Censored Data Analysis:

A Statistical Tool for Efficient and Information-Rich Testing

V. Bram Lillard

A Follow-on to….

Continuous Metrics for Efficient and Effective Testing

Laura J. Freeman
&
Bram Lillard

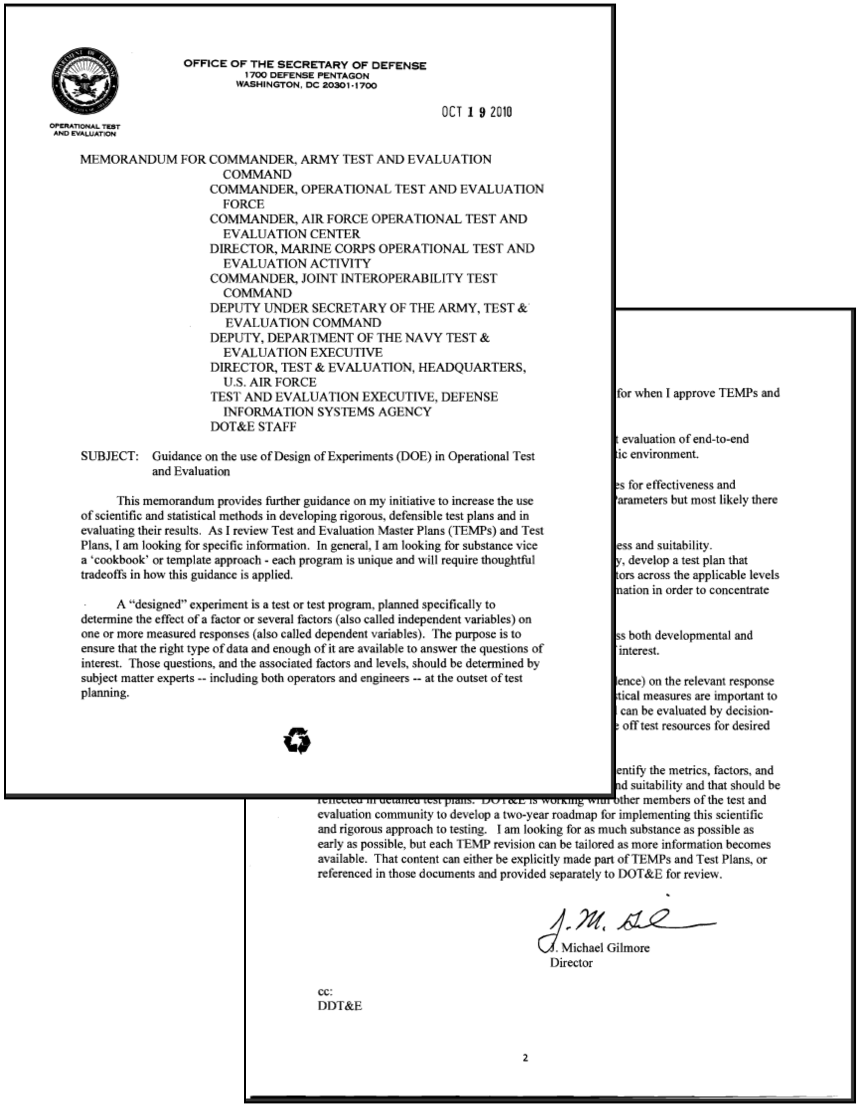
NDIA National Test and Evaluation Conference

March 15, 2012

IDA

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 1 9 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 ...f interest.

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

...entify 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

- **<u>The goal of the experiment</u>**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **<u>response variables</u>** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)
- **<u>Factors</u>** 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.
- **<u>A method for strategically varying factors</u>** across both developmental and operational testing with respect to responses of interest.
- **<u>Statistical measures of merit</u>** <u>(power and confidence)</u> 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 1 9 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.

