Empirical Signal-to-Noise Ratios from Operational Test Data

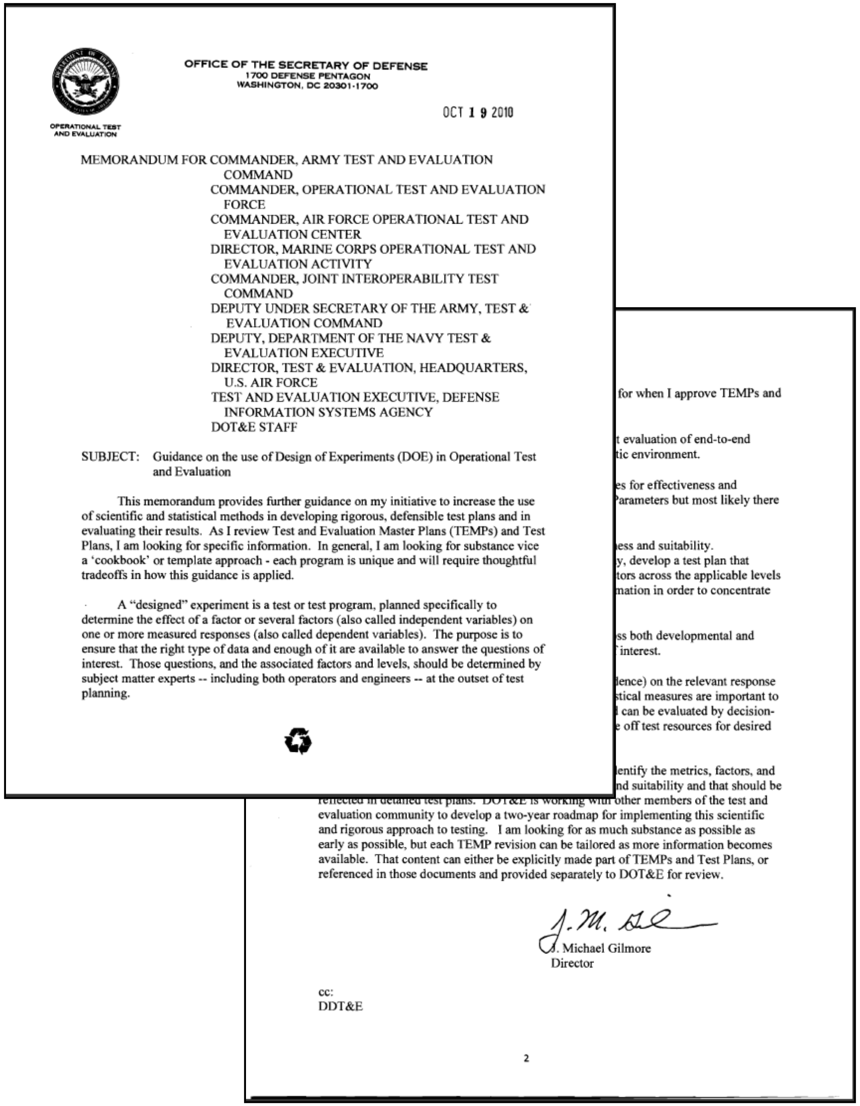
Dr. Matthew R Avery, Institute for Defense Analyses

Outline

- Using signal-to-noise ratios for operational test planning
- Signal-to-noise ratios for binary responses
- Summary of results
- Case Study: KC-46A
- Recommendations & next steps

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

ness 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
l can be evaluated by decision-
e off test resources for desired

lentify 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

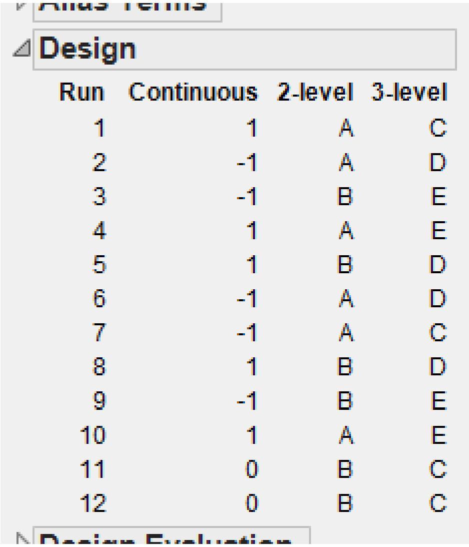
- ❑ **<u>The goal of the experiment</u>**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- ❑ Quantitative mission-oriented **<u>response variables</u>** for effectiveness and suitability. (These could be Key Performance Parameters but most 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 understanding "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.

