Practical Considerations in the Design of Experiments for Binary Data

Martin Bezener
martin@statease.com
Stat-Ease, Inc.

Agenda

• Introduction
• Some Simple Examples
• Practical Tips and Tricks
• Conclusion
Introduction

A Typical Experimentation Sequence:

• Choose factors and define the design space
• Determine the appropriate experimental design
• Perform the experiment
• Analyze data and check for problems
  • If the response continuous, usually use Ordinary Least Squares, possibly with a transform
• Optimize
  • Follow up experiments needed?
• Confirm
  • Fix problems resulting from confirmation?
• Finish
• As mentioned on the previous slide, in most cases the response(s) of interest in an industrial experiment are **continuous** and can be analyzed using OLS.

• A continuous response is one that can (theoretically) take any value in some feasible range:

![Histogram of Continuous Data](image1)

---

What is Binary Data?

• However, in some cases, at least one response may be a **binary response**. This is a response that can take one of two outcomes:

![Bar Chart of Binary Data](image2)
Introduction

Some Examples

• Some common examples of binary responses:
  • Quality Engineering: {Pass, Fail}
  • Pharma: {Survived, Didn’t Survive}
  • Medical: {Disease Present, Disease Absent}
  • Marketing: {Clicked on Ad, Didn’t Click on Ad}
  • Food: {Dough Rises, Dough Doesn’t Rise}
  • …

• Experiments may include both continuous and binary responses.
• A binary response is often considered a categorical response.

Introduction

Some Interesting Features of Binary Data:

• A primary use of binary data is to estimate a probability.
• Binary data is far less informative than continuous data.
  • Cohen (1983) shows that “discretizing” continuous data is equivalent to throwing away 40-60% of the observations.
  
  ✓ TIP: Designs that have a binary response will generally require many more runs compared to an equivalent design with a continuous response.
  ✓ TIP: If you have continuous data, analyze it as continuous data. If you need to choose a threshold on which to make a binary decision, do so after the data analysis, not before.
Introduction

Estimating a probability

• Why are probabilities hard to estimate?
• Suppose you want to estimate the proportion of parts defective in a manufacturing process. In reality, the proportion is 0.67.
• If you sample 5 parts, you can get the following estimated proportions of defectives: {0, 0.2, 0.4, 0.6, 0.8, 1}
  • In the best case scenario, you will underestimate the proportion defective by 0.67 – 0.60 = 0.07, or 7%. This will occur about 1/3 of the time.
• If you sample 10 parts, in the best case scenario you will estimate the proportion of defectives at 0.6 or 0.7, which occurs with a probability of about 1/2.

Logistic Regression

• Binary data is analyzed using logistic regression. This is a feature that is new to version 12 of the Design-Expert® software.
• With continuous data, we would model response $y_i$ at the $i^{th}$ run with a model like the following:
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_{12} x_{1,i} x_{2,i} = x_i^T \beta$$
• This model is not well-suited for binary data except in a few situations (more on this later). Instead, we model the probability of success $p_i$ at the $i^{th}$ run with a model like the following:
$$p_i = 1 / [1 + \exp(-x_i^T \beta)]$$
One Factor Illustration

- Consider the following data:

<table>
<thead>
<tr>
<th>Factor</th>
<th>Reps</th>
<th># Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

- A linear logistic model can be fit to this data (right).

How about OLS?
A Comparison of Two Tools:

- Both OLS and logistic regression give reasonably close predictions of probabilities between 0.2 and 0.8 (approximately).
- Logistic regression, however, will give reasonable predictions of probabilities between 0 and 1.
- Even if OLS and logistic regression models give reasonably close predictions, the standard errors (and hence inference) may be very different.

**TIP:** Use logistic regression to analyze binary data. Don’t try to force an OLS model to 0/1 data.
Other Types of Responses You May Encounter:

- Multinomial response
  - A multinomial response is a categorical response. Similar to binary, but can take more than two values.
  - Modelled using Multinomial Logit Models
  - Example: Discrete Choice Experiments
- Count response
  - A non-negative integer response: \( \{0, 1, 2, 3, 4, 5, \ldots\} \)
  - Modelled using Poisson Regression (coming in v13)
  - Example: Number of defects on a metal sheet
Some Simple Examples

One Factor Example

• Our first example involves only one factor, but easily illustrates a number of interesting features of binary data
• An aircraft fastener manufacturer wanted to study the ability of a new fastener to withstand different loads.
• Only one factor was varied: Load from 2500 psi to 4300 psi
• A number of fasteners were tested at 10 different loads in the factor range
• The goal was to see how Load affected the probability of a fastener failing.

