A/B Testing and Beyond: Designed Experiments for Data Scientists

A Continuing Education Certificate

at

The Univeristy of San Francisco's Data Institute

September 4 - October 16, 2018

Instructor: Nathaniel T. Stevens

ntstevens@usfca.edu

Preface

Over the last few decades, there has been an explosion in the amount of data that companies are using to inform decisions. Much of the insight drawn from this influx of data is correlational. Indeed data science is often associated with machine learning, which is powerful in its ability to find patterns and relationships in data for purposes of prediction and classification. However, the ease with which data can be collected provides an enormous opportunity to identify and quantify causal relationships, obtained via experimentation. When causal inference is required, a carefully designed experiment is necessary to evaluate the impact of altering one or more variables on some outcome of interest.

Designed experiments are key to the Scientific Method and are necessary for understanding the world around us. Historically experiments have been used in fields such agriculture, biology, physics, chemistry, pharmacology, epidemiology and industrial engineering, to name a few. More recently however, the utility of designed experiments has been recognized in the world of business and marketing as a tool to increase conversion, strengthen customer retention and improve the bottom line. Companies like Google, Amazon, Facebook, Netflix, Airbnb and Lyft have all adopted experimentation and A/B testing for these purposes. As such, data science practitioners and professionals are beginning to acknowledge experimentation as a foundational tenet of the field.

In this course participants will be exposed to the value experimentation; a strong emphasis is placed on the importance of thinking critically and carefully about the manner in which metrics should be selected and measured, and how data should be collected and analyzed in order to address and answer questions of interest. In particular, this course provides a thorough treatment of available methods and best practices in the design and analysis of experiments. Broad topics include A/B/n testing in which two or more variants are compared, multivariate experiments such as factorial and fractional-factorial designs, and optimization techniques such as multi-armed bandit experiments and response surface methodology.

What this course does not emphasize is third party experimentation platforms such as Optimizely, Google Analytics, Wasabi, Mixpanel, Apptimize, Split or AB Tasty. While the physical construction of variants and the collection of data is a necessary part of experimentation, there is no standard platform used by all data scientists at all companies. For this reason it would be a poor use of time to train participants in the use of any one platform in particular. The reality is that data scientists will use the experimentation platforms and data pipelines espoused by their own companies.

What this course does emphasize is the statistical principles and practical considerations that underlie effective experimentation. Specifically, participants will develop an appreciation for the careful navigation of the choices and nuances associated with the design of an experiment. Participants will also develop a mastery of the relevant hypothesis tests, power analyses, sample size calculations and analysis methods necessary to draw conclusions and make impactful statements about the question of interest. Participants will also become familiar with using `R` and `Python` to automate components of both the design and the analysis of experiments.

Introduction

In this chapter we discuss what an experiment is, how it differs from other data collection strategies, and why it is so useful. We will also discuss important concepts and important decisions that must be considered when planning an experiment, and we package all of this within a general framework for solving problems and answering questions with planned investigations. First, however, we will lay a foundation of notation and nomenclature which will help to make discussions in this course clear and concise.

1.1 Notation and Nomenclature

In all planned investigations interest lies in solving a problem or answering a particular question using data. The data available for such a task are typically composed of measurements on one or more variables. Here we make a distinction between two classes of variables, based on our interest in them.

The problem/question we wish to address is typically defined in the context of optimizing some metric of interest. In practice such metrics tend to be performance metrics or key performance indicators (KPIs) such as conversion rates, average purchase size, bounce rate, maximum page load time or average session duration, to name just a few. The variable whose measurements are used to calculate such a metric is referred to as the **response variable**. For example, an experiment may involve comparing different messages on a call-to-action button to find which message maximizes the click through rate (CTR). The metric of interested here is the CTR and the corresponding response variable is a binary indicator which identifies whether or not users click the button. As a second example, an experiment may involve the comparison of different webpage designs to decide which one maximizes the average time on page (TOP). Here the metric of interest is average TOP and the corresponding response variable is the continuous measurement of time on page for each user. Regardless of the type or goal of the experiment, the response variable is the one we are primarily interested in. Throughout this course we will use the letter y to denote response variables.

