Statistical Power to Support Test Adequacy Decisions

Part 1: Power Analysis Concepts

Tom Johnson, Laura Freeman, Jim Simpson

Rigorous T&E Knowledge Exchange Workshop, 13 Apr 16

IDA

Why a Power Tutorial?

- **All T&E organizations need to test adequately (i.e. just right) and maximize the knowledge gained**
- **Power is an important metric of test adequacy**
- **Power is a simple concept, the equation not as simple and easy to misapply**
- **Many power values for a single project can confuse – one per factor, per response, per design**
- **DOE software packages**
  - Critical to obtaining power estimates
  - The software platforms give different estimates for seemingly similar conditions!

Outline for Today

- Power Concepts
- Power for 2-level Designs
- Power for Multi-level Categorical Factor Designs
- Power for Binary Response Designs

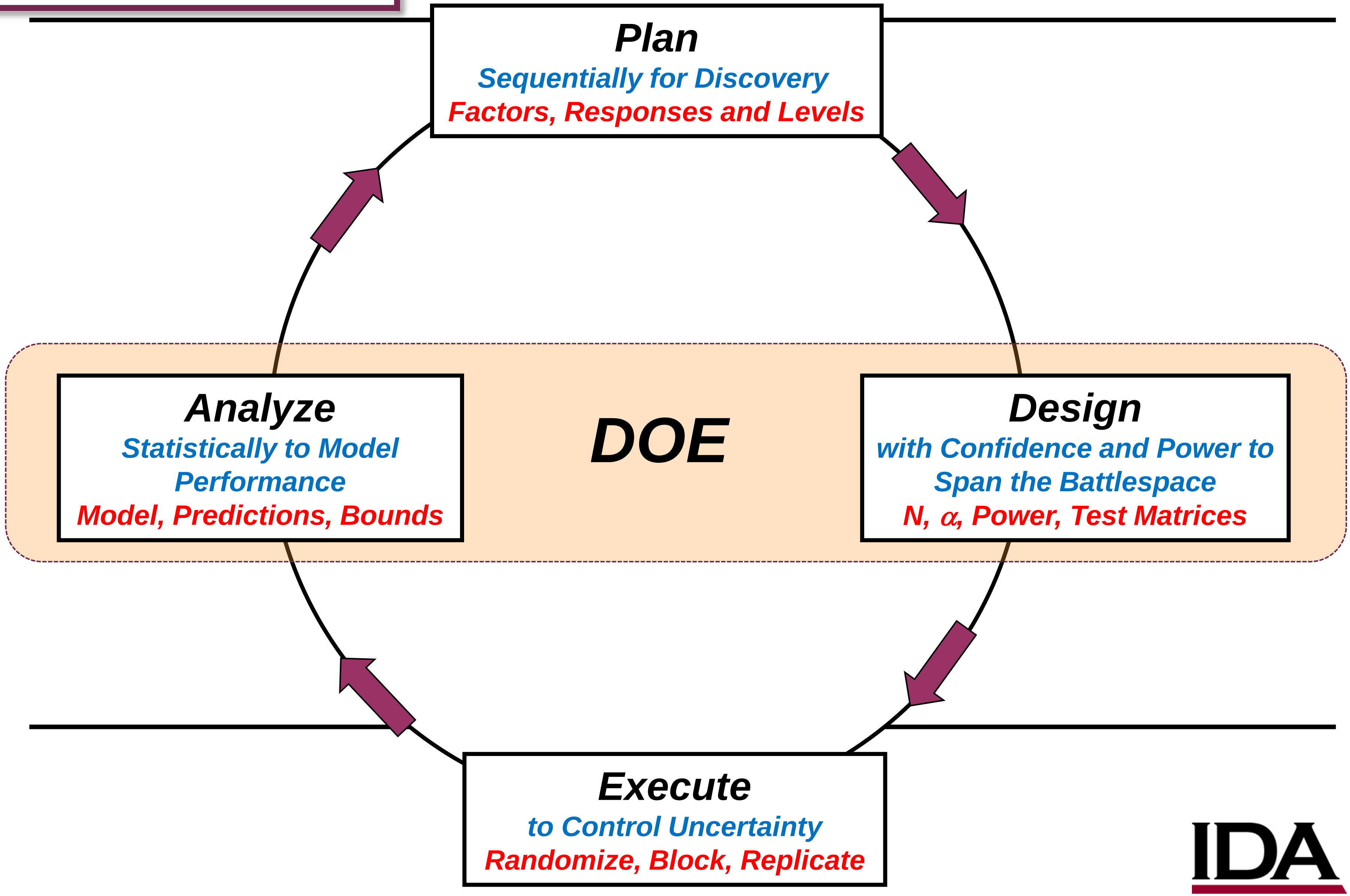
Simple Definition

Statistical Power *is a probability of uncovering active effects*

POWER CONCEPTS

DOE Process

Metrics of Note



Power in Context – One Phase and One Metric

Design Phase Key

- Statistical **power analysis** is performed to ensure very high chance of declaring factors of interest are important, given they really are important
- Design approach changes **all the relevant factors** simultaneously, spans the **factor level ranges**, permits estimating factor effects and factor interactions
- The **number of test events** (points) gradually increases as more factors are added
- **Test design developed** to gain efficiencies in total test resources allocated
- Design for **sequential testing** to leverage insight gained early in testing – ultimately maximizes knowledge gained for equal resources and flexes based on discovery – builds understanding in stages
- Provides the most potent allocation of test resources – by considering all relevant factors, **coverage of the test space**, right amount of **replication** for noise estimation, and only **feasible test combinations**

Power Analysis More Formally Defined

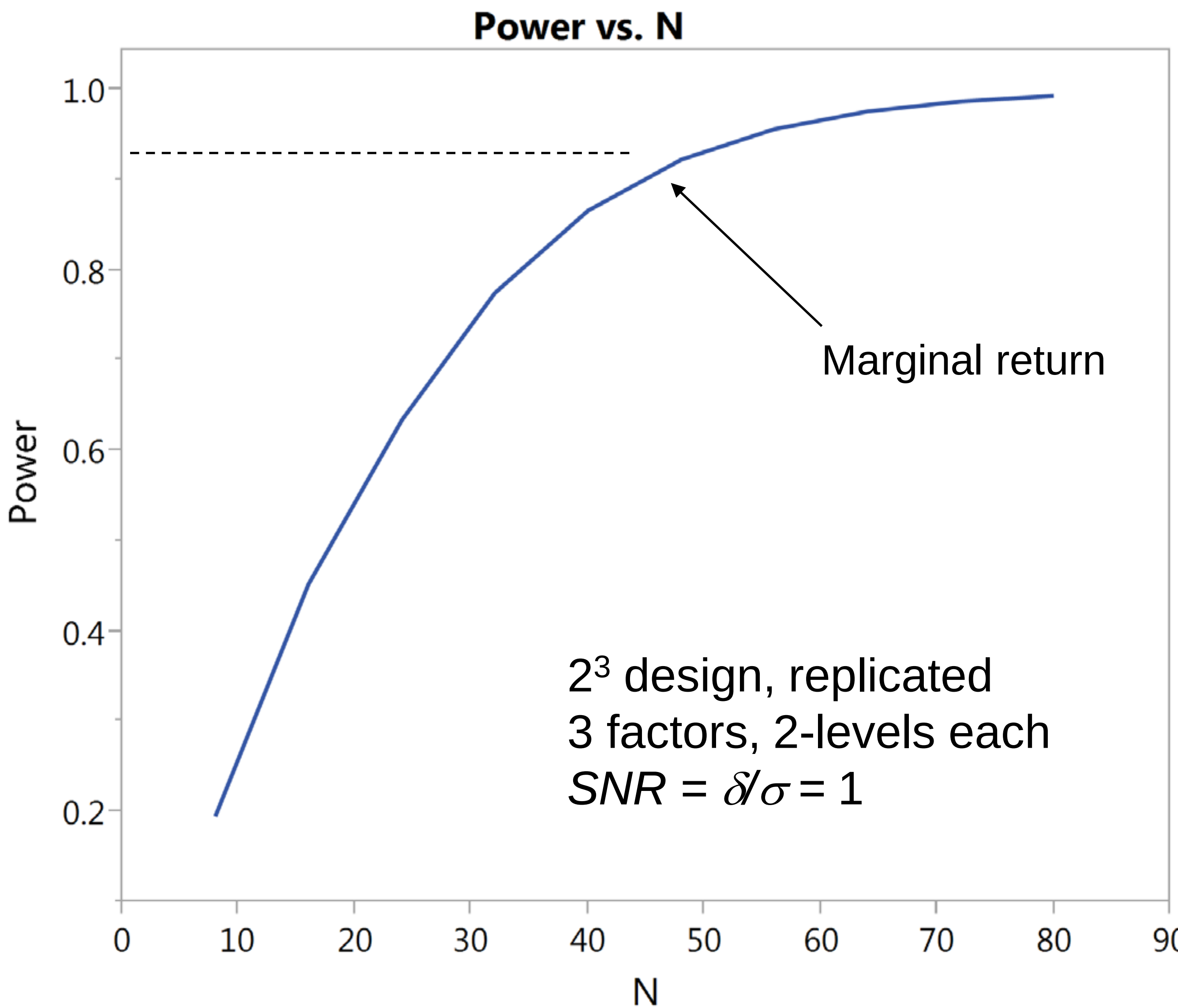
- **Statistical Power is a probability associated with making a correct conclusion about the system under study**
- **Specifically, when factors have been prescribed for a test, power is the probability that we will conclude that a factor is important, given it is really important**
- **More specifically, there are 2 types of error (and complements)**

α = Probability (the test conclusion is that a factor matters, given the factor has no effect)

β = Probability (the test conclusion is that a factor has no effect, given the factor matters)

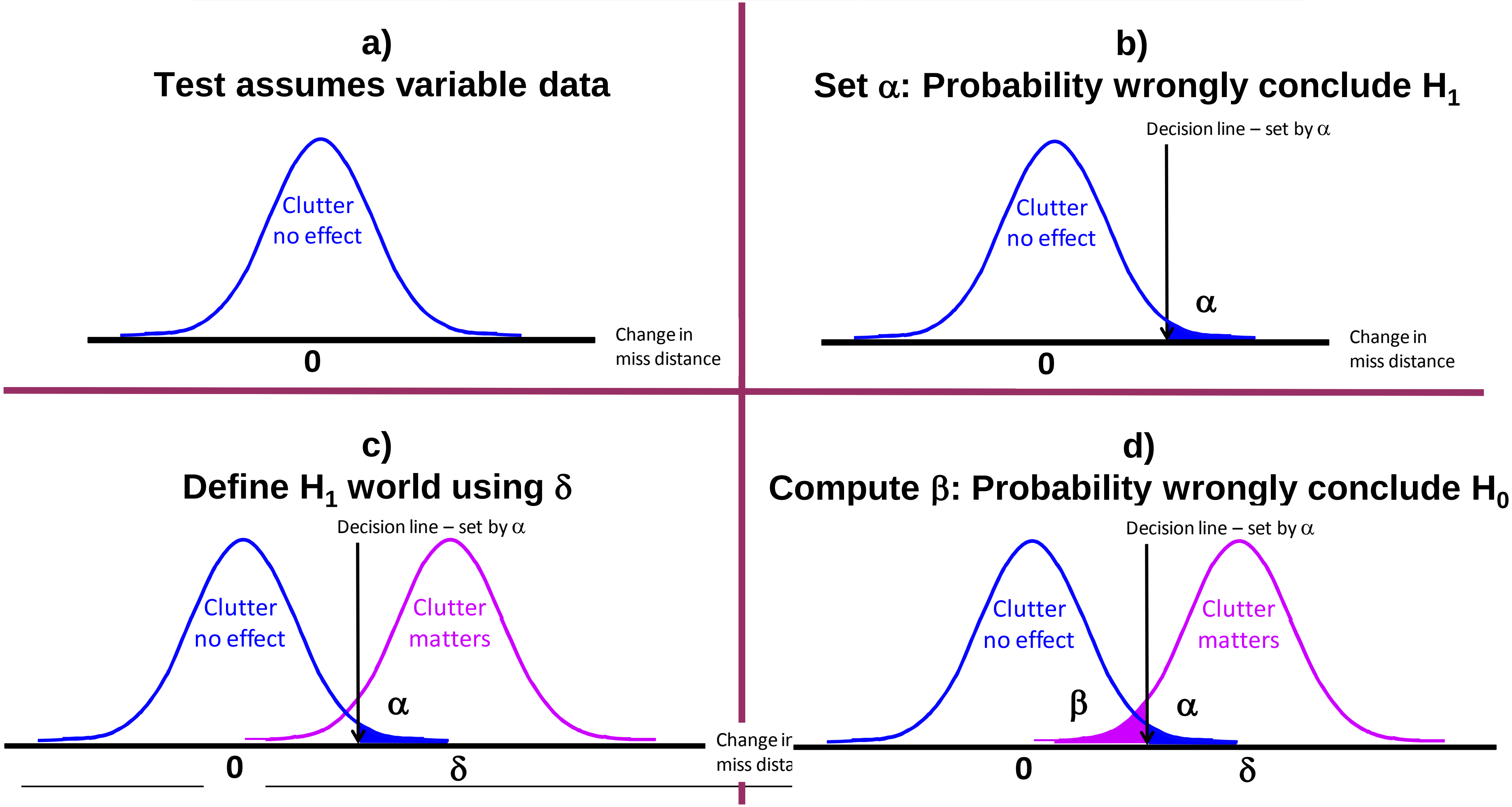
$1-\beta$ = Probability (the test conclusion is that a factor matters, given the factor matters)

Power vs. Sample Size (N) Relation



- *Example: **Does Clutter (High C vs. Low C) Degrade Missile Miss Distances (MD)?***
- *Form hypotheses: two possible worlds*

Hypothesis	Equation	In Words
H_0	$MD_{High\ C} - MD_{Low\ C} = 0$	Clutter no effect
H_1	$MD_{High\ C} - MD_{Low\ C} > 0$	Clutter matters



Decision Risks Illustrated

Example: Chemical agent detector

Truth Model: ***Detect Distance = Device + Agent***

Test Factors	*Hypotheses*	*Possible Conclusion*	*Error*
A: Humidity	**H_0: Humidity has no effect** H_1: Humidity matters	Humidity matters	α
B: Device	H_0: Device has no effect **H_1: Device matters**	**Device matters**	**None,** $1-\beta$
C: Agent	H_0: Agent has no effect **H_1: Agent matters**	Agent has no effect	β

**** Bold Blue reflects the truth***

α = Probability (the test conclusion is that a factor matters, given the factor has no effect)
β = Probability (the test conclusion is that a factor has no effect, given the factor matters)

