtain the latest versions of policy and guidance on the use of STAT in T&E.

Policy and Guidance for STAT in T&E
DoD

(Use as callout page)

DoDI 5000.02, January 2015

- "The Program Manager will…use ***scientific test and analysis techniques*** to design an effective and efficient test program that will produce the required data to ***characterize system behavior*** across an appropriately selected set of factors and conditions."
- "Resource estimates…will be derived from defensible statistical measures of merit (power and confidence) associated with ***quantification of the differences among the factors*** affecting operational performance ***as well as the risk to the government*** of accepting a poorly performing system or incorrectly rejecting a system with acceptable performance. Specifically, the TEMP must discuss and display, or provide a reference to, the ***calculations done*** to derive the content of testing…"

DT: Enclosure 4 ***OT and LF: Enclosure 5***

Policy and Guidance for STAT in T&E

DoD

(Use as callout page)

Defense Acquisition Guide, May 2013

- "A program applying STAT starts early in the acquisition process and assembles a team of subject matter experts to identify the primary evaluation metrics of interest against both the technical performance requirements, as well as the mission-oriented metrics that characterizes the performance of the system and its capabilities in the context of a mission-oriented evaluation."
- "The team identifies the factors, as well as the levels of these factors (i.e., the various conditions or settings that the factors can take), expected to drive the technical and operational performance of the system. The anticipated effects of each of the factors on the evaluation metrics are determined to aid in test planning."
- "To maximize test efficiency, the team uses experimental design techniques to strategically vary factors across the various developmental, operational, and live fire test activities."

Policy and Guidance for STAT in T&E DoD

(Use as callout page)

DOT&E Guidance on the Use of DOE in Operational Testing (OT), October 2010

- The following should be included in the TEMP and Test plans to be approved:
  - "The ***goal of the experiment***…should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment."
  - "***Quantitative mission-oriented response variables*** for effectiveness and suitability."
  - "***Factors*** that affect those measures of effectiveness and suitability….taking into account known information in order to concentrate on the factors of most interest."
  - "A method for ***strategically varying factors*** across both developmental and operational testing with respect to responses of interest. "
  - "***Statistical measures of merit (power and confidence)*** on the relevant response variables for which it makes sense."

Policy and Guidance for STAT in T&E
DoD

(Use as callout page)

TEMP Guidebook 3.0, November 2015

"The authors of the TEMP should employ scientific test and analysis techniques ***to develop a defensible analytical basis*** for the size and scope of the T&E program."

- "Scientific Test and Analysis Techniques (STAT) provide ***ideal tools*** for developing…integrated test events."
- "The ***testing strategy should accumulate evidence*** that the system performs across its operational envelope before and during IOT&E."
- "Often, multiple test designs will be necessary ***to fully characterize system under test mission performance***."
- The Guidebook also provides guidance (with examples) on DOE, observational studies, survey design and analyses, reliability test planning, and Bayesian analysis methods

Policy and Guidance for STAT in T&E

Army

(Use as callout page)

AR 73-1, November 2016

- "Scientifically-based test and analysis techniques and methodologies. Scientifically based test and analysis techniques and methodologies will be used for designing an effective and efficient test program, as well as analyzing the subsequent test data."
- "A top-level scientific and rigorous approach to designing an efficient test program that characterizes the system behavior across a variety of factors and conditions must be described starting at the initial and/or updated Test and Evaluation Master Plan and System Evaluation Plan (and in sufficient detail in subsequent test design plans) as appropriate."
- "Selected test design plan(s) should ensure more efficient integration of all types of testing up to and including a follow-on operational test (FOT)."

Policy and Guidance for STAT in T&E
Navy

(Use as callout page)

SECNAV 5000.2E / Navy Guidebook

- "Reusable application software derived from best value candidates reviewed by subject matter expert peers and selected based on **data-driven analyses and experimentation**."
- "Accelerated test methods (e.g., step stress testing, accelerated life testing, and reliability growth testing) should be used to assure design maturity prior to operational testing."

Policy and Guidance for STAT in T&E
Air Force

(Use as callout page)

AFI 99-103, 5.13.

- "Whenever feasible and consistent with available resources, STAT should be used for designing and executing tests, ***and for analyzing the subsequent test data.***"
- "The ITT should consult a ***STAT practitioner*** whenever test designs are considered."
- "The selected test design(s) should help ensure smoother, more efficient integration of all types of testing…including FOT&E."
- "The PM is responsible for the adequacy of the planned series of tests and reports on the expected decision risk remaining after test completion."
- "If a statistical analysis technique is not being used, discuss the analysis technique that is being used ***and provide rationale***."

Summary

- This lesson introduces an overview of STAT in T&E.

- You should now be able to :
  - Know the definition of STAT
  - Understand the value of STAT in T&E
  - Identify the policy and guidance for STAT in T&E

- The remaining lessons in this module build upon this foundation and provides information on specific scientific test and analysis techniques:
  - Design of Experiments (DOE)
  - Observational Studies
  - Survey Design and Analysis
  - Statistical Analysis

Lesson 2

Design of Experiments (DOE)

= hyper link to a call out

Lesson Objectives

- This module introduces design of experiments (DOE), explains the design phases, gives examples of classic experimental designs and optimal designs, and introduces design adequacy statistical measures of merit
- Upon completion of this lesson, you will be able to:
  - Know the definition of DOE
  - Understand the phases of DOE
  - Understand the 4 steps in the Plan and Design phases
  - Understand design options (classic experimental designs, optimal designs)
  - Understand Statistical Measures of Merit (statistical model supported, power, confidence, correlation coefficients)
  - Understand design evaluation
  - Understand common design mistakes

Experimental Study

- What is an experimental study?
  - In an experimental study, we make purposeful, controlled changes to a set of independent variables, recording the subsequent changes in an outcome (dependent variable)
  - An experimental study differs from an observational study (Lesson 3), as we are controlling conditions (e.g., altitude, threat type), rather than just observing them
  - The validity and the utility of an experiment is determined, in part, by the adequacy of its design
  - We aim to design experiments that answer our questions as rigorously and efficiently as possible

Design of Experiments (DOE)

- Design of experiments (DOE) is a process for planning, designing, executing and analyzing tests in which purposeful changes are made to the input variables in order to observe corresponding changes in outputs, which upon data analysis one can draw conclusions.

- DOE is a systematic, rigorous, statistical approach to problem-solving that applies principles and techniques at the test planning, execution and data collection stages so as to ensure the generation of valid, defensible, and supportable conclusions.

- Proper and early use of DOE produces tests yielding defensible results as well as answering the test objectives, identifying risks of making inaccurate conclusions, and reducing uncontrolled experimental error.

Design of Experiments (DOE)

- DOE provides the framework to help testers systematically address the four challenges that they typically face:

1. How many test points? Depth of testing.
   - Statistical power (the probability that we detect a difference that truly exists) is key to providing the answer. DOE provides tools to calculate statistical power.
   - DOE helps develop analytical trade space for test planning – balancing risk and resources.

2. Which points? Breadth of testing – spanning the operational envelope.
   - DOE provides test design techniques to develop efficient test designs.
   - DOE assists identification of reasonable subset of all possible test conditions.

Design of Experiments (DOE)

3. How to execute? Order of testing.
   - DOE provides techniques to execute tests and avoid systematic confounding of factors.
   - DOE testing sequence ensures protection against unknown background lurking variability (noise).

4. What conclusions? Test analysis – drawing objective, robust conclusions while controlling noise.
   - DOE enables tester to build statistical models or input/output relationships that also quantify noise.
   - In quantifying noise, DOE determines which inputs are significant statistically and quantifies uncertainty in the results.

Design of Experiments (DOE)

- DOE should be used to:
  - Screen for important factors driving performance
  - Optimize system performance with respect to a set of conditions
  - Predict performance, reliability, or material properties at use conditions
  - Test for problem cases that degrade system performance
  - Improve system reliability or performance by determining robust system configurations
  - Compare two or more systems across a variety of conditions
  - Determine whether a system meets requirements across a variety of operational conditions
  - Characterize performance across an operational envelope

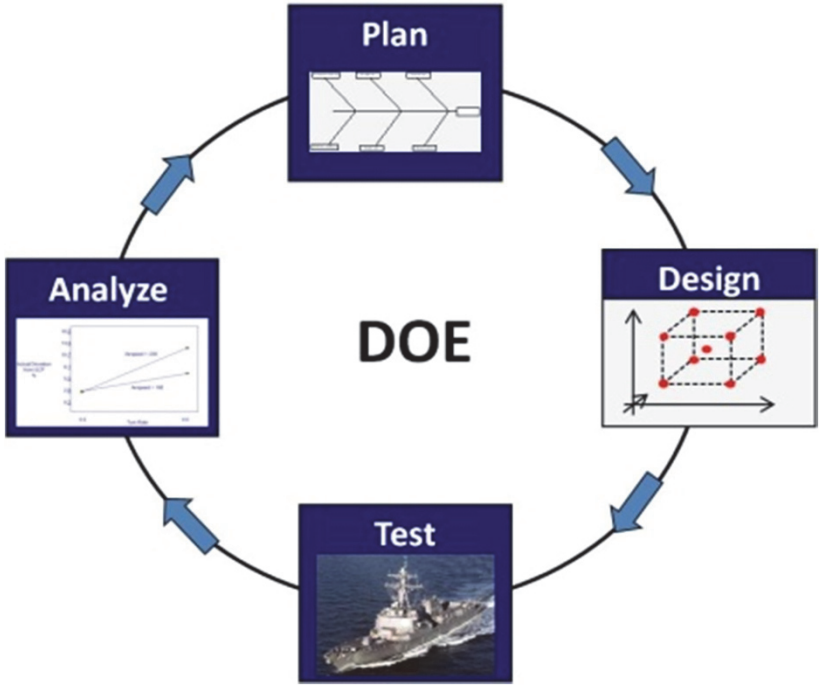
DOE: Common Test Objectives

- Screen for important factors driving performance
- Characterize performance across an operational envelope
  - Note this also implies data will be adequate to determine whether a system meets requirements across a variety of operational conditions
- Compare two systems (or more) across a variety of operating conditions
- Identify factors that degrade system performance
- Optimize system performance in the presence of the anticipated operational environment
  - The primary effects of interest are linear effects, two-factor interactions and second order quadratic factor effects
- Predict performance, reliability, or material properties at use conditions
- Test for problem cases that degrade system performance
- Determine robust system configurations to improve system reliability or performance

Phases of DOE

There are four phases in DOE

1. Plan
2. Design
3. Test
4. Analyze



Steps in DOE

There are seven steps in DOE:

Plan Phase

Step 1 - Define the test objective

Step 2 - Select appropriate response variables

Step 3 - Choose factors, levels, desired model

Design Phase

Step 4 - Choose the experimental design

- Disallowed combinations of factors (safety, operational realism)
- Realistic range for test resources
- Allowable test risk
- Analysis objectives

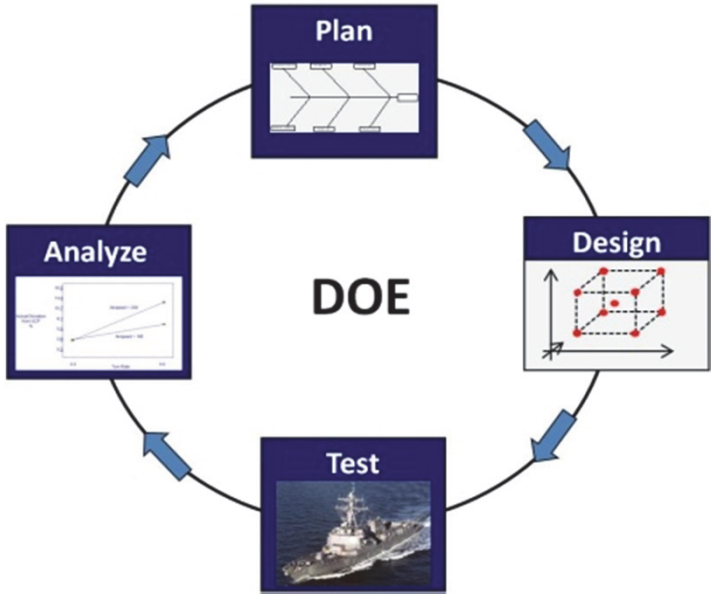
Test Phase

Step 5 - Perform the test

Analyze Phase

Step 6 - Statistically analyze the data

Step 7 - Draw conclusions



This lesson will cover Steps 1 - 4

Plan Phase - Step 1

- Step 1. **Define the test objective**
  - Determine the objective of the test.
  - The objective should be tied to the goals of the analyses that will contribute to assessing the systems progress toward achieving requirements.

Useful resource: Developmental Evaluation Framework and Operational Evaluation Framework, available with the TEMP at MS B

(Use as callout page)

- **The Developmental Evaluation Framework**
  - Identifies key data that will contribute to assessing progress toward achieving system requirements.
  - Shows the mapping between test events, key resources, and the decision(s) supported. It lists the tests, modeling and simulation, or other events that provides the data needed for system evaluation.

- **The Operational Evaluation Framework**
  - Shows how the major test events and test phases link together to form a systematic, rigorous, and structured approach to quantitatively evaluate system capability.
  - Identifies the goal of the test within a mission context, mission-oriented response variables, factors that affect those variables, and test designs for strategically varying the factors across the operational envelope.

Plan Phase - Step 2

- **Step 2. Select the appropriate response variables**
  - The output or response variable measures the outcome of interest for the test (measures of performance or effectiveness, or key performance parameters).
  - Requirements often inform response variable selection.
  - Response variables are best determined by developing a process flow diagram of test execution procedures and listing all possible opportunities to collect meaningful data.
  - Multiple response variables are common and generally necessary.

Useful resource: Developmental Evaluation Framework and Operational Evaluation Framework, available with the TEMP at MS B

Plan Phase – Step 2:
Good Response Variables

- Essential to a defensible experimental design
  - Measurable: they can be measured at a reasonable cost, with accuracy and precision, and without affecting the test outcome.
  - Valid: they directly address the test objective.
  - Informative: continuous responses provide 4-10 times the information per test point than pass/fail metrics (e.g., continuous detection range versus detect/non-detect).
- Provide adequate data to evaluate performance and inform how capabilities development document (CDD) requirements are met/not met (even if the response selected is not explicitly defined in the CDD).

Plan Phase - Step 3

- **Step 3. Choose factors, levels, and desired model**
  - Factors are all the independent variables that are expected to affect the outcome of a test and can be set at specific levels during test execution.
  - Levels are the specific values that the factors can assume. Factor levels are either numeric or categorical.
  - Brainstorm ALL the potential factors that could affect test outcomes – then decide which levels for each factor.
  - The desired statistical model form is chosen based on the test objectives.

Useful resource: Developmental Evaluation Framework and Operational Evaluation Framework, available with the TEMP at MS B

Plan Phase – Step 3: Good Factors

- Characteristics of good factors:
  - Important: choose factors and levels that are expected to have a large quantifiable effect on the test responses. When in doubt, include the factor in the design.
  - Controllable: factors are assumed to be able to be controlled (i.e., set to a specific level) at a reasonable cost.
  - Informative: numeric factors are preferred to categorical factors (e.g., if altitude is a factor, the preferable levels are 5,000, 10,000, and 15,000 as opposed to low, medium, and high)

Design Phase (Step 4)

- **Step 4. Choose the experimental design**
  - Build alternative designs and provide statistical measures of merit for each alternative to then compare and select the best design
  - Design options include:
    - Full Factorial Design
    - Fractional Factorial Design
    - General Factorial Design
    - Response Surface Design
    - Optimal Design
  - Considerations in choosing a design include:
    - Analysis objectives
    - Available test resources
    - Allowable test risk
    - Disallow combinations of factors to meet safety, operational realism

Design Phase (Step 4)

- **Step 4. Choose experimental design**
  - It is important to choose the right design(s) for your particular tests or series of tests for your program.
  - Each design has unique features, along with relative strengths and weaknesses, that will be more appropriate for the specific conditions of the test.
  - Test planning involves building and comparing alternative designs using a dozen or more design metrics.
  - It is common to consider more than one design type (e.g., response surface and an optimal design) for the same test program.

Factors, Levels and Test Points

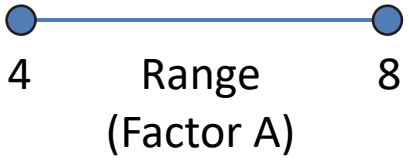
- To describe the notion of factors, levels and test points, we will use the following example
- Consider a simple test involving a gravity weapon being released from an aircraft toward a ground target with 3 factors: Range to the target (A), release Altitude (B) and release Airspeed (C).
- We will assume 2 numeric values for each factor.

Label	Factor	Low Level	High Level
A	Range (nm)	4	8
B	Altitude (ft)	5000	10000
C	Airspeed (knots)	450	650

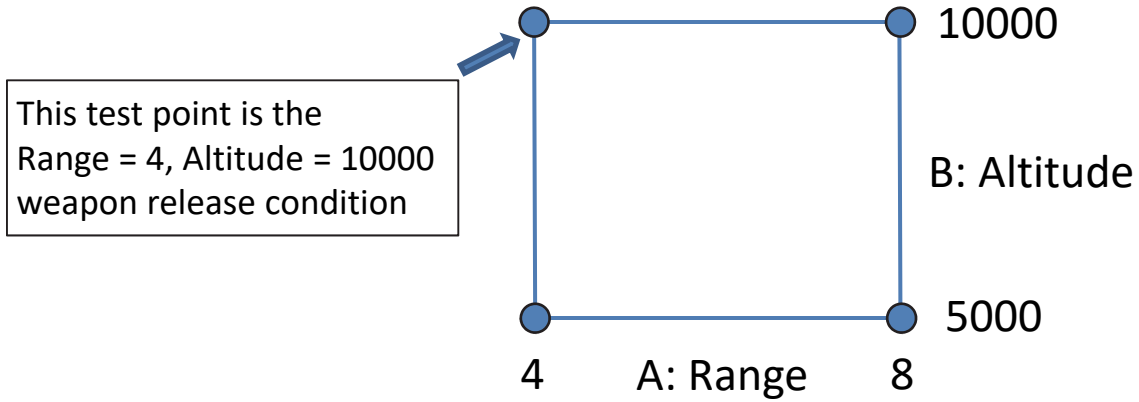
- This example will first serve to illustrate how to visualize a test design in what we call a **test space**. A test space is a geometrical representation of the test points, which are combinations of specific levels of each factor in the design. Each factor will be shown graphically as a dimension in the space.

Factors in a Test Space

- We compare designs with test points allocated to combinations of factors and levels geometrically, where each factor occupies a physical dimension.
- With only one factor (Range) with two levels (4,8) the design can be displayed as:

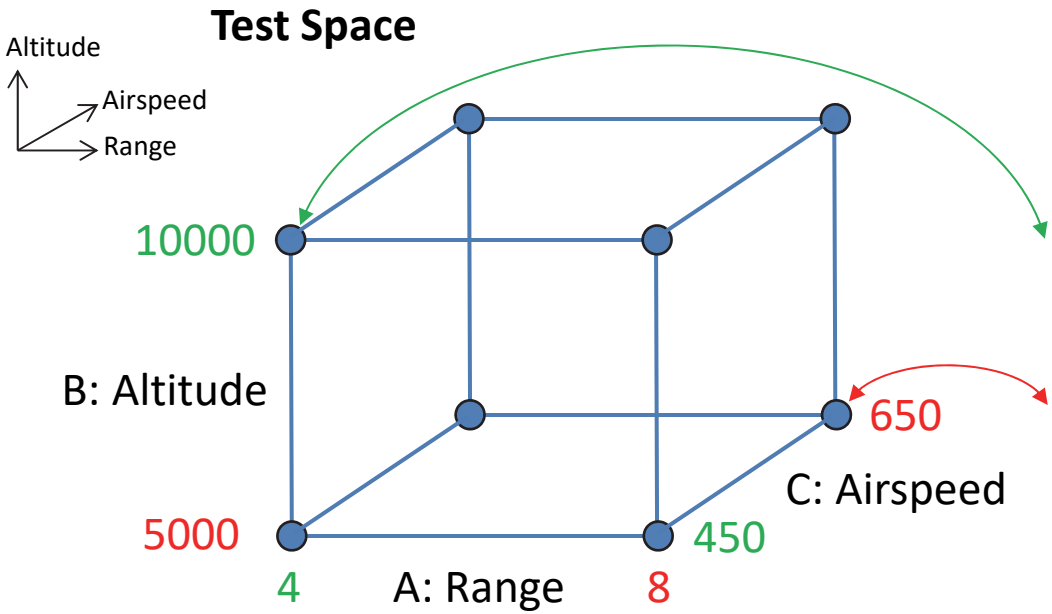


- With two factors (Range, Altitude), with two levels each (4,8 for Range; 5000,10000 for Altitude), the design can be displayed as:



3 Factor Test Space and Design

- Add the third factor (Airspeed) with two levels (450, 650), and the 3-factor test space is:



Test Design Table

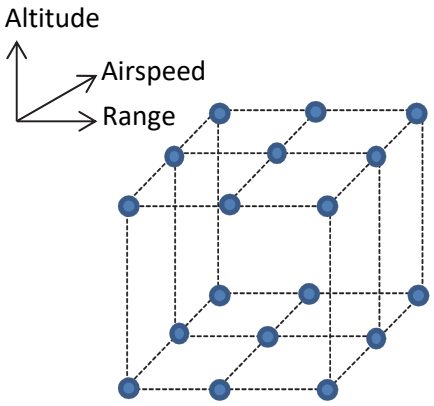
A: Range	B: Altitude	C: Airspeed
4	5000	450
8	5000	450
4	**10000**	**450**
8	10000	450
4	5000	650
8	**5000**	**650**
4	10000	650
8	10000	650

- There is one test space point for each run in the test design table

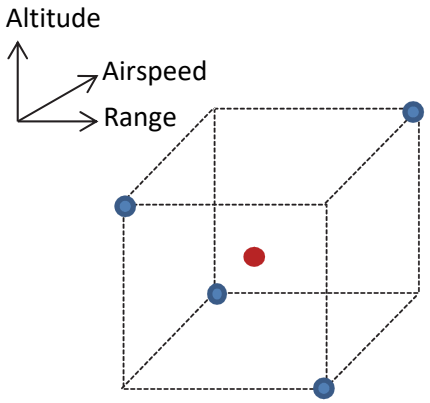
Designs in 3 Factors

- Designs can be illustrated by showing which factor combinations (test points) are included and located in the test space
- Points on the vertices of the test space are called corner points, points at the centroid are called center points, and points along the dimension axes are called axial points
- Additional levels can be added along a dimension (factor) as well

Numeric or categorical factors can have more than two levels. This design shows that the range and airspeed factor dimensions have 3 levels, while altitude has 2 levels.



