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

A Continuing Education Certificate

at

The Univeristy of San Francisco's Data Institute

September 6 - October 18, 2017

Instructor: Nathaniel T. Stevens

ntstevens@usfca.edu

Preface

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

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

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

multivariate experiments such as factorial and fractional-factorial designs, and optimization techniques such as multi-armed bandit experiments and response surface methodology.

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

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

Introduction

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

1.1 Notation and Nomenclature

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

We call a variable a **response variable** when it is the one of primary interest. This is the variable that the problem/question will be defined in terms of. In practice these tend to be performance metrics such as key performance indicators (KPIs) like conversion rates, average purchase size, bounce rate, page views or session duration, to name just a few. It is the response variable that one wishes to optimize. For example, an experiment may involve comparing different messages on a call-to-action button to find which message maximizes the click through rate (CTR). Or, an experiment may involve the comparison of different webpage designs to decide which one minimizes that page's bounce rate. Regardless of the type or goal of the experiment, the response variable is the one we are primiary interested in. Throughout this course we will use the letter y to denote response variables.

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

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

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

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

1.2 Experiments versus Observational Studies

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

As mentioned above, generally speaking, the goal in these sorts of investigations is to evaluate the change in response associated with a change in the factors. Strictly speaking one does not require an experiment to do this. Establishing these sorts of relationships can also be done with **observational studies**. The distinction between this and an experiment is that in an observational study there is no measure of control in the data collection process. Instead, data are recorded passively and any relationship between the response and factors is observed organically. While such an approach provides information about the association between these factors, it does not provide clear information about a causal relationship.

When **causal inference** (establishing causal connections between variables) is of interest, it is best if the data arise as a result of an experiment. While methods for establishing causal relationships from observational data do exist (see e.g., propensity score matching (Rosenbaum and Rubin, 1983)), they are much less sound and much more error prone than a carefully designed experiment.

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

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

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

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

1.3 QPDAC: A Strategy for Answering Questions with Data

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

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

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

is related to this question and whose variation can be quantified.

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

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

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

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

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

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

With the units defined, conditions established, and the response variable chosen, the final decisions to be made concern the number of units to assign to each condition, and the manner in which this assignment is made. Power analyses and sample size calculations are used to address the former concern and the sampling mechanism addresses the latter. While

random assignment is the standard approach, other hierarchical assignment strategies such as stratified or segmented sampling are also common. We elaborate on these topics later on in the course.

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

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

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

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

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

1.4 Fundamental Principles of Experimental Design

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

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

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

one another as can be, and thus any observed difference in response values can be attributed to the differences between the conditions themselves.

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

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

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

Thus we run the experiment at fixed levels of the nuisance factors, i.e., within **blocks**.

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

1.5 Exercise: The Instagram Experiment

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

How would you design such an experiment?

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?

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

$$H_0\colon\ \theta_1 \leq \theta_2 \text{ vs. } H_A\colon\ \theta_1 > \theta_2 \tag{2.1}$$

or

$$H_0\colon\ \theta_1 \geq \theta_2 \text{ vs. } H_A\colon\ \theta_1 < \theta_2 \tag{2.2}$$

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 (2.1) we will reject H_0 (and conclude that red is best), and if we test (2.2) we will not reject H_0 (and hence conclude that red is best).

Note that when comparing some metric of interest across two conditions, it may also be of interest to test the two-sided hypothesis

$$H_0\colon\ \theta_1 = \theta_2 \text{ vs. } H_A\colon\ \theta_1 \neq \theta_2. \tag{2.3}$$

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 (2.1) or (2.2) 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 (2.1), (2.2) and (2.3), 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

$$H_0\text{: } \mu_1 = \mu_2 \text{ vs. } H_A\text{: } \mu_1 \neq \mu_2 \tag{2.4}$$

$$H_0\text{: } \mu_1 \leq \mu_2 \text{ vs. } H_A\text{: } \mu_1 > \mu_2 \tag{2.5}$$

$$H_0\text{: } \mu_1 \geq \mu_2 \text{ vs. } H_A\text{: } \mu_1 < \mu_2 \tag{2.6}$$

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 (2.4), (2.5) and (2.6) we must first calculate a **test statistic**. Because $Y_{ij} \sim \text{N}(\mu_j, \sigma^2)$, it is also true that $\bar{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \sim \text{N}(\mu_j, \frac{\sigma^2}{n_j})$ and hence that

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \text{N}(0, 1). \tag{2.7}$$

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

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

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

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \tag{2.8}$$

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 (2.4), (2.5) and (2.6) 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 $\bar{y}_1 = \hat{\mu}_1$ and $\bar{y}_2 = \hat{\mu}_2$ and so the observed test statistic is given by

$$\begin{aligned} t &= \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(\hat{\mu}_1 - \hat{\mu}_2) - 0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \end{aligned} \tag{2.9}$$

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 (2.4) the p-value is calculated as p-value $= 2P(T \geq |t|)$; in the case of hypothesis (2.5) the p-value is calculated as p-value $= P(T \geq t)$, and in the case of hypothesis (2.6) the p-value is calculated as p-value $= P(T \leq t)$, 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.

Next we consider an example to illustrate these ideas more concretely.

2.1.2 Example: Instagram Ad Frequency

Suppose, again, 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 one ad every 8 posts in their social feed, but, in order to increase ad revenue, your manager is pressuring your team to show one 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 hypothesize 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 (2.5) 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 $\hat{\mu}_1 = \bar{y}_1 = 4.9162$ with a standard deviation of $s_1 = 0.9634$, and in the 7:1 condition the average session time is $\hat{\mu}_2 = \bar{y}_2 = 3.0518$

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 (2.9) gives

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

The p-value associated with this test is $P(T \geq 30.1013)$ 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 2.1.4.

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. In particular, you 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.

In fact, depending on the speed a user scrolls through their feed, this this increased ad frequency could actually reduce ad revenue; 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.

2.1.3 Power Analysis and Sample Size Calculations

When designing a two-condition experiment (i.e., an A/B test), 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

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

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

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

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 (2.7) which means that we will be working with the $\mathrm{N}(0, 1)$ 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 (2.4) 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 (2.5) and (2.6) 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 1.

