Continuous Metrics for Efficient and Effective Testing

Laura J. Freeman
&
Bram Lillard

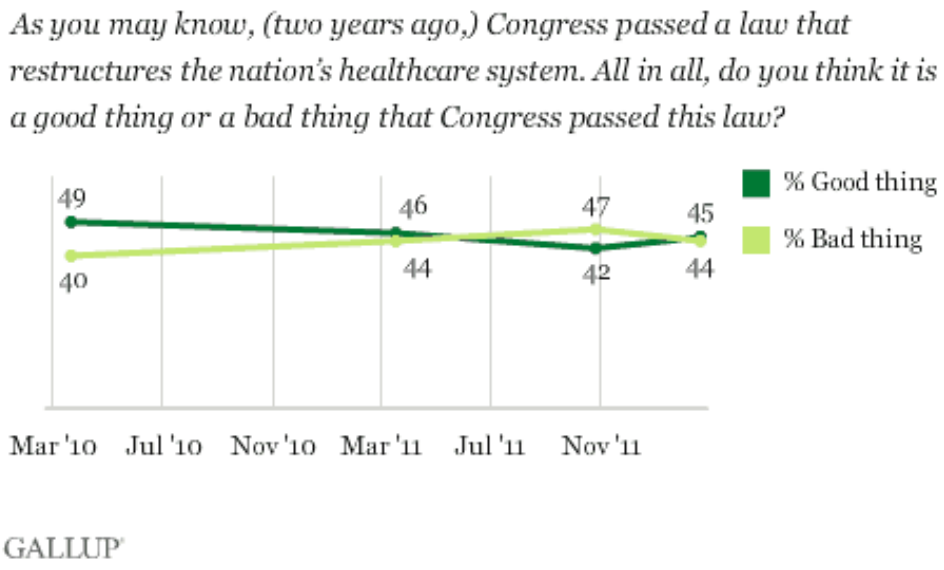
NDIA National Test and Evaluation Conference

March 15, 2012

IDA

Everyday Example

- Gallup Poll: “Americans Divided on Repeal of 2010 Healthcare Law…Americans divide evenly when asked if they favor (47%) or oppose (44%) a Republican president's repealing the 2010 healthcare law if elected this November.”



- Survey Methods: “a random sample of **1,040** adults, … For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is **4 percentage points**.

T&E cannot afford 1040 test points!

Outline

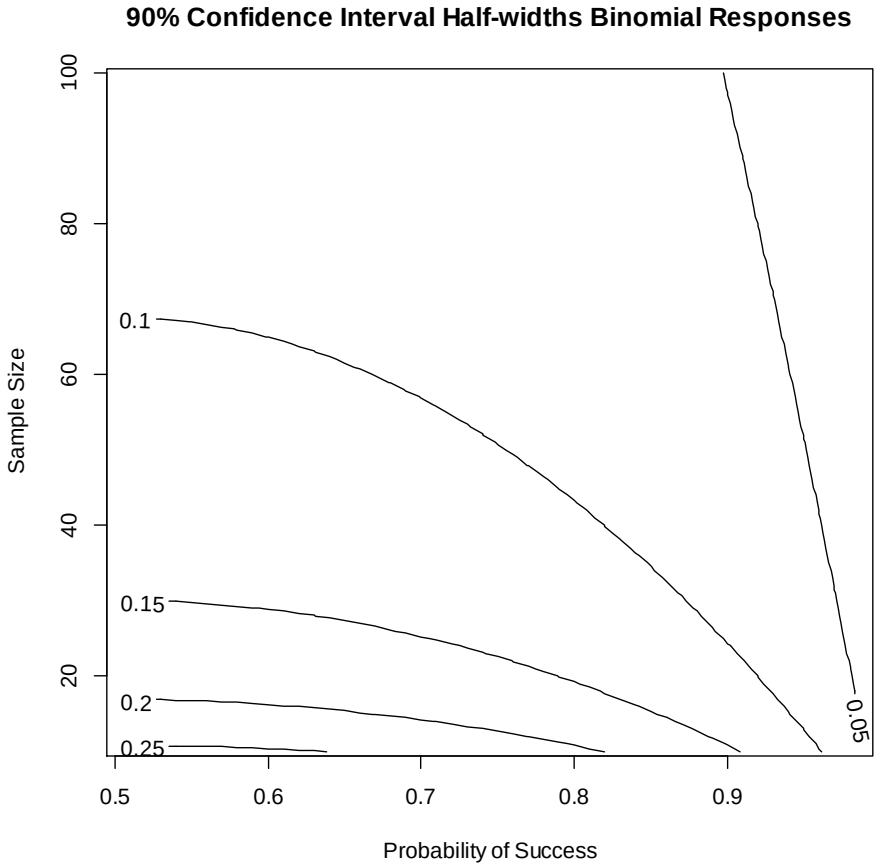
- **The binomial conundrum**
- **Continuous metrics: an informative test solution**
- **Efficient test examples**
  - Example 1: Chemical Agent Detector
    - » Verify a requirement within 10%
  - Example 2: Submarine Mine Detection
    - » Characterize performance drivers
- **Challenges**
- **Conclusions**