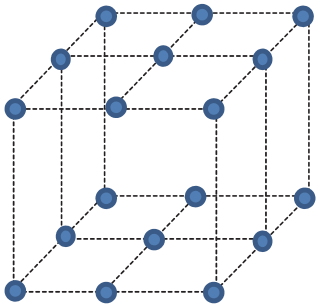
General Factorial
3x3x2 design



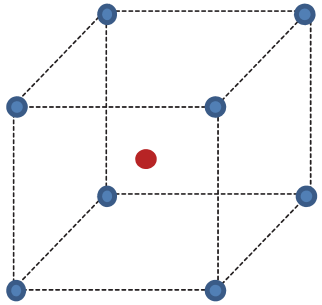
Fractional Factorial
2^{3-1} design

Designs can contain only a subset of the General Factorial corner point combinations. These subset points are chosen to enable estimation of important model terms. The center point is colored red to indicate 3-5 replicate points are located there.

Typical Test Design Alternatives

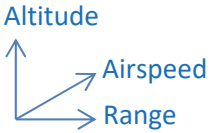


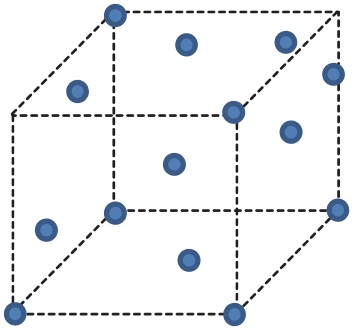
General Factorial
3x3x2 design



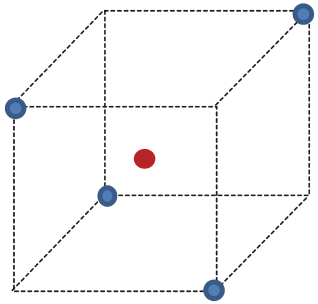
2-level Factorial
2^3 design

"Just Enough"
test points:
– most efficient

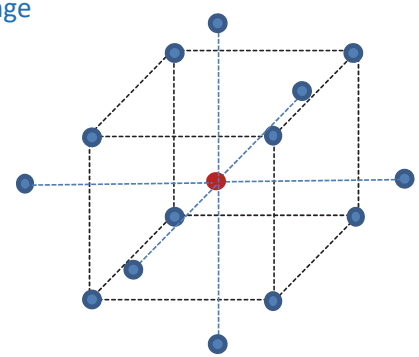




Optimal Design
IV-optimal



Fractional Factorial
2^{3-1} design



Response Surface
Central Composite design

single point
3-5 replicated points

Good Designs

- Characteristics of good designs:
  - **Correlation**: the test points should be placed in the test space to minimize the correlation among factors in the analysis of the data.
  - **Replication**: the design should include replicated points, such that factor combinations are run more than one time. Replication provides the critical estimate of system noise (unexplained variation). 3-5 replicate points are sufficient.
  - **Sequential**: the design should be capable of being run in stages of subsets of point so that meaningful analysis can be performed after each stage, allowing for valuable discovery to take place and improve the decisions for future test stages.

(Use as callout page)

Test Designs

Design	Strengths	Weaknesses

2-level Full Factorial Design (2^k, where k = number of factors)	• Design considers all possible combinations of factors and levels
	• Design strategy is simple. Design can be easily planned, executed and understood
	• Efficiency is gained by only considering 2 levels of each factor
	• Design results in a modest number of runs when the number of factors is less than 5
		• Design is often not practical when the number of factors is five or more because the number of runs doubles per factor added; i.e.,
		  • 4 Factors will result in 16 runs
		  • 5 Factors will result in 32 runs
		  • 6 Factors will result in 64 runs

2-level Fractional Factorial Design	• Excellent design choice when the number of factors is 5 or greater. • Also simple to understand and build • Easy design to augment with only a few runs to better know the true underlying statistical model • Typical fractional choice that is both efficient and informative allows for estimating all main effects and some or all 2-factor interactions	• Some fractions alias or confound too many model terms, thus difficult to determine which factors and interactions are influential • Additional knowledge/skill is needed to choose the better fractions for the given situation

(Use as callout page)

Test Designs

Design	Strengths	Weaknesses
General Factorial Design	• A design to consider when at least 1 factor requires more than 2 levels, often true with categorical factors • Contains all possible combinations of factors and levels • Simple design construct and easy to build, test and analyze when only 3 or less factors	• Design can require too many runs for as few as 3 or more factors and more than 2 levels. For example, a 3-factor test with a 3-level, a 4-level and a 5-level factor would require 3x4x5 = 60 runs
Response Surface Design	• Excellent design choice for tests with mostly numeric factors, and also if 1 or more factors may influence the response in a curvilinear fashion (second order model) • Most of the Response Surface Design choices are moderately efficient (fewer runs), especially considering the number of runs needed to estimate a second order model • Most of the Response Surface Designs can be built sequentially • The design points are placed in logical locations in the design space and each point contributes to a specific model term (linear, interaction, or quadratic) • Designs have impressive design metric values (support the intended model, low correlation, etc.) and design properties (robust to outliers, provides estimate of background variability or noise)	• Some design alternatives, such as optimal designs, for mostly numeric factors and a second order model are more efficient • If 2 or more factors are categorical in the design, more efficient design alternatives, including optimal designs should be considered

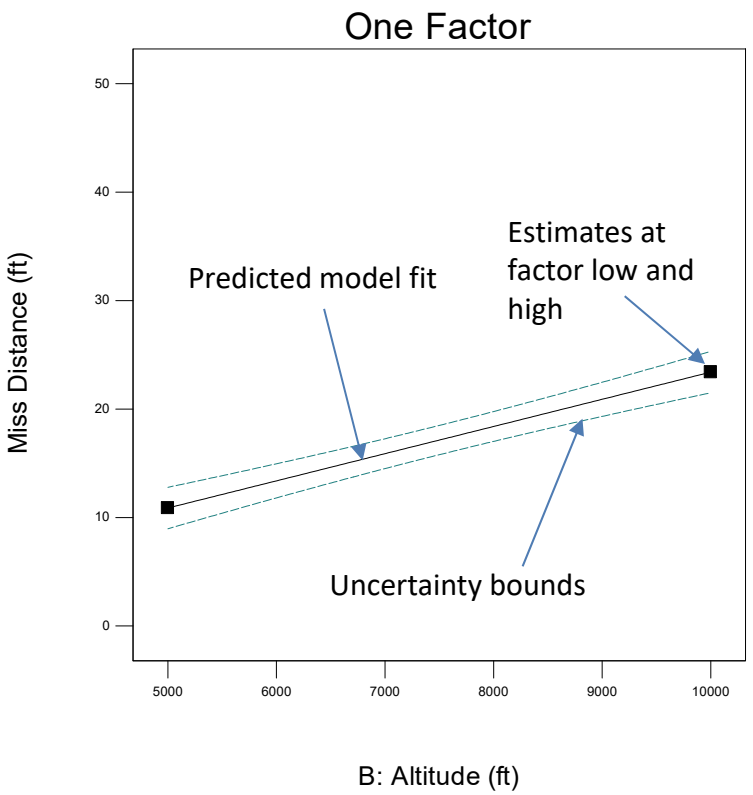
(Use as callout page)

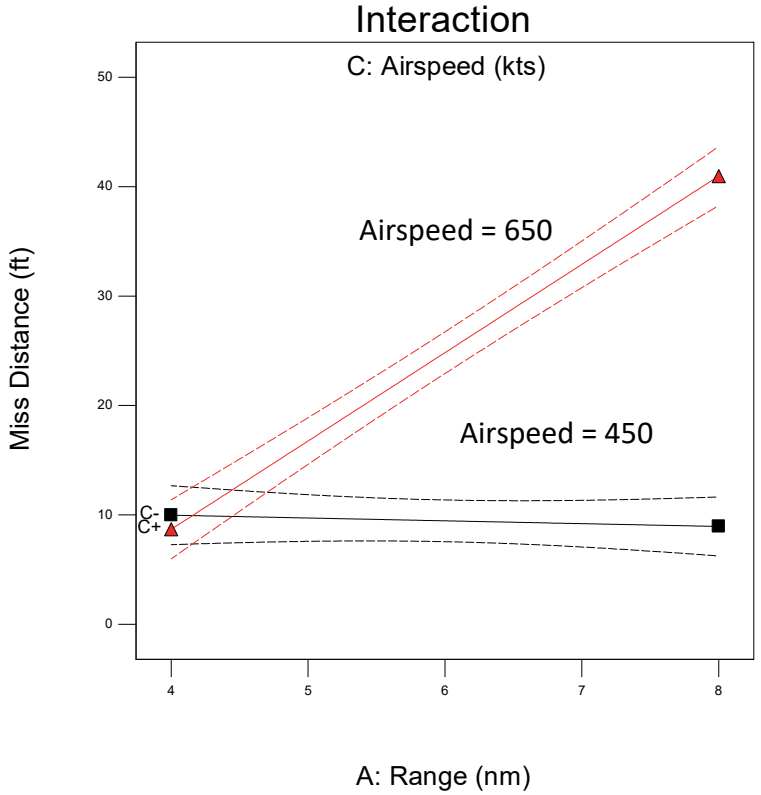
Test Designs

Design	Strengths	Weaknesses
Optimal Design	• The Optimal Design is the most flexible design alternative – it can handle any number of factors, with any numbers of levels, with a specified model to support and a certain number of runs • The analyst can also determine the number of levels for numeric factors to accommodate test planning and execution needs • The analyst can prescribe and build designs with disallowed factor combinations and numeric factor constraints	• Requires the most knowledge and experience to develop safe and robust optimal designs • Optimal Designs are sensitive to outliers and missing observations • One can easily build undesirable designs

Statistical Model Supported

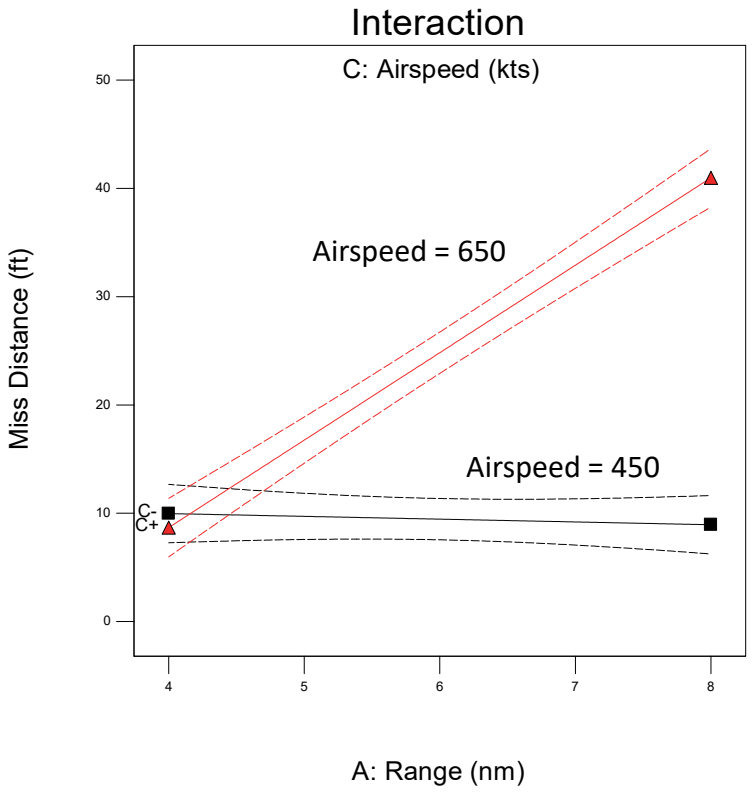
- The type of model supported by the design is the most important statistical consideration when assessing test adequacy
- Example (previously discussed): Miss distance for the gravity weapon
  - Three two-level factors: Range, Altitude, and Airspeed
- The objective is to characterize miss distance across the 3 factor test space. The chosen design is a 2-level full factorial with replicates at each of the 8 corners, resulting in a total of 16 runs, which allows for estimation of the main effects and two-way interactions in addition to providing excellent estimation of noise.
- The figures below depict the importance of estimating interaction effects. The one factor model (left) shows the main effect of altitude on miss distance and the two-factor interaction (right) shows the interaction between range and airspeed in explaining miss distance.





Statistical Model Supported – Interaction Explained

- The two-factor interaction is a very common effect type in military systems.
- Test designs should be capable of estimating two factor interactions. Points should be placed in the test space to allow for the estimation of these effects.
- Two-factor interactions reflect the reality in most systems that one factor (e.g. Range) can have a different effect on the response (e.g. miss distance) at different settings of another factor (e.g. Airspeed).



<u>Interaction Explained:</u>
When a weapon is released at Airspeed = 450 kts, Range has no effect on Miss Distance

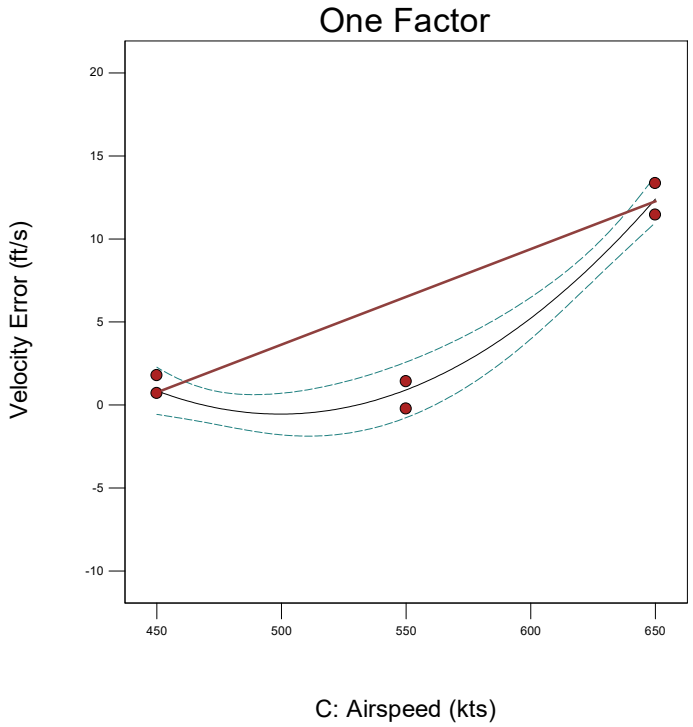
When a weapon is released at Airspeed = 650 kts, as Range increases, Miss Distance increases

The effect of Range on Miss Distance *depends on* the release Airspeed

Statistical Model Supported

- Let's look at another response variable: weapon velocity deviation from programmed velocity (called velocity error).
- It was anticipated that the relationship between factors and this response variable would show a curvilinear or nonlinear function, such that having only two levels for each factor is insufficient. A second order design and model are needed.
- The second order models include linear, interaction **and** quadratic effects.
- The second order design requires more levels per factor and more points in the test space.

The test team planned for second order model and discovered a curved relationship between airspeed and velocity error. The conclusion is that velocity error is nominal unless airspeed is > 600 kts



With only 2 levels, only a straight line would be possible. The linear fit would misrepresent the actual velocity error at airspeed > 450 and < 650 kts. Including another level allows us to estimate nonlinearity.

Power and Confidence

- Power and confidence are only meaningful in the context of a hypothesis test!
- Form two hypotheses to state two possible realities: one is H_0: the factor does not affect the response; the other is H_1: the factor does affect the response. For example:

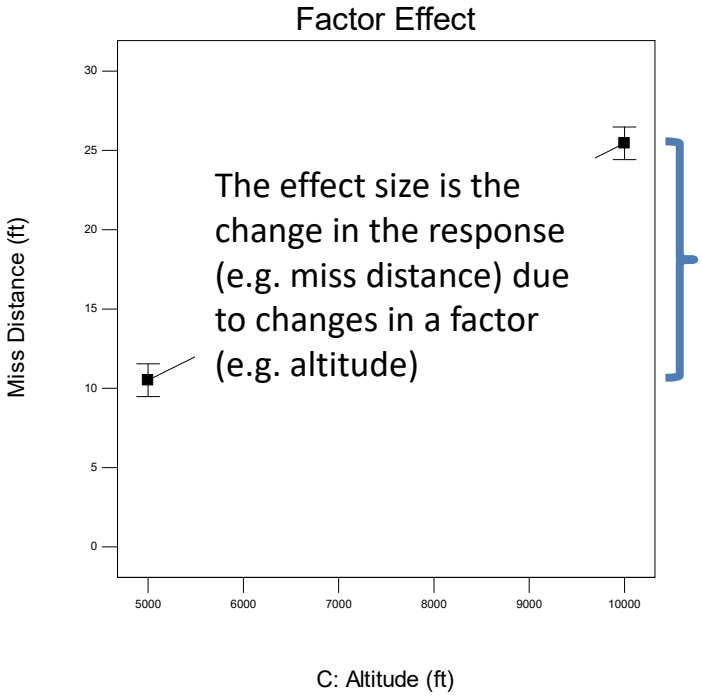
H_0: Weapon Miss Distance is the same for both levels of Airspeed
H_1: Weapon Miss Distance is different for each level of Airspeed

- We consider the probability of concluding a hypothesis based on test outcomes vs. the true state of the system as being either H_0 or H_1
- Power and Confidence are probabilities associated with the correct conclusions
- Power is the probability of correctly concluding that Miss Distance is affected by Airspeed, given that Airspeed truly affects Miss Distance.
- Confidence is the probability of correctly concluding that Airspeed has *no* effect on Miss Distance, if Airspeed truly *does not* influence Miss Distance

Power and confidence allow us to understand risks of incorrect conclusions

What Influences Power?

- For a given design, confidence ***is set by the tester*** and power is calculated for factor main effects and interactions.
- ***Power is the focus in building the test design.*** Power is affected by the number of runs, effect size, and system noise.



What Affects Power?

Parameter	Increasing Power
Number of Runs	More runs
Effect Size	Larger effect size
System Noise	Less noise

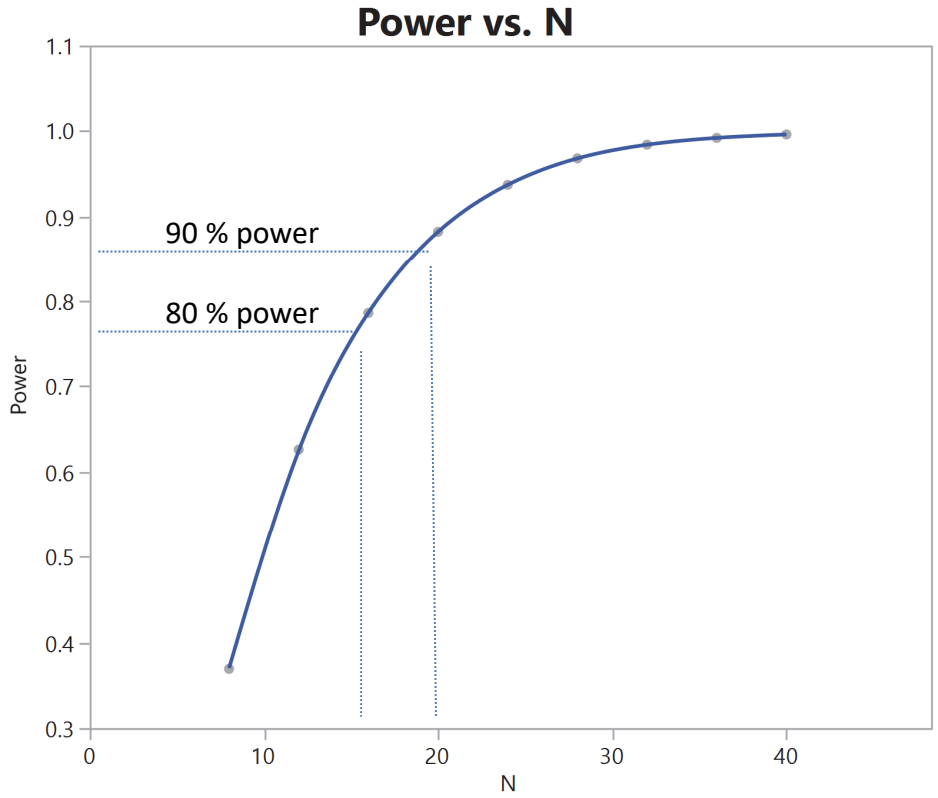
- The effect size is chosen by the tester and the system noise is estimated from historical data or pilot tests
- A signal-to-noise ratio is then calculated as an input to the power analysis

$$\text{Signal-to-noise (SNR)} = \frac{\text{Effect Size}}{\text{System Noise}}$$

- We typically size the test design with power > 90% to detect the specified effect size for the assumed level of noise (SNR).

Power versus Sample Size (N)

- Power increases rapidly initially as the number of test points (N) increases
- For this illustration, we use a notional signal to noise ratio (SNR) = 1.5
- Between 80% and 90% power, we see that only a small increase in our sample size will increase our power meaningfully. We want to select a sample size that achieves adequate power, but do not want to increase our sample size for very small increases in power.
  - For example, in the below plot, increasing our sample size from 35 to 40 would yield minimal increases in power and would waste test resources
- Test design sample sizes should provide > 80% power and we prefer > 90% power

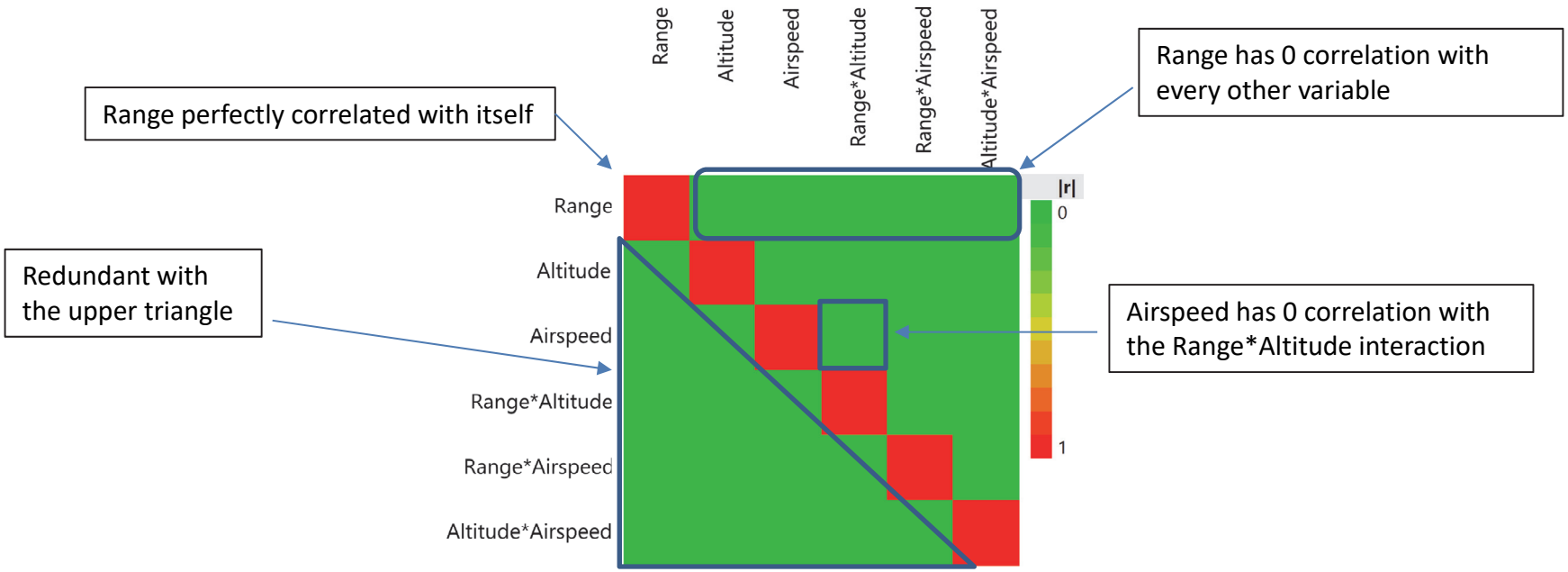


Correlation Coefficient

- To best determine which factor effects and interactions are important, the test design must have points placed in the test space such that the variables can be estimated independently.
- Correlation is a measure of the strength of the relationship between two variables. For example, the variables could be the model main effects and 2-factor interactions. The goal is to place test points in the space such that all main effects and 2-factor interactions are uncorrelated.
- The Correlation Coefficient (r) quantifies the degree of change of one variable based on the change of another variable. If two variables change independently, the r = 0. Perfect correlation gives r = 1.

