Optimal Verification Testing with Graphical Effects Analysis

Presented by Mark J. Anderson, PE, CQE, MBA
Stat-Ease, Inc., Minneapolis, Minnesota
mark@statease.com

Anecdote: Verifying tooth fairy
Optimal Verification Testing with Graphical Effects Analysis

Statistical Design of Experiments (DOE)

“A series of tests, in which purposeful changes are made to input factors, to identify causes for significant changes in the output responses.”

This talk will detail state-of-the-art tools for the design and analysis of verification tests that include both numeric and categoric factors.
Multi-Factor Testing (VS OFAT)
(Bearing life from accelerated test)

Factors:
A. Heat Treatment
B. Osculation
C. Material*
* Categorical

Relative efficiency = \( \frac{16}{8} \)
\( \bowtie 2 \text{ to } 1! \)

"To make knowledge work productive will be the great management task of this century."
-- Peter Drucker
Strategy of Experimentation

Focus on Verification

Screening

- Known Factors
- Unknown Factors

Characterization

- Factor effects and interactions
- Curvature?
  - yes
  - no

Optimization

- Response Surface Methods
  - yes
  - no

Verification

- Confirm?
  - yes
  - no
  - Backup

Celebrate!
Factorial Design Planning Process

1. State objective in terms of measurable responses. For each:
   a. Define the effect (response difference $\Delta y$) that is important to detect for each response. *(This is the signal, at a minimum, you are listening for.)*
   b. Estimate experimental error (response variation $\sigma$) for each response. *(The noise.)*

2. Select the input factors to study and establish their ranges. *(Wider the better for creating effects exceeding $\Delta y$).*

3. Select a design and evaluate it for:
   - Resolution of effects (beware of aliasing).
   - Power based on its signal to noise ratio ($\Delta y/\sigma$). *(For verification aim high: > 90% for every response.)*
   - Examine all runs for unsafe factor combinations. *(Pre-test any that may not work and/or create hazards.)*
Requisites of a Good Test Design*

- The test design matches the test objectives
- All the important responses are measured
- Factor ranges are practical
- Replication – measures experimental error
- Randomization – counteracts lurking variables
- Blocking – filters systematic variation
- Everyone is involved (teamwork)

Ideally these test matrices can accommodate any number of levels for any many factors, some or all of which might be categorical.

Typical problem*:

“We have 6 factors – 5 at 3 levels each and 1 at 2 levels. Each run is ~$600 (ouch). Our budget will likely support ~60 experimental runs, if needed – but I’d like to conduct this experiment in less runs and save costs.”

There are 486 possible combinations, which would cost almost $300,000 to perform. Is there a way to run only a fraction?

*(Correspondence to author on 2/8/10)
1. Specify a polynomial that you think is needed to get a decent approximation of the actual mechanism.
   - Do not overlook two-factor interactions (“2FIs”).

2. Select minimal points to estimate all coefficients in your design-for model.
   (Computer-based exchange algorithm.)

3. Consider augmenting the design with points for:
   - Replicates: To estimate pure error.
   - Lack-of-fit: To test how well the model represents actual behavior in our region of interest.
Criterion: D-optimal design minimizes the determinant of the \((X'X)^{-1}\) matrix. This minimizes the volume of the confidence ellipsoid for the coefficients and maximizes information about the polynomial coefficients.
Many are unclear on this difference!

<table>
<thead>
<tr>
<th>Important</th>
<th>Significant</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>🌼</td>
<td>😞</td>
</tr>
<tr>
<td>Yes</td>
<td>?️</td>
<td>⚠️</td>
</tr>
</tbody>
</table>

For example, let’s look at a two-level factorial verification test on a system that must be must not exceed 35 units of response due to factors varied within ranges that may be encountered in the field.
Half Normal Probability Paper

Sorting the vital few effects from the trivial many.

*Before looking at the two-level factorial verification test case, this vital tool for analysis must be explained.*

Significant effects: *The model terms!*

Negligible effects: *The error estimate!*

<table>
<thead>
<tr>
<th>Half-Normal % Probability</th>
<th>Standardized Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>5.38</td>
</tr>
<tr>
<td>10</td>
<td>10.75</td>
</tr>
<tr>
<td>20</td>
<td>16.13</td>
</tr>
<tr>
<td>30</td>
<td>21.50</td>
</tr>
<tr>
<td>50</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td></td>
</tr>
<tr>
<td>95</td>
<td></td>
</tr>
</tbody>
</table>

Edme section 1
Is anything statistically significant?

