STAT 430/830: Experimental Design

Nathaniel T. Stevens
nstevens@uwaterloo.ca
University of Waterloo, Waterloo, ON, Canada, N2L 3G1

TABLE OF CONTENTS

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 (Thomke (2020), Siroker and Koomen (2013), McFarland (2012)). Companies like Google, Amazon, Facebook, Netflix, Airbnb and Uber 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 students will be exposed to the value of 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 experimental conditions are compared, multivariate experiments such as factorial and fractional-factorial designs, and optimization techniques such as response surface methodology.

What this course does not emphasize is third party experimentation platforms such as Optimizely, Google Optimize, Mixpanel, Apptimize, Split or AB Tasty. While the physical construction of experimental conditions and the collection of data is a necessary part of experimentation, there is no standard platform used by all data scientists at all companies. Indeed, many companies have developed in-house solutions such as LinkedIn's XLNT platform, Netflix's ABlaze platform, Twitter's DDG platform, Microsoft's ExP, or Uber's XP, to name a few. For this reason it would be a poor use of time to train students 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, students 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, sample size calculations, and regression analyses necessary to draw conclusions and make impactful statements about the question of interest. Students will also become familiar with using `R` to automate components of both the design and the analysis of experiments.

1 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

For ease of discussion the following examples will serve as a source for context.

- **Example Experiment 1: List View vs. Tile View**

  Suppose that Nike, the athletic apparel company is experimenting with their mobile shopping interface and they are interested in determining whether changing the user interface from *list view* to *tile view* (see Figure 1) will increase the proportion of customers that proceed to checkout.

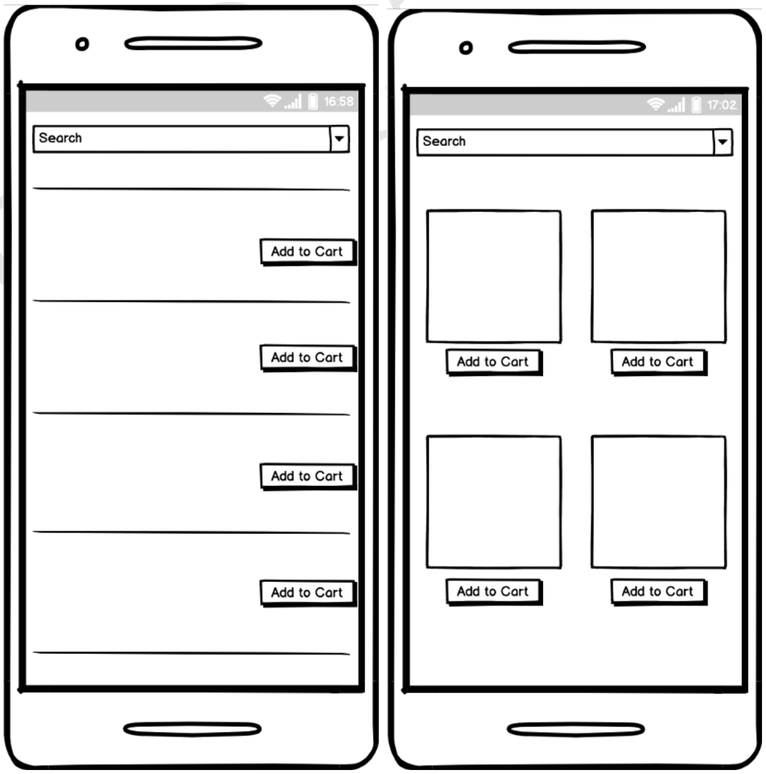

Figure 1: Nike Experiment. List View (left) vs. Tile View (right).

- **Example Experiment 2: Ad Themes**

  Suppose that Nixon, the watch and accessories brand, is experimenting with four different video ads that are to be shown on Instagram. The first has a surfing theme, the second has a rock climbing theme, the third has a camping theme, and the fourth has an urban professional theme. Interest lies in determining which of the four themes, on average, is watched the longest.

Figure 2: Nixon Experiment. Four different ad themes (from left to right: surfing, camping, rock climbing, urban professional).

In all data-driven 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**. In the Nike example above, the metric of interest is the percentage of customers that checked out, call this the checkout rate (COR). The corresponding response variable is the binary indicator which identifies whether or not users checked out. In the Nixon example, the metric of interest is the average viewing duration (AVD) of the ads and the corresponding response variable is the amount of time each user watches an ad. 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 examples above, the Nike app's layout and the Nixon ad themese are the factors that infuence COR and average AVD, respectively.

The different values that a factor takes on in an experiment are referred to as **levels**. In the Nike example {list view, tile view} are two different levels of the *layout* factor. In the Nixon example, {surfing, rock climbing, camping, professional} are four different levels of the *themes* factor. 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. In particular, Nike users are the experimental units in the first example and Instagram users are the experimental units in the second example. However, as will become clear in other examples and exercises throughout the course, an experimental unit need not be a human.

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 video ad's theme significantly influences its AVD, 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 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. Imagine one of the Nixon ads is unintentionally offensive and sours the public's attitude toward the brand. 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 ethical.

- **Unethical Experiment 1:** In evaluating whether smoking causes lung cancer, it would be unethical to have a *'smoking'* condition in which subjects are forced to smoke.

- **Unethical Experiment 2:** In dynamic pricing experiments, it would be unethical to show different users different prices for the same product.

- **Unethical Experiment 3:** In social contagion experiments, it would be unethical to show some social network users consistently negative content and others consistently positive content. But Facebook did this anyway (Shmueli, 2017).

- **Unethical Experiment 4:** Miroglio et al. (2018) describe an investigation that Mozilla conducted in which interest lied in determining whether Firefox users that installed an ad blocker were more engaged with the browser. However, it would have been unethical to force users to install an ad blocker and so they were forced to perform an observational study with propensity score matching instead.

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 data-driven investigations that formalizes the process by which data is collected to answer questions, and that is applicable 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, and it is used defined in terms of some metric of interest. For example, in the Nike experiment a question statement might look as follows: "*Relative to the list view layout, does a tile view layout increase checkout 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** and the metric of interest; it is through measurements of the response variable that the metric is calculated and the question is answered.

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. In the Nike and Nixon examples above, the factors previously identified were design factors. In both cases, possible nuisance factors include the country the experiment is taking place in (e.g., Canadian users vs. American users), the day of the week the experiment is run (e.g., weekday vs. weekend), etc. Examples of allowed-to-vary factors in this experiment inclue the unit's age, gender, device typpe, etc. It is important to note that what is considered a design factor vs. a nuisance factor vs. an allowed-to-vary factor depends entirely on the context; day-of-week could conceivably be a treated as a design factor *or* a nuisance factor *or* an allowed-to-vary factor, depending on the question being answered.

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 cluster-based 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.

Another important quality check is the *sample ratio mismatch test*. Experimental units are typically assigned randomly to the experimental conditions and this assignment is usually done in real-time. For instance, consider an Instagram user in the Nixon experiment: when they scroll to the Nixon ad a (proverbial) four-sided die is rolled to determine which theme will be shown. In the long run, if the assignment strategy is truly random then each experimental condition should contain roughly 25% of all of the units involved in the experiment. The sample ratio mismatch test is a formal hypothesis test to determine whether the observed sample ratios match what would be expected if assignment was truly done at random (Georgiev, 2019). This hypothesis test is discussed in more detail later in the course.

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 3.

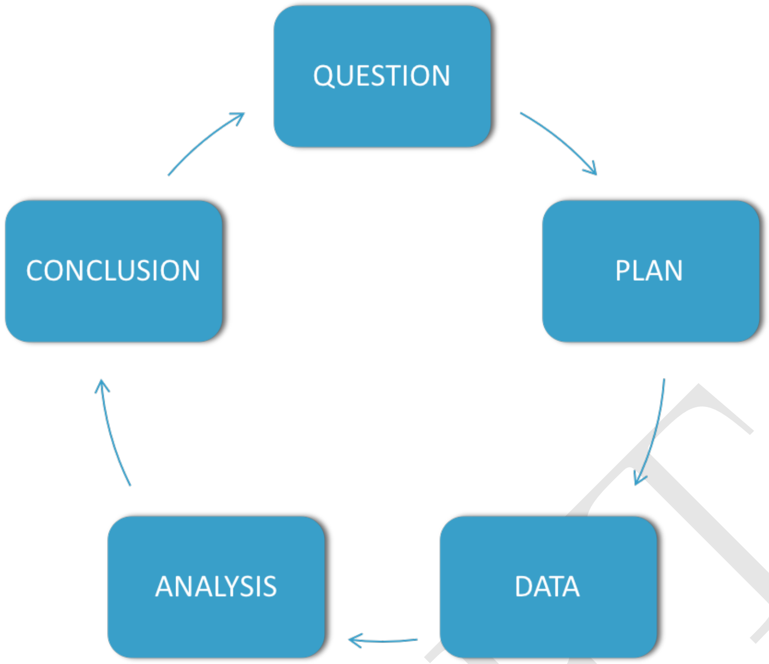

Figure 3: QPDAC Cycle

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, 2019). You will see that we have briefly mentioned each of 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 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 Nike experiment introduced previously. Suppose the CORs in the *list view* and *tile view* conditions were respectively 0.5 and 1. If these checkout 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 may be 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.

2 EXPERIMENTS WITH TWO CONDITIONS

We now consider the design and analysis of an experiment consisting of two experimental conditions – or what many data scientists broadly refer to as “A/B Testing”. Typically the goal of such an experiment is to decide which condition is optimal with respect to some metric of interest. For instance, the canonical A/B test is one in which two versions of a webpage are tested – one with a red button and the other with a blue button – and the ‘winning’ webpage is the one with the button that is clicked most frequently. Although this example is trivial, and it oversimplifies the difficulties, nuances, and importance of such an experiment, it serves as a tangible example of the question being answered: given two options, which one is best?

Figure 4: Canonical Button Colour Test

Formally, such a question is phrased as a statistical hypothesis that we test using the data collected from the experiment. In order to do so we must first define the two experimental conditions by selecting a single design factor and choosing two levels to experiment with. Once these choices are made, the experimental conditions are established and we must randomize n_1 experimental units to one condition and n_2 to the other condition. Next we measure the response variable (y) on each of these units and summarize these response measurements with some metric of interest, θ. Statistically speaking this metric might be a mean, a proportion, a variance, a percentile, or technically any statistic that can be calculated from sample data. Practically speaking such metrics might be things like average-time-on page, average-number-of-bookings, average-number-of-impressions, average-purchase-size, click-through-rate, bounce-rate, conversion-rate, retention-rate, etc. The exact metric chosen will depend on the question being answered and the type of data being collected.

Supposing the metric of interest has been chosen, interest lies in comparing this metric between the conditions and identifying the optimal condition as the one that optimizes (i.e., maximizes or minimizes) it. Because such a metric is calculated from sample data, which are drawn from a broader population, we view it as an estimate of the corresponding parameter in that population. For example, suppose in the red vs. blue button experiment the click-through-rates of the two conditions are 0.12 (red) and 0.03 (blue). These values are simply sample estimates of the true red vs. blue click-through-rates, which we denote by θ_1 and θ_2. Thus, $\hat{\theta}_1 = 0.12$ and $\hat{\theta}_2 = 0.03$. Although it is clear that $\hat{\theta}_1 > \hat{\theta}_2$ we must formally decide whether this sample data provides enough evidence to believe that regardless of the sample you might have drawn, the red button is superior to the blue one. In other words, that $\theta_1 > \theta_2$. As mentioned, such a statement is formally phrased as a statistical hypothesis of the form

$$H_0\colon\ \theta_1 \leq \theta_2 \text{ vs. } H_A\colon\ \theta_1 > \theta_2 \tag{1}$$

or

$$H_0\colon\ \theta_1 \geq \theta_2 \text{ vs. } H_A\colon\ \theta_1 < \theta_2 \tag{2}$$

Since it is the null hypothesis H_0 that is assumed to be true at baseline, which statement one wishes to test depends on this baseline assumption. However, notice that H_0 and H_A are complements of one another, and so only one of them is true. Furthermore, our decision is also binary: based on the observed data we choose to reject or not reject H_0. Thus, regardless of which direction you choose to state your hypothesis, the conclusion you draw will be the same. To make this clear, suppose that in the red vs. blue button experiment θ_1 and θ_2 respectively represent the click-through-rates for the red and blue buttons. If the data suggest red is better, it doesn't matter which hypothesis statement we test. If we test (1) we will reject H_0 (and conclude that red is best), and if we test (2) we will not reject H_0 (and hence conclude that red is best).

Note that when the data suggest that $\theta_1 = \theta_2$ we may fail to reject both of these null hypotheses, which appears to be a contradiction. This, however, is not a contradiction: $\theta_1 = \theta_2$ is included in both null hypotheses. One could rule out this scenario by first conducting the following the two-sided hypothesis

$$H_0\colon\ \theta_1 = \theta_2 \text{ vs. } H_A\colon\ \theta_1 \neq \theta_2. \tag{3}$$

This hypothesis provides no information about which of the two conditions is best, but does tell us whether they are different. As such, it may be used as an initial check of whether the conditions are different at all. If they aren't, then there is no reason to proceed. But if they are, then we would use hypothesis (1) or (2) to help determine the optimal condition. For a general review of statistical inference and hypothesis testing, please refer to Appendix A.2.

In the context of hypotheses such as (1), (2) and (3), we discuss in this chapter how to design an experiment to test them and we discuss how to analyze observed data to formally draw conclusions about them. In particular we discuss how to choose the number of units to assign to each condition, and we describe a variety of analysis techniques appropriate for different metrics of interest, and different types of response variables.

2.1 Comparing Means in Two Conditions

In this section we restrict attention to the situation in which the response variable of interest is measured on a continuous scale, although the associated methodology is also commonly applied when response variables are discrete and, for example, represent counts (as in the number of times an event of interest occurs). In these cases we assume that the response observations collected in the two conditions follow normal distributions, and in particular

$$Y_{i1} \sim \mathrm{N}(\mu_1, \sigma^2) \text{ and } Y_{i2} \sim \mathrm{N}(\mu_2, \sigma^2)$$

where $i = 1, 2, \ldots, n_j$ for $j = 1, 2$. Thus Y_{ij} represents the response observation for the i^{th} unit in the j^{th} condition, and we assume that the measurements in the two conditions could reasonably have been drawn from a normal distribution with mean μ_1 (in the first condition) or μ_2 (in the second) and common variance σ^2. Thus we believe that the distributions from which these samples were drawn only differ (if they differ at all) with respect to the mean, and in no other way. Thus a comparison of the two conditions corresponds to a comparison of the expected responses (i.e., the means) in each of them. Specifically we test hypotheses of the form

$$H_0\colon\ \mu_1 = \mu_2 \text{ vs. } H_A\colon\ \mu_1 \neq \mu_2 \tag{4}$$

$$H_0\colon\ \mu_1 \leq \mu_2 \text{ vs. } H_A\colon\ \mu_1 > \mu_2 \tag{5}$$

$$H_0\colon\ \mu_1 \geq \mu_2 \text{ vs. } H_A\colon\ \mu_1 < \mu_2 \tag{6}$$

In the following subsections we describe how to analyze data of this form and draw conclusions about such hypotheses and we also describe how to choose the sample size that allows one to be sufficently confident in their conclusions.

2.1.1 The Two-Sample t-Test

In order to test hypotheses (4), (5) and (6) we must first calculate a **test statistic**. Because $Y_{ij} \sim \mathrm{N}(\mu_j, \sigma^2)$, it is also true that $\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \sim \mathrm{N}(\mu_j, \frac{\sigma^2}{n_j})$ and hence that

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathrm{N}(0, 1). \tag{7}$$

Although we can substitute a hypothesized value for $\mu_1 - \mu_2$ into this expression, we do not have a hypothesized value for σ. As such, we replace it in the equation above using the following estimate

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(Y_{i1} - \overline{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{i2} - \overline{Y}_2)^2}{n_1 + n_2 - 2}.$$

Note that this quantity is simply a pooled estimate of σ^2 based on the sample variances in the two conditions.

Substituting $\hat{\sigma}$ for σ gives

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \tag{8}$$

which is the test statistic for these hypothesis tests and where $t_{(n_1+n_2-2)}$ is the **null distribution**. It is for this reason that the test is called a "t-test".

Hypotheses (4), (5) and (6) are formally tested by calculating the observed value of T from our sample data $\{y_{11}, y_{21}, \ldots, y_{n_1 1}\}$ and $\{y_{12}, y_{22}, \ldots, y_{n_2 2}\}$ and evaluating its extremity in the context of the $t_{(n_1+n_2-2)}$ distribution. Given the sample data, we have $\overline{y}_1 = \hat{\mu}_1$ and $\overline{y}_2 = \hat{\mu}_2$ and so the observed test statistic is given by

$$\begin{aligned} t &= \frac{(\overline{y}_1 - \overline{y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(\hat{\mu}_1 - \hat{\mu}_2) - 0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \end{aligned} \tag{9}$$

Notice that we have also substituted the hypothesized value $\mu_1 - \mu_2 = 0$ indicating a null assumption of 'no difference' between the two conditions.

To decide whether to reject H_0 or not reject H_0 we must calculate the **p-value** – the probability of observing a value of the test statistic at least as extreme as the one we observed, if H_0 really were true. In the case of hypothesis (4) the p-value is calculated as p-value $= 2\text{Pr}(T \geq |t|)$; in the case of hypothesis (5) the p-value is calculated as p-value $= \text{Pr}(T \geq t)$, and in the case of hypothesis (6) the p-value is calculated as p-value $= \text{Pr}(T \leq t)$, where here $T \sim t_{(n_1+n_2-2)}$. See Figure A.9 for a visual depiction of these calculations. We then decide to reject (or not reject) H_0 on the basis of a comparison between the calculated p-value and the **significance level** α. If p-value $\leq \alpha$ we reject H_0 in favor of H_A, and if p-value $> \alpha$ we do not reject H_0.

2.1.2 A Confidence Interval for the Difference in Means

The t-test statistic T may also be used as a *pivotal quantity* in the construction of a confidence interval. A consequence of the distributional result in equation (8) is that

$$\text{Pr}\left(-t_{(n_1+n_2-2,\alpha/2)} \leq \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{(n_1+n_2-2,\alpha/2)}\right) = 1 - \alpha$$

where $t_{(n_1+n_2-2,\alpha/2)}$ is the quantile of the $t_{(n_1+n_2-2)}$ distribution with upper tail probability $\alpha/2$. This equation can then be rearranged, isolating for $\mu_1 - \mu_2$ in the middle, to provide bounds which contain the true difference $\mu_1 - \mu_2$ with confidence $1 - \alpha$:

$$\text{Pr}\left((\overline{Y}_1 - \overline{Y}_2) - t_{(n_1+n_2-2,\alpha/2)}\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{Y}_1 - \overline{Y}_2) + t_{(n_1+n_2-2,\alpha/2)}\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$$

The endpoints in the probability statement above are random variables. Substituting the *random* quantities by their *observed* sample counterparts yields the following $(1 - \alpha) \times 100\%$ confidence interval for $\mu_1 - \mu_2$:

$$(\overline{y}_1 - \overline{y}_2) \pm t_{(n_1+n_2-2,\alpha/2)}\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

2.1.3 Example: Instagram Ad Frequency

Suppose that you are a data scientist at Instagram, and you are interested in running an experiment to learn about how user engagement is influenced by ad frequency. Currently users see an ad every 8 posts in their social feed, but, in order to increase ad revenue, your manager is pressuring your team to show an ad every 5 posts, under the assumption that users will not behave any differently under this new regime. You are justifiably nervous about this change and you worry that this will substantially decrease user engagement and hurt the overall user experience. As such you propose an experiment to test this new regime before rolling it out to all users. The experiment you propose is an A/B test in which average session time (i.e., the length of time a user engages with the app – in minutes) is compared between the two ad frequency conditions. You believe that the current ad frequency (condition 1) will correspond to a significantly longer average session time than the proposed ad frequency (condition 2).

Thus, in the language and notation of these notes, you're interested in testing a hypothesis such as (5) where μ_1 represents the average session time of a user in the 7:1 ad frequency condition and μ_2 represents the average session time of a user in the 4:1 ad frequency condition. The null hypothesis here assumes what your manager assumes – that increased ad frequency does not lead to reduced engagement (H_0: $\mu_1 \leq \mu_2$). Thus you expect to collect data that contradicts this statement so that it can be rejected in favor of the alternative that says that increased ad frequency significantly reduces the amount of time users are engaged with the app (H_A: $\mu_1 > \mu_2$).

In order to test this hypothesis you randomoize $n_1 = 500$ users to the 7:1 ad frequency condition and $n_2 = 500$ users to the 4:1 condition. The data you collect is summarized as follows: The average session time in the 4:1 condition is $\hat{\mu}_1 = \overline{y}_1 = 4.9162$ with a standard deviation of $s_1 = 0.9634$, and in the 7:1 condition the average session time is $\hat{\mu}_2 = \overline{y}_2 = 3.0518$ with a standard deviation of $s_2 = 0.9950$. The pooled standard deviation estimate is

$$\hat{\sigma} = \sqrt{\frac{499 \cdot 0.9634^2 + 499 \cdot 0.995^2}{998}} = 0.9793.$$

These summaries support your suspicion: session time appears to be negatively effected by an increased ad frequency.

To determine whether this difference is statistically significant, you formally test the hypothesis by calculating a p-value. To do this, you must first calculate the observed test statistic. Substituting these summaries into equation (9) gives

$$t = \frac{4.9162 - 3.0518}{0.9793\sqrt{\frac{2}{500}}} = 30.1013.$$

The p-value associated with this test is $\Pr(T \geq 30.1013)$ where $T \sim t_{998}$. When calculated this probability is equal to 1.84×10^{-142}, which is essentially 0. In `R` this probability is calculated using the command `pt(30.1013, df = 998, lower.tail = F)`. We can also use the `t.test()` function in `R` to do the whole test; you need only pass it the data and a few other arguments and it will calculate the necessary summaries, the test statistic and the p-value. Note that to replicate the results here we must set the logical argument `var.equal` to `TRUE`. We discuss an alternative approach to take when the variances are not assumed to be equal in Section 2.1.5.

In order to draw a conclusion, we must compare our calculated p-value to the significance level $\alpha = 0.05$. Since $1.84 \times 10^{-142} < 0.05$ we reject the null hypothesis in favor of the alternative. In the context of the experiment, this means that increased ad frequency significantly reduces the amount of time users engage with the app. We may also summarize the results fo this experiment with a point and interval estimate of the difference in engagement time. In particular, $\overline{y}_1 - \overline{y}_2 = 4.9162 - 3.0518 = 1.8644$ and the corresponding 95% confidence interval is $1.8644 \pm 1.9623 \times 0.9793\sqrt{\frac{2}{500}} = [1.7429, 1.9859]$. In other words, we can expect a 1 minute and 52 second reduction in average session time when you move from a 4:1 ad frequency to a 7:1 frequency. The lower and upper 95% confidence bounds on this estimate are respectively 1 minute 45 seconds and 1 minute 59 seconds.

As an alternative way of looking at the problem, depending on the speed a user scrolls through their feed, this increased ad frequency could actually *reduce* ad revenue rather than increase, as is the intention. Suppose that the typical user spends roughly 5 seconds looking at each post, which means they scroll through 12 posts per minute. In the 7:1 ad frequency condition a user would then see 1.5 ads per minute, and in the 4:1 frequency condition a user would see 2.4 ads per minute. Although a user in the 7:1 condition sees fewer ads per minute, they spend more time on the app. At an average session time of roughly 5 minutes, they see 7.5 ads per session, whereas a user in the 4:1 condition, whose session duration is roughly 3 minutes, will see 7.2 ads per minute. As such, it would be ill-advised to adopt this new ad regime from both the perspective of user engagement and ad revenue.

2.1.4 Power Analysis and Sample Size Calculations

When designing a two-condition experiment, the most important question (once the response variable and conditions have been chosen) is "*How many units do I need in each condition?*". The answer to this question is determined by the frequency with which we are comfortable drawing the wrong conclusion.

Recall that because H_0 and H_A are complements of one another, exactly one of them is correct. Thus, when we choose to reject or not reject H_0 we risk drawing the wrong conclusion. In this context we can make two types of errors:

- Type I Error: Reject H_0 when it is in fact true

- Type II Error: Do not reject H_0 when it is in fact false

Ideally these types of errors would happen very infrequently. Fortunately we are able to control the frequency with which such errors are made through the **significance level** and the **power** of the hypothesis test. The significance level is denoted by α where

$$\alpha = \Pr(\text{Type I Error}) = \Pr(\text{Reject } H_0 | H_0 \text{ is true})$$

and the power of the test is denoted by $1 - \beta$ where

$$\beta = \Pr(\text{Type II Error}) = \Pr(\text{Do not reject } H_0 | H_0 \text{ is false}).$$

Thus, a test that has a small significance level and large power is desirable as it simultaneously minimizes the chances of committing both Type I and Type II Errors.

In practice these values are chosen to be consistent with one's risk tolerance, though $\alpha = 0.05$ and $\beta = 0.2$ are the standard choices. As we will show here, the significance level and power of a hypothesis test are related to one another and also related to the sample size. In fact, for a given sample size, as the chances of a Type I Error decrease, the chances of a Type II Error increase, and vice versa. However, if we want to fix α and β at particular values, we can determine what sample size is necessary to do so. Thus, it is important to understand the nature of the relationship between these quantities – if you alter one of them, the others will also change.

In what follows, we will derive the formula that quantifies this relationship, and which can be used to determine the sample size necessary to keep Type I and Type II Errors at bay. We do so assuming that the parameter σ is known, or at least that we have a reasonable guess as to what it might be. Note that we do not need to make this assumption once the data are collected, but we do need to make it prior to data collection. Thus for the development below we define our test statistic T as in equation (7) which means that we will be working with the $\text{N}(0, 1)$ distribution.

We begin by precisely defining what it means (in terms of the test statistic) to reject H_0. In all cases this happens when p-value $\leq \alpha$. In the context of a two-sided hypothesis such as (4) this happens when $t \geq z_{\alpha/2}$ or $t \leq -z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)^{th}$ quantile of the standard normal distribution. Thus we can define a **rejection region** $R = \{t \mid t \geq z_{\alpha/2} \text{ or } t \leq -z_{\alpha/2}\}$ that describes all values of t for which H_0 would be rejected. Similar rejection regions can be defined for hypotheses (5) and (6) as well. These are respectively given by $R = \{t \mid t \geq z_{\alpha/2}\}$ and $R = \{t \mid t \leq -z_{\alpha/2}\}$. All of these rejection regions are depicted in blue in Figure 5.

Having defined these we now derive the formula which, for a given significance level and power, prescribes how many units should be assigned to each condition. Although it is very common to assign the same number of units to each of the conditions (i.e., $n_1 = n_2$), we will keep this derivation general and not make this specific requirement. What we do require, however, is an assumption about the relative sizes of n_1 and n_2. Specifically, we need to specify k where $n_1 = kn_2$. In the case that equal sample sizes are desired we would simply take $k = 1$. Furthermore, we provide this derivation under the assumption that we are dealing with a hypothesis that has a two-sided alternative as in (4). We indicate where and how this derivation changes

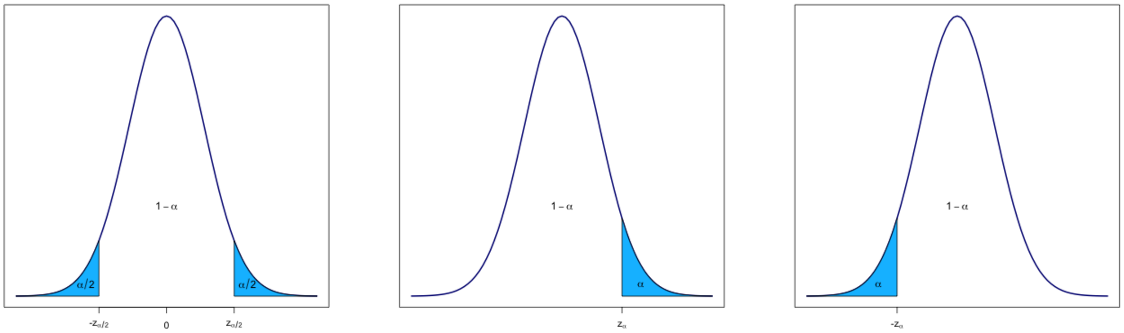

Figure 5: Rejection regions corresponding to one and two-sided hypotheses

if sample size calculations in the context of a one-sided hypothesis test is of interest.

We begin by considering the power of the hypothesis test:

$$\begin{aligned}
1-\beta &= \Pr(\text{Reject } H_0 \mid H_0 \text{ is false}) \\
&= \Pr(T \in R \mid H_0 \text{ is false}) \text{ where } R \text{ is the rejection region} \\
&= \Pr(T \geq z_{\alpha/2} \text{ or } T \leq -z_{\alpha/2} \mid H_0 \text{ is false}) \\
&= \Pr(T \geq z_{\alpha/2} \mid H_0 \text{ is false}) + \Pr(T \leq -z_{\alpha/2} \mid H_0 \text{ is false}) \\
&= \Pr\left(\frac{(\overline{Y}_1 - \overline{Y}_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq z_{\alpha/2} \mid H_0 \text{ is false}\right) + \Pr\left(\frac{(\overline{Y}_1 - \overline{Y}_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq -z_{\alpha/2} \mid H_0 \text{ is false}\right)
\end{aligned}$$

If $H_0 : \mu_1 = \mu_2$ were true, and hence $\mu_1 - \mu_2 = 0$ were true, then the ratios in the preceding line would follow a $\mathrm{N}(0,1)$ distribution. However, we know that H_0 is false which means that $\mu_1 - \mu_2 = \delta$ for some non-zero δ, and so it is

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - \delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

that follows a $\mathrm{N}(0,1)$ distribution. Let us make this substitution, being sure to replicate what is done on the left side of inequalities on the right. Also note that we no longer need to write "$|H_0$ is false" since we are now exploiting this fact.

$$\begin{aligned}
1-\beta &= \Pr\left(\frac{(\overline{Y}_1 - \overline{Y}_2) - \delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) + \Pr\left(\frac{(\overline{Y}_1 - \overline{Y}_2) - \delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq -z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) \\
&= \Pr\left(Z \geq z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) + \Pr\left(Z \leq -z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) \text{ where } Z \sim \mathrm{N}(0,1)
\end{aligned}$$

Note that depending on the sign of δ, just one of these terms will dominate. To see this, suppose $\delta > 0$; then $-z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ will be an extremely negative number and the probability that a standard normal random variable is smaller than an extremely negative number is effectively 0, and only the first term remains. Now suppose $\delta < 0$; then $z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ will be an extremely positive number and the probability that a standard normal random variable is larger than an extremely positive number is effectively 0, and only the second term remains. Assume, without loss of generality, that $\delta > 0$ in which case

$$1 - \beta \quad = \quad \Pr\left(Z \geq z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$$

Because this probability is equal to $1-\beta$ we know that $z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ must be equal to $z_{1-\beta}$, the β^{th} quantile of the standard normal distribution. Thus

$$z_{1-\beta} = z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

and we can rearrange this equation solving for the sample size. But first we must substitute $n_1 = kn_2$ so that there is just a single sample size to solve for:

$$z_{1-\beta} = z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{kn_2}+\frac{1}{n_2}}} = z_{\alpha/2} - \frac{\sqrt{n_2}\delta}{\sigma\sqrt{\frac{1}{k}+1}}$$

Solving for n_2 yields:

$$n_2 = \frac{(\frac{1}{k}+1)(z_{\alpha/2} - z_{1-\beta})^2\sigma^2}{\delta^2} \tag{10}$$

and then n_1 is found by computing kn_2. When equal sample sizes are desired ($k = 1$) each condition receives n units where

$$n = \frac{2(z_{\alpha/2} - z_{1-\beta})^2\sigma^2}{\delta^2}. \tag{11}$$

So when calculating a sample size we need to have chosen α and β (our Type I and Type II Error rates), we need a guess as to what σ is, and we need a value for δ. With all of this information one can readily use the formulae above to calculate n_1 and n_2.

But where does the δ value come from? We define δ to be the **minimum detectable effect** (MDE) of the test. The MDE for hypothesis tests like (4), (5) or (6) refers to the minimal difference between conditions (i.e., between μ_1 and μ_2) that we find to be practically relevant and that we would like to detect as being statistically significant. For instance, imagine we are comparing the average length of time users spend engaging with their Instagram apps as in Section 2.1.3. Suppose that condition 1 (7:1 ad frequency) corresponds to the current version of the app, and you know users engage with the app for an average of 5 minutes. Now suppose that condition 2 corresponds to the 4:1 ad frequency. Would it be practically relevant if users in condition 2 spend an average of 4.8 minutes engaged with the app? If not, would it be practically important if these users spent an average of 3.5 minutes engaged with the app? The answer to the question "*What is the minimal difference between μ_1 and μ_2 that is practically important?*" is the MDE, and is what is captured by $\delta = \mu_1 - \mu_2$.

Sometimes the MDE is defined on a standardized scale, and communicated in numbers of standard deviations as opposed to the absolute scale described above. In this case δ is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

and the sample size formula (10) simplifies to

$$n_2 = \frac{(\frac{1}{k} + 1)(z_{\alpha/2} - z_{1-\beta})^2}{\delta^2}$$

and the sample size formula (11) simplifies to

$$n = \frac{2(z_{\alpha/2} - z_{1-\beta})^2}{\delta^2}.$$

The advantage of defining the MDE on a standardized scale is that we do not require knowing or guessing σ in our sample size calculations.

It is important to also consider how these formulae change if we were performing sample size calculations for one-sided hypothesis tests. In these cases the rejection regions are also one-sided and based on the quantile z_α instead of $z_{\alpha/2}$. It turns out that this is the only difference, and when carried through the derivation yields sample size formulae equivalent to (10) and (11) but with $z_{\alpha/2}$ replaced by z_α.

As should be clear by looking at equations (10) and (11), there is an interdependent relationship between sample size, significance level, power, and MDE. These equations can be rearranged to isolate for any of these variables, which illustrates the fact that changing one variable leads to a change in all of the others. For an interactive demonstration of these interdependencies feel free to tinker with the sample size calculator found at the following link: https://nathaniel-t-stevens.shinyapps.io/SampleSizeCalculator/.

2.1.5 When Assumptions are Invalid

When testing hypotheses of the form (4), (5) and (6) using the two-sample t-test described in Section 2.1.1, we make two key assumptions. First, we assume that the variance in the two conditions are equal, and second, we assume that the response observations in each condition follow a normal distribution. In this subsection we describe alternative approaches when these assumptions are not valid. We begin with the equal variance assumption.

Welch's *t*-Test: When it is unreasonable to assume that the response variable measurements in each condition have equal variances, an approach that accommodates $Y_{ij} \sim \text{N}(\mu_j, \sigma_j^2)$ for $j = 1, 2$, and hence $\sigma_1^2 \neq \sigma_2^2$, is to be preferred. In this situation we may use the test statistic

$$t = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

where $\hat{\sigma}_j^2$ is the sample variance of the response measurements in condition $j = 1, 2$. However, this statistic does not follow a t-distribution exactly; it *approximately* follows a t-distribution with

$$\nu = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{(\hat{\sigma}_1^2/n_1)^2}{n_1-1} + \frac{(\hat{\sigma}_2^2/n_2)^2}{n_2-1}}$$

degrees of freedom. Carrying out the test using a $t_{(\nu)}$ null distribution (with ν as above) is referred to as *Welch's t-test* after Bernard L. Welch who devised this approximation (Welch, 1947). This test can be carried out in `R` using the `t.test()` function but with the logical argument `var.equal` set to `FALSE`.

In order to decide whether $\sigma_1^2 \neq \sigma_2^2$ and hence whether Welch's t-test is necessary, one might consider formally testing the hypothesis

$$H_0\colon\ \sigma_1^2 = \sigma_2^2 \text{ vs. } H_A\colon\ \sigma_1^2 \neq \sigma_2^2 \tag{12}$$

Such a hypothesis is commonly tested using an ***F*-test of equal variances**. If H_0 is true, and $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the F-test assumes that $Y_{ij} \sim \mathrm{N}(\mu_j, \sigma^2)$ which consequently means that

$$\frac{(n_j - 1)\hat{\sigma}_j^2}{\sigma^2} \sim \chi^2_{n_j-1}$$

and hence that

$$T = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{(n_1-1,n_2-1)}.$$

The observed value of the test statistic is

$$t = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

which is then compared to the null distribution $F_{(n_1-1,n_2-1)}$. Note that it is because the null distribution is an F-distribution that this test is known as an F-test.

Because the F-distribution is not symmetrical and not defined for negative values, in the context of the two-sided hypothesis above the p-value is calculated as

$$\text{p-value} = \Pr(T \geq t) + \Pr(T \leq 1/t)$$

since values greater than or equal to t and less than or equal to $1/t$ are what are considered "at least as extreme" in this situation. One-sided alternatives might also be considered where $H_A\colon\ \sigma_1^2 > \sigma_2^2$ or $H_A\colon\ \sigma_1^2 < \sigma_2^2$ in which case the p-values are respectively defined as p-value $= \Pr(T \geq t)$ and p-value $= \Pr(T \leq t)$, but these tests aren't relevant in the discussion on t-tests. Note that the F-test of equal variance can be carried out in `R` using the `var.test()` function.

Note that the null hypothesis in the statement above assumes $\sigma_1^2/\sigma_2^2 = 1$. In principle we could consider a more general hypothesis in which $H_0 : \sigma_1^2/\sigma_2^2 = \omega$. In this case we would exploit the fact that

$$\frac{(n_j - 1)\hat{\sigma}_j^2}{\sigma_j^2} \sim \chi^2_{n_j-1}$$

yielding the generalized test statistic

$$T = \frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2} \sim F_{(n_1-1,n_2-1)}$$

which clearly reduces to the one we used previously, if $\omega = 1$.

Permutation and Randomization Tests: All of the previous tests assume the response measurements are normally distributed. However, many situations exist in which a numeric response variable does not follow a normal distribution. Using the observed data, this assumption can be informally evaluated using QQ-plots or histograms, or formally using the Shapiro-Wilk test (Shapiro and Wilk, 1965). While both the Student's t-test and Welch's t-test are fairly robust to non-normality, it would be preferable to have a test that does not rely on this assumption. *Permutation* and *randomization tests* are nonparametric resampling approaches that may be used for this purpose in this context.

Suppose you collect response measurements $\{y_{11}, y_{21}, \ldots, y_{n_1 1}\}$ and $\{y_{12}, y_{22}, \ldots, y_{n_2 2}\}$ in conditions 1 and 2, respectively. Using these measurements you then estimate some metric of interest θ in the two conditions yielding $\hat{\theta}_1$ and $\hat{\theta}_2$. The goal, then, is to compare $\hat{\theta}_1$ to $\hat{\theta}_2$ in accordance with hypotheses such as (1), (2) or (3) to decide whether $\theta_1 = \theta_2$, $\theta_1 > \theta_2$ or $\theta_1 < \theta_2$. Note that despite where this discussion appears in these notes, permutation and randomization tests are actually much more widely applicable than the comparison of means; the can be used to compare *any* two metrics of interest.

If H_0 is true and there is truly no difference between the conditions, then the samples $\{y_{11}, y_{21}, \ldots, y_{n_1 1}\}$ and $\{y_{12}, y_{22}, \ldots, y_{n_2 2}\}$ should be very similar and permuting the labels 'condition 1' and 'condition 2' associated with each response measurement should not substantially change $\hat{\theta}_1$ or $\hat{\theta}_2$. In fact, if the null hypothesis is true, each of the

$$\binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2}$$

arrangements of the observed data are equally likely. A true **permutation test** calculates the test statistic t on the originally observed data and on *each* of the $\binom{n_1+n_2}{n_1} = \binom{n_1+n_2}{n_2}$ arrangements of data and then uses this distribution of values as the null distribution. A formal conclusion about H_0 is drawn on the basis of the extremity of t in the context of this null distribution. The p-value associated with such a test is calculated empirically as the proportion of resampled test statistics that were "at least as extreme" as t. Note that in the context of such tests t is typically defined simply and intuitively as a difference $t = \hat{\theta}_1 - \hat{\theta}_2$ or a ratio $t = \hat{\theta}_1 / \hat{\theta}_2$ comparing the relevant metrics of interest. For our discussion here we will use $t = \hat{\theta}_1 - \hat{\theta}_2$.

While conceptually appealing, the permutation test is not practical in most circumstances because the number of permutations of the data becomes enormous, even for relatively small sample sizes. For instance, if $n_1 = n_2 = 50$, there are $\binom{100}{50} = 1.09 \times 10^{29}$ distinct arrangements of the data. Thus, since true permutation tests tend to be computationally expensive, a practical approximation is the **randomization test** which simply investigates a large number of resamples, as opposed to all possible permutations. An algorithm for performing a randomization tests is as follows:

1. Calculate the test statistic $t = \hat{\theta}_1 - \hat{\theta}_2$ on the original sample.

2. Resample the data without replacement so that n_1 observations are randomly associated with a resam-

pled ‘condition 1’: $\{y^*_{11}, y^*_{21}, \ldots, y^*_{n_1 1}\}$ and n_2 observations are randomly associated with a resampled ‘condition 2’: $\{y^*_{12}, y^*_{22}, \ldots, y^*_{n_2 2}\}$.

3. Calculate the value of the test statistic, labeled t^*, on this resampled data.

4. Repeat steps 2 and 3 N times ($N = 1000$ or 2000 are common choices).

5. Compare t to the null distribution which is derived from the N resampled values of t^*, and calculate the p-value.

The p-values associated with tests of this sort are calculated differently depending on whether the alternative hypothesis, H_A, is one- or two-sided. These calculations are summarized below:

- H_A: $\theta_1 \neq \theta_2$: p-value = The proportion of resampled test statistics $t^* \geq |t|$ or $\leq -|t|$

- H_A: $\theta_1 > \theta_2$: p-value = The proportion of resampled test statistics $t^* > t$

- H_A: $\theta_1 < \theta_2$: p-value = The proportion of resampled test statistics $t^* < t$

See Edgington and Onghena (2007) for a more thorough and general treatment of randomization tests.

2.2 Comparing Proportions in Two Conditions

Very often the response variable in an A/B test is binary, indicating whether an experimental unit did, or did not, perform some action of interest. In cases like these we let

$$Y_{ij} = \begin{cases} 1 & \text{if unit } i \text{ in condition } j \text{ performs the action of interest} \\ 0 & \text{if unit } i \text{ in condition } j \text{ does not perform the action of interest} \end{cases}$$

for $i = 1, 2, \ldots, n_j$, $j = 1, 2$. Examples of “actions of interest” include opening an email, clicking a button, watching an ad, leaving a webpage without interacting with it, etc. In each case unit i’s response variable is recorded as a 1 if they perform the action and a 0 otherwise. Interest then lies in deciding which condition is optimal, where the optimal condition is the one for which the likelihood that a unit performs the action is highest (when maximization is of interest) or smallest (when minimization is of interest).

To formally decide which condition is optimal we must make an assumption about the distribution of the response variable. Because the Y_{ij}’s are binary, it is common to assume that they follow a Bernoulli distribution:

$$Y_{ij} \sim \text{BIN}(1, \pi_j)$$

where π_j represents the probability that $Y_{ij} = 1$, i.e., the probability that a unit in condition j performs the action of interest. The goal of the experiment then, is to determine whether $\pi_1 = \pi_2$, $\pi_1 > \pi_2$ or $\pi_1 < \pi_2$. This decision is formally made in association with the following hypotheses:

$$H_0\colon \pi_1 = \pi_2 \text{ vs. } H_A\colon \pi_1 \neq \pi_2 \tag{13}$$

$$H_0\colon\ \pi_1 \leq \pi_2 \text{ vs. } H_A\colon\ \pi_1 > \pi_2 \tag{14}$$

$$H_0\colon\ \pi_1 \geq \pi_2 \text{ vs. } H_A\colon\ \pi_1 < \pi_2 \tag{15}$$

In the subsections that follow we describe how to analyze data of this form and draw conclusions about hypotheses like these. We also describe power analyses and sample size calculations in this context as well.

2.2.1 The Z-test for Proportions

In order to test hypotheses (13), (14) and (15) we must calculate a test statistic. Due to the **Central Limit Theorem**[1] we know that for large enough n_j the random variable $\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \overset{\cdot}{\sim} \text{N}(\pi_j, \frac{\pi_j(1-\pi_j)}{n_j})$. Thus, with a large amount of replication $\overline{Y}_{ij}$ will approximately follow a normal distribution. Based on this result

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \overset{\cdot}{\sim} \text{N}(0,1). \tag{16}$$

As a general rule of thumb, this approximation may be very poor unless $n_j\pi_j \geq 10$ and $n_j(1-\pi_j) \geq 10$ for both $j = 1, 2$.

Although we can substitute a hypothesized value for $\pi_1 - \pi_2$ (i.e., zero) into the equation above, we have no hypothesized value for π_1 or π_2 individually, and so this equation is not calculable in practice. As such we replace instances of π_1 and π_2 in the denominator with estimates. This is akin to replacing σ by $\hat{\sigma}$ in equations (7) and (8). Because $\pi_1 = \pi_2 = \pi$ under the null hypothesis, we use the pooled estimate given by

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2} \tag{17}$$

where $\hat{\pi}_1$ and $\hat{\pi}_2$ are respectively equal to $\overline{Y}_1$ and $\overline{Y}_2$. We note that when the response variable is binary, means equate to proportions and so hypothesis tests in this setting amount to a comparison of proportions.

Making these substitutions gives

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{18}$$

which also approximately follows a $\text{N}(0,1)$ distribution. Thus T is the test statistic associated with hypotheses (13), (14) and (15) where $\text{N}(0,1)$ is the null distribution. It is for this reason that the test is called a "Z-test".

To formally test these hypotheses we calculate the observed value of the test statistic, t, from our sample data $\{y_{11}, y_{21}, \ldots, y_{n_1 1}\}$ and $\{y_{12}, y_{22}, \ldots, y_{n_2 2}\}$ and evaluate its extremity in the context of the $\text{N}(0,1)$ distribution. Given the sample data, we have $\overline{y}_1 = \hat{\pi}_1$ and $\overline{y}_2 = \hat{\pi}_2$ and so the observed test statistic is given

[1] The Central Limit Theorem states that for any sequence of random variables $X_1, X_2, \ldots, X_n$ with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$ for each $i = 1, 2, \ldots, n$, the random variable $\overline{X}$ follows a $\text{N}(\mu, \sigma^2)$ distribution for large enough n (i.e., as $n \to \infty$).

by

$$\begin{aligned} t &= \frac{(\overline{y}_1 - \overline{y}_2) - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \\ &= \frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \\ &= \frac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \end{aligned} \tag{19}$$

Notice that we have also substituted the hypothesized value $\pi_1 - \pi_2 = 0$ indicating a null assumption of 'no difference' between the two conditions.