The Binomial Conundrum

- **Testing for a binary response requires large sample sizes**

Table: Sample Size Requirements

Sample Size	90% Confidence Interval Width (p = 0.5)	90% Confidence Interval Width (p = 0.8)
10	26%	21%
50	11.6%	9.3%
100	8.2%	6.6%
500	3.7%	2.9%



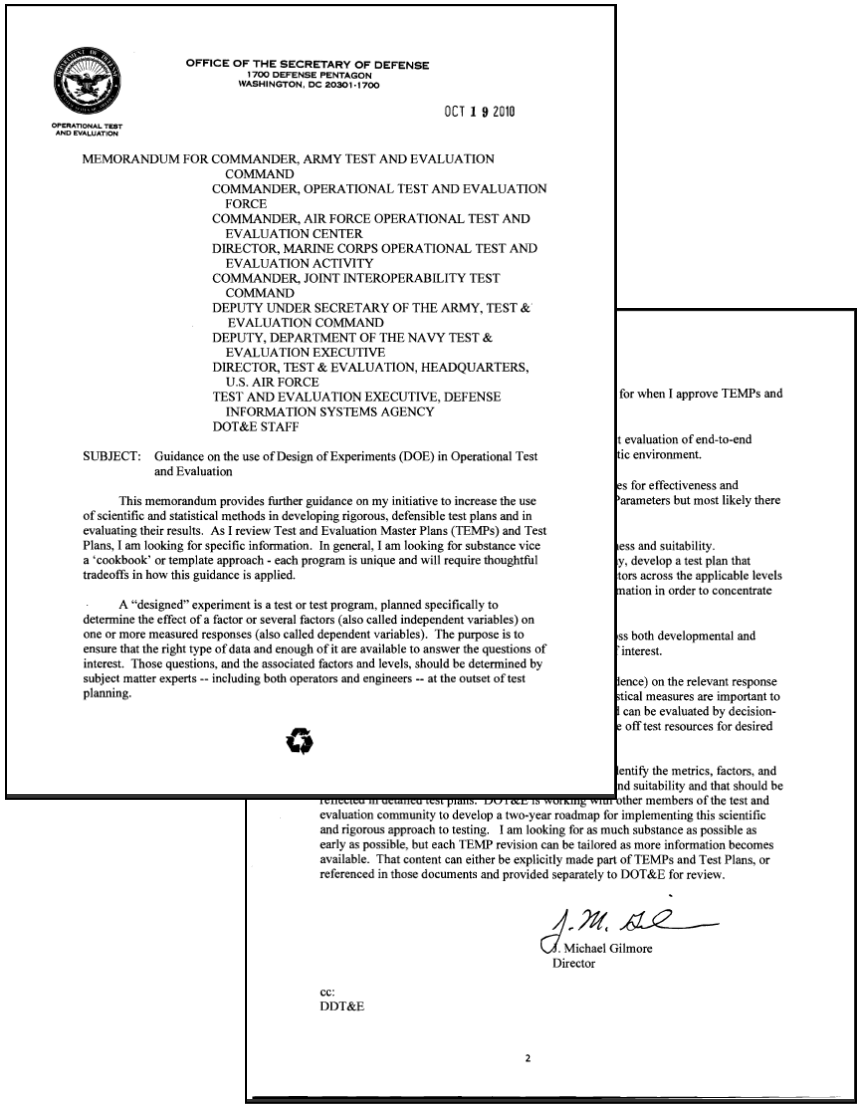
Continuous Metrics: An informative test solution