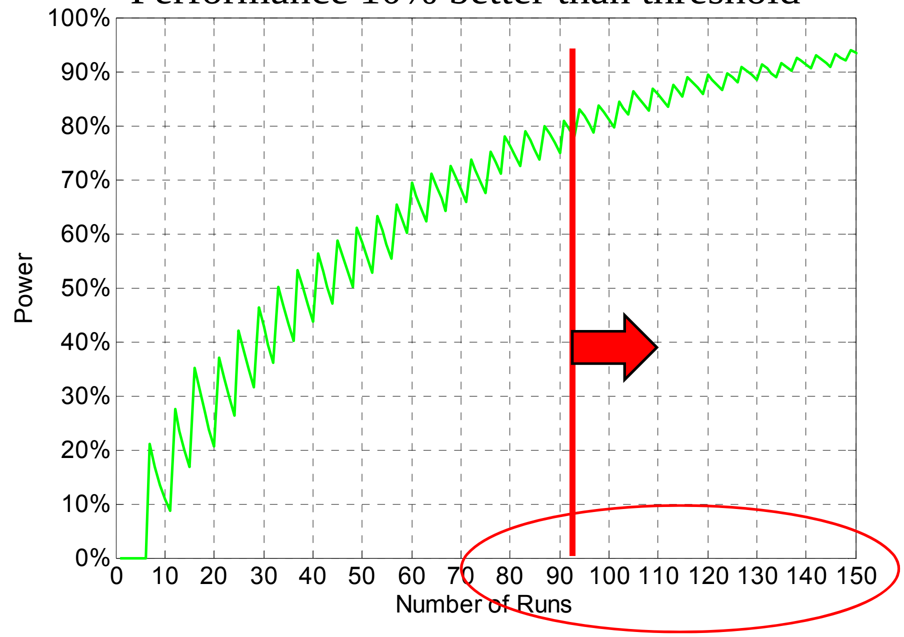
The Binomial Conundrum

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

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%

Power Calculation, 90% confidence, Performance 10% better than threshold

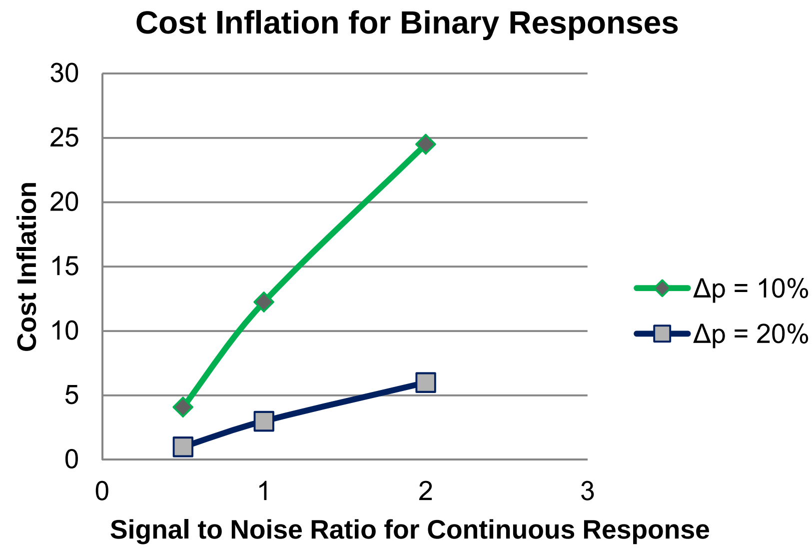


- **Difficult (impossible?) to achieve acceptable power for factor analysis unless many runs *(often >100)* can be resourced**
  - Non-starter for implementing DOE concepts (characterizing performance across multiple conditions)

Solutions

- **Recast Binomial metric (e.g., probability of detection) as a *continuous metric* (e.g., time-to-detect)**
  - Others: detection range, miss distance

- **Significant cost savings realized, plus the continuous metric provides useful information to the evaluator/warfighter**



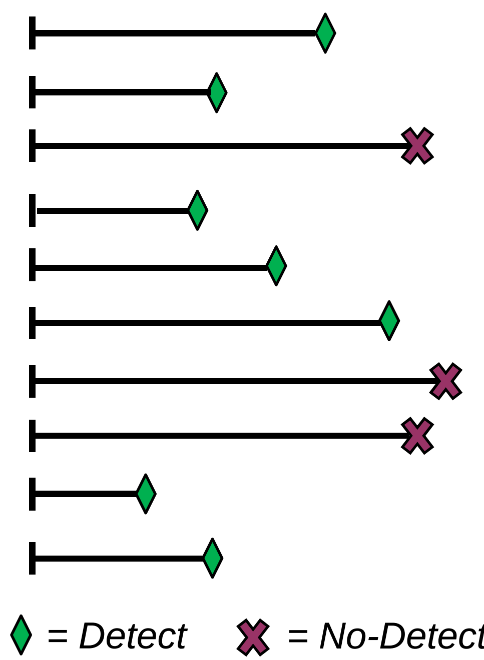
- **Challenges:**
  - How to handle ***non-detects***/misses?
    - » Typical DOE methods (linear regression) require an actual measurement of the variable for every event
    - » Can not force the test to get detection ranges – non-detects are important test results!
  - Common concern: Switching to the continuous measure seems to eliminate the ability to evaluate the requirement
    - » E.g., we measured time-to-detect and calculated a mean, how do we determine if the system met it's KPP: P_{detect}>0.50?)

Using Continuous Data (with non-detects)

- **Censored data = we didn't observe the detection directly, but we expect it will occur if the test had continued**
  - We cannot make an exact measurement, but there is information we can use!
  - Same concept as a time-terminated reliability trials (failure data)

Run No.	Result	Result Code
1	Detected Target	1
2	Detected Target	1
3	No detect	0
4	Detected Target	1
5	Detected Target	1
6	Detected Target	1
7	No detect	0
8	No detect	0
9	Detected Target	1
10	Detected Target	1

