A/B Testing

Jaime Arguello
INLS 509: Information Retrieval
jarguell@email.unc.edu

- Credits: these slides borrow heavily from examples and figures from Ron Kohavi’s presentations on A/B testing at Microsoft (available online)
Introduction

• Systems (e.g., search systems) are always trying to improve

• **Basic question:** If a specific change is introduced, will it improve key metrics?

• **Metrics:** measures that are believed to be correlated with the quality of the user experience

• Metrics are often things we want to minimize or maximize

• Examples?
A/B Testing

• Experiments where different populations of users are exposed to different versions of the system for a period of time

• **Control group:** group of users exposed to the “normal” or “baseline” version of the system

• **Experimental group:** group of users exposed to the experimental version of the system

• More often A/B/C/D/E… testing

• Search companies can have about 15 different A/B tests happening at once

• $5^{15} = 30,517,578,125$
The Alternative

• Make the change and measure the same metrics.
• Why is this a bad idea?
The Alternative

- Make the change and measure the same metrics.
- Why is this a bad idea?
  1. Temporal changes
  2. Good features lead to **incremental** improvements
  3. It’s difficult to assess the value of ideas
Temporal Changes

Source: http://exp-platform.com/2017abtestingtutorial/
Temporal Changes + Incremental Improvements

Source: http://exp-platform.com/2017abtestingtutorial/
Predicting the value of new features

• 1/3 of ideas improve the intended metric(s)
• 1/3 of ideas have no effect
• 1/3 of ideas degrade the intended metric(s)

Source: http://exp-platform.com/2017abtestingtutorial/
Predicting the value of new features

Source: http://exp-platform.com/2017abtestingtutorial/
(1) Predicting the value of new features

- Overall Evaluation Criterion: no. of searches
- A > B, A < B, or A = B?

Source: http://exp-platform.com/2017abtestingtutorial/
(2) Predicting the value of new features

Source: http://exp-platform.com/2017abtestingtutorial/
(2) Predicting the value of new features

10 search results  
A

8 search results  
B

- Overall Evaluation Criterion: clickthrough rate 1st SERP per query
- A > B, A < B, or A = B?

Source: http://exp-platform.com/2017abtestingtutorial/
(3) Predicting the value of new features

• Overall Evaluation Criterion: revenue
• 4 A ads for every 3 B ads
• A > B, A < B, or A = B?

Source: http://exp-platform.com/2017abtestingtutorial/
Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience

Source: http://exp-platform.com/2017abtestingtutorial/
Correlation does not Imply Causation

- Umbrellas cause rain
- People with smaller hands live longer
- A new feature (e.g., a new advanced search tool) increases retention rate

Source: http://exp-platform.com/2017abtestingtutorial/
Correlation does not Imply Causation

- Particularly important for understanding the impact of system features that are used more by certain types of users than others

Source: http://exp-platform.com/2017abtestingtutorial/
Correlation does not Imply Causation

- What are features used more by heavy users?

Source: http://exp-platform.com/2017abtestingtutorial/
Correlation does not Imply Causation

- What are features used more by **new** users?

Source: http://exp-platform.com/2017abtestingtutorial/
Challenges in A/B Testing

• Correlation does not imply causation
• Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
• Using the wrong metric
• Unexpected effects on important metrics
• Making claims not exactly tested
• Bugs in the experimental infrastructure
• Using sound statistical methods
• Hurting the user experience

Source: http://exp-platform.com/2017abtestingtutorial/
Short-term vs. Long-term Metrics

- An increase in ad clicks suggests an increase in revenue
- Showing lots of ads (often) hurts the user experience and decreases retention (i.e., long-term ad-click revenue)

Source: http://exp-platform.com/2017abtestingtutorial/
Using the wrong metric

- Hanoi’s French Quarter rat problem in 1902

**Rats Killed per day**

- 1,000/day
  - April, 1902
  - Week 1
- 4,000/day
  - April, 1902
  - Week 2
- ... (elided)
- 20,000/day
  - July, 1902

*Source: http://exp-platform.com/2017abtestingtutorial/
Using the wrong metric

• Hanoi’s French Quarter rat problem in 1902

Rats Killed per day

1,000/day
April, 1902
Week 1

4,000/day
April, 1902
Week 2

... ...

20,000/day
July, 1902

• What you do not measure, does not improve.

• Goodhart’s law: “when a measure becomes a target, it ceases to be a good measure”

Source: http://exp-platform.com/2017abtestingtutorial/
Unexpected Effects on Important Metrics

- **Example:** a hyperlink on the SERP was changed to open on a new browser tab.
- It increased avg. SERP load time by 8.32%
- **Why?**

The information is sourced from: [http://exp-platform.com/2017abtestingtutorial/](http://exp-platform.com/2017abtestingtutorial/)
Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience

Source: http://exp-platform.com/2017abtestingtutorial/
Making Untested Claims

• **Question:** What is the effect of SERP load-time on ad-click revenue?