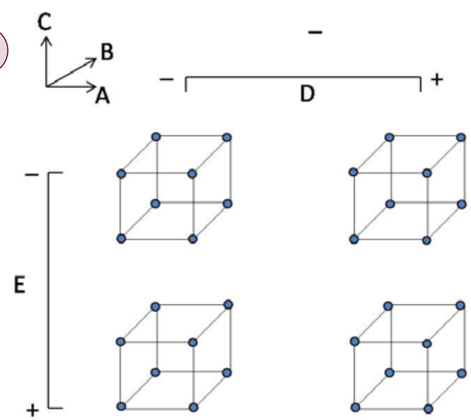
Decision Risks Illustrated – Part II

Another perspective: CV-22 Terrain Following/Avoidance

True Model: **Deviation = Ride + Turn**

Test Factors

- *A: Airspeed*
- *B: Turn*
- *C: SCP*
- *D: Ride*
- *E: Nacelle*



Run	Deviation
1	7.6
2	0.5
3	16.2
…	
32	9.3

	Model	*Error*
Deviation =	*~~Airspeed~~*	α
	Ride	**None, 1 - β**
	Noise	
	SCP	
	Nacelle	
	~~Turn~~	β

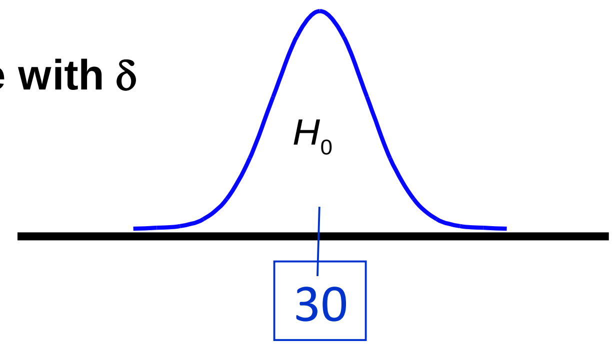
Power Analysis Parameters

Parameter	Description	How Obtained	Relevance in Planning
k: factors	Number of factors in the experiment	Determined in process decomposition	Key finding from process decomposition
df_{error}: model error	Amount of data reserved for estimating system noise	Desired model order (e.g. interaction, quadratic)	Estimate of complexity of input-output relation
α: alpha	Probability (declaring factor matters when it doesn't)	Set by test team	Fix and leave alone
δ: delta	Size of response change expert wants to detect	Experts and management determine	Some ability to vary
σ: sigma	System noise – run-to-run variability or repeatability	Historical data; pilot tests; expert judgment	System driven but can be improved by planning
1-β: power	Probability of declaring a factor matters when it does	Lower bound set by test team	Primary goal is to set N to achieve high power
N: test size		Usually computed based on all other parameters	Direct, should modify to satisfy power

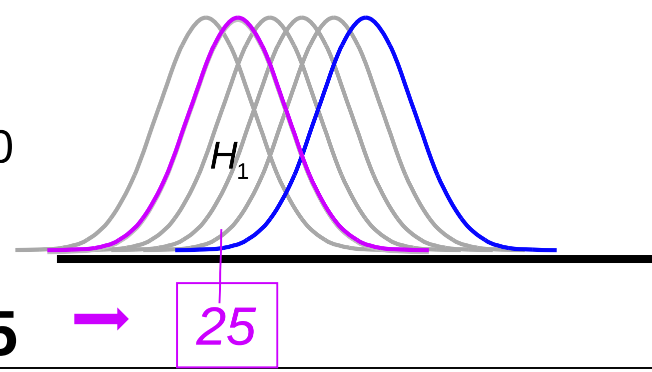
Hypothesis Testing – Set the Alternative with δ

- **Regardless of the distribution of the measure of performance, as *N* increases, the distribution of means becomes normal - CLT**
- **The means targeted in hypothesis testing have distributions**
- **The null hypothesis has a reference mean, but alternative has *infinite* means**
- **The δ is the difference between null and alternative means and is used to anchor the alternative**
- **Computing both α and β is possible with δ**
- **Example:**

$$H_0 : \mu_{TLE} \geq 30$$



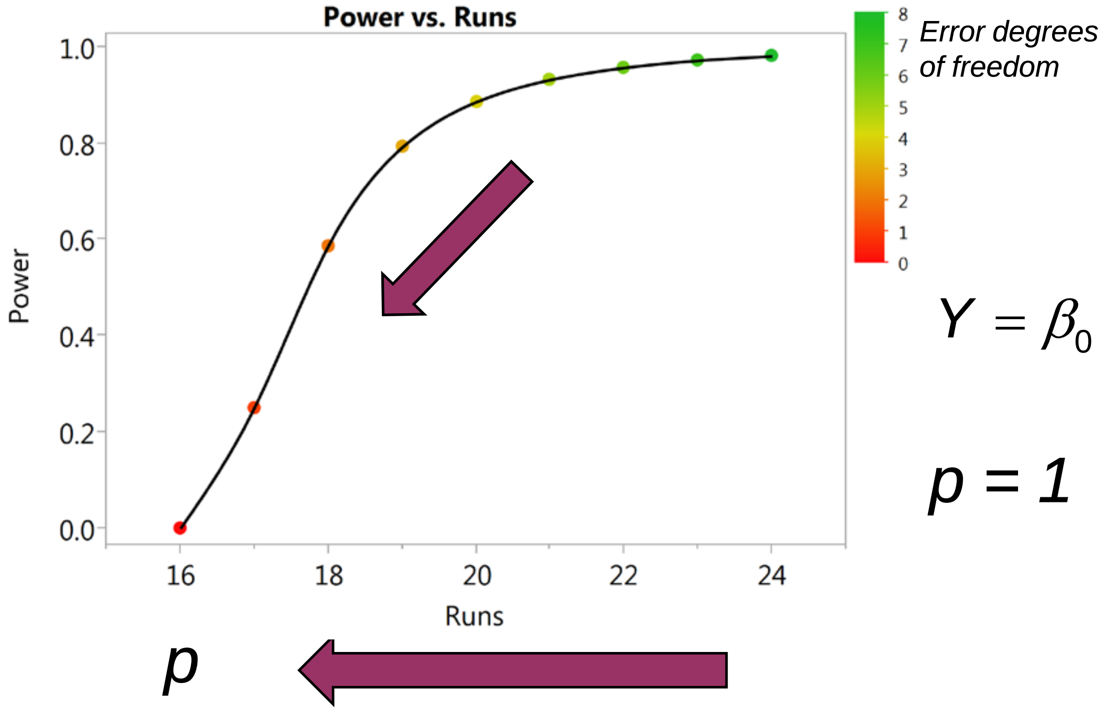
$$H_1 : \mu_{TLE} < 30$$



$\delta = 5$ ➡ 25

Sensitivity to df_{error}

- **As a design *N* approaches the number of model degrees of freedom, *p*, power drops drastically**

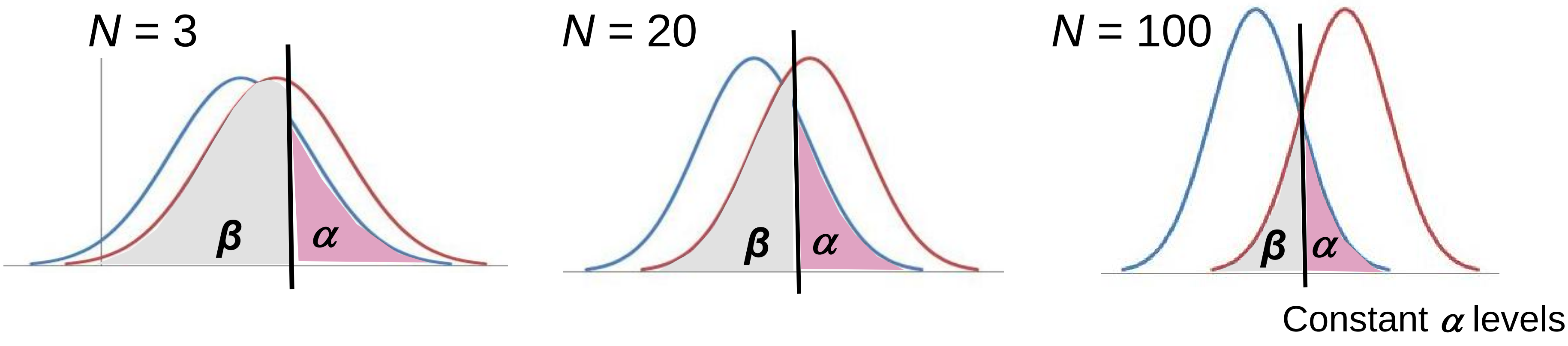


2^{5-1} design
ME + 2FI model

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$$

$$p = 1 \quad + \quad 5 \quad + \quad 10 = 16$$

Sensitivity to *N*



Adapted from: Osborne, Ken, Busby, Deborah, Schroeder, Kurt, Managing Test Risk During Design: Bushmaster II Testing, Eglin Technical Document, 2 Apr 2009.

Power Terminology in Software

- **Terminology in Software to Request or Report Delta (δ) and Sigma (σ) Estimates for Power Analysis**

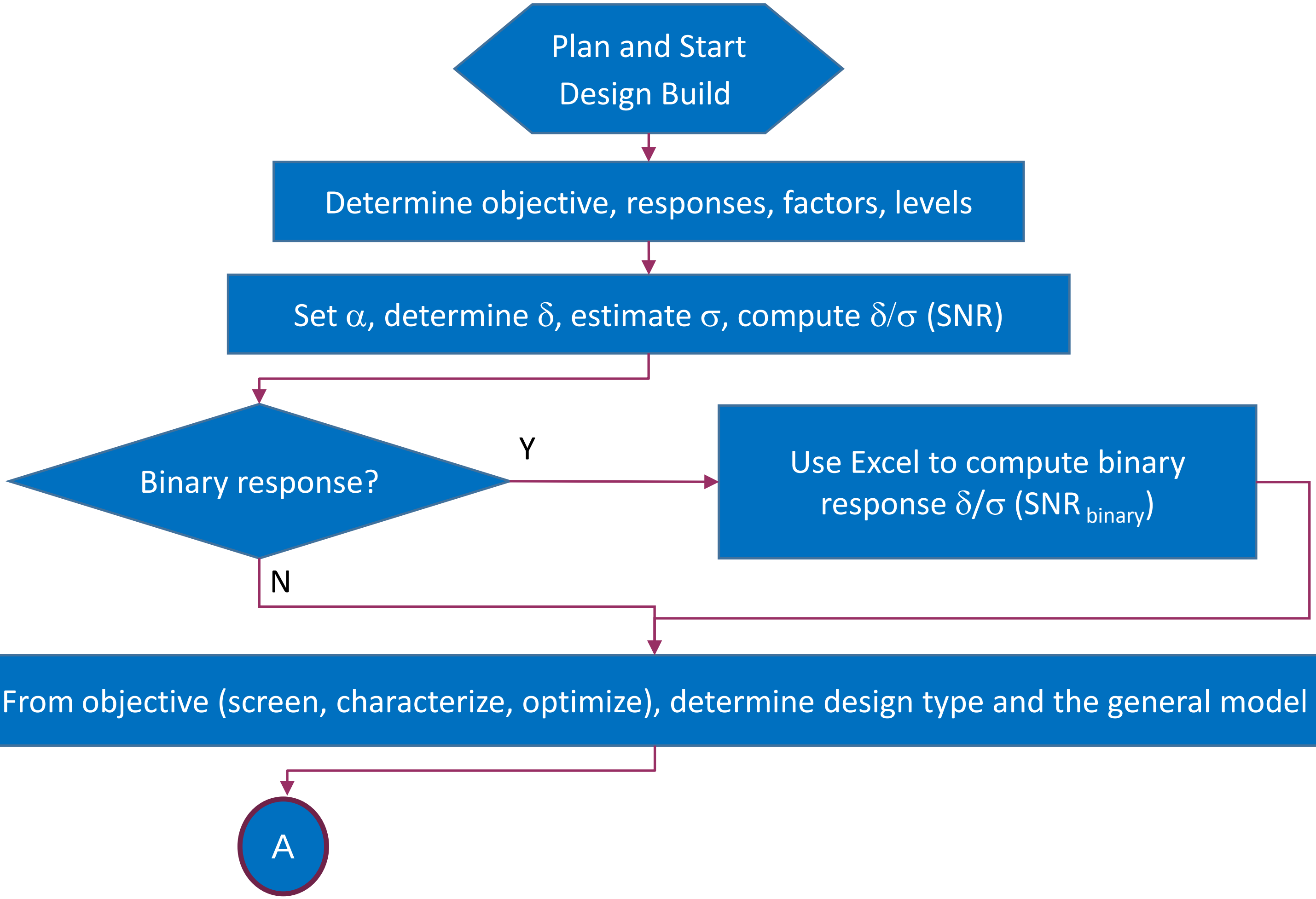
Software	Delta	Sigma	Delta/Sigma
Design Expert 8, 9, 10	**Delta**, Diff. to detect, “Signal”	**Sigma**, Est. Std. Dev., “Noise”	Delta/Sigma, Signal/Noise Ratio*
JMP 9	Implied, as Signal but can’t enter directly	Implied, as Noise but can’t enter directly	**Signal to Noise Ratio**
JMP 10	Implied, as Signal but can’t enter directly	Implied, as Noise but can’t enter directly	**Signal to Noise Ratio**
JMP 11, 12	Indirectly either using Anticipated Responses or Anticipated Coefficients, or directly using **Delta** under Advanced Options)	**Anticipated RMSE**	If using Advanced Options, and Power Analysis interface, then delta/RMSE, assuming RMSE = 1