Corresponding Timelines

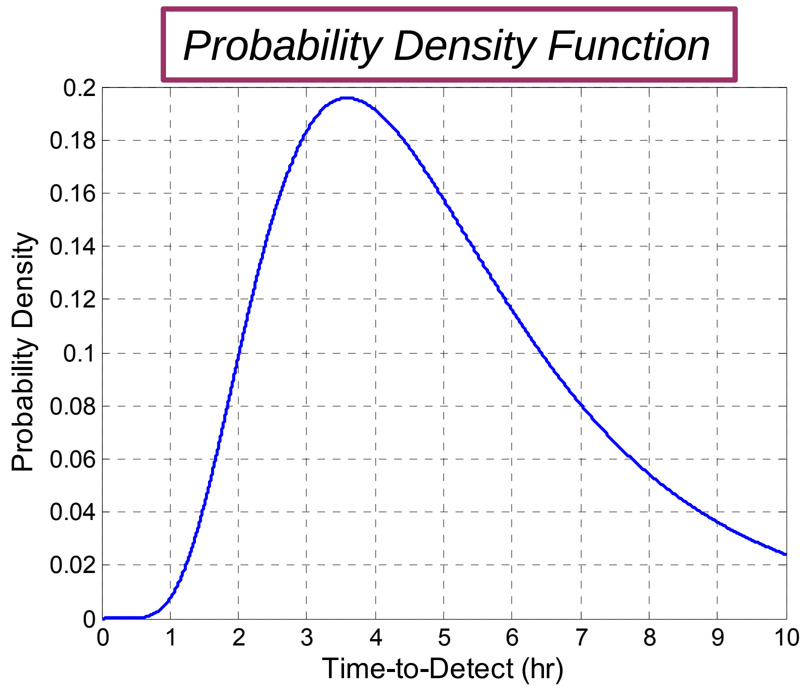


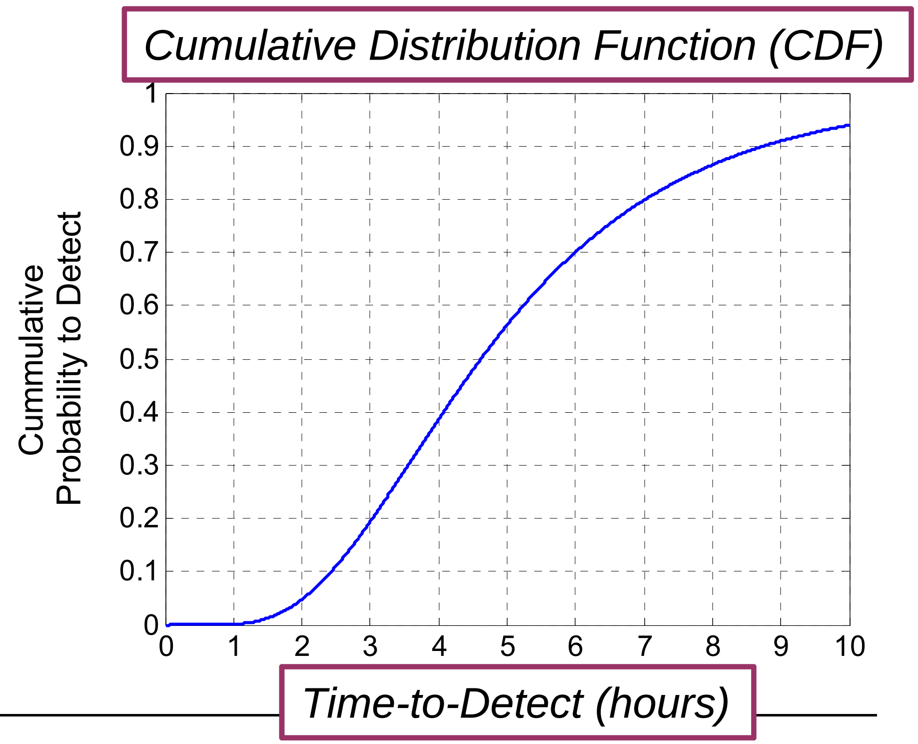
Run No.	Time of Detection (hours after COMEX)
1	4.4
2	2.7
3	>6.1
4	2.5
5	3.5
6	5.3
7	>6.2
8	>5.8
9	1.8
10	2.7

Parameterizing Data

- **Assume that the time data come from an underlying distribution, such as the log-normal distribution**
  - Other distributions may apply – *must consider carefully*, and check the assumption when data are analyzed (may have to find a better parameterization, or revert to binomial)

- **That parameterization will enable us to link the time metric to the probability of detection metric.**



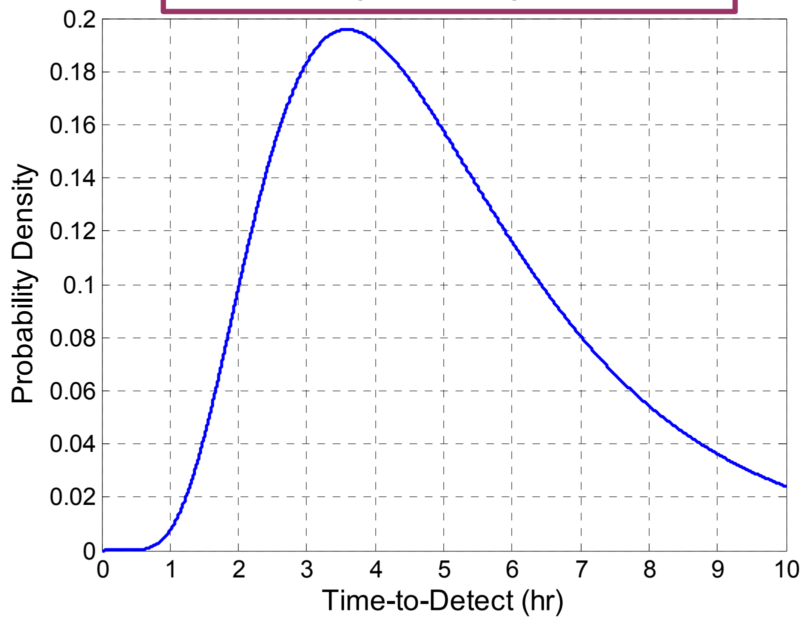


Parameterizing Data

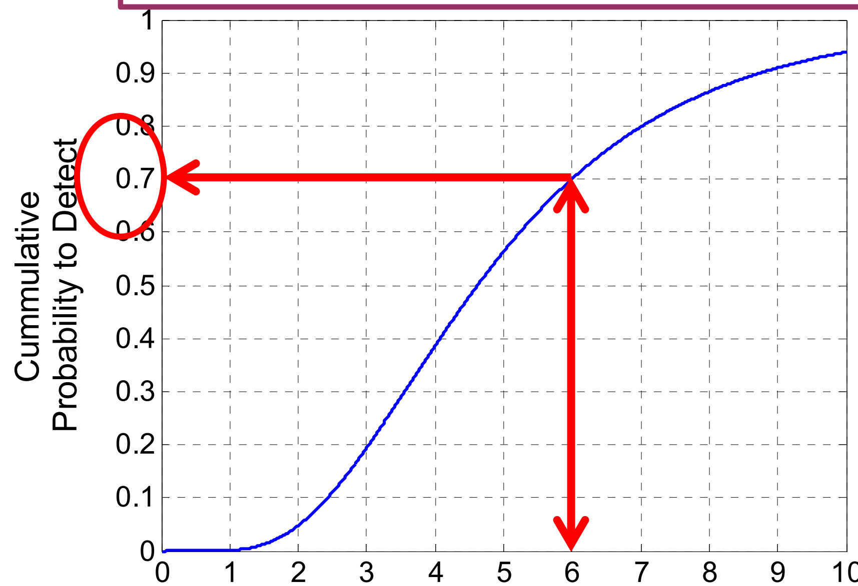
- **Example: Aircraft must detect the target within it's nominal time on station (6-hours)**
  - Binomial metric was detect/non-detect within time-on-station

