PEARSON BUSINESS ANALYTICS SERIES

Product Analytics

Applied Data Science Techniques for Actionable Consumer Insights

Joanne Rodrigues

Product Analytics

6

Why Are My Users Leaving? The Ins and Outs of A/B Testing

In the prior chapters, we discussed theory development and the process of metrics creation and development. All of the tools we learned in prior chapters are used in conjunction to further our understanding of user behavior. To test our theories, we need to establish hypotheses and test them. We use conceptualization to create quantitative metrics to measure qualitative concepts and test our hypotheses. Then, we use the results of these tests to either falsify or confirm our theories.

In this chapter, we focus on the process of testing our hypotheses. We assume that you have built a theory with an eye toward behavior change, used the metrics creation techniques to develop sound/interesting outcome metrics, and picked interesting treatments to test.

This chapter provides a practitioner's how-to guide to A/B testing. Metrics are often our outcome variables for A/B tests. This chapter covers the following topics:

- The differences between causal inference and correlation
- The nuts and bolts of A/B testing
- Building statistical tests to test our hypotheses

The first part of this chapter explains the need for A/B testing by exploring the difficulty of inferring *why* something happens from everyday data. The second part is a how-to guide to A/B tests, including conducting statistical tests to determine the effects of an intervention or treatment. The third part discusses common errors that occur with user data and how to avoid these in practice.

6.1 An A/B Test

We will start by defining an A/B test. An A/B test, also known as a split test, involves randomly picking some users for the first variant (usually the control) and some users for the second variant (the treatment).

For instance, in a medical setting, we randomly divide a group of subjects into two subgroups: one group will get the new drug (the treated group), and the other will get a placebo or sugar pill (the control group). In this setting, the A/B test is generally called a *randomized controlled* trial (RCT), but it is based on a similar statistical design as is applied in nonmedical settings. Next we provide some helpful definitions for treatment and control groups, which will be used throughout this chapter. We will build on these ideas in the section focused on advanced causal inference techniques.

In a medical setting, there are often strict guidelines placed on testing, since you are working with human subjects. In contrast, in the world of user analytics, the guidelines are relatively slim. Instead, the core problem in user analytics is selection. Selection means that the users coming to your site are nonrandom. Medical trials are often randomized, and many medical studies actively try to increase the sample representativeness. In A/B testing, you often have a strongly selected population to begin with. It is hard to get a representative sample of the general population in such a case.

When an A/B test takes place in a web context, that setting allows for more flexibility, repeated testing, and more variations. However, there is often limited follow-up compared to a medical trial.

- **Treatment:** Some element, action, or feature that could have a causal effect on something else. It could be a change to our website, a behavioral action, a medication, or something else.
- **Treated group:** A random group of participants who are given the treatment.
- **Control group:** A random group of participants who do not receive the treatment. They do not know that they did not receive treatment and often are given a placebo to prevent them from inferring that they did not receive the treatment. In the world of user analytics, it's easier to get away with no placebo, because you do not have to inform users that they are being tested.

We want all the conditions to be the same for both the treatment and the control, except for the treatment variable. In this case, we will change the cancer drug that we give to study participants.

We also need an outcome or some metric, which we can use to determine our treatment effect. In our medical trial, the outcome is life expectancy and the treatment effect is how much longer our cancer patients live based on receiving the new drug versus the placebo.

In a web context, suppose we want to test the effect of a new algorithm. We could add an algorithm to suggest friends to users in our snowmobile site. The random users assigned to the treatment will see the new algorithm's results. The users assigned to the control do not see the new algorithm's results.

We then consider the effect of viewing the results of the algorithm on retention. In this case, one outcome we may consider is the difference in average retention between the group who saw the new algorithm's results and the group who did not.

To make sure that our theories are indeed correct, they often need to be tested. This is not always readily apparent, so we'll spend some time in the next two subsections exploring why and when A/B testing is needed. The introductory sections will also help you recognize incorrect inferences from nonexperimental data.

6.2 The Curious Case of Free Weekly Events

Before we jump into A/B testing, let's explore why it's difficult to understand *why* something happens from regular "observational" data. **Observational data** is simply the data that we collect on what customers do. We don't need any special setup to collect this data. How, then, does observational data lead us astray? We'll explore a common example.

Angry Dodo Birds: The Sand Saga, a mobile gaming company, hires you as a data consultant. This company runs special events monthly to help bolster engagement in its product. The events allow free access to the gated web features such as higher, more complex levels.

The product managers think weekly events *cause* more purchasing because they find users who took part in these weekends buy more stuff on their site—free dodo treats and shiny weapons. Since the leadership team identifies these events as causing more purchasing, they want to increase the number of these free events and the number of users who have access to those events to drive more purchasing. Of course, this comes at a cost, so they ask you: Will this work?

SpellBook, another well-known mobile product, also needs your consulting wisdom. SpellBook is a social network that hosts weekly spelling competitions between users. Users can compete against other users for perks. Users who participate in these competitions make more purchases in the product, such as spelling bee videos and Scrabble games. Again, the executives think that greater engagement in these competitions *causes* greater purchasing. Due to this perceived causal relationship, they want to increase the number of competitions and randomly place new users into these competitions. They also ask you: Will this work?