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

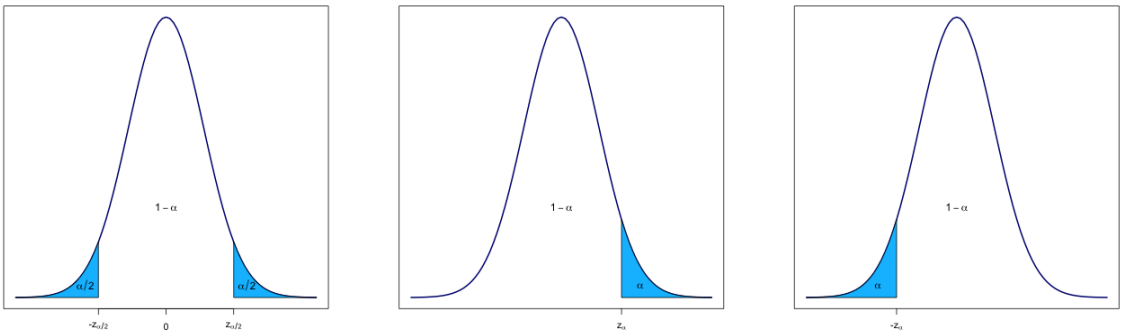

Figure 1: Rejection regions corresponding to one and two-sided hypotheses

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 (2.4). We indicate where and how this derivation changes 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:

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

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 none-zero δ, and so it is

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

that follows a $N(0,1)$ 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.

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

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

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

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:

$$n_2 = \frac{(\frac{1}{k} + 1)(z_{\alpha/2} - z_{1-\beta})^2\sigma^2}{\delta^2} \tag{2.10}$$

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

$$n = \frac{2(z_{\alpha/2} - z_{1-\beta})^2\sigma^2}{\delta^2}. \tag{2.11}$$

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 **effect size** of the test. The effect size for hypothesis tests like (2.4), (2.5) or (2.6) 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 2.1.2. 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 effect size, and is what is captured by $\delta = \mu_1 - \mu_2$.

Sometimes effect size 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 (2.10) simplifies to

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

and the sample size formula (2.11) simplifies to

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

The advantage of defining effect size 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 (2.10) and (2.11) but with $z_{\alpha/2}$ replaced by z_α.

As should be clear by looking at equations (2.10) and (2.11), there is an interdependent relationship between sample size, significance level, power, and effect size. 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/.

2.1.4 When Assumptions are Invalid

When testing hypotheses of the form (2.4), (2.5) and (2.6) 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 mea-

surements in each condition have equal variances, an approach that accommodates $Y_{ij} \sim \mathrm{N}(\mu_j, \sigma_j^2)$ 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

$$H_0\text{: } \sigma_1^2 = \sigma_2^2 \text{ vs. } H_A\text{: } \sigma_1^2 \neq \sigma_2^2 \tag{2.12}$$

Such a hypothesis is commonly tested using an ***F*-test of equal variances**. The F-test assumes that $Y_{ij} \sim \mathrm{N}(\mu_j, \sigma_j^2)$ which consequently means that

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

and hence that

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

Assuming H_0 is true, $\sigma_1^2/\sigma_2^2 = 1$ and so the observed value of the test statistic is

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

which we compare 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 to be

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

since values greater than or equal to t and less than or equal to $1/t$ are what is considered "at least as extreme" in this situation. One-sided alternatives might also be considered where H_A: $\sigma_1^2 > \sigma_2^2$ or H_A: $\sigma_1^2 < \sigma_2^2$ in which case the p-values are respectively defined as p-value $= P(T \geq t)$ and p-value $= P(T \leq t)$. This test can be carried out in `R` using the `var.test()` function.

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 techniques 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 (2.1), (2.2) or (2.3) to decide whether $\theta_1 = \theta_2$, $\theta_1 > \theta_2$ or $\theta_1 < \theta_2$. The philosophy behind the resampling approaches to testing such hypotheses is described below.

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 sub-

stantially 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** takes the test statistic to be $t = \hat{\theta}_1 - \hat{\theta}_2$ and then takes as the null distribution the set of test statistics calculated on each of the $\binom{n_1+n_2}{n_1} = \binom{n_1+n_2}{n_2}$ arrangements of data. 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.

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 resampled '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 B times ($B = 1000$ or 2000 are common choices).
5. Compare t to the null distribution which is derived from the B 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 $t^* < t$

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:

$$H_0\text{: } \pi_1 = \pi_2 \text{ vs. } H_A\text{: } \pi_1 \neq \pi_2 \tag{2.13}$$

$$H_0\text{: } \pi_1 \leq \pi_2 \text{ vs. } H_A\text{: } \pi_1 > \pi_2 \tag{2.14}$$

$$H_0\text{: } \pi_1 \geq \pi_2 \text{ vs. } H_A\text{: } \pi_1 < \pi_2 \tag{2.15}$$

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 (2.13), (2.14) and (2.15) we must calculate a test statistic. Due to the **Central Limit Theorem**[1] we know that for large enough n_j the random variable $\bar{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \dot{\sim} \text{N}(\pi_j, \frac{\pi_j(1-\pi_j)}{n_j})$. Thus, with a large amount of replication $\bar{Y}_{ij}$ will approximately follow a normal distribution. Based on this result

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \dot{\sim} \text{N}(0,1). \tag{2.16}$$

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 their estimates $\hat{\pi}_1$ and $\hat{\pi}_2$, which are respectively equal to $\bar{Y}_1$ and $\bar{Y}_2$. 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.

[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 $\bar{X}$ follows a $\text{N}(\mu, \sigma^2)$ distribution for large enough n (i.e., as $n \to \infty$).

Making these substitutions gives

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \tag{2.17}$$

which also approximately follows a $\mathrm{N}(0,1)$ distribution. Thus T is the test statistic associated with hypotheses (2.13), (2.14) and (2.15) where $\mathrm{N}(0,1)$ 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 $\mathrm{N}(0,1)$ distribution. Given the sample data, we have $\bar{y}_1 = \hat{\pi}_1$ and $\bar{y}_2 = \hat{\pi}_2$ and so the observed test statistic is given by

$$\begin{aligned} t &= \frac{(\bar{y}_1 - \bar{y}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \\ &= \frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \\ &= \frac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \end{aligned} \tag{2.18}$$

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 (2.13), (2.14) and (2.15) are respectively calculated as $2P(T \geq |t|)$, $P(T \geq t)$, and $P(T \leq t)$ where in each case $T \sim \mathrm{N}(0,1)$.

In the next subsection we consider an example to illustrate these ideas more concretely.

