# STAT 1793: Course Notes

**Introduction to Probability and Statistics I**

Dylan Spicker

2023-12-30

# Table of contents

# Preface

TODO: Write the front matter.

# Part I

# Part 1: Probability

# 1 Introduction

## 1.1 What is this Course?

TODO: Write this.

## 1.2 What is Probability?

At its core, statistics is the study of uncertainty. Uncertainty permeates the world around us, and in order to make sense of the world, we need to make sense of uncertainty. The language of uncertainty is **probability**. Probability is a concept which we all likely have some intuitive sense of: if there was a 90% probability of rain today, you likely considered grabbing an umbrella; you are not likely to wager your life savings on a game that has only a 1% probability of paying out; etc. We have a sense that probability provides a measure of *likelihood*. Defining probability is a non-trivial task, and there have been many attempts to formalize it throughout history. While we will spend a good deal of time formalizing notions of probability in this course, let us pause and emphasize the intuitive basis you are starting with.

Suppose that two friends, Charles and Sadie, meet for coffee once a week. During their meetings they have wide-ranging, deep, philosophical conversations spanning the important issues from "do we all really see green as the same colour" to "why is it that 'q' comes as early in the alphabet as it does? it deserves to be with UVWXYZ." Beyond making progress on the most pressing issues of our time, Charles and Sadie each adore probability. As a result, at the end of each of their meetings, they play a game to decide who will pay: they flip a coin three times. If two or more heads come up, Charles pays, otherwise, Sadie pays.

We can think about the strategy that they are using here and intuitively feel that this is going to be "fair". With two or more heads, Charles pays; with two or more tails, Sadie pays; there always has to be either two or more heads *or* two or more tails, and each is equally likely to come up [footnote: TODO: recent literature on fair coins]. The outcome of their game is uncertain before it begins, but we know that in the long-run neither of the friends is going to be disadvantaged relative to the other. We can say that the probability that either of them pays is equal, it's 50-50, everything is balanced.

One day, in the middle of their game, Charles gets a very important phone call (someone has just pointed out the irony in the fact that there is no synonym for synonym [footnote: TODO: is this actually ture?]) and he leaves abruptly after the first coin has been tossed. The first coin toss showed a heads. Sadie, recognizing the gravity of the phone call, pays for the both of them, but she also recognizes that Charles was well on the way to having to pay.

**?** Example: Basic Probability Enumeration

What is the probability that Sadie would have had to pay in the aforementioned scenario? That is, assuming that the first coin shows a head, what is the probability that at least two heads are shown on the first three coin tosses?

**Solution**

Sadie figures that any coin toss is equally likely to show heads or tails. because the first coin showed heads, then there are four possible sequences that could have shown up: $H, H, H$, $H, H, T$, $H, T, H$, $H, T, T$.
In three of these situations ($H, H, H$, $H, H, T$, $H, T, H$) there are two heads and so Charles would have to pay. In one of them there are two tails, and so Sadie would have to pay. As a result, Charles would have to pay with probability 0.25 and Sadie with probability 0.75.

We can see from this example that Sadie should likely have not paid. Only one times out of four would she have had to, given the first head. However, we can not be certain that, had all three tosses been observed, Sadie would not have paid. It is possible that we would have observed two tails, making her responsible for the bill. This possibility happens one time out of four, which is more likely than the probability of rolling a 4 on a six-sided die. Of course, 4s are rolled on six-sided dice quite frequently (about one out of every six rolls), and so it is not all together unreasonable for her to have paid.

This seemingly simple concept is the core of probability. Probability serves as a method for quantifying uncertainty. It allows us to make numeric statements regarding the set of outcomes that we can observe, quantifying the freqneucy with which we expect to observe those outcomes. However, probability does not *remove* the uncertainty. We still need to flip the coin or roll the die to understand what will happen. All probability has given us is a set of tools to exactly quantify this uncertainty. It turns out that these tools are critical for decision making in the face of the ever present uncertainty around us.

## 1.3 How to Interpret Probabilities (like a frequentist)

We indicated that, intuitively, probability is a measure of the frequency with which a particular outcome occurs. This intuition can be codified exactly with the **frequentist interpretation**

of probability. According to the frequentist interpretation (or frequentism, as it is often called), probabilities correspond to the proportion of time that whatever the event of interest is (be that flipping a heads, rolling a four, or observing rain on a given day) actually occcurs, in the long run. For a frequentist, you imagine yourself performing the experiment you are running over, and over, and over, and over, and over again. Each time you record "did the event happen?" and you count up those occurences. As you do this more and more and more, wherever that proportion lands corresponds to the probability.

TODO: Insert R Diagram for This (Coin Tosses)

To formalize this in terms of mathematics, we have to define several important terms. An **experiment** is any action whose outcome cannot be known with ceertainty, before it is performed. An **event** is a specified result that may or may not occur when an experiment is performed.

Suppose that an experiment is able to be performed as many times as one likes, limited only by your boredom. If you take $k_N$ to represent the number of times that the event of interest occurs when you perform the experiment $N$ times, then a frequentist would define the probability of the event as

$$\text{probability} = \lim_{N \to \infty} \frac{k_N}{N}.$$

In practice this means that, in order to interpret probabilities, we think about repeating an experiment many, many times over. As we do that, we observe the proportion of times that any particular outcome occurs, and take that to be the defining relation for probabilities. The reason that we say the probability of fllipping a heads is 0.5 is because if we were to sit around and flip a coin (experiment) over, and over, and over again, then in the long-run we would observea head (event) in 0.5 of cases.

Many situations in the real world are not able to be run over and over again. Think about, for instance, the probability that a particular candidate wins in a particualr election. There is uncertainty there, of course, but the election can only be run once. What then? There are several ways through these types of events.

First, we can rely on the **power of imagination**. There is nothing stopping us from envisioning the hypothetical possibility of running the election over, and over, and over, and over again. If we step outside of reality for a moment, we can ask "if we could play the day of the election many, many, many times, what proportion of those days would end with the candidate being elected?" If we say that the candidate has a 75% chance of being elected, then we mean that in 0.75 of those imaginaed worlds, the candidate wins. It is crucial to stress that in our imagination here, we need to be thinking about the **exact same day** over and over again. We cannot imagine a different path leading to the election, different speeches being given in advance, or different opposition candidates. If we start from the same place, and play it out many times over, what happens in each of those worlds?

The repeated imagination is not for everyone. In light of these types of shortcomings, alternative proposals to the definition of probability had been made. Most popularly, the **Bayesian interpretation** has become prominent in recent-ish years. To Bayesians, probability is a measure of subjective belief. To say that there is a 50% chance of a coin coming up is a statement about one's knowledge in the world: typically, coins show up heads half the time, so that's our belief about heads. The Bayesian view, though built around subjective confidence in the state of the world, can be formalized mathematically as well. A Bayesian considers the prior evidence that they have about the world, whatever evidence has been collected, and then they update their subject beliefs balancing these two sources of information.

> **i** Bayesian Probabilities and Belief Updating
>
> Suppose that a Bayesian is flipping a coin. Before any flips have been made the Bayesian, understandably believes that the coin will come up heads 50% of the time. However, when the coin starts to be flipped, the observations are a string of tails in a row.
> After having flipped the coin five times, the individual has observed five tails. Of course, it is totally possible to flip a fair coin five times and see five tails, but there is a level of skepticism growing.
> After 10 flips, the Bayesian has still not seen a head. At this point, the subjective belief is that there is likely something unfair about this coin: even though the experiment started with a baseline assumption that it was fair, the Bayesian no longer believes that the enxt flip will be a head.
> As this goes on, you can imagine the Bayesian continuing to update their view of the world. The probability is an evolving concept, capturing what was thought, and what has been observed.

For the election example, the Bayesian interpretation is easier to write out: based on everything that has been observed, and any prior beliefs about the viability of the candidates, the probability that a candidate wins the election is the subjective likelihood assigned to that outcome. Of course, if we disagree about prior beliefs, or have experienced different pieces of information, then we may disagree on the probability: that is okay.

In this course, we will focus on the frequentist interpretation. In part this is because frequentist probability is an easier introduction to the concepts that are necessary for taming uncertainty. In part this is because frequentist interpretations have been shown to give individuals a better grasp on understanding probabilities [CITE?]. However, it is important to know and recognize that there is a world beyond frequentist probability and statistics, one which can be very powerful once it is unlocked. (More on that in later years, if you so desire!)

If probability measures the long-term frequency of times that a particular event occurs, then how can we go about computing probabilities? Do we require to perform an experiment over and over again? Fortunately, the answer is no. The tools of probability we will cover allow us to make concrete statements about the probabilities of events without the need for repeated experimental runs. However, before dismissing the idea of repeatedly running an experiment

at face value it is worth considering a tool we have at our disposal that renders this more possible than it has ever been: computers.

## 1.4 R Programming for Probability and Statistics

Throughout this course we will make use of a programming language for statistical computing, called `R`. While clasically introductory statistics involved heavy computation of particular quantities, by hand, the use of a programming language (like R) frees us from the tedium of calculation, allowing for a far deeper focus on understanding, explanation, decision-making, and complex problem solving. While you will not be expected to write large R programs on your own, you will be expected to read simple scripts, make basic modifications, and to run code that is given to you.

Throughout these course notes, where relevant, R code will be provided to demonstrate the ideas being discussed. It may be useful to have R open alongside the notes to ensure that you can get the same results that are preinted throughout the notes. In this section we will cover some of the very basics of using R, and reading R code. If you are interested, there are plenty of resources to becoming a more proficient R programmer: this is a skill that will benefit you not only in this course, but in many courses to come, and far beyond your university training. If you have any programming in your background, R is a fairly simple language to learn; if not, R can be quite beginner friendly.

TODO: Write some basic intro to R stuff.

Recall that the motivation for the discussion of R was the frequentist interpretation of probability. One task that computers are very effective at is repeatedly performing some action. As a result, we can use computers to mimick the idea of repeatedly performing some experiment. Consider the simple case of flipping a coin over and over again.

We can use `sample(options, size)` as a function to select `size` realizations from the set contained in `options`. Thus, if we take T we can view this as flipping a coin one time. If we use the loop structure we talked before, then we can simulate the experience of repeatedly flipping a coin. Consider the following R code.

```
set.seed(3141592)                                    ①

number_of_runs <- 1000                               ②
tosses <- c()                                        ③

for(idx in 1:number_of_runs){                        ④
    toss <- sample(c("H","T"), 1)                    ⑤
    if(toss == "H"){                                 ⑥
        tosses <- c(tosses, 1)
```

```
    } else {
        tosses <- c(tosses, 0)
    }
}

mean(tosses)                                                    ⑦
```

① A seed ensures that the random numbers generated by the program are always the same. This helps to be able to reproduce our work.
② This is how many times we want to repeat the experiment.
③ This is where we are going to store the results of our tosses. It creates an empty list for us.
④ Here we are going to loop over the experiments, one for each run.
⑤ This is our coin toss. We are going to sample 1 from either 'H' or 'T'
⑥ If the coin toss is heads, then we add a 1 to the list. Otherwise, we add a zero to the list.
⑦ Return the mean of all of the tosses.

```
[1] 0.522
```

It is worth adjusting some of the parameters within the simulation, and seeing what happens. What if you ran the experiment only 5 times? Ten thousand times? What if instead of counting the number of heads, we wanted to count the number of tails? What if we wanted to count the number of times that a six-sided die rolled a 4? All of these settings can be investigated with simple modifications to the provided script.

# 2 Core Concepts of Probability

## 2.1 Basic Probability Models

While we gave the mathematical formulation for the frequentist interpretation of probability, we will typically require a more detailed mathematical model to work with probabilities. We want a description, framed in terms of mathematical objects, which will allow us to work out the probabilities of interest. In general, to form such a probability model we need both a list of all possible outcomes that the experiment can produce, as well as the probabilities of these outcomes.

We call the list of outcomes that can occur from an experiment the **sample space** of the experiment. The sample space is denoted $\mathcal{S}$, and is defined as the set of all possible outcomes from the experiment. For instance, if the experiment is flipping a coin we have $\mathcal{S} = \{\mathrm{H}, \mathrm{T}\}$. If the experiment is rolling a six-sided die then $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.

Example: Enumerating Sample Spaces

Write down the complete sample space, $\mathcal{S}$ for the game that Sadie and Charles play, based on flipping and observing a coin three times in sequence.

**Solution**

For Sadie and Charles their experiment involves tossing a coin three times in sequence. As a result each outcome is a three-dimensional list of values, given for instance by $(\mathrm{H}, \mathrm{H}, \mathrm{H})$. As a result, we can write down the full sample space as

$$\mathcal{S} = \{(\mathrm{H}, \mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{H}, \mathrm{T}), (\mathrm{H}, \mathrm{T}, \mathrm{H}), (\mathrm{H}, \mathrm{T}, \mathrm{T}), (\mathrm{T}, \mathrm{H}, \mathrm{H}), (\mathrm{T}, \mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{T}, \mathrm{T})\}.$$

With the sample space formally defined, we can introduce the formal concept of an event. Formally, an **event** is any collection of outcomes from an experiment. Mathematically, that means that an event is a subset of the sample space. Take for instance the experiment of a single coin. In this case, $\{\mathrm{H}\}$, $\{\mathrm{T}\}$, and $\{\mathrm{H}, \mathrm{T}\}$ are examples of possible events. [TODO: clarify $E_j$.] Here, $E_1$ corresponds to the event that a head is observed, $E_2$ corresponds to the event that a tail is observed, and $E_3$ corresponds to the event that either a tails or a heads was observed. Note that for each event we have $E_1 \subset \mathcal{S}$, $E_2 \subset \mathcal{S}$, and $E_3 \subseteq \mathcal{S}$.

While $E_1$ and $E_2$ each correspond to a simple outcome from the sample space, $E_3$ corresponds to a combined event. We call direct outcomes **simple events** and more complex outcomes like $E_3$ **compound events**. A simple event is an event that occurs in only one way: it is a direct outcome of the experiment. A compound event is an event that can occur in multiple, different ways, with multiple outcomes satisfying the event.

If we consider rolling a six-sided die, then an example of a simple event is that a four shows up, denoted $\{4\}$. A compound event could be that an even number is rolled, $\{2, 4, 6\}$, or that a number greater than or equal to four is rolled, $\{4, 5, 6\}$.

**?** Example: Specifying Events

What are the event(s) of interest for Charles and Sadie in the game that they are playing?

**Solution**

The event of interest in the game played by Sadie and Charles is either that at least two heads are rolled, or that at least two tails are rolled. These are both compound events as there are multiple possible outcomes that correspond to each event. For instance, we can take the event that at least two heads are rolled to be given by $\{(\text{H}, \text{H}, \text{H}), (\text{H}, \text{H}, \text{T}), (\text{T}, \text{H}, \text{H}), (\text{H}, \text{T}, \text{H})\}$.

We say that an event "occurs" if any of the outcomes comprising the event occur. As a result we can have more than one event occurring as the result of an experimental run. Consider the events "an even number was rolled" and "a number greater than or equal to four was rolled." If a four or a six are rolled, both of these events happen simultaneously.

Above, with $E_3$, we saw that $\mathcal{S}$ can be thought of as an event since $\mathcal{S} \subseteq \mathcal{S}$. This is the event that some (any) outcome is observed, which is certain to happen, and so we know it happens with probability 1. There is one further "special" event which is important to consider the so-called *null event*. Denote $\emptyset$, the null event is an event that corresponds to "nothing in the sample space". We know that everytime an experiment is run something in the sample space occurs, and so the null event is assigned probability zero.

Ultimately, we think of all events as being sets. These sets are subsets of the sample space, and can contain single or multiple outcomes. Every quantity that we are interested in can be expressed as some set of outcomes of interest. In building up these sets it is common to construct through the use of "and", "or", and "not" language.

Consider the example of drawing cards from a standard 52-card deck. In such a deck there are 13 card ranks, and four card suits, with one of each combination present. If we draw a single card we can think of the outcomes of the experiment as being any of the 52 possible combinations of rank and suit. However, we are often interested in an event such as "the card is red", which is the same as saying "the card is a heart **or** the card is a diamond." If we are interested in the event that the ace of spades was drawn, this can be expressed by saying that

"the card was a spade **and** the card was an ace." Perhaps we want to know whether the card was an ace through ten, this is the same as saying "the card is not a Jack, Queen, **or** King."

As you beginto pay attention to the linguistic representation of events that we use, you will notice more and more the use of these words to form compound events in particular. As a result, we give each of them a mathematical operation which allow us to quickly and compactly express these quantities in notation.

We incapsulate **or** through the use of an operation known as the union. The union between two sets, $A$ and $B$, is denoted mathematically as $A \cup B$, and is read $A$ union $B$. In words, the union of $A$ and $B$ is the set which contains all elements that are in either $A$ or $B$ (or both $A$ and $B$). When we wish to take the union of many sets, we write this as

$$\bigcup_{i=1}^{n} A_i.$$

Alongside the union, we introduce the intersection, which captures "and". The intersection of two sets, $A$, and $B$, is the set that contains all elements that belong to both $A$ and $B$. We write this as $A \cap B$, and say "$A$ intersect $B$." When we wish to take the union of many sets, we write this as

$$\bigcap_{i=1}^{n} A_i.$$

Finally, we can consider making formal the concept of "not." In particular, we take the **complement** of a set to be the set of all elements which occur in the universe but are not in the given set. We write this as $A^C$ and say "$A$ complement." When dealing with a sample space, $\mathcal{S}$, the complement of $A$ is the set of all elements in the sample space that are not in $A$.

TODO: example block with card hands.

These concepts allow us to more compactly express sets of interest, and in particular, will be quite useful when it comes to assigning probability. To practice, considering rolling a 6-sided die, and take the outcomes $A$ to be that a 6 is rolled, $B$ to be the outcome that the roll was at least 5, $C$ to be the outcome that the roll was less than a 4, and $D$ to be the outcome that the roll was odd.

If we consider $D^C$ this is the event that the roll was even; $A \cup C$ is the event that a 6 was rolled or that a number less than 44$ was rolled, which is to say anything other than a 4 or a 5; if we take $A \cup B$ then this will be the same as $B$, and $A \cap B$ will be $A$. If we take the event $A \cap C$, notice that no outcomes satisfy both conditions, and so $A \cap C = \emptyset$. We can also join together multiple operations. $D^C \cap C$ gives us even numbers than less than 4, which is to say the outcome 2. $(A \cap B)^C$ would represent the event that a number less than 6 is rolled.

Working with these basic set operations should become second nature. There are often very many ways of expressing the same event of interest using these different operations, and finding

the most useful method of representing a particular event can often be the key to solving challenging probability questions. The first step in making sure that these tools are available to you is in ensuring that the basic operations are fully understood, and this comes via practice. Remember, unions represent "ors", intersections represent "ands", and complements represent "nots".

The sample space is partitioned into outcomes, and the oujtcomes can be grouped together into events. These events are sets and can be manipulated via basic set operations. Sometimes it is convenient to represent this process graphically through the use of Venn diagrams. In a Venn diagram, the sample space is represented by a rectangle with the possible outcomes placed inside, and events are drawn inside of this as circles containing the relevant outcomes.

On the Venn diagram then, the overlap between circles represents their intersection, the combined area of two (or more) circles represents their union, and everything outside of a given circle represents the complement of a set. This can be a fairly useful method for representing sample spaces, and for visualizing the basic set operations that we use to manipulate events inside the sample spaces. A word of caution: Venn diagrams are useful tools, but they are not suitable as proof tools. It is possible to convince yourself of false truths if th e wrong diagrams are used, and as a result, Venn diagrams should be thought of as aids to understanding, rather than asa rigorous tool in and of themselves.

TODO: Add Example with Graphical Venn Diagrams

Sample spaces, events, and the manipulation of these quantities forms a critical component of understanding probability models. In particular, they describe the complete set of occurences in a statistical experiment that we could be interested in assigning proobability values to. To finish formalizing a probability model, however, we also need some rule for assigning probability values.

There are a plethora of ways to assign probabilities to diffferent events. At the most basic of levels any rule that maps from the space of possible events to numbers between 0 and 1 are possible candidates for probability assignment. That is, probability assignmnet at its core is simply a set of rules which says "for this event, assign this probability."

