

A/B Testing in Python

ODSC-East 2021

Mary C. Boardman, PhD

Senior Data Scientist, TI Health

Workshop Outline

- Theory
 - Designing Your A/B Test
 - Analyzing Your Results
 - Example Case Overview
- Code Walkthrough
- Results
 - Communicating the Specific Results
 - General Key Takeaways





A/B Testing (Experiment) Theory: Designing Your A/B Test

Why A/B Testing (Experiments)?

Machine learning: finding patterns in existing data.

Pros: very low risk, comparatively cheap, many people can put together a model at least competently

Cons: can only tell us about the past, can't show causality

Experiments: finding casual relationships, figuring out if doing a proposed thing will have the desired outcome

Pros: can tell us if the new thing worked or not, can show causality

Cons: can be high risk and/or high cost, need a deeper understanding of theory to do this well



Initial Conversations with Stakeholders

- Best place to start
- What is it we are trying to learn/actually do?
- Two good options:
 - Simplest intervention that might work
 - Most likely action that would be taken
- How big of an Average Treatment Effect (ATE) is big enough to matter?
 - Practical versus statistical significance
 - How much of an effect do we need to see in order for these results to be meaningful?
- Resource constraints
 - Timeframe
 - Budget
 - Human (usually engineering time)
- Risk to the business
 - If something goes wrong, what's the risk?
 - Are we all okay enough with this to proceed?



A/B Test Design (1/2)

- **Control and Test Groups**
 - Control group, nothing changes
 - Test group(s), one intervention per test
 - Multiple test groups are fine
- **Intervention**
 - What is it we are planning to do? Now is the time to get specific.
- **Sample Size Needed**
 - Cohen's D to determine how big of a sample you need given desired effect size
 - Any time you are dealing with human behavior, expect small effects
- **Stopping Conditions**
 - NEVER when the data looks right
 - Enough subjects
 - Duration



A/B Test Design (2/2)

- Who are the subjects?
 - Website visitors
 - Customers
- How will we recruit them?
 - Website visitors is easy
 - Actual recruiting can get costly (i.e. drug trials)
 - Are the people we recruit actually representative of the population we need to learn about?
- How can we guarantee randomization?
 - Simple code can do this if it's a human/computer interface (i.e. website visitors or this workshop's example)
 - More human involvement = double blind is best
- What resources do we need?
 - Recruitment
 - Software
 - Engineering time






A/B Testing (Experiment) Theory: Analyzing Your Results

Analyzing Your Results


This is the easy part!

- **Average Treatment Effect (ATE): test mean - control mean**
 - This is how we measure differences between test and control groups
 - **Attrition Check**
 - Were subjects more likely to drop out in one group or another?
 - Is this too large to have valid results?
 - **Covariate Balance Check**
 - Did randomization work?
 - **Significant Covariates**
 - Are there significant covariates, and if so, which ones?
 - **OLS Model and Hypothesis Test**
 - VERY basic model predicting ATE
 - Use ANOVA for multiple test groups, with t-tests for pairs if you find significance
 - Use independent, 2-sample t-tests for just 1 test group
- 




A/B Testing (Experiment) Theory: Example Case Overview

Detect the Fake Smile

- Key Question: Are people, on average, more likely to correctly spot a fake smile given information on how to detect this?
 - Control: Kajdasz viral survey, presented 20 pictures of smiles, some real some fake (Access it here: <https://www.surveymonkey.com/r/SmileRead>)
 - Document-Only Test: Participants read a 1 page document and then take the survey
 - Video-Only Test: Participants watch a 2 minute video and then take the survey
 - Document AND Video Test: Participants get both treatments then take the survey
 - Covariates: Age, gender, profession that involves reading body language, pre-questionnaire self-estimation of final score
- 

Design and Strategy

- Treatment Strategy: simplest intervention that might work
 - Significance Needed: statistical significance, given sample sizes of ~150 per group, no real practical application, and extremely simple treatment
 - Stopping Conditions: ~600 completed surveys or 2 weeks, whichever came first
 - Subject Recruitment Strategy: simple, website visitors taking an already viral survey (thanks Kajdasz!)
 - Resource Constraints: 4 week timeframe, \$500 budget, no engineering resources
 - Randomization: used survey software with randomization functionality. Website visitors were assigned a group at random
 - Risk: Minimal to none
- 

Code Overview



Communicating Your Results

No Significant Results, Now What?

- VERY common!
 - Set stakeholder expectations up front
 - If they aren't 100% on board with the possibility of the treatment doing nothing, back out of the experiment if you can
- No significant results are still VERY useful!
- Bottom Line up Front (BLUF)
 - Plain language, NO jargon.
 - None of the training options led to a significant increase in spotting fake smiles.
- Suggestions/Pivots
 - More thorough training (workshops, more detailed videos/documents)
 - This skill set may just not be trainable with standard knowledge/approaches



Caveats/Sticking Points

- Have a plan, especially if you will be telling someone something they don't want to hear
- Ask for advice on how the message would be best received
- Focus on the usefulness of the results
- Think through and present options
 - Keep asking yourself "so what"?
 - Keep asking yourself "what now"?
 - Present your favorite 3 ideas
 - Turn the meeting into a brief brainstorming session
 - Anticipate key stakeholders coming up with options and uses you haven't
 - Focus on collaboration and having a 2-way conversation, not dry reporting



Contact Information

Email: mary.boardman@gmail.com

LinkedIn: <https://www.linkedin.com/in/marycboardman/>

Twitter: @marycboardman



Thank You!

James Kajdasz
Sameed Musvee
Irene Seo