General Notes

Darian S. Martos

AB Testing Notes

These are some preliminary notes for A/B testing, using the online controlled experiments book as the main reference below. Some notes from a previously taken experimental design course taught by Dr. Art Owen, that are widely available, were also referenced for these notes. The chapter ordering for these notes is quite scattered, mostly due to me finding certain sections more fruitful than others.

Much of these notes are organized in a definition-based format. Examples are scarce unless I find them valuable.

Resources used (these are hyperlinks):

- Trustworthy Online Controlled Experiments [Kohavi, Tang, Xu]
- AB Testing Notes from Stanford Stats 263/363 [Owen]

Some baseline abbreviations:

- Controlled experiments - CEs
- Overall evaluation criterion - OEC (see Ch. 1.1)
- Return on investment - ROI
- Minimum viable product - MVP
- Expected value of information - EVI (see Ch. 1.6)

Preface

A mess of notes from the Owen A/B testing notes. What are sections?

Chapter 1. TOCE - Introduction and Motivation

Overall Evaluation Criterion (OEC) - Quantitative measure of an experiment's objective

1 Terms and Definitions

- Example - active days per user
- An ↑ in OEC implies an ↑ in site visits ⟶ good
- OEC should be short-term measurable, believed to causally drive long-term strategy

Parameter - Controllable experiment variable thought to influence OEC or other metrics, also known as factors or variables.

- The assigned values are known as *levels*.
- In simple A/B tests - univariate with two values
- Multivariate tests evaluate parameters together

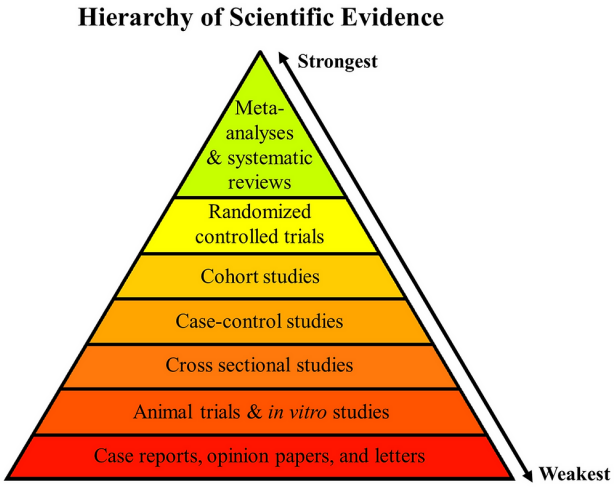
Variant - User experience being tested, A and B are two variants - usually a control and treatment.

Randomization unit - Pseudo-random process that is applied to units to map to variants. The recommended set of units is usually users.

Often, improvements in key metrics are achieved in small increments over time, "inch-by-inch."

2 Why experiment?

- Don't want to assume causality off observation alone
- Want to follow the hierarchy of evidence in terms of causal strength:



- Online CEs can establish causality with high probability, are able to detect small changes that are harder to detect with other techniques, such as changes over time (sensitivity), and are able to detect unexpected changes

3 Necessary Ingredients for Experiments

- There are experimental units that can be assigned to different variants with no interference.
- There are enough experimental units (ideally thousands).
- Key metrics, ideally an OEC, that are agreed upon and can be practically evaluated. Data should be reliable and can be easily collected.
- Changes should be easy to make, this is context dependent. For example - recommender systems are easier to adjust versus flight software.

4 Tenets for Organizations to Run Online CEs

- The organization wants to make data-driven decisions and has formalized an OEC.
- The organization is willing to invest in the infrastructure and tests to run CEs and ensure that the results are trustworthy.
- The organization recognizes that it is poor at assessing the value of ideas.

5 Improvements Over Time and Examples

- Improvements to key metrics are achieved over many small changes, from 0.1% to 2%.

- Examples come from Google and Bing – where relevant ad experiments (many of them!) drove up revenue month-to-month.
- Some interesting examples:
  - UI changes: 41 shades of blue were tested at Google, and the optimal shades of blue were shown to boost task completion and other key metrics across Google, such as time-to-success.
  - Greg Linden at Amazon prototyped personalized recommendation displays on the webpage. A marketing senior VP was against the idea, but Linden ran a CE that showed the success of the new feature.
  - An engineer at Bing changed the way JS was generated on a page, which reduced HTML sent to clients substantially. A CE over the change revealed a number of improved metrics. A 10 ms perforamance improvement in page loading paid for the annual cost of an engineer!

6 Strategy, Tactics, and Their Relationship to Experiments

- *Scenario 1* - A Business Strategy Exists and There are Enough Users to Experiment
  - Experiments can climb to a local optimumbased on the current strategy and product.
  - Experiments can help identify areas with high ROI (improving the OEC the most), help with optimizations that may not be obvious, help continuously iterate to better site redesigns, and can be crucial to optimize backend algorithms and infrastructure.
  - Having a strategy is necessary for experimentation, strategy drives the choice of OEC which cannot be gameable and should meet key characteristics.
  - Tying strategy with the OEC develops strategic integrity – a concept which emphasizes getting right strategies done and "matching top-down directed perspectives with bottom up tasks."
  - It's important to define guardrail metrics as well. What is the organization *not* willing to change?
- *Scenario 2* - You Have a Product and Strategy, but Results Point Towards a Pivot
  - Sometimes from internal data about the rate of change or external data from various benchmarks, we may have to change to a "different optimum" or even change the OEC or strategy outright.
  - It's always recommended to have a plethora of ideas, most of which optimize near the current experimental spot (or those which are big enough to "jump hills").

- Experiments could change in terms of: experiment duration and number of ideas tested.
- EVI is important to bear in mind as you experiment, this captures how additional info guides decision making. Running CEs allows for a reduction in uncertainty by trying a MVP, gathering data, and iterating.