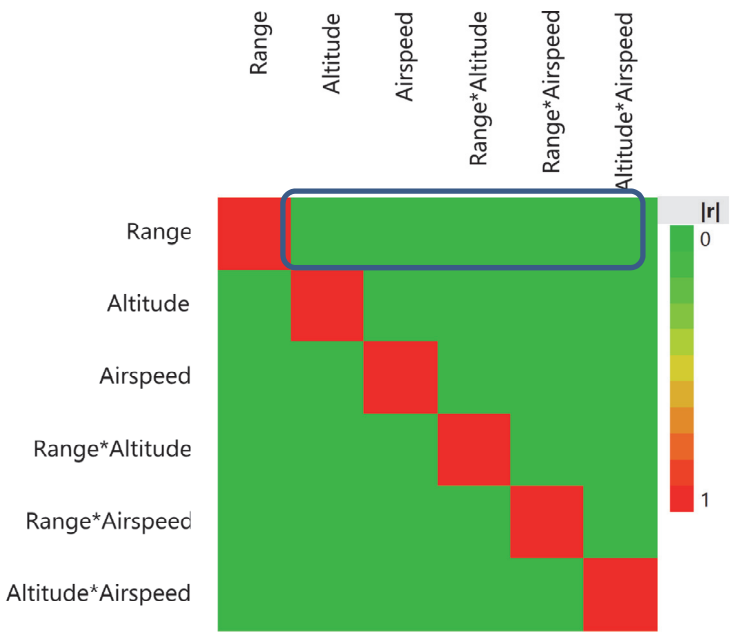
Correlation Coefficients - No Correlation

- One way to report Correlation Coefficients for a test design is via a color graph of all pairwise combinations of factor effects and their interactions.
- **Gravity Weapon Example**
  - Full Factorial 2-level Design
  - 3 Factors
  - 2^3 Design with 8 runs
- **In this design the main effects are all uncorrelated**
  - Green is perfectly uncorrelated (0%). Correlation Coefficient = 0
  - Red is perfectly correlated (100%). Correlation Coefficient = 1

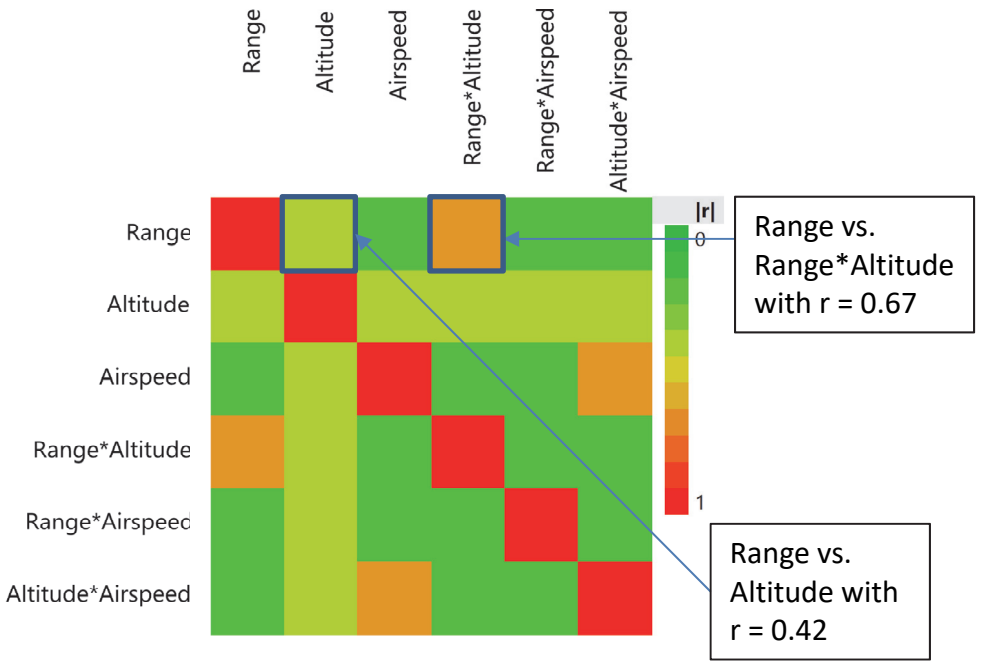


Evaluating Correlations Among Model Effects

2^3 Design (8 runs)



2^3 Design with 3 runs not executed (5 runs)



- Ideally, designed tests should support at least all linear effects and two way interactions.
  - When there are a large number of factors, it is often not possible to design tests to this standard. Therefore some correlation occurs.
  - Correlations > 0.5 are a bigger cause for concern.
- Correlation plots allow us to understand the tradeoffs in designs. For example, the 5-run design results in a 0.67 correlation between the Range main effect and the Range*Altitude interaction. Due to the high correlation, it will be more difficult during analysis to determine whether Range or Altitude truly affect miss distance.

Evaluation of the Test Design

1. Overall Design Approach
   - Is the test size proposed in the experimental design reasonable? Is it consistent with the resources section?
   - Are all the important factors included within the design?
2. Model Supported
   - Does the design supported at least a two-factor interaction model?
   - Are replicate points included to estimate noise?
   - Are quadratic effects (or at least center points) included for continuous factors?

Evaluation of the Test Design

3. Power and Confidence
   - Is confidence set at a reasonable level of risk?
     - Typically set confidence > 90%.
   - Is power calculated for each primary response variable?
   - Is there sufficient power for main effects?
     - Typically minimum power is 80% and the goal is power > 90%.
     - Power calculations should always be based on expected effect size and estimated noise variance (signal-to-noise ratio).
4. Correlations
   - Based on the model to be supported, the design correlations should be reported, considering all pairwise combinations of variables (main effects and 2-factor interactions)
   - The correlations are summarized by reporting the range of r values for all pairwise combinations of model effects. It is often possible to build designs with all $r = 0$. If correlations do exist, ensure the designs have $r < 0.50$.

Assessing the Adequacy of Test Designs: Statistical Measures of Merit

Statistical Measure of Merit	Experimental Design Utility	Usage
Statistical Model Supported (Model Resolution/Strength)	Describes the flexibility of the empirical modeling that is possible with the test design	Match to the test objective, and expected physical response of the system. (Second order is normally adequate for characterization.)
Power	Quantifies the likelihood in concluding a factor has an effect on the response variable when it really does.	Desire to maximize power
Confidence	Quantifies the likelihood in concluding a factor has no effect on the response variable when it really has no affect.	Desire to maximize confidence (confidence level is set by researcher)
Correlation Coefficients	Describes degree of linear relationship between individual factors.	Desire to minimize correlation between factors

Common Test Design Mistakes

- **Failure to link test objectives, responses, factors, levels, and resourcing**
  - All aspects of the planning (test objectives, responses, etc.) may be present but linkage is lacking. For instance, we may desire to characterize the relation between factors and responses, requiring a model that can support main effects and interactions. However, the chosen design can only estimate main effects.
  - **Solution**: Ensure responses answer the test objectives; the model supports the objective; and that the design spans the test space for the factors/levels and supports the desired model.

- **Failure to link analysis to the design**
  - If the design is constructed without recognizing that we want to adequately estimate the effects for **all** of the factors, as well as the 2-factor interactions, the power values for some of the factors or the interactions may be unacceptable due to small sample sizes.
  - **Solution**: Design tests with sample sizes that provide sufficient power for all the main effects and the 2-factor interactions. Report all power values.

Common Test Design Mistakes

- **Deciding not to include some factors in the test design, thinking the number of runs will be more reasonable/manageable**
  - It is true that adding a factor to a full factorial design (e.g. 4 to 5 factors) at least doubles the number of runs (e.g. 16 to 32) , but many other designs do not cause such a dramatic increase. Other designs will not increase the number of runs, or will only increase the number of runs slightly.
  - We know that not all factors and interactions will be important, but we don't know which ones ahead of time. For example, a design with 6 factors will result in 22 model variables (1 overall mean, 6 main effects and 15 2-factor interactions). Upon analysis, typically the number of model variables that have large effects is typically 10 or less.
  - Fractional factorial designs gain their efficiency by being able to estimate well the subset of important effects.
  - **Solution**: Fractional factorials and optimal designs can support many (e.g. 8) factors in a manageable number (e.g. 20) of runs.

Common Test Design Mistakes

- **Building a one-shot test design without considering a sequential approach.**
  - Consider 6 factors, where 3 factors have 2 levels and 3 factors have 4 levels each. An adequately powered (>90%) fractional factorial design supporting a model with main effects and 2-factor interactions requires 80 runs. It is highly unlikely that all 80 runs can be executed in one test period without unwanted consequences (e.g., operator fatigue).
  - **Solution**: Partition the design into subsets, where each subset stands alone as a test design capable of estimating some or most of the factor main effects and 2-factor interactions. Analyze the data after each subset and use that knowledge to refine the design in subsequent phases.

Motivating Example: Test Plan for Mine Susceptibility

- Goal:
  - Develop an adequate test to characterize the susceptibility of a cargo ship against a variety of mine types using the Mine Simulation System A.
- Responses:
  - Magnetic Signature, Acoustic Signature, Pressure
  - Slant Range at simulated detonation
- Factors:
  - Speed, Range, Degaussing System Status, Water Depth, Ship Direction
- Uncontrollable factors:
  - Sea State

Cargo Ship Mine Susceptibility Testing Example

Cargo Ship Mine Susceptibility Testing Example Plan Phase (Steps 1 – 3)

- Step 1 - Define the test objective
  - Develop an adequate test to assess the susceptibility of a cargo ship against a variety of mine types using the Mine Simulation System A.
- Step 2 - Select appropriate response variables
  - Magnetic Signature, Acoustic Signature, Pressure
  - Slant Range at simulated detonation
- Step 3 – Choose factors, levels, desired model
  - Factors (5):
  - Speed, Range, Degaussing System Status, Water Depth, Ship Direction
  - Levels (2 or 3): All numeric, except 1 categorical: Degaussing = Yes/No
    - Numeric values, generically:
      - Low and High, or
      - Low, Medium, and High
  - Uncontrolled factors:
    - Sea State
  - Desired model: Main effect plus 2-factor interaction required, and pure quadratic if the number of runs are feasible.

Cargo Ship Mine Susceptibility Testing Example Design Alternatives (Step 4)

- Step 4. Choose an experimental design
  - The table below compares design alternatives using the number of runs, power values, and correlation as criteria. Each design uses the same confidence of 95%, which is set by the test designer. Power is reported for a nominal SNR=2 and show the minimum and maximum power values for each of the variables in the model supported. Also listed are the ranges of pairwise correlations among model variables.

Model Supported	Design	Factors	Levels	Test Runs	Min Power	Max Power	Confidence	Correlation
Linear + Interaction	Full Factorial	5	2	32	99.9%	99.9%	95%	0
	General Factorial	5	2	32	99.9%	99.9%	95%	0
	Fractional Factorial	5	2	16	95.8%	95.8%	95%	0
Second order	Response Surface	5	3	29	93.7%	99.9%	95%	0 - 0.14
	Optimal	5	3	29	36.0%	96.9%	95%	0 - 0.52

- Interpreting the Table and Comparing Designs
  - For this example, each design accommodates all of the factors, and each factor has at least 2 levels. Likewise, confidence is set at 95% for each design.
  - Designs are usually compared by the model supported. The second order designs add points in the test space in order to estimate the quadratic effects, so 3 levels for each factor are required.
  - **Red** is less desirable, while **green** is desirable.

Cargo Ship Mine Susceptibility Testing Example Design Selection (Step 4)

Model Supported	Design	Factors	Levels	Test Runs	Min Power	Max Power	Confidence	Correlation
Linear + Interaction	Full Factorial	5	2	32	99.9%	99.9%	95%	0
	General Factorial	5	2	32	99.9%	99.9%	95%	0
	Fractional Factorial	5	2	16	95.8%	95.8%	95%	0
Second order	Response Surface	5	3	29	93.7%	99.9%	95%	0 - 0.14
	Optimal	5	3	29	36.0%	96.9%	95%	0 - 0.52

- Linear + Interaction Model Design Preference
  - The fractional factorial has half the number of runs as the full factorial and similar statistical measures of merit. The key advantage of the full factorial is that it can estimate higher order interactions. However, these model effects often are not large.
- Second Order Model Design Preference
  - Both designs have the same number of runs, so they are equally efficient. The response surface design has better power, especially for the quadratic effects. The response surface design has a low range (0 - 0.14) for pairwise correlations , whereas the optimal design correlation range (0-0.52) is of borderline concern.
- Design Selected
  - For this example, with our prior knowledge on the susceptibility of a cargo ship against mines, we would typically focus on the linear + interaction model. The **fractional factorial design** would be sufficient and requires 16 runs.

Summary

- This module introduced DOE, classic experimental designs, and their uses.
- You should now be able to:
  - Know the definition of DOE
  - Understand the phases of DOE
  - Understand the 4 steps in the Plan and Design phases
  - Understand design options (classic experimental designs, optimal designs)
  - Understand Statistical Measures of Merit (statistical model supported, power, confidence, correlation coefficients)
  - Understand design evaluation
  - Understand common design mistakes

Lesson 3
Observational Study


= hyper link to a call out

Lesson Objectives

- The observational study lesson introduces the role of observational studies in test and evaluation.
- Upon completion of this lesson, you will be able to:
  - Define Observational Study
  - Distinguish between DOE and an observational study
  - Understand why observational studies are used for data collection
  - Identify the various types of observational studies

Observational Study

- In Lesson 2, we learned about DOE where testers vary factors to determine the effects on a response variable.
- In this lesson, we will learn about observational study, another STAT technique to collect data.
- An observational study is a technique where the testers collect data on a subject(s) without affecting the subject(s). The key word is “Observe.”
- In an observational study, the testers:
  - Do not or cannot control/vary the factors (independent variables) during the event, or
  - Do not or cannot affect the subjects’ environment
- Observational studies can occur during large-scale training exercises and theater observations.

Observational Study

- Like DOE, observational studies also need a purposeful design.
- The testers still need to know:
  - What are the response variables?
  - What are the factors?
  - What are the levels?
  - What data need to be collected?
- The difference is that in the observational study the testers do not have control of the event. As a result, the following challenges occur:
  - Confounding data
  - Loss of ability to determine cause and effect
- Testers need to understand the pedigree of data collected from observational studies and the associated risk during analysis.

DOE and Observational Study – An Example

Objective: Measure the average time for the air traffic controllers to identify aircraft entering their controlled airspace.

Options to collect the data: designed experiment or observational study

In the DOE:

- Testers randomly assigned air traffic controllers to one of two groups:
  - One group was directed to drink beverages containing caffeine
  - One group was directed not to drink beverages containing caffeine
- Testers measured the average time for the air traffic controllers to identify aircraft entering their controlled airspace.
- This is **DOE** because the factor (caffeine) was controlled. One group was given the caffeine and the other group was not.

In the observational study:

- Testers took a random sample of air traffic controllers and collected data on their normal caffeine consumption.
  - Each controller was classified as light, moderate, or heavy caffeine user.
- Testers measured the average time for the air traffic controllers to identify aircraft entering their controlled airspace.
- This is an **Observational study** because the factor (caffeine) was not controlled. The air traffic controllers were not told how much caffeine to consume.

DOE and Observational Study

- The biggest difference between an observational study and DOE is the issue of association versus causation
  - **Association** - a relationship between two random variables which makes them statistically dependent. It refers to a general relationship.
  - **Causation** - change in one variable directly caused change in the other variable.
- Because factors are controlled in a designed experiment, testers can sometimes make conclusions of causation
- Because factors are not controlled in an observational study, the results can only be interpreted as associations

Why conduct observational studies?

- Testers conduct observational studies when:
  - They lack the ability to control a test/event.
  - Imposing conditions would violate safety or ethical standards, or potentially cause the system operators to execute a mission in a non-operationally realistic manner.
  - They want to verify correction of deficiencies/ corrective actions.
  - They want to collect data by leveraging availability of on-going events.

Types of Observational Studies

- There are three types of observational studies. Due to limited time to collect data during testing, testers most often conduct cross-sectional observational studies.
  - Case-Control Study: retrospective study where the testers look back in time and compares two or more populations in order to describe the association between the variables under study
  - Cohort Study: a prospective study where the testers select a population to study for a determined length of time (usually a long period of time, measured in years) to determine if a particular characteristic affects a variables under study.
  - Cross-Sectional Study: Testers study a variable or several variables for shorter, definite period of time to determine associations between certain characteristics and the variables

Challenges with Observational Studies

- The following are challenges that must be considered when conducting observational studies:
  - Problem to be studied must be clearly described before starting data collection
  - Formulate specific objectives of the study/observation
  - Decide on information to collect and corresponding data collection techniques
  - Consider sampling across the population or operational envelope
  - Consider possible confounding or lurking variables in data collection plan

Definitions of Lurking Variables and Confounding Variables

(Use as callout page)

- **Lurking variables:** variables not considered in a study that could influence the relationship between the variables in a study. For example, a study looks at the effect of diet and exercise on a person's blood pressure. Lurking variables could be a person's smoking status and stress level, as both variables likely relate to blood pressure.
- **Confounding variables**: variables that are considered in a study that could influence the relationship between the variables in the study. For example, suppose one group of navigators was given a lecture-based course on a new radar and another group was given hands on training. If the navigators are all tested on radar proficiency, we would be unable to determine whether the method of teaching or the instructor effectiveness influenced proficiency.

Observational Study Example
Aircraft Carrier Flight Deck Operations

- Situation:
  - Testing supported an assessment of changes to an aircraft carrier flight decks that are designed to increase the number of aircraft sorties per day from 120 to 135 in sustained operations
  - Test team needed to collect data from flight deck operations over which they exerted limited control on specific factors and conditions during the event
  - Data collection period is 6 days of sustained operations, at 12-hours per day of flight operations
  - Primary response variable for measuring success was Sortie Generation Rate (SGR) which measures the number of aircraft sorties launched in one day
- Selected STAT method:
  - Test team employed a cross-sectional observational study to collect and analyze data from flight deck operations because the test team (1) wanted to collect real-time data based on an operationally realistic aircraft carrier flight deck operation, and (2) had limited control of the factors

Observational Study Example
Aircraft Carrier Flight Deck Operations

- Question: In this Aircraft Carrier Flight Deck Operations example, controlling all factors to execute a designed experiment to evaluate SGR would provide the most operationally relevant data. **TRUE or FALSE** ?
- Response: False
- Rationale: During flight operations, a flight deck crew makes numerous real-time decisions that affect SGR.
  - Aircraft carriers have several catapults for launching aircraft.
  - A typical launch cycle involves multiple aircraft are launched from multiple and spare aircraft are available.
  - If a problem occurs during a launch cycle, such as a catapult or an aircraft breaking, the flight deck crew must consider multiple factors before making a decision on how to proceed.
  - Based on available information, the flight deck crew may wait until the repair is completed, move the aircraft to a different catapult, use a spare aircraft, cancel the launch, or select another option.
  - Artificially constraining the flight deck crew's options would not be realistic.
  - The observational study does not control these factors. Instead, the study design is based on collecting data from flight deck activities as an operationally realistic operation unfolds.

Summary

- This module introduced the role of observational studies in T&E.
- You should now be able to:
  - Define Observational Study
  - Distinguish between DOE and an observational study
  - Understand why observational studies are used for data collection
  - Understand when an observational study would be better than a DOE
  - Identify the various types of observational studies

Lesson 4
Survey Design and Analysis


= hyper link to a call out

Lesson Objectives

This module introduces surveys, survey types and their uses.

Upon completion of this lesson, you will be able to:

- Understand the purpose of a survey and how it can be used in T&E.
- Understand the parts of a survey, structured and unstructured questions, and survey quality.
- Understand empirically-vetted and custom-made surveys, and their use.
- Understand that surveys are part of a detailed test plan.

Definition & Purpose

- Surveys are instruments that measure people's thoughts and feelings
  - *For example*:
    - Usability of interfaces
    - Workload of tasks

- Surveys should be designed to address specific test questions

- Primary purpose of surveys is to collect quantitative data
  - Qualitative data collection is also possible

- Surveys should be developed according to established scientific principles, designed and included as part of the developmental and operational detailed test plans

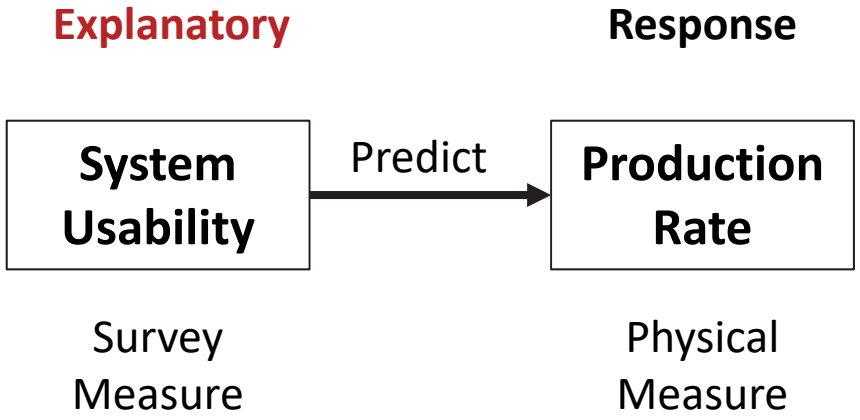
Surveys

Surveys can serve as explanatory or response variables.
Explanatory variable: a factor that explains changes in the response variable
Response variable: the outcome of interest
Example:

Survey as Explanatory Variable

Question:

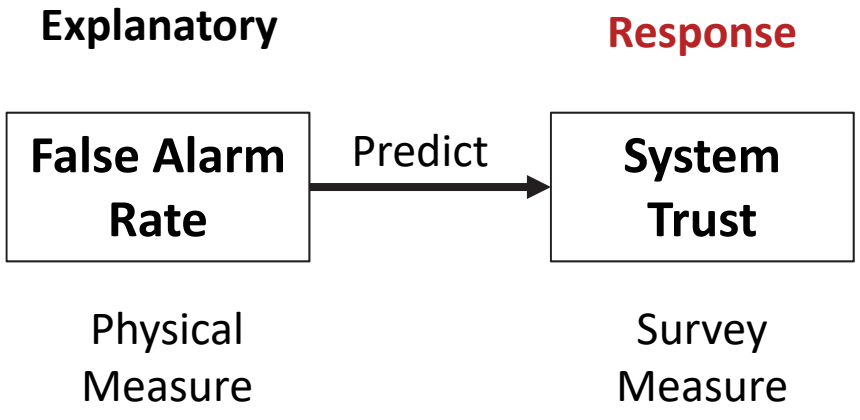
Does system usability affect image production rates?



