Udacity A/B Testing Lesson 1 Notes

Jung Ah Shin

Jan 2021

1 Overview of A/B Testing

1.1 Introduction

A/B testing is a general methodology used online when you want to test out a new product or feature.

Two sets of users: Existing product vs. New version

When NOT to use A/B Testing:

- A/B Testing is not useful for testing new experiences
  - What is the baseline for comparison?
  - How much time you need for users to adapt to the new experience?
- Time (e.g. apartment rentals → people don't look for apartments that often)
- Cannot tell you if you're missing something

Table 1: When to use A/B Testing Examples

Useful	Not Useful
Movie recommendation site - new ranking algorithm : clear control group and metrics **Change backend - page load time, results users see** : good if computing power available for both **Test layout of initial page** : clear control and metrics	**Online shopping company - Is my site complete?** : could try specific product,but cannot know in general **Add premium service** : could gather information but cannot fully test **Update brand, including main logo** : surprisingly emotional **Website selling cars** : too long and do not have data

Other techniques to use to gather information about users (Qualitative)

- Logs of what users did on the website - Analyze retrospectively to build hypothesis
- User experience research
- Focus groups
- Surveys
- Human evaluation

A/B Testing needs to have a consistent response from your control and experiment group

Goal of A/B Testing is to design an experiment that is going to be robust and give you repeatable results so that one can make a good decision

1.2 Business Example

E.g. Imagine an education company like Udacity called Audacity that focuses on creating finance courses
Goal: To increase student engagement User flow: Customer funnel (largest number of events at the top, where customers go back and forth the funnel)

- Homepage visits
- Exploring the site
- Create account
- Complete a purchase/class

Experiment Initial Hypothesis: Change the Start Now button from *orange* to *pink* will increase how many students explore Audacity's courses

Which metric to use?

- Total number of courses completed (BUT time consuming and not practical as it can take months for students to complete the course)
- Number of clicks (BUT if more total clicks in one version but with lower ration than other version)
- CTR (click-through-rate) = $\frac{Number\ of\ clicks}{Number\ of\ page\ views}$
- CTR (click-through-probability) = $\frac{Unique\ visitors\ who\ click}{Unique\ visitors\ to\ page}$

> *Use rate when you want to measure the usability of a site and a probability when you want to measure a total impact and disregard double-clicks, reloads, etc.*

Updated Hypothesis: Change the Start Now button from *orange* to *pink* will increase the click-through-probability of the button

Repeated measurement of click-through-probability

- visitors = 1000
- unique clicks = 100
- click-through-probability ≈ 10%

  Which results would surprise you if you repeated the measurement?

  - 100
  - 101
  - 110
  - 150 (above what I expected)
  - 900 (above what I expected)

Hypothesis Testing
P_{cont} and p_{exp}
Null hypothesis H_0: If changing button had no effect, where $p_{exp} - p_{cont} = 0$

Alternative Hypothesis H_A: $p_{exp} - p_{cont} \neq 0$

Steps

- Measure $\hat{p_{cont}}$ and $\hat{p_{exp}}$
- Compute the probability that this difference would have arisen by chance if the null hypothesis were true. $P(\hat{p_{exp}} - \hat{p_{cont}}|H_0)$
- Reject H_0 if the above probability is small enough ($p < 0.05$)

Question: Choosing H_0 and H_A

- Change checkout flow of online shopping site
- Test old flow vs. new flow
- Measure probability of completing checkout

Null hypothesis: The experimental and control groups have the same probability of completing a checkout

Alternative Hypothesis: The experimental and control groups have a different probability of completing a checkout

Then, what change in the click-through probability is substantive/practically significant?

Size your experiment appropriately, such that the statistical significance bar is lower than the practical significance bar

Audacity example: 2 percent change in the click-through probability would be practically significant

How many page views

- $\alpha = P(\text{reject null } |H_0 \text{ true})$

1.3 Statistics Review

- **Binomial Distribution**
  (Successes/Failures) e.g. (click = success, no click = failure)
  - biased user who has $p = \frac{3}{4}$ of clicking a page
  - success = click, failure = no click
  - As $N \to \infty$, binomial $\to$ normal
  - $mean = p$, $stddev = \sqrt{\frac{p(1-p)}{N}}$
  - Assume p not known
    - e.g. N = 20, clicks = 16, Estimate the bias $\hat{p} = \frac{4}{5}$
  - When to use binomial
    - 2 types of outcomes
    - independent events
    - identical distribution: p same for all

- **Confidence Intervals**
  For a 95% confidence interval, if we theoretically repeated the experiment over and over again, we would expect the interval we construct around the sample mean to cover the true value in the population 95% of the time
  - $\hat{p} = \frac{X}{N}$ where X = number of users clicked, N = number of users

- e.g. $\hat{p} = 0.1$
- *To use normal distribution if* $N\hat{p} > 5$ *and* $N(1-\hat{p}) > 5$
- standard error SE $= \sqrt{\frac{p(1-p)}{N}}$
- margin of error m $= zscore * SE = z * \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$
  The amount of random variation we expect in our sample is a proportion of both successes and the size of the sample. When the success probability is further from 0.5, then SE would be smaller, which means CI will be smaller.
  Similarly, if N is larger, the SE and CI will be smaller. For 95% CI, z-score will be 1.96 for two-tailed CI.
  m = 0.019 margin = 0.081 to 0.119

- **Hypothesis Testing**
  - Null hypothesis: There is no difference in click-through-probability between our control and experiment
  - Alternative hypothesis: Are we interested in the difference, or just higher or lower?
  -

- **Pooled Standard Error**
  - Number of users who click in each group: X_cont, X_exp
  - Total number of users in each group: N_cont,
  - First, calculate pooled probability of a click (total probability of a click across groups)
    $\hat{p_{pool}} = \frac{X_{cont}+X_{exp}}{N_{cont}+N_{exp}}$
  - Then, calculate pooled standard error
    $SE_{pool} = \sqrt{\hat{p_{pool}} * (1 - \hat{p_{pool}}) * (\frac{1}{N_{cont}} + \frac{1}{N_{exp}})}$
    Difference $\hat{d} = \hat{p_{exp}} - \hat{p_{cont}}$
  - Under H_0 :, true difference $d = 0$, $\hat{d} \sim N(0, SE_{pool})$
  - If $\hat{d} > 1.96 * SE_{pool}$ or $\hat{d} < -1.96 * SE_{pool}$, reject H_0

- **Size vs. Power Trade-Off**
  - Statistical power: e.g. How many page views needed in order to get a statistically significant result
  - Power has an inverse trade-off with size: The smaller the change that you want to detect or the increased confidence you want to have in the result, means larger experiment required, so more page views in control experiment