- **Chemical Agent Detector**
  - Requirement: Probability of detection greater than 85% within one minute
  - Original response metric: Detect/Non-detect
  - Replacement: Time until detection
- **Submarine Mine Detection**
  - Requirement: Probability of detection greater than 80% outside 200 meters
  - Original response metric: Detect/Non-detect
  - Replacement: Detection range
- **Missile System**
  - Requirement: Probability of hit at least 90%
  - Original response metric: Hit/Miss
  - Replacement: Missile miss distance

Continuous surrogate metrics provide additional information!

DOT&E Guidance

Dr. Gilmore's October 19, 2010 Memo to OTAs



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

OPERATIONAL TEST AND EVALUATION

OCT 19 2010

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY
DOT&E STAFF

SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.

for when I approve TEMPs and

t evaluation of end-to-end
tic environment.

es for effectiveness and
Parameters but most likely there

ess and suitability.
y, develop a test plan that
tors across the applicable levels
mation in order to concentrate

ss both developmental and
interest.

dence) on the relevant response
stical measures are important to
d can be evaluated by decision-
e off test resources for desired

dentify the metrics, factors, and
nd suitability and that should be
reflected in detailed test plans. DOT&E is working with other members of the test and evaluation community to develop a two-year roadmap for implementing this scientific and rigorous approach to testing. I am looking for as much substance as possible as early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

J. Michael Gilmore
Director

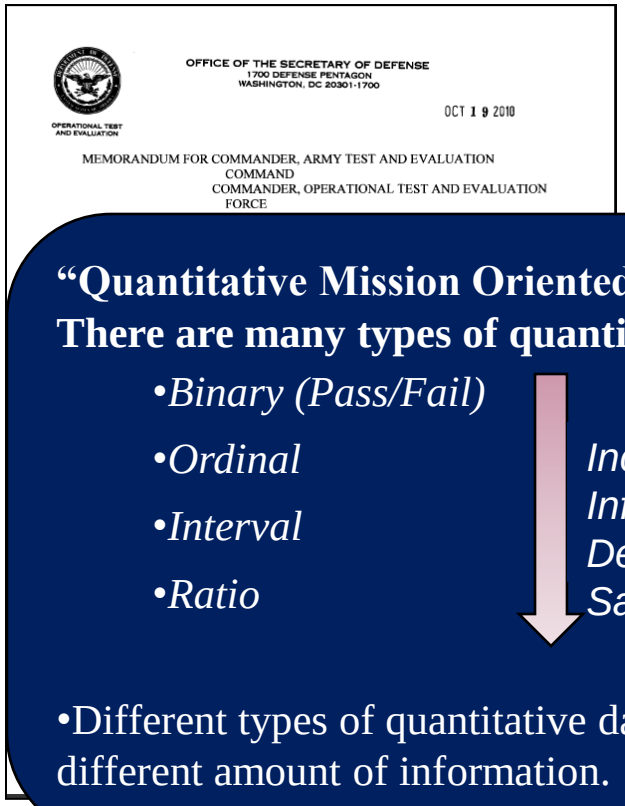
cc:
DDT&E

2

- **The goal of the experiment**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)
- **Factors** that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, 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. These statistical measures are important to understand "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.