** Note: In Design Evaluation, several default delta/sigma ratios (0.5, 1.0, 2.0) are shown as e.g.* ***2 Std. Dev****.*

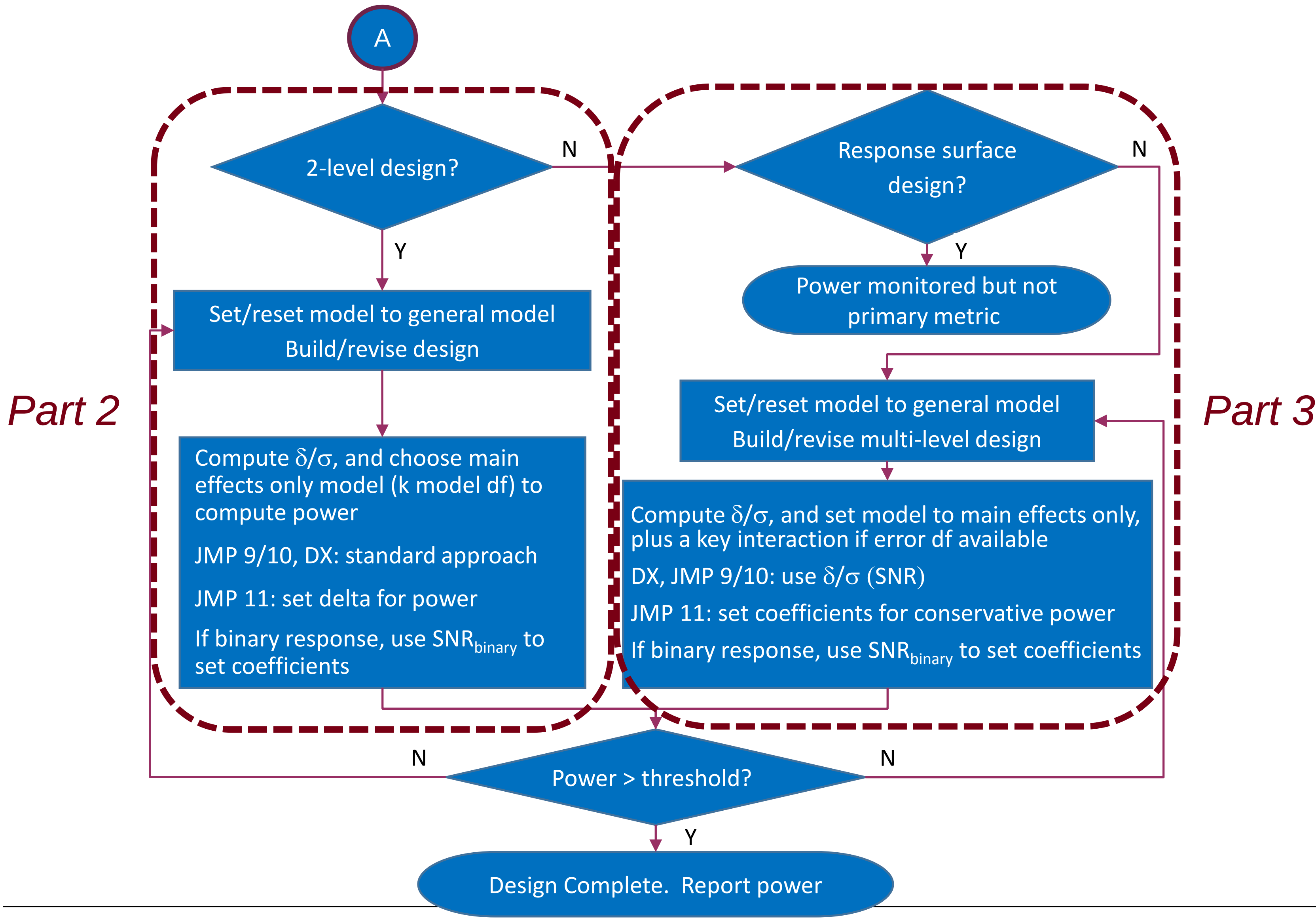
Power Is Only ONE Design Metric

Design Characteristics	Design Characteristic Descriptions and Notes	Design Metrics			
Design Name	Some names will not match any academic term since they are test specific.	D – Optimal (GA)	I – Optimal (GA)	Non-Orthogonal CCD (GA)	etc.
Design Size (N)	The total amount of data points (N) in the design.	14			
Design Cost	Cost of running the design.	$11[d] M			
# Factors Analyzed	Number of factors covered in the design	7			
Resolution/Aliasing	An indication of how difficult it will be to separate effects from one another	Bad Partial Aliasing			
α Error	The risk of declaring the M4E1 worse than the M4 or ICAD when in reality it is not (an incorrect fail.)	?			
(1 - α)		Decided by Requirements			
(1 - β) Power	The risk of declaring the M4E1 as good or better than the M4 or ICAD when in reality it is not (an incorrect fielding.)	? Decided by Test Team			
VIF (Average)	Variance Inflation Factor. Measures how much the variance of the model is inflated by a particular factor due to its lack of orthogonality.				
VIF (Max)					
Leverage (Average)	The potential for a design point to influence the fit of the model.				
Leverage (Max)					
FDS (@ 50%)	Fraction of the Design Space. The rank order of the change in prediction variance across the design space. A low flat curve is desired.				
FDS (@ 90%)					
Potential Model	The prediction model the design is capable of estimating.		ME + 2FI + Quadratics		

General Power Procedure – Stage 1



General Power Procedure – Stage 2



Power Demonstration

- **Using Monte Carlo simulation, we can illustrate the result of insufficient power – Monte Carlo can also be used to estimate power**

- **Consider three factors, 2-levels each, 1 replicate design**

Risk	Probability	Outcome from a DOE
α	P(conclude effect | no effect)	Effect significant but not in truth model
β	P(conclude no effect | effect)	Effect insignificant but in truth model

- **For the design for δ / σ = 2, 1 - β = 0.57**

- **Truth Model y = 50 + 4.5A - 5B, so effects are A = 9, B = 10**

- **Error standard deviation σ = 5**

Power Concepts Summary

- **The two test risks are probabilities associated with incorrect conclusions based on a pair of complementary hypotheses conjectured prior to test.**
- **Of the two risks, the α risk is set up front. Standard α = 0.05.**
- **The β risk is usually computed then iterated on by changing *N* until β is sufficiently small.**
- **Power is a probability (1- β) and is the complement of the β risk associated with test.**
- **Because of the way we address the two risks, power becomes the final risk typically addressed in design construction.**

POWER FOR 2-LEVEL DESIGNS

Introduction to Power for 2-level Designs

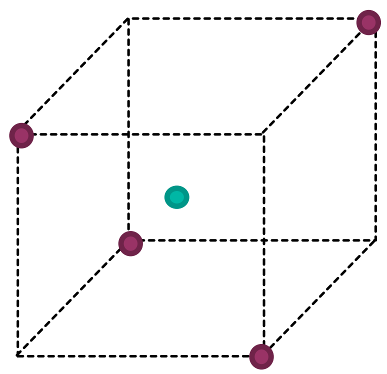
- **2-level Designs and Statistical Models**
  - Encourage these designs whenever practical
  - Designs ultra-efficient
  - Variables analyzed the same whether the factors are numeric or categorical
  - *df* concept vary simple too – all effects have 1 *df*
  - Model and effect interpretation very simple
  - Higher power designs

- **Statistical software tend to agree on power**

2-level Designs

Assumptions
Randomized
Numeric or Categorical
2 level

Design



Attributes
Single Replicate Corners
Center points
Orthogonal
Variance Optimal
Efficient

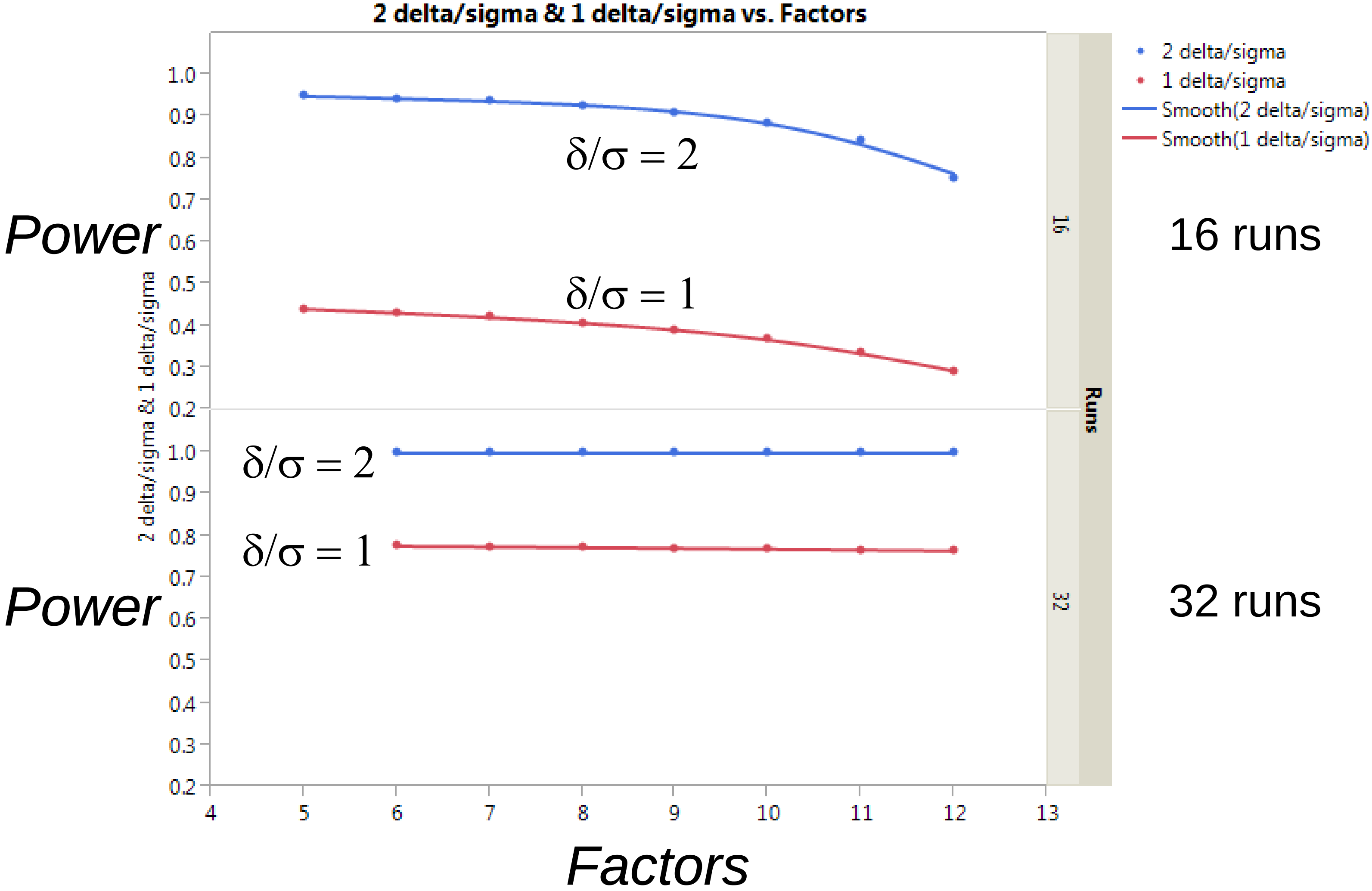
Assumptions
Errors NID (0, σ^2)
Model is adequate
Y well behaved

Model

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$$

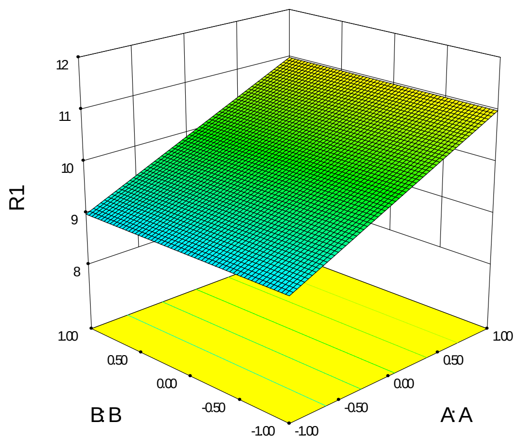
Attributes
Aliased terms – degree depends on resolution
Curvature estimate
Independent β estimates

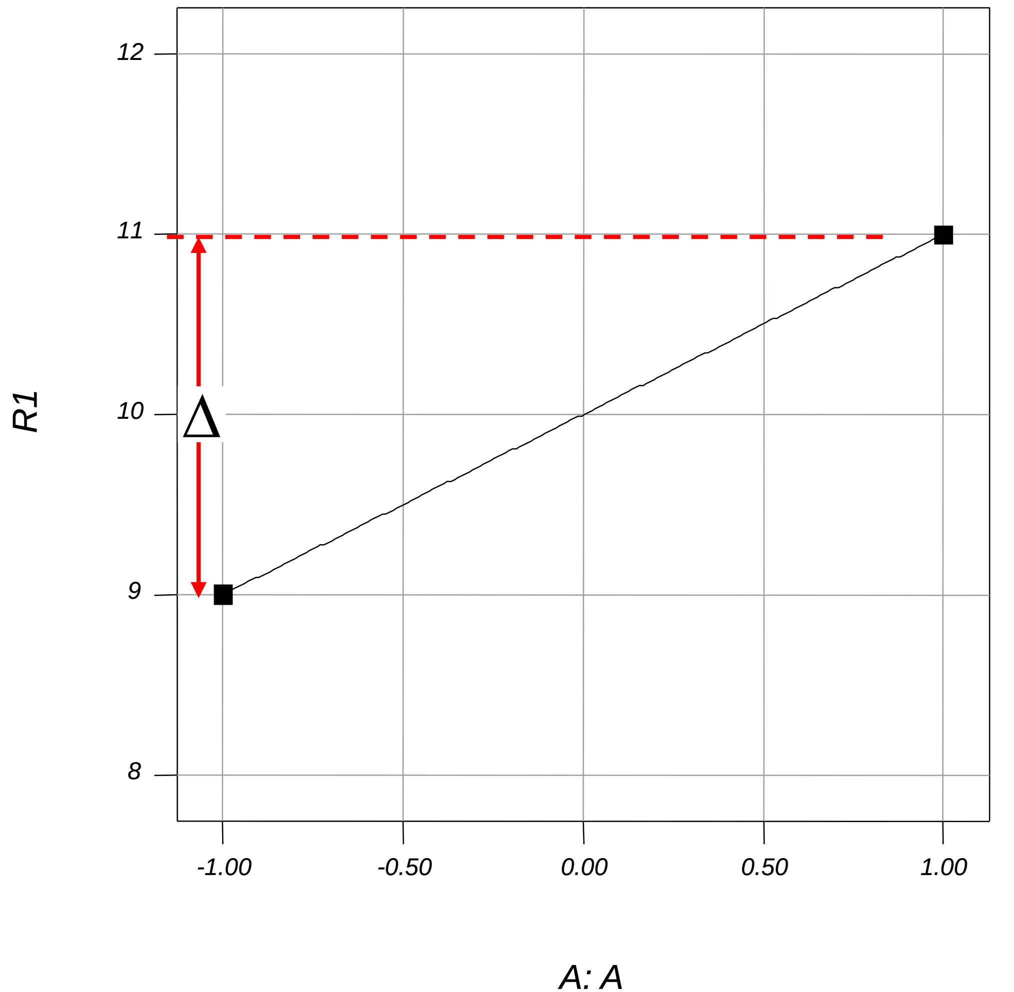
2^{k-p} Fractions
Statistical Power for Continuous Response



2-LEVEL DESIGN POWER CALCULATIONS

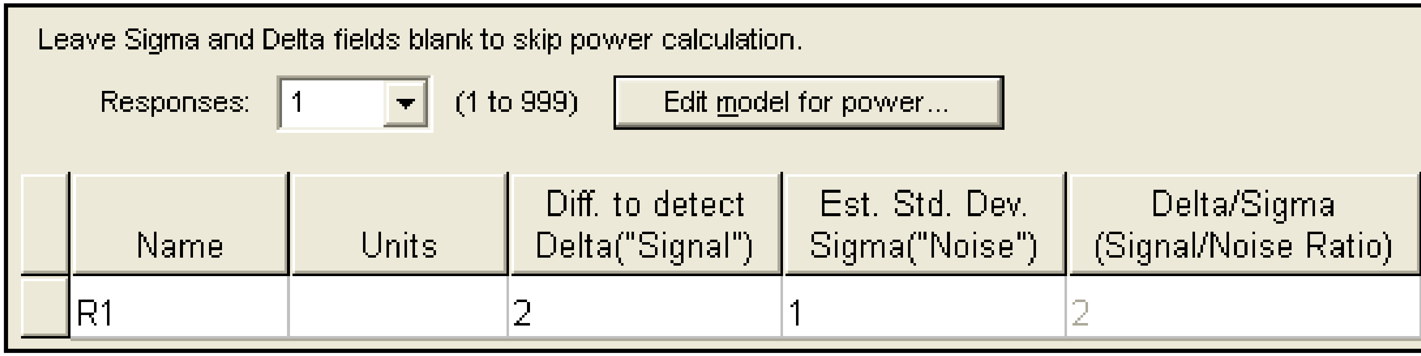
Factorial Design – Power
2^3 Full Factorial Δ=2 and σ=1, 2 Reps *N*=16