TODO: Add example of coin toss space.

Not every assignment of probability values is going to be valid. Suppose, for instance, that we have a six-sided die, each side labelled with a number from one to six. If I told you that there was a probability of 0.5 that it comes up 1, 0.5 that it comes up 2, 0.5 that it comes up 3, 0.5 that it comes up 4, 0.5 that it comes up 5 and 0.5 that it comes up 6, you would probably call me a liar (or else conclude that I was mistaken and likely should not be teaching probability). If, as we have previously seen, probabilities represent the long-run proportion of time that a particular event is observed, we cannot have 6 different outcomes each occuring in half of all caes.

beyond the requirements that we can impose on what consitutes a "valid" probability function, we have another concern: scalability. It is perfectly acceptable to indicate that in an experiment

with 3 outcomes, the first has a probability of 0.25, the second of 0.3, and the third of 0.45. What if the experiment has 100 possible outcomes? Or 1000?It quickly becomes apparent that enumerating the probabilities of each event in the sample space is an efficient way of assigning probabilities in practice.

A core focus of our study of probability will be on finding techniques that allow us to efficiently encode probability information into manageable objects. Once we have done this, we will be in a position where we can manipulate these (comparatively) simple mathematical quantities in order to make statements and conclusions about any of the events of interest, even if they have never been explicitly outlined as having an assigned probability.

While we will consider myriad methods for accomplishing these goals throughout our study of probability, we begin with a very useful model which simplifies probability assignment, without any added complexity, and creates a solid foundation for us to explore the properties of probability models. We start by considering **equally likely outcomes.** As the name suggests, the probability model considering equally likely outcomes assigns an equal probability to every possible outcome of the experiment. This is a probability model that we are already distinctly familiar with: flipping a coin, rolling a die, or drawing a card are all examples of experiments which rely on theequally likely outcomes framework.

TODO: Write aside about the use of urn models.

If we have an experiment with a sample space $\mathcal{S}$ which has $|\mathcal{S}| = k$ total elements, then each element of the sample space occurs with probability $\frac{1}{k}$. In the case of the coin toss example, $\mathcal{S} = \{\mathrm{H}, \mathrm{T}\}$, and so $k = 2$ and each outcome occurs with probability $\frac{1}{2}$. in the case of drawing a card at random, there are 52 different outcomes, and so $k = 52$, and the probability of drawing any particular card is $\frac{1}{52}$.

It is critically important to recognize that the equal probability model assigns equal likelihood to the possible outcomes of an experiment, not the possible events of interest. It will not be the case in general that all events have the same probability. To make this concrete, consider the events $A$ "the ace of spades is drawn" and $B$ "any spade is drawn". It is quite clear that $B$ happens more frequently than $A$, even though we have said that this is an experiment with equally likely outcomes. Remember: an outcome is an observation from a single experimental run, an event is any subset of these possible outcomes.

A core goal is then bridging the gap between the probability of an outcome - something which in the equally likely outcomes frameowkr we know - and the probability of an event - the quantity we are actually interested in. In ordder to do so, we will next consider the rules of probability, introducing the properties that are required for valid probability assignments and the techniques for manipulating probabilities to calculate the quantities of interest.

## 2.2 Rules of Probability

We have previously seen that not every probability assignment can be valid. For instance, assigning 0.5 probability to each outcome on a die leads to a nonsenical scenario. with just a little imagine we can conjure equally nonsensical scenarios in other regards. For instance, it would make very little sense to discuss the probability of an event being a negative value. What would it mean for an event to occur in a negative proportion of occurences? Alternatively, we can consider two events that are nested in one another: say event $A$ is that we draw the ace of spades, and event $B$ is that we draw any spade. Every single time that $A$ happens, we know that $B$ also happens. But there are ways that $B$ can occur wwhere $A$ does not (for instance, the Queen of spades being drawn). If I told you the probability of $A$ was 0.5 and the probability of $B$ was 0.2, this would violate our base instincts: how can it be more likely to draw the ace of spades than it would be to draw any spade at all?

Often in mathematics when we have an intuitive set of rules that particular quantities must obey, we work to add formality through defining properties of these concepts. To this end, we can define the key properties that probabilities must obey in order to be well-defined, valid probabilities.

It turns out that with three (fairly) basic properties, we can completely specify what must be true in order for a set of probabilities to be valid.

1. **Unitary** Every valid set of probabilities must assign a probability of 1 to the full sample space. That is, $P(\mathcal{S}) = 1$. This is an intuitive requirement as every time the experiment is run we observe an outcome in the sample space, and as a result, in every experimental run the event $\mathcal{S}$ occurs.
2. **Non-negative:** We require that every probability is non-negative. We can have probabilities of 0, but we can never have a probability less than zero. Again, this is sensible (what would it mean to have a negative probability?), but is important to include in our formalization. Specifically, for everyevent $E$, $P(X) \geq 0$.
3. **Additivity:** the final property requires slightly more parsing on first pass. Suppose that we define a sequence of events, $E_1, E_2, E_3, ...$ such that no two events have any overlap. That is, $E_j \cap E_\ell = \emptyset$ for all $\ell \neq j$. Then, the final property we require for probabilities is that
$$P\left(\bigcup_i E_i\right) = \sum_i P(E_i).$$
That is, the probability of the union of disjoint events is the summation of the probability of these events.

It is worth dwelling slightly on property (or axiom) 3 further. Consider the case of drawing a card at random from a deck of 52 cards. Using the equally likely outcome model for probability we know that the probability that any card is drawn is given by $\frac{1}{52}$. If I were to ask "what is the probability you draw that ace of spades?" under this model you can respond, immediately,

with $\frac{1}{52}$. Now, if I were to ask "what is the probability that you draw the ace of spades or two of spades?" intuitively you likely figure that this will be $\frac{2}{52}$. Note that the event $E_1$, "draw the ace of spades" and the event $E_2$ "draw the two of spades", are disjoint events. Moreover, recall that the union is the "or" and so $E_1 \cup E_2$ is the same as $E_1$ or $E_2$. Taken together then,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2).$$

The axiom of additivity simply extends this intuition to an arbitrary number of events.

TODO: insert example about additivity.

These three axioms fully define valid probabilities. Any mechanism that assigns probability values to events which conforms to these rules will assign valid probabilities. While it may seem counter intuitive that such basic rules fully define our notion of a probability, these rules readily give rise to many other properties that are indispensible when working with probabilities.

## 2.3 Secondary Properties of Probabilities

1. $P(E^C) = 1 - P(E)$, and equivalently, $P(E) = 1 - P(E^C)$.
2. $P(\emptyset) = 0$.
3. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.
4. $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3)$.
5. $P(\bigcup_i E_i) \leq \sum_i P(E_i)$.

TODO: write aside that proves these from the basic properties.

TODO: am I missing any?

These properties are immensely useful when computing probabilities. in fact, these secondary properties will be used with more frequency than the basic axioms when manipulating probabilities in practice. It is worth building comfort with these properties, early and often, as they will assist in manipulating all probability expressions in the future.

TODO: Example showing complement trick.

While these properties hold in general for all probability models, it is instructive to consider to focus on the equal probability model to begin building familiarity with probability broadly. These porperties allow us to take events – whether compound or simple – and combine, rewrite, and manipulate expressions to assist in the handling of the computations. Eventually, however, we require the ability to assign numerical values to these probabilities.

Consider a simple event, $A$. Recall that a simple event is defined as a possible outcome of an experiment, and so in this case, $A$ corresponds directly to an event that may be observed. If our sample space is $k$ elements large, then $P(A) = \frac{1}{k}$ in this framework. For instance, if $A$ is the event that a two is rolled on a six-sided fair die, then $P(A) = \frac{1}{6}$.

Now, suppose that a compound event is defined, $B$. By definition, a compound event can be expressed as a set of possible outcomes from the experiment. Suppose that we enumerate these possible events as $b_1, b_2, \ldots, b_\ell$. Then we know that $B$ occurs if any of $b_1, b_2, \ldots, b_\ell$ occur. Each $b_j$ are elements of the sample space and correspond to possible outcomes of the experiment. As a result, we know that $P(b_j) = \frac{1}{k}$, based on the equal probability assumption. Now, if we take any two distinct $b_j$, say $b_i$ and $b_j$, we know that they must be disjoint: $b_i \cap b_j = \emptyset$. This is because in an experiment run only one outcome can occur – there is nothing common between these two events, by definition. Moreover, we can say that $B = b_1 \cup b_2 \cup \cdots \cup b_\ell$.

Using the axioms of probability outlined above we therefore know that $P(B) = \sum_{j=1}^{\ell} P(b_j) = \sum_{j=1}^{\ell} \frac{1}{k} = \frac{\ell}{k}$. This holds in general for any compound event in this setting. If we take $B$ to be the event that an even number is rolled on a six-sided die, then we would have $b_1$ is the event that a two is rolled, $b_2$ is the event that a four is rolled, and $b_3$ is the event that a six is rolled. There are three such events, and so the probability that an even number is rolled must be $\frac{3}{6} = 0.5$, which matches our intuition.

If we consider what this process is doing at its core, we can reframe the calculation as counting up the number of ways that event can happen and dividing by the total number of events. In our previous discussion, there were $\ell$ ways of $B$ occuring, a total of $k$ outcomes, and so the probability becomes $\frac{\ell}{k}$. in the equal probability model, this will always be the case, the probability of any event $A$ occuring is given by

$$P(A) = \frac{N_A}{k},$$

where $N_A$ is the number of unique ways that $A$ can occur. In other words, $N_A$ is the size of the set $A$.

As a result of the realization, a large part of computing probabilities is in the counting of possible outcomes corresponding to different events. If we can determine the count of $A$, $N_A$, and the count of the total number of occurences, $k$, then we can determine the probability of $A$. This study of counting is known as **combinatorics**, and it is where we will turn our attention next.

TODO: include example that shows the utility of complements, again. Show specifically for counting where, for instance, the complement is very small and is easier to count.

## 2.4 Combinatorics

Fundamentally, counting is a matter of assessing the size of a collection of items. Sometimes, this is very straightforward: if you want to count the number of students in a classroom, you start at 1 and enumerate upwards through the integers. To count the number of days until the next Holiday, you do the same thin, sizing up the time that has to pass. If you really need to sleep, perhaps you will count imaginary sleep until you drift off. There is not much to this

type of counting, and it is certainly deeply familiar to you all. However, it is also quite limited in its utility.

Imagine that you are interested in determining how many possible ways there are of arranging a deck of \$524 cards. You could of course arrange them in a particular order, then count each of those. That would take a tremendous amount of time, so perhaps instead of using an actual deck you just write down the combinations. Still, each combination is going to be 52 cards long, and keeping track of that all will be a tremendous challenge. This seems like an approachable question, and yet, it illustrates how complicated (and large) these types of "counting" problems can become, very quickly.

Fortunately for us there are some strategies for simplifying these problems down, some of which you are likely already familiar with. Think about trying to form an outfit where you have 4 different sweaters, 3\$ pairs of pants, and 2 options for your shoes. Suppose that any combination of these will work well. How many are there? Well, if you have already picked your sweater and pants, then there are going to be 2 different outfits using these: one wiht each of the pairs of shoes. This is true for each ppossible sweater-pant combination, and so we can count 2 for each one these. In otherwords, to get the total number of outfits we multiply the number of sweater pant combinations by the number of shoe options (2). the same rationale can be applied to count the total number of sweater-pant combinations. For each sweater, there are 3 pairs of possible pants, and so to get the total number there we can take 3 for each possible sweeater, or in other words, #3×4\$.

Taken together then we have $4 \times 3 \times 2 = 24$ total possible outfits. Another way of framining this is that we have to make three sequential decisions: which of the 4 sweaters, which of the 3 pants, and which of the 2 shoes are to be worn. When we do this we multiply through the number of alternatives at each decision point to get the total number of combinations.

This is known as the **multiplication rule for counting**, and it is a very general rule for counting up total combinations. Any time that we have to make a sequence of $k$ choices, and each choices $j = 1, \dots, k$ has $n_j$ options, then the total number of combinations will be $N = n_1 \times n_2 \times \cdots \times n_k$.

TODO: add examples for the counting rule.

Sometimes it is helpful to express the counting rule graphically. To do so we rely on **tree diagrams**. A tree diagram puts each of the decisions in sequence, and draws a branch for each separate option. You start with the first choice, drawing one branch for each of the $n_1$ alternatives, labelling each. Then, at the second choice, you do the same process at the end of each of the branches you drew for choice 1, this time drawing $n_2$ branches there (so you will have just drawn $n_1 \times n_2$ branches). Then for each of those you draw the $n_3$ further branches, and so on and so forth until the end.

If you want to know the total number of choices, you simply count the end points at the very end of the diagram. Each branch corresponds to a single option. To determine which combination of choices it corresponds to, you simply read off the branch labels at each branch

you take. If you want to know how many possible combinations come with certain options selected, you canlook at only those branches which are downstream from the choices that you care about.

TODO: include tree diagram.

While tree diagrams can be quite useful for visualizing a problem, they often grow to be overly complex. As a result, we often need to fall back on the numerical representation afforded to us through the product rule for counting. Counting problems, in general, can very quickly become tremendously large and complex. For this reason, we have several tools to assist us in reducing this complexity based on common types of problems that we would like to count.

The first useful tool for simplifying these problems is the **factorial**. The factorial of an integer, denoted by $x!$ (factorials are exciting because they always look like they are shouting!) is given by the product of all integers from $x$ to 1. That is, $x! = x(x-1)(x-2)\cdots(2)(1)$. If we consider the product rule for counting then note that if $n_1 = 1$, $n_2 = 2$, … , $n_k = k$, then the total number of options is $k \times (k-1) \times \cdots \times 1 = k!$. The most common reason that this comes up is when we want to order a collection of items. Suppose that you have 10 books that you want to plae on a shelf. You can view this as making 10 sequential decisions: what book goes first, second, third, and so on. There are 10 options for the first book, then 9 for the second (any except for the first one), and then 8 for the third (any except for the first 2). This continues down to the last book, and so we conclude that there are $10 \times 9 \times 8 \times \cdots \times 1 = 10!$ ways of arranging these books.

TODO: Additional example on factorial.

Sometimes, we want to still order items from a collection, but we want to only use a subset of these tiems. That is, suppose that you have 20 books, only 10 of them will fit on the self, and so you want to know: how many ways can you put 10 books on the shelf, in order, from your collection of 20. Using the product rule of counting for this directly we can recognize that there are 20 options for the first, then 19, then 18, and so on until there are 11 choices for the 10th book to place. We can write this out in a seemingly strange way.

$$\frac{20(19)(18)(17)(16)(15)(14)(13)(12)(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}{10(9)(8)(7)(6)(5)(4)(3)(2)(1)}.$$

This expression is 20! divided by 10!, and gives the same as our argument from the product rule for counting directly. This is a more general result than our example with books would suggest. If we have $n$ items, and we want to choose $k$ of them and then order those choices, it will always be $n!$ divided by $(n-k)!$. We call this a **permutation**. Formally, we write

$$P_{n,k} = \frac{n!}{(n-k)!}.$$

Permutations arise when we select ordered subsets from a collection. We often, in combinatorial problems, talk about ordering, though sometimes what we mean by this is slightly more

abstract. Suppose that you want to form a committee with 5 different people, each of whcih occupies a different role: the president, vice president, treasurer, note taker, and critic. If there are 30 people to select for this, then there are $P_{30,5}$ total possible committees that can be formed. While there is not a sequential order here, we talk about this as being ordered since we can differentiate between the five roles. Instead of labelling them with their names, we could label them 1 through 5 and make the ordering more explicit.

TODO: include additional example on permutations. TODO: $0! = 1$.

Factorials compute the number of orderings for a set of objects, and permutations compute the number of ordered subsets from a collection of ojects. What about when we do not wish to differentiate the order of subsets? Suppose that you still need to form a 5 person committee, but you do not have explicit roles for the different members of the committee. Here we cannot use a permutation directly, as we know that this takes into account the order. Suppose that we formulate the ordered committee in a separate way: first, we select 5 people without concern for their order, then we choose which order they will have.

If $M$ represents the number of unordered sets of 5 from this population, the product rule for counting tells us that the total number of ordered committees will be $M \times 5!$, since there are 5! arrangements. Thus, we can write this down as

$$P_{30,5} = \frac{30!}{25!} = M \times 5! \implies M = \frac{30!}{25!5!}.$$

Once again, this will be true far more broadly than our committee example: if we want to select $k$ items from a collection of $n$, we will have $n!$ divided by the product of $k!$ and $(n-k)!$. We refer to these as **combinations**.

We will write $\binom{n}{k}$, which we read as "$n$ choose $k$", which translates to "select $k$ items from a population of $n$ total options, without concern for their order." Formally, a combination is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

To summarize: factorials allow us to order a complete collection, permutations allow us to select a subset with consideration of the ordering, and combinations allow us to select a subset from the collection without regard to the order. These three techniques can be used in combination with the product rule for counting to allow us to have very complex total summations.

TODO: introduce example for combinations. TODO: introduce extended examples for counting, mixing the different techniques.

There are other types of counting problems which ocassionally need to be considered. [TODO: include counting with replacement].

While combinatorics is a field of study on its own, with many intriguing tools and developments surrounding the neumeration of objects, for the purposes of simple probability models these

tools will suffice. Ultimately, we care about counting since in the equal probability model, the probability of any event can be determined by counting the number of ways that the event can occur and dividing by the total number of outcomes that are possible. That is, we use these tools to derive $N_A$, the total number of ways that $A$ can occur, and $N$, the total number of experimental outcomes, and then we conclude that

$$P(A) = \frac{N_A}{N}.$$

TODO: include brief probability example.

# 3 Probailities with More than One Event

## 3.1 Conditional Probability (and related theorems)

Regardless of whether we are using the equally likely outcome model, or not, there are several properties of probability which can be derived that are either of direct interest themselves, or else are useful for manipulating probability expressions. These rules apply to any probability model, and greatly increase the flexibility of working with these models.

Up until this point we have primarily focused on assigning probabilities to particular events. If we have some event of interest, $A$, then $P(A)$ is the probability that $A$ occurs in any manner. If we are using our equally likely probability model, then $P(A) = \frac{N_A}{N}$. This is the probability of the event $A$ where nothing else is known at all. If we smooth over anything which could alter the likelihood, if we have no additional information, if we want the best guess for the likelihood of occurence in a vaccuum, this is the probability of interest. We refer to such quantities as **marginal probabilities**.

While marginal probabilities are often of interest, and frequently are the best tool for summarizing the overall state of the world (or our knowledge regarding the state of the world), in pracctice we know that sometimes information that we have will change our understanding of the probability of an event. Suppose the event $A$ corresponds to the event that it snows tomorrow, in some particular city. It is possible to think about how often it snows on average, and report a value related to that as $P(A)$. Now, what if we know that it is currently the middle of summer? In this case, while $P(A)$ does not shrink to 0, it becomes far less likely than if we did not have that information. Similarly, if we know that it is winter, the likelihood that it snows tomorrow increases.

In order to formall capture this we can introduce the idea of **conditional probability**. Unlike in the cas of marginal probabilities, conditional probabilities allow us to *condition* on extra pieces of information. Instead of asking "what is the porbability of this event", they instead ask, "given that we know this piece of information, what is the probability of this event?" The subtle distinction becomes quite powerful, both in terms of manipulating and working with probabilities, but also in terms of expressing the correct events of interest for ourselves.

To make use of conditional probabilities, we will think of the process of **conditioning** on (one or more) events. That is, we will talk of the probability of $A$ conditional on $B$, where $A$ and $B$ are two events of interest. We write this quantity as $P(A|B)$, and typically will read this as

"the probability of $A$ given $B$." Intuitively, this it the probability of $A$ happening, supposing that we know that $B$ has already happened.

TODO: Include some examples.

Recall that $A$ and $B$, as events, are merely subsets of the sample space, $\mathcal{S}$. Each item in either $A$ or $B$ is one of the possible outcomes from the experiment or process that we are observing. Suppose that we know that $B$ has occurred. What this means is that, one of the outcomes in the set $B$ was the observed outcome from the experiment. Now, if we want to know $P(A|B)$, we want to know the probability, working from the assumption that $B$ has happened, that $A$ also happens. That is, knowing that $B$ has happened, what is the probability that $A$ and $B$ happen.

