Determination of Power for Complex Experimental Designs

Pat Whitcomb
Stat-Ease, Inc.
(612) 746-2036
pat@statease.com

Gary W. Oehlert
School of Statistics
University of Minnesota
(612) 625-1557
gary@umn.edu

DATAWorks 2018
Mini Tutorial

Two Part Agenda

- Part 1: Power
  - Two-level factorials
  - Multilevel factorials
  - Response surface designs
- Part 2: Precision
  - Response surface designs
Agenda Part 1: Power

- **Power**
  - Two-level factorials
    - Explanation of power
    - More runs: Pot pie example
    - Smaller sigma: Stent example
    - Split Plot designs
    - Binary response
    - Multilevel factorials
    - Response surface designs

Sizing Factorial Designs

**Power:**

- During screening and characterization the emphasis is on identifying factor effects.
- What are the important design factors?
- For this purpose power is an ideal metric to evaluate design suitability.
What is Power?
No Factor Effect; $H_0: \Delta = 0$

**Power** = $(1-\beta)\times 100\%$

*Power* is the probability of revealing an active effect of size delta ($\Delta$) relative to the noise ($\sigma$) as measured by signal to noise ratio ($\Delta/\sigma$).

It should be high (at least 80%!) for the effect size of interest.

<table>
<thead>
<tr>
<th>Truth:</th>
<th>Effect?</th>
<th>ANOVA says:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Type II Error (beta)</td>
<td><em>Failure to detect</em></td>
</tr>
<tr>
<td>No</td>
<td>Retain $H_0$</td>
<td>OK 😊</td>
</tr>
<tr>
<td></td>
<td>Reject $H_0$</td>
<td>Type I Error (alpha) <em>False Alarm</em></td>
</tr>
</tbody>
</table>
Factorial Design – Power
$2^3$ Full Factorial $\Delta=2$ and $\sigma=1$

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Diff. to detect Delta (&quot;Signal&quot;)</th>
<th>Est. Std. Dev. Sigma (&quot;Noise&quot;)</th>
<th>Delta/Sigma (Signal/Noise Ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>57.2%</td>
<td>57.2%</td>
<td>57.2%</td>
</tr>
</tbody>
</table>

Power is reported at a 5.0% alpha level to detect the specified signal/noise ratio.
Power should be approximately 80% or greater for the effects you want to detect.

\[
\text{noncentral } t_{\alpha=0.05, df=4} \text{ with noncentrality parameter of } 2.828
\]

Power = 57.2%
Power
How Much is Enough?

Not enough power:
- An experiment that is so small that we are unlikely to detect effects of interest is wasteful of resources.

Too much power:
- A too-large (and thus “too powerful”) experiment may also be wasteful, because it uses more resources than are likely to be needed for detecting the alternative.

The sweet spot for power is:
- 80% for screening and characterization.
- 95 – 99% for verification.

Factorial Design – Power
Two Replicates of $2^3$ Full Factorial $\Delta=2$ and $\sigma=1$

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Diff. to detect Delta(&quot;Signal&quot;)</th>
<th>Est. Std. Dev. Sigma(&quot;Noise&quot;)</th>
<th>Delta/Sigma (Signal/Noise Ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>95.6%</td>
<td>95.6%</td>
<td>95.6%</td>
</tr>
</tbody>
</table>

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

Power should be approximately 80% or greater for the effects you want to detect.
Factorial Design – Power
Two Replicates of $2^3$ Full Factorial $\Delta=2$ and $\sigma=1$

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

Power = 95.6%

Power depends on:

- The size of the difference $\Delta$: the larger the difference the higher the power.
- The size of the experimental error $\sigma$: the smaller $\sigma$ the higher the power.
- The $\alpha$ risk chosen: the larger $\alpha$ the higher the power (and more false alarms).
- Choose design appropriate to the problem: more orthogonal and larger designs have more power.
- The number of replicates: the more runs the higher the power.
Power
What do I enter for Delta and Sigma?

**Picking delta (Δ) the size of the difference:**
- What is important, e.g. when will you spend money or make a change.
- Remember smaller Δ means less power or more runs.

**Estimating sigma (σ) the size of the experimental error:**
- Historical data, e.g. control chart, records, etc.
- Literature, patents, journals, etc.
- Experience with similar processes.
- Remember smaller is better; is there a way to reduce σ? *Blocking, better process control, more precise measurement, etc.*

---

Power
Know the Basis *(page 1 of 2)*

(In coded units)

A ME coefficient (β) is equal to one-half of the factorial effect (Δ).

\[ \beta = \frac{1}{2} \Delta \]

Programs calculate power for one or the other.

Not knowing the basis (effect or coefficient) can cause confusion!
When calculating power for a 2-level factorials using either an effect or a coefficient is fine as long as you know what the basis is.

For multi-level factorials and response surface designs effects are less ambiguous than coefficients.

Design-Expert® software calculates power using effects.