The following 6 slides are from "Sizing Mixture (RSM) Designs, Pat Whitcomb, Stat-ease

Factorial Design – Power
Two Replicates of 2^3 Full Factorial Δ=2 and σ=1



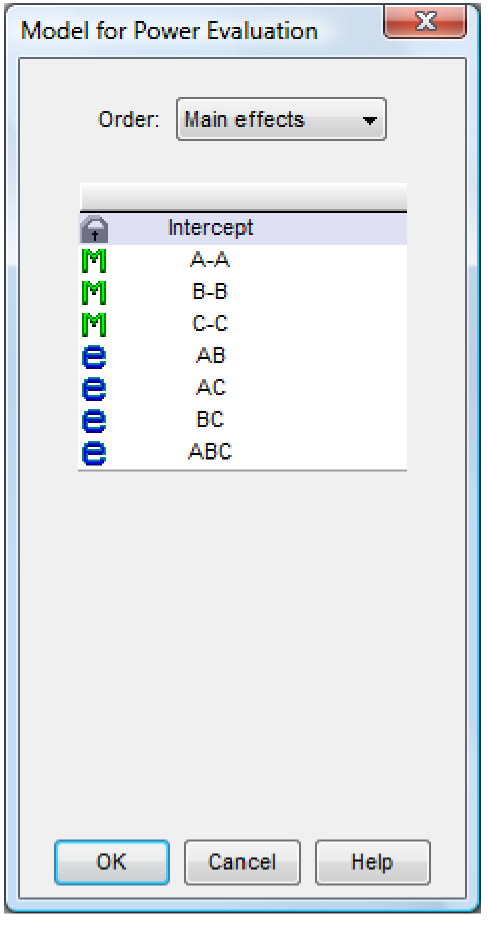
Power is reported at a 5.0% alpha level to detect the specified signal/noise ratio.
Recommended power is at least 80%.

R1

Signal (delta) = 2.00 Noise (sigma) = 1.00 Signal/Noise (delta/sigma) = 2.00

A	B	C
95.6 %	95.6 %	95.6 %

Factorial Design – Power
Two Replicates of 2^3 Full Factorial Δ=2 and σ=1



Assume a main effects model to estimate about the right number of significant model terms

Source	Degrees of Freedom (df)
Model	3
Error	12
Total	15 + 1 Intercept

So, df_{error} = 12 used to draw H_1 distribution

All df accounted for in budget

Two Replicates of 2^3 Full Factorial
$C = (X^TX)^{-1}$ matrix

The design determines the standard error of the coefficient:

$$C = \begin{pmatrix} 0.0625 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0625 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0625 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0625 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0625 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0625 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0625 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0625 \end{pmatrix}$$

$$\text{t-value}_i = \frac{\beta_i}{SE(\beta_i)} = \frac{\beta_i}{\sqrt{c_{ii}\hat{\sigma}^2}} = \frac{\beta_i}{\sqrt{(0.0625)\hat{\sigma}^2}}$$

NonCentrality Parameter
Two Replicates of 2^3 Full Factorial Δ=2 and σ=1

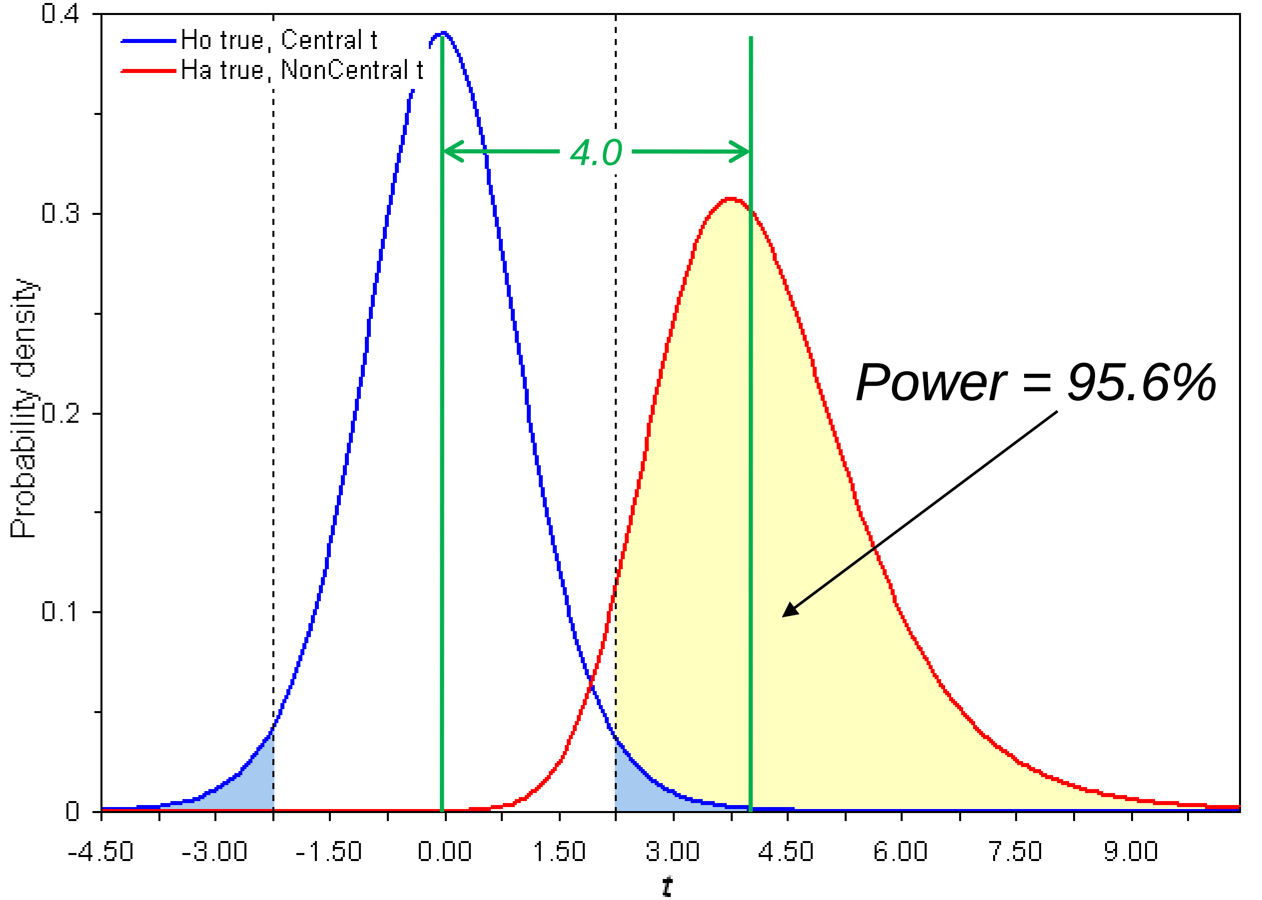
$$\text{noncentrality}_i = \frac{\beta_i}{\sqrt{c_{ii}\hat{\sigma}^2}} = \frac{\Delta_i/2}{\sqrt{c_{ii}\hat{\sigma}^2}}$$

$$= \frac{1}{\sqrt{(0.0625)(1)^2}}$$

$$= \frac{1}{0.25} = 4.0$$

Factorial Design – Power
Two Replicates of 2^3 Full Factorial Δ=2 and σ=1

noncentral $t_{\alpha=0.05,df=12}$ *with noncentrality parameter of 4.0*

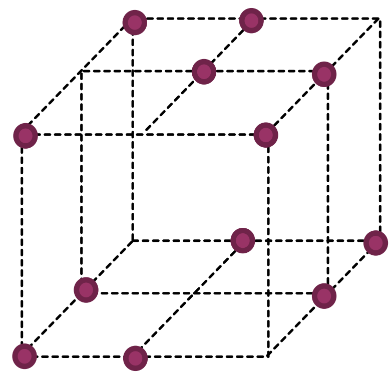


POWER FOR MULTI-LEVEL CATEGORICAL FACTOR DESIGNS

Multi-level Designs

Assumptions
Randomized
Some Categorical
Categorical > 2 level
Computer Generated

Design



Attributes
Replication
Target Model Variances
Single Criterion Designs
Efficient

Assumptions
Errors NID (0, σ^2)
Model is adequate
Y well behaved

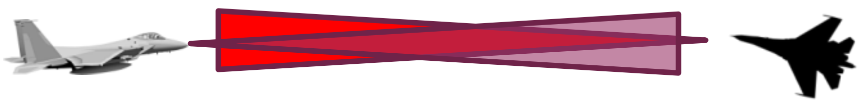
Model

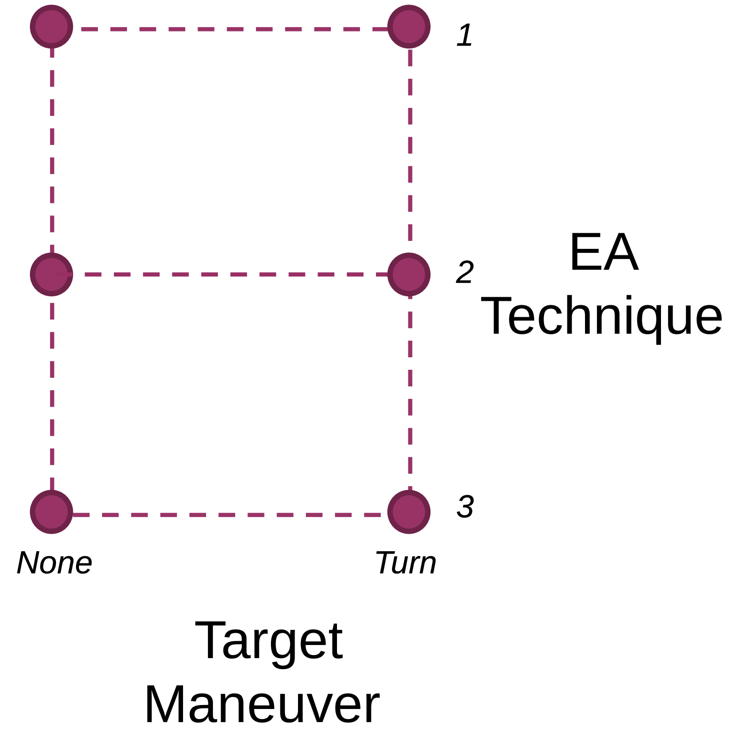
$$y_{ijk} = \beta_0 + \beta_{11}x_{11} + \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{33}x_{33} + \beta_{1121}x_{11}x_{21} + \beta_{1122}x_{11}x_{22} + \beta_{1133}x_{11}x_{33} + \beta_{2133}x_{21}x_{33} + \beta_{2233}x_{22}x_{33} + \varepsilon$$

Attributes
Some terms correlated
Pure error + LOF
Awkward ANOVA model

Multi-level Categorical Factor Analysis

- **Consider a factorial design dominated by categorical factors or containing at least one categorical factor requiring more than 2 levels**
- **Consider a sensor assessment study considering two factors that may impact electronic attack (EA, also known as electronic countermeasures)**





Regression Indicator Variable Model

$$y_{ijk} = \beta_0 + \sum_i \beta_{1i} x_{1i} + \sum_j \beta_{2j} x_{2j} + \sum_m \beta_{1i2j} x_{1i} x_{2j} + \varepsilon \begin{cases} i = 1, 2, \ldots, (a-1) \\ j = 1, 2, \ldots, (b-1) \\ m = 1, 2, \ldots, (a-1)(b-1) \end{cases}$$

Grand Mean

Main Effect A or X1

Main Effect B or X2

Interaction AB or X1*X2

Error

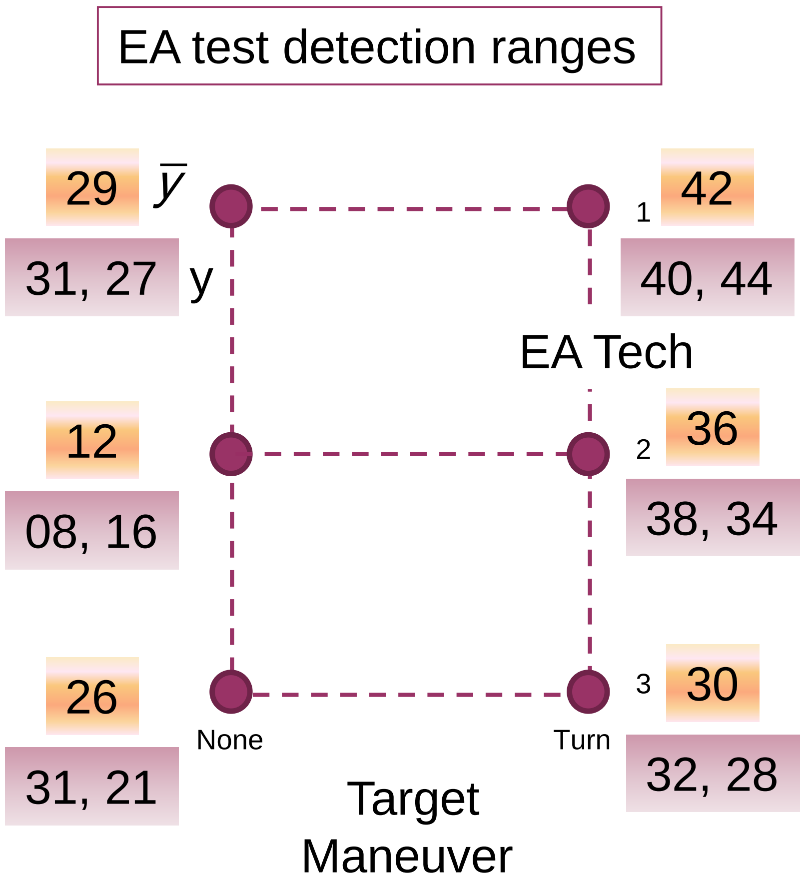
ANOVA Model Contrasts

- **As you know the coding of 2-level factors is -1, +1**
- **This coding is actually a contrast, a method for comparing different combinations of factor settings**
- **Contrasts are not unique, and some are better than others. Contrast coefficients must sum to 0**
- **Software: contrasts different for 2-level vs ≥ 3-level**

2-Level Factor A	β_{11} or A or X1
1	–1
2	+1

3-Level Factor B	β_{21} or B[1] or X2 1	β_{22} or B[2] or X2 2
1	+1	0
2	0	+1
3	–1	–1