Signal-to-noise Ratios

- **DOT&E requires power analysis to justify test size/duration for all operational tests**
  - JMP and Design Expert are common tools
    - » Both require Signal-to-Noise Ratio (SNR) as an input

- **Signal: Change in response per change in a factor's level**

- **Noise: Root Mean Square Error (RMSE)**

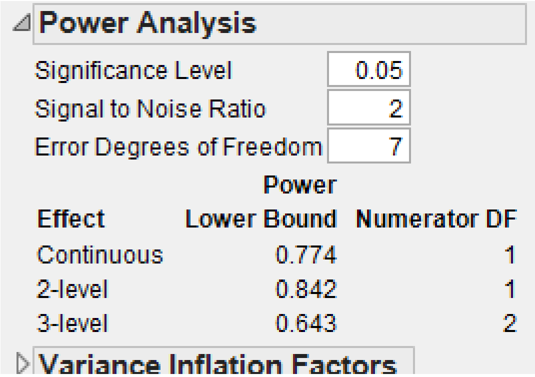


Alias Terms

Design

Run	Continuous	2-level	3-level
1	1	A	C
2	-1	A	D
3	-1	B	E
4	1	A	E
5	1	B	D
6	-1	A	D
7	-1	A	C
8	1	B	D
9	-1	B	E
10	1	A	E
11	0	B	C
12	0	B	C

Design Evaluation



Power Analysis

Significance Level	0.05
Signal to Noise Ratio	2
Error Degrees of Freedom	7

Effect	Power Lower Bound	Numerator DF
Continuous	0.774	1
2-level	0.842	1
3-level	0.643	2

Variance Inflation Factors

Aside: Power calculations can vary dramatically by software package and version

- Different assumptions
- Different coding
- Categorical factors particularly impacted

