Working with Random Systems

Mechanics, Meaning, and Modeling

by

Christopher R. Genovese

PREFACE

He who wishes to teach us truth should not tell it to us, but simply suggest it with a brief gesture, a gesture which starts an ideal trajectory in the air along which we glide until we find ourselves at the feet of the new truth.

— José Ortega y Gasset, "Preliminary Meditations"

These notes develop the theory of probability for undergraduate students, particularly for students in electrical engineering and computer science. Although this is the beginning of a text book, it is not yet a stand alone book. It will be supplemented by extended in-class examples and homework exercises during the course. You will also encounter case studies that work through interesting applications. Mastering the material requires an interplay between all these sources of information.

A common response to a mathematical presentation of probability is to recoil from the mathematics. One of the themes of this book is that in most important senses, mathematical notation and programming languages can be approached in the same way. In particular, mathematical objects have well-defined types and require definition much like those in a programming language. Mathematical expressions have scopes within which local variables can be defined and used. And most importantly, mathematical expressions represent an abstraction of algorithms in much the same way as a programming language. The syntax and some of the semantics differ, but "thinking like a compiler" can make the notation more easily approachable.

Mathematical notation can always be daunting, but I put in substantial effort to make it a help rather than a hindrance. The notation is designed to assign mnemonic meaning to the mathematical constructs we use. It is fanatically consistent throughout. Appendix N gives a summary of the notation for the reader's reference. The mathematical background we will build (which explains some of the notation) is described in Appendices B, F, M, and S. Appendix H gives a quick-reference help sheet that can be used during exams.

In terms of the content, I take a somewhat unusual approach to the theory but one that I believe has big payoffs. The notes begin with an emphasis on the foundations of a probability model and the physical meaning underlying it. This is helpful for ascribing meaning to the mechanics that we use for solving problems day to day. In contrast to most other treatments, I define and use random variables, expected values, and distributions almost immediately. Before proceeding further, you should have make sure you understand what physical quantities and operations these mathematical objects correspond to. These ideas are built up in layers of sophistication as we proceed and clarify the basically simple structure of a probability model from the outset. In the end, the entire class is about taking weighted averages. If you can take a weighted average, you can do this.

One often hears that the mathematics makes complicated what should be simple. Indeed, it is simple and intuitive enough to analyze Roulette or Coin Flipping without any edifice of theory. But, in general, for real applications it is complicated, and an understanding of what these mathematical objects *mean* is essential to constructing reasonable models.

This material exists at three levels. There is the *mechanics* of manipulating the mathematical objects. While this can be challenging, I believe that it is not enough for a course like this. The *meaning* of the mathematical constructs can be obscured by the mechanics but is critical for understanding the material in the long term. We need not wait until the mechanics have been mastered to develop this meaning; to do so is to lead to a course of mysterious, rote rules. Finally, the ultimate goal of this course is to teach students the mathematical *modeling* of uncertain or random systems to help them make better predictions and decisions.

These notes taken together with the examples, case studies, and supplementary exercises attempt to develop the material at all three levels. I have relied on many sources for examples, ideas, and inspiration in all of these components. (See Appendix R for a complete, annotated list.)

CONTENTS

FOUNDATIONS OF PROBABILITY MODELS

1

If you have built castles in the air, your work need not be lost; that is where they should be. Now put the foundations under them.
— *Henry David Thoreau,* Walden

1. Measurements, Prediction, and Decisions

Why do we need a theory of probability? Because we live in an uncertain world and must continually make decisions based on incomplete or imprecise information. The need for a systematic theory of randomness lies somewhere on the boundary between biology and quantum mechanics. On the one hand, the biological reality: we are all well-adapted measuring devices. Our sensory systems process immense quantities of information every second which is passed to the brain for organization, selection, and interpretation. Even so simple a task as following a moving animal with one's eyes requires a complicated circuitry in the brain, a progressive filtering and refiltering, separation and recombination of input stimuli. We take in the animal's path, our brain makes a prediction about where it will move, and our eyes follow.[1] What makes this system so powerful is our capacity to learn. While in any given moment there are virtually an infinite number of measurements we could take, not all of these are equally useful to us. We quickly learn to attend to some data and ignore others, to associate specific conditions with specific outcomes, to collect certain measurements that inform decisions necessary to our survival. This is the essence of probability

1 This process has especially high survival value if the animal is a grizzly chasing you.

theory in implicit and primal form. On the other hand, the physical reality: randomness is both pervasive and essential. In a Newtonian world, everything is determined and randomness is only an illusion; if we knew the exact position and velocity of every particle in the universe, we could read the future from Newton's laws. Any perceived randomness in such a world is only the by-product of what we do not know, so in principle, we could eliminate the randomness by measuring very precisely. Is the ultimate knowledge then reduced to an accounting of so many microscopic billiard balls? No! Quantum theory tells us that the randomness is present at a fundamental level. No matter how carefully we measure, we cannot make the randomness disappear, and we can never exactly determine the "true" state of the particles around us. Thus, we can never know the future exactly. Rather than ignore or fear the inherent randomness, we embrace it, describe it, use it. Probability theory is a calculus of the random.

Historically, probability theory was developed for a much more mundane pursuit—money. Gambling on games of chance can be traced back to the earliest civilizations.[2] Such games of chance illustrate both of the above ideas: the inherent randomness that defines the games and the ability to acquire knowledge and skill[3] in many games that can improve one's play. People developed a practical knack for the games, but many challenging puzzles remained unexplained. Mathematicians from the Renaissance to the 19th century attempted to tackle these and in the process devised a mathematical framework—a description of the games in terms of sets, functions, and measures—that enabled them to quantitatively analyze these games. With a little tweaking and a bit of generalization, their framework transfers from modeling gambling to almost any random system. This chapter describes this framework in detail. The key components are called Random Experiments, Elementary Outcomes, Events, and Random Variables.

2 The four-sided *astragali* of ancient Egypt made from animal heelbones are thought to be a precursor of dice.

3 Skill in games like bridge, poker, and blackjack can make a tremendous difference. The effectiveness of card counting in blackjack led casinos to increase the number of decks in the dealer's shoe. Skilled players in other games like bridge and poker will regularly trounce novices.

2. Abstraction

A probability model is a mathematical abstraction of the random system it describes. The abstraction requires some work to understand, but as we shall see, the work pays off in several ways. Before we think about probability models, let's see some examples of abstraction in other kinds of models.

In physics, Newtonian mechanics is based on an abstraction of space as a three-dimensional continuum and of matter as a collection of infinitesimal particles that have a definite position and momentum. Such assumptions about the continuum and particles need not be true (they aren't) to be useful. The abstraction provides a common set of tools and objects for any situations we will analyze with Newtonian methods. For example, both the position and momentum of a particle or body are represented by three-dimensional vectors, and we can use calculus to compute any desired feature of the system.[4] All of this is not exclusive to Newtonian physics: Quantum Mechanics, Relativity, even String theory have their own abstractions.

In biology, Mendelian genetics is based on the abstraction of genes, dominant and recessive, as basic and immutable units which are passed on according to specific rules. Modern genetic theory substantially refines these ideas[5] but itself abstracts the complicated chemical processes underlying genetic mutation and replication.

In computer science, programming languages are an abstraction of the complicated machine instruction set. This abstraction makes the essential concepts clearer (assignment, function call, etc.) and easier to generalize (to multiple languages), but it allows the same program to be used on machines with wholly different architectures and microcode. Algorithms provide a similar example. They abstract from the particulars of a computation to the essential structure of the operation, and are thus widely applicable. For example, the algorithm of sorting does not depend on the particulars of the items to be sorted; all that we require is a method with which to compare any two objects.

The hardest part of modeling a random system lies in describing the randomness itself. Each system has innumerable complexities that are specific to it, and any physical accounting for the various sources of randomness is often terribly difficult. As in the above examples, we

4 The equations that result, however, may be impossible to solve analytically, but we can then use numerical methods or simulation.

5 Through, for example, the idea of recombination.

can use an abstraction to make the unworkable workable. The principal abstraction that we use for a random system is that of picking a ticket out of a box at random. This does not seem like much to build a theory on, but this abstraction turns out to be powerful and, most importantly, it works.

The challenge we yet face is to map this "tickets in a box" abstraction to the real system we wish to model. To do this, we must add several additional layers of abstraction in support. At the top layer (closest to the application) are the observable quantities that we wish to measure. At the bottom layer (usually specified only implicitly) is a description of everything that can happen as an experiment unfolds. We take this all up in the ensuing sections.

3. Tickets in a Box

The canonical random system is the process of picking tickets out of a box. To understand the structure of a probability model, we need only start here. All of the fancy theory we will develop reduces the random systems we find in practice to this simpler form.

Consider a box filled with tickets that have been thoroughly mixed around. However, these are not simple, flimsy paper strips but rather the latest in ticket technology – the patented SmartTickets(TM). These SmartTickets have an embedded chip that enables us to store an arbitrary amount of information on them. Externally, the SmartTickets are physically indistinguishable in size, weight, color, and so forth.

To setup the SmartTickets, we put the entire collection of tickets into the SmartTicket Writer. We first name the fields of information to be stored on every ticket and then enter the information by name for each individual ticket. The SmartTickets also come with a SmartTicket Reader that allows us to retrieve any information stored on a ticket. We simply identify the name of the information we want to retrieve, scan a single ticket through the Reader, and the Reader displays the value of the requested information for that ticket.

We begin our experiment with randomness by entering the desired information onto our tickets. We put the tickets in the box,[6] and mix them thoroughly. We then allow "nature" to select a single ticket. The information stored on the selected ticket is the result of the random mechanism. The examples in the next section put flesh on the bones of

6 The SmartBox is available for an extra charge, but it is really not worth the price.

this metaphor.

Since the tickets are physically indistinguishable and are well mixed, there is nothing special about any one ticket. Symmetry suggests, therefore, that all tickets in the box are equally likely to be chosen.

The Components of a Probability Model

The Random Experiment

- We enter the information on the tickets.
- We put them in the box and mix them thoroughly.

Outcomes and Events

- Each ticket that might be chosen describes one way the experiment could turn out.
- An event is a set of tickets that all satisfy some particular property.

The Random Mechanism

- Nature draws one ticket at random from the box.
- An event is said to *occur* if the selected SmartTicket is a member of that set, i.e., satisfies the defining property of the event.

Random Variables

- We program the SmartTicket Reader to compute and display some quantity of interest using the information stored on a ticket.
- We scan the selected ticket through the SmartTicket Reader.
- The Reader computes and displays the requested value using the information stored on the selected ticket.

There are useful variations on this simple idea. We can arrange for some tickets to be bigger than others – and thus more likely to be chosen. We may also wish to allow for the possibility of an infinite (countably or uncountably) number of tickets to be placed in the box. Within that scope, we can describe any system we will encounter.

> **Abstraction.** The randomness in every system is analogous to the case of picking tickets randomly from a box.

4. Two Prototypical Examples

In this section, we describe two basic experiments – Roulette and Coin Flipping – that illustrate all of the essential elements of the theory. This will serve as motivation for the definitions in the following sections.

4.1. Roulette

A Nevada roulette wheel has 38 numbered and colored slots along its edge that can catch and hold the ball. The slots are numbered 0, 00, and 1 through 36. The first two are colored green, and half each of the remaining slots are red and black, respectively. (In the Figure below, the shaded squares are the black slots.)

The betting table for roulette contains an array of squares and tabs where players place their chips to determine their bet. The main portion of the array gives the numbers 0, 00, and 1 through 36 along with the colors of the corresponding slots.

00	3	6	9	12	15	18	21	24	27	30	33	36	2-to-1
	2	5	8	11	14	17	20	23	26	29	32	35	2-to-1
0	1	4	7	10	13	16	19	22	25	28	31	34	2-to-1
	1st-12				2nd-12				3rd-12				
	1-to-18		Even		Red		Black		Odd		19-to-36		

Below that are rectangles naming particular sets of numbers such as "Red" or "First 12" which correspond to composite bets. Before the spin, players place chips on the array to indicate the type of bet,[7]

7 Called a "play" by those in the know.

The wheel is spun and a ball released into it. Eventually the ball comes to rest in one of the slots. This slot determines the outcome of any bets placed on the spin. Players can bet on a variety of different possibilities as described in the table below.

Name of Bet	Winning Numbers	Payoff Odds
Even Money	A set of 18 numbers: red, black, even, odd, 1–18, 19–36	1 to 1
Dozen	One of 1–12, 13–24, or 25–36	2 to 1
Column	12 numbers in one row[8]	2 to 1
Line	6 consecutive numbers (2 columns)	5 to 1
House Special	0, 00, 1, 2, 3	6 to 1
Quarter	4 numbers that share a corner	8 to 1
Street	3 consecutive numbers (1 column)	11 to 1
Split	An adjacent pair of numbers	17 to 1
Straight	A single number	35 to 1

8 The table arrangement shown in the Figure is rotated 90 degrees from that used at the Nevada casinos. What we show as a row is usually displayed as a column, thus the name "Column" for this play.

The payoff odds for each play give the ratio of the amount won to the amount bet, should that bet succeed. For instance, the Straight Play is less likely to win than an Even Money Play, but a winning \$1 bet pays \$35 in the former case and \$1 in the latter.

THE RANDOM EXPERIMENT. When building a model for a random system, we need to describe what specific features of the system are essential to our model. We usually begin, then, with a description of the random experiment. This description should be clear and contain appropriate details; a good rule of thumb is that someone else reading your description should be able to run the experiment for you.

In this case, the experiment is simple: we and all other players place their bets, the croupier starts the ball spinning in a standard Nevada roulette wheel, and we watch it spin until it comes to rest at which time bets pay off. Notice that we have no control over how fast the wheel is spun or when the ball is released; uncertainty in these things is part of the experiment.

OUTCOMES AND EVENTS. We start with a simple question: what can happen during a spin of the roulette wheel? Actually, quite a lot. The seemingly meek old lady at the table could be discovered slipping another player's chips into her bag. The portly gentleman from Texas puffing on his cigar could collapse with severe stomach pains. A fight could break out on the casino floor to be ended roughly by several very large men in suits. Oh, and by the way, the ball could fall into one of the slots. Most of these events are neither likely nor too relevant for placing a bet on the spin, so we make a modeler's decision to ignore them for now.

The simplest description of the possible outcomes of a roulette spin considers only the 38 slots. In terms of our Tickets in a Box abstraction, this involves placing 38 SmartTickets in the box, each with the label of a single slot as the only information entered on the chip. For the purposes of our model, each of these 38 tickets represents one way the experiment can turn out; each ticket records on it all the interesting information that we may wish to measure in this experiment.

Mathematically, we represent these possibilities by a set,[9]

$$\Omega = \{-2, 0, 1, 2, \ldots, 36\},$$

whose elements each correspond to one of the slots in which the ball can end up. Because it will be easier (though not necessary) to work with sets of numbers, we use -2 here to represent the 00 slot. This choice is arbitrary, but using an even number seems appropriate and the "2" evokes the double-zero.

9 We depart from our naming convention for sets only for the set that plays this particular role in a probability model. For the corresponding set in any model, we will always use the Greek uppercase letter Ω.

Each element of $\omega \in \Omega$ is called an elementary outcome and embodies one way in which the experiment can turn out. Hence, an elementary outcome ω corresponds to exactly one SmartTicket. We call Ω the *outcome space* of the experiment because it describes all the possible outcomes. Ω corresponds to the entire collection of SmartTickets, which we configured to do the same thing.

Every ticket in the box describes the outcome of the experiment in complete detail, but sometimes we are interested in a coarser description. For example, if the ticket that is selected is any one of the numbers 1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, or 36, then red wins. It is thus reasonable for our description for the event that red wins to be the set of tickets with these slot numbers stored on them. Mathematically, the same idea works. Each elementary outcome ω embodies one way in which the experiment can turn out, so the set of elementary outcomes that satisfy some observable condition (e.g., the slot is colored red) describes all the ways the experiment can turn out such that that condition holds. Events, then, are subsets of the outcome space. From the above, we have that $\mathcal{R} = \{1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36\}$

is the event that the ball lands in a red slot. Notice that the complement of this set (in Ω) is set of non-red slots; we denote this by $\mathcal{N}$ for this example. Other examples in roulette might be $\mathcal{S} = \{36\}$ or $\mathcal{D}_1 = \{1, \ldots, 12\}$.

Thought Question What do the events $\mathcal{S}$ and $\mathcal{D}_1$ correspond to? What is the event that the House Special wins?

The Random Mechanism. So far, we have built a mathematical description of the game of roulette. What remains is to posit a mechanism by which one of the 38 possible outcomes is selected. By direct analogy to the Tickets in a Box abstraction, we assume the following basic mechanism:

1. The croupier spins the wheel,
2. Nature (or chance or fate or ...) reaches into the box and chooses one ticket at random, and
3. The slot in which the ball lands is determined by the information read off the selected ticket.

In terms of the outcome space Ω, this corresponds to choosing a single elementary outcome from Ω. This is called the *selected elementary outcome* and is denoted by ω^*.

Random Variables. In devising strategies for roulette, we are interested in the results of particular bets. The value of any bet is uncertain because the outcome of the spin is random, but once the final slot is known, the value of the bet can be computed explicitly. We can therefore exploit a new feature of the SmartTicket Reader: we can program it to display not just the information on the ticket but any quantity derived from that information by some algorithm. So, for example, instead of just the slot number, we might configure the reader to also display the value of a \$10 Even-Money play on red and a \$5 House Special play. We would get something like the following:

Slot Number:	16
Value of a \$10 Even-Money play on red:	10
Value of a \$5 House Special play:	-5

Other tickets in a box would show different values for these fields. Notice that all of these values are computed as a function of the slot number.[10] We call these random variables because the value we observe depends on which ticket is selected.

The mathematical analogy is again direct. A random variable in this experiment represents the value of one particular bet on the roulette spin.[11] Because each elementary outcome specifies precisely one way the experiment can turn out, the value of any bet can be computed as a function on the elementary outcomes. To account for every possible outcome, we need a rule that determines the value of the bet for any slot. This leads us to define a random variable as a *function* that takes an elementary outcome as an input argument and returns a number.[12]

Any function B that maps Ω to the set of real numbers[13] $\mathbb{R}$ is a random variable that embodies a particular bet. For any particular $\omega \in \Omega$, the value that the random variable takes is $B(\omega)$. This is what we would observe (how much we would win on our bet) if that ω (ticket, slot) is selected when the wheel is spun. When we talk about a random variable B, we refer to the function itself as a mathematical object. All statements made about B as an object must then be true for all values of its argument.

For example, if $\mathcal{R}$ is the event that a red slot occurs and if $\mathcal{N}$ is the complementary event that a red slot does not occur, a \$10 Even Money Play on red is a function B_{red} that maps Ω to the set of real numbers[14] $\mathbb{R}$ defined[15] by

$$B_{\text{red}} = 10 \cdot 1_{\mathcal{R}} - 10 \cdot 1_{\mathcal{N}},$$

and a \$10 Straight Play on the number 16 is a function

$$B_{16} = 350 \cdot 1_{\{\,16\,\}} - 10 \cdot 1_{\{\,-2,0,\ldots,15,17,\ldots,36\,\}}.$$

For example, if the ball ends up in slot 13, the values of these bets are given by

$$B_{\text{red}}(13) = 10 \cdot 1_{\mathcal{R}}(13) - 10 \cdot 1_{\mathcal{N}}(13) = 0 - 10 = -10$$
$$B_{16}(13) = 350 \cdot 1_{\{\,16\,\}}(13) - 10 \cdot 1_{\{\,-2,0,\ldots,15,17,\ldots,36\,\}}(13) = 0 - 10 = -10.$$

But if the ball ends up in slot 16, we have

$$B_{\text{red}}(16) = 10 \cdot 1_{\mathcal{R}}(16) - 10 \cdot 1_{\mathcal{N}}(16) = 10 - 0 = 10$$
$$B_{16}(16) = 350 \cdot 1_{\{\,16\,\}}(16) - 10 \cdot 1_{\{\,-2,0,\ldots,15,17,\ldots,36\,\}}(16) = 350 - 0 = 350.$$

10 The slot number trivially results from the identity function, the function that maps its argument directly to its return value.

11 Even if the casino only allows certain bets, you and I can bet on any combination, just between us. That applies even to sucker bets where you always give me one dollar.

12 In general, we also will allow a random variable to return a list of numbers, which is called a vector.

13 See Appendices S and F for more on sets and functions.

14 See Appendices S and F for more on sets and functions.

15 In these equations, $1_{\mathcal{A}}$, for any set $\mathcal{A}$, is the indicator function (see F.2.2) that returns 1 if its argument is in $\mathcal{A}$ and 0 if not. The equation for B_{red} simply states that $B_{\text{red}}(\omega)$ equals 10 if $\omega \in \mathcal{R}$ and equals -10 if $\omega \notin \mathcal{R}$.

Thought Question Define the sucker bet described in note 11 as a function B_{sucker} defined on Ω.

DESCRIBING THE RANDOMNESS. We have now built a probability model for a single spin of the roulette wheel, but to use this model in practice, we need some way to quantify the randomness. We could try to write down the physics of the situation, accounting for variables such as the velocity of the ball and the lubrication of the wheel, but it will be difficult to quantify the randomness this way. Remember the quantum lesson that the randomness is essential. We embrace it by *describing the randomness* rather than the physics, although the physics may inform our description.

Begin by considering the random mechanism described above. Because the tickets are well-mixed and physically indistinguishable, symmetry suggests that there is nothing special about any one ticket. It follows that all of the tickets (slots) are equally likely to be selected.

While this physical argument is intuitive (and correct), we have not yet precisely defined concepts like "likelihood" or "probability". In Chapter 2, we will develop these ideas in terms of the long-run average of a random variable over repeated spins of the wheel. If we spin the wheel again and again, we expect all 38 slots to come up roughly equally often. Hence, we will find that the long-run average value of a bet represented by a random variable B, which is denoted by $\mathsf{E}B$, is given by the number

$$\mathsf{E}B = \sum_{\omega \in \Omega} B(\omega) \cdot \frac{1}{38}.$$

We can think of this long-run average as the price we would have to pay so that both sides (casino and player) break even on average. If $\mathsf{E}B > 0$, then we win money on average and would need to pay up front to make things even. If $\mathsf{E}B < 0$, then we lose money, and the casino should pay us to make things even. (Don't hold your breath.) This also gives us a way to compare bets. For example, we find that $\mathsf{E}B_{\text{red}} = -10/19 = \mathsf{E}B_{16}$, so, in the long-run, there is little difference between the two plays. We also see that the casino has an edge, which is of course why they make money.

Thought Question Find B_{watch} that gives the dollar value of no bet (stealing chips not allowed) and compute $\mathsf{E}B_{\text{watch}}$.

4.2. Coin Flipping

Next, we consider a slightly less practical example: an infinite sequence of flips of a fair coin.[16] We call such a sequence a *Bernoulli sequence* after James Bernoulli, a seventeenth-century mathematician who helped develop the theory behind this example.

16 What makes a coin fair? Fairness is a basic symmetry between heads and tails in the coin and the flipping process. But certainly if I flip any coin, however well balanced, ten times, I will often get more heads than tails or vice versa. As we will explore more fully in later chapters, fairness can be defined in terms of long-run averages. A coin is fair if the relative frequency with which heads and tails come up approaches one as the sequence of flips gets longer and longer.

The Random Experiment. The random experiment we are studying here is simple to describe: we repeatedly flip a fair coin ad infinitum. When describing a random experiment, we want to be clear enough about the details so that someone else could run the experiment for us after reading our description. The idea behind this prescription is that a clear description of the experiment highlights the features of the system that we need to capture. There are, for example, important differences between the experiments of flipping a coin until the first heads appears and flipping a coin infinitely many times. So, we might wish to specify in addition that for our experiment the coin is to be flipped with the right thumb using an impulse between 1 and 5 Newtons/second, in a draft-free room with 12 foot high ceilings, ... Well, maybe not. Such details are only necessary to the degree that they help us use and interpret the model.

You might ask if this coin flipping example is actually a random experiment. For instance, how can it be run, let alone repeated, if it never ends? The answer is that it is often useful from a modeling perspective to idealize the experiment to make it easier both to study and to generalize to other contexts. While we do not ever need to consider a truly infinite sequence of flips, we may need to consider an arbitrarily long sequence of flips. (Consider for example the experiment in which we flip until the pattern HTHHTTTHTHTTHT appears for the first time.) Handling a collection of finite sequences of arbitrary length can turn out to be harder than handling infinite sequences. Even more importantly, this infinite coin flip model serves as a prototype for many other models that are not as easy to simplify.

Outcomes and Events. When a coin is flipped once, it can land heads or it can land tails. It can also land on edge in the slot at the base of the window or disappear down the sewer drain or get caught by the ceiling light fixture. But these unlikely possibilities simply confuse matters. A modeler's first task is to simplify the situation by assuming away those complications that one *thinks* will have little impact on one's decisions.[17] We will focus only on heads and tails.

We will represent each possible Bernoulli sequence by what is called a binary fraction.[18] A binary fraction is of the form $0.a_1a_2a_3\ldots$ where a_k is the result of the kth flip, 1 if it comes up heads and 0 if it comes up tails. It represents a number between 0 and 1 by the sum $\sum_{k=1}^{\infty} a_k 2^{-k}$. These are simply the binary analogue of the familiar decimal fractions; each place in the expansion relates the contribution of a particular power of 2.

As described in Section S.4, some numbers in $[0, 1]$ can be represented by two such binary fractions.[19] However, the set of sequences with a terminating expansion (ending all in 0's) is countable.[20] If we remove all of these sequences except 0, which has a unique expansion, then the set of remaining sequences, as binary fractions, can be explicitly identified with $[0, 1]$. How can we just ignore the terminating sequences, you might ask? This is another modeler's simplification: the sequences we removed collectively have a negligible chance to occur and can be safely ignored.[21] By removing them, we make our model easier to understand and use.

So, we are faced with a model in which we have an (uncountably) infinite number of tickets in our box. Fortunately, this does not bother us (yet). The information stored on each ticket is a number between 0 and 1 that can be analyzed by the SmartTicket Reader to find the result of any specified flips. Hence, in our mathematical model, the outcome space Ω is just the interval $[0, 1]$, the set of real numbers between 0 and 1. Each elementary outcome ω describes the results of all the flips and can be written in the form $\omega = 0.a_1a_2\ldots$. Given ω, we can read off any results of interest.

As in roulette, events correspond to sets of elementary outcomes. In this case, each event is a set of numbers whose binary digits satisfy certain properties.

Thought Question Describe the following events as subsets of Ω: (i) The first coin is a heads, (ii) The second coin is a heads, (iii) The first head appears on the fourth flip.

THE RANDOM MECHANISM. The random mechanism in this case again corresponds to picking one ticket from the box. We carry out the

17 Modeling is an iterative process, and such assumptions must be validated. If one turns out to have been optimistic, the model can be refined with more realistic assumptions.

18 See Cantor's diagonal argument in Section S.4.

19 Ex. $1/2 = 0.0111\ldots = 0.1000\ldots$

20 There are 2^n sequences that terminate after n digits. The function that maps terminating sequence $0.a_1\ldots a_n000\ldots$, with $a_n = 1$, to $\sum_{i=1}^{n} a_i 2^i$ is a one-to-one correspondence between this set of sequences and $\mathbb{Z}_+$.

21 This statement is true of any countable subset of sequences; we just applied it to the inconvenient ones.

infinite sequence of coin flips. The results specify ω bit by bit and in the end determine the selected elementary outcome ω^*.

RANDOM VARIABLES. The outcome space Ω and the elementary outcomes lie in the background, and the observable quantities – random variables – are what we work with instead. The coin-flipping model is an example of a very complicated outcome space on which we can define interesting random variables as explicit functions. Hopefully, this helps you understand the role of random variables in the model which carries over to situations in which we do not have such explicit expressions. Some interesting such functions in this case include the following for each $n \in \mathbb{Z}_+$:

1. An indicator of whether the nth flip comes up heads. This is a function[22] $H_n \colon \Omega \to \mathbb{R}$ defined for $\omega = 0.a_1a_2\ldots$ by

$$H_n(\omega) = a_n.$$

2. The number of heads seen up to and including the nth flip. This is a function[23] $S_n \colon \Omega \to \mathbb{R}$ defined for $\omega = 0.a_1a_2\ldots$ by

$$S_n(\omega) = \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} H_i(\omega).$$

3. The proportion, or frequency, of heads in the first n flips. This is a function $Q_n \colon \Omega \to \mathbb{R}$ defined by

$$Q_n(\omega) = \frac{1}{n} S_n(\omega).$$

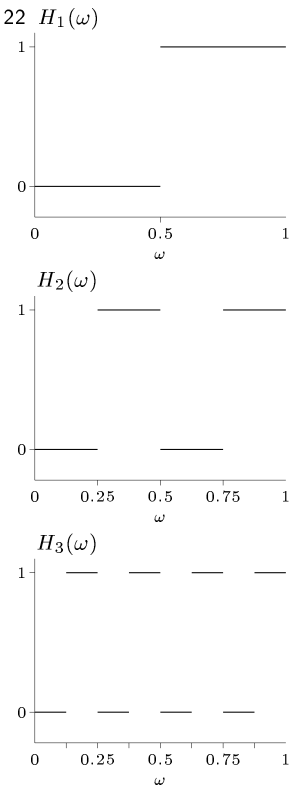
4. The amount by which the proportion of heads in the first n flips differs from $1/2$. This is a function $D_n \colon \Omega \to \mathbb{R}$ defined by

$$D_n(\omega) = \left| Q_n(\omega) - \frac{1}{2} \right|.$$

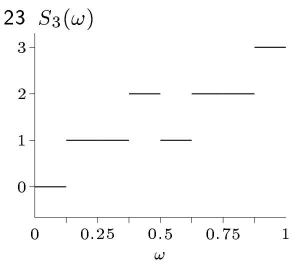
5. The number of consecutive heads starting at the nth flip, called the run length. This is a function $L_n \colon \Omega \to \mathbb{R}$ defined for $k \in \mathbb{Z}_\oplus$ by

$$L_n(\omega) = k \text{ if } H_n(\omega) = H_{n+1}(\omega) = \cdots = H_{n+k-1}(\omega) = 1$$
$$\text{and } H_{n+k}(\omega) = 0.$$

22



23



6. The number of flips until the first head appears. This is a function[24] $T: \Omega \to \mathbb{R}$ defined by

$$T(\omega) = k \text{ if } H_1(\omega) = H_2(\omega) = \cdots = H_{k-1}(\omega) = 0 \text{ and } H_k(\omega) = 1.$$

7. The number of flips until the pattern `head-tail-tail-head` first appears. This is a function $R: \Omega \to \mathbb{R}$ defined by

$$Z(\omega) = k \text{ if } H_{k-3}(\omega) = 1, H_{k-2}(\omega) = H_{k-1}(\omega) = 0, H_k(\omega) = 1.$$

Study these examples carefully to help you understand how a function represents an observable quantity. Notice, for instance, that some of the above functions are defined in terms of others in the list.

Thought Question What values do each of these functions take? In other words, what is their range?

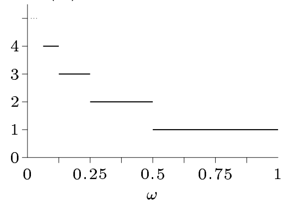
Describing the Randomness. In our tickets in a box abstraction, we hit a brick wall at full speed with this coin-flipping model. Fortunately, once we stop and shake off the dust, we will see that there is a way around. The problem is that there is an uncountable number of "tickets" in this outcome space $\Omega = [0, 1]$. We would like to be able to say something about the likelihood of choosing any ticket, as we did in the roulette example, but if we try to assign explicit positive probabilities to so many possibilities, we end up with a contradiction.[25]

To paraphrase Sherlock Holmes, if we are to assign probabilities in a consistent way, we must accept the inevitable fact that each individual elementary outcome in this Ω has zero probability of being selected. Shock. Disdain. Denial. Resignation.

The key to understanding this counterintuitive idea is to step back. There are simply too many elementary outcomes in Ω for us to enumerate them all. It is like trying to pick up a single water molecule with your hands; instead of doing so, we use a container to pick up a small drop of water. That is, instead of talking about the likelihood of individual elementary outcomes, we will instead consider bigger events. For example, looking at the pictures above and considering our notion of fairness, it seems intuitively necessary that the events that we get a heads on the nth flip for each n have a 50-50 chance of occurring.

To describe the randomness in this model, we will again use the idea of long-run averages. This means that we do the experiment again

24 $T(\omega)$



25 Specifically, the probabilities of most interesting events become infinite. Well-defined probabilities are between 0 and 1.

and again (flipping a coin infinitely many times, infinitely many times – ouch) and look at the average value of a random variable over the different runs.[26] It turns out that for any random variable X in this experiment, the long-run average of that random variable is given by

$$\mathsf{E}X = \int_0^1 X(\omega)\,d\omega.$$

We can view $\mathsf{E}X$ as a typical or representative value of the observable quantity embodied by X. Using this, we can compute typical values of the random variables we defined above.

Consider $\mathsf{E}H_n$ for any $n \in \mathbb{Z}_+$. What *should* this be on intuitive grounds? Since $\mathsf{E}H_n$ is the average of $H_n(\omega_i^*)$ over many random sequences $\omega_1^*, \omega_2^*, \ldots$, and since the coin is fair, we expect about as many heads as tails in any sequence of flips. In a long series of Bernoulli sequences $\omega_1^*, \omega_2^*, \ldots, \omega_M^*$, on average about half of the nth flips should be 1 and half should be 0. This suggests that $\mathsf{E}H_n = 1/2$ for any n.

The function H_n takes the value 1 on 2^{n-1} alternating intervals[27] of length 2^{-n} starting with $[2^{-n}, 2^{-n+1}[$, and it takes the value 0 on the 2^{n-1} alternating intervals of length 2^{-n}, starting with $[0, 2^{-n}[$. It follows that

$$\begin{aligned}\mathsf{E}H_n &= \int H_n(\omega)\,d\omega \\ &= 2^{n-1}\,2^{-n} \\ &= \frac{1}{2},\end{aligned}$$

as we had anticipated.

The function S_n is a bit more complicated, but we could compute an explicit expression for $S_n(\omega)$ to integrate if it were necessary. Fortunately, we have the relationship

$$S_n = H_1 + \cdots + H_n.$$

Notice that this is an equality between functions, which means that it holds for every possible value of the argument. In other words, no matter which sequence occurs, the observable quantity S_n (number of heads in first n flips) is equal to the sum of the observable quantities $H_1, \ldots, H_n$ (indicator of heads on each of first n flips). We thus have

$$\mathsf{E}S_n = \int S_n(\omega)\,d\omega$$

26 We are repeatedly drawing a ticket out of the box at random, recording information on that ticket, putting it back in the box, and then re-mixing the tickets.

27 The pictures of H_1 and H_2 help make this clearer. To see it mathematically, notice that $H_n(\omega) = 1$ if ω is of the form $0.a_1 \ldots a_{n-1} 1 a_{n+1} \ldots$. For any fixed $a_1, \ldots, a_{n-1}$, the set of ω for which $a_{n+1}, a_{n+2}, \ldots$ vary arbitrarily is the interval $[0.a_1 \ldots a_{n-1}011\ldots,$ $0.a_1 \ldots a_{n-1}111\ldots)$, which has length 2^{-n}. (Do not let the left endpoint confuse you; this interval was written using the non-terminating representation of ω.) As $a_1, \ldots, a_{n-1}$ vary arbitrarily (each taking values 0 or 1, of course), we have 2^{n-1} such intervals.

$$
\begin{aligned}
&= \int (H_1(\omega) + \cdots + H_1(\omega))\, d\omega \\
&= \sum_{i=1}^{n} \int H_i(\omega)\, d\omega \\
&= \sum_{i=1}^{n} \frac{1}{2} \\
&= \frac{n}{2},
\end{aligned}
$$

by the linearity properties of the integral. This says that for the typical sequence of flips, we get about half as many heads as tails.

Similarly, because of the relationship between Q_n and S_n, multiplication by a constant, we can write directly that

$$
\begin{aligned}
\mathsf{E}Q_n &= \int Q_n(\omega)\, d\omega \\
&= \int \frac{1}{n} S_n(\omega)\, d\omega \\
&= \frac{1}{n} \int S_n(\omega)\, d\omega \\
&= \frac{1}{2}.
\end{aligned}
$$

The function T offers a slightly different example. This takes the value k for all ω in the interval $]2^{-k}, 2^{-k+1}]$. Hence,

$$
\begin{aligned}
\mathsf{E}T_n &= \int T_n(\omega)\, d\omega \\
&= \sum_{k=1}^{\infty} \int_{]2^{-k},2^{-k+1}]} k\, d\omega \\
&= \sum_{k=1}^{\infty} k \int_{]2^{-k},2^{-k+1}]} d\omega \\
&= \sum_{k=1}^{\infty} k\, 2^{-k} \\
&= \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} 2^{-j} \\
&= \sum_{k=1}^{\infty} 2^{-k+1} \\
&= \sum_{k=0}^{\infty} 2^{-k}
\end{aligned}
$$

28 Write out the first few terms of both to see the pattern.

$$= 2.$$

Because the term 2^{-k} appears k times in the line above the double summation, it can be rewritten as a sum of sums, each starting later.[28] This means that we expect to get a head every two flips on average; since the coin is fair, this makes intuitive sense.

This approach to describing the randomness in the coin-flipping system gives exactly the intuitive results for all the variables we considered. We can test the approach in another way by showing that it preserves our idea of what a fair coin is. Note 16 defines fairness in terms of the sequences of coin flips that occur. Specifically, the coin is fair if for the "typical" sequence ω the relative frequency of heads and tails approaches 1 as we move out farther and farther in the sequence. For a typical sequence, we get both of the following equivalent results: (i) $Q_n(\omega) \to 1/2$ as $n \to \infty$ and (ii) $D_n(\omega) \to 0$ as $n \to \infty$. But what is a "typical" sequence?

To see that "typical" does not mean "every", consider a sequence of the form

$$\omega_{\text{bad}} = 0.a_1 11 a_4 11 a_7 11 a_{10} 11 \ldots,$$

for any $a_1, a_4, a_7, \ldots$. There are two 1's in every block of three digits, so for any $n \in \mathbb{Z}_+$, $Q_{3n}(\omega_{\text{bad}}) \geq 2/3$. This means that $Q_n(\omega_{\text{bad}})$ cannot go to $1/2$. In an uncountable set, there is plenty of room for exceptions; perhaps "typical" means "countable". Actually, no. Notice that as $a_1, a_4, a_7, \ldots$ the set of corresponding ω_{bad}'s is itself uncountable.

Define the event

$$\mathcal{N} = \left\{ \omega \in \Omega \text{ such that } Q_n(\omega) \to \frac{1}{2} \text{ as } n \to \infty \right\}.$$

If the randomly chosen ω^* lies in $\mathcal{N}$, then the observed sequence of flips will have the long run frequencies of heads and tails will both equal $1/2$. Otherwise, all bets are off. Let G be the random variable such that $G(\omega) = 1$ if $\omega \in \mathcal{N}$ and $G(\omega) = 0$ otherwise. It turns out that $\mathsf{E}G = 1$, meaning that the "typical" sequence is in $\mathcal{N}$. Thus, these bad sequences have no impact on what we observe and can thus be ignored completely. Our model thus makes consistent predictions with regard to fairness.

5. Random Experiments

The word "experiment" conjures the image of a dimly lit laboratory filled with an assortment of bubbling and sparking devices. A lone scientist sits huddled over the lab bench recording the data from painstaking measurements. Some critical question remains unanswered: What is the charge-to-mass ratio of the electron? What is the structure of DNA? What is the formula for Viagra? The scientist hopes that the data will reveal the answer.

But a single experiment by itself never provides a conclusive "answer". Because so much can go wrong, scientists rely instead on replication – repeating the experiment multiple times in a context as similar as possible to the first time. The more replications for which the conclusions hold, the more confidence we have in the results. The fundamental randomness we discussed above implies that the results of the experiment are different, even if only slightly, at every replication. This is even true of experiments in the physical sciences, which tend to give quite precise results.

An experiment is a controlled, repeatable situation for making careful measurements. Controlled means that any variables that might affect the measurements are either accounted for in the experiment or are held as constant as possible. Repeatable means that, in principle, someone else could do the experiment again with the same procedure. Part of the art of running an experiment is to adequately control the vast number of extraneous variables. In practice, it is impossible to do so, but with proper procedure we hope to minimize the effect of such variables. The challenge in repeating an experiment is that some (or all) of those extraneous variables might have changed when we run the experiment again. If they have changed too much, the experiment is qualitatively different and we cannot claim to have replicated the original.

Although we think of experimentation as a scientific enterprise – chemists, physicists, doctors, and psychologists do run experiments routinely – it is in fact much broader. For instance, a programmer attempting to find a bug in her employee's code might create two sets of input that differ in specific ways so that the difference in outputs will offer insights into the source of the problem. An industrial engineer overseeing a complex manufacturing process might systematically vary the conditions at the plant to find those that produce the best product. A marketing executive conducts a poll or focus group to evaluate products. All experiments are subject to randomness, but for many, the randomness is of such a magnitude that we must deal with it explicitly.

All probability models for random systems are founded on what is called a *random experiment*. Despite the fancy name, the idea is quite simple: A random experiment involves setting up a situation in a very specific way and then watching what happens. More precisely,

> **Definition.** A *random experiment* is a repeatable random process on which we can measure one or more quantities.

Any experiment is a random experiment, but a random experiment is somewhat more general because one need not control all the extraneous variables that affect the quantities to be measured. Indeed, sometimes we wish to understand how to predict the values of the measured quantities *in spite of* the extraneous variables. The word "random" also serves to emphasize the important role played by randomness in the system being studied.

The examples of the previous section, a single spin of the roulette wheel and an infinite sequence of coin flips, both represent random experiments. In each case, we can repeat the experiment simply (just spin or flip again) and there are a variety of possible measurements (the value of various bets or coin flips). In fact, almost any situation imaginable can serve as a random experiment. For example, the following all qualify: (i) go to a bus stop and wait until a bus comes, (ii) set up a target and shoot at it from a specified distance, (iii) run a computer system until it crashes, (iv) purchase a stock and watch it's price fluctuate, (v) start a chemical reaction and watch it progress.

A random experiment can involve any observable process but it must be specific. Imagine that you have an assistant who will do all the footwork for you. You write down in a notebook a description of the experiment and everything that is to be measured. The assistant then runs[29] the experiment and records the requested measurements. Your description must be precise enough that your assistant can run the

29 Running a random experiment does not necessarily mean doing anything other than observing the situation and recording measurements.

experiment unambiguously. That is,

> In describing a random experiment, you should provide enough specific details (units, procedures, etc.) so that someone else, reading your description from a piece of paper, could run the experiment for you, as you intend it to be run.

Of course, if we actually run the random experiment, then we get to see what happens. However, we often want to make predictions (e.g., whether the stock price will go up or down) or decisions (e.g., what bets to place on roulette or how to design the network to minimize crashes) *before* the random experiment takes place. It does no good to know that a stock price just went up because you cannot retroactively invest in that stock. It does no good to know that the ball was captured by slot 16 if we have not placed a bet. It does no good to know that the computer just crashed because we would rather prevent the crash in the first place. Probability theory is a tool for assessing the various outcomes of a random experiment *before* the outcomes of the experiment are known and using this knowledge to make effective predictions and decisions.

A schematic for how we use probability theory is as follows:

1. Design a random experiment. (That is, write down a specific situation clearly enough so that someone else could run the experiment for you.)
2. Decide what quantities (i.e., random variables) we are interested in measuring as we watch the experiment unfold.
3. Use probability theory to make predictions about the possible measurements we might obtain and how likely each possibility is.
4. Run the experiment and observe the actual outcome of our measurements.

Notice that we only run the experiment after we have specified how it is to be set up and made our predictions about what might happen. All of our calculations will be completed before the experiment is complete.

Thought Question Think of a situation in which we might need make a prediction using information that we obtain after the experiment has begun but while it is still running.

The random experiment is our basic description of the random system we wish to model. In specifying the details of the experiment, we must clearly state the assumptions we make, delineate which parts of the system we will model, and identify observable variables we wish to measure. The examples in the previous section did exactly this. We assumed that the roulette wheel and coin were fair; we decided to ignore possibilities such as fights in the casino or coins landing on their edge; and we identified several observable quantities of interest.

Any probability calculations we make will be done in the context of a specific random experiment. We must be clear about which experiment applies because differences in the experiment can change the interpretation of the observable quantities we are studying. For example, the following experiments are outwardly similar: flipping a coin once, flipping a coin ten times, and flipping a coin until the first head appears. However, the information available to us if we observe that the first flip is a heads can be used in vastly different ways in each case.

Sometimes we might also consider several random experiments and compare a specific prediction across the experiments. This is common, for example, when we are evaluating several strategies for action in the face of uncertainty. Each strategy represents a different manipulation of the system as the experiment unfolds. If these strategies must be chosen before the experiment begins, then it makes sense to consider these as different experiments and to then compare the efficacy of predictions based on each set of assumptions.

6. Outcomes and Events

Our presentation of the models for roulette and coin flipping started with a list of possible outcomes for the random experiment. We collected these possibilities in a set labeled Ω. In the first case, Ω contained thirty-eight elements because we had simplified away some of the more complicated possibilities in our description. In the second case, Ω contained an uncountable number of elements, but it still gave a vastly simplified list.

For a general probability model of a random experiment, we will denote the set of possible outcomes by Ω. We call this set the *outcome space* for our model of the experiment. Each $\omega \in \Omega$, one possible outcome, is called an *elementary outcome*. An elementary outcome describes one precise way in which the experiment can turn out. In terms of the Tickets in a Box abstraction, each elementary outcome corresponds to one ticket.

The content of the outcome space is a modeling decision; we need at minimum a description that allows us to represent everything we wish to measure. Sometimes, as in the roulette example, a very simple set will do, but there are observable quantities that we simply cannot represent by the model. For example, the total amount of money placed on any particular play is not represented in the outcome space with 38 elements, so the model we gave for roulette can say nothing about that. Very often, the process underlying a random experiment is inordinately complex and there is a staggering number of interesting potential measurements to consider. This suggests that we will have to expend a great deal of cleverness and energy to represent Ω properly.

Fortunately, this is not true. While the outcome space is a very important component of the model, it enters into our calculations in an implicit, abstract way. In most situations, we can proceed without explicitly writing down the structure of Ω. The reason we can do this is that mathematicians have shown that we can always find[30] an appropriate Ω. Here we present a way of thinking about Ω for general models that works in every situation we will encounter.

For any given random experiment, imagine *everything*, and that means everything, that can happen as that experiment unfolds. Let's reconsider the roulette example. We could record the exact path of the ball, the number of times the wheel spins, the types and amounts of bets placed by all the players, the number of decibels of the sound as the ball clicks into the slot, and so on. In fact, there is an infinite number of things that we *could* measure, an infinite number of things that actually happen when we run the experiment. Suppose we take all of the combinations for every possible measurement and list them out explicitly. Every combination on that list describes exactly what would be recorded for the entire set of possible measurements. We will let each combination on that list correspond to an elementary outcome $\omega \in \Omega$. Thus, *one* elementary outcome gives us all of the following (and more): a specific path for the ball $\gamma(t)$ in cm/s, a specific number of the slot that captures the ball (16), a specific number of revolutions

30 In fact, $[0, 1]$ will often do nicely because an uncountable number is a big number. Describe an algorithm by which you can encode a countable sequence of numbers in $[0, 1]$ within each number in $[0, 1]$. Can you think of a collection of information that you could not represent using the binary fractions in $[0, 1]$?

before the ball is captured (6.4), a specific initial angular velocity of the wheel (12 radians/s), and so on. This elementary outcome ω can be thought of as a tuple containing all this information:

$$\omega = (\gamma(t), 16, 6.4, 12, \ldots).$$

But that is only one way in which the experiment can happen. Other elementary outcomes on the list give different values for some or all of these measurements. Well, that's a big list to make, and in fact, we would never finish writing it out. So, instead of listing them out specifically, we just think about those possibilities in the abstract. Each elementary outcome $\omega \in \Omega$ gives a single combination on the list describing the values of all potential measurements, describes one specific way in which the random experiment might turn out, in all its complexity. The list of all elementary outcomes is usually unimaginably big, but we do not really need to write down. Hence, whenever you see an $\omega \in \Omega$, you can imagine that that ω contains all the information about one specific way in which an experiment occurred. This ω is not an observable quantity. Extracting information about specific observables is the job of *random variables*, which we discuss below.

Thought Question Consider the random experiment in which we perform an infinite sequence of roulette spins (analogously to the infinite sequence of coin flips). Find an explicit form for the outcome space Ω in this experiment. What advantage does it offer you to have this explicit form rather than using the catch-all version of Ω just described?

By thinking about each elementary outcome in this way, we have made it effectively impossible to ever write down the contents of any elementary outcome. But that's fine: the elementary outcome – indeed, the entire outcome space – is an abstraction that enables us to define the things we are interested in in a reasonable way. Suppose that you placed a bet in roulette that the ball would be captured by slot 16. When the spin is finished, you do not care what else happened, only whether or not the ball ends up in slot 16.

In other words, despite the myriad ways in which the experiment can unfold, you are only interested in one observable event. Among all the elementary outcomes, consider the set of them for which the ball ends up in slot 16. The different ω in this set can vary markedly in every other feature of the experiments – how many spins, the initial velocity of the ball, the total amount bet on red, and so forth – but they all share that one aspect. It follows that this set contains every possible way the experiment can unfold for which the event that the ball ends up in slot 16 occurs. In other words, that set of elementary outcomes *is* the event that the ball ends up in slot 16.

With this in mind, we define an *event* to be a subset of the outcome space, a set of elementary outcomes. We use events to specify sets of elementary outcomes for which some observable result occurs. For example, phrases like "the ball ends up in slot 16" or "there are 12 heads in the first 25 flips" describe events in the corresponding experiment. Each can occur in many different ways (that is, the contain many elementary outcomes) and each is defined by a specific observable outcome.

Because events are just sets, we name them with scripted capital letters like $\mathcal{E}, \mathcal{A}, \mathcal{U}$. Whenever something is called an event, you will know that it is simply a subset of the outcome space for the corresponding random experiment. However, in our study of probability theory, events will most often be defined by specifying conditions that certain observable results (random variables) must satisfy. The notation for this is important to understand and is described more fully in the next section.

You might be asking yourself the following questions: If elementary outcomes, as we will use them, contain so much information that we cannot write them down explicitly, then must it also be true that we can never specify the outcome space Ω explicitly? If both of these are true, then why do we need to think about Ω or ω at all? Moreover, if events can be specified with simple phrases, why do we need to define them mathematically at all?

These are good questions. For the first question, yes it is true that, as we will use them, outcome spaces and elementary outcomes will usually remain implicit. It is possible to construct a model with an explicit outcome space and well-specified elementary outcomes. We saw this in our simple model for roulette. However, this is rarely worthwhile. First, by specifying the outcome space, we limit the set of measurements we

31 We have not yet defined mathematically what it means for an event to occur, but this is a question of logic and language that can be answered without further mathematics.

can consider. Second, the specific nature of the outcome space is rarely of interest since elementary outcomes are by definition not observable. (Only random variables are observable, see below.) By keeping ω and Ω implicit and total, we save ourselves work, gain interpretability, and lose nothing.

For the second question, elementary outcomes are the tickets in the box, and Ω is the collection of all tickets. All other objects that we deal with – expected values, probabilities, random variables – are defined with respect to a particular Ω. Though they are implicit, ω and Ω are central to the mathematical abstraction that a probability model represents. We might care more about what is on the ticket than about the tickets themselves, but you cannot have one without the other.

For the third question, by representing observable events as sets, we have defined the medium and the basic logic under which we can work. For example, consider the thirty-eight events $\mathcal{S}_{00}, \mathcal{S}_0, \cdots, \mathcal{S}_{36}$ in the roulette experiment, where $\mathcal{S}_x$ is the event that the ball ends up in slot x. How many of these events can occur in any one spin of the wheel?[31] How can you express the event $\mathcal{R}$, namely that the ball ends up in a red slot, in terms of these events? What is the event that the ball ends up red and in the first 12? Think carefully about these questions before proceeding.

... Reader pausing to reflect ...

To see the correspondence between set operations and the day to day logic of events, we can see what kind of events are produced by the set operations. Below, we will let $\mathcal{E}$ and $\mathcal{F}$ be generic events (subsets of Ω). Consider the following:

> *Complements.* The set $\operatorname{compl}(\mathcal{E})$ is a subset of Ω and so is a well-defined event. It consists of all elementary outcomes not in $\mathcal{E}$; hence this is the event that the conditions defining $\mathcal{E}$ are *not* satisfied. We thus say that $\operatorname{compl}(\mathcal{E})$ is the event that $\mathcal{E}$ does not occur.

Unions. The set $\mathcal{E} \cup \mathcal{F}$ is a subset of Ω that consists of all elementary outcomes that are in $\mathcal{E}$ or $\mathcal{F}$ (and maybe both). Unions thus act like a logical or, and $\mathcal{E} \cup \mathcal{F}$ is the event that the conditions defining $\mathcal{E}$ or $\mathcal{F}$ or both are satisfied. We say that $\mathcal{E} \cup \mathcal{F}$ is the event that either $\mathcal{E}$ or $\mathcal{F}$ (or both) occurs.

Intersections. The set $\mathcal{E} \cup \mathcal{F}$ is a subset of Ω that consists of all elementary outcomes that are in both $\mathcal{E}$ and $\mathcal{F}$. Intersections thus act like a logical and, and $\mathcal{E} \cap \mathcal{F}$ is the event that both $\mathcal{E}$ and $\mathcal{F}$ occur.

There are other operations as well that are derived from these three. Take a shot at the following questions.

Thought Question

- Describe the meaning of the event obtained by taking the set difference $\mathcal{E} - \mathcal{F} = \mathcal{E} \cap \mathrm{compl}\,(\mathcal{F})$?
- How would you describe the event that corresponds to logical exclusive or of two events $\mathcal{E}$ and $\mathcal{F}$? That is, write an expression for the event that the conditions defining $\mathcal{E}$ or $\mathcal{F}$ but not both are satisfied.

Reasoning about events thus follows our intuitive logic, and we have the tools of set theory to help make that intuition precise. When examining the relationship between several events, we can usually think about an arbitrary elementary outcome and ask which of the events it belongs to.

For example, in day to day life, events are called *mutually exclusive* if at most one of them can occur. The roulette ball can only fall in one slot, for instance, so the events that it falls in slot 16 and slot 00 are mutually exclusive. To what does this property of mutual exclusivity correspond in the set abstraction? Suppose the events $\mathcal{E}$ and $\mathcal{F}$ are mutually exclusive. Now consider an arbitrary elementary outcome $\omega \in \mathcal{E}$. Suppose that ω were also in $\mathcal{F}$. Then, by definition, there would be at least one way in which the experiment can unfold such that the conditions that define both $\mathcal{E}$ and $\mathcal{F}$ are satisfied. This contradicts the stated mutual exclusivity of $\mathcal{E}$ and $\mathcal{F}$. Thus, if $\mathcal{E}$ and $\mathcal{F}$ are mutually exclusive and $\omega \in \mathcal{E}$, we know that $\omega \notin \mathcal{F}$. In other words, we have that

> Mutually exclusive events correspond to *disjoint* subsets of Ω.

(Remember that two sets are disjoint if they have no elements in common; in other words, their intersection is the empty set.)

Similarly, in day to day life, events are called *collectively exhaustive* if *at least one* of them *must* occur. We have assumed that the roulette ball must fall in some slot, so the events $\mathcal{S}_{00}, \mathcal{S}_0, \ldots, \mathcal{S}_{36}$ are collectively exhaustive. Suppose we have some collection $\mathcal{E}_i$ of collectively exhaustive events and consider an arbitrary $\omega \in \Omega$. Since one of the events must occur, every way that the experiment can unfold must satisfy the conditions defining one of these events. Thus, ω must be in at least one of the $\mathcal{E}_i$. It follows that $\bigcup_i \mathcal{E}_i = \Omega$.

> Collectively exhaustive events correspond to subsets of Ω whose union is Ω.

The same type of argument – reasoning about individual ωs – works in a wide range of cases.

7. The Random Mechanism

Most of the work in describing the random experiment goes into thinking about the possible outcomes of the experiment. This is the information on the tickets, so to speak. Once we have that our random mechanism is simple: we allow nature to choose one ticket from the box at random. In the roulette example, all the tickets were indistinguishable and were thus equally likely to be selected. In the coin-flipping example, we could not quite think about things in the same way because there were two many tickets, but by thinking about averages we got intuitive results. Other simple games of chance like cards and dice can be modeled similarly, although in cases such as rolling two dice, the basic outcomes (i.e., $2, 3, \ldots, 12$) need not all have the same chance of occurring. We can model this in several ways. For example, we can put 36 tickets in the box, all physically indistinguishable and each with a pair of integers from 1 through 6 describing the upward face on the two dice. Or we could put 11 tickets in the box and make them of different size or weight, thus adjusting the chance that each is selected. All of this changes how we quantify the randomness, but it does not affect the basic mechanism of selecting a ticket at random.

> **Fundamental Mechanism.** Each time a random experiment is run, nature randomly chooses one elementary outcome ω^*. This ω^* is called the selected elementary outcome. We say that ω^* occurred.

Once we have ω^*, we can determine whether any particular event has occurred. Specifically, if $\mathcal{A} \subset \Omega$ is an event, we say that $\mathcal{A}$ occurred (in this run of the experiment) if $\omega^* \in \mathcal{A}$. In other words, the selected elementary outcome satisfies the defining condition of the event.