Analysis – Signal: Grand Mean or Intercept



$$y_{ijk} = \beta_0 + \beta_{11}x_{11} + \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{1121}x_{11}x_{21} + \beta_{1122}x_{11}x_{22} + \varepsilon$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{N} \bar{y}_i}{N}$$

$$= (29 + 42 + 12 + \ldots 30)/6$$

$$= 29.2$$

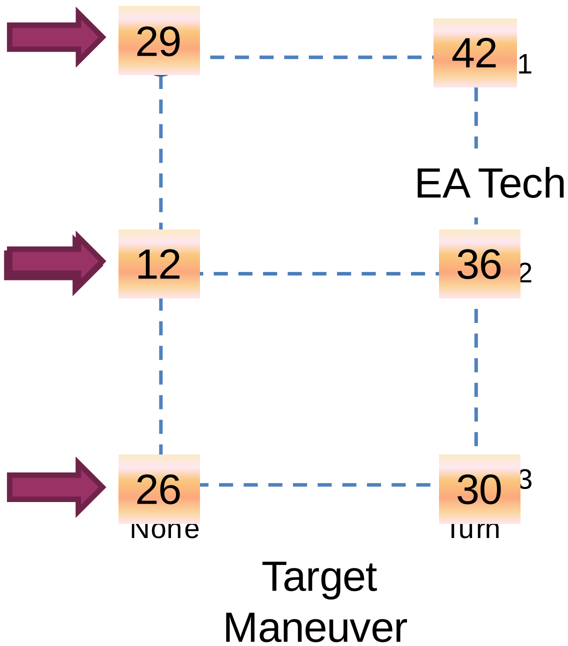
Analysis – 3-Level Main Effect

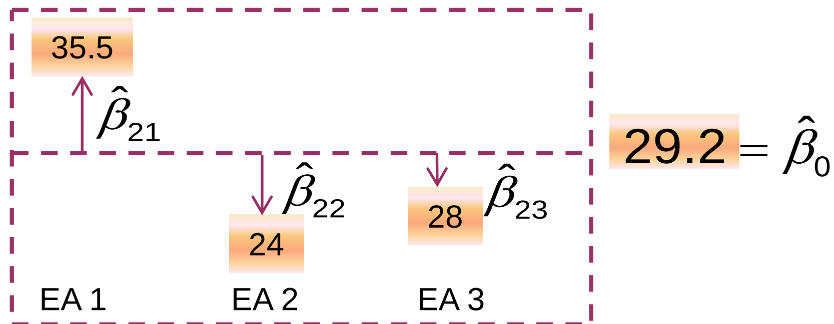
$$\hat{\beta}_{21} = \hat{\mu}_1 - \hat{\mu} = \hat{\mu}_1 - \hat{\beta}_0 = 35.5 - 29.2 = \mathbf{6.3}$$

$$\hat{\beta}_{22} = \{(12 + 36)/2\} - 29.2 = -\mathbf{5.2}$$

....

$$*\hat{\beta}_{23} = -1 * [\hat{\beta}_1 + \hat{\beta}_2] = -[6.3 - 5.2] = -1.1$$ * not in the model!





- **Factor has 3 levels and 2 parameters**
- **Each parameter estimate,** $\hat{\beta}_{2i} = \hat{\mu}_i - \hat{\mu}$
- **The level estimates from the parameters**

$$\hat{\mu}_1 = \hat{\mu} + \hat{\beta}_{21}$$

$$\hat{\mu}_2 = \hat{\mu} + \hat{\beta}_{22}$$

$$\hat{\mu}_3 = \hat{\mu} - \hat{\beta}_{21} - \hat{\beta}_{22}$$ Last level found using all coefficients

Statistical Model

General Model

$$y_{ijk} = \beta_0 + \beta_{11}x_{11} + \beta_{21}x_{21} + \beta_{22}x_{22} + \beta_{1121}x_{11}x_{21} + \beta_{1122}x_{11}x_{22} + \varepsilon$$

Coefficient Estimates

$$\hat{\beta}_0 = 29.2$$

$$\hat{\beta}_{11} = 6.8$$

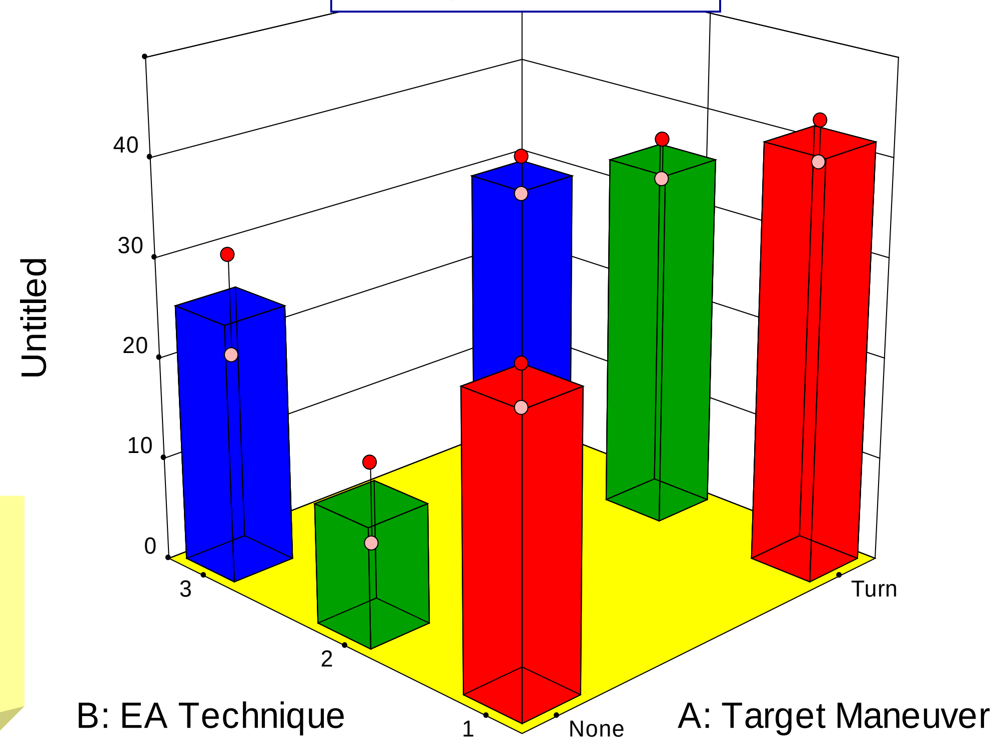
$$\hat{\beta}_{21} = 6.3$$

$$\hat{\beta}_{22} = -5.2$$

$$\hat{\beta}_{1121} = -0.3$$

$$\hat{\beta}_{1122} = 5.2$$

Response Plot



Note: All the parameter estimation complexity here is due to a categorical model – spikes in the battlespace. With continuous X variables we have first and second order slopes. Not required – preferred…

MULTI-LEVEL CATEGORICAL POWER

Categorical to Numeric Factors

- **Urge the planning team to consider numeric instead of categorical factors**

Factor	Categorical Levels	Numeric Factor	Numeric Levels
Weapon	GBU-10, GBU -16, GBU-12	Weapon Weight	500, 1000, 2000
Delivery	Loft, Level, Dive	Release Angle	+10, 0, -30 deg
Location	Eglin, Nellis	Visibility	5, 9 nm
Target Type	Car, Tractor Trailer	Target Size	60, 568 sq ft
Target Motion	Stationary, Moving	Target Speed	0, 30 mph
Time of Day	Day, Dusk, Night	Ambient Light	100, 500, 800 lumens
Range	Edge of LAR, Center of LAR	Range	5, 10 nm

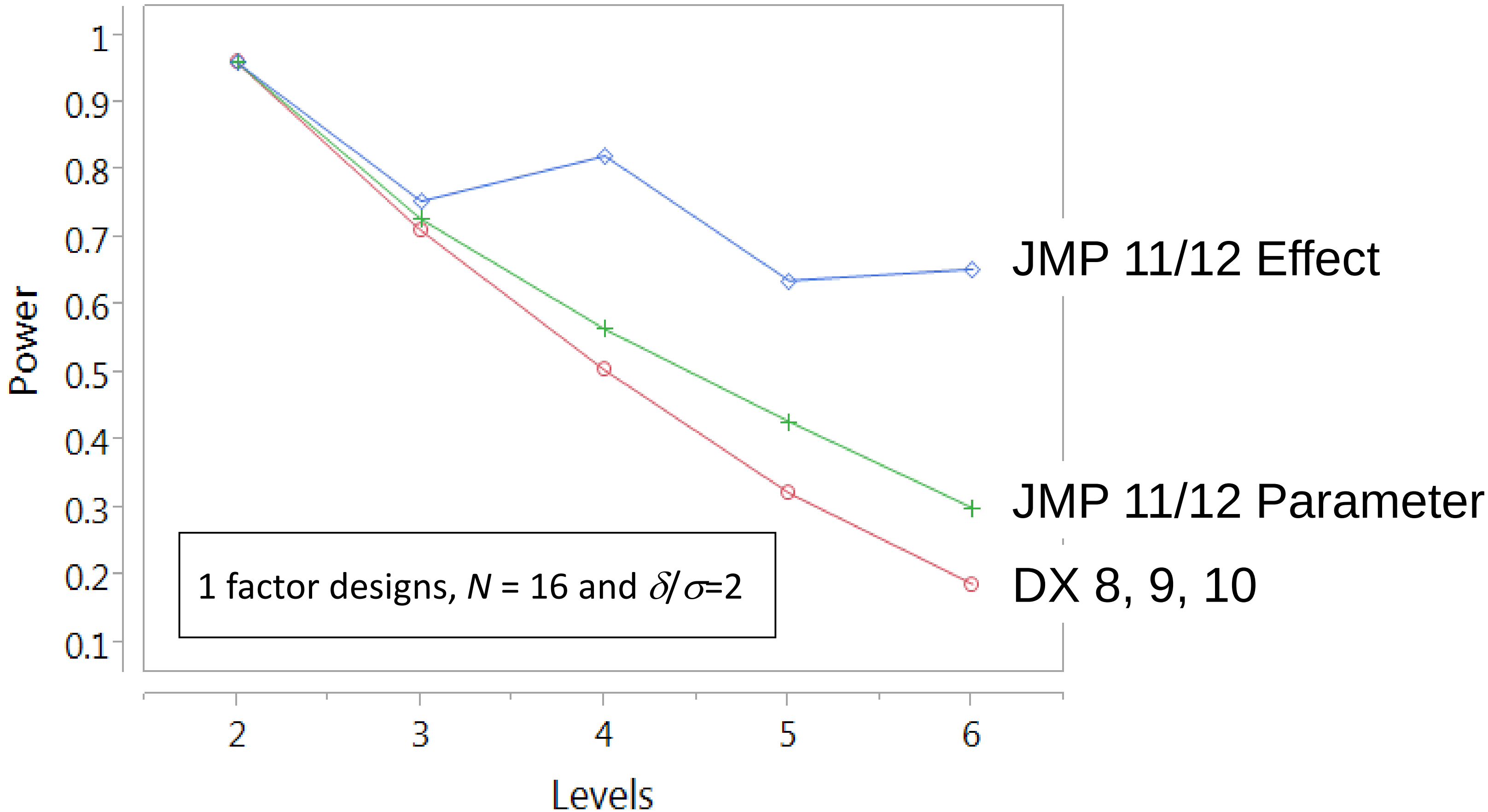
Issue with Multi-level Categorical Factors

- **Information is lost as the number of levels (q) increases**

Factor	Levels	Obs per level (N=20)	Obs per level (N=40)	Obs per level (N=60)
A	2	10	20	30
B	4	5	10	15
C	5	4	8	12
D	10	2	4	6

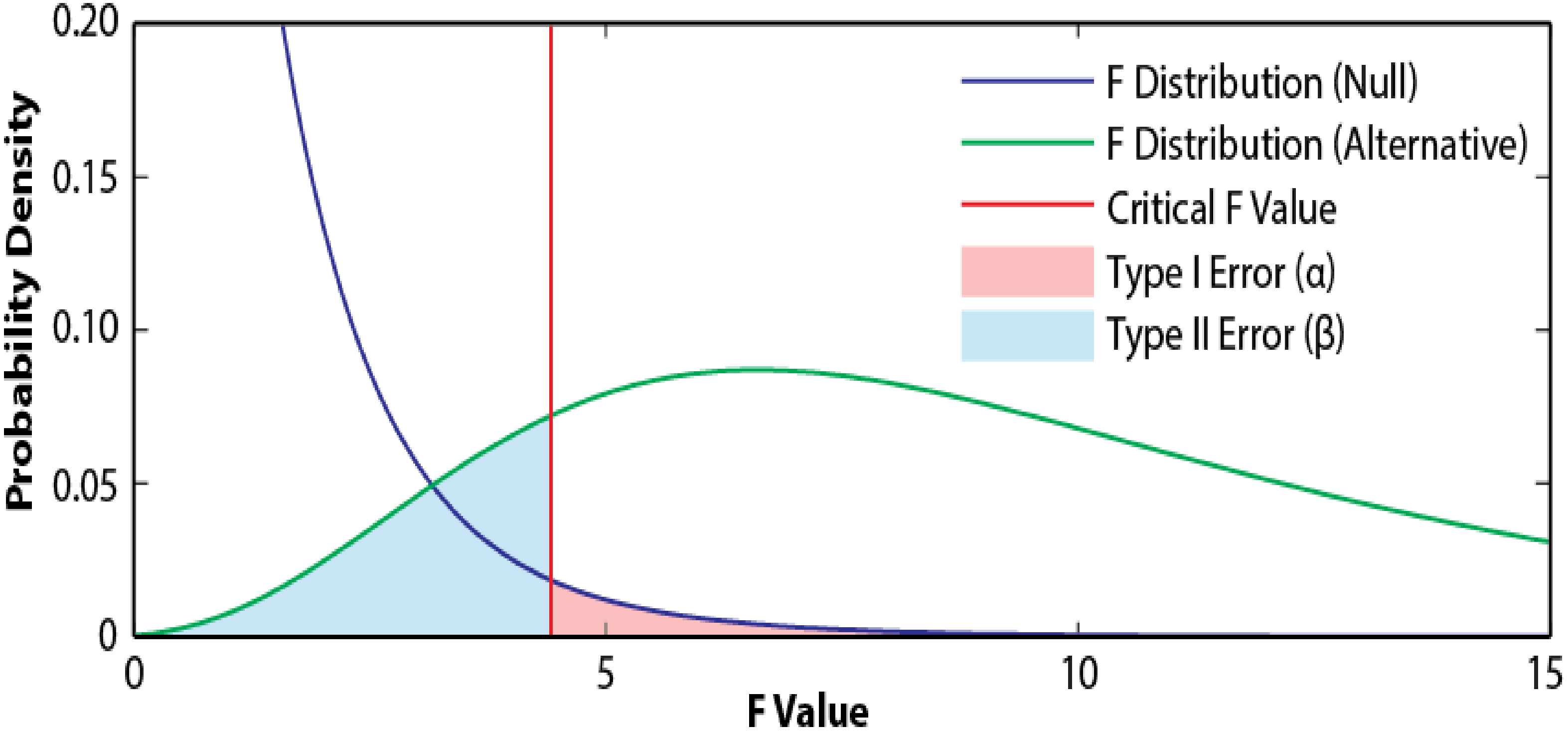
Power vs. Number of Levels

- **As the number of levels increases, power falls**
- **Less information per level for the same number of runs**