DOT&E Guidance

Dr. Gilmore's October 19, 2010 Memo to OTAs



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

OPERATIONAL TEST AND EVALUATION

OCT 19 2010

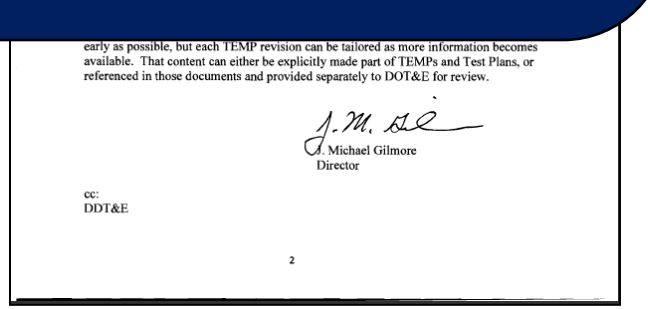
MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE

"Quantitative Mission Oriented Metrics"
There are many types of quantitative data:

- *Binary (Pass/Fail)*
- *Ordinal*
- *Interval*
- *Ratio*

Increasing Information: Decreasing Sample Size

- Different types of quantitative data contain a different amount of information.



early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

J. Michael Gilmore
Director

cc:
DDT&E

2

- **The goal of the experiment**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)
- **Factors** that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, 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. These statistical measures are important to understand "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.

Example 1: Chemical Agent Detector

- **Goal: Determine the probability of detection within one minute**
  - Threshold is least 85% within one minute
- **Metric (response variables) :**
  - Detect (Yes/No)
  - Detection time (seconds)
- **Factors to consider:**
  - Temperature, water vapor concentration, agent concentration, agent type
- **Notional test design: Full factorial (2^4)**

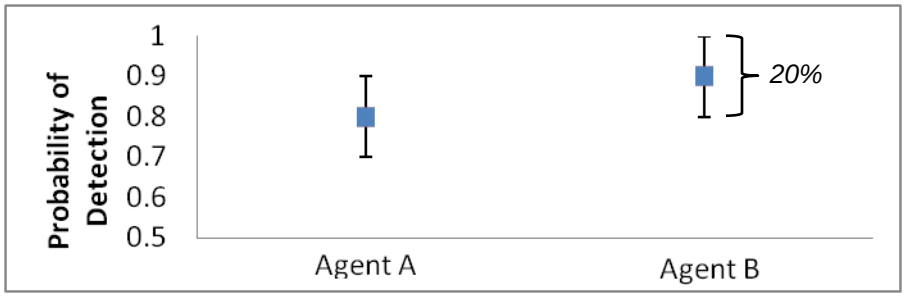
DOE Matrix

Agent Type	Agent Concentration	Low Temperature		High Temperature		Agent Type	Agent Concentration	Low Temperature		High Temperature	
		Low WVC	High WVC	Low WVC	High WVC			Low WVC	High WVC	Low WVC	High WVC
A	Low	?	?	?	?	B	Low	?	?	?	?
	High	?	?	?	?		High	?	?	?	?

What sample size is do we need to determine probability of detection?

Adequate Test Resources

- **Goal: Determine an adequate sample size to determine a 10% change in probability of detection across all factor levels (across the operational envelope)?**