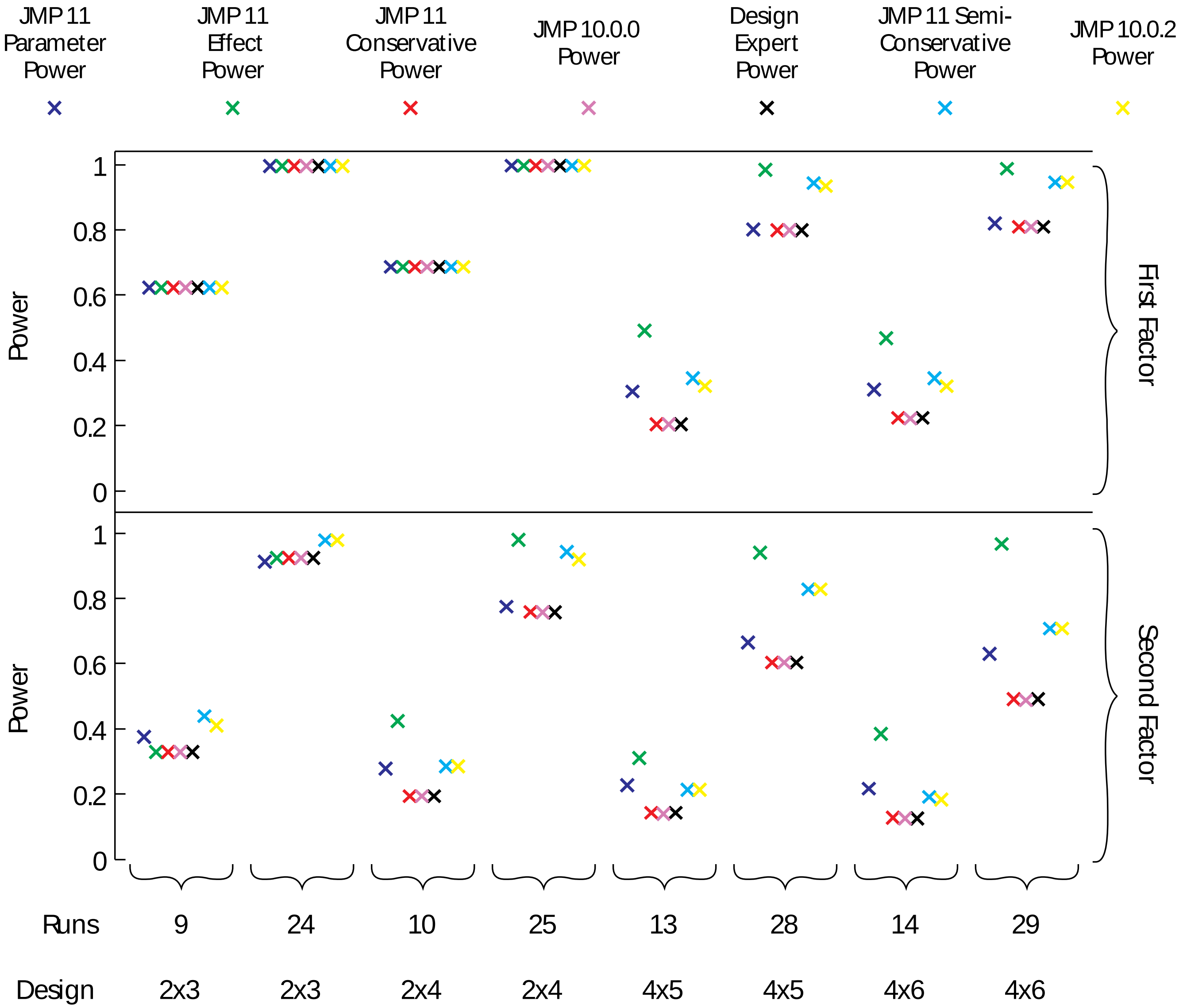


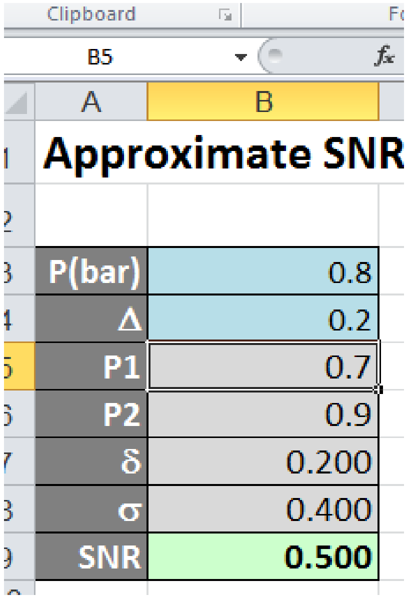
Chart courtesy of Dr. Tom Johnson (IDA) and Dr. Jim Simpson (UA Huntsville)

Power for binary responses

- **For some DOD systems, binary response variables are unavoidable**
  - Message completion rate
  - Torpedo hit/miss

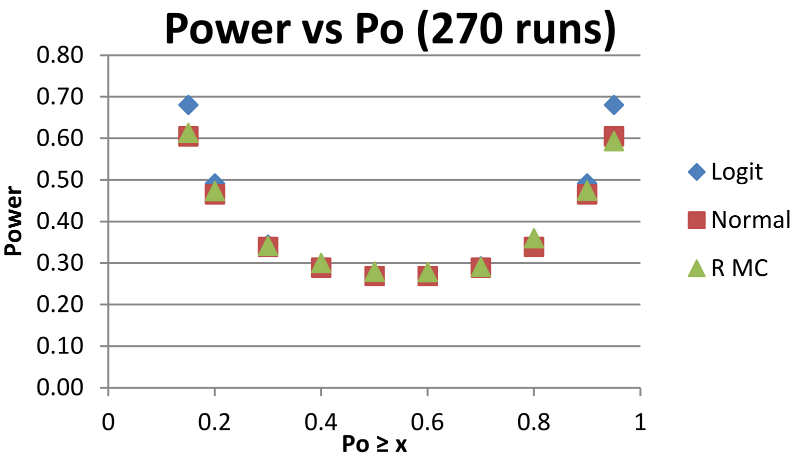
- **SNR framework doesn't apply well to binary response variables**
  - Signal
    - » Based on change in p?
    - » Based on log odds ratio?
  - Noise depends on $\bar{p}$
  - No software solution available

- **Work-around allows use of software[1]**
  - Normal approximation conservative relative to logit method
  - Resulting power estimates close to what you'd get through simulation



Approximate SNR

P(bar)	0.8
Δ	0.2
P1	0.7
P2	0.9
δ	0.200
σ	0.400
SNR	0.500



[1] *Dealing with Categorical Data Types in a Designed Experiment Part II: Sizing a Designed Experiment When Using a Binary Response*, Dr. Francisco Ortiz, AFIT STAT T&E COE; www.AFIT.edu/STAT

What SNR values are we currently using?

- **SNR**
  - STUAS: SNR of 0.5 for NIIRS, 2 for SPOI
  - AAV-SU: SNR of 1.3
  - AMISS: SNR of 2
  - Firescout: SNR of 1.5
  - MNRV: 2
  - JLTV: SNR=0.5, 1, 2

- **Effect Sizes**
  - APB 5: Δ=0.3, 0.2, 0.15
  - AMPV MS B TEMP: Δ=0.3, 0.25, 0.2
  - STUAS IOT Test Plan: Δ=0.2
  - MNRV: Δ=0.32

Are these values reasonable?