Power Calculations

- **Similar to the 2-level case, in the more general multi-level categorical case, power is measured as an area under a non-central *F*-distribution**



How Power is Calculated

- F-distribution based on a ratio of variances; two F-distributions here one for the null (no effect), other for the alternative.
- F-distribution for the null is central F, one for the alternative is non-central with parameter λ, which offsets the F
- The larger the non-centrality parameter, the more the alternative is offset, the larger the area to the right of the critical value = power probability
- The non-centrality parameter is used to define the alternative F, so that the area under this alternative distribution to the right of the critical value (based on a area to the right of that value under the null hypothesis F) is the power

Power Calculations - Equation

- **Power for the** i**th effect** (P_i) **is** $P_i = 1 - \tilde{F}\{\acute{F}_i, g_i, N - p, \lambda_i\}$

- **Non-centrality parameter:** $\lambda = (\boldsymbol{Lb})^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}\boldsymbol{Lb}$
  - where $\boldsymbol{L}$ is a matrix used to isolate the subset of coefficients under test
  - $\boldsymbol{b}$ is the coefficient vector of size $p\ x\ 1$
  - $\boldsymbol{X}$ is the design matrix of size $N\ x\ p$, N is the number of runs, p is the number of parameters in the model
  - $\boldsymbol{L}$ and $\boldsymbol{b}$ are used to generate the effect size specified by the anticipated coefficients

- $\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T$ **contains the variances of the effect estimates. This is important because the design orthogonality affects power**

- **So multicollinearity adversely affects power, such that $\lambda \cong 0$ even if the effect size is truly large, giving power $\cong$ 0**

Power Example

- **A 5-level, 1-factor design with 3 replicates, or 15 runs**

Level	JMP Parameter	Coefficient Estimate
1	X1 1	1
2	X1 2	1
3	X1 3	-1.5
4	X1 4	1
5		-1.5*

- **The non-centrality parameter** $\lambda = (\boldsymbol{Lb})^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}\boldsymbol{Lb} = 22.5$
  - $\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$, $\boldsymbol{b} = [1 \quad 1 \quad 1 \quad -1.5 \quad 1]^{\boldsymbol{T}}$
  - *critical F value is calculated as* $\acute{F} = F^{-1}\{1-\alpha, q-1, n-p\} = F^{-1}\{1-0.05, 4, 15-5\} = 3.48$
  - *Power is then computed as* $P = 1 - \tilde{F}\{\acute{F}, q-1, n-p, \lambda\} = 1 - \tilde{F}\{3.48, 4, 15-5, 22.5\} = \mathbf{0.87}$

Multi-level Categorical Example

- **Example: Integrated Defense Electronic Counter Measure (IDECM)**
  – on board jammer system and an ALE-55 towed decoy

- **Factors**

A: Aircraft Type	Countermeasure (Maneuver)	Threat
F/A-18 E/F	Dry	AA1
F-15E	Wet (none)	AA2
B-1B	Wet (M1)	SA1
	Wet (M2)	SA2
		SA3
		SA4

- **Responses: Miss Distance, Miss/Hit**

Step-by-step Power Analysis

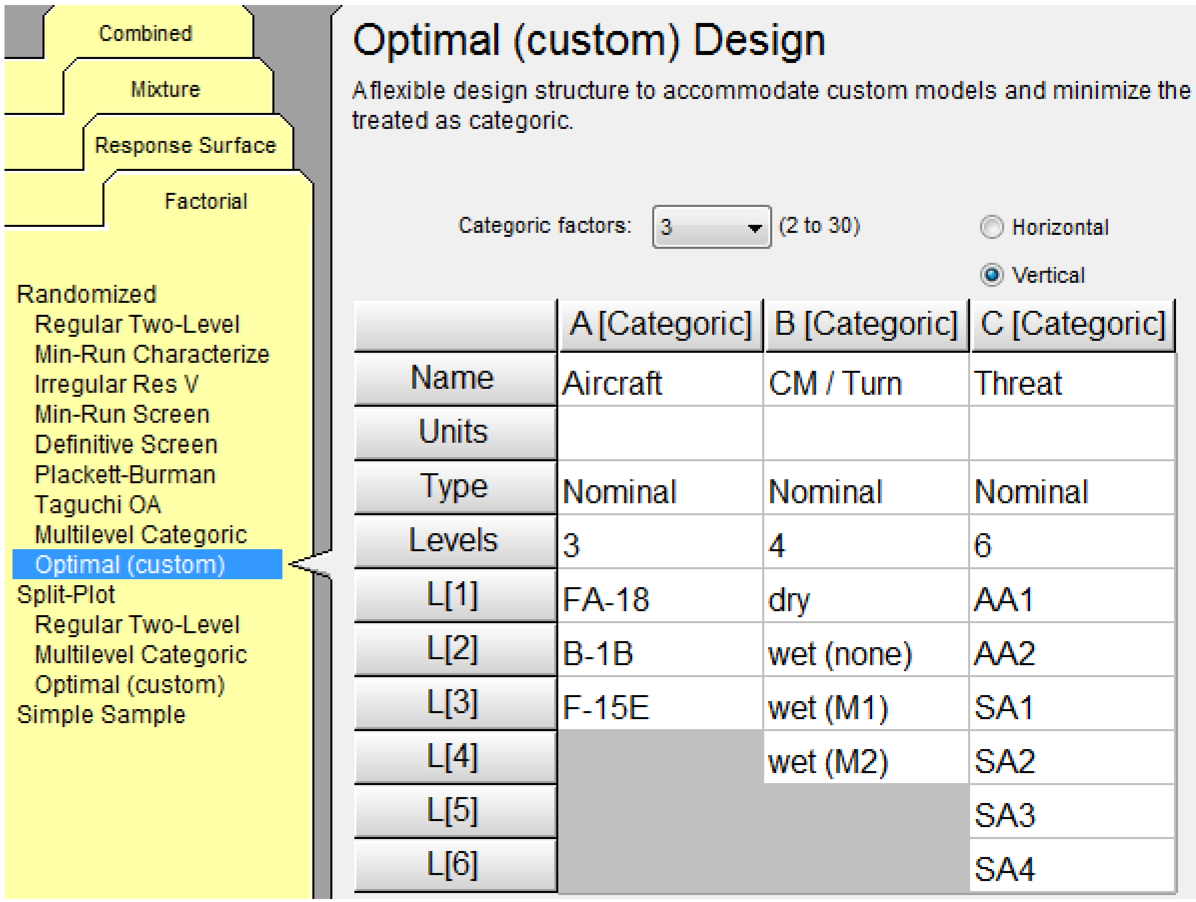
- **We will perform step-by-step Power Analysis using**
  - Design Expert
  - JMP 12
- **Stages of Power Analysis**
  - Power Parameter Estimates
  - Build Initial Design
  - Power Assessment
  - Design Modification and Re-assessment
  - Reporting Power
- **Also learn some capabilities of the software**

Power Parameters

Parameter	Description	IDECM Example
k: factors	Number of factors in the experiment	**3; $3^1 \times 4^1 \times 6^1$**
df_{error}: model error	Amount of data reserved for estimating system noise	**2 PE *df*** **Desired Model: ME + 2FI**
α: alpha	Probability (declaring factor matters when it doesn't)	**0.05**
δ: delta	Size of response change expert wants to detect	**20.0 ft**
σ: sigma	System noise – run-to-run variability or repeatability	**13.33 ft**
1-β: power	Probability of declaring a factor matters when it does	**solve**
N: test size		**31 initially**

Design Expert

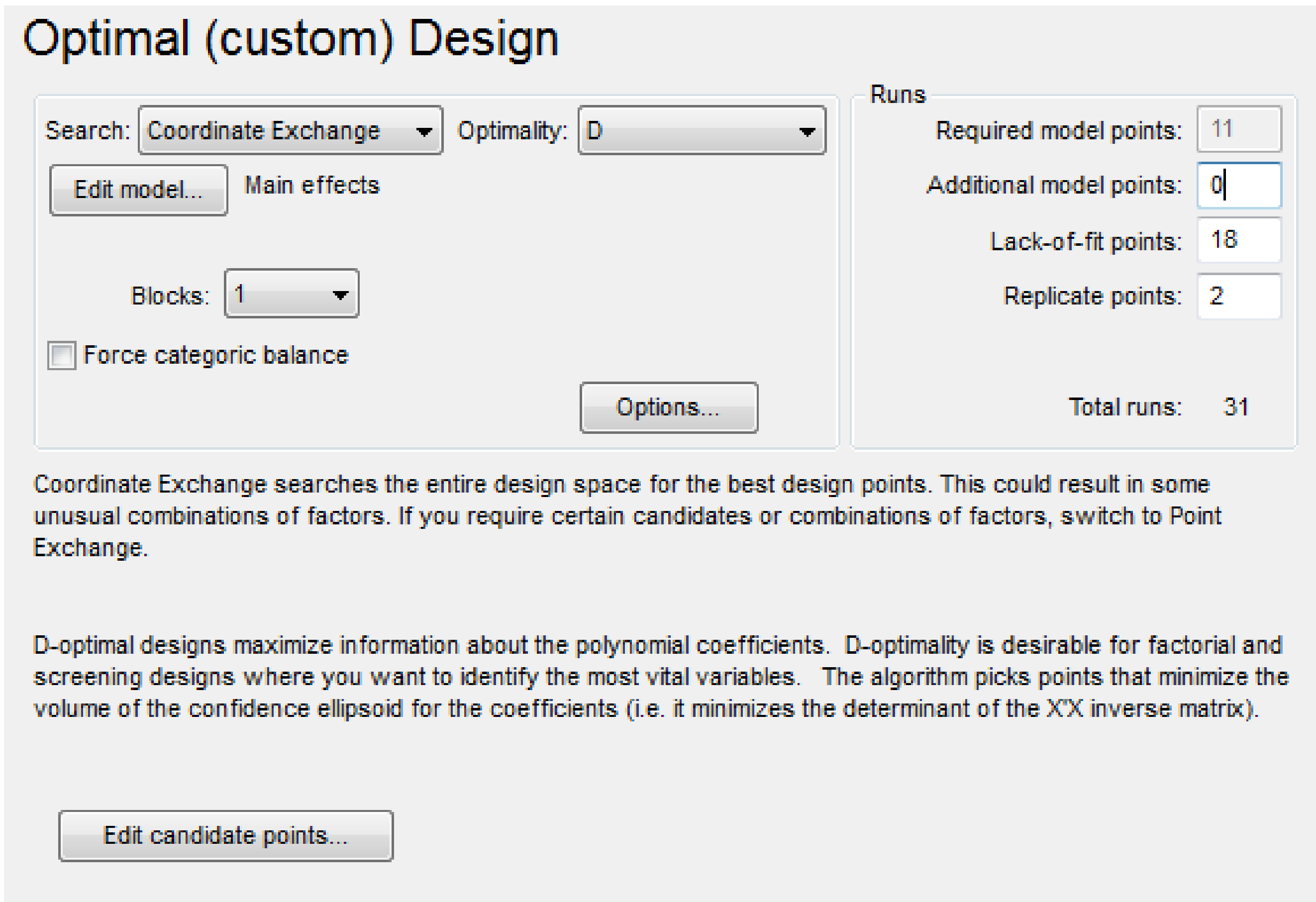
- **Initial Design Build – Optimal Factorial**



	A [Categoric]	B [Categoric]	C [Categoric]
Name	Aircraft	CM / Turn	Threat
Units			
Type	Nominal	Nominal	Nominal
Levels	3	4	6
L[1]	FA-18	dry	AA1
L[2]	B-1B	wet (none)	AA2
L[3]	F-15E	wet (M1)	SA1
L[4]		wet (M2)	SA2
L[5]			SA3
L[6]			SA4

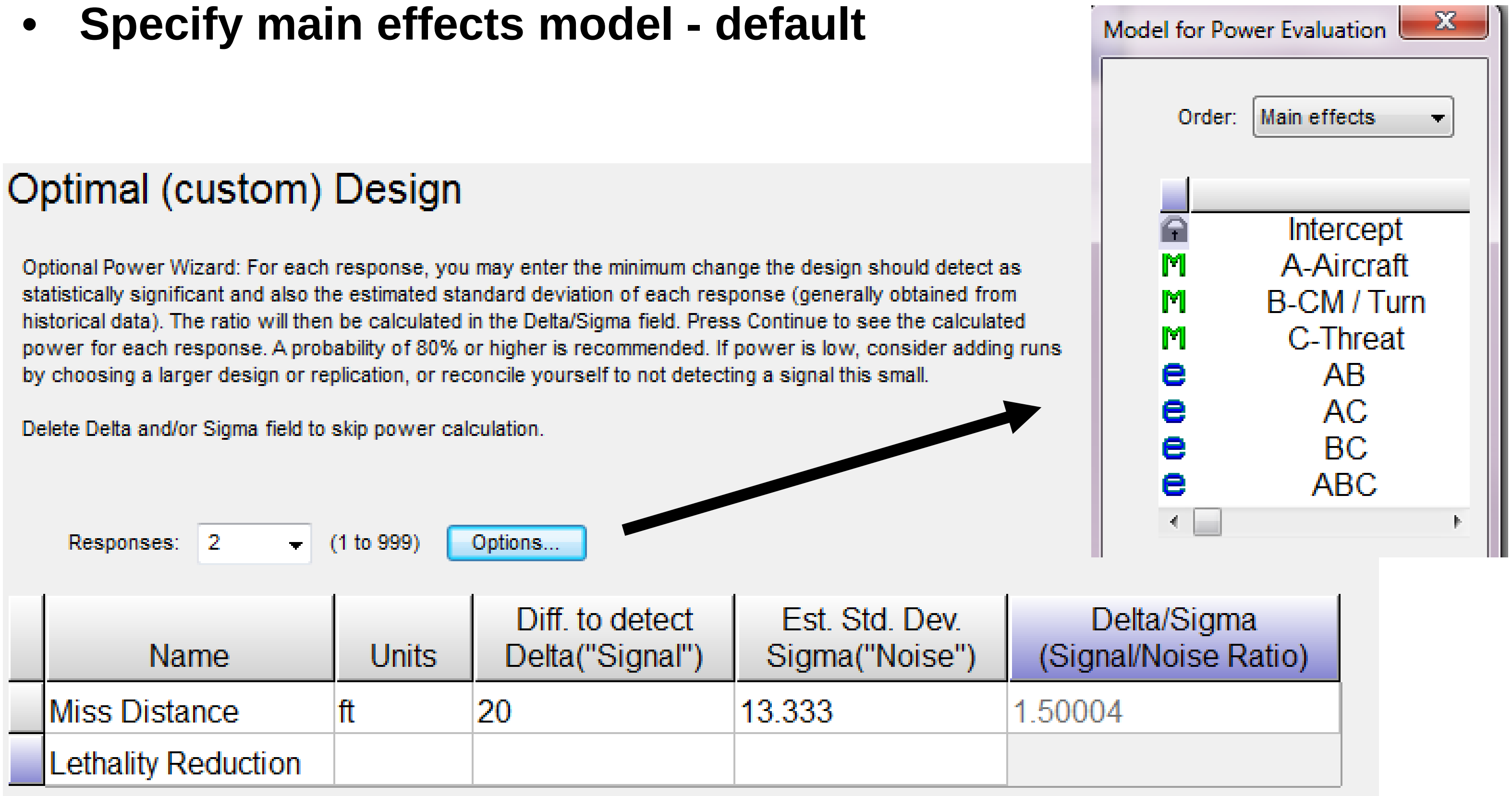
Specify Design Parameters

- **Customize the Design – Intended Model and Runs**



Power Parameters

- **Enter prescribed delta (δ) and estimated sigma (σ)**
- **Specify main effects model - default**



Power Report

- **Power reported per main effect**
- **Recall each factor has different number of levels (*q*)**

Optimal (custom) Design

Power is reported at a 5.0% alpha level to detect the specified signal/noise ratio.