Agenda Part 1: Power

- Power
  - Two-level factorials
    - Explanation of power
    - More runs: Pot pie example
  - Smaller sigma: Stent example
  - Split Plot designs
  - Binary response
  - Multilevel factorials
  - Response surface designs
A pot pie producer wants to determine the effects of process parameters on dough temperature:

- It is important to control dough temperature during processing. The goal is to keep the dough at 70°F. Even a few degrees difference in processing temperature can degrade the bottom crust of the pot pie.
- The producer wants to determine the effects of batch size, of scrap quantity and ice quantity on dough temperature.

1. Identify opportunity and define objective.
   The objective is to determine the effects of batch size, of scrap quantity & ice quantity on dough temperature.

2. State objective in terms of measurable responses.
   - Define the change ($\Delta y$) that is important to detect.
     A difference of 3 degrees Fahrenheit is of interest; $\Delta y \approx 3^\circ F$.
   - Estimate experimental error ($\sigma$) for each response. Historical data is used to estimate the standard deviation; $\sigma = 2.2^\circ F$.
   - Use the signal-to-noise ratio ($\Delta y/\sigma$) to estimate power.
     $\Delta y/\sigma = 3.0/2.2 = 1.36$
3. Select the input factors to study and what levels to test. (Levels chosen determine the size of \( \Delta y \!\) )

<table>
<thead>
<tr>
<th>Factor</th>
<th>(-1) level</th>
<th>(+1) level</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Batch size</td>
<td>590 lbs</td>
<td>750 lbs</td>
</tr>
<tr>
<td>B. Scrap Qty</td>
<td>0 lbs</td>
<td>340 lbs</td>
</tr>
<tr>
<td>C. Ice Qty</td>
<td>0 lbs</td>
<td>40 lbs</td>
</tr>
</tbody>
</table>

4. Select a design (a full 2\(^3\) two-level factorial) and evaluate:

- Aliases (fractional factorials and/or blocked designs)  
  \[ \text{Not an issue with this design choice (running all combinations).} \]

- All factor combinations for safety and reasonability (likelihood of producing meaningful information).  
  \[ \text{Assume the team knows from subject matter expertise and actual range-finding tests that all runs will be do-able and informative.} \]

- Power (ideally at least 80% probability for detection).  
  \[ \text{See following statistics on main-effect estimates (anticipating sparsity of effects)} \]
Dough Temperature
Replicated Factorial Design

For this study, the experimenter wants to replicate this $2^3$ design, but how many is enough? Evaluating “Power” gives us the answer!

Set up a three-factor full factorial design ($2^3$, 8 runs).

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Type</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>A [Numeric]</td>
<td>Batch Size</td>
<td>lbs</td>
<td>Numeric</td>
<td>590</td>
</tr>
<tr>
<td>B [Numeric]</td>
<td>Scrap Qty</td>
<td>lbs</td>
<td>Numeric</td>
<td>0</td>
</tr>
<tr>
<td>C [Numeric]</td>
<td>Ice Qty</td>
<td>lbs</td>
<td>Numeric</td>
<td>0</td>
</tr>
</tbody>
</table>

Dough Temperature
Evaluating Power

Dough temperature: $\Delta = 3.0$, $\sigma = 2.2$, $\Delta/\sigma = 1.36$

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Diff. to detect</th>
<th>Est. Std. Dev.</th>
<th>Delta/Noise Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dough Temp</td>
<td>deg F</td>
<td>3</td>
<td>2.2</td>
<td>1.36364</td>
</tr>
</tbody>
</table>

Design Power (1 replicate, 8 runs)

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dough Temp</td>
<td>deg F</td>
<td>3</td>
<td>2.2</td>
<td>1.36364</td>
<td>31.7%</td>
<td>31.7%</td>
<td>31.7%</td>
</tr>
</tbody>
</table>

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

Power should be approximately 80% or greater for the effects you want to detect.

Want power of at least 80% for effects of interest!
Dough Temperature
Evaluating Power

Design Power (2 replicates, 16 runs)

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dough Temp</td>
<td>deg F</td>
<td>3</td>
<td>2.2</td>
<td>1.36364</td>
<td>70.7%</td>
<td>70.7%</td>
<td>70.7%</td>
</tr>
</tbody>
</table>

Power is reported at a 5.0% alpha level to detect the specified signal/noise ratio.
Power should be approximately 80% or greater for the effects you want to detect.

Design Power (3 replicates, 24 runs)

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dough Temp</td>
<td>deg F</td>
<td>3</td>
<td>2.2</td>
<td>1.36364</td>
<td>88.8%</td>
<td>88.8%</td>
<td>88.8%</td>
</tr>
</tbody>
</table>

Power is reported at a 5.0% alpha level to detect the specified signal/noise ratio.
Power should be approximately 80% or greater for the effects you want to detect.

Power to detect main effects at 5% alpha:
- 8 unique design runs: 31.7% 😊
- 16 runs – 2nd replicate of original 2^3: 70.7% 😊
- 24 runs – 3rd replicate: 88.8% 😊

Calculating power allows us to determine how many replicates are necessary to give a high probability of finding effects that are large enough to be of interest.