- **If we determine the shape of this curve (i.e., determine the parameters of the PDF/CDF), we can use the time metric to determine the probability to detect!**

Probability Density Function



Cumulative Distribution Function (CDF)



Time-to-Detect (hours)

New Goal

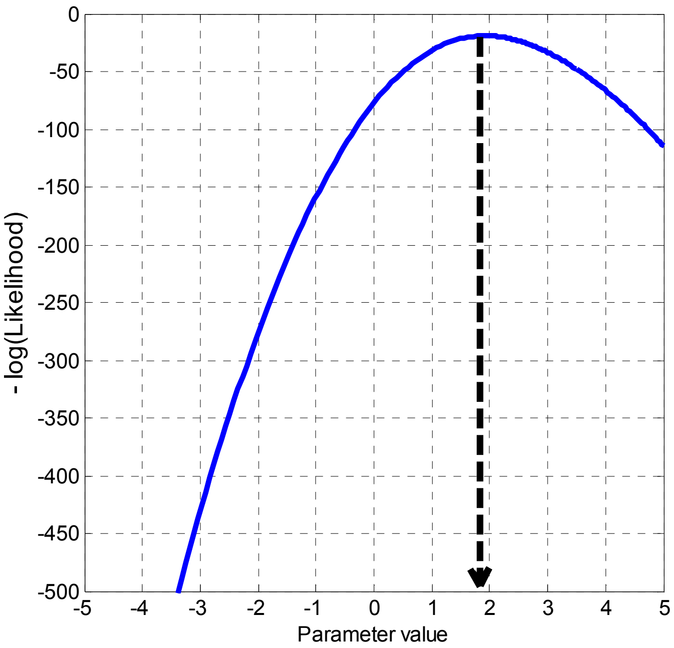
- **Goal of our data analysis: determine the parameters of the distribution**
  - Once the CDF's shape is known, can translate back to the binomial metric (probability to detect)

- **Most common and generalized technique for determining the parameters is via *maximum likelihood methodology***
  - **A Likelihood is simply a function** that defines how "likely" a particular value for a parameter is given the specific data we've observed

parameters TBD — *data*

$$L = \prod_{i=1}^{\#\,data\,points} f(\boldsymbol{\mu} | t_i)$$

- **Aside: these techniques are not difficult!**
  - **JMP** has these functions and maximization code built in (2 button clicks)
  - **R**
    - » Built-in PDFs/likelihoods and easy to write your own
  - **Matlab**
    - » Built-in PDFs/likelihoods and easy to write your own



For multiple parameters, imagine a multi-dimentional surface to be maximized

Likelihood for Censored Data

- **We construct our Likelihood function based on the desire to use censored data:**

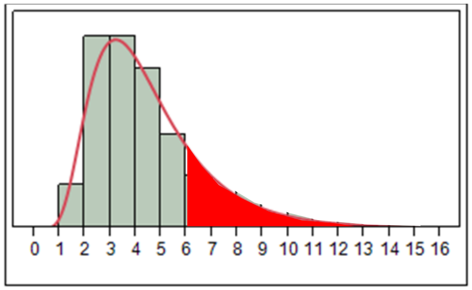
parameters TBD *data* *parameters TBD* *data*

$$L = \prod_{i=1}^{\#\,data\,points} [PDF(\boldsymbol{\mu},\boldsymbol{\sigma}|t_i)]^{(1-\delta_i)} \times [1 - CDF(\boldsymbol{\mu},\boldsymbol{\sigma}|t_i)]^{\delta_i}$$

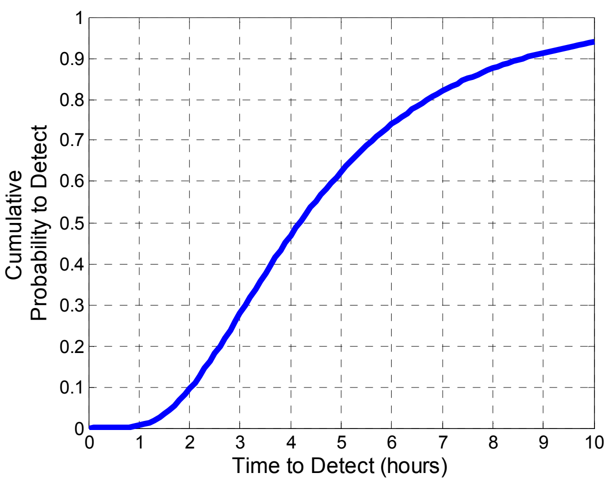
Non-Censored data (δ_i = 0) provide information to define the shape of the PDF!

Censored data (δ_i = 1) provide information to define the shape of the CDF!