As is typical, we decide whether to reject or not reject H_0 based on the size of the test's p-value in relation to the significance level α. If p-value $\leq \alpha$ we reject H_0 in favor of H_A, and if p-value $> \alpha$ we do not reject H_0. The p-values associated with hypotheses (13), (14) and (15) are respectively calculated as $2\Pr(T \geq |t|)$, $\Pr(T \geq t)$, and $\Pr(T \leq t)$ where in each case $T \sim \mathrm{N}(0,1)$.

2.2.2 A Confidence Interval for the Difference in Proportions

The Z-test statistic T may also be used as a *pivotal quantity* in the construction of a confidence interval. A consequence of the distributional result in equation (18) is that

$$\Pr\left(-z_{\alpha/2} \leq \frac{(\overline{Y}_1 - \overline{Y}_2) - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

where $z_{\alpha/2}$ is the quantile of the $\mathrm{N}(0,1)$ distribution with upper tail probability $\alpha/2$. This equation can then be rearranged, isolating for $\pi_1 - \pi_2$ in the middle, to provide bounds which contain the true difference $\pi_1 - \pi_2$ with confidence $1 - \alpha$:

$$\Pr\left((\overline{Y}_1 - \overline{Y}_2) - z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \leq \pi_1 - \pi_2 \leq (\overline{Y}_1 - \overline{Y}_2) + z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right) = 1-\alpha$$

The endpoints in the probability statement above are random variables. Substituting the *random* quantities by their *observed* sample counterparts yields the following $(1-\alpha) \times 100\%$ confidence interval for $\pi_1 - \pi_2$:

$$(\overline{y}_1 - \overline{y}_2) \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

2.2.3 Example: Optimizing Optimizely

Siroker and Koomen (2013) discuss an A/B test they ran on the Optimizely website. In particular they were in the midst of a complete website redesign, and they were interested in how new versions of certain pages influenced things like conversion and engagement relative to the old version. One such metric they were interested in was whether or not the redesigned homepage lead to a significant increase in the number of new accounts created.

Thus, in the language and notation of these notes, they were interested in testing a hypothesis such as (15) where π_1 represents the probability that a user would create an account on the old homepage and π_2 represents the probability that a user would create an account while viewing the redesigned homepage. The null hypothesis here assumes that the redesigned webpage is not better than the original since H_0: $\pi_1 \geq \pi_2$. Thus we hope to collect data that contradicts this statement so that it can be rejected in favor of the alternative that says the redesign is in fact superior (H_A: $\pi_1 < \pi_2$), and hence worth the expense and effort.

In order to test this hypothesis they randomized $n_1 = 8,872$ users to the original homepage and $n_2 = 8,642$ users to the redesigned one. In these conditions they observed 280 and 399 conversions, respectively. That is, 280 users in the control condition created accounts while 399 users in the redesign condition created accounts. This sample data is summarized numerically by $\hat{\pi}_1 = 280/8872 = 0.0316$ and $\hat{\pi}_2 = 399/8642 = 0.0462$ which in practical terms means that 3.16% of users in the control condition created accounts and 4.62% of users in the redesign condition created accounts – corresponding to a 46% increase over the control. We also find $\hat{\pi} = (280+399)/(8872+8642) = 0.0388$ meaning that the overall account creation rate is 3.88%. A 95% confidence interval for the difference in conversion rates is given by

$$0.0316 - 0.0462 \pm 1.96 \times \sqrt{0.0388(1-0.0388)(\frac{1}{8872}+\frac{1}{8642})} = [-0.0203, -0.0089].$$

To determine whether the difference in account creation rates between the two conditions is statistically significant, we must formally test the hypothesis by calculating a p-value. To do this, me must first calculate the observed test statistic. Substituting these summaries into equation (19) gives

$$t = \frac{0.0316 - 0.0462}{\sqrt{(0.0388)(0.9612)\left(\frac{1}{8872}+\frac{1}{8642}\right)}} = -5.0075.$$

The p-value associated with this test is $\Pr(T \leq -5.0075)$ where $T \sim \mathrm{N}(0,1)$. When calculated this probability is 2.76×10^{-7}, which is effectively 0. In R this probability is calculated using the command `1-pnorm(-5.0075)`.

In order to draw a conclusion, we must compare this value to the significance level $\alpha = 0.05$. Since $2.84 \times 10^{-7} < 0.05$ we reject the null hypothesis in favor of the alternative. In the context of the experiment, this means that the redesigned homepage has a significantly larger likelihood of user account-creation than does the original homepage. Specifically, a 46% increase in account-creation can be expected with the redesigned homepage relative to the original.

2.2.4 Power Analysis and Sample Size Calculations

Here we derive sample size formulae in a manner similar to the development presented in Section 2.1.4 but here we do it in the context of hypothesis tests such as (13), (14) and (15). As in Section 2.1.4 we perform the derivation assuming a two-sided hypothesis is being tested, but we indicate where and how the derivation would change if it were a one-sided hypothesis that was of interest. We also present the derivation in a general manner that does not require equal sample sizes in each condition, and so we assume $n_1 = kn_2$.

As we saw in Section 2.2.1, the null distribution in this scenario is the standard normal distribution – just like it was for the sample size calculations in Section 2.1.4. A convenient consequence of this is that the rejection regions defined in that section are appropriate here as well. The only difference is that we use a different test statistic here. In particular we use the quantity given in equation (16). Recall that we preferred not to use this particular equation when actually testing the hypothesis because it required knowing π_1 and π_2, which are typically not known in practice. However, we note that based on historical data, a data scientist will typically have a good idea of what π_1 is if condition 1 corresponds to the existing product/ platform/ process/ page, etc. Also, when planning the experiment the data scientist will define $\delta = \pi_1 - \pi_2$ to be the MDE (in a manner similar to Section 2.1.4). Thus, with these two pieces of information π_2 can be defined as $\pi_2 = \pi_1 - \delta$, which means we can treat both π_1 and π_2 as known.

As before, we begin by considering the power of the hypothesis test:

$$\begin{aligned}
1-\beta &= \Pr(\text{Reject } H_0 \mid H_0 \text{ is false}) \\
&= \Pr(T \in R \mid H_0 \text{ is false}) \text{ where } R \text{ is the rejection region} \\
&= \Pr(T \geq z_{\alpha/2} \text{ or } T \leq -z_{\alpha/2} \mid H_0 \text{ is false}) \\
&= \Pr(T \geq z_{\alpha/2} \mid H_0 \text{ is false}) + \Pr(T \leq -z_{\alpha/2} \mid H_0 \text{ is false}) \\
&= \Pr\left(\frac{(\overline{Y}_1 - \overline{Y}_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \geq z_{\alpha/2} \mid H_0 \text{ is false} \right) \\
&\quad + \Pr\left(\frac{(\overline{Y}_1 - \overline{Y}_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \leq -z_{\alpha/2} \mid H_0 \text{ is false} \right)
\end{aligned}$$

If $H_0 : \pi_1 = \pi_2$ were true, and hence $\pi_1 - \pi_2 = 0$ were true, then the ratios in the preceding line would follow a $\text{N}(0, 1)$ distribution. However, we know that H_0 is false which means that $\pi_1 - \pi_2 = \delta$ for some none-zero δ, and so it is

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - \delta}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$$

that follows a $\text{N}(0, 1)$ distribution. Let us make this substitution, being sure to replicate what is done on the left side of inequalities on the right. Also note that we no longer need to write "$|H_0$ is false" since we are now exploiting this fact.

$$\begin{aligned}1-\beta &= \Pr\left(\frac{(\overline{Y}_1-\overline{Y}_2)-\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \geq z_{\alpha/2}-\frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right)\\ &\quad + \Pr\left(\frac{(\overline{Y}_1-\overline{Y}_2)-\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \leq -z_{\alpha/2}-\frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right)\\ &= \Pr\left(Z \geq z_{\alpha/2}-\frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right)\\ &\quad + \Pr\left(Z \leq -z_{\alpha/2}-\frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right) \text{ where } Z\sim \mathrm{N}(0,1)\end{aligned}$$

As in Section 2.1.4 only one of these two terms will dominate, depending on the sign of δ. Assume, without loss of generality, that $\delta > 0$ in which case only the first term remains.

$$1-\beta = \Pr\left(Z \geq z_{\alpha/2}-\frac{\delta}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\right)$$

Because this probability is equal to $1-\beta$ we know that $z_{\alpha/2}-\delta/\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}$ must be equal to $z_{1-\beta}$, the β^{th} quantile of the standard normal distribution. Thus

$$z_{1-\beta} = z_{\alpha/2}-\delta/\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}$$

and we can rearrange this equation solving for the sample size. But first we must substitute $n_1 = kn_2$ so that there is just a single sample size to solve for:

$$z_{1-\beta} = z_{\alpha/2}-\delta/\sqrt{\frac{\pi_1(1-\pi_1)}{kn_2}+\frac{\pi_2(1-\pi_2)}{n_2}} = z_{\alpha/2}-\delta\sqrt{n_2}/\sqrt{\frac{\pi_1(1-\pi_1)}{k}+\pi_2(1-\pi_2)}$$

Solving for n_2 yields:

$$n_2 = \frac{(z_{\alpha/2}-z_{1-\beta})^2[\frac{\pi_1(1-\pi_1)}{k}+\pi_2(1-\pi_2)]}{\delta^2} \tag{20}$$

and then n_1 is found by computing kn_2. When equal sample sizes are desired ($k=1$) each condition receives n units where

$$n = \frac{(z_{\alpha/2}-z_{1-\beta})^2[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}{\delta^2}. \tag{21}$$

If it were a one-sided hypothesis being tested, the only difference between the formulae in that setting relative to equations (20) and (21) above is that in the one-sided case instances of $z_{\alpha/2}$ would be replaced by z_α.

Note that the interactive sample size calculator found at https://nathaniel-t-stevens.shinyapps.io/SampleSizeCalculator/ can also be used to explore the interdepencies between sample size, significance level, power, and the MDE in this setting as well.

2.2.5 Another Way to Think About Comparing Proportions

In order to test the hypotheses (13), (14) and (15) using the Z-test of Section 2.2.1 we required the assumption that the response measurements followed a Bernoulli distribution (i.e., $Y_{ij} \sim \text{BIN}(1, \pi_j)$) and the sample sizes were large enough to ensure the Central Limit Theorem was applicable. Here we describe an equivalent method of testing these hypotheses, but motivated in a slightly different manner. In particular we discuss the **chi-squared test of independence** (also known as Pearson's χ^2-test).

The chi-squared test of independence is typically used as a test for 'no association' between two categorical variables that are summarized in a contingency table. We apply this methodology here to test the independence of the binary outcome (whether a unit performs the action of interest) and the particular condition they are in. If the likelihood of performing the action is the same in each condition (i.e., $\pi_1 = \pi_2$) then the response and conditions are not associated. As such, this test is appropriate for evaluating whether $\pi_1 = \pi_2$, $\pi_1 > \pi_2$ or $\pi_1 < \pi_2$. To motivate this, consider the Optimizely data from Section 2.2.3 which have been arranged in a 2×2 contingency table shown in Table 1. When arranged in this fashion we clearly see that there were $n_1 = 8872$ units in condition 1, $n_2 = 8642$ units in condition 2 and there were respectively 280 and 399 conversions in these conditions (and hence 8592 and 8243 non-conversions).

Table 1: 2×2 contingency table for Optimizely's homepage experiment

		Condition		
		1	2	
Conversion	Yes	280	399	679
	No	8592	8243	16835
		8872	8642	17514

If $\pi_1 = \pi_2 = \pi$ then we would expect the conversion rate in each condition to be the same. An estimate of the pooled conversion rate in this case is $\hat{\pi} = 679/17514 = 0.0388$ since there were 679 conversions in total, and an overal sample size of 17514 users. Thus, we would expect $n_1\hat{\pi} = 8872 \cdot 0.0388 = 344.23$ conversions in condition 1 and $n_2\hat{\pi} = 8642 \cdot 0.0388 = 335.31$ conversions in condition 2. Clearly this is not what we observed, but the chi-squared test formally evaluates if the difference between what was observed and what is expected under the null hypothesis is large enough to be considered *significantly* different.

Table 2: A general 2×2 contingency table

		Condition		
		1	2	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	O_0
		n_1	n_2	$n_1 + n_2$

We formalize this process by considering the general 2×2 contingency table in Table 2, where we let $O_{1,j}$ and $O_{0,j}$ respectively represent the observed number of conversions and non-conversions in condition

$j = 1, 2$. Also, O_1 and O_0 represent the overall number of conversions and non-conversions (between both conditions) and so

$$\hat{\pi} = \frac{O_1}{n_1 + n_2} \text{ and } 1 - \hat{\pi} = \frac{O_0}{n_1 + n_2}$$

represent the proportions of units that did or did not convert. As demonstrated above, we use these pooled estimates to calculate the expected number of conversions/non-conversions in each condition. Specifically, we let $E_{1,j}$ and $E_{0,j}$ represent the expected number of conversions and non-conversions in condition $j = 1, 2$ which we calculate as

$$E_{1,j} = n_j \hat{\pi} \text{ and } E_{0,j} = n_j(1 - \hat{\pi}).$$

The χ^2 test statistic compares the observed count in each cell to the corresponding expected count, and is defined as

$$T = \sum_{l=0}^{1} \sum_{j=1}^{2} \frac{(O_{l,j} - E_{l,j})^2}{E_{l,j}}.$$

Assuming H_0 is true, it can be shown that T approximately follows a χ^2 distribution with $\nu = 1$ degree of freedom (i.e., $T \sim \chi^2_{(1)}$). As a general rule of thumb, this approximation may be very poor unless the observed and expected cell frequencies are all greater than 5.

Then, to draw a conclusion about the hypothesis, we compare the observed value of the test statistic t to the $\chi^2_{(1)}$ distribution. The p-value associated with this test is calculated differently depending on whether H_A is one- or two-sided. These calculations are summarized below:

- H_A: $\pi_1 \neq \pi_2$: p-value $= \Pr(T \geq t)$
- H_A: $\pi_1 > \pi_2$: p-value $= 1 - \Pr(T \geq t)/2$
- H_A: $\pi_1 < \pi_2$: p-value $= \Pr(T \geq t)/2$

Although it may not look like it, the value of this χ^2 test statistic will always be the square of the value produced by equation (19). For this reason, and because squaring a $N(0, 1)$ random variable yields a $\chi^2_{(1)}$ one, the p-values of this χ^2 test will always be identical to those of the Z-test for proportions, and so the conclusions drawn by either method will be the same.

Returning to the Optimizely example, the observed number of conversions in condition 1 and 2 are $O_{1,1} = 280$ and $O_{1,2} = 399$ and the observed number of non-conversions in each condition are $O_{0,1} = 8592$ and $O_{0,2} = 8243$. Using this information we calculate the overall conversion and non-conversion rates to be $\hat{\pi} = (280 + 399)/(8872 + 8642) = 0.0388$ and $1 - \hat{\pi} = (8592 + 8243)/(8872 + 8642) = 0.9612$. With this we calculate the expected number of conversions in conditions 1 and 2: $E_{1,1} = 343.96$ and $E_{1,2} = 335.04$ and the the expected number of non-conversions in conditions 1 and 2: $E_{0,1} = 8528.04$ and $E_{0,2} = 8306.96$. The observed test statistic is then calculated as

$$t = \frac{(280 - 343.96)^2}{343.96} + \frac{(399 - 335.04)^2}{335.04} + \frac{(8592 - 8528.04)^2}{8528.04} + \frac{(8243 - 8306.96)^2}{8306.96} = 25.0755.$$

Then $\Pr(T \geq 25.0755) = 5.52 \times 10^{-7}$, where $T \sim \chi^2_{(1)}$, and the p-values associated with hypotheses (13), (14) and (15) are respectively 5.52×10^{-7}, 0.9999997, and 2.76×10^{-7}. The first p-value suggests that we would reject H_0: $\pi_1 = \pi_2$, suggesting that the conversion rates on the two versions of the homepage are indeed different. The second and third p-values both suggest that $\pi_1 < \pi_2$ is true, indicating that the likelihood of creating an account on the redesigned homepage is higher than on the original version of the homepage.

Note that this test may be implemented in R using the `prop.text()` function.

2.3 The Trouble with Peeking

In Sections 2.1.4 and 2.2.4 we developed sample size calculations to determine the necessary number of units in each condition to ensure the Type I and Type II Error rates are held fixed at the predetermined values α and β. However, in practice, the experimentation platform used by a data scientist may provide a dashboard which displays whether, at that current point in time, there is a significant difference between conditions – or that one condition is significantly better than the other.

Often a data scientist may feel external (and/or internal) pressure to stop the experiment when they see this. After all, the results tell us that a winner has been found, right? Wrong. Well, maybe, but by stopping the experiment early you have not observed enough data to be confident in your conclusion. By stopping the experiment you are in effect rejecting the null hypothesis (that the conditions are not different) and so you risk making a Type I Error. And by stopping the experiment early the chances you make a Type I Error are much higher than the value α you chose when doing your sample size calculation.

This phenomenon whereby you regularly check the results of the experiment, waiting for a significant result, is known as "peeking". Peeking is certainly tempting, and depending on your experimentation dashboard, it may be impossible to avoid. In some circumstances, when several metrics are being tracked (in addition to your primary metric of interest) it is in fact a good idea to 'peek' to ensure the experiment is not negatively impacting other important *guardrail* metrics.

The problem, however, arises when, as a result of peeking, you decide to end the experiment early. Just because the results suggest a winner or a significant difference at one point in time does not mean that the results won't change as more data is collected. For instance, I might peek at my experiment now and see that condition 1 is significantly out-performing condition 2. But if I peek again in an hour I might find that condition 2 is significantly out-performing condition 1. Only until you have observed the pre-specified amount of data should you be sure of your conclusions.

To illustrate the dire consequences of peeking and ending an experiment early, consider the following simulated situation. Imagine condition 1 response measurements truly follow a $\mathrm{N}(0, 1)$ distribution and condition 2 response measurements also follow a $\mathrm{N}(0, 1)$ distribution and a t-test is performed to decide whether or not $\mu_1 = \mu_2$. In this case the null hypothesis H_0: $\mu_1 = \mu_2$ is true and the data collected should not reject it. However, simply due to chance we may obtain a dataset which leads us to reject H_0 and make a Type I Error. However, if the sample sizes n_1 and n_2 were determined so that $\alpha = 0.05$, for example, we would not expect to make this type of error more than 5% of the time.

By repeatedly simulating n_1 and n_2 data points independently from the $\mathrm{N}(0, 1)$ distribution, and each time testing the null hypothesis H_0: $\mu_1 = \mu_2$ we can empirically quantify the likelihood of making a Type I Error. For illustration we do this $N = 100,000$ times, each time with samples of size $n_1 = n_2 = 1000$. In addition to quantifying the Type I Error rate if we waited for all $n_1 = n_2 = 1000$ data points to be observed, we also calculate the Type I Error rate when the experiment is ended early by peeking at regular intervals.

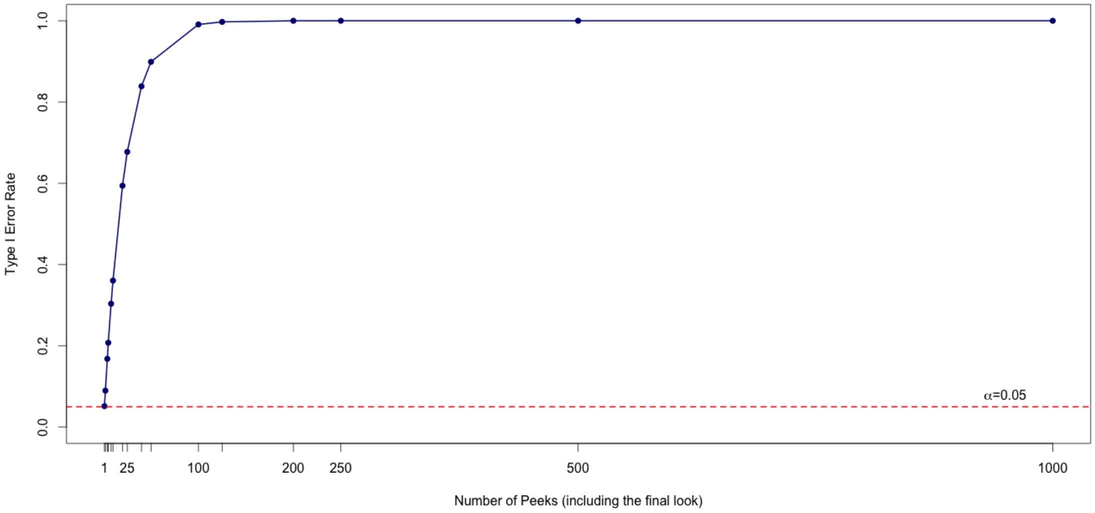

Figure 6: Type I Error rate for different levels of peeking.

Here we consider peeking (and ending the experiment if a significant result is indicated) after every successive data point and at intervals of every 2nd, 4th, 5th, 8th, 10th, 20th, 25th, 40th, 50th, 100th, 125th, 200th, 250th, 500th and 1000th data point. This corresponds to peaking 1000 times, 500 times, 250 time, 200 times, 125 times, 100 times, 50 times, 40 times, 25 times, 20 times, 10 times, 8 times, 5 times, 4 times, two times and no peeking at all. For each case in the simulation we peek at the results at the specified interval and end the experiment if the results are statistically significant. We then calculate the Type I Error rate as the proportion of the $N = 100,000$ times that a Type I Error was made. The plot shown in Figure 6 demonstrates how the chances of making a Type I Error increase dramatically for increased levels of peeking. Indeed, it is only in the case with no peeking that the Type I Error rate is actually equal to 0.05, and with enough peeking, committing a Type I Error becomes certain.

We note here that **sequential analysis** and **sequential testing** are important statistical topics/disciplines that are concerned with devising statistically sound methods for performing repeated significance tests as more data becomes available. Essentially, sequential testing corresponds to a host of techniques that allow you to peek and end an experiment early without increasing Type I Error rates. However, without adopting one of these techniques, peeking (and ending experiments early) should be avoided at all costs. For a more in-depth treatment of sequential testing see Tartakovsky et al. (2014).

3 EXPERIMENTS WITH MORE THAN TWO CONDITIONS

In the previous chapter we considered the situation in which the experiment contained just two experimental conditions. In the language of designed experiments, this corresponds to the investigation of a single design factor at two levels. We motivated the situation by discussing the canonical A/B test in which two versions of a webpage were compared – one with a red button and the other with a blue button – and we were interested in identifying the 'winning' webpage – the one with the button that is clicked most frequently. But what if we want investigate three button colours rather than just two? Or what about 10 button colours?

Figure 7: Canonical Button Colour Test

In many real-life scenarios it is reasonable to believe that a data scientist may be interested in comparing more than just two conditions. For instance, one might be interested in comparing 6 different ads to determine which is most profitable; or, one might be interested in comparing 3 different sign-up promotions to determine which has the highest conversion rate. In general, the question being answered now is: given several options, which is best?

The types of experiments that are used to answer this question are colloquially referred to by data scientists as "A/B/C", "A/B/C/D", or more generally, "A/B/n" tests. Formally, these experiments are designed in a very similar manner to A/B tests; a metric of interest θ is chosen to be optimized and the corresponding response variable (y) is determined. Recall that the metric may be any statistic that can be calculated from observed data, with various user engagement and conversion metrics being commonly used in practice. The response variable, then, is the variable on which mneasurements are required to actually calculate the metric of interest.

What is different, relative to a traditional A/B test, is the number of levels of the design factor, and hence number of experimental conditions. As before we index experimental conditions by j, but rather than $j = 1, 2$, now we have $j = 1, 2, \dots, m$, where m is the total number of conditions. In this case the metric of interest is calculated in each condition, giving $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$, and interest lies in comparing them to determine which condition is optimal. Condition j would be considered optimal if the observed data provided enough evidence to believe $\theta_j > \theta_{j'}$ for all $j' \neq j$ (when maximizing the metric is important) or $\theta_j < \theta_{j'}$ for all $j' \neq j$ (when minimizing the metric is important).

In this chapter we describe the statistical tests that are used to draw conclusions of this sort, and we discuss practical and statistical problems that must be considered in this situation. Like the previous chapter we consider the comparison of means and the comparison of proportions.

3.1 Comparing Means in Multiple Conditions

As in Section 2.1, we assume that our response variable follows a normal distribution and we assume that the mean of the distribution depends on the condition in which the measurements were taken, and that the variance is the same across all conditions. Mathematically, we assume $Y_{ij} \sim \mathrm{N}(\mu_j, \sigma^2)$ for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, m$. Thus, the only difference between these distributions (if there is a difference) is in their means. As such, formal hypothesis tests in this scenario concern only the μ_j's. While interest ultimately lies in finding the condition with the highest (or smallest) μ_j, a common (and sensible) starting point is to decide whether there is a difference at all between the conditions. To answer this question formally, the following hypothesis is tested.

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_m \text{ vs. } H_A\colon \mu_j \neq \mu_{j'} \text{ for some } j' \neq j \tag{22}$$

Failing to reject H_0 means that the expected response does not differ significantly from one condition to another, and so no single condition is optimal. However, if the observed data provide enough evidence to reject H_0, then we would conclude that the expected response in at least one of the conditions is not the same as the others. Follow-up hypothesis tests can then be used to determine which condition(s) is (are) optimal. These follow-up tests are typically performed in a pairwise manner, comparing a given condition to each of the other conditions. The two-sample methods discussed in Section 2.1 are useful for this task. However, it is important to note that when doing multiple comparisons and hence testing a series of hypothesis tests, the overall Type I Error rate becomes inflated, and so modifications to the testing procedure must be made. We discuss this further in Section 3.3.

In the next subsection we discuss how to formally test hypothesis (22). As we will see, this test can be performed using the ***F*-test for overall significance** in an appropriately defined linear regression model. For a primer on linear regression, see Appendix A.3.

3.1.1 The *F*-test for Overall Significance in a Linear Regression

The "appropriately defined linear regression model" in this situation is one in which the response variable depends on $m - 1$ indicator variables where, for example, the indicator variables may be defined as

$$x_{ij} = \begin{cases} 1 & \text{if unit } i \text{ is in condition } j \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, 2, \ldots, m - 1$. For a particular unit i, we adopt the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{m-1} x_{i,m-1} + \varepsilon_i$$

where Y_i is the response observation for unit $i = 1, 2, \ldots, N = \sum_{j=1}^{m} n_j$ and ε_i is the corresponding random error term assumed to follow a $\mathrm{N}(0, \sigma^2)$ distribution.

In this model the β's are unknown parameters which we interpret in the following manner. The intercept β_0 is the expected response when each of the indicator variables is equal to zero: $\mathrm{E}[Y_i|x_{i1} = x_{i2} = \cdots = x_{i,m-1} = 0] = \beta_0$. By design, if all of the indicator variables are equal to zero, this means that unit i was in condition m. Thus β_0 is the expected response in condition m.

Similarly, since $\mathrm{E}[Y_i|x_{ij} = 1] = \beta_0 + \beta_j$, we define $\beta_0 + \beta_j$ to be the expected response in condition $j = 1, 2, \ldots, m-1$, and interpret β_j as being the expected change in response in condition $j = 1, 2, \ldots, m-1$ relative to in condition m. As such condition m represents the baseline against which all other conditions are compared. Note that condition m was chosen here as the baseline for notational convenience. In principle *any* condition can serve as the baseline; in practice it us useful to define the 'control' condition (if there is one) as the baseline.

Based on these assumptions we have:

$$\begin{aligned} \mu_1 &= \beta_0 + \beta_1 \\ \mu_2 &= \beta_0 + \beta_2 \\ &\vdots \\ \mu_{m-1} &= \beta_0 + \beta_{m-1} \\ \mu_m &= \beta_0 \end{aligned}$$

As can be seen, H_0 in (22) is true if and only if $\beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$. Thus testing (22) is equivalent to testing

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ vs. } H_A\colon\ \beta_j \neq 0 \text{ for some } j$$

in the context of the linear regression model above. Such a test is known as the F-test for overall significance in a linear regression. The test statistic is defined to be the ratio of the regression mean squares (MSR) to the mean squared error (MSE) in a standard regression-based analysis of variance (ANOVA):

$$t = \frac{MSR}{MSE}.$$

Note that MSE is an estimate of σ^2, as described in Appendix A.3, and MSR is related to the MSE of the *reduced model* that assumes H_0 is true (i.e., $\beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0$). For details on this test statistic and the *additional sum of squares principle* that gives rise to it, see Abraham and Ledolter (2006)

Assuming the null hypothesis is true, this test statistic should look as if it comes from an F-distribution with $\nu_1 = m - 1$ and $\nu_2 = N - m$ degrees of freedom. The p-value associated with this test is calculated as p-value $= \Pr(T \geq t)$ where $T \sim F_{(m-1,N-m)}$ and is commonly displayed in regression summaries provided by statistical software. For example, both the `summary()` and the `anova()` output of an `lm()` object in `R` provides the results of this test. The We illustrate the use of this test with an example in the next subsection.

3.1.2 Example: Candy Crush Boosters

Candy Crush is experimenting with three different versions of in-game "boosters": the lollipop hammer, the jelly fish, and the colour bomb. Users are randomized to one of these three conditions ($n_1 = 121$, $n_2 = 135$, $n_3 = 117$) and they receive (for free) 5 boosters corresponding to their condition. Interest lies in evaluating the effect of these different boosters on the length of time a user plays the game. Let μ_j represent the average length of game play (in minutes) associated with booster condition $j = 1, 2, 3$. While interest lies in finding the condition associated with the longest average length of game play, here we first rule out the possibility that booster type does not influence the length of game play (i.e., $\mu_1 = \mu_2 = \mu_3$). In order to do this we fit the linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where x_1 and x_2 are indicator variables indicating whether a particular value of the response was observed in the jelly fish or colour bomb conditions, respectively. The lollipop hammer is therefore the reference condition. By using the `lm()` function in `R` we obtain the following output

```
Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 5.01281 0.08664 57.859 <2e-16 *** factor
```

From this output we see that $\hat{\beta}_0 = 5.0128$, $\hat{\beta}_1 = 1.1753$ and $\hat{\beta}_2 = 4.8828$ indicating the average length of game play is estimated to be $\hat{\mu}_1 = \hat{\beta}_0 = 5.0128$ minutes in the lollipop hammer condition, $\hat{\mu}_2 = \hat{\beta}_0 + \hat{\beta}_1 = 5.0128 + 1.1753 = 6.1881$ minutes in the jelly fish condition, and $\hat{\mu}_3 = \hat{\beta}_0 + \hat{\beta}_2 = 5.0128 + 4.8828 = 9.8956$ minutes in the colour bomb condition. These estimates suggest that the average length of game play differs depending on which booster condition a user is in. To formally draw this conclusion we perform the F-test of overall significance of the regression. The output shown above indicates that the observed test statistic is calculated to be $t = 851.9$ and the p-value $= \Pr(T \geq 851.9)$ is less than 2.2×10^{-16}, where T follows an F-distribution with $\nu_1 = 2$ and $\nu_2 = 370$ degrees of freedom. Such a small p-value provides very strong evidence against H_0 and so we conclude that the average length of game play is not the same for each of the boosters. To determine which booster is optimal – the one that maximizes game play duration – we must use a series of pairwise t-tests. This is left as an exercise for the reader.

3.2 Comparing Proportions in Multiple Conditions

Like in Section 2.2, we assume that our response variable is binary:

$$Y_{ij} = \begin{cases} 1 & \text{if unit } i \text{ in condition } j \text{ performs an action of interest} \\ 0 & \text{if unit } i \text{ in condition } j \text{ does not perform an action of interest} \end{cases}$$

for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, m$. As before we define $\pi_j = \Pr(Y_{ij} = 1)$ to be the probability that a unit in condition j performs the action – whether this means that they click a button, open an email, create an account, etc. Of interest, then, is a comparison of the likelihood that the action is performed across all conditions, with the ultimate goal of finding the condition with the highest (or smallest – whichever corresponds to optimal) π_j. While this is the ultimate goal, a sensible first step is to decide whether there is a difference between the conditions at all. In order to formally make this decision, the following hypothesis is tested.

$$H_0\text{: } \pi_1 = \pi_2 = \cdots = \pi_m \text{ vs. } H_A\text{: } \pi_j \neq \pi_{j'} \text{ for some } j' \neq j \tag{23}$$

Failing to reject H_0 means that the action of interest is no more probable in one condition than any other, and so no single condition is optimal. However, if the observed data provide enough evidence to reject H_0, then we would conclude that there is at least one condition in which units behave differently. Follow-up hypothesis tests can then be used to determine which condition(s) is (are) optimal. These follow-up tests are typically performed in a pairwise manner, comparing a given condition to each of the other conditions. The two-sample methods discussed in Section 2.2 are useful for this task. However, we remark again that performing multiple comparisons can lead to an increased Type I Error rate, which we discuss further in Section 3.3.

In the next subsection we discuss how to formally test hypothesis (23). As we will see, the χ^2 test from Section 2.2.5 generalizes to the comparison of any number of conditions and so we will apply it again in this scenario.

3.2.1 The Chi-squared Test of Independence

In Section 2.2.5 we introduced the χ^2 test of independence as a means to evaluate whether $\pi_1 = \pi_2$, $\pi_1 > \pi_2$, or $\pi_1 < \pi_2$ in the context of two experimental conditions. However, we noted that the test is more generally thought of as a test of 'no association' between two categorical variables. Therefore, we used it as a test of 'no association' between the binary outcome (whether a unit performs the action of interest) and the particular condition they are in. In that setting we considered just two conditions – but the test itself imposes no such restriction; here we consider the test more generally, in the context of m experimental conditions.

As before we are interested in comparing observed and expected frequencies of each outcome in each condition. The information associated with this test can be summarized in a $2 \times m$ contingency table, where rows correspond to the binary outcome, conversion, and the columns correspond to the different conditions. The value in the $(l, j)^{th}$ cell of this table, denoted $O_{l,j}$, corresponds to the observed number of conversions ($l = 1$) or non-conversions ($l = 0$) in condition $j = 1, 2, \ldots, m$. An example of such a table is shown in Table 3.

As in the 2×2 case, each of these observed frequencies is contrasted with an expected frequency where $E_{1,j}$ is the expected number of conversions in condition j and $E_{0,j}$ is the expected number of non-conversion in condition j. These expected frequencies are found by multiplying condition j's sample size by the pooled

Table 3: A general $2 \times m$ contingency table

		Condition				
		1	2	$\cdots$	m	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	$\cdots$	$O_{1,m}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	$\cdots$	$O_{0,m}$	O_0
		n_1	n_2	$\cdots$	n_m	$N = \sum\limits_{j=1}^{m} n_j$

conversion and non-conversion rates:

$$E_{1,j} = n_j \hat{\pi} \text{ and } E_{0,j} = n_j(1 - \hat{\pi})$$

where

$$\hat{\pi} = \frac{O_1}{N} \text{ and } (1 - \hat{\pi}) = \frac{O_0}{N}$$

are the sample estimates of homogenous probabilities of conversion and non-conversion.

The test statistic for this test is defined exactly as it was in the 2×2 case except that now we are summing over more cells and the null distribution has a different number of degrees of freedom. In particular, the test statistic is

$$T = \sum_{l=0}^{1} \sum_{j=1}^{m} \frac{(O_{l,j} - E_{l,j})^2}{E_{l,j}}$$

where, if H_0 is true, an observed value, t, should look as if it comes from a χ^2 distribution with $\nu = m - 1$ degrees of freedom. The p-value associated with this test is calculated as p-value $= \Pr(T \geq t)$ where $T \sim \chi^2_{(m-1)}$. As in the 2×2 case, this test can be carried out automatically using the `prop.test()` function in `R`. We illustrate the use of this test with an example in the next subsection.

3.2.2 Example: Nike SB Video Ads

Suppose that Nike is running an ad campaign for Nike SB, their skateboarding division, and the campaign involves $m = 5$ different video ads the are being shown in Facebook newsfeeds. A video ad is 'viewed' if it is watched for longer than 3 seconds, and interest lies in determining which ad is most popular and hence most profitable by comparing the viewing rates of the five different videos. Each of these 5 videos is shown to $n_1 = 5014$, $n_2 = 4971$, $n_3 = 5030$, $n_4 = 5007$, and $n_5 = 4980$ users and in each condition the videos are viewed 160, 95, 141, 293 and 197 times, respectively, yielding watch rates of $\hat{\pi}_1 = 0.0319$, $\hat{\pi}_2 = 0.0191$, $\hat{\pi}_3 = 0.0280$, $\hat{\pi}_4 = 0.0585$, and $\hat{\pi}_5 = 0.0396$.

Based on these estimates it would suggest that not all of the videos are equally popular, but to formally decide this we will conduct a χ^2-test. The 2×5 contingency table for these data are shown in Table 4.

The expected cell frequencies are found by multiplying n_j by $\hat{\pi}$ and $(1 - \hat{\pi})$, $j = 1, 2, 3, 4, 5$, where $\hat{\pi}$ is calculated using these data to be $\hat{\pi} = 886/25002 = 0.0354$. Table 5 displays these frequencies. The observed

Table 4: A 2 × 5 observed contingency table for the Nike example

		Condition					
		1	2	3	4	5	
View	Yes	160	95	141	293	197	886
	No	4854	4876	4889	4714	4783	24116
		5014	4971	5030	5007	4980	25002

Table 5: A 2 × 5 expected contingency table for the Nike example

		Condition					
		1	2	3	4	5	
View	Yes	177.68	176.16	178.25	177.43	176.48	886
	No	4836.32	4794.84	4851.75	4829.57	4803.52	24116
		5014	4971	5030	5007	4980	25002

test statistic for these data is calculated to be

$$t = \sum_{l=0}^{1} \sum_{j=1}^{5} \frac{(O_{l,j} - E_{l,j})^2}{E_{l,j}} = 129.1686.$$

The p-value, then, is $\Pr(T \geq 129.1686) = 5.86 \times 10^{-27}$ where $T \sim \chi^2_{(4)}$. Such a small p-value provides very strong evidence against H_0 in this case, and so we conclude that the likelihood that someone 'views' a video is not the same for all of the videos. To determine which video is optimal – the one with the highest likelihood of viewing – we must use a series of pairwise Z-tests or χ^2-tests. This is left as an exercise for the reader.

3.3 The Problem of Multiple Comparisons

As the examples in Sections 3.1.2 and 3.2.2 illustrate, the hypothesis of overall equality (see e.g., (22) or (23)) is often rejected. In these situations, a series of follow-up pairwise comparisons are necessary to determine which condition(s) is (are) optimal. From a practical standpoint, we are already armed with the statistical machinery to do this; we need only perform several two-sample t-tests, Z-tests, χ^2-squared tests or randomization tests (whatever the situation calls for). However, when doing multiple comparisons like this, it is important to recognize that if each individual test has a Type I Error rate of α, the likelihood that a Type I Error is made *somewhere*, is much larger than α. This is known as the **multiple comparison** or **multiple testing problem**.

To frame this discussion let us define some notation. Suppose that M hypothesis tests are performed, in M_0 of which the null hypothesis H_0 is true and $M_A = M - M_0$ of which the null hypothesis is false. Note that M is observable, but M_0 and M_A are not – they unknowable, yet fixed constants. For each of the M tests, on the basis of the observed data, we choose to reject or not reject H_0. In particular, suppose we choose to reject H_0 in R of the tests and hence not reject H_0 in $M - R$ of the tests. The outcomes of these M decisions are nicely summarized in the following table:

Each of these quantities may be interpreted as follows:

Table 6: Outcomes from M simultaneous hypothesis tests

		Decision		
		Reject H_0	Accept H_0	
Truth	H_0 is True	V	U	M_0
	H_0 is False	S	T	M_A
		R	$M-R$	M

- R is the number of null hypotheses that were rejected
- $M-R$ is the number of null hypotheses that were accepted
- M_0 is the number of true null hypotheses
- M_A is the number of false null hypotheses
- V is the number of true null hypotheses that were incorrectly rejected (i.e., number of Type I Errors)
- S is the number of false null hypotheses that were incorrectly rejected
- U is the number of true null hypotheses that were correctly accepted
- T is the number of false null hypotheses that were incorrectly accepted (i.e., number of Type II Errors)

Like M, the numbers R and $M-R$ are observable. Like M_0 and M_A, V, U, S, T are unobservable, but these are in fact random variables; their values are determined by the random process of collecting data and testing the M hypotheses.

Ideally we would like V and T to be small. Unsurprisingly, T is related to power and thus may be controlled by sample size. In this section we discuss methods of controlling V, or more precisely, of controlling functions of V. Two such functions are the **family-wise error rate** (FWER) and the **false discovery rate** (FDR).

3.3.1 Family-Wise Error Rate

The family-wise error rate is defined as the probability of committing a Type I Error in *any* of the M hypothesis tests. Mathematically:

$$FWER = \Pr(V \geq 1)$$

Although each of the M tests may be carried out with a significance level α, the FWER will be much greater than α, even for modest M. Boole's inequality may be used to derive an upper bound for FWER, but the bound is not very optimistic.

$$\begin{aligned}
FWER &= \Pr(V \geq 1) \\
&= \Pr(\text{At least one Type I Error in } M \text{ tests}) \\
&= \Pr\left(\bigcup_{k=1}^{M} \text{Type I Error on test } k\right) \\
&\leq \sum_{k=1}^{M} \Pr(\text{Type I Error on test } k) \\
&= \sum_{k=1}^{M} \alpha \\
&= M \times \alpha
\end{aligned}$$

If we are willing to assume the M tests are independent (i.e., their test statistics are statistically independent), we can show that $FWER = 1 - (1-\alpha)^M$.

$$\begin{aligned}
FWER &= \Pr(V \geq 1) \\
&= \Pr(\text{At least one Type I Error in } M \text{ tests}) \\
&= 1 - \Pr(\text{No Type I Error in } M \text{ tests}) \\
&= 1 - \Pr\left(\bigcap_{k=1}^{M} \text{No Type I Error on test } k\right) \\
&= 1 - \prod_{k=1}^{M} \Pr(\text{No Type I Error on test } k) \\
&= 1 - \prod_{k=1}^{M} (1-\alpha) \\
&= 1 - (1-\alpha)^M
\end{aligned}$$

Figure 8 illustrates the dependence of this family-wise error rate on the number of hypothesis tests, M. As we can see, as M increases so also does FWER. In the limit (i.e., as $M \to \infty$) this error rate goes to 1, and so as the the number of tests increases it becomes certain that a Type I Error will have been made somewhere. In practice, a common value of M is $\binom{m}{2}$: the number of pairwise comparisons necessary to compare each condition to every other condition. Supposing the experiment consists of $m = 5$ experimental conditions, then $M = \binom{5}{2} = 10$ which, if the significance level on each test is $\alpha = 0.05$, results in a family-wise error rate of 0.4013 – much higher than the Type I Error rate we are comfortable with.

A number of statistical procedures have been developed to combat this. See Wright (1992) for an overview. Using these techniques it is possible to carry out M tests while ensuring the FWER does not exceed some threshold (i.e., $FWER \leq \alpha^\star$). Here we consider three of the most common: the Bonferroni correction (Dunnett, 1955), the Šidák correction (Šidák, 1967), and Holm's procedure (Holm, 1979). For purposes of the discussion below, denote the M null hypotheses by $H_{0,1}, H_{0,2}, \ldots, H_{0,M}$ and the corresponding p-values

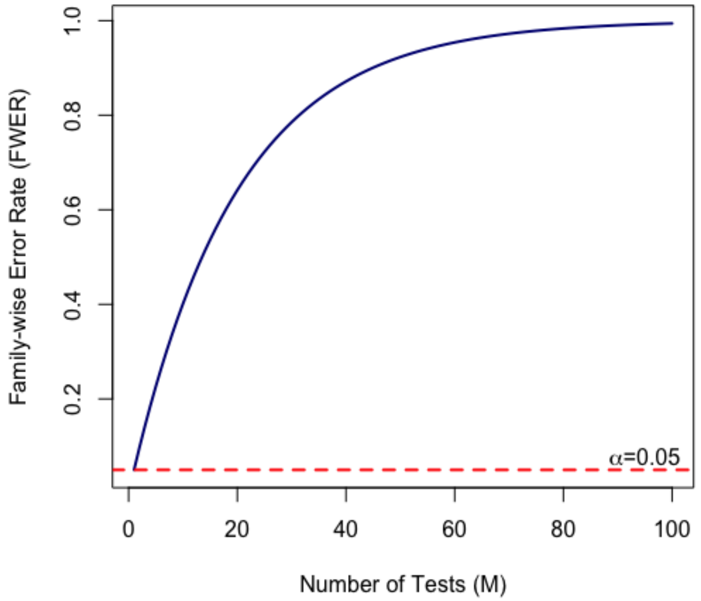

Figure 8: Family-wise error rate versus the number of hypothesis tests, M.

as $p_1, p_2, \ldots, p_M$. Also, as a running example, consider a situation in which four hypotheses are tested and the resulting p-values are $p_1 = 0.015$, $p_2 = 0.029$, $p_3 = 0.008$, and $p_4 = 0.026$.

The Bonferroni Correction:

The simplest method among the three, the Bonferroni correction mandates that null hypothesis $H_{0,k}$ be rejected only if

$$p_k \leq \frac{\alpha^\star}{M}$$

for $k = 1, 2, \ldots, M$. In other words, the Bonferroni correction involves testing each of the M hypotheses at a significance level of $\alpha^\star/M$. Doing so ensures that $FWER \leq \alpha^\star$ as is readily seen by substituting α with $\alpha^\star/M$ in the expression derived by Boole's inequality above. When independence is assumed, and the FWER was shown to be $1-(1-\alpha)^M$, the Bonferroni-corrected FWER becomes $1-(1-\frac{\alpha^\star}{M})^M$. Notice now that as M increases the FWER no longer approaches 1. Instead

$$\lim_{M\to\infty} 1 - \left(1 - \frac{\alpha^\star}{M}\right)^M = 1 - e^{-\alpha^\star}$$

which for typical values of $\alpha^\star$ in the range $(0, 0.1]$ is approximately equal to $\alpha^\star$. Figure 9 illustrates this.