The variable(s) we believe may influence the response variable are called **explanatory variables** and we tend to think of them as having secondary importance relative to the response variable. In a sense, these are independent variables whereas the response is a dependent variable. In the context of experimentation we refer to explanatory variables as **factors** and we denote them with the letter x. In the simple examples above, the button's message and the webpage's design are the factors that infuence CTR and average TOP, respectively.

The different values that a factor takes on in an experiment are referred to as **levels**. Suppose in the button message experiment the following three messages are being tested: *"Submit"*, *"Go"*, and *"Let's Go!"*. In this case the factor 'button message' has three levels: {*Submit*, *Go*, *Let's Go!*}. In the webpage design experiment, suppose two designs are being considered: one with a static image and one with a rotating carousel of static images. In this case the factor 'webpage design' has two levels: {*photo*, *carousel*}. It is plain to see that factor levels are what define different **experimental conditions**.

In general, the purpose of an experiment is to alter the levels of one or more factors, and then observe and quantify the resultant effect on the response variable. In order to do this, we must expose **experimental units** to different levels of the factor(s) under study (i.e., to different conditions) and measure their corresponding response value. In the context of online experiments like the examples above, the units are typically users or customers. Suppose that the button in the button message experiment must be clicked in order to complete a digital survey. The users that are exposed to the three different 'button message' conditions are the experimental units.

We note briefly that an experiment is not the only way to learn about the relationship between a response variable and one or more factors. In the next section we consider two different data collection strategies and discuss the advantages and disadvantages of each with respect to understanding the relationship between y and one or more x's.

1.2 Experiments versus Observational Studies

An **experiment** is composed of a collection of conditions defined by purposeful changes to one or more factors. The goal is to identify and quantify the differences in response variable values across conditions. In other words, the goal is to evaluate the change in response elicited by a change in the factors. In determining whether a factor significantly influences a response, like whether a button's message significantly influences CTR, it is necessary to understand how experimental units respond when exposed to each of the corresponding conditions. However, we cannot simultaneously expose the *same* set of units to each condition; a group of units can be exposed to just one condition. Unfortunately, then, we do not observe how the units respond in the conditions to which they were not exposed. Their hypothetical and unobservable response in these conditions is what we call a **counterfactual**. Because counterfactual outcomes cannot be observed, we require a proxy. Thus, instead, we randomly assign a *different* set of units to each condition and we compare the response variable measurements across conditions. When the units are assigned to the conditions at random, it is reasonable to believe that the only difference between the units in each condition is the fact that they are in different conditions. Thus, if there is a marked difference in the response between the conditions, then this difference can be attributed to the conditions themselves. In this way, we conclude that the observed difference in response values was **caused** by the condition the units were in, and hence by the controlled changes that were made to the factors. The key here is that the the factors are purposefully controlled in order to observe the resulting effect on the response.

As mentioned above, generally speaking, the goal in these sorts of investigations is to evaluate the change in response associated with a change in the factors. Strictly speaking one does not require an experiment to do this. Establishing these sorts of relationships can also be done with **observational studies**. The distinction between this and an experiment is that in an observational study there is no measure of control in the data collection process. Instead, data are recorded passively and any relationship between the response and factors is observed organically. While such an approach provides information about the association between these factors, it does not provide clear information about a causal relationship. When **causal inference** (establishing causal connections between variables) is of interest, it is best if the data arise as a result of an experiment. While methods for establishing causal relationships from observational data do exist (see e.g., propensity score matching (Rosenbaum and Rubin, 1983)), they are much less sound and much more error prone than a carefully designed experiment.

Thus, experiments are advantageous because causal inference is easier than in the context of an observational study. However, experiments can be risky and costly. Consider the situation in which an experimental condition very negatively effects the user experience and results in a revenue loss. This is an outcome, that if at all possible, one would like to avoid.

Another drawback to experimentation is that some experimental conditions may not be eithical. For example, in evaluating whether smoking causes lung cancer, it would be unethical to have a *'smoking'* condition in which subjects are forced to smoke. As a second example, in a pricing experiment it may be perceived as unethical to randomize users to different pricing conditions in which some users pay more money for the same product than others. Shmueli (2017) discusses ethics in online experimentation and points to a recent and controversial emotional contagion experiment at Facebook as being unethical.