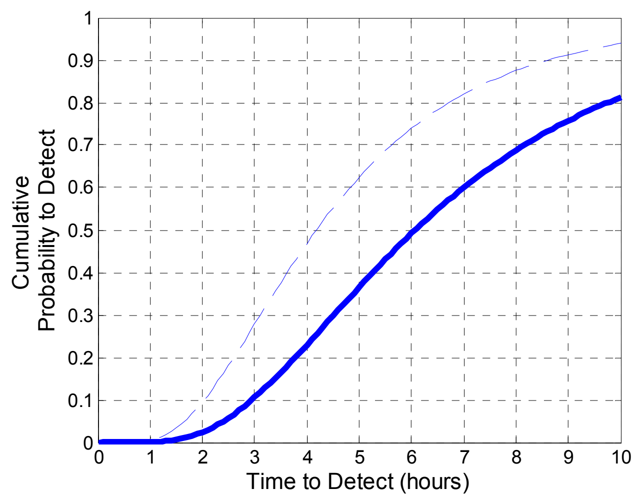
Conceptualizing the Censored-Data Fit



- **For non-censored measurements, the PDF fit is easy to conceptualize**
- **For censored measurements, the data can't define the PDF, but we know they contribute to the probability density beyond the censor point**
- **Example event from an OT: Time > 6 hours – that data point cannot increase the probability to the left of t=6.0 in the CDF!**
  - Detect will occur at some time in the future, so it must contribute to the probability beyond t=6.0



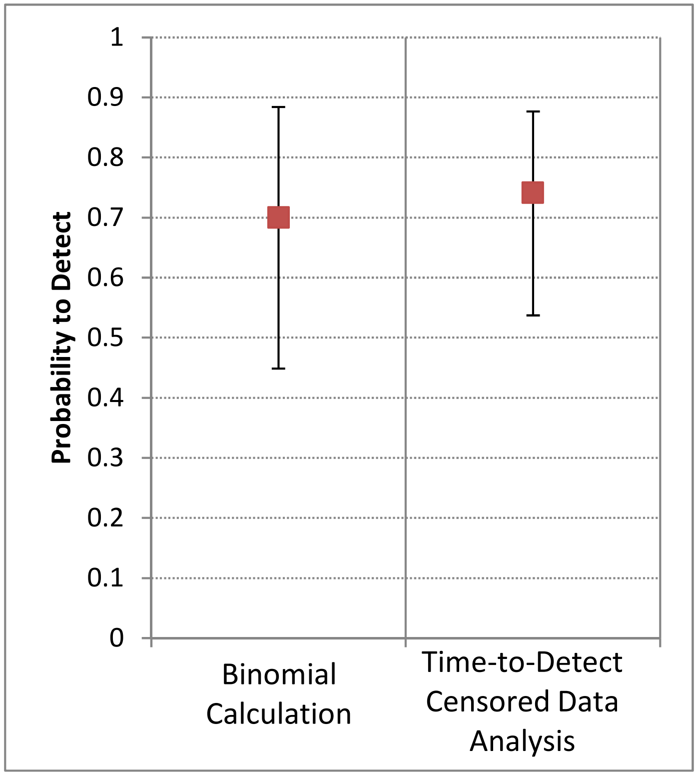
Including a bunch of censored (Time > 6 hour) events will push the CDF to the right (see how probability to detect is lower at 6 hours)

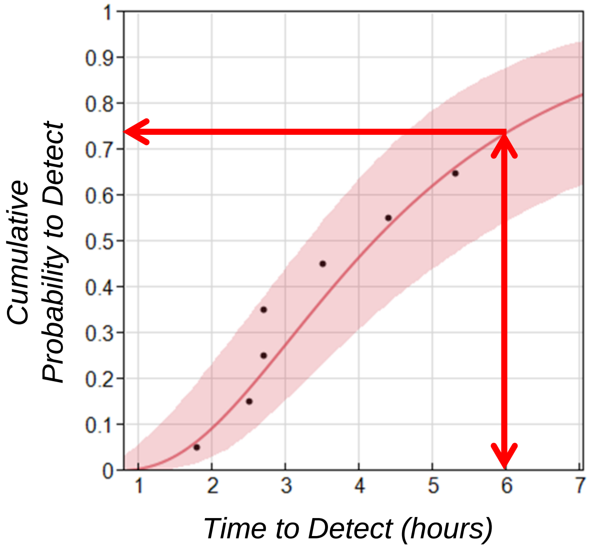


IDA

Simplest Example

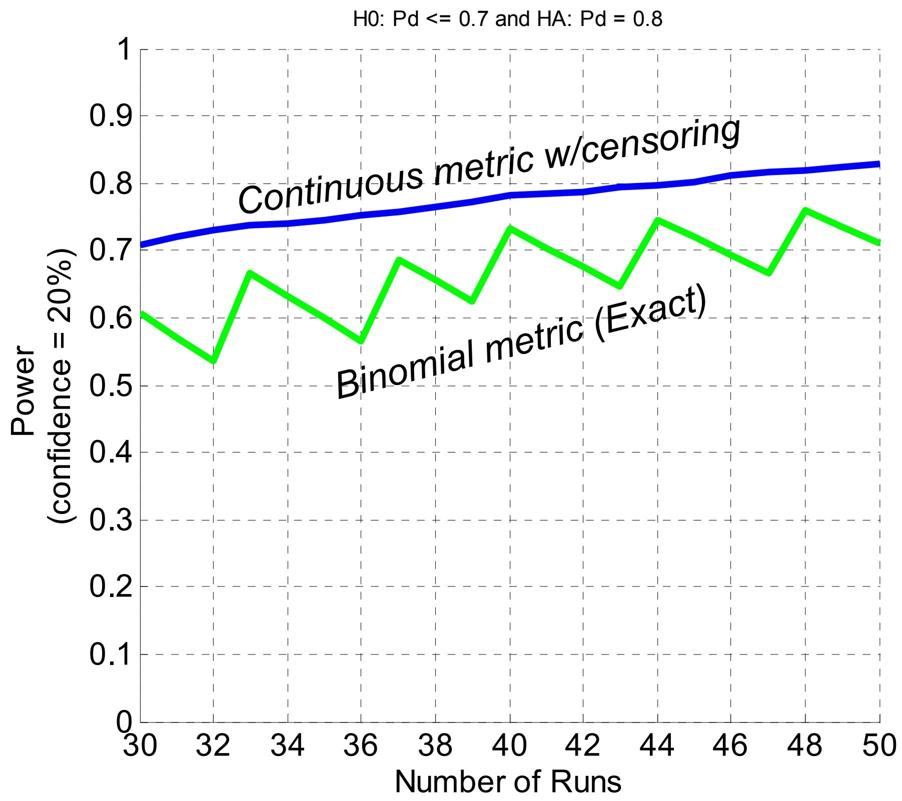
- **Consider data from slide 7…..**
- **With only 10 data points, the censored data approach provides smaller confidence intervals**
  - 16% reduction in interval size
  - Better estimate of the probability to detect
- **More confident system is meeting requirements, but with same amount of data**





