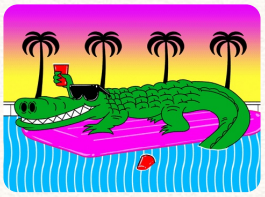
Date: Tuesday, March 19, 2019

EXPERIMENTS IN DATA SCIENCE



Examples of the need for experimentation

1) Economic indicators based on a country's poverty, employment rate, happiness, etc.
2) Medical treatments: will treatment A help control ailment B?

* **Focus**: Assessing cause & effect (aka causal analysis)

Purposeful data collection:

The collection of data in a planned, systematic way so that causal relationships can be inferred.

Notation and Nomenclature:

* **Dependent variable y:**

  Measures the outcome that we want to optimize over.
  Ex: CTR, session duration, bounce rate, etc.
  ↳ click-through rate

* **Explanatory variables $x_1, x_2, \dots, x_p$:**

  Variables that we expect to influence our dependent variable y.
  ↳ In an experiment, explanatory variables are referred to as **factors**.
  ↳ The values they can take on (eg. domain) are called **levels**.

Primary aim: Understand which combinations of explanatory variables have a causal relationship with y.

This inference gives us an action for future design/ engineering.

* **Experimental conditions:**

Unique combinations of the levels of one or more factors.

* **Experimental units:**

Applied to each condition and response value is recorded

Example 1: Button message

y_i: 1{ind i click button}

x_{i1} = message = $\alpha_1 1\{\text{"submit"}\} + \alpha_2 1\{\text{"go"}\} + \alpha_3 1\{\text{"let's go"}\}$ ← button i states

x_{i2} = color = $\beta_1 1\{\text{button i is red}\} + \beta_2 1\{\text{button i is blue}\}$

Conditions: {"submit", R} {"submit", B}

{"go", R} {"go", B}

{"let's go", R} {"let's go", B}

Experimental Units: Individuals that we're assigned each condition above.

Experiments vs Observational Studies

* In an **experiment**, we control and know how units are assigned to a condition. We can then assess causal relationships between conditions and the response.

* In an **obs. study**, we have no control over assignment to conditions. Instead, the data is observed passively. It is difficult to test for causality here, though methods do exist.
  - Ex: DAGs, propensity score matching, Granger causality.
    ↳ Directed acyclic graph

Example: A/B testing of user activity in seconds on version A + B of a website.

Conditions: {version A}, {version B} (2 conditions)

Dependent variable: y_i = time in seconds user i stays on the site

Experimental unit: The users!

Note: Assignment of units to conditions is done using various forms of randomization. The choice of randomization is typically referred to as the "**design**".

* Usually we cannot (or do not want to) assign units to multiple conditions.
  (Drug treatment; version of a webpage: seeing too many confuse or frustrate an user).

* Because of this, we do not measure the dependent variable for a user on at least one condition. The unobserved response for that user/condition is called a **counterfactual**.

* The primary aim of design is to ensure that the only difference(s) we see in response are due to difference in condition. (Thus, we need to control for other intrinsic features).