Survey as Response Variable

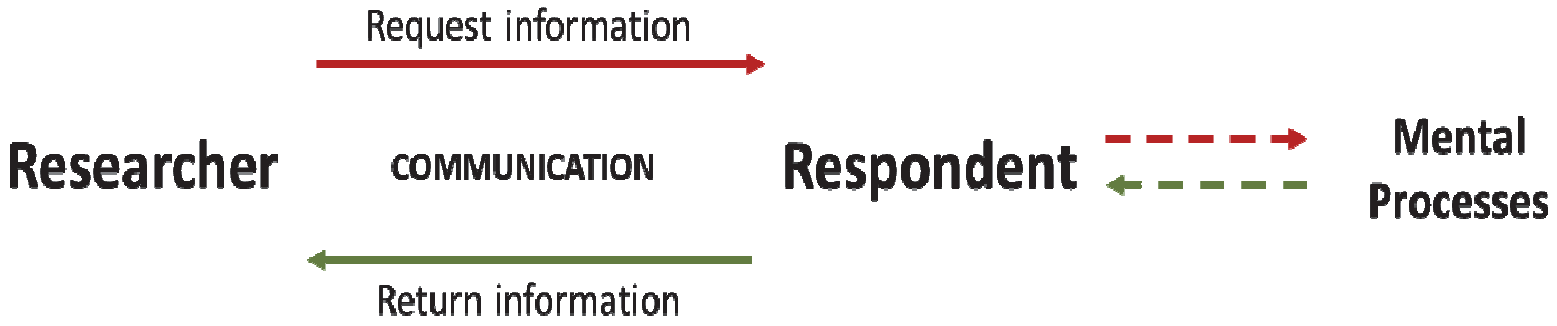
Question:

Does the false alarm rate affect users' trust in the system?



Surveys are Conversations

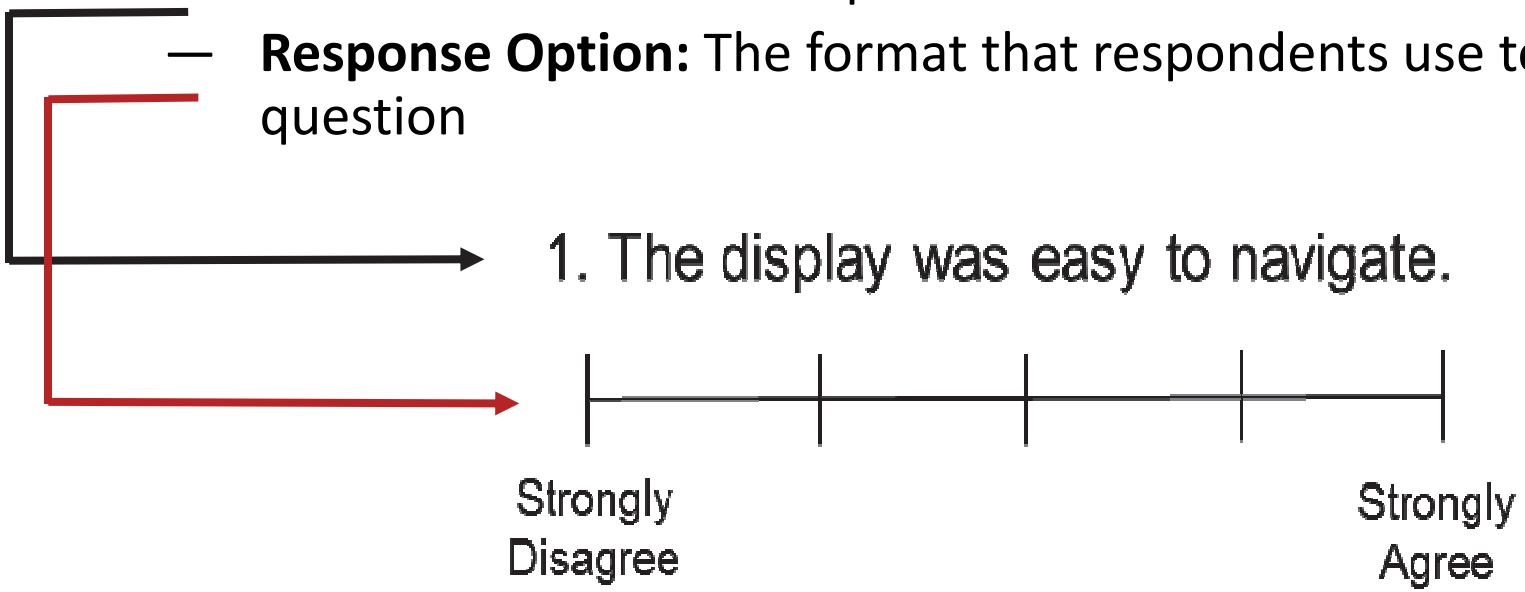
- Surveys are conversations between the tester and the respondent
  - Testers request information from respondents
  - Respondents answer the tester request through the use of the survey



- To answer each question the respondent must:
  1. Understand the question
  2. Recall relevant information
  3. Form a judgment
  4. Respond using the format provided
- Question construction and survey structure shapes these mental processes and ultimately, impacts data quality

The Survey

- Surveys are comprised of a series of questions.
- Questions are comprised of 2 parts:
  - **Item:** The words that respondents address
  - **Response Option:** The format that respondents use to answer the question

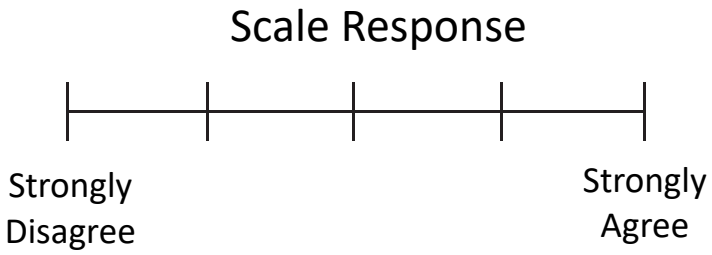


- Item and response option formats differ depending upon research goals.

Question Types

Structured Questions

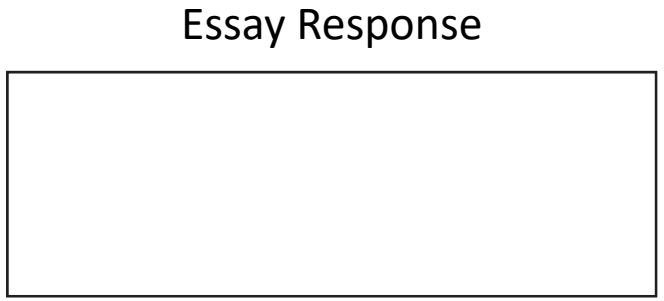
- The researcher constrains how respondents answer the question
- *For example:*



- Quantitative analysis
- Most of your survey questions
- Useful for **measuring** subjective experiences

Unstructured Questions

- Respondents answers are not constrained by the tester
- *For example:*



- Qualitative analysis
- Most appropriate for interviews, but it is okay to use sparingly in surveys
- Useful for understanding unanticipated events and discovering problems

(Use as callout page)

Structured Questions

- Structured questions can take many forms.

Dichotomous

☒ No
☐ Yes

Response-option format that forces respondents to select one of two options.

Multiple Choice

☐ Active Duty
☐ National Guard
☐ Civilian
☐ Contractor

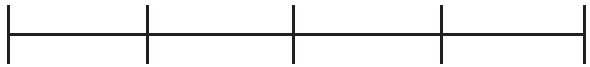
Response-option format that forces respondents to select among several options.

Rank

1 ___ Killer Robots
2 ___ Aliens
4 ___ Zombies
3 ___ Vampires

Response-option format that forces respondents to rank order their preferences.

Scale Response



Strongly Disagree — Strongly Agree

Survey Quality

Your survey is reliable if it produces similar scores under similar conditions.

- *For instance*, if you ask the same respondents to complete Task A and fill out the workload survey each Monday for a month. A reliable survey should produce similar workload scores each week.
- If the survey produces very different workload scores each week, the scale is unreliable.
- Critically, a scale that is unreliable is, by definition, not valid.

Your survey is valid if it measures what it is designed to measure.

- *For instance*, if you can demonstrate a correlation between the workload survey you designed and other measures of workload.

(Use as callout page)

Survey Quality

Types of Reliability

- **Test-Retest Reliability:** Consistency of a respondent's results from across time
  - Respondents who complete the survey multiple times, under similar conditions, should produce similar scores.
- **Inter-Rater Reliability:** Consistency of results among similar respondents
  - Similar respondents who complete the survey under similar conditions should produce similar scores.
- **Internal Reliability:** Consistency of results across items within a test
  - Questions that are designed to measure the same concept – for instance, workload – should be highly correlated with each other.

Types of Validity

- **Predictive Validity:** Degree to which the survey can predict a measure it should theoretically be able to predict
  - A survey that measures trust in automation should, for instance, predict the extent to which operators rely on automated systems.
- **Convergent Validity:** Degree to which the results are correlated with measures of theoretically related concepts
  - A survey that measures workload, for instance, should be correlated with theoretically similar measures like stress and fatigue.
- **Discriminant Validity:** Degree to which results are uncorrelated with measures of theoretically dissimilar concepts
  - A survey that measures trust in automation, for instance, should be uncorrelated with theoretically dissimilar concepts such as trait aggression.

Survey Types

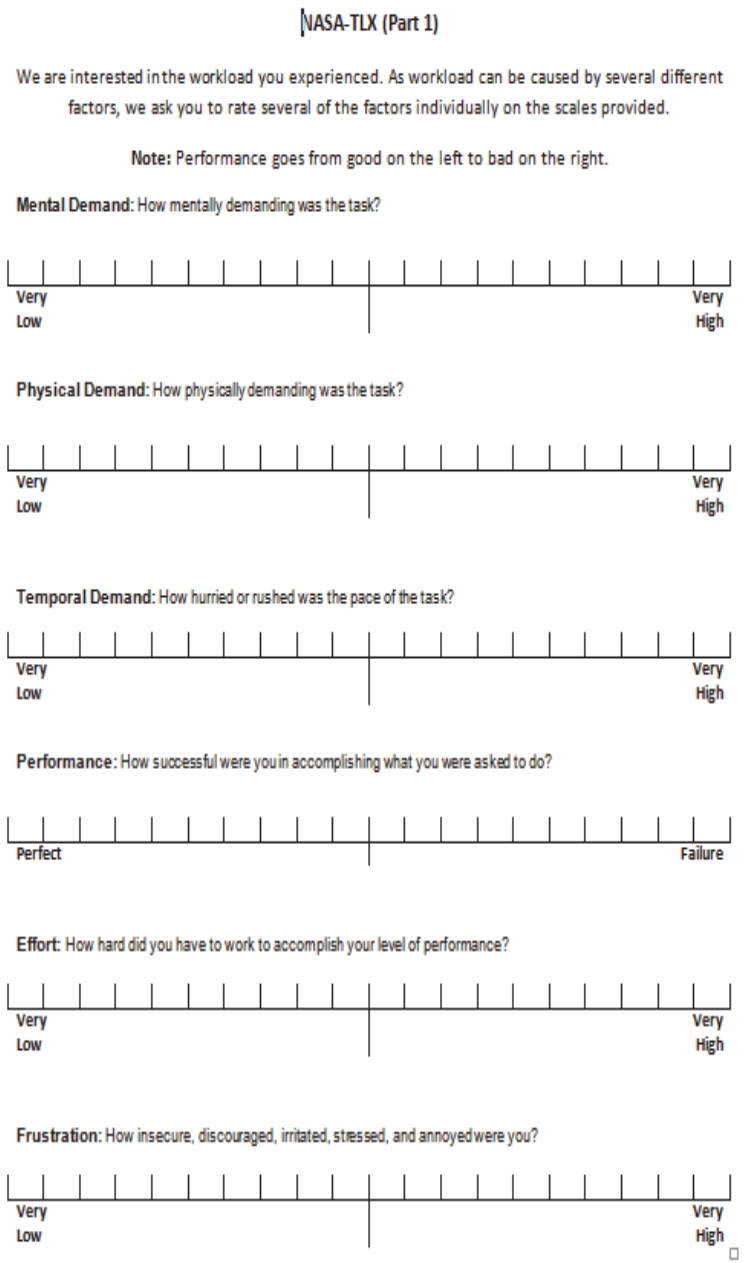
- Surveys that undergo reliability and validity testing differ from surveys that don't
  - Reliability and validity testing ensures that surveys consistently measure the concepts they are intended to measure

- **Empirically-Vetted Surveys** have undergone reliability and validity testing
  - Established levels of reliability and validity
  - Effect sizes and variances are available to aid in power analyses
    - Effect sizes are statistics that quantify the difference in scores observed between groups.
    - Variances are statistics that quantify how widely scores in a group vary.
  - Average scores can be used as standards for comparison
  - Empirically-vetted surveys are available for measuring concepts such as workload, usability, trust in automation, fatigue, and self-efficacy.
- **Custom-Made Surveys** have not undergone reliability and validity testing
  - Level of reliability and validity is unknown
  - Increase the likelihood that surveys are reliable and valid by using established survey design principles
  - Necessary to use when empirically-vetted surveys are not available

(Use as callout page)

Empirically-Vetted Survey for Workload NASA-TLX

- NASA-TLX has 2 parts:
  - **Part 1:** respondents rate their workload on a set of balanced, bipolar scales
  - **Part 2:** respondents make pairwise comparisons across the 6 dimensions to weight ratings obtained in Part 1
- We recommend using only Part 1
  - Part 2 does not increase the reliability or validity of results above and beyond Part 1
- 21 tick marks divide the 0-100 scale into increments of 5
- Simply average the ratings of each subscale to obtain the workload score



NASA-TLX (Part 1)

We are interested in the workload you experienced. As workload can be caused by several different factors, we ask you to rate several of the factors individually on the scales provided.

Note: Performance goes from good on the left to bad on the right.

Mental Demand: How mentally demanding was the task?

Very Low — Very High

Physical Demand: How physically demanding was the task?

Very Low — Very High

Temporal Demand: How hurried or rushed was the pace of the task?

Very Low — Very High

Performance: How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort: How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

Very Low — Very High

(Use as callout page)

Empirically-Vetted Survey for Workload

Crew Status Survey

- Single-item measure of workload
- Respondents rate their workload during the task on a 7 point scale
- Less sensitive than NASA-TLX
  - Unable to examine which aspect of workload drives scores
- Ideal for tasks where time is scarce and too much interference could be dangerous
- Also, useful to administer at regular intervals throughout task to measure changes in workload across time

Directions: Please circle the statement that best describes your workload during this task?

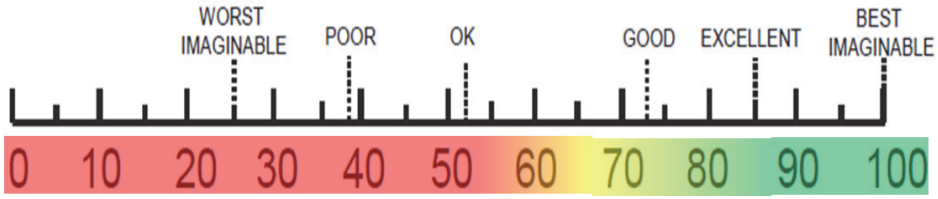
1. Nothing to do; No system demands.
2. Light Activity; minimal demands.
3. Moderate activity; easily managed considerable spare time.
4. Busy; Challenging but manageable; Adequate time available.
5. Very busy; Demanding to manage; Barely enough time.
6. Extremely Busy; Very difficult; Non-essential tasks postponed.
7. Overloaded; System unmanageable; Essential tasks undone; Unsafe.

(Use as callout page)

Empirically-Vetted Survey for Usability

System Usability Scale

- The System Usability Scale is a 10-item measure of usability
- Respondents rate the extent to which they agree with each item using a 5-point scale
- The scoring procedure is somewhat complex
    1. Subtract 1 from odd ratings for each respondent
    2. Subtract even ratings from 5 for each respondent
    3. Add up converted scores for each respondent
    4. Multiply each respondent's score by 2.5 to place score on a 0-100 scale
- Established threshold for evaluating usability



	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5. I found the various functions in this system were well integrated	1	2	3	4	5
6. I thought there was too much inconsistency in this system	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8. I found the system very awkward to use	1	2	3	4	5
9. I felt very confident using the system	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

Survey Design Basics

- Survey design impacts the quality of data collected from respondents
- Survey design includes:
  - The structure of the survey
  - How questions (item and response) are written
- Applying established survey design principles increases the likelihood that respondents will…
  - Understand the question you asked
  - Recall the event and form a judgment using relevant information
  - Understand how to answer the question
  - Successfully navigate from one question to the next
- Ensuring that respondents can easily complete these steps reduces measurement error and increases respondent motivation

(Use as callout page)

Survey Structure

- Define your survey structure before beginning to generate specific questions
- Properly structured surveys are based on well-formulated test questions
- Begin structuring your survey by:
  1. Writing down each test question
  2. Ordering the test questions logically
  3. Defining the type of information you need to answer each test question

For example:

Test question: Does the X application improve operators' workflow?

Information: Measure productivity and collect feedback from operators for each of the following concepts:

1. Usability of the application
2. Workload of routine tasks
3. Usefulness of information provided
4. Visual design
5. Comparison to previous workflow

(Use as callout page)

Survey Structure

- Now, you can create several focused questions for each concept you've defined
  - Each question should **directly** tie into one of the concepts you defined
  - This approach ensures that the survey is logically ordered and that each question directly supports the larger test design
- It is critical that you order each question logically within each concept
  - Well-ordered question lists coax respondents to be more honest and decrease the likelihood that they will skip questions
- Within each concept order questions from broad to specific, leaving difficult open-ended questions until the end.

(Use as callout page)

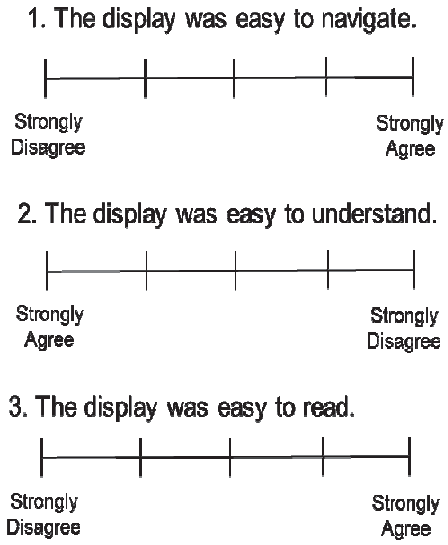
Survey Structure Tips

- Always include an introduction
  - Well-written introductions motivate respondents to produce thoughtful, honest answers
  - Introductions should:
    1. Thank respondents for participating and emphasizes the importance of their input
    2. Explain why you are conducting the survey and what you will do with the data
    3. Indicate how much time the survey should take to complete
    4. Assurance of confidentiality
- Place instructions where they are needed not simply at the beginning
  - For instance, placing a short set of directions between each section of the survey alerts respondents to the fact that the survey is changing categories and identifies the purpose of the questions that follow. Such transparency builds trust with the respondent and places them in the right mindset for answering the next set of questions

(Use as callout page)

Survey Structure Tips

- Place questions with the same response option in a matrix format and be consistent in the direction that scales are displayed to make the survey easier to navigate



VS.

	Strongly Disagree				Strongly Agree
1. The display was easy to navigate.	1	2	3	4	5
2. The display was easy to understand.	1	2	3	4	5
3. The display was easy to read.	1	2	3	4	5

- Only emphasize words that introduce important, but **easy to miss** changes in item wording or instructions.

(Use as callout page)

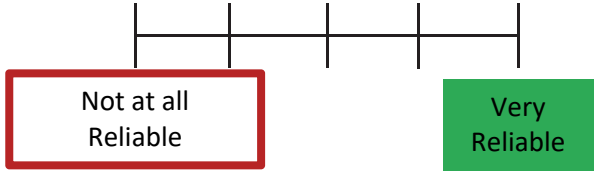
Survey Structure Tips

- Provide a consistent format to encourage respondents to read all the words on the page.

★ 1. How dependable is the system?

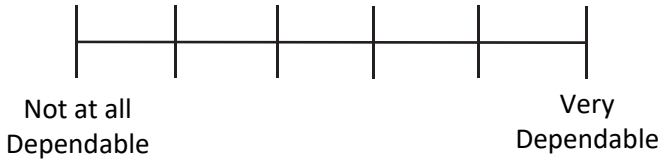
- ☐ Not dependable
- ☐ Sort of dependable
- ☐ Dependable

2. How **RELIABLE** is the system?

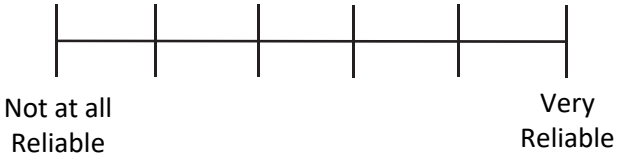


VS.

1. How **dependable** is the system?



2. How **reliable** is the system?



(Use as callout page)

Item Writing

- Only ask one question at a time
  - This will reduce uncertainty among respondents and assure that you can interpret answers

    Bad question: The interface was visually pleasing and easy to use.

    Good question: The interface was easy to use.

- Questions should be clear, concise, and grammatically simple

  Bad question: If the satellite stops working, would you consider or not consider using X system to improve your ability to communicate with others?

  Good question: If the satellite stops working, would you use system X to communicate?

- Questions shouldn't lead respondents to a specific answer

  Bad question: Shouldn't experienced operators be able to identify the target easily?

  Good question: How easily did you identify the target?

(Use as callout page)

Item Writing

- Avoid using “loaded” or emotive language

  Bad question: Do you believe the vehicle will protect soldiers from being maimed?

  Good question: How helpful was the vehicle in accomplishing task X?

- Questions should avoid absolutes and extremes

  Bad question: Do you always create unique passwords for each system?

  Good question: Do you create unique passwords for each system?

- Avoid asking respondents to do unnecessary computations

  Bad question: How many hours have you spent operating the system?

  Good question: How many months have you spent operating the system?

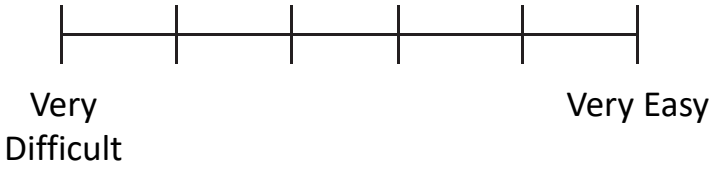
(Use as callout page)

Writing Response-Options

- The response option should clearly match the item

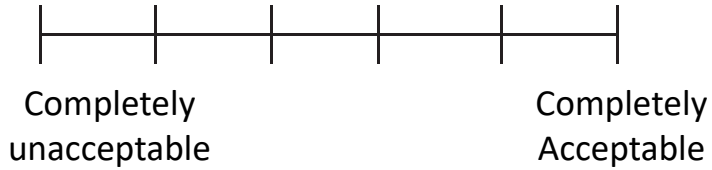
Good Question

How easy was the interface to use?



