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

A Continuing Education Certificate

at

The Univeristy of San Francisco's Data Institute

September 4 - October 16, 2018

Instructor: Nathaniel T. Stevens

ntstevens@usfca.edu

Preface

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

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

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

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

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

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

Introduction

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

1.1 Notation and Nomenclature

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

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

age time on page (TOP). Here the metric of interest is average TOP and the corresponding response variable is the continuous measurement of time on page for each user. Regardless of the type or goal of the experiment, the response variable is the one we are primarily interested in. Throughout this course we will use the letter y to denote response variables.

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

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

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

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

respect to understanding the relationship between y and one or more x's.

1.2 Experiments versus Observational Studies

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

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

is observed organically. While such an approach provides information about the association between these factors, it does not provide clear information about a causal relationship. When **causal inference** (establishing causal connections between variables) is of interest, it is best if the data arise as a result of an experiment. While methods for establishing causal relationships from observational data do exist (see e.g., propensity score matching (Rosenbaum and Rubin, 1983)), they are much less sound and much more error prone than a carefully designed experiment.

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

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

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

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

1.3 QPDAC: A Strategy for Answering Questions with Data

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

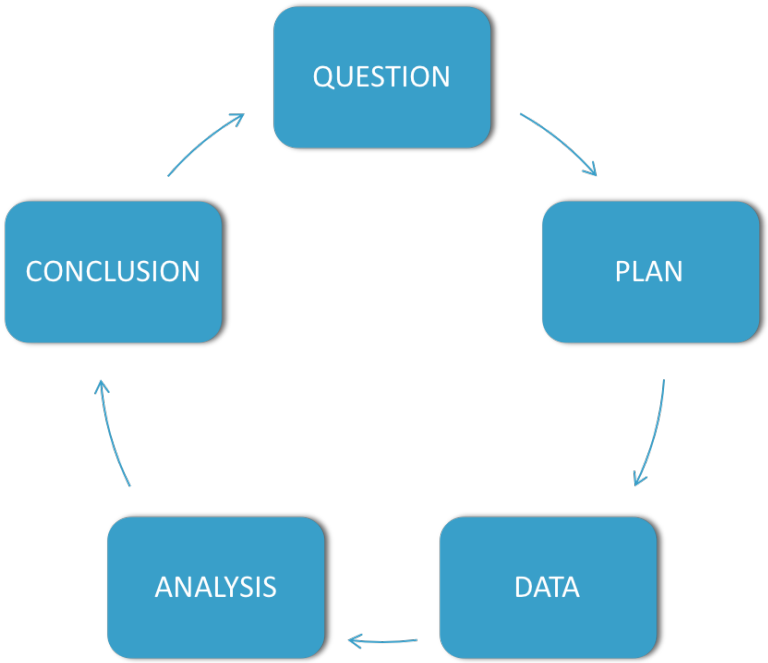

Figure 1: QPDAC Cycle

1.4 Fundamental Principles of Experimental Design

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

Randomization refers both to the manner in which experimental units are selected for inclusion in the experiment and the manner in which they are assigned to experimental

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

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

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

The importance of replication likely seems obvious, but the answer to the question *"how much replication is needed?"* is likely less obvious and is just as important. More directly, this question is equivalent to asking *"how many units should be assigned to each condition?"*. The **sample size** for a given condition, denoted by n, is defined to be the number of units

exposed to that condition. We use power analyses and sample size calculations to determine how many units to include in the study, and hence how many response variable observations are necessary to be sufficiently confident in your results. In the context of online experiments, where website traffic is heavy and predictable, replication is often communicated in terms of time as opposed to number of units. For instance, a common question is *"how long does the experiment need to run for?"*. Intuitively, the more confident one wishes to be in the experiment's results, the larger the sample size needs to be and hence the longer the duration of the experiment. We will formalize these reflections in the chapters to come.

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

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

1.5 Exercise: The Instagram Experiment

We end this chapter by pretending we are data scientists at Instagram that need to design an experiment concerning sponsored ads. While ads serve as a source of revenue for Instagram, they also serve as a source of frustration and annoyance to users. Thus,

we would like to run an experiment to gain insight into the interplay between ad revenue, user engagement and factors such as ad frequency, ad type (photo/video), whether the ad's content is targeted or not, etc. Ultimately the goal is to identify a condition that maximizes ad revenue without simultaneously plummeting user engagement below some minimally acceptable threshold.