Recommended power is at least 80%.

Miss Distance ft

Signal (delta) = 20.00 **Noise (sigma) = 13.33** **Signal/Noise (delta/sigma) = 1.50**

A	B	C
78.1 %	56.6 %	30.3 %

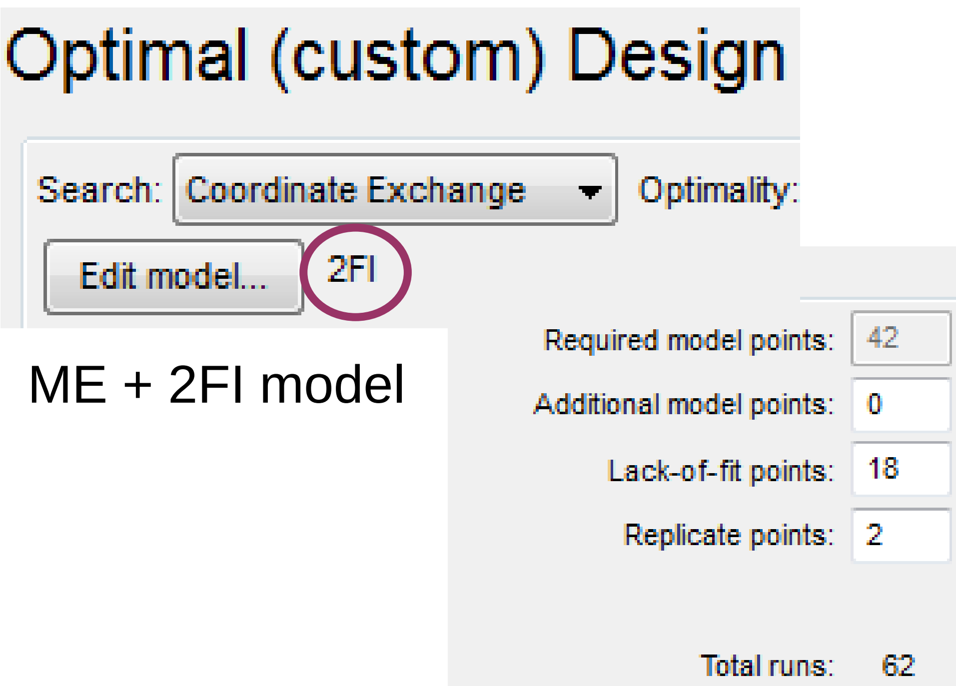
Initial Design

- **Design Runs**

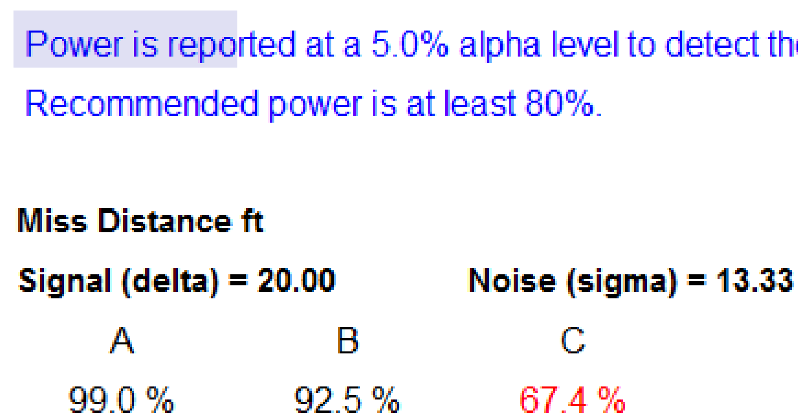
Select	Run	Factor 1 A:Aircraft	Factor 2 B:CM / Turn	Factor 3 C:Threat	Response 1 Miss Distance ft
	1	FA-18	wet (none)	AA2	
	2	B-1B	wet (none)	SA1	
	3	FA-18	wet (none)	AA1	
	4	F-15E	dry	AA2	
	5	F-15E	wet (M1)	AA1	
	6	F-15E	dry	SA3	
	7	F-15E	wet (M2)	SA2	
	8	F-15E	wet (none)	SA4	
	9	FA-18	wet (M2)	SA2	
	10	FA-18	dry	SA4	
	11	F-15E	dry	SA4	
	12	F-15E	wet (none)	SA3	
	13	FA-18	dry	AA2	
	14	FA-18	wet (M1)	SA1	
	15	B-1B	dry	AA1	
	16	B-1B	wet (none)	SA2	
	17	F-15E	wet (M2)	SA1	
	18	B-1B	wet (M2)	AA1	
	19	FA-18	wet (M2)	SA4	
	20	B-1B	wet (M1)	AA2	
	21	F-15E	wet (M2)	AA2	
	22	FA-18	dry	SA1	
	23	B-1B	wet (none)	SA1	
	24	F-15E	wet (M1)	AA1	
	25	B-1B	wet (M1)	SA4	
	26	B-1B	wet (M2)	SA3	
	27	B-1B	dry	SA2	
	28	FA-18	wet (M1)	SA3	
	29	F-15E	wet (M1)	SA2	
	30	FA-18	wet (none)	AA1	
	31	B-1B	wet (M2)	SA3	

Modify Design for Increased Power

- Based on the 31 run design, consider 2 alternatives: 46, 62 run
- For 46 run, choose ME + 2FI model, 2 LoF, 2 replicate runs
- For 62 run, choose ME + 2FI model, 18 LoF, 2 replicate runs



ME + 2FI model



Improved Power!

Design Alternative Comparison

- **Power is not the only metric, but is important**
- **Considering test objective, B (dry/wet) is the primary factor of interest, along with interaction BC (ECM success robust to threats)**

Metric	Design 1	Design 2	Design 3
Model Supported	ME + some 2FI	ME + 2FI	ME + 2FI
LoF df	n/a	2	18
PE df	2	2	2
Std error: B	0.34	0.27	0.23
BC	2.80	0.73	0.54
Power: 3-lvl	78	93	99
4-lvl	56	80	93
6-lvl	30	47	67
Runs	31	46	62

JMP 12

- **Initial Design Build – Custom Design**

Custom Design

Responses

Factors

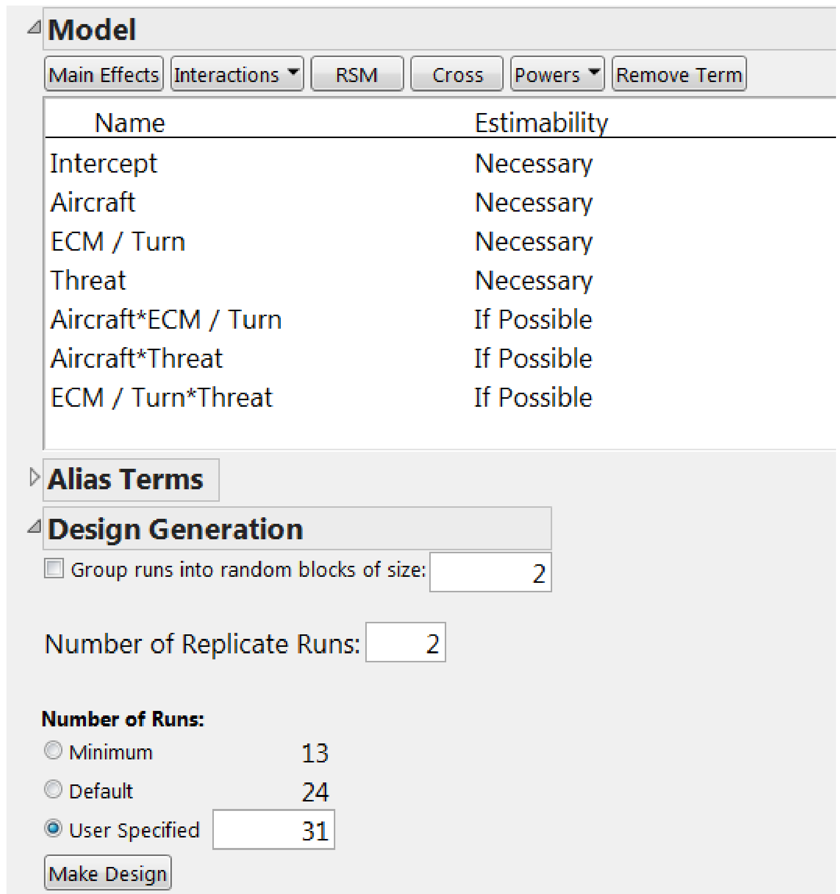
Add Factor | Remove | Add N Factors 1

Name	Role	Changes	Values					
Aircraft	Categorical	Easy	F/A-18	B-1B	F-15E			
ECM / Turn	Categorical	Easy	dry	wet (none)	wet (M1)	wet (M2)		
Threat	Categorical	Easy	AA1	AA2	SA1	SA2	SA3	SA4

IDA

JMP 12 Design Creation

- **Customize the Design – Intended Model and Runs**



JMP 12 Default Power

- **Set model for power**

1.

Model

Main Effects | Interactions | RSM | Cross | Powers | Remove Term

Name	Estimability
Intercept	Necessary
Aircraft	Necessary
ECM / Turn	Necessary
Threat	Necessary
Aircraft*ECM / Turn	If Possible
Aircraft*Threat	If Possible
ECM / Turn*Threat	If Possible

2.

Model

Main Effects | Interactions | RSM | Cross | Powers

Name	Estimability
Intercept	Necessary
Aircraft	Necessary
ECM / Turn	Necessary
Threat	Necessary

3.

Apply Changes to Anticipated Coefficients

- **Default power**

Design Evaluation

Power Analysis

Significance Level 0.05

Anticipated RMSE 1

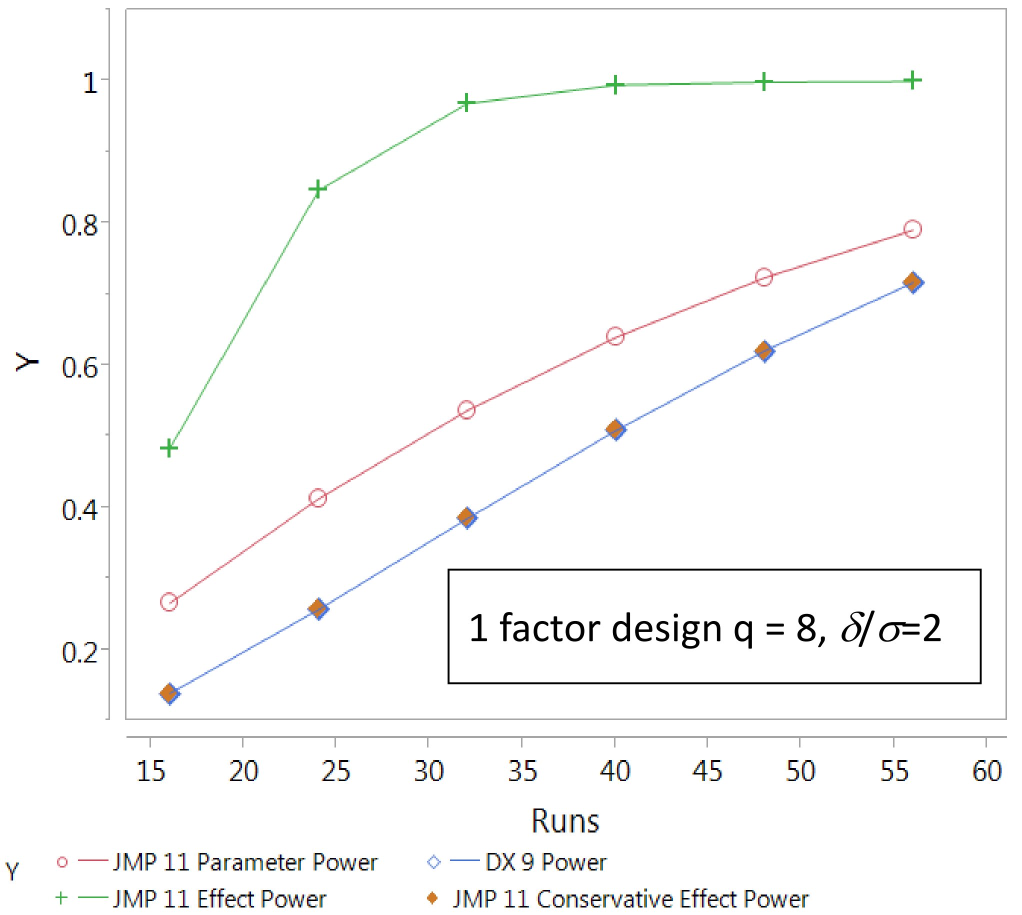
Parameter	Anticipated Coefficients	Power
Intercept	1	0.999
Aircraft 1	1	0.946
Aircraft 2	-1	0.946
ECM / Turn 1	1	0.868
ECM / Turn 2	-1	0.818
ECM / Turn 3	1	0.839
Threat 1	1	0.627
Threat 2	-1	0.622
Threat 3	1	0.636
Threat 4	-1	0.63
Threat 5	1	0.621

Apply Changes to Anticipated Coefficients

Effect	Power
Aircraft	0.958
ECM / Turn	0.988
Threat	0.963

Default JMP 11/12 Power and Another Option

- **Clearly the power estimates differ depending on your choice**
- **JMP 11 Conservative Power agrees with other software**