> **Definition.** An event $\mathcal{A}$ is said to have occurred in the experiment if the selected elementary outcome ω^* belongs to $\mathcal{A}$.

The selected elementary outcome also determines the observed value of all random variables in the model, as we now discuss.

8. Random Variables

A *random variable* is any quantity that we measure during a random experiment. Random variables are the basic objects with which we work in probability theory. The dollar value of all of the different bets in roulette are random variables as are the observable quantities H_n, S_n, and T in the coin-flipping experiment. In the other random experiments mentioned on page 20, the following are random variables that might be measured: (i) the time in minutes until the first bus arrives, the number of passengers who get on the bus, or the distance on the first bus's odometer when it arrives; (ii) the distance in cm from the bullet's landing place to the center of the target; (iii) the time in seconds until the computer system crashes; (iv) the price in US$ of the stock one day after purchase, the price of the stock two days after purchase, the price of the stock 5 days after purchase; (v) the time in milliseconds it takes until the reaction is complete, the weight in grams of the reaction's product, the amount of energy in joules produced by the reaction. Notice that for every random experiment we can measure many different quantities; some of them are related to each other, some are not.

Each of these is an observable quantities that we could measure if we were to run the corresponding random experiments. We call them "random" variables to emphasize the fact their observed value can be different on different runs of the same experiment. The value we measure is the result of a random process, and until the experiment is run, the value is uncertain. We use probability theory to understand what values of the random variable are possible and how likely each possible value is.

The most important thing to remember about random variables is the following:

> **Critical Connection:** Every random variable we define corresponds to a measurement that we could take during the random experiment and vice versa.

This connects the abstract mathematical framework to a real dirt and sweat experiment. We will spend a lot of time and effort defining and using random variables as mathematical objects,[32] and our mathematical tools will make this easy. Maintaining this connection between the mathematical and the practical allows your intuition about the system to help guide the mathematics and allows your mathematical results to help guide your intuition.

32 On a philosophical note, you will often notice that we define random variables that would be very difficult to measure in practice. For example, we could define a random variable to describe the initial velocity of a roulette ball. Chances are that there is not a casino in the land that would allow you to measure this. So, does this invalidate anything we do with such a random variable? No. Our predictions and decisions will be based on random variables that we work with in practice, measurements that are both interesting and accessible. We lose nothing if in doing our calculations we consider other, harder to measure quantities.

The second most important thing to remember about random variables is their mathematical definition. A random variable is a *function* whose domain is the outcome space Ω. This definition reflects the fact that the observable quantity can take different values, depending on which ω actually occurs, but that each ω determines exactly the value of that observable. A random variable plays the same role as the SmartTicket Reader when it computes and displays a particular quantity of interest using the selected ticket. We always denote random variables by capital, unslanted Latin letters.

All the terminology of functions we defined in Appendix F carries over to random variables. We add, multiply, compose, compare, and integrate random variables. We speak of the range of a random variable or say that it "takes a value". We talk about non-negative random variables or write $X \geq Y$. We define events in terms of the possible values of a random variable. (See below.)

Nonetheless, in most situations, our notation is designed so that manipulating random variables will seem just like manipulating familiar algebraic variables. If X is a random variable we are working with, almost every reference to it will be to the function itself rather than the value the function takes on a particular argument – X rather than $X(\omega)$. Statements made about X refer to the function itself, the function as an object,[33] and must therefore hold *for every value of the function's argument*. Thus, if we write $X = Y$, we mean this as an equality between functions; that is, $X(\omega) = Y(\omega)$ for *every* possible argument to the function. The same goes for statements like $X \geq Y$ or $Z = X + Y$ which mean $X(\omega) \geq Y(\omega)$ or $Z(\omega) = X(\omega) \cdot Y(\omega)$ for all $\omega \in \Omega$. In the coin-flipping example, we defined random variables $H_1, H_2, \ldots$ and $S_1, S_2, \ldots$, and we saw that

$$S_n = H_1 + \cdots + H_n,$$

which again means that for every $\omega \in \Omega$, $S_n(\omega) = H_1(\omega) + \cdots + H_n(\omega)$. The distinction between an algebraic variable and a random variable will be significant when we consider properties of the random variable that hold only for *some* elementary outcomes. We cannot, for instance, say that $S_4 > S_3$ or $S_4 \leq S_3$ in the coin-flipping example; both statements are false. However, there are ω for which $S_4(\omega) > S_3(\omega)$, and there are ω for which $S_4(\omega) \leq S_3(\omega)$.

Similarly, we have the following general rule:

Mapping Rule
A function of a random variable is itself a random variable.

Suppose for example that $X\colon \Omega \to \mathbb{R}$ is a random variable and $f\colon \mathbb{R} \to \mathcal{Y}$ is some function defined on the range of X. Then, the composition $f \circ X$ is just a function $\Omega \to \mathcal{Y}$, which maps $\omega \in \Omega$ to $f(X(\omega))$. This is by definition a random variable because it is a function defined on the outcome space. We will usually write this composition $f(X)$, and if f has an explicit form like $f(x) = x^2$ or $f(x) = e^x$ we write X^2 or e^X.

33 Think of "random variable" as a class in an object oriented programming language, a subclass of functions of a single variable. It has methods for assignment, comparison, and so forth. We will explore this connection further in class.

These latter are just the functions that map $\omega \in \Omega$ to $(X(\omega))^2$ and $e^{X(\omega)}$. Notice, however, we keep the argument (ω) implicit when we are speaking about the function as a whole.

Thought Question Referring to Section 4, let S_3 and H_3 be the random variables defined for the coin-flipping experiment. Derive and graph the following random variables: $2S_3^2$, $\sin(\pi\, S_3)$, and $S_3 - H_3$.

We will primarily use two classes of random variables: real-valued random variables, which are functions $\Omega \to \mathbb{R}$, and vector-valued random variables, which are functions $\Omega \to \mathbb{R}^k$ for some $k > 1$. All the random variables defined in Section 4, for example, (e.g., B_{red} and T) are real-valued random variables. Since any point in $\mathbb{R}^k$ is a k-tuple of real numbers, a vector valued random variable $X\colon \Omega \to \mathbb{R}^k$ can always be written in the form[34] $X = (X_1, \ldots, X_k)$, where each X_i is a real-valued random variable. Conversely, if $Y_1, Y_2, \ldots, Y_k$ are real-valued random variables, we can define a vector-valued random variable $Y\colon \Omega \to \mathbb{R}^k$ by

$$Y = (Y_1, \ldots, Y_k).$$

Thus, vector-valued random variables are just tuples (vectors) of real-valued random variables. They are just convenient packages for dealing all at once with a collection of related random variables.

Since in most models the set Ω is only defined implicitly, we will often be unable to give a mathematical expression for our random variables. In these cases, we define the random variable with language, saying, for example, that "H_n is the indicator of a heads on the nth flip". In order to define a random variable properly, we must describe the observable value clearly enough so that someone else (e.g., your assistant) could measure the value of the variable for you. Imagine that in specifying your random experiment, you write down a list of observable quantities to measure[35] and give specific instructions for how they are to be recorded (e.g., in what units). You then hand this list to your assistant who runs the experiment and records those values in a notebook, exactly as you instructed. You will receive a lot of practice doing this type of description.

Two of the simplest and most important types of random variables we will encounter are constants and indicator random variables. A constant random variable is any function $\Omega \to \mathbb{R}$ that takes the same value

34 This equality between vector-valued functions means that for any ω,

$$X(\omega) = (X_1(\omega), \ldots, X_k(\omega)).$$

35 The word measurement here is used loosely; even an observation with a pair of eyes counts as a measurement.

for every $\omega \in \Omega$.[36] Thus, 1, 2, 0, c for some constant c are all random variables; of course, they are much simpler to deal with than many other variables because their value is always the same. These are deterministic random variables! An indicator random variable is just an indicator function $1_{\mathcal{E}}$ for some event $\mathcal{E} \subset \Omega$. This random variable takes the value 1 for $\omega \in \mathcal{E}$ and 0 for $\omega \notin \mathcal{E}$. For short, we call this the "indicator of $\mathcal{E}$". Notice that the indicator of Ω, 1_Ω, is just the constant function 1 by definition.

36 And consequently for every repetition of the random experiment.

For most of the events we will use, the conditions that define those events are stated in terms of the values of particular random variables. Consider the random variable B_{red} in the roulette example which represents the value of a \$10 bet on red. The event that the bet is won is the set of elementary outcomes for which B_{red} takes the value 10; that is,[37]

37 This event is, by definition, just the inverse image of $\{10\} \subset \mathbb{R}$ under the function B_{red}.

$$\{\omega \in \Omega \text{ such that } B_{\text{red}}(\omega) = 10\}\,.$$

When we write such events, we drop the $\omega's$ and Ω's and leave them implicit; we write the above as

$$\{\, B_{\text{red}} = 10 \,\}\,.$$

The two forms are exactly equivalent, but the second form is preferable for two reasons: (i) we need not write $\omega \in \Omega$ etc. over and over and over again, and (ii) the meaning of the event is not obscured by needless notation.

So, we describe most events as follows:

$$\{\text{condition on the values of a random variable}\}.$$

This condition can take a variety of forms: $\{\, X > 10 \,\}$, $\{Y = 10 \text{ and } Z < 14\}$, $\{U \in [0,1] \;\text{ or } U > 100\}$, and so forth. In general, the expression

$$\{\,\text{condition about } Z\,\}$$

is a shorthand for

$$\{\,\omega \in \Omega \text{ such that condition about } Z \text{ is true}\,\}\,.$$

Whenever you see such a condition surrounded by braces, you are dealing with an event, a set of elementary outcomes. We use braces to remind us that these are sets.

Thought Question Referring to Section 4, let S_3 and H_3 be the random variables defined for the coin-flipping experiment. Derive and graph the indicator random variable $1_{\{S_3>1\}}$.

Thought Question What is the difference between the following statements:

- $X > 10$,
- $\{X > 10\}$,
- $X(\omega) > 10$.

When we talk about X as a function or refer to $X(\omega)$, we are considering the values that the random variable can take for generic elementary outcomes. However, when we run our experiment, we will eventually observe one and only one value for that measurable quantity. This is $X(\omega^*)$, the function X evaluated at the selected elementary outcome ω^*. In most cases, we will not be able to use $X(\omega^*)$, though we may refer to it, because when we are making our predictions, ω^* has not yet been selected.

One of the tricky distinctions relating to random variables is that we use the name of random variable in three ways:

- As a function, e.g., X.
- As the value a function takes on a given argument ω, e.g., $X(\omega)$.
- As the value of a measurement that we take when the experiment is run, e.g., $X(\omega^*)$.

The difference between the first two is an issue of "units analysis"; X is a function, $X(\omega)$ a number.[38] If you find yourself writing an equation that mixes these two things, you can detect the mistake by asking if both sides of the equation are of the correct type. An expression of the form $X = X(\omega)$, for instance, does not make sense since one side is a function and the other a number.[39] The difference between the second two is more subtle and depends on context. When we are working with a random variable as a function, we will always use ω as a generic elementary outcome and ω^* as the selected elementary outcome. When we are referring to the value we actually measure, we will say $X(\omega^*)$. This is only one possible value that the random variable can take, but it is the one we will eventually observe. we will suppose that the particular elementary outcome ω^* was selected.

38 For a real-valued random variable. It would be a vector of numbers for a vector-valued random variable.

39 Assume that the compiler (i.e., the grader) does *not* automatically promote the type of $X(\omega)$ to a a constant function. Even if it did, of course, the equation would still be wrong since now the two sides are not equal as functions.

Thought Question Consider the following sentence: "This random variable takes the value 1 if the event $\mathcal{E}$ occurs when the experiment is run and takes the value 0 if the event $\mathcal{E}$ occurs when the experiment is run." Explain how this sentence blurs the distinction between $X(\omega)$ and $X(\omega^*)$ and use your knowledge of indicators to define $1_{\mathcal{E}}$ more properly.

9. Describing the Randomness

The only step remaining in our description of a probability model is a way to quantify the randomness in the system. The examples above have hinted that we will do this by considering the long-run averages of random variables as we repeat the experiment again and again. This is indeed the case, and we develop these ideas fully in the next chapter.

2

LONG-RUN THINKING: EXPECTED VALUES AND PROBABILITY

In the long run, we're all dead.
— John Maynard Keynes

1. What is an average?

We begin with a mundanely familiar idea – the average. If you are given a list of numbers, say 10, 4, -6, 21, 115, and 0, it is a simple matter to compute the average of the list: just add up the numbers (144) and divide by the length of the list (6). But what is the meaning of the value (24) that we get from this procedure?

The average is a summary of the list that gives a central location for the numbers. Even though the numbers in the list can vary – and widely – the average is chosen to be simultaneously as close as possible (in some sense) to all of them. Even though the average need not be one of the numbers in the list, it acts as a representative or typical value for those numbers. We talk of the average consumer, the average test score, the average family size because these allow us to form a general picture without all the complexity of the details.

Consider three examples of averages from everyday life:

- A baseball player's batting average.
- The Standard & Poors 500 Index.
- The averages of homework and exam scores used to compute a student's grade in a class.

In the first example, the average is computed from a list of 0s and 1s, where a 1 is recorded for each at bat in which the batter gets a base hit and a 0 is recorded otherwise. The batting average is the *proportion* of hits to "at bats", that is, the ratio of the number of hits to the number of at bats. *Averaging a list of 0s and 1s always gives a proportion, a value between 0 and 1.*

In the second example, the average summarizes the prices[1] of the 500 listed companies, but the companies do not contribute equally to the average. This is an example of a *weighted average.*[2] The weight of each company's contribution to the index increases with its market capitalization.[3] If it were a standard arithmetic average, each company's price would be weighted by one over the number of companies (500). Instead, it is weighted by each company's market capitalization over the total market capitalization of all 500 companies. *An average need not weight all numbers in the list equally.*

1 Adjusted for technicalities like stock splits and mergers and such.

2 This is a bit of an ill-fitting term. The standard arithmetic average is weighted; it is just *equally* weighted.

3 The total value of a company's outstanding stock.

In the third example, several different averages are used. Suppose that the student was assigned ten homeworks and given three in-class exams and a final exam. The professor computes three separate averages for the homeworks, the in-class exams, and the final, all of which were computed using equal weights within each category. (Of course the list of final exam scores has only one number in it, so the average is the number itself.) The final grade, however, is a weighted average where the homework average, in-class exam average, and final exam (average) are weighted respectively by 0.25, 0.4, and 0.35. *The average of averages is an average.*

There is a physical interpretation of an average. Consider the numbers in our list as representing the coordinates along a massless line where we place weights of equal mass. The average is the balancing point for these weights – the position of a fulcrum that balances the line of weights. No matter where we place the weights, the fulcrum can never be outside the range of the weights (otherwise, the line could not balance). The analogue of an unequally-weighted average is an arrangements of unequal weights at the coordinates of the different numbers on the list. This weighted average is again the position of the fulcrum. The weight in the average assigned to each number on the list is the weight at that position over the total weight of the line.[4] From this pic-

4 Notice that the weight can never be negative. It can however be zero if we do not put any weight at that point on the line.

ture, we have the following property of an average (equally weighted or not): *An average always lies between the smallest and biggest numbers in the list.*

Let's formalize these ideas mathematically. Define the averaging operator A as follows:

$$A_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} \frac{1}{n} x_i.$$

This gives the (equally-weighted average) for the list of numbers $x_1, \ldots, x_n$. If $w = (w_1, \ldots, w_n)$ is a list of numbers[5] such that[6] $\sum_{i=1}^{n} w_i = 1$, then the averaging operator with weights w is given by

$$A_{n,w}(x_1, \ldots, x_n) = \sum_{i=1}^{n} w_i\, x_i.$$

Notice that $A_{n,w} = A_n$ if $w = (1/n, \ldots, 1/n)$, so we need only work with $A_{n,w}$. If we also write the list of numbers in vector notation as $x = (x_1, \ldots, x_n)$, then we have

$$A_{n,w}(x) = \sum_{i=1}^{n} w_i\, x_i.$$

If we have two lists of numbers $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, we can multiply all the numbers in a list by the same constant $c \in \mathbb{R}$ to get a new list $cx = (cx_1, \ldots, cx_n)$, or we can sum the lists elementwise getting a new list $x + y = (x_1 + y_1, \ldots, x_n + y_n)$. Averages behave in the way we might expect under these operations. We have

$$\begin{aligned} A_{n,w}(cx) &= \sum_{i=1}^{n} w_i c x_i = c \sum_{i=1}^{n} w_i x_i \\ &= c A_{n,w}(x) \\ A_{n,w}(x+y) &= \sum_{i=1}^{n} w_i (x_i + y_i) = \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i y_i \\ &= A_{n,w}(x) + A_{n,w}(y). \end{aligned}$$

Thus, *scaling all the numbers by the same amount scales the average in the same way*, and *adding two lists together gives a list whose average is the sum of the averages of the original list.* The point raised above about the balancing point being between the extremes can be expressed as

$$\min_{1 \le i \le n} x_i \le A_{n,w}(x) \le \max_{1 \le i \le n} x_i.$$

5 This is an example of a vector. See Appendix S Section 2.4 for details.

6 If we have weights $v_1, \ldots, v_n$ that do not sum to 1, we can construct weights for the average by

$$w_i = \frac{v_i}{\sum_{j=1}^{n} v_j}.$$

This is, for example, how the weights in the S&P 500 Index are computed where v_i is the market capitalization of the ith company.

Taken together, we have the following properties of averages:

- The average of a list need not be one of the numbers on the list.
- The average always lies between the smallest and biggest numbers in the list.[7]
- If we multiply all the numbers in the list by the same constant to make a new list, then the average of the new list equals that constant times the average of the old list.
- If we take two lists of the same length and add corresponding entries together to make a new list, then the average of the new list is the sum of the averages of the two old lists.

7 It follows that if all of the numbers on the list are non-negative, so is the average.

Almost everything we do in this course can be reduced to computing averages, the only catch is that we will not have direct access to the lists of numbers. These properties will thus become important – and frequently recurrent – tools.

2. Expected Values Are Long-Run Averages over Replications of an Experiment

You stand at the gaming table, chips in hand. With the flick of a wrist, you can risk your money on the slim chance of winning the big payoff. Is the potential reward worth the risk? How much should you bet? How much can you expect to win? Should you play at all?

What does it mean to say that some result is likely or that we expect to win some amount on average? When the game is over, money changes hands, whether the outcome was likely or not. You could win the big money, or you could lose everything, or something in between. Anything that could happen in a random experiment, might.

It is unclear, then, how to weight the different possible outcomes when making a decision in the face of randomness. What we need is some way to describe and quantify randomness. Our approach in these notes will use long-run averages of random variables when we replicate an experiment many times.

To *replicate* an experiment is to repeat it[8] so that (i) the setting is, for all practical purposes, identical to that of earlier runs, and (ii) the outcome is independent[9] of what happens in any other runs of the experiment. When a random experiment is replicated, the outcome can

8 When we defined random experiments, we emphasized that the random process being modeled must, in principle at least, be repeatable.

9 This independence of the replication means that knowing what happened one run of the experiment gives you no new information that can help you predict what will happen in another run.

vary. In particular, the observable quantities (i.e., random variables) in that experiment will take on different values.

The roulette example from Chapter 1 provides a good illustration. The random experiment here is a single spin of the wheel, and you can place a variety of possible bets. Before placing a bet, however, you would like to determine the best bet to make. But best in what sense? Certainly you wish to maximize your winnings, but with the ball capable of landing in any slot, no bet is guaranteed to win on this single spin. Long-run averages arise here naturally as a way to make this kind of decision. Imagine the following hypothetical situation: you will place an identical bet on a single roulette spin every day for the rest of your life, but you must decide *now* which bet that will be. Since you will be repeating the experiment over and over, your winnings are no longer as sensitive to the results of a single spin. Instead, you want to choose that bet that wins as much as possible *on average*.

Let's take a closer look at what this averaging means in a less hypothetical situation. In any casino, our roulette experiment is replicated day after day with every spin of the wheel. Although some of the details of the experiment (e.g., who has bet and how much) change from spin to spin, the interesting features – the results of the spin – do not change; the casino provides a guarantee that one spin is as good as another.

Suppose you enter the casino with $10 to bet. As you take your seat at the roulette table, you notice the woman seated next to you studying a piece of paper that shows the winning slots over the past 100 spins.

13	19	25	0	00	15	23	11	34	26	12	32	8	21	36	15	27	33	24	36
35	33	14	0	0	18	00	31	34	00	0	26	26	17	15	3	13	22	23	12
15	33	27	27	5	22	0	34	6	25	34	9	31	27	28	0	7	3	17	9
4	36	29	21	17	19	16	00	20	22	2	8	32	00	12	4	26	20	31	14
10	16	34	13	27	00	8	23	23	27	27	31	20	16	3	13	1	30	13	00

How do you use this information to evaluate the available plays?

Suppose, to begin with, that you saw only the last spin (00) on the woman's paper and process it into the following table describing how each of the nine standard plays fared on that spin.

Name of Play (bet)	Proportion of Wins	Average Value of a $10 bet
Even Money (red)	0	-10
Dozen (3rd 12)	0	-10
Column (1,4,7,...)	0	-10
Line (19–24)	0	-10
House Special	1	60
Quarter (2,3,5,6)	0	-10
Street (34–36)	0	-10
Split (8,11)	0	-10
Straight (00)	1	350

As the table shows, a Straight Play on 00 fared best, and the House Special is the only other bet with positive return. What does this information suggest you do on the next spin? If taken seriously, the table suggests making another Straight Play on 00. But there is something unsatisfying about this suggestion: when the wheel is next spun, the ball will very likely fall into a slot other than 00. Moreover, the fact that 00 occurred on the last spin should have no effect on what happens next. The problem is that the outcome of a single spin is not very representative of what can occur on a typical spin and therefore provides little useful information for making a decision.

Suppose now that you (quickly) used all 100 spins listed on the woman's paper to produce the following table.

Name of Play (bet)	Proportion of Wins	Average Value of a $10 bet
Even Money (red)	0.45	-1.00
Dozen (3rd 12)	0.34	0.20
Column (1,4,7,...)	0.30	-1.00
Line (19–24)	0.15	-1.00
House Special	0.18	2.60
Quarter (2,3,5,6)	0.06	-4.60
Street (34–36)	0.09	0.80
Split (8,11)	0.04	-2.80
Straight (00)	0.07	15.20

With one hundred spins, each slot gets several chances to win, and our assessments of the different plays are more balanced and less sensitive to the outcome of any one spin. Interestingly, the Straight play and the House Special continued to perform well in this set of spins, although the differences among the plays are smaller and two other bets had positive average return. The entries in this table are more representative than those in the previous table of what can happen on a typical spin. However, even the results with 100 spins are subject to chance fluctuations that make it harder to use this table for decision making. For example, one or two 00s beyond what might be expected in one hundred spins are sufficient to give that Straight play a (large) positive average return. The odds of seeing a chance fluctuation of that magnitude in one hundred spins are not too long.

This suggests that we look at the averages over many more spins. As the number of spins increases, short-lived chance fluctuations in the frequency with which any slot or group of slots occur have a decreasing impact on the average results. The following is a table of the same form as the above but based on 100,000 past spins.

Name of Play (bet)	Proportion of Wins	Average Value of a $10 bet
Even Money (red)	0.47255	-0.55
Dozen (3rd 12)	0.31396	-0.58
Column (1,4,7,...)	0.31603	-0.52
Line (19–24)	0.15884	-0.47
House Special	0.13257	-0.72
Quarter (2,3,5,6)	0.10534	-0.52
Street (34–36)	0.07877	-0.55
Split (8,11)	0.05205	-0.63
Straight (00)	0.02613	-0.59

Now all of the plays, except for the House Special, appear yield comparable average gains. (The winning proportions are very different, but remember that the payoff on different plays also differ.)

As we use more and more spins to compute the table, the proportion of wins and the average value of a $10 bet become more and more representative of a typical spin. With one or even one hundred spins, chance fluctuations lead certain slots to be preferred. But as the number of spins grows larger, the averages become less sensitive to small differences in the frequencies of the slots, and the effects chance fluctuations go to zero. Eventually, in the limit of an infinite number of spins, we get the following table:

Name of Play (bet)	Proportion of Wins	Average Value of a \$10 bet
Even Money (red)	$18/38 \approx 0.474$	$-20/38 \approx -0.526$
Dozen (3rd 12)	$12/38 \approx 0.316$	$-20/38 \approx -0.526$
Column (1,4,7,...)	$12/38 \approx 0.316$	$-20/38 \approx -0.526$
Line (19–24)	$6/38 \approx 0.158$	$-20/38 \approx -0.526$
House Special	$5/38 \approx 0.132$	$-30/38 \approx -0.789$
Quarter (2,3,5,6)	$4/38 \approx 0.105$	$-20/38 \approx -0.526$
Street (34–36)	$3/38 \approx 0.079$	$-20/38 \approx -0.526$
Split (8,11)	$2/38 \approx 0.053$	$-20/38 \approx -0.526$
Straight (00)	$1/38 \approx 0.026$	$-20/38 \approx -0.526$

So, in terms of long-run average payoff, all of the plays except the House Special are identical. The "House Special" certainly is special – for the house – so you may wish to avoid it. Interestingly, from the player's viewpoint the "special" quality of the latter arises from insufficient payoff odds relative to the other plays; it should pay 6.2 to 1 instead of 6 to 1 to balance the 5 opportunities it gives to win.

This final table indicates the average winnings per bet that we can secure if we play long enough. The fact that these average winnings are negative for all these plays tells us what we already knew at some level – the game is stacked in the casino's favor, if only slightly. We can think of this as a %5.26 charge per bet (or %7.89 for the House Special) applied by the casino. To make this a fair game the casino should pay you that %5.26 in advance to compensate for the average loss you will accrue. The long-run averages in the table therefore provide an assessment of the fair value of each these bets. Or to put it the other way, they give the fair price you should pay (or be paid if negative) to play the game.

We have been considering many spins of the roulette wheel, but for our purposes, we view these as replications of the random experiment in which the wheel is spun once. In terms of the tickets-in-a-box abstraction from Chapter 1, replicating the experiment ad infinitum corresponds to the following steps:

1. Draw a ticket,
2. Read and record the values of any interesting quantities from information stored on the ticket,
3. Replace the ticket in the box and re-mix the tickets.
4. Goto step 1.

In our probability model, each drawn ticket corresponds to an elementary outcome in the outcome space Ω chosen by nature. Each observable quantity of interest corresponds to a random variable, and measuring that quantity in the experiment corresponds to reading and recording the value of that variable. Hence, replicating an experiment over and over yields a sequence of elementary outcomes $\omega_1^*, \omega_2^*, \omega_3^*, \ldots \in \Omega$ chosen by the same random mechanism. For any random variable of interest X, this generates a list of numbers $X(\omega_1^*), X(\omega_2^*), X(\omega_3^*), \ldots \in \Omega$. The long-run averages are computed from that list.

Suppose we would like to assess the value of a specific \$10 play in the roulette example. As we saw in Chapter 1, the value of this play corresponds to some random variable B, and the indicator of whether this play wins corresponds to a random variable W. In forming the first three tables above, we computed the following averages

$$B(\omega_1^*), \qquad \frac{B(\omega_1^*) + \cdots + B(\omega_{100}^*)}{100}, \qquad \frac{B(\omega_1^*) + \cdots + B(\omega_{100000}^*)}{100000}$$

$$W(\omega_1^*), \qquad \frac{W(\omega_1^*) + \cdots + W(\omega_{100}^*)}{100}, \qquad \frac{W(\omega_1^*) + \cdots + W(\omega_{100000}^*)}{100000}.$$

The numbers $B(\omega_1^*), B(\omega_2^*), \ldots$ are the gains (and losses) from the specified bet on successive spins of the wheel. The averages on the first line give increasingly representative values for that bet. Similarly, the numbers $W(\omega_1^*), W(\omega_2^*), \ldots$ are 0s (for losses) and 1s (for wins) for the same sequence of spins. The averages on the second line are consequently proportions. The closer they are to 1, the more often the specified play won; the closer they are to 0, the more often the specified play lost.

As the number of replications goes to infinity, these averages – these *long-run averages* – converge to constant values:

$$\mathsf{E}B = \lim_{n\to\infty} \frac{B(\omega_1^*) + \cdots + B(\omega_n^*)}{n}$$
$$\mathsf{E}W = \lim_{n\to\infty} \frac{W(\omega_1^*) + \cdots + W(\omega_n^*)}{n}.$$

These numbers, which correspond to the entries in the final table above, are called the expected values of B and W. $\mathsf{E}W$, as a long-run proportion, is a special case; below, we will call it the probability of winning the bet.

This is the long-run thinking that underlies probability theory. If we replicate an experiment many many times, we use what *would* happen "in the long run" as a good guess for what will happen when we run the experiment now. This guess need not, and will not, be exact, but it is representative of the possible outcomes and their relative likelihood.

In general, the *expected value* of a random variable[10] is the long-run average value of the measured observable over many, many replications of the experiment. Specifically, generating a sequence of replications for the experiment means drawing a sequence of elementary outcomes $\omega_1^*, \omega_2^*, \omega_3^*, \ldots \in \Omega$ by the same mechanism. For any random variable X defined in that experiment, the expected value is given by

$$\mathsf{E}X = \lim_{n\to\infty} \frac{X(\omega_1^*) + \cdots + X(\omega_n^*)}{n}. \tag{2.1}$$

This number has many different interpretations. We can think of $\mathsf{E}X$ as a "good guess" for the value the random variable X will take. We can also think of $\mathsf{E}X$ as the fair price (in \$) we should pay to enter a game whose payoff is \$$X$. It is fair because if we pay that price, then on average, we will break even.

Interpretation. The expected value $\mathsf{E}X$ of a random variable X can be interpreted as the fair price to pay for entering a game that gives payoff X.

We will see other interpretations as we proceed.

If ω^* is the elementary outcome chosen by nature in our run of the experiment, then $X(\omega^*)$ is the value of the observable X that we *will*

10 Remember each random variable corresponds to a particular observable quantity (or quantities in the vector case) that we can measure in a random experiment.

observe. A critical implication of equation (2.1) is that $\mathsf{E}X$ *does not depend on the specific value of* $X(\omega_*)$ *for our particular experiment*; it only depends on the possible values of X and their relative likelihood. It follows that we can (in principle) compute $\mathsf{E}X$ *before* the experiment, which is exactly our goal.

Finally, the idea of replicating an experiment in identical circumstances borders on (if not crosses into) philosophical terrain. There is in fact no realistic situation that we could replicate to this standard; no matter how hard we try, some aspect of the situation will be different when we re-run the experiment. However, this is a model, so we are content to allow some variation as long as the important features of the experiment are the same. Over different roulette spins at the casino, many factors will change, but the basic process by which a ball ends up in a slot remains the same. There is actually no need to consider the replications to be sequential. For example, physicists use the idea of an *ensemble*, a large (or infinite) number of identical (and hypothetical) experiments proceeding in parallel.[11] This works just as well for our purposes. Fortunately, the abstractness of these replications and the impracticality of exactly replicating an experiment do not crack the foundations upon which probability theory rests. The reason is that we need never actually perform these experiments to do our calculations.

11 In the statistical theory of gases, we can take many identical volumes of a gas at the same temperature and pressure as an ensemble.

Indeed, a reasonable question to ask is how we can compute anything at all using long-run averages. If we needed to replicate an experiment, even several times let along an infinite number, just to make predictions about the experiment, then probability theory would be useless. It turns out, however, that the mere possibility of replicating the experiment is enough. We have methods for computing expected values before the experiment is ever run.[12] In general, our calculation of the expected value of a random will depend on the details of the random experiment and of our probabilistic assumptions. Section 8 in Chapter 3 shows how to do this using what we will call the distribution of a random variable. This can often be very difficult, especially when working with complicated systems. Fortunately, as we will see in the next section, expected values must satisfy certain rules that often enable us to simplify our calculations immensely.

12 In fact, probability theory is always used in advance. Most of our calculations will be complete before the experiment begins and all of them before the experiment ends. There is no point in making predictions when the results are known.

3. The Expected Value Operator E

If X is a random variable, we denote the expected value of X by $\mathsf{E}X$. This E is more than an abbreviation, it is an *operator*[13] that acts on a single random variable to produce[14] a number or a vector.

The operator E has one input slot to be filled by the random variable that follows it.[15] If that random variable is real-valued, the operator E returns a number; if that random variable is vector-valued, the operator E returns a vector with the same number of components. Specifically, if X is a real-valued random variable, then $\mathsf{E}X$ is a real-number. If $X = (X_1, \ldots, X_n)$ is a vector-valued random variable with real-valued components $X_1, \ldots, X_n$, then

$$\mathsf{E}X = (\mathsf{E}X_1, \ldots, \mathsf{E}X_n), \tag{2.2}$$

the vector of expected values of the components.

The operator E has higher precedence in the order of operations than addition and subtraction but lower precedence than multiplication, division, or exponentiation. Parentheses are only necessary to clarify the order in which operators are to be applied. Thus, we have the following notational equivalences for real-valued random variables X and Y:

$$\begin{aligned} \mathsf{E}X &\text{ is equivalent to } \mathsf{E}(X) \\ \mathsf{E}XY &\text{ is equivalent to } \mathsf{E}(XY) \\ \mathsf{E}X^2 &\text{ is equivalent to } \mathsf{E}(X^2) \\ \mathsf{E}\frac{X}{Y} &\text{ is equivalent to } \mathsf{E}\left(\frac{X}{Y}\right). \end{aligned}$$

On the other hand, we write $\mathsf{E}(X+Y)$ for the expected value of the sum; $\mathsf{E}X + Y = Y + \mathsf{E}X$ is the random variable Y plus the constant $\mathsf{E}X$. We write $(\mathsf{E}X)^2$ for the square of the expected value of X, not $\mathsf{E}X^2$, which is the expected value of the random variable X^2.

The advantage of thinking of the expected value as an operator that acts on random variables is that we can use the properties of averages to compute expected values that we might not be able to find from first principles. This is a powerful problem-solving tool that we will use often, and the rules described in the next subsection are worth memorizing.

13 An operator is a mathematical or algorithmic processor that acts on objects of a certain class to yield an object of another (not necessarily distinct) class. Two operators that are already familiar are the derivative d/dx and integral (anti-derivative) operators $\int_{-\infty}^{x} du$; both operate on functions to produce functions. Operators in programming languages also qualify. For example, dereferencing in C++ is an operator that takes an object of type `T*` and returns an associated object of type `T`. Operators can have one or more input slots for the objects they act upon. Those with one input slot, such as dereferencing and E, are called *unary* operators.

14 You might ask: Are operators functions? Technically, they are, because they return one and only one value for each possible input. For example, E is a function that takes another function (a random variable) as its argument and returns a number. However, this can be somewhat abstract because operators such as E are defined on a more complicated domain than the typical functions we deal with. So, we treat operators – whether E, addition, or multiplication – as special constructs. For example, we usually do not require explicit parentheses around the argument to E as described in the text.

15 Remember E acts only on random variables – including constant random variables – so it makes no sense to apply E to an event or a function that is not a random variable.

The Basic Expected Value Rules

Constancy Rule

Given: The constant random variable 1.

Yields: $\mathsf{E}1 = 1$.

In words: A constant is its own expected value.

Analogy: The average is 1 if all numbers in the list equal 1.

Non-negativity Rule

Given: A random variable X that is always non-negative, i.e., $X \geq 0$.

Yields: $\mathsf{E}X \geq 0$.

In words: The expected value of a non-negative random variable is non-negative.

Analogy: The average is non-negative if all numbers in the list are non-negative.

Scaling Rule

Given: Random variable X and a constant $c \in \mathbb{R}$.

Yields: $\mathsf{E}(cX) = c\mathsf{E}X$.

In words: Constants can be taken out of the expected value.

Analogy: Scaling all numbers in a list by the same constant scales the average of the list by that constant as well.

Additivity Rule

Given: Random variables $X_1, \ldots, X_n$ for positive integer n.

Yields: $\mathsf{E}(X_1 + \cdots + X_n) = \mathsf{E}X_1 + \cdots + \mathsf{E}X_n$.

In words: The expected value of a sum is the sum of the expected values.

Analogy: The average of an (elementwise) sum of lists is the sum of the averages.

Monotone Limits Rule

Given: A random variable X and random variables $X_1 \leq X_2 \leq \cdots$ such that $\lim_{i\to\infty} X_i = X$.

Yields: $\lim_{i\to\infty} \mathsf{E}X_i = \mathsf{E}(\lim_{i\to\infty} X_i) = \mathsf{E}X$.

In words: The order of limits and E can be exchanged if the sequence is increasing.

Analogy: There is one, but it is too technical to be worth stating.

3.1. The Basic Rules of Expected Values

We saw earlier that $\mathsf{E}X$, the expected value of X, is the long-run average of the values that X takes over many replications of the experiment. Being an average, the expected value necessarily has all of the familiar properties of averages that we discussed at the beginning of the chapter. This is reflected in the properties of the operator E, which must obey certain rules. By applying these rules, we can often compute expected values more quickly and without the need to return to first principles every time.

The basic rules[16] E must obey are listed on the facing page. We will use them all repeatedly throughout this course. Each of the rules apply in a certain situation (labeled "Given") and gives a particular result (labeled "Yields"). We also describe this result in words and make an explicit analogy to the properties of averages of lists of numbers.

16 In fact, these five properties can be taken as axioms that define E. If we started with these five rules, the rest of the theory would follow.

The *Constancy Rule* tells us the expected value of the random variable that always takes the value 1. Notice that 1 is used in two ways in the equation, as a random variable in the input slot of E and as the number returned by the operator. This is a bit loose; we should write the rule as $\mathsf{E}1_\Omega = 1$ to make the distinction clearer. However, the meaning of a constant is clear from context, and constants are used so frequently that it becomes tedious to write the random variable that is always c as $c1_\Omega$ every time.

The *Non-negativity Rule* has two useful consequences that are applied often enough to become named rules themselves. First, a non-negative random variable with expected value zero must be the constant random variable 0.

$$X \geq 0 \text{ and } \mathsf{E}X = 0 \implies \mathsf{P}\{\, X = 0 \,\} = 1. \qquad \text{(Nullity Rule)}$$

Second, if one random variable Y is always at least as big as another random variable X, then $\mathsf{E}Y$ is at least as big as $\mathsf{E}X$.

$$X \leq Y \implies \mathsf{E}X \leq \mathsf{E}Y. \qquad \text{(Monotonicity Rule)}$$

Recall that $X \leq Y$ means $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. Hence, $X \leq Y$ implies that the random variable $Y - X \geq 0$, and the Monotonicity Rule follows from the Non-negativity, Additivity, and Scaling rules combined. In applications, the condition that $X \leq Y$ is derived either from a specific mathematical relationship between X and Y or from a specific physical relationship between the associated observable quantities being measured in the experiment. Both of these derived rules correspond to properties of averages: (i) the only non-negative list of numbers with average zero is the list of all zeroes, and (ii) the average always lies between the smallest and biggest numbers in the list.

The *Scaling Rule* tells us simply that we can move any multiplicative constants outside the E. It follows as well that we can move any constant inside the E as well. A useful trick is to multiply an expected value by 1 in the form c/c for some constant, moving one constant inside the E (to simplify something) and keeping the other outside.

The *Additivity Rule* is the rule that we will use most often. It's primary value comes from decomposing a problem into simpler subproblems. It is usually easier to compute $\mathsf{E}X_i$ from first principles than $\mathsf{E}(X_1+\cdots+X_n)$. Remember this, we will use it many times. Incidentally, another useful trick in the same spirit as multiplying by 1 is to add 0 to an expected value, using the additivity rule to move $-c$ inside the E and c outside. That is, $\mathsf{E}X = \mathsf{E}(X - c) + c$ for any constant c.

The *Monotone Limits* rule is rather technical, but it is necessary for some of the finer points of the theory. For example, notice that the Additivity Rule applies only to *finite* sums of random variables. To deal with an infinite sum, we need to take limits, and the Monotone Limits Rule tells us about the interaction between the two operations of taking limits and taking expected values. As an example of how it is used, suppose that $\mathcal{A}_1, \mathcal{A}_2, \ldots \subset \Omega$ are disjoint events and let $\mathcal{A} = \bigcup_{i=1}^{\infty} \mathcal{A}_i$. Then, no two indicators $1_{\mathcal{A}_i}$ are non-zero at the same ω, and the random variables $X_n = \sum_{i=1}^{n} 1_{\mathcal{A}_i}$ are increasing in n while converging to $X = 1_{\mathcal{A}}$. Hence,

$$\sum_{i=1}^{\infty} \mathsf{E}1_{\mathcal{A}_i} = \lim_{n\to\infty} \mathsf{E}X_n = \mathsf{E}X = \mathsf{E}1_{\mathcal{A}}.$$

This result gives an important property of probabilities, called countable additivity, as described in the next section. Another implication (though somewhat indirect) of the Monotone Limits Rule is called the Dominated Convergence Rule. Let X_i be a sequence of random variables with $\lim_{i\to\infty} X_i = X$ and where $|X_i| \leq Y$ for all $i \in \mathbb{Z}_+$. Then, $\lim_{i\to\infty} \mathsf{E}X_i = \mathsf{E}X$. This gives somewhat different conditions under which limits and expected values can be interchanged.

When one averages a list of numbers, the average is a summary that is representative of the list, but it need not be in the list. Similarly, if X is a real-valued random variable, then the number $\mathsf{E}X$ need not be a possible value of X. We will see in Chapters 3 and 4 why $\mathsf{E}X$ can be considered a "typical" value for X.

It is also worth commenting on a property that the E operator does *not* possess. Let's start again with averages of lists of numbers. If our list is $-1, 1, 1, -1, -1, -1, 1, 1$, then its average is zero. If we first square every number in the list and then average, we get one, and yet $0^2 \neq 1$. Therefore, *the average of a list obtained by taking a function of each number in another list is not equal to the result of that function on the average of the other list.* More succinctly, we cannot interchange the operations of taking a function and taking an average. The same proviso works for expected values. If X is a random variable and g is a function defined on Range(X), then

$$\mathsf{E}g(X) \quad \text{is usually} \neq \quad g(\mathsf{E}X). \tag{2.3}$$

As an example, if X is a random variable that takes values 1 and -1 then $X^2 \equiv 1$, and $\mathsf{E}X^2 \neq (\mathsf{E}X)^2$.

4. Probabilities are Expected Values of Indicator Random Variables

In baseball, a player's batting average is the proportion of hits, the ratio of the number of hits to the total number of attempts ("at bats"). It is always a number between 0 and 1, where a value closer to one indicates a higher proportion of hits, and is computed by averaging a list of 0s and 1s where 1 represents a hit. We saw likewise that if W is the indicator that a particular bet wins in a single spin of roulette then $\mathsf{E}W$ is (the limit of) an average of a list of 0s and 1s where 1 represents a win. Thus, $\mathsf{E}W$ is a proportion and always between 0 and 1. A value close to 1 indicates that a win is very likely as the experiment is replicated; a value close to 0 indicates that a win is unlikely. The proportion $\mathsf{E}W$ measures how likely that bet is to win on a single spin. We call this the *probability* that the bet will win. The same logic works for any indicator random variable, and in general, we call the expected value of an indicator a probability.

The probability of an event $\mathcal{A}$ (in all the variants we will encounter) is defined to be the expected value of the indicator $1_{\mathcal{A}}$ of that event. This is a measure of how likely an event is to occur, with 1 representing certainty that the event will occur and 0 representing certainty that the event will not occur.

Thus, probabilities are nothing more than expected values of 0-1 random variables. There is consequently nothing new to the mathematics of probability that is not subsumed by the rules of expected values. Nonetheless, we use and compute probabilities so frequently that it is useful to have a special notation for the probability of an event. If $\mathcal{A}$ is an event, then

$$\mathsf{P}(\mathcal{A}) \equiv \mathsf{E}1_{\mathcal{A}} \tag{2.4}$$

is the probability of that event. This is a real-valued function on subsets of Ω that always gives a result between 0 and 1. By the rules of expected values, this function P must itself satisfy certain properties; P is what we call a *probability measure* in Appendix M.

For example, because $1_\Omega \equiv 1$ and $1_\emptyset \equiv 0$, $\mathsf{E}1_\Omega = 1$ and $\mathsf{E}1_\emptyset = 0$. In other words,

$$\mathsf{P}(\Omega) = 1 \quad \text{and} \quad \mathsf{P}(\emptyset) = 0. \tag{2.5}$$

This is often called the Rule of Total Probability. It tells us that *something* must happen in an experiment; some ticket is selected from the box. If $\mathcal{A}$ and $\mathcal{B}$ are two events

$$1_{\mathcal{A}\cup\mathcal{B}} = 1_{\mathcal{A}} + 1_{\mathcal{B}} - 1_{\mathcal{A}\cap\mathcal{B}}.$$

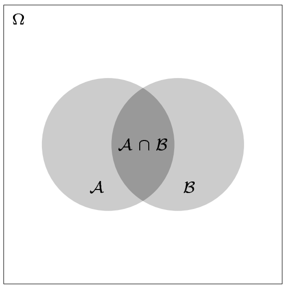
Hence, by the Additivity Rule, we have

$$\mathsf{P}(\mathcal{A}\cup\mathcal{B}) = \mathsf{P}(\mathcal{A}) + \mathsf{P}(\mathcal{B}) - \mathsf{P}(\mathcal{A}\cap\mathcal{B}). \tag{2.6}$$

We get the probability of a union by adding the probability of the sets but subtracting the probability for the points we have added twice – those in both sets. This is easiest to see in a picture.[17] Since $\mathcal{A}$ and $\operatorname{compl}(\mathcal{A})$ are disjoint and have union equal to Ω, we can take $\mathcal{B} = \operatorname{compl}(\mathcal{A})$ in equation 2.5 to get,

$$\mathsf{P}(\operatorname{compl}(\mathcal{A})) = 1 - \mathsf{P}(\mathcal{A}). \tag{2.7}$$

17 A useful conceptual tool is an adapted version of Venn's diagrams. We take Ω to be a box of area 1 within which we measure the probability of an event by its area. We can then represent events as subset of the box. This can reveal useful properties quite intuitively. For example, the following shows how $\mathcal{A}\cap\mathcal{B}$ contributes twice the expression $\mathsf{P}(\mathcal{A}) + \mathsf{P}(\mathcal{B})$:



If $\mathcal{A}$ and $\mathcal{B}$ are events that satisfy $\mathcal{A} \subset \mathcal{B}$, then $1_{\mathcal{B}} - 1_{\mathcal{A}} \geq 0$. The Non-negativity Rule then implies that

$$\mathcal{A} \subset \mathcal{B} \implies \mathsf{P}(\mathcal{A}) \leq \mathsf{P}(\mathcal{B}). \tag{2.8}$$

The condition that $\mathcal{A} \subset \mathcal{B}$ tells us that if the selected elementary outcome $\omega^* \in \mathcal{A}$, then we necessarily have $\omega^* \in \mathcal{B}$ as well. Hence, if $\mathcal{A}$ occurs, so must $\mathcal{B}$; the event $\mathcal{A}$ implies the event $\mathcal{B}$. The discussion of the Monotone Limit Rule in the last section showed that if $\mathcal{A} = \bigcup_i \mathcal{A}_i$ (the union can be finite or countably infinite) where the $\mathcal{A}_i$ are *mutually exclusive* sets[18]

$$\mathsf{P}\left(\bigcup_i \mathcal{A}_i\right) = \sum_i \mathsf{P}(\mathcal{A}_i). \tag{2.9}$$

This can be interpreted as follows: if you divide an event into a (countable) number of disjoint pieces, then no probability is lost; the probability of the pieces add up to the probability of the whole. All five of the preceding properties of probabilities will be useful throughout this course. But the key point is that there is nothing new here; if you know the expected value rules, you can find these easily.

18 This means that the $\mathcal{A}_i$ are *pairwise disjoint* sets, $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ if $i \neq j$.

> If $\mathcal{A}$ is an event, then the probability of that event is defined by
>
> $$\mathsf{P}(\mathcal{A}) \equiv \mathsf{E}1_{\mathcal{A}}.$$
>
> This is a real-valued function on subsets of Ω that, by the Expected Value Rules, satisfies the following:
>
> 1. $\mathsf{P}(\emptyset) = 0 \leq \mathsf{P}(\mathcal{A}) \leq 1 = \mathsf{P}(\Omega)$
> 2. For any countable[19] collection of pairwise disjoint events $\mathcal{A}_1, \mathcal{A}_2, \ldots \subset \Omega$, such that $\mathcal{A} = \bigcup_i \mathcal{A}_i$, then,
>
> $$\mathsf{P}(\mathcal{A}) = \sum_i \mathsf{P}(\mathcal{A}_i).$$
>
> 3. $\mathsf{P}(\mathcal{A} \cup \mathcal{B}) = \mathsf{P}(\mathcal{A}) + \mathsf{P}(\mathcal{B}) - \mathsf{P}(\mathcal{A} \cap \mathcal{B})$.
> 4. $\mathsf{P}(\mathrm{compl}\,(\mathcal{A})) = 1 - \mathsf{P}(\mathcal{A})$.
> 5. $\mathcal{A} \subset \mathcal{B} \implies \mathsf{P}(\mathcal{A}) \leq \mathsf{P}(\mathcal{B})$.

19 From Appendix S, countable means finite or countably infinite.

5. Illustration: Computing Expected Values

In this section, we show how the properties of the expected value operator E make it possible to compute expected values without replicating the experiment, and indeed before the experiment is even run. We focus here on the important class of random variables that can take at most a countable[19] number of possible values. Such random variables are called *discrete*.

The simplest discrete random variables are constants, and the properties of E tell us the expected values of constants directly via the Constancy and Scaling Rules.

Next simplest are are indicators. If $\mathcal{A} \subset \Omega$ is an event, then $\mathsf{E}1_{\mathcal{A}} \equiv \mathsf{P}(\mathcal{A})$ is just the long-run proportion of repliations in which $\mathcal{A}$ occurs. We can compute this probability in one of two ways: (i) by direct recourse to the assumptions of our probability model or (ii) by decomposing $\mathcal{A}$ into smaller sets whose probability we can compute by direct recourse to the assumptions of our model. The following examples illustrate both approaches, though this approach is fundamental to the theory.

► EXAMPLE 2.1. Flips of a Fair Coin

As in the Coin Flipping example in Section 4, let $H_1, H_2, H_3, \ldots$ be the indicators of heads on successive flips of a fair coin. Then, by the assumption that the coin is fair, we have

$$\mathsf{E}H_i = \frac{1}{2} = \mathsf{P}\{\, H_i = 1 \,\}.$$

This is a direct recourse to our model – if the coin is fair, it is equally likely to come up heads and tails.

Consider next the event that the first heads comes up after the third flip but before the 10th. This is the event $\mathcal{A}$ given by

$$\mathcal{A} = \bigcup_{i=4}^{9} \{\, H_1 = 0 \text{ and } \cdots \text{ and } H_{i-1} = 0 \text{ and } H_i = 1 \,\}.$$

Notice that the events in the union are mutually exclusive – at most one can occur. We can write this in terms of indicator random variables as the equality

$$1_{\mathcal{A}} = \sum_{i=4}^{9} 1_{\{\, H_1=0 \text{ and } \cdots \text{ and } H_{i-1}=0 \text{ and } H_i=1 \,\}}.$$

We know the expected value of the indicators within the sum by our assumptions. Using the Additivity Rule we have that

$$\begin{aligned}
\mathsf{P}(\mathcal{A}) &= \mathsf{E}\sum_{i=4}^{9} 1_{\{\, H_1=0 \text{ and } \cdots \text{ and } H_{i-1}=0 \text{ and } H_i=1 \,\}} \\
&= \sum_{i=4}^{9} \mathsf{E}1_{\{\, H_1=0 \text{ and } \cdots \text{ and } H_{i-1}=0 \text{ and } H_i=1 \,\}} \\
&= \sum_{i=4}^{9} \mathsf{P}\{\, H_1 = 0 \text{ and } \cdots \text{ and } H_{i-1} = 0 \text{ and } H_i = 1 \,\} \\
&= \sum_{i=4}^{9} 2^{-i}.
\end{aligned}$$

▶ EXAMPLE 2.2. Flush

Assume that you are dealt five cards from the top of a well-shuffled, standard deck. Our probability model is based on one assumption of symmetry: that every one of the 52! orderings of the cards is equally likely to occur.

Suppose we list all 52! permutations of the cards in some order and define $S_1, \ldots, S_{52!}$ to be the indicators that the corresponding permutation has occurred. Our model tells us directly that $\mathsf{E}S_i = 1/52!$ for $i = 1, \ldots, 52!$. These indicators all mutually exclusive and collectively exhaustive: for any elementary outcome, exactly one of the S_is will equal 1.

Let A be the indicator that the first card dealt is an ace. A is the sum of those S_is corresponding to a permutation with an ace in the first card. There are exactly $4 \cdot 51!$ such permutations because with each of the four aces as the first card the other 51 cards can be completely permuted. Thus,

$$\mathsf{E}A = \frac{1}{52!}\, 4 \cdot 51! = \frac{4}{52},$$

as we would have derived intuitively. While this seems like a lot of work for such a simple problem with an intuitive answer, it actually shows us a plan of attack for much more complicated problems for which intuition is a less reliable guide. The key strategy here was to decompose A as a sum of indicators about which we know something.

Now, let F be the indicator that five cards are all of the same suit (a flush). Note that $F = F_{\clubsuit} + F_{\heartsuit} + F_{\spadesuit} + F_{\diamondsuit}$, where each of the random variables on the right is the indicator that we get a flush of a particular suit. Since there is nothing special about one suit over another symmetry and the additivity rule tell us that $\mathsf{E}F = 4\mathsf{E}F_\spadesuit$. Moreover, $F_\spadesuit$ is the sum of the S_is for which the first five cards in the deck are spades. For each of the $5!\binom{13}{5}$ arrangements of 5 spades in the first 5 cards there are 47! permutations of the other cards. Hence,

$$\mathsf{E}F_\spadesuit = \frac{1}{52!}\,5!\binom{13}{5}\,47! = \frac{13\cdot 12\cdot 11\cdot 10\cdot 9}{52\cdot 51\cdot 50\cdot 49\cdot 48}.$$

Next, suppose that S is a random variable that can take on the values $v_1, v_2, \ldots, v_m$ for some positive integer m. We can re-express S in terms of indicators as follows:

$$S = v_1\cdot 1_{\{\,S=v_1\,\}} + \cdots + v_m\cdot 1_{\{\,S=v_m\,\}} = \sum_{i=1}^{m} v_i\cdot 1_{\{\,S=v_i\,\}}. \tag{2.10}$$

For example, if S takes only the values -1, 1, and 2, we have $S = 2\cdot 1_{\{\,S=2\,\}} + 1_{\{\,S=1\,\}} - 1_{\{\,S=-1\,\}}$. To see that the equation makes sense, notice that the events $\{\,S=v_i\,\}$ for $i=1,\ldots,m$ are both mutually exclusive and collectively exhaustive. For any $\omega\in\Omega$, $S(\omega)$ equals exactly one of the v_is. Hence, no two events $\{\,S=v_i\,\}$ and $\{\,S=v_j\,\}$ for $i\neq j$ can contain the same ω, and every $\omega\in\Omega$ must be contained in one such event. For any ω, exactly one of the indicators in equation (2.10) equals 1 and the rest equal 0, yielding $S(\omega)$ on both sides of the equation.

Applying the Additivity and Scaling Rules to equation (2.10) gives

$$\begin{aligned}
\mathsf{E}S &= \mathsf{E}\sum_{i=1}^{m} v_i\cdot 1_{\{\,S=v_i\,\}}\\
&= \sum_{i=1}^{m}\mathsf{E}\left(v_i\cdot 1_{\{\,S=v_i\,\}}\right)\\
&= \sum_{i=1}^{m} v_i\,\mathsf{E}1_{\{\,S=v_i\,\}}\\
&= \sum_{i=1}^{m} v_i\,\mathsf{P}\{\,S=v_i\,\}.
\end{aligned}$$

This is a weighted average of the values that S can take where the weights are the probabilities that S takes each value.

▶ EXAMPLE 2.3.

Two balanced, standard, six-sided dice – one blue and one red – are rolled once. Let D_r and D_b be, respectively, the face that comes up on the red and blue die. Let $D = D_r + D_b$ be the sum of the two dice.

By the assumptions of the model, each of the thirty-six possible values for the pair (D_r, D_b) has probability $1/36$. Exactly 1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1 of them give the respective value 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 of D. Hence,

$$\begin{aligned} \mathsf{E}D &= \sum_{k=2}^{6} k\,\frac{k-1}{36} + 7\cdot\frac{1}{6} + \sum_{k=8}^{12} k\,\frac{13-k}{36} \\ &= 7. \end{aligned}$$

Notice also that since both dice are balanced

$$\begin{aligned} \mathsf{E}D_r &= \mathsf{E}D_b \\ &= \sum_{k=1}^{6} k\,\frac{1}{6} \\ &= 3.5. \end{aligned}$$

Therefore, by additivity, we get the same result $\mathsf{E}D = 3.5 + 3.5 = 7$.

The latter point reflects the fact that the properties of the E operator tend to be easier to use than this direct weighted average. If necessary, we can compute the weighted average, but we try to do everything we can to avoid it.