While observational studies do not facilitate causal inference as easily as experiments do, they enjoy protection from these other issues since nothing is being manipulated or controlled. Users behave as they normally would and are not forced to participate in something which may be costly or which may be unethical. Thus there is a trade-off between experiments and observational studies: experiments facilitate causal inference, but they can be costly and unethical whereas observational studies are the exact opposite. Thus a data scientist planning an investigation should consider the goals of the investigation and choose their data collection strategy carefully.

In the next section we discuss a framework for planning investigations that formalizes the process by which data is collected to answer questions, regardless of the data collection strategy.

1.3 QPDAC: A Strategy for Answering Questions with Data

In this section we discuss a framework for planning and executing an investigation whose results are in turn analyzed so that conclusions may be drawn about some question of interest. This framework is referred to as QPDAC, an acronym that stands for *Question*, *Plan*, *Data*, *Analysis* and *Conclusion* (Steiner and MacKay, 2005). While this approach is suitable for any formal data-driven investigation, here we emphasize its utility in designing and analyzing experiments. We describe each step of this framework in turn.

Question: Develop a clear statement of the question that needs to be answered. This statement will correspond to some hypothesis that you would like to prove or disprove with an experiment. For example, in the webpage design experiment a question statement might look as follows: "*Relative to the original webpage design with a static image, does a rotating carousel of images decrease bounce rate?*". It is important that this statement is clear, concise and quantifiable because it will influence many decisions associated with the design and analysis of the experiment. It is also important that everyone involved in the experiment - from data scientists and analysts to product managers and engineers - is aware of the question of interest and hence the goal of the experiment. Experiments may have many goals including, for example, factor screening, optimization or confirmation (we will elaborate on each of these types of experiments as the course progresses). But no matter the goal, it is important that everyone involved is aware of it, and committed to the success of the experiment. Siroker and Koomen (2013) stress the importance of building a culture of testing and experimentation within your organization. When such a culture exists, experimentation is highly valued and can become maximally beneficial. Clearly communicating the question is an excellent first step toward this end.

Plan: In this stage the experiment is designed and all pre-experimental questions should be answered. For example, it is at this stage that the response variable and experimental factors must be chosen. This may seem trivial, but it is arguably the most important step in any experiment and careful consideration should be given to these choices. When choosing the response variable it is important to consider the **Question**; it is through measurements of this variable that the question is answered and so it is necessary to choose a metric that is related to this question and whose variation can be quantified.

The choice of which factor(s) to manipulate in the experiment will also be guided by the **Question**. Recall that factors are the variables we expect to influence the response. It is important at this stage to brainstorm all such factors that might influence the response and make decisions about whether and how they will be controlled in the experiment. We classify factors into one of three types:

i. **Design factors:** factors that we will manipulate in the experiment and that define the experimental conditions

ii. **Nuisance factors:** factors that we expect to influence the response, but whose effect we do not care about. These factors are typically held fixed during the experiment so as to eliminate them as a source of variation in the response variable.

iii. **Allowed-to-vary factors:** factors that we *cannot* control and factors that we are unaware of. In either case these factors are ones that we do not control in the experiment.

Once these choices have been made it is necessary to define the experimental conditions by deciding which levels of the design factor(s) you will experiment with.

Related to the choice of response variable and design factors is the choice of experimental units. After all, it is the units that are exposed to the different conditions and on which the response variable is measured. In many situations this will be an obvious choice, like an app's users or a company's customers. However, in other situations this decision is not so straightforward. For example, consider online marketplaces like Ebay, Etsy or Airbnb in which it is conceivable that the experimental unit could be the seller/owner or the buyer/renter. The type of question being posed and the particular response variable being measured will typically influence this choice.

With the units defined, conditions established, and the response variable chosen, the final decisions to be made concern the number of units to assign to each condition, and the manner in which this assignment is made. Power analyses and sample size calculations are used to address the former concern and the sampling mechanism addresses the latter. While random assignment is the standard approach, other hierarchical assignment strategies such as stratified or segmented sampling are also common. We elaborate on these topics later on in the course.