Thinking back to our discussion of events, $A$ and $B$ is denoted by $A \cap B$, and it corresponds to the set of events inside the set $B$, which also belong to the set $A$. Now, instead of considering the event $P(A \cap B)$ directly, we need to acknowledge that for $A|B$, only the events in $B$ were possible. That is, instead of being divided by the whole space, we can only divide by the space of $B$. In some sense, we can view conditioning on $B$ as treating $B$ as though it is the full sample space, and finding probabilities within that. In general, $B$ will be smaller than $\mathcal{S}$, and so $P(B) < 1$. Instead of the conditional probability being "out of" 1, it will instead be "out of" $P(B)$, which gives $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

To make this more clear, let's consider a simple example. Suppose that we take $A$ to be the event that a 2 is rolled on a fair, six-sided die, and $B$ to be the event that an even number was rolled. This is an equal probability model, and so each outcome gets $\frac{1}{6}$ probability. The original sample space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, the event $A$ is $\{2\}$, and the event $B$ is $\{2, 4, 6\}$. In order for both $A$ and $B$ to occur, we note that we need $A \cap B = \{2\}$. If we know that $B$ has occured, then we know that either a 2, 4, or 6 has been rolled, with equal probability for each. Thus, intuitively, we can view $B$ as the new sample space, and say that rolling a 2 has a $\frac{1}{3}$ probability, given that there are 3 outcomes and 1 of them is the event of interest. Another way to consider this is to note that $P(B) = \frac{1}{2}$, and so we need to scale each event by $\frac{1}{1/2}$ in order to make sure that the total probability of our reduced sample space equals 1. Then $P(A \cap B) = \frac{1}{6}$, so

$$P(A|B) = \frac{1/6}{1/2} = \frac{1}{3}.$$

TODO: try rewriting this a bit.

Suppose that, instead of a fair die, it was weighted so that 6 comes up more frequently than the other options, coming up with probability 0.5, and the other fives each coming up with probability 0.10. If $A$ and $B$ are the same events as above, then $P(B) = 0.7$ now. If we know that $B$ has occurred, then the new sample space is $\{2, 4, 6\}$ where $P(2) = \frac{0.1}{0.7} = \frac{1}{7}$, $P(4) = \frac{0.1}{0.7}$, and $P(6) = \frac{0.5}{0.7} = \frac{5}{7}$. Note that these three probabilities sum to 1 still, which constitutes a valid probability model, and so $P(A|B) = \frac{1}{7}$.

TODO: Add example about this.

Ultimately, the conditional probability of $A$ given $B$ is gtgiven the definition of

$$P(A|B) = \frac{PA \cap B}{P(B)}.$$

This is defined only for when $P(B) > 0$. The numerator in the expression, $P(A \cap B)$ is called the **joint probability** of $A$ and $B$, and may also be written as $P(A, B)$. Sometimes, we wish to condition on more than one event. To do so, the same process extends naturally. For instance, suppose we want to know the probability of $A$ given $B$ and $C$. This would be written

$$P(A|B, C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(A, B, C)}{P(B, C)}.$$

Moving beyond two events occurs in the expected way.

Conditional probability is a mechanism for capturing our knowledge of the world, and using that to update our sense of the uncertainties at play. For instance, suppose that we are interested in drawing a random card from a deck of 52, and we want to know the probability that it is a heart. Without any additional knowledge, the probability of this event is $\frac{1}{4}$. Now, suppose that you know that it is a red card. In this case, we now know that itis either a heart or a diamond, and there are equal numbes of each, meaning that the new probability is 0.5. We can work this out directly

$$P(\text{Heart}|\text{Red}) = \frac{P(\text{Heart}, \text{Red})}{P(\text{Red})} = \frac{P(\text{Heart})}{1/2} = \frac{1/4}{1/2} = 0.5.$$

Suppose instaead that wwe had been told that the card was an ace. Here we now know that there are four possible outcomes that correspond to an ace, and only one of these is a heart, meaning the probability is $\frac{1}{4}$. In this case, $P(A|B) = P(A)$, and our beliefs did not update.

What if instead we had considered the second event to be "the card was a spade." In this case if we want to know $P(A|B)$ then, given a spade being drawn, we know that the probability of drawing a heart is 0.

TODO: Add further examples of conditional probabilities.

While sometimes we will want to work out the conditional probability, knowing the joint and marginal probabilities, there are other times where it is easier to determine the conditional probability directly, but where we wish to understand the marginal probability. That is, we may be able to know $P(A|B)$, but we want ot make statements regarding $P(A)$ or $P(A, B)$.

In this case, we can simply rearrange the defining relationship of conditional probability, to solve for the quantities of interest. Because of the importance of this procedure, we actually give the most straightforward rearrangment a special name. Note that, by multiplying both sides of the definition of $P(A|B)$ by $P(B)$, wwe get that

$$P(A \cap B) = P(A|B)P(B).$$

This is known as the **multiplication rule**. In particular, it states that we can solve for the joint probability of $A$ and $B$ by multiplying the conditional probability of $A$ given $B$, by the marginal probability of $B$. Note that this is symmetric in $A$ and $B$ so that

$$P(A \cap B) = P(B|A)P(A).$$

This is useful as sometimes it is easier to determine $B$ given $A$.

TODO: Example regarding the multiplication rule.

While the multiplication rule gives us the capacity to solve for joint probabilities, often we wish to make statements regarding marginal probabilities. Fortunately, we can extend this process outlined in the multiplication rule to solve directly for marginal probabilities as well. To do so, we first introduce the concept of a **partition**.

A **partition** is simply a collection of sets which divide up the sample space such that all of the sets are disjoint from one another, and there is no overlap between any of the sets. For instance, if the sample space were all the positive integers, we could partition this space into all the even numbers as one set and all the odd numbers as a second. We could also partition this into the set of numbers which are less than 10, the set of numnbers that are greater than 10, and then 10. In both examples we have sets that form the full sample space with no overlap. In notation, if our partition is formed of $B_1, B_2, B_3, ...$, then we require $\bigcup_i B_i = \mathcal{S}$ and that $B_i \cap B_\ell = \emptyset$. Note that we could not partition the set into multiples of 2 and multiples of 3, since (i) not all values are contained between these two sets, and (ii) there is overlap between these two sets.

TODO: Include further partition examples

Given a partition, $B_1, B_2, ...$, suppose that we can work out both $P(B_i)$ for all $i$, and $P(A|B_i)$ for all $i$. In this case we can work out the marginal probability of $A$ using **the law of total probability**. To do so we take

$$P(A) = \sum_i P(A|B_i)P(B_i).$$

Intuitively, since the whole sample space is divided into the different $B_i$s, this rule breaks down the calculation of $A$ happening into manageable chunks. Each term in the summation is "the probability that $A$ happens, given $B_i$ happening" weighted by how likely it is that $B_i$ happens. Then by summing over all possible $B_i$, we know that we must be capturing all possible ways that $A$ can occur since all parts of the sample space are contained in exactly one of the sets of our partition.

The law of total probability is an indispensible tool for computing probabilities in practice. Suppose that an online retailer uses two different services to send out their packages, the first one delivers later with probability 0.05 and the second delivers late with probability 0.1. If the chances that a customer selects the first supplier are 0.75, then with this information we can work out the probability that any randomly selected package will be late. First note that in

this case our sample space is comprised of $\mathcal{S} = \{(L, A), (L, B), (O, A), (O, B)\}$ where $L$ stands for late, $O$ for on time, $A$ for the first compnay, and $B$ for the second. We want to know $P(L)$, where $L = \{(L, A), (L, B)\}$. Note that we can partition the space into those sent by $A$ and those sent by $B$, and we know that $P(A) = 0.75$ and $P(B) = 0.25$. Moreover, we know that $P(L|A) = 0.05$ and $P(L|B) = 0.10$ and so combining this all together we get

$$P(L) = P(L|A)P(A) + P(L|B)P(B) = (0.05)(0.75) + (0.10)(0.25) = 0.0625.$$

TODO: Write another example on the utility of L.o.T.P.

We have seen the direct computation of marginal probabilities (while using an equally likely outcome model), the computation of conditional probabilities, the use of the multiplication rule for joint probabilities, and the use of the law of total probability to indirectly calculate marginal probabilities through conditioning arguments. Throughout these discussions we have been primarily concerned with keeping events $A$ and $B$ arbitrary. Everything that we have indicated for $P(A)$ holds for $P(B)$, as does $P(A|B)$ and $P(B|A)$. In reality, it will often be the case that conditioning on one of the events will be natural, while conditioning on the other will be more tricky. In these events, it can be useful to be able to transform statements regarding $P(A|B)$ into statements regarding $P(B|A)$, and vice versa.

Note that the symmetry of definitions gives the fact that

$$P(A|B)P(B) = P(A, B) = P(B|A)P(A).$$

This is an application of the multiplication rule in two different orientations. If we divide both sides of the equality by $P(B)$, assuming that it is not 0, then we get

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Now, if we form a partition, say $A, A_2, A_3, \cdots$, then we can rewrite $P(B)$ using the law of total probability as

$$P(B) = P(B|A)P(A) + P(B|A_2)P(A_2) + \cdots,$$

which can replace $P(B)$ in the expression. Taken together, we refer to this relationship as **Bayes' Theorem**, and it is typically expressed as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + \sum_i P(B|A_i)P(A_i)}.$$

Bayes' Theorem allows us to convert statements regarding $P(B|A)$ into statements regarding $P(A|B)$. Note that, as we derived above, Bayes' Theorem is really just an application of the multiplication rule and an application of the law of total probability. Sometimes we may have $P(B)$ directly, rendering the law of total probability in the denominator unnecessary. Often, the natural partition to select when we do need the law of total probability is to take $A, A^C$.

Note that any set with its complement forms a partition, since by definition they occupy the entire space and are non-overlapping. When this is done we get the slightly more compact relationship of

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}.$$

Unlike our previous relationships, Bayes' Theorem allows us to translate one set of conditional knowledge into another. The most common example application is in medical testing. Suppose that we know the performance characteristics of a particular medical test: it is 99% accurate for positive cases, and 95% accurate for negative cases. That is, wit h probability 0.99 it correctly returns positive when an individual is infected, and with probability 0.95 it returns negative when an individual is not infected. These are both statements of conditional probability. If we take $A$ to be the event that the test returns positive, and $B$ to be the event that the patient is infected, then we are saying that $P(A|B) = 0.99$ and $P(A^C|B^C) = 0.95$ which means that $P(A|B^C) = 0.05$. Suppose that we know that, across the entire population, one in a thousand individuals is likely to be infected. This means that $P(B) = 0.001$. Now suppose that a random individual goes into a doctor's appointment, and tests positive for the disease. How likely are they to actually be infected?

Note that in this case we want to know the probability of them being infected given that they have tested positive. In notation, this is $P(B|A)$. We do not know this quantity directly, but given a simple application of Bayes' Theorem, we can work it out. Here using the natural partition of $B, B^C$, we get

$$P(B|A) = \frac{P(A|B)P(A)}{P(A|B)P(B) + P(A|B^C)P(B^C)} = \frac{(0.99)(0.001)}{(0.99)(0.001) + (0.05)(0.999)} \approx 0.019.$$

That is, despite the fact that this test is exceptionally effective at detecting this disease, a positive test still means that an individual has a probability of only 0.019 of actually having the illness. This counter intuitive fact was an intensely frustrating reality for statisticians everywhere during the height of the COVID-19 pandemic, when politicians and the population at large turned away from testing owing to its perceived ineffectiveness.

TODO: Add another exmaple regarding Bayes' Theorem

Bayes' Theorem highlights a key lesson when considering conditional probabilities, and it's a common mistake to make which should be avoided at all costs. Namely, we cannot interchange $P(A|B)$ with $P(B|A)$. These probabilities are not necessarily highly correlated with one another, and it is important to distinguish clearly which is the probability of interest. Mixing up $P(B|A)$ and $P(A|B)$ is often called "confusion of the inverse," and it can lead to very faulty conclusions when ignored. In the medical testing example above, it is important to not confuse "the probability that the test returns positive, assuming you have theillness" with "the probability that you have the illness, given that the test returns positive." The stakes of these types of confusion can be quite high, and it is tremendously important to ensure that

you are conditioning on the correct events. Fortunately, Bayes' Theorem allows us to translate ebtween events for conditioning, giving a mecchanism from translating between the two.

TODO: Include Another Example (perhaps Smoking vs. lung cancer)

TODO: Fix Bayes' to Bayes'

TODO: Add in subsections

## 3.2 Independence

We have seen that, in most cases, when we have conditioned on an event it has changed the probability of that event. For instance, if we want to know the probability it is raining, if we condition on knowing that it is a day full of gray skies, then the probability likely increases. By considering how tehse probabilities change when we condition, we are in effect indicating a dependence of the events on one another. In terms of probability, this dependence is captured by an influence on the degree of uncertainty present, depending on what we know.

It is of course, totally possible that two events do not influence one another. The weeather outside today is likely not influenced by your favourite sports team's performance last night, for instance. In this case, we would be suggest that $P(A|B) = P(A)$. We saw an example where this was concerned previously when we wanted to know the probability of a randomly selected card being a heart ($A$), given that it was an ace ($B$). We found tat this was $\frac{1}{4}$, exactly the same as the probability if we did not know that it was an ace. Thus here we have $P(A|B) = P(A)$. Note that in this case, we could have also said that $P(B|A) = P(B) = \frac{1}{13}$.

The symmetry of these events makes it somewhat more convenient to express this relationship differently. If we multiply both sides of the $P(A|B) = P(A)$, by $P(B)$, then applying the multiplication rule gives

$$P(A, B) = P(A)P(B).$$

Any two events that satisfy this relationship are said to be **independent**. If $A$ and $B$ are independent we write $A \perp B$.

Note that independence is always a symmetric property: if $A$ is independent of $B$, then $B$ is independent of $A$. To check whether two evetns are independent, we simply need to check whether their joint probability (that is, the proobability of $A$ and $B$) is equal to the product of their marginal probabilities. If $P(A) \neq 0$, then we can divide both sides by $P(A)$ to gives $P(B|A) = P(B)$. Similarly, if $P(B) \neq 0$, then we can divide both sides by $P(B)$ to give $P(A|B) = P(A)$.

This expression in terms of conditional probabilities is the more intuitive expression of independence. It directly captures the idea that "knowing $B$ does not change our beleif about $A$", which is how independence is best thought of. However, we must be careful. This conditional

argument is only valid when the event that is being conditioned on is not probability 0, where the defining relationship, $P(A, B) = P(A)P(B)$, will hold for any events.

TODO: Example assessing the independence of several events.

When two events are not independent, we say that they are dependent, and may sometimes write $A \not\perp B$. Note that we saw that, in general, $P(A \cap B) = P(A) + P(B) - P(A \cup B)$. It is only when assuming independence that this simplifies further.

Importantly, if $A \perp B$, then $A \cap B \neq \emptyset$ unless either $A = \emptyset$ or $B = \emptyset$ or both. To see this recall that $P(\emptyset) = 0$, and so if $A \cap B = \emptyset$ then $P(A \cap B) = P(A)P(B) = P(\emptyset) = 0$. This only holds if either $P(A) = 0$ or $P(B) = 0$. This may seem to be a rather technical point, however, it is the source of much confusion regarding independence. In particular, it is common to mistake independent events for **mutually exclusive events**.

Suppose that my partner and I always eat dinner together. It will either be the case that I cook at home, that they cook at home, that we order in, or that we go out to eat. If we take these to be four events, $A$, $B$, $C$, and $D$, then it reasonable to suggest that only one of these can happen on any given night. Whenever this is the case, we refer to the events as being mututally exclusive: if one happens, we know that the others did not. Mutually exclusive events are always dependent since knowing that $A$ occurs dramatically shifts our belief about $B$,$C$, and $D$ (namely, we now know that they are impossible).

The primary concern with mututally exclusive and independent events is a linguistic one. We often use words like independent to mean unrelated, and in a sense, mututally exclusive events are unrelated in that one has nothing to do with another. However, in statistics and probability, when we discuss independence, it is not an independence of the events themselves but rather an independence relating to our beliefs regarding the uncertainty associated with the events. In this sense, mutually exclusive events are very informative regarding the uncertainty associated with them.

TODO: Write example regarding mututally exclusive and independent events.

## 3.3 Contingency Tables

Through to this point we have discussed probabilities in the abstract, either through an enumeration of equally likely outcomes, or else by directly specifying the likelihood of various events. While these are useful in many regards, we are often looking for more concise manners of summarizing information of interest. One tool for accomplishing this is a **contingency table**. A contingency table summarizes teh frequency of occurence of a variable (or multiple variables) across a population of interest.

For instance, you could form a contingency table relaying the frequency with which undergraduate students are enrolled in various faculties at a particular university. This tells you, of the

whole population of students at the university, what is the faculty breakdown. By dividing the number in each faculty by the total number of students, you convert the frequencies to proportions, and these proportions can be viewed as probabilities.

TODO: write-up frequency and prop table.

The interpretation of proportions as probabilities implies a very specific statistical experiment. In particular, the proportion represents the probability that an individual selected at random from the entire population has the given trait. This is frequently a probability of interest, which makes these summary tables a useful tool. Often, when a single trait is displayed we will not refer to them as contingency tables but rather as frequency tables or frequency distributions. A contingency table will typically plot two or more traits on the same table, with each cell representing the frequency of both traits occurring simultaneously in the population. For instance, we may further include the student's years to see the breakdown of both faculty and year of study, in one table.

TODO: Include updated table

By including two (or more) factors on the table we are able to capture not only the marginal probabilities for the population, but also the joint probabilities for the population, and in turn, the conditional probabilities for the population. Being able to concisely summarize all of these concepts regarding traits in a population of interest renders contingency tables immensely useful in the study of uncertainty broadly.

TODO: Include example(s) of contingency tables.

Suppose we consider a two-way contingency table. Each cell consists of frequnecy with which a combination of two traits occurs in the population. If we take events corresponding to each of the levels of the two varaibles of interest, then these central cells represent the frequncy of joint events. That is, each interior cell gives the total number of observations with a set level for variable one AND a set level for variable two. Each row is summed, with the total number following into the corresponding row recorded in the right hand margin. each column is summed, with the total number corresponding to the given column recorded in the bottom margin. Then the margin totals are summed and the total is recorded in the lower right margin space.

TODO: Include graphic with summation of margins

Whether the rows or columns are summed, they should sum to the same total, which is the total of the population under consideeartion. This is the same as simply adding all of the observed interior frequencies. To turn a frequency into a probability, you need only divide the correct frquency by the correct total.

For the standard joint probabilities, you take the interior cell count and divide by the population total. Here we are saying that some fixed number, $m$, of the $N$ total individuals have both traits under consideration. If instead you wish to find a marginal probability, you have to consider the value in the corresponding margin: this is the total number of individuals with

the given trait, ignoring the level of the other variable. These marginal values are also divided by the total population size. For conditional probbailities, we restrict our focus to either only one row, or one column. Then, we can take teh joint cell and divide by the value in the margin, which gives the conditional probability of interest.

TODO: show probability caluclations for the various scenarios that we saw.

Note that these procedures are exactl in line with what we had seen before. The conditional probability is defined as the joint probability divided by the marginal probability. The process of computing the marginal can be seen as an application of the law of total probability. As a result, contingency tables can be a useful, tangible tool for investigating the techniques we have been discussing: they are not a substitutefor direct manipulation of the mathematical objedts, but they can present insight into the underlying processes where it may be hard to derive that insight otherwise.

TODO: show example of conditional probability derivation.

TODO: show example of law of total probabilitty derivation.

It is important to note that there is redundant information within a contingency table. For instance, the margins need not be listed explicit, as they can be directly calculated from the interior points. Same goes for interior points, given the margins of the table (assuming other interior points are also presented). TThis can be useful for a campact representation of the information, and manipulating these tables, finding the required information in many places, should become second anture as you continue to work with them more and more.