It’s sure beats guessing!
MR5 Design
Stent Delivery System

A stent is a wire mesh tube used to prop open an artery that's recently been cleared using angioplasty. The stent is collapsed to a small diameter over a balloon catheter. It's then moved into the area of the blockage.

When the balloon is inflated, the stent expands, locks in place and forms a scaffold. This holds the artery open.

The stent stays in the artery permanently, holding it open to improve blood flow to the heart muscle.
1. Identify opportunity and define objective.
   Relate stent safety and deliverability to process factors.

2. State objective in terms of measurable responses.
   Safety is quantified by Burst pressure. Deliverability is quantified by Pushability and Trackability. Want to estimate 2FI model.
   a. Define the change ($\Delta y$) that is important to detect for each response. $\Delta_{\text{Burst}} = 6$ psig, $\Delta_{\text{Push}} = 15$ g/cm and $\Delta_{\text{Track}} = 10$ g*cm.
   b. Estimate error ($\sigma$):
   c. Calculate signal to noise:
      
      | Response | $\sigma$ | $\Delta/\sigma$ |
      |----------|---------|----------------|
      | Burst    | 7 psig  | 0.857          |
      | Push     | 28 g/cm | 0.536          |
      | Track    | 6 g*cm  | 1.67           |

3. Select the input factors to study (there are ten):

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Type</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Balloon wall</td>
<td>in</td>
<td>Numeric</td>
<td>0.0011</td>
<td>0.0015</td>
</tr>
<tr>
<td>B Waist length</td>
<td>mm</td>
<td>Numeric</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C Waist wall</td>
<td>in</td>
<td>Numeric</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>D Wing fold</td>
<td>#</td>
<td>Categoric</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>E Inner ID</td>
<td>in</td>
<td>Numeric</td>
<td>0.016</td>
<td>0.018</td>
</tr>
<tr>
<td>F Inner wall</td>
<td>in</td>
<td>Numeric</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>G Inner weld</td>
<td>cm</td>
<td>Numeric</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>H Inner length</td>
<td>mm</td>
<td>Numeric</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>J Tip length</td>
<td>mm</td>
<td>Numeric</td>
<td>2</td>
<td>4.2</td>
</tr>
<tr>
<td>K Tip material</td>
<td>durometer</td>
<td>Numeric</td>
<td>43</td>
<td>69</td>
</tr>
</tbody>
</table>
4a Select a design:

- Evaluate aliases (fractional factorials and/or blocked designs)  
  *During build*

- Evaluate power (desire power > 80% for effects of interest)  
  *During build (default order is main effects)*

- Examine the design layout to ensure all the factor combinations are safe to run and are likely to result in meaningful information (no disasters)

**A Minimum-Run Resolution V (MR5)**
irregular fraction requiring 56 runs was used.

---

### MR5 Design: Power
Stent Delivery System

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Diff. to detect Delta(“Signal”)</th>
<th>Est. Std. Dev. Sigma(“Noise”)</th>
<th>Delta/Sigma (Signal/Noise Ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burst</td>
<td>psig</td>
<td>6</td>
<td>7</td>
<td>0.857143</td>
</tr>
<tr>
<td>Push</td>
<td>g/cm</td>
<td>15</td>
<td>28</td>
<td>0.535714</td>
</tr>
<tr>
<td>Track</td>
<td>g*cm</td>
<td>10</td>
<td>6</td>
<td>1.66667</td>
</tr>
</tbody>
</table>

**Design Power**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Burst</td>
<td>psig</td>
<td>6</td>
<td>7</td>
<td>0.85714</td>
<td>87.9%</td>
<td>87.7%</td>
<td>87.4%</td>
<td>87.7%</td>
<td>87.4%</td>
<td>87.7%</td>
<td>87.4%</td>
<td>87.7%</td>
<td>87.4%</td>
<td>87.7%</td>
<td>87.4%</td>
</tr>
<tr>
<td>Push</td>
<td>g/cm</td>
<td>15</td>
<td>28</td>
<td>0.53571</td>
<td>49.9%</td>
<td>49.3%</td>
<td>49.1%</td>
<td>49.5%</td>
<td>49.7%</td>
<td>49.3%</td>
<td>49.5%</td>
<td>49.7%</td>
<td>49.3%</td>
<td>49.1%</td>
<td>49.5%</td>
</tr>
<tr>
<td>Track</td>
<td>g*cm</td>
<td>10</td>
<td>6</td>
<td>1.66667</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

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

Power should be approximately 80% or greater for the effects you want to detect.
Power is low (~50%) for Push; to increase power:

1. Increase design size: a $2^{10-3}$ resolution V design (128 runs) gives adequate power (85.2%), but there are too many runs to be practical.

