Enhancing Power Of Marketing Experiments Using Observational Data

Ron Berman, The Wharton School
Elea McDonnell Feit, Drexel University
Executive Summary

- Marketing experiments comparing customers exposed and not-exposed to ads often show insignificant results.
- Post-stratification is an analysis method that can increase statistical power if customers can be divided into high- and low-responsiveness groups.
- CRM data about past marketing response can be used to post-stratify low powered experiments.
- The gain in variance reduction may be upwards of 20%, leading to more accurate ROI estimates.
Marketing response experiments
Marketing experiments are a popular way to estimate the incremental sales attributable to a marketing communication.

Customers randomly assigned to treatment or to control.

Following medical literature, sometimes called randomized controlled trial (RCT).

Source: Proving Marketing Impact, Think with Google
Implementations vary across media

- Direct mail and email – randomized holdout – Control group not sent any ad (Zantedeschi et al. 2016, Sahni et al. 2016)
- Display advertising – Control group exposed to public service ad (PSA) (at a cost), or randomized holdout not exposed at all (Lewis and Reiley 2014)

Source: Designing with Science, medium.com
Disadvantages of using control groups

Control groups are great, but can have disadvantages:

• Very large control groups often needed to reach statistical significance.

• When PSAs are used as the control, the control is obviously wasted.

• If the marketing treatment works, there is an opportunity cost of not marketing to the held out customers.
Our contribution

- We show how marketers can use additional past data (e.g., CRM data) to change how an experiment is analyzed.

- The analysis method applies the classic approach of **stratification**, on a new variable that is constructed from the past data.

- The gains from the method reduces the required sample size in one of our applications by up to 20%.
How this works: An example application
A US multi-channel specialty retailer. Maintains a CRM database that tracks sales and catalog mailings to individual customers.

Why catalogs?

- Creating a holdout group is easier than in display advertising.
- Direct costs are substantial (~$0.60 per exposure) making it important to determine whether the sales lift exceeds the cost.

However, our method can be used in any experiment where prior data is available to rank customers.
The method works in two stages:

1. Estimating and ranking consumers by their past response to marketing experiments.
2. Using the ranked estimates in a post-stratified estimator to analyze the experiment.
Stage 1 - Ranking consumers by their past response to marketing

Suppose that for each consumer $i$ there is data about past sales in time $t$ ($y_{it}$) and past marketing exposure ($x_{it}$).

1. Run the linear regression:

$$\log(y_{it} + 1) = \alpha_t + \beta_{0i} + \beta_{1i}x_{it} + \varepsilon_{it} \tag{1}$$

2. Rank consumers by $\hat{\beta}_{1i}$.

3. Assign $w_i = 1$ for consumers above the median (H stratum) and $w_i = -1$ for consumers below the median (L stratum).
Stage 2 - Post-stratified regression

Given data $y_i$, $x_i$ and $w_i$ about sales, exposure and strata in the experiment to analyze:

Run the linear regression

$$y_i = \alpha + \beta_w w_i + \beta_x x_i + \beta_{xw} x_i w_i + \varepsilon_{it}$$ (2)

The resulting sampling variance of $\beta_x$ will be smaller if post-stratification is effective.
We will compare standard unstratified estimates to stratified estimates in six randomized holdout experiments.

We stratify customers based on their ranking of responsiveness based on observational data prior to the experiments.
6 experiments to analyze

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Sample Size</th>
<th>Holdout Rate</th>
<th>Average Sales</th>
<th>St. Dev. of Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Catalog: (n(1))</td>
<td>Holdout: (n(0))</td>
<td>Catalog: (\bar{Y}(1))</td>
<td>Holdout: (\bar{Y}(0))</td>
</tr>
<tr>
<td>Oct 13</td>
<td>906</td>
<td>36</td>
<td>0.038</td>
<td>20.50</td>
</tr>
<tr>
<td>Nov 13</td>
<td>858</td>
<td>36</td>
<td>0.040</td>
<td>25.54</td>
</tr>
<tr>
<td>Dec 13</td>
<td>582</td>
<td>51</td>
<td>0.081</td>
<td>31.77</td>
</tr>
<tr>
<td>Jan 14</td>
<td>783</td>
<td>32</td>
<td>0.039</td>
<td>19.02</td>
</tr>
<tr>
<td>Feb 14</td>
<td>648</td>
<td>125</td>
<td>0.162</td>
<td>7.75</td>
</tr>
<tr>
<td>Mar 14</td>
<td>723</td>
<td>49</td>
<td>0.063</td>
<td>8.21</td>
</tr>
</tbody>
</table>