Do you think that these strategies were successful? Basically, if the executives are right and either the special events or competitions *cause* greater purchasing, then these strategies of expanding access would greatly increase their revenue.

However, neither company could tell from the data that either of these events actually causes greater purchasing in their product. Just because two variables are highly related, this does not mean that one **caused** the other.

The examples described here are true stories (though obviously different company names and products were used), and this scenario happens often in industry. An A/B test was run at both companies, and executives at Angry Dodo Birds were right: Monthly special events had a **huge causal effect** on purchasing. In contrast, the executives at SpellBook made an erroneous assumption: In fact, greater engagement in these competitions did not lead to more purchasing. There was almost **no causal effect** of the competitions, only a strong correlation.

With the data currently available to either Angry Dodo Birds and SpellBook, there is no way for you to know as the data consultant if there is a causal effect of either the greater access on weekends or the spelling competitions on purchasing. *We need to either run an experiment (an A/B test or other type) or try to use statistical techniques to properly assess causality from observational (nonrandomized) data.*

This example shows us that a "relationship" between two variables need not be causal. Two variables could, indeed, just be related, without one causing the other.

To better understand noncausal relationships, we need to understand the concept of correlation. Correlation is the directionless relationship between two things. An example is the gross domestic product (GDP) and employment. As the GDP increases, the population employment rate also goes up; conversely, as the population employment rate goes up, GDP increases. Measures of these two economic ideas are related. We don't know if there is a causal relationship, but we know that they generally move together.

The next section discusses why making conclusions from observed data often doesn't result in inferences that are causal, but rather those that are correlative in nature. In later sections, we'll explore what we need to make causal inferences.

To understand how a relationship can be solely correlative, we'll discuss spurious correlation (a correlation driven by an outside variable), selection bias (patterns in nonrandomized data that make inferences difficult), and randomness (data devoid of built-in patterns). These concepts will help us understand the need for A/B testing.

6.2.1 Spurious Correlation

Let's start with the concept of spurious correlation. A **spurious correlation** is a relationship between two variables driven by a third outside variable.

Most people struggle to understand the concept of spurious correlation when they first see it. It's absolutely *not intuitive*. Consider the following thought experiment. Figure 6.1 is the number of "likes" for a popular movie and the number of downloads of that movie on a company's website. Do you think that the number of movie "likes" causes the number of downloads, based on looking at Figure 6.1?

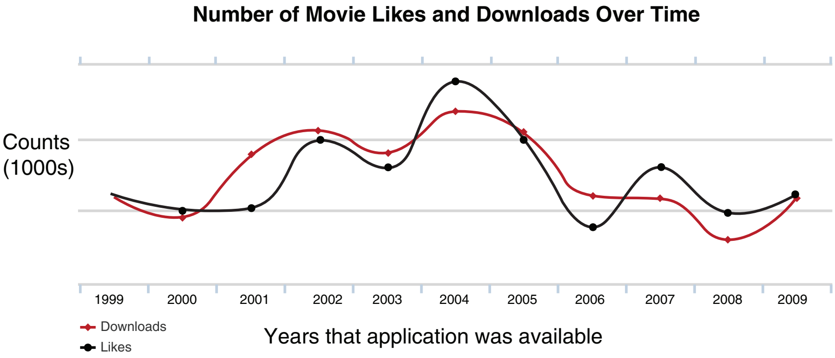

Figure 6.1 The relationship between "likes" and downloads.

When asked this question in workshops, participants will often say, "Yes, of course—don't you see the relationship?" The follow-up question is then, What is the mechanism? A mechanism is how one factor causes the outcome through any mitigating or intermediate variables. *Why do you think "likes" cause more downloads?* Think about your answer. There are lots of potential explanations. Here are some common ones: "Likes" lead others to think the movie is good, so more people download it. "Likes" lead to more people viewing the movie page. Since items with more views come up in more feeds, that leads more people to view the movie and download. The reasoning goes on and on.

Figure 6.2 shows the real names for two variables: the number of people killed by venomous spiders and the number of letters in the winning word at the Scripps National Spelling Bee. How many people would think the number of letters in the winning word *caused* venomous spiders to attack? None, of course.

Lots of variables can be related (in this case, letters and spider killings are highly correlated) even if one did not cause the other. *Correlation is a linear relationship between two variables. It does not imply that one causes the other.*

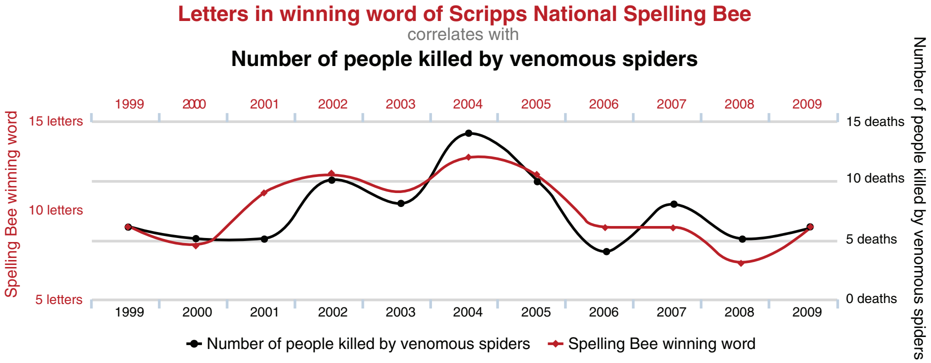

Figure 6.2 The relationship between spelling bees and venomous spider deaths.