2.2.2 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 (2.15) 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 randomoized $n_1 = 8,872$ 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.

To determine whether this difference 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 (2.18) gives

$$t = \frac{0.0316 - 0.0462}{\sqrt{\frac{0.0316 \times 0.9684}{8872} + \frac{0.0462 \times 0.9538}{8642}}} = -4.9992.$$

The p-value associated with this test is $P(T \leq -4.9992)$ where $T \sim \text{N}(0, 1)$. When calculated this probability is 2.88×10^{-7}, which is effectively 0. In `R` this probability is calculated using the command `pnorm(-4.9992, mean = 0, sd = 1)`.

In order to draw a conclusion, we must compare this value to the significance level $\alpha = 0.05$. Since $2.88 \times 10^{-7} < 0.05$ 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.

2.2.3 Power Analysis and Sample Size Calculations

Here we derive sample size formulae in a manner similar to the development presented in Section 2.1.3 but here we do it in the context of hypothesis tests such as (2.13), (2.14) and (2.15). As in Section 2.1.3 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 2.1.3. 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 (2.16). 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 effect size (in a manner similar to Section 2.2.1). 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:

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

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 $\mathrm{N}(0,1)$ 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

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

that follows a $\mathrm{N}(0,1)$ 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.

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

As in Section 2.1.3 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.

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

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:

$$n_2 = \frac{(z_{\alpha/2} - z_{1-\beta})^2[\frac{\pi_1(1-\pi_1)}{k} + \pi_2(1-\pi_2)]^2}{\delta^2} \tag{2.19}$$

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

$$n = \frac{(z_{\alpha/2} - z_{1-\beta})^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]^2}{\delta^2}. \tag{2.20}$$

If it were a one-sided hypothesis being tested, the only difference between the formulae in that setting relative to equations (2.19) and (2.20) 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.

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

shinyapps.io/SampleSizeCalculator/ can also be used to explore the interdepencies between sample size, significance level, power, and effect size in this setting as well.

2.2.4 When Assumptions are Invalid

In order to test the hypotheses (2.13), (2.14) and (2.15) 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. If either of these assumptions is invalid we can alternatively perform a **chi-squared test of independence** (also known as Pearson's χ^2-test), which does not require them.

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 2.2.2 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).

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

Table 2: A general 2×2 contingency table

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

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.

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

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 coversions 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 $P(T \geq 25.0755) = 5.52 \times 10^{-7}$, where $T \sim \chi^2_{(1)}$, and the p-values associated with hypotheses (2.13), (2.14) and (2.15) 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 2.1.3 and 2.2.3 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 my 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 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 an A/B 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 $N = 100000$ 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.

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 $N = 100000$ times that a Type I error was made. The plot shown in Figure 2 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, and elaborate further later on, 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

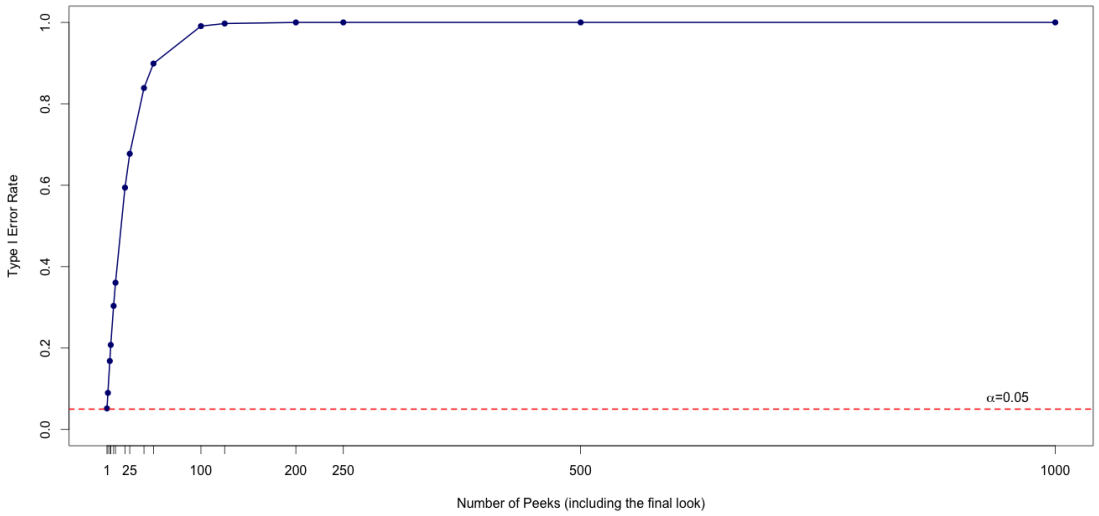

Figure 2: Type I error rate for different levels of peeking.

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. We will discuss sequential testing later on in the context of multi-armed bandit experiments.

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 colors rather than just two? Or what about 10 button colors?

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 response variable (y) is chosen and some metric of interest θ that summarizes the response measurements is also selected. Recall that this metric may be any statistic that can be calculated from observed data, with various user engagement and conversion metrics being commonly used in practice.

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 $j = 1, 2, \ldots, m$, where m is the total number of conditions. In this case the metric of interest is calculated in each condition, giving $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$, 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 $\theta_j > \theta_k$ for all $k \neq j$ (when maximizing the metric is important) or $\theta_j < \theta_k$ for all $k \neq j$ (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} $\mathrm{N}(\mu_j, \sigma^2)$ 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.

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_m \text{ vs. } H_A\text{: } \mu_j \neq \mu_k \text{ for some } k \neq j \tag{3.1}$$

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 (3.1). 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 variables 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

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

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 thet 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. It is practically useful to treat condition m as the 'control' condition, if there is one, since it represents the baseline against which all other conditions are compared.

Thus,

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

As can be seen, H_0 in (3.1) is true if and only if $\beta_1 = \beta_2 = \cdots = \beta_m = 0$. Thus testing (3.1) is equivalent to testing

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_m = 0 \text{ vs. } H_A\text{: } \beta_j \neq 0 \text{ for some } j$$

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) that are associated with 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., $\beta_1 = \beta_2 = \cdots = \beta_m = 0$).