**Table 1:** Descriptive statistics for catalog experiments
Using $t = 1, \ldots, 18$ months of data on catalog mailings and sales to individual customers $i$, we estimate the responsiveness of each customer using the following regression:

$$\log(sales_{it} + 1) = \alpha_t + \beta_{0i} + \beta_{1i} x_{it} + \varepsilon_{it}$$

- $x_{it}$ indicates whether customer $i$ received a catalog in month $t$.
- $\beta_{1i}$ is the (possibly biased) estimate of average catalog responsiveness for customer $i$.
- We also look at $\beta_{0i}$ as an alternative stratification variable.
These estimates are potentially biased and should not be used to assess catalog performance. But we can use them to re-analyze the holdout experiments.
Divide the customers into high- and low-responsiveness groups based on the estimate of $\beta_{1i}$ from the regression.

Re-analyze the catalog effect for each experiment separately for high- and low-responsiveness group.

Average the estimates for the high- and low-responsiveness groups to obtain the stratified estimate of the ATE. This is a more precise estimate of the advertising effect.
Stratified estimate of the catalog effect

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Standard ATE</th>
<th></th>
<th></th>
<th>Stratified ATE</th>
<th></th>
<th></th>
<th>Variance Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\tau}_{sd}$</td>
<td>Var($\hat{\tau}_{sd}$)</td>
<td>sd($\hat{\tau}_{sd}$)</td>
<td>$\hat{\tau}_{st}$</td>
<td>Var($\hat{\tau}_{st}$)</td>
<td>sd($\hat{\tau}_{st}$)</td>
<td></td>
</tr>
<tr>
<td>Oct 13</td>
<td>6.53</td>
<td>107.38</td>
<td>10.36</td>
<td>8.13</td>
<td>86.46</td>
<td>9.30</td>
<td>0.19</td>
</tr>
<tr>
<td>Nov 13</td>
<td>1.76</td>
<td>369.03</td>
<td>19.21</td>
<td>-0.26</td>
<td>456.91</td>
<td>21.38</td>
<td>-0.24</td>
</tr>
<tr>
<td>Dec 13</td>
<td>12.86</td>
<td>106.84</td>
<td>10.34</td>
<td>13.72</td>
<td>84.34</td>
<td>9.18</td>
<td>0.21</td>
</tr>
<tr>
<td>Jan 14</td>
<td>12.34</td>
<td>39.88</td>
<td>6.32</td>
<td>12.72</td>
<td>36.76</td>
<td>6.06</td>
<td>0.08</td>
</tr>
<tr>
<td>Feb 14</td>
<td>2.15</td>
<td>12.97</td>
<td>3.60</td>
<td>2.32</td>
<td>11.99</td>
<td>3.46</td>
<td>0.08</td>
</tr>
<tr>
<td>Mar 14</td>
<td>-29.79</td>
<td>379.22</td>
<td>19.47</td>
<td>-32.65</td>
<td>459.18</td>
<td>21.43</td>
<td>-0.21</td>
</tr>
</tbody>
</table>

**Table 2:** Comparison of standard and stratified estimates of the average treatment effect for catalog experiments. $\hat{\tau}_{sd}$: standard ATE. $\hat{\tau}_{st}$: stratified ATE.

We achieve up to a **21% reduction in variance**. Experiment sample size is linear in Variance, hence can achieve a **21% reduction in required sample size**.
Why this works: Theory
Standard Average Treatment Effect (ATE)

Standard approach to estimate the effect of an experiment – use the simple difference in mean response between the treatment and control:

\[
\hat{\tau}_{sd} = \frac{1}{n(1)} \sum_i T_i Y_i(1) - \frac{1}{n(0)} \sum_i (1 - T_i) Y_i(0)
\]

\(Y_i(1)\) are the sales with marketing (treatment) and \(Y_i(0)\) are sales w/o marketing (holdout/control) for customer \(i\).