Consider the example introduced above, in which we have four hypotheses tested $H_{0,1}, H_{0,2}, H_{0,3}, H_{0,4}$ yielding four p-values: $p_1 = 0.015, p_2 = 0.029, p_3 = 0.008, p_4 = 0.026$. Suppose that we wish to ensure $FWER \leq \alpha^\star = 0.05$. Thus the Bonferroni-corrected significance threshold is $\alpha^\star/M = 0.05/4 = 0.0125$. Based on this criteria, we see that we would reject only $H_{0,3}$, despite *all four* p-values individually being

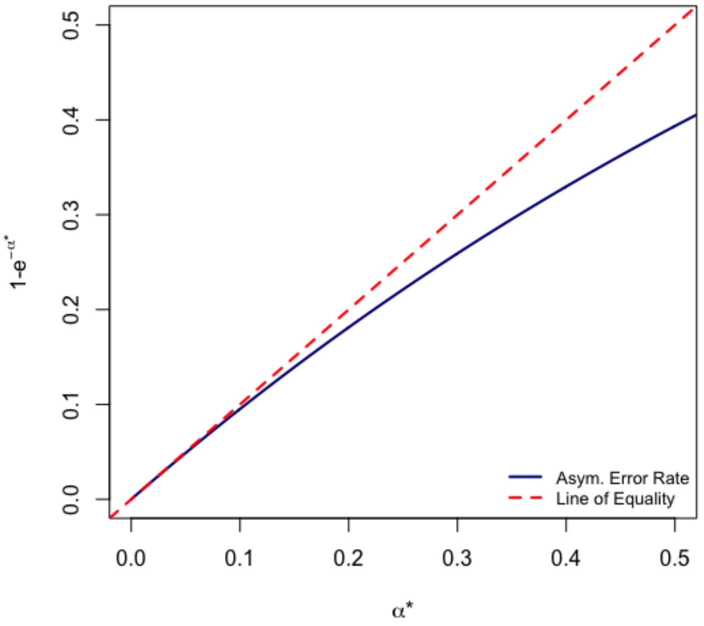

Figure 9: Illustration of the Bonferroni correction for asymptotically large M.

less than $\alpha^\star = 0.05$.

The Šidák Correction:

We have seen that if the M hypothesis tests are independent, and each tested at a significance level α, then $FWER = 1 - (1 - \alpha)^M$. If we wish to ensure that $FWER = \alpha^\star$ for some $\alpha^\star \in [0, 1]$ then it is straightforward to show that this is achieved when each of the tests is carried out with significance level $1 - (1 - \alpha^\star)^{\frac{1}{M}}$. This correction, known as Šidák's correction, mandates that null hypothesis $H_{0,k}$ be rejected only if

$$p_k \leq 1 - (1 - \alpha^\star)^{\frac{1}{M}}$$

for $k = 1, 2, \ldots, M$. Strictly speaking the Šidák correction is slightly less conservative than the Bonferroni correction since

$$1 - (1 - \alpha^\star)^{\frac{1}{M}} > \frac{\alpha^\star}{M}$$

But in practice the difference is immaterial (especially for large M) since the Bonferroni threshold is the first-order Taylor approximation of the Šidák threshold. However, even for reasonably small M the difference between the two is small. For instance, with $\alpha^\star = 0.05$ and $M = 10$, the Bonferroni threshold is 0.005 and the Šidák threshold is 0.005116197.

Let us illustrate this approach with our four-test example. Suppose, as we before, that we wish to ensure $FWER \leq \alpha^\star = 0.05$. Thus the Šidák-corrected significance threshold is $1 - (1 - \alpha^\star)^{\frac{1}{M}} = 1 - (1 - 0.05)^{\frac{1}{4}} =$

0.01274146. Unsurprisingly, this threshold is very similar to the Bonferroni threshold, and so we draw the same conclusions: only $H_{0,3}$ is rejected.

Holm's "Step-Up" Procedure:

This method is motivated by the desire for a correction method that is less strict (than the Bonferroni approach), but that still controls the FWER at some $\alpha^\star$. After all, the more strict the correction (i.e., the smaller the rejection threshold), the less likely *any* null hypothesis will be rejected. Holm's procedure is a surprisingly simple sequential solution which can be described as follows:

1. Order the M p-values from smallest to largest:

$$p_{(1)}, p_{(2)}, \ldots, p_{(M)}$$

   where $p_{(k)}$ is the k^{th} smallest p-value.

2. Starting from $k = 1$ and continuing incrementally, compare $p_{(k)}$ to $\frac{\alpha^\star}{M-k+1}$. Determine $k^\star$ the smallest value of k such that

$$p_{(k)} > \frac{\alpha^\star}{M-k+1}$$

3. Reject the null hypotheses $H_{0,(1)}, \ldots, H_{0,(k^\star-1)}$ and do not reject $H_{0,(k^\star)}, \ldots, H_{0,(M)}$.

To gain an intuition for how this procedure works, it is instructive to consider the comparison in Step 2 above, p-value by p-value. In particular, the smallest p-value $p_{(1)}$ is compared to $\frac{\alpha^\star}{M}$, the second smallest p-value $p_{(2)}$ is compared to $\frac{\alpha^\star}{M-1}$, and so on until the largest p-value $p_{(M)}$ is compared to $\alpha^\star$. At each stage, the next smallest p-value is being compared to a Bonferroni-corrected threshold that accounts for the number of null hypotheses that remain to be tested. A formal proof that this sequential procedure ensures $FWER \leq \alpha^\star$ is not included in these notes. The interested reader is referred to Holm (1979).

We illustrate this procedure with our four-test example. The ordered p-values in this case are $p_{(1)} = 0.008, p_{(2)} = 0.015, p_{(3)} = 0.026, p_{(4)} = 0.029$. Assuming $\alpha^\star = 0.05$, the Holm threshold against which each of these p-values are compared is

$$\left\{\frac{\alpha^\star}{M}, \frac{\alpha^\star}{M-1}, \frac{\alpha^\star}{M-2}, \frac{\alpha^\star}{M-3}\right\} = \left\{\frac{0.05}{4}, \frac{0.05}{3}, \frac{0.05}{2}, \frac{0.05}{1}\right\} = \{0.0125, 0.0167, 0.025, 0.05\}$$

Starting with $p_{(1)}$ we see that $p_{(1)} = 0.008 < 0.0125$ and so we reject $H_{0,(1)} = H_{0,3}$. Sequentially "stepping up" we next examine $p_{(2)}$. We see that $p_{(2)} = 0.015 < 0.0167$ and so we reject $H_{0,(2)} = H_{0,1}$. Sequentially "stepping up" we next examine $p_{(3)}$. We see that $p_{(3)} = 0.026 > 0.025$ and so we *do not* reject $H_{0,(3)} = H_{0,4}$. According to the Holm's procedure we now stop and also do not reject $H_{0,(4)} = H_{0,2}$, despite the fact that $p_{(4)} = 0.029 < 0.05$. Thus, the Holm's procedure is less stringent than either the Bonferroni correction or the Šidák correction: with this approach we reject both $H_{0,1}$ and $H_{0,3}$, whereas with the other methods we only rejected $H_{0,3}$.

To aid in developing an intuition for the three correction-methods described above, we may plot the ordered p-values $p_{(k)}$ vs. their ranks $k = 1, 2, \ldots, M$ and overlay the significance thresholds. Such a plot, for our four-test example, is shown in Figure 10. According to the Bonferonni and Šidák corrections, null hypotheses with p-values lower than their significance threshold are rejected, and the null hypotheses with p-values above their thresholds are not rejected. These thresholds are respectively indicated in blue and green. The Holm threshold $\alpha^\star/(M-k+1)$ is shown in red and, unlike the Bonferroni and Šidák thresholds, has curvature. This threshold is used as follows: moving from left to right, locate the *first* p-value which falls *above* the threshold: this is $p_{(k^\star)}$. Thus the hypotheses corresponding to p-values smaller than this one are rejected and all others are not rejected.

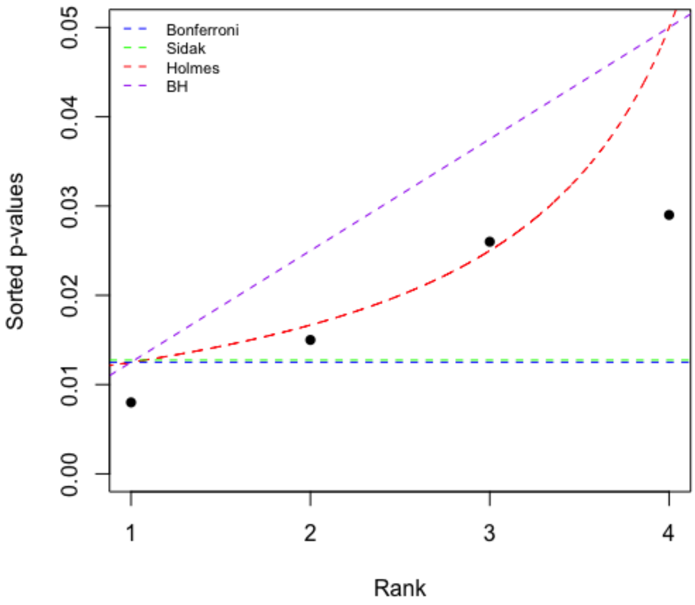

Figure 10: Significance thresholds for several methods of correction.

Adjusted p-values:

Until now we have framed each of the correction procedures above as an adjustment to the significance threshold against which each p-value is compared. However, we could alternatively (and equivalently) frame the correction procedures in terms of *adjusted p-values*. As we have seen, each correction method involved a comparison between the observed p-values and some function of $\alpha^\star$. We could, instead, re-write these comparisons so that in each case the p-values are adjusted and then compared to the nominal $\alpha^\star$. This is in some sense more familiar, since if feels like the traditional decision framework: reject the null hypothesis if the p-value is less than or equal to $\alpha^\star$. The difference here, of course, is that the p-values must first be adjusted according to whichever correction method is used.

For each of the three correction methods discussed, it is trivial to show that the decisions made with the following adjusted p-values is identical to that achieved by comparing unadjusted p-values to the method's adjusted significance threshold.

- Bonferroni: Reject $H_{0,k}$ if $p_k^\star \leq \alpha^\star$ where
$$p_k^\star = Mp_k$$
- Šidák: Reject $H_{0,k}$ if $p_k^\star \leq \alpha^\star$ where
$$p_k^\star = 1 - (1 - p_k)^M$$
- Holm: Reject $H_{0,(k)}$ if $p_{(k)}^\star \leq \alpha^\star$ where
$$p_{(k)}^\star = \max_{j \leq k}\{p_{(j)}(M - j + 1)\}$$

In our four-test example with p-values $p_1 = 0.015, p_2 = 0.029, p_3 = 0.008, p_4 = 0.026$, we can calculate the adjustments by each of these methods and then compare the adjusted p-values to $\alpha^\star = 0.05$. The Bonferroni-adjusted p-values are $p_1^\star = 0.06, p_2^\star = 0.116, p_3^\star = 0.032, p_4^\star = 0.104$, and the Šidák-adjusted p-values are $p_1^\star = 0.0587, p_2^\star = 0.1111, p_3^\star = 0.0316, p_4^\star = 0.1$. As before, in both cases we only reject $H_{0,3}$ because in both cases only $p_3^\star < \alpha^\star$. The Holm-adjusted p-values are $p_1^\star = 0.045, p_2^\star = 0.052, p_3^\star = 0.032, p_4^\star = 0.0.052$. As before, we see that we reject $H_{0,1}$ and $H_{0,3}$ since both $p_1^\star$ and $p_3^\star$ are less than $\alpha^\star = 0.05$. This approach to p-value adjustment is automated in R with the function `p.adjust()`.

3.3.2 False Discovery Rate

The development of methods to control the family-wise error rate largely took place in the mid-1900's. More recently, however, there has been a shift of focus to the control of another error metric, the *false discovery rate*. This change of emphasis is due in large part to the increased size and complexity of modern-day data-driven efforts. The methods developed in the mid-1900's were suitable for upward of $M = 20$ simultaneous tests, but for much larger numbers of tests the methods becomes increasingly, and prohibitively, strict. Nowadays it is not unreasonable to carry out hundreds, thousands or even tens of thousands of tests simultaneously. The is especially true in the context of feature selection in high-dimensional machine learning models and also in the realm of genetics studies (Efron and Hastie, 2016). In the context of designed experiments, however, we rarely encounter values of M that large and so control of the FWER is typically sufficient. Nonetheless, for completeness, we provide a brief overview of the false discovery rate.

Controlling $FWER = \Pr(V \geq 1)$ effectively controls the *number* of false positives. However, in light of such large values of M there has been a recognition that it is more sensible to control the *rate* that Type I Errors (i.e., *false discoveries*) occur. The *false discovery proportion* (FDP), defined as

$$Q = \frac{V}{R}$$

is the proportion of all rejected null hypotheses that were rejected in error. Because V is a random variable, the FDP is unobservable. However, many statistical methods (see Ramdas et al. (2019)) have been developed to control the expected value of Q, known as the *false discovery rate* (FDR):

$$\mathrm{E}[Q] = \mathrm{E}\left[\frac{V}{R}\right]$$

In settings with very large M, controlling the FDR is seen as a much more practically viable policy than controlling the FWER. Controlling the FDR acknowledges that, for example, 2 Type I Errors in 10 tests may be unbearable, but 2 Type I Errors in 100 tests may be bearable. The FDR is therefore adaptive in the sense that the number of Type I Errors V has different implications depending on the size of M. Methods that control the FDR (the expected proportion of Type I Errors) are therefore less stringent than methods that control the FWER (the probability of at least one Type I Error). Though a variety of such methods exist, we discuss here only the most common: the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Benjamini-Hochberg Procedure:

The Benjamini-Hochberg (BH) procedure is a sequentially rejective procedure much like Holm's procedure discussed in the previous section. The principle difference is in the adjusted significance level that the ordered p-values are compared to at each step. The BH procedure, which aims to ensure $FDR \leq \alpha^\star$, may be summarized as follows:

1. Order the M p-values from smallest to largest:

$$p_{(1)}, p_{(2)}, \ldots, p_{(M)}$$

   where $p_{(k)}$ is the k^{th} smallest p-value.

2. Starting from $k = 1$ and continuing incrementally, compare $p_{(k)}$ to $\frac{k\alpha^\star}{M}$. Determine $k^\star$ the largest value of k such that

$$p_{(k)} \leq \frac{k\alpha^\star}{M}$$

3. Reject the null hypotheses $H_{0,(1)}, \ldots, H_{0,(k^\star)}$ and do not reject $H_{0,(k^\star+1)}, \ldots, H_{0,(M)}$.

The decision process associated with this procedure is best visualized with a plot of the ordered p-values $p_{(k)}$ vs. their ranks $k = 1, 2, \ldots, M$ with the significance threshold overlaid, like we did for the FWER-controlling methods previously. For the Benjamini-Hochberg procedure the significance threshold is the line with intercept 0 and slope $\alpha^\star/M$. This threshold is used as follows: moving from left to right, locate the *first* p-value which falls *above* the line: this is $p_{(k^\star+1)}$. The hypotheses corresponding to p-values smaller than this one are rejected and all others are not rejected. This decision boundary, for the four-test example discussed in Section 3.3.1, is visualized by the purple line in Figure 10. Since all of the p-values fall below this line, all four of the null hypotheses are rejected.

It is clear to see that this threshold is much less strict than any of the FWER-control thresholds, but this is the appeal of the approach. More Type I Errors may occur with this method than the others, but this is viewed as acceptable when M is very large. The proof that this procedure guarantees $FDR \leq \alpha^\star$ is outside the scope of this course, but the interested reader is referred to Benjamini and Hochberg (1995) and Storey et al. (2004).

Just like for the three methods in Section 3.3.1, the Benjamini-Hochberg procedure can be inverted and performed by comparing suitably adjusted p-values to $\alpha^\star$. Within this paradigm $H_{0,(k)}$ is rejected if $p^\star_{(k)} \leq \alpha^\star$ where

$$p^\star_{(k)} = \min_{j \geq k} \left\{ \frac{Mp_{(j)}}{j} \right\}$$

The BH-adjusted p-values for our four-test example are $p^\star_1 = 0.029, p^\star_2 = 0.029, p^\star_3 = 0.029, p^\star_4 = 0.029$ and since each of these is smaller than $\alpha^\star = 0.05$ we reject all four hypotheses, which is the same conclusion derived from Figure 10. This p-value adjustment may also be automated using R's `p.adjust()` function.

3.3.3 Sample Size Determination

So what does all of this mean for power analyses and sample size calculations?

In Sections 2.1.4 and 2.2.4 we showed that sample size formulae could be derived which accounted for the desired power and significance level of the test. However, this did not account for the multiple comparison problem. If, at the time of designing the experiment, you know that you intend to do M pairwise hypothesis tests in order to find a 'winning' condition, then the significance level you use in your sample size calculations should be adjusted to account for this.

This is easier to do with *some* correction methods than others. For instance, if applying the Bonferroni correction, one should use $\alpha^\star/M$ as the significance level in sample size calculations, if a family-wise error rate of $\alpha^\star$ is to be maintained. Or, if applying the Šidák correction, one should use $1 - (1 - \alpha^\star)^{\frac{1}{M}}$ as the significance level in these calculations. However, accounting for a correction by the Holm's or Benjamini-Hochberg "step-up" procedures is not so straightforward because different significance thresholds are used for different tests, depending on the size of a test's p-value – but the p-value will not be known during the design of the experiment prior to data collection. Therefore, if one wishes to use these methods of correction once the data are observed, some strong assumptions need to be made during the design phase.

In general, sample sizes that account for the multiple comparison problem are much larger than those that do not. This is not surprising; the adjusted significance levels associated with each of the correction methods described here are *smaller* than $\alpha^\star$. Thus, by plugging smaller significance levels into our sample size calculations, we are being more strict and hence will require larger sample sizes. This is a consequence of the duality between significance level and power. All else equal, reducing the significance level will increase the Type II Error rate and hence decrease power. Thus, all of the correction procedures discussed (which decrease the effective significance level) negatively impact power. Then, in order to maintain power at some pre-specified level, we must compensate by increasing the sample size. Therefore, the more complicated your experiment (i.e., the more conditions it has), the larger your sample sizes will need to be.

4 BLOCKING

Recall that in the context of a designed experiment we categorize factors as either a *design* factor, an *allowed-to-vary* factor, or a *nuisance* factor. Design factors are those that we purposefully manipulate in an experiment to assess the resultant effect on the response variable. These define our experimental conditions. Allowed-to-vary factors are those that are not controlled in the experiment – they are either unknown to us, or known by uncontrollable. The effects of such variables are intended to be dealt with via the random assignment of units to conditions. Nuisance factors, on the other hand, are factors that are expected to influence the response, but we do not want to quantify these effects, we simply want to eliminate them.

It is important to remember that in practice, context dictates whether a (known) factor should be considered a design factor, a nuisance factor, or if it should be allowed to vary. For instance, *browser* (i.e., Google Chrome vs. Microsoft Edge vs. Firefox vs. Safari) might be considered a design factor in one experiment, a nuisance factor in another, and an allowed-to-vary factor in yet another. To make this clear, consider the following three scenarios:

Figure 11: Four levels of the *browser* factor.

Experiment 1: Usability testing involves studying the ease with which an individual uses a product or service for some intended purpose. Suppose investigators are performing a usability test to determine with which browser 70-80 year old users find it easiest to look-up the phone number of the nearest pharmacy. In this example, experimental units (70-80 year olds) are randomly assigned to one of four browser conditions, and the time it takes them to complete the task is measured. Clearly, browser is a *design* factor here.

Experiment 2: Suppose that Netflix is experimenting with server-side modifications to improve (reduce) the latency of Netflix.com. Suppose that the current infrastructure serves as a control condition and the modified infrastructure is hypothesized to reduce median page load time. It is possible that a user's browser may also effect page load time, but this effect is not of interest to the investigators. To control for the potential impact of one's browser, Netflix initially experiments with only Firefox users. Clearly, browser is a *nuisance* factor here.

Experiment 3: Suppose that Amazon.ca is experimenting with the width of their search bar. They hypothesize that a wider search bar will minimize the amount of mouse movement required to navigate to it, thereby minimizing the average time-to-query. The experimenters do not care which browser a customer uses and so this factor is not controlled and hence is *allowed-to-vary* during their experiment.

The previous three examples illustrate that a factor's categorization is context-dependent and may change from one experiment to another. In the context of a single experiment however, it is also important to understand the subtle distinction between nuisance factors and design factors. Both factors are controlled during the experiment, but with the design factor we wish to quantify its influence on the response variable. With a nuisance factor, however, we do not care to quantify its effect, we wish only to *eliminate* it. As was discussed in Chapter 1, we do so with *blocking*. Broadly speaking, blocking refers to the principle in which the effect of one or more design factors is investigated, while holding one or more nuisance factors fixed. We refer to the fixed levels of a nuisance factor as *blocks*. By holding a nuisance factor fixed, it cannot cannot vary and hence cannot influence the response. This is how the *browser* factor was handled in Experiment 2 above.

In this chapter we discuss several strategies for the design and analysis of experiments that use blocking as a means to eliminate the influence of one or more nuisance factors.

4.1 Randomized Complete Block Designs

The randomized complete block design (RCBD) is a simple experimental design that may be applied when we wish to investigate a single design factor while at the same time controlling for a single nuisance factor. In such a design, each of the experimental conditions is carried out within every one of the blocks. If the design factor has m levels (i.e., the experiment has m conditions) and the nuisance factor has b levels (i.e., the experiment has b blocks), the RCBD runs all m conditions inside all b blocks. The *observed* data in such an experiment may be denoted by y_{ijk}, which represents the response observation for experimental unit $i = 1, \ldots, n_{jk}$ in condition $j = 1, \ldots, m$ in block $k = 1, \ldots, b$. Note that we assume that there are n_{jk} units in (condition, block) = $(j,\, k)$ and thus an overall total of $N = \sum_{k=1}^{b} \sum_{j=1}^{m} n_{jk}$ units. Response data of this form may be tabulated as shown below.

Table 7: Response observations in a randomized complete block design

		Block				
		1	2	$\cdots$	b	
Condition	1	$\{y_{i11}\}_{i=1}^{n_{11}}$	$\{y_{i12}\}_{i=1}^{n_{12}}$	$\cdots$	$\{y_{i1b}\}_{i=1}^{n_{1b}}$	$\overline{y}_{\cdot 1 \cdot}$
	2	$\{y_{i21}\}_{i=1}^{n_{21}}$	$\{y_{i22}\}_{i=1}^{n_{22}}$	$\cdots$	$\{y_{2b}\}_{i=1}^{n_{2b}}$	$\overline{y}_{\cdot 2 \cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	m	$\{y_{im1}\}_{i=1}^{n_{m1}}$	$\{y_{im2}\}_{i=1}^{n_{m2}}$	$\cdots$	$\{y_{mb}\}_{i=1}^{n_{mb}}$	$\overline{y}_{\cdot m \cdot}$
		$\overline{y}_{\cdot\cdot 1}$	$\overline{y}_{\cdot\cdot 2}$	$\cdots$	$\overline{y}_{\cdot\cdot b}$	$\overline{y}_{\cdot\cdot\cdot}$

Note that if every cell contained n units, the design would be called "balanced" and $N = m \times b \times n$. Also note that the term *complete* here refers to the fact that *every* condition is represented within *every* block. That is, $n_{jk} > 0 \ \forall j, k$. Experimental designs that do not have this property are referred to as *incomplete* block designs. We discuss one such design in Section 4.2.

In Table 7 the row means $\overline{y}_{\cdot j \cdot}$ represent condition-specific averages of the response observations, and the column means $\overline{y}_{\cdot\cdot k}$ represent block-specific averages. The overall mean is denoted by $\overline{y}_{\cdots}$. These three means are calculated as follows:

$$\overline{y}_{\cdot j \cdot} = \frac{1}{b}\sum_{k=1}^{b}\overline{y}_{\cdot jk} \qquad \overline{y}_{\cdot\cdot k} = \frac{1}{m}\sum_{j=1}^{m}\overline{y}_{\cdot jk} \qquad \overline{y}_{\cdots} = \frac{1}{N}\sum_{k=1}^{b}\sum_{j=1}^{m}\sum_{i=1}^{n_{jk}} y_{ijk} = \frac{1}{N}\sum_{k=1}^{b}\sum_{j=1}^{m} n_{jk}\overline{y}_{\cdot jk}$$

where $\overline{y}_{\cdot jk}$ is the average of the response observations in cell (j, k). Simple summaries such as these provide a crude assessment of whether the condition-to-condition and block-to-block variation is large. If we were correct to include the nuisance factor we would expect there to be variation in the block-specific averages. And if there is truly a difference in expected response from one condition to another we would expect to see variation in the condition-specific averages.

The primary analysis goal in a RCBD is to determine whether the expected response differs significantly from one condition to another, and if so, to identify the optimal condition – all while controlling for the potential effect of the nuisance factor. Previously such a goal was achieved by testing a gatekeeper test such as

$$H_0\colon\ \theta_1 = \theta_2 = \cdots = \theta_m \text{ vs. } H_A\colon\ \theta_j \neq \theta_{j'} \text{ for some } j' \neq j$$

The same will be true here, except we must now account for the presence of the nuisance factor in the experiment. To formally analyze data arising from a RCBD and hence address our primary goal, we use an *appropriately defined* linear or logistic regression – the type of regression depends on whether the response variable is continuous or binary. In both cases, the structure of the *linear predictor* is identical; it contains:

- an intercept,
- $m - 1$ indicator variables for the design factor's levels, and
- $b - 1$ indicator variables for the nuisance factor's levels

This linear predictor may be written as

$$\alpha + \sum_{j=1}^{m-1}\beta_j x_{ij} + \sum_{k=1}^{b-1}\gamma_k z_{ik} \tag{24}$$

where $x_{ij} = 1$ indicates that unit i is in condition $j = 1, 2, \ldots, m-1$ and $z_{ik} = 1$ indicates that unit i is in block $k = 1, 2, \ldots, b-1$. With this notation we arbitrarily treat the m^{th} condition and the b^{th} block as references against which the other conditions and blocks will be compared. As is clear from the linear predictor above, the β's jointly quantify the effect of the design factor and the γ's jointly quantify the effect of the nuisance factor. Note that the effect of the design factor is called *additive* because it does not depend on the level of the nuisance factor.

Two relevant hypotheses that we may wish to test in the context of such a model are shown below.

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_{m-1} = 0 \text{ vs. } H_A\colon\ \beta_j \neq 0 \text{ for some } j \tag{25}$$

$$H_0\colon\ \gamma_1 = \gamma_2 = \cdots = \gamma_{b-1} = 0 \text{ vs. } H_A\colon\ \gamma_k \neq 0 \text{ for some } k \tag{26}$$

Hypothesis (25) is a test of overall equality of the condition-specific expectations and may be used to evaluate whether or not the design factor significantly influences the response (controlling for the influence of the nuisance factor). This is the main hypothesis of interest here, as it helps to achieve our primary analysis goal. Hypothesis (26) is of secondary interest, but important nonetheless. If H_0 cannot be rejected then it suggests the nuisance factor does not significantly influence the response, and hence that blocking was unnecessary. As such, this test may be used to confirm that the nuisance factor does indeed influence the response and that blocking was necessary.

Hypotheses such as (25) and (26) are tested by comparing a *full* model, with linear predictor given by equation (24), to a *reduced* model that arises when H_0 is true. Evidence is sought to determine whether the full model fits the data significantly better than the reduced one. The specific test used (i.e., the specific test statistic and null distribution) depends on the form the regression; in the remaining subsections we discuss the use of partial F-tests and likelihood ratio tests in the case of linear and logistic regressions, respectively.

4.1.1 RCBD to Compare Means

Here interest lies in testing the following hypothesis, all while accounting for the influence of the nuisance factor:

$$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_m \text{ vs. } H_A\colon\ \mu_j \neq \mu_{j'} \text{ for some } j' \neq j$$

As discussed above, this hypothesis is equivalent to hypothesis (25) in the context of the following linear regression model

$$Y_i = \alpha + \sum_{j=1}^{m-1} \beta_j x_{ij} + \sum_{k=1}^{b-1} \gamma_k z_{ik} + \varepsilon_i$$

Here Y_i is the response observation for unit $i = 1, 2, \ldots, N = \sum_{k=1}^{b} \sum_{j=1}^{m} n_{jk}$ and ε_i is the corresponding random error term assumed to follow a $\mathrm{N}(0, \sigma^2)$ distribution. The hypothesis above is tested with a two-way analysis of variance[2]. The test may be motivated either from a regression framework and the additional sum of squares principle, or from an *effects model* framework and the partitioning of the total response variation into variation due to conditions, variation due to blocks and variation due to error. Both approaches yield the same test and the same ANOVA table, which is shown below.

The sums of squares given in the table are:

[2] This is called a *two-way* ANOVA because we are dealing with two factors. This should be contrasted with the *one-way* ANOVA discussed in Section 3.1.1 which involved just one factor.

Table 8: Two-way ANOVA table associated with a randomized complete block design

Source	SS	*df*	MS	Test Stat.
Condition	SS_C	$m-1$	$MS_C = \frac{SS_C}{m-1}$	$\frac{MS_C}{MS_E}$
Block	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_E}$
Error	SS_E	$N-m-b+1$	$MS_E = \frac{SS_E}{N-m-b+1}$	
Total	SS_T	$N-1$		

- The total sum of squares, which quantifies overall variation in the response observations:

$$SS_T = \sum_{k=1}^{b}\sum_{j=1}^{m}\sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y}_{...})^2$$

- The condition sum of squares, which quantifies condition-to-condition variability in the response observations:

$$SS_C = \sum_{k=1}^{b}\sum_{j=1}^{m}\sum_{i=1}^{n_{jk}} (\overline{y}_{.j.} - \overline{y}_{...})^2$$

- The block sum of squares, which quantifies block-to-block variability in the response observations:

$$SS_B = \sum_{k=1}^{b}\sum_{j=1}^{m}\sum_{i=1}^{n_{jk}} (\overline{y}_{..k} - \overline{y}_{...})^2$$

- The error sum of squares, which quantifies variability unaccounted for by the conditions and blocks:

$$SS_E = \sum_{k=1}^{b}\sum_{j=1}^{m}\sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y}_{.j.} - \overline{y}_{..k} + \overline{y}_{...})^2$$

It is an easy but tedious calculation to show that $SS_T = SS_C + SS_B + SS_E$. It can also be shown that if H_0 in hypothesis (25) is true then $\mathrm{E}[MS_C] = \mathrm{E}[MS_E] = \sigma^2$, and if H_0 in hypothesis (26) is true then $\mathrm{E}[MS_B] = \mathrm{E}[MS_E] = \sigma^2$. This provides intuition for form of the test statistics $t_C \equiv \frac{MS_C}{MS_E}$ and $t_B \equiv \frac{MS_B}{MS_E}$; if t_C and t_B are much larger than 1, this provides evidence against H_0 in hypotheses (25) and (26), respectively. Conclusions are formally drawn on the basis of a comparison of t_C to the $F_{(m-1,N-m-b+1)}$ distribution and t_B to the $F_{(b-1,N-m-b+1)}$ distribution where p-values are defined as the right-tail probabilities

$$\Pr(F_{(m-1,N-m-b+1)} \geq t_C) \qquad \text{and} \qquad \Pr(F_{(b-1,N-m-b+1)} \geq t_B)$$

Although the sums of squares are simple enough to calculate by hand, this analysis is typically automated using the `lm` and `anova` functions in R, as we will see in the next subsection.

4.1.2 Example: Promotions at The Gap

The Gap has three versions of an online weekday promotion that a customer sees when they go to gapcanada.ca:

- Version 1: 50% discount on 1 item

- Version 2: 20% discount on your entire order
- Version 3: Spend \$50 and get a \$10 gift card

Interests lies in determining whether there is a difference in the average purchase total (i.e, the average dollar value of a customer's purchase) between promotion versions. However, the amount of money one spends may also be influenced by the nuisance factor, day of week. As such, a randomized complete block design was run with $m = 3$ experimental conditions (corresponding to the three promotions) and $b = 5$ blocks (corresponding to the day of the week). Here $n_{jk} = 50\ \forall j, k$, and so the design was "balanced". For each visitor to gapcanada.ca, their purchase total (in dollars) was recorded. The regression model fit to these response observations is

$$Y_i = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \gamma_4 z_{i4} + \varepsilon_i$$

where x_{i2} and x_{i3} are condition indicators for promotions 2 and 3 (promotion 1 is the baseline) and $z_{i1}, \ldots, z_{i4}$ are block indicators for Monday-Thursday (Friday is the baseline). The ANOVA table obtained from the observed data is shown in Table 9.

Table 9: The Gap RCBD ANOVA table

Source	*SS*	*df*	*MS*	Test Stat.
Condition	49618.34	2	24809.17	2165.39
Block	19258.30	4	4814.58	420.22
Error	8512.67	743	11.46	
Total	77389.32	749		

To determine whether there is a difference in average purchase total from one promotion-version to another we test

$$H_0 : \beta_2 = \beta_3 = 0$$

and we do so using the test statistic $t_C = 2165.39$. The p-value associated with this test is $P(T \geq 2165.39) = 1.101 \times 10^{-310}$, where $T \sim F_{(2,743)}$. This incredibly small p-value leads us to reject the null hypothesis and conclude that the expected purchase total in each condition is *not* the same. Follow-up pairwise tests may be used to determine *which* promotion maximizes average purchase total, but in lieu of this the plots in Figure 12 suggest that Promotion 2 (20% discount on your entire order) is best.

To confirm that blocking was a good thing to do, we could formally test

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0$$

and hence whether the z's (i.e., the nuisance factor) really needed to be in the model. The test statistic for this test is $t_B = 420.22$ with corresponding p-value $P(T \geq 420.22) = 4.345 \times 10^{-189}$, where $T \sim F_{(4,743)}$. This extraordinarily small p-value, and the plots in Figure 13 suggest that day-to-day differences are significant, and thus it was appropriate to account for them via blocking.

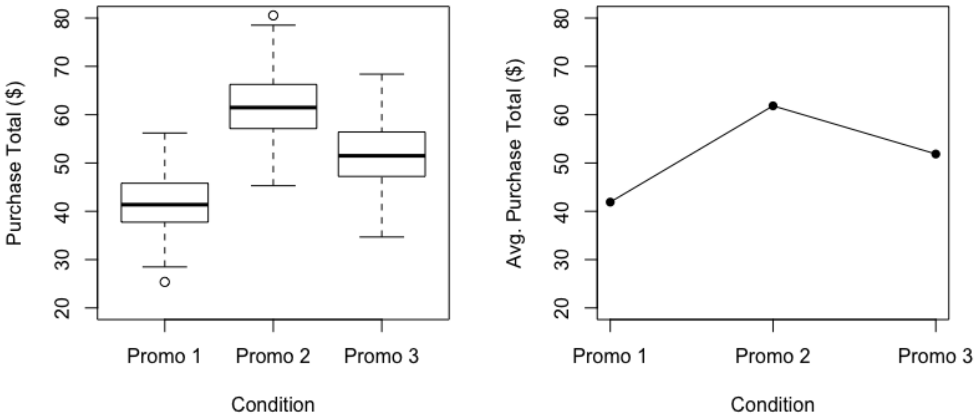

Figure 12: Visualization of promo-to-promo differences in The Gap experiment

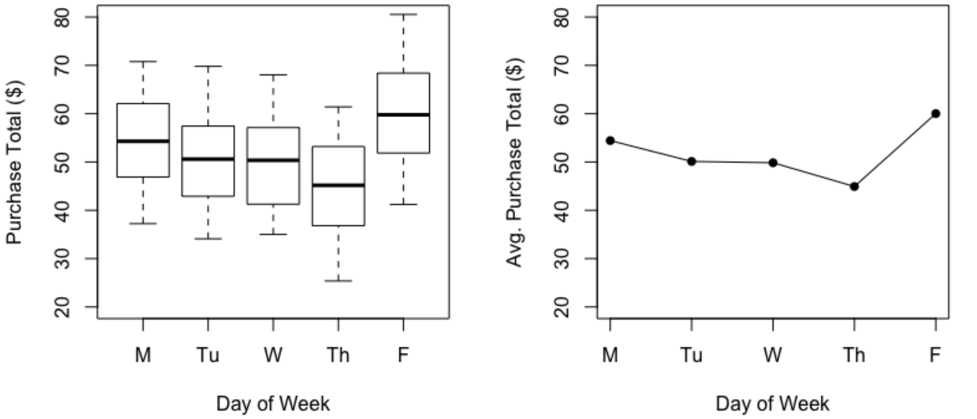

Figure 13: Visualization of day-to-day differences in The Gap experiment

4.1.3 RCBD to Compare Proportions

When our metric of interest is a probability/proportion our response variable is binary, in which case logistic regression is the appropriate analysis method. Here interest lies in testing the following hypothesis, all while accounting for the influence of the nuisance factor:

$$H_0\colon \pi_1 = \pi_2 = \cdots = \pi_m \text{ vs. } H_A\colon \pi_j \neq \pi_{j'} \text{ for some } j' \neq j$$

As discussed previously, this hypothesis is equivalent to hypothesis (25) in the context of the following logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{j=1}^{m-1} \beta_j x_{ij} + \sum_{k=1}^{b-1} \gamma_k z_{ik}$$

Here Y_i is the binary response observation for unit $i = 1, 2, \ldots, N = \sum_{k=1}^{b} \sum_{j=1}^{m} n_{jk}$ and $\pi_i = \text{E}[Y_i]$ is the corresponding expectation (i.e., the probaility that unit i performs some action of interest). Within this modeling framework, hypothesis (25) is tested with a likelihood ratio test (LRT) that compares the full model (shown above) to the reduced one without the x's. Similarly, hypothesis (26) is tested with a LRT that compares the full model to the reduced one without the z's. The observed test statistic for these tests is

$$\begin{aligned} t &= 2 \times \log\left(\frac{\text{Likelihood}_{\text{Full Model}}}{\text{Likelihood}_{\text{Reduced Model}}}\right) \\ &= 2 \times [\text{Log-Likelihood}_{\text{Full Model}} - \text{Log-Likelihood}_{\text{Reduced Model}}] \end{aligned}$$

which, if the null hypothesis is true, should follow a $\chi^2_{(\nu)}$ distribution, where

$$\nu = (\text{\# of coefficients in full model}) - (\text{\# of coefficients in reduced model})$$

The corresponding p-value is a right-tailed probability in the $\chi^2_{(\nu)}$ distribution: $\Pr(T \geq t)$ where $T \sim \chi^2_{(\nu)}$

The calculation of the likelihood values cannot be done in closed form, and so the analyses described above must be computed in R. We use the `glm` and `lrtest` functions for this task.

4.1.4 Example: Enterprise Banner Ads

Enterprise is experimenting with $m = 3$ banner ads as a mechanism to drive traffic to their website. Since there are known regional differences in consumer preferences in the US, they wish to control for the nuisance factor "region" with $b = 4$ blocks corresponding to the four major US geographic regions: Northeast (NE), Northwest (NW), Southeast (SE), and Southwest (SW). A total of $n_{jk} = 5000\ \forall j, k$ people were randomized to each ad condition in each region.

Interest lies in determining whether or not the different ads perform similarly with respect to click-through-rate (CTR) – and we wish to determine which one maximizes CTR – but we want to control for the effect of region. We do so with the following logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3}$$

where x_{i2} and x_{i3} are condition indicators for ads 2 and 3 (ad 1 is the baseline) and z_{i1}, z_{i2}, z_{i3} are block indicators for the NW, SE, SW regions (NE is the baseline).

To determine whether there is a difference in click-through-rate (CTR) from one ad to another we test

$$H_0 : \beta_2 = \beta_3 = 0$$

and we do so using the test statistic $t_C = 249.92$ (which was calculated using the `lrtest` function in R). The p-value associated with this test is $P(T \geq 249.92) = 5.367 \times 10^{-55}$, where $T \sim \chi^2_{(2)}$. This incredibly small p-value leads us to reject the null hypothesis and conclude that the CTR in each condition is *not* the same. Follow-up pairwise tests may be used to determine *which* ad maximizes CTR, but in lieu of this the left plot in Figure 14 suggest that Ad 1 is best.

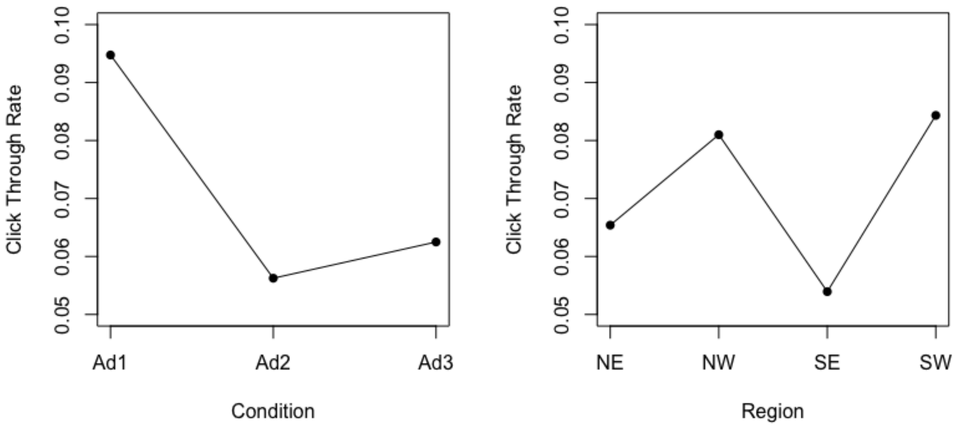

Figure 14: Left: Visualization of ad-to-ad differences in CTR in the Enterprise experiment Right: Visualization of region-to-region differences in CTR in the Enterprise experiment

To confirm that blocking was a good thing to do, we could formally test

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$$

and hence whether the z's (i.e., the nuisance factor) really needed to be in the model. The test statistic for this test is $t_B = 139.82$ (calculated using `lrtest`) with corresponding p-value $P(T \geq 139.82) = 4.126 \times 10^{-30}$, where $T \sim \chi^2_{(3)}$. This extraordinarily small p-value, and the right plot in Figure 14 suggest that region-to-region differences are significant, and thus it was appropriate to account for them via blocking.

4.2 Balanced Incomplete Block Designs

In Section 4.1 we discussed the randomized complete block design in which *every* experimental condition was carried out inside *every* block. However, in practice there are situations in which not *every* condition can be carried out inside *every* block. Consider again The Gap experiment from Section 4.1.2 in which $m = 3$ different promotional offers are being investigated:

- Version 1: 50% discount on 1 item
- Version 2: 20% discount on your entire order
- Version 3: Spend \$50 and get a \$10 gift card

As before, the experimenters would like to control for a possible day-of-week effect and so they want to block by day. However, now suppose that due to logistical concerns only two of the three promotions may be offered in a single day. Thus a RCBD is not possible here. In general, when practical constraints will not allow all experimental conditions to be run within each block, we require an *incomplete* block design (IBD). Consider the situation in which The Gap experimenters wish to define $b = 6$ blocks, and hence run the experiment across 6 different days. Table 10 illustrates both the RCBD that would have been ideal, and an IBD that may be performed as an alternative. Clearly, data is observed for every block-condition combination in the RCBD, but this is not the case for the IBD. This illustrates the hallmark of *complete* versus *incomplete* block designs.

Table 10: Complete (left) vs. incomplete (right) block designs. The ✓and × symbols indicate for which block-condition combinations data is observed.

		Block					
		1	2	3	4	5	6
Condition	1	✓	✓	✓	✓	✓	✓
	2	✓	✓	✓	✓	✓	✓
	3	✓	✓	✓	✓	✓	✓

		Block					
		1	2	3	4	5	6
Condition	1	✓	✓	×	×	✓	✓
	2	×	×	✓	✓	✓	✓
	3	✓	✓	✓	✓	×	×

The most common IBD is the *balanced incomplete block design* (BIBD). Like a RCBD, the BIBD facilitates the investigation of m experimental conditions in b blocks. However, unlike a RCBD, in the BIBD only $m^\star < m$ conditions may be run within each block. The number $m^\star$ is commonly referred to as the *block size*. Further, $r < b$ is the number of blocks that each experimental condition appears in, and λ is the number of blocks that any pair of conditions appear in together. To better understand this notation, consider the incomplete block design visualized in the righthand table of Table 10. This is in fact a BIBD. The numbers $(m, b, m^\star, r, \lambda)$ for this design are as follows:

- There are $m = 3$ experimental conditions (promotions).
- There are $b = 6$ blocks (days).
- Only $m^\star = 2$ conditions can be run in each block (since only two promotions can be run each day).
- Each condition appears in $r = 4$ blocks (i.e., each promotion is experimented with on 4 separate days). To very this, count the number of ✓'s in each row.
- Any given pair of conditions appear together in $\lambda = 2$ blocks (i.e., the number of days that any pair of promotions are experimented with together is 2).

The term "balanced" in the BIBD refers to the fact that the number of conditons in each block ($m^\star$) is the same for every block; the number of blocks each condition appears in (r) is the same for every condition; the number of blocks each pair of conditions appear in together (λ) is the same for every possible condition pairing. Despite the incompleteness, this balance allows for the comparison of a metric of interest across m conditions while still accounting for a nuisance factor with b levels. However, as we see in the next subsection, there are tradeoffs that must be made.

4.2.1 General Comments on the Design of a BIBD

It is important to recognize that great care must go into planning a balanced incomplete block design to ensure all forms of balance. In particular, there are a variety of restrictions that must be accounted for; not just any haphazard combination of $(m, b, m^\star, r, \lambda)$ values will yield a BIBD. One such restriction is that

$$mr = bm^\star$$

This is easy to justify as it is the number of block-condition combinations for which data is observed, counted in two different ways. Likewise,

$$r(m^\star - 1) = \lambda(m - 1)$$

is based on two different ways of counting the number of times (for a given condition) that the other conditions are paired with it across all of the blocks.

One way of using these equations is to specify m, $m^\star$, and λ based on the context and the number of times you wish each pair of conditions to coexist in the same block, and then use

$$r = \frac{\lambda(m-1)}{m^\star - 1} \quad \text{and then} \quad b = \frac{mr}{m^\star}$$

to determine the number of blocks required, noting that all of the numbers in these equations must be integers.

For instance, in The Gap BIBD we had $m = 3$, $m^\star = 2$, and $\lambda = 2$. Substituting these values into the equations above yield $r = 4$ and $b = 6$, indicating the need for 6 blocks (which is what was done). Now suppose that instead we wished for each pair of conditions to coexist in the same block exactly once ($\lambda = 1$). In that case, $r = (1 \times 2)/1 = 2$ and $b = (3 \times 2)/2 = 3$, indicating that three blocks would be required. And if instead we took $\lambda = 3$, then $b = 9$. Clearly, given these constraints, no BIBDs exist which investigate $m = 3$ conditions with $b = 1, 2, 4, 5, 7,$ or 8 blocks of size $m^\star = 2$.