	Binomial Probability Calculation	Time-to-Detect Censored Data Analysis
Confidence Threshold $P_{detect} > 0.5$ is met	82%	93%

Sizing the Test
(Confirming Threshold Performance)



Total Sample Size required to detect 10% improvement over threshold with 80% confidence, 80% power

Threshold Requirement	Binomial metric	Continuous metric w/censoring
80%	39	26
70%	55	43
60%	70	56
50%	77	63

20-30% reduction in test size

Benefits are greater for higher threshold requirements (most common in requirements documents)

Characterizing Performance

- **Now let's employ DOE…**

- **Consider a test with 16 runs**
  – **<u>Two</u>** factors examined in the test
  – Run Matrix:

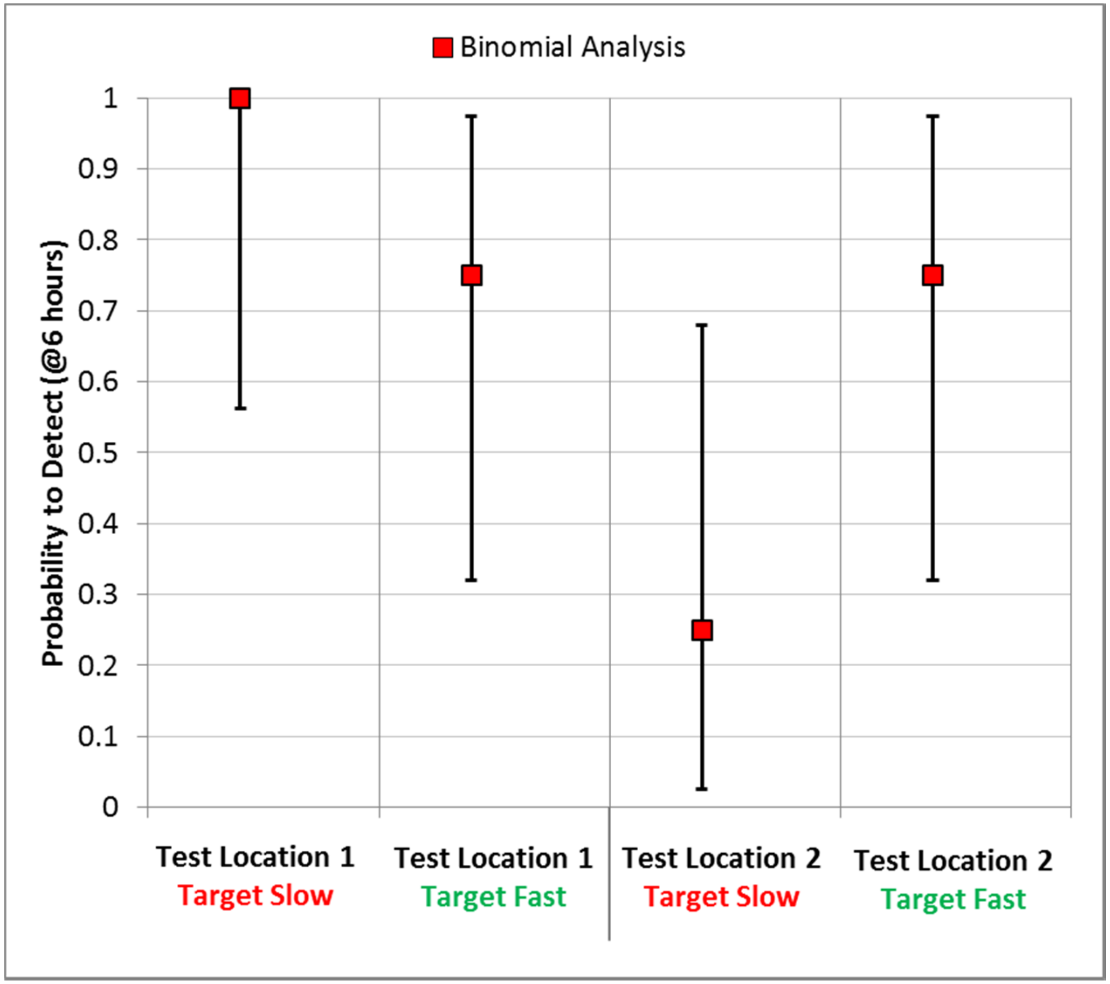
	Target Fast	Target Slow	Totals
Test Location 1	4	4	8
Test Location 2	4	4	8
	8	8	**16**

  – Detection Results:

	Target Fast	Target Slow	Totals
Test Location 1	3/4	4/4	**7/8 (0.875)**
Test Location 2	3/4	1/4	**4/8 (0.5)**
	6/8 (0.75)	**5/8 (0.63)**	

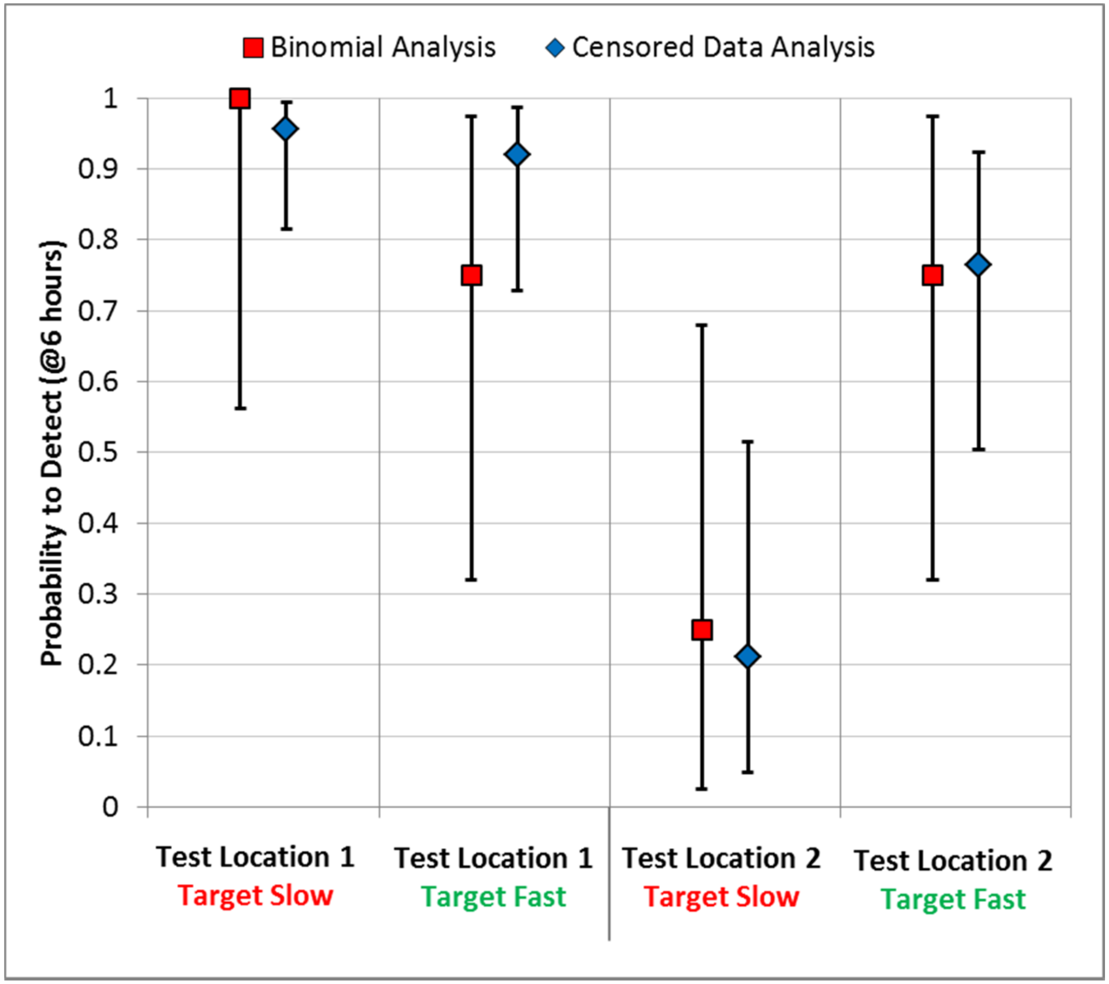
Attempt to Characterize Performance

- As expected, 4 runs in each condition is *insufficient* to characterize performance with a binomial metric
- Cannot tell which factor drives performance or which conditions will cause the system to meet/fail requirements
- Likely will only report a ‘roll-up’ of 11/16
  - 90% confidence interval: [0.45, 0.87]



Characterizing Performance Better

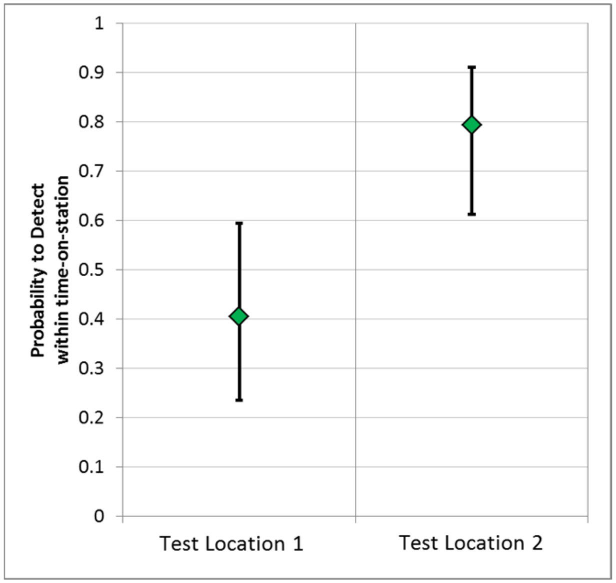
- **Measure *time-to-detect* in lieu of binomial metric, employ censored data analysis…**
- **Significant reduction in confidence intervals!**
  - Now can tell significant differences in performance
    - » E.g., system is performing **poorly** in Location 2 against slow targets
  - We can confidently conclude performance is above threshold in three conditions
    - » Not possible with a "probability to detect" analysis!