We can generalize this result to arbitrary discrete random variables.[20] Suppose that the real-valued random variable C takes values in a *countable* set which is enumerated $u_1, u_2, \ldots$. Then, $C = C_+ - C_-$ where $C_+ = \max(C, 0)$ and $C_- = \max(-C, 0)$. Each of the random variables C_+ and C_- can be written as a sum of indicators: $C_+ = \sum_{i=1}^{\infty} u_i^+ \cdot 1_{\{\,C_+ = u_i^+\,\}}$ and $C_- = \sum_{i=1}^{\infty} u_i^- \cdot 1_{\{\,C_- = u_i^-\,\}}$, where the u_i^+s and u_i^-s are just the positive and negative u's listed in some order. By the Monotone Limits rule, $\mathsf{E}C_+ = \sum_{i=1}^{\infty} u_i^+ \mathsf{P}\{\,C_+ = u_i^+\,\}$ and $\mathsf{E}C_- = \sum_{i=1}^{\infty} u_i^- \mathsf{P}\{\,C_- = u_i^-\,\}$. Hence, putting the pieces back together

$$\mathsf{E}C = \sum_{i=1}^{\infty} u_i\,\mathsf{P}\{\,C = u_i\,\}.$$

20 You can skip this paragraph on first reading and go right to the boxed result below.

The general result is worth highlighting.

> Suppose X is a *discrete* random variable, meaning that it takes values in some countable set $\mathcal{C}$, then
>
> $$\mathsf{E}X = \sum_{u \in \mathcal{C}} u\, \mathsf{P}\{\, X = u \,\}\,.$$
>
> $\mathsf{E}X$ can be computed as a weighted average of the values X can take with the probabilities of these values as weights.

That we can compute the expected value of a random variable as a weighted average of the variable's values with the probabilities at (or near) each value as weights will turn out to be quite general, as we will see in the next two chapters.

6. Aside: A Provisional Definition of Independence

Independence is a critically important idea in probability theory. Two random variables are *independent* if knowing the value that one takes in the experiment (i.e., the actual observed measurement) gives no useful information about the value that the other will take. Similarly, we say that a collection $X_1, X_2, \ldots, X_n$ of random variables are independent if knowing the values that any subset of them take in the experiment (i.e., the actual observed measurements) gives no useful information about the value the remaining ones will take.

Independence is tied up with how well one random variable can be used to predict another. It is thus commonly used as a simplying assumption in probability models, to allow different quantities to be analyzed separately. We will explore this idea – both its intuition and mathematical manifestation – more fully when we think about prediction. But for now, it is so useful an idea that we will state a provisional definition to hold us over.

Provisional Definition. Real-valued random variables $X_1, \ldots, X_n$ are independent if for any functions $g_1, \ldots, g_n$ defined on the respective ranges of the X_is, we have

$$\mathsf{E}\left(\prod_{i=1}^{n} g_i(X_i)\right) = \prod_{i=1}^{n} \mathsf{E} g_i(X_i).$$

Vector-valued random variables are independent if the corresponding components are independent.

This definition is a bit abstract, but at heart, it gives circumstances in which independent random variables can be analyzed separately. The most common form in which this is used is when $g_i = 1_{\mathcal{A}_i}$ for some sets of values $\mathcal{A}_1, \ldots, \mathcal{A}_n$ in the respective ranges of $X_1, \ldots, X_n$. Then, $\mathsf{E} g_i(X_i) = \mathsf{P}\{\, X_i \in \mathcal{A}_i \,\}$, the probability that X_i takes a value in the set $\mathcal{A}_i$, and the definition reduces to the following: $X_1, \ldots, X_n$ are independent if and only if

$$\mathsf{P}\{\, X_1 \in \mathcal{A}_1 \text{ and } \cdots \text{ and } X_n \in \mathcal{A}_n \,\} = \prod_{i=1}^{n} \mathsf{P}\{\, X_i \in \mathcal{A}_i \,\}$$

for any sets $\mathcal{A}_i$ in the ranges of the X_is.

Note that it is not sufficient for independence that $\mathsf{E}XY = \mathsf{E}X\mathsf{E}Y$.

▶ EXAMPLE 2.4. Coin Flips (again)

We typically model a sequence of coin flips as independent, meaning for instance that the value of one flip tells us nothing useful about the values of other flips.

This shows why in the example above

$$\begin{aligned}
&\mathsf{P}\{\, H_1 = 0 \text{ and } \cdots \text{ and } H_{i-1} = 0 \text{ and } H_i = 1 \,\} \\
&\qquad\qquad = \mathsf{P}\{\, H_1 = 0 \,\} \cdot \mathsf{P}\{\, H_{i-1} = 0 \,\} \cdot \mathsf{P}\{\, H_i = 1 \,\}.
\end{aligned}$$

▶ EXAMPLE 2.5. Constants are independent of any others

If X is a constant random variable and Y is any other random variable, then the Scaling Rule gives $\mathsf{E}g(X)g(Y) = g(X)\mathsf{E}g(Y)$. Thus, X and Y are independent. This makes sense in terms of prediction: our prediction for a constant random variable does not change – it is always perfect – no matter what information we have about Y.

DESCRIBING DISTRIBUTIONS

3

The first approached the Elephant
And, happening to fall
Against his broad and sturdy side
At once began to bawl
'God bless me, but the Elephant
is very like a wall.'

The second, feeling the tusk,
Cried 'Ho! What have we here,
So very round and smooth and sharp
To me 'tis very clear
This wonder of an Elephant
Is very like a spear'
— from The Parable of the Blind Men
and the Elephant by J. G. Saxe

---

1. Where and How Much

Suppose that you had access to the exact number of lightning strikes and the exact amount of rainfall at every point in the continental United States during 1999.[1]

If you were asked to describe this information, to communicate it to someone else, how would you do it?

One way to start would be to obtain two blank maps of the continental US. You could then mark a point on one map at the location of each lightning strike and shade the other map with greyscale tones that

1 But we would never have all of that data in practice, you scoff. True. But because we are trying to understand mathematical ideas, we can afford to think like mathematicians and ignore such practical realities.

correspond to different amounts of rainfall. Both maps produced this way impart two features of the information about rainfall and lightning strikes: where and how much. We say that the maps describe the *distributions* of rainfall and lightning in the continental US in 1999.

These two maps differ in one important respect: they display different quantities. The lightning strike map shows a discrete set of points where the strikes occurred,[2] and the map enables us to count the strikes in any region. We say that the lightning strike distribution is *discrete* because the "where" information is concentrated on the discrete set where the strikes occurred. The rainfall map, on the other hand, shows the *density* of rain at each point. (When meteorologists speak of 2.4″ of rain, for example, they mean 2.4″ of rain per square inch – a density.) Since rain diffuses along the ground hitting every point in the area, at least a little, we can only understand the amount of rainfall locally, as through a density. We cannot measure the rain that hits a single point in isolation. We say that the rainfall distribution is *continuous* because the "where" information is spread over the entire continuum of points represented by the maps.

Because the two maps display different quantities, they provide different representations of a distribution.[3] Both distributions answer the same questions about where and how much, but to answer those questions, we use the maps in a slightly different – though abstractly equivalent – ways. For example, suppose we wanted to know how much rain had fallen in Pennsylvania in 1999. We can do this by integrating the densities on the rainfall map over the area of Pennsylvania.[4]

$$\text{Total Rain(Pennsylvania)} = \int_{\text{Pennsylvania}} \text{Rainfall Density}(x, y)\, dx\, dy.$$

We could answer the analogous question for lightning strikes as well: we simply add up the number of strikes within the state of Pennsylvania. Mathematically, if the lightning strikes occurred at points $(x_1, y_1), \ldots, (x_n, y_n)$, then we have

$$\text{Total Strikes(Pennsylvania)} = \sum_{i s.t. (x_i, y_i) \in \text{Pennsylvania}} \#\ \text{Strikes at}(x_i, y_i).$$

It turns out that this sum can also be viewed as an integral.

We can answer other questions with these distributions as well. We can find the distribution of rainfall in Pennsylvania. (Think about how

2 We're being mathematical again, here, assuming that lightning hits at a single point among a continuum of points. Still, it is an intuitive abstraction.

3 We will call these the mass function and the density function below in a probability context.

4 If the density were flat, then it would be the value of the density times the area of the state. An integral just does this locally around each point and adds up the results.

you would do this.) We can determine the regions where more than 20″ of rain fell and the regions where less than 1″ fell. We can decide if there was more rain to the west of the Mississippi river than east of it. We can find the city with the most lightning strikes and all the places where lightning strikes occurred within an inch of each other. Indeed, we can answer any question about where and how much because the distribution embodies *all* the information about that.

The word "distribution" arises in several practical senses in daily life. Corporations carefully track the distribution of their products around the world. Presidential candidates and economists talk about the distribution of income. Professors describe the distribution of grades after an exam. Astronomers study the distribution of mass within a galaxy while evolutionary biologists study the distribution of mass extinctions across history. The key point is that all of these notions of distribution specify a "where" and a "how much" and thus exemplify the same class of object.[5] The particulars of the different examples are certainly important, but they can distract one from seeing the common structure. So, let's take a closer look.

First, the "where"s in these examples have different meanings. Corporations distribute their products to locations on the Earth, or to be more specific, to commercial or residential addresses. This is a where that feels like a where. In contrast, the distribution of income describes how many people earn each amount of money and the distribution of grades describes how many students earn each score. The where's in these examples are income levels and grade scores respectively, with incomes in pennies and grades in points. The distribution of mass in a galaxy describes how much matter is present at each location in a region of space; this is a three-dimensional where as opposed to the two-dimensional where of the products example. Mass extinctions are distributed over time rather than space – a where in time.[6] Notice that products and grades and extinctions have more discrete where's, like lightning, while income and mass are more easily considered as spread out continuously, like rainfall.

Second, the units of "how much" differ as well. The amount of a product from a corporation can be intrinsically integer-valued (ball bearings) or can be more generally a real number (jet fuel), though

5 What are methods does such an object support? Consider the following:

```
template< Set Domain,
          Subset<R> Range >
class Distribution
{
  public:

  ...

  // Describe how much
  // for each where

  Range
  howMuch( Subset<Domain> where );

  // Other representations
  // (see below)

  ...
}
```

6 We tend to think of mass extinctions as occurring at discrete points in time even though they can take millions of years to occur. This is a bit loose, but there is always some idealization in thinking of time or space as a continuum of points. Besides, geological time is vast. If you prefer, think of the times of earthquakes or neutrino's interacting with the earth, etc.

7 With this in mind and referring to the code template above, which of the following declarations corresponds to which example?

```
Distribution<EarthsSurface, Z_+>
Distribution<IncomeLevels, Z_+>
Distribution<0..100, Z_+>
Distribution<R^3, R_+>
Distribution<R_+, Z_+>
```

8 Computer scientists love code reuse; mathematicians love idea reuse.

they are always positive. The income and grade distributions measure the number of people at each level of income or grade. Mass can be considered discrete but in most cases is easiest to measure as a continuous amount of stuff. Extinctions on the other hand either occur or do not occur at a given time and are thus measured with a 0 or 1 only.

Despite these differences in context, all of the examples allow us to answer the same kinds of questions about the quantity being distributed.[7] Thus, we can use the same abstraction to capture all of these cases and more.[8]

Let's return for a moment to the rainfall and lightning distributions discussed above. In both cases, we constructed maps that seemed to be natural, but it is best to keep in mind the story from the introductory quote to this chapter. The information in these maps can be represented in several different ways. For example, suppose that at each point in the continental United States we plot the total number of lightning strikes or the total volume of rainfall in the continental US to the south and west of the given point. To see that the resulting maps contain the same information as the original maps, consider the value on the map at the points $(x, y), (x + \Delta, y), (x + \Delta, y + \Delta), (x, y + \Delta)$ where the distances are measured say in kilometers. Suppose the value of the new map at a point (u, v) is given by $F(u, v)$. Then, $F(x+\Delta, y+\Delta) - F(x+\Delta, y) - F(x, y+\Delta) + F(x, y)$ is, depending on the map, the number of lightning strikes or total volume of rainfall in the Δ^2 square kilometer region bounded by the four points $(x, y), (x+\Delta, y), (x+\Delta, y+\Delta), (x, y+\Delta)$. By letting $\Delta \to 0$, we can get the number of strikes or rainfall volume in an arbitrarily small region around (x, y). For the lightning map, this limit as $\Delta \to 0$ gives us the original map. For the rainfall map, we need to compute

$$\lim_{\Delta \to 0} \frac{F(x + \Delta, y + \Delta) - F(x + \Delta, y) - F(x, y + \Delta) + F(x, y)}{\Delta^2},$$

which gives us the density of rainfall at (x, y) – the original map.

In this class we will use distributions of probability instead of income or mass. In probability theory, the idea of a distribution is fundamental; it represents our state of knowledge about the values a random variable will take. Since we cannot specify the exact value of most random variables before an experiment is completed, we use distributions to describe the likelihood of observing different possible values.

A probability distribution for a random variable contains all the information for predicting the value of the variable that is intrinsic in the experiment.[9] It gives us a predicted answer to any question about the random variable. We will begin with a useful, if somewhat abstract, definition of a probability distribution, but we will soon see that this reduces to "where and how much". We will also see that, as in the examples above, there are many different representations of the *same* information. We can convert among the representations, and we will find that each is useful in certain circumstances.

9 When information about other, related random variables becomes available, we can update this distribution accordingly as we will see in Chapter 4.

2. The Distribution of a Random Variable

Like the elephant in the opening quote of this chapter, a probability distribution can be described in many different ways. When applying probability theory to real problems, we will find that we use a variety of different descriptions, each best suited to the particular situation we face.

In this section, we view the whole elephant. We define the distribution of a random variable as an abstract object from which all other representations will be derived. While, in practice, we will not use this abstract representation very often, it will always lurk in the background, telling us how the side, tusk, trunk, and leg relate to each other. We will also see how the distribution embodies the questions of "where" and "how much" described in the last section.

2.1. The Distribution Motivated

It is in the nature of a random experiment that the values of some random variables are unknown until the experiment is complete. In practice, the best we can do when faced with a random system is to use the available information to make well-reasoned guesses about the answer to any question concerning a random variable of interest. At any moment before or during a random experiment, the distribution of a random variable embodies our current (uncertain) knowledge about the value that random variable will take. In mathematical terms, the distribution is the tool with which we compute our guesses about that random variable.

Suppose that X is a random variable (scalar or vector valued) and that g is a real-valued function that is defined for all values of X. We

10 The function g_3 is defined on $\mathbb{R}^2$. As described in Appendix F, it can be written in two equivalent forms: $g_3(x)$ for $x \in \mathbb{R}^2$ or $g_3(x_1, x_2)$ for $x_1, x_2 \in \mathbb{R}$, whichever is most convenient at any moment.

11 One might wonder why we cannot simply find a good guess for X, $\mathsf{E}X$, and use $g(\mathsf{E}X)$ as our guess for the result of algorithm g applied to X. Unfortunately, this approach loses too much information for the guesses to be good. In general, $g(\mathsf{E}X) \neq \mathsf{E}g(X)$; see equation (2.3).

can think of g as an *algorithm* that takes an input and computes some corresponding result. The random variable $g(X)$ is consequently the result of executing the algorithm g on the value of the random variable X.

For example, consider the algorithms for determining the following:

1. Whether a single input value is between 0 and 1,
2. How far a single input value deviates from the constant c, and
3. Whether the first component of an input ordered pair is less than the second component.

These algorithms can be written as functions: $g_1(u) = 1_{[0,1]}(u)$, $g_2(u) = |u - c|$, and $g_3(u, v) = 1_{]-\infty,0[}(u - v)$ for inputs[10] $u, v \in \mathbb{R}$. If X and Y are real-valued, the random variables $g_1(X)$, $g_2(X)$, and $g_3(X, Y)$ describe the results of the three algorithms applied to the measurements represented by X and Y.

As we saw in Chapter 2, the expected value operator E returns a typical or representative value – in other words, a good guess – of any random variable. Thus, for any algorithm g, $\mathsf{E}g(X)$ is a good guess for the result of applying that algorithm to the value of the random variable X. Everything we know about X before the random experiment begins is contained in the collection of these guesses $\mathsf{E}g(X)$ over all algorithms g. The distribution of X is just a mathematical package containing all of these guesses. If we know the distribution of X, then we can make a good guess of the answer to any question regarding the value that X will take.

Interpretation. The distribution of a random variable gives us a guess as to the result of any algorithm applied to the value of the random variable.

We will define the distribution of a random variable X as the collection of numbers $\mathsf{E}g(X)$ for all real-valued functions g that are defined on the range of X. To make it easier to refer to the distribution of X as an object, we perform a little notational slight of hand. We take all the numbers $\mathsf{E}g(X)$ and index them by the algorithm (function) g, treating the X as understood.[11] We write this as an operator D_X that has one input slot to be filled by an algorithm g and that returns as an output our best guess $\mathsf{E}g(X)$ of the result of that algorithm applied to the value of the random variable X. That is,

$$\mathsf{D}_X g \equiv \mathsf{E}g(X).$$

This form puts the emphasis on the fact that the g is varying here, not the X or E, but there is no other difference between writing $\mathsf{D}_X g$ and $\mathsf{E}g(X)$.

Here is the plan for the rest of the chapter. First, we explore some of the properties of the distribution using the operator D_X. Second, we show that knowing D_X is equivalent to answering the questions "where" and "how much" concerning the probabilities of different values of X. Finally, we show how to derive from D_X several alternative representations of the distribution. (See the table below.) All of these are equivalent to D_X in the sense that if you had any one of them and no other information, you could recover D_X. Nonetheless, because each emphasizes a particular feature of the distribution, they are simpler and often easier to use than D_X itself. Many calculations can be made easier by working with the right representation.

Representation	Notation	For Which RVs	Section
Distribution	D_X	Any	2
Measure	m_X	Any	2
PMF	p_X	Discrete	3
PDF	f_X	Continuous	4
CDF	F_X	Any	6
SF	S_X	Any	6
PGF	G_X	Integer-Valued	7
MGF	M_X	Non-negative	7
CGF	C_X	Any	7

The above table also highlights a convention that we will use consistently. When working with any representation of a distribution, we use the principal symbol to denote which type of object it is (e.g., D for the Distribution Operator, p for a PMF, f for a PDF, and so forth), and we use a subscript to denote what random variable(s) this distribution relates to. For example, random variables X, Y, and Z have distribution (operators) D_X, D_Y, and D_Z and CDF's F_X, F_Y, and F_Z, respectively. The principal symbol is also set a sans-serif typeface to further set it apart from an arbitrary function. As we proceed, we will see some variations on what can appear in the subscript, but in every case, random variables in the subscripts are labels only; their presence does not imply that the function is random in any way.

A common mathematical theme here is the use of functions as packages for collections of related information. This puts the information together in one handy object that can be manipulated and compared while still enabling us to access each piece of information separately. For example, when we want to work with the collection of numbers $\mathsf{E}g(X)$, we package them together in the function (operator) D_X. We can work with all of the numbers simultaneously by manipulating D_X, or we can study a particular $\mathsf{D}_X g$. All of the alternative representations listed above are functions that are created in a similar way. You should keep your eyes open for this use of functions in the coming sections.

12 In fact, we limit ourselves to functions that satisfy a technical condition called measurability, but this is not an important distinction for our purposes. See Appendix D.

2.2. The Distribution Defined

We now give a more complete definition of the distribution operator, D_X.

> If X is a random variable, the distribution of X is the operator D_X that operates on functions $g\colon \mathcal{X} \to \mathbb{R}$, for any $\mathcal{X}$ containing range(X), by
>
> $$\mathsf{D}_X g = \mathsf{E}g(X). \tag{3.1}$$

Notice that the this definition of D_X transparently handles the case of vector-valued X.[13] If $X = (X_1, \ldots, X_n)$ for some integer $n > 1$, then[14]

$$\mathsf{D}_X g = \mathsf{E}g(X_1, \ldots, X_n). \tag{3.2}$$

This is just a more explicit, though completely equivalent, form of equation (3.1).

13 We can extend the definition of D_X to vector-valued functions g using the same trick as with the E operator. If $g = (g_1, \ldots, g_m)$ is an $\mathbb{R}^m$-valued function for some integer $m > 1$, we can define

$$\mathsf{D}_X g = \left(\mathsf{D}_X g_1, \ldots, \mathsf{D}_X g_m\right).$$

However, this generality will only rarely, if ever, be useful.

14 See Section F.2.7.

In general, we allow an ambiguity between vector and component forms for functions and random variables. For example, as in Section F.2.7, if $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and if g is a function defined on $\mathbb{R}^n$, then we can write $g(x)$ or $g(x_1, \ldots, x_n)$ to mean the same thing: the value of the function g at the point x. Convenience determines which we use. A similar situation occurs for random variables. If Y is a vector-valued random variable that takes values in $\mathbb{R}^n$, then Y can be written in component form as $Y = (Y_1, \ldots, Y_n)$ for real-valued random variables Y_i, $i = 1, \ldots, n$. The distribution of Y, D_Y, provides good guesses for the results $g(Y) \equiv g(Y_1, \ldots, Y_n)$ of any algorithm g that takes the values of $Y_1, \ldots, Y_n$ as input.

Sometimes, however, we have a collection of random variables Z_1, ..., Z_m that we have not explicitly grouped as a vector. To describe the probabilistic information in these variables *jointly*, we could simply define a vector-valued random variable $Z = (Z_1, \ldots, Z_n)$ and talk about the distribution of Z. However, we want to be able to talk about the distribution of $Z_1, \ldots, Z_n$ jointly without obscuring the individual identity of these random variables. This gives rise to the idea of the joint distribution of a collection of random variables.

The *joint distribution* of any collection of random variables $Z_1, \ldots, Z_n$, denoted by $\mathsf{D}_{Z_1,\ldots,Z_n}$, is the operator D_Z where $Z = (Z_1, \ldots, Z_n)$. This is a notational conceit only; we have simply skipped the step of defining Z explicitly. The word "joint" here is a mnemonic that indicates we are considering several random variables simultaneously. The operator $\mathsf{D}_{Z_1,\ldots,Z_n}$ acts on a single function of n variables, $g(u_1, \ldots, u_n)$. As we saw earlier, any such function corresponds to an algorithm that takes n input values; $\mathsf{D}_X g$ is a guess for the result of that algorithm on the values of $Z_1, \ldots, Z_n$. The joint distribution packages the simultaneous predictions about these random variables into the distribution of an implicitly defined vector when there is no handy symbol for the vector of those variables.[15] There is no requirement that joint distribution refer to random variables that have the same root symbol; we can consider the joint distribution of any random variables be they called Z_1, Z_2, Z_3 or X, U, S.

15 Perhaps a better notation would be $\mathsf{D}_{(Z_1,\ldots,Z_n)}$, but all those extra parentheses get tiresome.

If $Z_1, Z_2, \ldots, Z_n$ is a collection of random variables, their *joint distribution* is the operator $\mathsf{D}_{Z_1,\ldots,Z_n}$ given by

$$\mathsf{D}_{Z_1,\ldots,Z_n} \equiv \mathsf{D}_Z$$

where $Z = (Z_1, \ldots, Z_n)$.

Often, when we are dealing with several random variables simultaneously, we find that we also want to consider the distribution of one of those variables in isolation. For example, if we are working with the joint distribution of X and Y, we may wish to compute the distribution of X. There is no ambiguity here: The distribution of X is just the distribution of X and the joint distribution of X and Y is just the distribution of the vector (X, Y). However, when in the context of several random variables, it is common to refer to the distribution of one random variable as the *marginal* distribution of that variable. An example motivating the name "marginal" is shown in the next section. We will keep the use of this term to a minimum.

Two random variables X and Y have the same distribution if $\mathsf{D}_X = \mathsf{D}_Y$ as operators. This means that both operators act on the same set of functions and that, for every suitable function g, we must have $\mathsf{D}_X g = \mathsf{D}_Y g$. In words, *two random variables have the same distribution if our predicted answers to every question posed about the two variables are the same.* This means that we have precisely the same probabilistic information about the two variables before the experiment is run. The random variables need not be equal as functions, however. For $\mathsf{D}_X = \mathsf{D}_Y$ to hold, it is necessary (though not sufficient) that X and Y have the same range. As an example, suppose that Z_1, Z_2, and Z_3 are real-valued random variables. Then, the distributions D_{Z_1}, D_{Z_1,Z_2}, and D_{Z_1,Z_2,Z_3} are *not* equal because they act on different types of functions. Notice however that we can get either of the first two operators from the last one. Let g_1 and g_2 be arbitrary real-valued functions defined on $\mathbb{R}$ and $\mathbb{R}^2$ respectively. Define

$$\begin{aligned} g_I(x, y, z) &= g_1(x) \\ g_{II}(x, y, z) &= g_2(x, y). \end{aligned}$$

Then,

$$\begin{aligned} \mathsf{D}_{Z_1,Z_2,Z_3} g_I &= \mathsf{D}_{Z_1} g_1 \\ \mathsf{D}_{Z_1,Z_2,Z_3} g_{II} &= \mathsf{D}_{Z_1,Z_2} g_2. \end{aligned}$$

In other words, *if we know the joint distribution of a collection of random variables, we can find the (joint) distribution of any subset of that collection.*

Once we have the the distribution D_X of X, we can make any specific predictions we need. For example, if we want to know how likely X is to be between 0 and 2, we compute $\mathsf{D}_X 1_{[0,2]} = \mathsf{P}\{\, 0 \le X \le 2 \,\}$. If we want to know how far on average X will be from its expected value, we define $g(u) = |u - \mathsf{E}X|$ and compute[16] $\mathsf{D}_X g = \mathsf{E}|X - \mathsf{E}X|$. D_X contains all of the probabilistic information about X that we have at our disposal.

Nonetheless, we do not usually need to compute D_X explicitly. There are special collections of functions such that if we know $\mathsf{D}_X g$ for every one of these functions g, it is enough to determine D_X. For example, to specify D_X precisely, it is enough know $\mathsf{D}_X 1_{\mathcal{A}}$ for every subset $\mathcal{A}$ of codomain(X). This gives us what is called the *measure* of the distribution of X, m_X, which maps a set $\mathcal{A} \subset$ codomain(X) to the probability

$$\mathsf{m}_X(\mathcal{A}) = \mathsf{D}_X 1_{\mathcal{A}} = \mathsf{P}\{\, X \in \mathcal{A} \,\}\,.$$

This object m_X has the properties to make it a probability measure as defined in Appendix M. Notice that this tells us where the probability is (what sets $\mathcal{A}$) and how much (the size of $\mathsf{P}\{\, X \in \mathcal{A} \,\}$). If you know m_X, then we can compute $\mathsf{D}_X g$ explicitly for any g.[17] In other words, all of the information in the distribution reduces to where and how much. As we will see in later sections, even simpler objects are sufficient to describe the distribution completely.

16 This g is a perfectly well defined function; $\mathsf{E}X$ is a number that does not depend on the specific value X will take in our run of the experiment.

17 Referring to Appendix M, we have that

$$\mathsf{D}_X g = \int g(t)\, d\mathsf{m}_X(t).$$

So, having one is as good as having the other.

18 The more standard and rigorous order of definition is to define distribution operators (measures) first and then relate them to random variables. We have reversed that order here in the hope that it is clearer and because we will use the distributions of random variables much more often than distributions in the abstract.

2.3. Distributions in General

So far, we have only considered probability distributions that are associated with specific random variables. Although this is how we will most often use distributions, probability distributions make sense as stand-alone objects, and it will be useful to be able to define them as such.[18] The key to doing this is to notice that the distribution D_X inherits certain properties as an operator from the E operator used to define it. This leads to the following definition.

Suppose D is an operator with one input slot to be filled by a real-valued function that is defined on $\mathbb{R}^k$ for some $k \in \mathbb{Z}_+$ Suppose also that D satisfies the following for constant c and real-valued functions $g, h, g_1, g_2, \ldots$ on $\mathbb{R}^k$:

1. $\mathsf{D}_X 1 = 1$.
2. $\mathsf{D}_X g \ge 0$ if $g \ge 0$.
3. $\mathsf{D}_X(cg) = c\mathsf{D}_X g$.
4. $\mathsf{D}_X(g + h) = \mathsf{D}_X g + \mathsf{D}_X h$.

5. If $g_1 \le g_2 \le \cdots$ converge to g, then $\lim_{n\to\infty} \mathsf{D}_X g_n = \mathsf{D}_X g$.

Then, D is a probability distribution on $\mathbb{R}^k$. From this, we can derive alternative representations as with any other distribution; see the following sections.[19]

If D is a probability distribution, we say then that a random variable X has distribution D if and only if $\mathsf{D}_X = \mathsf{D}$.

19 For those referring back: the PMF and PDF of D would be functions p and f. We will use generic identifying subscripts in such cases as necessary.

2.4. Summary

The distribution of a random variable contains all the knowledge that we have about that random variable before the experiment begins. All of our predictions about a random variable will be based on its distribution. However, the distribution has a great deal of built-in redundancy, and in practice, it will usually be inconvenient to derive the operator explicitly. As with the maps of lightning strikes and rainfall, we will find that we can describe all of the information in the distribution in several more convenient ways. We turn next to describing these.

3. The Probability Mass Function for Discrete Distributions

A distribution is *discrete* if it assigns all of its probability to a countable set; a random variable is said to be discrete if its distribution is discrete. Formally, the distribution of a random variable X is discrete if there is a countable set[20] $\mathcal{C}$ such that $\mathsf{P}\{ X \in \mathcal{C} \} = 1$. A random variable X is discrete if we can explicitly enumerate all of its possible values in a (perhaps infinitely long) list $x_1, x_2, x_3, \ldots$. Any random variable that takes values in a finite set or in any subset of the integers is thus necessarily discrete.

20 Recall this means finite or countably infinite.

▶ EXAMPLE 3.1. Balanced Die.

Consider an experiment in which we roll a balanced, standard six-sided die. Let D be the number on the face that ends up on top.

D is discrete because it takes values in the finite set $\{1, 2, 3, 4, 5, 6\}$.

▶ EXAMPLE 3.2. Radioactive Decay.

Radioactive substances such as Uranium undergo random radioactive decay. The decay events are scattered randomly over time. Let the random variable N be the number of radioactive decay events from a one kilogram block of Uranium during a specified second.

N is discrete because it takes integer values.

► EXAMPLE 3.3. Random Point on a Circle.

Let Z be a randomly selected point on the unit circle at one of the angles $2\pi k/n$ for integers n and $1 \le k \le n$. (Angles are measured counterclockwise from the positive x direction.)

Z is a discrete vector-valued random variable because it takes values in a finite set. We can write Z in vector form by $Z = (Z_1, Z_2)$. The real-valued random variable Z_1 takes values $\cos(2\pi k/n)$ for $k = 1, \ldots, n$, and the real-valued random variable Z_2 takes values $\sin(2\pi k/n)$ for $k = 1, \ldots, n$ Both of these are also discrete.

For random variables like these, it seems we should be able to specify directly how the probability is distributed among the possible observed values, without all the trouble of creating a distribution operator as in the last section. Indeed, if X is discrete and takes values in the countable set $\{x_1, x_2, \ldots\}$, all we need do is to specify the probability $p_i \ge 0$ (how much) assigned to each value x_i (where). That is, we have $\mathsf{P}\{\, X = x_i \,\} = p_i$ for $i = 1, 2, \ldots$. Since X must take *some* value, we must check that $\sum_i p_i = 1$ to ensure that we have accounted for all the possibilities.

The distribution of a discrete random variable X is therefore relatively simple to specify. It is necessary to give the list of possible values and the probabilities that X takes each value. We will package this information in a *function*, p_X, called the Probability Mass Function (PMF). The PMF takes as input a single candidate value of X and returns the probability that X takes that value: $\mathsf{p}_X(u) = \mathsf{P}\{\, X = u \,\}$. It is convenient for working with PMFs to allow as input any u in the codomain of X – any real number for a real-valued random variable, any k-vector for an $\mathbb{R}^k$-valued random variable – even if range(X) is a strict subset of the codomain. By definition, $\mathsf{p}_X(u)$ necessarily returns 0 for any value outside range(X). (It may return zero for some values in the range as well.) For example, we will often work with random variables that take only integer values. The range in this case is a subset of integers, but we take the codomain to be the real numbers and define the PMF on $\mathbb{R}$.

Practical Definition of the PMF

The PMF of a random variable X satisfies

$$\mathsf{p}_X(u) = \mathsf{P}\{\, X = u \,\} \tag{3.3}$$

for all $u \in \text{codomain}(X)$.

The words mass in "probability MASS function" evokes a useful metaphor: you can think about probability as a kind of massive stuff that is distributed around space. We begin with 1 unit (kg, say) of total mass and "distribute" it across the possible values of the random variable. In the discrete case, we break off pieces of our original supply and place them at *precise* locations. In the continuous case, to be described next section, we smear the mass around. The distribution of the random variable describes where the mass is placed and how much is placed at any location. Finding the probability that a random variable takes values in a particular set of possibilities corresponds to measuring the mass within that set. Finding expected values will correspond to finding balancing points, as we will see below. The PMF embodies the information contained in two lists of numbers – the list of locations, and the list of amounts at those precise locations. It only makes sense to define the PMF when we are distributing probability in discrete chunks. The PMF is therefore only defined for discrete random variables.

▶ EXAMPLE 3.1 Revisited.

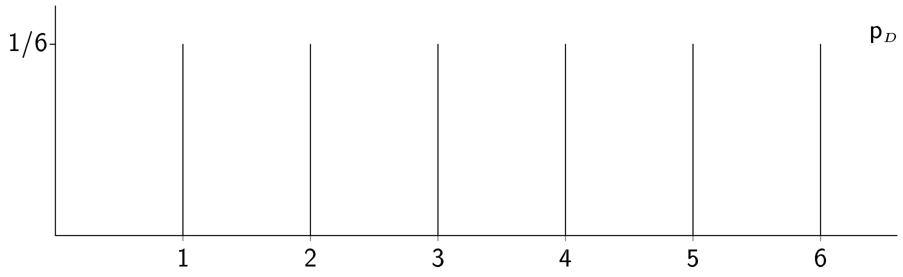
Because the die is balanced, we expect all six sides to come up equally often in the long run. Put another way, we have by symmetry that $\mathsf{P}\{\, D = k \,\} = 1/6$ for $k = 1, 2, 3, 4, 5, 6$. The PMF of D, p_D, is a function on $\mathbb{R}$, and we can write it in two equivalent ways:

$$\mathsf{p}_D(u) = \begin{cases} \frac{1}{6} & \text{if } k \in \{1, 2, \ldots, 6\} \\ 0 & \text{otherwise.} \end{cases}$$

or

$$\mathsf{p}_D(u) = \frac{1}{6}\, 1_{\{1,2,\ldots,6\}}(u).$$

You can use whichever form makes the most sense to you.



▶ EXAMPLE 3.2 Revisited.

Under a commonly used model for the radioactive decay events, we have that, for a fixed constant $\lambda > 0$

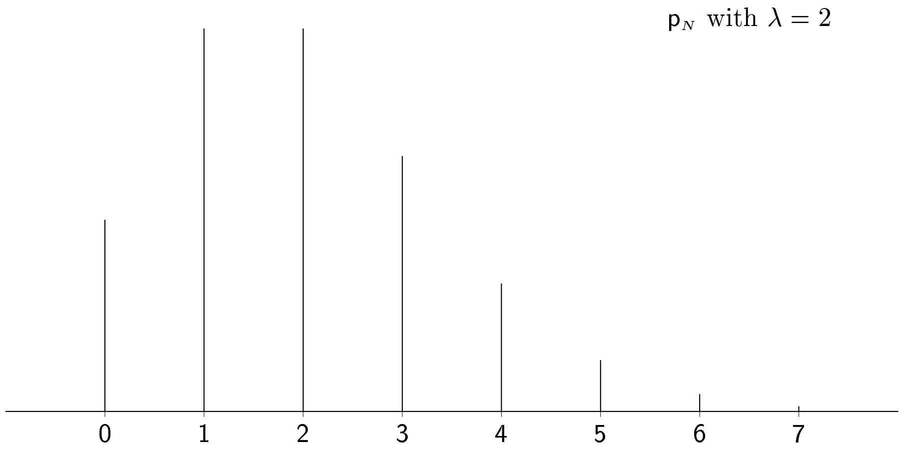
$$\mathsf{P}\{\, N = k \,\} = e^{-\lambda}\, \frac{\lambda^k}{k!},$$

when $k \in \mathbb{Z}_\oplus$. This random variable can thus take arbitrarily large values (at least in our model). We can again write the PMF in two equivalent ways

$$\mathsf{p}_N(u) = \begin{cases} e^{-\lambda}\, \frac{\lambda^u}{u!} & \text{if } u \in \mathbb{Z}_\oplus \\ 0 & \text{otherwise.} \end{cases}$$

or

$$\mathsf{p}_N(k) = e^{-\lambda}\, \frac{\lambda^u}{u!}\, 1_{\mathbb{Z}_\oplus}(u).$$



Notice that N must take *some* value, so

$$1 = \mathsf{P}\{\, N \in \mathbb{R} \,\} = \sum_{k=0}^{\infty} \mathsf{P}\{\, N = k \,\} = \sum_{k=0}^{\infty} \mathsf{p}_N(k).$$

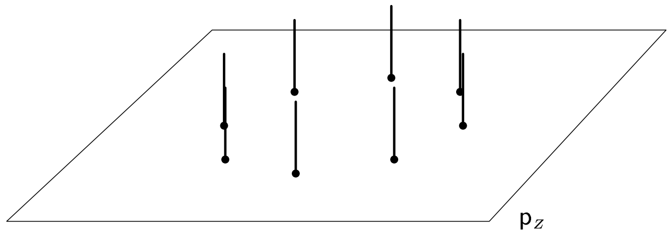
The second equality is true because $\{\, N \in \mathbb{R} \,\}$ is a disjoint union of the events $\{\, N = k \,\}$ for $k \in \mathbb{Z}_\oplus$ and because $\sum_{k=0}^{\infty} \lambda^k/k! = e^{\lambda}$.

▶ EXAMPLE 3.3 Revisited.

That Z is a vector-valued random variable does not change the basic logic of the PMF. Z can be one of eight possible points, and all are equally likely by assumption. It follows that

$$\mathsf{p}_Z(v) = \begin{cases} \frac{1}{8} & \text{if } v = (\cos 2\pi k/8, \sin 2\pi k/8) \text{ for } k = 0, \ldots, 7, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that the argument to p_Z must be a vector in $\mathbb{R}^2$.



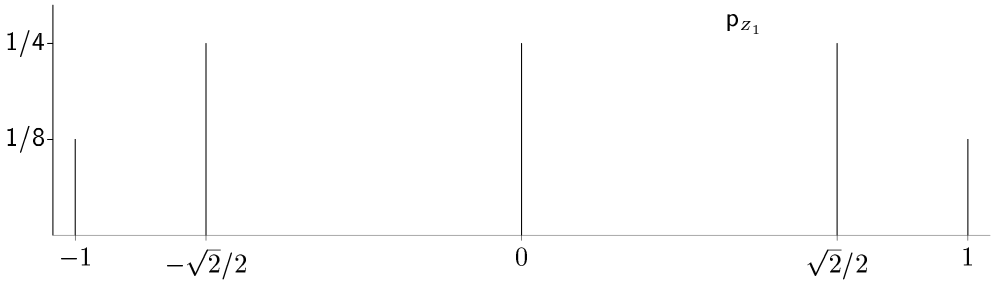
We can also use p_Z to compute p_{Z_1} as follows

$$\mathsf{p}_{Z_1}(u) = \sum_w \mathsf{p}_Z(u, w),$$

and similarly for p_{Z_2}. For example, $\mathsf{p}_{Z_1}(\cos 2\pi 5/8) = \mathsf{p}_Z(\cos 2\pi 5/8, \sin 2\pi 5/8)+$ $\mathsf{p}Z(\cos 2\pi 5/8, \sin 2\pi 3/8) = 1/4$. We have then

$$\mathsf{p}_{Z_1}(t) = \begin{cases} \frac{1}{8} & \text{if } t = \pm 1 \\ \frac{1}{4} & \text{if } t = \cos 2\pi k/8 \text{ for } k = 1, 2, 3 \\ 0 & \text{otherwise,} \end{cases}$$

and that $\mathsf{p}_{Z_2} = \mathsf{p}_{Z_1}$. Notice that knowing p_{Z_1} and p_{Z_2} is not enough to derive p_Z; p_Z contains information not only about Z_1 and Z_2 separately but about the values they take jointly.



So far, we have motivated the PMF without any connection to the distribution as defined in the last section. We now describe the connection between D_X and p_X and what it means. This is important for two reasons: (i) it easily gives us the properties that a PMF must satisfy, and (ii) it reveals the common structure underlying all our calculations, whether or not the random variables are discrete.

Formally, the PMF of a random variable X is derived from the distribution D_X. It is a function, uniquely determined by D_X, that tells us how to compute expected values as weighted averages.

Formal Definition of the PMF

The probability mass function (PMF) of a random variable X is the unique function p_X defined on codomain(X) that is zero outside range(X) and such that

$$\mathsf{D}_X g = \sum_u g(u)\, \mathsf{p}_X(u) \tag{3.4}$$

for all real-valued functions g for which $\mathsf{D}_X g$ is defined. The PMF exists for any discrete distribution.

Because p_X is zero outside range(X), the sum in equation (3.4) can be taken to be over $u \in$ range(X) (or any superset thereof).

It is important to keep in mind that the PMF is a feature of the distribution; it does not depend on the value $X(\omega^*)$ that X will take in our experiment. To be strictly accurate, we should call p_X the probability mass function (PMF) of the *distribution* of X. We tend to use the shorter phrasing for convenience.

Equation (3.4) shows that if you have p_X, then we can recover D_X: given a g, we compute the weighted average of the values $g(u)$ with $\mathsf{p}_X(u)$ as weights. The reverse is true as well. If $g_u(x) = 1_{\{u\}}(x)$, i.e., $g(x) = 1$ if $x = u$ and 0 otherwise, then $\mathsf{p}_X(u) = \mathsf{D}_X g_u = \mathsf{P}\{\, X = u \,\}$.

This equivalence shows that the PMF p_X contains as much information as the distribution operator D_X. Therefore, if we want to find the distribution of a discrete random variable, it is sufficient to find the PMF. There is no need to compute the distribution operator explicity.

Equation (3.4) also reveals the properties of any PMF. For example, PMFs must be non-negative functions. By the Non-negativity Rule, $1_{\{u\}} \ge 0$ for every u, so $\mathsf{D}_X 1_{\{u\}} \ge 0$ as well. Hence, for all values u, $\mathsf{p}_X(u) \ge 0$. Similarly, since $\mathsf{D}_X 1 = 1$, we have that $\sum_u \mathsf{p}_X(u) = 1$. Indeed, if $\mathcal{A}$ is any set of possible values of X, we have

$$\mathsf{P}\{\, X \in \mathcal{A} \,\} = \mathsf{D}_X 1_{\mathcal{A}} = \sum_{u \in \mathcal{A}} \mathsf{p}_X(u).$$

This tells us how to compute the probability of any event relating only to the values of X.

Properties of the PMF

- Non-negativity. $\mathsf{p}_X(u) \geq 0$ for every u.
- Total Probability. $\sum_u \mathsf{p}_X(u) = 1$.
- Probability in a set. $\mathsf{P}\{\, X \in \mathcal{A} \,\} = \sum_{u \in \mathcal{A}} \mathsf{p}_X(u)$.

It is worth restating (and remembering) the immediate implication of equation (3.4):

How to compute expected values with the PMF

If X is a discrete random variable and g is a real-valued function defined on range(X), then

$$\mathsf{E} g(X) = \sum_u g(u)\, \mathsf{p}_X(u),$$

where the sum can be taken over any set containing range(X).

This tells us a rather startling fact. To get the long-run average of $g(X)$ over many replications of our experiment, we need only compute a weighted average. We average the values $g(X)$ can take using the probabilities of each value of X as weights.

As an example, we have

$$\mathsf{E} X = \sum_u u\, \mathsf{p}_X(u)$$

taking $g(y) = y$ for all y. Referring back to our analogy of probability as mass, we can get some insight into the meaning of $\mathsf{E}X$. The quantity $\mathsf{p}_X(u)$ represents the amount of mass we have placed at the point u along the real line. The sum

$$\sum_u (u - c)\, \mathsf{p}_X(u) = \sum_u u\, \mathsf{p}_X(u) - \sum_u c\, \mathsf{p}_X(u) = \sum_u u\, \mathsf{p}_X(u) \;-\; c$$

is proportional to the total torque on the line generated by the distribution of mass. Hence, $c = \mathsf{E}X$ is the point for placing a fulcrum at which the total torque is zero; or, $\mathsf{E}X$ is the balancing point of the distribution of X.

In general, we will find that *expected values can be computed by taking weighted averages of the values of a random variable using the probabilities of these values as weights.* Whenever possible we will compute expected values without using this equation – using the properties of the E operator for instance – because it usually involves messy details. Nonetheless, we often have no choice.

Finally, equation (3.4) works transparently for vector-valued random variables. If $X = (X_1, \ldots, X_n)$ is a vector-valued random variable and $u = (u_1, \ldots, u_n)$ be a possible value of X, then we can compute expected values of random variables such as $g(X) \equiv g(X_1, \ldots, X_n)$ by averaging values $g(u_1, \ldots, u_n)$ with weights $\mathsf{p}_X(u)$. But what does $\mathsf{p}_X(u)$ mean? Remember that two vectors are equal if and only if all their components are equal, so the event $\{ X = u \}$ is just the intersection of the events $\{ X_i = u_i \}$ for $i = 1, \ldots, n$. Put another way,

$$\mathsf{p}_X(u) \equiv \mathsf{p}_X(u_1, \ldots, u_n) = \mathsf{P}\{ X_1 = u_1 \text{ and } \cdots \text{ and } X_n = u_n \}. \quad (3.5)$$

By direct analogy, we define the *joint* PMF of real-valued random variables $Z_1, \ldots, Z_m$ by

$$\mathsf{p}_{Z_1,\ldots,Z_m}(u_1, \ldots, u_m) = \mathsf{P}\{ Z_1 = u_1 \text{ and } \cdots \text{ and } Z_m = u_m \}. \quad (3.6)$$

This is again just notational slight of hand that makes it easier to keep track of which PMF we are using. The joint PMF of these Z_is is exactly p_Z where $Z = (Z_1, \ldots, Z_m)$ is the vector with the Z_is as components. It follows that the joint PMF must satisfy all the properties of PMFs listed above. For instance, it is non-negative, sums to 1 over all $u_1, \ldots, u_n$, and gives $\mathsf{P}\{ (Z_1, \ldots, Z_n) \in \mathcal{A} \}$ by summing $\mathsf{p}_{Z_1,\ldots,Z_n}$ over values $(u_1, \ldots, u_n) \in \mathcal{A}$.

▸ EXAMPLE 3.4. Demographic Table

The horned slimy-toed lizard comes in several varieties: they can have from 1–4 horns and can have from 1–3 slimy toes. Field biologists have done a careful census of the population of these lizards and have found the following demographics:

		Horns 1	2	3	4	
Toes	1	0.15	0.01	0.04	0.05	0.25
	2	0.10	0.25	0.00	0.04	0.39
	3	0.02	0.04	0.15	0.15	0.36
		0.27	0.30	0.19	0.24	1.00

Suppose that the random variables T and H represent the number of toes and number of horns of a randomly chosen lizard from the population.

The joint PMF of T and H is the function $\mathsf{p}_{T,H}(u, v) = \mathsf{P}\{T = u \text{ and } H = v\}$. To clarify the notation, notice that the first (second) argument is a possible value of the first (second) listed random variable in the subscript. The table above gives the values of $\mathsf{p}_{T,H}(u, v)$ for $u = 1, 2, 3$ and $v = 1, 2, 3, 4$. Each cell in the table gives the probability of a particular combination of values for T and H. This is the sense that $\mathsf{p}_{T,H}$ represents the *joint* distribution.

We obtained the numbers in the margins of the table by summing the values in each row or column. For example,

$$\begin{aligned}
0.25 &= \mathsf{p}_{T,H}(1, 1) + \mathsf{p}_{T,H}(1, 2) + \mathsf{p}_{T,H}(1, 3) + \mathsf{p}_{T,H}(1, 4) \\
&= \mathsf{P}\{\, T = 1 \text{ and } H \in \{1, 2, 3, 4\} \,\} \\
&= \mathsf{P}\{\, T = 1 \,\} \\
&= \mathsf{p}_T(1).
\end{aligned}$$

These numbers in the margin give the PMFs of T and H considered separately, what we called the *marginal* distributions of T and H above.

4. The Probability Density Function for Continuous Distributions

The result of a dice roll is definite, exact; only specific values are possible. Continuous quantities like times, distances, and angles, on the other hand, are necessarily inexact, and we can measure them with arbitrarily high precision.[21] Because in practice, we record our data with limited precision, it might be tempting to treat such variables as discrete. However, from a modeling viewpoint, it is very difficult to deal with even even 10^{10} possible values.[22] The PMF of one such variable

21 Within the limits dictated by the Heisenberg Uncertainty Principle, of course.

22 Approximately the number of single-precision values between 0 and 1.

becomes unwieldy to work with; doing probabilistic calculations in any realistic problem becomes infeasible. It is far easier and more effective to treat these variables as taking a continuum of possible values. Times and distances, for example, are modeled as arbitrary, non-negative real numbers; angles can be modeled as arbitrary real numbers in the interval $[0, 2\pi[$.

23 See Appendix S, Section 4.

By making this continuum approximation, we get a theory of greater scope and power. The cost is that we have to pay attention to the difference between discrete and continuous random variables and distributions. Whereas discrete distribution places concentrated chunks of probability at specific individual values, a continuous distribution *smears* the probability over an uncountable range of values. In terms of our probability-as-mass metaphor, a discrete distribution assigns a mass of probability to each point; a continuous distribution assigns a density of probability locally around a point. The latter is necessary because a continuous distribution assigns probabilities over an uncountable set of values. Because we cannot explicitly enumerate the elements in an uncountable set,[23] we cannot assign definite, non-zero probabilities to each specific point and still get meaningful results (e.g., a finite total probability).

A distribution is *continuous* if it assigns zero probability to every countable set; a random variable is said to be continuous if its distribution is continuous. Formally, the distribution of a random variable X is continuous if $\mathsf{P}\{\, X \in \mathcal{C} \,\} = 0$ for every countable set $\mathcal{C}$. A continuous random variable can take uncountably many different values. For example, a real-valued random variable that can take any value in some interval (e.g., times, distances, angles) is necessarily continuous.

The definition of a continuous distribution has a striking implication which is easily the most counter-intuitive result we will see: if X is a random variable with a continuous distribution,

$$\mathsf{P}\{\, X = u \,\} = 0 \tag{3.7}$$

for every value u, even though

$$\mathsf{P}\{\, X \in \mathrm{range}(X) \,\} = 1.$$

In other words, the probability that some value is measured is 1, but the probability that any specific value is measured is 0.

The coin-flipping example in Chapter 1 provides an interesting example of this phenomenon. Each elementary outcome ω in that model corresponds to a number in $[0,1]$. Let S be the random variable that gives this numerical value for each ω; S is the random variable that encodes the entire infinite sequence of flips. Notice that S can take any value in the interval $[0,1]$, so it is a continuous random variable. Since all of the flips are fair and independent, the probability of any particular sequence of n flips is 2^{-n}, As $n \to \infty$, the probability of any particular sequence of n flips approaches 0, so the inevitable conclusion must be that the probability of any particular infinitely long sequence of flips is 0. However, we can still get meaningful probabilities if instead of looking at single elementary outcomes, we look at events corresponding to ranges of $\omega \in [0,1]$. For example, the events $\{\lfloor 4S \rfloor = k\}$ for $k = 0,1,2,3$ corresponds to the intervals $[0,0.25]$, $]0.25,0.5]$, $]0.5,0.75]$, and $]0.75,1]$ in Ω, and each has probability 1/4. For $s \in [0,1]$, even though $\mathsf{P}\{S = s\} = 0$, any small interval containing s has positive probability.

In general, if the distribution of X assigned positive probability to uncountably many points, there would simply be too much probability,[24] and nonsense would result. Put another way, there is nothing in our basic expected value rules that tell us how to compute the expected value of an *uncountably* infinite sum. So do we face a basic contradiction? Does our entire abstraction fall apart? Fortunately, no. Equation (3.7) simply tells us that we have to be more careful in adding up the probabilities. We will use the calculus, which was invented for precisely this problem. Instead of specifying the mass of probability *at* each particular point, we will specify the density of probability *near* each particular point. Instead of explicitly summing over a list of possible values to compute expected values, we will use integrals over ranges of possible values.

24 It is possible to have (countably) infinite sums that are finite, but with uncountably many terms, any sum with positive summands is always infinite.

▶ EXAMPLE 3.5. Uniform Random Numbers.

We wish to choose a random number in the unit interval $[0,1]$ in a way that treats every point in the same way. Let U be the number chosen when the experiment is run. Intuitively, we want the distribution of U to guarantee that no point is "more likely" to be chosen than any other, but it is not clear what this means: no matter how we construct the distribution of U, we will have $\mathsf{P}\{U = x\} = 0$ for all $x \in [0,1]$.

► EXAMPLE 3.6. Light-bulb Lifetimes.
Let L be the working lifetime of a light bulb that is turned on at time 0 and left on until it fails. We model L as a continuous random variable that can take any non-negative real value.[25] Although L can be arbitrarily large, any reasonable distribution for L should account for aging by assigning smaller probabilities to longer lifetimes.

► EXAMPLE 3.7. Random Point in a Triangle.
Consider the experiment in which we choose a random point uniformly within an equilateral triangle with vertices at $(-1, 0)$, $(1, 0)$, and $(0, \sqrt{3})$. Let $Z = (X, Y)$ be the coordinates of this randomly chosen point. The word "uniformly" is a statement of symmetry that no point in the triangle is special; the probability is spread evenly.

To find the distribution of a random variable, we need to answer two questions – where and how much – about the probability of observing different possible values. But with a continuous distribution, it is unclear how to answer the second question because the "how much" seems to be zero at every point. The key to finding a better answer is to think about how much probability is assigned near, not just at, each point.

Suppose that X is a random variable with a continuous distribution. Let u be a possible value of X and let $\Delta > 0$ be a constant.[26] Since we cannot use $\mathsf{P}\{\, X = u \,\}$ to quantify the probability around u, we will instead consider the probability in an interval $[u, u + \Delta[$ containing u.[27] This is $\mathsf{P}\{\, u \le X < u + \Delta \,\}$, which can certainly be non-zero. Unfortunately, because $[u, u + \Delta[$ contains many points not-equal to u, $\mathsf{P}\{\, u \le X < u + \Delta \,\}$ does not seem characteristic of the point u itself. The natural first approach to this problem is to make Δ small to eliminate extraneous values. For small Δ, we have the probability of values close to u; but how close is close enough. As we take Δ smaller and smaller, $\mathsf{P}\{\, u \le X < u + \Delta \,\}$ goes to 0:

$$\lim_{\Delta \to 0} \mathsf{P}\{\, u \le X < u + \Delta \,\} = 0.$$

We could all agree on some definite but small value for Δ and quantify the probability at u with the probability in $[u, u + \Delta[$. However, this is less than satisfying, and it defeats the practical purpose of using continuous distributions in the first place. So, we are back where we started.

25 L is 0 if it fails when first turned on.

26 If X is $\mathbb{R}^n$-valued, then we take $\Delta \in \mathbb{R}^n$ as well. In this case, the condition $\Delta > 0$ implies that every component of Δ is positive. See Appendix S.2.4.

27 If X is vector-valued, the set of points $[u, u + \Delta[$ defines a hyper-rectangle. See also the discussion of hyper-rectangles in Appendix S. ??.

Our probability-as-mass analogy suggests a way out of this conundrum. Consider for a moment a gas in a box. This gas consists of discrete molecules whizzing around at high speed, but unless we are interested in the details of those molecules, its easier to think of a gas as a continuous smear of substance throughout the box. Naturally, as we measure the mass of gas in smaller and smaller regions, we get values going to zero, but the *density* of the gas – the mass per unit volume – does not go to zero. So, if a continuous distribution of probability smears the probability over some volume,[28] this suggest we look at the density of probability around each point.

28 The word volume here is generic. One dimensional volume is length; two dimensional volume is area; three dimensional volume is, well, volume; and so forth even into higher dimensions.

The probability density function of X, f_X, which we usually call the density or the PDF, gives at each point the density of probability around that point. If X is a real-valued random variable, the PDF at a point u is given by

$$\mathsf{f}_X(u) = \lim_{\Delta \to 0} \frac{\mathsf{P}\{\, u \le X < u + \Delta \,\}}{\Delta}. \tag{3.8}$$

The ratio within the limit is the density (mass per unit length) of probability in the neighborhood $[u, u+\Delta[$. By taking the limit, we eliminate nearby values from consideration and get a number that is truly a function of u. And by considering the density, the limit is not necessarily 0.

In the general case of $\mathbb{R}^k$-valued X, Δ is a vector with positive components $\Delta_1, \ldots, \Delta_k$, and the ($k$-dimensional) volume of the hyper-rectangle $[u, u+\Delta[$ is just the product $\Delta_1 \cdots \Delta_k$. Thus, in this case, we can write

$$\mathsf{f}_X(u) = \lim_{\Delta \to 0} \frac{\mathsf{P}\{\, u \le X < u + \Delta \,\}}{\Delta_1 \cdots \Delta_k},$$

where the equation holds for any $k \ge 1$.