The Design

• The following experimental design was used. No information was provided to justify the choice of this design.

<table>
<thead>
<tr>
<th>Force</th>
<th># Replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>2500</td>
<td>50</td>
</tr>
<tr>
<td>2700</td>
<td>70</td>
</tr>
<tr>
<td>2900</td>
<td>100</td>
</tr>
<tr>
<td>3100</td>
<td>60</td>
</tr>
<tr>
<td>3300</td>
<td>40</td>
</tr>
<tr>
<td>3500</td>
<td>85</td>
</tr>
<tr>
<td>3700</td>
<td>90</td>
</tr>
<tr>
<td>3900</td>
<td>50</td>
</tr>
<tr>
<td>4100</td>
<td>80</td>
</tr>
<tr>
<td>4300</td>
<td>65</td>
</tr>
</tbody>
</table>

Total = 690
Some Simple Examples

The Results

- The number of failures was recorded, along with the proportion that failed:

<table>
<thead>
<tr>
<th>Force</th>
<th># Replicates</th>
<th># Failures</th>
<th>prop Failures</th>
</tr>
</thead>
<tbody>
<tr>
<td>2500</td>
<td>50</td>
<td>10</td>
<td>0.20</td>
</tr>
<tr>
<td>2700</td>
<td>70</td>
<td>17</td>
<td>0.24</td>
</tr>
<tr>
<td>2900</td>
<td>100</td>
<td>30</td>
<td>0.30</td>
</tr>
<tr>
<td>3100</td>
<td>60</td>
<td>21</td>
<td>0.35</td>
</tr>
<tr>
<td>3300</td>
<td>40</td>
<td>18</td>
<td>0.45</td>
</tr>
<tr>
<td>3500</td>
<td>85</td>
<td>43</td>
<td>0.51</td>
</tr>
<tr>
<td>3700</td>
<td>90</td>
<td>54</td>
<td>0.60</td>
</tr>
<tr>
<td>3900</td>
<td>50</td>
<td>33</td>
<td>0.66</td>
</tr>
<tr>
<td>4100</td>
<td>80</td>
<td>60</td>
<td>0.75</td>
</tr>
<tr>
<td>4300</td>
<td>65</td>
<td>51</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Some Simple Examples

Data Analysis

- A linear logistic model was fit to the data.
- This model was selected by both forward and backward selection using AICc as the criterion.

<table>
<thead>
<tr>
<th>Coefficients in Terms of Actual Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>A-Load</td>
</tr>
</tbody>
</table>

Practical DOE for Binary Data
Some Simple Examples

A Mixture Experiment with a Binary Response

- The process of separating mixtures of compounds is called chromatography. Ultra high performance liquid chromatography (UHPLC) is the gold standard for commercially available chromatograph techniques, but it uses very expensive equipment to provide the ultra high pressure.

- A research team tries to create a polymeric column that will have a pore size and skeleton thickness to achieve the same efficiency as UHPLC techniques but with lower pressure and hence lower cost.
The Problem

- A formulation does not always create a polymer that is completely solid, which is called a "homogeneous" column.
- The pores are not always interconnected, meaning the compound will not flow through the column.

- **Immediate Goal:** Design an experiment to model how the blending properties of water, monomer, and surfactant affect the probability of creating an acceptable column (one that is homogeneous and capable of flowing through the column).

Some Simple Examples

Mixture Components, Ranges, and Constraints

<table>
<thead>
<tr>
<th>Component</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Water</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>B: Monomer</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>C: Surfactant</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>D: Initiator</td>
<td>0.02</td>
<td>0.025</td>
</tr>
<tr>
<td>E: Salt</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Total = 100 weight %

Also, we need $0.002 < \frac{D}{B} < 0.01$
Some Simple Examples

The Design

• The design was built using Design-Expert version 12.