Bad Question

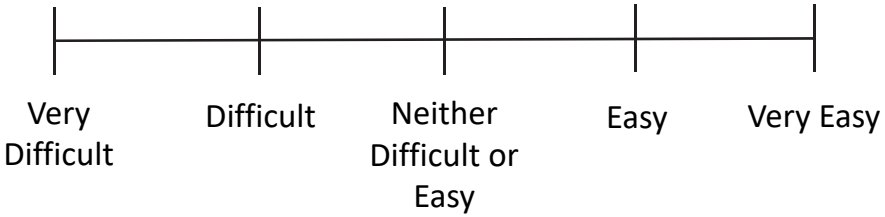
How easy was the interface to use?



- Use balanced, bipolar scales

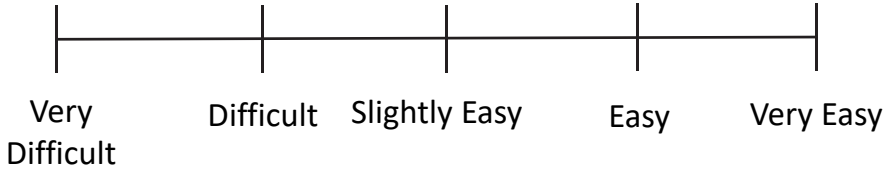
Good Question

How easy was the interface to use?



Bad Question

How easy was the interface to use?



(Use as callout page)

Writing Response-Options

- Eliminate check-all-that apply questions.
  - The likelihood that respondents will check a response option differs by position in the list.

Bad Question

Identify areas where you encountered a problem. (Check all that apply)

- ☐ Option A
- ☐ Option B
- ☐ Option C
- ☐ Option D

Designing Surveys into Tests

- Developing a quality survey does not guarantee that the data you collect will aid in answering your research question. Surveys must be administered systematically to be useful

- Research on humans differs from research on physical systems
  - Respondents are not interchangeable due to differences such as:
    - Past experience
    - Expertise
    - Motivation
    - Cognitive and physical capacities
  - Respondents are sensitive to social and contextual factors – for example:
    - Respondents may try to conform to testers' expectations
    - Respondents answer differently depending upon whether they are alone or in a group
    - Respondents answer differently depending upon temperature, mood, and fatigue

Designing Surveys into Tests

- The most common methods include:
  - Correlational Designs
  - Between-Subjects Designs
  - Within-Subjects Designs
  - Mixed Designs (combination of between- and within-subjects designs)

- Each method has strengths and weaknesses and dictates the types of analyses that are possible with the survey data you collected
- The method that is most appropriate depends upon your test question
- The method you select should be integrated into the larger test design

(Use as callout page)

Correlational Designs

- Correlational designs measure the extent to which two variables are related
- Correlational designs should be used when you want to understand if:
  - Two sets of respondent ratings are related

    For example: Are usability and workload related?
  - Demographic information is related to respondent ratings

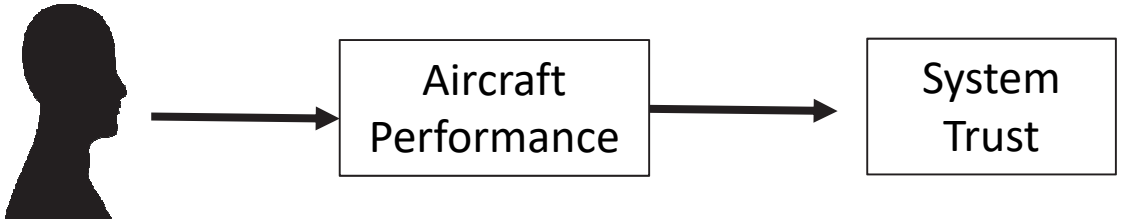
    For example: Are years of experience and usability scores related?
  - Respondent ratings are related to performance

    For example: Is system usability related to productivity?
- Respondents need scores on both measures
- Scores should be obtained under the same test conditions for all respondents
  - The goal of correlational designs is **not** to compare across conditions in the test
  - Cause and effect cannot be established

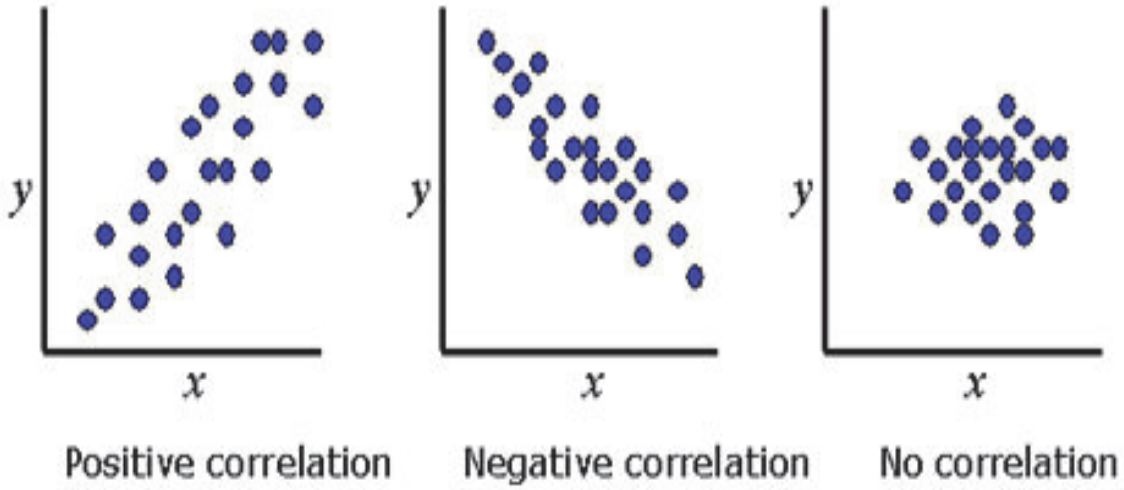
(Use as callout page)

Correlational Design

Example:



Analyze data from correlational designs with correlation or regression equations



(Use as callout page)

Between-Subjects Designs

- Between-subjects designs compare data from two or more separate groups
- Between-subjects designs should be used when you want to understand if:
  - Ratings or performance differ by demographic factors

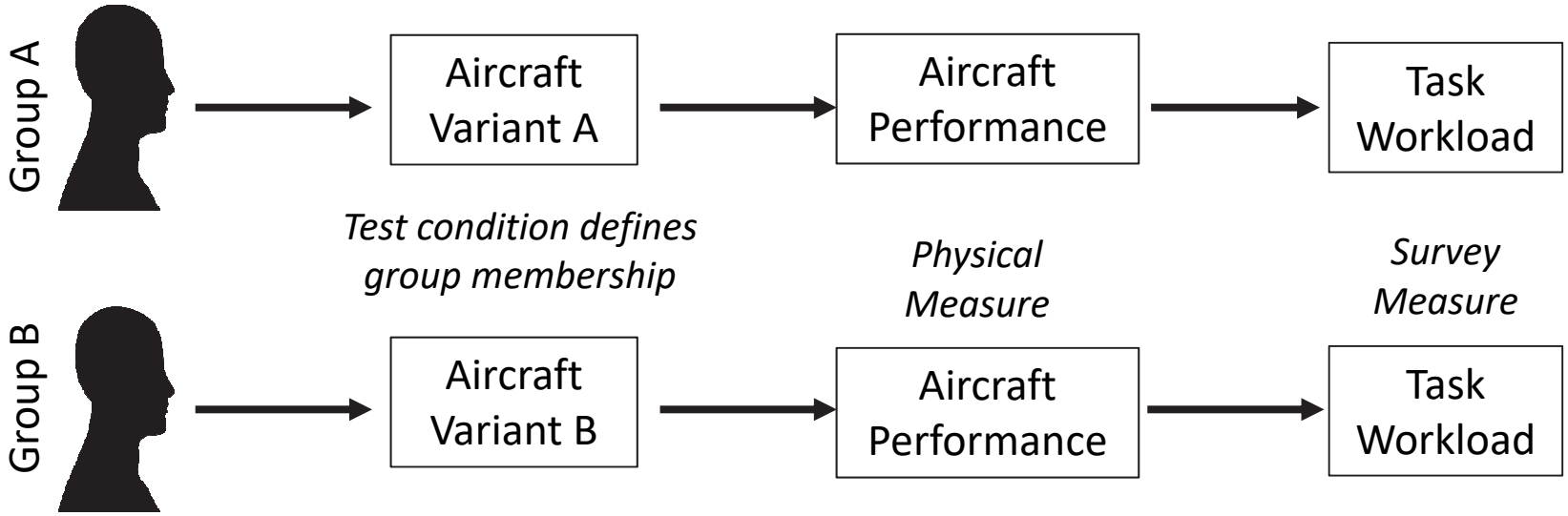
    For example: Are men more likely to trust systems than women?
  - Ratings or performance differ by test conditions

    For example: Does workload differ during day vs. night?
- Between-subjects designs are:
  - **Experimental designs** if group membership is manipulated
  - **Quasi-experimental designs** if group membership is naturally occurring
- Each respondent is assigned to a unique group
- Tasks and context should be the same for all respondents

(Use as callout page)

Between-Subjects Designs

Example:



- Between-subjects designs protect against order effects (learning) and are less demanding for respondents than other designs
- Between-subjects designs require more respondents than other designs

(Use as callout page)

Within-Subjects Designs

- Within-subjects designs compare data from the same respondents across time or test conditions
- Within-subjects designs should be used when you want to understand if:
  - Ratings differ across time

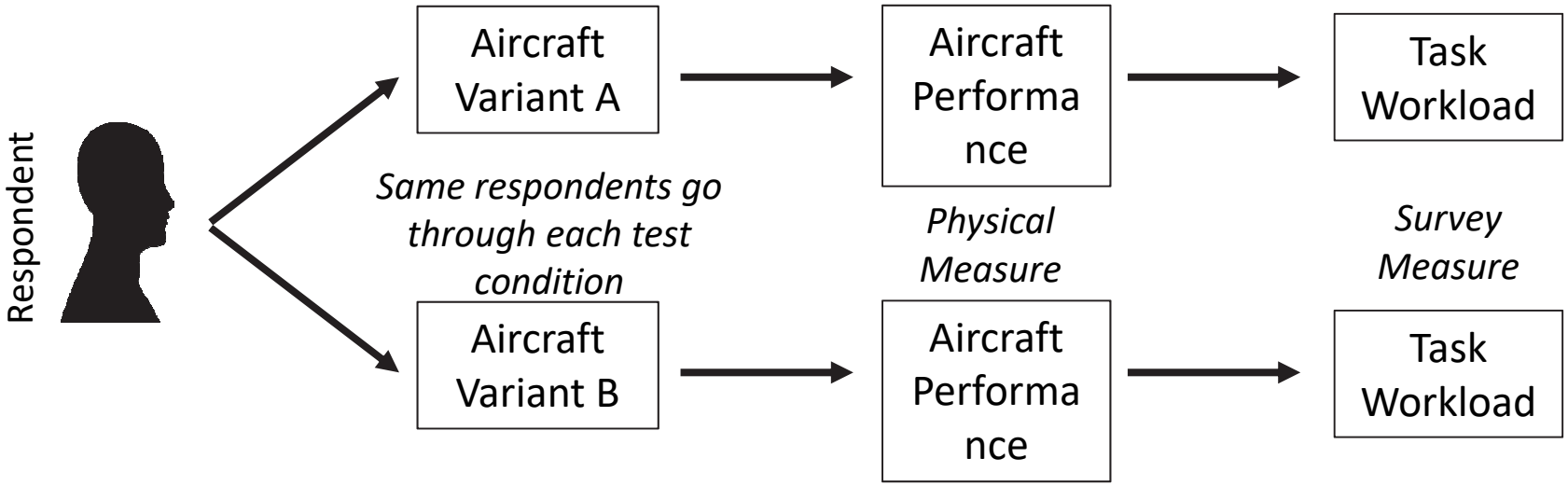
    For example: Do usability scores improve once users gain experience on the system?
  - Ratings differ across test conditions

    For example: Does communication quality decrease in forested vs. urban terrain?
- Within-subjects designs are experimental designs if test conditions are manipulated and quasi-experimental designs if test conditions are naturally occurring
- Every respondent goes through the same test conditions

(Use as callout page)

Within-Subjects Designs

Example:



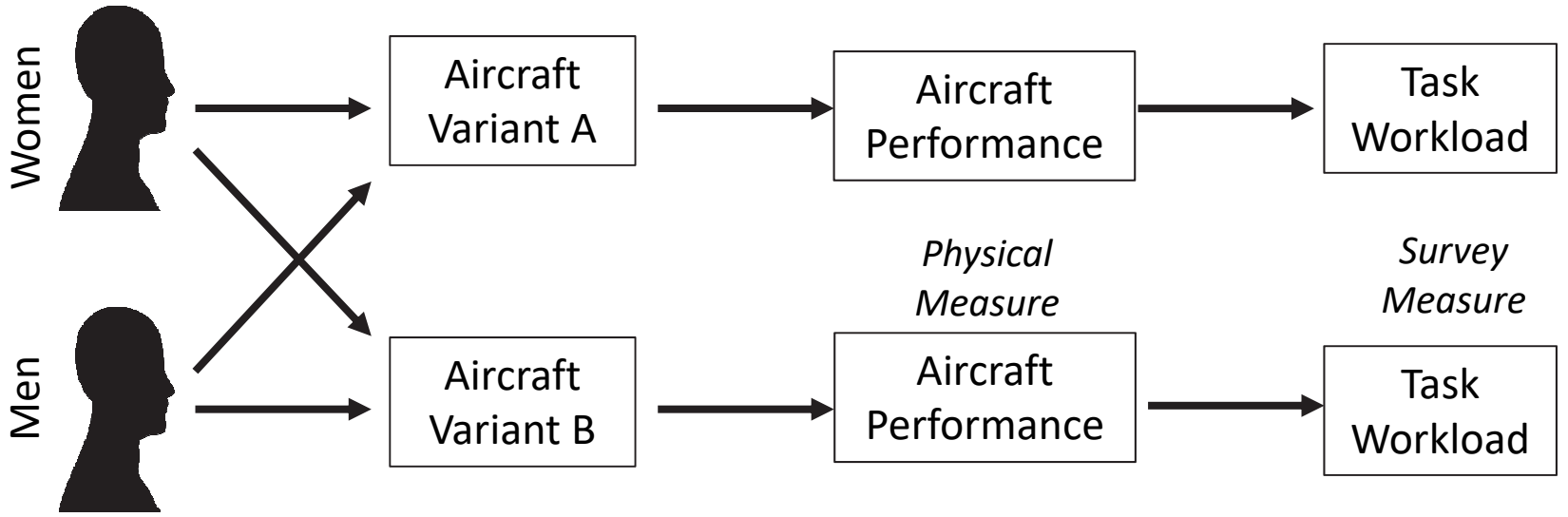
- Within-subjects designs susceptible to order effects (learning) and are more demanding for respondents than other designs
- Within-subjects designs require fewer respondents than other designs

(Use as callout page)

Mixed Designs

- Mixed designs are a combination of between- and within-subjects designs

Example:



Respondents from two separate groups go through the same set of test conditions

Surveys – Part of the Detailed Test Plan

- Surveys are part of the detailed test plan.
- Detailed test plans should include:
  - The type of survey method
  - The number of respondents
  - How respondents are separated into groups (if applicable)
  - The test conditions that each group of respondents will undergo
  - When surveys will be administered throughout the test
  - How surveys will be administered (paper and pencil, electronic, alone, in groups)
  - How survey data from respondents will be mapped to test conditions and performance measures (if applicable)

Summary

This module introduced surveys, survey types and their uses.

You should now be able to:

- Understand the purpose of a survey and how it can be used in T&E.
- Understand the parts of a survey, structured and unstructured questions, and survey quality.
- Understand empirically-vetted and custom-made surveys, and their use.
- Understand that surveys are part of a detailed test plan.

Lesson 5
Statistical Analysis – Regression Modeling


= hyper link to a call out

Lesson Objectives

This module introduces analysis, statistical models and their uses.

At the end you will be able to:

- Define statistical modeling and its key elements
- Understand the analysis checklist
- Understand pre-modeling and exploratory data analysis.
- Understand that different types of statistical models exist
- Understand model fitting and checking assumptions

Overview

- What is a Statistical Model?
- Statistical Analysis Checklist
  - Pre-Modeling
  - Select Analysis Method
  - Fit Model Regression
    - Simple Linear Regression
    - Multiple Linear Regression
    - Categorical Variables
  - Model Fitting & Checking Assumptions
  - Summarize Results & Make Inferences
- Example Analyses

What is a statistical model?

- **There are two general classes of models: deterministic and statistical**
  - Statistical models have a random error terms included in addition to a deterministic component.

- **Deterministic Model**
  - Based in understanding of the physical sciences
  - Make no attempt to explain variability
  - Examples:

    $P = \frac{nRT}{V}$, Ideal Gas Law

    $F = ma$, Newton's Second Law
  - Unfortunately, data collected in the real world exhibit variability which obscures underlying physical processes

- **Statistical Model**
  - A probabilistic model which formally takes into account random variations from the values given by the deterministic model
  - Examples:

    $P = \frac{nRT}{V} + \epsilon$, where ϵ is a random error

    $F = ma + \epsilon$, where ϵ is a random error

Statistical Models

- A very common statistical model we will use is a linear model, where we express our outcome variable (y) as a linear function of our independent variable (X): $y = \beta_0 + \beta_1 X_1 + \epsilon$
  - Where y is a measured response variable, X_1 is an independent variable or factor, β_0 and β_1 are model parameters, and ϵ is a random error
  - Note: this is just the equation for a line! β_0 is the intercept and β_1 is the slope. We will extend to more complex models later.

- Why are statistical models beneficial?
  - Provide a structure for interpreting data and accounting for variability
  - Provide a basis for making predictions about the system capabilities based on test data
  - For complex systems or systems of systems (such as those used in the DoD) the underlying physical relationships are often unknown, statistical models provide a methodology for developing and understanding of system capability

Hypothesis Tests are Important to Statistical Models

- CLM 035, Introduction to Probability and Statistics, covers 1-sample and 2-sample hypothesis tests in detail. It also discusses their relationship to confidence intervals.

- Remember that hypothesis tests are a tool to determine whether a result we've observed (e.g., a difference in two group means) is due to chance
  - If we conclude that the difference is not due to chance, we conclude that is "statistically significant"

- Common simple hypothesis tests used in test and evaluation:
  - Comparison of average performance to a requirements (1-sample hypothesis test)
  - Comparing average performance between new and legacy systems (2-sample hypothesis test)

- Hypothesis tests provide the basis for statistical models, which we will cover in this section.
  - In a statistical model, we will not just have one hypothesis test, but instead a hypothesis test for each model effect

Pre-Modeling &
Exploratory Data Analysis

Analysis Checklist

We will momentarily show an example of fitting a regression model to test data. The most important aspect of statistical analyses is picking the right analysis model to represent your data.

All statistical analyses follow a similar check list:

1. Pre-modeling analysis
2. Specify analysis model
   - Select distribution
   - Specify relationship between explanatory variables and distribution parameters
3. Fit Model
4. Model Selection & Checking Assumptions
5. Summarize Results & Make Inferences
   - Confidence intervals
   - Performance predictions

Pre-Modeling

1. Determine analysis goals
2. Review variables collected
   - What are the types of variables: continuous, discrete?
   - What is the scale of measurement: nominal, ordinal, interval, ratio?
   - Do additional variables need to be considered that were not part of the test plan?
3. Investigate different graphical displays to understand what models make sense for the data
4. Think about prior knowledge, historical data, subject matter expertise to inform what models make sense.

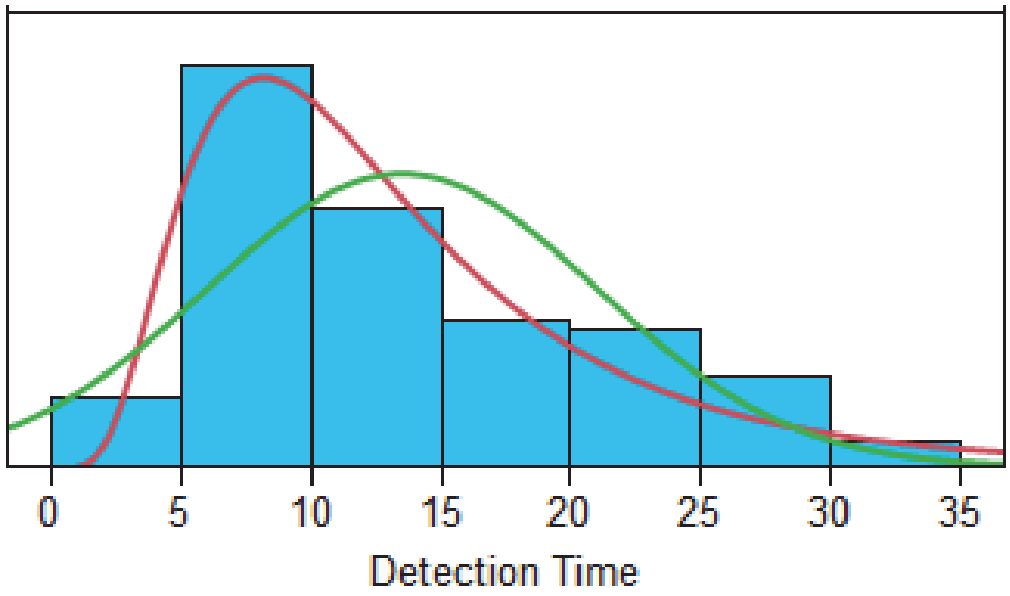
Before attempting to fit a statistical model, or summarize the data, an analyst should become familiar with the data set – Always look at multiple visualizations of your data before conducting a statistical analysis.

Pre-Modeling: Variable Review

- The Test Plan should have outlined the critical response variables (often KPPs, CTPs, etc.), but new outcomes may be observed during testing. This information should be considered in the analysis. Analysts should review and update:
  - Response variables
  - Independent variables (Factors and levels)
  - Covariates (recordable variables that should be accounted for in the model but either aren't of primary interest or cannot be easily controlled)
- Keep in mind that the factors included in your model may change from your original plan due to limitations discovered during testing and additional variables captured during the test.