• Artificially **increase** SERP load-time and measure **decrease** in ad-click revenue

• Make the claim that **decreasing** the SERP load-time will have a comparable **increase** in ad-click revenue

• What’s wrong with this?

Source: http://exp-platform.com/2017abtestingtutorial/
Making Untested Claims

• **Question:** What is the effect of SERP load-time on ad-click revenue?

• Artificially **increase** SERP load-time and measure **decrease** in ad-click revenue

• Make the claim that **decreasing** the SERP load-time will have a comparable **increase** in ad-click revenue

• What’s wrong with this?

• Assumes (bi-directional) linear relationship

*Source: http://exp-platform.com/2017abtestingtutorial/
Making Untested Claims

Source: http://exp-platform.com/2017abtestingtutorial/
Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience

Source: http://exp-platform.com/2017abtestingtutorial/
Bugs in the Experimental Infrastructure

- User sampling + measurement + statistics
- How can we debug this infrastructure without opening the “black box”?

Source: http://exp-platform.com/2017abtestingtutorial/
Bugs in the Experimental Infrastructure

- Run lots of A/A tests (no differences between experimental and control conditions)
- How often should we observe a \( p \)-value of 0.05 or less?

Source: http://exp-platform.com/2017abtestingtutorial/
Sound Statistical Methods

- Even when there is no difference between the two systems, it is still possible to observe a $p$-value of less than 0.05
- Why?

Source: http://exp-platform.com/2017abtestingtutorial/
By definition, the $p$-value is the probability of the observed difference in means (or a more extreme difference) under the null hypothesis!
A/A Testing

- Run lots of A/A tests (no differences between experimental and control conditions)
- We should only observe $p$-values of 0.05 or less about 5% of the time
- The $p$-value distribution should be uniform rather than skewed to low or high values

*Source: http://exp-platform.com/2017abtestingtutorial/*
Challenges in A/B Testing

• Correlation does not imply causation
• Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and/or revenue
• Using the wrong metric
• Unexpected effects on important metrics
• Making claims not exactly tested
• Bugs in the experimental infrastructure
• Using sound statistical methods
• Hurting the user experience

Source: http://exp-platform.com/2017abtestingtutorial/
Causes of Type I Errors (False Positives)

- Running the same A/B test many times until we observe a significant difference
- Using 100+ metrics and focusing on the ones that are significant
- Running an experiment for as long as it takes to reach significance
- Running an experiment and stopping early because we reached significance

Source: http://exp-platform.com/2017abtestingtutorial/
Causes of Type I Errors (False Positives)

- **Bonferroni correction**: multiplying the $p$-value by the number of comparisons

*Source: http://exp-platform.com/2017abtestingtutorial/*
Hurting the User Experience

- Less manual monitoring of experiments
- Buggy features or bad ideas
- Interactions between concurrent experiments: the whole is less than the sum of its parts

Source: http://exp-platform.com/2017abtestingtutorial/
Cautionary Steps: Starting Small

- Starting internally (within the company)
- Starting with only a few users
- Starting with only partial exposure (1/10 queries)

Source: http://exp-platform.com/2017abtestingtutorial/
Cautionary Steps: Different types of Metrics

- **Data quality metrics**: ensure that the feature was implemented correctly
- **Overall evaluation criteria**: single metric that measures improvement in user experience (e.g., number of satisfied clicks)
- **Guardrail metrics**: metrics used to shutdown an experiment (e.g., queries with no clicks)
- **Local metrics**: metrics that measure what the user is doing less of (because of the new feature)

*Source: http://exp-platform.com/2017abtestingtutorial/*
Cautionary Steps: Measuring interactions

Source: http://exp-platform.com/2017abtestingtutorial/
Cautionary Steps: Measuring interactions

Source: http://exp-platform.com/2017abtestingtutorial/
Cautionary Steps: Measuring interactions

Source: http://exp-platform.com/2017abtestingtutorial/
Ethical Considerations

• System development is influenced by the majority

• Certain communities may be under-represented in the data

• While there is an “average user”, there is also high variance (nobody is close to the average)

• Metrics used in A/B tests are crude measures of “user experience”

• Users may need to experience extreme differences to show (positive or negative) changes in behavior

• A/B tests are done without considering whether the user is in a vulnerable state

Source: http://exp-platform.com/2017abtestingtutorial/
Challenges in A/B Testing

- Correlation does not imply causation
- Understanding how short-term metrics (measured during A/B tests) lead to long-term improvements in user experience and revenue
- Using the wrong metric
- Unexpected effects on important metrics
- Making claims not exactly tested
- Bugs in the experimental infrastructure
- Using sound statistical methods
- Hurting the user experience