Which design is selected is based on a trade-off between larger λ values and smaller b values. Larger λ values are to be preferred because it increases the amount of information available when conducting pairwise comparisons of the conditions. Smaller b values are to be preferred because the larger the number of the blocks, the larger (and hence more expensive) the experiment.

4.2.2 General Comments on the Analysis of a BIBD

In this section we (very) briefly, and rather informally, discuss the analysis of a balanced incomplete block design. For a more detailed treatment of this topic, the interested reader is referred to Section 3.8 of Wu and Hamada (2011) or Section 4.4 of Montgomery (2019).

As with a randomized complete block design, we wish to determine whether their exist significant differences among the expected response values from one experimental condition to another. In a RCBD we do this by comparing the condition-specific means $\overline{y}_{\cdot j \cdot}$ to the overall mean $\overline{y}_{\cdots}$. However, in a BIBD we need to be careful making such a comparison. Because condition j did not appear in every block, a fairer comparison would be to compare $\overline{y}_{\cdot j \cdot}$ with the average response from the blocks that condition j *did* appear in:

$$\frac{1}{r}\sum_{k=1}^{b}\overline{y}_{\cdot\cdot k}\mathbf{1}[\text{condition } j \text{ appears in block } k]$$

In general, the analysis of BIBDs involves an *adjustment* of this form when evaluating the effect of the design factor.

4.3 Latin Square Designs

In Section 4.1 we considered the design of an experiment that investigates *one* design factor while controlling for *one* nuisance factor. Here we discuss *Latin square designs* which may be used to experiment with *one* design factor while simultaneousy controlling for *two* nuisance factors. Note that if one wishes to control for the effects of three or four nuisance factors, Graeco-Latin and Hyper-Graeco-Latin square designs may be employed, but we will not discuss those extensions here. The interested reader is referred to Section 3.7 of Wu and Hamada (2011) or Section 4.3 of Montgomery (2019) for more details.

In combinatorics, a Latin square of order p is a $p \times p$ grid containing p unique symbols with the property that each of these symbols occurs exactly once in each column and exactly once in each row. The term *Latin* is used to describe such a square because Latin letters (i.e., A, B, C, D, ...) are typically used as the symbols. Examples of 3×3, 4×4, and 5×5 Latin squares are shown in Table 11. Note that if you have ever played a Sudoku puzzle, you have encountered a Latin square of order $p = 9$ with "symbols" $\{1, 2, \ldots, 9\}$[3].

Table 11: 3×3, 4×4, and 5×5 Latin Square Examples

A	C	B
C	B	A
B	A	C

A	B	C	D
C	D	A	B
B	C	D	A
D	A	B	C

A	B	C	D	E
E	A	B	C	D
D	E	A	B	C
C	D	E	A	B
B	C	D	E	A

In the field of experimental design we exploit this combinatorial structure to help us design experiments that facilitate blocking by two nuisance factors. In particular, we randomly associate the p rows with the levels of the first nuisance factor, we randomly associate the p columns with the levels of the second nuisance factor, and we randomly associate the p Latin letters with the levels of the design factor. A visualization of a Latin square design with $p = 4$ is shown in Table 12. Each cell in this table represents a "block" in which the nuisance factors' levels are held fixed, and the Latin letter indicates which experimental condition is being executed. For instance, the $(3, 2)$ cell of Table 12 defines a block where nuisance factors 1 and 2 are

[3] Note: a Sudoku puzzle has the added constraint that the 9×9 grid be broken up into nine 3×3 grids which each contain the numbers $\{1, 2, \ldots, 9\}$

respectively held fixed at their 3rd and 2nd levels and within which we run experimental condition D, which is the condition defined by the design factor's 4th level.

Table 12: 4×4 Latin square design

		Nuisance Factor 2			
		1	2	3	4
Nuisance Factor 1	1	A	B	C	D
	2	D	A	B	C
	3	C	D	A	B
	4	B	C	D	A

An obvious limitation to such a design is that our design and nuisance factors must each be experimented with at p levels; if the factors do not all have p levels that you wish to experiment with, a Latin square design is not appropriate. However, when it is reasonable to experiment with these factors, each at p levels, a Latin square design allows for the comparison of a metric of interest across all p experimental conditions (defined by the p levels of the design factor) while accounting for the influence of the two nuisance factors. Due to the structure of the Latin square, the rows, columns and letters are all orthogonal, allowing us to separately estimate the effects of the design factor and each of the two nuisance factors.

In such an experiment we let y_{ijkl} represent the response observation for unit $i = 1, 2, \ldots, n$ in condition $j = 1, 2, \ldots, p$, within block (k, l), $k = 1, 2, \ldots, p$, $l = 1, 2, \ldots, p$. For notational convenience only, we assume the experiment is balanced and that there are n experimental units assigned to each block. We remark that because each block contains just one condition, each pair (k, l) uniquely determines the value of j. Consequently, there exist just p^2 tuples (j, k, l), and we denote them by the set $\mathcal{S}$. The 16 elements of $\mathcal{S}$ corresponding to the design in Table 12 are shown in Table 13.

Table 13: Index set $\mathcal{S}$ for the 4×4 Latin square design shown in Table 12. Each entry has the form (j, k, l)

(1,1,1)	(2,1,2)	(3,1,3)	(4,1,4)
(4,2,1)	(1,2,2)	(2,2,3)	(3,2,4)
(3,3,1)	(4,3,2)	(1,3,3)	(2,3,4)
(2,4,1)	(3,4,2)	(4,4,3)	(1,4,4)

It will also be useful to define $\mathcal{S}_j$, $\mathcal{S}_k$ and $\mathcal{S}_l$ which are subsets of $\mathcal{S}$ and which respectively contain all tuples for which the design factor level is j, nuisance factor 1's level is k, and nuisance factor 2's level is l. Each of these subsets contain exactly p tuples. For instance, $\mathcal{S}_{l=3}$ contains the tuples shown in column 3 of Table 13, $\mathcal{S}_{k=1}$ contains the tuples shown in row 1 of Table 13, and for the same example $\mathcal{S}_{j=2} = \{(2, 1, 2), (2, 2, 3), (2, 3, 4), (2, 4, 1)\}$.

These notational subtleties are important to be aware of when, for example, computing average response values. For instance, the overall mean is given by

$$\overline{y}_{\cdots\cdot} = \frac{1}{N} \sum_{(j,k,l) \in \mathcal{S}} \sum_{i=1}^{n} y_{ijkl}$$

where $N = np^2$. The level-specific means of the design and two nuisance factors are respectively given by:

$$\overline{y}_{\cdot j \cdot\cdot} = \frac{1}{np} \sum_{(j,k,l)\in\mathcal{S}_j} \sum_{i=1}^{n} y_{ijkl} \qquad \overline{y}_{\cdot\cdot k \cdot} = \frac{1}{np} \sum_{(j,k,l)\in\mathcal{S}_k} \sum_{i=1}^{n} y_{ijkl} \qquad \overline{y}_{\cdot\cdot\cdot l} = \frac{1}{np} \sum_{(j,k,l)\in\mathcal{S}_l} \sum_{i=1}^{n} y_{ijkl}$$

Comparison of these averages to the overall average $\overline{y}_{\cdot\cdot\cdot\cdot}$ provides insight into the potential heterogeneity among response values observed at the different levels of each factor.

The primary analysis goal in a Latin square design is to determine whether the expected response differs significantly from one condition to another, and if so, to identify the optimal condition – all while controlling for the potential effect of the nuisance factors. In other words, we wish to test hypotheses such as

$$H_0\colon\ \theta_1 = \theta_2 = \cdots = \theta_p \text{ vs. } H_A\colon\ \theta_j \neq \theta_{j'} \text{ for some } j' \neq j$$

while accounting for the presence of the nuisance factors in the experiment. To formally analyze data arising from a Latin square design, we use an *appropriately defined* linear or logistic regression – the type of regression depends on whether the response variable is continuous or binary. In both cases, the structure of the linear pedictor is identical; it contains:

- an intercept,
- $p - 1$ indicator variables for the design factor's levels, and
- $p - 1$ indicator variables for nuisance factor 1's levels, and
- $p - 1$ indicator variables for nuisance factor 2's levels

This linear predictor may be written as

$$\alpha + \sum_{j=1}^{p-1} \beta_j x_{ij} + \sum_{k=1}^{p-1} \gamma_j z_{ik} + \sum_{l=1}^{p-1} \delta_l w_{il} \tag{27}$$

where $x_{ij} = 1$ indicates that unit i is in condition $j = 1, 2, \ldots, p-1$, $z_{ik} = 1$ indicates that unit i is in a block with nuisance factor 1 at level $k = 1, 2, \ldots, p-1$, and $w_{il} = 1$ indicates that unit i is in a block with nuisance factor 2 at level $l = 1, 2, \ldots, p-1$. With this notation we arbitrarily treat the p^{th} level of each factor as references against which the other levels are compared. As is clear from the linear predictor above, the β's jointly quantify the effect of the design factor, the γ's jointly quantify the effect of nuisance factor 1, and the δ's jointly quantify the effect of nuisance factor 2. Just as in the linear predictor (24), the effect of the design factor is *additive*, because it is the same for any block.

Three relevant hypotheses that we may wish to test in the context of such a model are shown below.

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \text{ vs. } H_A\colon\ \beta_j \neq 0 \text{ for some } j \tag{28}$$

$$H_0\colon\ \gamma_1 = \gamma_2 = \cdots = \gamma_{p-1} = 0 \text{ vs. } H_A\colon\ \gamma_k \neq 0 \text{ for some } k \tag{29}$$

$$H_0\colon\ \delta_1 = \delta_2 = \cdots = \delta_{p-1} = 0 \text{ vs. } H_A\colon\ \delta_l \neq 0 \text{ for some } l \tag{30}$$

Hypothesis (28) is a test of overall equality of the condition-specific expectations and may be used to evaluate whether or not the design factor significantly influences the response (controlling for the influence of the nuisance factors). This is the main hypothesis of interest here, as it helps to achieve our primary analysis goal. Hypotheses (29) and (30) are of secondary interest, but important nonetheless. In either of them, if H_0 cannot be rejected then it suggests the nuisance factor does not significantly influence the response, and hence that blocking by this factor was unnecessary. As such, these tests may be used to confirm that the nuisance factors do indeed influence the response and that blocking was necessary.

Hypotheses such as (28), (29) and (30) are tested by comparing a *full* model, with linear predictor given by equation (27), to a *reduced* model that arises when H_0 is true. Evidence is sought to determine whether the full model fits the data significantly better than the reduced one. The specific test used (i.e., the specific test statistic and null distribution) depends on the form the regression; in the remaining subsections we discuss the use of partial F-tests and likelihood ratio tests in the case of linear and logistic regressions, respectively.

4.3.1 Latin Squares to Compare Means

Here interest lies in testing the following hypothesis, all while accounting for the influence of the two nuisance factors:

$$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_p \text{ vs. } H_A\colon\ \mu_j \neq \mu_{j'} \text{ for some } j' \neq j$$

As discussed above, this hypothesis is equivalent to hypothesis (28) in the context of the following linear regression model

$$Y_i = \alpha + \sum_{j=1}^{p-1} \beta_j x_{ij} + \sum_{k=1}^{p-1} \gamma_k z_{ik} + \sum_{l=1}^{p-1} \delta_l w_{il} + \varepsilon_i$$

Here Y_i is the response observation for unit $i = 1, 2, \ldots, N = np^2$ and ε_i is the corresponding random error term assumed to follow a $\mathrm{N}(0, \sigma^2)$ distribution. The hypothesis above is tested with a three-way analysis of variance. The test may be motivated either from a regression framework and the additional sum of squares principle, or from an *effects model* framework and the partitioning of the total response variation into variation due to the design factor, variation due to the nuisance factors and variation due to error. Both approaches yield the same test and the same ANOVA table, which is shown in Table 14.

The sums of squares given in the table are:

- The total sum of squares, which quantifies overall variation in the response observations:

$$SS_T = \sum_{(j,k,l)\in\mathcal{S}} \sum_{i=1}^{n} (y_{ijkl} - \overline{y}_{....})^2$$

Table 14: Three-way ANOVA table associated with a Latin square design

Source	SS	df	MS	Test Stat.
Design Factor	SS_C	$p-1$	$MS_C = \frac{SS_C}{p-1}$	$\frac{MS_C}{MS_E}$
Nuisance Factor 1	SS_{B_1}	$p-1$	$MS_{B_1} = \frac{SS_{B_1}}{p-1}$	$\frac{MS_{B_1}}{MS_E}$
Nuisance Factor 2	SS_{B_2}	$p-1$	$MS_{B_2} = \frac{SS_{B_2}}{p-1}$	$\frac{MS_{B_2}}{MS_E}$
Error	SS_E	$N-3p+2$	$MS_E = \frac{SS_E}{N-3p+2}$	
Total	SS_T	$N-1$		

- The condition sum of squares, which quantifies variability in the response observations between different levels of the design factor:

$$SS_C = \sum_{(j,k,l)\in\mathcal{S}} \sum_{i=1}^{n} (\overline{y}_{.j..} - \overline{y}_{....})^2 = np\sum_{j=1}^{p} (\overline{y}_{.j..} - \overline{y}_{....})^2$$

- The first block-to-block sum of squares, which quantifies variability in the response observations between the different levels of nuisance factor 1:

$$SS_{B_1} = \sum_{(j,k,l)\in\mathcal{S}} \sum_{i=1}^{n} (\overline{y}_{..k.} - \overline{y}_{....})^2 = np\sum_{k=1}^{p} (\overline{y}_{..k.} - \overline{y}_{....})^2$$

- The second block-to-block sum of squares, which quantifies variability in the response observations between the different levels of nuisance factor 2:

$$SS_{B_2} = \sum_{(j,k,l)\in\mathcal{S}} \sum_{i=1}^{n} (\overline{y}_{...l} - \overline{y}_{....})^2 = np\sum_{l=1}^{p} (\overline{y}_{...l} - \overline{y}_{....})^2$$

- The error sum of squares, which quantifies variability unaccounted for by the conditions and blocks:

$$SS_E = \sum_{(j,k,l)\in\mathcal{S}} \sum_{i=1}^{n} (y_{ijkl} - \overline{y}_{.j..} - \overline{y}_{..k.} - \overline{y}_{...l} + 2\overline{y}_{....})^2$$

It is an easy but tedious calculation to show that $SS_T = SS_C + SS_{B_1} + SS_{B_2} + SS_E$. It can also be shown that if H_0 in hypothesis (28) is true then $\mathrm{E}[MS_C] = \mathrm{E}[MS_E] = \sigma^2$, and if H_0 in hypotheses (29) and (30) are true then $\mathrm{E}[MS_{B_1}] = \mathrm{E}[MS_E] = \sigma^2$ and $\mathrm{E}[MS_{B_2}] = \mathrm{E}[MS_E] = \sigma^2$ also. This provides intuition for form of the test statistics $t_C \equiv \frac{MS_C}{MS_E}$, $t_{B_1} \equiv \frac{MS_{B_1}}{MS_E}$ and $t_{B_2} \equiv \frac{MS_{B_2}}{MS_E}$; if t_C, t_{B_1} and t_{B_1} are much larger than 1, this provides evidence against H_0 in hypotheses (28), (29) and (30), respectively. Conclusions are formally drawn on the basis of a comparison of these test statistics to the $F_{(p-1,N-3p+2)}$ distribution. For each test, the p-value is defined as the right-tail probability $\Pr(F_{(p-1,N-3p+2)} \geq t)$. Note that the null distribution is the same for all three tests, because all three factors have the same number of levels p.

Although the sums of squares are simple enough to calculate by hand, this analysis is typically automated using the `lm` and `anova` functions in R, as we will see in the next subsection.

4.3.2 Example: Netflix Latency

Consider the latency experiment described at the beginning of the chapter in which Netflix is experimenting with server-side modifications to improve (reduce) the latency of Netflix.com. In particular, they have four different experimental conditions (A,B,C,D) that are intended to reduce average latency (in milliseconds). Two nuisance factors that may also influence latency are browser (Google Chrome, Microsoft Edge, Firefox, Safari), and time of day (00:01-06:00, 06:01-12:00, 12:01-18:00, 18:01-00:00). The design of the experiment is the 4×4 Latin square shown in Table 15. In order to determine whether the expected latency in each condition differs significantly, $n = 500$ users are randomized to each of the $p^2 = 16$ blocks.

Table 15: 4×4 Latin square design for the Netflix experiment

		Browser			
		Chrome	Edge	Firefox	Safari
Time	00:01-06:00	A	B	C	D
	06:01-12:00	D	A	B	C
	12:01-18:00	C	D	A	B
	18:01-00:00	B	C	D	A

The data is analyzed with the following linear regression model

$$Y_i = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \delta_2 w_{i2} + \delta_3 w_{i3} + \delta_4 w_{i4} + \varepsilon_i$$

where x_{i2}, x_{i3}, x_{i4} are indicators for conditions B,C,D (condition A is the baseline), z_{i1}, z_{i2}, z_{i3} are browser indicators for Microsoft Edge, Firefox and Safari (Google Chrome is the baseline), and w_{i2}, w_{i3}, w_{i4} are time indicators for time periods 2-4 (time period 1 is the baseline). The ANOVA table associated with this model is shown in Table 16.

Table 16: Netflix Latin Square ANOVA table

Source	SS	df	MS	Test Stat.
Condition	203903.38	3	67967.79	679.14
Browser	32.95	3	10.98	0.1097
Time	333242.01	3	111080.67	1109.92
Error	799636.18	7990	100.08	
Total	1336815	7999		

To determine whether there is a difference in average latency from one condition to another we test

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

and we do so using the test statistic $t_C = 679.14$. The p-value associated with this test is $P(T \geq 679.14) \approx 0$, where $T \sim F_{(3,7990)}$. This incredibly small p-value leads us to reject the null hypothesis and conclude that the average latency in each condition is *not* the same. Follow-up pairwise tests may be used to determine *which* condition minimizes average latency, but in lieu of this the top plots in Figure 15 suggest that condition A is best.

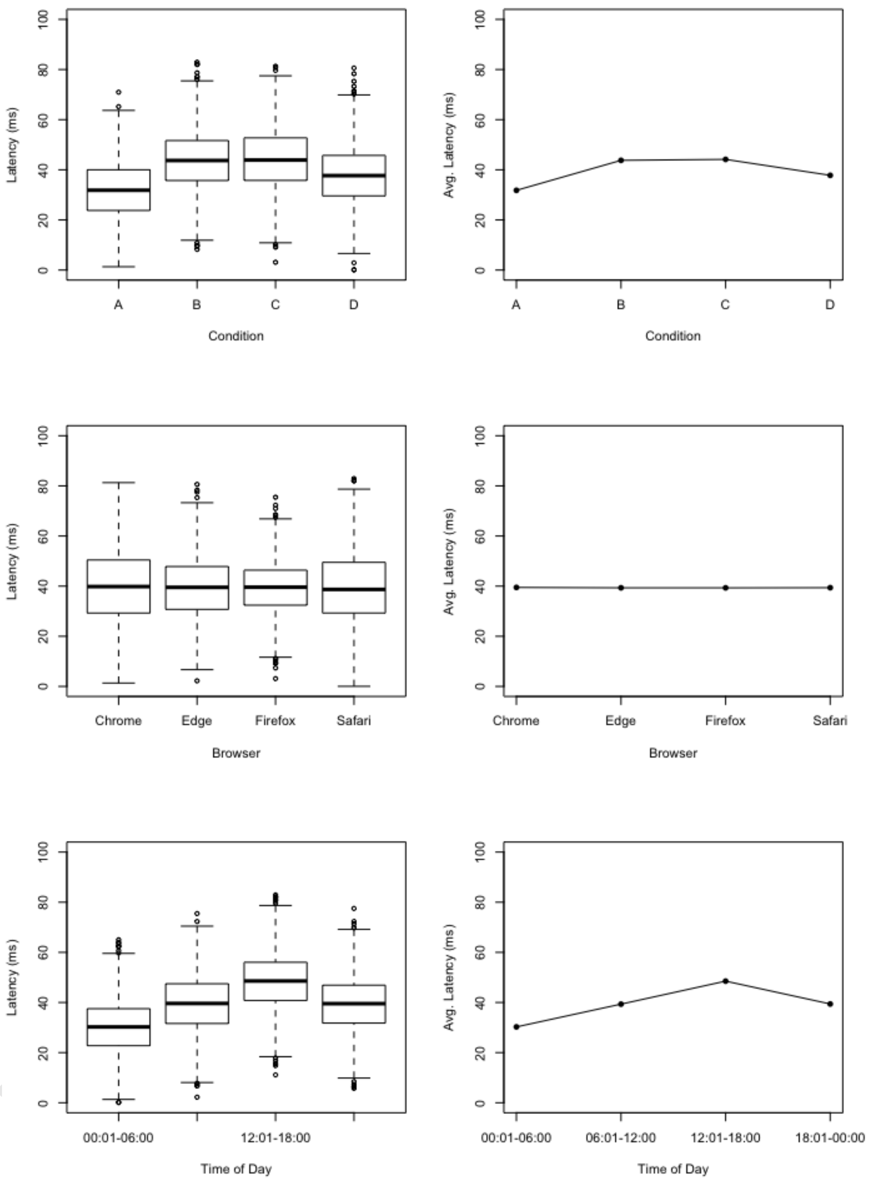

Figure 15: Top: Visualizations of condition-to-condition differences in latency; Middle: Visualizations of browser-to-browser differences in latency; Bottom: Visualizations of timeframe-to-timeframe differences in latency

To confirm that blocking was a good thing to do, we could formally test

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$$

and hence whether the z's (i.e., the nuisance factor "browser") really needed to be in the model, as well as

$$H_0 : \delta_2 = \delta_3 = \delta_4 = 0$$

and hence whether the w's (i.e., the nuisance factor "time of day") really needed to be in the model. The test statistics for these tests are respectively $t_{B_1} = 0.1097$ and $t_{B_2} = 1109.92$ with corresponding p-values

$P(T \geq 0.1097) = 0.9545$ and $P(T \geq 1109.92) \approx 0$ where $T \sim F_{(3,7990)}$. These p-values, and the middle and bottom plots in Figure 15, suggest that timeframe-to-timeframe differences are significant, but browser-to-browser differences are not. Thus blocking by time of day seems to have been appropriate, but blocking by broswer was probably not necessary.

4.3.3 Latin Squares to Compare Proportions

When our metric of interest is a probability/proportion our response variable is binary, in which case logistic regression is the appropriate analysis method. Here interest lies in testing the following hypothesis, all while accounting for the influence of the two nuisance factors:

$$H_0\colon \pi_1 = \pi_2 = \cdots = \pi_p \text{ vs. } H_A\colon \pi_j \neq \pi_{j'} \text{ for some } j' \neq j$$

As discussed previously, this hypothesis is equivalent to hypothesis (28) in the context of the following logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{j=1}^{p-1} \beta_j x_{ij} + \sum_{k=1}^{p-1} \gamma_k z_{ik} + \sum_{l=1}^{p-1} \delta_l w_{il}$$

Here Y_i is the binary response observation for unit $i = 1, 2, \ldots, N = np^2$ and $\pi_i = \mathrm{E}[Y_i]$ is the corresponding expectation (i.e., the probaility that unit i performs some action of interest). Within this modeling framework, hypothesis (28) is tested with a likelihood ratio test (LRT) that compares the full model (shown above) to the reduced one without the x's. Similarly, hypotheses (29) and (30) are tested with LRTs that compare the full model to the reduced ones without the z's and w's. The observed test statistic for these tests is

$$\begin{aligned} t &= 2 \times \log\left(\frac{\text{Likelihood}_{\text{Full Model}}}{\text{Likelihood}_{\text{Reduced Model}}}\right) \\ &= 2 \times [\text{Log-Likelihood}_{\text{Full Model}} - \text{Log-Likelihood}_{\text{Reduced Model}}] \end{aligned}$$

which, if the null hypothesis is true, should follow a $\chi^2_{(p-1)}$ distribution. The corresponding p-values are right-tail probabilities in the $\chi^2_{(p-1)}$ distribution: $\Pr(T \geq t)$ where $T \sim \chi^2_{(p-1)}$. As with the F-tests discussed in Section 4.3.1, the null distribution is the same for all three tests because all three factors have the same number of levels p.

The calculation of the likelihood values cannot be done in closed form, and so the analyses described above must be computed in R. We use the `glm` and `lrtest` functions for this task.

4.3.4 Example: Uber Weekend Promos

Consider an experiment in which Uber is investigating the influence of three different promotional offers on ride-booking-rate (RBR).

- Promo A: None

- Promo B: One free ride today
- Promo C: Book a ride today and get 50% off your next 2 rides

The experimenters would like to control for a possible day-of-week effect and so they want to block by day. They would also like to control for possible city-to-city differences and so they also want to block by city. To do so they run a 3×3 Latin square design as illustrated in Table 17. Interest lies in determining whether or not the different promotions perform similarly with respect to RBR – and they wish to determine which one maximizes RBR – while controlling for the effects of day and city. In order to do this they randomize $n = 1000$ users to each of the $p^2 = 9$ blocks.

Table 17: 3×3 Latin square design for the Uber experiment

		City		
		Toronto	Vancouver	Montreal
Day	Friday	A	B	C
	Saturday	C	A	B
	Sunday	B	C	A

The data is analyzed with the following logistic regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \delta_1 w_{i1} + \delta_2 w_{i2}$$

where x_{i2} and x_{i3} are condition indicators for promos B and C (promo A is the baseline), z_{i1} and z_{i2} are day indicators for Saturday and Sunday (Friday is the baseline), and w_{i1} and w_{i2} are city indicators for Toronto and Vancouver (Montreal is the baseline). To determine whether there is a difference in ride-booking-rate from one promotional offer to another we test

$$H_0 : \beta_2 = \beta_3 = 0$$

and we do so using the test statistic $t_C = 16.648$ (which was calculated using the `lrtest` function in R). The p-value associated with this test is $P(T \geq 16.648) = 0.00024$, where $T \sim \chi^2_{(2)}$. This small p-value leads us to reject the null hypothesis and conclude that the RBR in each condition is *not* the same. Follow-up pairwise tests may be used to determine *which* promo maximizes RBR, but in lieu of this the left plot in Figure 16 suggest that Promo B is best.

To confirm that blocking was a good thing to do, we could formally test

$$H_0 : \gamma_1 = \gamma_2 = 0$$

and hence whether the z's (i.e., the nuisance factor "day") really needed to be in the model, as well as

$$H_0 : \delta_1 = \delta_2 = 0$$

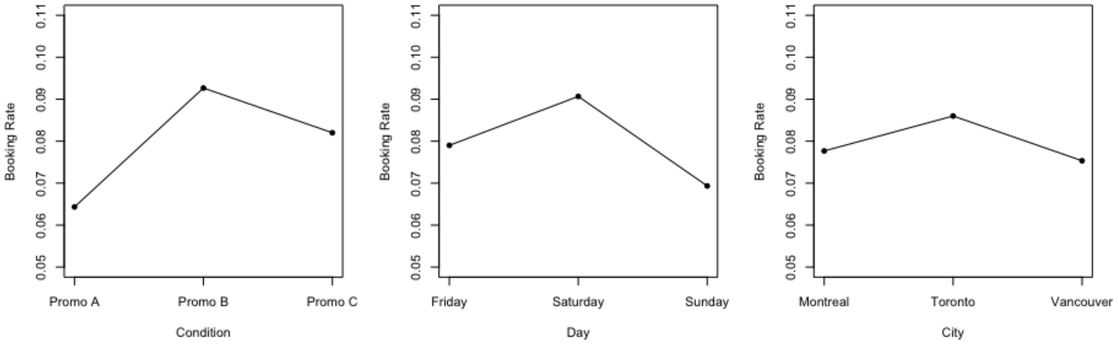

Figure 16: Left: Visualization of promo-to-promo differences in RBR in the Uber experiment; Middle: Visualization of day-to-day differences in RBR in the Enterprise experiment; Right: Visualization of city-to-city differences in RBR in the Uber experiment

and hence whether the w's (i.e., the nuisance factor "city") really needed to be in the model. The test statistics for these tests are respectively $t_{B_1} = 8.9107$ and $t_{B_2} = 2.1193$ (calculated using `lrtest`) with corresponding p-values $P(T \geq 8.9107) = 0.0116$ and $P(T \geq 2.1193) = 0.3466$ where $T \sim \chi^2_{(2)}$. These p-values, and the middle and right plots in Figure 16, suggest that day-to-day differences are significant, but city-to-city differences are not. Thus blocking by day seems to have been appropriate, but blocking by city was probably not necessary.

5 EXPERIMENTS WITH MULTIPLE DESIGN FACTORS

Thus far we have considered experiments with two or more experimental conditions. But no matter the number of conditions, they have always been derived from the levels of a single design factor. However, in most practical circumstances, there might be several factors that are expected to impact a response variable, and hence the metric of interest. As such, in this chapter, we consider experimentation with multiple design factors – sometimes referred to colloquially as *multivariate testing* (MVT). In previous chapters we have motivated the discussion by considering a toy example in which an experiment is performed to determine the influence of a button's colour on the likelihood that the button gets clicked. But what if you also wish to investigate the influence of the button's size or location, or the button's message, on the likelihood that it gets clicked? Is it a large blue button that says "ENTER" that is most likely to get clicked, or is it a medium-sized red one that says "SUBMIT"?

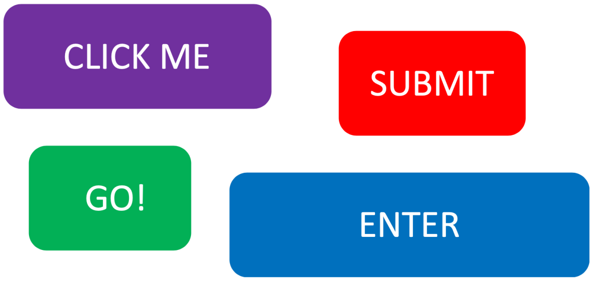

Figure 17: Canonical Multivariate Button Test

In this chapter we describe how to design and analyze experiments that efficiently investigate multiple factors. Similar to Chapters 3 and 4, experiments in this setting involve multiple conditions where the goal is to find the optimal one(s), but the difference here lies in how the conditions arise. In the prior chapters each level of the *single* design factor defined an experimental condtion. Here, an experimental condition is defined by a specific combination of the levels of *multiple* design factors. Then, interest lies in determining which combination of the factors' levels optimize the metric of interest.

5.1 The Factorial Approach

The key to multi-factor experiments is to efficiently investigate different combinations of the factor levels. A common (and simple) method of doing this is the **one-factor-at-a-time** approach in which a sequence of experiments is performed, each with just one factor being varied. Such an approach is manifested as a sequence of single-factor multi-level experiments in which the winning level in a given experiment is retained in future experiments while some other factor is manipulated.

As an example, consider the change Twitter implemented a few years ago in which “favorites” (expressed as stars) were replaced by “likes” (expressed as hearts) and the colour of the heart was chosen to be red (as opposed to say, yellow). These decisions may have came about through a series of experiments in which the icon’s shape and colour were changed. The factors, then, may have been shape and colour with levels respectively given by {*star*, *heart*} and {*red*, *yellow*}. This represents the simplest possible multi-factor situation: two factors, each with two levels. To investigate this using the one-factor-at-a-time approach an initial ‘shape’ experiment would be conducted (i.e., an A/B test to determine which shape – yellow stars or yellow hearts – was most popular). Then, supposing the heart was the winning shape, a follow-up ‘colour’ experiment to determine the optimal colour would be performed (i.e., a second A/B test to determine whether red hearts or yellow hearts are preferable).

While sequential and iterative learning are important tenets of experimentation, a single, more efficient and more informative, experiment could have been performed in place of this less efficient sequence of two experiments. Note that with the one-factor-at-a-time approach, a red star was never investigated. What if the red star truly out-performs the three shape-colour combinations that were investigated? The only way to determine this would have been to formally run a *red star* condition.

The **factorial** approach to multi-factor experimentation is protected against this shortcoming – potentially optimal conditions are not missed because *every* combination of factor levels is considered. A factorial approach to the Twitter experiment would consist of four conditions stemming from the four possible shape-colour combinations: yellow star, red star, yellow heart, red heart. And assuming the one-factor-at-a-time approach consisted of two A/B tests, a total of four conditions were investigated anyway. So the one-factor-at-a-time approach takes as many conditions as the factorial approach! This represents a missed opportunity; why not simply investigate all possible combinations so that there is no loss of information? Note that as the number of factors and number of levels increase beyond two, the number of conditions associated with a factorial experiment will necessarily be larger than with the one-factor-at-a-time approach. However, one does not risk missing an optimal combination with this approach.

Another reason the factorial approach should be preferred to the one-factor-at-a-time approach is because it allows for the quantification of *both* **main effects** and **interaction effects**. The main effect of factor A (let’s call it), represents the change in the response variable produced by a change in that factor. If the factor has a large influence on the response variable, then the magnitude of its main effect will be large. However, sometimes the main effect of one factor, A, depends on the level of another factor, B. In this case we say that factors A and B **interact** and a large interaction effect will lead to the main effect of A being very different for different levels of B. In order to estimate such an interaction it must be observed – and the only way it can be observed is with a factorial approach.

In Section 5.3 we will discuss the notion of main and interaction effects in more detail, and we will make their definitions more precise. But first, in Section 5.2, we discuss the practical considerations that need to be made when designing a factorial experiment.

5.2 Designing a Factorial Experiment

Conceptually, the design of a factorial experiment is simple: select a number of design factors you believe influence the response, select the levels you'd like to experiment with for each of these design factors, and then define the experimental conditions to be each of the possible combinations of these factors' levels. For instance, suppose you are interested in investigating the factors labeled '1', '2', '3', and '4' which have m_1, m_2, m_3 and m_4 levels, respectively. The full factorial experiment that investigates these factors consists of $m_1 m_2 m_3 m_4$ experimental conditions, representing all possible combinations of the four factors' levels. In general, a full factorial experiment with K factors requires $m = m_1 m_2 \cdots m_K$ conditions.

Clearly, as the number of factors and/or levels increases, the number of experimental conditions gets large. This is the primary drawback of factorial experiments: they get big, quickly. So as not to design an experiment that is unmanageably large, careful thought should be given to both the selection of design factors and their associated levels. In particular, it would be a waste of effort to investigate factors that are highly correlated, or factors that are difficult to manipulate outside the confines of a small controlled experiment. It would also be a waste of effort to investigate several levels of a factor that is uninfluential.

In general, when choosing factors and levels in the context of a factorial experiment, these choices should be made with the "keep it simple" principle in mind. A special class of factorial experiments is the one in which each factor is investigated at just *two* levels. Such a factorial experiment minimizes the number of experimental conditions required to investigate multiple factors, and as such is useful for **screening** factors to determine which ones are influential and which ones are not. Further experimentation can then be performed with the influential factors at more than two levels. We briefly revisit such screening designs in Section 5.4 and we consider them in great depth in Chapters 6 and 7.

Once the factors, levels, and hence experimental conditions have been established, experimental units must be randomized to each of the m conditions. The number of experimental units assigned to each condition is denoted n_j ($j = 1, 2, \ldots, m$). If the design is *balanced* the number of units is the same in each condition (i.e., $n_1 = n_2 = \cdots = n_m = n$). How many units should be assigned to each condition may be determined using the sample size calculations discussed in Chapter 2. To justify this we note that this type of experiment can have two goals:

(1) Identify which combination of factor levels (i.e., which condition) is optimal

(2) Identify which factors are influential and quantify their influence

In Section 5.3 we will see that drawing conclusions about (1) require a sequence of pairwise comparisons similar to those discussed in Section 3.3 and drawing conclusions about (2) above requires regression (linear or logistic). As such, sample size calculations associated with two-sample tests that account for the multiple comparison problem are appropriate here.

5.3 Analyzing a Factorial Experiment

In the previous section we mentioned that factorial experiments may be used for two purposes: (1) to find the optimal combination of factor levels, with respect to the metric of interest, and (2) to identify and quantify the influence of the factors. Determining the optimal condition requires little sophistication and can simply be achieved by pairwise hypothesis tests. Quantifying the influence of the factors, and determining which factors' influence is significant requires an evaluation of main effects and interaction effects and the use of regression models. Because pairwise testing was discussed at length in Chapter 2, in this section we focus on regression analyses associated with factorial experiments.

As always, the type of regression depends on whether the response variable is continuous or binary. However, no matter which type of regression is appropriate, we use a linear predictor which contains the following terms:

- An intercept
- Main effect terms
- Two-factor interaction terms
- Three-factor interaction terms

$$\vdots$$

- K-factor interaction terms

The main effect for each factor $k = 1, 2, \dots, K$ is represented by $m_k - 1$ indicator variables. This is akin to the manner in which the design and nuisance factors were represented in the linear predictors of Chapter 4. However, unlike the previous linear predictors we have discussed, we now also include interaction terms between each of the factors. In general, an h-factor interaction is represented by the h-way product of the main effect terms associated with each of the $\binom{K}{h}$ combinations of the h factors. The inclusion of such interaction terms accommodates the possibility that the influence of one factor may be modulated by another. This is in direct contrast to the *additive* models of Chapter 4.

The following example illustrates this general structure more clearly.

Example: Suppose we have $K = 3$ factors with $m_1 = 2$, $m_2 = 2$ and $m_3 = 3$ levels and we wish to analyze the factorial experiment that arises from the $m = m_1 m_2 m_3 = 12$ combinations these factor's levels.

The required linear predictor will contain main effects, two factor interactions and three factor interactions. The main effects will be represented by indicator variables, and the interaction effects will be represented by specific products of these indicator variables. In particular, let x_1 and x_2 be the indicator variables associated with design factors 1 and 2, respectively, and let x_3 and x_4 be the indicator variables associated with design factor 3. The linear predictor is thus given by,

$$\beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4}_{\text{main effects}} + \underbrace{\beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_2 x_3 + \beta_9 x_2 x_4}_{\text{two-factor interactions}} + \underbrace{\beta_{10} x_1 x_2 x_3 + \beta_{11} x_1 x_2 x_4}_{\text{three-factor interactions}}$$

Hypotheses concerning the main effects will therefore involve $\beta_1, \beta_2, \beta_3, \beta_4$, hypotheses concerning the two-factor interactions will involve $\beta_5, \beta_6, \beta_7, \beta_8, \beta_9$, and hypotheses concerning the three-factor interactions will involve β_{10}, β_{11}. Such tests will be performed by comparing full versus reduced models via partial F-tests (in the case of linear regression) and likelihood ratio tests (in the case of logistic regression). We consider both scenarios in the next two subsections.

5.3.1 Continuous Response – The Instagram Example

We illustrate the topics discussed in this section in the context of the following example. Consider a multi-factor, multi-level version of the Instagram experiment from Section 2.1.3 in which the influence of ad frequency on session duration was evaluated. Now suppose that ad frequency has levels {*9:1*, *7:1*, *4:1*, *1:1*} corresponding to ad frequencies of 1 in 10, 1 in 8, 1 in 5, and every other. Additionally, suppose ad type is introduced as a second design factor with levels {*photo*, *video*}. Thus we have two factors with 4 and 2 levels, respectively, thereby creating $2 \times 4 = 8$ experimental conditions – the four ad frequencies with photo ads and the four ad frequencies with video ads.

Now suppose that $n = 1000$ units (users) are randomized to each condition and their session duration is recorded. These data, and the influence of ad freqency and ad type both marginally and jointly are depicted in Figures 18, and 19. In particular, we see in the left plot of Figure 18 that session duration decreases steadily as ad frequency increases, and in the right plot we see (perhaps unsurprisingly) that session duration is slightly longer when video ads are displayed instead of photo ads. In addition to this information, these plots suggest that ad frequency is a more influential factor than ad type since the change in session duration across different ad frequencies is much larger than the change produced by different ad types. The average session durations for each ad frequency and each ad type are provided in Table 18.

However, discussing main effects like this can be uninformative and potentially misleading if there is a signficant interaction between the factors. In particular, if the main effect of ad frequency (depicted in the left plot of Figure 18) is different for photo ads versus video ads, then we would say that there is an interaction between the factors. Visually, such an interaction would be indicated if the pattern seen in the

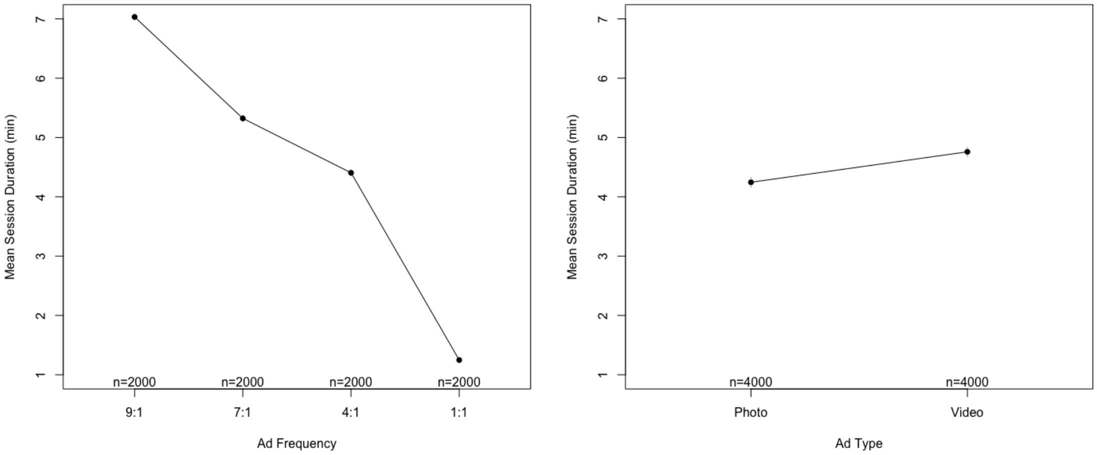

Figure 18: Left: Main effect plot for ad frequency; Right: Main effect plot for ad type.

Table 18: Average session duration for different ad frequencies

Frequency	Average Session Duration (min)
9:1	7.03
7:1	5.32
4:1	4.41
1:1	1.25
Type	Average Session Duration (min)
Photo	4.25
Video	4.76

left plot of Figure 18 is roughly the same (i.e., parallel line segments) for both ad types. Or, equivalently, if the pattern seen in the right plot of Figure 18 is roughly the same (i.e., parallel lines) for all ad frequencies. Figure 19 depicts the interaction plots for the two factors. The left plot corresponds to a plot of the main effect of frequency for the two different ad types, and the right plot depicts the main effect of ad type for the four different frequencies. Non-parallel line segments on these plots would indicate the presence of an interaction since this would correspond to the main effect of one factor depending on the levels of the other factor. The larger the departure from parellelism, the stronger the interaction effect.

As we can see in Figure 19, the lines are not perfectly parallel, indicating the presence of a small interaction effect. However, the departure from parallelism is not drastic, and so we would not expect the interaction effect to be large. The average session durations in each condition are provided in Table 19

Table 19: Average session duration in each ad frequency-type condition.

		Ad Type	
		Photo	Video
Frequency	9:1	6.98	7.09
	7:1	5.01	5.64
	4:1	3.98	4.83
	1:1	1.02	1.48

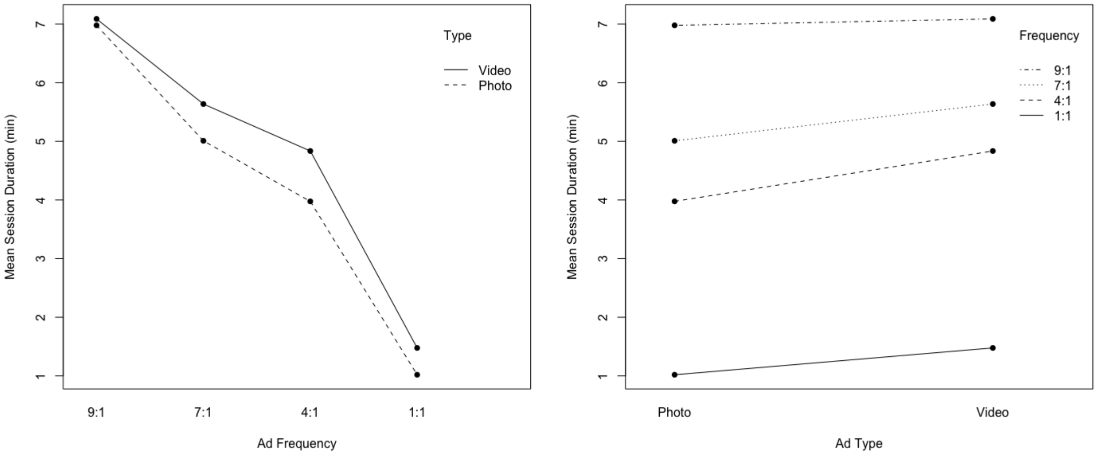

Figure 19: Interaction plot for ad frequency and ad type.

In order to formally evaluate whether the frequency-type main or interaction effects are significant we turn to linear regression. Here we assume that our response variable measurements (in this case, session duration) follow a normal distribution and can be modeled by an ordinary linear regression model with linear predictor as discussed above. In particular, for this experiment, the appropriately defined linear regression model is given by

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i1} x_{i4} + \beta_6 x_{i2} x_{i4} + \beta_7 x_{i3} x_{i4} + \epsilon_i \tag{31}$$

where the x's are indicator variables such that $x_{i1} = 1$ if unit i is in a condition with 7:1 ad frequency, $x_{i2} = 1$ if unit i is in a condition with 4:1 ad frequency, $x_{i3} = 1$ if unit i is in a condition with 1:1 ad frequency, and $x_{i4} = 1$ if unit i is in a condition with video ads, for $i = 1, 2, \dots, n$. Table 20 summarizes the expected response in each of the experimental conditions for this model.