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 $= P(T \geq t)$ where $T \sim F(m-1, N-m)$ and is commonly displayed in regression summaries provided by statistical software. For example, the `summary()` of an `lm()` object in `R` provides the results of this test. 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 color 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

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

where the x's are indicator variables indicating whether a particular value of the response was observed in the jelly fish or color bomb conditions. By using the `lm()` function in `R` we obtain the following output

```
Call:
lm(formula = time ~ factor(booster), data = candy)
Residuals:
     Min       1Q   Median       3Q      Max
-2.84231 -0.69476  0.02617  0.65326  2.76681
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.01281    0.08664  57.859   <2e-16 ***
```

```
factor(booster)2  1.17528    0.11931   9.851   <2e-16 ***
factor(booster)3  4.88279    0.12357  39.515   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.953 on 370 degrees of freedom
Multiple R-squared:  0.8216,Adjusted R-squared:  0.8206
F-statistic: 851.9 on 2 and 370 DF,  p-value: < 2.2e-16
```

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 color 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 $= P(T \geq 851.9)$ 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 $\pi_j = P(Y_{ij} = 1)$ 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 conditionds at all. In order to formally make this decision, the following hypothesis is tested.

$$H_0\text{: } \pi_1 = \pi_2 = \cdots = \pi_m \text{ vs. } H_A\text{: } \pi_j \neq \pi_k \text{ for some } k \neq j \tag{3.2}$$

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 (3.2). As we will see, the χ^2 test from Section 2.2.4 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 2.2.4 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 was more generally used as a test of 'no association' between two categorical variables, and so we could think of 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. However, before 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

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	$N = \sum_{j=1}^{m} n_j$

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 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

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

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

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 $= P(T \geq t)$ 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 test statistic for these data is calculated to be

$$t = \sum_{l=0}^{1} \sum_{j=1}^{5} \frac{(O_{l,j} - E_{l,j})^2}{E_{l,j}} = 129.1761.$$

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

The p-value, then, is $P(T \geq 129.1761) = 5.84 \times 10^{-27}$ 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., (3.1) or (3.2)) 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 practucal 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 overall Type I error rate associated with this family of tests, is much larger than α.

This problem – where a series of independent hypothesis tests lead to an inflated family-wise error rate – is known as the **multiple comparison** or **multiple testing problem**. It can be shown that if a family of k hypothesis tests are performed, each with significance level α, the family-wise error rate (the probability of making a Type I error in any of these

tests) is $1-(1-\alpha)^k$. To see this:

$$\begin{aligned}
P(\text{Type I Error}) &= 1 - P(\text{No Type I Error}) \\
&= 1 - P(\{\text{No Type I Error on test 1}\} \text{ and } \{\text{No Type I Error on test 2}\} \text{ and} \\
&\quad \cdots \text{ and } \{\text{No Type I Error on test } k\}\ ) \\
&= 1 - P(\{\text{No Type I Error on test 1}\} \cap \{\text{No Type I Error on test 2}\} \cap \\
&\quad \cdots \cap \{\text{No Type I Error on test } k\}\ ) \\
&= 1 - \prod_{i=1}^{k} P(\text{No Type I Error on test } i) \\
&= 1 - \prod_{i=1}^{k} (1-\alpha) \\
&= 1 - (1-\alpha)^k
\end{aligned}$$

Figure 3 illustrates the dependence of this family-wise error rate on the number of pairwise comparisons, k. As we can see, as k increases so also does the error rate. In the limit (i.e., as $k \to \infty$) this error rate goes to 1, and so as the the number of pairwise comparisons increases it becomes certain that a Type I error will have been made somewhere. In practice, a common value of k 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 $k = \binom{5}{2} = 10$ 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.

In order to combat this, a variety of statistical approaches have been developed. Here we consider the simplest and most commonly used: the **Bonferroni correction** (Dunnett, 1955). In order to keep the family-wise error rate maintained at α, the Bonferroni correction involves performing each of the k hypothesis tests at a significance level of α/k. Doing so yields a family-wise error rate of $1-(1-\frac{\alpha}{k})^k$ which, for typical values of α in the range $(0, 0.05]$ is approximately equal to α. Figure 4 illustrates this; across a range of values for k the family-wise error is held fixed at the specified level of significance α.

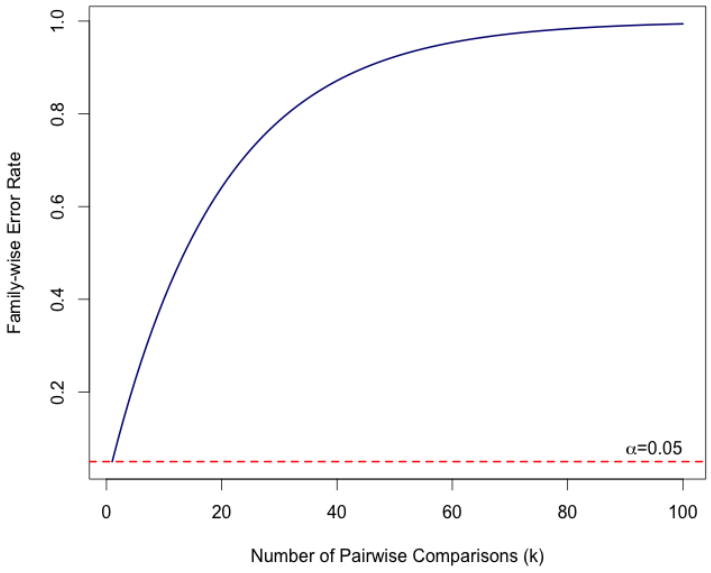

Figure 3: Family-wise error rate versus the number of pairwise comparisons, k.