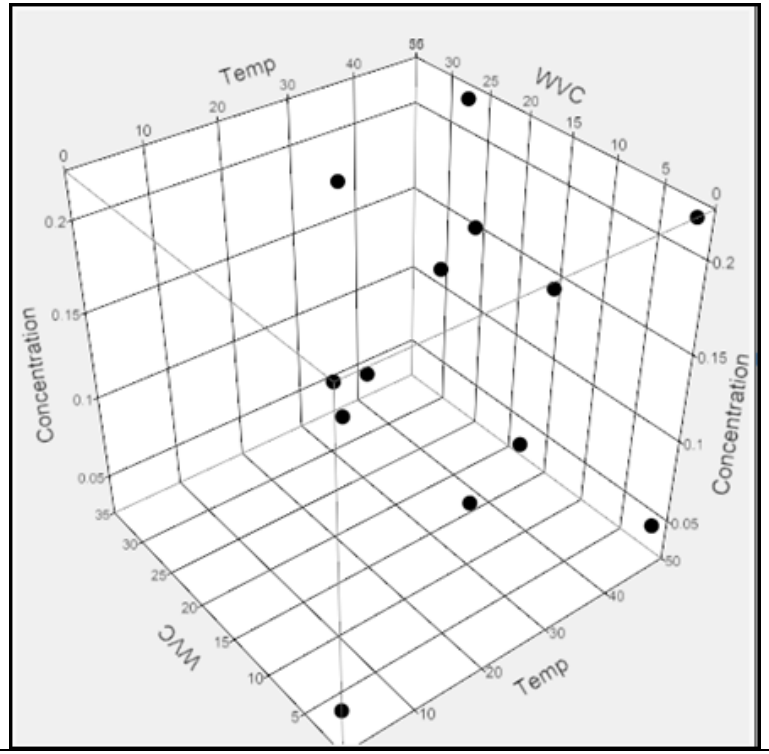


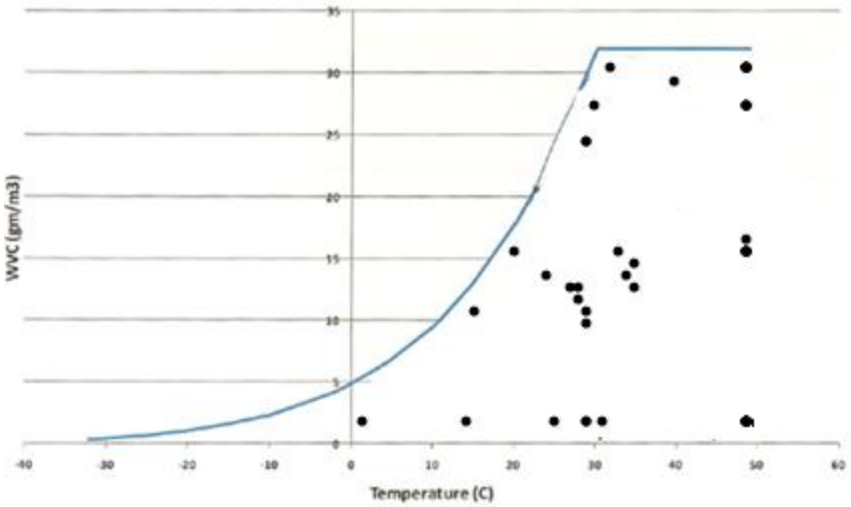
- **Steps**
  - Determine detectable difference for binary response (10%)
  - Calculate sample size for binary response variable
  - Determine the appropriate continuous response (detection time)
  - Determine equivalent effect size of interest using percentiles of appropriate continuous response distribution (e.g. lognormal)
  - Calculate sample size for continuous response variable & compare
- **Results**
  - Detectable difference = 10%
  - 90% Confidence Level, 80% Power
    - » Binomial response (detect/non-detect): 14 replications of full factorial (224 total test points)
    - » Continuous response (time until detection): 5 replications of full factorial (80 total test points) – 65% reduction in test costs!

This example results in a 65% reduction in test cost!