2. Increase $\Delta_{\text{Push}} = 15 \text{ g/cm}$: No – we are interested in a difference of 15 g/cm.

3. Decrease $\sigma_{\text{Push}} = 28 \text{ g/cm}$: By partitioning the variance, we determine that the push measurement contributes most (75%) of the variation. Repeating the test (not the experimental run) to reduce $\sigma$ is the answer.

See next slide.

---

$\sigma_{\text{Push}} = 28$ & $\sigma_{\text{Push}}^2 = \sigma_{\text{Process}}^2 + \sigma_{\text{Measurement}}^2$

$784 = 200 + 584 \quad \therefore \quad \%\text{Contribution} = \frac{584}{784} \approx 75\%$

Make three independent push measurements for each run. Enter the average of the measurements as the response:

Then by the CLT $\left( \sigma_{\text{Average}}^2 = \frac{\sigma_{\text{Measurement}}^2}{n} \right)$:

$\sigma_{\text{Push}}^2 = 200 + \frac{584}{3} = 395 \quad \%\text{Contribution} \approx 50\%$

$\sigma_{\text{Push}} = \sqrt{395} \approx 19.9 \quad \frac{\Delta}{\sigma} = \frac{15}{19.9} = 0.75$
Go back to the responses and enter $\sigma_{\text{Push}} = 19.9$:

### Design Power

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
<th>Power for D</th>
<th>Power for E</th>
<th>Power for F</th>
<th>Power for G</th>
<th>Power for H</th>
<th>Power for I</th>
<th>Power for J</th>
<th>Power for K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burst</td>
<td>psig</td>
<td>6</td>
<td>7</td>
<td>0.857143</td>
<td>87.9%</td>
<td>87.4%</td>
<td>87.3%</td>
<td>87.6%</td>
<td>87.7%</td>
<td>87.4%</td>
<td>87.6%</td>
<td>87.7%</td>
<td>87.4%</td>
<td>87.4%</td>
<td></td>
</tr>
<tr>
<td>Push</td>
<td>g/cm</td>
<td>15</td>
<td>19.9</td>
<td>0.753769</td>
<td>78.6%</td>
<td>77.9%</td>
<td>77.8%</td>
<td>78.2%</td>
<td>78.4%</td>
<td>78.0%</td>
<td>78.2%</td>
<td>78.4%</td>
<td>78.0%</td>
<td>77.9%</td>
<td></td>
</tr>
<tr>
<td>Track</td>
<td>g*cm</td>
<td>10</td>
<td>6</td>
<td>1.66667</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td></td>
</tr>
</tbody>
</table>

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

Power should be approximately 80% or greater for the effects you want to detect.

Close enough to 80%

**Summary:**

- Replicating runs will reduce the system error; from both process and measurement.

- Repeating the measurement reduces only the measurement error.

- The magnitude of each of these errors and the relative cost of replicating runs versus repeating measurements dictates which will give the most “bang” for your buck.
Agenda Part 1: Power

- Power
  - Two-level factorials
    - Explanation of power
    - More runs: Pot pie example
    - Smaller sigma: Stent example
  - Split Plot designs
    - Binary response
    - Multilevel factorials
    - Response surface designs

Problem
Often in designed experiments some of the factors, e.g., temperature, are more difficult or expensive to vary than others. In some cases conducting a completely randomized design isn’t practical.

Solution
Restrict the randomization so it is practical to conduct the design. If you have to sort a factor to make a DOE easier to run, this restriction in randomization results in a “split-plot” design.

Are there factors in your process that are difficult to randomize?
The “split-plot” design originated in the field of agriculture. Experimenters applied one treatment (e.g. herbicide) to a large area of land, called a “whole plot” and other treatments (e.g. crop) to smaller areas of land within the whole plot called a “subplot”.

Split-Plot Designs:

- Split plots have two types of factors: “Hard-to-change” (HTC) and “Easy-to-change” (ETC).
- The randomization of the HTC factor is restricted.
- Split plots naturally arise in many DOE studies.
- Building and analyzing a split-plot design is tricky, unless you have a good DOE package.

**Luckily, split plots are easy using modern DOE software!**
Wind tunnel experiment:

- Two HTC factors:
  - Offset Gap
  - Deflection angle

- Three ETC factors:
  - Attack angle
  - Sideslip angle
  - Reynolds number
Factors:

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Change</th>
<th>Type</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gap</td>
<td>Hard</td>
<td>Numeric</td>
<td>-1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Deflection</td>
<td>Hard</td>
<td>Numeric</td>
<td>-1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Attack angle</td>
<td>Easy</td>
<td>Numeric</td>
<td>-1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sideslip angle</td>
<td>Easy</td>
<td>Numeric</td>
<td>-1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Reynolds</td>
<td>Easy</td>
<td>Numeric</td>
<td>-1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Split-Plot Power:

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Diff. to detect</th>
<th>Est. Subplot</th>
<th>Delta/Sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coeff. of Lift</td>
<td>200</td>
<td>150</td>
<td>1.33333</td>
<td></td>
</tr>
</tbody>
</table>

DATAWorks 2018

Split-Plot Design Power

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for a</th>
<th>Power for b</th>
<th>Power for C</th>
<th>Power for D</th>
<th>Power for E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coeff. of Lift</td>
<td>200</td>
<td>212.132</td>
<td>0.942809</td>
<td>72.8%</td>
<td>72.8%</td>
<td>72.8%</td>
<td>72.8%</td>
<td>72.8%</td>
<td></td>
</tr>
</tbody>
</table>

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

Power should be approximately 80% or greater for the effects you want to detect.