Pre-Modeling

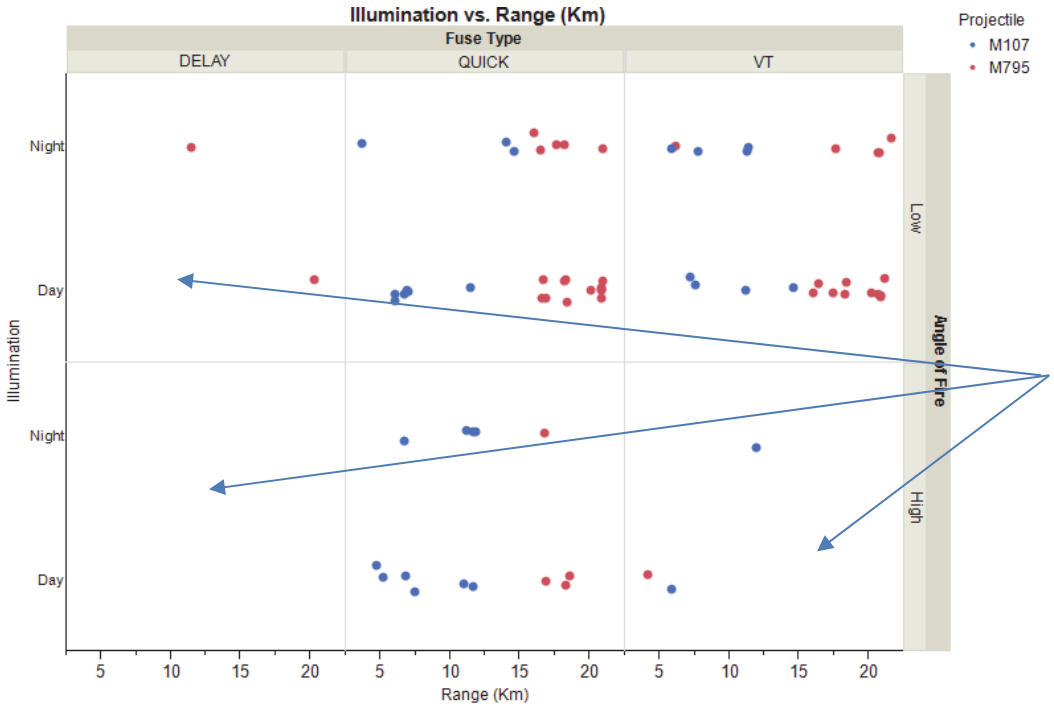
- Choose the most appropriate distribution for the response variable
  - Continuous (e.g. normal, exponential, lognormal) vs. discrete (e.g. binomial)
  - Viewing a histogram can inform the decision – in the example below, the lognormal distribution is clearly a better representation of the data.



= Lognormal
= Normal

Exploratory Data Analysis

- Create factor *x* factor plots
  - Look for gaps in the data (certain factor combinations where no data exists)
  - These will potentially impact the ability to statistically estimate the effects of (or interactions between) those factors with “missing” data
  - A good DOE should prevent unwanted gaps in your data. But examining these plots is important because designs do not always go as planned.



- **Howitzer Example**
  - Data plotted across 5 factors of interest
- **Empty spaces mean some model terms cannot be estimated**

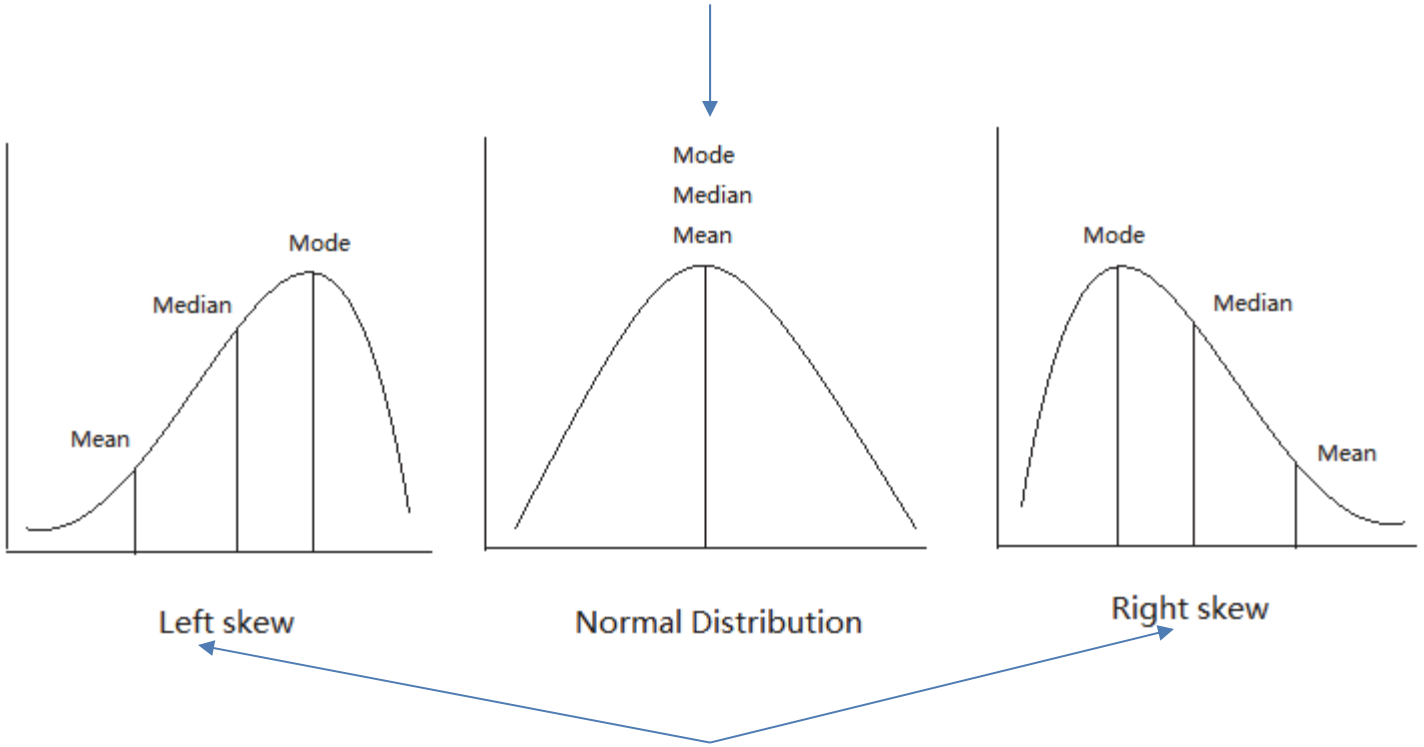
Basic Descriptive Statistics

- Before proceeding to model-fitting, it is useful to summarize your data using descriptive statistics. Descriptive statistics aim to describe basic features of the data in quantitative terms.
- Measures of *central tendency* provide a single number to represent the "typical value" of a variable:
  - Mean: the sum of observations divided by the number of observations. Preferred when a variable's distribution is approximately symmetrical
  - Median: the value that separates the upper 50% of the distribution from the lower 50% of the distribution. Preferred when a variable's distribution is skewed
  - Mode: the most commonly occurring value. Preferred when a variable is measured on a nominal scale (e.g., a categorical variable)
  - Visualizing central tendency and your distribution
- Measures of *variability* provide a single number (or range of numbers) to represent the typical spread or dispersion of a variable:
  - Standard deviation: the typical spread of scores away from the mean
  - Variance: the squared standard deviation
  - Range: the difference between the maximum and minimum value
  - Visualizing variability

Visualizing Central Tendency

(Use as callout page)

For a symmetric distribution, the mean is our preferred measure of central tendency



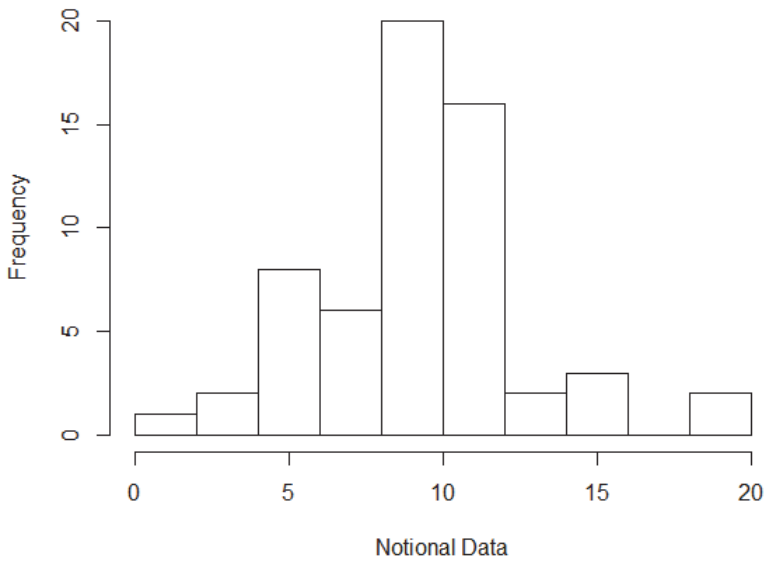
For a skewed distribution, we prefer the median as a measure of central tendency. The mean is influenced by extreme scores, and therefore is not a very representative value for our sample.

Visualizing Variability

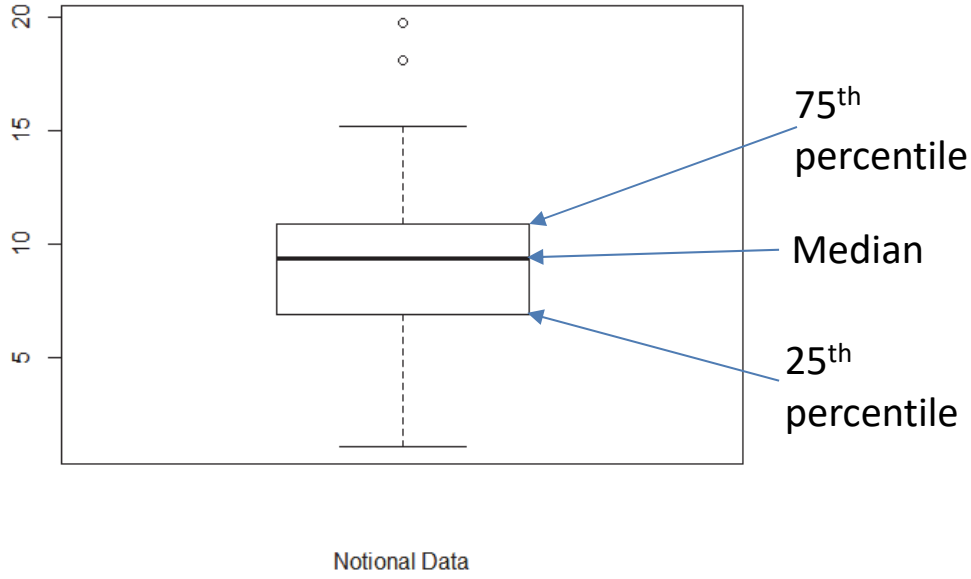
(Use as callout page)

- Several options exist for plotting your data to visualize the amount of variability. Two popular options are shown below that help visualize both central tendency and variability:

Histogram: "bin" the range of values into set intervals and plot the frequency of scores within each interval



Box plot: depict median (50th percentile), 25th percentile, 75th percentile.



**Note: the whiskers on the box plot usually represent 1.5*(75th percentile – 25th percentile)*

Simple Linear Regression

Simple Linear Regression

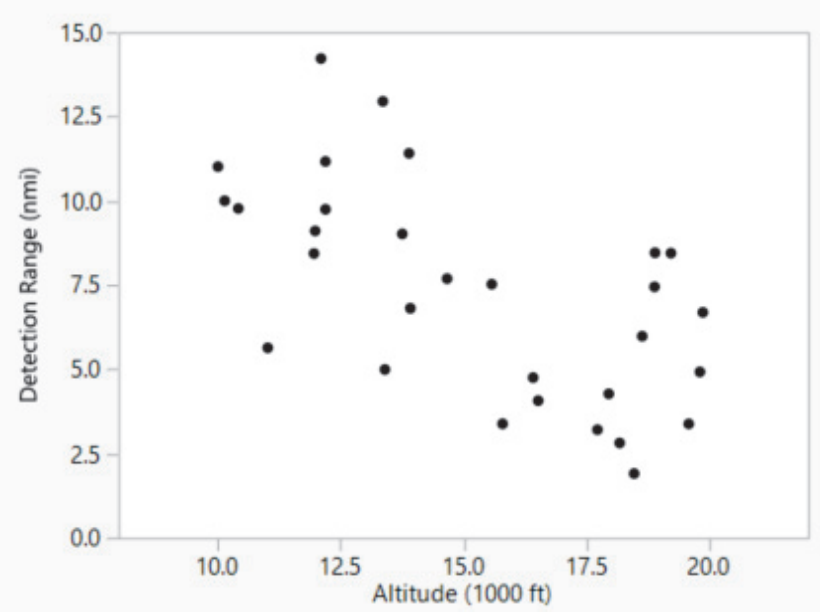
- Simple linear regression (SLR) is a suitable analysis technique when you want to compare the impact of one continuous variable (called a independent variable or factor) on another continuous variable (called a response variable or dependent variable).

- We will use the terminology:
  - the independent variable (x) and the response variable (y) in this course.

- Simple linear regression (SLR) fits a straight line to data with one response variable and one independent variable

Example of Simple Linear Regression

- From our detection range example, we are interested in characterizing the detection range for a new unmanned aerial vehicle (UAV) as a function of altitude
  - Response variable – detection range
  - Independent variable – altitude
- Observational Study approach
  - The figure below shows observational data, where the UAV patrols an area of interest filled with targets of opportunity at a variety of altitudes.
- A scatter plot (left) of the data shows there is likely a linear relationship between detection range and altitude



Important points to note:

- Detection ranges tend to increase as altitude decreases
- The relationship looks somewhat linear, but with a lot of variability
- The altitudes are not fixed to pre-set levels, rather they vary across a range – this is because of the observational study approach to collecting data.

How is the regression line calculated?

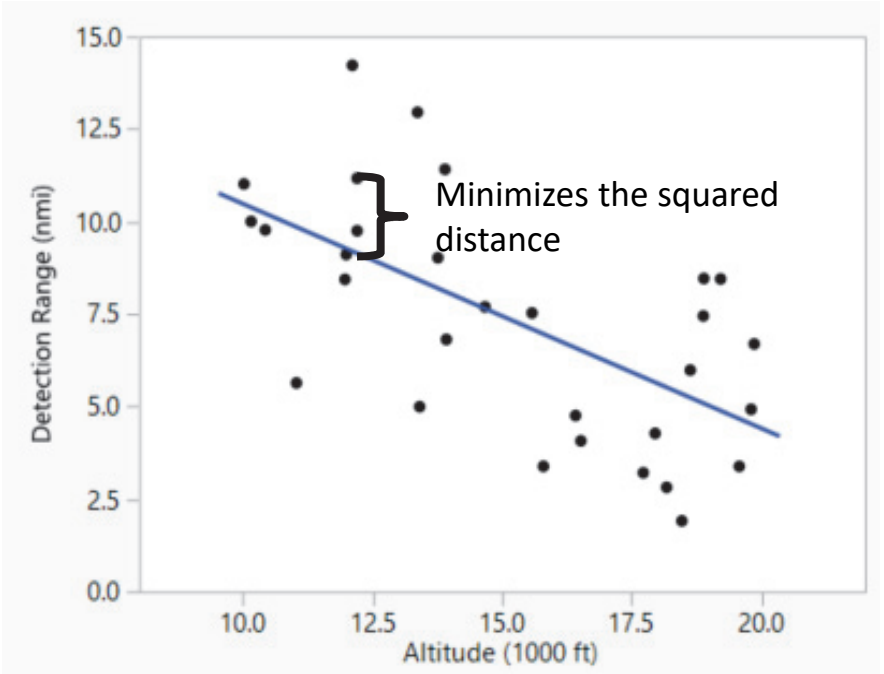
- The prediction equation (solid line plotted in the graph below) is of the form:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  - note: $\widehat{\ }$ (pronounced hat) above a coefficient/variable indicates an estimated quantity
  - For every value of the independent variable (altitude), x_i, we will get a predicted value for the response variable (detection range), y_i.
  - The error term is not shown in the prediction expression.

- The regression line minimizes the average squared distance of the points from the line. This criteria is called "least squares" so the regression is often referred to as least squares regression.

- Maximum likelihood estimation is another type of estimation method that is used in many software packages, but not covered in this introduction.



Prediction equation:
Detection Range$= 16.5 + 0.61 * Altitude$

How is the regression line calculated? Link to next slide

How is the regression line calculated? Mathematical Details

Link to call out on previous slide.

- Recall from calculus, that minimums and maximums of functions can be found using 1st partial derivatives.

- For simple linear regression, $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the sums of the squared differences between the response and the regression line if they satisfy:

$$\frac{\partial}{\partial \beta_o} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0$$

- Taking derivatives yields:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- Solving the equations simultaneously yields $\hat{\beta}_0$, $\hat{\beta}_1$ the estimates of β_0, β_1:

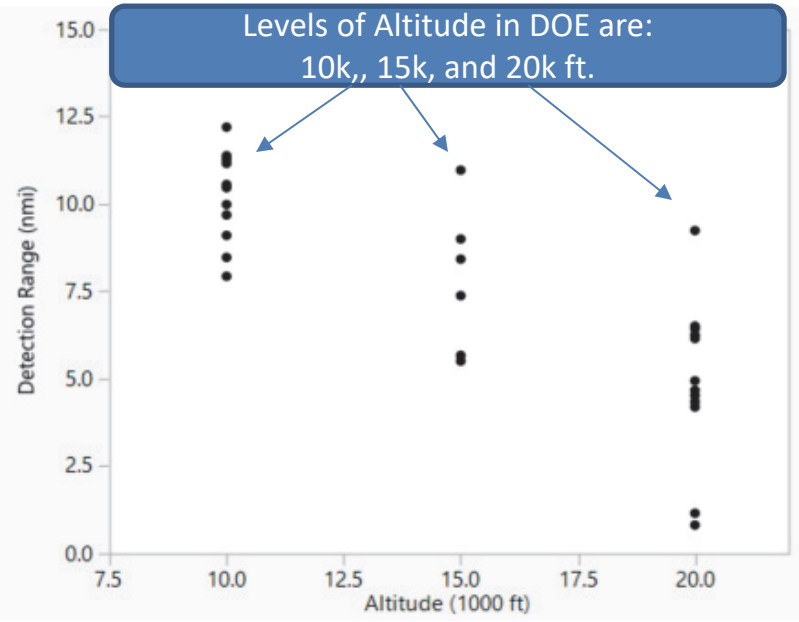
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

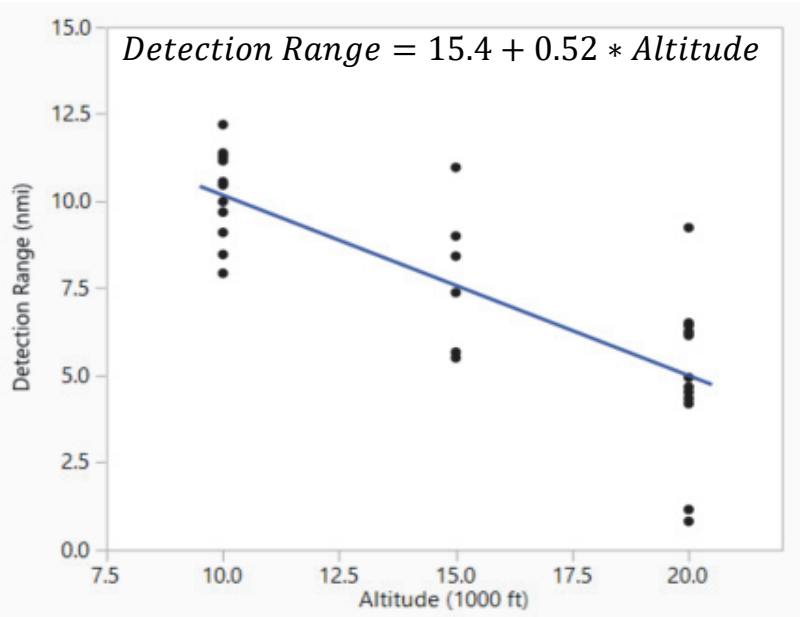
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

  - Where $\bar{x}$ is the average of the independent variable (x) and $\bar{y}$ is the average of the response variables

Regression Analysis for a Designed Experiment

- The analysis is the same for data from designed experiments
- Experimental Design approach
  - The figure below shows a single factor design, where the UAV is send to three different altitudes and patrols an area of interest filled with fixed targets.
  - The experimental design maximizes statistical power by placing test points at high and low altitudes with center points to check for curvature
- We can fit the same model using the same math, but now our x values are purposefully controlled and set to pre-defined levels.





Interpretation of Simple Linear Regression

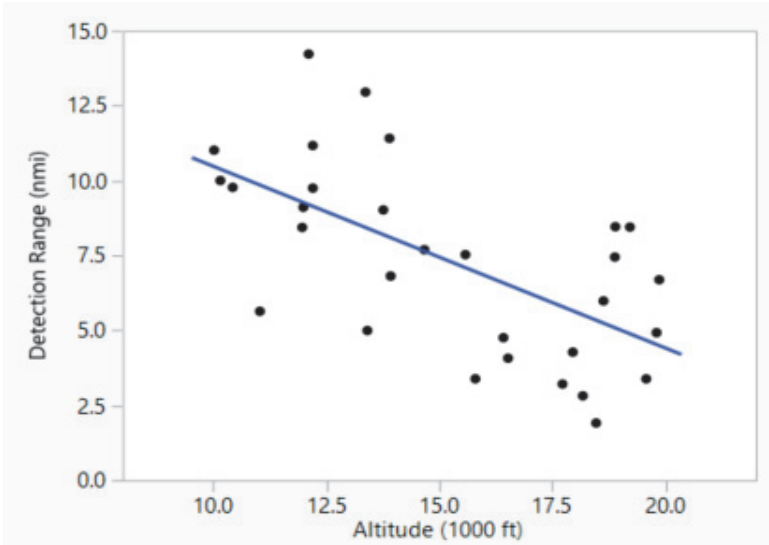
- Remember that statistical models require hypothesis tests on model parameters
  - Hypothesis tests tell us if the differences or relationships we see in our statistical models are due to chance.

- Just estimating a line does not tell us how well the line fits the data! Two elements are useful for understanding if the model is significant.

  1 – Test of Slope

  2 – Coefficient of determination

- **Test of Slope:**
  - We can do a hypothesis test on the slope of the prediction equation. If the slope is 0, then there is no relationship between the response variable and predictor variable. A non-zero slope indicates some relationship.
  - Statistical significance quantifies whether the result (change in the outcome) is likely due to chance or the change in the independent variable (altitude in our example)
    - H_0: $\beta_1 = 0$ (no relationship)
    - H_1: $\beta_1 \neq 0$ (relationship)
  - Statistical software provides a comparison of the slope to unexplained variability. If a "statistically significant" portion of the variability is explained, then we reject the null hypothesis and concludes there is a relationship.



Interpretation of Simple Linear Regression

Just estimating a line does not tell us how well the line fits the data! Two elements are useful for understanding if the model is significant.