TODO: include fill-in-the-blank contingency table

It is also important to recognize that independence and mutually exclusive events can be codified via the table as well. Zeros on the interior points indicate events which are mutually exclusive: if we know that one of them occurred, we also know that the other one did not. For independence, it requires a degree of solving proportions. We can either check that the joint probability ($N_{A,B}/N$) is equal to the product of the two marginal probabilities, ($N_A N_B/N^2$), or else (assuming that the events are all non-zero), that the conditional probability ($N_{A,B}/N_A$) equals the marginal probability $N_B/N$. Either way this is represented by $N \times N_{A,B} = N_A N_B$, and when this holds, we can conclude that the events are independent.

TODO: Example on mututally exclusive / independent events on a contingency table.

# 4 Random Variables

When introducing probability originally we worked from a sample space and then corresponding events. This is a very general framework which allows us to effectively capture any statistical experiment that we may want. Sample spaces are not resitrcted to be numeric, for instance, and events are simply subsets of the sample space. As a result, this framewokr provides the tools for capturing uncertainty quantification in just about every setting that we may desire. Still, the need to enumerate sample sapces adn events over complex sets of arbitrary items is cumbersome and may prevent succinct representations of the underlying phenomenon. Often, rather than caring about the entire space out outcomes from an experiment of interest, we are primarily concerned we a summary of the experiment.

When we can summarize the experiment using a numerical quantity, we are able to define a **random variable**. Arandom variable is simply a numeric quantity whose speciifc value depends on chance, through the outcome of a statistical experiment. Specifically, a random variable is a mapping from the result of an experiment to a set of numbers, and by reporting the numeric value of the random variable, we are able to summarize the key part of the experiment, succinctly.

For instance, suppose that we are repeatedly tossing a coin. If we toss the coin 1100 times, then the sample space is going to consist of $2^{100}$ total possible outcomes, each of which is a sequence of 100 heads and tails. Instead, it may be more convenient to assign a random varaible to be the number of heads that show up on the 100 tosses of the coin. In this case, the random variable takes on a non-negative integer between 0 and 100. In many situations, such a summary may be all the is relevant from the experiment.

TODO: Include an example of random variables.

because of their ability to summarize effectively the components of a random experiment that we care about, random variables will become the default paradigm for discussing randomness going forward. Wehn discussing a random variable we will typically use capital letters, $X$, to represent the random quantity with an unknown value. In the event that an experiment is actually performed, and a value is realized for a random variable, we will record this value as a lower case letter, such as $x$. For instance, the number of heads showing in 100 flips of a coin is an unknown quantity depending on chance, which we call $X$. Once we have flipped the coin 100 times and observed 57 heads, we denote this as $x = 57$.

The importance of this notation is merely to emphasize what values are unknown and random, and what values are simply numeric quantities. because $x$ is a known value, taking on some

set number, we will not often speak of probabilities involving $x$. Instead, we wish to translate the language of probability that we have built to statements regarding the random variable $X$.

The random variable $X$ has a corresponding sample space of possible values that it can take on. This sample space, which we can think of as directly analogous to the sample space of arbitrary elements from before, will be dictated by the possible realizations from the underlying experiment. Then, after the experiment has been performed, the random variable will take on a single value from this set. Often we are able to very compactly describe the set of possible values for a random variable, for instance, by stating all of the integers, or integers between 5 and 10, or even vvalues less than 1000. The probability of realizing any of these outcomes is then, just as in the case of the arbitrary sample space, dictated by the underlying probability model.

When introducing the concepts of probability we indicated that probability was assigned to events. When using random variables we still use this convention, and as a result, we need to define events in terms of random variables. When we have a random variable, $X$, an event is defined as any set of values that it can take on. For instance, we may have the event $X = 4$, or the event $X \geq 18$, or the event $2 \leq X \leq 93$, or the event $X \in \{2, 4, 6, 8, 10\}$. In each case these are simply subsets of the possible outcomes that can be observed on the experiment, once summarized through the lens of the random variable.

Note that, just as was the case before, these events can be though of as being compound events (comprising of multiple outcomes) or simple events (comprised of a single outcome). The event $\{X = 5\}$ is a simple event, whereas the event $\{X \geq 25\}$ is a compound event. With events defined in this way, we can translate all of the other concepts over from before. Specifically, we merely think of the experiment as an experiment producing a numerical outcome, and then use the same sets of tools as we did before, applied to numeric events.

TODO: Include example recapping previous discussions with regarding the random variables.

When considering random variables there is a key distinction between two types of random variables: discrete and continuous. A discrete random variable is a random variable which takes values from a countable set of possiblevalues. When we say countable we mean that there are either a finite number of possible values that it can take on (say $\{1, 2, \ldots, 31415\}$) or an infinite number of possible values, but values that can all be written out in sequence (say $1, 2, 3, 4, \ldots$ or $2, 4, 5, 6, 8, 10, \ldots$). Continuous random variables, on the other hand, are random variables which take on values from an uncountably infinite set of values. Typically this will be expressed as random varaibles which can take on any value in some interval (or set of intervals), such as $X \in [0, 1]$ or $X \in (0, \infty)$ or $X \in (-\infty, 129]$. In all of these cases there is no way to enumerate the possible set of values that $X$ can take on, and so we say that there are an uncountable number of them.

TODO: Include example regarding discrete and continuous random variables.

We will discuss continuous random variables, and how they differ from the discussions we have had up until this point, shortly. For now, we turn our focus to discrete random variables. One of the major utilities of random variables, as has been previously mentioned, is that they provide a shorthand for summarizing the results of a statistical experiment. To this end, there are also several tools which have been developed relating to random variables which help to expedite the manipulation of ocncepts related to probability calculations.

Chief among these tools is the concept of a **probability distribution**. A probability distribution is a summary of the probabalistic behaviour of a random variable. Distributions capture the underlying random behaviour of the random variables of interest, and in so doing, summarize information regarding the experiment or process that is being considered. When concerned with discrete random variables, we typically summarize probability distributions through the use of a **probability mass function**.

A probability mass function is a mathematical function which maps possible values for a discrete random variable to the probability correspond to that event. That is, if a random variable $X$ can take on the values $x_1, x_2, \dots, x_k$, a probability mass function is a function $p(x)$ such that $p(x_1) = P(X = x_1)$, $p(x_2) = P(X = x_2)$, ..., $p(x_k) = P(X = x_k)$. As a result, once a probability mass function is known, all of the probabalistic behaviour of the random variable can be fully described.

TODO: Include example regarding pmf.

The conditions that we previously outlined for probabilities still must hold when assessing probabilities through a probability mass function. As a result, we know that probabilities are all between 0 and 1, and so we must have $0 \leq p(x) \leq 1$, for all $x$. Moreover, we know that the probabilities of the sample space must sum to one, and so we must have that

$$\sum_{x \in \mathcal{X}} p(x) = 1,$$

where $\mathcal{X}$ is the set of possible values that $X$ can take on.

TODO: include example regarding finding the PMF.

When solving questions related to probabilities using a probability mass function, the same secondary properties apply. Notably, if we want to know $P(X \in A)$, for some set of possible values $A$, then we can write

$$P(X \in A) = \sum_{x \in A} p(x).$$

This can be particularly helpful, for instance, if we want to know $P(X \leq c)$ for some constant value $c$. In this case we know that the possible values range from teh smallest value $X$ ccan take on, through to $c$, giving for instance,

$$P(X \leq c) = \sum_{x=0}^{c} p(x),$$

if $X \geq 0$. Rules regarding the complements of events continue to hold as well, where for instance, $P(X > c) = 1 - P(X \leq c)$, giving a useful avenue for simplifying probability calculations.

TODO: Include probability calculations.

With events defined in terms of random variables, we can talk about events as being independent of eachother or mututally exclusive using the same definitions as before. Likewise, we can talk of joint and conditional probabilities, relating to multiple events. With joint probabilities, it is often easiest to combine the event into a single, compound event and find the marginal probability of that event. For instance, if you have the events $X$ is even and $X \leq 15$, then the intersection of these events is $X \in \{2, 4, 6, 8, 10, 12, 14\}$ (supposing $X > 0$).

TODO: include example regarding conditional and joint event probabilities.

While we can discuss independence, joint probabilities, and conditional probabilities relating to events on the same random variable, it is often of interest to combine multiple random variables. Sometimes these random variables will be multiple versions coming from the same distribution, and other times they will be coming from multiple different distributions. In either event, frequently our main concern is in summarizing the probabalisitc behaviour of two or more random quantities.

Note that when we talk of a random variable `following'' a particular distribution, we are saying that the probability mass function of the random variable is described by that distribution's mass function. Thus, if two random variables`share a distribution'', we just mean that their probabilistic behaviour is described by the same underlying mass function. For instance, if I flip a coin 10 times, and you flip a different coin 10 times, and we each count the number of heads that show up, we can say that the random quantity for the number of heads I observe will have the same distribution as the random quantity for teh number of heads that you observe. These two quantities are not equal, in general, but they aredescribed by the same random processes.

When we describe the distribution of a particular random variable, we are implicitly describing the marginal probabilities associated with that quantity. Just as before, the marginal probabilities describe the behaviour of the radnom variable alone. What happens when we want to be able to describe multiple components of an experiment, together? For this we require extending the idea of a joint probability beyond the concept of events.

Suppose that we roll two six-sided fair dice. Let $X$ denote the sum of the two dice, and let $Y$ denote the maximum value showing on the two dice. $X$ is a discrete random variable taking on values between 2 and 12, while $Y$ is a discrete random variable taking on values between 1 and 6. The sample spaces for the two random variables are different from one another and so immediately we know that their probabalistic behaviour must be different, despite the fact that both random variables summarize the same statistical experiment. We can also immediately see that the two random variables, while not equal to eachother, certainly dependent on one

another: if you know that $Y = 1$ then you know that $X = 2$. If you know that $Y = 3$, then you know that $X \leq 6$.

To begin to capture the joint behaviour of $X$ and $Y$ we introduce a **joint probability mass function**. The joint probability mass function describes the behaviour of the **joint distribution**, in an otherwise analogous manner as the marginal probability mass function. That is, the joint probability mass function assigns a probability value for every pair of values that $(X, Y)$ can take on. Then, once you know the joint behaviour of $X$ and $Y$, you can fully summarize the combined behaviour of the experiment.

TODO: Include an example of a joint PMF.

In practice, joint probability mass functions can be thought of as analogous to the contingency tables we previously saw. If the first variable represents the first random variable being considered, and the second variable represents the second random variable, then each cell of the contingency table assigns a probability to one of the joint events that could be observed in the experiment. Joint distributions are a useful generalization of contingency tables as they allow us to compactly represent not only two different random variables, but sometimes many more. All of the definitions used for the case of two random variables extend naturally to three, four, and beyond.

TODO: Include example of contingency table and joint PMF.

If we continue to consider the case of a bivariate (two variable) joint distribution, we can use this setting to introduce the concept of independence of random variables. Recall that the joint probability mass function of $X$ and $Y$ is a function, $p(x, y) = P(X = x, Y = y)$. We have also introduced the marginal mass functions, $p_X(x) = P(X = x)$ and $p_Y(y) = P(Y = y)$. When dealing with events said that two events were independent if their joint probability was equal to the product of their marginal probabilities, that is $P(A, B) = P(A)P(B)$.

If we imagine taking $A = \{X = x\}$ and $B = \{Y = y\}$, then if $A \perp B$ we can write $p(x, y) = p_X(x)p_Y(y)$. If this holds for every possible $x$ and every possible $y$, then we say that $X$ and $Y$ are independent random variables, and we write $X \perp Y$. [TODO: fix this statement to use $x'$ and $y'$]

In words, two random variables are independent whenever all possible combinations of events between them are independent. When this happens we can write that

$$p(x, y) = p_X(x)p_Y(y),$$

which is to say that the joint probability mass function is the product of the two marginal probability mass functions.

TODO: iinclude example regarding independence of PMFs.

There is an equivalence between the described definition, and a slightly more intuitive definition for independence. Whenever $X \perp Y$ we can say that any two events corresponding to $X$ and $Y$, say $X \in A$ and $Y \in B$ are independent. The subtle distinction is that in our previous

definition, we were only concerned with simple events of the form $X = x'$, whereas here we allow any two arbtirary events. However, note that if we have the above definition holding for simple events, then

$$\begin{aligned} P(X \in A, Y \in B) &= \sum_{x \in A} P(X = x, Y \in B) \\ &= \sum_{x \in A} \sum_{y \in B} P(X = x, Y = y) \\ &= \sum_{x \in A} \sum_{y \in B} p_X(x) p_Y(y) \\ &= \left( \sum_{x \in A} p_X(x) \right) \left( \sum_{y \in B} p_Y(y) \right) \\ &= P(X \in A) P(Y \in B). \end{aligned}$$

That is, even by only making the assumption for simple events, the conclusionregarding compound events follows naturally. Whenever any two random variables are known to be independent we know that any two events corrresponding to these random variables will be independent. Moreover, we can directly write down the joint probability mass function , simply by taking the product of the two marginals.

TODO: include example of independent functions.

When introducing events, we discussed how the concepts of independence and dependence could be understood more intuitively through the use of conditional probabilities. The same is true for random variables. The conditional distribution of a random variable captures the behaviour of a random variable when we have information about another. The conditional probability mass function will output the probability associated with any conditional event between multiple random variables. That is, if we wanted to characterize events of the form $X$ given $Y = y$, which is to say $P(X = x|Y = y)$, then we would use the conditional probability mass function of $X$ given $Y$,

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

This definition is analogous to the formula for conditional probabilities more generally, taking the joint over the marginal. To then determine the probability of any event (for $X$) given some information about $Y$, you simply plug in $X = x$ and $Y = y$ into the conditional probability mass function. If you want to condition on more than one random variable, the quantities extend in exactly the same way, where for instance

$$P(X = x|Y = y, Z = z) = \frac{P(X = x, Y = y, z = z)}{P(Y = y, Z = z)}.$$

TODO: include example of condition al pmf

As was discussed, the joint probability mass function of two random variables is given by the product of the margianls whenever those variables are independent. If we take $X \perp Y$, then plugging $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ in to the expression for the conditional probability mass function gives $p_{X|Y}(x|y) = p_X(x)$. That is, whenever two variables are independent, the conditional probability mass function is exactly equal to the marginal probability mass function. We saw that this was true for events, and the same reasoning applies here. This result gives a more intuitive method for interpretting the independence of random variables. Two random variables are independent whenever any information about the one does not provide information about theother, which is to say when they are completely uninformative for one another. With this intuitive concept in mind, it becomes easier to infer when independence of random variables seems reasonable, which becomes a useful skill for manipulating probability expressions.

TODO: includeexamples for independence.

Seeing as the joint, marginal, and conditional probability mass functions are exactly analogous to the corresponding concepts when they were introduced regarding events, it is reasonable to assume that we canextend the multiplication rule, the law of total probability, and Bayes' theorem to the framework of probability functions as well. Indeed, each of these relationships continues to hold for random variables in much the way that would be exptected.

The multiplication rule simply states that $p_{X,Y}(x, y) = p_{X|Y}(x|Y)p_Y(y) = p_{Y|X}(y|x)p_X(x)$. This can be seen by rearranging the relationship defining the conditional probabilities. Bayes' Theorem also can be extended in nearly an identical fashion giving

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}.$$

These rules give the ability to compute the joint distribution and the other conditional information, when we have information regarding some of the marginals andsome of the conditionals. These properties are used less explicitly when dealing with probability mass functions directly, instead becoming absorbed into the fabric of thedefining relationships themselves. That is to say, you are less likely to see Bayes' theorem invoked directly when moving ebtween conditional distributions, however, moving between conditional distributions is an important skill which is very often required.

TODO: include examples for multiplication rule and Bayes' theorem

Unlike the multiplication rule and Bayes' theorem, the extension of the law of total probability is frequently cited when manipulating probability mass functions. It is a process which is important enough so as to warrant its own name: **marginalization**. The idea with marginalization is that we are going to take a joint probability mass function and **marginalize** it, turning it into a marginal distribution. This is analogous to the law of total probability. Note that when dealing with a random variable, $Y$, there is some set of numeric values that $Y$ can take on. Suppose that wwe refer to these values as $\mathcal{Y}$. A natural partition of the space is to

then take each possible value for $Y$ as the event, and simply enumerate through the elements of $\mathcal{Y}$.

Using this partition, we can ask about the ways that $X$ can take on any particular value. In order for $X$ to be $x$, we know that one of the $Y = y'$ for some $y' \in \mathcal{Y}$ must have ocurred. Thus, if we add up all the possible combinations, $(X = x, Y = y_1)$, $(X = x, Y = y_2)$, and so forth, then we will have covered every possible way of making $X = x$. This is the exact same process we used when looking at contigency tables, where we summed a row or column to get the marginal probabilities.

TODO: include partition graphic

Taking this argument and encoding it with mathematical notation we get that

$$P(X = x) = \sum_{y \in \mathcal{Y}} P(X = x, Y = y) \implies p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y).$$

That is, the process of marginalization involves summing over one of the random variables in a joint distribution function, leaving behind only the marginal. This is often a very effective way of determining a marginal distribution when information about two random variables is easy to disceren than information about only one.

To complete the analogy to the law of total probability, recall that the multiplication rule tells us that $p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y)$, and so we may alsomarginalize by taking

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x|y)p_Y(y).$$

This makes it clear that marginalization is often accomplished via arguments based on conditioning.

When confronted with questions from statistics and probability, it will often be the case that the natural answer to the question is `it depends.'' For instance, if asked`what si the probability that a student passes their next exam?'' the likely response is `it depends.'' One very useful technique for solving these questions in a satisfactory way is to continue that line of thought and explicitly specify`on what.'' For the students, for instance, you may say `it depends on how much they study.'' The conceit in this situation is that, if you were to know how much the student has studied, then you would better understand the outcomes of the student's test. In our mathematical terms this means you have a firm belief about the conditional distribution between the random quantities of exam performance and study time. The process of marginalization, and the law of total probability before that, provide useful ways of being able to translate these`it depends'' statements into concrete beliefs about the marginal probabilities. Recall that marginal distributions are distributions which do not depend on any otherquantitity, and instead, thy capture the overall behaviour. They have, in some sense, averaged out all other factors and give you the beliefs which do not depend on anything else at

all. The technique for accomplishing this is marginalization, and other forms of conditioning arguments.

TODO: include worked example on conditioning.

## 4.1 Independent and Identically Distributed: A Framework for Interpretation

A very common assumption when addressing questions in statistics and in probability is that we have a set of random variables which are independent and identically distributed (iid). We now have the tools to understand concretely what this means. Specifically, a set of random variables, $X_1, X_2, \ldots, X_n$ are said to be independent and identically distributted (denoted iid almost all the time) whenever (i) every subset of random variables in the collection is independent of every other subset of random variables in the collection, and (ii) the marginal distribution for each of the random variables are exactly the same. The assumption of iid random quantities will often come up when we are repeating a process many times over, and thinking about what observations will arise from this. Suppose that $X_1$ is a random variable that takes the value 1 if a flipped coin comes up heads, and 0 otherwise. If we imagine flipping this coin 100 times then it is reasonable to assume that each sequential coin flip will be independent of eacch other coin flip, since the result on one flip of a coin should not influence the result of any other flip of a coin. Moreover, every time the coin is flipped, it is reasonable to assume that the probability it shows up heads remains the same. As a result, these 100 random quantities, $X_1, X-2, \ldots, X_{100}$ can be said to be iid.

TODO: include other iid examples

While we will use the assumption of iid random variables later, they also provide an intuitive method for interpretting probability functions and distributions. Suppose that we wwere to take a distribution function, $p_X(x)$. If we were able to generate independent and identically distributed realizations from this probability mass ufnction, then the function $p_X(x)$ describes the behaviour for these repeated realizations. Specifically, $p(x)$ will give the long-run proportion of realizations of the iid random variables which take the value $x$.

TODO: Include interpretation statement

This type of statement is always the flavour of interpretation statements that are made with respect to probability and statistics. It will always be the case that, in order to understand what is meant specifically by a statement of probability will involve the repetition of some statistical experiment over and over again. When we were discussing sample spaces and experiments directly, we talked about repeating the experiment over and over again. When we begin to work with random variables instead, it becomes more natural to think about the replication procedures coming through the use of independent and identically distributed random variablles.