Randomized Design Power

The power of an equivalent completely randomized design.

The power(s) shown below are not for the current design: It is provided only for comparison to the Split-Plot case.

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)*</th>
<th>Signal/Noise</th>
<th>Power for a</th>
<th>Power for b</th>
<th>Power for C</th>
<th>Power for D</th>
<th>Power for E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coeff. of Lift</td>
<td>200</td>
<td>212.132</td>
<td>0.942809</td>
<td>72.8%</td>
<td>72.8%</td>
<td>72.8%</td>
<td>72.8%</td>
<td>72.8%</td>
<td></td>
</tr>
</tbody>
</table>

* Adjusted for randomized design using the variance ratio.
Split Plot Power
Split Plot versus CRD

**Disadvantage:**
- Less power for HTC factors. This is the price for restricting the randomization.

**Advantage:**
- Easier to run, randomization of HTC factors is restricted.
- Higher precision estimating ETC factors.

\[
\sigma_{ETC} = 150 \rightarrow \sigma_{\text{subplot}} = 150
\]
\[
\text{variance ratio } \left( \frac{\sigma_{ETC}^2}{\sigma_{HTC}^2} \right) = 1 \rightarrow \sigma_{HTC} = 150
\]
\[
\sigma_{\text{whole plot}} = \sqrt{150^2 + 150^2} = 212
\]
Binary Response

Background

Binary response:

- Binary response is when the test result is either Pass or Fail; i.e. it is one of two complementary outcomes.

- Measured by the proportion \( p \) defective in a sample of parts \( n \) made during each DoE run:

\[
p = \frac{\text{#defective}}{n}
\]

- Need to approximate SNR:

![Image of proportion power averaging options]

DATAWorks 2018 45
**Binary Signal to Noise Approximation Methods**

**Normal Transformation**

\[
\delta = |p_1 - p_2| \\
p_1 = \bar{p} + \frac{\Delta}{2} \\
p_2 = \bar{p} - \frac{\Delta}{2} \\
\sigma = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}} \\
\text{SNR} = \frac{\delta}{\sigma}
\]

**Arcsine Transformation**

\[
\delta = \arcsin \left( \sqrt{\frac{\bar{p} + \Delta}{2}} \right) - \arcsin \left( \sqrt{\frac{\bar{p} - \Delta}{2}} \right) \\
\sigma = \frac{1}{\sqrt{4n}} \\
\text{SNR} = \frac{\delta}{\sigma}
\]

**Logit Transformation**

\[
\delta = \ln \left( \frac{p_1}{1 - p_1} \right) - \ln \left( \frac{p_2}{1 - p_2} \right) \\
p_1 = \bar{p} + \frac{\Delta}{2} \\
p_2 = \bar{p} - \frac{\Delta}{2} \\
\sigma = \frac{1}{\sqrt{n\bar{p}(1 - \bar{p})}} \\
\text{SNR} = \frac{\delta}{\sigma}
\]

**Continuous versus Proportion Power**

2⁵ Full Factorial (SNR = 1, \(\bar{p} = 0.10 \& \Delta = 0.02\))

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Response Type</th>
<th>Diff. to detect (Delta(&quot;Signal&quot;))</th>
<th>Est. Std. Dev. (Sigma(&quot;Noise&quot;))</th>
<th>Samples per Run</th>
<th>Proportion (p-bar)</th>
<th>Run Variation (% of p-bar)</th>
<th>Delta/Sigma (SNR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Continuous</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>R2</td>
<td>proportion</td>
<td>0.02</td>
<td>226</td>
<td>0.1</td>
<td>2</td>
<td>0.99862</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

226 samples/run are required to achieve an equivalent SNR of 1

**Continuous Response Power**

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
<th>Power for D</th>
<th>Power for E</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>77.7%</td>
<td>77.7%</td>
<td>77.7%</td>
<td>77.7%</td>
<td>77.7%</td>
</tr>
</tbody>
</table>

**Proportion Response Power**

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Proportion</th>
<th>Run Variation</th>
<th>Samples/Run</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
<th>Power for D</th>
<th>Power for E</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2</td>
<td>proportion</td>
<td>0.02</td>
<td>0.1</td>
<td>2</td>
<td>226</td>
<td>0.99862</td>
<td>77.6%</td>
<td>77.6%</td>
<td>77.6%</td>
<td>77.6%</td>
<td>77.6%</td>
</tr>
</tbody>
</table>

(32 runs) (226 samples/run) = 7,232 total samples for the DoE
Continuous versus Proportion Data

Problems with Proportion Data

- It takes a ton of data to estimate a small probability or small probability change.