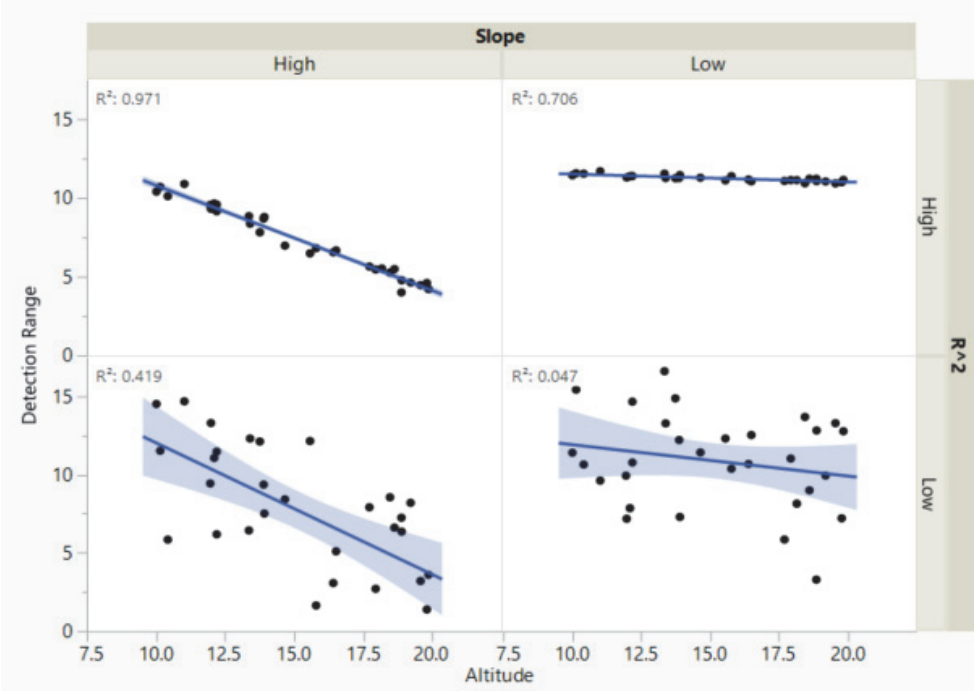
1 – Test of Slope

2 – Coefficient of determination

- **R^2 (Coefficient of determination)** - is a measure of how well the model fits the data. R^2 is always between 0% and 100%. It is the percent of variation in the data that is explained by the regression line. A higher R^2 means that the regression line is explaining a lot of variation in the data.

- Using the Slope and Coefficient of Determination - The figure to the right shows different combinations of slopes and R^2. In only the low slope, low R^2 case (bottom right panel) is the test of slope nonsignificant. This means that the slope is barely different from zero, and there is a lot of variability around the slope.

- Practical Significance - An additional consideration is practical significance. In the case of a low slope, but high R^2 (top right panel), the result may be statistically significant but not of practical significance because the slope is so close to zero.



Extend to Multiple Regression

DoD T&E Data with Multiple Independent Variables

- DoD Testing often requires more than one independent variable.
  - For example, DOE is often used to cover many independent variables (factors) simultaneously.
  - In Lesson 2 on DOE, we learned about several design options involving multiple factors, and potential interaction effects

- The corresponding statistical analyses ensure that we can determine the significant factors (variables) and answer the following questions:
  - Which independent variables out of many possible affect performance the most? By how much?

Hypothesis Tests – Multiple Independent Variables

- UAV Example:
  - For the same UAV previously considered, we now want to detect both slow and fast moving targets. The detection range of the UAV is investigated using a designed experiment.
  - Recall from the DOE section, a full factorial design covers all combinations of independent variables (factors). Testers run 30 trials:
    - 2 Target Types*3 altitudes*5 replications each = 30 trials
- Multiple Hypothesis Tests to Consider (one for each effect):

Overall Test

H_0: *Detection range is the same everywhere considered in the test*

H_1: *Performance is significantly different at least one location in the operational envelope (e.g. slow targets can be detected at longer ranges)*

Mathematically:

$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_n$, where μ is the average detection range at a specific test condition

$H_1: At\ least\ one\ \mu_i$ is $different,$

Test for Each Factor (i.e., target type, altitude):

H_0: *Detection range is the same for each target type*

H_1: *Detection range is different for one target type*

Or

H_0: *Detection range is the same at all altitudes*

H_1: *Detection range changes with altitude*

Multiple Linear Regression

- Simple linear regression is only appropriate when there is a single independent variable. When you have multiple independent variables (Xs), multiple regression should be used.

- It is more difficult to create a useful graph in multiple regression because we are dealing with multiple dimensions. For example, with two predictors, we would have to visualize the relationships in 3-dimensional space.

- Scatterplots or fitted line plots can be created separately for each predictor variable with the outcome, but these may be misleading. This is because the interrelationship between predictor variables, or potential interaction effects, would not be captured.

- Model for multiple linear regression
  - The model for multiple regression will have a coefficient for each independent variable and one coefficient for the intercept term.
  - The coefficient is the expected change in the response variable, *holding all the other variables constant*.
  - If we have 2 predictor variables and one response variable then the prediction expression is:
    - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$

Important Concept – Dummy Variables

- Often in the DoD, we have variables that are not continuous in nature.
  - For example, in our UAV test we have vehicle type (tracked and wheeled) as well as the presence/absence of countermeasures.

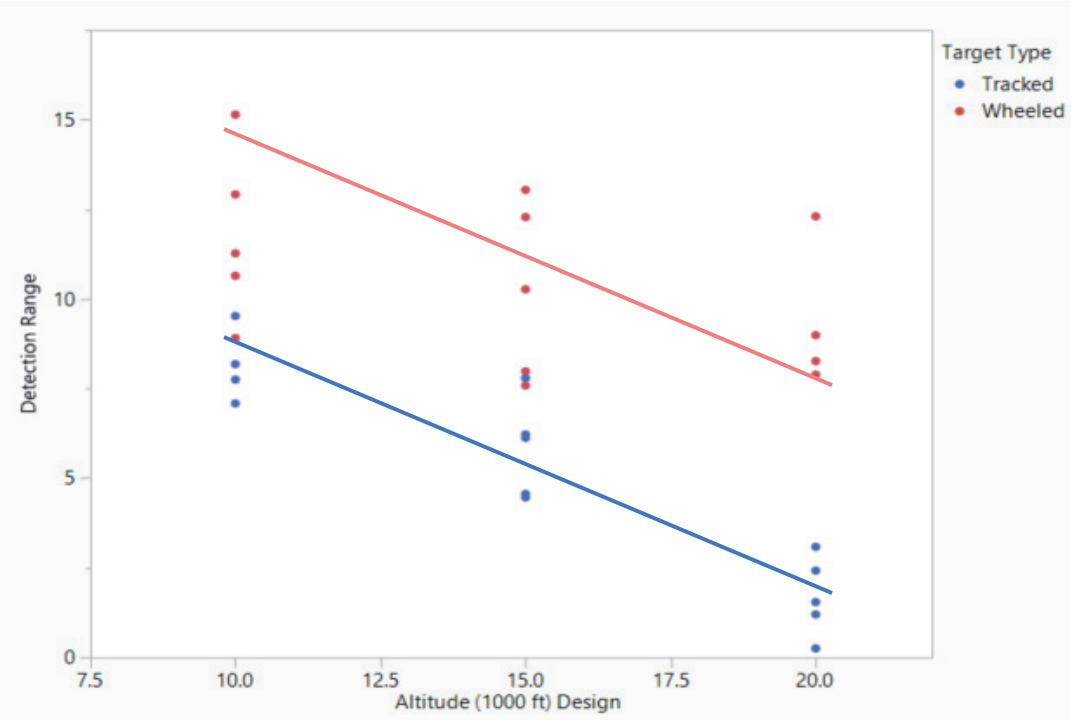
- To represent these categorical variables in the model, we use what is known as a "dummy variable." Dummy variables allow for any categorical factor with multiple levels (k levels) to be recoded as (0 or 1) indicator variables using k-1 new variables. These dummy variables are then used in the regression analysis in place of the original categorical variable.

- For example, we can define $x_2 = \begin{cases} 1, if\ wheeled \\ 0, if\ tracked' \end{cases}$
  - This above example uses k-1, or 1 indicator variable.

- If there were actually three vehicle types of interest: tracked, wheeled, hover (k=3), we could make 2 dummy (indicator) variables to represent these 3 levels.
  - We define $x_2 = \begin{cases} 1, if\ wheeled \\ 0, if\ tracked \\ 0, if\ hover \end{cases}$
  - And $x_3 = \begin{cases} 0, if\ wheeled \\ 1, if\ tracked \\ 0, if\ hover \end{cases}$
  - Thus, our three levels would be represented as follows for a single individual:
    - Wheeled vehicle: x_2 = 1 and x_3 = 0
    - Tracked vehicle: x_2 = 0 and x_3 = 1
    - Hovering vehicle: x_2 = 0 and x_3 = 0

Extension to Multiple Linear Regression

- Consider the UAV example now with two types of targets, fast (wheeled) and slow (tracked) vehicles
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$
  - Where $x_2 = \begin{cases} 1, if\ wheeled \\ 0, if\ tracked \end{cases}$
- The result in this simple analysis is two regression equations
  - However, notice that they are parallel lines (this may not be reasonable)



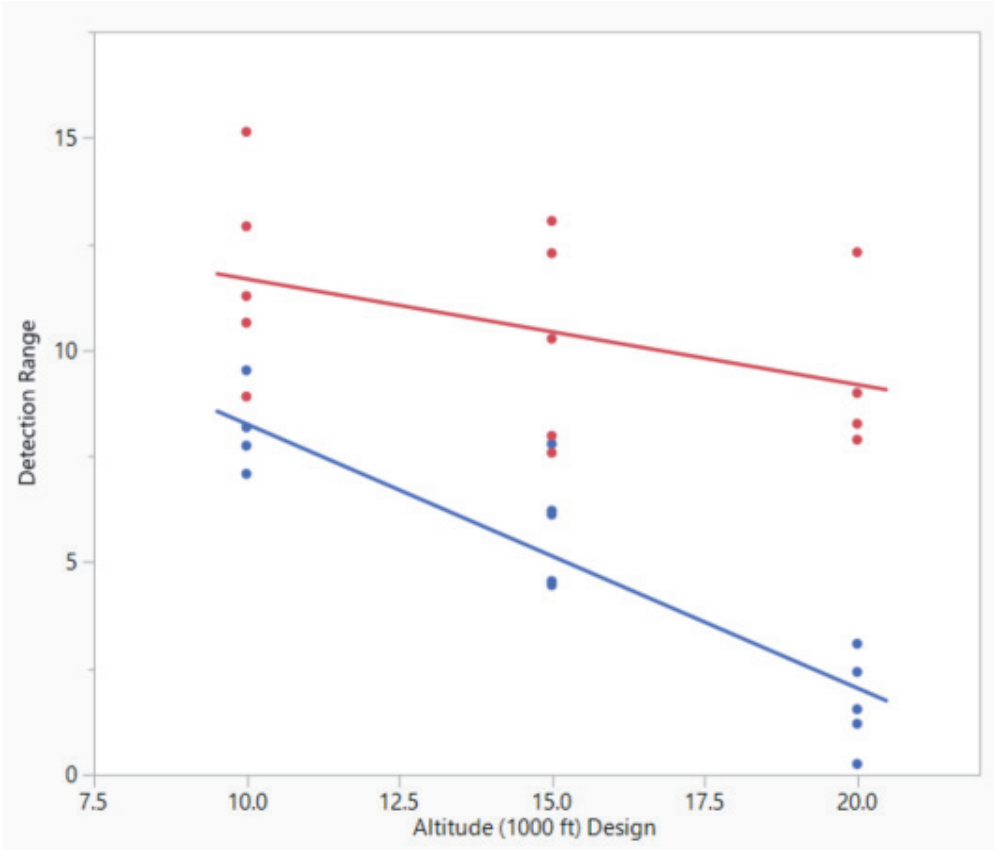
Multiple Regression with Interactions

- Consider the UAV example now with two types of targets, fast (wheeled) and slow (tracked) vehicles
- A multiple regression with interactions allows for the relationship between detection range and altitude to be different for tracked vehicles than wheeled vehicles. This is clear in the non-parallel lines below.
- The model is now:
  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_{12} x_{1i} x_{2i}$
  - Where $x_2 = \begin{cases} 1, if\ wheeled \\ 0, if\ tracked \end{cases}$

Interaction term in model

The result in this simple analysis is two regression equations that do not have to be parallel.

We had to include the two-factor interaction to see the difference in slope between tracked and wheeled vehicles. This is why two-factor interactions are important!



Explain the Math Link

More Complex Model Structures

- We can leverage low-order polynomials to fit the data.
  - A first order model (only main effects):
    - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ji}$
    - For k different independent variables
  - A second order model:
    - $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ji} + \sum_{j=1}^{k} \hat{\beta}_{jj} x_j^2 + \sum_{j=1}^{k} \sum_{l>j}^{k} \hat{\beta}_{jl}\, x_{ji} x_{li}$
  - Analysis seeks to account for major sources of variation using low order polynomials

Quadratic Terms

Two-way Interactions

- Higher order models are possible, but rarely necessary to explain the variation in the data from designed experiments.
  - For example, third order models adds three-way interactions and cubic terms)
  - Importantly, the test design must support higher order model terms
    - For example, you cannot fit a quadratic model if you do not have sufficient data to estimate a quadratic model
    - Need two levels to fit a line, three levels to fit a quadratic curve, etc.

Potential Pitfalls

- Do not extrapolate
  - The prediction equation is only valid for the range of data that you collected. Trying to extrapolate outside that range may lead to misleading conclusions.

- Beware nonsensical relationships
  - Do not do regression on variables that don't make sense.

- Correlation does not mean causation
  - Just because you see a strong relationship, that doesn't mean that the predictor variable *causes* changes in the response variable.

Regression Assumptions and Diagnostics

Regression Assumptions

Regression is a powerful tool for T&E. Its responsible used is based on several core assumptions being met:

1. Independence – The responses are statistically independent of one another
2. Normally Distributed – For a fixed level of the factor, the response has a normal distribution
3. Homoscedasticity – The variance of the response is the same for any value of the independent variable; in other words, variance is constant
4. Linearity – the model is linear in the parameters.
5. No Outliers

Checking Assumptions

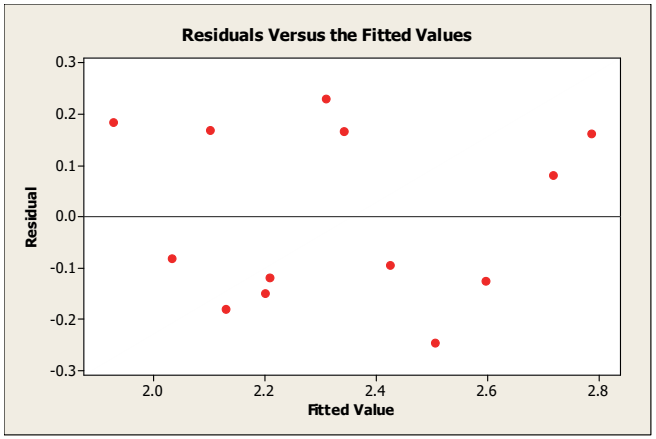
- Residual analysis is used to check the assumptions made when we are doing regression
- Every data point has a residual that is simply the difference between the actual value and the predicted value (also known as the fitted value).
- What does a residual analysis consist of?
  - Making plots of residuals that check each of the assumptions:
    1. Residuals vs. predicted values to check constant variance assumption
    2. Residuals vs. time to check independence assumption
    3. Normal probability plot to check normality assumption
    4. All 3 of the previous plots can be used to check for existence of outliers
  - Good news! These plots are created by most software programs
- Recall, for a simple linear regression the predicted value is:

  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, so the residual is:

  $e_i = y_i - \hat{y}_i$ (observed data minus predicted value)

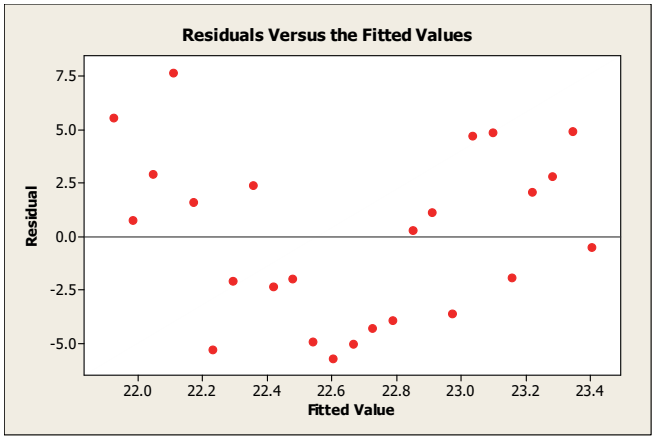
Residuals vs. Predicted Values

Good (assumptions meet) – equal variance

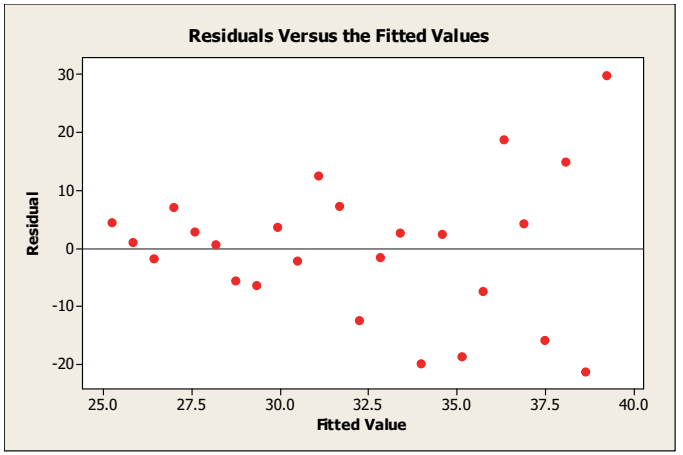


Generally, no patterns in the residuals by the predicted variable indicates that the model explains the functional relationships in the data well and all that is left is noise. A good plot has no clear pattern. Any pattern is an indication of a problem.

Bad (assumptions not meet) – parabolic shape



Bad (assumptions not meet) – funnel shape indicates increasing variance

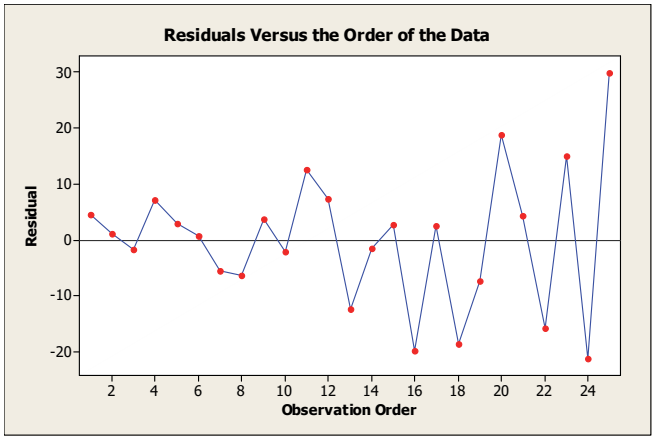


Residual vs. Time Order

- Again, no pattern is a good indication of no time order dependence in the data.
- A good experimental design should help ensure that the assumption of independence is not violated (e.g., time will be randomized)

Good – independence, stable

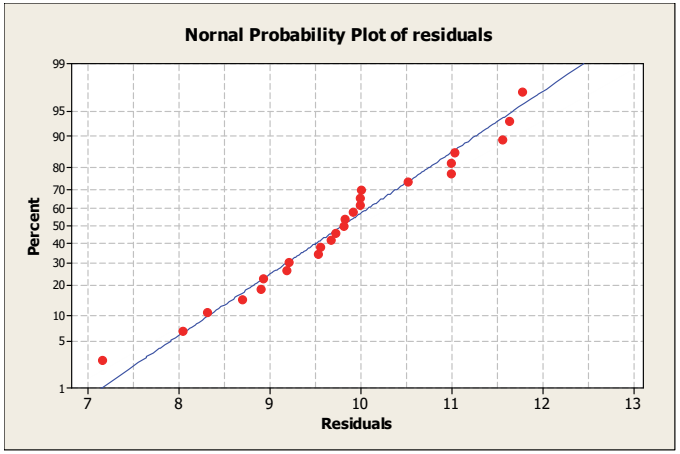
Bad – not independent, the residuals are getting larger as time goes on



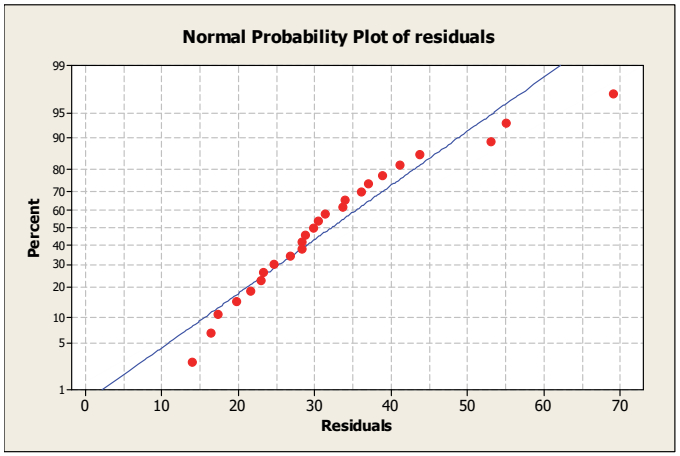
Normal Probability Plot of Residuals

- The plot should follow a straight line, any curvature from the line indicates a departure from normality

Good – data follow the straight line

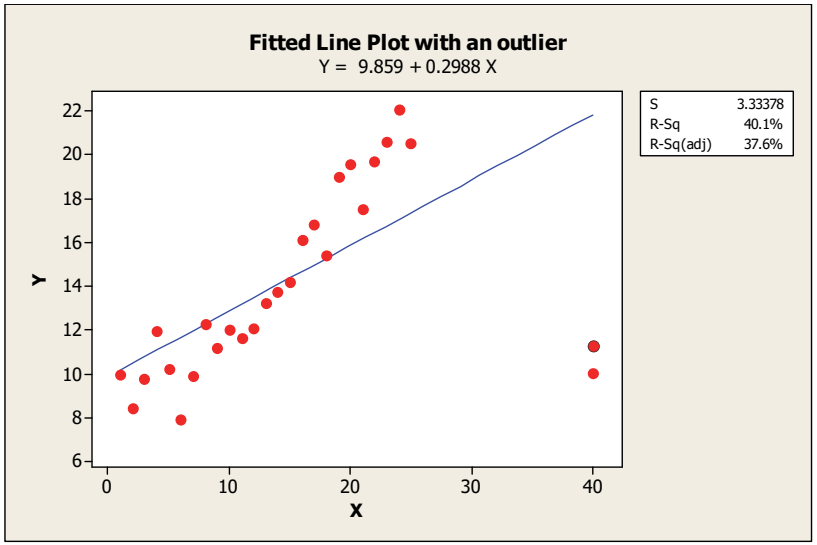


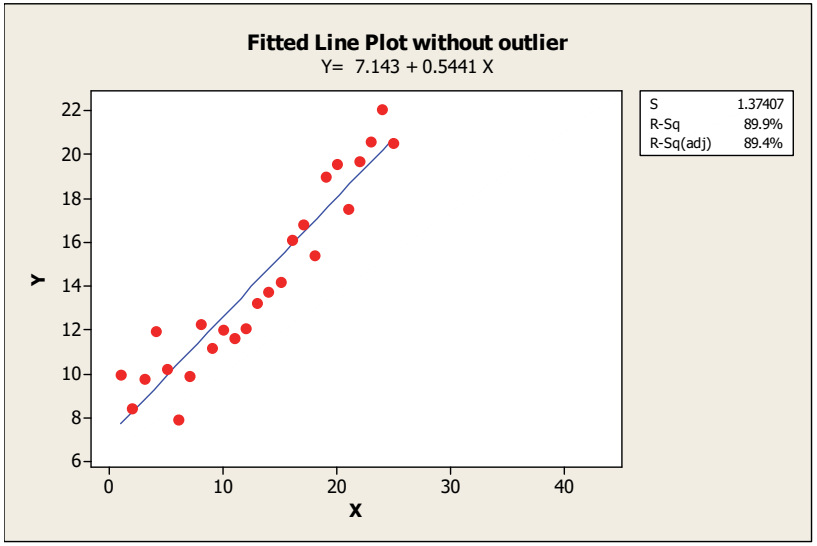
Bad – curvature indicates non-normality



The Impact of Outliers

- Don't ignore outliers – they can have a huge impact on the regression analysis (especially with a small sample size).
- Notice in the example, the slope of the line changes dramatically when an outlier is included.
- However, outliers should not be arbitrarily removed from the analysis. If at all possible, a root cause for the outlier should be identified before removing it from the analysis. If no root cause can be found, then it should still be reported on, but potentially not included in the regression analysis.