Estimating Empirical SNRs

Goal: Determine what size effects are observed in real test data

Fitting the model

- Fit a plausible, fully estimable model
- All two-way interactions if possible
- Reduce model if necessary (estimability, degrees of freedom, model over-fit, etc.)
  - Note: Goal *is not* to fit optimal model

For continuous response variables:

- Noise is RMSE
- Signal:
  - For categorical factor, the signal is β (R default 0-1 coding used)
  - For continuous factor, the signal is $\beta(\mu_{75} - \mu_{25})$
    - μ_n is the nth percentile for that factor
    - Many data sets have a few “extreme” data points

Estimating Empirical Δs

For categorical response variables:

- Using “workaround”, all we need is to estimate Δ
- Begin by computing $\bar{p}$:
  - Literally estimated by taking average over all effects:
  - $\bar{p} = \beta_0 + \frac{1}{m}\sum\beta_i^*$, where m is the number of effects estimated, and $\beta^* = \frac{1}{\mathrm{m_i}}\sum\beta_j^i$
- Estimating Δ:
  - For categorical factor, the signal is $\mathrm{inverse\_logit}(\mathrm{\bar{p}} + \beta)$
  - For continuous factor, the signal is $\mathrm{inverse\_logit}(\bar{p} + \beta(\mu_{75} - \mu_{25}))$
    - » μ_q is the qth percentile for that factor

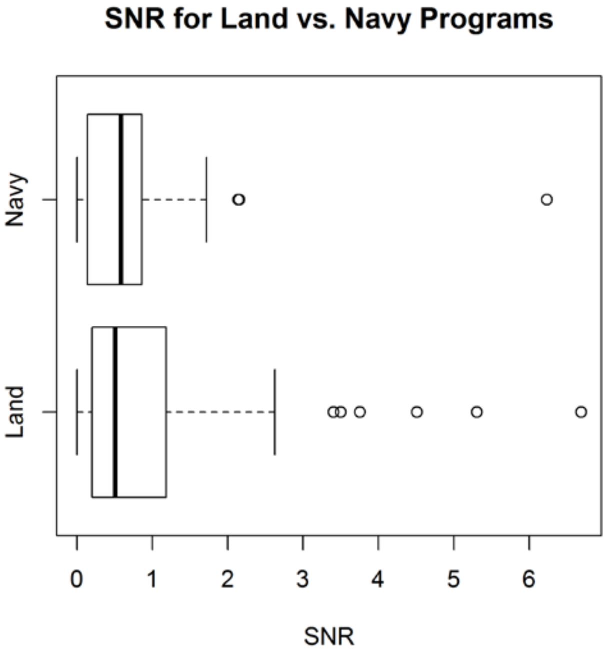
Summary of programs involved in this study

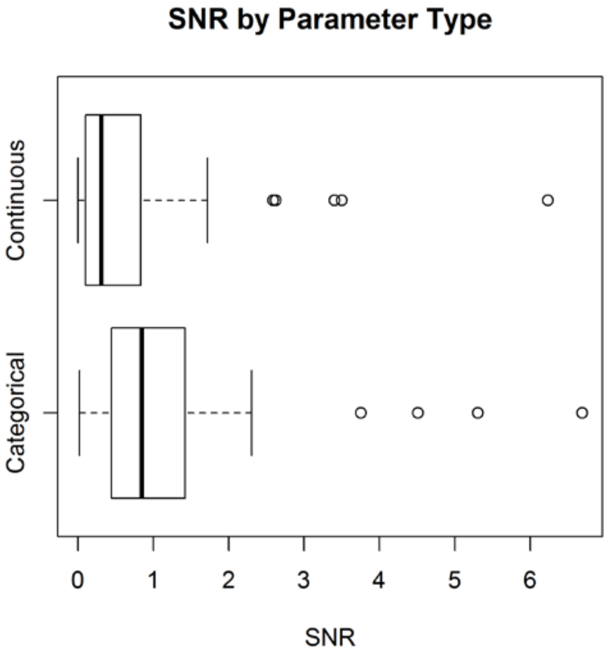
System	Response Variable	n	
Aegis	P(Raid Annihilation)	22	
Airborne Mine Neutralization System	Time to neutralize	33	
Virginia Class Submarine	Bearing Prediction Error	147	256
Chemical Agent Detector	Time to Detection	9,461	
LPD-17 (amphibious combat ship)	P(Impact)	296	
Mk54 CBASS Torpedo	P(Hit)	115	
Mk48 Torpedo	P(Hit)	35	
ARC-I Sonar	Difference in detection time	100	
Patriot	P(Intercept)	3,472	
RQ-21a Tactical UAV	Target Location Error	32	
Stryker Mobile Gun System	Correct Target Classification	464	
Global Broadcast Service	P(Successful Communication)	358	87
Paladin Self-Propelled Howitzer	Miss Distance	71	
Shadow Tactical UAV	Target Location Error	285	