Data: In this stage the data are collected according to the **Plan**. It is extremely important that this step be done correctly; the suitability and effectiveness of the analysis relies on the data being collected correctly. Computer scientists often use the phrase "garbage in, garbage out" to describe the phenomenon whereby poor quality input will always produce faulty output. This sentiment is true here also. If the data quality is compromised, the resulting analysis may be invalid in which case any conclusions drawn will be irrelevant.

One particularly important data quality check is to ensure the assignment strategy is working properly. If the **Plan** requires that units be randomly assigned to conditions, it is prudent to confirm whether condition assignment does appear to be random. A common approach for this is an A/A test, where units are assigned to one of two *identical* conditions. If the assignment was truly random, characteristics of the two groups of units (i.e., measurements of the response variable or demographic composition) should be indistinguishable. If they aren't, then there is likely something wrong with the assignment mechanism or the manner in which the data are being recorded. Either way, there is a problem that needs to be fixed prior to running the actual experiment.

Analysis: In this stage the **Data** are statistically analyzed to provide an objective answer to the **Question**. This is most typically achieved by way of estimating parameters, fitting models, and carrying out statistical hypothesis tests. If the experiment was well-designed and the data were collected correctly, this step should be straightforward. Throughout the course we will discuss, at length, a variety of statistical analyses whose suitability will depend on the design of the experiment and the type of data that were collected.

Conclusion: In this stage the results of the **Analysis** are considered and one must draw conclusions about what has been learned. These conclusions should then be clearly communicated to all parties involved in - or impacted by - the experiment. Clearly communicating your "wins" or what you learned from your "losses" will help to foster the culture of experimentation Siroker and Koomen (2013) suggest organizations should strive for.

It is very common that these results will precipitate new questions and new hypotheses that further experimentation can help answer. As we will emphasize routinely throughout the course, effective experimentation is sequential; information learned from one experiment helps to inform future experiments and knowledge is generated through a sequence of planned investigations. In this way, the QPDAC framework can be viewed as an ongoing cycle of knowledge generation as illustrated in Figure 1.

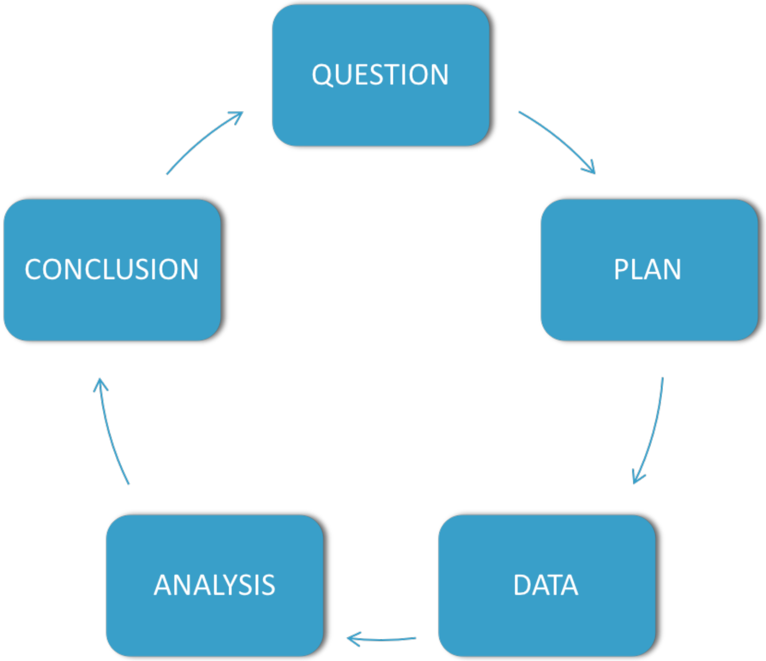

Figure 1: QPDAC Cycle

1.4 Fundamental Principles of Experimental Design

Having now described the merits and utility of experimentation, and having provided a framework for planning and executing such an investigation, we now describe three fundamental experimental design principles that should be considered when planning any experiment: *randomization*, *replication*, and *blocking* (Montgomery, 2017). You will see that we have briefly mentioned these concepts previously, but we formalize them here.

Randomization refers both to the manner in which experimental units are selected for inclusion in the experiment and the manner in which they are assigned to experimental conditions. Note that to avoid the risk of underperforming conditions or conditions with negative side effects, online experiments typically do not include all possible units (users). Instead, some fraction of them is selected for inclusion in the study. Then, once selected, the experimental units are assigned to one of the experimental conditions. Thus we have two levels of randomization.