What if the assumptions are not met?

- There are many statistical models available
- Common assumptions that are violated and some fixes:
  - Independence
    - Account for the dependence in the model:
      - Repeated measures (multiple measurements on an individual)
      - Add a time based variables to the model (could be useful to account for learning curves)
  - Unexplained patterns in the residual by predicted graph:
    - Add variables
    - Add interactions
    - Add higher powers of variables (e.g., x^2, x^3)
  - Normality & Homoscedasticity
    - Transformations of response variables
      - For example, take the log() of a detection time variable
    - Pick a different distribution - Generalized linear models (GLM)
      - Exponential, Gamma, Binomial (Logistic Regression)

Model Selection

Model Selection Overview

- The goal of model selection is to choose a sparse model (i.e., fewest number of variables) that adequately explains the data
  - Model selection includes selecting the right outcome distribution and which factors, interactions, and higher order terms should be including in the model
- Statistical/empirical model can then be used to:
  - Make **statements** about changes in performance across the operational envelope (e.g., performance during the day was better than performance at night)
  - Make **predictions** of system performance (i.e., characterize performance) across the operational envelope
- Multiple model selection methods exist, with various evaluation criteria to decide on the best model

There may be multiple correct solutions to model selection!
It is important to assess the robustness of the conclusion to the analysis.

Model Selection Methods

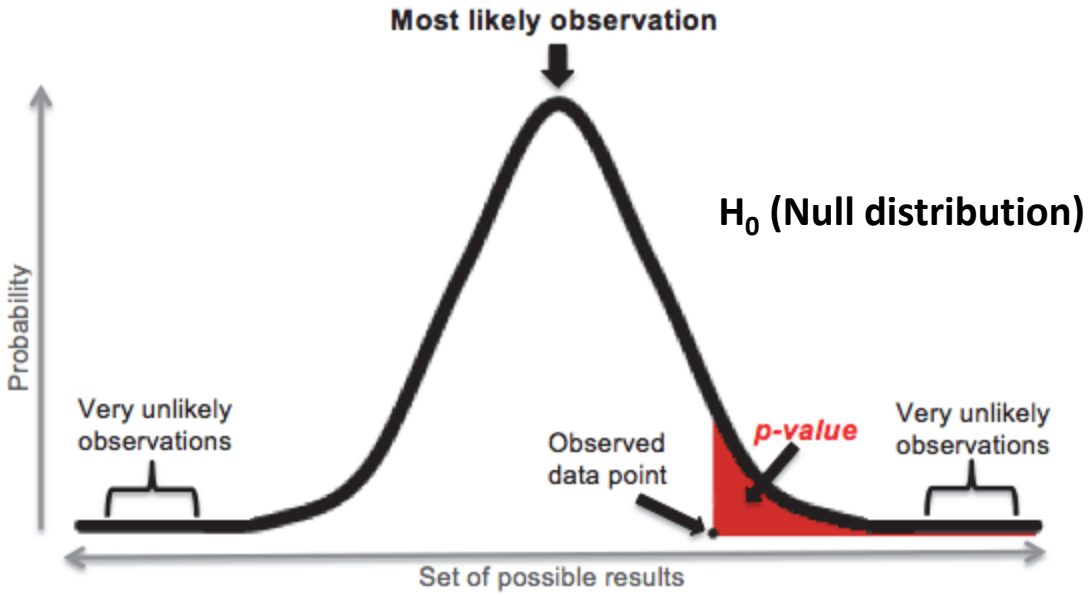
(Use as callout page)

- Forward Selection: Start with nothing but an intercept in the model; test the addition of each variable using a chosen criterion; add the variable (if any) that improves the model the most; repeat adding variables one at a time until none improve the model
- Backward Selection: Start with all possible variables in the model; test the deletion of each variable using a chosen criterion; remove the variable (if any) that improves the model the most by being deleted; continue removing variables until no further improvement is possible
- Stepwise Selection: A combination of the above methods; test at each stop for variables to be added OR removed
- All of these methods are automated in most statistical software using different selection criteria

Model Selection Criteria

(Use as callout page)

- p-value
  - Probability that the effect due to a particular factor (or interaction) occurred by chance alone
  - Smaller p-value = stronger evidence of factor effect (or interaction)
  - **We want to keep factors in our model with small p-values**



A *p-value* (shaded red area) is the probability of an observed (or more extreme) result arising by chance

p-value in a nutshell: the null distribution represents all of the values of our test statistic we would expect due to chance.

A high p-value means that there is a high probability of observing that result just due to chance alone.

A low p-value means that there is a low probability of observing that result due to chance alone.

Model Selection Criteria (cont.)

Information Criteria

Methods for comparing various candidate subsets of factors are based on a tradeoff between 2 things:

1. Lack of fit (measured by model likelihood)
2. Complexity (measured by number of parameters in the model)

- Information criteria are used when we want to compare the relative fit of competing models. Two popular criteria are the AIC and BIC
  - Akaike Information Criterion (AIC)
    - AIC = $-2\,ln(\text{likelihood}) + 2p$, where p is the # of parameters in the model
    - Smaller is better
    - Discourages over-fitting
  - Bayes Information Criterion (BIC)
    - BIC = $-2\,ln(\text{likelihood}) + p\,ln(n)$, where p is the # of parameters in the model and n is the number of observations in the dataset
    - Larger penalty for more terms than AIC, usually resulting in a sparser model

(Use as callout page)

Model Selection Conclusions

- Model selection is a critical part of statistical analysis
  - Goal is to obtain a *sparse* model that adequately *explains* the data
  - Always think about what you will do with the modeling results
- Get to know your data before fitting models – do pre-modeling
  - Choose appropriate distribution of response variable
  - Plot response variables and independent variables.
- Various model selection methods and criteria to choose from
  - There is no ONE correct answer!
  - Use automated procedures in software to narrow down terms of interest
  - Closely examine competitive models and incorporating subject matter expertise as appropriate
- Consider both *statistical* and *practical (operational)* significance
  - Consider the implications for reporting

Summarize Results & Make Inferences

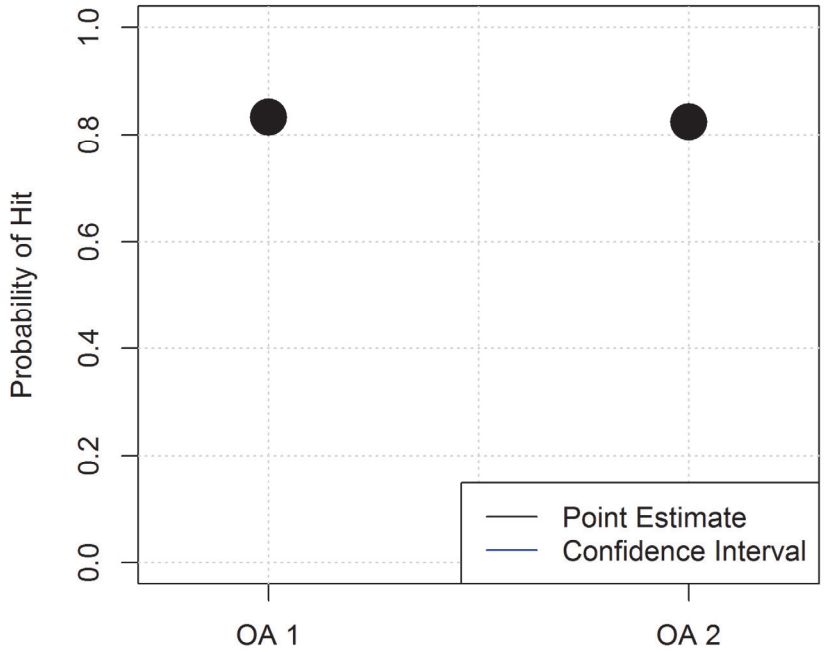
Reporting on a Regression Analysis

- Once you select the final regression model, it can be used to make predictions and provide a level of confidence in those predictions.
- All factors used to construct the DOE should be identified, and there should be a discussion of which factors were used in the analysis and why.
- Significant factors are often best presented in a graph with confidence intervals.
- Nonsignificant factors may also be important to discuss because they can inform future testing planning. Additionally, they can provide operational context on where performance is robust.
  - For example, "Detection range did not depend on the vehicle speed suggesting the sensor is robust to vehicle speeds."
- Interval estimates are extremely important because they allow you to make decisions based on what results you might see in the future. Types include:
  - Confidence intervals – tell you about the expected value of the parameter of interest
  - Prediction intervals – tell you about an individual future observation

Interval Estimates Are Important

- Results from a single test event cannot predict exactly how a system will perform in the field
  - Confidence intervals tell us how precise our test results are
  - More data → tighter confidence bounds
- Example: New turret for LAV Anti-Tank variant (notional data)
  - Shoots TOW missiles
  - Operational Assessment (OA) 1: 12 shots
    - 10 hits
  - OA 2: 40 shots
    - 33 hits

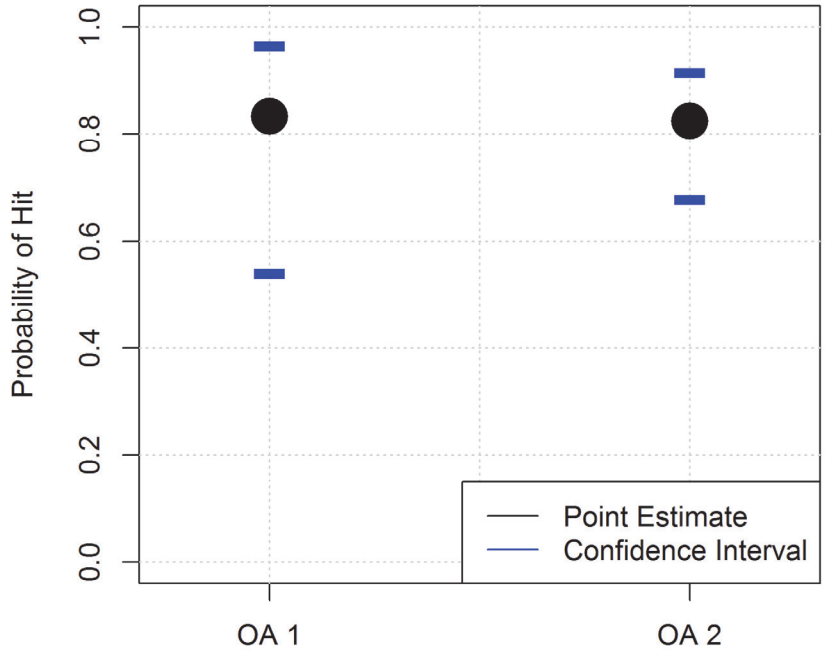
Call out for why interval estimates are important



Interval Estimates Are Important

- Results from a single test event cannot predict exactly how a system will perform in the field
  - Confidence intervals tell us how precise our test results are
  - More data → tighter confidence bounds
- Example: New turret for LAV Anti-Tank variant (notional data)
  - Shoots TOW missiles
  - OA 1: 12 shots
    - Interval Width: 42.5%
  - OA 2: 40 shots
    - Interval Width: 23.8%

Interval estimates show the range of likely values for the probability of hit if the same number of shots are conducted again under similar conditions.



Putting It All Together:
Air to Ground Missile Example

Concluding Regression Example: Air to Ground Missile

- Air to Ground Missile Test
  - Objectives:
    1. Characterize performance of a new air-to-ground missile
    2. Compare the new missile to legacy
  - Response variable: miss distance
  - Factors: range to target, altitude, speed, variant (new versus legacy)
- Test Design
  - Full factorial, recall from the DOE section this is a 2^4 = 16 run design

The 16 run test design and miss distance outcomes is on the right.

- Note that the values for the continuous variables match the specific test design values. If during the test the altitude was actually 24,000 ft. on one run, that should be recorded and used in the analysis.
- Note: range is scaled to values of -1 and positive 1

Run	Variant	Range	Altitude (1000 ft)	Airspeed (Mach)	Miss Distance (ft.)
1	New	-1	35	0.85	1.14
2	Legacy	1	35	0.95	41.47
3	New	1	25	0.85	18.45
4	Legacy	-1	35	0.95	13.76
5	New	1	35	0.95	39.81
6	Legacy	1	25	0.85	5.23
7	New	-1	25	0.85	13.04
8	Legacy	1	35	0.85	5.63
9	New	1	25	0.95	41.90
10	New	-1	35	0.95	8.58
11	Legacy	1	25	0.95	40.09
12	Legacy	-1	25	0.85	4.65
13	Legacy	-1	35	0.85	26.55
14	Legacy	-1	25	0.95	10.58
15	New	1	35	0.85	10.44
16	New	-1	25	0.95	3.44

Data Analysis

- The full factorial design allows us to consider multiple models from the 4 independent variables.
- A full second order model with linear and 2nd order interaction terms were considered. The estimated coefficients for each model term are shown to the right.
- Model selection is next to reduce the model to only the significant terms

Model Term	Estimate
Intercept	-114.765
Variant	0.6975
Range	7.58
Altitude	0.125
Airspeed	143.125
Variant*Range	-2.97
Variant*Altitude	0.5465
Range*Altitude	-0.333
Variant*Airspeed	16.475
Range*Airspeed	165.675
Altitude*Airspeed	1.305

Full model:

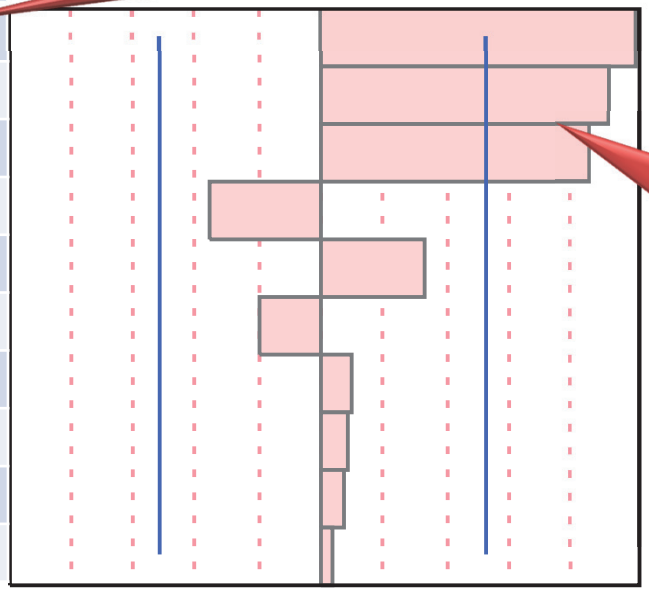
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \cdots + \beta_{34} x_3 x_4$$

$$\widehat{Miss\ Distance} = -114.765 + (0.6975 \text{ x } Variant) + (7.58 \text{ x } Range) + \cdots$$

$$(-2.97 \text{ x } Variant * Range) + \cdots + (1.305 \text{ x } Altitude * Airspeed)$$

Model Selection- Important Variables

Model Term	P-Value
range*air speed	**0.0043***
range	**0.0062***
air speed	**0.0079***
variant*range	0.1363
variant*altitude	0.1636
range*altitude	0.3657
variant*air speed	0.6436
variant	0.6943
altitude	0.7242
altitude*air speed	0.8532



Small p-value means there's little chance the change in performance when changing this factor is due to chance alone

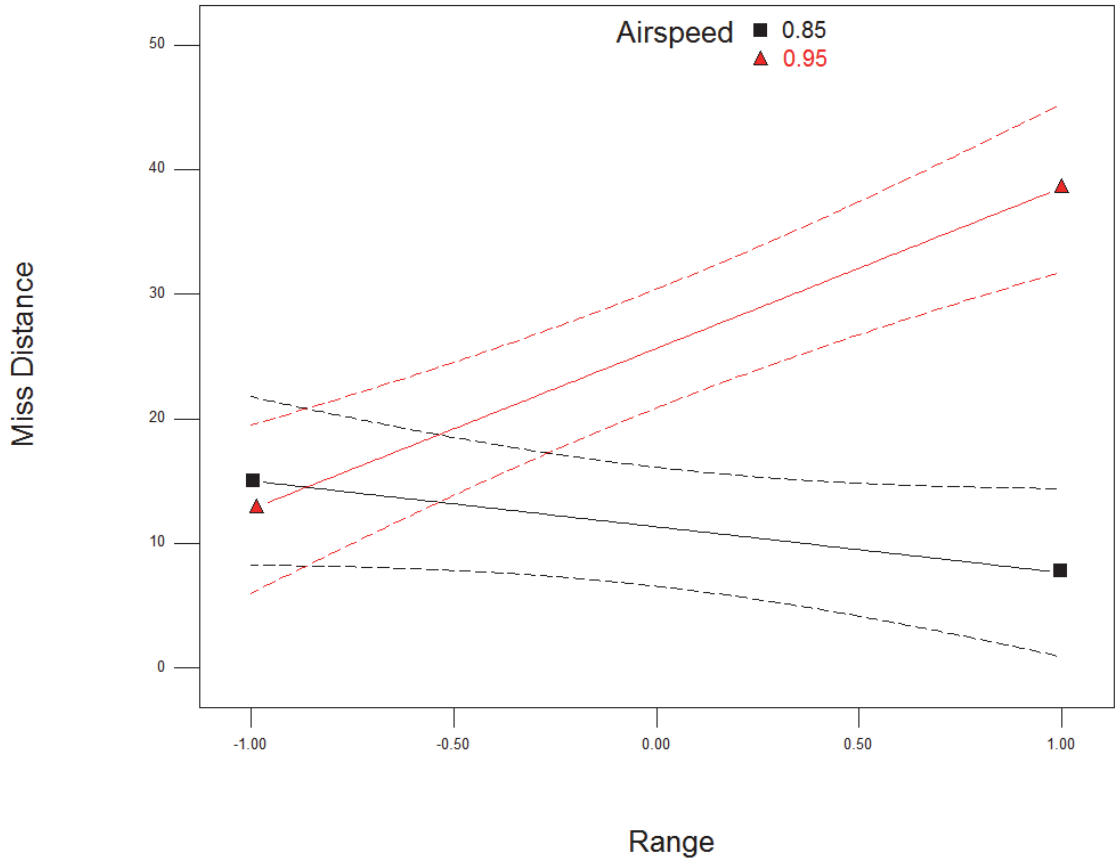
This chart shows us how significant a factor is on performance relative to the noise in the data

- Conclusion: Range, airspeed, and their interaction are the most important factors in characterizing performance for both the new and legacy air to ground missiles
- On average, there is no statistically distinguishable difference between the two variants across the operational envelope investigated in this test.
- Final Model:

$$\widehat{MissDistace} = -114.765 + 7.58 * Range + 143.125 * Airspeed + 165.675 * Range * Airspeed$$

Graphical Presentation of Results

- Plots provide meaningful insights
  - Miss distance increases with range at the higher airspeed



Summary of Example

- Several important conclusions that are important to the evaluation of the new system come from the regression analysis.
- First, the new system is not significantly better than or worse than the legacy system.
  - Therefore, if we were buying the system to improve on the accuracy of the legacy system, we would conclude the system performance is unacceptable.
  - However, if the new system was only required to replace the old system (maybe because of the loss of a manufacturing facility) without any decrement in performance, then this system meets the requirement of "no change" in terms of weapon accuracy.
- Next, we note that miss distance is generally low, except in the case of fast speed, long range shots.
  - This information could be used to refine the weapon's algorithms
  - Or this information could inform tactics (e.g., slow down before taking a shot if possible).

Summary

- This module introduced analysis, statistical models and their uses.
- You should now be able to:
  - Define statistical modeling and its key elements
  - Understand the analysis checklist
  - Understand pre-modeling and exploratory data analysis.
  - Understand that different types of statistical models exist
  - Understand model fitting and checking assumptions

Math Behind the Curtain

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_{12} + \varepsilon$$

Write the model equation as a matrix (one row for each run)

$$\mathbf{y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

"Design Matrix"

Number of observations

Model terms →

Number of runs

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{12} \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & (x_1 x_2)_1 \\ 1 & x_{1,2} & x_{2,2} & (x_1 x_2)_2 \\ 1 & x_{1,3} & x_{2,3} & (x_1 x_2)_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,12} & x_{2,12} & (x_1 x_2)_{12} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix}$$

- Typically put $x_{i,j}$ in "coded" units: i.e., the point in the design space where you make a measurement
  - Example, run number 1 was done at the (x_1 = +1, x_2 = +1) part of the DOE matrix

Linear Regression

- Goal is to find values of $\widehat{\boldsymbol{\beta}}$ = the least squares estimators of β

$$\mathbf{y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Minimize:

$$\boldsymbol{\varepsilon}' \cdot \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \cdot (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}' \cdot \boldsymbol{y}$$

- **Also need to calculate the "mean square error" = sum of the squares divided by degrees of freedom**

$$\boldsymbol{MSE} = \sigma^2 = \frac{Sum\ of\ squares}{dof} = \frac{\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y}}{(N - \#\mathrm{ModelTerms})}$$

DOE Estimates and Confidence Intervals

- Define what point in the test envelope you want the estimate of performance (mean value in a bin)

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_{12} \end{bmatrix}$$

- Using the regression model, mean response at that point is:

$$\mathbf{y}(\boldsymbol{x}_0) = \boldsymbol{x}_0{}' \cdot \widehat{\boldsymbol{\beta}}$$

- Variance for the estimate at that point is:

$$Var[\mathbf{y}(\boldsymbol{x}_0)] = \boldsymbol{MSE} \cdot (\boldsymbol{x}_0{}' \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{x}_0)$$

Calculation of Confidence Intervals

- Consider the pieces of the Variance:

If the design is a balanced factorial, this is diagonal;
all diagonal terms = 1/N

$$Var[\mathbf{y}(\boldsymbol{x}_0)] = \boldsymbol{MSE} \cdot (\boldsymbol{x}_0{}' \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{x}_0)$$

If the design if balanced factorial, this is just
the average variance across the space.

- Confidence Intervals:

$$\mathbf{y}(\boldsymbol{x}_0) \pm \boldsymbol{t}_{\boldsymbol{\alpha/2,N-p}} \cdot \sqrt{\boldsymbol{MSE} \cdot (\boldsymbol{x}_0{}' \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{x}_0)}$$