The equations above that give the PDF in terms of limits are rigorously correct and meaningful. But in practice, it helps to have a looser, intuitive way to think about the PDF; for this, we turn to the old idea of an infinitesimal. Although the calculus was developed independently by Newton and Leibniz, much of the notation we use today comes from Leibniz. The "infinitesimals" like dx that appear in derivatives and integrals – df/dx and $\int f(x)\, dx$ – serve today as semantic placeholders, but they have another useful meaning. An infinitesimal quantity is one that is not zero but is arbitrarily – infinitely – small. We think of it as bigger than zero but smaller than any positive number. From a rigorous mathematical viewpoint, this is nonsense; there is nothing between 0 and the positive real numbers. What is actually meant is a more complicated limiting process, but infinitesimals provide a conceptually powerful short hand.

If we replace Δ above by an infinitesimal du, then we no longer need a limit, and equation (3.8) becomes

$$\mathsf{f}_X(u) = \frac{\mathsf{P}\{\, u \leq X < u + du \,\}}{du},$$

or equivalently

$$\mathsf{f}_X(u)\, du = \mathsf{P}\{\, u \leq X < u + du \,\}.$$

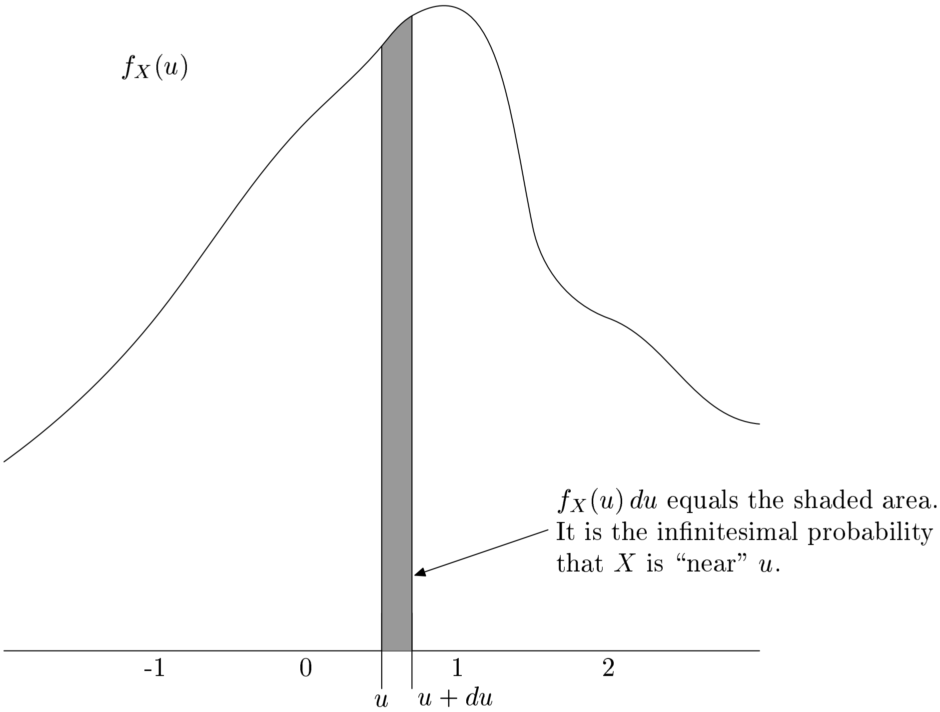
The vector-valued case works as above. If X takes values in $\mathbb{R}^k$, then we use du to stand for the vector of infinitesimals $(du_1, \ldots, du_k)$ and the infinitesimal volume $du_1 \cdots du_k$, letting the ambiguity be made clear from context. Hence, in general, we can write

$$\mathsf{f}_X(u)\, du = \mathsf{P}\{\, u \leq X < u + du \,\}, \tag{3.9}$$

or more explicitly,

$$\begin{aligned} &\mathsf{f}_X(u_1, \ldots, u_k)\, du_1 \cdots du_k = \\ &\qquad \mathsf{P}\{\, u_1 \leq X_1 \leq u_1 + du_1 \text{ and } \cdots \text{ and } u_1 \leq X_1 \leq u_1 + du_1 \,\}. \end{aligned}$$

The above equations have the right units. The infinitesimal du is an infinitesimal unit of volume and $\mathsf{f}_X(u)$ is a density (mass per unit volume), so the product $\mathsf{f}_X(u)\, du$ has units of mass. The latter is the infinitesimal probability that X is "near" u.



For example, in one dimension, du, is an infinitesimal unit of length and $\mathsf{f}_X(u)$ measures the probability per unit length around u. In the graph of the function $\mathsf{f}_X(u)\,du$ when X is real valued, we can think of this infinitesimal probability as an infinitesimal slice of area under the curve f_X.

To capture the above in a simple package, we define the near operator. indicates whether two values are within an infinitesimal distance of each other. We say that x near u if $x \in [u, u + du[$ for some infinitesimal du. Therefore, $x = u$ implies that x near u, but x near u also allows a little smearing. We can write f_X in terms of the near operator with $\mathsf{f}_X(u)\,du = \mathsf{P}\{\, X \text{ near } u \,\}$. Although infinitesimal lengths and the near operator are mythical creatures – like the unicorn, the honest politician, and the delta function – they have consistent logic and semantics that work. We will use the near operator to help clarify how we use PDFs and continuous random variables.

Practical Definition of the PDF

The PDF of a random variable X satisfies

$$\mathsf{f}_X(u)\,du = \mathsf{P}\{\,X \text{ near } u\,\} \tag{3.10}$$

for all $u \in \text{codomain}(X)$.

Just as with the PMF, the PDF of a random variable X is formally derived from the distribution D_X. It is a function, uniquely determined by D_X, that tells us how to compute expected values as weighted averages.

Formal Definition of the PDF

The probability density function (PDF) of a random variable X is the unique function f_X defined on codomain(X) that is zero outside range(X) and such that

$$\mathsf{D}_X g = \int g(u)\,\mathsf{f}_X(u)\,du \tag{3.11}$$

for all real-valued functions g for which $\mathsf{D}_X g$ is defined. The PDF exists for any continuous distribution.

Because f_X is zero outside range(X), the integral in equation (3.11) can be taken to be over range(X) or any superset thereof.

As with the PMF, the PDF is a feature of the distribution; it does not depend on the value $X(\omega^*)$ that X will take in our experiment. To be strictly accurate, we should call f_X the PDF of the *distribution* of X, but we tend to use the shorter phrasing for convenience.

Equation (3.11) shows one can find D_X given only f_X and vice versa. Given any function g, we can get $\mathsf{D}_X g$ by integrating g against $\mathsf{f}_X(u)\,du$. In the other direction, we can define $g_{u,\Delta}(x) = (1/\Delta)\,1_{[u,\Delta[}(x)$. Then, as in equation (3.8) and its generalizations,

$$\mathsf{f}_X(u) = \lim_{\Delta\to 0} \mathsf{D}_X g_{u,\Delta}.$$

This equivalence shows that the PDF contains as much information as the distribution operator D_X, just packaged in a convenient form. Therefore, if we want to find the distribution of a continuous random variable, it is sufficient to find the PDF. There is no need to compute the distribution operator explicitly.

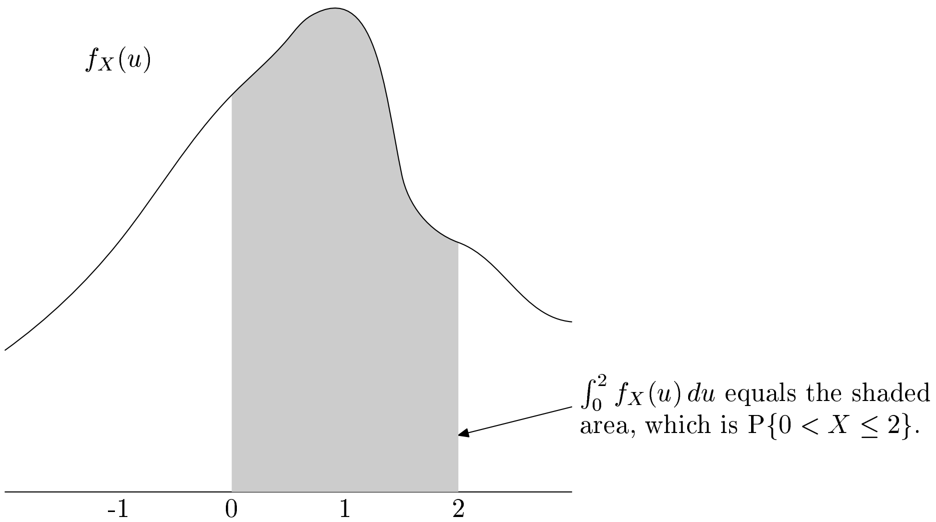
Equation (3.11) also reveals the properties of any PDF. For example, by the Non-negativity rule of expected values and the above limits, any PDF must be a non-negative function. Similarly, since $\mathsf{D}_X 1 = 1$, we have that $\int \mathsf{f}_X(u)\, du = 1$, where the integral is over the range of X. Indeed, if $\mathcal{A}$ is any set of possible values of X, we have

$$\mathsf{P}\{\, X \in \mathcal{A} \,\} = \mathsf{D}_X 1_{\mathcal{A}} = \int_{\mathcal{A}} \mathsf{f}_X(u)\, du = \int_{u \in \mathcal{A}} \mathsf{P}\{\, X \text{ near } u \,\}. \qquad (3.12)$$

Integrals are defined precisely to add up infinitesimal quantities, and the first integral sums up the infinitesimal probabilities $\mathsf{f}_X(u)\, du$ near u over elements of the set $\mathcal{A}$. For example, if X is a real valued random variable, we can use this to compute $\mathsf{P}\{\, 0 < X \le 2 \,\}$ by

$$\mathsf{P}\{\, 0 < X \le 2 \,\} = \mathsf{P}\{\, X \in\ ]0, 2]\ \} = \int_0^2 \mathsf{f}_X(u)\, du.$$

This integral adds up all the infinitesimal slices of area under the graph of f_X between 0 and 2.



The second integral in equation (3.12) expresses $\mathsf{P}\{\, X \in \mathcal{A} \,\}$ in terms of the near operator to emphasize the commonality of the operations in the discrete and continuous cases. We will see more about this in the next section.

Properties of the PDF

- Non-negativity. $\mathsf{f}_X(u) \geq 0$ for every u.
- Total Probability. $\int \mathsf{f}_X(u)\,du = 1$.
- Probability in a set. $\mathsf{P}\{\, X \in \mathcal{A} \,\} = \int_{\mathcal{A}} \mathsf{f}_X(u)\,du$.

It is worth restating (and remembering) the immediate implication of equation (3.11):

How to compute expected values with the PDF

If X is a continuous random variable and g is a real-valued function defined on range(X), then

$$\mathsf{E}g(X) = \int g(u)\,\mathsf{f}_X(u)\,du = \int_u g(u)\,\mathsf{P}\{\, X \text{ near } u \,\},$$

where the integral can be taken over any set containing range(X).

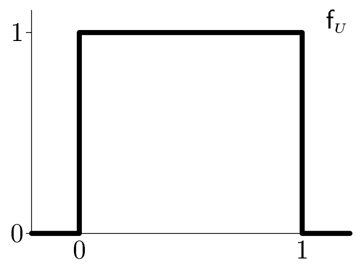
Just as in the discrete case, we see that we can compute expected values as weighted averages the values a random variable can take using the probability at or near each value as weights.

▶ EXAMPLE 3.5 Revisited.

In terms of the PDF, it is now clear how to define the distribution of U. To get total symmetry among the points in $[0,1]$, we need the same density of probability near each point. That is, $\mathsf{f}_U(t) = c$ for some constant c whenever $0 \leq t \leq 1$. (We must have $\mathsf{f}_U(t) = 0$ for t outside $[0,1]$ because U can only take values from 0 to 1.) We find the constant c using the rule of total probability: since $c = \int \mathsf{f}_U(t)\,dt = 1$, c is determined. Hence, we define

$$\mathsf{f}_U(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

as shown in the following figure

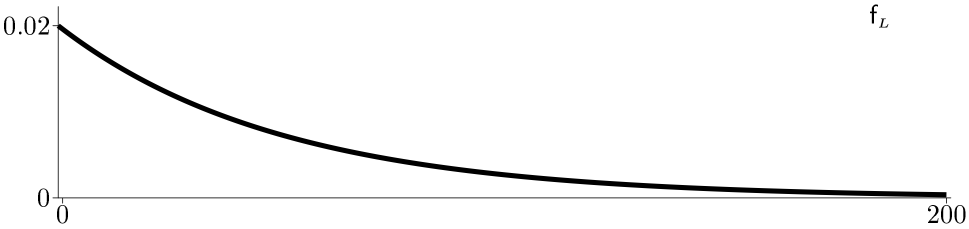


▶ EXAMPLE 3.6 Revisited.

We will model the lifetime L of the bulb as having a distribution whose PDF f_L is given by

$$\mathsf{f}_L(t) = \begin{cases} \frac{1}{50}\, e^{-t/50} & \text{if } t \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where L is measured in days.



This distribution assigns exponentially decreasing probability to longer and longer lifetimes. For instance, we can compute $\mathsf{P}\{\, L > 100 \,\}$ directly from the PDF.

$$\begin{aligned} \mathsf{P}\{\, L > 100 \,\} &= \int_{100}^{\infty} \mathsf{f}_L(t)\, dt \\ &= \int_{100}^{\infty} \frac{1}{50}\, e^{-t/50}\, dt \\ &= 0 - (-e^{-100/50}) \\ &= e^{-2} \approx 0.135. \end{aligned}$$

Similarly, we can compute the expected lifetime

$$\begin{aligned} \mathsf{E}L &= \int_{-\infty}^{\infty} t\, \mathsf{f}_L(t)\, dt \\ &= \int_{0}^{\infty} \frac{t}{50}\, e^{-t/50}\, dt \\ &= 50 \int_{0}^{\infty} x\, e^{-x}\, dx \\ &= 50. \end{aligned}$$

▶ EXAMPLE 3.7 Revisited.

As in Example 3.5, we are choosing a point uniformly from a set. This symmetry assumption implies that f_Z will be some non-zero constant c over that set and 0 elsewhere. We can find this by the rule of total probability. In this case, $1 = \int \mathsf{f}_Z(u)\, du$ equals c times the area of the triangle $\sqrt{3}$. To see this, notice that we can write $\mathsf{f}_Z(u)\, du$ as $\mathsf{f}_Z(u_1, u_2)\, du_1\, du_2$, and the double integral is over all of $\mathbb{R}^2$. However, all densities are zero except those for points in the triangle. The PDF of Z is thus given by

$$\mathsf{f}_Z(u) = \begin{cases} \frac{1}{\sqrt{3}} & \text{if } u \text{ is in the triangle} \\ 0 & \text{otherwise.} \end{cases}$$

The definitions of the PDF above all work as is for vector-valued random variables. Single integrals become multiple integrals and the PDFs take vectors as arguments rather than single numbers.

By direct analogy, we define the *joint* PDF of real-valued random variables $Z_1, \ldots, Z_m$ by

$$\mathsf{f}_{Z_1,\ldots,Z_m}(u_1, \ldots, u_m) = \mathsf{f}_Z, \tag{3.13}$$

where Z is the vector-valued random variable given by $(Z_1, \ldots, Z_m)$. Heuristically, we have

$$\mathsf{f}_{Z_1,\ldots,Z_m}(u_1, \ldots, u_m)\, du_1 \ldots du_m = \\ \mathsf{P}\{\, Z_1 \text{ near } u_1 \text{ and } \cdots \text{ and } Z_m \text{ near } u_m \,\}.$$

As usual, the joint version of the corresponding object for vector-valued random variables only differs in that we mention the components explicitly. Consequently, the joint PDF must satisfy all the properties of PMFs listed above. For instance, it is non-negative, integrates to 1 over all $u_1, \ldots, u_n$, and gives $\mathsf{P}\{\, (Z_1, \ldots, Z_n) \in \mathcal{A} \,\}$ by integrating $\mathsf{f}_{Z_1,\ldots,Z_n}$ over values $(u_1, \ldots, u_n) \in \mathcal{A}$.

▶ EXAMPLE 3.7 Once More.

The joint PDF of the coordinates X and Y is given by

$$\mathsf{f}_{X,Y}(u, v) = \mathsf{f}_Z(u, v).$$

We can use this to get the PDF of Y alone:

$$\begin{aligned} \mathsf{f}_Y(v)\, dv &= \mathsf{P}\{\, Y \text{ near } v \,\} \\ &= \int_u \mathsf{P}\{\, X \text{ near } u \text{ and } Y \text{ near } v \,\} \\ &= \int_u \mathsf{f}_{X,Y}(u, v)\, du\, dv. \end{aligned}$$

Hence, canceling the dvs on both sides gives

$$\mathsf{f}_Y(v) = \int \mathsf{f}_{X,Y}(u, v)\, du.$$

Carrying out the integration over the triangle at vertical coordinate v yields

$$\mathsf{f}_Y(v) = \begin{cases} \frac{2}{\sqrt{3}}(1 - v/\sqrt{3}) & \text{if } 0 < v < \sqrt{3} \\ 0 & \text{otherwise.} \end{cases}$$

Here are a few things to keep in mind about the PDF:

- The PDF $\mathsf{f}_X(u)$ is a probability density; we get a probability only when we scale this by a measure of volume, as in $\mathsf{f}_X(u)\,du$. The latter is an infinitesimal probability, which we also write as $\mathsf{P}\{\, X \text{ near } u \,\}$.
- The PDF must be non-negative for every value of its argument, and it must integrate to 1 over its domain. However, because it returns probability densities rather than probabilities, it is possible for the PDF to take values greater than 1.
- In some cases, like in the examples above, it is possible to derive the PDF directly. However, it is often necessary and usually convenient to derive the PDF by first computing a related function, such as the Cumulative Distribution Function (CDF) or Survival Function, and deriving the PDF from this. We describe how to do in the sections to follow.

Events of the form $\{\, X \text{ near } u \,\}$ follow the same logical rules as other events, with one technical but important exception. If X is a continuous (possibly vector-valued) random variable and h is an invertible function taking values in the range of X, then

$$\mathsf{P}\{\, X \text{ near } h(u) \,\} = \mathsf{f}_X(h(u))\, |h'(u)|\, du,$$

where $|h'(u)|$ is the Jacobian determinant of h and du is the volume element. Since the near operator depends on values in an infinitesimal neighborhood around the selected value, the extra factor of the Jacobian determinant can be viewed as the additional stretching of the neighborhood caused by the function h.

This is useful for calculating the distributions of random variables derived as a function of other continuous random variables. Indeed, he above is just the regular change of variables formula for smooth functions. For example, if $Z = h^{-1}(X)$,

$$\mathsf{P}\{\, Z \text{ near } u \,\} = \mathsf{P}\{\, X \text{ near } h(u) \,\},$$

so knowing the function h and the distribution of X is enough to find the distribution of Z.

5. Aside: Discrete vs Continuous Distributions

The PMF and PDF are distinct types of objects; one gives probabilities, one gives densities of probabilities. On first look, then, it seems as if we work with these objects in entirely distinct ways as well. Fortunately, this is not the case: only the notation we use for discrete and continuous random variables is different, the essence of the operations is identical.

The unifying thread is the distribution itself. Both the PMF and the PDF are derived from the corresponding distributions and contain the same basic information. Indeed, almost every quantity we compute in this course will be of the form $\mathsf{D}_X g$ for some function g.

For the remainder of this section, suppose that S is a discrete random variable and that T is a continuous random variable. For convenience, let $\mathcal{R}_S = \text{range}(S)$ and $\mathcal{R}_T = \text{range}(T)$.

The similarity between D_S and D_T

For any g defined on $\mathcal{R}_S \cup \mathcal{R}_T$, we have

$$\mathsf{E}g(S) = \sum_{u \in \mathcal{R}_S} g(u)\, \mathsf{p}_S(u)$$

$$\mathsf{E}g(T) = \int_{u \in \mathcal{R}_T} g(u)\, \mathsf{f}_T(u)\, du.$$

Thus, despite the differences between PMFs and PDFs, the calculations we are doing have basically the same form.

We can write both the expressions in another, equally revealing way,

$$\mathsf{E}g(S) = \sum_{u \in \mathcal{R}_S} g(u)\, \mathsf{P}\{\, S \ \mathsf{near}\ u \,\}$$

$$\mathsf{E}g(T) = \int_{u \in \mathcal{R}_T} g(u)\, \mathsf{P}\{\, T \ \mathsf{near}\ u \,\}.$$

This uses the fact that in the discrete case $\{\, S \ \mathsf{near}\ u \,\} = \{\, S = u \,\}$. As long as we remember the fact that events with the near operator give infinitesimal probabilities in the continuous case (that's why we need the integral), we can work with S and T in essentially the same manner. In particular, events of the form $\{\, T \ \mathsf{near}\ u \,\}$, though they are mythical, obey the same logical rules as events like $\{\, S = u \,\}$ do.

It follows that to transform an expression that works in the discrete case (i.e., with S) to an expression that works in the continuous case (i.e., with T), we need only carry out three easy steps.

How to transform from a discrete to a continuous calculation.

1. Replace $\sum$ by $\int$.
2. Replace $\mathsf{p}_S(u)$ by $\mathsf{f}_T(u)\,du$.
3. Replace the $=$ operator with the near operator in events for conditions relating to T.

Thought Question Suppose S is an integer-valued random variable. We can express $\mathsf{P}\{\, S > 2 \,\}$ as follows:

$$\mathsf{P}\{\, S > 2 \,\} = \sum_{k>2} \mathsf{p}_S(k).$$

Using the transformation rules above, write the corresponding expression for $\mathsf{P}\{\, T > 2 \,\}$.

6. The Cumulative Distribution Function and The Survival Function

The PMF and the PDF are each defined for specific types of distributions, discrete and continuous respectively. We can define two alternative representations for the distribution that are each well-defined in both the discrete and continuous cases. These are the Cumulative Distribution Function (CDF) and the Survival Function (SF).

If X is a real-valued random variable, then Cumulative Distribution Function or CDF of X is a function $\mathsf{F}_X\colon \mathbb{R} \to [0,1]$ defined by

$$\mathsf{F}_X(u) = \mathsf{P}\{\, X \le u \,\}.$$

The Survival Function or SF of X is a function $S_X\colon \mathbb{R} \to [0,1]$ defined by

$$S_X(u) = \mathsf{P}\{\, X > u \,\} = 1 - F_X(u).$$

Thus, the two functions are exactly complementary in that $\mathsf{F}_X(u) + S_X(u) = 1$ for every u. It is nonetheless useful to give each a name as it is sometimes more convenient to use one than the other.

We begin with the CDF. The word "cumulative" in the name comes from the special properties of the function. As the value of its argument increases, the CDF cannot decrease. (If $u_2 > u_1$, $\{ X \le u_1 \} \subset \{ X \le u_2 \}$.) Thus, as u goes from smaller to larger values ($-\infty$ to ∞), the probability returned by F_X aCUMULATEs.

Properties of the CDF

- For all u, $0 \le \mathsf{F}_X(u) \le 1$.
- If $u < v$, then $\mathsf{F}_X(u) \le \mathsf{F}_X(v)$.
- $\mathsf{F}_X(-\infty) = 0$ and $\mathsf{F}_X(\infty) = 1$.

One of the most important features of the CDF is that it can be used to derive the PMF and PDF. Suppose first that X is discrete. If u is a value for which $\mathsf{p}_X(u) > 0$,

$$\mathsf{P}\{ X \le u \} - \mathsf{P}\{ X < u \} = \mathsf{p}_X(u).$$

In other words, F_X has a *jump* at the value u, and the value of that jump is $\mathsf{p}_X(u)$.[29] In this case, $\mathsf{P}\{ X < u \}$ is the limit of F_X as one approaches u from below; that is,

$$\mathsf{P}\{ X < u \} = \lim_{s \uparrow u} \mathsf{F}_X(s) \equiv \mathsf{F}_X(s-),$$

where the last expression is a definition. Hence,

$$\mathsf{p}_X(u) = \mathsf{F}_X(u) - \mathsf{F}_X(u-) = \lim_{\Delta \downarrow 0} \left[\mathsf{F}_X(u) - \mathsf{F}_X(u - \Delta) \right],$$

and we can find the PMF from the CDF. We can go the other way as well:

$$\mathsf{F}_X(u) = \sum_{s \le u} \mathsf{p}_X(s),$$

accumulating all of the probabilities no bigger than u.[30]

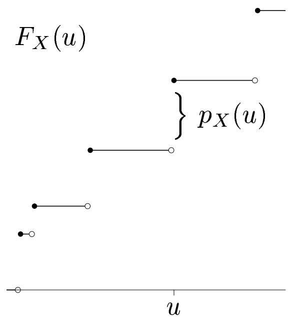
If, on the other hand, X is continuous, then F_X has no jumps; it is continuous over its entire domain.[31] We have that

$$\mathsf{P}\{ u \le X < u + \Delta \} = \mathsf{F}_X(u + \Delta) - \mathsf{F}_X(u).$$

Hence, by the definition of the PDF,

$$\mathsf{f}_X(u) = \lim_{\Delta \to 0} \frac{\mathsf{P}\{ u \le X < u + \Delta \}}{\Delta}$$

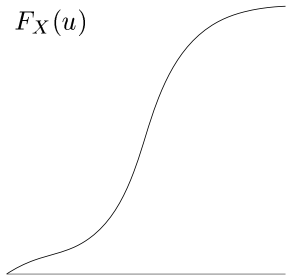
29 Discrete case



The empty dots correspond to where the jumps occur; the filled dots show the actual values of F_X at those points.

30 This sum is well defined because at most a countable number of values are non-zero.

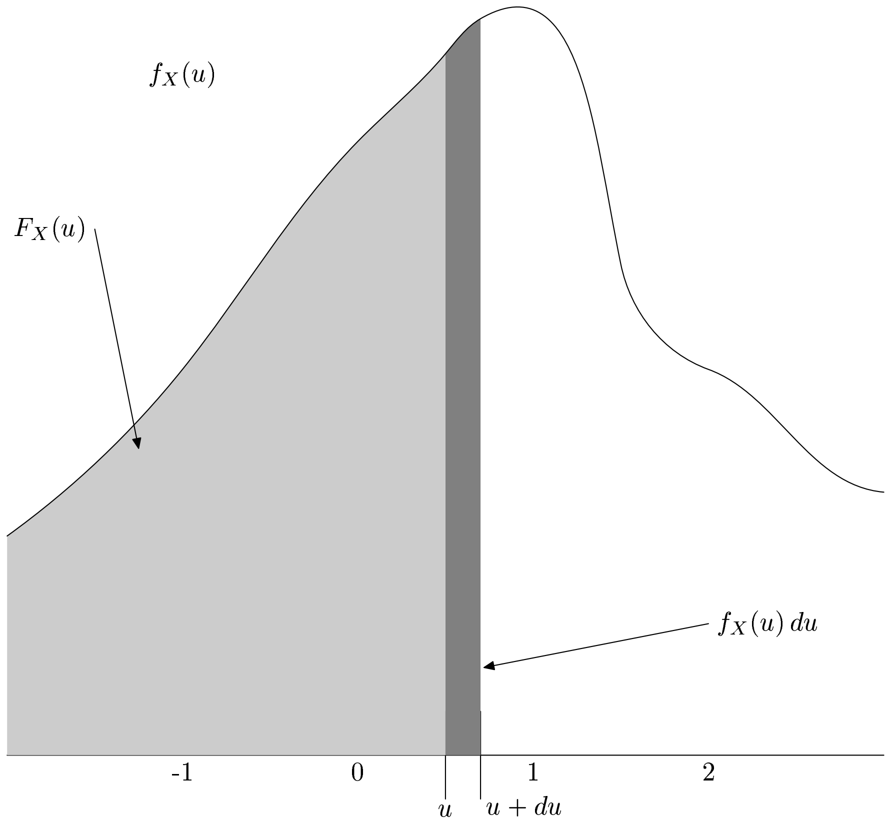
31 Continuous case



$$= \lim_{\Delta \to 0} \frac{\mathsf{F}_X(u + \Delta) - \mathsf{F}_X(u)}{\Delta}$$
$$= \frac{d}{du}\mathsf{F}_X(u).$$

We can go the other direction here as well, again accumulating the probability:

$$\mathsf{F}_X(u) = \int_{-\infty}^{u} \mathsf{f}_X(s)\, ds.$$



The Survival Function (SF) gets its name from a specific scenario in which it arises frequently. Suppose that X represents the lifetime (time until failure) of some organism or system. Then $\mathsf{P}\{ X > u \}$ is the probability that the organism or system survives past the time u.

The SF has properties that are complementary to the CDF.

Properties of the SF

- For all u, $0 \le \mathsf{S}_X(u) \le 1$.
- If $u < v$, then $\mathsf{S}_X(u) > \mathsf{S}_X(v)$.
- $\mathsf{S}_X(-\infty) = 1$ and $\mathsf{S}_X(\infty) = 0$.

In particular, the SF cannot increase as its argument increases. As for the CDF, the SF has jumps for discrete distributions and is continuous for continuous distributions. Moreover, one can find the PMF or PDF from the SF, and vice versa.

$$\mathsf{S}_X(u) = \sum_{t>u} \mathsf{p}_X(t) \qquad \mathsf{p}_X(u) = \mathsf{S}_X(u-) - \mathsf{S}_X(u)$$
$$\mathsf{S}_X(u) = \int_u^\infty \mathsf{f}_X(t)\,dt \qquad \mathsf{f}_X(u) = -\frac{d}{du}\mathsf{S}_X(u)$$

Notice also that $\mathsf{S}_X = 1 - \mathsf{F}_X$.

For vector valued random variables, the CDF and SF generalize appropriately. If $Z = (X_1, \dots, X_n)$ is an $\mathbb{R}^n$-valued random variable and if $u = (u_1, \dots, u_n) \in \mathbb{R}^n$, then we define

$$\mathsf{F}_Z(u) = \mathsf{P}\{\, X_1 \le u_1 \text{ and } \cdots \text{ and } X_n \le u_n \,\}$$
$$\mathsf{S}_Z(u) = \mathsf{P}\{\, X_1 > u_1 \text{ or } \cdots \text{ or } X_n > u_n \,\}.$$

The first is the probability that the vector Z is in the lower-left orthant below u; the second is the complementary probability that Z lies outside the lower-left orthant. This is a bit less natural than in one dimension, but it is sometimes easier to compute derive the PMF and PDF from the CDF or SF. For instance,

$$\mathsf{f}_Z(u) = \frac{d^n}{du_1 \cdots du_n}\mathsf{F}_Z(u) = -\frac{d^n}{du_1 \cdots du_n}\mathsf{S}_Z(u).$$

▶ EXAMPLE 3.8. A Discrete Uniform Distribution

Suppose that D is a random integer between 1 and m, chosen so that all m possibilities are equally likely. We know that

$$\mathsf{p}_D(k) = \begin{cases} \frac{1}{m} & \text{if } k \in \{1, \dots, m\} \\ 0 & \text{otherwise.} \end{cases}$$

We have that $\mathsf{P}\{\, D \le u \,\} = 0$ for any $u < 1$ and $\mathsf{P}\{\, D \le u \,\} = 1$ for any $u \ge m$, and $\mathsf{P}\{\, D \le u\}$ for $1 \le u \le m$ is just the number of the points $1, \dots, m$ that are $\le u$. Hence,

$$\mathsf{F}_D(u) = \begin{cases} 0 & \text{if } u < 1 \\ \frac{k}{m} & \text{if } k \le u < k+1,\ k \in \{1, \dots, m-1\} \\ 1 & \text{if } u \ge m. \end{cases}$$

The boundary points in this definition matter. For example, if $m \geq 4$, $\mathsf{F}_D(4) = 4/m$ but $\mathsf{F}_D(3.99999) = 3/m$. Draw a picture of this function in the case $m = 5$. We could also write this CDF more simply as

$$\mathsf{F}_D(u) = \begin{cases} 0 & \text{if } u < 0 \\ \frac{\lfloor u \rfloor}{m} & \text{if } 0 \leq u \leq m \\ 1 & \text{if } u > m. \end{cases}$$

The SF is complementary:

$$\mathsf{S}_D(u) = \begin{cases} 1 & \text{if } u < 0 \\ \frac{m - \lfloor u \rfloor}{m} & \text{if } 0 \leq u \leq m \\ 0 & \text{if } u > m. \end{cases}$$

▶ EXAMPLE 3.9. A Continuous Uniform Distribution

Let U be the random number from Example 3.5, chosen uniformly from the interval $[0, 1]$. then

$$\mathsf{f}_U(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

We have $\mathsf{P}\{\, U \leq 0 \,\} = 0$ and $\mathsf{P}\{\, U \leq 1 \,\} = 1$ and for $0 \leq x \leq 1$,

$$\begin{aligned} \mathsf{F}_U(x) &= \mathsf{P}\{\, U \leq x \,\} \\ &= \int_{-\infty}^{x} \mathsf{f}_U(t)\, dt \\ &= \int_{0}^{x} dt \\ &= x. \end{aligned}$$

Hence,

$$\mathsf{F}_U(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1, \end{cases}$$

and

$$\mathsf{S}_U(x) = \begin{cases} 1 & \text{if } x < 0, \\ 1 - x & \text{if } 0 \leq x \leq 1, \\ 0 & \text{if } x > 1. \end{cases}$$

► EXAMPLE 3.10. An Exponential Lifetime Distribution

Let L be the random variable in Example 3.6; this represents the lifetime of a light bulb. We have that

$$\mathsf{f}_L(t) = \begin{cases} \frac{1}{50}\, e^{-t/50} & \text{if } t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Notice that $\mathsf{P}\{\, L > 0 \,\} = 1$. For $t \geq 0$, we have

$$\begin{aligned} \mathsf{S}_L(t) &= \mathsf{P}\{\, L > t \,\} \\ &= \int_t^\infty \mathsf{f}_L(s)\, ds \\ &= \int_t^\infty \frac{1}{50}\, e^{-s/50}\, ds \\ &= e^{-t/50}. \end{aligned}$$

Hence,

$$\mathsf{S}_L(t) = \begin{cases} 1 & \text{if } t < 0, \\ e^{-t/50} & \text{if } t \geq 0. \end{cases}$$

The CDF satisfies $\mathsf{F}_L(t) = 1 - \mathsf{S}_L(t)$:

$$\mathsf{S}_L(t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 - e^{-t/50} & \text{if } t \geq 0. \end{cases}$$

7. Generating Functions for When You Need More Power

All of the representations we have seen so far for the distribution of a random variable have given us to direct access to the probabilities at each location. A generating function is a package for the information in the distribution with three important properties:

1. It uniquely determines the distribution.
2. It offers easy access to features of the distribution that are difficult to compute by other means.
3. It allows a variety of manipulations that make it possible to find generating functions in complicated situations.

We will make use of three types of generating functions which are defined for successively broader classes of distributions.

7.1. The Probability Generating Function for Integer-Valued Random Variables

One of the most powerful tools for working with sequences of numbers is to manipulate a power series that "generates" that sequence. Consider the function that results from a simple geometric series:

$$G(z) = \frac{1}{1-z} = 1 + z + z^2 + z^3 + \cdots .$$

We call this the "generating function" for the sequence $(1, 1, 1, \ldots)$ because if we know $G(z)$, we can get each term of the sequence by a specific operation on the function.

To get a hint of how generating functions make it easy to manipulate sequences, consider the functions

$$\begin{aligned} zG(z) &= z + z^2 + z^3 + z^4 + \cdots \\ \frac{G(z)}{1-z} &= 1 + 2z^2 + 3z^3 + 4z^4 + \cdots \\ G(z^2) &= 1 + z^2 + z^4 + z^6 + \cdots . \end{aligned}$$

These are, respectively, the generating functions of the shifted sequence $(0, 1, 1, \ldots)$, the cumulative sum sequence $(1, 2, 3, \ldots)$, and the spaced sequence $(1, 0, 1, 0, \ldots)$. Many other types of manipulations are available. Suppose that $F(z) = \sum_k a_k z^k$ and $G(z) = \sum_k b_k z^k$. Then,

$$\begin{aligned} z^j G(z) &= \sum_k b_{k-j} z^k \\ zG'(z) &= \sum_k k b_k z^k \\ G(cz) &= \sum_k b_k c^k z^k \\ F(z)G(z) &= \sum_k \left(\sum_j a_j b_{k-j} \right) z^k \\ uF(z) + vG(z) &= \sum_k (u a_k + v b_k) z^k, \end{aligned}$$

where all the sums are over the integers. These will be the most useful manipulations for our work; see the sources below for others.

In general, if we have a sequence of numbers $(a_k)_{k\in\mathbb{Z}}$, the generating function of this sequence is the power series

$$G(z) = \sum_{k\in\mathbb{Z}} a_k \, z^k .$$

For most purposes, we treat this as a *formal* power series. That is, we do not worry terribly much about the convergence of the series. The only time we have to worry about such details is when we evaluate these functions, which usually comes at the end of our calculation.

Generating functions make many calculations easier carry out. One important example is their usefulness in solving recurrence relations. Consider the recurrence relation that defines the Fibonacci numbers F_n, which is given by

$$F_n = \begin{cases} 0 & \text{if } n \leq 0 \\ 1 & \text{if } n = 1 \\ F_{n-1} + F_{n-2} & \text{if } n \geq 2. \end{cases}$$

Let's put these terms as coefficients is a generating power series

$$F(z) = \sum_{k \in \mathbb{Z}} F_k\, z^k = \sum_{k=0}^{\infty} F_k\, z^k.$$

Notice that

$$F(z) - zF(z) - z^2 F(z) = \sum_{k \in \mathbb{Z}} (F_k - F_{k-1} - F_{k-2}) z^k = z$$

because all of the coefficients in the sum are zero except when $k = 1$. Thus,

$$F(z) = \frac{z}{1 - z - z^2}$$

is the generating function for the Fibonacci numbers. We can use this to derive a closed form expression for F_n. See *Concrete Mathematics* by Graham, Knuth, and Patashnik for a superb introduction to generating functions and their uses in solving general recurrences.

For now, we turn to the uses of generating functions in probability. If X is an integer-valued random variable with PMF p_X, we consider the function that generates the sequence of numbers $\mathsf{p}_X(k)$ for $k \in \mathbb{Z}$. That is,

$$\mathsf{G}_X(z) = \sum_{k \in \mathbb{Z}} \mathsf{p}_X(k)\, z^k.$$

By construction, this contains all the information in the PMF and thus in the distribution itself. Notice that by the definition of p_X, $\mathsf{G}_X(z) = \mathsf{E} z^X$.

The Probability Generating Function

If X is an integer-valued random variable, the *probability generating function* (PGF) of X, G_X, is defined by

$$\mathsf{G}_X(z) = \mathsf{E}z^X.$$

The most important property of the PGF is that it uniquely determines the distribution of a random variable. That is: If two random variables have the same PGF, they necessarily have the same distribution. This means in practice that to find the distribution of a random variable it is sufficient to find the PGF.

Two other important properties of the PGF are that we can find expected values by taking derivatives and that we can find the distribution of sums of independent random variables by multiplying their generating functions. The latter follows from the fact that the coefficient of z^k in $\mathsf{G}_X(z)\mathsf{G}_Y(z)$ is $\sum_j \mathsf{p}_X(j)\mathsf{p}_Y(k-j)$ which sums the probabilities of combinations of X and Y that sum to k.

Properties of the PGF

- Integer-valued random variables X and Y have the same distribution (i.e., $\mathsf{D}_X = \mathsf{D}_Y$) if and only if $\mathsf{G}_X = \mathsf{G}_Y$
- $\mathsf{E}X = \mathsf{G}'_X(1)$.
- $\mathsf{E}X(X-1)\cdots(X-k+1) = \mathsf{G}_X^{(k)}(1)$, for $k \in \mathbb{Z}_+$.
- If $X_1, \ldots, X_n$ are independent,

$$\mathsf{G}_{X_1+\cdots+X_n} = \prod_{i=1}^{n} \mathsf{G}_{X_i}.$$

- $\mathsf{p}_X(k)$ is the coefficient of z^k.

▶ EXAMPLE 3.11. An Integer Constant

The simplest integer-valued random variable is one that takes the same value for every elementary outcome. If $X \equiv k$ for some integer k, then $z^X = z^k$, and $\mathsf{G}_X(z) = z^k$.

▶ EXAMPLE 3.12. Indicators

If I is an indicator random variable with $\mathsf{p}_I(1) = p$ for some $0 \le p \le 1$, then, the random variable z^I takes the values 1 and z with respective probabilities $1-p$ and p. Hence,

$$\mathsf{G}_I(z) = (1-p) + pz.$$

By the above, $\mathsf{E}I = \mathsf{G}_I'(1) = p$. Since the PGF of I determines its distribution, we see that knowing $p = \mathsf{E}I$ does as well. Hence, because indicator random variables take only two possible values, the distribution of an indicator is determined by its expected value.

▶ EXAMPLE 3.13. The "Binomial" Distribution

Let $H_1, \ldots, H_n$ be a sequence of n independent indicator random variables with the same expected value $\mathsf{E}H_i = p$ for $i = 1, \ldots, n$. For example, H_i might be the indicator of heads on the ith coin flip out of n, where the coin has probability p of coming up heads.

Let $S = H_1 + \cdots + H_n$ be the number of "1s" in the sequence (the number of heads if the H_is represent coin flips). Then, by the above properties

$$\begin{aligned} \mathsf{G}_{H_i}(z) &= (1-p) + pz \quad \text{for all } i \\ \mathsf{G}_S(z) &= \prod_{i=1}^{n} \mathsf{G}_{H_i}(z) \\ &= (1-p+pz)^n. \end{aligned}$$

That $\mathsf{G}_S'(1) = np(1-p+p1)^{n-1} = np$ confirms the result of the additivity rule for expected values. Notice that by the Binomial Theorem

$$(1-p+pz) = \sum_{k=0}^{n} \binom{n}{k} (pz)^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} z^k.$$

Hence,

$$\mathsf{p}_S(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k \in \{0, \ldots, n\}, \\ 0 & \text{otherwise.} \end{cases}$$

In Chapter 6, we will say that any random variable with this PMF (or PGF for that matter) has a Binomial$\langle n, p\rangle$ distribution.

32 See Appendix B.

▶ EXAMPLE 3.14. Attempts until Success

Suppose now that $H_1, H_2, \ldots,$ is a sequence of independent and identically distributed indicators. Let N be the first i such that $H_i = 1$. The only way for N to equal an positive integer k is if there are $k-1$ 0's before the last 1. Since all of the H_is are independent, we have – as we saw by examples in class – that

$$\mathsf{p}_N(k) = \begin{cases} (1-p)^{k-1}p & \text{if } k \in \mathbb{Z}_+, \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} \mathsf{G}_N(z) &= \sum_{k=1}^{\infty} (1-p)^{k-1} p z^k \\ &= pz \sum_{k=1}^{\infty} (1-p)^{k-1} z^{k-1} \\ &= pz \sum_{k=0}^{\infty} ((1-p)z)^k \\ &= \frac{pz}{1-(1-p)z}, \end{aligned}$$

using the value of an infinite geometric sum.[32] It follows that

$$\mathsf{E}N = \mathsf{G}'_N(1) = \left.\frac{p}{(1-(1-p)z)^2}\right|_{z=1} = \frac{1}{p}.$$

In Chapter 6, we will say that any random variable with this PMF or PGF has a Geometric$\langle p \rangle$ distribution.

▶ EXAMPLE 3.15. Double Heads

The following is a really nifty example of some of the challenges that can arise in even simple situations. It also introduces a new technology.

Describe the Experiment. We flip a coin until two consecutive heads appear.

Specify your assumptions.

We assume that the coin flips are all independent. We assume that the coin comes up heads with probability $0 < p < 1$, though we are most interested in the case of $p = 1/2$.

Define relevant random variables.

Let H_i be the indicator of heads on the ith roll, for $i \in \mathbb{Z}_+$.

Let N be the number of flips (inclusive) until heads first appears on two consecutive flips.

State what you know.

We know that the H_is are independent and have $\mathsf{E}H_i = p$.

We know that $N \geq 2$.

State what you want to find.

We want to find the PMF of N, p_N.

Find it.

This is an interesting problem in many ways. The naive approach would be make a direct analogy to the modem connection problem. We might reason as follows: "Consider pairs of flips. Heads followed by heads occurs with probability p^2 and not with probability $1-p^2$. Using the modem connection problem solution, we have that the probability that we get two consecutive heads on the first time in $2n$ flips is $(1-p^2)^{n-1}p^2$." Unfortunately, this reasoning is wrong. Why?

A more careful analysis of the sequences that can occur tells us that the observed sequences consist of a series of `T`s and `HT`s followed by `HH`. (In regular expression terms, this is `(T|HT)*HH`.) We can use the modem connection problem here in one way. Since `T` occurs with probability $1-p$, `HT` with probability $p(1-p)$, and `HH` with probability p^2 (the three terms sum to 1), the probability that we see `HH` after k `T`s and `HT`s is $(1-p^2)^k p^2$. This is right, but unfortunately, it answers the wrong question. To determine the number of *flips*, we have to carefully count all the ways that `T`s and `HT`s can occur in this sequence. Messy!

Generating functions give us a fast and generalizable approach. Consider all the sequences that can possibly be observed in this experiment. We can express this as a formal sum that simply accumulates the possibilities

$$\mathsf{G}_N = \mathtt{HH} + \mathtt{THH} + \mathtt{HTHH} + \mathtt{TTHH} + \mathtt{HTTHH} + \cdots .$$

Now, replace every `HT` by $p(1-p)z^2$, every `HH` by p^2z^2 and every remaining `T` by $(1-p)z$. Then, when we collect terms with common powers of z, we find that the coefficient of z^k is just $\mathsf{p}_N(k)$.

Define the *probability generating function* G_N by

$$\mathsf{G}_N(z) = \sum_{k\in\mathbb{Z}} \mathsf{p}_N(k)\, z^k.$$

This is a packaging of the PMF of N as a formal power series that we can manipulate somewhat differently.

Notice that

$$\mathsf{G}_N(1) = \sum_{k=2}^{\infty} \mathsf{p}_N(k) = 1,$$

and

$$\mathsf{G}'_N(1) = \sum_{k=2}^{\infty} k\mathsf{p}_N(k) = \mathsf{E}N.$$

So, if we can compute G_N, we can easily find many quantities of interest. (In fact, we can recover p_N. Do you see how?)

Going back to the formal sum above we have that every final sequence starts with `T` every

$$\begin{aligned}\mathsf{G}_N &= \mathtt{HH} + \mathtt{THH} + \mathtt{HTHH} + \mathtt{TTHH} + \mathtt{HTTHH} + \cdots \\ &= \mathtt{HH} + \mathtt{T}(\mathtt{HH} + \mathtt{THH} + \cdots) + \mathtt{HT}(\mathtt{HH} + \mathtt{THH} + \cdots) \\ &= \mathtt{HH} + \mathtt{T}\mathsf{G}_N + \mathtt{HT}\mathsf{G}_N.\end{aligned}$$

Or to put this more formally, for every z we have

$$\mathsf{G}_N(z) = p^2 z^2 + (1-p)z\mathsf{G}_N(z) + p(1-p)z^2\mathsf{G}_N(z).$$

Solving for $\mathsf{G}_N(z)$ we have

$$\mathsf{G}_N(z) = \frac{p^2z^2}{1-(1-p)z-p(1-p)z^2}.$$

By taking $\mathsf{G}'_N(1)$, we can easily find that that $\mathsf{E}N = (1+p)/p^2$. Nifty.

In the case $p = 1/2$, we can write $\mathsf{G}_N(z) = (z/2)F(z/2)$, where

$$F(z) = \frac{z}{1-z-z^2}.$$

This $F(z)$ is also a generating function, though of an arbitrary sequence not a probability distribution. You might recognize it from above as the generating function for the Fibonacci numbers. To derive this in the reverse direction, write

$$F(z) = \sum_{k\in\mathbb{Z}} F_k z^k.$$

Notice that $F(z)1 - zF(z) - z^2F(z) = z$. Hence,

$$\sum_{k\in\mathbb{Z}} (F_k - F_{k-1} - F_{k-2})z^k = z,$$

so

$$F_k = 1_{(k=1)} + F_{k-1} + F_{k-2}.$$

Since $z^jF(0) = 0$ for all integers $j \geq 0$, we have

$$F_k = \begin{cases} 0 & \text{if } k \leq 0 \\ 1 & \text{if } k = 1 \\ F_{k-1} + F_{k-2} & \text{otherwise.}\end{cases}$$

which are the Fibonacci numbers $1, 1, 2, 3, 5, 8, 13, \ldots$.

It follows that

$$\mathsf{G}_N(z) = \sum_{k=2}^{\infty} \frac{F_{k-1}}{2^k} z^k,$$

so

$$\mathsf{p}_N(k) = \begin{cases} \frac{F_{k-1}}{2^k} & \text{if } k \in \mathbb{Z}_+ \\ 0 & \text{otherwise.}\end{cases}$$

7.2. The Moment Generating Function

For any random variable X, the number $\mathsf{E}X^k$ is called the kth *moment* of X. If $X \geq 0$, as k gets larger, the kth moment gives greater weight to large values of X than $\mathsf{E}X$. Thus, loosely speaking, the kth moment is a measure of how much probability the distribution assigns to extreme values.[33]

33 We will see more about moments in Chapter 5.

The moment generating function is roughly the power series that "generates" the moments of the distribution:

$$1 + \mathsf{E}Xz + \mathsf{E}X^2\frac{z^2}{2!} + \mathsf{E}X^3\frac{z^3}{3!} + \cdots.$$

The factorial terms here are included to make the series converge for reasonable values of z. The precise definition of the moment generating function is in this spirit but is a bit cleaner.

The Moment Generating Function

The *moment generating function* (MGF) of a random variable X, M_X, is defined by

$$\mathsf{M}_X(t) = \mathsf{E}e^{tX}.$$

The MGF is also known Laplace Transform of the distribution. If X is integer valued, then $\mathsf{M}_X(t) = \mathsf{G}_X(e^t)$.

A technical point: Because the exponential function grows so quickly, the MGF might not be finite for every value of its arguments. For most purposes, we only need that it be well defined for values in a neighborhood of 0. If the MGF is not properly defined around zero, then we cannot use it.

The properties of the MGF exactly mimic those of the PGF. The MGF determines the distribution uniquely; we get moments of the distribution by taking derivatives; and the MGF of a sum of independent random variables is the product of the MGFs.

Properties of the MGF

- Random variables X and Y have the same distribution (i.e., $\mathsf{D}_X = \mathsf{D}_Y$) if and only if $\mathsf{M}_X = \mathsf{M}_Y$.
- For $k \in \mathbb{Z}_+$,

$$\mathsf{E}X^k = \frac{d^k}{dt^k}\mathsf{M}_X(t)\Big|_{t=0}.$$

- If $X_1, \ldots, X_n$ are independent random variables, then

$$\mathsf{M}_{X_1+\cdots+X_n} = \prod_{i=1}^n \mathsf{M}_{X_i}.$$

► EXAMPLE 3.16. Lifetime Distribution Again

Let L be the random variable in Example 3.6 and let $\lambda = 1/50$. Then,

$$\begin{aligned}
\mathsf{M}_L(u) &= \int_{-\infty}^{\infty} e^{ux}\, \mathsf{f}_L(x)\, dx \\
&= \int_0^{\infty} e^{ux}\, \lambda\, e^{-\lambda x}\, dx \\
&= \frac{\lambda}{\lambda - u}\int_0^{\infty} (\lambda - u)\, e^{-(\lambda-u)x}\, dx \\
&= \begin{cases} \frac{\lambda}{\lambda-u} & \text{if } u < \lambda \\ \infty & \text{otherwise.} \end{cases}
\end{aligned}$$

So, M_L is defined for all $u < \lambda$ and thus in an interval containing 0. For such u, we can write

$$\begin{aligned}
\mathsf{M}_L(u) &= \frac{\lambda}{\lambda - u} \\
&= \frac{1}{1 - \frac{u}{\lambda}} \\
&= \sum_{k=0}^{\infty} \frac{u^k}{\lambda^k} \\
&= \sum_{k=0}^{\infty} k!\, \lambda^{-k}\, \frac{u^k}{k!}.
\end{aligned}$$

It follows that $\mathsf{E}L^k = k!\, \lambda^{-k}$. Specifically, $\mathsf{E}X = 1/\lambda$.

7.3. The Characteristic Generating Function[34]

34 This section can be skipped on first reading.

The primary difficulty with using the MGF is that it is not always well-defined; it can return ∞ for certain values of its argument. By replacing the exponential e^{tX} with the *complex* exponential e^{itX} where $i = \sqrt{-1}$, we get a function that is well defined for any random variable whatsoever. This is the Characteristic Generating Function.

The Characteristic Generating Function

The *characteristic generating function* (CGF) of a random variable X, C_X, is defined by

$$\mathsf{C}_X(t) = \mathsf{E}e^{itX}.$$

The CGF is also known Fourier Transform of the distribution. The properties of the CGF mimic those of the MGF; we must only carry around the extra factor of i.

Properties of the CGF

- Random variables X and Y have the same distribution (i.e., $\mathsf{D}_X = \mathsf{D}_Y$) if and only if $\mathsf{C}_X = \mathsf{C}_Y$.
- For $k \in \mathbb{Z}_+$,

$$\mathsf{E}X^k = i^{-k} \frac{d^k}{dt^k} \mathsf{C}_X(t) \Big|_{t=0} .$$

- If $X_1, \ldots, X_n$ are independent random variables, then

$$\mathsf{C}_{X_1+\cdots+X_n} = \prod_{i=1}^{n} \mathsf{C}_{X_i}.$$

Calculation of the CGF looks very much like that of the MGF except that we carry around factors if i.

▶ EXAMPLE 3.17. Lifetime Distribution Yet Again

Let L be the random variable in Example 3.6 and let $\lambda = 1/50$. Then,

$$
\begin{aligned}
\mathsf{C}_L(u) &= \int_{-\infty}^{\infty} e^{iux}\, \mathsf{f}_L(x)\, dx \\
&= \int_0^{\infty} e^{iux}\, \lambda\, e^{-\lambda x}\, dx \\
&= \int_0^{\infty} \lambda\, e^{-(\lambda - iu)x}\, dx \\
&= \frac{\lambda}{\lambda - iu} \int_0^{\infty} (\lambda - iu)\, e^{-(\lambda - iu)x}\, dx \\
&= \frac{1}{1 - i\frac{u}{\lambda}}.
\end{aligned}
$$

Notice that this is well-defined for all u.

35 If Y and X are two random variables related by $Y = g(X)$ for some function g, then we can compute $\mathsf{E}Y$ in two ways: $\mathsf{E}Y = \mathsf{D}_Y \mathrm{Id}$, where $\mathrm{Id}(u) = u$ is the identity function, and $\mathsf{E}Y = \mathsf{D}_X\, g$. Both of these are equivalent. To compute every prediction about Y, we would need to find its distribution, but to find expected values, we can use the distribution of X. Indeed, we can compute the distribution of Y in terms of the distribution of X: $\mathsf{D}_Y\, h = \mathsf{D}_X h \circ g$. If g turns out to be an invertible function, we can find D_X given only D_Y by computing $\mathsf{D}_Y h \circ g^{-1}$.

36 This logic works just as well in the vector-valued case.

8. Expected Values Revisited

The expected value of a random variable is property of the distribution of that random variable. It does not depend on what we will observe in the experiment, only on the possible values and their likelihood. We can, fortunately, compute it before the experiment is ever run.

Whenever we compute expected values directly from the distribution, the calculation – though it might differ in details – always has the same structure. Namely, we compute weighted averages of the possible values with the corresponding probabilities, possibly infinitesimal probabilities, as weights.

> **Calculating expected values in general.**
> The expected value of a random variable can be calculated as the weighted average of the values that random variable can take with the probabilities at (or near) each value as weights.

This heuristic provides a useful check on our mathematics, but it also gives another useful interpretation of the expected value. Thinking of probability as mass, a distribution is a set of instructions for how to spread that mass through space. Suppose we imagine the real line[36] as a weightless platform for our mass. To find the point under which to place a fulcrum so that the mass and its platform are balanced, we need to find the point with net torque equal to zero. Suppose that this is the distribution of a random variable X. That is, we want to find the point z such that

$$\begin{aligned} 0 &= \text{Net Torque} \\ &= \sum_u (u - z)\, \mathsf{p}_X(u) \quad \text{or} \quad \int_u (u - z)\, \mathsf{f}_X(u)\, du. \end{aligned}$$

In the discrete case, we have

$$\begin{aligned} 0 &= \sum_u (u - z)\, \mathsf{p}_X(u) \\ &= \sum_u u\, \mathsf{p}_X(u) - z \sum_u \mathsf{p}_X(u) \\ &= \mathsf{E}X - z. \end{aligned}$$

In the continuous case, we have

$$\begin{aligned} 0 &= \int_u (u - z)\, \mathsf{f}_X(u)\, du. \\ &= \int_u u\, \mathsf{f}_X(u)\, du - z \int_u \mathsf{f}_X(u)\, du. \\ &= \mathsf{E}X - z. \end{aligned}$$

Hence, $z = \mathsf{E}X$ is the balancing point.[37]

37 Notice again that the details – integration or sum, PMF or PDF – are different but that the essence of the operation is the same.

We have seen that $\mathsf{E}X$ can be interpreted as the long-run average of X over replications of the experiment. We have seen that $\mathsf{E}X$ can be interpreted as the fair entry price for a game with payoff X. We now have the following interpretation as well:

Interpretation. The expected value of a random variable $\mathsf{E}X$ can be interpreted as the balancing point of the distribution of X.

