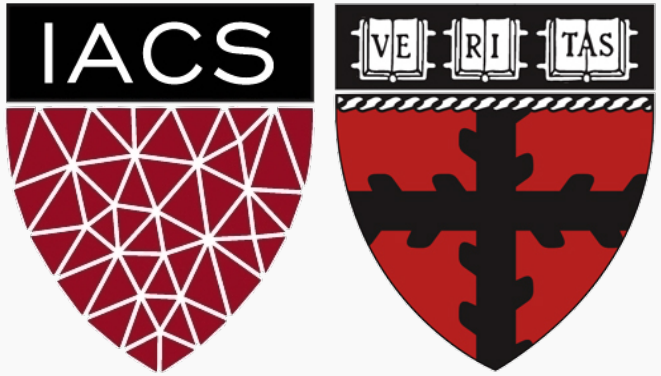
Lecture 23: AB Testing 1

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner



Announcements

HW 7 Clarifications:

- Don't get tripped up on the notation (what *Z* represents).
- Reporting: do not multiply by 100 (leave in decimal form)
- Scoring: not just the leaderboard (because there is a 'hidden' test set)
- Kaggle submissions: be sure to accept the terms and then join the competition

HW 8: will be **short** and on solely on Ed. Very little coding.

Outline

- Causal Effects
- Experiments and *AB*-testing
- *t*-tests, binomial *z*-test, fisher exact test, oh my!
- Obama 2008

Association vs. Causation

In many of our methods (regression, for example) we often want to measure the association between two variables: the response, *Y*, and the predictor, *X*. For example, this association is modeled by a β coefficient in regression, or amount of increase in R^2 in a regression tree associated with a predictor, etc...

If β is *significantly different* from zero (or amount of R^2 is greater than by chance alone), then there is evidence that the response is associated with the predictor.

How can we determine if β is *significantly different* from zero in a model?

Association vs. Causation (cont.)

But what can we say about a *causal association*? That is, can we manipulate *X* in order to influence *Y*?

Not necessarily. Why not?

There is potential for confounding factors to be the driving force for the observed association.

Controlling for confounding

How can we fix this issue of confounding variables?

There are 2 main approaches:

1. Model all possible confounders by including them into the model (multiple regression, for example). Or use fancy methods ('causal methods') to account for the confounders.
2. An *experiment* can be performed where the scientist manipulates the levels of the predictor (now called the *treatment*) to see how this leads to changes in values of the response.

What are the advantages and disadvantages of each approach?

Controlling for confounding: advantages/disadvantages

1. Modeling the confounders

- Advantages: cheap
- Disadvantages: not all confounders may be measured.

2. Performing an experiment

- Advantages: confounders will be *balanced*, on average, across treatment groups
- Disadvantages: expensive, can be an artificial environment

Experiments and *AB*-testing

Completely Randomized Design

There are many ways to design an experiment, depending on the number of treatment types, number of treatment groups, how the treatment effect may vary across subgroups, etc...

The simplest type of experiment is called a Completely Randomized Design (CRD). If two treatments, call them treatment *A* and treatment *B*, are to be compared across *n* subjects, then *n*/2 subject are randomly assigned to each group.

- If $n = 100$, this is equivalent to putting all 100 names in a hat, and pulling 50 names out and assigning them to treatment *A*.

Experiments and *AB*-testing

In the world of Data Science, performing experiments to determine causation, like the completely randomized design, is called *AB*-testing.

AB-testing is often used in the tech industry to determine which form of website design (the treatment) leads to more ad clicks, purchases, etc... (the response). Or to determine the effect of a new app rollout (treatment) on revenue or usage (the response).

Assigning subject to treatments

In order to balance confounders, the subjects must be properly randomly assigned to the treatment groups, and sufficient enough sample sizes need to be used.

For a CRD with 2 treatment arms, how can this randomization be performed via a computer?

You can just sample $n/2$ numbers from the values 1, 2, ..., n without replacement and assign those individuals (in a list) to treatment group *A*, and the rest to treatments group *B*. This is equivalent to sorting the list of numbers, with the first half going to treatment *A* and the rest going to treatment *B*.

This is just like a 50-50 test-train split!

Beyond just A vs. B

How can an AB test be expanded to include more than two options? What if there are more than just one type of treatment?

The **multivariate experimental design** generalizes this approach. If there are two treatment types (font color, and website layout), then both treatments' effects can (and should) be tested simultaneously. Why?

In a **full factorial experimental** design, each and every combination of treatments are considered different treatment groups. Experiments online are cheap. Full factorial designs are often possible and feasible.


I DON'T ALWAYS RUN
A WEBSITE TEST
BUT WHEN I DO, IT'S
ALWAYS MULTIVARIATE
made on imgur

t-tests, binomial *z*-test, fisher exact test, oh my!

Analyzing the results

Just like in statistical/machine learning, the analysis of results for any experiment depends on the form of the response variable (categorical vs. quantitative), but also depends on the design of the experiment.

For *AB*-testing (classically called a 2-arm CRD), this ends up just being a 2-group comparison procedure, and depends on the form of the response variable (aka, if *Y* is binary, categorical, or quantitative).

Analyzing the results (cont.)

For those of you who have taken Stat 100/101/102/104/111/139:

If the response is quantitative, what is the classical approach to determining if the means are different in 2 independent groups?

- a 2-sample *t*-test for means

If the proportions of successes are different in 2 independent groups?

- a 2-sample *z*-test for proportions

2-sample *t*-test

Formally, the 2-sample *t*-test for the mean difference between 2 treatment groups is:

$H_0: \mu_A = \mu_B$ vs. $H_0: \mu_A \neq \mu_B$

$$t = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

The *p*-value can then be calculated based on a $t_{\min(n_A, n_B)-1}$ distribution.

The assumptions for this test include (i) independent observations and (ii) normally distributed responses within each group (or sufficiently large sample size).