As the study of probability foes on, we begin to need to work with random quantities in a strictly theoretical sense. In introductory level problems, we are often holding in mind very concrete examples to illustrate the procedures and concepts. In this setting it is easy enough to hold in mind the experiment of interest: for instance, we may have a random variable representing the result of a coin toss, and you can envision repeatedly tossing a coin. As the concepts become less concrete, more abstract, and harder to draw directl parallels to tangible scenarios, it becomes more and more important to rely on the interpretations rooted in a series of independent and identically distributed random variables. A large component of statistics as an area of study is making explicit the assumptions we are working with, and doing our best to ensure that these are reasonable. By interpreting probability mass functions as the proportion of independent and identically distributed random variables that take on a particular value, when we repeatedly take realizations of these random variables ad infinitum, this philosophy is made clear and explicit.

## 4.2 Expectation

Until this point in our discussions of probability we have relied upon characterizing the behaviour of a random variable via the use of probability mass functions. In some sense, a probability mass function captures all of the probabilistic behaviour of a discrete random variable. Using the mass function you are able to characterize how often, in the long-run, any particular value will be observed, and any questions associated with this. As a result, the mass function remains a critical area of focus for understanding how random quantities behave.

However, these functions need to be explored and manipulated in order for useful information to be extracted from them. They do not summarize this behaviour effectively, as they are not intended to be a summary tool, and understandably we often wish to have better numeric quantities which are able to concisely indicate components of the behaviour of a distribution. Put differently, provided with a priobability mass function it is hard to immediately answer "what do we expect to happen, with this random variable?'' despite the fact that this is a very obvious first question.

To address questions related to expectations, we turn towards the statistical concept of an **expected value**. We refer to expected values as expectations, averages, and means of a distribution, interchangeably. The idea with an expected value is that we are trying to capture, with one single number, what value is expected when we make observations from the random quantity. There are many ways one might think to describe our expectations, and it is worth exploring these concepts in some detail.

One way that we may think to define our expected value is by asking what value is the most probable. This is a question which can be directly answered using the probability mass function. The process for this would require looking at the function and determining which value for $x$ corresponds to the highest probability: this is the value that we are most likely to see. Sometimes this procedure is fairly straightforward, sometimes it is quite complicated.

No matter the complexity of the specific situation, the underlying process is the same: what value has the highest probability of being seen, and that is the most likely one.

TODO: Example of mode

As intuitive as this may seem, this is not the value that will be used as the expectedvalue generally. Instead, this quantity is referred to as the **mode**. While the mode is a useful quantity, and for some decisions will be the most pertinent summary value, there are some major issues with it as a general measure which make it less desirable. For starters, consider that our most common probability model considered until this point has been that of equally likely outcomes. Here, there is no well-defined mode (convention tends to be taking it as the set of all the most probable values). Presenting the mode is equivalent to presenting the full mass function in this setting.

While the case of equally likely outcomes is a fairly strong explanation for some issues with the mode, it need not be so dramatic to undermine its utility. It is possible for a distribution to have several modes which are quite distinct from one another, even if it's not all values. Moreover, it is quite common for the modal value to be not particularly likely itself. Consider a random variable that can take on a million different values. If all of the probabilities are approximately 0.000001 then presenting the mode as the most probable value does not translate to saying that the mode is particularly probable.

TODO: incllude example where the mode is slightly more likeely than a string of similar values.

If the mode has these shortcomings, what else might work? Another intuitive concept is to try to select the "middle' ' oof thedistribution. One way to define the middle would be to select the value such that half of observations are beneath it, and half of observations are above it. That way, when you are told this value, you immediately know that it is equally likely to observe values on either side of this mark. This is also a particularly intuitive definition for expected value, and is important enough to be named: **the median**.

The median is the midpoint of a distribution, and is very important for describing the behaviour of random ariables. Medians are often the most helpful single value to report to indicate the typical behaviour of a distribution, and they are frequently used. When people interpret averages, in general, it is often the median that they are actually interpretting. It is very intuitive to be given a value and know that it is the middle of all the possible values for a distribution.

TODO: include examples on medians

Despite the advantages of medians, they have their own drawbacks as well. For starters, the median can be exceptionally challenging to compute in certain settings. As a result, even when a median is appropriate, it may not be desirable if it is too challenging to determine. Beyond the difficulties in computation, medians have some properties which may be undesirable, depending on the specific use case. One concern which arises frequently is that medians are not translated to totals, which can make them challenging in certain use cases.

Suppose that you are a store and you know that your median quantity of items sold in a day is 50 and the median cost of these items is \$10, you cannot simply multiply the 50 and the \$10 to suggest that your median revenue in a day is \$500. Doing this ttype of unit conversion or basic arithmetic with medians can be challenging, and as a result they are not always the most useful when reporting values that are going to be interpretted as rates.

TODO: Expand on the above example.

Beyond the basic manipulation medians have a feature which is simultaneously a major benefit in some settings, and a major fallback in others. Specifically, medians are less influenced by extreme values in the probability distribution. Consider two different distributions: one of them is equally likely to take any value between 1 and 10, where the other is equally likely to take any value between 1 and 9 or $1,000,000$. In both of these settings, we can take the median to be 5.5 since half of the probability mass falls above 5.5 and half falls below it.

The median, in some sense, ignores the extreme value in the probability distribution and remains stable throughout it. In certain settings, this can be very desirable. For instance, in the distribution of household incomes, the median may be an appropriate measure seeing as there are a few families who have very extreme incomes which otherwise distort the picture provided by most families. In this sense, the median's robustness to extreme values is a positive feature of it in terms of a summary measure for distributional behaviour.

Suppose instead that you work for an insurance company and are concerned with understanding the value of insurance claims that your company will need to pay out. Thedistribution will look quite similar to the income distribution: most of the probability will be assigned to fairly small claims, with a small chance of a very large one. As an insurance company, if you use the median this large claim behaviour will be smoothed over, perhaps leaving you unprepared for the possibility of extremely large payouts. In this setting, the extreme values are informative and important, and as a result the median's robustness becomes a hindrance to correctly describing the behaviour.

TODO: Another median example?

Between the median and the mode we have two measures which capture some sense of expected value, each with their own set of strengths and drawbacks. Neithercapture what it is that is referred to as *the* expected value. For this, we need to take inspiration from the median, and consider another way that we may think to find the center of the distribution.

If the median gives the middle reading along the values sequentially, we may also wish to think about trying to find the "center of gravity'' of the numbers. Suppose you take a pen, or marker, or small box of chocolates, and you wish to balance this object on a finger or an arm. To do so, you do not place the item so that half of it sits on one side of the appendage and half on the other: you adjust the location so that half of the masssits on either side of the appendage.

Throughout our probability discussions, we have always referred to probability as mass itself. We use the probability mass function to generate our probability values. This metaphor can

be extneded when we try to find the center of the distribution. If we imagine placing a mass with weight equal to the probability mass functions value at each value that a random variable can take on, we may ask: where would we have to place a fulcrum to have this number line be balanced? The answer to this question serves as another possible measure of center.

It turns out that this notion of center is the one that we are all most familiar with: the simple average. And this simple average is also the conception of expectation which gets bestowed with the name "expected value''. Mathematically, the expeccted value is desirable for many reasons, some of which we will study in more depth later on. One of these desirable features, which stands in contrast with the median, is the comparative ease with which expected values can be computed. For a random variable, $X$, we write the expected value of $X$ as $E[X]$, and assume that $X$ takes values in $\mathcal{X}$ with a probability mass function $p_X(x)$, we get

$$E[X] = \sum_{x \in \mathcal{X}} x p_X(x).$$

TODO: Example compute simple expected value.

In the case of an equally likely probability model, the expected value becomes the standard average that is widely used. Suppose that there are $n$ options in the sample space, denoted $x_1, \ldots, x_n$, then we can write

$$E[X] = \sum_{i=1}^{n} x_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

When the probability models are more complex, the formula is not precisely the stnadard average - instead, it becomes a weighted average.

TODO: Add basic average example.

While less commonly applied than the simple average, a weighted average is familiar to most students for a crucial purpose: grade calculations. If you view the weight of each item in a course as a probability mass, and the grade you scored as the value, then your final grade in the course is exactly the expected value of this distribution. The frequency with which expectedvalues are used make them attractive as a quick summary for the center of a distribution.

TODO: Include pricing calculation showing mean versus median.

While the mean provides a useful, intuitive measure of center of the distribution, it is perhaps counter intuitive to name it the expected value. To understand the naming convention it is easiest to consider the application which has likely spurred more development of statistics and probability than any other: gambling.

Suppose that there is some game of chance that can pay out different amounts with different probabilities. A critical question for a gambler in deciding whether or not to play such a game is "how much can I expect to earn, if I play?'' This is crucial to understanding, for instance,

how much you should be willing to pay to participate, or if you are the one running the game, how much you should charge to ensure that you make a profit.

If you want to understand what you expect to earn, the intuitive way of accomplishing this is to weight each possible outcome by how likely it is to occur. This is exactly the expected value formula that has been provided, and so the expected value can be thought of as the expected payout of a game of chance where the outcomes are payouts corresponding to each probability.

TODO: Include expected payout calculation.

This also represents the cost at which a rational actor should be willing to pay to participate. If a game ofchance costs more than the expected value to play, in the long run you will lose money. If a game of chance costs less than the expected value, in the long run you will earn money. It is hard to overstate the utility of gambling in developing probabiility theory, and as such these types of connections are expected.

To interpret the expected value of a random variable, one possibility is using the intuition that we used to derive the result. Notably, the expected value is the center of mass of the distribution, where the masses correspond to probabilities. This means that it is not necessarily an actual central number over the range, but rather that it sits in the weighted middle. While this interpretation is useful in many situations, there are times where the point of balance is a less intuitive description. For these, it can sometimes be useful to frame the expected value as the long-term simple average from the distribution.

If we imagine observing many independent and identically distributed random variables, then as the number of samples tends to infinity, the expected value of $X$ and the simple average will begin to coincide with one another. That is the distance between $E[X]$ and $\frac{1}{n}\sum_{i=1}^{n} X_i$ will shrink to 0. As a result, we can view the expected value as the average over repeated experiments. This interpretation coincides nicely with the description based on games of chance. Specifically, if you were to repeatedly play the same game of chance, the average payout per game will be equal to the expected value, if you play for long enough.

TODO: Include convergence graphic.

Somtimes the value of a random variablle needs to be mapped through a function to give the value which is most relevant to us. Consider, for instance, a situation wherein the side lengths of boxes being manufactured by a specific supplier are random, due to incorrectly calibrated tolerances in the machines. The resulting boxes are cubes, but what is of more interest is the volume of the produced box, not the side length. If a box has side length $x$, then its volume will be $x^3$, and so we may desire some way of computing $E[X^3]$ rather than $E[X]$.

In general, for some function $g(X)$, we may want to compute $E[g(X)]$. It is important to recognize that, generally speaking, $E[g(X)] \neq g(E[X])$. This is a common mistake, and an attractive one, but a mistake nonetheless. If we are unable to simply apply the function to the epxected value, then the question of how to compute the expected value remains. Instead of

applying the function to overall expected value, instead, we simply apply the function to each value in the defining relationship for the expected value. That is,

$$E[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x).$$

This is sometimes referred to as the "law of the unconscious statistician,'' a name which may be aggressive enough to help remember the correct way to compute the expectation.

TODO: Move the next thing up. Where the median demonstrated robustness against extreme values in the distribution, the mean (or expected value) does not. For instance, if we consider the distribution of incomes across a particular region, the mean will be much higher than the median, as those families with exceptionally high incomes will not be smoothed over as they were with medians. In this case, the lack of robustness for the expected value will render the mean a less representative summary for the true behaviour of the random quantity.

To see this concretely, consider the difference between a random variable which with equal probability takes a value between 1 and 10. This will have $E[X] = 5.5$. Now, if the 10 is made to be $1,000,000$, the expected value will now be $E[X] = 100,004.5$. This is a far cry from the median which does not change from 5.5 in either case. This lack of robustness is desirable in the event of the insurance example from the median discussion, but will be less desirable in other settings.

The mean, median, and mode are the three standard measures of central tendency. They are also referred to as measures of location, and in general, are single values which describe the standard behaviour of a random quantity. Each of the three has merits as a measure, and each has drawbacks for certain settings. The question of which to use and when depends primarily on the question of interest underconsideration, rather than on features of the data alone. Often, presenting more than one measure can give a better sense of the distributional behaviour that any one individual will.

TODO: Include example of choosing measures.

Despite the utility of all three measures, the expected value holds a place of more central importance in probability and statistics. A lot of this has to do with further mathematical properties of the mean. Because of its central role, it is worth studying the expected value in some more depth. % END OF MOVING SECTION.

TODO: include example using LOTUS.

These functions applied to random variables are often thought of as "transformations'' of the random quantities. For instance, we *transformed* a side length into a volume. While the law of the unconscious statistician will apply to any transformation for a random variable, we can sometimes use shortcuts to circumvent its application. In particular, when $g(X) = aX + b$, for

constant numbers $a$ and $b$, we can greatly simplify the expected value of the transformation. To see this note

$$\begin{aligned} E[aX+b] &= \sum_{x\in\mathcal{X}} (ax+b)p_X(x) \\ &= \sum_{x\in\mathcal{X}} axp_X(x) + bp_X(x) \\ &= a\sum_{x\in\mathcal{X}} xp_X(x) + b\sum_{x\in\mathcal{X}} p_X(X) \\ &= aE[X]+b. \end{aligned}$$

That is, in general, we have that $E[aX+b] = aE[X]+b$.

This is particularly useful as linear transformations like $aX+b$ arise very commonly. For instance, most unit conversions are simple linear combinations. If a random quantity is measured in one unit then this result can be used to quickly convert expectations to another.

TODO: include example of temperature or weight conversion.

This type of linear transformation also frequently comes up with games of chance and payouts, or with scoring more generally. For instance, suppose you are betting a certain amount on the results of a coin toss, or that you are taking a multiple choice test that gives 2 points for a correct answer.

Measures of central tendency are important to summarize the beahviour of a random quantity. Whether using the mean, median, or mode, these measures of location describe, on average, what to expect from observations of the random quantity. However, understanding a distribution requires understanding far more than simply the measures of location. As was discussed previously, the probability mass function captures the complete probabilistic behaviourof a discrete random variable, it is only intuitive that some information would be lost with a single numeric summary.

TODO: Example with equivalent mean, median, and mode.

A key characteristic of the behaviour of a random variable which is not captured by the measures of location is the variability of the quantity. If we imagine taking repeated realizations of a random variable, the variability of the random variable captures how much movement there will be observation to observation. If a random variable has low variability, we expect that the various observations will cluster together, becoming not too distant from one another. If a random variable has high variability, we expect the observations to jump around each time.

Just as was the case with measures of location, there are several measures of variability which may be applicable in any given setting. One fairly basic measure of the spread of a random variable is simply the range of possible values: whati s the highest possible value, what is the lowest possible value, and how much distance is there between those two points? This is a fairly intuitive notion, and is particularly useful in the equal probability model over a sequence

of numbers. Consider, for instance, dice. dice are typically defined by the range of values that they occupy, say 1 to 6, or 1 to 20. Once you know the values present on any die, you have a sense for how much the values can move observation to observation.

TODO: Include example for the range.

While the range is an important measure to consider to determine the behaviour of a random variable, it is a fairly crude measurement. It may be the case that, while the extreme values are possible, they are sufficiently unlikely so as to come up very infrequently and not remain representative of thelikely spread of observations. Alternative, many random variables have a theoretically infinite range. In these cases, providing the range will likely not provide much utility.

TODO: Include example.

To rememdy these two issues, we can think of some techniques for modifying the range. Instead of taking the start and end points to be the lowest and highest values, we can instead consider ranges of values which remain more plausible. A common way to do this is to extend our concept of a median beyond the half-way point. The median of a random variable $X$, is the value, $m$, such that $P(X \leq m) = 0.5$. While there is good reason to care about the midpoint, we can think of generalizing this to be *any* probability.

That is, we could find a number $z$, such that $P(X \leq z) = 0.1$. We could then use this value to conclude that 10% of observations are below $z$, and 90% of observations are above $z$. (TODO: Change this to be "probability of observation' '). These values are referred to, generally, as percentiles and they are the natural extension of medians. We will typically denote the $100p$th percentile as $\zeta(p)$, which is the value $P(X \leq \zeta(p)) = p$. Thus, the median of a distribution is $\zeta(0.5)$.

TODO: Include examples

We can leverage percentiles to remedy some of the issues with the range as a measure of variability. Framed in terms of percentiles, the minimum value is $\zeta(0)$, and the maximum value is $\zeta(1)$. Instead of considering theextreme endpoints, if we consider the difference between more moderate percentiles, we can overcome the major concenrs outlined with the range. The most common choices would be to take $\zeta(0.25)$ and $\zeta(0.75)$; these are referred to as the first and third quartiles, respectively. They are named as, taking $\zeta(0.25)$, $\zeta(0.5)$ and $\zeta(0.75)$, the distribution is cut into quarters.

TODO: Include examples.

With the first and third quartiles computed, we can compute the interquartile range, which is given by $\zeta(0.75) - \zeta(0.25)$. Typically, we denote the interquarite range simply as IQR, and like the overall range, it gives a measure of how much spread there tends to be in the data. Unlike the range, however, we can be more certain that both the first and third quartiles are reasonable values around which repeated observations of the random variable would be observed. Specifically, there is a oribability of 0.5 that a value between the first and third

quartile will be observed. The larger the IQR, the more spread out these moderate observations will be, and as a result, the more variable the distribution is.

TODO: Write examples.

Both the range and the interquartile range give a sense of the variation in the distribution irrespective of the measures of location for that distribution. Another plausible method for assessing the variability of a distribution is to assess how far we expect observations to be from the cneter. Intuitively, if observations of $X$ are near the center with high probability, then the distribution will be less variable than if the averagedistance to the center is larger.

This intuitive measure of variability is useful for capturing the behaviour of a random variable, particularly when paired iwth a measure of location. However, we do have to be careful: not all measures of dispersion based on this notion will be useful. Consider the most basic possibility, to consider $X - E[X]$. We might ask, for instance, what is the expected value of this quantity. If we take $E[X - E[X]]$ then note that this a linear combination in expectation since $E[X]$ is just some number. Thus, $E[X - E[X]] = E[X] - E[X] = 0$. In other words, the expected difference between a random variable and its mean is exactly 0. We thus need to think harder about how best to turn this intuition into a useful measure of spread as the first idea will result in 0 for all random quantities.

The issue with this procedure is that some realizations are going to be below the mean, making the difference negative, and some will be above the mean, making the difference positive. Our defining relationship for the mean relied on balancing these two sets of mass. However, when discussing the variability of the random variable, we do not much care whether the observations are lower than expected or higher than expected, we simply care how much variability there is around what is expected. To remedy this, we should consider only the distance between the observation and the expectation, not the sign. That is, if $X$ is 5 below $E[X]$ we should treat that the same as if $X$ is 5 above $E[X]$.

There are two common ways to turn value into its magnitude in mathematics generally: squaring the number and using absolute values. Both of these tactics are useful approaches to defining measures of spread, and they result in the **variance** when using the expected value of the squared deviations, and the **mean absolute deviation** when using the absolute value. While $E[|X - E[X]|]$ is perhaps the more intuitive quantity to consider, generally speaking it will not be the one that we use.

In general when we need a positive quantity in mathematics it will typically be preferable to consider the square to the absolute value. The reasons for this are plentiful, but generally squares are easier to handle than absolute values, and as a result become more natural quantities to handle. The variance is the central measure of deviation for random variables, so much so that we give it its own notation,

$$\text{var}(X) = E[(X - E[X])^2].$$

Note that if we take $g(X) = (X - E[X])^2$, then the variance of $X$ is the expected value of a transformation. We have seen that to compute these we apply the law of the unconscious statistician, and substitute $g(X)$ into the defining relationship for the expected value, which for the variance gives

$$\text{var}(X) = \sum_{x \in \in \mathcal{X}} (x - E[X])^2 p_X(x).$$

Prior to computing the variance, we must first work out the mean as the function $g(X)$ relies upon this value.

TODO: Include example for calculating variance.