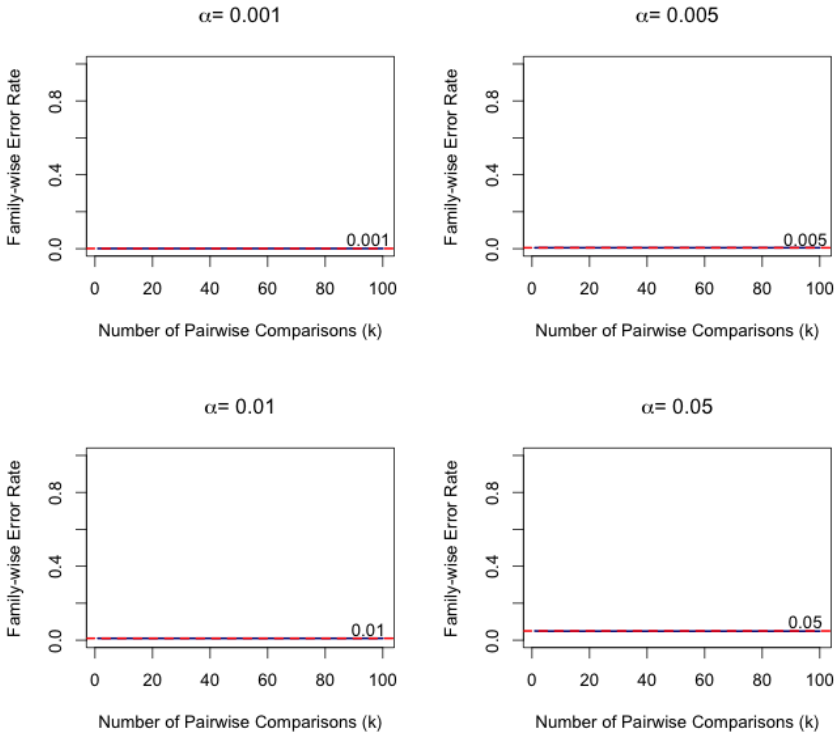

Figure 4: Illustration of the Bonferroni correction for selected values of α.

Furthermore, we can see that as k gets asymptotically large the family-wise error rate no longer approaches 1. Instead

$$\lim_{k\to\infty} 1 - (1 - \frac{\alpha}{k})^k = e^{-\alpha},$$

which for typical values of α in the range $(0, 0.05]$ is approximately equal to α. Figure 5 illustrates this.

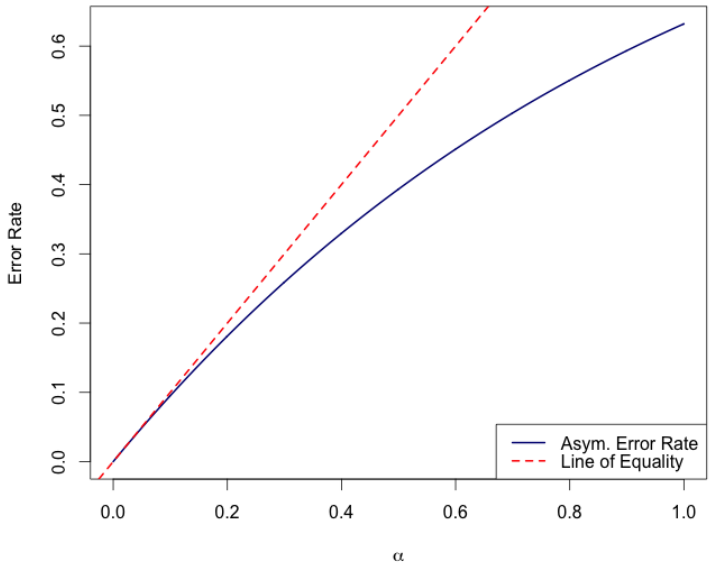

Figure 5: Illustration of the Bonferroni correction for asymptotically large k.

So what does this mean for sample size calculations and power analyses? In Sections 2.1.3 and 2.2.3 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 k 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. In particular, if applying the Bonferroni correction, one should use α/k as the significance level in these calculations, if a family-wise error rate of α is to be maintained.

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, and linear 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., $P(Y = a)$) or lies in some interval (i.e., $P(a < Y < b)$). 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.

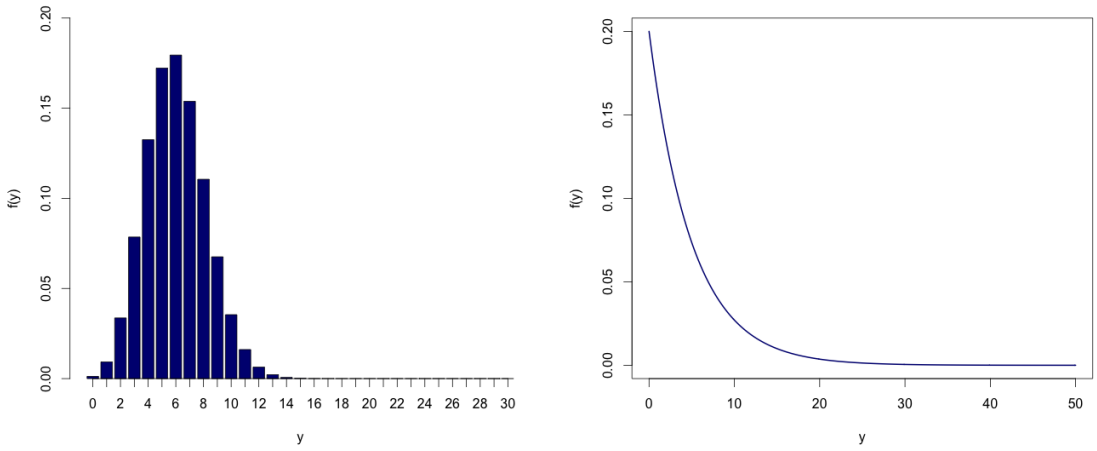

Figure A.1: Left: Distribution Characterizing Survey Respondents; Right: Distribution Characterizing iPhone Battery Lifetimes

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

$$f(y) = P(Y = y)$$

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 ($P(Y = 6)$), or the probability that 10 or more individuals respond to the survey ($P(Y \geq 10)$). By summing the heights of the bars corresponding to all values of y consistent with these events, we find that $P(Y = 6) = 0.1795$ and $P(Y \geq 10) = 0.0611$. 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 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 ($P(Y \leq 10)$) or more than 20 hours ($P(Y > 20)$), 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

