Foundations of Data Science: A/B testing
Overview

- Principle of A/B testing
  - what it is, estimation and hypothesis testing approaches with the bootstrap

- Increasing certainty in A/B testing

- Theoretical, large-sample approach to A/B testing

- Issues in A/B testing

- Comparing numeric samples
Foundations of Data Science: A/B testing -
The principle of A/B testing
A/B Testing

Welcome to our website
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Click rate: 52% 72%

1. Is A significantly better or worse than B?
2. How much better or worse is A than B?
Fast growing companies use VWO for their A/B testing

Thousands of brands across the globe use VWO as their experimentation platform to run A/B tests on their websites, apps and products.

name@yourcompany.com

TRY VWO FOR FREE
## Approaches

<table>
<thead>
<tr>
<th>Parameter estimation</th>
<th>Hypothesis Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0.</strong> Decide underlying parameter to infer</td>
<td><strong>0.</strong> Decide on ( H_0 ) and ( H_a )</td>
</tr>
<tr>
<td><strong>1.</strong> Construct formula for estimator in terms of data</td>
<td><strong>1.</strong> Define test statistic in terms of data</td>
</tr>
<tr>
<td><strong>2.</strong> Find approx. sampling distribution of estimator using bootstrap or large sample theory</td>
<td><strong>2.</strong> Find distribution of test statistic under ( H_0 )</td>
</tr>
<tr>
<td><strong>3.</strong> Return confidence interval</td>
<td><strong>3.</strong> Reject / not reject ( H_0 ) of find p-value</td>
</tr>
</tbody>
</table>
A/B testing example: Estimation approach

**Parameters**
- \( p_A \) - parameter for proportion of click-throughs from A/B
- \( p_B \) - parameter for difference.
\[ d = p_A - p_B \]

**Data**
- \( n = 1000 \) \# presentations of A \& B
- \( n_A = 700 \) \# click-throughs on A
- \( n_B = 720 \) \# click-throughs on B

**Estimators**
- \( \hat{p}_A = \frac{n_A}{n} \)
- \( \hat{p}_B = \frac{n_B}{n} \)
- \( \hat{d} = \hat{p}_A - \hat{p}_B \)
Sampling distribution of $\hat{\theta}$ with bootstrap

$B$ - # repetitions

for $i$ in 1, ..., $B$

- Sample $n^*_i$ from $\text{Binom}(n, \hat{p}_a)$
- $n^*_i$ from $\text{Binom}(n, \hat{p}_B)$
- Compute difference and store it.

\[ d^*_i = \frac{n^*_a - n^*_B}{n} \]

Compute quantiles, std error in estimator.
Results

\[ \hat{d} = 0.70 - 0.72 = -0.02 \]

\[ \hat{d} = -0.02 \]

95\% CI = (-0.06, 0.02)

\[ p(\hat{p}_A - \hat{p}_B) > 0 = (0.1547) \]
Exercise

How would you apply the hypothesis testing approach to A/B testing?

1. $H_0$:

2. Test statistic:

3. Distribution of test statistic:
Foundations of Data Science: A/B testing - Increasing certainty
A/B Testing

\[ n = 1000 \]

Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Click rate:

\[ 52\% \quad 70\% \quad 72\% \]

Maxime Lorant, Wikimedia, CC SA 4.0
Bootstrap results

\[ \hat{d} = -0.02 \]

95% CI = (-0.06, 0.02)

\[ p(\hat{p}_A - \hat{p}_B) > 0 = 0.1547 \]

\[ \hat{d} = \hat{p}_A - \hat{p}_B = 0.70 - 0.72 = -0.02 \]

15% chance A is better than B
Getting a more certain result

\( \hat{d} = -0.02 \)
95% CI = (-0.06, 0.02)
\( p(p_{A} - p_{B}) > 0 = 0.1547 \)

\( n=1000 \)

\( \hat{d} = -0.02 \)
95% CI = (-0.04, -0.00)
\( p(p_{A} - p_{B}) > 0 = 0.0143 \)

\( n=5000 \)

\( \hat{d} = -0.02 \)
95% CI = (-0.03, -0.01)
\( p(p_{A} - p_{B}) > 0 = 0.0012 \)

\( n=10000 \)
Question: Is a big enough sample good enough?

We can run more experiments to get lower p-values, but could we still have the wrong answer?
Foundations of Data Science:
A/B testing -
Large sample theory
Confidence level: $1 - \alpha$

$$CI = \left( \hat{\rho}_a - z_{\alpha/2} \hat{\sigma}_a, \hat{\rho}_a + z_{\alpha/2} \hat{\sigma}_a \right)$$

Eg. $\hat{\rho}_a - \hat{\rho}_b = 0.70 - 0.72 = -0.02$

$$\hat{\sigma}_a = \sqrt{\hat{\rho}_a(1-\hat{\rho}_a) + \hat{\rho}_b(1-\hat{\rho}_b)}$$

$$= \sqrt{0.70(1-0.70) + 0.72(1-0.72)} = 0.02$$
95\% CI \Rightarrow z_{\alpha/2} = 2 \cdot 0.025 = 1.96