The higher that an individual random variables variance is, the more spread we expect there to be in repeatedly realizations of that quantity. Specifically, the more spread out around the mean value the random variable will be. A random variable with a low variance will concentrate more around the mean value than one with a higher variance. One confusing part of the variance of a random variable is in trying to assess the units. Suppose that a random quantity is measured in a particular set of units - dollars, seconds, grams, or similar. In this case, our interpretations of measures of location will all be in the same units, which aids in drawing connections to the underlying phenomenon that we are trying to study. However, because the variance is squared, we cannot make the same extensions to it: variance is not measured in the regular units, but in the regular units squared.

Suppose you have a random time being measured, perhaps the reaction time for some treatment to take effect in a treated patient. Finding the mean or median will give you a result that you can read off in seconds as well. The range and interquartile range both give you the spread in seconds. However, if you work out the variance of this quantity it will be measured in seconds squared - a unit that is challenging to have much intuition about. To remedy this we will often use a transformed version of the variance, called the **standard deviation**, returning the units to be only the original scale. The standard deviation of a random variable is simply given by the square root of the variance, which is to say

$$\text{SD}(X) = \sqrt{\text{var}(X)}.$$

We do not often consider computing the standard deviation directly, and so will most commonly refer to the variance when discussing the behaviour of a random variable, but it is important to be able to move seamlessly between these two measures of spread.

TODO: Standard deviation.

When computing the variance of a random qauntity, we often use a shortcut for the formula.

Consider

$$\begin{aligned}
\text{var}(X) &= \sum_{x \in \mathcal{X}} (x - E[X])^2 p_X(x) \\
&= \sum_{x \in \mathcal{X}} (x^2 - 2xE[X] + E[X]^2) p_X(x) \\
&= \sum_{x \in \mathcal{X}} x^2 p_X(x) - 2E[X] \sum_{x \in \mathcal{X}} x p_X(x) + E[X^2] \sum_{x \in \mathcal{X}} p_X(x) \\
&= E[X^2] - 2E[X]E[X] + E[X]^2 \\
&= E[X^2] - E[X]^2.
\end{aligned}$$

This result gives us the identity that the variance of $X$ can be found via $E[X^2] - E[X]^2$. Generally, this is moderately more straightforward to calculate since $X^2$ is an easier transformation than $(X - E[X])^2$. This identity will come back time and time again, with a lot of versatility in the ways that it can be used. Typically, when a variance is needed to be calculated the process is to simply compute $E[X]$ and $E[X^2]$, and then apply this relationship.

TODO: include variance calculation example.

With expectations, we saw that $E[g(X)]$ needed to be directly computed from the definition. The same is true for variances of transformations. Specifically, $\text{var}(g(X))$ is given by $E[(g(X) - E[g(X)])^2]$ which can be simplified with the previous relationship as $E[g(X)^2] - E[g(X)]^2$. Just as with expectations, it is important to realize that $\text{var}(g(X)) \neq g(\text{var}(X))$, and so dealing with transformations requires further work.

TODO: Transformation example. TODO: Move the following discussion up. Beyond being linear over simple transformations, summations in general behave nicely with expectations. Specifically, for any quantities separated by addition, say $g(X) + h(X)$, the expected value will be the sum of each expected value. Formally,

$$\begin{aligned}
E[g(X) + h(X)] &= \sum_{x \in \mathcal{X}} (g(X) + h(X)) p_X(x) \\
&= \sum_{x \in \mathcal{X}} g(X) p_X(x) + h(x) p_X(x) \qquad = \sum_{x \in \mathcal{X}} g(x) p_X(x) + \sum_{x \in \mathcal{X}} h(x) p_X(x) \\
&= E[g(X)] + E[h(X)].
\end{aligned}$$

Behaving well under linearity is one of the very nice properties of expectations. It will come in useful when dealing with a large variety of important quantities, and as we will see shortly, this linearity will also extend to multiple different random quantities.

TODO: add example of linearity. %End section to move.

With expectations, we highlighted linear transformations as a special case, with $g(X) = aX + b$. For the variance, the linear transformations are also worth distinguishing from others. To this

end, we can apply the standard identity for the variance, giving

$$\begin{aligned} E[(aX+b)^2] &= E[a^2X^2 + 2abX + b^2] \\ &= E[a^2X^2] + E[2abX] + E[b^2] \\ &= a^2E[X^2] + 2abE[X] + b^2. \end{aligned}$$

TODO:Move upwards Note that part of the property of the linearity of expectation that we can immediately see if that the expected valuye of any constant is always that constant. If we take $a = 0$, then we see that $E[aX + b] = E[b] = b$. Thus, any time that we need to take the expected value of any constant number, we know that it is just that number. %End upwards movement

Next, we note that $E[aX + b] = aE[X] + b$ and so

$$\begin{aligned} E[aX+b]^2 &= (aE[X] + b)^2 \\ &= a^2E[X]^2 + 2abE[X] + b^2. \end{aligned}$$

Differencing these two quantities gives

$$a^2E[X^2] + 2abE[X] + b^2 - a^2E[X]^2 - 2abE[X] - b^2 = a^2(E[X^2] - E[X]^2).$$

By noting that $E[X^2] - E[X]^2$, we can complete the statement that

$$\text{var}(aX + b) = a^2\text{var}(X).$$

Thus, when applying a linear transformation, only the multiplicative constant matters , and it transforms the varaince by a squared factor. This should make some intuitive sense that the additive constant does not change anything. If we consider that variance is a measure of spread, adding a constant value to our random quantity will not make it more or less spread out, it will simply shift where the spread is located. This is not true of the mean, which measures where the center of the distribution is, which helps explain why the result identities are different.

TODO: Example using this.

In the same way that the linearit of expectation demonstrates that the expected value of any constant is that constant, we can use this identity to show that the variance of constant is zero. However, we can also reason to this based on our definitions so far. Suppose that we have a random variable which is constant. This seems to be an oxymoron, but it is perfectly well defined. A constant $b$ can be seen as a random variable with probability distribution $p_X(x) = 1$ if $x = b$ and $p_X(x) = 0$ otherwise. In this case, the expected value is going to be $E[X] = 1(b) = b$, and $E[X^2] = 1(b)^2 = b^2$. As a result, we see that $E[b] = b$, as previously stated, and $\text{var}(b) = E[X^2] - E[X]^2 = b^2 - b^2 = 0$. From an intuitive perspective, there is no variation around the mean of a constant: it is always the same value. As a result, when taking the variance, we know that it should be 0.

Unlike the expecctation, the variance of additive terms will not generally be the addition of the varainces themselves. That is, we cannot say that $\text{var}(g(X) + h(X)) = \text{var}(g(X)) + \text{var}(h(X))$, as a general rule. Writing out the definition shows issue with this:

$$E[(g(X) + h(X))^2] = E[g(X)^2] + 2E[g(X)h(X)] + E[h(X)^2].$$

The first and third terms here are nicely separated and behave well. However, the central term is not going to be easy to simplify, in general. You can view $g(X)h(X)$ as a function itself, and so $E[g(X)h(X)] \neq E[g(X)]E[h(X)]$,$ in general. Instead, this will typically need to be worked out for any specific set of functions.

## 4.3 Conditional and Joint Expectations and Variances

Up until this point we have considered the marginal probability distribution when exploring the measures of central tendency and spread. These help to summarize the marginal behaviour of a random quantity, cpaturing the distribution of, for instance, $X$ alone. When introducing distributions, we also made a point to introduce the conditional distribution as one which is particularly relevantwhen there is extra information. The question "what do wwe expect to happen, given that we have an additional piece of information?'' is not only well-defined, but it is an incredibly common type of question to ask. To answer it, we require **conditional expectations**.

TODO: Include set of questions relating to conditional expectation.

In principle, a conditional expectation is no more challenging to calculate than a marginal expectation. Suppose we want to know teh expected value of $X$ assuming that we know that a second random quantity, $Y$ has taken on the value $y$. We write this as $E[X|Y = y]$, and all we do is replace $p_X(x)$ with $p_{X|Y}(x|y)$ in the defining relationship. That is

$$E[X|Y = y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x|y).$$

In a sense, we can think of the conditional distribution of $X|Y = y$ as simply being a distribution itself, and then work with that no differently. The conditional variance, which we denote $\text{var}(X|Y = y)$ is also exactly the same.

TODO: Include an example.

Above we supposed that we knew that $Y = y$. However, sometimes we want to work with the conditional distribution more generally. That is, we want to investigate the behaviour of $X|Y$, without yet knowing what $Y$ equals. We can use the same procedure as above, however, this time we leave $Y$ unspecified. We denote this as $E[X|Y]$, and this expression will be (in general) a function of $Y$. Then, whenever a value for $Y$ is observed, we can simply specify $Y = y$, deriving the specific value. In practice, we will typically compute $E[X|Y]$ rather than $E[X|Y = y]$, since once we have $E[X|Y]$ we can easily find $E[X|Y = y]$ for *every* value of $y$.

TODO: Show example.

Since $E[X|Y]$ is a function of an unknown random quantity, $Y$, $E[X|Y]$ is also a random variable. It is a transformation of $Y$, and as such, it will have some distribution, some expectation, and some varaince itself. This is often a confusing concept when it is first introduced, so to recap: $X$ and $Y$ are both random variables; \$E[X] and $E[Y]$ are both constant, numerical values describing the distribution of $X$ and $Y$; $E[X|Y = y]$ and $E[Y|X = x]$ are each numeric constants which summarize the distribution of $X|Y = y$ and $Y|X = x$ respectively; $E[X|Y]$ and $E[Y|X]$ are functions of $Y$ and $X$, respectively, and can as such be seen as transformations of (and random quantities depending on) $Y$ and $X$ respectively.

We do not often think of the distribution of $E[X|Y]$ directly, however, there is a very useful result about both its expected value and its variance, which will commonly be exploited. Specifically, if we take the expected value of $E[X|Y]$ we will find that $E[E[X|Y]] = E[X]$. Note that since $E[X|Y] = g(Y)$ for some transformation, $g$, the outer expectation is taken with respect to the distribution of $Y$. Sometimes when this may get confusing we will use notation to emphasize this fact, specifically, $E_Y[E_{X|Y}[X|Y]] = E_X[X]$. This notation is not necessary, but it can clarify when there is much going on, and is a useful technique to fallback on.

$$\begin{aligned}
E_Y[E[X|Y]] &= \sum_{y \in \mathcal{Y}} E[X|Y] p_Y(y) \\
&= \sum_{y \in \mathcal{Y}} \left( \sum_{x \in \mathcal{X}} x p_{X|Y}(x|Y) \right) p_Y(y) \\
&= \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} x \frac{p_{X,Y}(x,y)}{p_Y(y)} p_Y(y) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x p_{X,Y}(x,y) \\
&= \sum_{x \in \mathcal{X}} x p_X(x) \\
&= E[X].
\end{aligned}$$

TODO: Include example.

This property, that $E[E[X|Y]] = E[X]$ is important enough that it receives its own name: **the law of total expectation**. In the same way that it is sometimes easier to first condition on $Y$ in order to compute the marginal distribution of $X$ via applications of the law of total probability, so too can it be easier to first work out conditional expectations, and then take the expected value of the resulting expression. This adds on to the so-called "conditioning arguments'' that were discussed previously, allowing a technique to work out the marginal mean indirectly.

TODO: Show example use case of LOTE.

While the conditional expectation is used quite prominantly, the conditional variance is less central to the study of random variables. As discussed, briefly, the conditional variance is given by the same variance relationship, replacing the marginal probability distribution with the conditional one (just as with expectations). Just as with expectations, $\text{var}(X|Y=y)$ is a numeric quantity given by $E[(X-E[X|Y=y])^2|Y=y]$ and $\text{var}(X|Y)$ is a random variable given by $E[(X-E[X|Y])^2|Y]$. This means that when working with the general, $\text{var}(X|Y)$, we can also consider taking expectations of the resulting transformation.

TODO: Include examples.

A final result relating to conditional expectations and varainces connects the two concepts. This is known as **the law of total variance**. For any random variables $X$ and $Y$, we can write

$$\text{var}(X) = E[\text{var}(X|Y)] + \text{var}(E[X|Y]).$$

This result can be viewed as decomposing thevariance of a random quantity into two separate components, and comes up again in later statistics courses. At this point we can view this as a method for connecting the marginal distribution through the conditional variance nad expectation.

TODO: Examples of this.

The final set of techniques to consider for now relate to making use of the joint distribution between $X$ and $Y$. Specifically, if we have any function of two random variables, say $g(X,Y)$ and we wish to find $E[g(X,Y)]$. This follows all of the expected derivations that we have used so far, this time replacing the marginal with the joint distribution. That is,

$$E[g(X,Y)] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g(x,y)p_{X,Y}(x,y).$$

For instance, if we want to consider the product of two random variables, we could use this technique to determine $E[XY]$. The variance extends in the same manner as well.

TODO: Include example.

This defining relationship allows us to work out the expected value of a linear combination of two random variables. That is

$$\begin{aligned}
E[X+Y] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(x+y)p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}} x\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) + \sum_{y\in\mathcal{Y}} y\sum_{x\in\mathcal{X}} p_{X,Y}(x,y) \\
&= \sum_{x\in\mathcal{X}} xp_X(x) + \sum_{y\in\mathcal{Y}} yp_Y(y) \\
&= E[X] + E[Y].
\end{aligned}$$

The same property does not apply with varainces, at least not in general. To see this, consider that

$$\begin{aligned} E[(X+Y-E[X]-E[Y])^2] &= E[((X-E[X])+(Y-E[Y]))^2] \\ &= E[(X-E[X])^2]+E[(Y-E[Y])^2]+2E[(X-E[X])(Y-E[Y])] \\ &= \text{var}(X)+\text{var}(Y)+2E[(X-E[X])(Y-E[Y])]. \end{aligned}$$

The term that impedes the linear relationship, $E[(X-E[X])(Y-E[Y])]$ can be computed as any joint function can be. This quantity, however, is particularly important when considering the relationship between two random variables. This is called the **covariance** and it is a measure of the relationship between $X$ and $Y$. Typically we write $E[(X-E[X])(Y-E[Y])] = \text{cov}(X,Y)$ so that

$$\text{var}(X+Y) = \text{var}(X)+\text{var}(Y)+2\text{cov}(X,Y).$$

TODO: Include example.

The covariance behaves similarly to the varaince. We can see directly from the definition that $\text{cov}(X,X) = \text{var}(X)$. Moreover, using similar arguments to those used for the varaince, we can show that

$$\text{cov}(aX+b,cY+d) = ac\text{cov}(X,Y).$$

Covariances remain linear, so that

$$\text{cov}(X+Y,X+Y+Z) = \text{cov}(X,X)+\text{cov}(X,Y)+\text{cov}(X,Z)+\text{cov}(Y,X)+\text{cov}(Y,Y)+\text{cov}(Y,Z).$$

These make covariances somewhat nicer to deal with than variances, and on occasion it may be easier to think of variances as covariances with themselves.

TODO: Example? Maybe.

It is worth considering, briefly, the ways in which conditional and joint expectations interact. Namely, if we know that $Y = y$, then the transformation $g(X,y)$ only has one random component, which is the $X$. As a result, taking $E[g(X,Y)|Y=y] = E[g(X,y)|Y=y]$. If instead we use the conditional distribution without a specific value, we still have that $Y$ is fixed within the expression, it is just fixed to an unknown quantity. That is $E[g(X,Y)|Y]$ will be a function of $Y$. We saw before that $E[E[X|Y]] = E[X]$, and the same is true in the joint case. Thus, one technique for computing the joint expectation, $g(X,Y)$ is to first compute the conditional expectation, and then compute the marginal expectation of the resulting quantity.

TODO: example.

## 4.4 Independence in all of this

Whenever we can assume independence of random quantities, this allows us to greatly simplify teh expressions we are dealing with. Recall that the key defining relationship with independence is that $p_{X,Y}(x,y) = p_X(x)p_Y(y)$. Suppose then that we can write $g(X,Y) = g_X(X)g_Y(Y)$. For instance, for the covariance we have $g(X,Y) = (x - E[X])(Y - E[Y])$ and so $g_X(X) = X - E[X]$ and $g_Y(Y) = Y - E[Y]$. If we want to compute $E[g(X,Y)]$ then we get

$$\begin{aligned} E[g(X,Y)] &= E[g_X(X)g_Y(Y)] \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g_X(x)g_Y(y)p_{X,Y}(x,y) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g_X(x)g_Y(y)p_X(x)p_Y(y) \\ &= \sum_{x\in\mathcal{X}} g_X(x)p_X(x)\sum_{y\in\mathcal{Y}} g_Y(y)p_Y(y) \\ &= E[g_X(X)]E[g_Y(Y)]. \end{aligned}$$

Thus, whenever random variables are independent, we have the ability to separate them over their expectations.

TODO: example.

Consider what this means, in particular, for the covariance between independent random variables. If $X \perp Y$ then

$$\begin{aligned} \text{cov}(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[(X - E[X])]E[(Y - E[Y])] \\ &= (E[X] - E[X])(E[Y] - E[Y]) \\ &= 0. \end{aligned}$$

That is to say, if $X$ and $Y$ are indepdent, then $\text{cov}(X,Y) = 0$. As a result of this, for independent random variables $X$ and $Y$ we also must have that $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$. It is critical to note that this relationship does not go both ways: you are able to have $\text{cov}(X,Y) = 0$ even if $X \not\perp Y$.

TODO: Include example of independence.

While we have primarily focused on joint and conditional probabilities with two random variables, the same procedures and ideas apply with three or more as well. The relevant joint distribution, or conditional distribution would simply need to be substituted in definitions. Often the complexity here becomes a matter of keeping track of which quantities are random, and which are not. For instance, if we have $X, Y, Z$ as random variables, then $E[X|Y,Z]$ is a random function of $Y$ and $Z$. We will still have that $E[E[X|Y,Z]] = E[X]$, however, the outer

expectation is now the joint expectation with respect to $Y$ and $Z$. As a result, we can also write $E[E[X|Y,Z]|Y]$. The first expectation will be with respect to $X|Y,Z$, while the outer expectation is with respect to $Z|Y$. This becomes a useful demonstration for when making the distribution of the expectation explicit helps to clarify what is being computed. As a general rule of thumb, the innermost expectations will always have more conditioning variables than the outer ones: each time we step out, we peel back one of hte conditional variables until the outermost is either a marginal (or joint). This will help to keep things clear.

# 5 The Named Discrete Distributtions

So far, our discussion of probability distributions and their summaries has centered on general results for arbitrary probability mass functions. The basic premise has been that, by knowing a probability mass function, you are able to understand the complete behaviour of a random quantity. Directly from this mass function we are able to derive summaries for the behaviour, for instance describing the location and variability of the random variable. In short by knowing the probability mass function (and to a lesser extent, the expected value and variance), we immediately understand how the random variable behaves. We have not, however, spent much time discussing where the probability mass functions come from.

We have, indirectly, seen one fairly general probability mass function, the one deriving from the equally likely outcomes model. This probability mass function is completely defined by the set of possible values that the random variable can take on. Suppose that we restrict our attention to the sample space being a set of $k$ integers from $a$ through to $a + k$. Note that this assumption is not actually restrictive: if you have any $k$ items you can simply label each of the items one of the numbers between 1 and $k$ and take $a = 1$. When setup in this way, this distribution is often referred to as the **discrete uniform distribution**. Typically, we will use the values for the lower bound ($a$) and the upper bound ($b = a + k$) to define *which* discrete uniform we are discussing. If we say that $X$ follows a discrete uniform distribution with parameters $a$ and $b$ we are say that $X$ is a random variable which has an equal probability of taking any of the integers from $a$ to $b$. Put different, we have that

$$p_X(x) = \begin{cases} \frac{1}{b-a+1} & x \in \{a, a+1, \ldots, b\} \\ 0 & \text{otherwise.} \end{cases}$$

Knowing the probability mass function of $X$, we also immediately can work out the expectation and variance for the random variable. Doing this results in $E[X] = \frac{a+b}{2}$ and $\text{var}(X) = \frac{(b-a+1)^2-1}{12}$ (see the following derivation!). This means that just by knowing that a random variable follows a discrete uniform distribution, we immediately know (or can look up) any of the properties that we have discussed up until this point.

TODO: Include derivation.