Chemical Agent Detector: Test Design

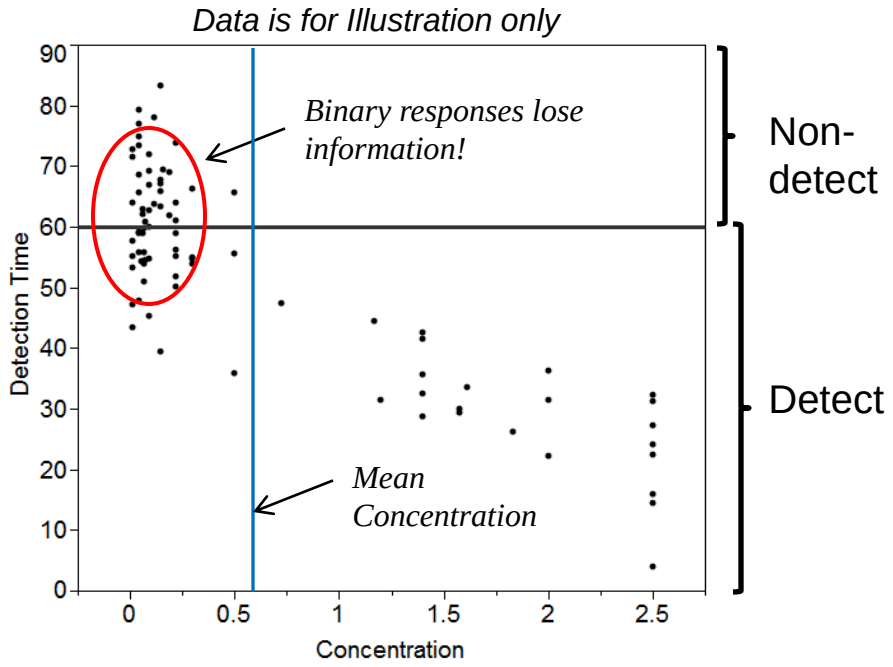
- **Design from Joint Chemical Agent Detector**
  - Employed an optimal design methodology
  - Responses times are hypothetical
  - What is the implication in test analysis?





Chemical Agent Detector Results

- Estimate the probability of detection at 60 seconds at the mean concentration
- Detection times and detect/non-detect information recorded
- Binary analysis results in **400% increase** in confidence interval width



Response	Probability of Detection within 60 seconds at mean	Lower 90% Confidence Bound	Upper 90% Confidence Bound	Confidence Interval Width
Binary (Detect: Yes/No)	83.5%	60.5%	94.4%	33.9%
Continuous (Time)	91.0%	86.3%	94.5%	8.2%

Example 2: Submarine Mine Detection

- **Goal: Characterize performance (detection ability) across the operational envelope**
  - Threshold probability of detection is 80%
- **Metric (response variables) :**
  - Detect (Yes/No)
  - Detection range (meters)
- **Factors to consider:**
  - Mine type, pulse type, array type
- **Notional test design: General Factorial**

DOE Matrix									
Mine Type	Pulse Type 1		Pulse Type 2		Mine Type	Pulse Type 1		Pulse Type 2	
	Array 1	Array 2	Array 1	Array 2		Array 1	Array 2	Array 1	Array 2
A	?	?	?	?	B	?	?	?	?

What sample size is do we need to characterize performance?