Warning! No terms are selected.

No!

Could largest effect be important?

Warning! No terms are selected.

No!

\(\Delta \) of 28.5

Must not exceed 35 units of response due to factors varied within ranges that may be encountered in the field.

Optimal Verification Testing with Graphical Effects Analysis
Verification Failed 😞: Not Significant, but Important (?)

**Is anything statistically significant?**

Warning! No terms are selected.

**Could largest effect be important?**

Yes! \( \Delta \) of 42.75

Optimal Verification Testing with Graphical Effects Analysis
Is anything statistically significant?

Yes!

Could largest effect be important?

No! ∆ of 22

Optimal Verification Testing with Graphical Effects Analysis
Verification Failed: Significant and Important

Is anything statistically significant?

Must not exceed 35 units

Could largest effect be important?

Optimal Verification Testing with Graphical Effects Analysis

Verification Failed: Significant and Important

Is anything statistically significant?

Must not exceed 35 units

Could largest effect be important?

Optimal Verification Testing with Graphical Effects Analysis
Hydraulic gear pumps are vital for many machines including vehicles and airplanes. However, they tend to lose efficiency due to internal leakage. OEM engineers must verify that one such device will remain within a specified range of performance regardless of normal production variations. They settle on nine factors, primarily categorical -- shown at right with number of levels each.

Based on subject-matter knowledge, the engineers are most concerned about interactions involving the float (factor C).

Customized pumps are costly, so a minimal-run design is desired.

Choose a d-optimal design for a reduced 2FI model: intercept, 9 main effects and 8 two-factor interactions (those involving C).

- 39 model points (builds) must be picked from the 3888 (3x3x3x2x2x2x2x3x3) possible combinations, in other words, approximately a 1/10th fraction.
- To estimate lack-of-fit pick 5 more unique combinations
- From these 44 builds, select 4 to replicated for pure error.

48 pumps will be built in total.

Differences in leak-back of 2 units are of interest. It must not exceed 10 units overall.
Pump Case

Will the design provided proper power?

1. Define the change ($\Delta y$) that is important. *Signal*
2. Estimate experimental error ($\sigma$). *Noise*
3. From signal-to-noise ratio ($\Delta y/\sigma$) estimate power.

*If runs suffice, the averaging provided by the matrix will cut the grass (noise) to reveal the snake (effect)!*

<table>
<thead>
<tr>
<th>Signal (delta) = 2.00</th>
<th>Noise (sigma) = 1.00</th>
<th>Signal/Noise (delta/sigma) = 2.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.6 %</td>
<td>99.6 %</td>
<td>99.6 %</td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>99.9 %</td>
<td>99.9 %</td>
<td>99.9 %</td>
</tr>
<tr>
<td>99.9 %</td>
<td>99.5 %</td>
<td>99.5 %</td>
</tr>
</tbody>
</table>

*Yes, power at 5% risk to see signal exceeds the 90% guideline for verification testing.*
**Pump Case Results**

Significant effects, but is the largest less than 10 leak-back units?  
We cannot tell with this graph because the effects are in a “normal” scale similar to a Z score.

This is not a two-level factorial. Some trickery is required (see Ref. #2) to plot multiple contrasts on a comparative scale.
Three significant main effects (one-factor), plus one two-factor interaction. =>

Does the predicted response range more than 10 units? If so, it fails verification.
Minimum Leak-Back

Difference < 10 units

Maximum Leak-Back

*Predicted response ranges less than 10 units, so it passes verification.*
Take-Home Messages

By way of a case study on verification testing of an hydraulic gear pump, this presentation on design of experiments (DOE) provided insights into statistically-optimal designs involving many categorical factors at multiple levels.

Upon collection of the response data, an innovative new graphical approach to assessing effects – the half-normal plot – revealed at-a-glance the likely significance and, for two-level factorials the importance of the signal generated by the experiment. *(General factorials require special scaling per Ref #2 shown on following slide.)*

**Multifactorial** DOEs like this are more efficient than traditional one-factor-at-a-time (OFAT) testing, which would never reveal an interaction such as the one that came to light in this case.
Further Reading
for More Detail on Methodology


For a copy of updated “Optimal Verification Testing with Graphical Effects Analysis,” e-mail Mark@StatEase.com. Feel free to provide comments and suggestions!

**Thanks for listening!**

“It may happen that small differences in the initial conditions produce great ones in the final phenomena.”

- Henri Poincare