- For binary data, the variance of each outcome is a function of its mean. The only way to get a more precise value is to take more data; you cannot reduce the variance of its measurement.

- Less sensitive than continuous; i.e. only two complementary outcomes versus a continuous scale.
  - If you have no successes you have no information.

- Fewer options for analysis compared to continuous data.

- Little indication of sources of variability.

  Continuous response measurement is always preferred over binary!

Agenda Part 1: Power

- Power
  - Two-level factorials
    - Explanation of power
    - More runs: Pot pie example
    - Smaller sigma: Stent example
  - Split Plot designs
  - Binary response
  - Multilevel factorials
  - Response surface designs
Multilevel Categoric General Factorials

$2^3 \Rightarrow 2\times2\times2$ Factorial

2x3x4 Factorial

When fitting a main effects model:

- How do we define power for factors having more than two levels?
- Will factors A, B and C have the same power?

Multilevel Categoric Power Calculations
Recall:

Power is the probability of revealing an active effect of size delta (Δ) relative to the noise (σ) as measured by signal to noise ratio (Δ/σ).
Multilevel Categoric – Power
2x3x4 Factorial $\Delta=40$ and $\sigma=20$

Probability of revealing $\Delta$, if the largest $\Delta = 40$

Power for B = 91.5%

Power for C = 73.8%

Multi-Level Factorial
Power is Unambiguous for $\Delta$

Same Fit $\Rightarrow$ Same $\Delta$!

Different Coefficients

Nominal contrasts

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Intercept</td>
<td>7.33</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>A[1]</td>
<td>-3.33</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td>A[2]</td>
<td>2.67</td>
</tr>
</tbody>
</table>

Ordinal contrasts

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>Intercept</td>
<td>7.33</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>A[1]</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>A[2]</td>
<td>-1.33</td>
</tr>
</tbody>
</table>
Combining Categoric Factors
CWA detection Case Study

Goal: Prove effectiveness of a remote detection system for a chemical-warfare agent (CWA).

Factors and levels:

A. CWA (Threshold, Objective)
B. Interferent (None, Burning diesel, Burning plastic)
C. Time (Day, Night)
D. Distance (1 km, 3 km, 5 km)
E. Environment (Desert, Tropical, Arctic, Urban, Forest)
F. Season (Summer, Winter)
G. Temperature (High, Low)
H. Humidity (High, Low)

Looking at the last four factors, there are too many combinations (40) and not all of these combinations are meaningful:

- E. Environment (Desert, Tropical, Arctic, Urban, Forest)
- F. Season (Summer, Winter)
- G. Temperature (High, Low)
- H. Humidity (High, Low)

*Not meaningful:* Tropical with low temperature and low humidity  
Arctic with high temperature and high humidity

Solution is to combine these four categorical factors into one.
Combining Categoric Factors
CWA detection Case Study

The eight most meaningful combinations:

<table>
<thead>
<tr>
<th>Background</th>
<th>Temperature</th>
<th>Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desert Winter</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Desert Summer</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Tropical</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Arctic</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Urban Winter</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Urban Summer</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Forest Winter</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Forest Summer</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

The 2-factor interaction model has 69 terms (including the intercept) add 5 lack-of-fit runs and you have a DoE with 74 runs.

Power for the main effects is:

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Delta (Signal)</th>
<th>Sigma (Noise)</th>
<th>Signal/Noise</th>
<th>Power for A</th>
<th>Power for B</th>
<th>Power for C</th>
<th>Power for D</th>
<th>Power for E</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td></td>
<td>40</td>
<td>20</td>
<td>2</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td>99.9%</td>
<td><strong>83.8%</strong></td>
</tr>
</tbody>
</table>

Remember: Factors having lots of levels have lower power!
Conclusions: Combining Categoric Factors

Advantages:

- Number of combinations reduced.
  - Combinations reduced by 80% (1,440 to 288) for CWA.
- All the combinations can be run and result in meaningful results.
- Can fit a full model.
  
  Preventing meaningless (or impossible) runs that result in missing data thereby creating an aliased model.

Agenda Part 1: Power

- Power
  - Two-level factorials
    - Explanation of power
    - More runs: Pot pie example
  - Smaller sigma: Stent example
  - Split Plot designs
  - Binary response
  - Multilevel factorials
  - **Response surface designs**
Response Surface Designs
Power of Two-Factor FCCD

Central Composite Design
Each numeric factor is set to 5 levels: plus and minus alpha (axial points), plus and minus 1 (factorial points) and the center point. If categorical factors are added, the central composite design will be duplicated for every combination of the categorical factor levels.