Table 20: Expected response in each ad frequency-type condition

		Ad Type	
		Photo	Video
Freq.	9:1	$E[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 0] = \beta_0$	$E[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 1] = \beta_0 + \beta_4$
	7:1	$E[Y_i \mid x_{i1} = 1, x_{i4} = 0] = \beta_0 + \beta_1$	$E[Y_i \mid x_{i1} = 1, x_{i4} = 1] = \beta_0 + \beta_1 + \beta_4 + \beta_5$
	4:1	$E[Y_i \mid x_{i2} = 1, x_{i4} = 0] = \beta_0 + \beta_2$	$E[Y_i \mid x_{i2} = 1, x_{i4} = 1] = \beta_0 + \beta_2 + \beta_4 + \beta_6$
	1:1	$E[Y_i \mid x_{i3} = 1, x_{i4} = 0] = \beta_0 + \beta_3$	$E[Y_i \mid x_{i3} = 1, x_{i4} = 1] = \beta_0 + \beta_3 + \beta_4 + \beta_7$

In the presence of a significant interaction effect, it no longer makes sense to discuss the main effect of a factor, because doing so ignores the fact that this effect changes depending on the level of another factor. As such it is typical to first decide whether the interaction is statistically significant. Notice that if $\beta_5 = \beta_6 = \beta_7 = 0$ in model (31) then it would presume that an interaction between ad frequency and ad type does not exist. As such, a formal test of

$$H_0\colon\ \beta_5 = \beta_6 = \beta_7 = 0 \text{ vs. } H_A\colon\ \beta_j \neq 0$$

for $j = 5, 6, 7$ would evaluate the significance of the interaction effect. As can be seen in Table 20, if the interaction is significant then any conclusions regarding the effect of one factor must be made in the context of the levels of the other factor.

Alternatively, if the interaction is not significant, then the interaction terms can be removed from the model yielding the following simplified *main effects* model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i \tag{32}$$

which can be used to evaluate the significance of the main effect of each factor (in this case ad frequency and ad type). Table 21 summarizes the expected response (based on this main effects model) in each of the experimental conditions.

Table 21: Expected response in each ad frequency-type condition, based on the main effects model

		Ad Type	
		Photo	Video
Freq.	9:1	$E[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 0] = \beta_0$	$E[Y_i \mid x_{i1} = x_{i2} = x_{i3} = 0, x_{i4} = 1] = \beta_0 + \beta_4$
	7:1	$E[Y_i \mid x_{i1} = 1, x_{i4} = 0] = \beta_0 + \beta_1$	$E[Y_i \mid x_{i1} = 1, x_{i4} = 1] = \beta_0 + \beta_1 + \beta_4$
	4:1	$E[Y_i \mid x_{i2} = 1, x_{i4} = 0] = \beta_0 + \beta_2$	$E[Y_i \mid x_{i2} = 1, x_{i4} = 1] = \beta_0 + \beta_2 + \beta_4$
	1:1	$E[Y_i \mid x_{i3} = 1, x_{i4} = 0] = \beta_0 + \beta_3$	$E[Y_i \mid x_{i3} = 1, x_{i4} = 1] = \beta_0 + \beta_3 + \beta_4$

Notice that if $\beta_1 = \beta_2 = \beta_3 = 0$ in this model there is no difference between the expectations in the four rows. Thus the hypothesis

$$H_0\colon\ \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_A\colon\ \beta_j \neq 0$$

for $j = 1, 2, 3$ tests whether ad frequency is a significant factor, and performing this test indicates whether the main effect of ad frequency is significant. Next, notice that if $\beta_4 = 0$ there is no difference between the expectations in the two columns. Thus the hypothesis

$$H_0\colon\ \beta_4 = 0 \text{ vs. } H_A\colon\ \beta_4 \neq 0$$

tests whether ad type is a significant factor, and performing this test indicates whether the main effect of ad type is significant. It is important to emphasize, however, that these tests and the interpretation of main effects are only appropriate in the absence of interaction.

Each of the null hypotheses just described generates a **reduced model** with fewer terms relative to a **full model** with all terms. Each of the hypotheses, then, is tested by comparing the reduced model to the full one. In particular, to decide whether the frequency-type interaction is significant we compare model (31) to the reduced version that arises when $\beta_5 = \beta_6 = \beta_7 = 0$. To decide whether the ad frequency main effect is significant we compare model (32) to the reduced version when $\beta_1 = \beta_2 = \beta_3 = 0$, and to decide whether the ad type main effect is significant we compare model (32) to the reduced version when $\beta_4 = 0$. Each of these comparisons is done using a **partial *F*-test** which compares the mean squared errors between the full and reduced models – similar to the F-test for overall significance in a linear regression. If the number of β's in the full model is p, and $q < p$ in the reduced model, then the numerator degrees of freedom for these tests is $\nu = p - q$. The denominator degrees of freedom is the error degrees of freedom in the full model. The test statistics and p-values of these tests are provided in standard linear regression ANOVA tables.

The linear regression output from R's `lm()` function applied to these data is shown below. In this output we can see, based on the displayed p-values, that each individual term in the model is statistically significant. However, when judging the overall significance of a main or interaction effect it is inappropriate to draw conclusions from the individual t-tests from which these p-values arise. Instead we must examine the p-values associated with the partial F-tests associated with the model's analysis of variance.

```
         Coefficients: Estimate Std.Err  t value   Pr(>|t|)

           (Intercept)  6.97785 0.02824  247.104  < 2e-16 ***
          Frequency7:1 -1.96929 0.03994  -49.312  < 2e-16 ***
          Frequency4:1 -3.00204 0.03994  -75.173  < 2e-16 ***
          Frequency1:1 -5.95856 0.03994 -149.206  < 2e-16 ***
             TypeVideo  0.10993 0.03994    2.753  0.00592 **
Frequency7:1:TypeVideo  0.51768 0.05648    9.166  < 2e-16 ***
Frequency4:1:TypeVideo  0.74924 0.05648   13.266  < 2e-16 ***
Frequency1:1:TypeVideo  0.34731 0.05648    6.150 8.14e-10 ***

--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.893 on 7992 degrees of freedom
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8496
F-statistic: 6455 on 7 and 7992 DF, p-value: < 2.2e-16
```

The ANOVA table from R's `anova()` function is shown below. The p-values in this table are all sufficiently small (smaller than any reasonable significance level), allowing us to conclude that the main effects of ad frequency and ad type, and the interaction between them, are all statistically significant. The size of the test statistics corresponding to these p-values provides insight into the relative size of these effects. In particular we see that ad frequency is substantially more influential than ad type, which is more significant than the interaction effect. Thus we conclude that both factors are important, and so is their interaction, which means that both factors should be considered when trying to optimize session duration, and the influence of one cannot be separated from the influence of the other.

‘

```
Analysis of Variance Table Response:

         Time  Df  Sum Sq Mean Sq F value   Pr(>F)
     Frequency  3   35353 11784.3 14778.187 < 2.2e-16 ***
          Type  1     527 527.3   661.318   < 2.2e-16 ***
Frequency:Type  3     149 49.8    62.398    < 2.2e-16 ***
     Residuals  7992 6373 0.8

--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In order to determine which frequency-type combination (and hence experimental condition) is optimal, a series of one-sided *t*-tests may be used. This is left as an exercise for the reader. In the next section we discuss the analysis of a similarly designed factorial experiment, but where the response variable is binary as opposed to continuous.

5.3.2 Binary Response – The TinyCo Example

The informal and formal evaluation of main and interaction effects can be performed in the context of a binary response variable as well. However, in this situation, main effect and interaction effect plots are based on observed proportions and logistic regression is used instead of ordinary linear regression. Like linear regression, the linear predictor in a logistic regression model is composed of indicator variables corresponding to the levels of each design factor, and products of these indicator variables.

The difference between logistic regression and ordinary linear regression is that this linear predictor is equated to $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ where $\pi_i = \text{E}[Y_i|\mathbf{x_i}]$ is the probability that unit i, with explanatory variable values given by $\mathbf{x_i}$, has response value equal to 1.

As in the case of linear regression, interest lies in determining whether subsets of the β's are equal to zero (as assumed by H_0) to evaluate the significance of various main and interaction effects. Whereas partial F-tests are used for this task in the context of linear regression models, we use **likelihood ratio tests** for this purpose with logistic regression models. To perform such a test, we compare the maximized log-likelihood of the full model to that of a reduced model (that arises when H_0 is true). The test statistic for the likelihood ratio test is the same here as it was in Chapter 4:

$$
\begin{aligned}
t &= 2 \times \log\left(\frac{\text{Likelihood}_{\text{Full Model}}}{\text{Likelihood}_{\text{Reduced Model}}}\right) \\
&= 2 \times [\text{Log-Likelihood}_{\text{Full Model}} - \text{Log-Likelihood}_{\text{Reduced Model}}]
\end{aligned}
$$

which, if H_0 is true, approximately follows a χ^2 distribution with ν degrees of freedom, where ν is the number of restrictions implied by H_0. For instance, the null hypothesis H_0: $\beta_5 = \beta_6 = \beta_7 = 0$ places restrictions on three of the β's and so $\nu = 3$ in this case. Recall that a simple way to determine the value of ν is to compare the number of regression coefficients in the two models: if the number of β's in the full model is p, and $q < p$ is the number of β's in the reduced model, then $\nu = p - q$. The p-value for this test is calculated as p-value $= P(T \geq t)$ where $T \sim \chi^2_{(\nu)}$.

While such tests are useful for identifying significant main and interaction effects, a series of pairwise Z-tests (χ^2-tests) may be performed to determine which of the experimental conditions is optimal. We illustrate this process with an example concerning TinyCo.

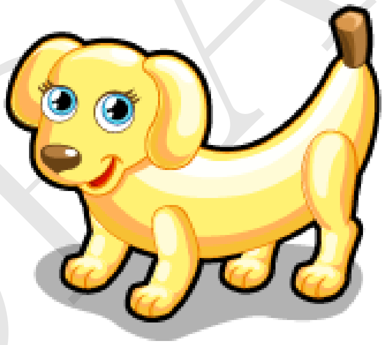

Figure 20: The “Bananimal”

TinyCo is a mobile video game studio that develops the Tiny Zoo game. In this game users own zoos and collect animals to put in their zoos. An experiment is performed in which a new animal, the “bananimal” (see Figure 20), is released for purchase in the game. Interest lies in understanding the relationship between conversion (purchase rate) and two factors: the bananimal's colour (yellow or gold) and the bananimal's price (\$10, \$20, or \$30 of in-game currency). A factorial experiment with 6 conditions was performed to investigate these relationships. A summary of the data resulting from this experiment is tabulated in Table 22 and visualized in Figure 21.

The top left plot in Figure 21 indicates that the purchase rate is higher for gold as opposed to yellow bananimals. In fact, Table 22 indicates that the three highest purchase rates are observed when the bananimal is gold. The bottom left plot indicates that as the bananimal's price increases, the purchase rate decreases.

Table 22: Purchase rates in each condition of the Bananimal experiment

Condition	Sample Size	Purchase Rate
\$10 + Yellow	500	0.1720
\$20 + Yellow	483	0.0973
\$30 + Yellow	488	0.0492
\$10 + Gold	500	0.2260
\$20 + Gold	500	0.1840
\$30 + Gold	487	0.1992

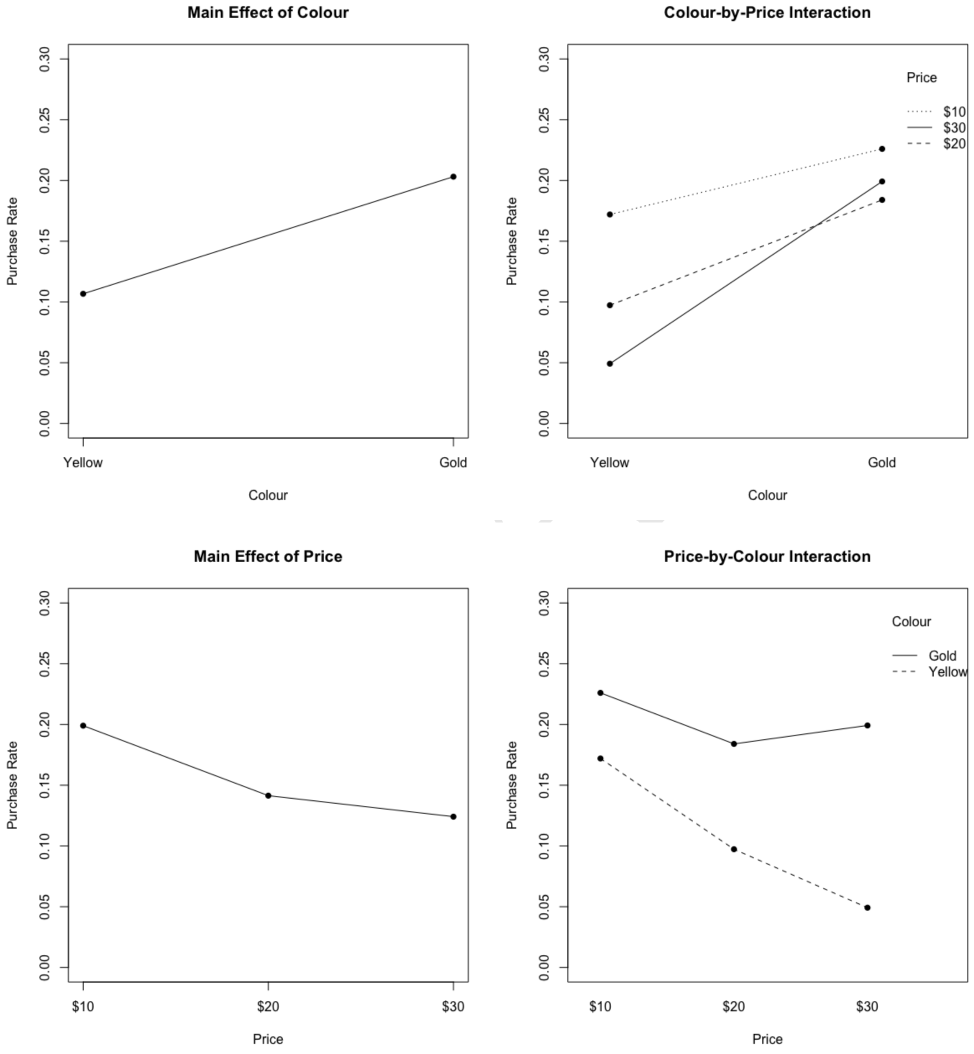

Figure 21: Main and interaction effect plots for the Bananimal experiment

However, interpreting the main effecs based on the left-most plots may be misleading because, as the right-most plots indicate, there appears to be an interaction between colour and price. This interaction effect is most apparent in the bottom-right plot; we see that when we do not ignore colour, purchase rate does not necessarily decrease as price increases. Despite what the main effect plot suggests, purchase rate actually *increases* when a gold bananimal's price increases from \$20 to \$30.

To formally analyze this data we fit a logistic regression model with linear predictor based on the indicator variables x_{i1} (which is 1 if unit i is in a gold bananimal condition), x_{i2} (which is 1 if unit i is in a \$20 bananimal condition), and x_{i3} (which is 1 if unit i is in a \$30 bananimal condition). The linear predictor associated with the full model, which quantifies both main and interaction effects, is

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} \tag{33}$$

The corresponding main effects linear predictor is

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} \tag{34}$$

An LRT-based comparison of these models formally tests the significance of the interaction effect via the hypothesis

$$H_0 : \beta_4 = \beta_5 = 0 \text{ vs. } H_A : \exists\ j \in \{4, 5\} \text{ s.t. } \beta_j \neq 0$$

The test statistic associated with this test is $t = 19.918$ and the p-value is $\Pr(T \geq 19.918) = 4.731 \times 10^{-5}$, where $T \sim \chi^2_{(2)}$. Since this p-value is quite small, we reject H_0 and conclude that the interaction between colour and price is significant.

If we wish, we could also confirm the significance of the main effects of colour and price by respectively testing $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = \beta_3 = 0$ in the context of the main effects model (34). The test statistics and p-values for these tests are respectively given by $t = 53.757$; p-value $= 2.269 \times 10^{-13}$ and $t = 23.324$; p-value $= 8.614 \times 10^{-6}$. These p-values provide sufficient evidence against both null hypotheses, suggesting that both colour and price significantly influence purchase rate. However, because a significant interaction effect was identified, the main effects should not be interpreted in isolation. In particular, we expect gold bananimals to have a higher purchase rate than yellow bananimals. We also find that increasing price from \$10 to \$20 elicits a decrease in purchase rate, and increasing price from \$20 to \$30 elicits a decrease in purchase rate for yellow bananimals, but an increase for gold bananimals.

In order to determine the optimal condition we compare the observed purchase rates. A series of χ^2-tests of independence may be used to show that the purchase rates in gold bananimal conditions do not significantly differ from one another. In other words, as long as the bananimal is gold, TinyCo can charge either \$10 or \$20 or \$30 without significantly impacting purchase rate. Presumably, TinyCo wants to make as much money as possible, and therefore would choose to sell gold bananimals at \$30 each.

5.4 Two-Level Factorial Experiments

Throughout this chapter we have seen that factorial experiments are an informative and efficient way of investigating the influence of multiple factors on some response variable of interest. The key advantage to this method of experimentation (relative to the one-factor-at-a-time approach) is that every possible combination of factor levels is considered, and so we do not risk missing an optimal combination. However, as we have discussed, this advantage is also the main drawback of such an experiment. The problem is that as the number of factors and the number of levels increase, the number of unique combinations, and hence experimental conditions, gets very large.

A simple special case of a general factorial experiment is the **two-level** factorial experiment. In these experiments we consider investigating K design factors, each at *only two levels*. While several levels may be plausible, just two must be chosen for the purpose of these experiments. Such experiments are typically used for factor **screening**. When a large number of factors may potentially influence the response, screening experiments may be used to determine which of them is most influential. In practice, the **Pareto principle**[4] often applies and interest lies in determining the small number of important factors. Once these factors have been identified, follow-up experiments that investigate just these factors at a larger number of levels can be performed. Here we think of these two-level experiments as providing an efficient method of homing in on truly influential factors.

With K factors, each at two levels, there are a total of 2^K unique combinations of the factors' levels. In Chapter 6 we consider **2^K factorial experiments** which investigate K factors using 2^K experimental conditions defined by each of the unique combinations of the factors' levels. In Chapter 7 we discuss **2^{K-p} fractional factorial** experiments which investigate K factors with just 2^{K-p} specifically chosen experimental conditions (i.e., just a *fraction* of all possible conditions). Such two-level designs efficiently explore the relationship between a response variable and K design factors.

[4] The Pareto principle states that only a *vital few* factors are important relative to the *trivial many*.

6 2^K FACTORIAL EXPERIMENTS

In this chapter we discuss the design and analysis of 2^K factorial experiments – that is, factorial experiments involving K design factors, each at two levels. Such experiments are typically used for factor screening, as discussed in Section 5.4. Thus, the *primary* goal of these experiments is, with 2^K conditions, to determine which among the K factors significantly influence the response variable. A *secondary* concern, if all of the levels of interest are the ones experimented with, is to determine which combination of levels is optimal.

6.1 Designing 2^K Factorial Experiments

As with any experiment, the first things to choose in the design phase of a 2^K factorial experiment are the of metric of interest and the corresponding response variable. These choices are typically dictated by the context of the experiment and the business/research question that the experiment is being designed to answer. Next, one must choose the design factors. In particular, one must choose K factors that are expected to influence the response variable in some way. Then, two levels for each factor are chosen to experiment with. Given that this form of factorial experiment is usually used for screening, and hence identifying influential factors, it is important to choose levels that provide the largest opportunity for an influential factor to be noticed. Recall that the main effect of a factor is defined to be the change in response produced by a change in the factor; thus levels should be chosen that are quite different from one another; even a very influential factor may not appear to be influential if the factor levels are too similar.

With the factors and factor levels chosen, the experimental conditions are defined to be the unique combinations of the factor levels. Since each factor is investigated at just two levels, there are 2^K distinct combinations and hence 2^K experimental conditions. Experimental units are then assigned to each condition. For notational simplicity only, we assume that the experiment is balanced and n units are assigned to each of the 2^K conditions. Thus a total of $N = n2^K$ units and hence response observations are collected in such an experiment. To be clear, the balanced sample size assumption is not necessary; the analysis techniques discussed in the next section are applicable even when sample sizes are not balanced.

In the context of two-level experiments, it is common to regard the two levels of a factor as *low* and *high* values of that factor. If the factor is numeric this interpretation is sensible, but if a factor is categorical and there exists no natural ordering for its levels, we arbitrarily label one as *low* and the other as *high*. This labeling is formalized mathematically by associating -1 and $+1$ with the factor's low and high levels, respectively. In particular, we represent each factor by a binary variable x defined as follows:

$$x = \begin{cases} -1 & \text{if the factor is at its "low" level} \\ +1 & \text{if the factor is at its "high" level} \end{cases}$$

Previously, we used binary *indicator* variables to represent factors with two levels. However, as will become clear in the next section, representing factors in this way gives rise to several statistical conveniences when analyzing 2^K factorial experiments.

Table 23: Design matrices for 2^2 (left) and 2^3 (right) factorial experiments.

Condition	Factor 1	Factor 2
1	-1	-1
2	$+1$	-1
3	-1	$+1$
4	$+1$	$+1$

Condition	Factor 1	Factor 2	Factor 3
1	-1	-1	-1
2	$+1$	-1	-1
3	-1	$+1$	-1
4	$+1$	$+1$	-1
5	-1	-1	$+1$
6	$+1$	-1	$+1$
7	-1	$+1$	$+1$
8	$+1$	$+1$	$+1$

With the factor levels coded in this way, each experimental condition can be identified by a unique combination of plus and minus ones. For instance, when $K = 2$, the $2^2 = 4$ conditions may be identified by the pairs $(x_1, x_2) \in \{(-1,-1), (+1,-1), (-1,+1), (+1,+1)\}$. The design of such an experiment may be succinctly described by what is known as the **design matrix**, which contains 2^K rows and K columns of plus and minus ones. The ± 1 entries are organized such that each row corresponds to a unique condition and the columns correspond to each of the factors. Thus, the design matrix provides a prescription for running the 2^K factorial experiment; for a given condition (row) the ± 1 entries indicate whether each of the factors is to be run at their low or high level. Examples of design matrices for 2^2 and 2^3 factorial experiments are shown in Table 23. As can be seen, the general pattern for construction is to alternate -1's followed by $+1$'s in powers of 2, from one column to the next. Doing so ensures that every condition is represented exactly once, and it ensures that the columns are *orthogonal*. This point will be elaborated up on the next section.

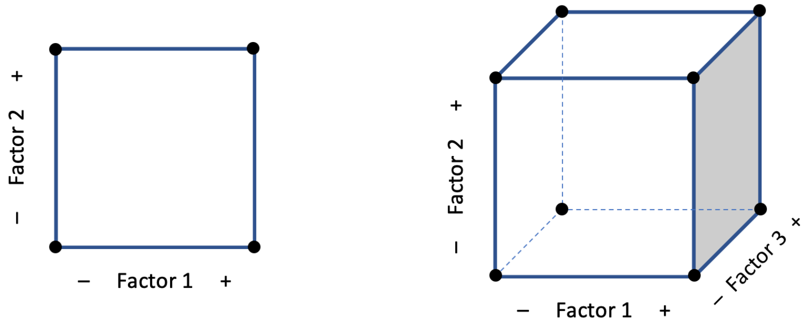

Figure 22: Cuboidal representation of 2^2 (left) and 2^3 (right) factorial designs.

A 2^K experiment may also be visualized geometrically as a K-dimensional hypercube where the vertices correspond to the unique configurations of the K factors' levels. Thus, when $K = 2, 3$, or 4 the 2^K design may be visualized as a square, cube, or tesseract, respectively. Such visualizations for 2^2 and 2^3 factorial experiments are shown in Figure 22.

Having now discussed the design of a 2^K experiment, we next discuss how the $N = n2^K$ response observations may be analyzed in order to achieve our analysis goals.

6.2 Analyzing 2^K Factorial Experiments

Given that the primary goal of a 2^K factorial experiment is factor screening, interest lies primarily in estimation of main and interaction effects for the K design factors. Because each factor is investigated at just two levels, the main effect of a factor is simply defined as the expected change produced by changing that factor from its low to its high level. As defined in Chapter 5, an interaction is said exist between two factors if the main effect of one depends on the levels of the other. As we will see, the estimation and evaluation of such effects may be carried out with linear or logistic regression modeling, depending on whether the response is continuous or binary. We discuss the regression framework for estimation and significance testing in Section 6.2.2. But first, in Section 6.2.1, we provide an analysis of a 2^2 factorial experiment based on intuition which, as we will see in Section 6.2.2, is closely related to the regression-based approach. Then, in Section 6.2.3, we conduct an analysis of a 2^4 factorial experiment that was performed to investigate the influence of four factors on conversion in a credit card promotion.

6.2.1 An Intuition-Based Analysis

As a toy example, consider a 2^2 experiment in which two factors A and B are investigated. The $n = 3$ response observations in each condition are provided in Table 24. An intuitive estimate of the main effect of factor A (i.e., the expected change in response produced by changing factor A from its low to its high level) is the difference between the average response when A is at its high vs. its low level. In particular:

$$\widehat{ME}_A = \overline{y}_{A^+} - \overline{y}_{A^-} = \frac{\overline{y}_{A^+\cap B^-} + \overline{y}_{A^+\cap B^+}}{2} - \frac{\overline{y}_{A^-\cap B^-} + \overline{y}_{A^-\cap B^+}}{2} \tag{35}$$

which for these data is given by $\frac{(12/3)+(8/3)}{2} - \frac{(4/3)+(6/3)}{2} = \frac{10}{6}$.

Similarly, we estimate the main effect of factor B by

$$\widehat{ME}_B = \overline{y}_{B^+} - \overline{y}_{B^-} = \frac{\overline{y}_{A^-\cap B^+} + \overline{y}_{A^+\cap B^+}}{2} - \frac{\overline{y}_{A^-\cap B^-} + \overline{y}_{A^+\cap B^-}}{2} \tag{36}$$

which for these data is $\frac{(6/3)+(8/3)}{2} - \frac{(4/3)+(12/3)}{2} = -\frac{1}{3}$.

To evaluate whether factors A and B interact, we should compare the main effect of A when B is at its high level ($ME_{A|B^+}$) to the main effect of A when B is at its low level ($ME_{A|B^-}$). Estimates of these conditional main effects are defined as

$$\widehat{ME}_{A|B^+} = \overline{y}_{A^+\cap B^+} - \overline{y}_{A^-\cap B^+}$$

and

$$\widehat{ME}_{A|B^-} = \overline{y}_{A^+ \cap B^-} \; - \; \overline{y}_{A^- \cap B^-}$$

Analogous definitions exist for the main effect of B conditional on the levels of A. Note that for these data $\widehat{ME}_{A|B^+} = (8/3) - (6/3) = 2/3$ and $\widehat{ME}_{A|B^-} = (12/3) - (4/3) = 8/3$. Since $\widehat{ME}_{A|B^+} \neq \widehat{ME}_{A|B^-}$ we conclude that there is an interaction between factors A and B. The interaction effect is defined as the average difference between the conditional main effects:

$$\widehat{IE}_{AB} = \frac{\widehat{ME}_{A|B^+}}{2} - \frac{\widehat{ME}_{A|B^-}}{2} = \frac{\widehat{ME}_{B|A^+}}{2} - \frac{\widehat{ME}_{B|A^-}}{2} = \frac{\overline{y}_{A^+ \cap B^+} + \overline{y}_{A^- \cap B^-}}{2} - \frac{\overline{y}_{A^+ \cap B^-} + \overline{y}_{A^- \cap B^+}}{2} \quad (37)$$

For this data we find $\widehat{IE}_{AB} = (2/6) - (8/6) = -1$.

Note that if a third factor C were involved, we may define the three-way ABC interaction as:

$$\widehat{IE}_{ABC} = \frac{\widehat{IE}_{AB|C^+}}{2} - \frac{\widehat{IE}_{AB|C^-}}{2}$$

By induction, an H-way interaction effect is defined as the average difference between the $(H-1)$-way interaction effects when the H^{th} factor is at its high vs. low level.

Table 24: Design matrix and response observations for the toy 2^2 experiment.

Condition	Factor A	Factor B	Response (y)	Average Response ($\overline{y}$)
1	−1	−1	{1,1,2}	4/3
2	+1	−1	{3,4,5}	12/3
3	−1	+1	{2,1,3}	6/3
4	+1	+1	{1,2,5}	8/3

These calculations may be visualized in reference to the 2^2 design's square. Figure 23 visualizes the comparisons that are made in the main and interaction effect calculations. In particular, the main effect of factor A is the difference between the average of response averages calculated at the high vs. low level of A. As we can see in the lefthand figure, this is exactly the average of the response averages at the rightmost corners of the square, minus the average of the response averages at the leftmost corners of the square. Similarly, as is visualized in the middle figure, the main effect of B is the average of the response averages at the topmost corners of the square, minus the average of the response averages at the bottommost corners of the square. Finally, the A:B interaction effect is the difference of the average response averages in opposing corners of the square. This is visualized in the righthand figure.

In this example we have used a non-binary response variable for illustration. However, the comparisons visualized in Figure 23 are still relevant when the response variable is binary. The main difference is that at each corner we calcluate the *odds* that $Y = 1$, not the average response. A second difference is that rather than defining effects as differences of arithmetic means as we have done with the non-binary data, we define effects as ratios of geometric means of the odds:

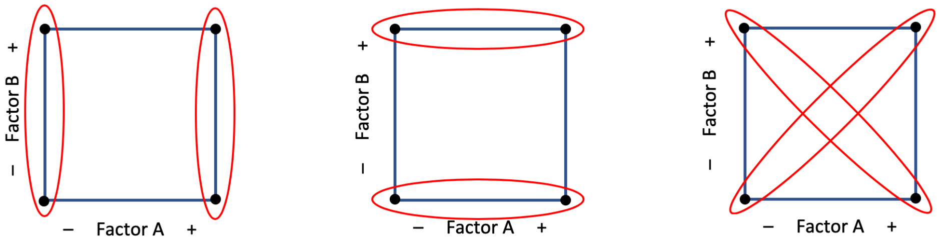

Figure 23: Visualization of main and interaction effects in a 2^2 factorial experiment.

$$\widehat{ME}_A = \sqrt{\frac{\overline{y}_{A+\cap B-}}{1-\overline{y}_{A+\cap B-}} \times \frac{\overline{y}_{A+\cap B+}}{1-\overline{y}_{A+\cap B+}}} \div \sqrt{\frac{\overline{y}_{A-\cap B-}}{1-\overline{y}_{A-\cap B-}} \times \frac{\overline{y}_{A-\cap B+}}{1-\overline{y}_{A-\cap B+}}} \tag{38}$$

$$\widehat{ME}_B = \sqrt{\frac{\overline{y}_{A-\cap B+}}{1-\overline{y}_{A-\cap B+}} \times \frac{\overline{y}_{A+\cap B+}}{1-\overline{y}_{A+\cap B+}}} \div \sqrt{\frac{\overline{y}_{A+\cap B-}}{1-\overline{y}_{A+\cap B-}} \times \frac{\overline{y}_{A-\cap B-}}{1-\overline{y}_{A-\cap B-}}} \tag{39}$$

$$\widehat{IE}_{AB} = \sqrt{\frac{\overline{y}_{A+\cap B+}}{1-\overline{y}_{A+\cap B+}} \times \frac{\overline{y}_{A-\cap B-}}{1-\overline{y}_{A-\cap B-}}} \div \sqrt{\frac{\overline{y}_{A+\cap B-}}{1-\overline{y}_{A+\cap B-}} \times \frac{\overline{y}_{A-\cap B+}}{1-\overline{y}_{A-\cap B+}}} \tag{40}$$

6.2.2 A Regression-Based Analysis

The Model

Formal analysis of a 2^K factorial experiment is carried out via regression: linear regression (when Y is continuous) and logistic regression (when Y is binary). The corresponding fitted models provide an estimate of the **response surface** that relates the response variable to the K design factors. As discussed in Section 6.1, each of these factors is represented by the binary variables

$$x_j = \begin{cases} -1 & \text{if factor } j \text{ is at its ``low'' level} \\ +1 & \text{if factor } j \text{ is at its ``high'' level} \end{cases}$$

for $j = 1, 2, \ldots, K$. Since each factor is represented by a single term, the linear predictor for these models contains an intercept, K main effect terms, $\binom{K}{2}$ two-factor interaction terms, $\binom{K}{3}$ three-factor interaction terms and so on, up to and including the $\binom{K}{K} = 1$ K-factor interaction term. Thus, the linear predictor contains $\sum_{k=0}^{K} \binom{K}{k} = 2^K$ terms. For instance, the linear predictor associated with a 2^3 factorial experiment is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$$

Estimation

Estimation of the β's is carried out by ordinary least squares (in the case of linear regression) and maximum likelihood (in the case of logistic regression). In both cases we find that there is a one-to-one connection between the β estimates and the expressions for the main and interaction effects defined in Section 6.2.1.

We illustrate the connection here for the case of linear regression and leave the logistic regression case as an exercise for the reader.

Consider, again, the toy example summarized in Table 24. The linear predictor for that experiment is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

where x_1 and x_2 respectively represent factors A and B. The linear regression model is thus

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

for $i = 1, 2, \ldots, N = 12$. This model may be re-expressed in matrix-vector notation as

$$\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon$$

where, in general, $\mathbf{Y}$ is an $N \times 1$ vector of response observations, ε is an $N \times 1$ random vector of error terms, $\boldsymbol{\beta}$ is a $2^K \times 1$ vector of regression coefficients, and X is the $N \times 2^K$ **model matrix** containing plus and minus ones. Note that each column represents a different term in the linear predictor, and interaction columns are obtained by the elementwise multiplication of the main effect columns involved in the interaction. Constructed in this way, the model matrix (like the design matrix) is orthogonal. It can be verified that $X^T X$ is a diaginal matrix, and in particular that $X^T X = N I_{2^K}$ and hence

$$(X^T X)^{-1} = \frac{1}{N} I_{2^K}$$

where I_p is the $p \times p$ identity matrix. Thus, the least squares estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} = \frac{X^T \mathbf{Y}}{N}$$

For the toy example we have:

$$\mathbf{Y} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 2 \\ 1 \\ 3 \\ 1 \\ 2 \\ 5 \end{bmatrix} \qquad X = \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & -1 & -1 & +1 \\ +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & +1 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix}$$

The least squares estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 5/2 \\ 10/12 \\ -1/6 \\ -1/2 \end{bmatrix}.$$

Notice that

$$2\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 5 \\ 10/6 \\ -1/3 \\ -1 \end{bmatrix} = \begin{bmatrix} 2\overline{y} \\ \widehat{ME}_A \\ \widehat{ME}_B \\ \widehat{IE}_{AB} \end{bmatrix}$$

This is no coincidence. In general, due to the orthogonality of the model matrix, we have for any β

$$\widehat{\beta} = \frac{\mathbf{x}^T\mathbf{Y}}{N} = \frac{\mathbf{x}^T\mathbf{Y}}{n2^K}$$

where $\mathbf{x}$ is the column of X corresponding to the particular β. Thus, any effect (whether main or interaction) is estimated as

$$\widehat{\text{Effect}} = 2\widehat{\beta} = \frac{\mathbf{x}^T\mathbf{Y}}{n2^{K-1}}$$

where $\mathbf{x}$ is the column of X corresponding to the effect of interest, and β is the corresponding regression coefficient. That an effect is equal to $2\widehat{\beta}$ should be unsurprising. Ordinarily β quantifies the change in expected response when the corresponding x is increased by *one unit*. Since we have defined a factor's effect to be the change in response expected when the factor moves from its low (-1) to its high $(+1)$ level, we are considering a *two unit* increase.

The preceding discussion assumed a continuous response and linear regression. Though not derived explicitly here, when the response variable is binary, closed-form estimates of the β's in a logistic regression are available, and also related to the effect estimates discussed in Section 6.2.1. It can be shown that the effect estimates defined by the ratios of geometric means of odds in equations (38)-(40) may also be computed by:

$$\widehat{\text{Effect}} = e^{2\widehat{\beta}}$$

where β is the regression coefficient corresponding to the effect of interest.

Hypothesis Testing

As in general factorial experiments, the significance of main and interaction effects is determined by testing hypotheses that set the relevant β's equal to 0. This is still the manner in which significance is established, but the hypotheses and tests are simplified in the context of the 2^K factorial experiment due to the fact that each factor is experimented with at just two levels, and is hence represented by just one term in the model. For this reason, the hypotheses we tend to consider involve just a single β. For instance, the linear predictor for the toy example is $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$. If we wish to determine the significance of factor A we simply test

$$H_0 : \beta_1 = 0$$

or if we want to determine whether the A:B interaction is significant, we test

$$H_0 : \beta_{12} = 0.$$

Hypotheses of this form may be tested with the usual significance tests for individual regression coefficients. In the case of linear regression t-tests are used to evaluate the significance of individual β's while Z-tests are used for this purpose in the context of logistic regression. If one wishes to test the simultaneous elimination of several terms all at once, we may, as usual, use partial F-tests in the case of linear regression, and likelihood ratio tests in the case of logistic regression.

6.2.3 The Credit Card Example

To illustrate a complete analysis of a 2^K factorial experiment, we consider an example from Montgomery (2019) in which an experiment was performed to test new ideas to improve the conversion rate of credit card offers. For this example, the response is binary – indicating whether an individual signed up for a credit card as a result of the offer – and so an analysis based on logistic regression is performed.

A 2^4 factorial experiment was carried out to investigate four factors and their influence on credit card sign ups. The four factors and each of their levels are summarized in Table 25. The $2^4 = 16$ unique combinations of these factor levels produced 16 experimental conditions, each of which was assigned $n = 7500$ units. Practically speaking, 16 credit card offers were devised (one corresponding to each condition) and each was mailed to 7500 customers. The design matrix and a summary of the conversion rates are provided in Table 26.

Table 25: Factors and levels for the credit card example.

Factor	Low (−)	High (+)
Annual Fee (x_1)	Current	Lower
Account-Opening Fee (x_2)	No	Yes
Initial Interest Rate (x_3)	Current	Lower
Long-term Interest Rate (x_4)	Low	High

Table 26: Design matrix and response summary for the 2^4 factorial credit card experiment.

Condition	Factor 1	Factor 2	Factor 3	Factor 4	Sign-ups	Conversion Rate
1	−1	−1	−1	−1	184	2.45%
2	+1	−1	−1	−1	252	3.36%
3	−1	+1	−1	−1	162	2.16%
4	+1	+1	−1	−1	172	2.29%
5	−1	−1	+1	−1	187	2.49%
6	+1	−1	+1	−1	254	3.39%
7	−1	+1	+1	−1	174	2.32%
8	+1	+1	+1	−1	183	2.44%
9	−1	−1	−1	+1	138	1.84%
10	+1	−1	−1	+1	168	2.24%
11	−1	+1	−1	+1	127	1.69%
12	+1	+1	−1	+1	140	1.87%
13	−1	−1	+1	+1	172	2.29%
14	+1	−1	+1	+1	219	2.92%
15	−1	+1	+1	+1	153	2.04%
16	+1	+1	+1	+1	152	2.03%

Using this data we fit a logistic regression model with the following linear predictor

$$\begin{aligned}
\beta_0 \;&+\; \beta_1 x_1 \;+\; \beta_2 x_2 \;+\; \beta_3 x_3 \;+\; \beta_4 x_4 \\
&+\; \beta_{12} x_1 x_2 \;+\; \beta_{13} x_1 x_3 \;+\; \beta_{14} x_1 x_4 \;+\; \beta_{23} x_2 x_3 \;+\; \beta_{24} x_2 x_4 \;+\; \beta_{34} x_3 x_4 \\
&+\; \beta_{123} x_1 x_2 x_3 \;+\; \beta_{124} x_1 x_2 x_4 \;+\; \beta_{134} x_1 x_3 x_4 \;+\; \beta_{234} x_2 x_3 x_4 \\
&+\; \beta_{1234} x_1 x_2 x_3 x_4
\end{aligned}$$

We do this in `R` using the `glm()` function with a logit link. A summary of this model is provided below.

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -3.739697   0.019342 -193.347  < 2e-16 ***
x1           0.080845   0.019342    4.180 2.92e-05 ***
x2          -0.106211   0.019342   -5.491 3.99e-08 ***
x3           0.058248   0.019342    3.011  0.00260 **
x4          -0.108086   0.019342   -5.588 2.29e-08 ***
x1:x2       -0.055164   0.019342   -2.852  0.00434 **
x1:x3       -0.004794   0.019342   -0.248  0.80426
x2:x3       -0.006967   0.019342   -0.360  0.71868
x1:x4       -0.013178   0.019342   -0.681  0.49566
x2:x4        0.010625   0.019342    0.549  0.58280
x3:x4        0.038079   0.019342    1.969  0.04899 *
x1:x2:x3    -0.009646   0.019342   -0.499  0.61799
x1:x2:x4     0.010629   0.019342    0.550  0.58265
x1:x3:x4    -0.002543   0.019342   -0.131  0.89539
x2:x3:x4    -0.020946   0.019342   -1.083  0.27885
x1:x2:x3:x4 -0.009496   0.019342   -0.491  0.62347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we can see, all main effects are significant but there are several insignificant terms; in particular, all three-factor and four-factor interactions are insignificant, and only two of the two-factor interactions are significant. In light of this, we fit a reduced model with $\beta_{13} = \beta_{14} = \beta_{23} = \beta_{24} = \beta_{123} = \beta_{124} = \beta_{134} = \beta_{234} = \beta_{1234} = 0$. The output from this reduced model is shown below. While the Z-test p-values in the full model suggested this reduced model is appropriate, it is wise to formally test

$$H_0: \; \beta_{13} = \beta_{14} = \beta_{23} = \beta_{24} = \beta_{123} = \beta_{124} = \beta_{134} = \beta_{234} = \beta_{1234} = 0$$

with a likelihood ratio test. The maximized log-likelihood for the full model is -13370.45 and for the reduced model is -13371.91 (obtained using the `logLik()` function in `R`) which gives a likelihood ratio test statistic of $t = 2.9244$ that corresponds to a p-value of $P(T \geq 2.9244) = 0.9672$, where $T \sim \chi^2_{(9)}$. Thus we do not reject H_0, implying that the reduced model is adequate and that only the main effects and two of the two-factor interactions are significant.

```
Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -3.73961    0.01934 -193.316  < 2e-16 ***
x1           0.08214    0.01920    4.279 1.88e-05 ***
x2          -0.10834    0.01920   -5.644 1.66e-08 ***
x3           0.05886    0.01916    3.072  0.00212 **
x4          -0.11068    0.01916   -5.777 7.61e-09 ***
x1:x2       -0.05706    0.01920   -2.972  0.00296 **
x3:x4        0.04051    0.01916    2.115  0.03447 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Main effect and interaction effect plots are shown in Figures 24 and 25. Unsurprisingly, the main effect plots demonstrate that lower annual fees, no account-opening fee, lower initial interest rates and lower long-term interest rates are all associated with increased conversion rates. Furthermore, the interaction plots indicate that a lower annual fee is especially effective at increasing the conversion rate when there is no account-opening fee; when there is an account opening fee, the annual fee is not as influential. We can also see that as long as the long-term interest rate is low, the initial interest rate doesn't really matter.

These plots and the regression results provide very useful insight into determining which factors most influence credit card initiation, and which combination of factor levels is optimal. In the context of these 16 experimental conditions, the three best conversion rates occurred in conditions 2, 6 and 14 where the conversion rates were 3.36%, 3.39% and 2.92%, respectively. The p-value associated with a χ^2-test of $H_0: \pi_2 = \pi_6 = \pi_{14}$ is 0.1917, indicating no significant difference between these conditions. Presumably the credit card company would choose to implement the one that is most profitable for them.

6.3 A Note on 3^K Factorial Experiments

A larger alternative to two-level screening experiments are three-level screening experiments in which each factor is experimented with at three levels. The 3^K factorial experiment is one such three-level experiment. Just like a 2^K factorial experiment, the 3^K factorial experiment investigates K design factors and the experimental conditions are defined by the unique combinations of these factors' levels.

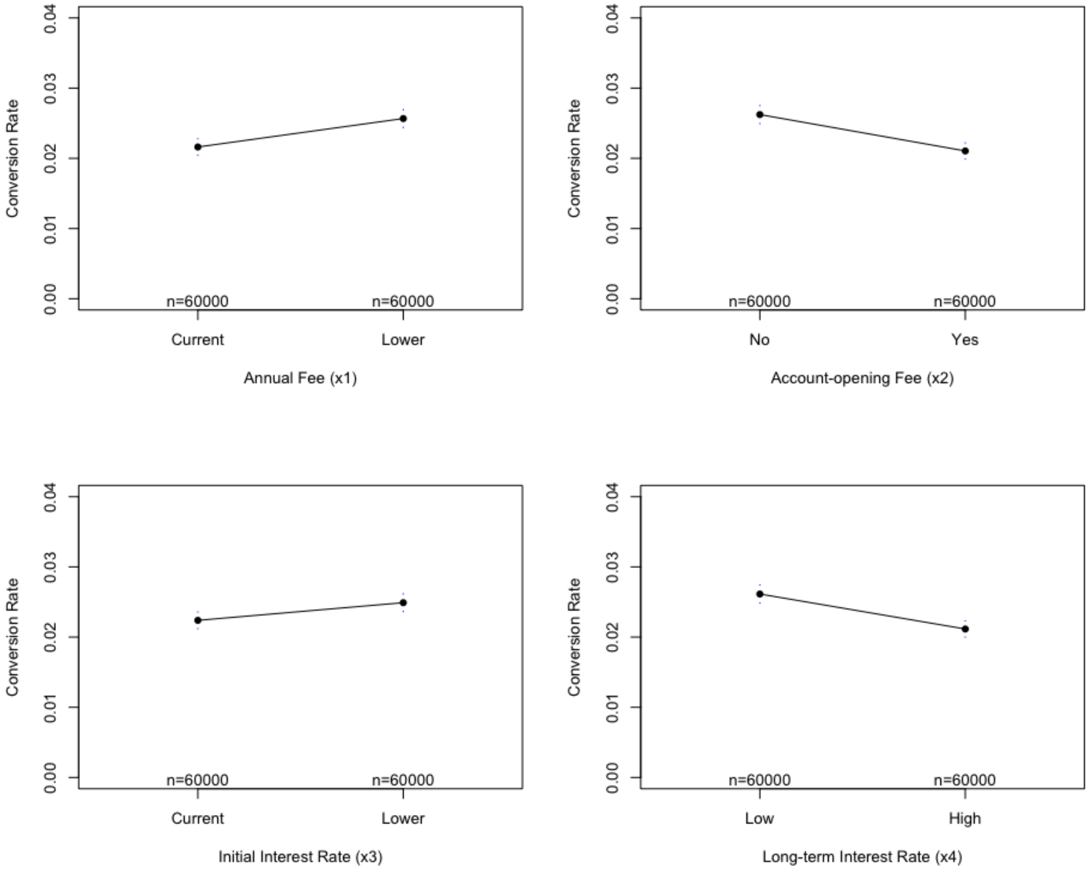

Figure 24: Main effect plots for the credit card example.