How would you design such an experiment?

Appendix

In this Appendix we review some of the statistical prerequisites for the material discussed throughout the notes. In particular we review random variables and probability distributions, point and interval estimation, hypothesis testing, linear regression, and logistic regression.

A.1 Random Variables and Probability Distributions

A.1.1 Random Variables and Probability Functions

A **random variable** $Y : \Omega \to \mathbb{R}$ is a function that assigns real numbers to outcomes of a random process, such as flipping a coin or measuring some quantity of interest. We refer to the possible values a random variable can take on as the **support set**, and we dichotomize random variables based on the type of values they assume. A **discrete** random variable is one whose support set is finite or countably infinite such as $y = 0, 1, 2, \ldots, n$ or $y = 0, 1, 2, \ldots$. We typically use discrete random variables when counting events is of interest. A **continuous** random variable, on the other hand, takes on a continuum of values and so its support set is a subinterval of the real numbers such as $y \geq 0$, $y \in [0, 1]$ or $-\infty < y < \infty$. We typically use continuous random variables when measuring some continuous quantity is of interest. Note that for clarity we denote random variables with upper case letters and the values they take on with lower case letters.

Example 1: Suppose we send an email survey to $n = 30$ individuals and we're interested in the the number of these individuals that respond to the survey. Let Y represent the number of survey responses. In this case the support set is $y = 0, 1, 2, \ldots, 30$, and so Y is a discrete random variable.

Example 2: Interest often lies in measuring lifetimes of people, products, and processes. Suppose that, in particular, we are interested in the lifetime of an iPhone's battery. Let Y represent the lifetime (in hours) of an iPhone battery. In this case the support set is theoretically $y \geq 0$, which is a continuous subinterval of the real numbers, and so Y is a continuous random variable.

Because random variables take on values randomly, interest lies in quantifying the probability that Y assumes a particular value (i.e., $P(Y = a)$) 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.

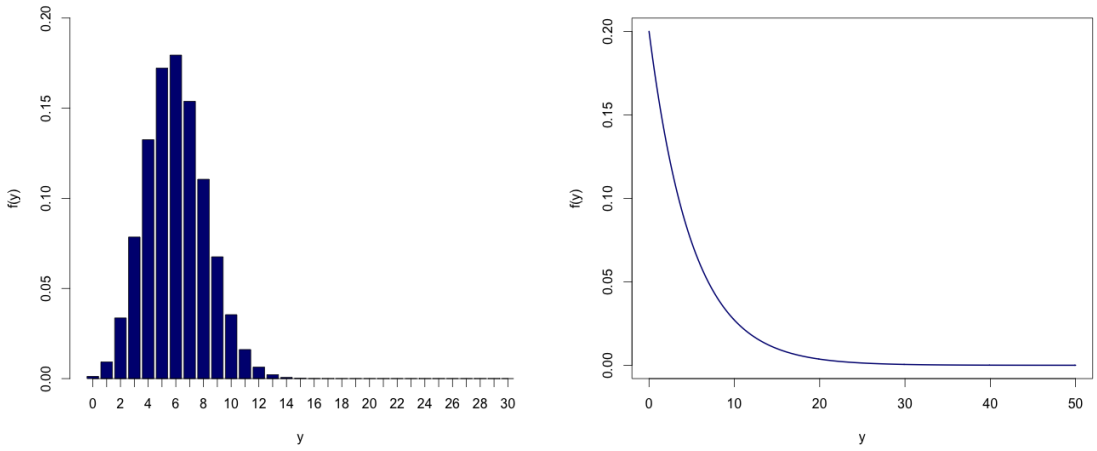

Figure A.1: Left: Distribution Characterizing Survey Respondents; Right: Distribution Characterizing iPhone Battery Lifetimes