The variables that you care about in a web product are more likely to be causally related than letters and spider bites, or windstorms in the Sahara and the stock market. Even so, that doesn't mean that thinking a causal relationship exists without empirical support is any less false.

Regarding whether "likes" *cause* more downloads in our movie example, there could be a spurious relationship with page views. Page views can cause more "likes" of the movie, since people need to see the movie page to "like" it. Page views can also cause more downloads of the movie, as people need to see a movie to download it. More people viewed the movie, and therefore more people got interested and downloaded it. The number of likes had no effect on the number of downloads. Perhaps users never even noticed the "like" icon.

Many correlations in your product are driven by **spurious relationships**. A spurious relationship between B and C occurs when some variable A is causing B and C to go up. When you see B and C go up, you might assume that they are causally related. However, B and C are only related through A, so your inference about a relationship between B and C is incorrect. In Figure 6.3, we can see a visual representation of this idea. If we only look at B and C, we incorrectly assume a causal relationship between them. However, if we widen the view, we notice that A causes both.

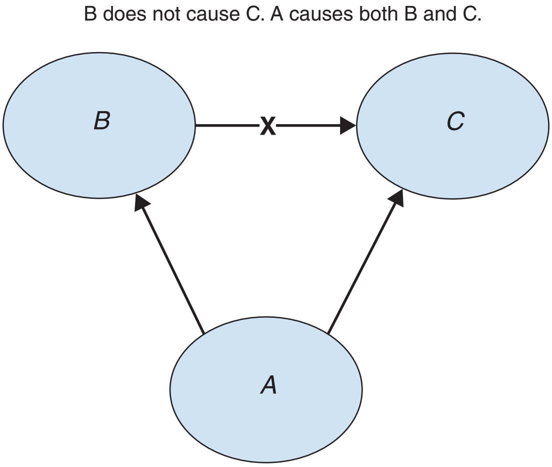

Figure 6.3 Spurious relationship. A causes B and C. However, one could mistakenly think B causes C.

To flesh out the idea of spurious correlation, we need to understand another concept, selection bias. Selection bias occurs when a set of data has preset selection patterns that prevent us from making valid causal inferences. It's related to spurious correlation because it prevents us from differentiating between spurious and causal relationships in real data.

6.2.2 Selection Bias

The reason it's difficult to identify spurious relationships is because of selection bias. **Selection bias** occurs when people self-select (or are nonrandomly selected) into certain behavioral patterns. An example will help us understand this definition. Suppose your company puts you in charge of analyzing the data from historical health awareness days. The managers ask you, Does having a company health awareness day increase company retention? During the annual health awareness day, your company holds a 5K run for employees and their families. A small number of employees sign up and do the 5K run. Those who do sign up to stay at the company two years longer, on average, than those that do not.

When the company managers see this data, they argue that everyone should be forced to do a 5K run on the next health awareness day. The CEO agrees, and you analyze the data again and find that the retention number does not go up. In fact, it goes down. Why? The effect that you see on retention could be selection bias, meaning that there is no causal relationship between the run and retention.

Selection bias arises when users are not randomly picked from the population, but are selected by some other force. In this example, those more invested in the company are more likely to run the 5K race when it's optional. The fact that employees who run the 5K race are more invested than the average employee is driving the lower turnover of these employees. Company interest is our hidden A variable driving the correlation between our participation in the 5K run and higher retention. We can see that the selection bias is the nonrandom way in which employees are selecting into the 5K run; it is driving the difference in retention. The treatment of running the 5K run is not causing the increase in retention.

The key to understanding why we cannot make inferences from this data is to understand how observational data differs from randomized data. Randomness is like rainbows. This metaphor might seem over the top, but it's really not. Randomness to a statistician is like rainbows to a small child: It will brighten your day.

Of course, you don't believe me yet, so let's go back and reframe our example. Let's say that instead of letting anyone sign up for the company's 5K run, we randomly assigned participants. Judy from accounting was included, while Steve in billing had to cheer Judy on from the sidelines. Now, we find the same result: Participation in the 5K run increased company retention. Well, that's it, folks, we're done. We can all go home now.

Really, it's that simple. When the B variable we're interested is randomized, we can assume that the effect is causal. Why? Because randomness eliminates the spurious relationships. So, you can tell Jim, the CEO, to have more health awareness days and encourage everyone to run a 5K. It will have a positive causal effect.

Since it's much easier to determine correlation than causation, in the next section, we'll nail down the technical definition of correlation. It's important not only to understand the concept, but also how it's traditionally measured.