Summary Statistics for Empirical SNRs

Mean	0.888
Median	0.534
75th percentile	1.151
90th percentile	2.026

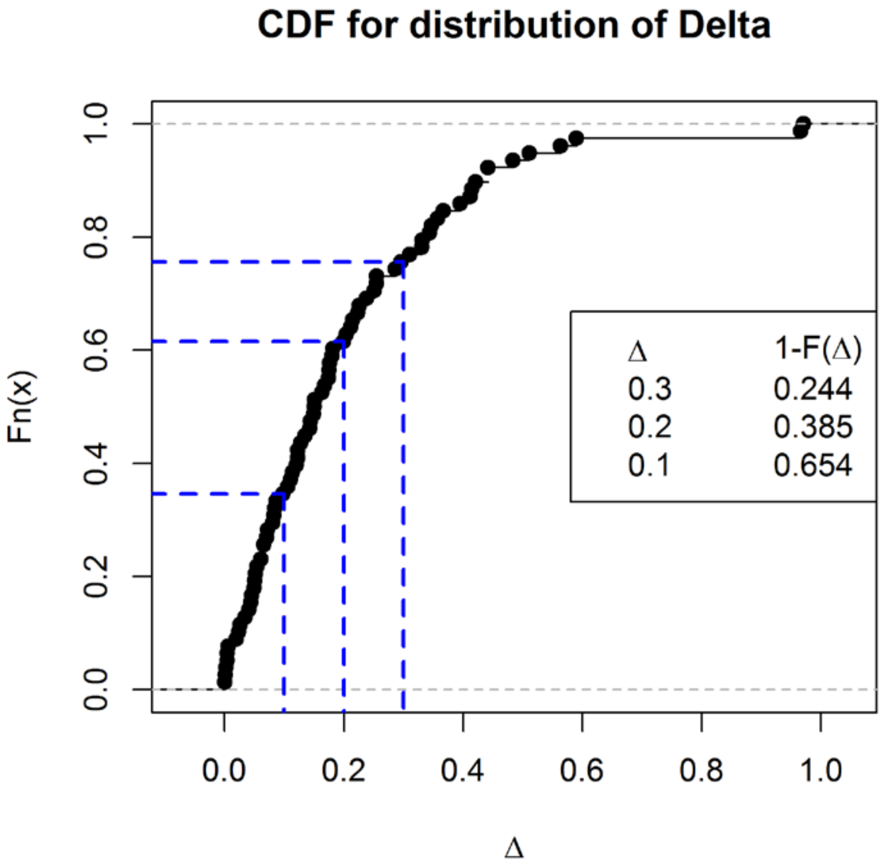
- Over 90% of observed effects have $SNR < 2$
- Minimal variation across warfare group
- Categorical factors had higher SNR
  - » Possibly an artifact of estimation method





CDF for Categorical Responses

- **Some effects are very large**
  - Largest come from continuous factors observed over large ranges
- **Typical values for Δ when sizing tests: 0.3, 0.2, 0.1**
  - Median effect size: 0.151
- **Many effect sizes very close to 0**
  - Most (11/14) with $\Delta < 0.05$ are interactions
  - How many are just "noise"?



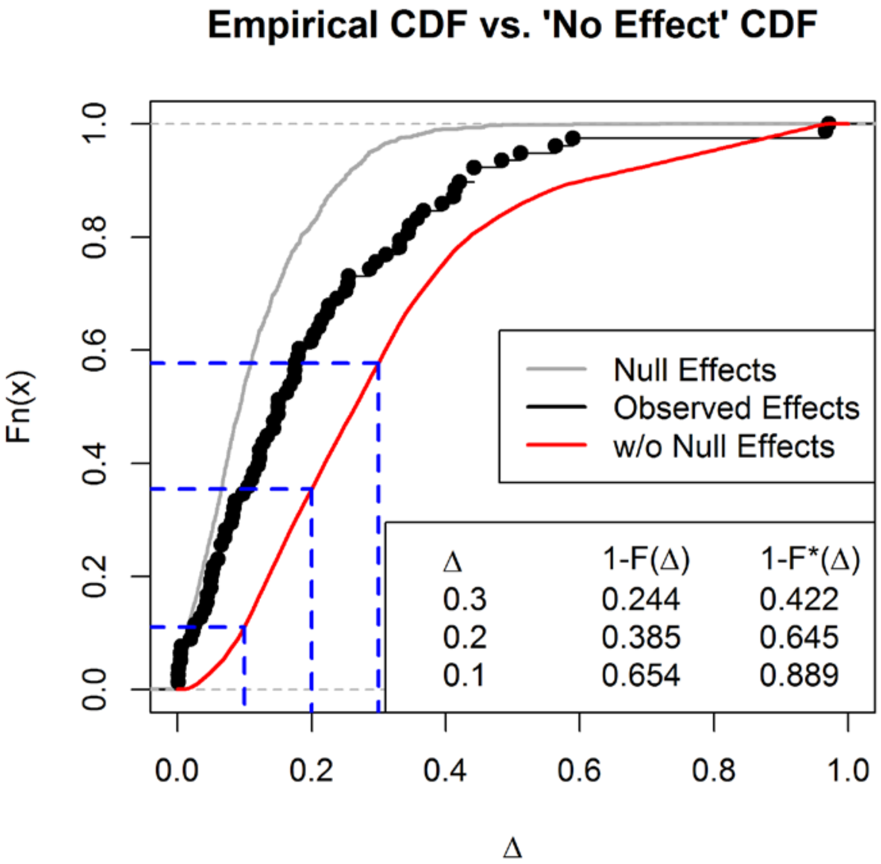
Δ	1-F(Δ)
0.3	0.244
0.2	0.385
0.1	0.654

Comparison to Null Model

- **Gray curve: Simulated data where “null” model is true**
  - Most effects are small
  - Median=0.093