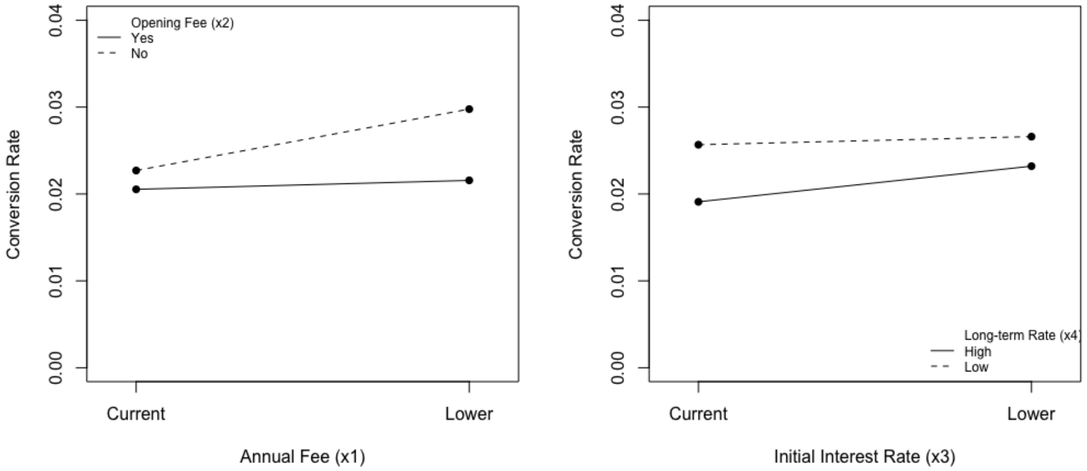

Figure 25: Interaction effect plots for the credit card example.

In such experiments it is common practice to label the three levels as *low*, *intermediate* and *high*, and represent each factor by a ternary variable such as:

$$x = \begin{cases} -1 & \text{if the factor is at its low level} \\ 0 & \text{if the factor is at its intermediate level} \\ +1 & \text{if the factor is at its high level} \end{cases}$$

As such, like a 2^K design, a 3^K design can be visualized in terms of K-dimensional hypercubes, but where the experimental conditions correspond not only to the vertices, but also the centers of each edge and each face. Figure 26 presents a visualization of 3^2 and 3^3 factorial designs.

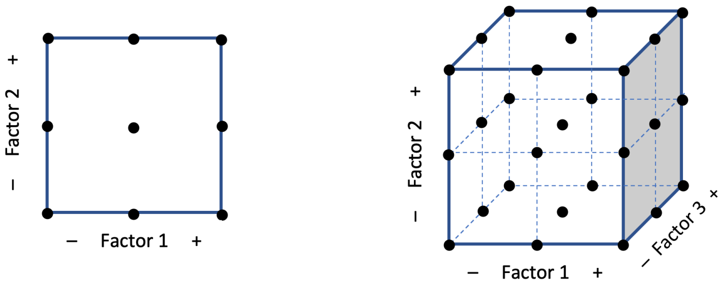

Figure 26: Cuboidal representation of 3^2 (left) and 3^3 (right) factorial designs.

A clear disadvantage of the 3^K factorial experiment is that it requires many more experimental conditions (3^K vs. 2^K). However, if such an experiment is practically feasible it provides much more information about the nature of the relationship between the response and design factors. It allows one to model curvature via quadratic terms in the linear predictor. Such models are the focus of Chapter 8 and quite useful in the context of response optimization. We defer further discussion of these designs and these models until then.

7 2^{K-p} FRACTIONAL FACTORIAL EXPERIMENTS

As we have seen, the 2^K factorial experiment is an economical special case of a general factorial experiment since it minimizes the number of levels being investigated, and hence reduces the overall number of experimental conditions. However, the 2^K factorial experiment still investigates *all possible* combinations of the factor levels, and 2^K can still be a very large number of conditions even for moderate K. For instance, an experiment with just $K = 8$ factors requires $2^8 = 256$ distinct experimental conditions – which may be unmanageably large.

A 2^{K-p} fractional factorial experiment, on the other hand, similarly investigates K factors but in just a fraction of the conditions; specifically, $(1/2)^p$ as many. Thus, rather than experimenting with all 2^K conditions, we specially select 2^{K-p} of them. When $p = 1$ we investigate K factors in half as many conditions – we call this a *one-half fraction* of the 2^K design. When $p = 2$ we investigate K factors in a quarter of the conditions – we call this a *one-quarter fraction*. The value p dictates the degree of *fractioning* and is typically chosen to minimize the number of experimental conditions, given a fixed number of design factors. However, it it may also be chosen to maximize the number of design factors explored in a fixed number of conditions. Note that for a given value of K the value of p must satisfy the constraints $p \in \mathbb{Z}^+$, $1 \leq p < K$, and $2^{K-p} > K$.

From either perspective, the choice of *which* 2^{K-p} conditions to execute is of particular importance. Given that the primary goal of such an experiment is factor screening we want to choose the conditions in such a way that ensures all important main effects and interaction effects can be identified. But as we will see, our effects will no longer be separately estimable. This is the sacrifice we make in order to investigate a relatively large number of factors with a relatively small number of conditions.

To further motivate the utility of a fractional factorial experiment, consider the linear predictor from the full 2^K factorial experiment, which contains K main effect terms, $\binom{K}{2}$ two-factor interaction terms, $\binom{K}{3}$ three-factor interaction terms and so on, up to and including the $\binom{K}{K} = 1$ K-factor interaction term; this is a total of $\sum_{k=1}^{K} \binom{K}{k} = 2^K - 1$ estimated effects. However, just $K + \binom{K}{2}$ of these are main effects and two-factor interactions – the remaining regression coefficients correspond to higher order interaction effects. In the case that $K = 8$, there are 8 main effects, 28 two-factor interactions and 219 higher order interactions, many of which are likely to be insignificant.

The principle of **effect sparsity** says that in the presence of several factors, variation in the response is likely to be driven by a small number of main effects and low-order interactions. You may recall we saw this in the credit card example in Section 6.2.3: none of the three-factor or four-factor interactions were significant, and in fact many of the two-factor interactions were insignificant. Based on a meta-analysis of over 100 published studies, Li et al. (2006) found empirically that roughly 41% of main effects were significant, only 11% of two-factor interactions were significant, and just 6% of interactions involving three or more factors were significant. In light of effect sparsity, it may be viewed as a waste of resources to estimate higher order interaction terms, and it would be a better use of resources to estimate the main effects and low-order interactions of a larger number of factors. This is the driving philosophy behind fractional factorial experiments.

With this in mind, let us define three illustrative examples that will be referred back to regularly throughout this chapter.

- **The 2^{3-1} Example:** In this example we consider a one-half fraction of the 2^3 design which explores $K = 3$ factors (A,B,C) in $m = 4$ conditions rather than 8. The design matrix associated with a full 2^3 design is shown in Table 27 and the full 2^3 design is visualized in Figure 27. The question of primary interest is: *which* $m = 4$ conditions do we choose for the 2^{3-1} experiment?

Table 27: The design matrix for a full 2^3 factorial experiment.

Condition	Factor A	Factor B	Factor C
1	-1	-1	-1
2	$+1$	-1	-1
3	-1	$+1$	-1
4	$+1$	$+1$	-1
5	-1	-1	$+1$
6	$+1$	-1	$+1$
7	-1	$+1$	$+1$
8	$+1$	$+1$	$+1$

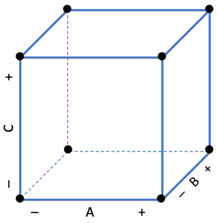

Figure 27: Cuboidal visualization of full 2^3 factorial design.

- **The 2^{4-1} Example:** In this example we consider a one-half fraction of the 2^4 design which explores $K = 4$ factors (A,B,C,D) in $m = 8$ conditions rather than 16. The design matrix associated with a full 2^4 design is shown in Table 28 and the full 2^4 design is visualized in Figure 28. Similar to the 2^{3-1} example, the question of primary interest is: *which* $m = 8$ conditions do we choose for the 2^{4-1} experiment?

- **The 2^{5-2} Example:** In this example we consider a one-quarter fraction of the 2^5 design which

Table 28: The design matrix for a full 2^4 factorial experiment.

Condition	Factor A	Factor B	Factor C	Factor D
1	-1	-1	-1	-1
2	$+1$	-1	-1	-1
3	-1	$+1$	-1	-1
4	$+1$	$+1$	-1	-1
5	-1	-1	$+1$	-1
6	$+1$	-1	$+1$	-1
7	-1	$+1$	$+1$	-1
8	$+1$	$+1$	$+1$	-1
9	-1	-1	-1	$+1$
10	$+1$	-1	-1	$+1$
11	-1	$+1$	-1	$+1$
12	$+1$	$+1$	-1	$+1$
13	-1	-1	$+1$	$+1$
14	$+1$	-1	$+1$	$+1$
15	-1	$+1$	$+1$	$+1$
16	$+1$	$+1$	$+1$	$+1$

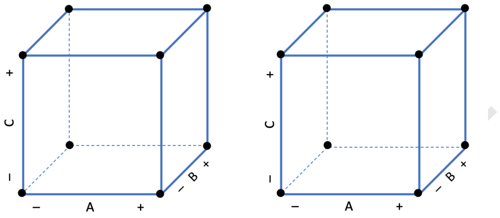

Figure 28: Cuboidal visualization of full 2^4 factorial design.

explores $K = 5$ factors (A,B,C,D,E) in $m = 8$ conditions rather than 32. The design matrix associated with a full 2^5 design is shown in Table 29 and the full 2^5 design is visualized in Figure 29. Similar to the previous two examples, the question of primary interest is: *which* $m = 8$ conditions do we choose for the 2^{5-2} experiment?

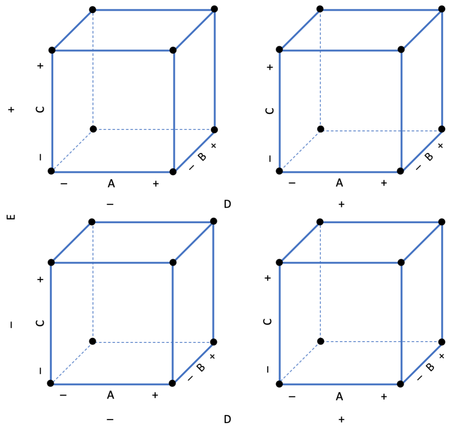

Figure 29: Cuboidal visualization of full 2^5 factorial design.

Table 29: The design matrix for a full 2^5 factorial experiment.

Condition	Factor A	Factor B	Factor C	Factor D	Factor E
1	-1	-1	-1	-1	-1
2	$+1$	-1	-1	-1	-1
3	-1	$+1$	-1	-1	-1
4	$+1$	$+1$	-1	-1	-1
5	-1	-1	$+1$	-1	-1
6	$+1$	-1	$+1$	-1	-1
7	-1	$+1$	$+1$	-1	-1
8	$+1$	$+1$	$+1$	-1	-1
9	-1	-1	-1	$+1$	-1
10	$+1$	-1	-1	$+1$	-1
11	-1	$+1$	-1	$+1$	-1
12	$+1$	$+1$	-1	$+1$	-1
13	-1	-1	$+1$	$+1$	-1
14	$+1$	-1	$+1$	$+1$	-1
15	-1	$+1$	$+1$	$+1$	-1
16	$+1$	$+1$	$+1$	$+1$	-1
17	-1	-1	-1	-1	$+1$
18	$+1$	-1	-1	-1	$+1$
19	-1	$+1$	-1	-1	$+1$
20	$+1$	$+1$	-1	-1	$+1$
21	-1	-1	$+1$	-1	$+1$
22	$+1$	-1	$+1$	-1	$+1$
23	-1	$+1$	$+1$	-1	$+1$
24	$+1$	$+1$	$+1$	-1	$+1$
25	-1	-1	-1	$+1$	$+1$
26	$+1$	-1	-1	$+1$	$+1$
27	-1	$+1$	-1	$+1$	$+1$
28	$+1$	$+1$	-1	$+1$	$+1$
29	-1	-1	$+1$	$+1$	$+1$
30	$+1$	-1	$+1$	$+1$	$+1$
31	-1	$+1$	$+1$	$+1$	$+1$
32	$+1$	$+1$	$+1$	$+1$	$+1$

7.1 Designing 2^{K-p} Fractional Factorial Experiments

In this section we answer the question: *how do we choose the 2^{K-p} fractional factorial conditions from the full 2^K design?*

7.1.1 Aliasing

The first step in constructing a 2^{K-p} fractional factorial design is to write out the model matrix (when $n = 1$) for a *full* 2^{K-p} design. For instance, in the 2^{3-1} example, to figure out which four conditions to choose from the full 2^3 design, we begin by considering the model matrix (when $n = 1$) for a full 2^2 design with factors A and B. This matrix is shown in Table 30.

We now frame the question as: in the full 2^2 design, in which of the four conditions will factor C be run at its low level and in which will factor C be run at its high level? We use the ±1's in the AB interaction column to dictate, for a given condition, whether to run factor C at its low or high levels. In conditions 1

Table 30: The model matrix (when $n = 1$) associated with a full 2^2 factorial experiment.

Condition	I	A	B	AB
1	$+1$	-1	-1	$+1$
2	$+1$	$+1$	-1	-1
3	$+1$	-1	$+1$	-1
4	$+1$	$+1$	$+1$	$+1$

and 4, $AB = +1$ so we know to run factor C at its high level in these conditions. Similarly, since $AB = -1$ in conditions 2 and 3 we know that factor C should be run at its low level in conditions 2 and 3. What results is a prescription for experimenting with $K = 3$ factors in $2^{3-1} = 4$ conditions. The resulting 2^{3-1} fractional factorial design is visualized in Figure 30 with red points indicating which among the possible 2^3 conditions are actually experimented with. We can see that these are conditions 2, 3, 5, 8. We refer to these four conditions as the **principal fraction**. The other four conditions (1,4,6,7) also constitute a 2^{3-1} fractional factorial design – they are referred to as the **alternate** or **complementary fraction** (see the black points in Figure 30). Whereas the principal fraction arose by associating the levels of C with the ±1's in the AB column, the complementary fraction would have arisen by associating C with $-AB$.

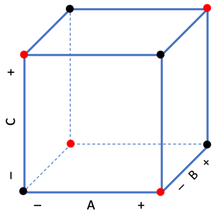

Figure 30: Cuboidal visualization of the 2^{3-1} fractional factorial design. Red points indicate the conditions associated with the principal fraction; black points indicate the conditions associated with the complentary fraction.

This association of new factors to existing interactions is referred to as **aliasing**. In the example just discussed, we aliased C with the AB interaction. We denote this aliasing as $C = AB$ and we call '$C = AB$' the **design generator**. When we do this, we **confound** the interaction effect with the main effect of the new factor, meaning that they cannot be separately estimated. As was discussed in the previous chapter, the orthogonality of the model matrix ensures that each column is individually responsible for estimating a single effect. In an ordinary 2^2 experiment with factors A and B, the A and B columns in Table 30 are used to estimate ME_A and ME_B, and the AB column is used to estimate IE_{AB}. But due to the aliasing in the experiment described above, the $AB = C$ column now jointly quantifies the main effect of C and the AB interaction effect – we cannot estimate these effects separately. As is discussed in the next section, this confounding is not restricted to just C and AB ; all effects are now confounded. This is the price we pay for using fewer conditions than what is prescribed by the full 2^K design.

7.1.2 The Defining Relation

In the 2^{3-1} example discussed in the previous section we aliased C with the AB interaction. However, the aliasing (and hence confounding) doesn't stop there. Upon closer inspection we find that the main effects of A and B are now also aliased with interaction effects. To see this, notice that the design generator $C = AB$ implies $I = ABC$, where I is the identity columns of +1's in Table 30. This implication is seen by recognizing that the elementwise multiplication of any column with itself yields the identity column. The equation $I = ABC$ is referred to as the **defining relation** and may be used to uncover all aliases. For instance, to determine the aliasing associated with the main effect of A, we perform elementwise multiplication between column A and the defining relation:

$$\begin{aligned} A \times I &= A \times ABC \\ A &= A^2BC \\ A &= BC \end{aligned}$$

Similar calculations with B and C yield $B = AC$ and $C = AB$. Thus, every main effect is aliased with a two factor interaction. Clearly, introducing aliasing anywhere causes confounding everywhere. In this design we cannot separately estimate the main effect of A from the BC interaction, the main of effect of B from the AC interaction, or the main effect of C from the AB interaction.

2^{4-1} Example

To illustrate some slightly more complicated defining relations (and hence aliasing structures) we consider the 2^{4-1} and 2^{5-2} fractional factorial designs which both use $m = 8$ conditions to explore $K = 4$ and $K = 5$ factors, respectively. To construct both factorial designs we consider the model matrix (when $n = 1$) associated with a full 2^3 design. This matrix is shown in Table 31.

Table 31: The model matrix (when $n = 1$) associated with a full 2^3 factorial experiment.

Condition	I	A	B	C	AB	AC	BC	ABC
1	+1	−1	−1	−1	+1	+1	+1	−1
2	+1	+1	−1	−1	−1	−1	+1	+1
3	+1	−1	+1	−1	−1	+1	−1	+1
4	+1	+1	+1	−1	+1	−1	−1	−1
5	+1	−1	−1	+1	+1	−1	−1	+1
6	+1	+1	−1	+1	−1	+1	−1	−1
7	+1	−1	+1	+1	−1	−1	+1	−1
8	+1	+1	+1	+1	+1	+1	+1	+1

For the 2^{4-1} design we must choose one interaction column to alias a new factor D with, and hence decide which column will dictate factor D's levels. Our choices are AB, AC, BC and ABC. Since the effect sparsity principle claims that higher order interaction terms are likely to be neglible, it seems sensible to alias D with the highest order interaction (i.e., $D = ABC$) thereby avoiding confounding main effects with two-factor interactions, which are more likely to both be significant. Doing so yields the defining relation $I = ABCD$ which gives rise to the **complete aliasing structure** for the design:

$$A = BCD$$

$$B = ACD$$

$$C = ABD$$

$$D = ABC$$

$$AB = CD$$

$$AC = BD$$

$$AD = BC$$

However, we could have chosen $D = AB$ or $D = AC$ or $D = BC$ as design generators instead of $D = ABC$, and each would have yielded a different (but valid) 2^{4-1} fractional factorial design. These four designs are visualized in Figure 31, with red points indicating the principal fraction and black points indicating the complementary fraction That would have arisen by taking $D = -ABC$, $D = -AB$, $D = -AC$, or $D = -BC$). Two relevant questions to ask here are: is it possible for one design to be *better* than another? And if so, which design is best? We answer these questions in Section 7.1.3.

2^{5-2} Example

For the 2^{5-2} design, in addition to choosing an alias for factor D like we just did for the 2^{4-1} design, we also need to choose an alias for factor E. Suppose that we choose $E = BC$. The resulting design is visualized in Figure 32. We now have *two* design generators $D = ABC$ and $E = BC$. Note that the number of design generators will always equal p, the degree of fractioning. These design generators each give rise to a defining relation: $I = ABCD$ and $I = BCE$. Notice that these two defining relations can be combined into a single *complete* defining relation by multiplying them:

$$I = ABCD = BCE = ABCD \times BCE$$
$$I = ABCD = BCE = AB^2C^2DE$$
$$I = ABCD = BCE = ADE$$

This complete defining relation may be used as usual to determine the complete aliasing structure. Note that the defining relation here has three terms (in addition to the identity term), implying that in this design every effect is aliased with three other effects. This is verified in the complete aliasing structurue below.

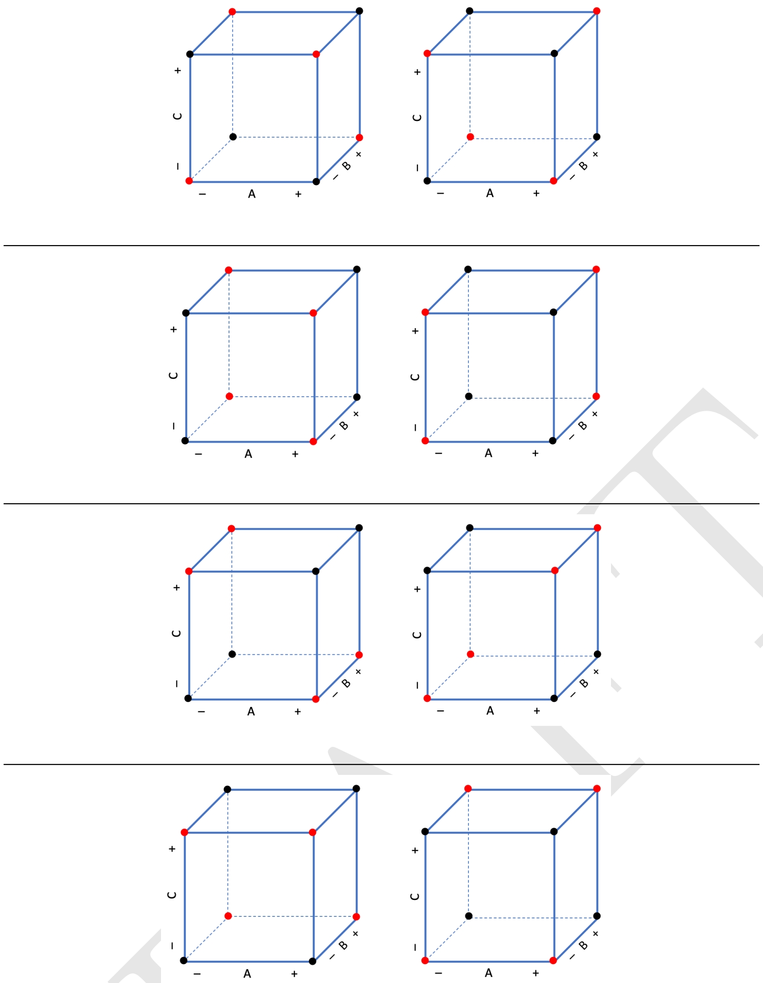

Figure 31: Cuboidal visualization of the 2^{4-1} fractional factorial designs. From the top row to the bottom row the design generators are $D = ABC$, $D = AB$, $D = AC$, $D = BC$. Red points indicate the $m = 8$ conditions included in the design.

Note that in general the number of effects aliased with a given effect is $2^p - 1$. Thus, in a 2^{K-p} fractional factorial design, every effect estimate actually jointly quantifies 2^p effects. We elaborate on the implications of this fact in Section 7.2.

$$A = BCD = ABCE = DE$$

$$B = ACD = CE = ABDE$$

$$C = ABD = BE = ACDE$$

$$D = ABC = BCDE = AE$$

$$E = ABCDE = BC = AD$$

$$AB = CD = ACE = BDE$$

$$AC = BD = ABE = CDE$$

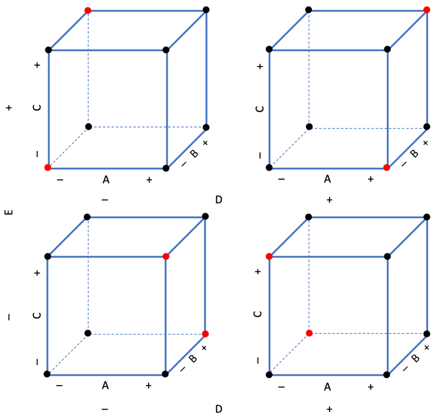

Figure 32: Cuboidal visualization of a 2^{5-2} fractional factorial design. Red points indicate the $m = 8$ conditions included in the design.

Thus, a 2^{K-p} fractional factorial experiment is designed by looking at the model matrix (with $n = 1$) for a full 2^{K-p} design with $K - p$ factors, and choosing p interaction columns to alias an additional p factors with. The ±1's in these columns are then used to dictate, for each condition, whether the p additional factors are run at their low or high levels. But the question still remains, how do we know *which* interactions to choose?

7.1.3 Resolution

Due to the confounding that results from aliasing a new main effect with an existing interaction, it is important to think carefully about *which* interaction to choose as an alias. If at all possible, it is best to avoid aliasing a new factor with an interaction that is likely to be significant since separately estimating these significant effects is desirable. As such, high order interaction terms (that are unlikely to be significant) are good choices for aliases.

This notion is quantified by the **resolution** of the fractional factorial design. In general, a design is of resolution R (typically denoted by Roman numerals) if main effects are aliased with interaction effects involving at least $R - 1$ factors. In the 2^{3-1} and 2^{5-2} designs discussed above, main effects are aliased with some two-factor interactions and so these designs are resolution III. In the 2^{4-1} design discussed above, main effects are aliased with some three-factor interactions and so this design is resolution IV.

The easiest way to determine R is by looking at the defining relation. Each of the terms in the equivalence is referred to as a *word*. The length of the shortest word is the resolution of the design. Recall the defining relations for the 2^{3-1}, 2^{4-1}, and 2^{5-2} designs:

$$I = ABC$$

$$I = ABCD$$

$$I = ABCD = BCE = ADE$$

The shortest words in these three relations have respectively 3, 4, and 3 letters. Thus the resolutions are III, IV, and III. Using the notation 2_R^{K-p} these three designs may be completely described as 2_{III}^{3-1}, 2_{IV}^{4-1}, and 2_{III}^{5-2}.

In general, higher resolution designs are to be preferred over lower resolution designs. For instance, resolution IV and V designs are to be preferred over a resolution III design since in these cases main effects will not be confounded with two-factor interactions. Since two-factor interactions are often significant, it is best if their effects are not confounded with main effects. As we have seen, the resolution of a fractional factorial experiment is determined by two things:

1. The degree of fractioning desired (i.e., the size of p relative to K).

2. The design generators chosen for aliasing.

The number of factors (K) is typically pre-specified, and the degree of fractioning (p) is typically determined by resource constraints; i.e., how many experimental conditions can you feasibly manage? Thus, given K and p, we should choose design generators that *maximize resolution*. For an excellent guide that simplifies this decision, Wu and Hamada (2011) (Chapter 5) and Montgomery (2019) (Chapter 8) provide tables which for given values of K and p provide the design generators necessary to achieve a particular resolution.

Let us return to the 2^{4-1} example. As depicted in Figure 31, the four choices of the design generator give rise to four different 2^{4-1} designs. The defining relation for each choice of generator is shown in Table 32. We see that the design arising from $D = ABC$ is resolution IV, whereas the designs arising from any other choice of generator are resolution III. Thus, to maximize resolution (and hence avoid confounding main effects with two-factor interactions), $D = ABC$ should be chosen as the design generator and the design depicted at the top of Figure 31 should be used. Note that the complementary fraction ($D = -ABC$) would also be a resolution IV design.

Table 32: Design generators and defining relations for various 2^{4-1} designs.

Design Generator	Defining Relation
$D = ABC$	$I = ABCD$
$D = AB$	$I = ABD$
$D = AC$	$I = ACD$
$D = BC$	$I = BCD$

Another way to justify the maximum resolution criterion is by the **projective property** of fractional factorial designs. It can be shown that a resolution R fractional factorial design can be projected into a full factorial design on *any subset* of $R-1$ factors. To visualize this, consider the 2^{3-1}, 2^{4-1}, and 2^{5-2} designs visualized in Figures 30, 31, 32. As discussed above, these designs are respectively resolution III, IV, and III. The red points in these figures, when projected onto $R-1$ dimensional space, correspond to full 2^2, 2^3 and 2^2 factorial designs. This property can be exploited when analyzing the experimental data – if at most

$R-1$ factors are significant, they can be analyzed as a full factorial experiment without confounding. Thus, maximizing R maximizes the size of the projected full factorial design.

7.1.4 Minimum Aberration

In the previous subsection we introduced the maximum resolution criterion for choosing design generators. However, a secondary criterion, the **minimum aberration** criterion, may be used to select among various designs that all have the same resolution. Consider a 2^{7-2}_{IV} design which is resolution IV and explores $K = 7$ factors in $m = 32$ conditions. Three design generator configurations that all give rise to a 2^{7-2}_{IV} design are shown in Table 33.

Table 33: Design generators and defining relations for various 2^{7-2} designs.

Design	Design Generators	Defining Relation
1	$F = ABC$, $G = ABD$	$I = ABCF = ABDG = CDFG$
2	$F = ABC$, $G = CDE$	$I = ABCF = CDEG = ABDEFG$
3	$F = ABCD$, $G = ABCE$	$I = ABCDF = ABCEG = DEFG$

How should we choose among these? Is one better than the others? To answer these questions, consider the lengths of the words in each of these defining relations: they are (4,4,4), (4,4,6) and (4,5,5). We can compare these designs on the basis of how many words of length 4 appear in the defining relation. Design 3 minimizes this number, and hence minimizes the number main effects aliased with the lowest-order interactions (in this case three-factor interactions). In general, for a given resolution R the **minimum aberration design** is one which minimizes the number minimum-length words in the defining relation. These designs are preferred since they minimize the number times main effects are aliased with the lowest order ($(R-1)$-factor) interactions.

7.2 Analyzing 2^{K-p} Fractional Factorial Experiments

The analysis of a 2^{K-p} fractional factorial experiment is in fact not very different from the analysis procedures for 2^K factorial experiments discussed in Section 6.2. In particular, we visually summarize effects of interest via main and interaction effect plots, and regression models (linear or logistic, depending on the response variable) are used to test hypotheses of the form $H_0 : \beta = 0$ to determine whether a given effect is significantly different from zero. Such tests are carried out with t-tests in the case of linear regression and Z-tests in the case of logistic regression.

However, as discussed in the previous section, the fractional factorial design introduces confounding; two effects that are confounded cannot be separately estimated. In particular, just 2^{K-p} effects (and hence β's) can be estimated. Furthermore, each of these β's jointly quantifies 2^p different effects. As such, we need to be careful to account for this in our analysis. It is therefore important to know the complete aliasing structure of the design so as to be fully aware of *which* effects are confounded.

Accounting for this confounding is particularly important when interpreting effect estimates and evaluating their significance. For instance, suppose that in the 2^{5-2} design discussed in the previous section we find that the main effect of factor A is significant. Unfortunately, because A is aliased with BCD, $ABCE$, and DE we cannot be 100% certain that the significance of the effect is truly due to the main effect of factor A; it could also be due a significant BCD interaction, a significant $ABCE$ interaction, or a significant DE interaction. It is also possible that the individual A, BCD, $ABCE$, and DE effects are small, but cumulatively they are significant. The opposite is also true: two effects of similar magnitudes but opposing signs may cancel each other out when confounded. Therefore, it is possible that significant main effects could be masked by significant interaction effects, and vice versa.

It is therefore crucial to avoid confounding effects that are likely to be significant with other ones that are also likely to be significant. Thankfully, due to the effect sparsity principle, high resolution designs with minimal aberration should guard against issues like the ones raised in the previous paragraph. Such problems cannot be eliminated altogether, but they can be mitigated.

7.2.1 The Chehalem Example

To make these ideas more clear we consider an example from Montgomery (2019) in which a fractional factorial experiment was used in the production of wine to study the influence of a variety of factors on a particular vintage of Pinot Noir. In this experiment $K = 8$ factors were investigated each at two levels (the factors and their levels are shown in Table 34) which, if a full factorial experiment was used, would have required 256 conditions. However a 2^{8-4}_{IV} fractional factorial experiment was performed that required only 16 conditions. The response variable in this case is the rating of the wine as determined by 5 raters.

Table 34: Factors and levels for the wine example.

Factor	Low (−)	High (+)
Pinot Noir clone (A)	Pommard	Wadenswil
Oak type (B)	Allier	Troncias
Age of barrel (C)	Old	New
Yeast/skin contact (D)	Champagne	Montrachet
Stems (E)	None	All
Barrel toast (F)	Light	Medium
Whole cluster (G)	None	10%
Fermemtation temperature (H)	Low (75°F max)	High (92°F max)

Thus, 16 different wines were produced (based on the 16 unique combinations of these factors' levels) and $n = 5$ raters tasted and rated each of them (low scores are good, large scores are bad). The design matrix and a summary of the response is provided in Table 35. It is easily verified that the $p = 4$ design generators are $E = BCD$, $F = ACD$, $G = ABC$ and $H = ABD$.

Because the response variable in this setting is continuous, we use linear regression to analyze the data from this experiment. And because only $2^4 = 16$ conditions were used, we can only fit a model with 16

Table 35: Design matrix and response summary for the 2^{8-4} fractional factorial wine experiment.

Condition	A	B	C	D	E	F	G	H	Average Rating
1	−1	−1	−1	−1	−1	−1	−1	−1	9.6
2	+1	−1	−1	−1	−1	+1	+1	+1	10.8
3	−1	+1	−1	−1	+1	−1	+1	+1	12.6
4	+1	+1	−1	−1	+1	+1	−1	−1	9.2
5	−1	−1	+1	−1	+1	+1	+1	−1	9.0
6	+1	−1	+1	−1	+1	−1	−1	+1	15.0
7	−1	+1	+1	−1	−1	+1	−1	+1	5.0
8	+1	+1	+1	−1	−1	−1	+1	−1	15.2
9	−1	−1	−1	+1	+1	+1	−1	+1	2.2
10	+1	−1	−1	+1	+1	−1	+1	−1	7.0
11	−1	+1	−1	+1	−1	+1	+1	−1	8.8
12	+1	+1	−1	+1	−1	−1	−1	+1	2.8
13	−1	−1	+1	+1	−1	−1	+1	+1	4.6
14	+1	−1	+1	+1	−1	+1	−1	−1	2.4
15	−1	+1	+1	+1	+1	−1	−1	−1	9.2
16	+1	+1	+1	+1	+1	+1	+1	+1	12.6

regression coefficients. In the context of a full 2^4 factorial experiment, this would be the model with 4 main effects, 6 two-factor interactions, 4 three-factor interactions and 1 four-factor interaction. A summary of this fitted model is shown below.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5000     0.2658  31.985  < 2e-16 ***
A             0.8750     0.2658   3.293 0.001619 **
B             0.9250     0.2658   3.481 0.000906 ***
C             0.6250     0.2658   2.352 0.021772 *
D            -2.3000     0.2658  -8.655 2.27e-12 ***
A:B          -0.3500     0.2658  -1.317 0.192532
A:C           1.3000     0.2658   4.892 7.07e-06 ***
B:C           0.4500     0.2658   1.693 0.095261 .
A:D          -0.8750     0.2658  -3.293 0.001619 **
B:D           1.2250     0.2658   4.610 1.98e-05 ***
C:D           0.3750     0.2658   1.411 0.163063
A:B:C         1.5750     0.2658   5.927 1.35e-07 ***
A:B:D        -0.3000     0.2658  -1.129 0.263168
A:C:D        -1.0000     0.2658  -3.763 0.000367 ***
B:C:D         1.1000     0.2658   4.139 0.000104 ***
A:B:C:D       0.4750     0.2658   1.787 0.078613 .
---
```

Notice this output does not involve the factors E, F, G or H – it only directly references factors A, B, C and D. However, because of the confounding associated with the aliasing in this experiment the BCD interaction estimate also corresponds to the main effect of E, the ACD interaction estimate also corresponds to the main effect of F, the ABC interaction estimate also corresponds to the main effect of G, and the ABD interaction estimate also corresponds to the main effect of H. While we cannot technically separate these effects, we assume that the three-factor interactions are neglibible, and hence any significant effect observed is due to the aliased main effect. The same model summary is shown again below, but this time with factors E, F, G and H referenced instead of the three-factor interactions.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5000     0.2658  31.985  < 2e-16 ***
A             0.8750     0.2658   3.293 0.001619 **
B             0.9250     0.2658   3.481 0.000906 ***
C             0.6250     0.2658   2.352 0.021772 *
D            -2.3000     0.2658  -8.655 2.27e-12 ***
E             1.1000     0.2658   4.139 0.000104 ***
F            -1.0000     0.2658  -3.763 0.000367 ***
G             1.5750     0.2658   5.927 1.35e-07 ***
H            -0.3000     0.2658  -1.129 0.263168
A:B          -0.3500     0.2658  -1.317 0.192532
A:C           1.3000     0.2658   4.892 7.07e-06 ***
A:D          -0.8750     0.2658  -3.293 0.001619 **
A:E           0.4750     0.2658   1.787 0.078613 .
A:F           0.3750     0.2658   1.411 0.163063
A:G           0.4500     0.2658   1.693 0.095261 .
A:H           1.2250     0.2658   4.610 1.98e-05 ***
---
```

Based on this output it appears as though all of the main effects but H (fermentation temperature) are significant. Furthermore, factors D, E, F, G (yeast/skin contact, stems, barrel toast, whole cluster) are most influential. Additionally, the AC, AH and AD interactions also appear to be significant. Note that because of the aliasing structure imposed by the four design generators, these two-factor interactions are respectively confounded with DF, FG and EG – and because factors D, E, F and G are most influential, it is likely that the DF, FG and EG interactions are responsible for the significant effect – not the AC, AH and AD interactions. However, this is simply speculation – we cannot know for certain what causes a significant effect – this is the sacrifice that is made when performing a fractional factorial experiment.

Note that a partial F-test of

$$H_0: \ \beta_H = \beta_{AB} = \beta_{AE} = \beta_{AF} = \beta_{AG} = 0$$

which compares the full model above to the one that is reduced by H_0 has an associated p-value of $P(T \geq 2.2124) = 0.06375$ where $T \sim F(5, 64)$. Thus, at a 5% level of significance we do not reject H_0 and we conclude that all factors other than H have significant main effects, and the two factor interactions DF, FG and EG are also statistically significant. Figure 33 depicts ‘main effect plots for all eight factors and Figure 34 depicts the interaction effect plots for the three signifcant interactions.

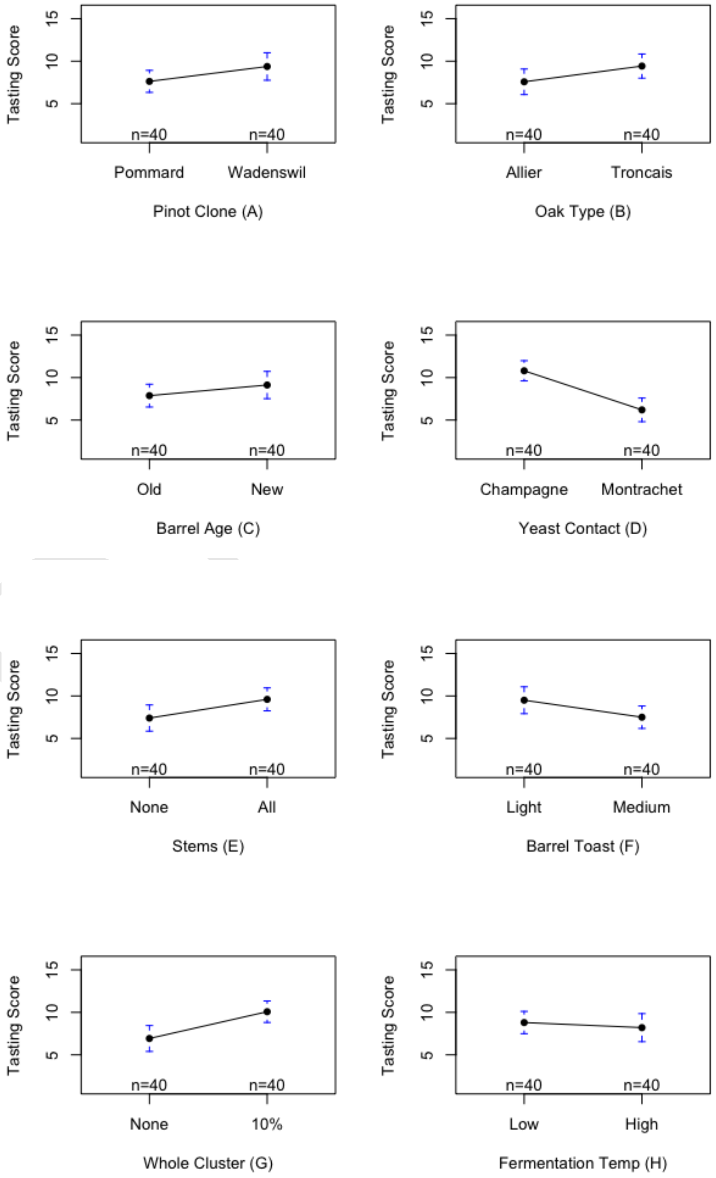

Figure 33: Main effect plots for the wine example.

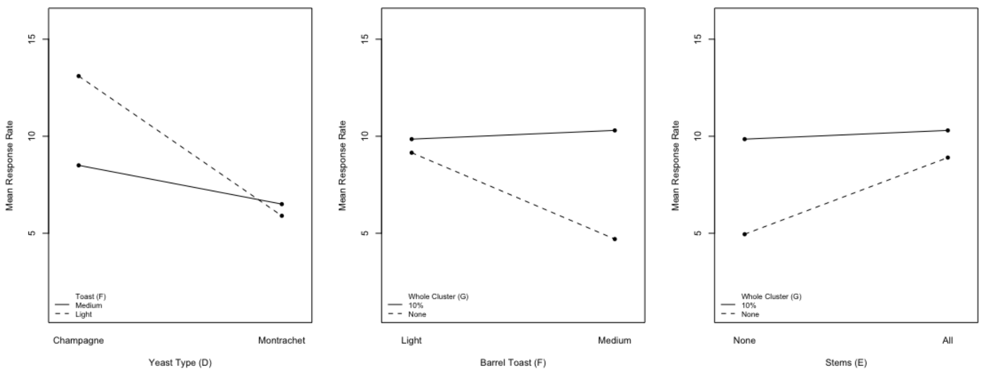

Figure 34: Interaction effect plots for the wine example.

By examining the main effect plots it becomes clear that yeast type (D) and the amount of whole clusters (G) used during fermentation are most important, with no whole clusters and Montrachet yeast producing a better tasting Pinot Noir. As well, medium barrel toast (F) and no stems (E) also seem to correspond to a better tasting wine. The interaction plots indicate that if yeast type is Montrachet, the level of barrel toasting doesn't matter much, but if yeast type is Champagne, a medium barrel toast is best. Also, if barrel toast is chosen to be medium, then not including any whole-clusters is best, and similarly, if using none of the stems, then it is also best not to include any whole-clusters. However, it is important to reiterate that due to the confounding present in the design these conclusions should be drawn tentatively.

8 RESPONSE SURFACE METHODOLOGY

Effective experimentation is sequential; information gained in one experiment can help to inform future experiments. This is the philosophy of **response surface methodology**. In the previous two chapters we discussed the notion of screening experiments, whose primary purpose is to identify which among a large number of factors are the ones that significantly influence the response variable. We saw that two-level experiments such as 2^K factorial and 2^{K-p} fractional factorial experiments could be used for this purpose. In this chapter we discuss how such screening experiments may be followed-up by further experiments whose primary purpose is response optimization. In particular, the **method of steepest ascent/descent** and **response surface experiments** may be used to locate optimal settings of the factors that were identified as significant in the screening phase.

8.1 Overview of Response Optimization

Once the important factors have been identified, it is prudent to determine which levels are optimal. In other words, we aim to determine which combination of the factors' levels will optimize the response variable and hence the metric of interest. Here we consider $K' \leq K$ design factors which are a subset of the K factors investigated during the screening phase. The set of possible values these factors can take on is referred to as the **region of operability**. It is this region that we explore and in which we run our experiments to determine the *optimal* operating condition.

While this region specifies acceptable factor values in their natural units (such as dollars, minutes, percent, etc.), we typically work on a transformed scale. Just as in the regression models in Chapters 5 and 6, we represent each factor by a coded variable x that takes on the values -1 and $+1$ when the factor is at its *low* and *high* levels. When the factor is categorical this coding is arbitrary. However, when the factor is numeric the coding arises through the following transformation

$$x = \frac{U - (U_H + U_L)/2}{(U_H - U_L)/2} \tag{41}$$

where U_H and U_L correspond to the high and low values of the factor as recorded in its natural units, and x corresponds to the coded version of U (a particular value of the factor in its natural units). It can be readily seen that substituting $U = U_H$ and $U = U_L$ into this equation gives $+1$ and -1, respectively. This equation may also be inverted allowing for conversion from the coded units back to the natural units as follows:

$$U = x \times \frac{(U_H - U_L)}{2} + \frac{(U_H + U_L)}{2} \tag{42}$$

Conversion back and forth between the coded and natural units is useful as it facilitates experimentation at conditions outside the ± 1 cuboidal regions used in the two-level screening designs. It also allows for interpolation, which is especially useful when translating the location of the optimum from the coded units to the natural units. We will see equations (41) and (42) used for both of these purposes in Sections 8.2 and 8.3.

Using this notation, the objective of response optimization may be stated as determining the value of $\mathbf{x} = (x_1, x_2, \ldots, x_{K'})^T$ (and hence $\mathbf{U} = (U_1, U_2, \ldots, U_{K'})^T$) at which we expect the response to be optimized. Whether optimization corresponds to maximization or minimization depends on the context of the problem. In either case, this goal may be achieved via **response surface experimentation** where one seeks to characterize the relationship between the expected response $\text{E}[Y]$ and the K' design factors. In the case of a continuous response, we may write this relationship generally as

$$\text{E}[Y] = f(x_1, x_2, \ldots, x_{K'})$$

and in the case of a binary response

$$\log\left(\frac{\text{E}[Y]}{1 - \text{E}[Y]}\right) = f(x_1, x_2, \ldots, x_{K'}).$$

In both cases, the function $f(x_1, x_2, \ldots, x_{K'})$ respresents the *true* but *unknown* **response surface**. If this surface was known, one could simply find the point $(x_1, x_2, \ldots, x_{K'})$ at which $f(\cdot)$ is optimized. However, because $f(\cdot)$ is unknown, we must fit models that approximate this surface. As usual, we use linear and logistic regression.

Although many different models may be used to approximate the response surface we exploit Taylor's Theorem and use low-order polynomials. In particular, we fit first and second-order models which rely on main effects, two-factor interactions and quadratic effects. The linear predictor in a main-effect-only (i.e., first-order) model is

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{K'} x_{K'}$$

and in a main-effect-plus-interaction model the linear predictor is

$$\eta = \beta_0 + \sum_{j=1}^{K'} \beta_j x_j + \sum_{j<l} \beta_{jl} x_j x_l$$

and in a full quadratic (i.e., second-order) model the linear predictor is given by

$$\eta = \beta_0 + \sum_{j=1}^{K'} \beta_j x_j + \sum_{j<l} \beta_{jl} x_j x_l + \sum_{j=1}^{K'} \beta_{jj} x_j^2.$$

We must acknowledge that the approximation of $f(x_1, x_2, \ldots, x_{K'})$ by η (regardless of whether η is first or second order) is likely to be poor when considered across the entire x-space. However, in the small localized region of an experiment, such low-order polynomials should well-approximate $f(\cdot)$.

