Experimental Design

Sean Sall

May 18th, 2016

Objectives

- Understand the difference between an *experimental setting* and an *observational setting*
- Understand what a *confounding* factor is, and how to identify one
- Think critically . . . always and forever

Agenda

- Experimental Design
  - Experimental Setting
  - Observational Setting
- Case studies & discussion

Why does this matter?

In a perfect world, we'd be able to set up state-of-the-art experiments where we are able to control for all factors in such a way that we can establish cause-effect or predict with one hundred percent accuracy. In reality, this **never** happens.

One of the best skills you can develop as a data scientist moving forward is knowledge of how to think about different data sets and the error, bias, etc. inherent in those data sets. This will help you in subsequently determining what to do with those data sets in order to achieve your desired results.

Experimental Design Overview

Experimental Design

The goal of **experimental design** is to establish causality, estimate effect size, and avoid bias

Experimental versus Observational

- An **experimental** setting is one where:
  - Subjects are randomly assigned to experimental groups (often control and treatment) and then we observe the effect of some treatment
  - **Can be used to establish causality**
    - ⋆ Example: We randomly assign homework to students and measure the performance of the two groups (homework versus no homework)
- An **observational** setting is one where:
  - Subjects are observed without assignment to experimental groups
  - **Can't be used to establish causality**
    - ⋆ Example: We simply find students who did and didn't do their homework and measure the performance of both groups (homework versus no homework)

Confounding Factors

- A **confounding factor** is an external factor that correlates with the dependent variable and independent variable
  - For our homework example, whether or not a student is hard-working is a confounding factor, since it might affect performance
- In an **observational** setting, there are often many confounding factors that we can't control for
- In an **experimental** setting, the random assignment helps to alleviate the need to control for confounding factors

Confounding Factors - Questions

- What are some confounding factors when we are looking at weight gain (dependent variable) and activity level (independent variable)?
- What are some confounding factors when we are looking at recovery (dependent variable) and drug usage (independent variable)?

Experimental Design Examples

Example 1 - The Gold Standard

Let's assume that we're Warren Buffet, and we're sparing no expense to set up an experiment that tests whether or not a new drug is effective. How do we approach this?

Given Warren Buffet's piggy bank, we should be able to get as close to a **gold standard** study as possible (e.g. one that allows us to most accurately measure any underlying cause-effect relationships).

Example 2 - A/B Testing

Let's pretend we're testing two different websites against each other (commonly referred to as **A/B testing**). Further, let's pretend that we know we're going to randomize what visitor sees what page. That is, when they visit our site, we'll flip a non-biased coin, and if it's heads they see page A, and if it's tails, they see page B. But, let's consider that when we randomize, we might be sending some visitors to a terrible version of the site where they never click through. So, we decide that we're only going to follow this randomization procedure for a **subset** of our visitors.

Given this knowledge, is there anything you would make sure to suggest to those overseeing this testing strategy? Is there a way that you could imagine them running this randomization strategy such that they aren't getting a good idea of which site is truly more effective?

Example 3 - We get some data!

Let's imagine that you're at work, and your boss walks in and hands you some data, telling you there is a presentation over the data in two hours (and that's it). What do you do?

Creating ground truth motivation

Sometimes in data science, we end up with data that doesn't have explicit labels handed to us, but have to in some way establish labels (e.g. what we'll call **ground truth**) given what's in the data. For example, you'll work with a data set later in the program where you have to decide what you consider to be a fraudulent transaction versus what isn't.

Other times, we have ideas for projects that involve supervised learning, but don't have explicit labels for our data. In this case, we have to figure out how we might label the data.

In both of these cases, we need to be aware of:

- The errors/biases that are present in the raw data
- What kinds of errors/biases can be introduced in producing the labels in one way over another.

Example 4 - Forest Fires Problem

Consider having two different data sets:

1. A data set of observations that are geographically located (by lat/long), and have been identified as fires by some satellite imagery.
2. A data set of observations that are forest-fire perimeter boundaries and also happened to be geographically located.

Using data set 2, how do you go about establishing the ground truth of which observations from data set 1 are actually forest-fires? Some questions to consider:

1. Are there particular questions that you ask about the data set, especially when considering other variables that might be useful to have in the labeling process?
2. What error/bias might exist in each of the original data sets?
3. What error/bias might you be introducing in each of the given labeling processes that you consider?

Example 4 - Forest Fires Graphic

The following graphic might be helpful. The green dots are individual observations for a given day, and the blue boundary is a forest-fire perimeter boundary for that same day.

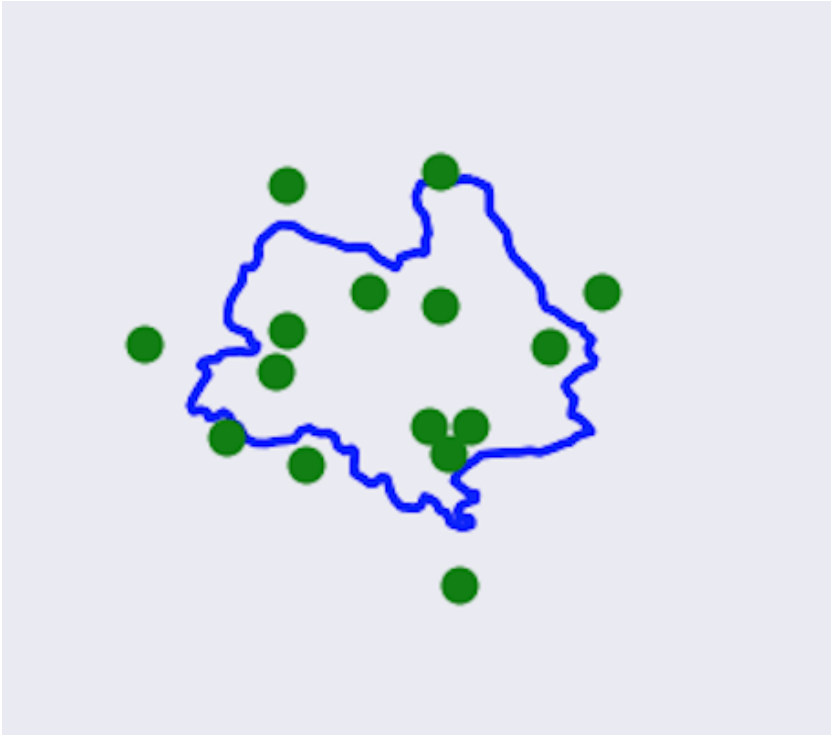

Figure 1:Fires

Example 5 - Eye Cancer

Imagine you're a parent whose son/daughter has had a special type of eye cancer, and you think you can build a machine learning model to try to identify that type of eye cancer. How do you go about this?

Questions to consider:

1. Where do you get your data?
2. What sources of bias are there in your data?
3. If you have to hand label data, what are some potential complications that might arise in that labeling process?