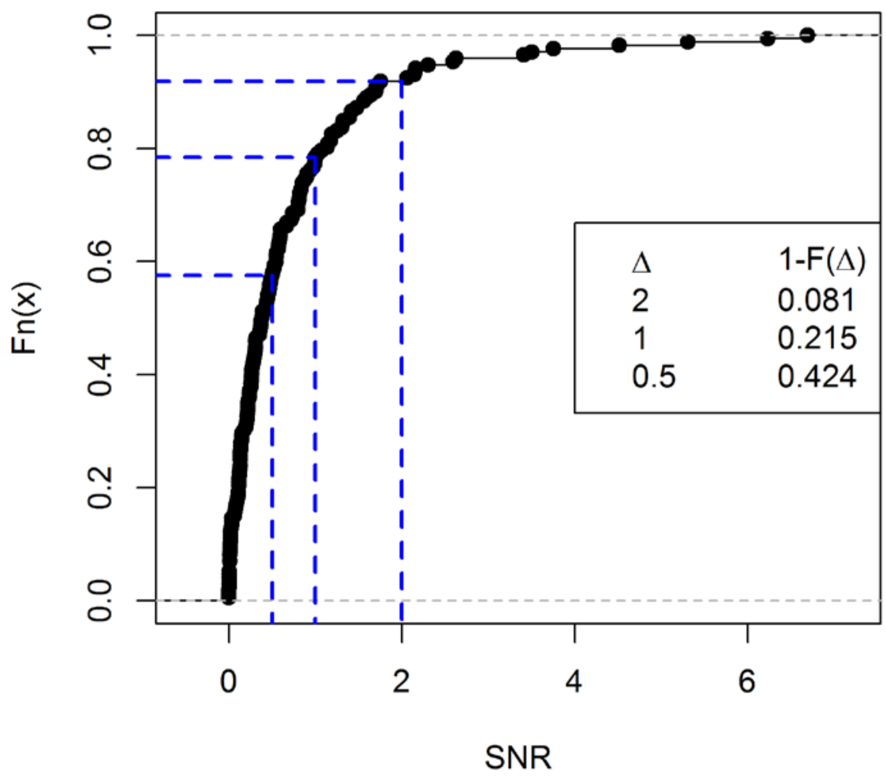
- **Subtracting “null” effects and normalizing yields red curve**
  - Distribution of true effects
  - Most are greater than 0.2
  - Nearly all greater than 0.1