9. Summary

The distribution of a random variable describes the possible values the random variable can take (where) and the probability at or near each value (how much). A distribution is essentially a set of instructions for placing probability among these possible values.

We started with a more abstract definition, however. The distribution of a random variable embodies all of our predictions about functions of that random variable. If we know the distribution, we can make good guesses, though in general we cannot do more since the random variable is unobserved. From this idea, we derived the distribution operator D which is the mathematical object that contains all these predictions.

In practice, we do not compute the operator D explicitly but rather compute some other – simpler – representation of the distribution such as those described in the previous sections. Still, it is worthwhile to remember that all of these functions that we have described contain the same information that D does and can be derived from it.

Representation	Notation	How to get from D_X
Measure	m_X	$\mathsf{m}_X(\mathcal{A}) = \mathsf{D}_X 1_{\mathcal{A}}$
PMF	p_X	$\mathsf{p}_X(u) = \lim_{\Delta\to 0} \mathsf{D}_X 1_{[u,u+\Delta[}$
PDF	f_X	$\mathsf{f}_X(u) = \lim_{\Delta\to 0}(1/\Delta)\, \mathsf{D}_X 1_{[u,u+\Delta[}$
CDF	F_X	$\mathsf{F}_X(u) = \mathsf{D}_X 1_{]-\infty,u]}$
SF	S_X	$\mathsf{S}_X(u) = \mathsf{D}_X 1_{]u,\infty[}$
PGF	G_X	$\mathsf{G}_X(z) = \mathsf{D}_X g_z \quad$ where $g_z(u) = z^u$
MGF	M_X	$\mathsf{M}_X(z) = \mathsf{D}_X h_s \quad$ where $h_s(u) = e^{su}$
CGF	C_X	$\mathsf{C}_X(z) = \mathsf{D}_X r_t \quad$ where $r_t(u) = e^{itu}$

Each of these forms make different information accessible, and we will find situations where each is the easiest to use to solve our problem.

4

PREDICTION AND CONDITIONAL EXPECTED VALUES

Each morning the day lies like a fresh shirt on our bed; this incomparably fine, incomparably tightly woven tissue of pure prediction fits us perfectly. The happiness of the next twenty-four hours depends on our ability, on waking, to pick it up.

— *Walter Benjamin*

It's tough to make predictions, especially about the future.

— *Yogi Berra*

---

1. Prediction, Information, and Algorithms

We cannot avoid randomness, so we have to learn to live with it. Up to this point in the course, we have developed the tools for *describing* the randomness in a system. This is has proved useful for making decisions and predictions involving simple random systems. However, for modeling most realistic systems we need something more: a mechanism for updating our decisions and predictions for new information that we acquire as the experiment is unfolding. In this chapter, we develop this mechanism and, in the process, increase the power of the tools we already have.

During most random experiments of any realistic complexity, we get new information all the time. While playing blackjack, the dealer reveals cards at each stage and the player gets new cards. For predicting the weather (a notoriously difficult task), we have get new information about current conditions. For predicting the value of stocks, we get a steady stream of financial (and other) information about a company, other companies in the same sector, and the economy as a whole. To make good predictions, we must be able to use this information effectively.

Consider, for example, the problem of deciding whether to purchase stock in a particular company. To inform your decision, we could build a rather simple probability model for the system. You would purchase a fixed number of shares of the stock at the designated price and arrange to sell the stock at a pre-specified time, whatever the price. The random variable of interest in this experiment is your net return from the transaction – the ratio of the selling price to the purchase price of the stock, perhaps adjusted for taxes and commissions. With some effort (and several assumptions), we could determine the distribution of your net return, and in principle, could attempt to optimize the expected return over number of shares and the designated selling date. This distribution would embody the effect on the stock's price of all the information that comes to light between the purchase and sale of the stock, whether it be a new patent or an indictment against the company's CEO.

The only strength of the simple model is that we know how to analyze it. All in all, it is not a very satisfying model on which to base investment decisions. The primary reason is not an inadequacy in the derived distribution on the net return; in practice, this distribution could be constructed to more than adequately describe what would happen. Rather, it is that the investor must make all decisions in advance of the experiment. Information about new patents and indicted CEOs will affect the stock's price because it is unexpected by investors, and such information, may change what is the best course of action – to buy, sell, or hold. A better model would allow the investor to adjust his or her decisions to account for information as it becomes available.

Unfortunately, we do not yet have the tools to analyze such a model.

How does new information should change our calculations? To see that it does change, consider that the distribution of the net return in the simple model averages over many different possible outcomes, good and bad, mild and extreme. If we get new information about the company's prospects on the day before the designated sale of the shares, for instance, some of those possibilities are immediately eliminated, and the originally specified distribution is no longer valid. New information thus updates the possibilities in our original outcome space, and we would like to construct a "distribution" for the net return that ignores the eliminated outcomes. We can then base decisions on the updated distribution, which accounts for what we know at any moment.

In this view, the distribution of a random variable represents our state of knowledge about that random variable. At the beginning of the experiment, we consider all possibilities, weighted appropriately by their likelihood. But as new information comes in, some possibilities are eliminated and others refined, concentrating what we know about the random variable, until at last, we observe the random variables value and know it exactly.

The missing piece in our theory is, then, the mechanism for updating our knowledge for new information acquired as the experiment is running. We will cast this in terms of prediction: using the information in random variables that we have observed to make good guesses of the values of random variables that are yet unobserved. If we can make good predictions, then we can make good decisions, and we can incorporate new information effectively.

The basic schema is as follows. We observe the values of one or more random variables during the course of an experiment. These random variables are related in some way, possibly implicit, to another random variable in which we are interested. We will generate an algorithm that uses the observed random variables as inputs and produces a guess for the random variable of interest. Our goal is to find the best algorithm – the one that gives the most accurate predictions.

A key theme that will arise repeatedly is that any algorithm used for prediction corresponds to a function that takes the observed random variables as input and produces a guess for the unobserved random variables. So, if Y is a random variable we wish to predict and if X

(possibly vector-valued) is the random variable with which we will predict it, then any function g that takes values of X as input represents an algorithm. It takes an input (the observed values of X) and produces an output (the guess for Y). The random variable $g(X)$, then, measures the result of running that algorithm on the observed value of X, whatever it may be. By searching over all functions g, we can then find the best algorithm for prediction.

2. Expected Values As An Optimal Prediction

Before we approach the general problem of using one random variable to predict another, we will consider the special case of predicting a random variable without any special information beyond the distribution of that random variable. In other words, we want to find our best prediction of a random variable *at the beginning of the experiment*, when we have done all our calculations and before any new information is acquired from observing other, potentially related, random variables. The word "best" here is charged; the accuracy of a prediction depends in part on how we measure accuracy.

▶ EXAMPLE 4.1. Balanced Die

A standard, balanced, six-sided die is rolled once. Let D be the value showing on the upwards face after the roll. What's our best guess for D?

All six possibilities are equally likely, so our chance of making a mistake with any guess in $1, \dots, 6$ is the same. Alternatively, we might make a guess that will be close on average to the value of D that is observed. This suggests guessing a value between 3 and 4. We will see below that, using a sensible criterion, 3.5 is the value that is closest to D on average.

Suppose that Y is a real-valued random variable.[1] We want to find the *number* c that is closest to Y on average, in the sense that $\mathsf{E}(Y-c)^2$ is small.

The number $\mathsf{E}(Y-c)^2$ is called the *mean square error* of our guess c. To understand its meaning, it is helpful to look at the expression from the inside out. The random variable $Y - c$ takes a value near 0 whenever Y takes a value near c. This random variable does not really measure how close Y is to c because it can be both positive and negative. So, we use a function[2] of $Y - c$, in this case $(Y - c)^2$. The random variable $(Y - c)^2$ is small when Y is close to c and large when Y is far from c, so it measures the right notion – the deviation of Y

1 For our result to make sense, we will require the assumption that $\mathsf{E}Y^2 < \infty$. This technical assumption will hold in most cases we will encounter, but it can fail in practice.

2 It might seem more natural to consider $|Y - c|$ rather than $(Y - c)^2$, since the latter measures the distance between the two values. This is true, and it is reasonable; however, the absolute values make $|Y - c|$ hard to work with, whereas $(Y - c)^2$ is much more convenient. For example, we can always expand the quadratic. For this reason, we tend to prefer mean square error rather than mean absolute error.

from c. We then consider $\mathsf{E}(Y-c)^2$, the average deviation of Y from c. If $\mathsf{E}(Y-c)^2$ is small, then Y tends to be close to c, on average over identical replications. If $\mathsf{E}(Y-c)^2$ is large, then Y tends to be far from c, on average.

So, we want to find the value of the constant c that minimizes $\mathsf{E}(Y-c)^2$. We expand the quadratic to obtain

$$\begin{aligned}\mathsf{E}(Y-c)^2 &= \mathsf{E}(Y-\mathsf{E}Y+\mathsf{E}Y-c)^2\\ &= \mathsf{E}(Y-\mathsf{E}Y)^2+\mathsf{E}(\mathsf{E}Y-c)^2+2\,\mathsf{E}(Y-\mathsf{E}Y)(\mathsf{E}Y-c)\\ &= \mathsf{E}(Y-\mathsf{E}Y)^2+(\mathsf{E}Y-c)^2.\end{aligned}$$

The last equality follows because $\mathsf{E}Y-c$ is a constant. In particular, the cross term is zero:

$$\begin{aligned}\mathsf{E}(Y-\mathsf{E}Y)(\mathsf{E}Y-c) &= (\mathsf{E}Y-c)\,\mathsf{E}(Y-\mathsf{E}Y)\\ &= (\mathsf{E}Y-c)\,(\mathsf{E}Y-\mathsf{E}\mathsf{E}Y)\\ &= (\mathsf{E}Y-c)\,(\mathsf{E}Y-\mathsf{E}Y)\\ &= 0,\end{aligned}$$

and $\mathsf{E}(\mathsf{E}Y-c)^2=(\mathsf{E}Y-c)^2$.

The first term in

$$\mathsf{E}(Y-c)^2=\mathsf{E}(Y-\mathsf{E}Y)^2+(\mathsf{E}Y-c)^2$$

does not depend on c, and the second term is minimized by choosing $c=\mathsf{E}Y$. It follows that the best prediction[3] of Y at the beginning of the random experiment is the value $\mathsf{E}Y$.

3 In the mean-square sense.

> The optimal mean-square predictor of a random variable Y *at the beginning of the experiment*, that is before any extra information about Y is available, is the constant $\mathsf{E}Y$.

▶ EXAMPLE 4.1 Revisited.

$\mathsf{E}D = 3.5$ is, as hinted above, the mean-square optimal prediction of D. But what kind of prediction is this if D can never take the value 3.5? Why not choose 3 or 4, at least you might be right?

One way to understand these questions, and thus the mean square error criterion, is to think of a two guessing games. In the first game, we are penalized \$1 if we make an incorrect guess of D. In the second game, we are penalized an amount that is zero at the value of D and increases as the squared distance between our guess and D.

In the first game, a guess of 3.5 will always be penalized there can never be any payoff from guessing 3.5 because that value cannot happen. In the second game, the difference in average penalty between, say, 3 and 3.001 is very small, and 3.5 has the smallest average penalty. These two guessing games represent different criteria for thinking about the losses that arise from inaccuracy in our guesses. Using the expected value as a prediction corresponds to using the second game.

▶ EXAMPLE 4.2. Predicting a Coin Flip

A coin has probability $0 < p < 1$ of coming up heads on any given flip. What is the best prediction of the coin flip before we flip the coin?

Let Y be the indicator of heads on a single flip of the coin; then $\mathsf{E}Y = p$. This is our optimal prediction. Y can only take the values 0 or 1, and our prediction can be neither. How can that be a good prediction? If we had to guess one value or the other, it would be a bad prediction. But recall that our goal is to get *close* on average, and we ignore the option of guessing exactly. Notice that $\mathsf{E}Y$ adjusts for the bias in the coin in a reasonable way. If p is large, 1 is more likely, and our prediction is large. If p is small, 0 is more likely, and our prediction is small.

2.1. Not as Strange as It May Seem

As an additional motivation for the mean square error criterion, consider the case where we have a list of numbers $x_1, x_2, \ldots, x_n$, all known precisely. There is nothing random here. What is the constant c that minimizes the "mean square error" given by

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2.$$

By an argument similar to that for expected values, it turns out that the optimal $c = \bar{x}$ where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

the arithmetic average of the list. Hence, averages are the minimizers of mean square error; the result holds for expected values because they are simply averages, as we have seen.[4]

4 Incidentally, any median of the list minimizes the "mean absolute error",

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - c|.$$

This suggests the result that a median of the distribution should minimize $\mathsf{E}|Y - c|$, which is true quite generally.

2.2. Variance

Above, we saw that the constant $\mathsf{E}Y$ minimizes the mean square error of prediction at the beginning of an experiment. We also saw that the minimum value of this mean square error is the number

$$\mathsf{E}(Y - \mathsf{E}Y)^2.$$

This measures how good of a prediction $\mathsf{E}Y$ is. If Y tends to be close (in a mean square sense) to $\mathsf{E}Y$, then $\mathsf{E}Y$ is a good guess, and $\mathsf{E}(Y - \mathsf{E}Y)^2$ will be small. If Y tends to be far (in a mean square sense) from $\mathsf{E}Y$, then $\mathsf{E}Y$ is a poor guess, and $\mathsf{E}(Y - \mathsf{E}Y)^2$ will be large. An extreme example of the latter is the case where Y is a constant; then, $Y = \mathsf{E}Y$, so $\mathsf{E}(Y - \mathsf{E}Y)^2 = 0$.

We have the following important definition.

> If Y is any real-valued random variable, then
>
> $$\mathsf{Var}\, Y = \mathsf{E}(Y - \mathsf{E}Y)^2$$
>
> is called the *variance* of Y. The variance is a non-negative number that is zero only if Y is a constant. It measures how spread out the probability is around $\mathsf{E}Y$ in the distribution of Y. Or, put another way, it gives the minimum possible mean square error of prediction of Y at the beginning of the experiment.

Variance will be discussed more fully in Chapter 5.

2.3. Balancing Point Interpretation Revisited

One interpretation of expected values that we have discussed is that $\mathsf{E}Y$ is the "balancing point" of the distribution of Y. Specifically, we can think of the distribution of Y as representing a mass distribution. The total torque generated by that distribution when the fulcrum is placed at position c is proportional to $\mathsf{E}(Y - c)$ and the balancing point is the point c with $\mathsf{E}(Y - c) = 0$. If we throw caution to the wind and allow for the interchangeability of the E and derivative operators, then, solving the equation $\mathsf{E}(Y-c)=0$ is equivalent to minimizing $\mathsf{E}(Y-c)^2$ because $\frac{\partial}{\partial c}\mathsf{E}(Y-c)^2 = 2\mathsf{E}(Y-c)$.

▶ EXAMPLE 4.3. The Standard Normal Distribution

Let Y be a continuous random variable with PDF

$$\mathsf{f}_Y(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \qquad -\infty < t < \infty.$$

This distribution is symmetric about 0 assigns non-zero probability near every real number. The balancing point of the distribution is thus 0, so $\mathsf{E}Y = 0$, which is consequently our best prediction of Y.

3. Optimal Mean Square Prediction Algorithms

A gambler sits at the blackjack table and begins a new game. With every turn of the cards, the gambler receives new information, information that can alter his strategy and determine his next move. In an ideal world, at each turn of the card, the gambler could stop, pull out his laptop, and do some calculations as to the best choice to make for the next card: hold, hit, split, double down, and so forth. Unfortunately, casinos are a bit annoyed by such behavior; several large men without necks would soon appear to escort him from the casino. Since he cannot halt the game to calculate his best move, the gambler's only recourse is to encode his prediction algorithm – the rules that will map a particular set of cards to a particular action – before the game is played.[5] The gambler can construct this algorithm in advance because he has a complete description of the experiment, its outcomes, and the distributions of all relevant random variables. When the information actually does become available during the game, he can then act on it.

Our goal is to show how to construct good prediction algorithms in general. Like the blackjack player, we will specify our algorithms *before* the random experiment begins so that we can act quickly as new information becomes available. Mathematically, prediction is the process of guessing the value of one random variable in advance using the value of another random variable that we observe while the experiment runs. Throughout this chapter, we will use Y to denote the scalar random variable[6] to be predicted and X to denote the random variable[7] we will use to predict Y.

5 Of course, he has to be able to compute it in real time, in his head, as well. Blackjack is one of the few casino games for which such systems can give one an advantage; indeed, it is one of the few that a gambler can win consistently. However, the casinos do as much as possible to make this hard, such as using six or seven decks in a typical shoe and removing anyone who is doing too well.

6 Once we learn how to predict scalar-valued random variables, we can easily predict vector-valued random variables as well. Suppose that $Y = (Y_1, \ldots, Y_m)$ is an $\mathbb{R}^m$-valued random variable. Then, if $g_1, \ldots, g_m$ are our optimal predictors given X of respective scalar-valued random variables $Y_1, \ldots, Y_m$, $g = (g_1, \ldots, g_m)$ is our optimal predictor of Y given X.

7 X can be vector valued.

Our goal is to use the information in X that relates to Y as effectively and economically as possible. We seek an algorithm that takes the observed value of X as input and produces the best available guess for the value that Y will take. As we have seen, we can think of that algorithm as a *function* whose argument is a possible value of X and whose return value is the guess for Y.

A predictor of Y given X is a rule that associates to each possible value of X a guess for the value of Y. We encode such a rule as a function that takes a possible value of X and returns a guess for the value of Y. Hence, we say that a *predictor* of Y given X is any function g that maps the range of X to $\mathbb{R}$ such that $g(x)$ is a guess for Y when X is observed to take the value x. There is nothing in this definition that requires the predictor to yield accurate guesses; finding a good predictor is our job.

A *prediction* of Y given X is the random variable that represents the guess we will make, under a specific predictor, when the value of X is actually observed. If g is a predictor of Y given X, then the random variable $g(X)$ is the corresponding prediction. It is a random variable because, at the beginning of the experiment, we do not know what value X will take. Once we see the value of X, we know exactly what the prediction (guess for Y) will be. Notice that each prediction corresponds to a specific predictor.

▶ EXAMPLE 4.4. Biased Coins

Two coins lay on a table. One comes up heads with probability 1/4 on any flip, and the other comes up heads with probability 3/4 on any flip. We do not know which is which. We choose one at random (both equally likely to be chosen) and flip the chosen coin six times. Assume that all flips are independent. Let X be the number of heads that occur in the first five flips, and let Y be the indicator that the coin comes up heads on the sixth flip.

We know before the experiment proceeds that X can be 0, 1, 2, 3, 4, or 5. (In fact, we can compute the distribution of X.) We will learn the value of X before the value of Y, and this value helps us guess Y because it has information about whether the chosen coin is biased towards heads or tails. Consider the following predictors of Y given X, defined for each $k \in \{0, \dots, 5\}$ by

$$g_1(k) = 1$$

8 Try to work this out.

$$g_2(k) = \begin{cases} 1 & \text{if } k \geq 3 \\ 0 & \text{if } k \leq 2 \end{cases}$$

$$g_3(k) = \frac{\frac{3}{4}\,3^k + \frac{1}{4}\,3^{5-k}}{3^k + 3^{5-k}}.$$

The first predictor g_1 always guesses heads regardless of the value of X. This is a perfectly valid predictor; it just ignores X. The corresponding prediction is the constant random variable $g_1(X) \equiv 1$; because g_1 ignores X, we know the value of the predictions ahead of time. The second predictor g_2 guesses heads if we have seen more heads than tails and tails if we have seen more tails than heads. The logic here is that the $3/4$ coin is more likely to produce more heads than tails and the $1/4$ coin more likely to produce more tails than heads. Having guessed the coin from X, our guess for Y is then the most likely outcome. The prediction $g_2(X)$ is an indicator random variable which turns out[8] to take the value 1 with probability $1/2$. The third predictor g_3 seems more mysterious on first glance because it produces values *between* 0 and 1. As k ranges from 0 to 5, $g_3(k)$ is approximately 0.252, 0.268, 0.375, 0.625, 0.732, and 0.748. These values can only be understood as guesses for Y if we judge the accuracy of the guesses by how close they are to the true value rather than by the likelihood that they are exactly correct. More on this later. The prediction $g_3(X)$ is a random variable that takes the six values listed above. Can you find the PMF of this random variable?

The Biased Coins example shows that if we are to find accurate predictors, we first must define a criterion for accuracy. A natural criterion in the example is that we want to choose a predictor g to maximize $\mathsf{P}\{\, g(X) = Y \,\}$, the probability that our prediction is exactly correct. It turns out that

$$\mathsf{P}\{\, g_1(X) = Y \,\} = \frac{1}{2}$$
$$\mathsf{P}\{\, g_2(X) = Y \,\} = 0.698$$
$$\mathsf{P}\{\, g_3(X) = Y \,\} = 0.$$

With this criterion for accuracy, the third predictor is useless because it never guesses 0 or 1. The second predictor improves markedly over the first.

An alternative criterion judges not whether the guess is exactly correct but whether it is close: choose a predictor g to minimize $\mathsf{E}(g(X) -$

$Y)^2$. This is called the *mean square error* criterion.[9] We have

$$\mathsf{E}(g_1(X) - Y)^2 = \frac{1}{2}$$
$$\mathsf{E}(g_2(X) - Y)^2 = 0.302$$
$$\mathsf{E}(g_3(X) - Y)^2 = 0.208.$$

So here, g_2 again improves on g_1, but now g_3 is the clear winner. In fact, g_3 was designed to minimize the mean square error, which explains its mysterious structure. It was eliminated with the first criterion, not because it is a bad predictor, but because it does not strive for exactness.

So which criterion should we use? In practice, our measure of accuracy for prediction will depend on the context in which we are using it. Different applications have different needs, and no criterion is better in all circumstances. However, mean square error has many strengths to recommend it as a default: it is intuitive, useful in a wide variety of applications, works for discrete and continuous random variables, leads to a powerful theory, and has strong connections to averaging that we can exploit. For all of these reasons, we will use the mean square error criterion as our measure of accuracy.[10]

The *optimal (mean square) predictor* of Y given X is the function g defined on the range of X that minimizes $\mathsf{E}(Y - g(X))^2$. We will denote the optimal mean square predictor by $\mathsf{pred}_{Y|X}$. Thus, if g is the function minimizing $\mathsf{E}(Y - g(X))^2$, $\mathsf{pred}_{Y|X} = g$; this is an equality between functions meaning that $\mathsf{pred}_{Y|X}(u) = g(u)$ for all u.[11] The subscript is only a label to designate what random variable is being predicted and what random variable is being used to predict it. The $|$ is read "given"; anything to the right of it in an expression is being taken as a known input for prediction. For example, if W and U are random variables, $\mathsf{pred}_{W|U}$ is the optimal predictor of W given U and $\mathsf{pred}_{U|W}$ is the optimal predictor of U given W. The *optimal (mean square) prediction* of Y given X is the random variable denoted by the mind-blowing expression $\mathsf{pred}_{Y|X}(X)$. Do not be confused by the multiple uses of X in this expression: $\mathsf{pred}_{Y|X}$ is some function g as above and $\mathsf{pred}_{Y|X}(X)$ is simply the random variable $g(X)$; the subscript is again only a label. Actually, both of these notations will soon pass by the wayside. In the next section, we will make an important connection

9 One might reasonably ask: what happened to absolute differences $\mathsf{E}|g(X)-Y|$. This is called the mean absolute error. It is a very reasonable criterion, but it is difficult to work with because the absolute values brook little manipulation. The mean square error is similar in spirit and allows for a more flexible theory. One advantage of mean square error in some contexts is that very small errors are lightly penalized (relative to absolute error) while large errors are more severely penalized.

10 From here on, "optimal predictor" or "optimal prediction" should be implicitly taken to be in the mean square sense.

11 In general, the value of the optimal predictor is undefined for values that X cannot possibly take. We can supply an arbitrary value in those cases such as defining $\mathsf{pred}_{Y|X}(u) = 0$ for values of u such that $\mathsf{P}\{\, X \text{ near } u \,\} = 0$.

and see that the mean square optimal predictor and predictions can be written in a more useful and intuitive way.

First, however, we must learn to understand and calculate optimal predictors. In the following subsections, we study the optimal predictors in increasingly complicated situations.

Summary of Terms
Let Y be a scalar-valued random variable and X be an arbitrary (possibly vector-valued) random variable.
• A predictor of Y given X is a function that maps each possible value of X (that is, each value in the range of X) to a real number. This number represents our guess of the value of Y if the corresponding value of X is observed.
• A prediction of Y given X is the random variable that represents the guess that will be made, using a particular predictor, when X is eventually observed.
• The optimal (mean square) predictor, $\mathsf{pred}_{Y\|X}$, is the function g that minimizes $\mathsf{E}(Y - g(X))^2$.

3.1. Prediction Given Whether An Event Occurs

The simplest information we can obtain during an experiment is whether or not a particular event has occurred. In playing blackjack, we see whether the dealer's exposed card is a face card or an ace. In predicting the price of a stock, we learn whether a company's most recent earnings exceeded analysts expectations. When predicting tomorrow's whether, we learn today whether it rained.

Let $\mathcal{A}_1$ be some event and, for convenience, let $\mathcal{A}_0 = \mathrm{compl}\,(\mathcal{A}_1)$. Observing whether $\mathcal{A}_1$ occurred is equivalent to observing the value of the random variable $X = 1_{\mathcal{A}_1}$.

To find the optimal predictor of Y given X, we need to search over all functions g from $\{0, 1\}$ to $\mathbb{R}$ for that which minimizes $\mathsf{E}(Y - g(X))^2$. Fortunately, this is not as hard as it sounds because any function g defined on the two-valued set $\{0, 1\}$ can take at most two distinct values. If we define $c_0 = g(0)$ and $c_1 = g(1)$, then we can write

$$g(X) = c_0 1_{\mathcal{A}_0} + c_1 1_{\mathcal{A}_1}.$$

We need only search over all possible values of c_0 and c_1 to find the optimal g.

Because $\mathcal{A}_1$ and $\mathcal{A}_0$ are disjoint, can also write that

$$(g(X))^2 = c_0^2 1_{\mathcal{A}_0} + c_1^2 1_{\mathcal{A}_1}.$$

Therefore, we can write

$$\begin{aligned}
\mathsf{E}(Y - g(X))^2 &= \mathsf{E}Y^2 + \mathsf{E}(g(X))^2 - 2\mathsf{E}Y g(X) \\
&= \mathsf{E}Y^2 + \mathsf{E}\left(c_0^2 1_{\mathcal{A}_0} + c_1^2 1_{\mathcal{A}_1}\right) - 2\mathsf{E}Y\left(c_0 1_{\mathcal{A}_0} + c_1 1_{\mathcal{A}_1}\right) \\
&= \mathsf{E}Y^2 + c_0^2\,\mathsf{E}1_{\mathcal{A}_0} + c_1^2\,\mathsf{E}1_{\mathcal{A}_1} - 2c_0\,\mathsf{E}Y1_{\mathcal{A}_0} - 2c_1\,\mathsf{E}Y1_{\mathcal{A}_1}.
\end{aligned}$$

To minimize this, let's take derivatives with respect to the c_is and set them to 0:

$$\frac{\partial}{\partial c_0}\mathsf{E}(Y - g(X))^2 = 0 + 2\,c_0\,\mathsf{E}1_{\mathcal{A}_0} - 2\mathsf{E}Y1_{\mathcal{A}_0} = 0$$

and

$$\frac{\partial}{\partial c_1}\mathsf{E}(Y - g(X))^2 = 0 + 2\,c_1\,\mathsf{E}1_{\mathcal{A}_1} - 2\,\mathsf{E}Y1_{\mathcal{A}_1} = 0.$$

Solving these equations gives

$$\begin{aligned}
c_0 &= \frac{\mathsf{E}Y1_{\mathcal{A}_0}}{\mathsf{E}1_{\mathcal{A}_0}} = \frac{\mathsf{E}Y1_{\mathcal{A}_0}}{\mathsf{P}(\mathcal{A}_0)} \\
c_1 &= \frac{\mathsf{E}Y1_{\mathcal{A}_1}}{\mathsf{E}1_{\mathcal{A}_1}} = \frac{\mathsf{E}Y1_{\mathcal{A}_1}}{\mathsf{P}(\mathcal{A}_1)}.
\end{aligned}$$

That is, the optimal predictor is given by[11]

$$\mathsf{pred}_{Y|1_{\mathcal{A}_1}}(u) = \begin{cases} \dfrac{\mathsf{E}Y1_{\mathcal{A}_0}}{\mathsf{E}1_{\mathcal{A}_0}} & \text{if } u = 0 \\ \dfrac{\mathsf{E}Y1_{\mathcal{A}_1}}{\mathsf{E}1_{\mathcal{A}_1}} & \text{if } u = 1. \end{cases} \tag{4.1}$$

▶ EXAMPLE 4.5. Nonstandard Dice

Two balanced six sided dice are labeled in a non-standard way. The first has faces labeled 1, 1, 2, 2, 3, 3, and the second has faces labeled 2, 2, 4, 4, 6, 6. We choose one of the two dice at random, both being equally likely to be chosen, and roll the chosen die. Let D be the value on the upward face of the chosen die and let T be the indicator that we chose the first die. Then,

$$\mathsf{pred}_{D|T}(u) = \begin{cases} 4 & \text{if } u = 0 \\ 2 & \text{if } u = 1. \end{cases}$$

We can see this in two ways. First, the intuition of prediction, says that if you knew that the first die was chosen, your best guess for the value of D would be the expected value of a single roll of that die, or 2. Similarly, if you knew the second die was chosen, your best guess for the value of D would be the expected value of a single roll of that die, or 4. This follows because the expected value is the mean square optimal predictor of a random variable without additional information.

Second, we could establish this more formally. Plugging into equation (4.1), we have that

$$\mathsf{pred}_{D|T}(u) = \begin{cases} \dfrac{\mathsf{E}Y(1-T)}{\mathsf{E}(1-T)} & \text{if } u = 0 \\ \dfrac{\mathsf{E}YT}{\mathsf{E}T} & \text{if } u = 1. \end{cases}$$

We know that $\mathsf{E}T = 1/2$ and $\mathsf{E}(1-T) = 1/2$ by the stated assumptions. To find $\mathsf{E}YT$ and $\mathsf{E}Y(1-T)$, we need only find the joint PMF of Y and T:

$$\mathsf{p}_{Y,T}(j,k) = \begin{cases} \dfrac{1}{6} & \text{if } k = 0 \text{ and } j \in \{2,4,6\} \\ \dfrac{1}{6} & \text{if } k = 1 \text{ and } j \in \{1,2,3\} \\ 0 & \text{otherwise.} \end{cases}$$

We then have

$$\begin{aligned} \mathsf{E}YT &= \sum_{j=1}^{6}\sum_{k=0}^{1} j \cdot k \cdot \mathsf{p}_{Y,T}(j,k) \\ &= \sum_{j=1,2,3} j \cdot \mathsf{p}_{Y,T}(j,1) \\ &= 1 \\ \mathsf{E}Y(1-T) &= \sum_{j=1}^{6}\sum_{k=0}^{1} j \cdot (1-k) \cdot \mathsf{p}_{Y,T}(j,k) \end{aligned}$$

$$= \sum_{j=2,4,6} j \cdot \mathsf{p}_{Y,T}(j,0)$$
$$= 2.$$

The optimal predictor given above then follows.

▸ EXAMPLE 4.6. Two Independent Coin Flips

We flip a coin twice. Assume both flips are independent. Let H_1 and H_2 be respectively the indicators that the first and second flips come up heads. What is the best predictor of H_2 given H_1?

Again, we can approach this problem in two ways. First, the random variables H_1 and H_2 are independent by assumption; this means that knowing the value of one of them does not change our predictions about the other. Thus, our best guess for H_2 given H_1 should be the same as our best guess for H_2 without any information, namely $\mathsf{E}H_2$. That is,

$$\mathsf{pred}_{H_2|H_1}(u) = \mathsf{E}H_2,$$

for any u.

Second, we can see this formally by using the definition in equation (4.1) above. We have that

$$\mathsf{pred}_{H_2|H_1}(u) = \begin{cases} \dfrac{\mathsf{E}H_2(1-H_1)}{\mathsf{E}(1-H_1)} = \mathsf{E}H_2 & \text{if } u = 0 \\ \dfrac{\mathsf{E}H_2H_1}{\mathsf{E}H_1} = \mathsf{E}H_2 & \text{if } u = 1. \end{cases}$$

Since H_1 and H_2 are independent, $\mathsf{E}H_2H_1 = \mathsf{E}H_2\mathsf{E}H_1$ and $\mathsf{E}H_2(1-H_1) = \mathsf{E}H_2\mathsf{E}(1-H_1)$; the denominators thus cancel, leaving $\mathsf{E}H_2$ in both cases.

Equation (4.1) shows us how to update our predictions when we obtain one bit of new information – whether $\mathcal{A}_0$ or $\mathcal{A}_1$ occurred. It is important to understand that this result extends the result we derived in the last section. Specifically, at the beginning of the experiment, we know the value of the random variable $1_\Omega \equiv 1$. We found that $\mathsf{pred}_{Y|1_\Omega}(u) = \mathsf{E}Y$. Notice that this result falls out of equation (4.1) by taking $\mathcal{A}_1 = \Omega$. In this case, $\mathcal{A}_0$ can never happen,[12] and $\mathsf{E}Y1_\Omega/\mathsf{E}1_\Omega = \mathsf{E}Y$. Equation (4.1) takes us beyond this by allowing us to update our predictions for other events whose occurrence is uncertain at the beginning of the experiment.

We can gain further insight into the meaning of equation (4.1) with a simple metaphor. During practice, a long jumper took twenty jumps, and her coach recorded the distances, marking down 0 if she faulted[13]

12 Put another way, 1_Ω represents 0 bits of information, since it is always 1.

13 A fault occurs if the jumper's foot crosses a delineated boundary while taking a jump.

on a jump. Suppose the final series of recorded jumps is as follows (in feet): 15, 16, 0, 18, 0, 12, 0, 0, 14, 15, 20, 18, 17, 16, 0, 10, 14, 15, 13, 12. The coach does a quick calculation and finds that the average of this list is 11.25 feet and that the jumper faulted on 25% of the jumps (5 out of 20). Suppose that the coach wished to calculate the jumper's average lengths for the non-faulted jumps, perhaps to assess what she is capable of. He could of course do this by adding up the non-zero numbers in the list and dividing by 15. However, he can also calculate it using only the overall average of the list and the proportion of non-faulted jumps. Take a moment to see how this could be done.

... Reader Pausing To Think ...

The 0s in the list bring down the overall average because they add nothing to the sum in the numerator but add 1 each to the count in the denominator. To compute the average that excludes the 0s, we need only compensate for this damping effect. The adjusted average is $11.25/0.75 = 15$. To see this, note that

$$\begin{aligned} 15 = \frac{11.25}{0.75} &= \frac{\text{sum of non-zero jumps } + 0 + 0 + 0 + 0 + 0}{20} \cdot \frac{20}{15} \\ &= \frac{\text{sum of non-zero jumps}}{15}. \end{aligned}$$

In other words, all we have to do to eliminate the effect of the zeros is to renormalize the sum by the proportion of cases we are averaging.

Now, consider the expression

$$\frac{\mathsf{E}Y1_{\mathcal{A}_1}}{\mathsf{E}1_{\mathcal{A}_1}}$$

in equation (4.1). Both the numerator and denominator are long-run averages of the corresponding random variables over replications of the experiment. For any elementary outcome ω^* in the sequence of replications, the random variable $Y1_{\mathcal{A}_1}$ takes the value 0 unless $\omega^* \in \mathcal{A}_1$ in which case it takes the value $Y(\omega^*)$. It follows that the long-run average $\mathsf{E}Y1_{\mathcal{A}_1}$ is the average[14] of a list of values of Y with 0s replacing any replication for which $\mathcal{A}_1$ did not occur. The 0s correspond here to the faulted jumps, and $\mathsf{E}Y1_{\mathcal{A}_1}$ to the overall average of the list. The denominator $\mathsf{E}1_{\mathcal{A}_1}$, however, is the long-run proportion of the time that

14 Actually, a limit of averages of increasingly long lists, but the idea is the same.

$\mathcal{A}_1$ occurs; this corresponds to the proportion of non-faulted jumps. By dividing $\mathsf{E}Y1_{\mathcal{A}_1}$ by $\mathsf{E}1_{\mathcal{A}_1}$, we are renormalizing the average to be the long-run average of values of Y *only for cases in which* $\mathcal{A}_1$ *occurs*. This is the analogue of the average over non-faulted jumps.

This bears repeating. The optimal predictor of Y with the information that $\mathcal{A}_1$ has occurred is the long-run average of Y only over those replications for which $\mathcal{A}_1$ occurs. We can see this more formally as

$$\begin{aligned}
\frac{\mathsf{E}Y1_{\mathcal{A}_1}}{\mathsf{E}1_{\mathcal{A}_1}} &= \frac{\lim_{n\to\infty} \frac{1}{n}\ (Y(\omega_1^*)1_{\mathcal{A}_1}(\omega_1^*) + \cdots + Y(\omega_n^*)1_{\mathcal{A}_1}(\omega_n^*))}{\lim_{n\to\infty} \frac{1}{n}\ (1_{\mathcal{A}_1}(\omega_1^*) + \cdots + 1_{\mathcal{A}_1}(\omega_n^*))} \\
&= \lim_{n\to\infty} \frac{(Y(\omega_1^*)1_{\mathcal{A}_1}(\omega_1^*) + \cdots + Y(\omega_n^*)1_{\mathcal{A}_1}(\omega_n^*))}{(1_{\mathcal{A}_1}(\omega_1^*) + \cdots + 1_{\mathcal{A}_1}(\omega_n^*))} \\
&= \lim_{n\to\infty} \text{Average of } Y(\omega_k^*) \text{ over } 1 \le k \le n \text{ with } \omega_k^* \in \mathcal{A}_1.
\end{aligned}$$

The resulting long-run average quacks like an expected value, suitably adjusted for the specific knowledge that ω^* (the elementary outcome that occurs in our actual experiment) is in $\mathcal{A}_1$. If it quacks like an expected value, well .... See the next section.

Finally, while it may not seem like it, the result in the special case where X is an indicator immediately provides us with the general form of the optimal predictor for arbitrary X. The rule in equation (4.1) can be interpreted as telling us that if an event $\mathcal{A}$ is known to have occurred, the optimal prediction Y with that information is

$$\frac{\mathsf{E}Y1_{\mathcal{A}}}{\mathsf{P}(\mathcal{A})}.$$

So, suppose X be an arbitrary discrete random variable, and let u be a value for which $\mathsf{p}_X(u) > 0$. If we observe that the event $\{\, X = u \,\}$ has occurred – in other words, that X takes the value u – then our best guess for Y should then be $\mathsf{E}Y1_{\{\, X=u \,\}}/\mathsf{P}\{\, X = u \,\}$. Since u is arbitrary here, subject to $\mathsf{p}_X(u) > 0$, this gives us our best prediction for any observed value of X. Notice that in the discrete case, $\{\, X = u \,\} = \{\, X \text{ near } u \,\}$, so we can write this as $\mathsf{E}Y1_{\{\, X \text{ near } u \,\}}/\mathsf{P}\{\, X \text{ near } u \,\}$. Why do this? Consider the case where X is a continuous random variable and u is a value with $\mathsf{f}_X(u) > 0$. It follows from this that $\mathsf{P}\{\, X \text{ near } u \,\} > 0$, so if we observe that the event $\{\, X \text{ near } u \,\}$ has occurred, our best guess for Y should then be $\mathsf{E}Y 1_{\{X \text{ near } u\}}/\mathsf{P}\{X \text{ near } u\}$, which is of the same form. We thus have a quite general rule.

The Optimal Predictor of *Y* given *X*

For scalar-valued random variable Y and arbitrary random variable X, the optimal predictor of Y given X is defined by

$$\mathsf{pred}_{Y|X}(u) = \frac{\mathsf{E}Y\, 1_{\{X \text{ near } u\}}}{\mathsf{P}\{X \text{ near } u\}}, \tag{4.2}$$

for all u for which $\mathsf{P}\{X \text{ near } u\} > 0$. If X is discrete, then we can write $\{X \text{ near } u\}$ as $\{X = u\}$ for convenience. The optimal predictor can be given an arbitrary value, such as 0 or $\mathsf{E}Y$, for other values u.

The next two subsections formalize this derivation and show how to compute the resulting predictor in some common special cases, such as when Y and X are both discrete or both continuous.

3.2. Prediction Given a Discrete Random Variable

Let X be a discrete, real-valued random variable with PMF p_X, and for the moment, let Y be an arbitrary real-valued random variable. The argument at the end of subsection 3.1 tells us that

$$\mathsf{pred}_{Y|X}(u) = \frac{\mathsf{E}Y\, 1_{\{X=u\}}}{\mathsf{p}_X(u)}.$$

Our goal in this subsection is twofold: to rederive the optimal predictor directly, thus validating the argument above, and to show what form the predictor takes in an important case.

First, suppose that X takes the values $x_1, x_2, \ldots$ where the x_is are distinct real numbers.[15] We can write X in the simple (but self-referential) form

$$X = \sum_i x_i\, 1_{\{X=x_i\}}. \tag{4.3}$$

It follows that for any function g defined on all the x_is, we can write

$$g(X) = \sum_i g(x_i)\, 1_{\{X=x_i\}}. \tag{4.4}$$

Make sure you understand why this is true.[16]

15 For any discrete random variable, we can enumerate its possible values in precisely this way.

16 Notice that by taking expected values on both sides of this equation, we get the expression for $\mathsf{E}g(X)$ that occurs in the formal definition of the PMF.

The function g defined on range(X) is determined by the values $c_i = g(x_i)$, so we can find the minimizing g by searching over the constants c_i, for $i = 1, 2, \ldots$. We have

$$\begin{aligned}
\mathsf{E}(Y - g(X))^2 &= \mathsf{E}Y^2 + \mathsf{E}(g(X))^2 - 2\mathsf{E}Yg(X) \\
&= \mathsf{E}Y^2 + \sum_i g^2(x_i)\, \mathsf{p}_X(x_i) - 2\sum_i g(x_i)\, \mathsf{E}(Y1_{\{\,X=x_i\,\}}) \\
&= \mathsf{E}Y^2 + \sum_i c_i^2\, \mathsf{p}_X(x_i) - 2\sum_i c_i\, \mathsf{E}(Y1_{\{\,X=x_i\,\}}),
\end{aligned}$$

where the second term follows from the definition of the PMF and the last term follows by the additivity of E and equation (4.4). To minimize over all the c_i, we take partial derivatives and set them to 0, yielding

$$\frac{\partial}{\partial c_i}\mathsf{E}(Y - g(X))^2 = 2c_i\, \mathsf{p}_X(x_i) - 2c_i\, \mathsf{E}(Y1_{\{\,X=x_i\,\}}).$$

Setting the derivatives equal to zero[17] gives

$$\frac{\partial}{\partial c_i}\mathsf{E}(Y - g(X))^2 = 0 \implies c_i = \frac{\mathsf{E}(Y1_{\{\,X=x_i\,\}})}{\mathsf{p}_X(x_i)}.$$

Therefore, the optimal predictor is exactly that given above by our earlier arguments. Again, we see that our best prediction of Y if we know that $X = x_i$ is the long-run average of Y over only those replications for which $\{\, X = x_i \,\}$ occurs.

Second, we turn to an important special case. Suppose that X and Y are both discrete with join PMF $\mathsf{p}_{X,Y}$. With these assumptions, we can give the optimal predictor a more specific form. Notice that $Y\,1_{\{\,X=u\,\}} = h(X, Y)$ where

$$h(s,t) = \begin{cases} t & \text{if } s = u, \\ 0 & \text{if } s \neq u. \end{cases}$$

Hence,

$$\begin{aligned}
\mathsf{E}(Y\,1_{\{\,X=u\,\}}) &= \sum_{v,t} h(v,t)\, \mathsf{p}_{X,Y}(t,v) \\
&= \sum_v v\, \mathsf{p}_{X,Y}(u,v),
\end{aligned}$$

by the definition of h. Similarly,

$$\mathsf{E}1_{\{\,X=u\,\}} = \mathsf{P}\{\, X = u \,\} = \mathsf{p}_X(u) = \sum_v \mathsf{p}_{X,Y}(u,v).$$

17 Notice that there can be countably infinitely many c_is, but this poses no problems since all of the partial derivatives are well defined

It follows that the optimal predictor can be written as

$$\mathsf{pred}_{Y|X}(u) = \frac{\sum_v v\, \mathsf{p}_{X,Y}(u,v)}{\sum_v \mathsf{p}_{X,Y}(u,v)} = \frac{\sum_v v\, \mathsf{p}_{X,Y}(u,v)}{\mathsf{p}_X(u)},$$

for u with $\mathsf{p}_X(u) > 0$. (For other u, we can define $\mathsf{pred}_{Y|X}(u) = \mathsf{E}Y$. This is arbitrary but makes the predictor a function defined for all u.) Because $\mathsf{p}_X(u) = \sum_v \mathsf{p}_{X,Y}(u,v)$, the denominator can be written as $\mathsf{p}_X(u)$ as well, yielding:

$$\mathsf{pred}_{Y|X}(u) = \sum_v v\, \frac{\mathsf{p}_{X,Y}(u,v)}{\mathsf{p}_X(u)} = \sum_v v\, \frac{\mathsf{p}_{X,Y}(u,v)}{\sum_w \mathsf{p}_{X,Y}(u,w)}.$$

By moving the denominator inside the sum reveals that $\mathsf{p}_X(u)$ is exactly the factor by which we must divide $\mathsf{p}_{X,Y}(u,v)$ so that it sums to 1 over all v. This leads us to define a family of PMFs, called the *conditional* PMFs of Y given X, as follows

$$\mathsf{p}_{Y|X}(v \mid u) = \mathsf{P}\{\, Y = v \mid X = u \,\} = \frac{\mathsf{p}_{X,Y}(u,v)}{\mathsf{p}_X(u)},$$

for u with $\mathsf{p}_X(u) > 0$. If $\mathsf{p}_X(u) = 0$, we define $\mathsf{p}_{Y|X}(v \mid u) = 0$. For each possible value of u, $\mathsf{p}_{Y|X}(v \mid u)$ is a PMF as a function of v; it describes the updated distribution of Y given the information that X equals the value u. These conditional PMFs are described fully in Section 7 below.

Notice that in terms of the conditional PMFs, we can write the optimal predictor of Y given X as a weighted average that looks suspiciously like an expected value:

$$\mathsf{pred}_{Y|X}(u) = \sum_v v\, \mathsf{p}_{Y|X}(v \mid u). \tag{4.5}$$

Indeed, in the next section, we will define a conditional expected value $\mathsf{E}(Y \mid X = u)$ which it turns out is equal to $\mathsf{pred}_{Y|X}(u)$. That is,

$$\mathsf{E}(Y \mid X = u) = \sum_v v\, \mathsf{p}_{Y|X}(v \mid u).$$

▶ EXAMPLE 4.7. A Random Random Number

Draw a random integer N uniformly from $\{0, 1, 2, 3, 4\}$ and then draw a random integer M uniformly between 0 and N. Find the best predictor of M given N.

For any integer k from 0 to 4, we have from equation (4.5) that

$$\begin{aligned}
\mathsf{pred}_{M|N}(k) &= \sum_{j=0}^{4} j\,\mathsf{p}_{M|N}(j \mid k) \\
&= \sum_{j=0}^{k} j\,\mathsf{P}\{\, M = j \mid N = k \,\} \\
&= \sum_{j=0}^{k} j\,\frac{1}{k+1} \\
&= \frac{k(k+1)}{2(k+1)} \\
&= \frac{k}{2},
\end{aligned}$$

because if $N = k$, then we know that M is uniformly distributed over $\{0, 1, \ldots, k\}$.

▶ EXAMPLE 4.8. A Random Sum of Indicators

Let N be a positive, integer-valued random variable and let $I_1, I_2, \ldots$ be a sequence of independent indicators with $\mathsf{E}I_k = p$ for all $k \in \mathbb{Z}_+$ and some $0 < p < 1$. Assume that N is independent of all the I_ks. Define

$$S = \sum_{j=1}^{N} I_j.$$

Notice that S is a sum of a random number of indicators and is consequently a bit more complicated than the typical finite sums we have dealt with.

Find the best predictor of S given N.

$$\begin{aligned}
\mathsf{pred}_{S|N}(n) &= \frac{\mathsf{E}S\,1_{\{\,N=n\,\}}}{\mathsf{p}_N(n)} \\
&= \frac{\mathsf{E}\sum_{j=1}^{N} I_j\,1_{\{\,N=n\,\}}}{\mathsf{p}_N(n)} \\
&= \frac{\mathsf{E}\sum_{j=1}^{n} I_j\,1_{\{\,N=n\,\}}}{\mathsf{p}_N(n)} \\
&= \frac{\mathsf{E}1_{\{\,N=n\,\}}\,\mathsf{E}\sum_{j=1}^{n} I_j}{\mathsf{p}_N(n)}
\end{aligned}$$

$$= \frac{\mathsf{p}_N(n) \sum_{j=1}^{n} \mathsf{E} I_j}{\mathsf{p}_N(n)}$$
$$= \frac{\mathsf{p}_N(n)\, np}{\mathsf{p}_N(n)}$$
$$= np.$$

It is worthwhile to mention again that the random variable X on which our predictions are based can be vector valued. If $X = (X_1, \ldots, X_n)$, then all the same formulas hold because equality of vectors is equivalent to equality component by component.

Suppose we have two random variables X_1 and X_2 that we would like to use for prediction. We can group them into a vector $X = (X_1, X_2)$ and compute the predictor of Y (say) given X. It is just as easy and perfectly reasonable to call this the predictor of Y given X_1 and X_2.

In general, we can treat several random variables as the components of some implicit vector. If $X_1, \ldots, X_n$ are real-valued random variables, the optimal predictor of Y given $X_1, \ldots, X_n$, $\mathsf{pred}_{Y|X_1,\ldots,X_n}$, is simply the optimal predictor of Y given the vector-valued random variable $X = (X_1, \ldots, X_n)$. This is a purely semantic issue, identical to that which arises in talking about joint distributions. We always want to make it transparently easy to use either vector or component notation.

▶ EXAMPLE 4.9. A Random Sum of Indicators Plus

In the previous example, find the optimal predictor of S given N and I_1.

Let $X = (N, I_1)$; this is an $\mathbb{R}^2$-valued random variable. Then, for $x = (n, i)$, we have

$$\begin{aligned}
\mathsf{pred}_{S|N,I_1}(n,i) &= \mathsf{pred}_{S|X}(n,x) \\
&= \frac{\mathsf{E} S\, 1_{\{X=x\}}}{\mathsf{p}_X(x)} \\
&= \frac{\mathsf{E} S\, 1_{\{N=n \text{ and } I_1=i\}}}{\mathsf{p}_{N,I_1}(n,i)} \\
&= \frac{\mathsf{E} \sum_{j=1}^{N} I_j\, 1_{\{N=n \text{ and } I_1=i\}}}{\mathsf{p}_{N,I_1}(n,i)} \\
&= \frac{\mathsf{E} \sum_{j=1}^{n} I_j\, 1_{\{N=n \text{ and } I_1=i\}}}{\mathsf{p}_{N,I_1}(n,i)} \\
&= \frac{i \mathsf{E} 1_{\{N=n \text{ and } I_1=i\}} + \mathsf{E} \sum_{j=2}^{n} I_j\, 1_{\{N=n \text{ and } I_1=i\}}}{\mathsf{p}_{N,I_1}(n,i)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{i\mathsf{p}_{N,I_1}(n,i) + \sum_{j=2}^{n} \mathsf{E} I_j\, 1_{\{\, N=n \text{ and } I_1=i\,\}}}{\mathsf{p}_{N,I_1}(n,i)} \\
&= \frac{i\mathsf{p}_{N,I_1}(n,i) + \sum_{j=2}^{n} \mathsf{E} I_j\, \mathsf{E}1_{\{\, N=n \text{ and } I_1=i\,\}}}{\mathsf{p}_{N,I_1}(n,i)} \\
&= \frac{i\mathsf{p}_{N,I_1}(n,i) + \sum_{j=2}^{n} \mathsf{E} I_j\, \mathsf{p}_{N,I_1}(n,i)}{\mathsf{p}_{N,I_1}(n,i)} \\
&= i + \sum_{j=2}^{n} \mathsf{E} I_j \\
&= i + (n-1)p.
\end{aligned}
$$

3.3. Prediction Given a Continuous Random Variable

Let X and Y be a continuous, real-valued random variable with joint PDF $\mathsf{f}_{X,Y}$. The argument at the end of subsection 3.1 tells us that

$$
\mathsf{pred}_{Y|X}(u) = \frac{\mathsf{E} Y\, 1_{\{\, X \text{ near } u\,\}}}{\mathsf{P}\{\, X \text{ near } u\,\}}
$$

Our goal in this subsection is to derive the form of this optimal predictor in simpler terms and to confirm that it is optimal.

First, notice that $Y\, 1_{\{\, X \text{near } u\,\}} = h(X,Y)$ where

$$
h(s,t) = \begin{cases} t & \text{if } s \text{ near } u \\ 0 & \text{otherwise.} \end{cases}
$$

Hence,

$$
\begin{aligned}
\mathsf{E} Y\, 1_{\{\, X \text{ near } u\,\}} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(s,t)\, \mathsf{f}_{X,Y}(s,t)\, dt\, ds \\
&= \int_{-\infty}^{\infty} t\, \mathsf{f}_{X,Y}(u,t)\, dt\, du.
\end{aligned}
$$

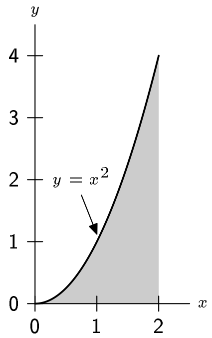
Similarly,

$$
\mathsf{P}\{\, X \text{ near } u\,\} = \mathsf{f}_X(u)\, du = \int_{-\infty}^{\infty} \mathsf{f}_{X,Y}(u,t)\, dt\, du.
$$

It follows that

$$
\begin{aligned}
\mathsf{pred}_{Y|X}(u) &= \frac{\mathsf{E} Y\, 1_{\{\, X \text{ near } u\,\}}}{\mathsf{P}\{\, X \text{ near } u\,\}} \\
&= \frac{\int_{-\infty}^{\infty} t\, \mathsf{f}_{X,Y}(u,t)\, dt\, du}{\int_{-\infty}^{\infty} \mathsf{f}_{X,Y}(u,t)\, dt\, du}
\end{aligned}
$$

18



$$= \frac{\int_{-\infty}^{\infty} t\,\mathsf{f}_{X,Y}(u,t)\,dt\,du}{\mathsf{f}_X(u)\,du}$$

$$= \frac{\int_{-\infty}^{\infty} t\,\mathsf{f}_{X,Y}(u,t)\,dt}{\mathsf{f}_X(u)}$$

$$= \int_{-\infty}^{\infty} t\,\frac{\mathsf{f}_{X,Y}(u,t)}{\mathsf{f}_X(u)}\,dt$$

$$= \int_{-\infty}^{\infty} t\,\mathsf{f}_{Y|X}(t \mid u)\,dt.$$

Several comments are in order. First, the indicator selects only those values of s near u, so the integral over s goes away and leaves the function evaluated at u and a du. Those dus appear in the numerator and denominator eventually cancel. Finally, we move the denominator inside the integral and, in the last line, define

$$\mathsf{f}_{Y|X}(t \mid u) = \frac{\mathsf{f}_{X,Y}(u,t)}{\mathsf{f}_X(u)} = \frac{\mathsf{f}_{X,Y}(u,t)}{\int_{-\infty}^{\infty} \mathsf{f}_{X,Y}(u,t)\,dt},$$

for u such that $\mathsf{f}_X(u) > 0$. For any such u, $\mathsf{f}_{Y|X}(t \mid u)$ is a PDF as a function of t. If $\mathsf{f}_X(u) = 0$, we define $\mathsf{f}_{Y|X}(t \mid u) = 0$. We call $\mathsf{f}_{Y|X}(t \mid u)$ the conditional PDFs of Y given X; we have the identity that

$$\mathsf{f}_{Y|X}(t \mid u)\,dt = \mathsf{P}\{\,Y \text{ near } t \mid X \text{ near } u\,\}.$$

These conditional PDFs are described fully in Section 7 below.

Notice that the last line in the derivation of $\mathsf{pred}_{Y|X}(u)$ above shows that the optimal predictor is a weighted average of the values of Y with respect to the conditional PDF for each u. This looks like an expected value, as indeed it is. See the next section.

▶ EXAMPLE 4.10.

A random point (U, V) is chosen from the shaded region in the figure.[18] The joint PDF of the random variables U and V is given by

$$\mathsf{f}_{U,V}(x,y) = \begin{cases} \frac{x}{4} & \text{for } 0 < x < 2 \text{ and } 0 < y < x^2 \\ 0 & \text{otherwise.} \end{cases}$$

For a given $0 < x < 2$, we want to find the best predictor of V given U. We will first find the conditional PDF $\mathsf{f}_{V|U}(y \mid x)$ and then take a weighted average. We have that

$$\mathsf{f}_{V|U}(y \mid x) = \frac{\mathsf{f}_{U,V}(x,y)}{\mathsf{f}_U(x)}.$$

To find the denominator $\mathsf{f}_U(x)$, we integrate out values of V from the joint PDF. For $0 < x < 2$,

$$\begin{aligned}\mathsf{f}_U(x) &= \int_{-\infty}^{\infty} \mathsf{f}_{U,V}(x,y)\,dy \\ &= \int_0^{x^2} \mathsf{f}_{U,V}(x,y)\,dy \\ &= \int_0^{x^2} \frac{x}{4}\,dy \\ &= \frac{x}{4}\int_0^{x^2} dy \\ &= \frac{x^3}{4}.\end{aligned}$$

For other values of x, $f_U(x) = 0$. As a check, we can see that this PDF integrates to 1.