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

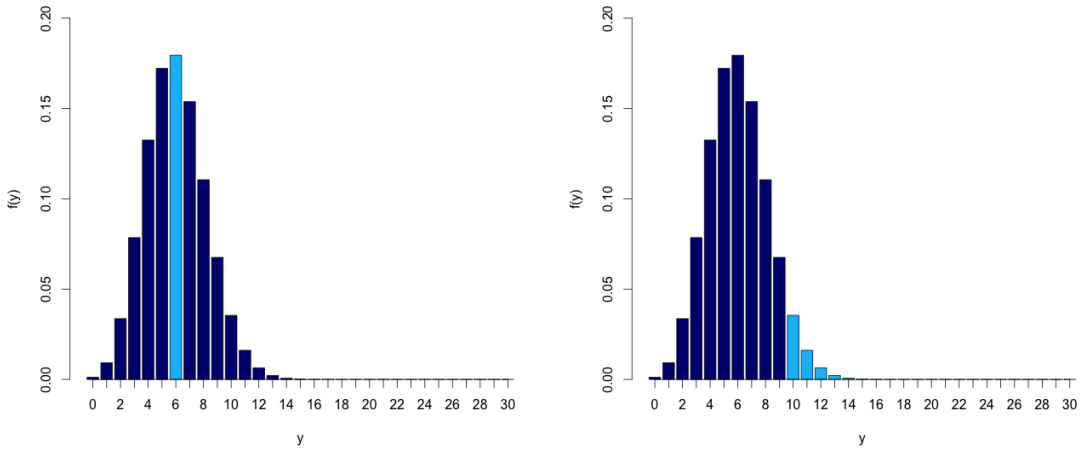

Figure A.2: Left: $P(Y = 6) = f(6)$; Right: $P(Y \geq 10) = \sum_{y=10}^{30} f(y)$

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.

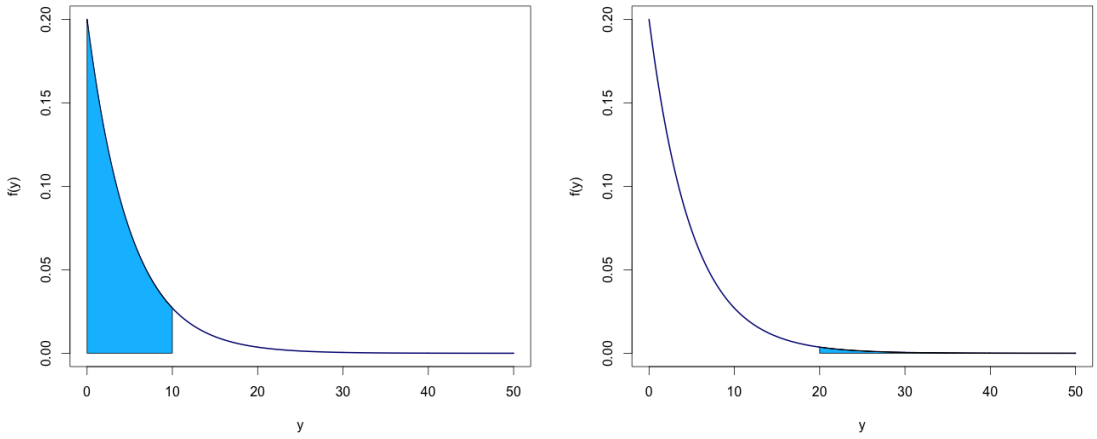

Figure A.3: Left: $P(Y \leq 10) = \int_0^{10} f(y)dy$; Right: $P(Y > 20) = \int_{20}^{\infty} f(y)dy$