Empirical CDF vs. 'No Effect' CDF

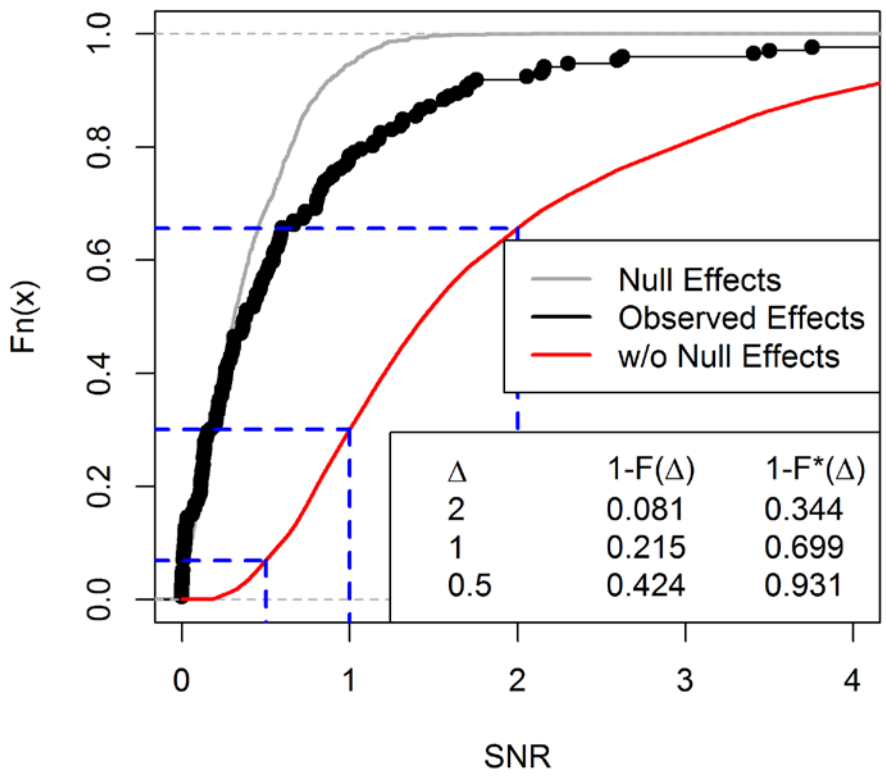


Δ	1-F(Δ)	1-F*(Δ)
0.3	0.244	0.422
0.2	0.385	0.645
0.1	0.654	0.889

Empirical SNR for continuous data



CDF for distribution of SNR



Empirical CDF vs. 'No Effect' CDF

So what?

- **After normalizing:**
  - **59%** of SNRs between **0.5** and **2**
  - **46%** of Δs between **0.1** and **0.3**

- **How do these values compare to what we've used for test planning?**
  - Planning for SNR=2 or Δ=0.3 is probably optimistic
    - Only 34.4% of effects have SNR>2
    - Only 42.4% of effects have Δ>0.3

- **Look at the ranges**
  - Compare power estimates over range of SNRs/Δs with likelihood of observing effects of that size
    - Ranges should at least cover 0.5 (SNR) or 0.1 (Δ)

- **Is it appropriate to generalize across all systems?**
  - Possibly….

Customization: Case Study for KC-46A

- **KC-46 GWEF testing**
  - KC-46 is new in-flight refueler
    - Replacing KC-135
  - Objective: Characterize performance for LAIRCM on KC-46 against representative surface-to-air threats

- **Test planning using empirical SNR distributions**
  - Identify similar tests
    - Response variable
    - Number of factors/levels
    - Test size
  - Compute “null” distribution based on these tests
  - Estimate CDF for SNRs
    - Difference between distribution of SNRs from similar tests and “null” distribution

Null distribution for KC-46 test design

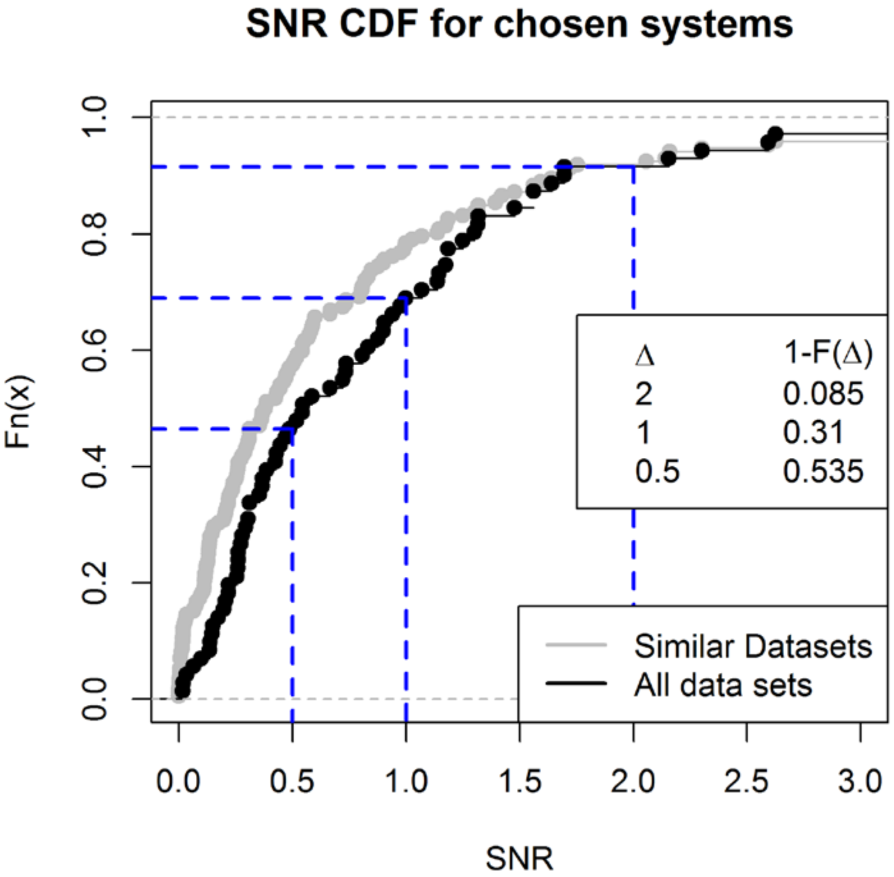
- **Response Variable: Miss distance (continuous)**
- **Factors**
  - IRCM status (Wet vs. Dry)
    - 2 levels
  - Scenario
    - 3 levels (categorical)
  - Declare Time
    - 5 levels (continuous)
  - Range
    - 5 levels (continuous)
  - Azimuth
    - 7 levels (categorical)
- **Total of n=500 data points**
- **Most similar data sets:**
  - PIM, JCAD, ARC-I

What is “similar”?