For $0 < x < 2$ and $0 < y < x^2$, we then have

$$\begin{aligned}\mathsf{f}_{V|U}(y \mid x) &= \frac{\mathsf{f}_{U,V}(x,y)}{\mathsf{f}_U(x)} \\ &= \frac{x/4}{x^3/4} \\ &= \frac{1}{x^2}.\end{aligned}$$

For $0 < x < 2$, $\mathsf{f}_{V|U}(y \mid x) = 0$ if y is outside the range $]0, x^2[$. Hence, given U, the vertical coordinate is uniformly distributed from 0 to the value of U^2.

It follows that

$$\mathsf{pred}_{V|U}(x) = \int_0^{x^2} y\,\frac{1}{x^2}\,dy = \frac{x^4}{2x^2} = \frac{x^2}{2}.$$

▶ EXAMPLE 4.11.

Suppose continuous random variables S and T have joint PDF given by

$$\mathsf{f}_{S,T}(x,y) = \begin{cases} \frac{e^{-x/y}\, e^{-y}}{y} & \text{if } x > 0 \text{ and } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

For $y > 0$, find $\mathsf{pred}_{S|T}(y)$.

We will start by finding $\mathsf{f}_{S|T}(x \mid y)$. For $x, y > 0$,

$$\begin{aligned}
\mathsf{f}_{S|T}(x \mid y) &= \frac{\mathsf{f}_{S,T}(x,y)}{\mathsf{f}_T(y)} \\
&= \frac{\dfrac{e^{-x/y}\, e^{-y}}{y}}{\displaystyle\int_{t=0}^{\infty} \frac{e^{-t/y}\, e^{-y}}{y}\, dt} \\
&= \frac{\dfrac{e^{-x/y}\, e^{-y}}{y}}{e^{-y} \displaystyle\int_{t=0}^{\infty} \frac{1}{y}\, e^{-t/y}\, dt} \\
&= \frac{\dfrac{e^{-x/y}\, e^{-y}}{y}}{e^{-y}} \\
&= \frac{1}{y}\, e^{-x/y}.
\end{aligned}$$

The second-to-last equality follows because the integral in the denominator is the integral of a PDF, and thus one. If $x \le 0$ or $y \le 0$, then $\mathsf{f}_{S|T}(x \mid y) = 0$.

Consequently,

$$\mathsf{pred}_{S|T}(y) = \int_0^{\infty} \frac{x}{y}\, e^{-x/y}\, dx = y.$$

As above, we can in general allow X to be vector-valued. This changes neither the notation nor the meaning, but does mean that all the functions take a multi-dimensional argument.

▶ EXAMPLE 4.12.

A random point (X, Y, Z) is chosen uniformly within the hemisphere $\{(x, y, z)$ such that $x^2 + y^2 + z^2 \le 1$ and $z > 0\}$. Find the best predictor of X given Y and Z.

To write this in vector form, we could define the $\mathbb{R}^2$-valued random variable $W = (Y, Z)$ and consider for $r = (b, c) \in \mathbb{R}^2$,

$$\mathsf{pred}_{X\,|\,W}(r) = \int_{-\infty}^{\infty} a\, \mathsf{f}_{X\,|\,W}(a \mid r)\, da,$$

where, by definition,

$$\mathsf{f}_{X\,|\,W}(a \mid r) = \frac{\mathsf{f}_{X,W}(a, r)}{\mathsf{f}_W(r)} = \frac{\mathsf{f}_{X,Y,Z}(a, b, c)}{\mathsf{f}_{Y,Z}(b, c)},$$

and where

$$\mathsf{f}_{Y,Z}(b, c) = \int_{-\infty}^{\infty} \mathsf{f}_{X,Y,Z}(a, b, c)\, da.$$

Fortunately, there is no need to define the random variable W explicitly; we can simply list the components out.

$$\mathsf{pred}_{X\,|\,Y,Z}(b, c) = \int_{-\infty}^{\infty} a\, \mathsf{f}_{X\,|\,Y,Z}(a \mid b, c)\, da,$$

where, as above,

$$\mathsf{f}_{X\,|\,Y,Z}(a \mid b, c) = \frac{\mathsf{f}_{X,Y,Z}(a, b, c)}{\mathsf{f}_{Y,Z}(b, c)},$$

Since the point is chosen uniformly within the hemisphere we have

$$\mathsf{f}_{X,Y,Z}(a, b, c) = \begin{cases} \frac{2}{3}\,\pi & \text{if } a^2 + b^2 + c^2 \le 1 \text{ and } c > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} \mathsf{f}_{Y,Z}(b, c) &= \int_{-\infty}^{\infty} \mathsf{f}_{X,Y,Z}(a, b, c)\, da \\ &= \int_{-\sqrt{1-b^2-c^2}}^{\sqrt{1-b^2-c^2}} \frac{2}{3}\,\pi\, da \\ &= \frac{4}{3}\,\pi\, \sqrt{1 - b^2 - c^2}. \end{aligned}$$

It follows that

$$\mathsf{f}_{X\,|\,Y,Z}(a \mid b, c) = \frac{1}{2\sqrt{1 - b^2 - c^2}},$$

and

$$\mathsf{pred}_{X\,|\,Y,Z}(b, c) = \int_{-\infty}^{\infty} a\, \mathsf{f}_{X\,|\,Y,Z}(a \mid b, c)\, da$$

19 You can skip this derivation if you like; it is here for the skeptics.

$$= \int_{-\sqrt{1-b^2-c^2}}^{\sqrt{1-b^2-c^2}} a\,\mathsf{f}_{X\mid Y,Z}(a \mid b,c)\,da$$
$$= \frac{1}{2\sqrt{1-b^2-c^2}} \int_{-\sqrt{1-b^2-c^2}}^{\sqrt{1-b^2-c^2}} a\,da$$
$$= \frac{(1-b^2-c^2)}{2\sqrt{1-b^2-c^2}}$$
$$= \frac{1}{2}\sqrt{1-b^2-c^2},$$

as desired

Next, we confirm that this is indeed the optimal preditor[19] For convenience define the function g by $g(u) = \mathsf{pred}_{Y\mid X}(u)$, the best mean-square predictor of Y given X. We need to show that g minimizes $\mathsf{E}(Y - h(X))^2$ over all functions h defined on range(X).

Let h be any function defined on the range of X:

$$\begin{aligned}
\mathsf{E}g(X)h(X) &= \int_{-\infty}^{\infty} h(u)\,g(u)\,\mathsf{f}_X(u)\,du \\
&= \int_{-\infty}^{\infty} h(u) \int_{-\infty}^{\infty} v\,\mathsf{f}_{X,Y}(u,v)\,dv\,du \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(u)\,v\,\mathsf{f}_{X,Y}(u,v)\,dv\,du \\
&= \mathsf{E}Yh(X).
\end{aligned}$$

The second equality follows by cancelling the denominator in the optimal predictor above. Notice that $\mathsf{f}_X(u) = 0$ implies that $\mathsf{f}_{X,Y}(u,v) = 0$, so the cancellation is valid.

Now, write

$$\begin{aligned}
&(Y - h(X))^2 \\
&\quad = (Y - g(X) + g(X) - h(X))^2 \\
&\quad = (Y - g(X))^2 + (g(X) - h(X))^2 + 2(Y - g(X))(g(X) - h(X)).
\end{aligned}$$

Applying the E operator, we have

$$\mathsf{E}(Y-h(X))^2 = \mathsf{E}(Y-g(X))^2 + \mathsf{E}(g(X)-h(X))^2 + 2\mathsf{E}(Y-g(X))(g(X)-h(X)).$$

Consider the last term,

$$\mathsf{E}(Y - g(X))(g(X) - h(X))$$

$$= \mathsf{E}\,Y\,(g(X) - h(X)) - \mathsf{E}g(X)\,(g(X) - h(X))$$
$$= 0,$$

by the result above.[20] Hence,

$$\mathsf{E}(Y - h(X))^2 = \mathsf{E}(Y - g(X))^2 + \mathsf{E}(g(X) - h(X))^2.$$

20 The function $g-h$ plays the role of h in that result.

The first term cannot be affected by choice of h, and the second term, which is non-negative, is minimized by taking $h = g$.

4. Conditional Expected Values and Conditional Probabilities

In the previous sections, we found the form of optimal predictors of Y given X. What is most striking about these predictors is that they are simply long-run averages of Y – but only over those replications whose outcome is consistent with our information about X. This became clearer in the special cases where X and Y were both discrete or both continuous; we found that the expected values could be computed as weighted averages using a particular PMF or PDF as weights – conditional PMFs and PDFs – which reflect our knowledge of Y updated for our information about X.

All of this suggests that we should view the optimal predictors as *conditional* expected values based on the given information. Indeed, we will define

$$\mathsf{E}(Y \mid X \text{ near } u) \equiv \mathsf{pred}_{Y|X}(u). \tag{4.6}$$

The expression on the left hand side is called the *conditional expected value* of Y given that X is near u. The $|$ within the expected value is read "given", and the conditions listed after it describe the information that is taken as known. Although the conditional expected value is just another name for the optimal predictor, there is power in this renaming. The notation reinforces our interpretation of the optimal predictor as a long-run average over selected replications, and as we will see below, it enables us to exploit the properties of expected values in all of the usual manipulations.

By the derivations in the last section, for each value u such that $\mathsf{P}\{\, X \text{ near } u \,\} > 0$, we have

$$\mathsf{E}(Y \mid X \text{ near } u) = \frac{\mathsf{E}Y1_{\{\, X \text{ near } u \,\}}}{\mathsf{P}\{\, X \text{ near } u \,\}}. \tag{4.7}$$

For all other u, $\mathsf{E}(Y \mid X \text{ near } u)$ is arbitrary since such values of X will never be observed; we usually define the conditional expected value to be 0 in these cases because it is convenient to have a function that is well-defined for all values of u. Equation (4.7) works just as well for continuous and discrete random variables. Note, however, that if X is discrete, then X near u if and only if $X = u$, so we can replace near with $=$ in equation (4.7). Hence, in the discrete case, we will often write the conditional expected value as $\mathsf{E}(Y \mid X = u)$.

If X and Y are both discrete or both continuous, then we have an alternative way to compute $\mathsf{E}(Y \mid X \text{ near } u)$, as derived in the last section:

- If X and Y are discrete random variables with joint PMF $\mathsf{p}_{X,Y}$, then

$$\mathsf{E}(Y \mid X \text{ near } u) = \sum_v v\, \mathsf{p}_{Y|X}(v \mid u),$$

where the sum is over all values for which $\mathsf{p}_Y(v) > 0$ and where $\mathsf{p}_{Y|X}(v \mid u) = \mathsf{p}_{X,Y}(u, v)/\mathsf{p}_X(u)$ as defined in the last section. Recall that in the discrete case $\mathsf{E}(Y \mid X \text{ near } u)$ and $\mathsf{E}(Y \mid X = u)$ are interchangeable.

- If X and Y are discrete random variables with joint PMF $\mathsf{p}_{X,Y}$, then

$$\mathsf{E}(Y \mid X \text{ near } u) = \int_{v=-\infty}^{\infty} v\, \mathsf{f}_{Y|X}(v \mid u)\, dv,$$

where $\mathsf{f}_{Y|X}(v \mid u) = \mathsf{f}_{X,Y}(u, v)/\mathsf{f}_X(u)$ as defined in the last section.

Note that, in both cases above, X can be a vector valued random variable here, in which case u is also a vector.

These expressions further justify are using an expected value notation: they are weighted averages of the values of Y with weights being the probabilitity that Y is at or near that value. The only twist is that probabilities are derived from a PMF or PDF that accounts for observed information about X.

▶ EXAMPLE 4.13. More on Biased Coins

In the Biased Coins Example above, Example 4.4, let C be the indicator that the coin with probability 3/4 of coming up heads is chosen. Consider what information X, the number of heads in the first five flips, gives us about C. For integer k from 0 to 5, we have

$$\begin{aligned}
\mathsf{E}(C \mid X = k) &= \frac{\mathsf{E}C1_{\{X=k\}}}{\mathsf{P}\{X = k\}} \\
&= \frac{\mathsf{P}\{C = 1 \text{ and } X = k\}}{\mathsf{P}\{X = k\}} \\
&= \frac{\mathsf{P}\{C = 1 \text{ and } X = k\}}{\mathsf{P}\{C = 1 \text{ and } X = k\} + \mathsf{P}\{C = 0 \text{ and } X = k\}} \\
&= \frac{\frac{1}{2}\binom{5}{k}\left(\frac{3}{4}\right)^k\left(\frac{1}{4}\right)^{5-k}}{\frac{1}{2}\binom{5}{k}\left(\frac{3}{4}\right)^k\left(\frac{1}{4}\right)^{5-k} + \frac{1}{2}\binom{5}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{5-k}} \\
&= \frac{3^k}{3^k + 3^{5-k}}.
\end{aligned}$$

We will revisit this example in later sections.

Two important special cases occur frequently enough to merit a short-hand notation. During an experiment, we often obtain the information that some event $\mathcal{A}$ has occured; this is the information that the random variable $1_{\mathcal{A}}$ takes the value 1, i.e., that $1_{\mathcal{A}} = 1$. Because "$1_{\mathcal{A}} = 1$" is both confusing to read[21] and tedious to write, we define the conditional expected value of Y given an event $\mathcal{A}$ by

$$\mathsf{E}(Y \mid \mathcal{A}) \equiv \mathsf{E}(Y \mid 1_{\mathcal{A}} = 1) = \frac{\mathsf{E}Y1_{\mathcal{A}}}{\mathsf{P}(\mathcal{A})}. \tag{4.8}$$

This is our best guess for Y given the information that the event $\mathcal{A}$ has occured; or equivalently, the long-run average of Y over those replications of the experiment for which $\mathcal{A}$ occurs.

Recall that the expected value of an indicator of an event is, by definition, the probability of that event. If $Y = 1_{\mathcal{B}}$ for some event $\mathcal{B}$, then $\mathsf{E}(Y \mid \mathcal{A})$ should be the probability that $\mathcal{B}$ occurs given that $\mathcal{A}$ has occurred. We define the conditional probability of $\mathcal{B}$ given $\mathcal{A}$ accordingly by

$$\mathsf{P}(\mathcal{B} \mid \mathcal{A}) \equiv \mathsf{E}(1_{\mathcal{B}} \mid \mathcal{A}) = \frac{\mathsf{P}(\mathcal{B} \cap \mathcal{A})}{\mathsf{P}(\mathcal{A})}. \tag{4.9}$$

21 Recall that $\{1_{\mathcal{A}} = 1\}$ means $\{\omega \in \Omega$ such that $1_{\mathcal{A}}(\omega) = 1$ $= \mathcal{A}$.

This is our updated probability of the event $\mathcal{B}$ given the information that the event $\mathcal{A}$ has occured; or equivalently, the long-run proportion of those replications in which $\mathcal{A}$ occurs for which $\mathcal{B}$ also occured. We discuss conditional probabilities in more detail in Subsection 4.2 below.

▶ EXAMPLE 4.14. Center of Mass

A random point (X, Y) is chosen uniformly from within the unit disk in $\mathbb{R}^2$ $\{(x, y)$ such that $x^2 + y^2 \leq 1\}$. Find $\mathsf{E}(X \mid \{Y > 0\})$ and $\mathsf{E}(Y \mid \{Y > 0\})$.[22]

Intuitively, we guess that $\mathsf{E}(X \mid \{Y > 0\}) = 0$, since the upper half-disk is symmetric about 0 in the horizontal coordinate, and we guess that $\mathsf{E}(Y \mid \{Y > 0\})$ will be the center of mass of the upper half-disk. Indeed,

$$\begin{aligned}
\mathsf{E}(X \mid \{Y > 0\}) &= \frac{\mathsf{E}X\, 1_{\{Y>0\}}}{\mathsf{P}\{Y > 0\}} \\
&= 2\mathsf{E}X\, 1_{\{Y>0\}} \\
&= 2\int_{v=0}^{\infty}\int_{u=-\infty}^{\infty} u\, \mathsf{f}_{X,Y}(u, v)\, du\, dv \\
&= 2\int_{v=0}^{1}\int_{u=-\sqrt{1-v^2}}^{\sqrt{1-v^2}} u\, \frac{1}{\pi}\, du\, dv \\
&= 0.
\end{aligned}$$

The second equality occurs because $\mathsf{P}\{Y > 0\} = 1/2$, and the last equality occurs by symmetry of the inner integral for all v.

Similarly,

$$\begin{aligned}
\mathsf{E}(Y \mid \{Y > 0\}) &= \frac{\mathsf{E}Y\, 1_{\{Y>0\}}}{\mathsf{P}\{Y > 0\}} \\
&= 2\mathsf{E}Y\, 1_{\{Y>0\}} \\
&= 2\int_{v=0}^{\infty}\int_{u=-\infty}^{\infty} v\, \mathsf{f}_{X,Y}(u, v)\, du\, dv \\
&= 2\int_{v=0}^{1}\int_{u=-\sqrt{1-v^2}}^{\sqrt{1-v^2}} v\, \frac{1}{\pi}\, du\, dv \\
&= \frac{2}{\pi}\int_{v=0}^{1} v \int_{u=-\sqrt{1-v^2}}^{\sqrt{1-v^2}} du\, dv \\
&= \frac{4}{\pi}\int_{v=0}^{1} v\sqrt{1 - v^2}\, dv \\
&= \frac{4}{3\pi}.
\end{aligned}$$

22 Notice that $\mathsf{E}(Y \mid X$ near $u)$ is equivalent to $\mathsf{E}(Y \mid \{X \text{ near } u\})$. So, we might similarly, just write $\mathsf{E}(Y \mid Y > 0)$ for $\mathsf{E}(Y \mid \{Y > 0\})$. We will stick to the event notation in these notes, but it is permissible to drop the extra braces if they become tedious.

► EXAMPLE 4.13 Revisited.

Above, we computed $\mathsf{P}\{\, C = 1 \mid X = k \,\}$ for $k = 0, \ldots, 5$. Notice that $\mathsf{P}\{\, C = 1 \mid X = k \,\}$ is *not* the same thing as $\mathsf{P}\{\, X = k \mid C = 1 \,\}$. The two quantities are computed on the basis of fundamentally different information: in the former case, we observe how many heads we get and want to predict which coin we are using; in the latter case, we observe which coin we are using and want to predict how many heads we will get.

If we know that we have the 3/4 coin, then by a direct counting argument,[23] we have

$$\mathsf{P}\{\, X = k \mid C = 1 \,\} = \binom{5}{k} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{5-k}.$$

We did not have to compute this by the definition of conditional probability but could instead use the information about what coin we were flipping to reduce it to a simpler problem.[24] Notice that the above expression is very different from the $\mathsf{P}\{\, C = 1 \mid X = k \,\}$ we computed earlier.

Notice also that by the definition of conditional probability

$$\mathsf{P}\{\, X = k \mid C = 1 \,\} \; \mathsf{P}\{\, C = 1 \,\} = \mathsf{P}\{\, C = 1 \mid X = k \,\} \; \mathsf{P}\{\, X = k \,\}.$$

This implies that

$$\mathsf{P}\{\, C = 1 \mid X = k \,\} = \mathsf{P}\{\, X = k \mid C = 1 \,\} \; \frac{\mathsf{P}\{\, C = 1 \,\}}{\mathsf{P}\{\, X = k \,\}}.$$

This inversion relation is what we will call Bayes' rule below.

It is also worth pointing out that the event $\{\, X = k \,\}$ contains outcomes that lead to each value of C. To compute $\mathsf{P}\{\, X = k \,\}$ above, we decomposed $\{\, X = k \,\}$ as follows

$$\{\, X = k \,\} = \{\, X = k \text{ and } C = 1 \,\} \cup \{\, X = k \text{ and } C = 0 \,\},$$

where this union is disjoint. We can then use the definition of conditional probability to compute

$$\begin{aligned}\mathsf{P}\{\, X = k \text{ and } C = 1 \,\} &= \mathsf{P}\{\, X = k \mid C = 1 \,\} \; \mathsf{P}\{\, C = 1 \,\} \\ &= \binom{5}{k} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{5-k} \cdot \frac{1}{2}.\end{aligned}$$

23 See the section on the Binomial distributions in Chapter 6.

24 Each sequence of 5 heads and tails with k heads has probability $p^k\,(1-p)^{5-k}$ by independence. There are exactly $\binom{5}{k}$ such sequences.

4.1. The $\mathsf{E}(\cdot \mid X \text{ near } u)$ Operator

In both interpretation (as a long run average) and computation (as a weighted average of values weighted by probabilities), conditional expected values act in every way like the (unconditional) expected values we used in Chapters 2 and 3, with the only difference is that extra information is assumed known when computing the expected values in the conditional case.

We therefore define an operator $\mathsf{E}(\cdot \mid X \text{ near } u)$. This operator has a single input slot that takes a random variable. If that random variable is real-valued, the operator returns a number; if that random variable is vector-valued, the operator returns a vector of numbers. Specifically, for real-valued Y, $\mathsf{E}(Y \mid X \text{ near } u)$ is a number, and if $Y = (Y_1, \dots, Y_n)$ is $\mathbb{R}^n$-valued,

$$\mathsf{E}(Y \mid X \text{ near } u) = (\mathsf{E}(Y_1 \mid X \text{ near } u), \dots, \mathsf{E}(Y_n \mid X \text{ near } u)).$$

This exactly parallels the definition of the E operator.

The operator $\mathsf{E}(\cdot \mid X \text{ near } u)$ satisfies all the same rules as the unconditional E operator.

- *Constancy Rule.* $\mathsf{E}(1 \mid X \text{ near } u) = 1$.
- *Non-negativity Rule.* If $Y \geq 0$, $\mathsf{E}(Y \mid X \text{ near } u) \geq 0$.
- *Scaling Rule.* $\mathsf{E}(cY \mid X \text{ near } u) = c\mathsf{E}(Y \mid X \text{ near } u)$, for $c \in \mathbb{R}$.
- *Additivity Rule*
$$\mathsf{E}(\sum_{i=1}^{n} Y_i \mid X \text{ near } u) = \sum_{i=1}^{n} \mathsf{E}(Y_i \mid X \text{ near } u).$$
- *Monotone Limits Rule.* For $Y_1 \leq Y_2 \leq \cdots \to Y$,
$$\lim_{n\to\infty} \mathsf{E}(Y_n \mid X \text{ near } u) = \mathsf{E}(Y \mid X \text{ near } u).$$

One way to see these relationships is to notice that, for appropriate u,

$$\mathsf{E}(Y \mid X \text{ near } u) = \mathsf{E}Y \frac{1_{\{X \text{near}\, u\}}}{\mathsf{P}\{X \text{near}\, u\}}.$$

The second random variable in the product on the right-hand side is the same for all arguments to the operator, so the rules for the E operator generally apply. For example,

$$\mathsf{E}(Y_1 + \cdot + Y_n \mid X \text{ near } u) = \mathsf{E}(Y_1 + \cdots + Y_n) \frac{1_{\{X \text{near}\, u\}}}{\mathsf{P}\{X \text{near}\, u\}},$$

and the additivity rule follows from the additivity of E.

Dealing with conditional expected values becomes easy – we can rely on the familiar properties of expected values in every case – when you keep in mind the role of the conditions listed after the given bar. This information is a form of book-keeping; it tells us what is to be taken as known for the calculating that expected value. Specifically, any condition that follows the given bar $|$ is taken as a *known fact* for all the calculations *within* the input slot of the conditional expected value operator.[25] Formally, this manifests itself in an Enhanced Scaling Rule satisfied by the conditional expected value operator:

$$\mathsf{E}(g(X)Y \mid X \text{ near } u) = g(u)\,\mathsf{E}(Y \mid X \text{ near } u), \tag{4.10}$$

for any function g defined on the range of X. Within the operator, the $g(X)$ takes on the value $g(u)$ because X is known, and then this constant can pass outside the expected value by the regular Scaling Rule. This is an extremely potent property.

▶ EXAMPLE 4.15. Independent Random Variables
If Y and X are independent, $\mathsf{E}(Y \mid X \text{ near } u) = \mathsf{E}Y$.[26] An important special case of this is when X is a constant random variable. Since constants are independent of every random variable[27] It follows that, in mathematical terms, conditional expected values as we've defined them are extensions of the idea of an expected value.

4.2. Conditional Probabilities

If $\mathcal{A}$ and $\mathcal{B}$ are two events, we defined conditional probabilities above in equation (4.9) as follows:

$$\mathsf{P}(\mathcal{B} \mid \mathcal{A}) = \frac{\mathsf{P}(\mathcal{B} \cap \mathcal{A})}{\mathsf{P}(\mathcal{A})} = \mathsf{E}(1_{\mathcal{B}} \mid \mathcal{A}).$$

In this subsection, we explore the intuition behind this definition and show a few tricks for working with conditional probabilities.

First, it is important to point out that all probability measures we encounter define the probability of an event to be the (corresponding) expected value of the indicator of that event. This is true for conditional as well as unconditional probabilities. Consequently, as a function of the event $\mathcal{B}$, the conditional probability measure $\mathsf{P}(\cdot \mid \mathcal{A})$ satisfies all

25 In general, the conditions after the given bar cannot be taken as known anywhere outside the operator. They are not, for instance, inherited as common knowledge for any other part of a problem solution. on the same line. We may for example use several different conditional expected values on the same line. The purpose of the book-keeping information in the conditional expected value notation is to help us keep these all straight.

26 If $\mathsf{P}\{\, X \text{ near } u \,\} = 0$, then this is true only if we give the conditional expected values (and similarly the conditional PMFs and PDFs) appropriate definitions for impossible values of X. This is both easy to do and practically irrelevant, so the claim should just taken to be an unqualified fact.

27 No information can ever change our predictions about constants since they are already known.

of the usual properties. For example, $\mathsf{P}(\emptyset \mid \mathcal{A}) = 0$, $0 \leq \mathsf{P}(\mathcal{B} \mid \mathcal{A}) \leq 1$, $\mathsf{P}(\text{compl}\,(\mathcal{B}) \mid \mathcal{A}) = 1 - \mathsf{P}(\mathcal{B} \mid \mathcal{A})$, and

$$\mathsf{P}(\mathcal{B}_1 \cup \mathcal{B}_2 \mid \mathcal{A}) = \mathsf{P}(\mathcal{B}_1 \mid \mathcal{A}) + \mathsf{P}(\mathcal{B}_2 \mid \mathcal{A}) - \mathsf{P}(\mathcal{B}_1 \cap \mathcal{B}_2 \mid \mathcal{A}).$$

Similarly, if a collection of events $\mathcal{B}_i$, are disjoint (mutually exclusive), then

$$\mathsf{P}(\bigcup_i \mathcal{B}_i \mid \mathcal{A}) = \sum_i \mathsf{P}(\mathcal{B}_i \mid \mathcal{A}).$$

Finally, while it is true that $\mathsf{P}(\Omega \mid \mathcal{A}) = 1$, we can state a more stringent rule of Total Probability,

$$\mathsf{P}(\mathcal{A} \mid \mathcal{A}) = 1. \tag{4.11}$$

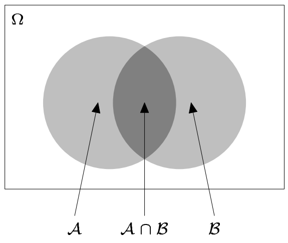
This should not be surprising: if we know for a fact that $\mathcal{A}$ has occurred, then our updated probability of $\mathcal{A}$ must be 1. Indeed, the denominator of $\mathsf{P}(\mathcal{A})$ in the definition of conditional probability is chosen precisely to make this common sense statement true.[28]

The intuition underlying conditional probability is contained in the stylized figure[29] in the margin. Here, we see events $\mathcal{A}$ and $\mathcal{B}$ within Ω, and area relative to Ω in the figure corresponds to probability. If we are told that $\mathcal{A}$ has occurred, then all elementary outcomes outside of $\mathcal{A}$ are immediately irrelevant because they cannot possibly be the chosen elementary outcomes. The only outcomes that can happen in light of this information are those in $\mathcal{A}$, so the only part of $\mathcal{B}$ that can happen is the part of $\mathcal{B}$ within $\mathcal{A}$, that is, $\mathcal{B} \cap \mathcal{A}$. To assess the probability of $\mathcal{B}$ updated for the information that $\mathcal{A}$ has occurred, we use the relative area within $\mathcal{A}$ in the picture because nothing outside $\mathcal{A}$ can happen. The probability $\mathsf{P}(\mathcal{B} \cap \mathcal{A})$, however, is computed within Ω; to make $\mathcal{A}$ the proper frame of reference, we must normalize the unconditional probability $\mathsf{P}(\mathcal{B} \cap \mathcal{A})$ by the unconditional probability $\mathsf{P}(\mathcal{A})$.

One point that often causes confusion is that we are defining the conditional probability in terms of unconditional probabilities. One can only take $\mathcal{A}$ as known to have occurred within the argument of $\mathsf{P}(\cdot \mid \mathcal{A})$; it is not taken as known in the probabilities on the right-hand side of the definition. Notice also that $\mathsf{P}(\mathcal{B} \mid \mathcal{A})$ is not in general equal to $\mathsf{P}(\mathcal{A} \mid \mathcal{B})$; though they share the same intersection, they are normalized differently. See the subsection below on Bayes rule for more on this issue.

28 As we will see below, this normalizing factor has the same interpretation as described earlier with the long-jumper metaphor. $\mathsf{P}(\mathcal{B} \mid \mathcal{A})$ represents the long-run proportion of replications for which $\mathcal{A}$ has occurred in which $\mathcal{B}$ also occurrs. The factor $\mathsf{P}(\mathcal{A})$ is the proportion of all replications for which the precondition holds that $\mathcal{A}$ occurs.

29



► EXAMPLE 4.16. Blackjack!

In a game of blackjack, a player gets A♠ and 9♣ for his first two cards and sees that the dealer has a K♡. Find the probability that the dealer gets a blackjack given this information. (To do this, she needs to have an ace as her other card.)

Let A_P be the number of aces the player receives in his first two cards. Let A_D be the number of aces the dealer receives in his first two cards. With the information provided, the dealer gets Blackjack if he has an ace in his other card. Hence, we want to find $\mathsf{P}\{\, A_D = 1 \mid A_P = 1 \text{ and } A_D \le 1 \,\}$.

$$\begin{aligned}
\mathsf{P}\{\, A_D = 1 \mid A_P = 1 \text{ and } A_D \le 1 \,\} &= \frac{\mathsf{P}\{\, A_D = 1 \text{ and } A_P = 1 \text{ and } A_D \le 1 \,\}}{\mathsf{P}\{\, A_P = 1 \text{ and } A_D \le 1 \,\}} \\
&= \frac{\mathsf{P}\{\, A_D = 1 \text{ and } A_P = 1 \,\}}{\mathsf{P}\{\, A_P = 1 \text{ and } A_D \le 1 \,\}} \\
&= \frac{\mathsf{P}\{\, A_D = 1 \text{ and } A_P = 1 \,\}}{\mathsf{P}\{\, A_D = 0 \text{ and } A_P = 1 \,\} + \mathsf{P}\{\, A_D = 1 \text{ and } A_P = 1 \,\}} \\
&= \frac{48 \cdot 4 \cdot 47 \cdot 3}{48 \cdot 4 \cdot 47 \cdot 3 + 48 \cdot 4 \cdot 47 \cdot 46} \\
&= \frac{3}{49}.
\end{aligned}$$

► EXAMPLE 4.17. Prime Dice

Two fair, six-sided dice are rolled. Assume for convenience that one is red and one is blue.

Let D_r and D_b be the number on the upward face of the red and blue dice respectively. Let $S = D_r + D_b$ be the sum of the dice.

Find $\mathsf{P}\{\, D_r = 6 \mid S \text{ prime} \,\}$, the probability that the red die comes up 6 given that the sum is prime.

$$\begin{aligned}
\mathsf{P}\{\, D_r = 6 \mid S \text{ prime} \,\} &= \frac{\mathsf{P}\{\, D_r = 6 \text{ and } S \text{ prime} \,\}}{\mathsf{P}\{\, S \text{ prime} \,\}} \\
&= \frac{\mathsf{P}\{\, D_r = 6 \text{ and } S \in \{2,3,5,7,11\} \,\}}{\mathsf{P}\{\, S \in \{2,3,5,7,11\} \,\}} \\
&= \frac{\mathsf{P}\{\, D_r = 6 \text{ and } D_b \in \{1,5\} \,\}}{\mathsf{P}\{\, S \in \{2,3,5,7,11\} \,\}} \\
&= \frac{\frac{1}{6} \cdot \frac{1}{3}}{\frac{1}{36} + \frac{2}{36} + \frac{4}{36} + \frac{6}{36} + \frac{2}{36}} \\
&= \frac{\frac{1}{18}}{\frac{15}{36}} \\
&= \frac{2}{15}.
\end{aligned}$$

The denominator in the fourth equality follows by counting configurations of the two dice, all of which are equally likely.

4.2.1. The Essential Conditional Heuristic

There has been a common pattern in the way we have defined conditional quantities in this chapter. For any type of object, be it expected value, probability, PMF, or PDF, we have the following Essential Conditional Heuristic:

$$\text{Conditional} = \frac{\text{Joint}}{\text{Marginal}}.$$

This means that the conditional version of the object (e.g., conditional expected value or conditional PMF) is obtained by normalizing the joint version (that which involves both the predicted and the predicting random variables) by the marginal version (that which involves only the predicting random variable). The definition of conditional probabilities clearly displays the heuristic: the numerator is the probability of both events jointly and the denominator is the probability of the given event only.

The heuristic is a reminder rather than a mnemonic device, but it is useful to keep in mind when you are working with conditionals.

4.2.2. The Multiplication Rule

An equivalent form of the Essential Conditional Heuristic is that Joint = Conditional $\times$ Marginal. We have seen this in the examples above when we write

$$\mathsf{P}(\mathcal{B} \cap \mathcal{A}) = \mathsf{P}(\mathcal{A})\, \mathsf{P}(\mathcal{B} \mid \mathcal{A}).$$

The intuition here is that for both $\mathcal{A}$ and $\mathcal{B}$ to occur, we need $\mathcal{A}$ to occur and then $\mathcal{B}$ to occur given that $\mathcal{A}$ has occurred.[30]

30 This statement is not meant to imply that $\mathcal{A}$ occurs first in time; think of it instead as a checklist.

This relationship generalizes into a useful *Multiplication Rule* for conditional probabilities. If $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are events, then

$$\mathsf{P}(\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_n) = \mathsf{P}(\mathcal{A}_1)\, \mathsf{P}(\mathcal{A}_2 \mid \mathcal{A}_1)\, \mathsf{P}(\mathcal{A}_3 \mid \mathcal{A}_1 \cap \mathcal{A}_2) \cdots \mathsf{P}(\mathcal{A}_n \mid \mathcal{A}_1 \cap \cdots \cap \mathcal{A}_{n-1}). \tag{4.12}$$

This is true because the right-hand side of equation (4.12) can be written as

$$\mathsf{P}(\mathcal{A}_1) \frac{\mathsf{P}(\mathcal{A}_1 \cap \mathcal{A}_2)}{\mathsf{P}(\mathcal{A}_1)} \frac{\mathsf{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3)}{\mathsf{P}(\mathcal{A}_1 \cap \mathcal{A}_2)} \cdots \frac{\mathsf{P}(\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_n)}{\mathsf{P}(\mathcal{A}_1 \cap \cdots \mathcal{A}_{n-1})},$$

and all the extraneous terms cancel.

▶ EXAMPLE 4.18. Charitable Giving

Ten members of a family each decide whether to give to charity at the end of the year by the following rule. The oldest gives with probability 1/2. For every younger family member, if any of the older members gave to charity, the individual gives with probability 0.9; otherwise, the individual gives with probability 0.2.

Let $G_1, \ldots, G_{10}$ be the indicators that the family members, ordered from oldest to youngest, give to charity. Find

$$\mathsf{P}\{\, G_1 = 1 \text{ and } \cdots \text{ and } G_{10} = 1 \,\},$$

the probability that the whole family gives.

$$\begin{aligned}
&\mathsf{P}\{\, G_1 = 1 \text{ and } \cdots \text{ and } G_{10} = 1 \,\} \\
&\quad = \mathsf{P}\{\, G_1 = 1 \,\}\, \mathsf{P}\{\, G_2 = 1 \mid G_1 = 1 \,\} \cdots \\
&\qquad \mathsf{P}\{\, G_{10} = 1 \mid G_1 = 1 \text{ and } \cdots \text{ and } G_9 = 1 \,\} \\
&\quad = \frac{1}{2} \cdot 0.9 \cdots 0.9 \\
&\quad \approx 0.19.
\end{aligned}$$

4.2.3. Bayes Rule

By the definition of conditional probability, we have that

$$\mathsf{P}(\mathcal{A} \mid \mathcal{B})\, \mathsf{P}(\mathcal{B}) \;=\; \mathsf{P}(\mathcal{B} \mid \mathcal{A})\, \mathsf{P}(\mathcal{A}).$$

It follows that

$$\mathsf{P}(\mathcal{B} \mid \mathcal{A}) = \mathsf{P}(\mathcal{A} \mid \mathcal{B})\, \frac{\mathsf{P}(\mathcal{B})}{\mathsf{P}(\mathcal{A})}. \tag{4.13}$$

Hence, $\mathsf{P}(\mathcal{A} \mid \mathcal{B}) = \mathsf{P}(\mathcal{B} \mid \mathcal{A})$ only in the special case where $\mathsf{P}(\mathcal{A}) = \mathsf{P}(\mathcal{B})$.

The relationship in equation (4.13) is the simplest form of what is called *Bayes Rule*. More complicated forms of Bayes Rule are derived from this by decomposing $\mathsf{P}(\mathcal{A})$ into pieces that are easier to compute. For example, because $\mathcal{B}$ and $\operatorname{compl}(\mathcal{B})$ are mutually exclusive and collectively exhaustive, we have

$$\begin{aligned}
\mathsf{P}(\mathcal{A}) &= \mathsf{P}(\mathcal{A} \cap \mathcal{B}) + \mathsf{P}(\mathcal{A} \cap \operatorname{compl}(\mathcal{B})) \\
&= \mathsf{P}(\mathcal{A} \mid \mathcal{B})\, \mathsf{P}(\mathcal{B}) + \mathsf{P}(\mathcal{A} \mid \operatorname{compl}(\mathcal{B}))\, \mathsf{P}(\operatorname{compl}(\mathcal{B})),
\end{aligned}$$

so Bayes Rule above can be written as

$$\mathsf{P}(\mathcal{B} \mid \mathcal{A}) = \frac{\mathsf{P}(\mathcal{A} \mid \mathcal{B})\,\mathsf{P}(\mathcal{B})}{\mathsf{P}(\mathcal{A} \mid \mathcal{B})\,\mathsf{P}(\mathcal{B}) + \mathsf{P}(\mathcal{A} \mid \operatorname{compl}(\mathcal{B}))\,\mathsf{P}(\operatorname{compl}(\mathcal{B}))}.$$

Because the same term appears in the numerator and denominator, this ratio is between 0 and 1 as required.

If the (finite or infinite) sequence of events $\mathcal{B}_1, \mathcal{B}_2, \dots$ are mutually exclusive and collectively exhaustive, we have similarly that

$$\begin{aligned}\mathsf{P}(\mathcal{A}) &= \sum_i \mathsf{P}(\mathcal{A} \cap \mathcal{B}_i)\\ &= \sum_i \mathsf{P}(\mathcal{A} \mid \mathcal{B}_i)\,\mathsf{P}(\mathcal{B}_i).\end{aligned}$$

Thus, again we can rewrite Bayes Rule:

$$\mathsf{P}(\mathcal{B}_i \mid \mathcal{A}) = \frac{\mathsf{P}(\mathcal{A} \mid \mathcal{B}_i)\,\mathsf{P}(\mathcal{B}_i)}{\sum_j \mathsf{P}(\mathcal{A} \mid \mathcal{B}_j)\,\mathsf{P}(\mathcal{B}_j)}.$$

▶ EXAMPLE 4.19. Blood Work

Describe the random experiment. Suppose a blood test is designed to test for the presence of a particular disease. Choose an individual at random from the population and give them the blood test.

Specify your assumptions. The disease inflicts 0.5% of the population. If someone with the disease is tested, the test is positive with probability 0.95. If someone without the disease is tested, the test is positive with probability 0.01.

Define and name the relevant random variables.

Let T be the indicator that the test is positive.

Let D be the indicator that the chosen person has the disease.

State what you know.

$\mathsf{P}\{\, D = 1 \,\} = 0.005$; $\mathsf{P}\{\, T = 1 \mid D = 1 \,\} = 0.95$; and $\mathsf{P}\{\, T = 1 \mid D = 0 \,\} = 0.01$. Notice that $PT = 0 \mid D = 1 = 1 - \mathsf{P}\{\, T = 1 \mid D = 1 \,\}$ and similarly for $\mathsf{P}\{\, D = 0 \,\}$ and $\mathsf{P}\{\, T = 0 \mid D = 0 \,\}$.[31]

State what you want to find.

We want to find $\mathsf{P}\{\, D = 1 \mid T = 1 \,\}$, the probability that the patient has the disease given a positive result on the test.

Find it. By Bayes Rule,

$$\begin{aligned}&\mathsf{P}\{\, D = 1 \mid T = 1 \,\}\\ &\quad = \mathsf{P}\{\, T = 1 \mid D = 1 \,\}\,\frac{\mathsf{P}\{\, D = 1 \,\}}{\mathsf{P}\{\, T = 1 \,\}}\end{aligned}$$

31 Remember that $\mathsf{P}\{\, T = 1 \mid D = 0 \,\}$ is *not* equal to $1 - \mathsf{P}\{\, T = 1 \mid D = 1 \,\}$. Conditional probabilities satisfy the probability rules with respect to the events before the given bar using the same information after the given bar.

$$= \frac{\mathsf{P}\{\, T=1 \mid D=1 \,\}\ \mathsf{P}\{\, D=1 \,\}}{\mathsf{P}\{\, T=1 \mid D=1 \,\}\ \mathsf{P}\{\, D=1 \,\} + \mathsf{P}\{\, T=1 \mid D=0 \,\}\ \mathsf{P}\{\, D=0 \,\}}$$
$$= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} \approx 0.32.$$

Notice that the event $T = 1$, which is the given information, is used in every term on the right hand side. There is no need to consider the event that $T = 0$ since we know that this did not happen. The form of Bayes rule we used here is determined by what we know: we do not know $\mathsf{P}\{\, T = 1 \,\}$ but we do know $\mathsf{P}\{\, T = 1 \mid D = j \,\}$ and $\mathsf{P}\{\, D = j \,\}$ for $j = 0, 1$.

▶ EXAMPLE 4.20.

Describe the random experiment. A bucket contains three dice: one with four sides, one with six, and one with eight. We reach into the bucket, choose one die at random (all three equally likely to be chosen), and roll it.

Specify your assumptions.

The dice are all fairly balanced and are each numbered from one to its number of sides.

Define and name the relevant random variables.

Let D be the value of the die roll. (For a four-sided die, the downward face determines the roll.) Let W be represent which die was chosen, with values 4, 6, or 8.

State what you know.

We know that $\mathsf{P}\{\, W = j \,\} = 1/3$ for $j = 4, 6, 8$, and that

$$\mathsf{P}\{\, D = k \mid W = j \,\} = \begin{cases} \frac{1}{j} & \text{if } 1 \le k \le j \text{ integer} \\ 0 & \text{otherwise.} \end{cases}$$

State what you want to find.

Find the probability that we chose the six-sided die given that the die roll shows k for some $k = 1, \ldots, 8$. This is the conditinal probability $\mathsf{P}\{\, W = 6 \mid D = k \,\}$.

Find it.

We apply Bayes rule as follows:

$$\mathsf{P}\{\, W=6 \mid D=k \,\} = \frac{\mathsf{P}\{\, D=k \mid W=6 \,\}\ \mathsf{P}\{\, W=6 \,\}}{\sum_{j=4,6,8} \mathsf{P}\{\, D=k \mid W=j \,\}\ \mathsf{P}\{\, W=j \,\}}$$
$$= \frac{\mathsf{P}\{\, D=k \mid W=6 \,\}}{\sum_{j=4,6,8} \mathsf{P}\{\, D=k \mid W=j \,\}}$$
$$= \frac{\frac{1}{6}\, 1_{\{\, k \le 6 \,\}}}{\sum_{j=4,6,8} \frac{1}{j}\, 1_{\{\, k \le j \,\}}}.$$

If $k = 7, 8$, for example, we get 0; if $k = 5, 6$, we get $(1/6)/(1/6 + 1/8) \approx 0.57$; if $k \leq 4$, we get $(1/6)/(1/4 + 1/6 + 1/8) \approx 0.31$.

Summary of Section 4

$\mathsf{E}(Y \mid X \text{ near } u)$ is defined to be the optimal (mean square) predictor of Y given X as a function of u. Equivalently, it is the long-run average of Y over replications for which X near u.

$$\mathsf{E}(Y \mid X \text{ near } u) = \frac{\mathsf{E}Y1_{\{X \text{near} u\}}}{\mathsf{P}\{X \text{ near } u\}}$$

for all u for which $\mathsf{P}\{X \text{ near } u\} > 0$. For other values of u, we take $\mathsf{E}(Y \mid X \text{ near } u) = 0$. If X and Y are both discrete or both continuous, we have respectively that

$$\mathsf{E}(Y \mid X \text{ near } u) = \sum_v v\, \mathsf{p}_{Y|X}(v \mid u),$$
$$\mathsf{E}(Y \mid X \text{ near } u) = \int_{v=-\infty}^{\infty} v\, \mathsf{f}_{Y|X}(v \mid u)\, dv.$$

We also have three short-hand notations. For events $\mathcal{A}$ and $\mathcal{B}$, we define

- Conditional Expected Value of Y given $\mathcal{A}$

$$\mathsf{E}(Y \mid \mathcal{A}) \equiv \mathsf{E}(Y \mid 1_{\mathcal{A}} = 1) = \frac{\mathsf{E}Y1_{\mathcal{A}}}{\mathsf{P}(\mathcal{A})}$$

- Conditional Probability of $\mathcal{B}$ given $\mathcal{A}$

$$\mathsf{P}(\mathcal{B} \mid \mathcal{A}) \equiv \mathsf{E}(1_{\mathcal{B}} \mid \mathcal{A}) = \frac{\mathsf{P}(\mathcal{B} \cap \mathcal{A})}{\mathsf{P}(\mathcal{A})}$$

- Vector and Component forms equivalent
  If $X = (X_1, \ldots, X_n)$ and $u = (u_1, \ldots, u_n)$, then:

$$\mathsf{E}(Y \mid X \text{ near } u) \equiv \mathsf{E}(Y \mid X_1 \text{ near } u_1, X_2 \text{ near } u_2, \ldots, X_n \text{ near } u_n).$$

Finally, for any u, $\mathsf{E}(\cdot \mid X \text{ near } u)$ is an operator that takes a single random variables in its input slot and returns a number. This operator satisfies all the rules of an expected value: constancy, non-negativity, scaling, additivity, montone limits. In addition: for any function g, $\mathsf{E}(g(X) \mid X \text{ near } u) = g(u)$ and if Y and X are independent, $\mathsf{E}(Y \mid X \text{ near } u) = \mathsf{E}Y$.

5. Conditional Expected Values as Predictions Given a Random Variable

The mean-square optimal predictor of Y given X is just the function that maps u, a possible value of X, to the number $\mathsf{E}(Y \mid X \text{ near } u)$. This tells us the best prediction for any observed value of X. The mean-square optimal *prediction*, on the other hand, is the random variable that tell us what prediction *we will make in our experiment* when X is eventually observed. We will call this random variable the conditional expected value of Y given X and denote it by $\mathsf{E}(Y \mid X)$.

> **Practical Definition**
>
> The *conditional expected value of* Y *given* X, denoted by $\mathsf{E}(Y \mid X)$, is the *random variable* that gives the mean-square optimal prediction of Y given X. In other words, it is the random variable $\mathsf{pred}_{Y|X}(X)$ or, equivalently, the random variable $g(X)$ where $g(u) = \mathsf{E}(Y \mid X \text{ near } u)$. Formally, as X maps $\omega \in \Omega$ to the number or vector $X(\omega)$, $\mathsf{E}(Y \mid X)$ maps $\omega \in \Omega$ to the number or vector $\mathsf{E}(Y \mid X \text{ near } X(\omega))$.

Computing $\mathsf{E}(Y \mid X \text{ near } u)$ for every u is like coding our prediction algorithm in advance of the experiment. The random variable $\mathsf{E}(Y \mid X)$ is then the result of running that algorithm on the observed value of X: if X near u, then $\mathsf{E}(Y \mid X)$ takes the value $\mathsf{E}(Y \mid X \text{ near } u)$. It is a function of X, and it is a random variable because at the beginning of the experiment, when we do our probabilistic calculations, we do not yet know the value of X. As a random variable, $\mathsf{E}(Y \mid X)$ is a function of X.

Among all notational conventions in the field, there is none more confusing than $\mathsf{E}(Y \mid X)$. It is written as an expected value yet it gives a random variable rather than a number, like all other forms of expected value. It is a function of X but not necessarily a simple one, and that function itself has a tricky notation. It has something after the given bar but not an observable condition, just a sole random variable. However, we will see that when we pull this idea apart there is nothing new or surprising about it. By writing the optimal prediction as a conditional expected value (albeit a strange one), we will also see that we can use it as an extremely powerful tool for making complicated problems simple. Let us begin our exploration with a few simple examples.

▶ EXAMPLE 4.21. Given an Indicator

Let $\mathcal{A}$ be an event and Y a real-valued random variable. Then, the random variable $\mathsf{E}(Y \mid 1_{\mathcal{A}})$ can take only two possible values depending on the value of $1_{\mathcal{A}}$. If $\mathcal{A}$ occurs (i.e., $1_{\mathcal{A}} = 1$), then our prediction will be $\mathsf{E}(Y \mid \mathcal{A})$; otherwise, our prediction will be $\mathsf{E}(Y \mid \text{compl}\,(\mathcal{A}))$. We can write

$$\mathsf{E}(Y \mid 1_{\mathcal{A}}) = \mathsf{E}(Y \mid \mathcal{A})\, 1_{\mathcal{A}} \,+\, \mathsf{E}(Y \mid \text{compl}\,(\mathcal{A}))\, 1_{\text{compl}(\mathcal{A})}.$$

This tells us that $\mathsf{E}(Y \mid 1_{\mathcal{A}})$ has PMF

$$\mathsf{p}_{\mathsf{E}(Y \mid 1_{\mathcal{A}})}(x) = \begin{cases} \mathsf{P}(\mathcal{A}) & \text{if } x = \mathsf{E}(Y \mid \mathcal{A}) \\ 1 - \mathsf{P}(\mathcal{A}) & \text{if } x = \mathsf{E}(Y \mid \text{compl}\,(\mathcal{A})). \end{cases}$$

Such an equation seems odd when you first look at it, but other than the extra book-keeping it is just the PMF of a simple, two-valued random variable. As stated above, we know the values that the random variable takes and the probabilities that it takes those values.

Notice also that

$$\begin{aligned}
\mathsf{E}(\mathsf{E}(Y \mid 1_{\mathcal{A}})) &= \mathsf{E}(\mathsf{E}(Y \mid \mathcal{A})\, 1_{\mathcal{A}} \,+\, \mathsf{E}(Y \mid \text{compl}\,(\mathcal{A}))\, 1_{\text{compl}(\mathcal{A})}) \\
&= \mathsf{E}(Y \mid \mathcal{A})\, \mathsf{E}1_{\mathcal{A}} \,+\, \mathsf{E}(Y \mid \text{compl}\,(\mathcal{A}))\, \mathsf{E}1_{\text{compl}(\mathcal{A})} \\
&= \mathsf{E}(Y \mid \mathcal{A})\, \mathsf{P}(\mathcal{A}) \,+\, \mathsf{E}(Y \mid \text{compl}\,(\mathcal{A}))\, \mathsf{P}(\text{compl}\,(\mathcal{A})) \\
&= \frac{\mathsf{E}(Y\, 1_{\mathcal{A}})}{\mathsf{P}(\mathcal{A})}\, \mathsf{P}(\mathcal{A}) \,+\, \frac{\mathsf{E}(Y\, 1_{\text{compl}(\mathcal{A})})}{\mathsf{P}(\text{compl}\,(\mathcal{A}))}\, \mathsf{P}(\text{compl}\,(\mathcal{A})) \\
&= \mathsf{E}(Y\, 1_{\mathcal{A}}) \,+\, \mathsf{E}(Y\, 1_{\text{compl}(\mathcal{A})}) \\
&= \mathsf{E}Y\, (1_{\mathcal{A}} + 1_{\text{compl}(\mathcal{A})}) \\
&= \mathsf{E}Y.
\end{aligned}$$

Again, this only *looks* strange – with expected values of expected values and so forth – but it is only the expected value derived from the simple PMF above. Note that $\mathsf{E}(Y \mid \mathcal{A})$ and $\mathsf{E}(Y \mid \text{compl}\,(\mathcal{A}))$ are just numbers, so they come outside the outer expected value by the Scaling Rule. The rest follows by the definition of these numbers and the fact that $1_{\mathcal{A}} + 1_{\text{compl}(\mathcal{A})} = 1$. We will have plenty to say about the relationship $\mathsf{E}(\mathsf{E}(Y \mid 1_{\mathcal{A}})) = \mathsf{E}Y$, which is quite general, in the next section.

▶ EXAMPLE 4.7 Revisited.

Draw a random integer N uniformly from $\{0, 1, 2, 3, 4\}$ and then draw a random integer M uniformly between 0 and N. Above, we found that for integers k from 0 to 4,

$$\mathsf{E}(M \mid N = k) = \frac{k}{2}.$$

$\mathsf{E}(M \mid N)$ is the random variable that takes the value $\mathsf{E}(M \mid N = k)$ if and only if $N = k$. Put another way, the random variable $\mathsf{E}(M \mid N)$ can be written as

$$\mathsf{E}(M \mid N) = \sum_{k=0}^{4} \mathsf{E}(M \mid N = k)\, 1_{\{\, N=k \,\}}.$$

This is the prediction we will make when we observe N, whatever it may be.

Substituting in the form of $\mathsf{E}(M \mid N = k)$, we have

$$\mathsf{E}(M \mid N) = \sum_{k=0}^{4} \frac{k}{2}\, 1_{\{\, N=k \,\}} = \frac{N}{2}.$$

So, $\mathsf{E}(M \mid N) = N/2$. $\mathsf{E}(M \mid N)$ is clearly a random varible and a function of N: Whatever N turns out to be in our experiment, our best prediction of M is obtained by dividing N in half.

▶ EXAMPLE 4.9 Revisited.

We derived above that

$$\mathsf{E}(S \mid N = n) = np.$$

It follows that

$$\mathsf{E}(S \mid N) = Np.$$

This is just the random variable that takes the value $np = \mathsf{E}(S \mid N = n)$ when $N = n$. Our best prediction algorithm for S takes whatever value of N we observe and multiplies it by p. Because N is not known at the beginning of the experiment – when we do these calculations – the output of this prediction algorithm is also not known and is consequently a random variable.

▶ EXAMPLE 4.10 Revisited.

We constructed the optimal predictor in this example, showing that

$$\mathsf{E}(V \mid U \text{ near } x) = \frac{x^2}{2}.$$

It follows by definition that

$$\mathsf{E}(V \mid U) = \frac{U^2}{2}.$$

We simply plug the value of U into the function that defines the optimal predictor.

▶ EXAMPLE 4.11 Revisited.

In this example, we showed that

$$\mathsf{E}(S \mid T \text{ near } y) = y,$$

so again,

$$\mathsf{E}(S \mid T) = T,$$

our best prediction of T given the value of T, whatever it may be, is T itself.

These examples may have an unsatisfying similarity about them – after all, all we do each time is compute the optimal predictor and take that function of the predicting random variable – but that is the point. The optimal prediction is derived directly from the optimal predictor; it is not as complicated as it appears. The random variable $\mathsf{E}(Y \mid X)$ takes the value $\mathsf{E}(Y \mid X \text{ near } u)$ if and only if the event $\{X \text{ near } u\}$ occurs, which happens with probability $\mathsf{P}\{X \text{ near } u\}$.

We thus have an immediate characterization of the random variable $\mathsf{E}(Y \mid X)$, expressed as a function of X, and we know its distribution. For instance, if X is discrete, then so is $\mathsf{E}(Y \mid X)$, and it takes the value $\mathsf{E}(Y \mid X = u)$ with probability $\mathsf{p}_X(u)$. If X is continuous, then so is $\mathsf{E}(Y \mid X)$, and its PDF at the value $\mathsf{E}(Y \mid X \text{ near } u)$ is $\mathsf{f}_X(u)$. (In other words, the probability that $\mathsf{E}(Y \mid X)$ is near $\mathsf{E}(Y \mid X \text{ near } u)$ equals $\mathsf{P}\{X \text{ near } u\} = \mathsf{f}_X(u)\,du$.)