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

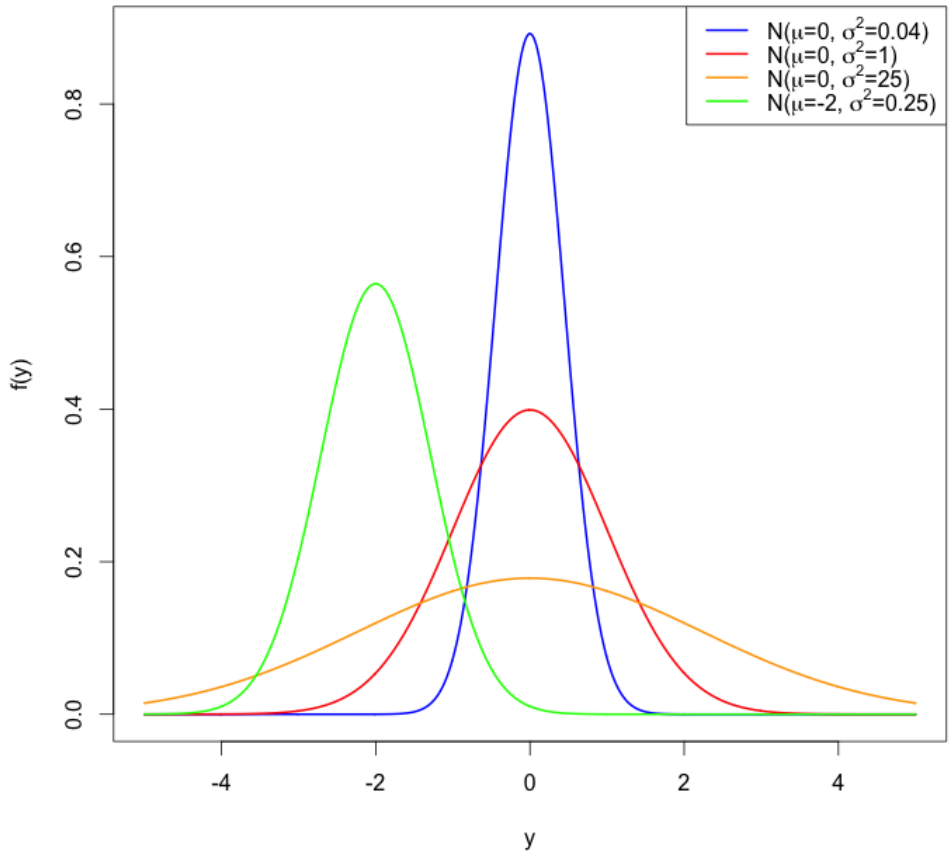

Figure A.4: A variety of normal density curves based on different values of μ and σ

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

[1]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

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

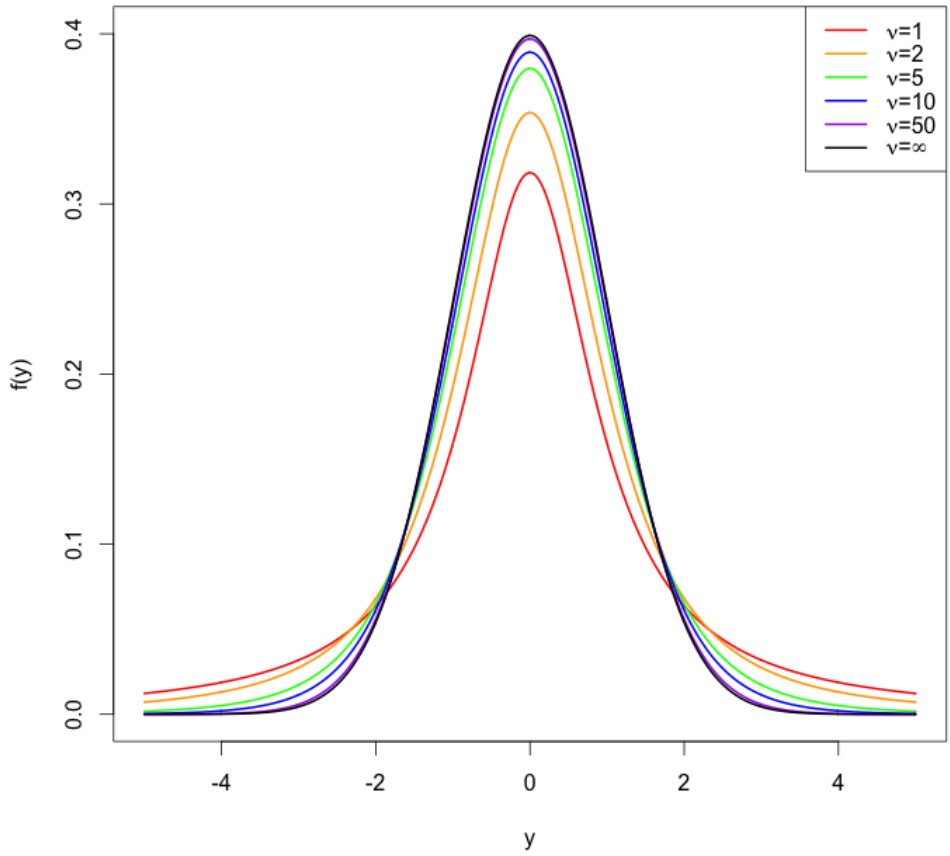

Figure A.5: A variety of t-distribution density curves based on different numbers of degrees of freedom ν