Sizing Tests

- **Why size a test based on ability to detect differences in P_{detect}?**
  - This is standard way to employ power calculations to detect factor effects in DOE methodology

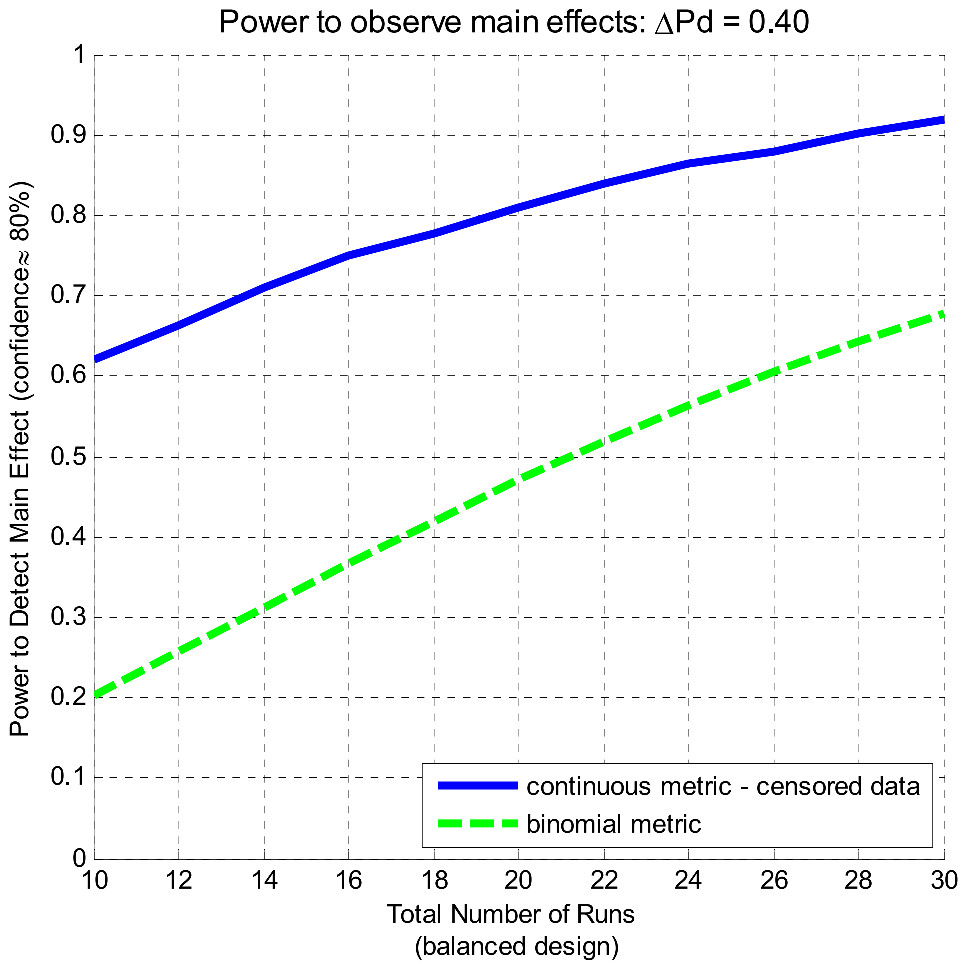
  - We <u>are</u> interested in performance differences – this is how we *characterize performance* across the operational envelope
  - This is also how we ensure a level of precision occurs in our measurement of P_{detect} (size of the “error bars” will be determined)



If we size the test to detect this difference, then the confidence intervals on the results will be approx. this big

If the measured delta is different than assumed, still ensure a level of accuracy in the measurement

Sizing Tests



Total Sample Size required to detect Factor Effects with 90% confidence, 80% power

ΔP detectable	Binomial metric	Continuous metric w/censoring
40%	44	24
30%	74	38
20%	166	98

40-50% reduction in test size

IDA

How to Calculate Power

- **No closed form equation to determine in this case**
- **Standard method when no closed-form exists is to conduct a Monte Carlo**
- **Method:**
  - Establish the parameters (μ and σ) under the null hypothesis (e.g., $P_{detect} \leq 0.50$)
  - Establish the parameter to be tested (μ in this case) under the alternate hypothesis
    - » Assume some effect size of interest for probability-to-detect; this equates to a shift in μ
  - Simulate data under the alternate hypothesis
    - » For times that occur beyond the nominal event duration (e.g., 6-hour on-station time), the censor value is set to "1."
  - Conduct the analysis on the simulated dataset
    - » i.e., MLE determines fitted values of μ and σ
  - Determine the standard errors (or confidence intervals) for the parameters (and P_{detect}). Based on the standard errors and the selected alpha (1 – confidence) value chosen, determine if the fitted P_{detect} value is statistically different than the null hypothesis P_{detect} value
    - » If so, it's a "correct rejection" of the null
  - Repeat the above steps 10,000 times.
  - Power equals the fraction of correct rejections
- **Note that Type 1 Error does not necessarily equal the alpha value you chose! Must check when doing power calculations….**
  - For censored data analyses, type 1 error (chance of wrongly rejecting null when it's true) is higher than alpha when:
    - » Small data sets
    - » High censoring

Conclusions

- **Many binary metrics can be recast using a continuous metrics**
  - Care is needed, does not always work, but…
  - Cost saving potential is too great not to consider it!
- **With Censored-data analysis methods, we retain the binary information (non-detects), but gain the benefits of using a continuous metric**
  - Better information for the warfighter
  - Maintains a link to the "Probability of…" requirements
- **Converting to the censored-continuous metric maximizes test efficiency**
  - In some cases, as much as 50% reduction in test costs for near identical results in percentile estimates
  - Benefit is greatest when the goal is to identify significant factors (characterize performance)