This is worth highlighting. The notation here – particularly the E – can make it difficult to remember that $\mathsf{E}(Y \mid X)$ is a random variable, but thinking of this in terms of optimal predictions and predictors helps make the connection clearer.

> **Useful Facts**
>
> • $\mathsf{E}(Y \mid X)$ is a random variable that is a function of X. • If X is discrete, then $\mathsf{E}(Y \mid X)$ is a discrete random variable that takes the value $\mathsf{E}(Y \mid X = u)$ with probability $\mathsf{p}_X(u)$. • If X is continuous, then $\mathsf{E}(Y \mid X)$ is a continuous random variable that is near the value $\mathsf{E}(Y \mid X \text{ near } u)$ with probability $\mathsf{P}\{\, X \text{ near } u \,\}$. In other words, the PDF of $\mathsf{E}(Y \mid X)$ takes the value $\mathsf{f}_X(u)$ at the value $\mathsf{E}(Y \mid X = u)$.

Part of what makes this all confusing is the apparently self-referential notation. After all, one sentence in the previous paragraph could be written as

$$\mathsf{P}\{\, \mathsf{E}(Y \mid X) = \mathsf{E}(Y \mid X = k) \,\} = \mathsf{P}\{\, X = k \,\}.$$

Ahhh! That is enough to knock anyone for a loop. The sentence above is more than sufficient, so you can stick to that. Still, this only looks bad. The key to understanding this is to keep in mind what type of object everything is: $\mathsf{E}(Y \mid X)$ and X are random variables and $\mathsf{E}(Y \mid X = k)$ is a number. The equation could be written without the probabilities as an event equality; it shows that the equivalence of two random variables taking on corresponding values.

The relationship between $\mathsf{E}(Y \mid X)$ and $\mathsf{E}(Y \mid X \text{ near } u)$ is another confusing point. I like to write it in two steps: (i) $g(u) = \mathsf{E}(Y \mid X \text{ near } u)$, (ii) $\mathsf{E}(Y \mid X) = g(X)$. This expresses $\mathsf{E}(Y \mid X)$ as a function of X while keeping it clear what that function is. Resist the temptation to plug in X directly as in $\mathsf{E}(Y \mid X \text{ near } X)$. This is wrong. Remember that for any function h, the random variable $h(X)$ maps $\omega \in \Omega$ to $h(X(\omega))$. Similarly, $\mathsf{E}(Y \mid X)$ maps $\omega \in \Omega$ to $g(X(\omega)) = \mathsf{E}(Y \mid X \text{ near } X(\omega))$, where the $X(\omega)$ is now a number and thus a permissible right-hand side to the near operator.

5.1. A Programmer's Perspective

We motivated the optimal predictor of Y given X as the *algorithm* we use to predict the value of Y given the value of X. Since it is an algorithm, it must work for every possible value of the input X. We then consider the random variable obtained by running the algorithm on the observed value of X. This is the prediction of Y given X.

The key to the usefulness of these ideas is that we can construct the algorithm *before* running the experiment. As a result, we can also consider the distribution of the random variable we get by putting X (whatever it may be) into that algorithm. That is, we can study the properties of the prediction $\mathsf{E}(Y \mid X)$ before we know its value; this is sometimes useful.

Keeping all of these things straight can be very difficult. It may be helpful to see this translated into the terms of a programming language. We have three components. First, the predictor, which is just a function we have written:

```
double predictor_Y_given_X( double u );
```

It takes a single number as input and returns a single number that represents the prediction of Y that results when X is near u. (In the vector-valued X case, the argument to the function would be of type `vector<double> u`.)

Second, we have the distribution of the random variable X, represented by a function

```
double prob_X_near( double u );
```

that gives $\mathsf{P}\{\, X \text{ near } u \,\}$ for each input value u. This too can be computed before the experiment is run.

Third, we have the random experiment which gives us a value for X:

```
double observe_X( );
```

With these three pieces in hand, we can compute anything we have discussed above. For example, the prediction $\mathsf{E}(Y \mid X)$ is obtained by the call:

```
double prediction = predictor_Y_given_X( observe_X() );
```

which is the realized value of a random variable because it depends on the value of X.

In the examples to follow, a useful exercise for understanding conditional expected values is to write the first two functions explicitly, taking appropriate license with the syntax of your favorite programming language.

5.2. A Few of the Finer Points

First, the definition of $\mathsf{E}(Y \mid X \text{ near } u)$ automatically tells us what to do in the cases in which X or Y is a vector. If $Y = (Y_1, \ldots, Y_n)$ is a vector, then

$$\mathsf{E}(Y \mid X) = (\mathsf{E}(Y_1 \mid X), \ldots, \mathsf{E}(Y_n \mid X)),$$

which is an $\mathbb{R}^n$-valued random variable.

If $X = (X_1, \ldots, X_m)$ is a vector, then $u = (u_1, \ldots, u_m)$ is a vector as well, with the same number of components, and nothing changes. It is, however, sometimes convenient to work explicitly with the components instead of with the vector. Then,

$$\mathsf{E}(Y \mid X) \equiv \mathsf{E}(Y \mid X_1, \ldots, X_m).$$

If we define $g(u_1, \ldots, u_m) = \mathsf{E}(Y \mid X_1 \text{ near } u_1, \ldots, X_m \text{ near } u_m)$, then

$$\mathsf{E}(Y \mid X_1, \ldots, X_m) = g(X_1, \ldots, X_m).$$

Second, we always define the probability of an event to be the expected value of the indicator of the event. It therefore makes sense to define the random variable

$$\mathsf{P}(\mathcal{A} \mid X) = \mathsf{E}(1_{\mathcal{A}} \mid X).$$

This is sometimes useful.

Finally, it is worth pointing out[32] that for technical reasons $\mathsf{E}(Y \mid X)$ is defined formally a bit differently than how we defined it above.

32 In other words, you can skip to the next subsection unless the details interest you.

Formal Definition

The *conditional expected value of* Y *given* X, denoted by $\mathsf{E}(Y \mid X)$, is the *random variable* defined as a function of X that satisfies

$$\mathsf{E}Yh(X) = \mathsf{E}\mathsf{E}(Y \mid X)h(X),$$

or equivalently $\mathsf{E}h(X)(Y - \mathsf{E}(Y \mid X)) = 0$, for all functions h defined on range(X). If $\mathsf{E}Y^2 < \infty$, then $\mathsf{E}(Y \mid X)$ equals the optimal prediction of Y given X.

This Formal Definition of $\mathsf{E}(Y \mid X)$ implies the Practical Definition given earlier. For any function g defined on range(X), define $h(u) = g(u) - \mathsf{E}(Y \mid X \text{ near } u)$, so $h(X) = g(X) - \mathsf{E}(Y \mid X)$. Then, by the Formal Definition above,

$$\begin{aligned}\mathsf{E}(Y - g(X))^2 &= \mathsf{E}(Y - \mathsf{E}(Y \mid X) + \mathsf{E}(Y \mid X) - g(X))^2 \\ &= \mathsf{E}(Y - \mathsf{E}(Y \mid X))^2 + \mathsf{E}(\mathsf{E}(Y \mid X) - g(X))^2 + \\ &\quad 2\mathsf{E}(Y - \mathsf{E}(Y \mid X))(\mathsf{E}(Y \mid X) - g(X)) \\ &= \mathsf{E}(Y - \mathsf{E}(Y \mid X))^2 + \mathsf{E}(\mathsf{E}(Y \mid X) - g(X))^2,\end{aligned}$$

which is minimized by taking $g(X) = \mathsf{E}(Y \mid X)$.

The advantage of the Formal Definition is that it is more general and that it makes $\mathsf{E}(Y \mid X)$ a bit easier to analyze when it comes to the technical details. For our purposes, the Practical Definition of $\mathsf{E}(Y \mid X)$ as the optimal prediction is enough.

5.3. The $\mathsf{E}(\cdot \mid X)$ Operator

As we did with $\mathsf{E}(\cdot \mid X \text{ near } u)$, we can define for each random variable X an operator $\mathsf{E}(\cdot \mid X)$. This operator has a single input slot that takes a random variable, and it returns a random variable that is a function of X. The operator applied to a random variable Y gives the random variable $\mathsf{E}(Y \mid X)$.

The operator $\mathsf{E}(\cdot \mid X)$ satisfies all the same rules as the other expected value operators:

- *Constancy Rule.* $\mathsf{E}(1 \mid X) = 1$.
- *Non-negativity Rule.* If $Y \geq 0$, $\mathsf{E}(Y \mid X) \geq 0$.
- *Scaling Rule.* $\mathsf{E}(cY \mid X) = c\mathsf{E}(Y \mid X)$, for $c \in \mathbb{R}$.
- *Additivity Rule*

$$\mathsf{E}(\sum_{i=1}^{n} Y_i \mid X) = \sum_{i=1}^{n} \mathsf{E}(Y_i \mid X).$$

- *Monotone Limits Rule.* For $Y_1 \leq Y_2 \leq \cdots \to Y$,

$$\lim_{n\to\infty} \mathsf{E}(Y_n \mid X) = \mathsf{E}(Y \mid X).$$

Finally, the given bar in $\mathsf{E}(\cdot \mid X)$ tells us to take any random variables after it as *known constants*, but only for the purposes of calculations within the argument of the operator. It follows that any function of X in the argument to the operator can be taken as known and removed as though it were a constant. Put another way, the $\mathsf{E}(\cdot \mid X)$ operator satisfies the following *Enhanced Scaling Rule*:

$$\mathsf{E}(g(X)Y \mid X) = g(X)\,\mathsf{E}(Y \mid X), \tag{4.14}$$

for any function g defined on the range of X. Within the operator, the $g(X)$ is a constant and can be taken outside, but as it passes outside the $\mathsf{E}(Y \mid X)$, it is no longer known and reverts – still as $g(X)$ – to a random variable. As with its analogue with conditional expected values in the last section, this is an extremely potent property.

▶ EXAMPLE 4.22. Independent Argument
If Y and X are independent, then $\mathsf{E}(Y \mid X) = \mathsf{E}Y$.

▶ EXAMPLE 4.23. Symmetric Sum
We roll two standard, balanced six-sided dice, one red and one blue. Let D_r and D_b be the number that shows up on the upward face after the roll for the red and blue dice respectively. Let $S = D_r + D_b$.

We have that

$$\begin{aligned}
\mathsf{E}(S \mid D_r) &= \mathsf{E}(D_r + D_b \mid D_r) \\
&= \mathsf{E}(D_r \mid D_r) + \mathsf{E}(D_b \mid D_r) \\
&= D_r + \mathsf{E}(D_b \mid D_r) \\
&= D_r + \mathsf{E}D_b \\
&= D_r + 3.5.
\end{aligned}$$

Similarly, we know by the symmetry of the problem that $\mathsf{E}(D_r \mid S) = \mathsf{E}(D_b \mid S)$, namely that our predictions of D_r and D_b using S must be the same. It follows then that

$$\begin{aligned}
S &= \mathsf{E}(S \mid S) \\
&= \mathsf{E}(D_r + D_b \mid S) \\
&= \mathsf{E}(D_r \mid S) + \mathsf{E}(D_b \mid S) \\
&= 2\mathsf{E}(D_r \mid S).
\end{aligned}$$

Hence, $\mathsf{E}(D_r \mid S) = S/2$.

6. The Mighty Conditioning Identity

If $\mathsf{E}(Y \mid X)$ is the optimal prediction we *will* make for Y when we observe X and $\mathsf{E}Y$ is the optimal prediction we make for Y before we observe X, then $\mathsf{E}(\mathsf{E}(Y \mid X))$, the optimal prediction (before the experiment) of what our best prediction will be (when we see X), can be no better a prediction than $\mathsf{E}Y$. It turns out, however, that the two predictions are exactly the same: our best prediction of our best prediction is our best prediction. This is the Mighty Conditioning Identity.

The Mighty Conditioning Identity

In its simplest form, the Mighty Conditioning Identity states that

$$\mathsf{E}(\mathsf{E}(Y \mid X)) = \mathsf{E}(Y). \tag{4.15}$$

More generally, for any random variable Z, real or vector valued, we have

$$\mathsf{E}(\mathsf{E}(Y \mid X, Z) \mid X) = \mathsf{E}(Y \mid X). \tag{4.16}$$

We obtain equation (4.15) by taking Z to be a constant random variable in equation (4.16).
$\mathsf{E}(Y \mid X, Z)$ is in general a function of X and Z, but *if* it ends up being a function of X alone, then $\mathsf{E}(Y \mid X, Z) = \mathsf{E}(Y \mid X)$.

The Mighty Conditioning Identity is one of the best tools in our arsenal. It is the basis for the powerful technique of conditioning which is discussed more fully in Chapter 7. Both forms of the identity can be proven using the Formal Definition of conditional expected values given a random variable that was given in the last section. But for our purposes, it suffices to explain equation (4.15) in the case in which X and Y are either both discrete or both continuous.

In the special case where X and Y are both discrete, then $\mathsf{E}(Y \mid X)$ is the random variable that takes the value $\mathsf{E}(Y \mid X = u)$ with probability $\mathsf{p}_X(u)$. Hence,

$$\mathsf{E}(\mathsf{E}(Y \mid X)) = \sum_u \mathsf{E}(Y \mid X = u)\, \mathsf{p}_X(u)$$

$$
\begin{aligned}
&= \sum_u \sum_v v\,\mathsf{p}_{Y|X}(v \mid u)\,\mathsf{p}_X(u)\\
&= \sum_u \sum_v v\,\mathsf{p}_{X,Y}(u,v)\\
&= \sum_v v \sum_u \mathsf{p}_{X,Y}(u,v)\\
&= \sum_v v\,\mathsf{p}_Y(v)\\
&= \mathsf{E}Y,
\end{aligned}
$$

where sums over u are over values of X for which $\mathsf{p}_X(u) > 0$ and sums over v are over values of Y for which $\mathsf{p}_Y(v) > 0$. The cancelation in the third equality follows from the definition of $\mathsf{p}_{Y|X}$.

Similarly, if X and Y are both continuous, then $\mathsf{E}(Y \mid X)$ is the random variable that takes the value $\mathsf{E}(Y \mid X \text{ near } u)$ with infinitesimal probability $\mathsf{P}\{\, X \text{ near } u \,\} = \mathsf{f}_X(u)\,du$. Hence,

$$
\begin{aligned}
\mathsf{E}(\mathsf{E}(Y \mid X)) &= \int_{u=-\infty}^{\infty} \mathsf{E}(Y \mid X \text{ near } u)\,\mathsf{f}_X(u)\,du\\
&= \int_{u=-\infty}^{\infty} \int_{v=-\infty}^{\infty} v\,\mathsf{f}_{Y|X}(v \mid u)\,dv\,\mathsf{f}_X(u)\,du\\
&= \int_{u=-\infty}^{\infty} \int_{v=-\infty}^{\infty} v\,\mathsf{f}_{X,Y}(u,v)\,dv\,du\\
&= \int_{v=-\infty}^{\infty} v \int_{u=-\infty}^{\infty} \mathsf{f}_{X,Y}(u,v)\,du\,dv\\
&= \int_{v=-\infty}^{\infty} v\,\mathsf{f}_Y(v)\,dv\\
&= \mathsf{E}Y,
\end{aligned}
$$

where similar cancellations occur.

▶ EXAMPLE 4.4 and 4.13 Revisited.

In the Biased Coins examples above, if you know the value of C (which coin is being flipped), then X (the number of heads in the first five flips) and Y (the indicator of heads on the sixth flip) are independent.[33] It follows that $\mathsf{E}(Y \mid X, C)$ is a function only of C, hence, by the above $\mathsf{E}(Y \mid X, C) = \mathsf{E}(Y \mid C)$.

Now, $\mathsf{E}(Y \mid C)$ takes the value $\mathsf{E}(Y \mid C = 1)$ with probability $1/2 = \mathsf{P}\{\, C = 1 \,\}$ and $\mathsf{E}(Y \mid C = 0)$ with probability $1/2 = \mathsf{P}\{\, C = 0 \,\}$. But $\mathsf{E}(Y \mid C = 1) = 3/4$ and $\mathsf{E}(Y \mid C = 0) = 1/4$ because Y is the indicator of heads and C indicates the coin. Hence,

$$\mathsf{E}(Y \mid C) = \frac{3}{4}\,C + \frac{1}{4}\,(1 - C).$$

33 X and Y are not independent without this information because, for instance, X gives information about how likely heads is, which can in turn be used to predict Y. We say that X and Y are *conditionally* independent given C. See Section 7.1 below.

We have then, by the Mighty Conditioning Identity, that

$$\begin{aligned}
\mathsf{E}(Y \mid X) &= \mathsf{E}(\mathsf{E}(Y \mid X, C) \mid X) \\
&= \mathsf{E}(\mathsf{E}(Y \mid C) \mid X) \\
&= \mathsf{E}(\frac{3}{4}C + \frac{1}{4}(1-C) \mid X) \\
&= \frac{3}{4}\mathsf{E}(C \mid X) + \frac{1}{4}\mathsf{E}((1-C) \mid X) \\
&= \frac{\frac{3}{4}3^X + \frac{1}{4}3^{5-X}}{3^X + 3^{5-X}}.
\end{aligned}$$

The last step follows by the calculations in Example 4.13 and the fact that $\mathsf{E}((1-C) \mid X) = 1 - \mathsf{E}(C \mid X)$.

7. Conditional Distributions

The distribution of a random variable encodes our state of knowledge about the value that random variable will take when we run the experiment. Put another way, it embodies all our best predictions about that random variable. As we obtain new information about a random variable, our knowledge and our best predictions both change. It makes sense, then, that the distribution of the random variable should change as well. At the beginning of the experiment, before any information is available, the distribution is as we computed it in Chapter 3, but as we get more and more information – by observe the values of other, related random variables – the distribution should get more concentrated, the predictions more accurate. Eventually, when we observe the random variable itself, the distribution becomes perfectly concentrated at the observed value. At that point, we have total knowledge about the value of the random variable.

The techniques we have developed in the previous sections of this chapter comprise the calculus for updating our knowledge for new information. For example, we defined the Conditional PMFs of Y given X, $\mathsf{p}_{Y|X}$, for discrete random variables X and Y by

$$\mathsf{p}_{Y|X}(t \mid s) = \mathsf{P}\{\, Y = t \mid X = s \,\}. \tag{4.17}$$

For each s, this is a PMF as a function of t that represents the distribution of Y updated for the information that $X = s$. Similarly, we defined the Conditional PDFs of Y given X, $\mathsf{p}_{Y|X}$, for continuous random variables X and Y by

$$\mathsf{f}_{Y|X}(t \mid s)\, dv = \mathsf{P}\{\, Y \text{ near } t \mid X \text{ near } s \,\}. \tag{4.18}$$

For each s, $\mathsf{f}_{Y|X}(t \mid s)$ is a PDF as a function of t that represents the distribution of Y updated for the information that X near s.

The conditional PMFs and PDFs are thus representatives of a *conditional distribution* just as the PMF and PDF are representatives of a distribution. Indeed, the conditional PMF and conditional PDF can be defined formally just as we defined the unconditional versions:

Formal Definitions

Let h be any function defined on range(Y).

- If X and Y are both discrete, then the conditional PMFs of Y given X are the functions $\mathsf{p}_{Y|X}(v \mid u)$ that satisfy, for all u,

$$\mathsf{E}(h(Y) \mid X = u) = \sum_v h(v)\, \mathsf{p}_{Y|X}(v \mid u). \tag{4.19}$$

- If X and Y are both continuous, then the conditional PDFs of Y given X are the functions $\mathsf{f}_{Y|X}(v \mid u)$ that satisfy, for all u,

$$\mathsf{E}(h(Y) \mid X \text{ near } u) = \int_{v=-\infty}^{\infty} h(v)\, \mathsf{f}_{Y|X}(v \mid u)\, dv. \tag{4.20}$$

Thus, the conditional PMFs and PDFs tell us how to compute conditional expected values, that is optimal predictors, for any function of Y.

The bottom line is that conditional distributions are just probability distributions, and everything we know about distributions applies to them as well. For instance, we can define conditional CDFs, SFs, PGFs, MGFs, and CGFs in the analogous way

$$\mathsf{F}_{Y|X}(v \mid u) = \mathsf{P}\{\, Y \le v \mid X \text{ near } u \,\}$$
$$\mathsf{S}_{Y|X}(v \mid u) = \mathsf{P}\{\, Y > v \mid X \text{ near } u \,\}$$

$$\mathsf{G}_{Y|X}(z \mid u) = \mathsf{E}(z^Y \mid X \text{ near } u)$$
$$\mathsf{M}_{Y|X}(s \mid u) = \mathsf{E}(e^{sY} \mid X \text{ near } u)$$
$$\mathsf{C}_{Y|X}(t \mid u) = \mathsf{E}(e^{itY} \mid X \text{ near } u),$$

where each generating function is defined only for Y of appropriate type. The only difference with what we are used to is the bookkeeping associated with the information X near u. This can change the value of these quantities, but the meaning of the various functions is exactly the same.

In the same way, we can define the conditional distribution as an operator:

$$\mathsf{D}_{Y|X}(h \mid u) = \mathsf{E}(h(Y) \mid X \text{ near } u),$$

where the operator takes functions h defined on range(Y) and a potential value u of X. The conditional PMF, PDF, and so forth are defined from this operator in exact analogy to the unconditional case.

7.1. Independence

Two random variables are independent if knowing the value of one does not change our best predictions about the other. Thus, X and Y are independent, if our best prediction of $g(X)$ for any function g is unchanged if we know the value of Y, and vice versa for any $h(X)$. That is, X and Y are independent if

$$\mathsf{E}(g(X) \mid Y) = \mathsf{E}g(X)$$

or

$$\mathsf{E}(h(Y) \mid X) = \mathsf{E}h(Y),$$

for any function g defined on range(X) and any function h defined on range(Y).

It follows that if two random variables are independent then

$$\begin{aligned}\mathsf{E}g(X)h(Y) &= \mathsf{E}(\mathsf{E}(g(X)h(Y) \mid X)) \\ &= \mathsf{E}(g(X)\mathsf{E}(h(Y) \mid X)) \\ &= \mathsf{E}g(X)\mathsf{E}h(Y),\end{aligned}$$

which was our provisional definition of independence given in Chapter 2.

Finally, X and Y are independent if and only if the conditional distribution of Y given X equals the unconditional distribution of Y and vice versa. In terms of the conditional PMF and PDF we have, as appropriate,

$$\mathsf{p}_{Y|X}(v \mid u) = \mathsf{p}_Y(v)$$
$$\mathsf{p}_{X|Y}(u \mid v) = \mathsf{p}_X(u)$$

or

$$\mathsf{f}_{Y|X}(v \mid u) = \mathsf{f}_Y(v)$$
$$\mathsf{f}_{X|Y}(u \mid v) = \mathsf{f}_X(u).$$

This follows because under independence we get one of the following:

$$\mathsf{p}_{Y|X}(v \mid u) = \frac{\mathsf{p}_{X,Y}(u,v)}{\mathsf{p}_X(u)} = \frac{\mathsf{p}_X(u)\,\mathsf{p}_Y(v)}{\mathsf{p}_X(u)}$$
$$\mathsf{f}_{Y|X}(v \mid u) = \frac{\mathsf{f}_{X,Y}(u,v)}{\mathsf{f}_X(u)} = \frac{\mathsf{f}_X(u)\,\mathsf{f}_Y(v)}{\mathsf{f}_X(u)}.$$

In the most general form, the equality of the conditional distributions is expressed as follows:

$$\mathsf{D}_{Y|X}(g \mid x) = \mathsf{D}_Y g$$

for all values x in range(X) and all functions g defined on range(Y), or, equivalently,

$$\mathsf{D}_{X|Y}(h \mid y) = \mathsf{D}_X h$$

for all values y in range(Y) and all functions h defined on range(X),

It is possible for two random variables to be dependent, but to become independent if some other variable is known. This is called conditional independence. In the Biased Coins examples above, we saw that X and Y were in general dependent because knowing X helped us predict which coin was chosen originally which in turn helps us predict Y. But once we know C, the coin chosen, X and Y contain no other information about each other. We say that X and Y are conditionally independent given C.

In general, X and Y are conditionally independent given another random variable Z, possibly vector valued, if

$$\mathsf{D}_{Y|X,Z}(g \mid x, z) = \mathsf{D}_{Y|Z}(g \mid z),$$

or equivalently

$$\mathsf{D}_{X|Y,Z}(h \mid y, z) = \mathsf{D}_{X|Z}(h \mid z),$$

for functions g and h defined on range(Y) and range(X) respectively and possible values x, y, z of X, Y, and Z.

VARIANCE

5

Certainty is the mother of quiet and repose,
and uncertainty the cause of variance and contentions
— Sir Edward Coke
from The Institutes of the Laws of England

1. Risk and Uncertainty

The Internet company `ferrets.com` offers three "money-saving" options to each customer who makes a purchase of \$10 or more. At the time of purchase, the customer chooses from one of the three options described in Table 1 below; the value of the randomly chosen reward is then credited or charged to the customer's credit-card.[1] Assuming that you actually wish to purchase a ferret or ferret supplies, consider for a moment which option you would choose in this situation.

Option I		Option II		Option III	
Value	Probability	Value	Probability	Value	Probability
1	1	1000000	0.00001	10	0.600
		-10	0.90000	-1000	0.005
		0	0.09999	0	0.395

Table 1. Distributions of savings in options offered by `ferrets.com`

Why did you choose as you did? What are the differences among the three options?

We can model this situation in terms of the following simple random experiment: the customer chooses an option, a single random number is drawn uniformly from $[0, 1]$, and that number determines[2] the value

1 Assume that, by prior agreement, the credit-card companies will not allow one to stop payment on a loss, even if the purchase is returned.

2 To see how the value is determined, suppose the values and probabilities are enumerated as $v_1, p_1, \ldots, v_k, p_k$ for some integer $k \geq 1$. If u is the value of the uniform random number in $[0, 1]$, then the customer receives v_i if

$$\sum_{j=1}^{i-1} p_j < u < \sum_{j=1}^{i} p_j.$$

Convince yourself that this works.

the customer receives under the chosen option, which is then credited or charged to the customer's credit card.

Define random variables R_I, R_{II}, and R_{III} to be the monetary value (in dollars) received by the customer if the customer had chosen option I, II, or III respectively. Negative values correspond to losses. Table 1 gives the probability distributions of these three random variables. From this information, we can compute the expected values of these random variables, which represent the long-run average values received under each option:

$$\begin{aligned} \mathsf{E}R_I &= 1 = 1 \cdot 1 \\ \mathsf{E}R_{II} &= 1 = 1000000 \cdot 0.00001 - 10 \cdot 0.9 + 0 \cdot 0.09999 \\ \mathsf{E}R_{III} &= 1 = 10 \cdot 0.6 - 1000 \cdot 0.005 + 0 \cdot 0.395. \end{aligned}$$

All three options have the same expected value. This means that if you make many, many, many purchases, you will on average save \$1 per purchase.[3]

So in the sense that the expected value represents the typical return from choosing an option, all three options are equivalent. However, most people – and very probably you as well – would find the three options anything but interchangeable. Option I gives a small but completely certain reward. Option II offers a faint promise of a very large reward but makes a moderate loss very likely. Option III provides a good chance of moderate reward and a small but nontrivial chance of a significant loss.

Put another way, the three options have different levels of risk. The risk of an investment measures the degree to which the actual return on the investment tends to deviate from one's best prediction of the return when the investment is made.[4] We can think of the potential wins and losses that result from the random experiment as the outcome of an investment. There is no risk if one receives exactly the expected return, as in Option I. Risk increases with the probability of the actual return deviating from the expected return and with the size of that deviation. Thus, Option II and Option III both have greater risk than Option I. Which do you think is riskier? In financial terms, a person is said to be risk neutral if they do not need to be paid a premium to assume extra risk. A person is said to be risk averse if they do require a premium,

3 This savings is only helpful if you actually want the items purchased. If instead you make the minimum \$10 purchase just to play the game, then you lose \$9 on average each time.

4 We are taking some liberties here; risk usually refers to deviations from the expected return that result in a loss. However, if a distribution is symmetric about the expected return, as it often can be modeled in practice, then deviations from expected return are just as likely to lead to a loss as a gain. This motivates our use of the term risk.

and the more they would need to be paid, the more risk averse they are.[5] Which category do you fall in? How much would `ferrets.com` have to pay you (or would you pay) to choose either Option II or III?

5 And what do we call someone who would pay to assume more risk? Risk seeking, or perhaps crazy.

Another way to think about the three options is that they offer different degrees of uncertainty about what one will win or lose. Option I has no uncertainty. The uncertainty increases with the probability of values deviate from the expected value and with the size of that deviation. Thus, Option II and Option III both have greater uncertainty than Option I. Which is more uncertain? The probability distribution of a random variable can be understood as our state of knowledge about the values which that random variable will take. The more concentrated the distribution, the more certain our knowledge; the more diffuse the distribution, the more uncertain our knowledge.

As you might have noticed, Risk and Uncertainty are alternative ways of capturing the same idea, but as we have defined them, both are rather abstract. To make these ideas useful, we must be able to quantify them. The *spread* of a distribution is a quantity that describes the Risk and the Uncertainty in that distribution.

In order to capture the ideas of Risk and Uncertainty, any measure of spread must have the following properties:

- If a probability distribution is perfectly concentrated at a single point (what we called a point mass), then it has spread 0.
- The spread of a distribution is a non-negative number in the same units as the random variables with that distribution.
- The spread of a distribution does not depend on where the distribution is centered (i.e., the expected value). Thus, the distributions of the random variables X and $X + c$ have the same spread for every constant c.
- The spread of a distribution scales with constant dilation. In other words, if X is a random variable whose distribution has spread s, then the distribution of the random variable cX has spread cs.

Next, we will turn to defining some specific ways to measure the spread of a distribution.

6 The tails, sometimes called the wings, of the distribution, refers to the values extremely far from the distribution's expected value.

2. Measuring the Spread of a Probability Distribution

One very natural way to measure the spread of a distribution is to determine how far on average a random variable with this distribution deviates from its expected value. That is, we measure the spread of the distribution of a random variable X by $\mathsf{E}|X - \mathsf{E}X|$. Consider this expression from the inside out. The random variable $X - \mathsf{E}X$ has expected value 0; its distribution looks the same as the distribution of X but is simply shifted to be centered on 0. The random variable $|X - \mathsf{E}X|$ gives the distance between the value that X actually takes and its expected value. The expected value $\mathsf{E}|X - \mathsf{E}X|$ tells us what we can expect this difference to be on average – before we observe the value of X. This quantity $\mathsf{E}|X-\mathsf{E}X|$ satisfies all of the properties listed above.

Thought Question Confirm that $\mathsf{E}|X-\mathsf{E}X|$ satisfies the intuitive properties of spread listed above.

Other possibilities also satisfy our intuitive requirements. For example, for any positive integer k, we can define a measure of spread by $\left(\mathsf{E}|X - \mathsf{E}X|^k\right)^{1/k}$. (Notice that this is not the same as $\mathsf{E}|X - \mathsf{E}X|$.) The higher the value of k we choose, the faster $|X - \mathsf{E}X|^k$ grows with $|X - \mathsf{E}X|$, and thus the greater the impact of the "tails" of the distribution[6] on the computed spread.

▶ EXAMPLE 5.1. Ferrets (cont'd)

Consider the spread of the distribution in the Ferrets example using $k = 1, 2, 4$.

X	$\mathsf{E}\|X - \mathsf{E}X\|$	$\sqrt{\mathsf{E}(X - \mathsf{E}X)^2}$	$(\mathsf{E}(X - \mathsf{E}X)^4)^{1/4}$
R_I	0	0	0
R_{II}	20	3162	56234
R_{III}	10.8	71.13	266.18

For example,

$$\mathsf{E}|R_{II} - \mathsf{E}R_{II}| = 1000001 \cdot 0.00001 + 11 \cdot 0.9 + 1 \cdot .09999 = 20.$$

Notice that R_{II} has the greatest spread under all three measures, resulting from the probability, albeit small, assigned to 1000000. Also, as

k grows, the measure of spread gives more and more weight to extreme values of the random variables because they are raised to higher powers.

Alternatively, we can measure spread by considering the distance between two points that contain a fixed amount of probability between them. For $0 < \alpha < 1$, the α quantile of a probability distribution with CDF F is the value q_α such that $F(q_\alpha) = \alpha$. Thus, the 0.25 quantile has one quarter of the probability to the left of it, and the 0.75 quantile has one quarter of the probability to the right of it.[7] The 0.25 and 0.75 quantiles are called the upper and lower *quartiles*. The 0.5 quantile is called the *median*. The $0.1, 0.2, \ldots, 0.9$ quantiles are called *deciles* (DESS-eye-lls). If one looks at a specific inter-quantile range, one obtains a measure of spread. For example, the distance $q_{0.75} - q_{0.25}$ between the upper and lower quartiles of a distribution (called the inter-quartile range) satisfies all of the desired properties. It is unaffected by the very extremes of a distribution. Similarly, for any $0 < \alpha < 1/2$, we could use at $q_{1-\alpha} - q_\alpha$.

7 It is also common to use percentiles instead of quantiles. The 100α percentile is the same as the α quantile; it is just named by a percentage rather than by a fraction.

▶ EXAMPLE 5.2.

Let U have a Uniform$\langle -1, 1\rangle$ distribution. Then, U has CDF

$$\mathsf{F}_U(t) = \begin{cases} 0 & \text{if } t < -1 \\ \frac{t+1}{2} & \text{if } -1 \le t \le 1 \\ 1 & \text{if } t > 1. \end{cases}$$

The α quantile of this distribution is the value $-1 \le q_\alpha \le 1$ that solves $\mathsf{F}_U(q_\alpha) = \alpha$. Hence, $q_\alpha = 2\alpha - 1$. The interquartile range for this distribution is $q_{0.75} - q_{0.25} = 1/2 - (-1/2) = 1$.

Let V have a Uniform$\langle -2, 2\rangle$ distribution. Then, V has CDF

$$\mathsf{F}_V(t) = \begin{cases} 0 & \text{if } t < -2 \\ \frac{t+2}{4} & \text{if } -2 \le t \le 2 \\ 1 & \text{if } t > 2. \end{cases}$$

The α quantile of this distribution is the value $-2 \le q_\alpha \le 2$ that solves $\mathsf{F}_V(q_\alpha) = \alpha$. Hence, $q_\alpha = 4\alpha - 2$. The interquartile range for this distribution is $q_{0.75} - q_{0.25} = 1 - (-1) = 2$.

All of the above are meaningful and useful ways to quantify the spread of a distribution. They each emphasize slightly different features of a distribution. While $\mathsf{E}|X - \mathsf{E}X|$ is an intuitive and direct definition, it suffers from a serious practical problem: in all but the simplest cases, it is difficult to compute. The problem is that absolute values are not amenable to manipulation or simplification. Inter-quantile ranges are also useful in many situations but can be equally difficult to compute and complicated with discrete distributions. What we need is an interpretable definition of spread that is mathematically tractable and easy to compute. This leads us directly to the notion of variance and standard deviation.

3. Standard Deviation and Variance

The most commonly used measure of a distribution's spread is called standard deviation.

Definition. If X is a random variable, then its distribution has *standard deviation* defined by

$$\mathsf{SD}(X) = \sqrt{\mathsf{E}(X - \mathsf{E}X)^2}. \tag{5.1}$$

We usually refer to this – speaking loosely – as "the standard deviation of X", but in fact, it is the standard deviation of the distribution of X.[8] It can be computed before the value of X is known and is the same for all random variables with the same distribution, even if they eventually take different values.

The quantity inside the square root in equation (5.1) is used so often that it is given a name.

Definition. The *variance* of the distribution of X is defined by

$$\mathsf{Var}\,(X) = \mathsf{E}(X - \mathsf{E}X)^2. \tag{5.2}$$

The variance of a distribution is the average *squared* distance between the value of a random variable with that distribution and the expected value of that distribution.[9] Variance is not technically a measure of spread as we defined it above because it is in squared-units. Nonetheless, if one distribution has a higher variance than the other then it has a greater spread.[10] The importance of variance is that it is easy to work with mathematically. Square roots and absolute values, on the other

8 The same sloppiness is evident in the notation for expected value. $\mathsf{E}X$ looks like a function of X but is a function only of the distribution of X. However, we tend to work with random variables, keeping the distributions in the background, so it would become cumbersome to carry around the extra symbol if we were to write something like ED_X. This is particularly true when manipulating expected values. For example, writing $\mathsf{E}(X+Y) = \mathsf{E}X + \mathsf{E}Y$ would have to be written as $\mathsf{ED}_{X+Y} = \mathsf{ED}_X + \mathsf{ED}_Y$. Ahhhhhhhhhhhhh!

9 This was first defined in Chapter 4 as the minimum achievable mean square error of prediction, in the absence of extra information. We will come back to this point below.

10 Put another way: The square root function is monotonically increasing, so it preserves the order of comparisons.

hand, are difficult to work with. So, we will often work with variances with the understanding that one takes a square root of the value to get a proper measure of spread.

As with expected values, we thing of SD and Var as operators. Both SD and Var have a single input slot to be filled by a real-valued random variable,[11] and both return a non-negative number. That standard deviations and variances are non-negative follows from the positivity rule of expected values since $(X - \mathsf{E}X)^2 \geq 0$ for any random variable X. The precedence of the SD and Var operators is the same as for E: they bind more tightly than addition or subtraction, less tightly than multiplication, division, and subtraction.

Example 5.1 gives the standard deviations of R_I, R_{II}, and R_{III}. The variances of these random variables are just the square of the standard deviations 0, 10000129, and 5059. For example, because $\mathsf{E}R_{III} = 1$, we have that

$$\begin{aligned}\mathsf{Var}\,(R_{III}) &= \mathsf{E}(R_{III} - 1)^2 \\ &= 9^2 \cdot 0.6 + 1001^2 \cdot 0.005 + 1^2 \cdot 0.395 \\ &= 5059.\end{aligned}$$

The standard deviation of a random variable X has two interesting interpretations that can be helpful to keep in mind. First, in Chapter 4, we motivated $\mathsf{E}X$ as our best (i.e., minimum mean square error) prediction of X at the beginning of the experiment. In other words, $\mathsf{E}X$ is a good guess for X that will tend to be close on average to the value X takes. We saw that the minimized value of the mean squared error for this prediction is $\mathsf{E}(X - \mathsf{E}X)^2$, the variance. By taking a square root to return this measure of error to the same of units as X,[12] we get the standard deviation. Standard deviation is therefore a measure of how accurately we can guess X without any additional information. If $\mathsf{SD}(X)$ is small, our best guess will be close to the actual value *on average*; if $\mathsf{SD}(X)$ is large, our best guess will be far from the actual value on average.

A second interpretation is that the standard deviation provides a scale by which to compare possible values of X. To see this, consider

11 We can define the variance of a vector valued random variable as described in a note at the end of the chapter. The result is a matrix, and we will not make much use of this case.

12 Which is often called the *root mean squared error*.

the following inequality called Chebychev's inequality

$$\mathsf{P}\{\,|X - \mathsf{E}X| \geq k\mathsf{SD}(X)\,\} \leq \frac{1}{k^2}.$$

That is, no matter what the distribution, as long as it has a finite expected value, we can always bound the amount of probability that is farther than k standard deviations from the expected value. This inequality is derived in a later chapter.

▶ EXAMPLE 5.3. Constants

A constant random variable has a distribution which is perfectly concentrated at the constant value. Hence, any constant has variance zero.

▶ EXAMPLE 5.4. Indicators

Suppose I is an indicator with $\mathsf{E}I = p$. Then, because $I^2 = I$, we have

$$\begin{aligned}
\mathsf{Var}\,(I) &= \mathsf{E}(I - \mathsf{E}I)^2 \\
&= \mathsf{E}(I - p)^2 \\
&= \mathsf{E}(I^2 - 2pI + p^2) \\
&= \mathsf{E}I^2 - 2p\mathsf{E}I + p^2 \\
&= \mathsf{E}I - 2p^2 + p^2 \\
&= p - p^2 \\
&= p(1 - p)
\end{aligned}$$

▶ EXAMPLE 5.5. By the Book

If X is a discrete random variable and Z is a continuous random variable, then

$$\begin{aligned}
\mathsf{Var}\,(X) &= \sum_u (u - \mathsf{E}X)^2\, \mathsf{p}_X(u) \\
\mathsf{Var}\,(Z) &= \int_{v=-\infty}^{\infty} (v - \mathsf{E}Z)^2\, \mathsf{f}_Z(v)\, dv.
\end{aligned}$$

4. Computing Variances

We use the standard deviation instead of $\mathsf{E}|X - \mathsf{E}X|$ because the square term in the variance makes it much easier to compute. Still, computing variances can be a challenge, so it will be helpful to have a few tricks at hand to smooth the way. You should work to remember each of the "rules" below; you will find yourself using them almost every time you compute variances. Most of these follow directly from the intuitive properties of spread.

The variance of any constant is zero. If c is any constant, then $\mathsf{Var}\,(c) = 0$.

From the intuition above, this rule should be clear because a constant takes the same value no matter the outcome of the random experiment. Its value is therefore completely certain and so it's distribution has zero spread, however spread is defined. But to see it formally, notice that $\mathsf{E}c = c$, so $\mathsf{E}(c - \mathsf{E}c)^2 = \mathsf{E}0 = 0$.

Constants taken out of Var *are squared.* If X is any random variable and c is any constant, then $\mathsf{Var}\,(cX) = c^2\mathsf{Var}\,(X)$.

This follows from the properties of spread given above, since $\mathsf{SD}(cX) = c\,\mathsf{SD}(X)$. We can show it directly as follows:

$$
\begin{aligned}
\mathsf{Var}\,(cX) &= \mathsf{E}(cX - \mathsf{E}(cX))^2 \\
&= \mathsf{E}(cX - c\mathsf{E}X)^2 \\
&= \mathsf{E}(c(X - \mathsf{E}X))^2 \\
&= \mathsf{E}c^2(X - \mathsf{E}X)^2 \\
&= c^2\,\mathsf{E}(X - \mathsf{E}X)^2 \\
&= c^2\,\mathsf{Var}\,(X).
\end{aligned}
$$

Variance is unchanged by constant shifts. If X is a random variable and c is any constant, then $\mathsf{Var}\,(X + c) = \mathsf{Var}\,(X)$.

Again, this follows from the properties of spread. Spread measures how concentrated a distribution is, independently of where it is centered. So shifting the center should have no affect on the standard deviation – or the variance. We can show it directly as follows:

$$
\mathsf{Var}\,(X + c) = \mathsf{E}(X + c - \mathsf{E}(X + c))^2
$$

$$\begin{aligned} &= \mathsf{E}(X + c - \mathsf{E}X - c)^2 \\ &= \mathsf{E}(X - \mathsf{E}X)^2 \\ &= \mathsf{Var}\,(X). \end{aligned}$$

A useful shortcut. $\mathsf{Var}\,(X) = \mathsf{E}X^2 - (\mathsf{E}X)^2$.

This form is often much easier to use than the definition itself. To see the equivalence, we simply expand the square, as follows:

$$\begin{aligned} \mathsf{Var}\,(X) &= \mathsf{E}(X - \mathsf{E}X)^2 \\ &= \mathsf{E}(X^2 - 2X\mathsf{E}X + (\mathsf{E}X)^2) \\ &= \mathsf{E}X^2 - \mathsf{E}(2X\mathsf{E}X) + \mathsf{E}(\mathsf{E}X)^2 \\ &= \mathsf{E}X^2 - 2\mathsf{E}X\mathsf{E}X + (\mathsf{E}X)^2 \\ &= \mathsf{E}X^2 - (\mathsf{E}X)^2. \end{aligned}$$

▶ EXAMPLE 5.6.

Let U have a Uniform$\langle -1, 1\rangle$ distribution. Because the distribution of U is symmetric about 0, $\mathsf{E}U = 0$, so $\mathsf{Var}\,(U) = \mathsf{E}U^2$. Hence,

$$\mathsf{Var}\,(U) = \int_{-1}^{1} u^2\,\frac{1}{2}\,du = \frac{1}{3}.$$

▶ EXAMPLE 5.7. Integer-Valued Random Variables

Suppose that K is an integer-valued random variable with PGF G_K. Notice that

$$\mathsf{G}_K''(z) = \sum_{j\in\mathbb{Z}} j(j-1)t^{j-2}\,\mathsf{p}_K(j),$$

and hence

$$\mathsf{G}_K''(1) = \mathsf{E}K(K-1).$$

Since, $\mathsf{E}K(K-1) = \mathsf{E}K^2 - \mathsf{E}K$, we have

$$\mathsf{Var}\,(K) = \mathsf{G}_K''(1) + \mathsf{E}K - (\mathsf{E}K)^2.$$

The variance of a sum of independent random variables is the sum of the variances. If $X_1, \ldots, X_n$ are *independent* random variables, then

$$\mathsf{Var}\,(X_1 + \cdots + X_n) = \mathsf{Var}\,(X_1) + \cdots + \mathsf{Var}\,(X_n).$$

To show that this is true, it is enough to show it for $n = 2$. The more general case follows by a recursive argument (induction). So, suppose that X_1 and X_2 are independent.

$$\begin{aligned}
\mathsf{Var}\,(X_1 + X_2) &= \mathsf{E}(X_1 + X_2 - \mathsf{E}(X_1 + X_2))^2 \\
&= \mathsf{E}(X_1 + X_2 - \mathsf{E}X_1 - \mathsf{E}X_2)^2 \\
&= \mathsf{E}((X_1 - \mathsf{E}X_1) + (X_2 - \mathsf{E}X_2))^2 \\
&= \mathsf{E}(X_1 - \mathsf{E}X_1)^2 + 2\mathsf{E}(X_1 - \mathsf{E}X_1)(X_2 - \mathsf{E}X_2) + \\
&\qquad \mathsf{E}(X_2 - \mathsf{E}X_2)^2 \\
&= \mathsf{Var}\,(X_1) + \mathsf{Var}\,(X_2).
\end{aligned}$$

The term $\mathsf{E}(X_1 - \mathsf{E}X_1)(X_2 - \mathsf{E}X_2)$ is called the covariance of X_1 and X_2 and is discussed in more detail in the next section. However, when X_1 and X_2 are independent,

$$\mathsf{E}(X_1 - \mathsf{E}X_1)(X_2 - \mathsf{E}X_2) = \mathsf{E}(X_1 - \mathsf{E}X_1)\,\mathsf{E}(X_2 - \mathsf{E}X_2) = 0 \cdot 0 = 0.$$

▶ EXAMPLE 5.8. The Binomial Distribution

Let $M = I_1 + \cdots + I_n$, where $I_1, \ldots, I_n$ are independent and identically distributed indicators with $\mathsf{E}I_j = p$ for all j.

Then, by the above,

$$\begin{aligned}
\mathsf{Var}\,(M) &= \mathsf{Var}\,(I_1 + \cdots + I_n) \\
&= \mathsf{Var}\,(I_1) + \cdots + \mathsf{Var}\,(I_n) \\
&= np(1-p).
\end{aligned}$$

Standardization. If X is *any* random variable, the random variable Z defined by

$$Z = \frac{X - \mathsf{E}X}{\mathsf{SD}(X)}$$

has $\mathsf{E}Z = 0$ and $\mathsf{SD}(Z) = 1$.

The operation of standardization is a simple linear transformation of X which shifts and scales it to a standard center (0) and a standard spread (1). To see why it works, notice

$$\mathsf{E}Z = \mathsf{E}\frac{X - \mathsf{E}X}{\mathsf{SD}(X)}$$

13 The expected value of the sample standard deviation, however, is not σ.

$$= \frac{1}{\mathsf{SD}(X)} \mathsf{E}(X - \mathsf{E}X)$$
$$= 0,$$

and

$$\begin{aligned} \mathsf{Var}\, Z &= \mathsf{Var}\, \frac{X - \mathsf{E}X}{\mathsf{SD}(X)} \\ &= \frac{1}{(\mathsf{SD}(X))^2} \mathsf{Var}\,(X - \mathsf{E}X) \\ &= \frac{1}{\mathsf{Var}\, X} \mathsf{Var}\, X \\ &= 1. \end{aligned}$$

Estimating Variance. Suppose $x_1, x_2, \ldots, x_n$ is a list of numbers. Let $\bar{x}$ be the average of the numbers. At some point, you have probably been told that the "sample standard deviation" of those numbers is

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

We are now in a position to understand this.

Let $X_1, \ldots, X_n$ be independent random variables with the *same* distribution. Let $m = \mathsf{E}X_1 = \cdots = \mathsf{E}X_n$ and $\sigma = \mathsf{SD}(X_1) = \cdots = \mathsf{SD}(X_n)$. Define the random variables

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n},$$

the sample average, and

$$s(X_1, \ldots, X_n) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2},$$

the sample standard deviation. Then,

$$\mathsf{E}\bar{X} = m$$
$$\mathsf{E}s^2(X_1, \ldots, X_n) = \sigma^2.$$

In other words, the sample average of the n numbers drawn from a distribution has expected value equal to the expected value of that distribution. Similarly, the sample *variance* of the numbers[13] has expected value equal to the variance of the distribution.

The first of these facts follows from the linearity of expected values:

$$
\begin{aligned}
\mathsf{E}\bar{X} &= \mathsf{E}\frac{X_1 + \cdots + X_n}{n} \\
&= \frac{1}{n}\mathsf{E}(X_1 + \cdots + X_n) \\
&= \frac{1}{n}(\mathsf{E}X_1 + \cdots + \mathsf{E}X_n) \\
&= \frac{1}{n}(m + \cdots + m) \\
&= m.
\end{aligned}
$$

The second is a bit more complicated because each of the terms in the sum involves all n of the random variables. The trick is to add and subtract zero inside the squared expression in each term.

$$
\begin{aligned}
\mathsf{E}s^2 &= \mathsf{E}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n-1}\,\mathsf{E}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n-1}\,\mathsf{E}\sum_{i=1}^{n}\left(\frac{n-1}{n}\,X_i - \frac{1}{n}\sum_{j\neq i} X_i\right)^2 \\
&= \frac{1}{n-1}\,\mathsf{E}\sum_{i=1}^{n}\left(\frac{n-1}{n}\,m - \frac{n-1}{n}\,m + \frac{n-1}{n}\,X_i - \frac{1}{n}\sum_{j\neq i} X_j\right)^2 \\
&= \frac{1}{n-1}\,\mathsf{E}\sum_{i=1}^{n}\left(\frac{n-1}{n}\,(X_i - m) - \frac{1}{n}\sum_{j\neq i}(X_j - m)\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\frac{n-1}{n}\right)^2 \mathsf{Var}\,X_i \;+\; \frac{1}{n^2}\sum_{j\neq i}\mathsf{Var}\,X_j\right] \qquad (*) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(n-1)^2 + (n-1)}{n^2}\,\mathsf{Var}\,X_i \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\frac{(n-1)^2 + (n-1)}{n^2}\,\sigma^2 \\
&= \sigma^2\,\frac{1}{n-1}\,\frac{(n-1)^2 + (n-1)}{n^2}\,n \\
&= \sigma^2\,\frac{(n-1) + 1}{n^2}\,n
\end{aligned}
$$

$$= \sigma^2.$$

The step in equation (*) follows from independence because all of the cross terms, the covariances of X_i and X_j, are zero. The rest is a matter of collecting common terms. While this calculation looks horrible, notice that each step is just a simple and familiar manipulation.

5. Covariances and Correlations

The variance measures the degree to which a distribution varies about its expected value. When we deal with more than one random variable jointly, we also need to consider how the random variables *co-vary*. That is, we consider for each elementary outcome whether the two random variables deviate from their expected values in the same or opposite direction. Loosely speaking, we say

- If the two random variables tend to deviate in the same direction from their expected values, they co-vary positively;
- If the two random variables tend to deviate in the opposite direction from their expected values, they co-vary negatively;
- If the two random variables are equally likely to deviate from their expected values in the same and opposite directions then they do not co-vary.

We first saw such a term arise in the process of computing the variance of a sum above.

Definition. The *covariance* of two random variables X and Y is defined by

$$\mathsf{Cov}(X, Y) = \mathsf{E}(X - \mathsf{E}X)\,(Y - \mathsf{E}Y). \tag{5.3}$$

Consider the random variable $(X - \mathsf{E}X)\,(Y - \mathsf{E}Y)$. If for a particular $\omega \in \Omega$, both $X(\omega)$ and $Y(\omega)$ are above their expected values, then $(X - \mathsf{E}X)\,(Y - \mathsf{E}Y)$ is positive. If, on the other hand, one of $X(\omega)$ and $Y(\omega)$ is above its expected value and one below, then $(X-\mathsf{E}X)\,(Y-\mathsf{E}Y)$ is negative. The covariance between X and Y then averages all of these positives and negatives, weighted by their likelihood as usual. A positive (or negative) covariance, then, implies that X and Y tend to deviate in the same (or opposite) direction from their respective expected values.

The covariance has several properties worth remembering:

- If X and Y are independent, $\mathsf{Cov}(X, Y) = 0$.
  If X and Y are independent, then we can factor the expected value
  $$\mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y) = \mathsf{E}(X - \mathsf{E}X)\,\mathsf{E}(Y - \mathsf{E}Y) = 0.$$
  Intuitively, this means that X and Y have no systematic relationship in their deviations from their expected values.[14]
- $\mathsf{Cov}(X, X) = \mathsf{Var}\, X$.
  In words: the variance of a random variable is the covariance of that random variable with itself.
- $\mathsf{Cov}(X + Z, Y) = \mathsf{Cov}(X, Y) + \mathsf{Cov}(Z, Y)$.
  This follows from: $(X + Z - \mathsf{E}(X + Z)) = (X - \mathsf{E}X) + (Z - \mathsf{E}Z)$.
- $\mathsf{Cov}(aX + b, cY + d) = ac\mathsf{Cov}(X, Y)$.
  The covariance is unaffected by shifts and scales linearly in multiples of each random variable.
- $\mathsf{Cov}(X, Y) = 0$ need not imply that X and Y are independent.
- $-\mathsf{SD}(X) \cdot \mathsf{SD}(Y) \le \mathsf{Cov}(X, Y) \le \mathsf{SD}(X) \cdot \mathsf{SD}(Y)$.
  Because of this bound, we can normalize the covariance into a unitless number between -1 and 1. See correlation below.

We have seen above that the covariance occurs in computing the variance of a sum of random variables. Specifically,

$$\mathsf{Var}\,(X + Y) = \mathsf{Var}\,(X) + 2\,\mathsf{Cov}(X, Y) + \mathsf{Var}\,(Y).$$

The covariance term can make the variance of the sum larger or smaller than the sum of the variances. If X and Y are independent however, $\mathsf{Cov}(X, Y) = 0$, and the variance of the sum is the sum of the variances. In general, if $X_1, \ldots, X_n$ are real-valued random variables

$$\mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{Var}\,(X_i) + 2 \sum_{1 \le i < j \le n} \mathsf{Cov}(X_i, X_j).$$

Thought Question Find two random variables X and Y for which $\mathsf{Var}\,(X) > 0$, $\mathsf{Var}\,(Y) > 0$, and yet $\mathsf{Var}\,(X + Y) = 0$.

14 Note that the converse relationship does not hold: two random variables may have covariance zero but be dependent.

15 The covariance also arises in defining the variance of a vector-valued random variable $X = (X_1, \ldots, X_n)$. We define $\mathsf{Var}\,(X)$ to be the matrix whose i, jth entry is $\mathsf{Cov}(X_i, X_j)$. Thus, $\mathsf{Var}\,(X_i)$s are on the diagonal, and the matrix is non-negative definite. We will not make use of this generalization in this course.

16 $\operatorname{sgn} x$ is -1, 0, or 1 if $x < 0$, $x = 0$, or $x > 0$ respectively.

The correlation between two random variables X and Y is a normalized, unitless number that

Definition. The *correlation* between two random variables X and Y is defined by

$$\mathsf{Cor}(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\mathsf{SD}(X)\,\mathsf{SD}(Y)}. \tag{5.4}$$

The correlation is always between -1 and 1.

The correlation measures the strength of the linear relationship between two random variables. If a random variable Y can be expressed as a linear function of the random variable X, e.g., $Y = aX + b$ for constants a, b, then the two are perfectly correlated, positively or negatively depending on the sign of a. To see this, note that by the properties above, we have that $\mathsf{Cor}(X, aX + b) = \operatorname{sgn} a$.[16] If Y can be accurately, but not perfectly, predicted by a linear function of X, then the correlation will be large in magnitude (close to 1 or -1).

BASIC MANEUVERS

B

1. Sequences and Limits

A *sequence* is an indexed collection of numbers. Think of a sequence as an array that may be infinitely long. As with an array, we use a single symbol adorned with an index to represent numbers in the sequence. For example, we might write a sequence $a_1, a_2, a_3, \ldots$; in this case, the base symbol is a and the indices are in $\mathbb{Z}_+$.

The set of possible indices for a sequence is called the *index set* for that sequence. The index set can be in any set whatsoever. However, we most often will encounter sequences whose indices are in $\mathbb{Z}_+$ or $\mathbb{Z}_\oplus$. Indeed, by default, we will assume that a sequence has index set that is some subset of the integers.

When defining a sequence, we usually specify an arbitrary element. For example, we might say let $a_n = 1/n$ for $n \in \mathbb{Z}_+$ or $b_k = 2^k/k!$. The n and k here are only used to define the sequence; they have no meaning outside that expression. When refering to the sequence as a whole, we put parentheses around the sequence element, as in (a_n), using the index set as a subscript if necessary for clarity, as in $(a_n)_{n\in\mathbb{Z}+}$. The parentheses here are supposed to evoke the parentheses in a vector, and indeed, a sequence can be viewed as a vector with infinitely many components.