\(T_i\) indicates if the consumer was assigned to treatment and \(n(0), n(1)\) are the number of consumers in the control and treatment, respectively.

Marketers can compare an estimate of the ATE to the cost of marketing to determine ROI.
Variance of the ATE estimator $\hat{\tau}_{sd}$

Because the sample is random, $\hat{\tau}_{sd}$ may be inaccurate.

The sampling variance of the ATE is:

$$V_{sd} = \text{Var}(\hat{\tau}_{sd}) = \frac{\sigma^2(0)}{n(0)} + \frac{\sigma^2(1)}{n(1)}$$

with $\sigma^2(0) = \text{Var}(Y_i(0))$ and $\sigma^2(1) = \text{Var}(Y_i(1))$.

Smaller samples ($n(0)$ and $n(1)$) and noisier data (larger $\sigma^2(0)$ and $\sigma^2(1)$) will make the estimate less accurate and will often result in non statistically-significant results.
If some customers are more responsive to ads than others, i.e., $Y_i(1) - Y_i(0)$ varies across the population, we can:

1. Stratify the sample into high- and low-responsiveness groups
2. Estimate ad response by strata
3. Take a weighted average of the strata estimates to estimate the ATE

Intuitively, by stratifying into more homogeneous groups, the within-group variance in response is reduced.

Reducing the variance of the ATE, allows us to reduce the required sample size.
Post-stratification

Suppose we can split the population to two strata: High (H) and Low (L) responsiveness.

In each stratum, we can estimate the treatment effect, e.g.,

\[ \hat{\tau}_H = \frac{\sum_{i \in H} T_i Y_i(1)}{n_H(1)} - \frac{\sum_{i \in H}(1 - T_i) Y_i(0)}{n_H(0)} \]

And then, combine them to a stratified ATE estimator using the fraction of the population in each stratum, \( f_H \) and \( f_L \):

\[ \hat{\tau}_{st} = f_H \hat{\tau}_H + f_L \hat{\tau}_L \]

\( \hat{\tau}_{st} \) is an unbiased estimator of the ATE \( \tau \).
The variance of the estimator is (Miratrix et al. 2013):

$$\nabla_{st} = \text{Var}(\hat{\tau}_{st}) = \frac{1}{n} \sum_{k \in \{H, L\}} f_k ((\beta_{1k} + 1)\sigma^2_k(1) + (\beta_{0k} + 1)\sigma^2_k(0))$$

when $\sigma^2_k(l)$ is the population variance of potential outcome $l$ in stratum $k$, and $\beta_{lk}$ are correction factors for post-stratification.
Why does it work?

By splitting the customers into high- and low-responsiveness groups, the firm generates two strata.

Each stratum has a lower variance of treatment effects.
1. The response needs to be heterogeneous in the population. Only certain distributions with heterogeneity lead to an improvement in power.

2. The advertiser needs some way to stratify customers into low- and high-responsiveness groups.
Our key insight is that past observational sales data can be used to rank customers by responsiveness.

- Many firms have this data from CRM systems or digital advertising platforms.
- (Assumption) Past response to marketing is likely to be correlated with the response in the experiment.
- (Claim) Even if estimates of responsiveness from observational data are biased due to targeting, they will maintain the ranking of customers on average.
Potential benefit of stratification

Based on one of the experiments in our application we simulate data with:
\[ \mu(1) = 7.75, \mu(0) = 5.6, \sigma(1) = 53.54, \sigma(0) = 32.69, n = 1000 \]

Red line: variance of standard estimator; Circles: variance of post-stratified estimator.

Improvement in variance: approximately 25%.
Conclusions
Summary

We can reduce the variance of our estimates of average treatment effects (and increase power) in randomized ad response tests by stratifying.

Not all stratification variables are good. The estimated response to marketing from past observational data seems to be an effective stratification variable that is available to many marketers.

This approach can substantially reduce the direct and opportunity costs of marketing response experiments by reducing sample sizes.