Numeric factors: 2
(2 to 50)
Categoric factors: 0
(0 to 10)

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Low</th>
<th>High</th>
<th>-alpha</th>
<th>+alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>A [Numeric]</td>
<td>Attack angle</td>
<td>deg</td>
<td>-6</td>
<td>4</td>
<td>-6</td>
</tr>
<tr>
<td>B [Numeric]</td>
<td>Sideslip angle</td>
<td>deg</td>
<td>-8</td>
<td>8</td>
<td>-8</td>
</tr>
</tbody>
</table>

Type: Full
Blocks: 1
Points:
Non-center points: 0
Center points: 5
alpha = 1

Std Error of Design

A: Attack angle

B: Sideslip angle
Response Surface Designs
Power of Two-Factor FCCD (13 runs)

Model Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Standard Error*</th>
<th>VIF</th>
<th>Rᵢ²</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.4082</td>
<td>1</td>
<td>0.0000</td>
<td>55.9 %</td>
</tr>
<tr>
<td>B</td>
<td>0.4082</td>
<td>1</td>
<td>0.0000</td>
<td>55.9 %</td>
</tr>
<tr>
<td>AB</td>
<td>0.5000</td>
<td>1</td>
<td>0.0000</td>
<td>40.8 %</td>
</tr>
<tr>
<td>A²</td>
<td>0.6017</td>
<td>1.16976</td>
<td>0.1451</td>
<td>81.2 %</td>
</tr>
<tr>
<td>B²</td>
<td>0.6017</td>
<td>1.16976</td>
<td>0.1451</td>
<td>81.2 %</td>
</tr>
</tbody>
</table>

* For a standard deviation of 1.

Power calculations are performed using response type "Continuous" and parameters:
Delta=2, Sigma=1
Power is evaluated over the -1 to +1 coded factor space.

Response Surface Designs
Power is Calculated for Equal Effects

RSM Equal Effects:

RSM Equal Coefficients:
Response Surface Designs
Power for Two-Factor FCCD \textit{(replicated 26 runs)}

Model Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Standard Error*</th>
<th>VIF</th>
<th>$R_i^2$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2887</td>
<td>1</td>
<td>0.0000</td>
<td>90.9%</td>
</tr>
<tr>
<td>B</td>
<td>0.2887</td>
<td>1</td>
<td>0.0000</td>
<td>90.9%</td>
</tr>
<tr>
<td>AB</td>
<td>0.3536</td>
<td>1</td>
<td>0.0000</td>
<td>76.7%</td>
</tr>
<tr>
<td>A²</td>
<td>0.4255</td>
<td>1.16976</td>
<td>0.1451</td>
<td>99.4%</td>
</tr>
<tr>
<td>B²</td>
<td>0.4255</td>
<td>1.16976</td>
<td>0.1451</td>
<td>99.4%</td>
</tr>
</tbody>
</table>

* For a standard deviation of 1.
Power calculations are performed using response type "Continuous" and parameters:
Delta=2, Sigma=1
Power is evaluated over the -1 to +1 coded factor space.

Adding more runs \textit{(replicating the FCCD)} increases power

Response Surface Designs
Power of Two-Factor Constrained Design

Let’s say the extremes of the factors are unattainable and the DoE space has to be restricted.

\textbf{Do these restrictions affect power?}
Build an I-optimal with multiple linear constraints:

- 6 Model points
- 5 Lack-of-fit points
- 5 Replicate points

Model Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Standard Error*</th>
<th>VIF</th>
<th>R²</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.5694</td>
<td>1.90469</td>
<td>0.4750</td>
<td>35.5%</td>
</tr>
<tr>
<td>B</td>
<td>0.5751</td>
<td>1.88174</td>
<td>0.4686</td>
<td>35.0%</td>
</tr>
<tr>
<td>AB</td>
<td>2.54</td>
<td>9.09444</td>
<td>0.8900</td>
<td>6.5%</td>
</tr>
<tr>
<td>A²</td>
<td>1.33</td>
<td>4.03542</td>
<td>0.7522</td>
<td>27.5%</td>
</tr>
<tr>
<td>B²</td>
<td>1.46</td>
<td>5.20911</td>
<td>0.8080</td>
<td>23.8%</td>
</tr>
</tbody>
</table>

* For a standard deviation of 1.

Power calculations are performed using response type "Continuous" and parameters:
Delta=2, Sigma=1
Power is evaluated over the -1 to +1 coded factor space.
**Response Surface Designs**

**Power for Two-Factor Constrained Design (26 runs)**

### Model Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Standard Error*</th>
<th>VIF</th>
<th>Rᵢ²</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.4403</td>
<td>2.26644</td>
<td>0.5588</td>
<td>58.0 %</td>
</tr>
<tr>
<td>B</td>
<td>0.4623</td>
<td>2.26951</td>
<td>0.5594</td>
<td>53.9 %</td>
</tr>
<tr>
<td>AB</td>
<td>1.74</td>
<td>11.2638</td>
<td>0.9112</td>
<td>8.5 %</td>
</tr>
<tr>
<td>A²</td>
<td>0.9005</td>
<td>4.14984</td>
<td>0.7590</td>
<td>56.1 %</td>
</tr>
<tr>
<td>B²</td>
<td>1.08</td>
<td>5.9073</td>
<td>0.8307</td>
<td>42.2 %</td>
</tr>
</tbody>
</table>

* For a standard deviation of 1.