\Rightarrow CI: \left( \hat{d} - \frac{z_{\alpha/2} \hat{\sigma}_d}{\sqrt{n}}, \hat{d} + \frac{z_{\alpha/2} \hat{\sigma}_d}{\sqrt{n}} \right)

= -0.02 - 1.96 \times 0.02, 0.02 + 1.96 \times 0.02

= (-0.06, 0.02)
Sample size calculation

\[
\frac{|\hat{d}|}{\hat{\sigma_d}} = \frac{2}{z_{0.01}}
\]

\[
\hat{\sigma_d} = \sqrt{\hat{\rho_a}(1-\hat{\rho_a}) + \hat{\rho_b}(1-\hat{\rho_b})}
\]

\[
= \frac{2}{z_{0.01}} \sqrt{\hat{\rho_a}(1-\hat{\rho_a}) + \hat{\rho_b}(1-\hat{\rho_b})}
\]

\[
\Rightarrow \quad n = \frac{2}{z_{0.01}^2} \frac{(\hat{\rho_a}(1-\hat{\rho_a}) + \hat{\rho_b}(1-\hat{\rho_b}))}{\hat{d}^2}
\]
Foundations of Data Science: A/B testing - Issues in A/B testing
Statistical versus practical significance

Which scenario is more statistically significant?
Which scenario could be more significant practically?

\[
\begin{align*}
\text{Scenario 1:} & \quad p \sim 0.001 \\
& \quad n = 10,000 \\
\text{Scenario 2:} & \quad p = 0.06 \\
& \quad n = 100
\end{align*}
\]
Ethical issues

- Informed consent
  - Remember the Facebook experiment from Semester 1

- Data protection

- Questions to ask
  - Would I feel comfortable if this change were tested on me?
  - What potential harms could be caused to users?

- Academic setting - ethics approval always needed
Foundations of Data Science: A/B testing - Comparing numeric samples
Same or different? (Hypothesis test)
How big is the difference in the means? (Estimation)

Estimator of difference: \( \hat{d} = \bar{x} - \bar{y} \)
Standard error of estimator: \( \hat{\sigma}_d = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}} \)

\( t = \frac{\hat{d}}{\hat{\sigma}_d} \)
Parameter estimation

Hypothesis test (t-test)

\( \hat{\alpha} \)

\( \hat{\alpha} \sim \text{scaled } z \sigma \text{ t-dist} \)

95\% CI:

\[
(\hat{\alpha} - \hat{\sigma}_\alpha z_{0.025}, \hat{\alpha} + \hat{\sigma}_\alpha z_{0.025})
\]
Effect size - Cohen's d

\[ d = \frac{\bar{x} - \bar{y}}{s} \]

\[ s = \sqrt{\frac{(n_x - 1)s^2_x + (n_y - 1)s^2_y}{n_x + n_y - 2}} \]
Interpretation of Cohen's $d$

$d=0.01$  very small  
$d=0.2$    small         
$d=0.5$    medium        
$d=0.8$    large         
$d=1.2$    very large    
$d=2.0$    huge          

A well-known use of Cohen's d

<table>
<thead>
<tr>
<th>Influence</th>
<th>Cohen's $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reported grades</td>
<td>1.33</td>
</tr>
<tr>
<td>Teacher credibility</td>
<td>0.9</td>
</tr>
<tr>
<td>Deliberate practice</td>
<td>0.79</td>
</tr>
<tr>
<td>Feedback</td>
<td>0.7</td>
</tr>
<tr>
<td>Spaced vs. mass practice</td>
<td>0.6</td>
</tr>
<tr>
<td>Note taking</td>
<td>0.5</td>
</tr>
<tr>
<td>Cooperative learning</td>
<td>0.4</td>
</tr>
<tr>
<td>Ability grouping for gifted students</td>
<td>0.3</td>
</tr>
<tr>
<td>Extra-curricula programs</td>
<td>0.2</td>
</tr>
<tr>
<td>Open vs. traditional classrooms</td>
<td>0.01</td>
</tr>
<tr>
<td>Lack of sleep</td>
<td>-0.05</td>
</tr>
<tr>
<td>Television</td>
<td>-0.18</td>
</tr>
<tr>
<td>Boredom</td>
<td>-0.49</td>
</tr>
</tbody>
</table>

https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/
Paired data

Paired t-test

\[ d_i = x_i - y_i \]

\[ \hat{\sigma}_d = \sqrt{\frac{1}{n} \sum (x_i - y_i)^2} \]

\[ t = \frac{\bar{d}}{\hat{\sigma}_d} \]

\[ d = x - y \]
Summary

1. A/B testing: controlled experiment, binary response

2a. Estimate confidence intervals between response rates in A and B, by bootstrap or theoretically
b. Test if response rate in A is different from B, by statistical simulation, or theoretically

3. Increasing sample size decreases confidence interval and decreases p-value

4. Issues: Ethics and effect size

5. Numeric samples – estimation, hypothesis testing, effect size (Cohen's d)