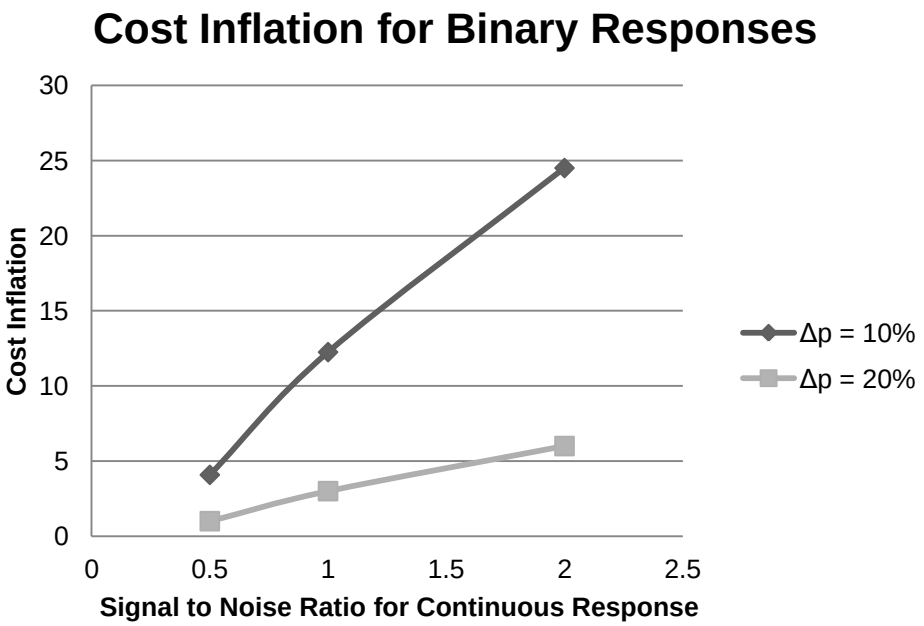
Adequate Test Resources

- **Determine an adequate sample size to characterize the systems ability to detect mines across the operational envelope.**
  - For example, how sensitive is the submarines detection ability to the type of sonar array? Does the submarines ability to detect mines vary based on the mine type?

- **Power Analysis**
  - 90% Confidence Level, 80% Power to detect factor effects

Detection Range (Continuous Response)		
Signal to Noise Ratio	Factorial Replicates	Total Detection Opportunities
0.5	12	96
1	4	32
2	2	16

Detect? Yes/No (Binary Response)		
Δp	Factorial Replicates	Total Detection Opportunities
10%	49	**392**
20%	12	96



Challenges

- **Accounting for non-detects**
  - Advanced statistical methods provide potential solutions
    - » Censored data analysis for unobservable non-detects
    - » Mixture distributions
- **Can require high fidelity instrumentation during data collection process**
  - For example , the ability to measure miss distance in operational testing
- **Pass/Fail may be a function of multiple (possibly correlated) continuous variables**
  - Advanced statistical methods provide potential solutions:
    - » Multivariate analyses
    - » Copulas, similar to the financial markets

Cost saving potential is to great to not tackle these challenges!

Conclusions

- **Most binary metrics can be recast using a continuous metrics**
- **Continuous metrics lead to more detailed insight than binary metrics**
  - Provides useful information to the evaluator and the warfighter
- **Converting to a continuous metric from a binary response metric maximizes test efficiency**
  - Conservatively, approximately 50% reduction in test costs for near identical results in percentile estimates
  - "Result in a reduction in statistical power equivalent to discarding 38% - 60% of the cases"
    - » Cohen, J. *The Cost of Dichotomization*
    - » Hamada, M. *The Advantages of Continuous Measures Over Pass/Fail Data*
  - Cost savings are much larger if the goal is to identify significant factors

Backup Material

Types of Data

- **Discrete**
  - Categorical:
    - » Nominal (or categorical) data consist of discrete labels, names or categories only. No ordering information (high-low, best-worst) is available. Examples include names, colors, vendors, and scenario names. Numeric values assigned to nominal data are meaningless.
  - Ordinal:
    - » Ordinal data are typically discrete values that imply some ordering relationship is possible, but lack information about the width of the intervals separating the values. Examples include rankings, place order in races, letter grades, and preference levels (best to worst). Numbers assigned to ordinal data values preserve order, but uneven intervals may pose problems in calculating averages and the like. The binary success/failure response is another example of ordinal data (assuming success is better than failure.)
- **Continuous**
  - Interval
    - » Interval data are measured on a continuous measurement scale such that the width of the interval between any two values can be determined, but the origin (zero) point of the scale is arbitrary. Examples include temperature, years, and possibly Likert scales in questionnaire responses. Differences of intervals are meaningful, but ratios of interval data are usually meaningless.
  - Ratio
    - » Ratio level data are the richest level of measurement comprising order, interval, and a true zero point. Most real physical values are ratio scales including length, weight, time, speed, target signatures, power levels, light levels, etc. All mathematical operations are meaningful on ratio data.
- **Definitions copied from Statistical T&E Glossary currently in final revisions for addition to DAU Glossary**