Power calculations are performed using response type "Continuous" and parameters:

Delta=2, Sigma=1

Power is evaluated over the -1 to +1 coded factor space.

Adding more runs (10 additional model points) does little to increase power

---

**Response Surface Designs**

**Standard Error of Predicted Mean Response**

- **Replicated 26 run FCCD**
- **I-optimal 26 run Design**

DATAWorks 2018
Two Part Agenda

- Part 1: Power
  - Two-level factorials
  - Multilevel factorials
  - Response surface designs

- Part 2: Precision
  - Response surface designs

Response Surface Designs
Sizing for Precision

Response Surface Methodology

- When the goal is optimization (usually the case for RSM) emphasis is on the fitted surface.
- How well does the surface represent true behavior?
- For this purpose precision (measured using Fraction of Design Space, FDS) is a good metric to evaluate design suitability.
Sizing for Precision

To illustrate sizing for precision we will start with a simple one-factor example.

The solid center line is the fitted model; \( \hat{y} \) is the expected value or mean prediction.

The curved dotted lines are the computer generated confidence limits, or the actual precision.

\( d \) is the half-width of the desired confidence interval, or the desired precision. It is used to create the outer straight lines.

Note: The actual precision of the fitted value depends on where we are predicting.
One Factor Experiment
Sizing Linear Model

- Want a linear surface to represent the true response value within ± 0.90 with 95% confidence.
- The overall standard deviation for this response is 0.55.

Enter:

\[ \begin{align*}
\text{d} &= 0.90 \\
\text{s} &= 0.55 \\
\alpha (\alpha) &= 1 - 0.95 = 0.05
\end{align*} \]

\[ \approx 50\% \text{ of the design space is precise enough to predict the mean within } \pm \ 0.90. \]
Fraction of Design Space (FDS)
The Black Curve & The Blue Line

\[ SE_{\hat{y}_0} = s\sqrt{x_0^T (X^T X)^{-1} x_0} \]

 Normalize: \[ \frac{SE_{\hat{y}_0}}{s} = \sqrt{x_0^T (X^T X)^{-1} x_0} \]

Calculate \( SE_{\hat{y}_0} \) for 150,000 random points, sort and plot cumulation from zero (smallest) to one (largest).

\[ CI = \hat{y} \pm d \]

\[ d = t_{\alpha/2, df} SE_{\hat{y}} \]

\[ SE_{\hat{y}} = \frac{d}{t_{\alpha/2, df}} \]

Normalize: \[ \frac{SE_{\hat{y}}}{s} = \frac{d}{t_{\alpha/2, df} s} \]

\[ \frac{SE_{\hat{y}}}{s} = \frac{d}{t_{0.05, 3} s} = 0.9 \]

\[ \frac{3.18(0.55)}{s} = 0.514 \]

One Factor Experiment
Linear Model

50% of the design space has the desired precision, i.e. is inside the solid straight (red) lines.
Fraction of Design Space:

- Calculates the volume of the design space having a prediction variance (PV) less than or equal to a specified value.
- The ratio of this volume to the total volume of the design volume is the fraction of design space.
- Produces a single plot showing the cumulative fraction of the design space on the x-axis (from zero to one) versus the PV on the y-axis.
Fraction of Design Space – FDS
Replicated 26 run FCCD (δ=1, σ=1)

Fraction of Design Space – FDS
I-optimal 26 run Design (δ=1, σ=1)
Sizing for Precision
What Level of FDS is Good Enough?

How good is good enough? Rules of thumb:
- For exploration want FDS ≥ 80%
- For verification want FDS of 95-100%

What can be done to improve precision?
- Manage expectations; i.e. increase d
- Decrease noise; i.e. decrease s
- Increase risk of Type I error; i.e. increase alpha
- Increase the number of runs in the design

Choose design appropriate for the problem:
- Size the design for the precision required.

Conclusions

<table>
<thead>
<tr>
<th>Factorial DOE</th>
<th>Response Surface Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>During screening and characterization (factorials) emphasis is on identifying factor effects.</td>
<td>When the goal is optimization (usually the case for RSM) emphasis is on the fitted surface.</td>
</tr>
<tr>
<td>What are the important design factors?</td>
<td>How well does the surface represent true behavior?</td>
</tr>
<tr>
<td>For this purpose power is an ideal metric to evaluate design suitability.</td>
<td>For this purpose precision (FDS) is a good metric to evaluate design suitability.</td>
</tr>
</tbody>
</table>
References


Determination of Power for Complex Experimental Designs

Pat Whitcomb  
Stat-Ease, Inc.  
(612) 746-2036  
pat@statease.com

Gary W. Oehlert  
School of Statistics  
University of Minnesota  
(612) 625-1557  
gary@umn.edu

Thank You for Attending!