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

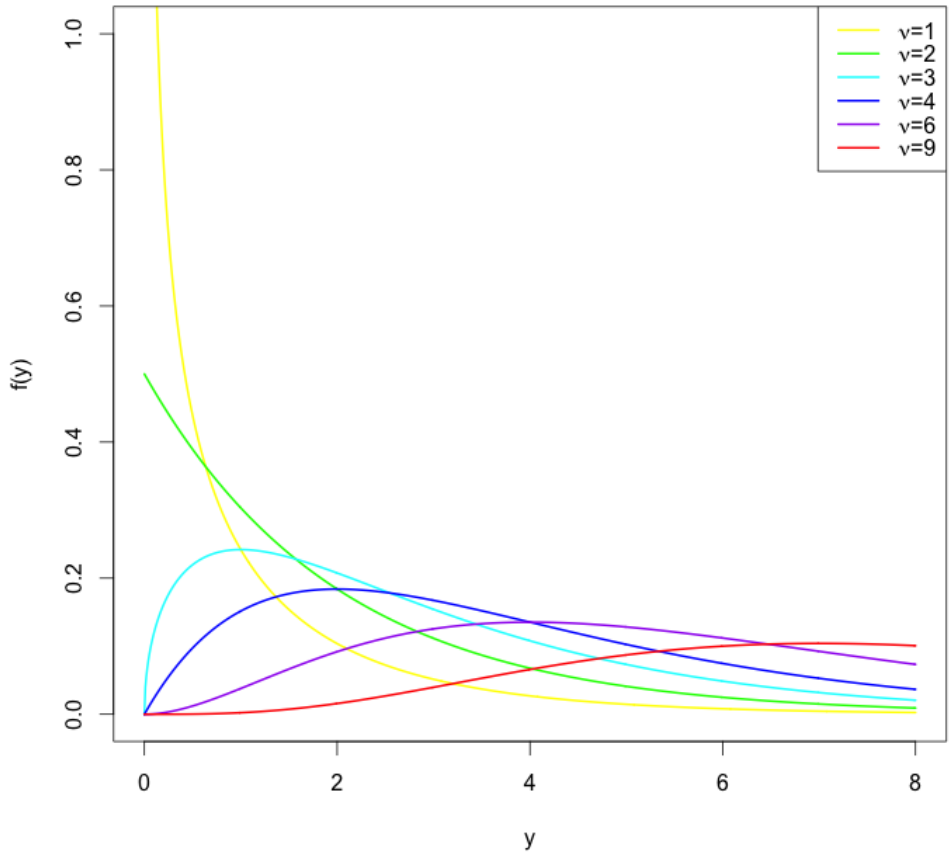

Figure A.6: A variety of χ^2-distribution density curves based on different numbers of degrees of freedom ν

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.