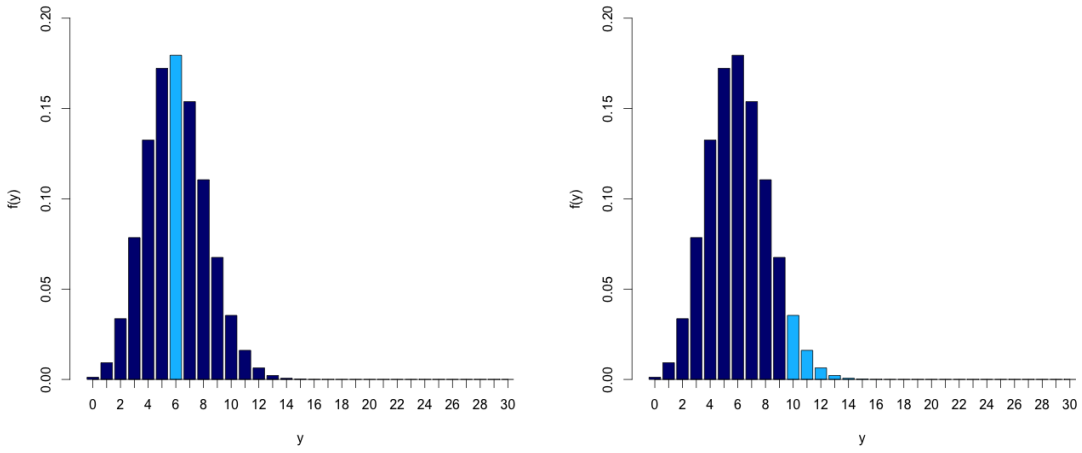

Figure A.2: Left: $P(Y = 6) = f(6)$; Right: $P(Y \geq 10) = \sum_{y=10}^{30} f(y)$

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.

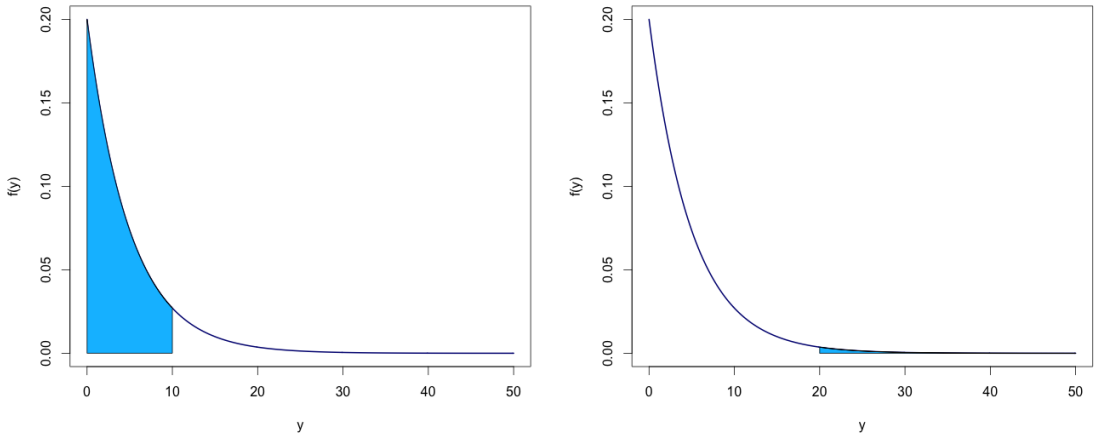

Figure A.3: Left: $P(Y \leq 10) = \int_0^{10} f(y)dy$; Right: $P(Y > 20) = \int_{20}^{\infty} f(y)dy$

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 dis-

cuss 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

$$f(y) = P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

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 a survey respondent was male or female.

The probability mass function for the Bernoulli distribution is given by

$$f(y) = P(Y = y) = \pi^y(1 - \pi)^{1-y}$$

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 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 necessary 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

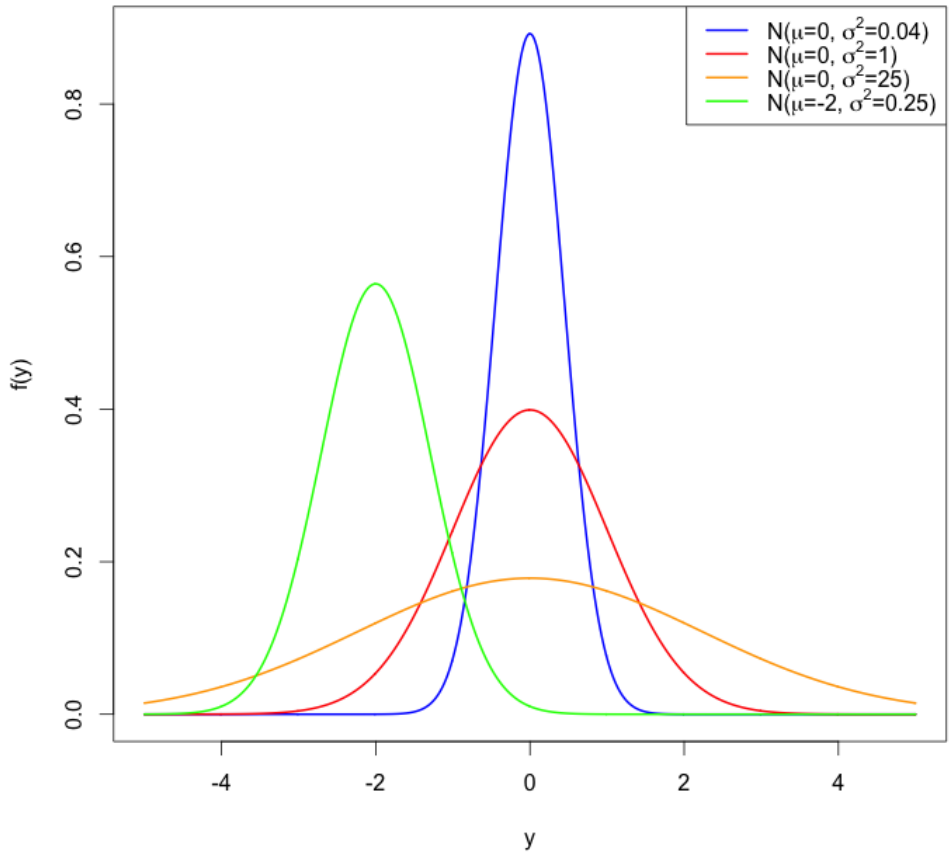

Figure A.4: A variety of normal density curves based on different values of μ and σ

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[2] of William Gosset, the statistician who first derived

[2]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*.