When $K' > 2$ we cannot visualize these response surfaces, but when $K' = 2$ we can visualize them either with 3-dimensional surface plots or 2-dimensional contour plots. In this case, the first-order linear predictor can be visualized as a plane, the first-order-plus-interaction linear predictor can be visualized as a twisted plane, and the second-order linear predictor can be visualized as a concave or convex surface (i.e., as a hill or a bowl). Figure 35 provides example surface and contour plots for each of these three types of models.

The complexity of the model is dictated by the goal of the experiment. In the context of factor screening we saw that first-order and first-order-plus-interaction models suited our needs. However, such models do not possess the concavity/convexity needed to locate the maximum/minimum – this is the purpose of second-order models. Unfortunately two-level designs do not provide enough information to fit a second order model. In order to do so we must explore the K' factors at more than just two levels each. The class of designs that facilitate this are called **response surface designs**. A variety of such designs exist, but in Section 8.3 we focus on the central composite design, which arises as a natural extension of the two-level designs that we are already familiar with.

Supposing that sufficient data is collected and the second order model may be fitted, we obtain the estimated response surface

$$\widehat{\eta} = \widehat{\beta}_0 + \sum_{j=1}^{K'} \widehat{\beta}_j x_j + \sum_{j<l} \widehat{\beta}_{jl} x_j x_l + \sum_{j=1}^{K'} \widehat{\beta}_{jj} x_j^2.$$

This expression may be re-written in vector-matrix notation as

$$\widehat{\eta} = \widehat{\beta}_0 + \mathbf{x}^T \mathbf{b} + \mathbf{x}^T \mathbf{B} \mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K'} \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_{K'} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \widehat{\beta}_{11} & \frac{1}{2}\widehat{\beta}_{12} & \cdots & \frac{1}{2}\widehat{\beta}_{1K'} \\ \frac{1}{2}\widehat{\beta}_{12} & \widehat{\beta}_{22} & \cdots & \frac{1}{2}\widehat{\beta}_{2K'} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\widehat{\beta}_{1K'} & \frac{1}{2}\widehat{\beta}_{2K'} & \cdots & \widehat{\beta}_{K'K'} \end{bmatrix}.$$

Thus $\mathbf{x}$ is a $K' \times 1$ vector of factor values, $\mathbf{b}$ is a $K' \times 1$ vector of estimates of the the first order coefficients, and $\mathbf{B}$ is a $K' \times K'$ symmetric matrix of estimates with the quadratic coefficients along the diagonal and one-half the interaction coefficients on the upper and lower triangle.

In order to identify the optimal factor settings, i.e., the value of $\mathbf{x} = (x_1, x_2, \ldots, x_{K'})^T$ that maximizes/minimizes the expected response, we must find the **stationary point** of the estimated response surface. This is found by solving

$$\frac{d\widehat{\eta}}{d\mathbf{x}} = \mathbf{b} + 2\mathbf{B}\mathbf{x} = \mathbf{0}$$

The stationary point is therefore

$$\mathbf{x}_s = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b}$$

and the optimal expected response is

$$\widehat{\eta}_s = \widehat{\beta}_0 + \frac{1}{2}\mathbf{x}_s^T \mathbf{b}$$

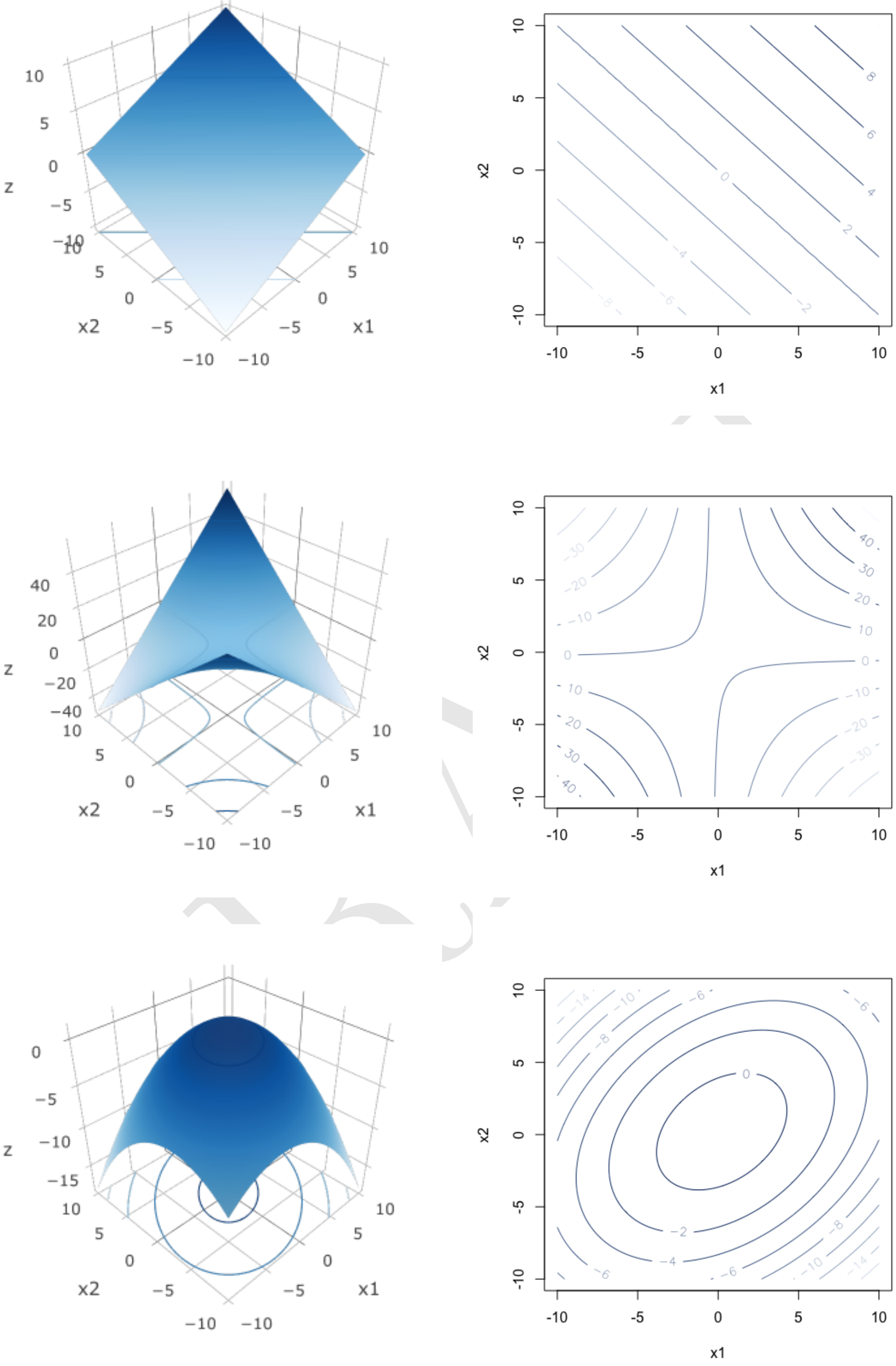

Figure 35: Example 3D surface and 2D contour plots of first-order (top), first-order-plus-interaction (middle) and second-order (bottom) response surfaces.

in the case of linear regression and

$$\frac{e^{\widehat{\eta}_s}}{1+e^{\widehat{\eta}_s}} = \frac{e^{\widehat{\beta}_0+\frac{1}{2}\mathbf{x}_s^T\mathbf{b}}}{1+e^{\widehat{\beta}_0+\frac{1}{2}\mathbf{x}_s^T\mathbf{b}}}$$

in the case of logistic regression. For practical implementation of this solution, the stationary point $\mathbf{x}_s$ must be translated into optimal operating conditions in natural units $\mathbf{U}_s$ using equation (42).

However, for us to be confident that $\mathbf{x}_s$ indeed optimizes $f(\cdot)$, we must be confident that η and, in particular, that $\widehat{\eta}$ adequately represents $f(\cdot)$. Since we only expect the second-order approximation to be adequate in a small localized region, it is important that this small localized region contains the true optimum. It is quite unlikely that the values of $x_1, x_2, \ldots, x_{K'}$ considered in the screening phase are close to the optimum. As such, an intermediate phase of experimentation is needed prior to fitting a second order model with a response surface design. In this intermediate phase a sequence of experimental conditions are performed to move from the initial region of experimentation toward region of the optimum. The process by which this is achieved is known as the **method of steepest ascent** (in the case of maximization) and the **method of steepest descent** (in the case of minimization). This procedure is discussed in more detail in Section 8.2.

We remark that the topic of response surface methodology goes much deeper than the treatment here; here we only begin to scratch the surface (no pun intended!). Other considerations include the plethora of response surface designs beyond the central composite design; response optimization when an optimal response is one that is held on target (rather than being maximized/minimized); or the problem of multiple-response variables and hence multi-objective optimization. For a more complete and thorough treatment of response surface methodology see Myers et al. (2016).

8.2 Method of Steepest Ascent/Descent

Prior to fitting a second order model with a response surface design, it is important to ensure that the location of such a design (in the x-space) is reasonably close to the optimum. Thus, while a response surface design and a second order model can pinpoint the optimum precisely, we must first possess a *rough* understanding of where in the x-space the optimum lies. A naive solution to this problem would be a grid search over the x-space. While this would almost certainly locate the region of the optimum, it would not be nearly as efficient (in terms of number of experimental conditions required) as the gradient-based method of steepest ascent/descent. For a given position in the x-space, the method of steepest ascent/descent identifies the direction that when traversed moves you toward the optimum as quickly as possible.

8.2.1 The Path of Steepest Ascent/Descent

The method of steepest ascent/descent is a gradient-based solution in which we successively define new experimental conditions along the gradient of the fitted first order surface. The gradient indicates the direction of steepest increase/decrease on the fitted surface, and hence the direction to move in order to optimize the response. Using a $2^{K'}$ factorial experiment we fit a first order model yielding the estimated linear predictor

$$\widehat{\eta} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_{K'} x_{K'}$$

The gradient of this surface is

$$\mathbf{g} = \nabla\widehat{\eta} = \begin{bmatrix} \frac{\partial\widehat{\eta}}{\partial x_1} & \frac{\partial\widehat{\eta}}{\partial x_2} & \cdots & \frac{\partial\widehat{\eta}}{\partial x_{K'}} \end{bmatrix}^T$$

If maximizing the response is of interest, then we should ascend the surface by moving in the direction of $+\mathbf{g}$. We call this the *path of steepest ascent.* Alternatively, if minimization is of interest, then we should descend the surface by moving in the direction of $-\mathbf{g}$. We call this the *path of steepest descent.*

Given a fixed step size λ, stepping a distance of λ away from $\mathbf{x}$ along the path of steepest ascent takes you to

$$\mathbf{x}' = \mathbf{x} + \lambda\mathbf{g} \tag{43}$$

and along the path of steepest descent takes you to

$$\mathbf{x}' = \mathbf{x} - \lambda\mathbf{g} \tag{44}$$

In either case we typically define the step size λ as

$$\lambda = \frac{\Delta x_j}{|\widehat{\beta}_j|}$$

where factor j is the one we know most about or that is hardest to manipulate. Its corresponding regression coefficient is β_j and Δx_j is its step size (in coded units), chosen by the experimenter.

Starting from the origin (center point) of the x-space, Equations (43)/(44) are used to determine the location $\mathbf{x}'$ of each new experimental condition. The coordinates of $\mathbf{x}'$ are translated into natural units using Equation (42) so that the experimenter knows how to define the new conditions in the real-world. The data collected in each new condition are used to calculate the metric of interest. If $\widehat{\eta}$ was an adequate representation of the true underlying response surface, then the MOI calculated at successive steps should gradually improve as we ascend/descend the response surface. We stop traversing the gradient when the MOI stops incrementally improving; this signifies that we've passed over the optimum. We should then return to the best location along the gradient and test for curvature (as discussed in the next section). If the test for curvature suggests that you are not yet in the vicinity of the optimum, fit a new first order model and traverese its gradient until no further improvement is possible. This process is repeated until a test for curvature suggests you have reached the vicinity of the optimum. Once this happens, use a response surface design to fit a full second order model and hence precisely identify the coordinates of the optimum.

8.2.2 Checking for Curvature

A test for quadratic curvature is an important component of the method of steepest ascent/descent, because the presence of quadratic curvature indicates that were are in the vicinity of the optimum. Such a test is possible when a $2^{K'}$ factorial experiment is augmented with a **center point** condition. The center point condition is defined (in coded units) as $x_1 = x_2 = \cdots = x_{K'} = 0$, and is located at the center of the cuboidal region defined by the $2^{K'}$ factorial conditions. As we have discussed previously, the data arising from a $2^{K'}$ factorial design is insufficient to estimate a second order linear predictor; we are able to estimate the main effects and the two-factor interaction effects, but *not* the quadratic effects. However, with the addition of the center point condition, one additional effect may be estimated: the **pure quadratic effect**

$$\beta_{PQ} = \sum_{j=1}^{K'} \beta_{jj}.$$

A test of $H_0 : \beta_{PQ} = 0$ is a test for *overall curvature.*

To gain an intuition for this effect and this test we will consider, as an illustrative example, the simple case when $K' = 2$. In this case the second order linear predictor is

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

In a 2^2 factorial design plus a center point, we have the following five unique experimental conditions: $(x_1, x_2) \in \{(-1,-1), (+1,-1), (-1,+1), (+1,+1), (0,0)\}$ which respectively give rise to five unique variants of the linear predictor, which we define as

$$\eta_{LL} = \beta_0 - \beta_1 - \beta_2 + \beta_{12} + \beta_{11} + \beta_{22}$$

$$\eta_{HL} = \beta_0 + \beta_1 - \beta_2 - \beta_{12} + \beta_{11} + \beta_{22}$$

$$\eta_{LH} = \beta_0 - \beta_1 + \beta_2 - \beta_{12} + \beta_{11} + \beta_{22}$$

$$\eta_{HH} = \beta_0 + \beta_1 + \beta_2 + \beta_{12} + \beta_{11} + \beta_{22}$$

$$\eta_C = \beta_0$$

Unfortunately with only these five conditions, we cannot separately estimate β_{11} and β_{22}, but we *can* estimate $\beta_{PQ} = \beta_{11} + \beta_{22}$. To gain an intuition for this pure quadratic effect, notice that

$$\beta_{PQ} = \frac{\eta_{LL} + \eta_{HL} + \eta_{LH} + \eta_{HH}}{4} - \eta_C.$$

Thus, β_{PQ} is the difference between the average linear predictor values at the factorial conditions, minus the linear predictor value at the center point condition. The estimate is therefore:

$$\widehat{\beta}_{PQ} = \frac{\widehat{\eta}_{LL} + \widehat{\eta}_{HL} + \widehat{\eta}_{LH} + \widehat{\eta}_{HH}}{4} - \widehat{\eta}_C.$$

If this difference, and hence $\widehat{\beta}_{PQ}$, is small then it suggests that the response values observed in the factorial conditions are similar to those observed in the center point condition and hence that there isn't significant curvature in the response surface. However, if $\widehat{\beta}_{PQ}$ is very different from zero it suggests that there is significant quadratic curvature.

In order to make this determination we must formally test $H_0 : \beta_{PQ} = 0$. We do so using t-tests (or Z-tests) in an *appropriately defined* linear (or logistic) regression model. The five linear predictors above can be unified into a single linear predictor that is entirely estimable based on the factorial and center point conditions:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{PQ} x_{PQ}.$$

Here x_{PQ} is a binary indicator that indicates whether an observation came from one of the factorial conditions or the center point condition:

$$x_{PQ} = \begin{cases} 1 & \text{if } (x_1, x_2) \in \{(-1,-1), (+1,-1), (-1,+1), (+1,+1)\} \\ 0 & \text{if } (x_1, x_2) = (0,0) \end{cases}$$

If β_{PQ} is significantly different from 0 then it suggests that *both* β_{11} and β_{22} are significantly non-zero, and therefore that there is significant quadratic curvature in the response surface in the region of the x-space at which the experiment was conducted.

In general, for values of K' other than 2, we conduct a $2^{K'}$ factorial experiment with a center point and then test for curvature using a regression model with linear predictor:

$$\eta = \beta_0 + \sum_{j=1}^{K'} \beta_j x_j + \sum_{j<l} \beta_{jl} x_j x_l + \beta_{PQ} x_{PQ}$$

where now $\beta_{PQ} = \sum_{j=1}^{K'} \beta_{jj}$ and x_{PQ} is again a binary indicator indicating whether a response value was observed in a factorial condition or the center point condition. No matter the value of K', the pure quadratic effect is always represented by a single term in the model. As such, the test for curvature is always a test of $H_0 : \beta_{PQ} = 0$ and is carried out with ordinary t-tests in a linear regression and Z-tests in a logistic regression.

The intuitive estimate for β_{PQ} in the $K' = 2$ case also generalizes. In general it may be written as

$$\widehat{\beta}_{PQ} = \overline{\widehat{\eta}}_F - \widehat{\eta}_C$$

where $\overline{\widehat{\eta}}_F$ is the average of the estimated linear predictor values in the factorial conditions and $\widehat{\eta}_C$ is the estimated linear predictor value at the center point.

Note that this test assumes that all of the β_{jj}'s, $j = 1, 2, \ldots, K'$, have the same sign. If they didn't, then it's possible that significantly large β_{jj}'s could cancel each other out, making $\sum_{j=1}^{K'} \beta_{jj}$ close to zero, incorrectly suggesting a lack of curvature. This assumption is fine in practice, as long as the experiment is not conducted near a saddle point on the response surface. The only way to rule out this possibility is to perform a response surface design and fit a second order model, thereby obtaining separate estimates of each quadratic effect.

8.2.3 The Netflix Example

Here we illustrate the *method of steepest descent* using the hypothetical Netflix experiment from your final project. We focus on the Preview Length and Preview Size factors only. We begin with a 2^2 factorial experiment with a center point condition. The factor levels in coded and natural units are shown in Table 36.

Table 36: Average browsing time by condition in the $2^2 + CP$ Netflix experiment.

Condition	Preview Length	x_1	Preview Size	x_2	Average Browsing Time
1	90 seconds	-1	0.2	-1	22.16 minutes
2	120 seconds	$+1$	0.2	-1	22.20 minutes
3	90 seconds	-1	0.5	$+1$	20.22 minutes
4	120 seconds	$+1$	0.5	$+1$	21.98 minutes
5	105 seconds	0	0.35	0	22.05 minutes

Prior to embarking down the path of steepest descent, a curvature test was performed to determine whether this experimental region was already in the vicinity of the optimum. The linear regression model with linear predictor

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{PQ} x_{PQ}$$

was fit. The resulting output is shown below. As we can see, a test of $H_0 : \beta_{PQ}$ *is not rejected*, suggesting that we are not in the vicinity of the optimum and hence that a steepest descent search is warranted.

```
Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 22.04607    0.19513 112.979  < 2e-16 ***

x1           0.44828    0.09757   4.595 4.78e-06 ***

x2          -0.53894    0.09757  -5.524 4.04e-08 ***

x1:x2        0.43118    0.09757   4.419 1.08e-05 ***

xPQ         -0.40466    0.21817  -1.855   0.0639 .

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To begin the procedure, we use the aforementioned data to fit the first order regression model with linear predictor

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The model summary is shown below. The lefthand plot in Figure 36 depicts the contours of the estimated first order response surface. It is clear from this plot that if we move toward the top left corner of the x-space the average browsing time will decrease.

```
Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.72234    0.08800 246.852  < 2e-16 ***
```

```
x1           0.44828    0.09838   4.556 5.71e-06 ***

x2          -0.53894    0.09838  -5.478 5.20e-08 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

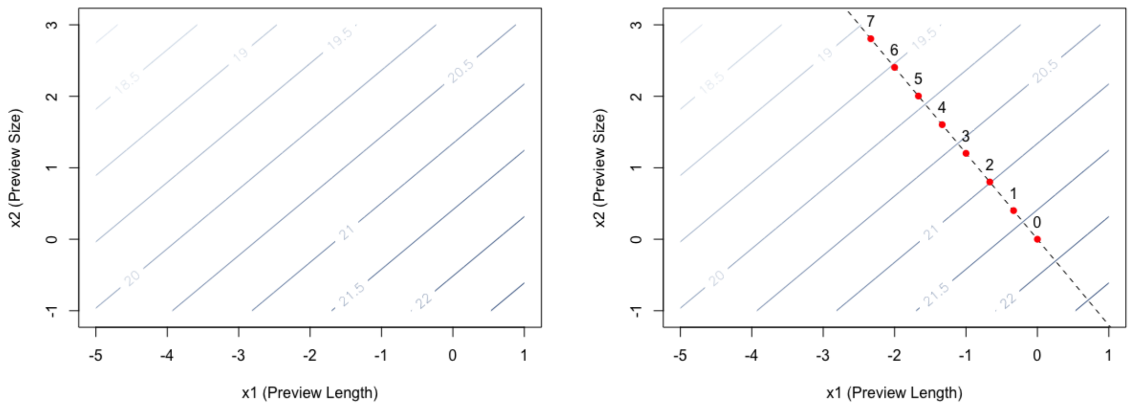

Figure 36: Contour plot for the estimated first order response surface (left) and the path of steepest descent (right) for the Netflix experiment.

To formalize this observation we calculate the gradient

$$\mathbf{g} = \begin{bmatrix} \widehat{\beta}_1 & \widehat{\beta}_2 \end{bmatrix}^T = \begin{bmatrix} 0.44828 & -0.53894 \end{bmatrix}^T$$

This path of steepest descent is depicted by the dashed black line in the righthand plot in Figure 36. The red dots signify the experimental conditions conducted along this path, beginning from the center point $(x_1, x_2) = (0, 0)$. The locations in coded and natural units for each of these conditions are provided in Table 37. Note that a step size of

$$\lambda = \frac{1/3}{|0.44828|}$$

was used, where the value 1/3 was chosen to ensure steps of 5 seconds in Preview Lengths.

Table 37: Average browsing time along the path of steepest descent.

Step	Preview Length	x_1	Preview Size	x_2	Average Browsing Time
0	105 seconds	0	0.3500000	0	22.00 minutes
1	100 seconds	$-1/3$	0.4101118	0.4007454	21.67 minutes
2	95 seconds	$-2/3$	0.4702236	0.8014908	21.26 minutes
3	90 seconds	-1	0.5303354	1.202236	19.11 minutes
4	85 seconds	$-4/3$	0.5904472	1.602982	18.24 minutes
5	80 seconds	$-5/3$	0.6505591	2.003727	15.94 minutes
6	75 seconds	-2	0.7106709	2.404472	14.89 minutes
7	70 seconds	$-7/3$	0.7707827	2.805218	17.16 minutes

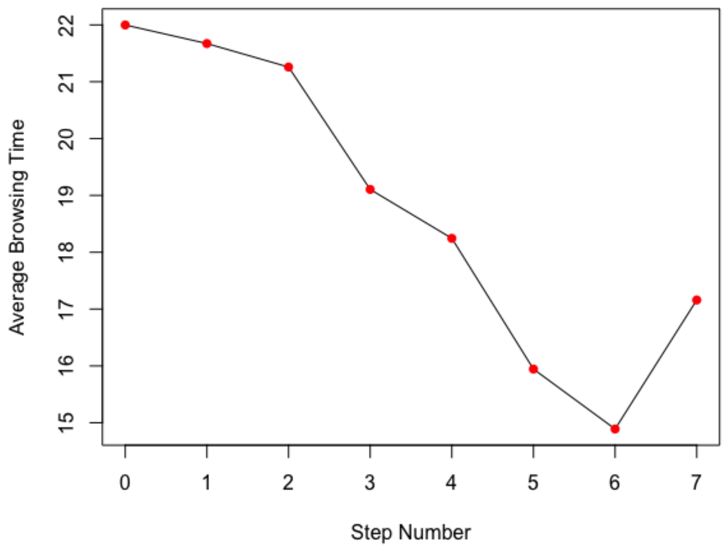

Figure 37: Average browsing time along the path of steepest descent.

The average browsing time in each condition is reported in Table 37 and visualized in Figure 37. We see clearly that Step 6 corresponded to the lowest observed average browsing time and so we should perform another test of curvature in this region to determine whether we've reached the vicinity of the optimimum. In order to do so, another 2^2 factorial experiment with a center point needs to be run. The factor levels in coded and natural units for this next experiment are shown in Table 38.

Table 38: Average browsing time by condition in the second $2^2 + CP$ Netflix experiment.

Condition	Preview Length	x_1	Preview Size	x_2	Average Browsing Time
1	60 seconds	-1	0.6	-1	18.22 minutes
2	90 seconds	$+1$	0.6	-1	14.57 minutes
3	60 seconds	-1	0.8	$+1$	14.83 minutes
4	90 seconds	$+1$	0.8	$+1$	18.65 minutes
5	75 seconds	0	0.7	0	18.17 minutes

Once again we fit a linear regression model with linear predictor

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{PQ} x_{PQ}$$

The resulting output is shown below. As we can see, a test of $H_0 : \beta_{PQ}$ *is rejected*, suggesting that there is significant quadratic curvature in this region of the response surface and that we have indeed arrived in the vicinity of the optimum. This investigation should now be followed up by a response surface experiment so that a full second order model may be fit and the optimum identified.

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.83111    0.18918  78.397   <2e-16 ***
x1           1.00913    0.09459  10.669   <2e-16 ***
x2           1.03291    0.09459  10.920   <2e-16 ***
x1:x2       -0.79191    0.09459  -8.372   <2e-16 ***
xPQ          2.57337    0.21151  12.167   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

8.3 Response Surface Experiments

The goal of a response surface experiment is to be able to fit a full second order response surface model. This requires estimating $(K'+1)(K'+2)/2$ effects. Several such designs exist, but here we study one in particular: the central composite design. It consists of $2^{K'}$ **factorial** conditions, 1 **center point** condition, and $2K'$ **axial** conditions. We elaborate on this design in the next section, and we illustrate its use in the section after that.

8.3.1 The Central Composite Design

Among all response surface designs, **central composite designs** (CCD) are some of the most commonly used in practice. A CCD is typified by three different types of experimental conditions:

i two-level factorial conditions

ii a center point condition

iii axial, or *star*, conditions

The factorial conditions constitute a full $2^{K'}$ factorial design, the center point condition sits at $x_1 = x_2 = \cdots = x_{K'} = 0$ in the center of the factorial ones, and the axial conditions sit ‘outside’ of the factorial ones at $\pm a$ on each of the K' factors' axes (we discuss the choice of the value a below). When investigating K' factors the central composite design therefore requires $2^{K'} + 2K' + 1$ distinct experimental conditions. Example design matrices for $K' = 1, 2, 3$ are shown in Figure 39. These designs may also be visualized geometrically. Figure 38 visualizes the CCD for $K' = 1, 2, 3$.

The value of a is determined by the experimenter, and may be chosen to balance both practical and statistical concerns. In particular, when choosing the value of a we must be mindful of the constraints imposed by the region of operability. If the CCD takes place in a corner of this region, it's possible that the natural-unit counterpart to the desired value of a (as determined by Equation (42)) lies outside of the region, in which case a different value will need to be chosen. Similarly, the natural-unit counterpart to the desired

Table 39: Design matrices associated with central composite designs on $K' = 1$ (left), $K' = 2$ (middle) and $K' = 3$ (right) factors.

Condition	x_1
1	-1
2	$+1$
3	$-a$
4	$+a$
5	0

Condition	x_1	x_2
1	-1	-1
2	$+1$	-1
3	-1	$+1$
4	$+1$	$+1$
5	$-a$	0
6	$+a$	0
7	0	$-a$
8	0	$+a$
9	0	0

Condition	x_1	x_2	x_3
1	-1	-1	-1
2	$+1$	-1	-1
3	-1	$+1$	-1
4	$+1$	$+1$	-1
5	-1	-1	$+1$
6	$+1$	-1	$+1$
7	-1	$+1$	$+1$
8	$+1$	$+1$	$+1$
9	$-a$	0	0
10	$+a$	0	0
11	0	$-a$	0
12	0	$+a$	0
13	0	0	$-a$
14	0	0	$+a$
15	0	0	0

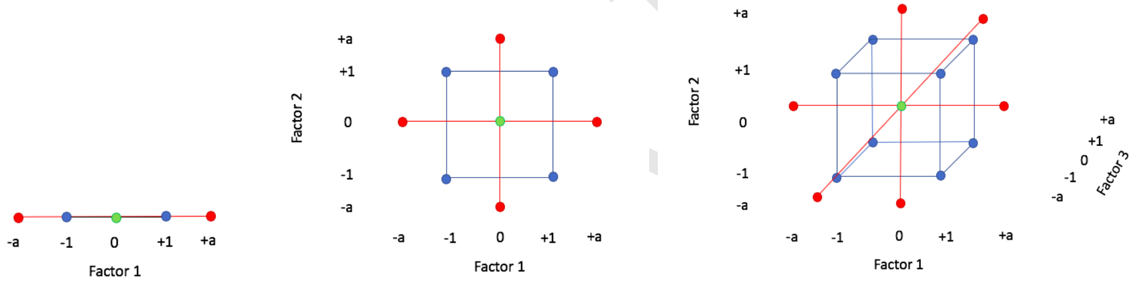

Figure 38: Central composite designs for $K' = 1$ (left), $K' = 2$ (middle) and $K' = 3$ (right) factors. Blue dots indicate factorial conditions, red dots indicate axial conditions and green dots indicate center point conditions.

value of a may be be practically inconvenient (like a 35.12% discount), in which case a more convenient value may be chosen (like a 35% discount).

While, in principle, an experimenter may choose any value of a they wish, the values $a = 1$ and $a = \sqrt{K'}$ are common choices, and for good reason. When $a = 1$, the CCD reduces to a $3^{K'}$ design, such as those discussed in Section 6.3, and is sometimes referred to as a *face-centered central composite design*. The benefit of such a design is that it requires just 3 (not 5) levels for every factor and, as depicted in Figure 26, is a cuboidal design and so it inherets all of the usual conveniences associated with orthogonal cuboidal designs.

The value $a = \sqrt{K'}$ is attractive as it sets the axial conditions at an equal distance from the center point as the factorial conditions. This distance is $\sqrt{K'}$ units. Such a design is referred to as *spherical* since it places all axial and factorial conditions on a sphere of radius $\sqrt{K'}$. The benefit of such equal spacing is that it ensures that the estimate of the response surface at each condition is equally precise. A design is called **rotatable** if it has this property (i.e., the estimation uncertainty associated with $\widehat{\eta}$ is the same for every condition that is the same distance from the center point).

No matter the choice of a, the CCD facilitates estimation of the full second order response surface model, and hence identification of the optimum.

8.3.2 The Lyft Example

In this section we illustrate the design and analysis of a central composite experiment in the context of a common ride-sharing problem. Suppose that Lyft is interesed in designing a promotional offer that maximizes ride-bookings during an experimental period. Previous screening experiments evaluated the influence of discount amount, discount duration, ride type, time-of-day, and the method of dissemination. It was found that the most important factors were discount amount (x_1) and discount duration (x_2). A previous steepest ascent exercise also suggested that the optimal discount duration is somewhere in the vicinity of 4.5 days and the optimal discount amount is somewhere in the vicinity of 50%.

To find optimal values of these factors a follow-up two-factor central composite design was run in order to fit a second-order response surface model. This experiment consisted of 4 factorial conditions, 4 axial conditions and 1 center point condition (as in the middle design matrix in Table 39 and the middle image in Figure 38). The factor levels in coded and natural units are shown in Table 40. Note that the experimenters had intended to perform axial conditions at $a = \sqrt{2}$ but the corresponding discount amounts and discount durations were (14.64466%, 85.35534%) and (0.9644661 days, 8.035534 days). In the interest of defining experimental conditions with practically convenient levels they opted for $a = 1.4$ yielding the discount amounts and durations shown in the table below.

Table 40: Booking rate by condition in the Lyft experiment.

Condition	Discount Amount	x_1	Discount Duration	x_2	Booking Rate
1	25%	-1	2 days	-1	0.71
2	75%	$+1$	2 days	-1	0.32
3	25%	-1	7 days	$+1$	0.71
4	75%	$+1$	7 days	$+1$	0.35
5	85%	$+1.4$	4.5 days	0	0.53
6	15%	-1.4	4.5 days	0	0.50
7	50%	0	8 days	$+1.4$	0.26
8	50%	0	1 day	-1.4	0.78
9	50%	0	4.5 days	0	0.72

Subsequently $n = 500$ users were randomized into each of these $m = 9$ conditions and for each user, whether they booked a ride in the experimentation period was recorded. The booking rates in each condition are shown in Table 40. The output from the fitted second order logistic regression model is shown below. Contour plots of the fitted response surface are shown in Figure 39. The left plot depicts the relationship between booking rate and x_1 and x_2 (i.e., the factors on the coded levels), and the right plot depicts the relationship between booking rate and discount amount and discount duration. The elliptical contours verify that we are indeed in the vicinity of the optimum.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.94284    0.09952   9.474  < 2e-16 ***
x1           0.03881    0.03307   1.174    0.241
x2          -0.80684    0.03568 -22.612  < 2e-16 ***
x1:x2        0.03392    0.04846   0.700    0.484
I(x1^2)     -0.44207    0.05788  -7.637 2.22e-14 ***
I(x2^2)     -0.41448    0.05931  -6.989 2.77e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

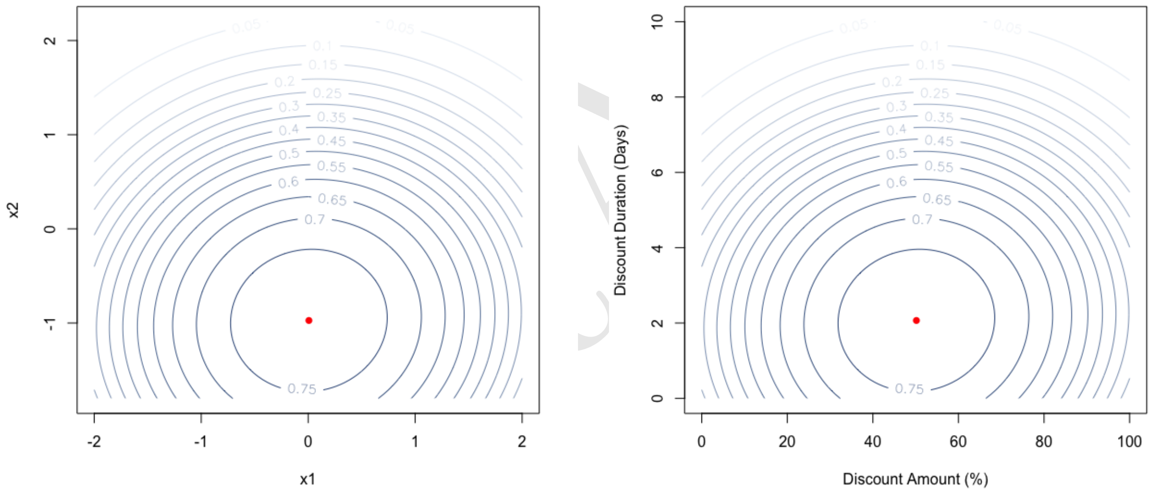

Figure 39: 2D contour plots of the second-order Lyft model. Left: Coded-Unit Factor Space. Right: Natural-Unit Factor Space

The stationary point for this second order model is located (in coded units) at $x_1 = 0.006565206$, $x_2 = -0.973047233$. In the natural units this corresponds to a discount rate of 50.16% that lasts for 2.07 days. The stationary point is identified in red in both plots in Figure 39. The predicted booking rate at this point is 0.7918, with a 95% prediction interval given by (0.7691, 0.8144). The appeal of this approach is that this particular combination of factor levels was not actually experimented with, yet we identified it as being optimal. Additionally, the predicted booking rate is higher than anything observed in the experiment. Follow-up confirmation experiments could be performed in order to confirm the good perfomance of this condition. See Stevens and Anderson-Cook (2019) and Jensen (2016) for more information on confirmation experiments.

Note that a slightly less optimal but more practically feasible promotion would be a 50% discount lasting 2 days which achieves a booking rate of 0.7917 with a 95% prediction interval of (0.7693, 0.8141). The marginal sacrifice in booking rate is offset by the gain in convenience, so we will take these values to define the optimal promotion.

8.4 RSM with Qualitative Factors

Everything that has been discussed thus far with respect to central composite designs and response surface optimization has assumed that the factors under experimentation are quantitive (i.e., the factors have numeric levels). In the presence of one or more categorical factors we need to take additional care. The interpolation between factor levels that takes place during the method of steepest ascent and when identifying the stationary point of the fitted response surface is not conducive to categorical factors.

When categorical factors are present, we can think of there being different response surfaces that relate the response to the quantitative factors at each of the factorial combinations of the categorical factors' levels. Thus, the general strategy is to enumerate all factorial combinations of the categorical factors' levels and employ the methods of response surface methodology independently within each. In particular, we may perform the method of steepest ascent/descent independently on each surface, perform CCDs independently on each surface, independently fit second order models for each surface, and independently identify the stationary point on each surface.

Then, when the stationary points have been identified we determine the optimum of each surface. Among all of the candidate surfaces, the one with the most optimal optimum is the 'winner' and the factor levels (numeric and categorical) that gave rise to it should be defined as the optimal operating conditions.

A APPENDIX

In this Appendix we review some of the statistical prerequisites for the material discussed throughout the notes. In particular we review random variables and probability distributions, point and interval estimation, hypothesis testing, linear regression, and logistic regression.

A.1 Random Variables and Probability Distributions

A.1.1 Random Variables and Probability Functions

A **random variable** $Y : \Omega \to \mathbb{R}$ is a function that assigns real numbers to outcomes of a random process, such as flipping a coin or measuring some quantity of interest. We refer to the possible values a random variable can take on as the **support set**, and we dichotomize random variables based on the type of values they assume. A **discrete** random variable is one whose support set is finite or countably infinite such as $y = 0, 1, 2, \ldots, n$ or $y = 0, 1, 2, \ldots$. We typically use discrete random variables when counting events is of interest. A **continuous** random variable, on the other hand, takes on a continuum of values and so its support set is a subinterval of the real numbers such as $y \geq 0$, $y \in [0, 1]$ or $-\infty < y < \infty$. We typically use continuous random variables when measuring some continuous quantity is of interest. Note that for clarity we denote random variables with upper case letters and the values they take on with lower case letters.

Example 1: Suppose we send an email survey to $n = 30$ individuals and we're interested in the the number of these individuals that respond to the survey. Let Y represent the number of survey responses. In this case the support set is $y = 0, 1, 2, \ldots, 30$, and so Y is a discrete random variable.

Example 2: Interest often lies in measuring lifetimes of people, products, and processes. Suppose that, in particular, we are interested in the lifetime of an iPhone's battery. Let Y represent the lifetime (in hours) of an iPhone battery. In this case the support set is theoretically $y \geq 0$, which is a continuous subinterval of the real numbers, and so Y is a continuous random variable.

Because random variables take on values randomly, interest lies in quantifying the probability that Y assumes a particular value (i.e., $\Pr(Y = a)$) or lies in some interval (i.e., $\Pr(a < Y < b)$). Such probabilities are described by the **probability distribution** of the random variable and quantified by the corresponding **probability function** $f(y)$. The form of this function will differ from one distribution to another, but in all cases, by substituting all values of $y \in A$ (where A is the support set of Y) into $f(y)$ and constructing a plot of $f(y)$ vs. y, we can visualize the probability distribution. Doing so provides insight into the shape of the distribution – most notably, the center and spread – and hence an idea of what values of y seem typical and which ones seem extreme.

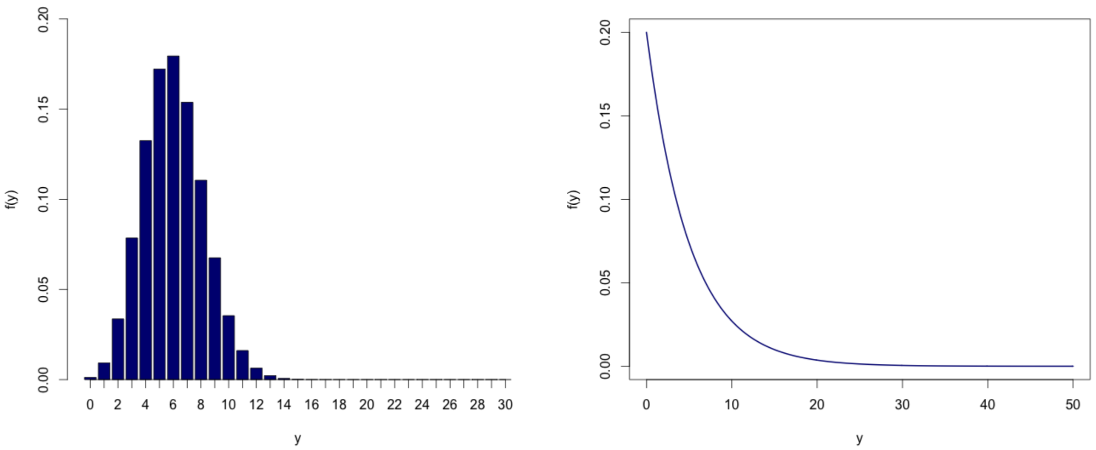

Figure A.1: Left: Distribution Characterizing Survey Respondents; Right: Distribution Characterizing iPhone Battery Lifetimes

Figure A.1 depicts hypothetical distributions for the random variables defined in Examples 1 (left panel) and 2 (right panel). We see that when Y is a discrete random variable the plot of $f(y)$ vs. y is a barplot, with bar heights equaling the probability the Y takes on a given value y. On the other hand, the plot of $f(y)$ vs. y for continuous Y is a smooth curve. In the left hand plot we see that one could reasonably expect 0 to 15 survey responses, with 4 to 8 responses being most likely, and anymore than 15 responses very unlikely. Similarly, the right plot suggests that it is quite likely that an iPhone will last up to 10 hours on a single charge, but it is not very likely to live past 20 hours on a single charge.

To formalize observations like these, we can use probability functions to calculate the probability that such events occur. However, the manner in which these functions are used to calculate probabilities depends on whether Y is discrete or continuous. A **probability mass function** (PMF) describes the probabilistic behavior of a discrete random variable Y, and is given by

$$f(y) = \Pr(Y = y)$$

for all $y \in A$. Thus, for a given value of y, the PMF is the probability that Y takes on that particular value. As such, the PMF allocates probability to every element in the support set, and hence every outcome of the random process for which it is defined. The left plot in Figure A.1 is a visual display of the probability distribution describing the random variable Y defined in Example 1. With this we can calculate things like the probability that exactly 6 individuals respond to the survey ($\Pr(Y = 6)$), or the probability that 10 or more individuals respond to the survey ($\Pr(Y \geq 10)$). By summing the heights of the bars corresponding to all values of y consistent with these events, we find that $\Pr(Y = 6) = 0.1795$ and $\Pr(Y \geq 10) = 0.0611$. These calculations are depicted visually in the left and right panels of Figure A.2.

A **probability density function** (PDF) describes the probabilistic behavior of a continuous random

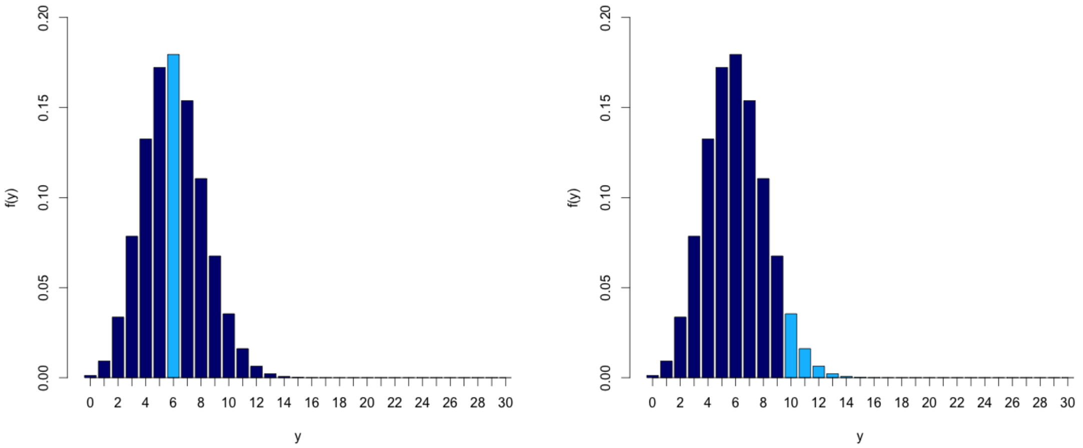

Figure A.2: Left: $\Pr(Y = 6) = f(6)$; Right: $\Pr(Y \geq 10) = \sum\limits_{y=10}^{30} f(y)$

variable Y. Unlike the probability mass function, which for a particular value of y is itself a probability, we think of the PDF $f(y)$ as being the equation of a **density curve** and probabilities concerning Y are calculated as areas beneath this curve. For instance, a hypothetical probability density function describing the lifetime of an iPhone battery (as in Example 2) is plotted in the right panel of Figure A.1. If we are interested in the probability that an iPhone battery will last up to 10 hours ($\Pr(Y \leq 10)$) or more than 20 hours ($\Pr(Y > 20)$), we calculate the area beneath the curve to the left of 10 and right of 20, respectively. Mathematically this requires integration of the PDF. The two probabilities of interest in this case are given by 0.8647 and 0.0183 and visualized in the left and right panels of Figure A.3, respectively.

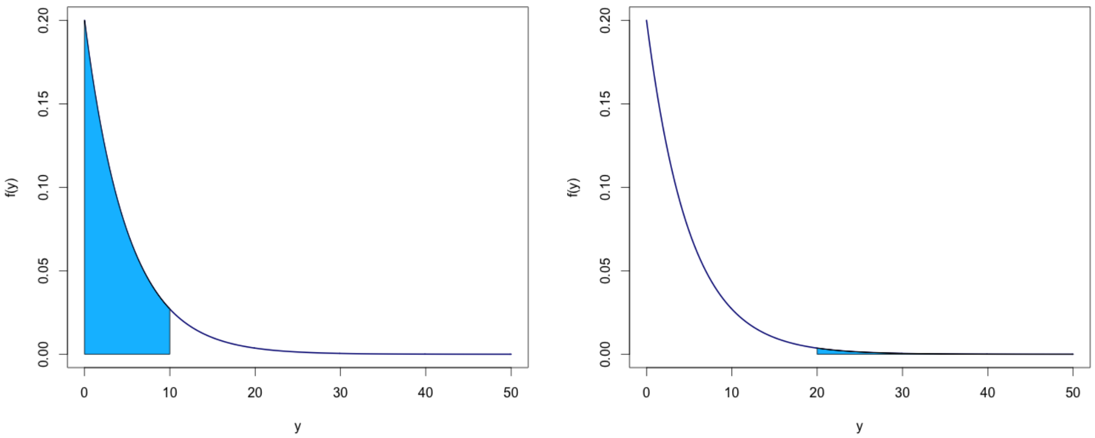

Figure A.3: Left: $\Pr(Y \leq 10) = \int\limits_{0}^{10} f(y)dy$; Right: $\Pr(Y > 20) = \int\limits_{20}^{\infty} f(y)dy$

While a probability distribution is most efficiently summarized by a plot, such as those given in Figure A.1, the probability function (and hence the distribution) may also be characterized by a closed-form expression. This is the case for several well-known probability distributions which are useful for describing a host of real-life random phenomenon. We discuss some of these distributions here, focusing on ones that are used routinely in the context of experimentation.