- Two binary responses were measured for each formulation:

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>homogeneous</td>
<td>0-No 1-Yes</td>
</tr>
<tr>
<td>flow</td>
<td>0-No 1-Yes</td>
</tr>
</tbody>
</table>

Some Simple Examples

Homogeneous Results…

Practical DOE for Binary Data
Some Simple Examples

Flow Results…

Joint Optimization

Practical DOE for Binary Data
Some Simple Examples

Joint Optimization

- The joint maximum probability for both responses occurs at:

  ![Graphs showing joint maximum probability](image)

  - Notice that two of the components (B and C) are at their maximums.
  - **TIP:** Consider expanding the ranges of B and C a bit in a follow-up experiment.

Some Simple Examples

Check Confidence Intervals

- The lower bound on the confidence interval for flow is 0.48 (less than half!).
- **TIP:** Consider doing additional follow up runs in a region around this optimum to improve precision.
Some of the things we’ve discussed so far:

• **TIP 1**: Designs that have a binary response will generally require many more runs compared to an equivalent design with a continuous response.

• **TIP 2**: If you have continuous data, analyze it as continuous data. If you need to choose a threshold on which to make a binary decision, do so after the data analysis, not before.

• **TIP 3**: Use logistic regression to analyze binary data. Don’t try to force an OLS model to 0/1 data.

Now we’ll go through issues related to designing experiments for binary data.
Practical Tips and Tricks

The Issues

• Building designs for binary data is generally **harder** than building designs for continuous data.
• This is due to the (previously mentioned) fact that binary data is less informative than continuous data.
• The **canned** designs for continuous data, such as Central Composite RSMs, or Simplex Lattice Mixture designs, will be too small for a binary responses.
• Optimal designs are also more tricky. Unlike designs for a continuous response, optimal designs for binary data depend on the actual values of the regression coefficients, which of course, are unknown.
• The construction of optimal experimental designs for binary data is still an active research area.

D-Optimal Designs

• A D-optimal design for a **continuous response** and linear model minimizes the determinant of the variance-covariance matrix of the regression coefficients:

\[
D(X) = |\sigma^2(X'X)^{-1}| 
\]

where \(X\) is the \(n \times p\) model matrix.
• Because of the constant variance assumption under linear model, the \(\sigma^2\) is constant and we simplify this to

\[
D(X) = |(X'X)^{-1}| 
\]
• Notice that the D-optimal design \(X\) does not depend on \(\sigma^2\) or the regression coefficients. This design will be D-optimal for all sets of regression parameters.
• An I-optimal design would suffer from these same issues.
D-Optimal Designs

• A D-optimal design for a binary response and logistic model minimizes the determinant of the variance-covariance matrix of the regression coefficients:

\[ D(X) = |(X^TWX)^{-1}| \]

where \( W \) is a matrix that depends on the regression parameters \( \beta \).

• Example (Chipman & Welch, 1996): Suppose we have two factors of interest and a binary response. We are interested in fitting a linear logistic model.

\[ \frac{1}{1 + \exp(-u)} \]  
where \( u = \beta_0 + \beta_1x_1 + \beta_2x_2 \)

• Four 12-run designs are generated for the following combinations of \( (\beta_1, \beta_2) \):

\[(1, 1) \quad (1, 3) \quad (3, 1) \quad (3, 3)\]
Practical Tips and Tricks

Practical Guidelines

• Having some idea of what the regression coefficients are beforehand will make building an experiment for binary data much easier.
• Bayesian methods have been developed that reduce the dependence of design optimality on the regression coefficients.
• Space-filling points are very useful.
• Sequential experimentation is key!
  • Pilot study to obtain rough estimate of coefficients
  • Re-define space based on the results of pilot study
  • Optimize, check confidence intervals widths, add runs if necessary
  • Continue until satisfied

Agenda

• Introduction
• Some Simple Examples
• Practical Tips and Tricks
• Conclusion
Conclusion

Take Home Points

- Binary data is usually analyzed using logistic regression. Don’t try to force familiar methods to binary data – take some time to learn the proper tools! Design-Expert 12 makes this easy to do.
- Don’t turn continuous data into binary data before analyzing your model. If you make your data binary, do so after analyzing the continuous response.
- Choosing a design is tricky – start small and iterate until you achieve your desired results. You may need to tweak component ranges and add more runs.

Future Developments

Additional Tools for Binary Data in Design-Expert 13 (tentative)

- Classification tools for predictions:

  ![Leave One Out Classification Table]

  ![ROC Curve Example]

- And more!
References

Some References (Analysis of Binary Data)

Practical DOE for Binary Data

Some References (Design Experiments for Binary Data)

Practical DOE for Binary Data