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 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 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

$$f(y) = \frac{y^{\frac{\nu}{2}-1}e^{-y/2}}{2^{\nu/2}\Gamma(\frac{\nu}{2})}$$

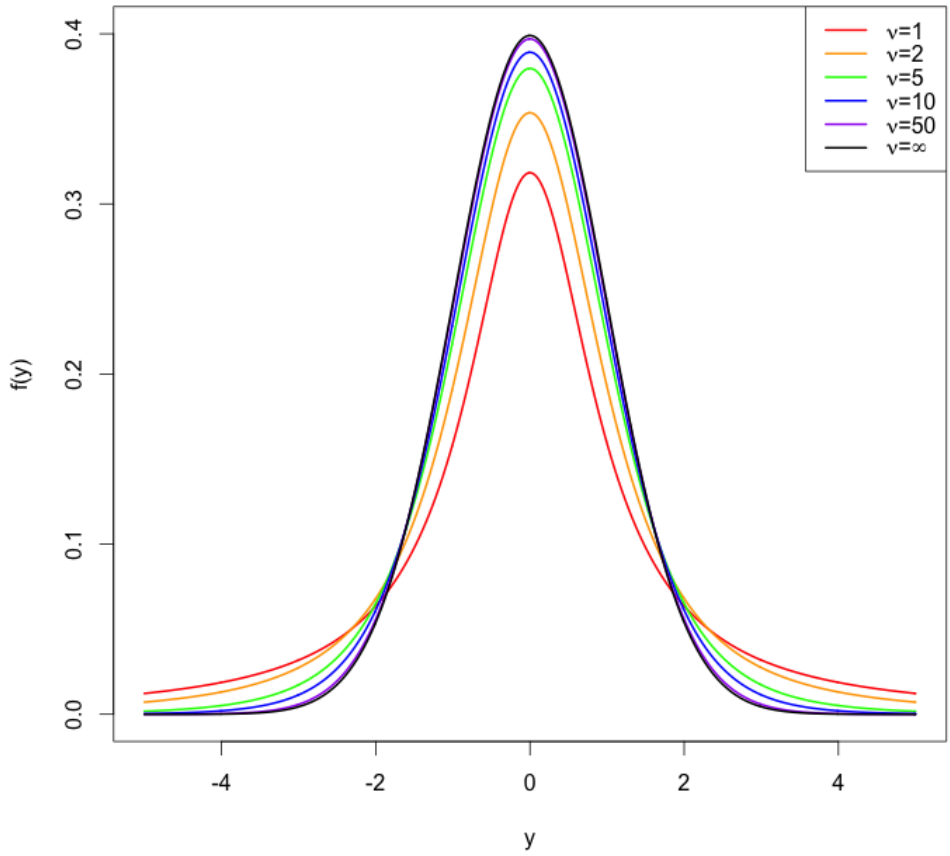

Figure A.5: A variety of t-distribution density curves based on different numbers of degrees of freedom ν

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 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 density function is

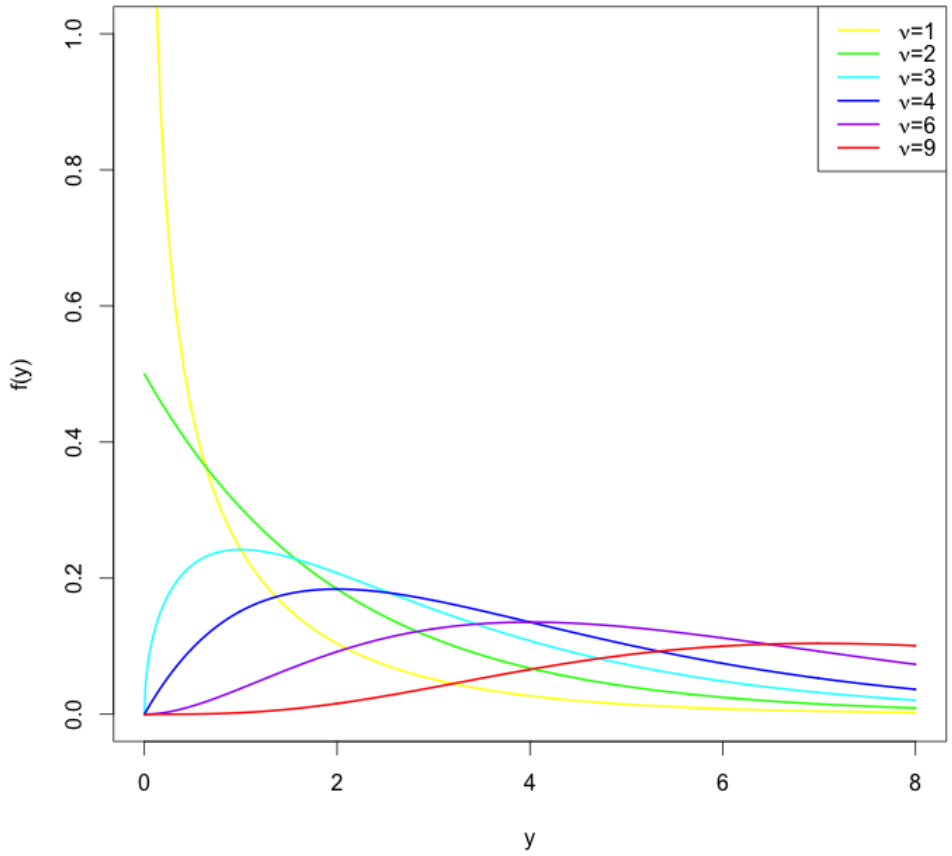

Figure A.6: A variety of χ^2-distribution density curves based on different numbers of degrees of freedom ν

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.