A.1.2 Relevant Distributions

The Binomial Distribution: As noted above, discrete distributions typically describe the randomness associated with counting events. The binomial distribution is one such distribution, and is relevant when counting events in the context of **Bernoulli trials**. Note that a Bernoulli trial is a random process in which there are just two possible outcomes, arbitrarily labelled *successes* and *failures*. Additionally, the occurrence of these outcomes must be independent of one another (i.e., the outcome of one trial does not influence the outcome of any other trial) and the probability of success π (and hence the probability of failure $1-\pi$), must be the same on each trial. Flipping a coin is a common example of a Bernoulli trial where, for example, the coin turning up 'heads' qualifies as a success and 'tails' qualifies as a failure. If the coin is fair, the probability of a success is $\pi = 0.5$ each time and whether the coin turns up 'heads' on one toss does not influence the outcome of any other toss.

In a sequence of n independent Bernoulli trials, each having probability of success π, the binomial random variable Y counts the number of successes, and we denote it by $Y \sim BIN(n, \pi)$. The probability mass function $f(y)$ for this distribution, which describes the probability of observing exactly y successes in a sequence of n Bernoulli trials, is given by

$$f(y) = \Pr(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$$

and is defined for $y = 0, 1, 2, \ldots, n$ and $\pi \in [0, 1]$. In practice, we obtain probabilities of interest by substituting particular values of y into this formula.

Note that as a special case, when $n = 1$, the binomial distribution simplifies to what is known as the **Bernoulli distribution** which is commonly used to describe response variables that are recorded on a binary scale, such as whether or not an experimental unit clicked or did not click a certain button, or whether an experimental unit did or did not bounce from a webpage. The probability mass function for the Bernoulli distribution is given by

$$f(y) = \Pr(Y = y) = \pi^y(1-\pi)^{1-y}$$

where $y = 0, 1$ and again $\pi \in [0, 1]$.

The Normal Distribution: The normal distribution is arguably the most important and most useful distribution in all of probability and statistics. The veracity of this bold claim will become evident as we work through the statistical analyses associated with different types of experiments. For now we motivate its utility in a practical way by simply stating that there are a remarkable number of real-life phenomenon that can be well-modeled by a normal distribution.

A random variable Y is said to be normally distributed if it takes on values $-\infty < y < \infty$ in accordance with the following probability density function

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

where $-\infty < \mu < \infty$ and $\sigma > 0$. We denote this random variable as $Y \sim N(\mu, \sigma^2)$ and remark that the shape of this distribution is completely determined by the parameters μ and σ. In particular, the distribution can qualitatively be described as 'bell-shaped and symmetrical' where μ determines the location of the axis of symmetry and σ determines the dispersion, or spread, of the distribution. Figure A.4 depicts a variety of normal density curves for various values of μ and σ and demonstrates that no matter the (μ, σ) combination, the distribution is always centered at μ and its dispersion is controlled by σ, with larger values corresponding to increased dispersion and smaller values corresponding to decreased dispersion. We note in passing that due to a constraint which says that the area beneath a density curve must equal 1, wider distributions are necessarily shorter than thinner distributions. This is also visualized in Figure A.4.

Note that an important special case exists when $\mu = 0$ and $\sigma = 1$; we call the $N(0,1)$ distribution the **standard normal distribution** and the corresponding random variable is typically denoted by the letter Z. It can be shown that the following transformation, which is known as *standardization*, can convert any normal random variable $Y \sim N(\mu, \sigma^2)$ into a standard normal random variable $Z \sim N(0,1)$:

$$Z = \frac{Y - \mu}{\sigma}$$

We will find the standard normal distribution very useful in the context of hypothesis testing.

The Student's *t*-Distribution: Another continuous distribution that is very useful in the context of hypothesis testing is the t-distribution, sometimes referred to as the "Student's" t-distribution (after the pseudonym[5] of William Gosset, the statistician who first derived it). Like the normal distribution, the t-distribution is 'bell-shaped and symmetrical', but unlike the normal distribution the t-distribution is always centered at 0 and its dispersion is determined by a parameter ν called the **degrees of freedom**. A random variable Y that follows a t-distribution with ν degrees of freedom is denoted $Y \sim t_{(\nu)}$ and the corresponding probability density function is given by

$$f(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{y^2}{\nu})^{-\frac{\nu+1}{2}}$$

for $-\infty < y < \infty$ and ν is a positive integer. Note that $\Gamma(a)$ is referred to as the "gamma function" and is

[5] Historical Note: William Gosset was an English statistician who worked at the Guiness Brewery in Dublin Ireland in the early 1900's. Due to a publication ban imposed by Guiness at the time (because of a previous leak of trade secrets), Gosset was forced to publish under the pseudonym *Student*.

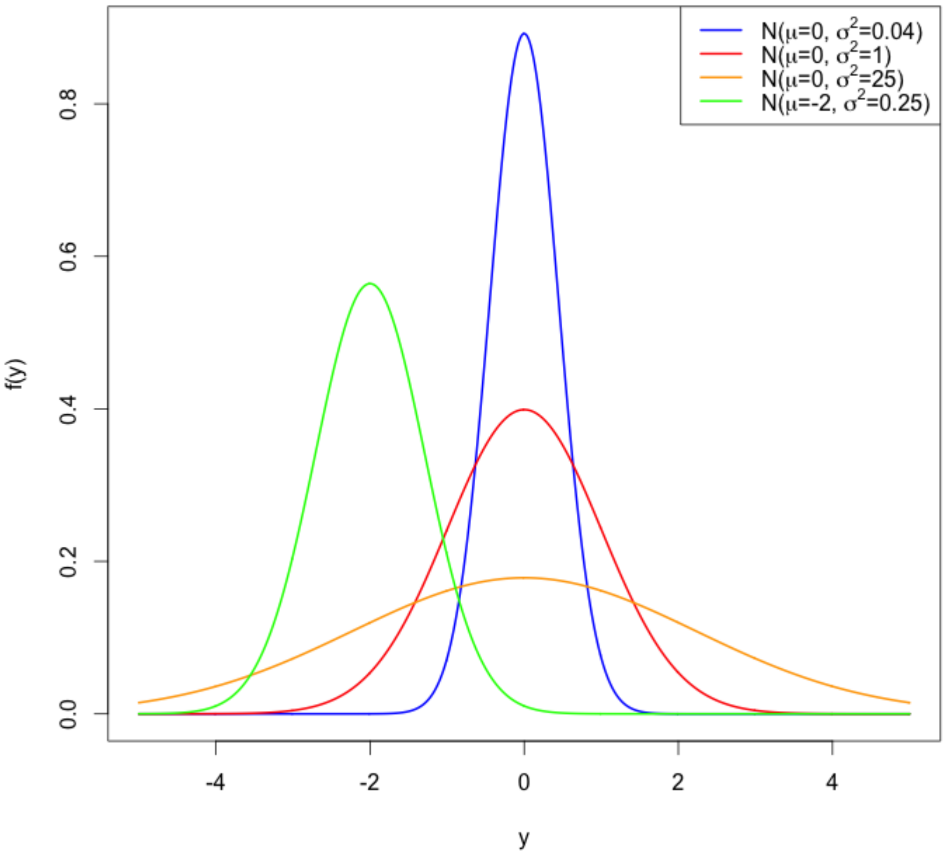

Figure A.4: A variety of normal density curves based on different values of μ and σ

evaluated as

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

which, if a is a positive integer, is $\Gamma(a) = (a-1)!$.

Figure A.5 depicts various t-distribution density curves and illustrates how dispersion depends on the degrees of freedom. Notably, as the number of degrees of freedom tends to infinity ($\nu \to \infty$), the t-distribution converges to the black curve. Although outside the scope of this Appendix, it can be shown that this black curve is the standard normal density curve. In other words

$$\lim_{\nu \to \infty} t_{(\nu)} = N(0, 1)$$

This will become a practically useful result in the context of various hypothesis tests when we are dealing with very large sample sizes, n.

The Chi-Squared Distribution: The chi-squared distribution (also called the χ^2-distribution) is another

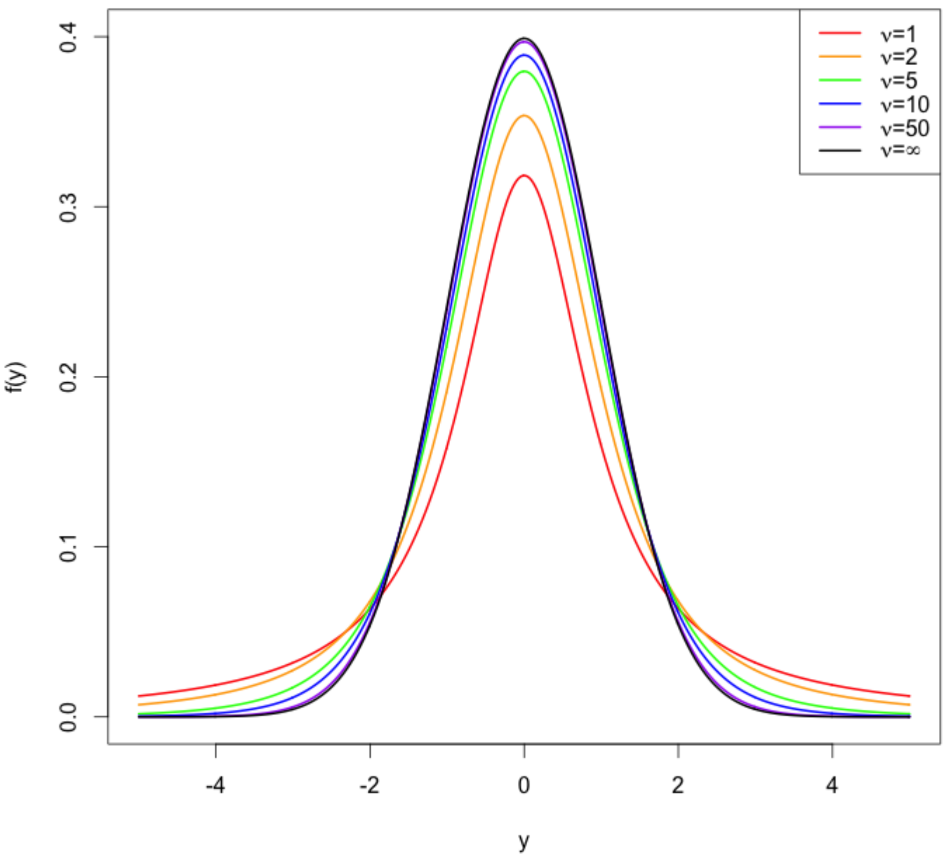

Figure A.5: A variety of t-distribution density curves based on different numbers of degrees of freedom ν

continuous distribution useful in the context of hypothesis testing whose shape is dependent upon a parameter ν called the degrees of freedom. A random variable Y that follows a chi-squared distribution with ν degrees of freedom is denoted $Y \sim \chi^2_{(\nu)}$, and its probability density function is given by

$$f(y) = \frac{y^{\frac{\nu}{2}-1}e^{-y/2}}{2^{\nu/2}\Gamma(\frac{\nu}{2})}$$

for $y \geq 0$ and where ν is a positive integer. Figure A.6 depicts a variety of chi-squared density curves corresponding to different values of ν. As we can see, the shape of the chi-squared distribution tends to be right-skewed, with a few special cases exhibiting exponential decay.

The F-Distribution: The F-distribution (also called Snedecor's F-distribution, after Ronald A. Fisher and George W. Snedecor) is another continuous distribution useful in the context of hypothesis testing whose shape is dependent upon two parameters ν_1 and ν_2 called the degrees of freedom. A random variable Y that follows an F-distribution with ν_1 and ν_2 degrees of freedom is denoted $Y \sim F_{(\nu_1,\nu_2)}$, and its probability

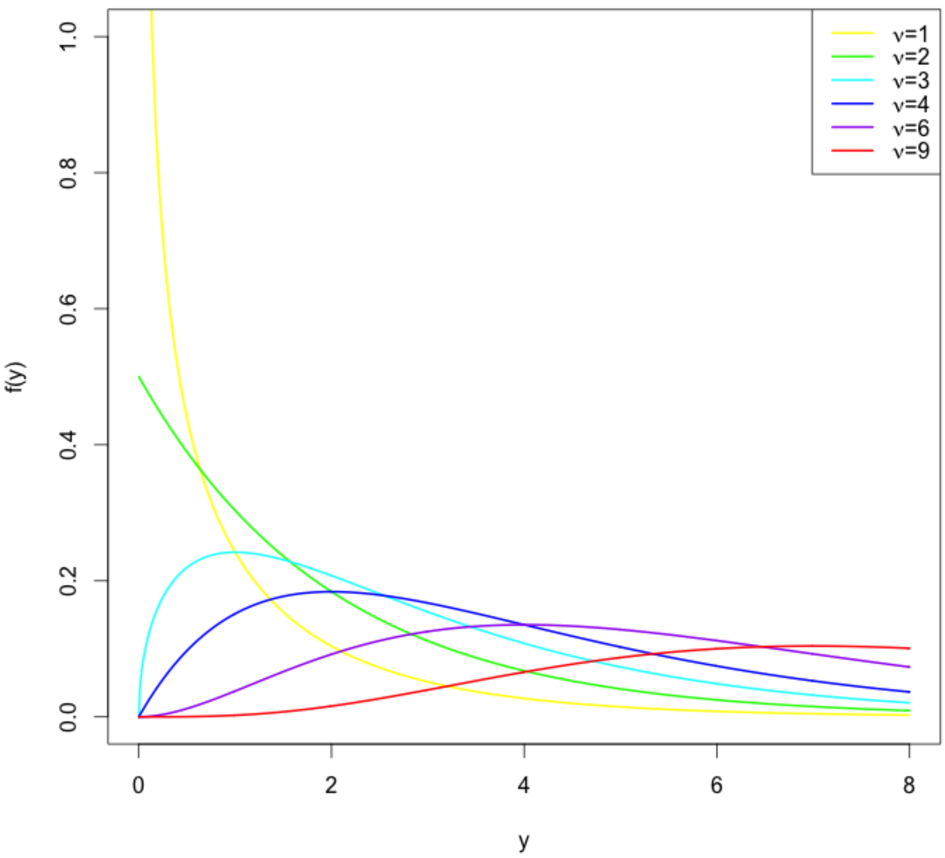

Figure A.6: A variety of χ^2-distribution density curves based on different numbers of degrees of freedom ν

density function is given by

$$f(y) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} y^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2} y\right)^{-\frac{\nu_1+\nu_2}{2}}$$

for $y \geq 0$ and where ν_1 and ν_2 are positive integers. Figure A.7 depicts are variety of F density curves corresponding to the different values of ν_1 and ν_2. As we can see, like the chi-squared distribution, the shape of the F-distribution tends to be right-skewed, with a few special cases exhibiting exponential decay.

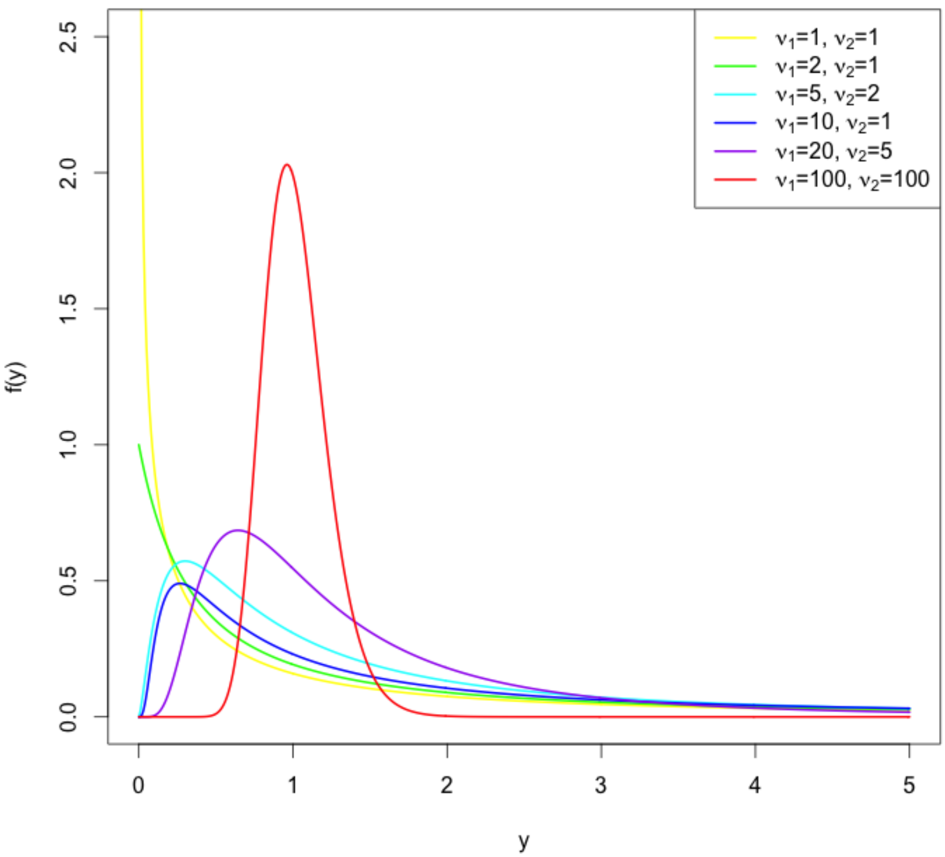

Figure A.7: A variety of F-distribution density curves based on different numbers of degrees of freedom ν_1 and ν_2

Relationships Between Random Variables:

- If $Z \sim \mathrm{N}(0,1)$, then $Z^2 \sim \chi^2_{(1)}$
- If $Z \sim \mathrm{N}(0,1)$ and $U \sim \chi^2_{(\nu)}$ and they're independent, then
$$\frac{Z}{\sqrt{U/\nu}} \sim t_{(\nu)}$$
- If $U_1 \sim \chi^2_{(\nu_1)}$ and $U_2 \sim \chi^2_{(\nu_2)}$ and they're independent, then
$$\frac{U_1/\nu_1}{U_2/\nu_2} \sim F_{(\nu_1,\nu_2)}$$
- If $X \sim t_{(\nu)}$ then $X^2 \sim F_{(1,\nu)}$
- If $W \sim F_{(\nu_1,\nu_2)}$ then $W^{-1} \sim F_{(\nu_2,\nu_1)}$

A.1.3 Expectation and Variance

Figures A.4, A.5, A.6, and A.7 demonstrate the variety of different shapes that a probability distribution can exhibit. Not only are these images visually pleasing, they are informative; with one glimpse we can tell which values of y seem typical and which seem extreme, we get sense of how dispersed the distribution is, and we can tell whether it is symmetrical or skewed. However, these observations – when gleaned from a plot – are informal. A quantitative method of communicating the shape of a distribution is with its **moments**. Before discussing moments, however, we must discuss the notion of **expectation**.

The **expected value** of a random variable Y, denoted $E[Y]$, is thought of as the 'average' value of Y and as a measure of center in Y's distribution. Mathematically, the expected value of Y is calculated as

$$E[Y] = \sum_{all\ y} yf(y)$$

if Y is a discrete random variable and as

$$E[Y] = \int_{all\ y} yf(y)dy$$

if Y is a continuous random variable.

Moments, then, are defined to be special expected values, which when taken together, completely specify the shape of a distribution. We define the k^{th} moment of Y to be $E[Y^k]$, which is calculated as in the preceding equations except that y^k (and not y) is multiplied by $f(y)$. Of particular importance in probability and statistics are the first four moments:

- The **first moment** $E[Y]$ quantifies the center of the distribution of Y
- The **second moment** $E[Y^2]$ quantifies the spread of the distribution of Y
- The **third moment** $E[Y^3]$ quantifies the skewness of the distribution of Y
- The **fourth moment** $E[Y^4]$ quantifies the kurtosis (or 'tailedness') of the distribution of Y

These four moments provide a tremendous amount of information about the distribution of Y. That said, in practice the first two moments are the ones used most frequently to describe a distribution's shape; relatively speaking more readily useful information is contained in the first two moments than in the others.

While the second moment $E[Y^2]$ itself provides information about the dispersion of a distribution, it is most commonly used in the calculation of the **variance** of Y, $Var[Y]$. The variance of a random variable Y is defined to be

$$Var[Y] = E[(Y - E[Y])^2]$$

and is interpreted as the expected squared deviation from the mean, with larger values indicating more dispersion and smaller values indicating less dispersion. It can be shown that the equation above can

Table A.1: Expected values and variances associated with some common distributions

Distribution	$E[Y]$	$Var[Y]$
$Y \sim BIN(n, \pi)$	$n\pi$	$n\pi(1-\pi)$
$Y \sim N(\mu, \sigma^2)$	μ	σ^2
$Y \sim t_{(\nu)}$	0	$\nu/(\nu-2)$
$Y \sim \chi^2_{(\nu)}$	ν	2ν
$Y \sim F_{(\nu_1, \nu_2)}$	$\frac{\nu_2}{\nu_2-2}$	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$

equivalently be expressed as

$$Var[Y] = E[Y^2] - E[Y]^2$$

which makes explicit the dependence of $Var[Y]$ on $E[Y^2]$. Note that the dispersion of a distribution is also commonly communicated in terms of the standard deviation of Y, denoted $SD[Y]$, and calculated as $SD[Y] = \sqrt{Var[Y]}$. Note that Table A.1 contains the expected values and variances of the five distributions described in the previous subsection. As can be seen, these rely entirely on the parameters associated with each distribution.

A.2 Statistical Inference

In practice, we often wish to study a particular characteristic, such as a response variable Y, in some **population** and make inferences about it. In most cases the population is too large to examine in its entirety and so we take a **sample** $\{y_1, y_2, \ldots, y_n\}$ from this population and generalize the conclusions drawn in the sample, applying them to the broader population. This process of generalizing sample information to the population from which it was taken is referred to as **statistical inference**. From a probabilistic point of view, we use probability distributions to model sample data, and assume that the chosen distribution is an accurate representation of Y at the population-level.

In the previous section we saw that much of a distribution's information is contained in its shape, and the shape of a given distribution relies entirely on one or more parameters. For instance, the binomial distribution depends on π, the normal distribution depends on μ and σ, both the t-distribution and the chi-squared distribution rely on degrees of freedom ν, and the F-distribution relies on two types of degrees of freedom, ν_1 and ν_2. In practice, however, the values of these parameters are unknown and interest typically lies in (i) estimating these parameters in light of the observed data, and/or (ii) testing hypotheses about the parameters. Here we discuss both types of statistical inference, but because the analysis of experiments typically involves testing one or more hypotheses of interest, we place more emphasis on (ii).

A.2.1 A Primer on Point and Interval Estimation

When a data scientist says that they are "fitting" a model to some data, what they really mean is:

- They've assumed a certain model or probability distribution is appropriate for describing some characteristic or relationship in a population.

- They have collected data (i.e., a sample from the population) with which they intend to study this characteristic or relationship.

- They intend to use the observed data to estimate the unknown parameters associated with the model or distribution.

Thus, the goal of **point estimation** is to use observed data to obtain reasonable values of a model's unknown parameters (call them θ) that are consistent with the data that were actually observed. Whereas we typically use Greek letters to denote unknown parameters we use Greek letters over scored by a circumflex (a "hat"[6]), i.e., $\hat{\theta}$, to denote its corresponding estimate. In general, a variety of estimation methods may be used to obtain parameter estimates: the method of moments, maximum likelihood estimation and least squares estimation, to name a few. All estimation procedures have advantages and disadvantages, and so it is important to choose the one that is appropriate for your data and your problem.

It is also important to distinguish between point estimation and **interval estimation**. In the context of point estimation we use our data to obtain a single estimate of θ. However, if we were to draw a second sample and repeat the exact same estimation procedure we would very likely obtain a slightly different value of $\hat{\theta}$ than before, simply due to sampling variation. Given this sampling variation, how would you know if your estimate is a good one? In other words, how do you know if your estimate is anywhere close to the true, unknown, value of θ? The reality is that we can't know this. However, rather than calculating just a point estimate of θ, we can also calculate an interval estimate, more commonly known as a **confidence interval**, for θ. Doing so acknowledges that a point estimate, although likely close to the parameter's true value, is probably not exactly equal to the parameter's true value. Such an interval provides a range within which we are reasonably certain the true value of θ lies. Thus in addition to reporting point estimates of a parameter θ it is most informative to also report a confidence interval for θ as well. For a thorough, but introductory, overview of point and interval estimation techniques see Bain and Engelhardt (1992).

A.2.2 A Primer on Hypothesis Testing

In the context of point and interval estimation we treat the parameter θ as completely unknown and something we need to estimate. However, in some circumstances we may have a belief about the value of θ, and we may wish to use sample data to evaluate whether or not that belief seems reasonable. Statistically speaking such a belief is called a **hypothesis** and the use of data to evaluate that belief is referred to as **hypothesis testing**.

Suppose we believe $\theta = \theta_0$. A formal hypothesis statement corresponding to this can be framed as

$$H_0\colon\ \theta = \theta_0 \text{ vs. } H_A\colon\ \theta \neq \theta_0$$

[6] The notation $\hat{\theta}$ is read "θ-hat".

We call H_0 the **null hypothesis** and it is the statement we believe to be true, and that we want to test using observed data. The statement denoted H_A is called the **alternative hypothesis** and it is the complement of H_0. Thus, exactly one statement is true – either the null hypothesis or the alternative hypothesis – and we use observed data to try and empirically uncover the truth. Note that according to H_A values of θ both larger and smaller than θ_0 correspond to H_0 being false, and so we call such a test **two-sided**. This is to be contrasted with **one-sided** tests for which values of θ larger than θ_0 *or* values of θ smaller than θ_0 (but not both) correspond to H_0 being false. One-sided hypotheses can be stated as

$$H_0\colon\ \theta \leq \theta_0 \text{ vs. } H_A\colon\ \theta > \theta_0$$

or

$$H_0\colon\ \theta \geq \theta_0 \text{ vs. } H_A\colon\ \theta < \theta_0$$

depending on the context of the problem and the question that the hypothesis test is designed to answer. No matter which hypothesis is appropriate, the goal is always the same: based on the observed data, we will decide to *reject* H_0 or *not reject* H_0.

In order to draw such a conclusion, we define a **test statistic** T which is a random variable that satisfies three properties: (i) it must be a function of the observed data, (ii) it must be a function of the parameter θ, and (iii) its distribution must not depend on θ. Assuming the null hypothesis is true, the test statistic T follows a particular distribution which we call the **null distribution**. We then calculate t, the observed value of the test statistic, by substituting the observed data and the hypothesized value of θ into the expression for T. Note that expressions for t commonly incorporate terms of the form $\hat{\theta} - \theta_0$ or $\hat{\theta}/\theta_0$. and so the data enter the expression through the parameter's estimate $\hat{\theta}$.

Next we evaluate the extremity of t relative to the null distribution. If t seems very extreme, as though it is very unlikely to have come from the null distribution, then this gives us reason to believe that the null distribution may not be appropriate. On the other hand, if t appears as though it could have come from the null distribution, then there is no reason to believe the null distribution is inappropriate. The left and right panels of Figure A.8 illustrate these two cases. On the left, the value of t is not at all unreasonable in the context of the null distribution. However, on the right, the value of t is very extreme and would have been very unlikely if the null distribution (and hence the null hypothesis) really were true. Thus when we observe very extreme values of a test statistic it provides evidence against the null hypothesis, and leads us to believe that perhaps H_0 is not true; and the more extreme t is, the more evidence we have against H_0. With enough evidence (i.e., extreme enough t) we will choose to reject the null hypothesis.

We formalize the extremity of t using the **p-value** of the test. Probabilistically speaking, a p-value is defined to be the probability of observing a value of the test statistic *at least as extreme* as the value we observed, if the null hypothesis is true. Thus the p-value formally quantifies how "extreme" the observed

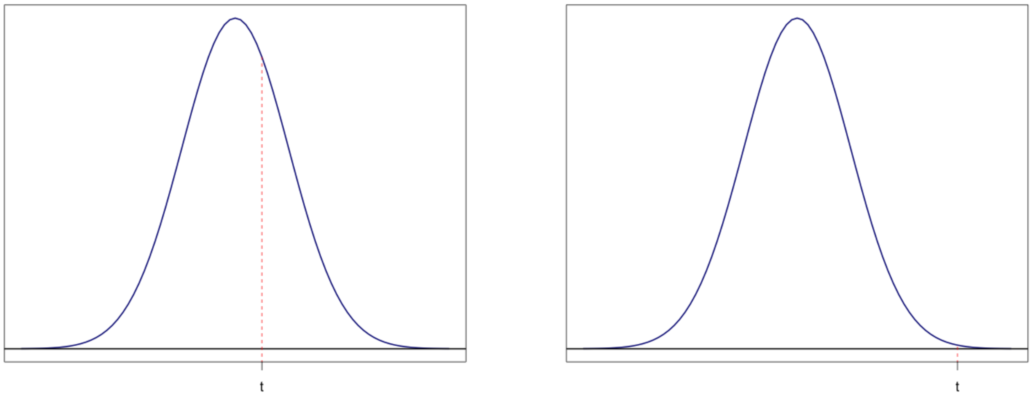

Figure A.8: Left: A non-extreme value of a test statistic; Right: An extreme value of a test statistic

test statistic is. Whether large values of t, small values of t, or both, are to be considered extreme depends on whether H_A is one- or two-sided. When H_A is two-sided, both large and small values of t are considered extreme and we define the p-value mathematically as

$$\text{p-value} = \Pr(T \geq |t|) + \Pr(T \leq -|t|)$$

which, if the null distribution is symmetrical, is equivalent to $2\Pr(T > |t|)$. The left panel of Figure A.9 provides a visual depiction of this calculation. Note that this particular calculation is relevant when t involves a difference such as $\hat{\theta} - \theta_0$. But when t involves a ratio such as $\hat{\theta}/\theta_0$ we typically calculate the p-value as

$$\text{p-value} = \Pr(T \geq t) + \Pr(T \leq 1/t)$$

The exact formulae used to calculate a p-value may differ depending the type of data observed and the nature of the hypothesis being tested, but in all cases one simply needs to consider which values of t would provide evidence against H_0. For instance, when H_A is one-sided then either large values of t or small values of t are considered extreme, and this depends on the direction of the inequality in H_A. If H_A: $\theta > \theta_0$, values of θ larger than θ_0 and hence large values of t will provide evidence against H_0. Thus in this case large values of t are considered extreme and the p-value is calculated as

$$\text{p-value} = \Pr(T \geq t).$$

The center panel of Figure A.9 provides a visual depiction of this calculation. If H_A: $\theta < \theta_0$, values of θ smaller than θ_0 and hence small values of t will provide evidence against H_0. Thus in this case small values of t are considered extreme and the p-value is calculated as

$$\text{p-value} = \Pr(T \leq t).$$

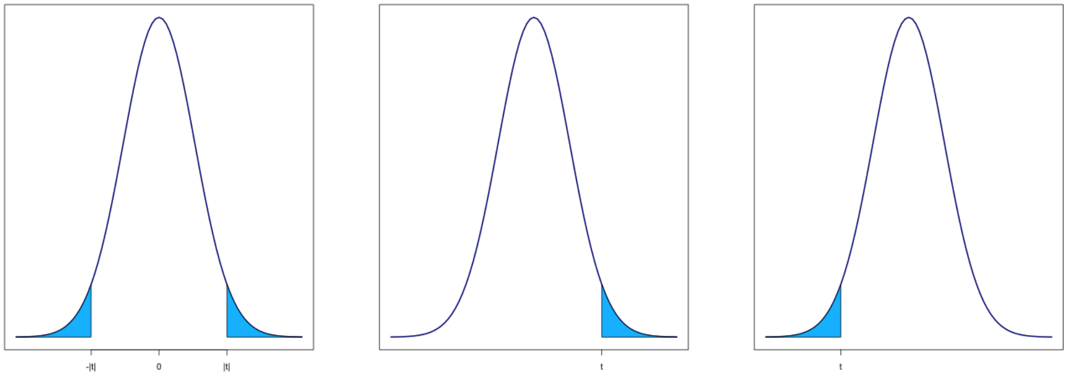

Figure A.9: Illustration of the calculation of p-values in one- and two-sided tests

The right panel of Figure A.9 provides a visual depiction of this calculation.

How "extreme" t must be, and hence how small the p-value must be to reject H_0, is determined by the **significance level** of the test, which we denote by α. In particular, if

- p-value $\leq \alpha$ we reject H_0
- p-value $> \alpha$ we do not reject H_0

Note that $\alpha = 0.01$ or 0.05 are common choices. In order to motivate these choices we need to discuss the two types of error that can be made when drawing conclusions in the context of a hypothesis test.

Recall that by design either H_0 is true or H_A is true. This means that there are four possible outcomes when using data to decide which statement is true:

(1) H_0 is true and we correctly do not reject it

(2) H_0 is true and we incorrectly reject it

(3) H_0 is false and we incorrectly do not reject it

(4) H_0 is false and we correctly reject it

Obviously scenarios (1) and (4) are ideal since in them we are making the correct decision, and (2) and (3) should be avoided since in those scenarios we are not making the correct decision. Scenarios (2) and (3) are respectively referred to as **Type I Error** and **Type II Error**. Clearly we would like to reduce the likelihood of making either type of error, but it is important to recognize that in practice there are different consequences to each type of error, and so we may wish to treat them differently. To make this point clear, consider a courtroom analogy where the defendant is assumed innocent until proven guilty. This hypothesis can be stated formally as

$$H_0\text{: the defendant is innocent vs. } H_A\text{: the defendant is guilty}$$

Within this analogy a Type I Error occurs when the defendant is truly innocent, but the evidence leads the jury to find the defendant guilty. Thus, this error leads to an innocent person being convicted of a crime they did not commit. A Type II Error, on the other hand, occurs when the defendant is truly guilty, but the evidence leads the jury to find the defendant innocent. In this case the error leads to a criminal being set free. In this analogy, and in any hypothesis testing setting, both types of errors lead to negative outcomes, but these negative outcomes may be prioritized differently.

Fortunately it is possible to control the frequency with which these types of errors occur. We do so by controlling the significance level and **power** of the test. We define a test's significance level to be $\alpha = \Pr(\text{Type I Error})$ and we define the power of a test to $1 - \beta$ where $\beta = \Pr(\text{Type II Error})$. Thus it is desirable to have a test with a small significance level and a large power since this corresponds to simultaneously reducing both types of errors.

In practice we choose α and β based on how often we are comfortable allowing Type I and Type II Errors to occur. For instance, if we can only tolerate a Type I Error 1% of the time, then we would choose $\alpha = 0.01$ and if we can only tolerate making a Type II Error 5% of the time, then we would choose $\beta = 0.05$. With these choices we would say that the corresponding hypothesis test has a 1% significance level and 95% power. Common choices for significance level and power are respectively 5% and 80%, corresponding to $\alpha = 0.05$ and $\beta = 0.2$.

As is now apparent, the significance level α (i.e., the probability of making a Type I Error), determines how small a p-value must be (and hence how extreme t must be) in order to reject a null hypothesis. This decision should be made prior to testing the hypothesis and in fact prior to collecting any data. We defer a discussion of controlling power until Chapter 2 where we will see that for a fixed value of α the power determines the sample size and so it is also a decision that should be made prior to collecting the data, else you will not know how much data to collect.

A.3 Linear Regression

In this section we provide a brief overview of **linear regression**. Linear regression is a form of statistical modeling that is appropriate when interest lies in relating a response variable (Y) to one or more explanatory variables ($x_1, x_2, \ldots, x_p$). The idea is that Y is influenced in some manner by the explanatory variables through an unknown function $f(\cdot)$:

$$Y = f(x_1, x_2, \ldots, x_p).$$

The purpose of statistical modeling in general, and linear regression in particular, is to approximate this function $f(\cdot)$. The typical linear regression model in this situation is given by

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

where Y is the response variable; the x_j's are explanatory variables which we treat as fixed (not random) quantities; the β's are unknown parameters that quantify the influence of a particular explanatory variable on the response; and $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ is a random error term that accounts for the fact that $f(x_1, x_2, \ldots, x_p) \neq \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ exactly. The distributional assumption for ε has several consequences. Chief among them is that Y is also a random variable and

$$Y \sim \mathrm{N}(\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2).$$

Thus, for particular values of the explanatory variables, we expect the response variable to be equal to $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, on average. Variation around this relationship is quantified by σ^2.

Based on this distributional assumption we find that $\mathrm{E}[Y|x_1 = x_2 = \cdots = x_p = 0] = \beta_0$, and so β_0 is interpreted as the intercept of the model – the expected response when all of the explanatory variables are equal to zero. Also, notice that

$$\begin{aligned}
\mathrm{E}[Y|x_j = x+1] - \mathrm{E}[Y|x_j = x] &= (\beta_0 + \beta_1 x_1 + \cdots + \beta_j(x+1) + \cdots + \beta_p x_p) \\
&\quad -(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \cdots + \beta_p x_p) \\
&= \beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \beta_j + \cdots + \beta_p x_p \\
&\quad -\beta_0 - \beta_1 x_1 - \cdots - \beta_j x - \cdots - \beta_p x_p \\
&= \beta_j
\end{aligned}$$

As such, β_j is interpreted as the expected change in response associated with a unit increase in x_j ($j = 1, 2, \ldots, p$), while holding all other explanatory variables fixed. Given the intuitively pleasing interpretations of these coefficients, it should be clear that linear regression models are well-suited for *explanatory modeling*, although they may also be used effectively for *predictive modeling*. See Shmueli (2010) for an interesting discussion of these two goals of statistical modeling.

Whether we wish to use a linear model for explanatory or predictive purposes, we need to estimate the regression coefficients. Recall the β's are unkown parameters. This is typically done with **least squares estimation** where the goal is to find the values of $\beta_0, \beta_1, \ldots, \beta_p$ that minimize the error, ε, associated with the model. Specifically, for observed data given by $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, 2, \ldots, n$, we wish to minimize

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2$$

with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$.

By writing the linear regression model above in matrix form as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

it can be shown that the least squares estimate of $\boldsymbol{\beta}$ and hence the individual β's is given by $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ where

- $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is an $n \times 1$ vector of response variable observations,
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ is an $n \times 1$ vector of random errors,
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $(p+1) \times 1$ vector of regression coefficients, and
- X is the following $n \times (p+1)$ matrix of explanatory variable observations

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

With the regression coefficients estimated we define the **fitted values** $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$ to be the estimated expected response for specific values of the explanatory variables. Next we define the **residuals** $e_i = y_i - \hat{\mu}_i$ to be the difference between the observed value of the response and what the model predicts the response to be. It can be shown that the least squares estimate of σ^2 is based on the residuals, and in particular is given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2}{n-p-1}.$$

This estimate is sometimes referred to as the **mean squared error** (MSE) of the model, since $\hat{\sigma}$ quantifies the typical squared distance (i.e., squared error) between an observed response value and the value predicted by the model.

Having estimated $\beta_0, \beta_1, \ldots, \beta_p$ and σ^2, the fitted linear regression model can be used for inference and prediction. Of particular importance are hypothesis tests for the individual β's. For instance, the hypothesis H_0: $\beta_j = 0$ vs. H_A: $\beta_j \neq 0$ may be used to formally evaluate whether the explanatory variable x_j signficantly influences Y and whether it belongs in the model. Also of importance are confidence and prediction intervals for predicted values of Y. For a much more thorough (yet approachable) treatment of linear regression see Abraham and Ledolter (2006).

A.4 Logistic Regression

As was shown in the previous section, linear regression is an effective method of modeling the relationship between a single response variable (Y), and one or more explanatory variables ($x_1, x_2, \ldots, x_p$). However, ordinary linear regression assumes that the response variable follows a normal distribution (i.e., $Y \sim N(\mu, \sigma^2)$); when the response variable is binary, this assumption is no longer valid. In the context of a binary response, the Bernoulli distribution (i.e., $Y \sim BIN(1, \pi)$) is a much more appropriate distributional assumption, but ordinary linear regression is no longer appropriate. Consequently, we use **logistic regression** to model the relationship between a binary response variable and one or more explanatory variables.

In the context of a linear regression model, the model is formulated so that the expected response (given the values of the explanataory variables) is equated to the **linear predictor** $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. Namely

$$\mathrm{E}[Y|x_1, x_2, \ldots, x_p] = \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

In the context of logistic regression we would also like to relate the expected response to the linear predictor, but in this case $\mathrm{E}[Y] = \pi$ which is a probability that must lie within $[0, 1]$. However, it is unrealistic to impose this constraint on the linear predictor and so equating π and $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ does not make sense. Instead, we relate the linear predictor to $\mathrm{E}[Y] = \pi$ through a monotonic differentiable **link function** that maps $[0, 1]$ to the real the numbers. Logistic regression arises when this link function is chosen to be the **logit** function:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{45}$$

Note that different link functions result in different **generalized linear models**. See McCullagh and Nelder (1989) for a thorough and more general treatment of generalized linear models.

Notice that by inverting the link function the expected response (given the values of the explanataory variables) in a logistic regression is

$$\mathrm{E}[Y|x_1, x_2, \ldots, x_p] = \pi = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}.$$

To interpret β_0 we consider setting each explanatory variable to zero (i.e., $x_1 = x_2 = \cdots = x_p = 0$) in equation (45). In doing so we see that β_0 is the **log-odds** that $Y = 1$, or in other words, that e^{β_0} is the **odds** that the response would equal 1 when $x_1 = x_2 = \cdots = x_p = 0$. The interpretation of β_j, for $j = 1, 2, \ldots, p$, is seen by considering equation (45) for different values of x_j. In particular, let π_x be the value of π when $x_j = x$ and let π_{x+1} be the value of π when $x_j = x + 1$, and notice that

$$\begin{aligned}
\log\left(\frac{\pi_{x+1}}{1-\pi_{x+1}}\right) - \log\left(\frac{\pi_x}{1-\pi_x}\right) &= (\beta_0 + \beta_1 x_1 + \cdots + \beta_j(x+1) + \cdots + \beta_p x_p) \\
&\quad -(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \cdots + \beta_p x_p) \\
&= \beta_0 + \beta_1 x_1 + \cdots + \beta_j x + \beta_j + \cdots + \beta_p x_p \\
&\quad -\beta_0 - \beta_1 x_1 - \cdots - \beta_j x - \cdots - \beta_p x_p \\
&= \beta_j
\end{aligned}$$

and simplifying the left hand side yields

$$\log\left(\frac{\pi_{x+1}}{1-\pi_{x+1}} \Big/ \frac{\pi_x}{1-\pi_x}\right) = \beta_j$$

which makes it clear that β_j is interpreted as a **log-odds ratio** and hence e^{β_j} is interpreted as the **odds ratio**, comparing the odds that $Y = 1$ when $x_j = x + 1$ vs. when $x_j = x$ (all else being equal). We note in passing that **maximum likelihood estimation** is typically used to estimate the β's and hence π. This derivation is outside the scope of these notes, so the interested reader is referred to McCullagh and Nelder (1989) for more details.

Just as in the case of linear regression we can perform hypothesis tests to determine whether subsets of the regression coefficients are equal to zero. Practically this amounts to fitting and comparing nested models (full models and reduced models that assume H_0 is true) to determine whether they differ significantly. Because maxmimum likelihood estimation is used to fit such models, this comparison is made using **likelihood ratio tests**. In order to test individual hypotheses about particular regression coefficients, like $H_0 : \beta_j = 0$, we can similarly apply a likelihood ratio test that compares the models with and without β_j. Alternatively, we may also use individual Z-tests for this purpose.

References

Abraham, B. and J. Ledolter (2006). *Introduction to regression modeling.* Thomson Brooks/Cole.

Bain, L. J. and M. Engelhardt (1992). *Introduction to probability and mathematical statistics* (2nd ed.). Brooks/Cole.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association 50*(272), 1096–1121.

Edgington, E. and P. Onghena (2007). *Randomization tests.* CRC Press.

Efron, B. and T. Hastie (2016). *Computer age statistical inference*, Volume 5. Cambridge University Press.

Georgiev, G. Z. (2019). *Statistical Methods in Online A/B Testing.* Self Published.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Jensen, W. A. (2016). Confirmation runs in design of experiments. *Journal of Quality Technology 48*(2), 162–177.

Li, X., N. Sudarsanam, and D. D. Frey (2006). Regularities in data from factorial experiments. *Complexity 11*(5), 32–45.

McCullagh, P. and J. A. Nelder (1989). Generalized linear models, no. 37 in monograph on statistics and applied probability.

McFarland, C. (2012). *Experiment!: Website conversion rate optimization with A/B and multivariate testing.* New Riders.

Miroglio, B., D. Zeber, J. Kaye, and R. Weiss (2018). The effect of ad blocking on user engagement with the web. In *Proceedings of the 2018 World Wide Web Conference*, pp. 813–821.

Montgomery, D. C. (2019). *Design and analysis of experiments* (10th ed.). John Wiley & Sons.

Myers, R. M., D. C. Montgomery, and C. M. Anderson-Cook (2016). *Response Surface Methodology: process and product optimization using designed experiments* (4 ed.). John Wiley Sons, Inc.

Ramdas, A. K., R. F. Barber, M. J. Wainwright, M. I. Jordan, et al. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics 47*(5), 2790–2821.

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.

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 52*(3/4), 591–611.

Shmueli, G. (2010). To explain or to predict? *Statistical Science 25*(3), 289–310.

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

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), 626–633.

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.

Stevens, N. T. and C. M. Anderson-Cook (2019). Design and analysis of confirmation experiments. *Journal of Quality Technology 51*(2), 109–124.

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205.

Tartakovsky, A., I. Nikiforov, and M. Basseville (2014). *Sequential analysis: Hypothesis testing and change-point detection*. CRC Press.

Thomke, S. H. (2020). *Experimentation Works: The Surprising Power of Business Experiments*. Harvard Business Press.

Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika 34*(1/2), 28–35.

Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics*, 1005–1013.

Wu, C. J. and M. S. Hamada (2011). *Experiments: planning, analysis, and optimization*, Volume 552. John Wiley & Sons.