As we will see later in the course, the validity of many methods of statistical analysis and statistical inference rely on the assumption that inclusion and assignment were done at random. However, there is a more intuitively appealing justifcation for randomization. The first level of randomization exists to ensure the sample of units included in the experiment is representative of those that were not. This way, the conclusions drawn from the experiment can be generalized to the broader population. The second level of randomization exists to balance out the effects of extraneous variables not under study (i.e., the allowed-to-vary factors). This balancing, in theory, ensures that the units in each condition are as similar to one another as can be, and thus any observed difference in response values can be attributed to the differences between the conditions themselves.

Replication refers to the existence of multiple response observations within each experimental condition and thus corresponds to the situation in which more than one unit is assigned to each condition. Assigning multiple units to each condition provides assurance that the observed results are genuine, and not just due to chance. And as the number of units in each condition increases (i.e., with more replication), we become increasingly sure of the results we observe. For instance, consider the button message experiment introduced previously. Suppose the CTRs in the *Submit*, *Go* and *Let's Go!* conditions were respectively 0.5, 0.5 and 1. If these click-through-rates were calculated from 2 users in each condition, the results would not be nearly as convincing as if they had been calculated from 1000 users in each condition.

The importance of replication likely seems obvious, but the answer to the question *"how much replication is needed?"* is likely less obvious and is just as important. More directly, this question is equivalent to asking *"how many units should be assigned to each condition?"*. The **sample size** for a given condition, denoted by n, is defined to be the number of units exposed to that condition. We use power analyses and sample size calculations to determine how many units to include in the study, and hence how many response variable observations are necessary to be sufficiently confident in your results. In the context of online experiments, where website traffic is heavy and predictable, replication is often communicated in terms of time as opposed to number of units. For instance, a common question is *"how long does the experiment need to run for?"*. Intuitively, the more confident one wishes to be in the experiment's results, the larger the sample size needs to be and hence the longer the duration of the experiment. We will formalize these reflections in the chapters to come.

Blocking is the mechanism by which nuisance factors are controlled for. Recall that nuisance factors are known to influence the response variable, but we are not interested in these relationships. Because we wish to ensure the only source of variation in response values is due to the experimental conditions (i.e., changing levels of design factors), we must hold the nuisance factors fixed during the experiment so that they do not impart any variation. Thus we run the experiment at fixed levels of the nuisance factors, i.e., within **blocks**.

For example, consider an email promotion experiment in which the primary goal is to test different variations of the message in the subject line with the goal of maximizing 'open rate'. However, suppose that it is known that 'open rate' is also influenced by the time of day and the day of the week that the email is sent. So as not to conflate the influence of the email's subject with these time effects, we may elect to send all of the emails at the same time of day and on the same day of the week. Here the block is the particular day and time of day in which the emails are sent. Blocking in this way eliminates these additional sources of variation, and guarantees that observed variation in the response variable is not due to time-of-day or day-of-week effects.

1.5 Exercise: The Instagram Experiment

We end this chapter by pretending we are data scientists at Instagram that need to design an experiment concerning sponsored ads. While ads serve as a source of revenue for Instagram, they also serve as a source of frustration and annoyance to users. Thus, we would like to run an experiment to gain insight into the interplay between ad revenue, user engagement and factors such as ad frequency, ad type (photo/video), whether the ad's content is targeted or not, etc. Ultimately the goal is to identify a condition that maximizes ad revenue without simultaneously plummeting user engagement below some minimally acceptable threshold.

How would you design such an experiment?

References

Montgomery, D. C. (2017). *Design and analysis of experiments* (9th ed.). John Wiley & Sons.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Shmueli, G. (2017). Analyzing behavioral big data: Methodological, practical, ethical, and moral issues. *Quality Engineering 29*(1), 57–74.

Siroker, D. and P. Koomen (2013). *A/B testing: The most powerful way to turn clicks into customers*. John Wiley & Sons.

Steiner, S. H. and R. J. MacKay (2005). *Statistical Engineering: an algorithm for reducing variation in manufacturing processes*. ASQ Quality Press.