JMP 11 Conservative Power

- **Set Delta for Power**

Please Enter a Number

Choose value for delta. Anticipated coefficients will be half of this value: 1.5

OK | Cancel

Design Evaluation

Power Analysis

Significance Level: 0.05

Anticipated RMSE: 1

Parameter	Anticipated Coefficients	Power
Intercept	0.75	0.976
Aircraft 1	0.75	0.778
Aircraft 2	-0.75	0.778
ECM / Turn 1	0.75	0.616
ECM / Turn 2	-0.75	0.62
ECM / Turn 3	0.75	0.591
Threat 1	0.75	0.405
Threat 2	-0.75	0.457
Threat 3	0.75	0.407
Threat 4	-0.75	0.407
Threat 5	0.75	0.407

2 minimum (Aircraft 1, Aircraft 2)

1 minimum (ECM / Turn 3)

2 minimum (Threat 1 to Threat 4)

Apply Changes to Anticipated Coefficients

- **Edit Anticipated Coefficients**

Design Evaluation

Power Analysis

Significance Level: 0.05

Anticipated RMSE: 1

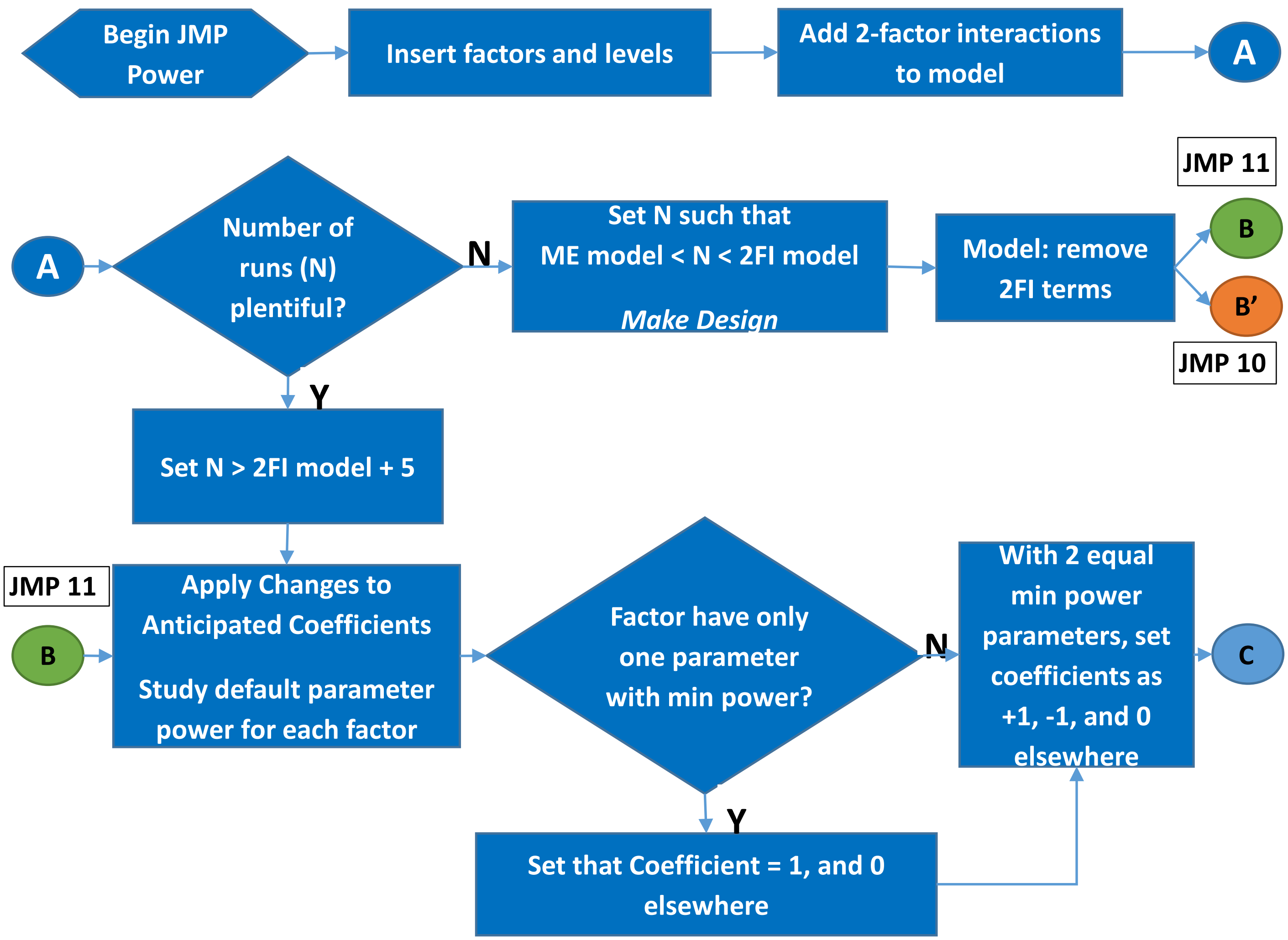
Parameter	Anticipated Coefficients	Power
Intercept	0.75	0.976
Aircraft 1	0.75	0.778
Aircraft 2	-0.75	0.778
ECM / Turn 1	0	0.05
ECM / Turn 2	0	0.05
ECM / Turn 3	0.75	0.591
Threat 1	0.75	0.405
Threat 2	0	0.05
Threat 3	0	0.05
Threat 4	-0.75	0.407
Threat 5	0	0.05

Apply Changes to Anticipated Coefficients

Effect	Power
Aircraft	0.781
ECM / Turn	0.566
Threat	0.304

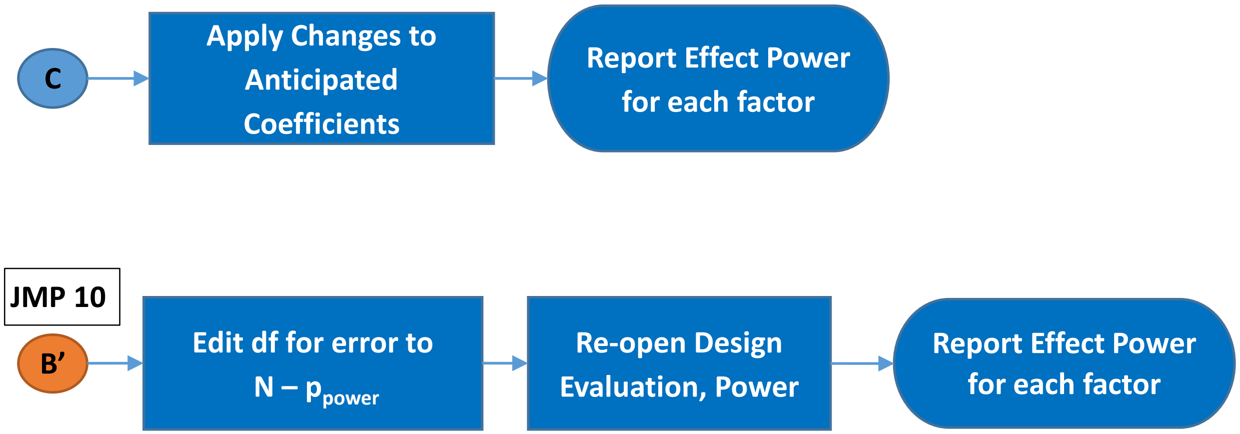
Report Effect Power

JMP 11 Conservative Power – Process Flow



JMP 11 Conservative Power – Process Flow

- **Continued**



AFOTEC Script for JMP 11/12

- **Conservative Power Script for JMP**

Power Summary - Multi-level Categorical

- Statistical power for categorical factors with > 2 levels requires an additional decision or assumption be made regarding the nature of the factor effect.
- Because each of the factor levels can be thought to stand on their own, a common modeling approach used is indicator variables.
- For a factor of this type, one must decide how many levels are active, assuming that the effect is real.
- Standard approaches historically (and currently in JMP 9/10 and DX) for active levels is to assume the most conservative scenario with only a pair of levels different by d.
- Conservative power is reported by default in JMP 9/10 and DX, whereas JMP 11/12 allows the user to specify the factor level effects.

Summary - Multi-level Categorical (cont'd)

- JMP 11/12 power analysis is purposefully adapted to provide the user flexibility in tailoring effect power for categorical factors with more than 2 levels.
- JMP 11/12 default anticipated coefficients make all factor levels active (with coefficient $\delta/2$), except the last level for factors with odd numbered levels.
- JMP 11/12 anticipated coefficients can be structured fairly easily for most conservative effect power.
- It is highly recommended, that for consistent reporting across software platforms, that users of JMP 11/12 configure the anticipated coefficients for most conservative power.

POWER FOR BINARY RESPONSES

Converting Binary Response to SNR

- **Binary response are 0, 1 outcomes, like pass/fail or detect/no-detect**
- **Use a binomial underlying distribution as the number of detects of *n, so a proportion p is used***
- **Binomial power requires we specify nominal *p*, p_1, and δ is how big of a change we wish to detect, p_2**
- **Three methods, logistic (logit) shown here**
  - **Model** $y^* = \ln\left(\frac{\pi}{1-\pi}\right) = \mathbf{X\beta}$
  - **Delta in transformed scale** $\delta = \left| ln\left(\frac{p_1}{1-p_1}\right) - ln\left(\frac{p_2}{1-p_2}\right) \right|$
  - **Sigma** $\sigma = \sqrt{n\bar{p}(1-\bar{p})} = \sqrt{\bar{p}(1-\bar{p})}$

SNR Method Comparison

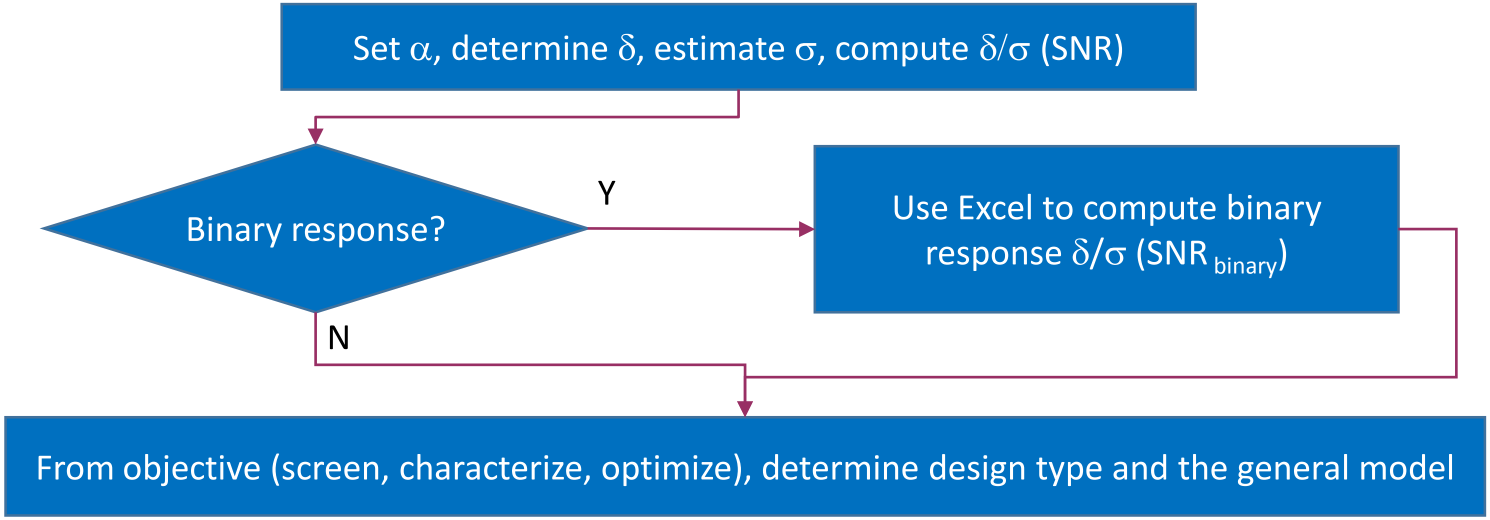
- **Results similar, but Normal or Arcsin more conservative**

p	D	SNR (arcsin)	SNR (logit)	SNR (normal)
0.9	0.100	0.3444	0.3630	0.3333
0.85	0.100	0.2838	0.2896	0.2801
0.8	0.100	0.2518	0.2544	0.2500
0.75	0.100	0.2320	0.2334	0.2309
0.7	0.100	0.2189	0.2198	0.2182
0.65	0.100	0.2102	0.2107	0.2097
0.6	0.100	0.2045	0.2050	0.2041
0.55	0.100	0.2014	0.2017	0.2010
0.5	0.100	0.2003	0.2007	0.2000
0.45	0.100	0.2014	0.2017	0.2010
0.4	0.100	0.2045	0.2050	0.2041
0.35	0.100	0.2102	0.2107	0.2097
0.3	0.100	0.2189	0.2198	0.2182
0.25	0.100	0.2320	0.2334	0.2309
0.2	0.100	0.2518	0.2544	0.2500
0.15	0.100	0.2838	0.2896	0.2801
0.1	0.100	0.3444	0.3630	0.3333

- **Note the difference in magnitude compared to SNR of 1 or 2**

Binary Response Power

- **If a response has two possible outcomes (e.g. miss/hit) it is a binary response and must be addressed separately to find the δ/σ or SNR**
- **The process involves finding SNR using Excel**



Binary Response Calculation

- **Assume a nominal success probability = 0.90 or 90% and a desire to detect a difference of 0.10 or 10%, and use confidence and power thresholds = 0.90 or 90%**

User Inputs

P(success)	0.9
Δ =	0.1
Confidence	0.9
Power	0.9

Method 2: Signal to Noise Calculations

Signal to Noise (Arcsin method)	0.344
Signal to Noise (Logit method)	0.363
Signal to Noise (Normal method)	0.333

Design Build for Binary Response

- **Iterate on the design size (using number of complete replicates) until desired power achieved**

		Power (%)		
Reps	***N* (hit/miss)**	**Aircraft**	**ECM**	**Threat**
1	72	16	11	8
3	216	41	27	15
5	360	63	44	24
7	504	78	59	34
10	720	91	76	47
15	1080	99	92	67

***N* (miss dist)**
46
62

Runs for equivalent power if miss distance response

Power Analysis Summary

- **Power is only one of the design goodness metrics, albeit important in characterization**
- **Both risks of wrong conclusions are handled directly, first set α then iterate on β**
- **Both risks are prior probabilities – assessments made before the test is conducted. After the test, it is difficult to retrospectively determine whether incorrect conclusions have been drawn**
- **Power is computed using area under H_1 using a non-central t- or F-reference distribution**

Power Summary – cont'd

- **Power depends on N, α risk, δ, σ, k, df_{error}**
- **Higher power values are desired, and while designs can be under-powered,** *right-sized* **or over-powered, we usually strive to right-size a test, left alone would be under-powered**
- **Continuous responses are vastly more informative than categorical responses, especially binary responses**
- **Power is only one of many design metrics, but one of the more important indicators of test design sufficiency**
- **Because many parameters need to be estimated in a power analysis, reported precision is usually at the decile level (e.g. 90% vs. 80%)**
- **Suggested Reference: Freeman, L. J., Johnson, T. H., and Simpson, J. R., "Power Analysis Tutorial for Experimental Design Software," *IDA Technical Document D-5205*, Nov 2014**