This is a partiuclarly powerful realization. There are many real-world quantities which we immediatley know follow a discrete uniform distribution. For instance, rolling a die. In the case of a die roll we take $a = 1$, $b = 6$, and immediately understand that $E[X] = 3.5$, that

$\text{var}(X) = \frac{17}{6}$, and that the probability of each value is $\frac{1}{6}$. Of course,we could do the same calculations for any die, with any sides labeled in consecutive order.

TODO: Include examples of real-world discrete uniform settings.

Despite the fact that the discrete uniform is fairly prominent in terms of real-world utility, it is also a comparatively simple distribution. However, the main point of this discussion is not actually to introduce the discrete uniform, but rather to introduce the concept of a **named distribution**. The basic premise here is that there are processes in the world which occur frequently enough, in a diverse array of settings, with the same underlying structure ot their uncertainty. If we can study one version of these general problems, then we can extract the mass function, expectation, and varaince and we are easily able to describe the probabilistic behaviour of these quantities. At that point, understanding the uncertainty of random quantities becomes a matter of matching the processes to the correct distribution, and then applying what we know about that distribution directly. While not every process will directly correspond to a known, named distribution, we can often get very close using just a handful of these [TODO: write aside on the pareto distribution].

For each named distribution there is an underlying structure describing in what scenarios it will arise. For instance, for the discrete uniform, this is when there is a set of equally likely outcomes which can be described using consecutive integers. Once matched, there will also be a probability mass function, an expected value, and a variance associated with the distribution. Importantly, all of tehse quantities will depend on some **parameters**. In the discrete uniform we used the parameters $a$ and $b$. These parameters specify which *version* of the distribution is relevant for the underlying scenario. It is best to think of the named distributions as families of distributions, with specific iterations being dictated by the parameter values. If two processes follow the same distribution with different parameters they will not be identically distributed, they are simply drawn from the same family. If two processes have the same parameter values and the same underlying distribution, they are identically distributed and their probabilistic behaviour will be exactly the same. For instance, there is no probabalistic differencebetween rolling a fair, six-sided die or drawing a card at random from a set of 6 cards labelled 1 through 6. There may be real-world differences which matter, but from a probabilistic point of view, they are exacttly the same.

This is a useful realization as it allows the use of simple models to understand more complex phenomenon. Perhaps the best way to demonstrate the effectiveness of these simple models is to introduce perhaps the most basic named probability distribution: **the Bernoulli distribution**. The Bernoulli distribution characterizes any statistical experiment with a binary outcome when these resultsare denoted 0 and 1. The parameter that indexes the distribution is $p$, which gives the probability of observing a 1.

Whenever we want to say that $X$ follows a particular distribution, wwe use a mathematical shorthand to do so. Specifically, we write $X \sim$ Distribution(parameters) to mean "$X$ follows

the Distribution with parameters.'' For instance, if $X$ represents the results of a fair six-sided die roll, we can write $X \sim$ Discrete Uniform$(1, 6)$. We will typically shorten this to be $X \sim$ D.Unif$(1, 6)$. %TODO: Move this up.

If $X \sim \text{Bern}(p)$ (where Bern is used for Bernoulli distributions) then we know that

$$p_X(x) = \begin{cases} p^x(1-p)^{1-x} & x \in \{0, 1\} \\ 0 & \text{otherwise.} \end{cases}$$

Further, we know that $E[X] = p$ and $\text{var}(X) = p(1-p)$. We typically call $X = 1$ a `success'' and $X=0$ a`failure'' when discussing Bernoulli random variables. The most straightforward application of a Bernoulli random variable is the flip of a coin. Take $X = 1$ if a head is shown, and $X = 0$ if a tails is shown. Then $X \sim \text{Bern}(p)$.

A coin flip is, by itself, not particularly interesting. However *any* statistic experiment with binary outcomes coded this way can be seen as a Bernoulli random variable. Suppose, for instance, you are interested in whether you will pass a particular course or not. There are two options, a `success'' (passing) and a`failure'' (failing), and the chances of this are governed by some probability $p$. Suppose you want to know whether the next flight you take will land safely, it is the same situation. Or whether a particular medical treatment will effectively treat an illness. Each of these scenarios is analyzed in exactly the same way as a coin toss: the probabilities change, but the underlying functions and mathematical objects do not. There is no probabalistic difference between determining whether a coin will come up heads or whether a plane will safely land.

TODO: Examples

Thediscrete uniform and Bernoulli are both distributions which map on to the real-world, but which are quite basic and which do not demonstrate the most meaningful applications of named distributions. These ideas become far more powerful when we begin to explore furthernamed distributions.

### 5.0.1 The Binomial Distribution

A natural extension to tossing a coin once and seeing if it comes up heads or not is tossing a coin $n$ times and counting how many times it comes up heads. If we take $X$ to be the number of successes in $n$ independent and identically distributed Bernoulli trials, then we say that $X$ has a **binomial distribution**. The binomial distribution is characterized by two different parameters, the number of trials that are being performed, denoted $n$, and the probability of a success on each trial, $p$. We write $X \sim \text{Bin}(n, p)$.

TODO: Simple example for bin.

If we know that $X \sim \text{Bin}(n, p)$ then we get that

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x \in \{0, 1, \dots, n\} \\ 0 & \text{otherwise.} \end{cases}$$

This is the first distribution we have seen which knowing the underlying distribution would not have immediately translated into knowing the probability mass function, which begins to illustrate why this is a useful area of study. We can work out that $E[X] = np$ and $\text{var}(X) = np(1-p)$.

TODO: Example with binomial probabilities.

Note that the binomial can be constructed by summing iid Bernoulli variables. Specifically, if $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(p)$ (note, $\overset{iid}{\sim}$ means "independent and identically distributed according to…'') then taking $Y = \sum_{i=1}^{n} X_i$ gives a binomial with $n$ and $p$ distribution. This should be intuitive from the underlying structure: if a Bernoulli comes up 1 when we get a heads on a single flip of teh coin, then if we flip the coin $n$ times and count the number of heads this is the same as counting the number of 1's from each corresponding Bernoulli trial, or in other words, summing them. Once we know this construction, we can use the properties we have previously seen about independent random variables to work out the mean and variance for the distribution.

TODO: Aside on the workout of this.

It is important to note and remember that binomial random variables make several key assumptions. First, we are counting the number of successes in a **fixed** number of trials. In order for something to be binomially distributed, we must know in advance how many trials there are under consideration. Second, each of these trials must be independent of one another: the outcome on one cannot impact any of the others. Third, there must be a constant probability of success across all trials. If the probabilities are shifting overtime, then a binomial is no longer appropriate.

TODO: identify several examples for this.

## 5.1 Geometric Random Variables

While the binomial counted the number of successes in a fixed number of trials, we may also be interested in questions relating to how many trials would be needed to see a success. Instead of `how many heads in $n$ flips of a coin?'' we may ask`how many flips of a coin to get a head?'' Like the binomial, quantities related to counting the number of trials until a success are related deeply to Bernoulli trials, where once more we are envisioning a sequence of independent and identically distributed trials being performed. However, instead of knowing that we will stop after $n$ trials, here we will only stop once we see a particular result.

Any random quantities following this type of procedure are said to follow a **geometric distribution**. The geometric distribution is parameterized with a single parameter, $p$, the probability of success. We write $X \sim \text{Geo}(p)$, and have that

$$p_X(x) = \begin{cases} (1-p)^{x-1}p & x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, $E[X] = \frac{1}{p}$ and $\text{var}(X) = \frac{1-p}{p^2}$.

TODO: geometric calculation.

The geometric distribution differs from other named distributions that we have considered in that the random variable can take on an infinite number of possible values. The probability that $X$ exceeds a very large threshold shrinks down to 0, however, there is no maximum value that can be observed. Beyond that assumption, the key set of assumptions are the same as for the binomial: independent and identically distributed Bernoulli trials are run, and are stopped only after the first observed success.

In the framing we are using here, the random variable $X$ counts the total number of trials *including* the trail upon which the first success was reached. Sometimes you may see this distribution parameterized slightly differently, taking $X$ to instead count the number of failures before the first success. To convert between the two framings we need only subtract 1: there is no meaningful difference in the underlying behaviour.

TODO: include several geometric examples.

## 5.2 Negative Binomial

A natural way to make the geometric distribution more flexible is to not stop after the first success, but rather after a set number of successes. That is, instead of flipping a coin until we see a head, we flip a coin until we see $r$ heads. Any random quantity which follows this general pattern is said to follow a **negative binomial distribution**. We use two parameters to describe the negative binomial distribution, $r$ the number of successes we are looking to achieve, and $p$ the probability of a success on any given trial. We write $X \sim \text{NB}(r, p)$.

TODO: Basic example

If we know that $X \sim \text{NB}(r, p)$, then we immediately get

$$p_X(x) = \begin{cases} \binom{x-1}{r-1}p^r(1-p) & x \geq r \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, we have $E[X] = \frac{r}{p}$ and $\text{var}(X) = \frac{r(1-p)}{p^2}$. Setting $r = 1$, wee get the same quantities explored in the case of the geometric distribution.

TODO: Include NB Examples

Like for the geometric distribution, we have taken $x$ to represent the total number of trials considered, including the $r$ successes. (TODO: fix geometric distribution range, start at 1). There are alternative parameterizations which would count how many failures occur prior to the $r$th success, which can be viewed as the value we consider minus $r$.

TODO: Include many NB examples

## 5.3 Hypergeometric Distributions

One of the use cases demonstrated for the binomial distribution is drawing *with* replacement. In order for the binomial distribution to be relevant it must be the case that the probability of a success is unchanging, and correspondingly, if the process under consideration cosntitutes random draws from a population then these draws must be with replacement as otherwise the probabilities would shift. Suppose that we are trying to draw the ace of apdes from a standard, shuffled deck of 52 cards. If we begin drawing cards without returning them to the deck after each draw, the probability that the next draw is the ace of spades is increasing over the draws. As a result, this type of scenario does not fit into the independent and identically distributed Bernoulli trials that we have been exploring.

Instead, suppose that we are drawing **without** replacement from a finite population. The population consists of two types of items: successes and failures. If we are interested in counting how many successes we seein a set number of draws, then this random quantity will follow a **hypergeometric distribution**. The hypergeometric distribution is parameterized using three different parameters: the number of items in the population, $N$, the number of these which are considered successes, $M$, adn the total number of items that are to be drawn without replacement, $n$. We write $X \sim \text{HG}(N, M, n)$.

TODO: Basic example

If $X \sim \text{HG}(N, M, n)$ then

$$p_X(x) = \begin{cases} \frac{\binom{N-M}{n-x}\binom{M}{x}}{\binom{N}{n}} & x \in \{\max\{0, N-M+n\}, \ldots, \min\{n, M\}\} \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, $E[X] = \frac{nM}{N}$ and the variance is given by

$$\text{var}(X) = n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1}.$$

TODO: Bigger calculation.

The hypergeometric is closely linked to the binomial distribution. If we consider the population described in the hypergeometric setup then the probability of a success on the first draw is $p = \frac{M}{N}$. Note that $E[X] = np$, exactly the same as in the binomial. However, plugging this in fro the variance we get $\text{var}(X) = np(1-p)\frac{N-n}{N-1}$. Notice that if $n = 1$, this extra term is simply 1, and for $n > 1$ it will be less than 1. As a result, the variance of the hypergeometric is smaller than the variance of the corresponding binomial. This makes intuitive sense: in a hypergeometric, the fact that draws are without replacement means that as more draws go on the probability of observing a success increases, reducing the likelihood of long runs of no observed successes. There is a cap on the behaviour of the random quantity thanks to the finiteness of the population. Correspondingly, the multiplicative term by which the varaince shrinks, $\frac{N-n}{N-1}$ is referred to as the **finite population correction** and it differentiates the behaviour of the hypergeometric and the binomial.

Note that as $N$ becomes very, very large, as long as $n$ is small by comparison, the finite population correction will approach 1. in other words: drawing without replacement in a large enough sample behaves almost exactly the same as drawing with replacement in the sample. The binomial distribution can be used to approximate hypergeometric distributions, so long as the population is very large. Again, this makes sense intuitively. If you have a deck with a million cards in it, and you are going to draw 2, whether or not you return the first one to the deck has very little bearing on the probabilities associated with this scenario. Generally, the binomial distribution is easier to work with, so this approximation can be useful in some settings.

TODO: aside on this for sampling.

TODO: Many examples of HG random variables.

## 5.4 Poisson Distribution

The hypergeometric broke from the pattern in the other distributions that have been discussed by not being represented as the sequence of several Bernoulli random trials. However, it was still characterized by a sequence of repeated trials. While many statistical experiments can be framed in this way, there are of course processes which do not fit well into this framing.

Consider, for instance, any process where something is observed for a set period of times and events may or may not occur during this interval. Perhaps you sit on the side of the road and count the number of cars travelling by a particular intersection over the course of an hour. Each car going by is an event, but in this setting the number of events is the random quantity itself. None of the distributions discussed until this point are suited to this type of process.

When we have events which occur at a constant rate, and our interest is in the number of events which are occuring, then we can make use of the **Poisson distribution**. The Poisson distribution takes a single parameter, $\lambda$, which is the average rate of occurence of the events over the time period we are interest in. We write $X \sim \text{Poi}(\lambda)$.

TODO: Basic example.

If $X \sim \text{Poi}(\lambda)$ then

$$p_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, $E[X] = \lambda$ and $\text{var}(X) = \lambda$. The Poisson distribution is interesting in that the mean and variance are always equal to one another.

TODO: Bigger example

While the most common applications for the Poisson distribution have to do with the occurences of events throughout time, it is also possible to view this as the occurences of events throughout space. For instance, if there is a manufcturer producing rope, then the number of defects in a set quantity of rope is likely to follow the Poisson distribution. Similarly, in a set geographic area, the number of birds of a particular species is likely to follow a Poisson distribution. For the Poisson distribution, we are typically thinking that there is a rate at which events of interest occur, and we can use the Poisson to model the total number of occurences over some specified interval.

TODO: many examples for the use of Poisson.

## 5.5 Using Named Distributions

While many other named, discrete distributions exist, these are likely the most common. When confronted with a problem in the real-world for which you wish to understand the uncertainty associated with it, a reasonable first step is to determine whether a named distribution is well-suited to representing the underlying phenomenon. Is it a situation with enumerated events which are equally likely? Use the discrete uniform. Is it a binary outcome? Use the Bernoulli. Are you counting the number of success in a fixed number of trials? Use the binomial. Are you running repeated trials until a (certain number of) success(es)? Use the geomtric (negative binomial). Are you sampling without replacement? Use the hyper geometric. Are you counting events over a fixed space? Use the Poisson.

Once identified, the distribution can be used in exactly the same way as any probability mass function. That is, we still require all of the probability rules, event descriptions, and techniques from before. The difference in these cases is that we immediately have access to the correct form of the probability mass function, the expected value, and the variance.

An additional utility with this approach to solving probability questions is that, over time and repeated practice, you can build-up an intuition as to the behaviour of random variables following these various distributions. Probabilities in general are deeply unintuitive: it can be hard to assess, without fomrally working it out, whether an evetn is likely or unlikely, let alone how likey any event is. However, the lack of intuition from our wider experience can

be negated almost entirely by building of intuition through the repeated application of these distributions. You can start to gain a sense of how binomial random variables behave, being able to determine just from inspection whether events seem plausible or not. Much of the study of probability and statistics is about building a set of tools that can overcome the flaws in our intuitive reasoning regarding uncertainty. This comes only through practice, however, this framework of named distributions provides a very solid foundation to perform such practice.

# 6 Continuous Random Variables

Our discussions of probability distributions, and their summaries have focused entirely on discrete random variables. To recap, a discrete random variable is any random quantity with a countable number of elements in the sample space. Discrete random variables are defined in contrast to continuous random variables, which take on values over the span of intervals in uncountably large sets. Suppose that $X$ can take any real number between 0 and 1. There is no way to enumerate the set of possible values for this random quantity, and so it must not be discrete.

Many quantities of interest are better treated as a continuous quantity rather than a discrete one, even if this is not technically correct. For instance, time measured in seconds is often best thought of as continuous, even though any stop watch used to grab these measurements will have some limit to the precision with which it can measure. Similarily, lengths (or heights) will often be better treated as continuous quantities, even though any measuring device will necessarily have some minimal threshold afterwhich it cannot discern distances. Deciding whether a quantity is continuous or discrete can thus, sometimes, be a judgment call. In general discrete quantities are harder to work with when the set of possibilities is very large. In these cases, not much is lost by treating the random variables as though they were continuous. This distinction is another area which requires the active development of intuition, but once present, it becomes second nature.

TODO: include examples for discrete versus continuous.

## 6.1 Continuous Versus Discrete

Distinguishing whether a random quantity is continuous or discrete is crucial as, broadly speaking, the two types of quantities are treated differently. The same underlying ideas are present, but the distinctions between the two settings require some careful thought. As a general rule, the use of continuous random variables necessitates an understanding of introductory calculus. This is not a pre-requisite for this course, and as a result, we will not focus as deeply on working with the underlying quantities. However, continuous random variables are also the dominant type of random variables outside of introductory courses. As a result, understanding the distinctions, and beginning to become familair with how they are to be manipulated is an important skill.

The key difference between discrete and continuous random variables is that, for discrete random variables the beahviour is governed by assessing $P(X = x)$ for all possible values of $x$, and for continuous random variables $P(X = x) = 0$ for every value of $x$. This is a surprising statement, and as such it is worth reiterating. With discrete random variables we discussed how all of the probabalistic behaviour was governed by the probability mass function which was defined as $p_X(x) = P(X = x)$. If $X$ is continuous, it will aalways be the case that $P(X = x) = 0$. Correspondingly, continuous random variables do not have probability mass functions, and to understand the behaviour of these random variables we must turn to other quantities.

While the fact that $P(X = x) = 0$ may seem unintuitive at first glance, it is worth exploring this even further. The nature of a continuous random variable is such that there is no possible way to enumerate all of the values which are possible to be realized by the random quantity. Suppose that we take a set of countably many possible observations and gave each of these a probability of greater than 0 of occuring directly. Even if we take an infinite number of them, there will still be an uncountably infinite number of events in the sample space which we have not accounted for. We know that the total probability of the sample space must be 1, and so we must have the total probability of the first set of events being less than 1 (if it summed to 1 then this would not be a continuous random variable, since there would only be a countable number of possible events). Now suppose we take another set of countably many events, again giving each of them a positive probability. Once more the sum of all of these probabilities must be less than 1, and specifically, the sum of both sets must also be less than 1. Even after these two sets there are still uncountably infinite events to go, and so we continue this process. Because we always need the total probability to be 1 once all events have been accounted for, and because we will always have an uncountably infinite number of events left to account for, we can **never** have a positive probability assigned to each event in a set of events. Even if we made the probability of each of these sets very, very small (say 1 in a million) after some fixed number of countable eevents the probabilities would be greater than 1 which cannot happen. As such, each event itself must have 0 probability.

An alternative technique for understanding this intuition is to think about how uunlikely it really would be to observe any specific value. Suppose that $X$ takes values on the interval $[0, 1]$. Recall that when we defined probabilities, we discussed them as being the long run proportion of time that an event occurs. Take some event, say $X = 0.5$. Suppose that we took repeated measurements of $X$ which are independent and identically distributed. Now suppose that at some point we exactly do observe $X = 0.5$. Should we expect that this will ever happen again? The next time we get near 0.5 might we instead not observe 0.51 or 0.49 or 0.5000000000000000001 or any of the other uncountably infinite values in the very near vicinity of 0.5? Each time that we make an observation the denominator of our proportion is growing, but if every value between 0 and 1 is truly possible, as time goes on the number of times that $X = 0.5$ must stay much, much smaller than the total number of trials. If we continue this off to infinity, in the limit, the probability must become 0.

This conclusion leads to a few different points. First, impossibility is not the same as probabil-

ity 0. Impossible events do have probability 0, but possible events may also have probability 0. Events which are outside of the sample space are impossible. Events inside the sample space, even probability zero events, remain possible. Second, we require alternative mathematical tools for discussing the probability of events in a continuous setting. Ideally this would be analogous to a probability mass function, but would somehow function in the case of continuity.