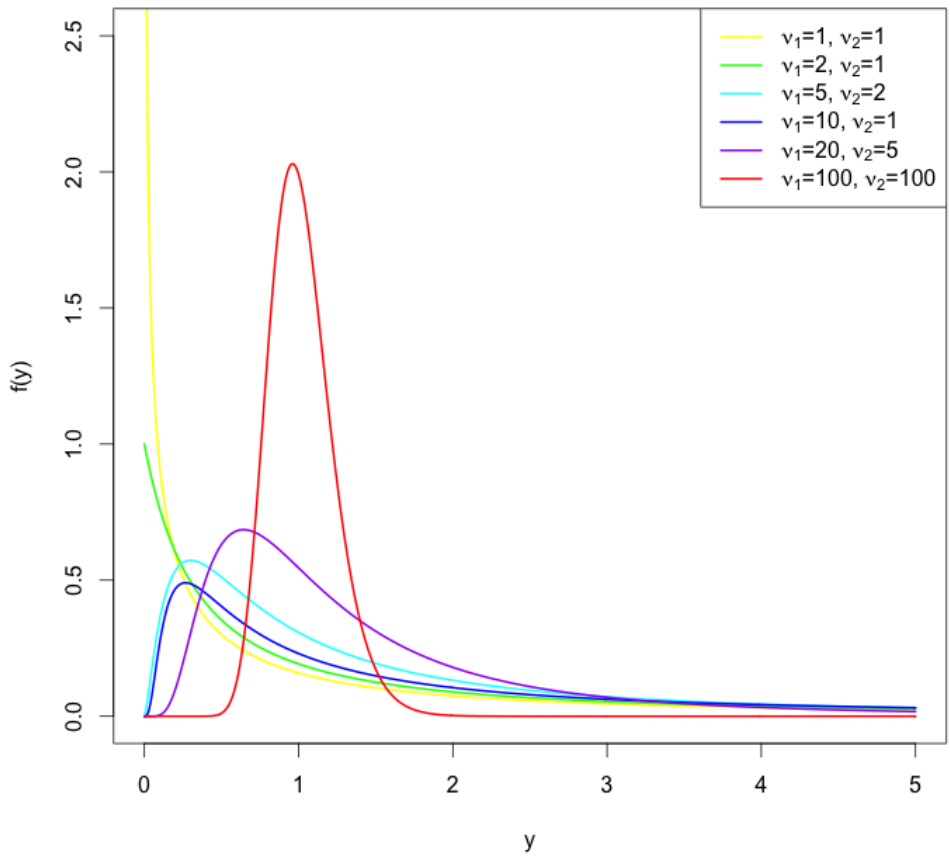

Figure A.7: A variety of F-distribution density curves based on different numbers of degrees of freedom ν_1 and ν_2

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

$$Var[Y] = E[(Y - E[Y])^2]$$

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
$Y \sim F(\nu_1, \nu_2)$	$\frac{\nu_2}{\nu_2-2}$	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$

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 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"[3]), 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.

[3]The notation $\hat{\theta}$ is read "θ-hat".

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

$$H_0\text{: } \theta = \theta_0 \text{ vs. } H_A\text{: } \theta \neq \theta_0$$

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

$$H_0\text{: } \theta \leq \theta_0 \text{ vs. } H_A\text{: } \theta > \theta_0$$

or

$$H_0\text{: } \theta \geq \theta_0 \text{ vs. } H_A\text{: } \theta < \theta_0$$

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.

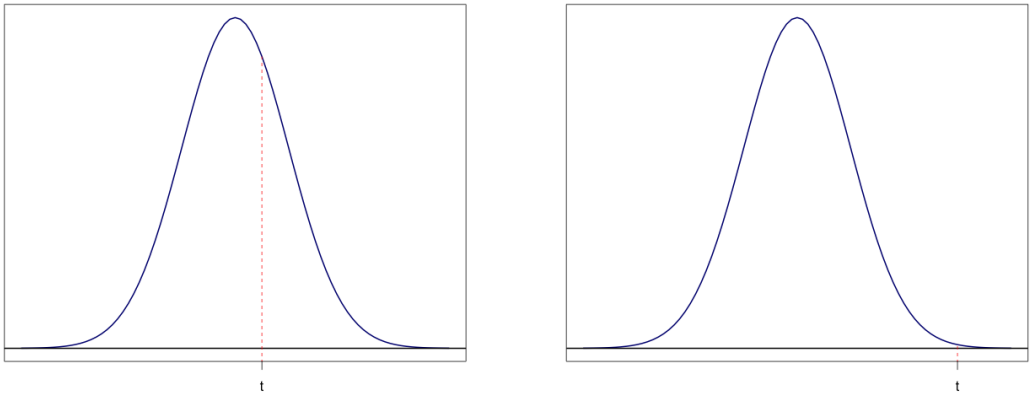

Figure A.8: Left: A non-extreme value of a test statistic; Right: An extreme value of a test statistic

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 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

$$\text{p-value} = P(T \geq |t|) + P(T \leq -|t|)$$

which, if the null distribution is symmetrical, is equivalent to $2P(T > |t|)$. The left panel of Figure A.9 provides a visual depiction of this calculation.

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 render H_0 false. Thus in this case large values of t are considered extreme and the p-value is calculated as

$$\text{p-value} = P(T \geq t).$$

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 render H_0 false. Thus in this case small values of t are considered extreme and the p-value is calculated as

$$\text{p-value} = P(T \leq t).$$

The right panel of Figure A.9 provides a visual depiction of this calculation.

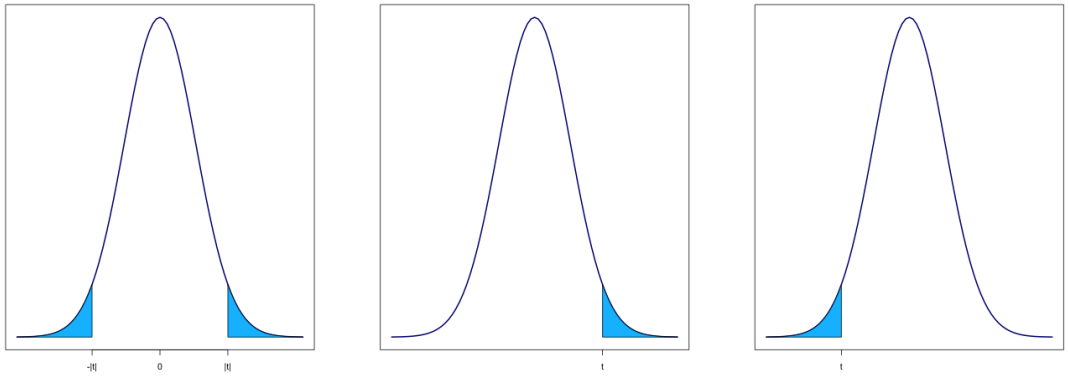

Figure A.9: Illustration of the calculation of p-values in one- and two-sided tests

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

$$H_0\text{: } \textit{the defendant is innocent} \text{ vs. } H_A\text{: } \textit{the defendant is guilty}$$

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 $\alpha = P(\text{Type I Error})$ and we define the power of a test to $1-\beta$ where $\beta = P(\text{Type II Error})$. 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

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

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 $\epsilon \sim \mathrm{N}(0, \sigma^2)$ 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 arounnd 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

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2$$

with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$.

By writing the linear regression model above in matrix form as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

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,
- $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ 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 distance (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).

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.

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.

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

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

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.

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.

Welch, B. L. (1947). The generalization ofstudent's' problem when several different population variances are involved. *Biometrika 34*(1/2), 28–35.