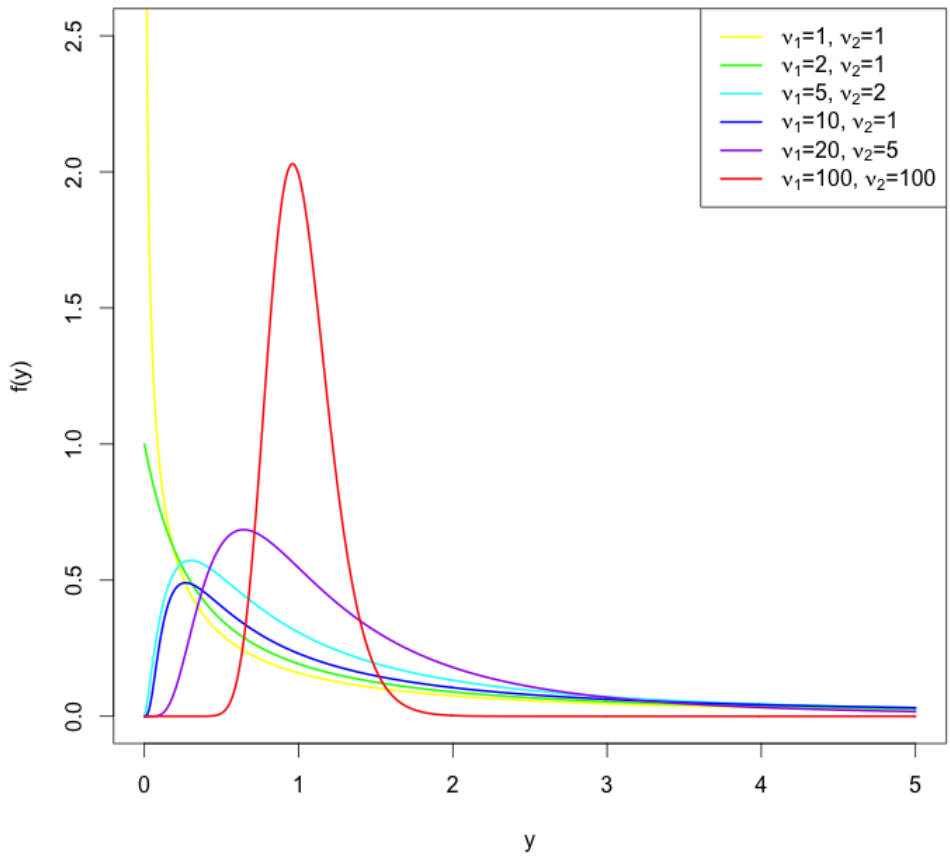

Figure A.7: A variety of F-distribution density curves based on different numbers of degrees of freedom ν_1 and ν_2

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"[2]), 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.

[2]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.

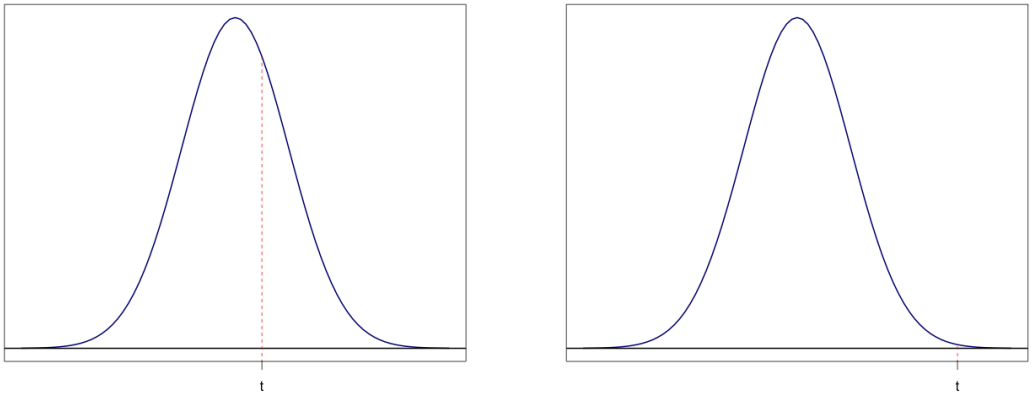

Figure A.8: Left: A non-extreme value of a test statistic; Right: An extreme value of a test statistic

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.

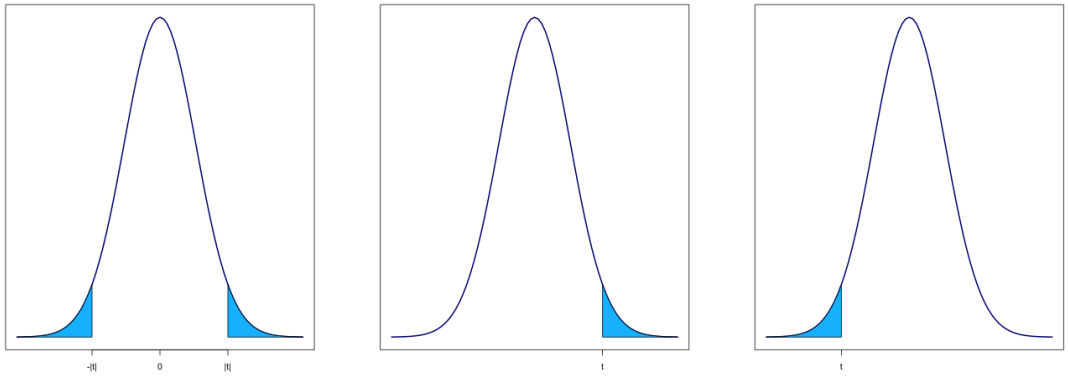

Figure A.9: Illustration of the calculation of p-values in one- and two-sided tests

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.

References

Bain, L. J. and M. Engelhardt (1992). *Introduction to probability and mathematical statistics* (2nd ed.). Brooks/Cole.

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

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

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

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

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