## 6.2 Cumulative Distribution Functions

To begin building up to the continuous analogue to the probability mass function, we will start by focusing on events that are easier to define in the continuous case. Suppose that $X$ is defined on some continuous interval. Instead of thinking of events relating to $X = x$, we instead turn our focus to events of the form $X \in (a, b)$ for some interval defined by the endpoints $a$ and $b$. Now note that, relying only on our knowledge of probabilities relating to generic events, we can rewrite $P(X \in (a, b))$ slightly. Specifically,

$$\begin{aligned} P(X \in (a,b)) &= 1 - P(X \notin (a,b)) \\ &= 1 - P(\{X < a\} \cup \{X > b\}) \\ &= 1 - (P(X < a) + P(X > b)) \\ &= 1 - P(X > b) - P(X < a) \\ &= P(X < b) - P(X < a). \end{aligned}$$

TODO: diagram showing this graphically.

In words we know that the probability that $X$ falls into any particular interval is given by the probability that it is less than the upper bound of the interval minus the probability that it is less than the lower bound of the interval. Notice that $X < a$ is an event, and if we knew how to assign probabilities to $X < a$ for arbitrary $a$, then we could assign probabilities to any interval. Also note that, even in the continuous case, it make sense to talk of $P(X < a)$ for some value $a$. These intervals will contain an uncountably infinite number of events, and as such, can certainly occur with greater than 0 probability. Using our common example of $X$ being defined on $[0, 1]$, then $P(X < 1) = 1$. Note that we could have written $P(X \leq 1) = 1$, which may have been more obviously true. However, $P(X \leq 1) = P(\{X < 1\} \cup \{X = 1\}) = P(X < 1) + P(X = 1)$ and we know that $P(X = 1) = 0$. In the continuous case we do not need to worry whether we use $X \leq a$ or $X < a$, and we will interchange them throughout.

TODO: uniform example

The centrality of events of the form $X < a$ prompts the definition of a mathematical function which we call the **cumulative distribution function**. We will typically denote the cumulative distribution function of a random variable $X$ as $F(x)$, sometimes using $F_X(x)$ to emphasize that this function relates to $X$ specifically. We may also refer to the cumulative distribution

function simply as the **distribution function**. By definition, we take $F_X(x) = P(X \leq x)$. Once we have defined the distribution function for a random variable, using the above derivation we are able to determine the probability associated with any events based on intervals.

It is worth noting that the cumulative distribution function can also be defined for discrete random variables. In the case of a discrete random variable, we would have

$$F_X(x) = \sum_{k \in \mathcal{X}; k \leq x} p_X(k).$$

Since it is simply the summation of the probability mass function it tends to be a less useful quantity. Still, the cumulative distribution functions for discrete random variables do come up on ocassion, and it is worth recognizing that they are defined in exactly the same way.

TODO: Include example with the CDF.

Supposing that, for some continuous random variable $X$ we have the cumulative distribution function, then thisknowledge actually permits us to compute *any* probability associated with the random variable that we can want. Consider any event associated with $X$ which we may wish to determine the probability of. We know that events are merely subsets of the sample space. Every one of these events can be written using our basic set operations (unions, intersections, and complements) applied to intervals of the form $(a, b)$ and **singelton sets** of the form $\{x\}$. Our basic axioms of probability allow us to compute probabilities across the set operations, and our knowledge of the cumulative distribution, the conversion of $P(X \in (a, b)) = F_X(b) - F_X(a)$, and the fact that $P(X = 0) = 0$ gives all of the results we need to derive probabilities for these events.

TODO: Include simple examples

## 6.3 The Probability Density Function

The distribution function will be the core object used to discuss the probabalistic behaviour of a continuous random variable. All of the beahviour of these random quantities will be described by the distribution function, and as such we will take the distribution function as a function which defines the distribution of a continuous random quantity. This is all that we need in orderto analyze the probabalistic behaviour of these random variables, however, it may be a little unsatisfying in contrast with the discrete case.

We had set out to find a quantity which was a parllel to the probability mass function, and instead concluded that the cumulative distribution function can eb made to play the same role in terms of describing the behaviour of the random quantity. Still, it may be of interest for us to have a function which takes into account the relative likelihood of being near some value. Suppose, for instance, that for a random variable defined on $[0, 1]$ we wanted to know how likely it was to be in the vicinity of $X = 0.5$. We could take a small number, say $\delta = 0.01$ and calculate $P(X \in (0.5 - \delta, 0.5 + \delta)) = F(0.5 + \delta) - F(0.5 - \delta)$. This is perfectly well defined

based on our discussions to this point. Now, if we assume that the probability isfairly evenly distributed throughout this interval, then if we wanted to assign a likelihood to each value we could divide this total probability by hte length of the interval, which is $2\delta$. As a result, we are saying that the probability that $X$ is nearly 0.5 will be approximately given by the expression $\frac{F(0.5+\delta)-F(0.5-\delta)}{2\delta}$.

We had taken $\delta = 0.01$, but the same process could be applied for smaller and smaller $\delta$, say 0.001 or 0.0001. Intuitively, as the size of this itnerval shrinks more and more we are getting a better and better estimate for the likelihood that the random variable is in the immediate vicinity of 0.5. Moreover, as $\delta$ gets smaller and smaller our assumption of a uniform probability over the interval becomes more and more reasonable. Now,, we cannot set $\delta = 0$ exactly, however, we can ask what happens in the limit as $\delta$ continues to get smaller and smalller. This question is in the purvue of calculus, and can in fact be answered. While working out the answer is beyond the scope of the course, we will provide the result anyway.

The resulting function is called the **probability density function**, and is related to the cumulative distribution function through derivatives (and integrals). If a random variable $X$ has a cumulative distribution function, $F(x)$, we typically denote the corresponding density function as $f(x)$. The density function also describes the behaviour of the random variable, and mirrors the behaviour of the probability mass function in the discrete case. Roughly speaking, the density function evaluates how likely (relatively speaking) it is for a continuous quantity to be in a small neighbourhood of the given value. Critically, **probability density functions do not give probabilities directly**. In fact, probability density functions may give values that are greater than 1!

TODO: Example of uniform PDF.

Still, if we see the shape of the probability density function, we can state how likely it is to make observations near the results of interest. We will often graph the density function. The high points of the graph indicate regions with more probability than the regions of the graph which are lower. Again, the specific probability of any event $X = x$ will always be 0, but some events fall in neighbourhoods which are more likely to observe than others.

TODO: include examples.

## 6.4 Using Continuous Distributions

With the exception of the differences indicated until this point, there is otherwise not much difference between continuous and discrete random variables. The tools to analyze them differ (in the continuous case, we cannot sum over the sample space, and so we must use techniques from calculus to mirror this process, for instance), but the fundamentals remain the same. It is still possible to compute expected values (and medians and modes) with roughly the same interpretations. It is still possible to describe the range, interquartile range, and variance, again with corresponding interpretations. The axioms of probability still underpin the manipulation

and analysis of these random variables. The distinction is merely that in place of elementary mathematics to complete the calculations, calculus is required.

Just as with discrete distributions, there are continuous named distributions. These are typically governed by either a density function or else a distribution function, alongside the expected value and variance. And just like the named discrete distributions, by matching the underlying scenario to the correct process, we are able to side step a lot of work in understanding the beahviour of the random quantities. Now, because there is no assumed knowledge of calculus, we will not work too widely with continuous distributions. We will introduce only two named continuosu distributions: the uniform distribution, which we have already started to see, and the normal distribution, which is far and away the most important distribution (discrete or continuous) in all of probability and statistics.

## 6.5 The Uniform Distribution

The uniform distribution, sometimes called the continuous uniform distribution to distinguish it from the discrete counterpart, is parameterized over a set interval specified as $(a, b)$. On this interval, equal probability density is given to every event, which is to say that the density function is constant. Specifically, for $X \sim \text{Unif}(a, b)$ we note that

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & \text{otherwise.} \end{cases}$$

From the density function we can work out that

$$F(x) = \frac{x - a}{b - a}$$

for $x \in (a, b)$, with $F(x) = 0$ for $x < a$ and $F(x) = 1$ for $x > b$. Moreover, we have $E[X] = \frac{a+b}{2}$ and $\text{var}(X) = \frac{(b-a)^2}{12}$.

TODO: Include some calculations

The uniform distribution is analogous to the discrete uniform. Any time there is an interval of possible outcomes which are all equally likely, the uniform distribution is the distribution to use. Compared with other distributions it is also fairly straightforward to work with, which makes it a useful demonstration of the concepts relating the continuous probability calculations.

TODO: Include examples.

# 7 The Normal Distribution

The normal distribution, also sometimes referred toas the Gaussian distribution, is a named continuous distribution function defined on the complete real line. The distribution is far and away the most prominently used distribution in all of probability and statistics. In fact, most people have heard of normal distributions even if they are not aware of this fact. Any time that there is a discussion of a bell curve, for instance, this is in reference to the normal distribution. Normally distributed quantities arise all over the place from measurements of heights, grades, or reaction times through to levels of job satisfaction, reading ability, or blood pressure. There is a tremendous number of normally distributed phenomena naturally occurring in the world, which renders the normal distribution deeply important across a wide range of domains.

Perhaps more important than the places where the normal distribution arises in nature are hte places where it arises mathematically. At the end of this course we will see a result, the central limit theorem, which is one of the core results in all of statistics. Most of the statistical theory that drives scentific inquiry sits ontop of the central limit theorem, and at the core of the central limit theorem is the normal distribution. It is virtually impossible to overstate the importance of the normal distribution, and as a result, we will spend a great deal of time investigating it.

## 7.1 The Specification of the Distribution

A normal distribution is parameterized by two different parameters: the mean, $\mu$, and the varaince $\sigma^2$. We write $X \sim N(\mu, \sigma^2)$. These parameters directly correspond to the relevant quantities such that $E[X] = \mu$ and $\text{var}(X) = \sigma^2$. The density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

This can be quite unwieldy to work with, however, when it is plotted we ssee that the normal distribution takes on a bell curve which is centered at $\mu$.

TODO: include plots

## 7.2 The Standard Normal Distribution

Normally distributed random variables are particularly well-behaved. One way in which this is true is that if you multiply a normally distributed random variable by a constant, it will remain normally distributed, and if you add a constant to a normally distributed random variable, it will remain normally distributed. Consider then if $X \sim N(\mu, \sigma^2)$, taking the quantity $X - \mu$. We have seen in our discussions of expected values that $E[X-\mu] = E[X]-\mu = 0$. Furthermore, adding or subtracting a constant will not change the variance. Thus, $X - \mu \sim N(0, \sigma^2)$.

Now, consider dividing this by $\sigma$, or equivalently, multiply by $\frac{1}{\sigma}$. The expected value of the new quantity will be $\frac{1}{\sigma} \times 0 = 0$, and the variance of the newquantity will be $\frac{1}{\sigma^2} \times \sigma^2 = 1$. Taken together then, if $X \sim N(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

This holds true for *any* starting normal distribution, with any mean or variance values. As a result, this straightforward transformation allows us to discuss any normal distribution in terms of $N(0, 1)$. We call this the **standard normal distribution**, and will typically use $Z$ to denote a random variable from the standard normal distriubtion.

TODO: Include example

If $Z \sim N(0, 1)$, then we use a special notation for the density function and distribution function of $Z$. Specifically, we take the density function to be denoted var$phi(z)$ and the cumulative distribution function to be given by $P(Z \leq z) = \Phi(z)$. The cumulative distribution function does not have a nice form to be written down, however, it is a commonly applied enough function that many computing languages have implemented it, including of course R.

TODO: Demonstration of using *norm.

The utility in this is demonstrated by realizing that events can be converted using the same transformations. Specifically, suppose we have $X \sim N(\mu, \sigma^2)$, and we want to find $P(X \leq x)$. Note that, $X \leq x$ must also mean that

$$\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma},$$

simply by applying the same transformation to both sides. But we *know* that the left hand side of this inequality is exactly $Z$, a standard normal random variable with cumulative distribution function $\Phi(z)$. Thus,

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Using this trick of **standardization** any normal probability can be converted into a probability regarding the standard normal, for which we can easily use computer software.

TODO: Include normal calculation TODO: Include discussion of R calculating normal probabilities

As a result, combining our knowledge of continuous random variables, with the process of standardization we are able to calculate normal probabilities for any events relating to normally distributed random quantities. Moreover, since the shape of the normal distribution is so predictable, it is often easy to draw out the density function, and indicate on this graphic the probabilities of interest, which in turn helps with the required probability calculations. Calculating probabilities from normal distributions will remain a central component of working with statistics and probabilities beyond this course. Developing the skills and intuition at this point, through repeated practice is a key step in successfully navigating statistics here and beyond.

TODO: another example.

When you have access to a computer, and your interest is in calculating a normal probability, as described above, there is not properly a need for standardization. However, it rmeains an important skill for several reasons. First, by always working with the same normal distribution, you will develop a much more refined intuition for the likelihoods of different events. It goes beyond working with the same family of distributions, you get very used to working with exactly the same distribution. Second, you will likely become quite familiar with certain key **critical values** of the standard normal distribution. These values arise frequently, and allow you to quickly approximate the likelihood of different events. Finally, as we begin to move away from studying probability and into studying statistics, the standard normal will feature prominently there.

## 7.3 The Empirical Rule

Another way in which the normal distribution is well behaved is summarized in the **emprical rule**. The shape of the distribution is such that, no matter the specific mean or variance, all members of the family remain quite similar. This enables the derivation of an easy, approximate result, to help intuitively gauge the probabilities of normal events. The emprical rule states that, if $X$ has a normal distribution, then the probability of observing a value within $\sigma$ or the mean is approximately 0.68, the probability of observing a value within $2\sigma$ of the mean is approximately 0.95, and the probability of observing a value within $3\sigma$ of the mean is approximately 0.997.

TODO: Include graphic

In words, the empirical simply states that almost all of the observations from a normal distribution will fall within $\mu \pm 3\sigma$. In mathematical terms, the empirical rule is summarized as $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$, and $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$. With the standard normal we can replace $\mu$ with 0, and $\sigma$ with 1 to get a version which

is slightly more concise to state. It is then possible to combine these different intervals by recognizing the symmetry in the normal distribution. That is, $P(\mu \leq X \leq \mu+\sigma) \approx \frac{0.68}{2} = 0.34$.

TODO: Calculations with the empirical rule.

The empirical rule is not exact, and again, when computing probabilities with access to statistical software, it is likely of limiteddirect utility. However, it is another tool to leverage to continue developing a refined intuition for the behaviour of random quantities. It is also a good check to have a sense of the likelihood of different events. If you compute an answer which seems out of line with the empirical rule, take a look moreclosely. If you have someone tell you that they have observed events which are out of line with the empirical rule, be skeptical.

### 7.3.1 Chebyshev's Inequality

The empirical rule is a useful result to aid in building intuition regarding the normal distribution. However, when quantitiesare not normally distributed, we cannot use the empirical rule. A related, though somewhat weaker result, does hold for *any* distribution, and it is a useful extension to the empirical rule. Stated in words, Chebyshev's inequality says that there is a probability of 0.75 or more of observing an observation within two standard deviations, and a probability of at least 0.8889 of observing a value within three standard deviations of the mean.

Formally, we can write that

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

Here $k$ can be any real number which is greater than 0. If $k \leq 1$, this result is uninteresting since the bound simply is 0. However, taking $k = 2$ gives the 0.75 lower bound outlined above, which is a more useful result. Additionally, there is no requirement for $k$ to be an integer here, and so, for instance, the probability of observing a value within $\mu \pm \sqrt{2}\sigma$ is at least 0.5, for all distributions.

TODO: Include some examples.

## 7.4 Closure of the Normal Distribution

We have seen a certain type of "closure property" for the normal distribution when we discussed standardization. That is, adding and multiplying by constants does not change the distribution when working with normally distributed quantities. This is an interesting property which does not hold for most distributions, and makes normally distributed random variables quite nice to work with. The normal distribution has an addition type of closure property which is frequently used, and is also somewhat surprising.

Suppose that $x$ and $Y$ are independent of one another, with $X \sim N(\mu_X, \sigma_X^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$. In this setting,

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

in words, the addition of two independent normally distributed random variables will also be normally distributed. This extends beyond two in the natural way, simply by applying and reapplying the rule (as many times as is required).

TODO: Include example of this.

This becomes particularly useful when we think of generating many iid realizations in an experiment from a normal population. For instance, if $X_1, \ldots, X_n$ are all iid from a $N(\mu, \sigma^2)$ distribution, then

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2).$$

If instead we consider the average of these $n$ independent variables, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}),$$

through an application of our standard expectation and variance transformation rules. This type of result is central to the practice of statistics, and this closure under addition further aids in the utility of the normal distribution.

## 7.5 Normal Approximations

A final utility to the normal distribution is in its ability to approximate other distributions. While several of these approximations exist, we will focus on the normal approximation to the binomial as an illustrative example. Historically, these approximations were critical for computing probabilities by hand in a timely fashion. Owing to the widespread use of statistical software, these usecases are more and more limited, which removes the necessity of these approximations directly. However, there are two major advantages to learning these approximations. First, with an approximation it becomes easier to leverage the intuition you will build regarding the normal distribution in order to better understand the behaviour of other random quantities. Second, the normal approximation has the same "flavour" as many results in statistics, and so it presents an additional path to familiarity with these types of findings.

Suppose that $X \sim \text{Bin}(n, p)$. Through knowledge of the binomial distribution, we know that $E[X] = np$ and $\text{var}(X) = np(1-p)$. If $n$ is sufficiently large then it is possible to approximate a binomial distribution using a normal distribution with the corresponding mean and variance. That is, for $n$ large enough, we can take $X \sim \text{Bin}(n, p)$ to have approximately the same distribution as $W \sim N(np, np(1-p))$.

TODO: Brief example.

One consideration that we need to make when applying this approximation has to do with the fact that the normal distribution is continuous while the binomial distribution is discrete. As a result, the normal distribution can take on any value on the real line, where the binomial is limited to the integers. A question that we must answer is what to do with the non-integer valued numbers. The natural solution is to rely on rounding. That is, for any value between $[1.5, 2.5)$ we would round to the nearest integer, which is 2.

This natural solution is in fact a fairly useful technique, and it is the one that we will make use of in the normal approximation. While rounding is quite natural, the process for leveraging this idea in probability approximation is somewhat backwards. That is, we typically will need to go from probabilities relating to $X$ and transform those into probabilities relating to $W$. So, for instance, if we wish to know $P(X \leq 2)$, then we need to be able to make this a statement regarding the random variable $W$. In order to do this we need to ask "what is the largest value for $W$ that would get rounded to 2?" The answer is 2.5 and so $P(X \leq 2) \approx P(W \leq 2.5)$.

A similar adjustment would be required if we instead wanted $P(X \geq 5)$. Here we would ask "what is the smallest value for $W$ which would get rounded to 5?" and note that the answer is 4.5. Thus, $P(X \geq 5) \approx P(W \geq 4.5)$. Once we have expressed the probability of interest in terms of the normla random variable, we can use the standard techniques previously outlined to compute the relevant probabilities.

TODO: Examples

It is very important to keep in mind that the two results discussed above were of the form $X \geq x$ and $X \leq x'$. If we instead had considered $X > x$ or $X < x'$, we would need to take an additional step. For continuous random variables whether $X \geq x$ or $X > x$ is considered it makes no difference. However, for discrete random variables this is not the case. As a result we should first convert the event the an equivalent event which contains the equality sign within the inequality, and then apply the continuity correction. (TODO: ensure that continuity correction was referenced before). That is, if we want $P(X > 3)$ first note that for $X > 3$ to hold, we could equivalently write this as $X \geq 4$. Alternatively, if the event of interest is $X < 8$, this is the same as $X \leq 7$.

TODO: Further examples.

When it is not necessary, it rarely makes sense to use an approximation. There will be cases where the approximation is directly useful, and in those moments it is great to be able to use it. This example of using the normal distribution to approximate a discrete random variable serves as a nice bridge from the study of probability to the study of statistics. In statistics we take a different view of the types of problems we have been considering to date, and we require the tools of probability that have been brought forth. As a result, a deep comfort with manipulating probability expressions is required to build a strong foundation while studying statistics.

# Part II

# Part 2: Statistics