- **Physically**
  - Response variable
  - System type
- **Statistically**
  - Sample size
  - Number of factors
  - Levels of factors

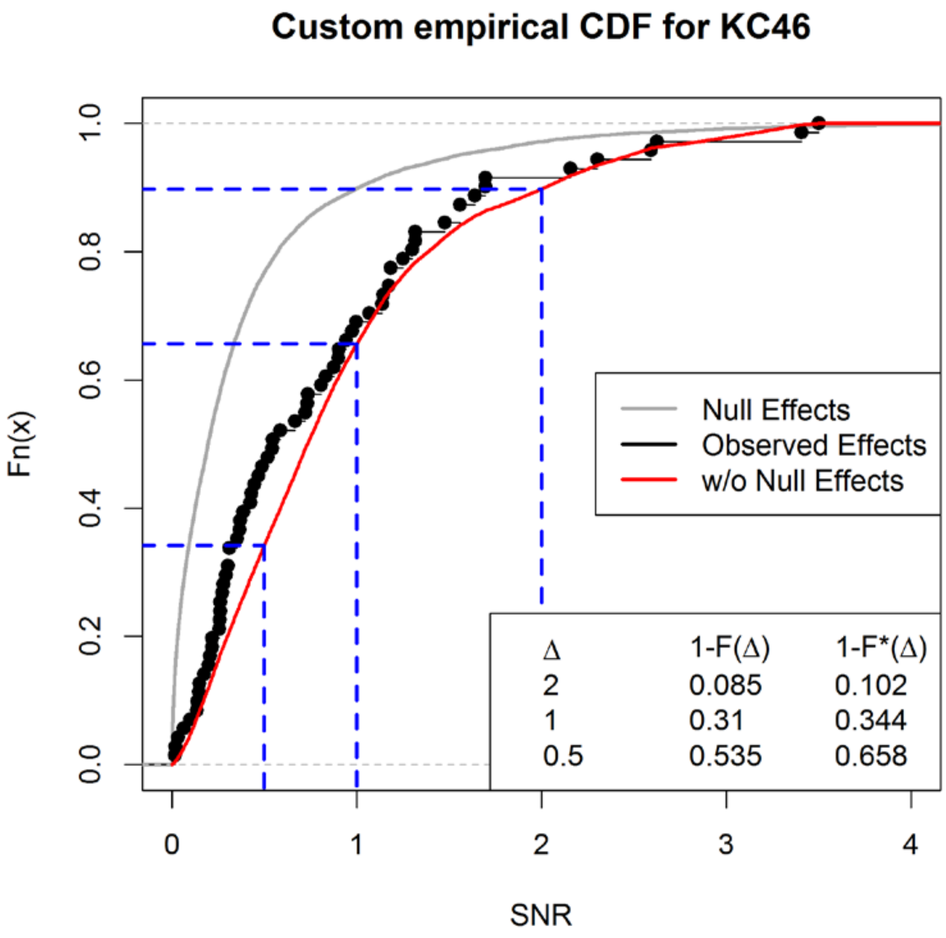
SNR distribution for similar systems to KC-46

- **SNR distribution from PIM, JCAD, and ARC-I**
  - Relatively few (~80) SNRs in the new curve
  - Fewer very small SNRs (SNR<0.5)
  - More mid-sized SNRs (0.5<SNR<1.5)



Custom SNR CDF for KC-46

- **Using custom CDF, we can estimate distribution of “real” effects for this test**
  - 25% have 1<SNR<2
  - 30% have 0.5<SNR<1
  - Based on this data, nearly 2/3 of SNRs from similar data sets to KC-46 are smaller than 1
    - » For all data sets, only 30% of effects have SNR<1
- **How much power does this design have for these SNRs?**



Δ	1-F(Δ)	1-F*(Δ)
2	0.085	0.102
1	0.31	0.344
0.5	0.535	0.658

Conclusions & next steps

- **Major Conclusions**
  - After normalizing:
    - **59%** of SNRs between **0.5** and **2**
    - **46%** of Δs between **0.1** and **0.3**
- **Future Work**
  - Additional data sets must be added for “customized” approach to be effective
  - Assess accuracy of *a priori* estimates of SNR
    - Are the values currently being used in test plans reflective of the SNRs observed once the tests have been conducted?
  - Assess uncertainty of estimates
    - Confidence intervals, sensitivity testing
- **Recommendations**
  - *Ceteris paribus*, use SNR no greater than 1 (70%) for power calculations
  - *Ceteris paribus*, use Δ no greater than 0.15 (76%) for power calculations
  - When power ranges reported, should include SNR=0.5 and Δ=0.1