For sequences indexed by integers, we are often interested in how the numbers in the sequence behave for very large indices. For $a_n = 1/n$, we know that the sequence elements get smaller and smaller while the sequence $c_n = n$ grows without bound as n gets bigger. In cases like the former, the elements of the sequence get closer and closer to some specific value. We say that the sequence converges to this value and that this value is the limit of the sequence.

A sequence (a_n) is said to *converge* to a value a, denoted $a_n \to a$, if for any positive number ϵ – no matter how small – the sequence elements a_n eventually get within ϵ of a and stay there. That is, for every $\epsilon > 0$, there is an index n_ϵ so that $|a_n - a| < \epsilon$ whenever $n \geq n_\epsilon$. If (a_n) converges to a, we say that a is the limit of the sequence, and we write $a = \lim_{n\to\infty} a_n$. If the sequence does not converge, we say that it *diverges*.

▶ EXAMPLE B.1.

(a) If $s_n = 1$, then $s_n \to 1$.
(b) If $s_n = 1/n^\alpha$, for $\alpha > 0$, then $s_n \to 0$.
(c) If $s_n = (1 + x/n)^n$, then $s_n \to e^x$.
(d) If $s_n = x^{1/n}$ and $x > 0$, then $s_n \to 1$.
(e) If $s_n = x^n$ and $|x| < 1$, then $s_n \to 0$.
(f) If $s_n = n$, then (s_n) diverges.
(g) If $s_n = (-1)^n$, then (s_n) diverges.

There are a few useful facts to keep in mind about limits:

1. Limits are unique: If $a_n \to a$ and $a_n \to a'$, then $a = a'$.
2. If $a_n \to a$ and $b_n \to b$, then $\lim_{n\to\infty}(a_n + b_n) = a + b$.
3. If $a_n \to a$ and $b_n \to b$, then $\lim_{n\to\infty} a_n \cdot b_n = a \cdot b$.
4. If $a_n \to a$ and $a \neq 0$, then $\lim_{n\to\infty} 1/a_n = 1/a$.
5. If $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nk}) \in \mathbb{R}^k$, then $\boldsymbol{x_n} \to \boldsymbol{x}$ if and only if $x_{nj} \to x_j$ for $j = 1, \ldots, k$.

2. Sums

Sums appear so frequently in probability theory that it is useful to master a few tools for dealing with them. This section is an overview only; an excellent and comprehensive introduction the mathematics of sums can be found in *Concrete Mathematics* by R. Graham, D. Knuth, and O. Patashnik.

2.1. The $\sum$-Notation

If I want to add up the numbers 1,17,-3, and 10, I can simply write $1+17+-3+10$. If I want to add up a much larger list of numbers or a variable number of them, it becomes tiresome or impossible to write down the entire list. For example, if we were to add up all the numbers from 1 to 100, we might write

$$1+2+\cdots+100.$$

The $\cdots$'s indicate that the pattern of adding the next number continues.

This $\cdots$ notation can be useful but even this gets tiresome (and sometimes unclear) for complicated sums. The $\sum$-notation is a shorthand representing sums that is very convenient and mimics the integral sign from calculus.

We denote the sum of the $x_1, x_2, \ldots, x_n$ by

$$\sum_{i=1}^{n} x_i.$$

This is equivalent to writing $x_1 + x_2 + \cdots + x_n$; we replace i in x_i with each value from 1 to n in turn and add up the results. The index i is called the "index of summation". It is a variable that is local to the expression inside the $\sum$; it has no meaning outside the summation. The values 1 and n here are called the *lower limit* and *upper limit* of summation respectively.

▸ EXAMPLE B.2.

$$\sum_{i=1}^{100} i = 1+2+\cdots+100 = 100*101/2 = 5050.$$

$$\sum_{i=1}^{10} 1 = 1+1+\cdots+1 = 10.$$

$$\sum_{i=1}^{10} i/5 = (1/5)\sum_{i=1}^{10} i = (1/5)*10*11/2 = 11.$$

In the most common case, the limits of summation will be integers a and b with $a \le b$,

$$\sum_{i=a}^{b} x_i = x_a + x_{a+1} + \cdots + x_{b-1} + x_b$$

We replace i in x_i with each value from a to b in turn (incrementing by 1 each time) and add up the results. However, if $b < a$, then the sum is an *empty sum* and has value 0.

▶ EXAMPLE B.3.

$$\sum_{i=0}^{5} i = 0 + 1 + 2 + 3 + 4 + 5 = 15.$$

$$\sum_{i=-1}^{8} 1 = 1 + 1 + \cdots + 1 = 10.$$

$$\sum_{i=3}^{3} 5e^i = 5e^3.$$

$$\sum_{i=3}^{1} g(i) = 0.$$

Sometimes, we want to sum over a more complicated set than a set of consecutive integers. In this case, we indicate the condition on the index of summation below the $\sum$. The sum over indices in a set $\mathcal{A}$ is denoted by

$$\sum_{i \in \mathcal{A}} x_i.$$

For example,

$$\sum_{i \in \mathbb{Z}_+} 2^{-i} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

Or, if the set is specified by a simple defining condition $C(i)$, we can give that as well

$$\sum_{C(i)} x_i.$$

For example,

$$\sum_{i=1}^{n} x_i = \sum_{1 \le i \le n} x_i,$$

where the index is understood to be an integer in this case.

2.2. A Programming Analogue[1]

Summation can be easily expressed in terms of a for-loop in programming. Here is a summation and a C++ analogue:

$$\sum_{i=a}^{b-1} x_i.$$

```
double sum = 0.0;     // The value of the sum; 0 if empty

for( int i = a; i < b; i++ )
    sum += x[i];
```

1 Skip this subsection if you are not familiar with the C, C++, or Java programming languages.

```
// Now sum contains the value of the summation
// Notice: i is not defined outside the loop
```

The array `x` can contain any values at all, however complicated it was to compute them.

Notice that the index `i` in the for-loop is in scope only within the loop; in other words, it is a dummy variable. If we had called it `j` instead, it would not change the value of the sum. Put another way,

$$\sum_{i=a}^{b-1} x_i = \sum_{j=a}^{b-1} x_j$$

because the two sums add up the same values $x_a, \ldots, x_{b-1}$, just in different orders.

Also, notice from the programming snippet that if the upper index is smaller than the lower index, the sum here is zero just as in the mathematical notation.

2.3. Multiple Sums

We need not restrict ourselves to sums over a single index only.

$$\sum_{C(i,j,\ldots)} x_{i,j,\ldots}.$$

For a simple example, consider

$$\sum_{1\le j,k\le 2} g(j)h(k) = g(1)h(1) + g(1)h(2) + g(2)h(1) + g(2)h(2).$$

In general, we need only one $\sum$ sign even when summing over multiple indices. However, it is often convenient to write a multiple sum as a sum of sums. In this case, we use one $\sum$ sign for each index we are summing over. For example,

$$\sum_{1\le j,k\le 2} g(j)h(k) = \sum_{j=1}^{2}\sum_{k=1}^{2} g(j)h(k).$$

When we use multiple $\sum$ signs, we can exchange the order of summation because addition can be done in any order. If $1_{C(i,j,\ldots)}$ is the indicator that the condition C is true for the given indices, we have

$$\sum_{i\in\mathbb{Z}}\sum_{j\in\mathbb{Z}} a_{ij} 1_{C(i,j)} = \sum_{C(i,j)} a_{ij} = \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} a_{ij} 1_{C(i,j)}.$$

2.4. Some Important Sums

A few sums are particularly important and will arise repeatedly.

The Constant Sum. When we sum a constant value, we get a multiple of that constant.

$$\sum_{k=a}^{b-1} c = (b-a)c.$$

Notice that c does not depend on the index of summation k, thus we call it a "constant" relative to the summation index. A "non-constant" random variable that does not depend on the summation index is still a "constant" relative to the sum. These two uses of the word can be confusing, so look out for it.

$$\sum_{k=0}^{n-1} X = nX.$$

The sum of integers.

$$\sum_{i=0}^{n-1} i = \frac{n(n-1)}{2}.$$

The Finite Geometric Sum. For any z,

$$\sum_{k=a}^{b-1} z^k = \frac{z^a - z^b}{1-z}.$$

To see this, let $S = \sum_{k=a}^{b-1} z^k$. Then, $(1-z)S = S - zS = z^a - z^b$, because the other terms cancel, and S is as given above.

The Infinite Geometric Sum. When $|z| < 1$, then $z^b \to 0$ as $b \to \infty$, and the finite geometric sum converges to

$$\sum_{k=a}^{\infty} z^k = \frac{z^a}{1-z}.$$

The Harmonic Sum. The harmonic numbers H_n are given by

$$H_n = \sum_{k=1}^{n} \frac{1}{k}.$$

For large n,

$$H_n = \ln n + \gamma + \frac{1}{2n} + r_n,$$

where $\gamma = 0.5772156649\ldots$ is Euler's constant and $|r_n| \le 1/n^2$. In particular, $H_n \to \infty$ as $n \to \infty$.

The harmonic numbers represent a discrete analogue of the logarithm. For example,

$$\sum_{k=0}^{n-1} H_k = nH_n - n$$
$$\sum_{k=0}^{n-1} kH_k = \frac{n(n-1)}{2}\left(H_n - \frac{1}{2}\right),$$

which mimic the corresponding integrals of logarithms.

Sums with Binomial Coefficients. There are many, many identities involving binomial coefficients (see below), but the most important is the Binomial Theorem

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k\, y^{n-k}.$$

In particular,

$$2^n = \sum_{k=0}^{n} \binom{n}{k}$$

where $x = 1 = y$.

2.5. A Few Useful Tricks

There are some common tricks for manipulating sums.

The Distributive Rule. Removing constants with respect to the summation index. If all the expression being summed are multiplied by a value that does not depend on the summation index, that value can be taken out of the sum to multiply the final value:

$$\sum_{j=a}^{b} c\,x_j = c\sum_{j=a}^{b} x_j.$$

Shifting the summation. Because the index of summation is a dummy variable, we can replace it with any other convenient dummy variable that leads us to sum exactly the same values. Thus,

$$\sum_{k=a}^{b} x_k = \sum_{k=a+j}^{b+j} x_{k-j},$$

for any j. We have shifted the limits and compensated by shifting the index. Check for yourself that this leads to the same expressions being added up.

For example,

$$\sum_{k=2}^{\infty} z^{k-2} = \sum_{k=0}^{\infty} z^k.$$

This can be useful for reducing a complicated summation to something we know how to do.

We can also use any transformation of the summation index that preserves the terms we are adding up. For example,

$$\sum_{k=0}^{48} \binom{48-k}{7} = \sum_{k=0}^{48} \binom{k}{7}.$$

Removing Factors. This is just a special case of #1 above. If we have terms of the form z^k for any z, we can always remove some factors of z from every term. Thus,

$$\sum_{k=10}^{\infty} z^k = \sum_{k=10}^{\infty} z^{k-10} z^{10}.$$

Now apply the above rules to get

$$\ldots = z^{10} \sum_{k=10}^{\infty} z^{k-10} = z^{10} \sum_{k=0}^{\infty} z^k.$$

Removing the tails. Sometimes it helps to write out the end cases explicitly:

$$\sum_{k=0}^{\infty} z^k = 1 + \sum_{k=1}^{\infty} z^k.$$

We can extract any individual terms from the summation. We can then work with the extracted terms individually and reduce the remaining sum to simpler form.

3. Integer Functions

If n is a non-negative integer, then $n!$ (read "n factorial") is defined by

$$n! = n(n-1)\cdots 1$$

where $0! = 1$ because the product is empty. The integer $n!$ represents the number of ways to permute n items. (There are n ways to choose the position of the first item, $n-1$ ways to choose the position of the second, and so forth.)

A generalization of the factorial function are the factorial powers. There are two varieties: the falling factorial powers $x^{\underline{k}}$ and the rising factorial powers $x^{\overline{k}}$ defined for integers k by

$$x^{\underline{k}} = \begin{cases} x(x-1)\cdots(x-k+1) & \text{for integer } k \geq 0 \\ \dfrac{1}{(x+1)(x+2)\cdots(x+|k|)} & \text{for integer } k < 0. \end{cases}$$

and

$$x^{\overline{k}} = \begin{cases} (x+k-1)\cdots(x+1)x & \text{for integer } k \geq 0 \\ \dfrac{1}{(x-1)(x-2)\cdots(x-|k|)} & \text{for integer } k < 0. \end{cases}$$

In general, the two varieties are related by

$$x^{\overline{k}} = (x+k-1)^{\underline{k}}.$$

We have $x^{\underline{0}} = x^{\overline{0}} = 1$ and $n! = n^{\underline{n}} = 1^{\overline{n}}$.

The *floor* and *ceiling* functions map a real number x to a nearby integer. The floor of x, denoted by $\lfloor x \rfloor$ is the largest integer that is less than or equal to x. For example, $\lfloor 1.5 \rfloor = 1 = \lfloor 1 \rfloor$ and $\lfloor -0.25 \rfloor = -1$. The ceiling of x, denoted by $\lceil x \rceil$, is the smallest integer that is greater than or equal to x. For example, $\lceil 1.5 \rceil = 2 = \lceil 2 \rceil$ and $\lceil -0.25 \rceil = 0$. Notice that if k is an integer, then $k = \lfloor k \rfloor = \lceil k \rceil$.

4. Binomial Coefficients

The *binomial coefficients* are a collection of special numbers that arise frequently in probability and discrete mathematics. The name "binomial coefficients" comes from the celebrated "Binomial Theorem" which re-expresses the power of a binomial (two variable) expression as follows:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \, y^{n-k},$$

for any numbers x and y and integer power $n \geq 0$. The numbers $\binom{n}{k}$ multiplying $x^k \, y^{n-k}$ is a binomial coefficient.

For any real number r and any *integer* k, the binomial coefficient $\binom{r}{k}$ (read "r choose k") is defined by

$$\binom{r}{k} = \begin{cases} \dfrac{r^{\underline{k}}}{k!} = \dfrac{r(r-1)\,\cdots\,(r-k+1)}{k!} & \text{if } k \geq 0 \\ 0 & \text{if } k < 0. \end{cases}$$

Three special cases are worth remembering:

$$\binom{r}{0} = 1 \qquad \binom{r}{1} = r \qquad \binom{r}{2} = \frac{r(r-1)}{2}.$$

The first of these follows from the fact that $r^{\underline{0}} = 1$. In general, notice that $\binom{r}{k}$ is a kth degree polynomial.

For us, the most important case by far is when $r = n$, a non-negative integer. Then,

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

As strange as it might seem, the binomial coefficient will always be an integer in this case.

The binomial coefficients satisfy a vast array of useful identities. For $n \in \mathbb{Z}_{\oplus}$ and $r \in \mathbb{R}$, the following arise quite frequently.

Binomial Coefficient Identities

$$\binom{n}{k} = \binom{n}{n-k}$$

$$\binom{-1}{k} = (-1)^k$$

$$\binom{r}{k} = \binom{r-1}{k-1} + \binom{r-1}{k}$$

$$\binom{r}{k} = \frac{r}{k}\binom{r-1}{k-1}$$

$$\binom{r}{k} = (-1)^k\binom{k-r-1}{k}$$

$$\sum_{k=0}^{n}\binom{n}{k} x^k\, y^{n-k} = (x+y)^n$$

$$\sum_{k\le n}\binom{r+k}{k} = \binom{r+n+1}{n}$$

$$\sum_{0\le k\le n}\binom{k}{m} = \binom{n+1}{m+1}$$

$$\sum_{k\in\mathbb{Z}}\binom{r}{k}\binom{s}{n-k} = \binom{r+s}{n}$$

Another useful feature of binomial coefficients is that every polynomial $p(x)$ of degree n can be written in the so-called *Newton Form*

$$p(x) = \sum_{k=0}^{n} c_k \binom{x}{k}.$$

The binomial coefficients have several useful interpretations that arise frequently in sampling problems. The most important one is that $\binom{n}{k}$ is the number of subsets of size k from a set of n elements. This idea can be expressed in several equivalent ways.

Interpretations of Binomial Coefficients

For integers n and k, with $n \geq 0$,

- $\binom{n}{k}$ is the number of subsets of size k contained in a set of size n.
- $\binom{n}{k}$ is the number of binary bit strings of length n with exactly k bits set to 1.
- $\binom{n}{k}$ is the number of ways to fill exactly k slots out of n, each with a single item.

Many of the binomial coefficient identities given above can be thought of directly in terms of these interpretations. For example, by the Binomial Theorem,

$$2^n = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n},$$

which decomposes the total number of subsets into the counts of subsets of each size.

To see where these interpretations come from, recall that $n!$ is the number of permutations of n items. Given this, consider how to count the number of subsets with cardinality k from a set with cardinality n. Here's an initial algorithm for generating a subset: List out the elements of the full set in some order and use the first k elements as the subset.

This procedure works well: in $n!$ possible lists, it generates all possible subsets of size k. Its problem is that the algorithm generates each subset multiple times. Any listing of the elements of the full set with the same first k elements (and same last $n - k$ elements) corresponds to the same subset. How many such listings are there per subset? Each listing corresponds to one of $k!$ permutations of the first k elements listed and one of $(n-k)!$ permutations of the last $n-k$ elements listed. Hence, each subset is listed $k!(n-k)!$ times. Thus, there are $n!/k!(n-k)!$ subsets of size k.

5. Counting and Sampling

Counting problems arise in probability theory in the following basic scenarios:

> We have a collection of n objects from which we draw a sample of k objects in such a way that every possible sample is equally likely. We want to find the probability that our sample has certain properties.

Suppose the event $\mathcal{A}$ represents the event that the sample has some pre-specified property (e.g., a flush in cards, no defective chips, etc.). Then, because every specific sample is equally likely to occur and has probability

$$\frac{1}{\text{Total \# of possible samples}},$$

it follows that

$$\mathsf{P}(\mathcal{A}) = \frac{\text{\# of samples represented in } \mathcal{A}}{\text{Total \# of possible samples}}.$$

Such problems therefore reduce to problems of counting configurations.

There are different sampling schemes and different ways to count. We can sample with or without replacement and we can count with or without order. All four combinations are possible, as shown below.

Replacement. If we sample with replacement, we "put each item back" after we sample it. If we sample without replacement, we do not. In other words, when sampling with replacement, each item can be sampled more than once; when sampling without replacement, each item can be sampled at most once. To determine which scheme applies in any given situation, you can thus ask yourself: is it possible to get repetitions with this sampling method?

Order. If we count *with order*, we distinguish—count as separate—samples with the same items that were sampled in a different order. If we count *without order*, a distinct sample is determined only by which objects it contains.

The key distinction here is between a *vector* and a *set*. In a vector, we pay attention to which element is first, which is second, etc.. In a set, we only pay attention to which elements are members.

Counting with orders is equivalent to counting the vectors $(x_1, x_2, \ldots, x_k)$ where x_i is the label of the item sampled on the ith draw. Note that (a, b, c, d) and (b, a, c, d) are differen vectors.

Counting without order is equivalent to counting the sets $\{x_1, x_2, \ldots, x_k\}$ where x_i is the label of one of the items sampled.

▶ EXAMPLE B.4.

Here are samples of size 2 from 4 items A, B, C, and D sampled/counted in different ways.

Without order, without replacement:

AB, AC, AD, BC, BD, CD

With order, without replacement:

AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC

With order, with replacement:

AA, AB, AC, AD, BA, BB, BC, BD,
CA, CB, CC, CD, DA, DB, DC, DD

Without order, with replacement:

AA, AB, AC, AD, BB, BC, BD, CC, CD, DD

In general, the number of samples of size k from a batch of n objects is as follows:

	With Replacement	Without Replacement
With Order	n^k	$\dfrac{n!}{(n-k)!}$
Without Order	$\dbinom{n+k-1}{k}$	$\dbinom{n}{k}$

The following mnemonics can help here. The number of vectors $(x_1, \ldots, x_k)$ where every x_i can be $1, \ldots, n$ is just n^k items. If we restrict repetitions, then x_2 can be any one of $n-1$ items, x_3 can be any of $n-2$ items, giving us $n(n-1)\cdots(n-k+1)$. These are the top two rows in the table.

We can compute the number of ways to choose k items out of n with replacement and without order by a cute mapping. Each such sample corresonds to exactly one arrangements of k x's and $n-1$ /'s. To see this, imagine a group of people selecting items from a menu with n items: the /'s separate adjacent menu items and each x marks a single selection of the corresponding ite. For example, the arrangement $xx//x/xxx/x$ has 2 people selecting item 1, nobody selecting item 2, 1 person selecting item 3, 3 people selecting item 4, and 1 person selecting item 5. There are $\binom{n-1+k}{k}$ ways of placing the k x's in the $n-1+k$ slots.

Finally, there are $\binom{n}{k}$ subsets of size k with each subset corresonding to exactly one unordered samples without replacement.

FUNCTIONS

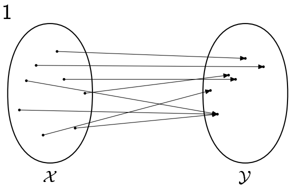
1. Fundamentals

Suppose that we have two non-empty sets $\mathcal{X}$ and $\mathcal{Y}$ (keep in mind that these two sets may be equal). A *function* that maps $\mathcal{X}$ to $\mathcal{Y}$ is rule that associates to each element $x \in \mathcal{X}$ one and only one element $y \in \mathcal{Y}$. (Functions are sometimes called mappings to emphasize this association between the two sets.) The figure[1] shows a "conceptual icon" to illustrate this idea. As usual, we like to give objects names. For example, we might say that f is a function. To indicate that f maps $\mathcal{X}$ to $\mathcal{Y}$ we write[2] $f\colon \mathcal{X} \to \mathcal{Y}$. When speaking, we read "$f\colon \mathcal{X} \to \mathcal{Y}$" as "f maps $\mathcal{X}$ to $\mathcal{Y}$" or "f from $\mathcal{X}$ to $\mathcal{Y}$" depending on the context.

You have seen functions before. In your math courses, you were probably asked to do something like the following: "Graph the curve $y = x^2$." What you were representing with your graph was a function because to each number on the "x-axis" you associated one and only one number on the "y-axis". (This would fail, for instance, if graph folded over on itself.) The y-coordinate is computed in this case by squaring the x-coordinate, which gives a definite rule for mapping one set of numbers (the real numbers) to another set of numbers (also the real numbers). That's all a function is.

Let's call this function f. What do we know about it? First, f maps real numbers to real numbers, so we can write $f\colon \mathbb{R} \to \mathbb{R}$. Moreover, we know that to any real number x, f associates the real number x^2. We write this in argument-result form as $f(x) = x^2$. The value of x is an argument to the function and the value x^2 is the result.[3]

There are a few things to notice about this function f. First, different values of the argument can give the same return value, for instance $x = \pm 1$ both give $f(x) = 1$, but each value of the argument only

1 

2 If $\mathcal{X}$ and $\mathcal{Y}$ are clear from context, we need not state the $\colon \mathcal{X} \to \mathcal{Y}$ part.

3 In C++ notation, we would write this function as

```
real f( real x )
{
    return( x*x );
}
```

which emphasizes the distinction between argument and return value.

4 See the discussion of domain, codomain, and range below.

has one associated return value. This uniqueness property is at the heart of the definition of a function. Second, the argument can be any real number whatsoever, but the result is always non-negative. Thus it would be just as correct to write $f: \mathbb{R} \to [0, \infty[$. Which is to be preferred is largely cosmetic; unless we have a good reason we go with the option that is simpler to write.[4] Note that it is not accurate to write $f: \mathbb{R} \to [0, 1[$ since $f(2) = 4$ is not an element of $[0, 1[$.

A quick aside on the terminology. The phrase "Graph the curve $y = x^2$" from your algebra and calculus classes is a bit too loose for our purposes. We will express this in two steps. First, we define the function: "Let f be the function $f(x) = x^2$." What we mean by this is that $f: \mathbb{R} \to \mathbb{R}$ is the function that takes x as an argument and returns x^2 for any $x \in \mathbb{R}$. It is important to note that the variable x here is a *local variable*; it is a placeholder for defining the function that has no meaning outside the definition. The variable `x` in the code of note 3 above has the same property; the compiler would not recognize as the same variable any reference to `x` outside the function. Second, having defined f we can "Graph the function f". This emphasizes that the object of interest is the function itself and deemphasizes the role played by dummy variables like y and x. Incidently, the graph itself is one way of identifying the function; it is useful for numerically valued functions like $f(x) = x^2$. Sometimes thinking of the two sets separately with the function represented by explicit mappings between points (as in the figures above) can be more helpful. Pictures can be a big help in thinking about mathematical ideas like functions. Remember in either case that, in general, functions are mappings from one set to another.

Thought Question Draw a curve on the x-y plane that does not represent the graph of a function.

Thought Question Can the graph of a valid function have breaks or jumps in it? Why or why not?

Thought Question Define a function that maps $\mathbb{Z}_+$ to $\mathbb{Z}$ so that every $k \in \mathbb{Z}$ is the value of the function for some $j \in \mathbb{Z}_+$.

Let $f\colon \mathcal{X} \to \mathcal{Y}$ be a general function. If $x \in \mathcal{X}$, we say that f "takes the value" or "maps to" $f(x)$ at x. The set $\mathcal{X}$ is called the *domain* of the function f and the set $\mathcal{Y}$ is called the *codomain* of the function f. The domain is the set of possible arguments of f, and the codomain is the set of possible values f can take. Often, as in the case of $f\colon \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$, the set of values that a function takes must be a subset of but need not be equal to the codomain. The set $\{f(x) \text{ such that } x \in \mathcal{X}\}$ is called the *range* of f.

The choice of codomain for a function is largely arbitrary as long as the chosen set contains the range of the function. Unless there is some compelling reason to choose a more specific set, in practice, we usually define functions so that the codomain is some familiar space, such as $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_\oplus$, $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{Z}_\oplus$, $\mathbb{C}$, or $\mathbb{R}^n$ for some $n > 1$. In this case, where the codomain of a function g is a named set $\mathcal{Y}$, we say that g is $\mathcal{Y}$-valued.[5] The common cases are as follows for a function g:

g takes values in	g is a ____ function
$\mathbb{R}$	real-valued
$\mathbb{R}_+$	positive real-valued
$\mathbb{R}_\oplus$	non-negative real-valued
$\mathbb{Z}$	integer-valued
$\mathbb{Z}_+$	positive integer-valued
$\mathbb{Z}_\oplus$	non-negative integer-valued
$\mathbb{Q}$	rational-valued
$\mathbb{C}$	complex-valued
$\mathbb{R}^n$	$\mathbb{R}^n$-valued

Notice that these categories are not necessarily exclusive. For example, an integer-valued function takes values in $\mathbb{R}$ since $\mathbb{Z} \subset \mathbb{R}$, so it is also a real-valued function. There is another, related distinction to be made. If g is $\mathbb{R}^n$-valued for integer $n > 1$, then we say that g is *vector valued*.[6] Otherwise, if g takes values in some subset of $\mathbb{R}$ or $\mathbb{C}$, we say that g is *scalar valued*.

Comparing Functions. Functions are mathematical objects in their own right. We can name them, operate on them, study sets or sequences of them. We can also compare them. Two functions are equal if they have the same domain and if for every element of the domain the two

5 When we want to be precise, we say that $g\colon \mathcal{X} \to \mathcal{Y}$ is a $\mathcal{Y}$-valued function defined on $\mathcal{X}$.

6 See subsection 2.7 for more details about vector-valued functions.

functions take the same value.

Conditions for Equality of Functions.
Two functions f and g are equal if and only if they have the same domain, and

$$f(x) = g(x) \text{ for all } x \text{ in their common domain.} \qquad \text{(F.1)}$$

Because real numbers can be ordered – there is a way of determining which of two numbers is bigger – some real-valued functions can also be ordered. Consider the functions $f(x) = e^x$ and $g(x) = 1 + x$ defined on $\mathbb{R}$. No matter which $x \in \mathbb{R}$ you consider it is always true that $f(x) \geq g(x)$. It makes sense, then, to consider the function f to be $\geq$ the function g.

Conditions for Ordering Real-Valued Functions.
For functions f and g defined on the same domain and taking values in $\mathbb{R}$, we have $f \leq g$ if and only if

$$f(x) \leq g(x) \text{ for all } x \text{ in their common domain.} \qquad \text{(F.2)}$$

The same conditions apply with $>$, $<$, or $\geq$ replacing $\leq$.

Thus, for example, a non-negative (positive) real-valued function f satisfies $f \geq 0$ ($f > 0$), where 0 is the constant function on the domain of f,

Function Images. If $\mathcal{A} \subset \mathcal{X}$, then the *image* of $\mathcal{A}$ is the subset of the codomain $\mathcal{Y}$ given by

$$f(\mathcal{A}) = \{f(x) \text{ such that } x \in \mathcal{A}\}\,.$$

The image of $\mathcal{A}$ tells us the set of all results we obtain by giving every element of $\mathcal{A}$ as a function of f. For $f(x) = x^2$, $f(\{-1, 1\}) = \{1\}$. To generate, a more complicated example, look at the image of a set of points in $\mathcal{X}$ in figure 1. The range of f is just the set $f(\mathcal{X})$ which is necessarily a subset of $\mathcal{Y}$.

If $\mathcal{B} \subset \mathcal{Y}$, then the *inverse image* of $\mathcal{B}$ is the subset of the domain $\mathcal{X}$ given by

$$f^{-1}(\mathcal{B}) = \{x \in \mathcal{X} \text{ such that } f(x) \in \mathcal{B}\}\,.$$

The inverse image of $\mathcal{B}$ tells us all the possible arguments whose values lie in $\mathcal{B}$. For $f(x) = x^2$, $f^{-1}(\{1\}) = \{-1, 1\}$. To generate, a more complicated example, look at the inverse image of a set of points in $\mathcal{Y}$ in figure 1.

Thought Question If $f: \mathcal{X} \to \mathcal{Y}$ is a function $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{X}$, is $f(\mathcal{A}_1 \cap \mathcal{A}_2) = f(\mathcal{A}_1) \cap f(\mathcal{A}_2)$? If so, why, and if not, given an example where it fails.

Thought Question If $f: \mathcal{X} \to \mathcal{Y}$ is a function, what is $f^{-1}(\emptyset)$? What is $f^{-1}(\mathcal{Y})$?

Thought Question If $f(x) = x^2$, what is $f^{-1}(\{-1, 1\})$?

Thought Question If $f: \mathcal{X} \to \mathcal{Y}$ is a function, $\mathcal{A} \subset \mathcal{X}$, and $\mathcal{B} \subset \mathcal{Y}$, then explain with a couple sentences, with a picture, or with an example why $f(f^{-1}(\mathcal{B})) \subset \mathcal{B}$ and $f^{-1}(f(\mathcal{A})) \supset \mathcal{A}$.

2. Some Important Examples

2.1. Constant Functions

The simplest function is a *constant function* which takes the same value for every argument. For example, $f(x) = 1$, $g(x) = 0$, and $h(x) = 117.017394$ for $x \in \mathbb{R}$ are all real-valued constants. A general function $f: \mathcal{X} \to \mathcal{Y}$ is constant if there is a $y \in \mathcal{Y}$ such that $f(x) = y$ for all $x \in \mathcal{X}$.

In practice, we treat the notation for constant functions a bit loosely, by allowing a constant value c to stand for the number and the function. For example, we use the symbol 1 for both the number 1 and the function that takes the value 1 for every point in its domain. Which meaning is being used should be clear from context. Thus, when we write $\mathsf{E}1 = 1$, the 1 on the left is a function because the expected value operator acts on functions and the 1 on the right is a number because the expected value operator returns a number.

2.2. Indicator Functions

An indicator function is any function that take only the values 0 and 1. These will be very important in our work.

If $\mathcal{A}$ is a subset of $\mathcal{X}$ for any set $\mathcal{X}$, then the indicator function $1_{\mathcal{A}}: \mathcal{X} \to \mathbb{R}$ is defined by

$$1_{\mathcal{A}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{A} \\ 0 & \text{if } x \notin \mathcal{A}. \end{cases} \tag{F.3}$$

The notation "1" is intended as a mnemonic; this function takes the value 1 on the set given in the subscript and is zero otherwise. We call this the "indicator of $\mathcal{A}$".

Thought Question To get a feel for indicators, sketch the graph of the indicators of the following sets: $[0,1]$, $[-2,-1] \cup [1,2]$, $[0,\infty[$, and $\mathbb{R}$.

Thought Question Describe the function $1_{\mathcal{X}}: \mathcal{X} \to \mathbb{R}$ more simply.

Thought Question If $\mathcal{A}, \mathcal{B} \subset \mathcal{X}$, define $1_{\mathcal{A} \cap \mathcal{B}}$ using $1_{\mathcal{A}}$ and $1_{\mathcal{B}}$.

Thought Question If $\mathcal{A}, \mathcal{B} \subset \mathcal{X}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$ (i.e., $\mathcal{A}$ and $\mathcal{B}$ are disjoint), define $1_{\mathcal{A} \cup \mathcal{B}}$ in terms of $1_{\mathcal{A}}$ and $1_{\mathcal{B}}$.

Thought Question Define the function $1_{\text{compl}(\mathcal{A})}$ in terms of $1_{\mathcal{A}}$.

Indicators are very useful for representing functions that take on a finite number of values. For example, suppose the function g is defined as follows:

$$g(x) = \begin{cases} 3 & \text{if } x > 4 \\ -2 & \text{if } 2 \leq x \leq 4 \\ 1 & \text{if } -1 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

Then, we can write

$$g = 3 \cdot 1_{]4,\infty[} - 2 \cdot 1_{[2,4]} + 1_{]-1,2[}.$$

To see this, plug in a variety of arguments on both sides and compare the values. Notice that only one of the indicators in the above expression is ever non-zero. In general, suppose g is a function that takes the value y_i on the set $\mathcal{A}_i$ for $i = 1, \ldots, n$, where $y_1, \ldots, y_n \in R$ and the sets $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are disjoint. We could write this g as

$$g(x) = \begin{cases} y_1 & \text{if } x \in \mathcal{A}_1 \\ y_2 & \text{if } x \in \mathcal{A}_2 \\ \ldots & \text{if } \ldots, \end{cases}$$

but this is both tedious and can make g hard to work with. Instead, we can write

$$g = \sum_{i=1}^{n} y_i \cdot 1_{\mathcal{A}_i}.$$

2.3. Polynomials

If x is a real number, then powers of x, such as $x \cdot x$, $x \cdot x \cdot x$, $x \cdot x \cdot x \cdot x$, and so forth, are also real numbers, which we denote respectively by x^2, x^3, x^4, and so forth. Notice that because multiplication is commutative $x^2 x^4 = x^6$; more generally for integers j, k, $x^k x^j = x^{k+j}$ for the same reason. Incidentally, by this rule, $x^4 x^{-1} = x^3$;. But what does x^{-1} mean? In order to cancel the extra factor of x, we must take $x^{-1} = 1/x$. It follows by similar logic that $x^{-k} = 1/x^k$ for any positive integer. Similarly, $x^0 = 1$ because $0+0+0+\cdots = 1$. As suggested by the above equation, we define $x^{-1} = 1/x$ Thus, for each $k \in \mathbb{Z}$, we define the kth degree *power function* which maps x to x^k. If k is even, then the power function x^k takes only non-negative values; if k is odd, then the power function x^k takes all real values. Whenever $k \neq 0$, the power function x^k takes the value 0 at 0. The function x^0 is just the constant 1.

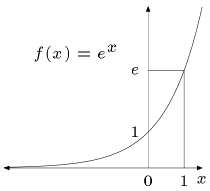
Suppose that $a_0, a_1, \ldots, a_m$ are real numbers. Then, the function that maps x to $a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$ is called an *mth degree polynomial.* A basic theorem of mathematics states that every mth degree polynomial has exactly m *complex* points at which the polynomial is zero.[7] There may be fewer than m such points in $\mathbb{R}$, however. Another important theorem states that any sufficiently smooth function can be approximated arbitrarily well by a polynomial of some (potentially large) degree.

2.4. Exponential Functions and Logarithms

One of the fundamental functions of mathematics is the exponential function, which maps x to e^x, where e is the irrational number 2.71828182846....[8] This function arises in so many different contexts in mathematics that can arguably be called the most important mathematical function. Here we give a few useful facts about the exponential

7 These are called roots of the polynomial.

8



function:

Properties of the Exponential Function

For any $x, u \in \mathbb{R}$,

1. $e^x > 0$, $e^0 = 1$, and $e^1 = e$.
2. $e^x \to \infty$ and $e^{-x} \to 0$, as $x \to \infty$.
3. $e^{x+u} = e^x\, e^u$.
4. $$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$
5. $\frac{d}{dx}\, e^x = e^x$ and $\int_{-\infty}^{x} e^t\, dt = e^x$.

Notice that from the above properties e^x increases as x increases, from 0 "at" $-\infty$ to ∞ "at" ∞. Suppose that $e^{x_1} = e^{x_2}$, then by property 3 above, $1 = e^{x_1}/e^{x_2} = e^{x_1}e^{-x_2} = e^{x_1-x_2}$ which implies that $x_1 - x_2 = 0$, or $x_1 = x_2$. In other words, given any $y > 0$, we can find some $x \in \mathbb{R}$ with $e^x = y$. That value x is called the *logarithm* of y, denoted by $\log y$. Whereas e^x maps $\mathbb{R}$ onto $]0, \infty[$, log maps $]0, \infty[$ onto $\mathbb{R}$. Here we give a few useful facts about the logarithm function:

Properties of the Logarithm Function

For any $y, v > 0$ and $p \in \mathbb{R}$,

1. $e^{\log(y)} = y$ and $\log(e^x) = x$.
2. If $0 < y < 1$, $\log(y) < 0$; $\log(1) = 0$; and if $y > 1$, $\log(y) > 0$.
3. As $y \to \infty$, $\log(y) \to \infty$ and as $y \to 0$, $\log(y) \to -\infty$.
4. $\log(y\, v) = \log(y) + \log(v)$.
5. $\log(1/y) = -\log(y)$.
6. $\log(y^p) = p \log(y)$.
7. $\frac{d}{dy} \log(y) = 1/y$ and $\int_0^y \log(t), dt = y \log(y) - y$.

We defined the logarithm of y as the number x such that $e^x = y$; we call this the logarithm "base e". We could just as easily define a logarithm "base 10" (the number x such that $10^x = y$) or logarithm "base 2" (the number x such that $2^x = y$) or logarithm "base b" for any $b > 0$.[9]

9 By the properties listed above, notice that $e^{x \log(b)} = e^{\log(b^x)} = b^x$.

Because of the importance of the exponential function, the logarithm base e is called the *natural logarithm*. If we want to talk about other bases, we will denote it with a subscript on the "log", as in $\log_{10}$, $\log_2$, and $\log_b$.

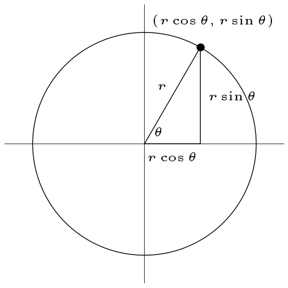
10 Taking $\theta = \pi$ yields the miraculous relationship

$$e^{\imath\pi} + 1 = 0$$

among the five most important numbers in mathematics.

2.5. Trigonometric Functions

11



The trigonometric functions sin and cos are intimately related to the exponential function through the relationship between complex numbers:

$$e^{\imath\theta} = \cos\theta + \imath\sin\theta.$$

As x changes, $e^{\imath x}$ moves around the unit circle in the complex plane.[10] What this tells us is that the coordinates of a point on the unit circle an angle θ from the positive x-axis are $(\cos\theta, \sin\theta)$.[11] This observation yields the key identity: $\cos^2\theta + \sin^2\theta = 1$. The functions cos and sin satisfy a huge variety of useful identities, including the following:

Trigonometric Identities
For any $\theta, \phi \in \mathbb{R}$,

1. $\cos^2\theta + \sin^2\theta = 1$,
2. $\cos\theta = \sin(\pi/2 - \theta)$ and $\sin\theta = \cos(\pi/2 - \theta)$,
3. $\sin(\theta + \phi) = \sin\theta\cos\phi + \cos\theta\sin\phi$ and $\cos(\theta + \phi) = \cos\theta\cos\phi - \sin\theta\sin\phi$,
4. $\sin' = \cos$ and $\cos' = -\sin$.

From sin and cos, we can produce the other trigonometric functions including $\tan = \sin/\cos$, $\cot = \cos/\sin$, $\sec = 1/\cos$, $\csc = 1/\sin$. For example, $\tan(\theta)$ is defined for $-\pi/2 < \theta < \pi/2$ and satisfies identities like $\sec^2\theta - \tan^2\theta = 1$.

Finally, over limited ranges these trigonometric functions are invertible. The inverses are usually obtained by putting the word "arc" in front of the name. Thus, we have arcsin, arccos, arctan, and so forth.

2.6. Other Commonly Used Functions

The *Gamma Function* $\Gamma(x)$ arises frequently in probabilistic calculations. It is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}\, e^{-t}\, dt.$$

While this may look daunting, you are actually already familiar with the Gamma Function: For an integer k,

$$\Gamma(k+1) = k! = k(k-1)\ldots 1, \tag{F.4}$$

where $k!$ is read "k factorial". Hence, $\Gamma(1) = \Gamma(2) = 1$, $\Gamma(4) = 6$, and so forth. This follows from the more general recurrence relationship which the Γ satisfies:

$$\Gamma(x+1) = x\,\Gamma(x).$$

We can also compute that $\Gamma(1/2) = \sqrt{\pi}$, so it follows that $\Gamma(3/2) = (1/2)\sqrt{\pi}$, $\Gamma(5/2) = (3/4)\sqrt{\pi}$, and so forth.

In general, there is no explicit expression for $\Gamma(x)$ in terms of elementary functions when x is not an integer or half integer. A formula called Stirling's Approximation does give us a useful approximation to Γ for large x, and hence to $n!$ for large n:

$$\lim_{x\to\infty} \frac{\Gamma(x+1)}{\sqrt{2\pi}\, x^{x+1/2}\, e^{-x}} = 1. \tag{F.5}$$

In other words, if n is large, we can approximate $n!$ by $\sqrt{2\pi}\, n^{n+1/2}\, e^{-n}$.

The *floor* and *ceiling* functions map a real number x to a nearby integer. The floor of x, denoted by $\lfloor x \rfloor$ is the largest integer that is less than or equal to x. For example, $\lfloor 1.5 \rfloor = 1 = \lfloor 1 \rfloor$ and $\lfloor -0.25 \rfloor = -1$. The ceiling of x, denoted by $\lceil x \rceil$, is the smallest integer that is greater than or equal to x. For example, $\lceil 1.5 \rceil = 2 = \lceil 2 \rceil$ and $\lceil -0.25 \rceil = 0$. Notice that if k is an integer, then $k = \lfloor k \rfloor = \lceil k \rceil$.

2.7. Vector Functions

2.7.1. $\mathbb{R}^n$-valued functions

A vector is, in essence, an indexed list of numbers. A vector valued function can therefore be viewed as an indexed list of functions; this view is helpful in working with such functions.

If, for some $n > 1$, g is an $\mathbb{R}^n$-valued function defined on a set $\mathcal{X}$, then we can write

$$g(x) \equiv (g_1(x), g_2(x), \ldots, g_n(x)), \tag{F.6}$$

where the g_i's are real-valued functions defined on $\mathcal{X}$. These are referred to as the *component functions* of g. Therefore, to specify an $\mathbb{R}^n$-valued function, we need only give its n component functions, and we can move freely between working with g or the g_is.

Most of the operators we use work transparently on vector-valued functions. For example, if we can define an integral $\int$ and derivative d/dx operators on $\mathcal{X}$, then we have

$$\int g(x)\,dx = (\int g_1(x)\,dx, \ldots, \int g_n(x)\,dx)$$
$$\frac{d}{dx}g(x) = (\frac{d}{dx}g_1(x), \ldots, \frac{d}{dx}g_n(x)).$$

That is, the operator just acts on each component, giving a new vector.

More importantly, the same story holds for the expected value and distribution operators. If $Y = (Y_1, \ldots, Y_n)$ is any $\mathbb{R}^n$-valued random variable and if X is any random variable (scalar or vector valued), we have that

$$\mathsf{E}Y = (\mathsf{E}Y_1, \ldots, \mathsf{E}Y_n)$$
$$\mathsf{D}_X g = \mathsf{E}g(X) = (\mathsf{E}g_1(X), \ldots, \mathsf{E}g_n(X)).$$

2.7.2. Real-valued functions defined on $\mathbb{R}^n$

We can also considered real-valued functions defined on $\mathbb{R}^n$ for some $n > 1$. If h is such a function and $x = (x_1, \ldots, x_n)$ is an n vector in $\mathbb{R}^n$ with components $x_1, \ldots, x_n \in \mathbb{R}$, then we can write the value of h at x in two ways:

$$h(x) \;\equiv\; h(x_1, \ldots, x_n). \tag{F.7}$$

We treat these as completely equivalent and interchangeable.

When integrating functions defined on $\mathbb{R}^n$, we use the analogous equivalence for infinitesimals

$$dx \equiv dx_1\, dx_2\, \cdots\, dx_n. \tag{F.8}$$

Thus, in $\mathbb{R}^2$, dx is an infinitesimal area element and in $\mathbb{R}^3$, dx is an infinitesimal volume element. These directly generalize the infinitesimal length element in $\mathbb{R}$.

Using these forms for the infinitesimal, we can write the integral of h in either of two equivalent ways

$$\int h(x)\, dx \;\equiv\; \int\int \cdots \int h(x_1, \ldots, x_n)\, dx_1\, dx_2\, \cdots\, dx_n. \tag{F.9}$$

The former is usually more convenient when working with integrals in the abstract. The operation of integrating a function like h has the same structure regardless of the dimension of the dimension. The latter is more convenient when actually doing the calculations.

2.7.3. Summary

Vector Function Equivalences

- A vector-valued function g can be defined in terms of real-valued component functions $g_1, \ldots, g_n$ as in equation (F.6).
- The operators E, D_X, as well as integration and differentiation, operate on vector-valued functions to produce a vector by acting on the real-valued component functions.
- A real-valued function h defined on $\mathbb{R}^n$ can be written with its argument in vector or component form as in equation (F.7).
- The integral of a real-valued function h defined on $\mathbb{R}^n$ can be written in either vector or component form as in equation (F.9).

3. How to make new functions from old ones

Most of the functions we will deal with are either in the list of examples above or are constructed from other functions by one of the methods described below.

Composition. If $f: \mathcal{X} \to \mathcal{Y}$ and $g: \mathcal{Y} \to \mathcal{Z}$ are functions, then the *composite function* $g \circ f: \mathcal{X} \to \mathcal{Z}$ defined by

$$(g \circ f)(x) = g(f(x)).$$

This function can only be defined if the range of f is a subset of the domain of g.

> **Thought Question** If f and g are as above, can $g \circ f$ and $f \circ g$ be different functions? Can they be the same function? Give examples.

Linear Combinations. Consider the power functions x^0, x, x^2, and x^3. If I have constants 1.2, 3.1, -3, and 10, I can form a new function $f(x) = 1.2x^0 + 3.1x - 3x^2 + 10x^3$, which is just a cubic polynomial. The operation of multiplying several function by constants and then adding the results together is called taking a *linear combination* of the functions. Suppose we have functions $f_1, \ldots, f_m$ with the same domain $\mathcal{X}$ and a codomain that is either $\mathbb{R}^k$ (or $\mathbb{C}^k$) for some $k \in \mathbb{Z}_+$. Given real (or complex) constants $a_1, \ldots, a_m$, we can form a linear combination of the f_i's

$$f = a_1 f_1 + \cdots + a_m f_m,$$

which is defined by

$$f(x) = a_1 f_1(x) + \cdots + a_m f_m(x), \quad x \in \mathcal{X}.$$

As you may have guessed, polynomials are just linear combinations of power functions.

Products. If I have two functions $f, g: \mathcal{X} \to \mathbb{C}$,[12] then the product $f \cdot g$ is a new function defined by

$$(f \cdot g)(x) = f(x)g(x), \quad x \in \mathcal{X}.$$

Notice that since all constants can be considered functions on any domain, we can always multiply a function by a constant to get a new function.

12 Any functions $\mathcal{X} \to \mathbb{R}$ satisfy this condition.

Cumulative Integrals. If $f\colon \mathbb{R} \to \mathbb{R}$, then we can define

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

Such a *cumulative integral function* has at least one derivative, and $F'(x) = f(x)$. Moreover, if $f(t) \geq 0$ for every $t \in \mathbb{R}$, $F(x)$ cannot decrease as x gets larger; that is, $y \geq x$ implies that $F(y) \geq F(x)$. To see this, think about the area under a curve accumulating as x increases; if the curve is always above zero, then the area is always increasing.

Thought Question Draw the graph of the cumulative integral of $1_{[0,1]}$.

Derivatives. If $f\colon \mathbb{R} \to \mathbb{R}$ is sufficiently smooth so that its derivatives are defined, then we can form new functions by taking derivatives f', f'', f''', $f^{(iv)}$.

4. One-to-one Correspondences

One-to-one and Onto. The only requirement on a function is that each argument can have one and only one value. However, it is possible for the function to take the same value for different arguments, cf. constant functions for example. Consider the functions from $\mathbb{R} \to \mathbb{R}$ given by $f_1(x) = x^2$, $f_2(x) = e^x$, $f_3(x) = x^3$, and $f_4(x) = x(x-1)(x+1)$. The first, f_1, maps the real line onto the non-negative numbers: for every $y \geq 0$, there is an $x \in \mathbb{R}$ such that $f_1(x) = y$. Moreover, for every $x \in \mathbb{R}$, f_1 takes the same value on $\pm x$. The function f_2 also maps the real line onto the non-negative numbers, but it does it "without repeats". If $f_2(x) = f_2(z)$, it must be true that $x = z$. Or put another way, if $x \neq z$, then $f_2(x) \neq f_2(z)$. Such a function is said to be "one-to-one" because every value is mapped from only one point. The function f_3 is also one-to-one, but in addition, it maps $\mathbb{R}$ onto $\mathbb{R}$ itself: for every $y \in \mathbb{R}$, there is an $x \in \mathbb{R}$ with $y = f_3(x)$. In contrast, f_4 maps $\mathbb{R}$ onto $\mathbb{R}$ but it is not on one-to-one as a quick sketch of its graph will show.

These different properties of a function are defined as follows. We say that a function $f\colon \mathcal{X} \to \mathcal{Y}$ is

- *onto* if the range and codomain are equal, that is $f(\mathcal{X}) = \mathcal{Y}$.
- *one-to-one* if $x_1 \neq x_2$ implies that $f(x_1) \neq f(x_2)$.

- *A one-to-one correspondence* if f is both onto and one-to-one.

As we saw above, two sets are said to have the same cardinality if there exists a one-to-one correspondence between them. If f is a one-to-one, then it is a one-to-one correspondence between its domain, $\mathcal{X}$, and its range, $f(\mathcal{X})$.

Inverse Functions. If f is a one-to-one correspondence between $\mathcal{X}$ and $\mathcal{Y}$, then, under f, each element of $\mathcal{X}$ corresponds to a unique element of $\mathcal{Y}$. Hence, we can define a function $f^{-1}\colon \mathcal{Y} \to \mathcal{X}$ that maps the other way. This is called the *inverse* of f, and if f has an inverse,[13] we say that f is *invertible.* In particular, we have that

- $f^{-1}(f(x)) = x$
- $f(f^{-1}(y)) = y$
- If $\mathcal{B} \subset \mathcal{Y}$, then the direct image of $\mathcal{B}$ under f^{-1}, $f^{-1}(\mathcal{B})$, is exactly the same as the inverse image of $\mathcal{B}$ under f defined above, also denoted by $f^{-1}(\mathcal{B})$.

13 Not every function has an inverse, only those that are one-to-one correspondences between their domain and codomain.

Despite all the notation, the idea is simple. If f maps $\mathcal{X}$ and $\mathcal{Y}$ by one-to-one correspondence, we think of each point in $\mathcal{X}$ connected by an arrow to one and only one point in $\mathcal{Y}$. The inverse function f^{-1} simply reverses the direction of the arrow.

> **Thought Question** Suppose f has an inverse function and also has at least one derivative. Use the chain rule to find the derivative of f^{-1}.

Monotonic Functions. Real-valued functions defined on some subset $\mathcal{A}$ of $\mathbb{R}$ that do not decrease as their arguments increase are called increasing. More precisely, if $f\colon \mathcal{A} \to \mathbb{R}$ and if $x > y$ implies that $f(x) \geq f(y)$, we say that f is *increasing.* If $x > y$ implies $f(x) > f(y)$, we say that f is *strictly increasing.* The function $-1_{]-\infty,0[} + 1_{[0,1]} + 2 \cdot 1_{]1,\infty[}$ is increasing but not strictly increasing; the functions e^x, $\tan x$, $x/1{+}x$ are strictly increasing. Real-valued functions defined on some subset $\mathcal{A}$ of $\mathbb{R}$ that do not increase as their arguments increase are called decreasing More precisely, if $f\colon \mathcal{A} \to \mathbb{R}$ and if $x > y$ implies that $f(x) \leq f(y)$, we say that f is *decreasing.* If $x > y$ implies $f(x) < f(y)$, we say that f is *strictly decreasing.* The function $2 \cdot 1_{]-\infty,0[} + 1_{[0,1]} - 1_{]1,\infty[}$ is decreasing while e^{-x}, $1/1 + x$, and the function that maps $x \in \,]0,\infty[$ to $1/x$ are strictly decreasing. A function that is either increasing or decreasing is called *monotonic*. The functions sin, x^2, and e^{-x^2} are not monotonic.

Thought Question If $f\colon \mathbb{R} \to \mathbb{R}$ is a strictly increasing function with at least one derivative, is it necessarily a one-to-one correspondence between its domain and range? Why or why not?

HELP SHEET

1. Useful Facts

Combinatorics

For integers n and k with $0 \leq k \leq n$

$$n! = n(n-1)\cdots 1$$
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Exponentials and Logarithms

$$\begin{array}{ccc} e^x \geq 0 & e^0 = 1 & e \approx 2.7182818 \\ \log(1) = 0 & \log(x \cdot y) = \log(x) + \log(y) & \log(x^a) = a\log(x) \\ e^{\log(x)} = x & \log(e^y) = y & x^a = e^{a\log x} \end{array}$$

Unless otherwise specified, log denotes the natural logarithm and $\log_b$ denotes the logarithm base b.

Trigonometric Functions

$$\begin{array}{cc} \tan\theta = \dfrac{\sin\theta}{\cos\theta} & \cot\theta = \dfrac{\cos\theta}{\sin\theta} \\ \sec\theta = 1/\cos\theta & \csc\theta = 1/\sin\theta \\ \sin^2\theta + \cos^2\theta = 1 & \tan^2\theta + 1 = \sec^2\theta \\ \sin(\theta) = \cos(\theta - \pi/2) & e^{\imath\theta} = \cos\theta + \imath\sin\theta \\ \sin(\theta \pm \phi) = \sin\theta\cos\phi \pm \sin\phi\cos\theta & \cos(\theta \pm \phi) = \cos\theta\cos\phi \mp \sin\theta\sin\phi \end{array}$$

The Gamma Function Γ

$$\Gamma(t) = \int_0^\infty x^{t-1}\, e^{-x}\, dx$$
$$\Gamma(t+1) = t\Gamma(t) \quad \Gamma(1) = 1 \quad \Gamma(1/2) = \sqrt{\pi}$$
$$\Gamma(n+1) = n! \text{ for integer } n \geq 0$$

The Standard Normal CDF Φ

Φ is smooth and increasing with $\Phi(0) = 1/2$, $\Phi(-x) = 1 - \Phi(x)$, and $\Phi'(x) = e^{-x^2/2}/\sqrt{2\pi}$. Approximate values of Φ are given by

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
-6	9.866×10^{-10}	-3.5	2.326×10^{-4}	-1	0.15866	0.5	0.69146
-5.5	1.899×10^{-8}	-3	1.350×10^{-3}	-0.75	0.22662	1	0.84134
-5	2.867×10^{-7}	-2.5	6.210×10^{-3}	-0.50	0.30854	2	0.97725
-4.5	3.398×10^{-6}	-2	2.275×10^{-2}	-0.25	0.40129	3	0.99865
-4	3.167×10^{-5}	-1.5	6.681×10^{-2}	0	0.5	4	0.99997

2. Sums

For real number $0 < r < 1$, integer $n \geq 0$, and arbitrary numbers x, y, and z, we have

$$\sum_{j=0}^{n} 1 = n + 1 \qquad \sum_{j=0}^{n} j = \frac{j(j+1)}{2}$$

$$\sum_{j=n}^{\infty} r^j = \frac{r^n}{1-r} \qquad \sum_{j=0}^{n} r^j = \frac{1 - r^{n+1}}{1-r}$$

$$\sum_{k=0}^{\infty} \frac{z^k}{k!} = e^z \qquad \sum_{k=0}^{n} \binom{n}{k} x^k\, y^{n-k} = (x+y)^n$$

3. Derivatives and Integrals

$$\frac{d}{dx} 1 = 0 \qquad \frac{d}{dx} x^a = a x^{a-1} \text{ for } a \neq 0$$

$$\frac{d}{dx} e^{ax} = a\, e^{ax} \qquad \frac{d}{dx} \log(x) = \frac{1}{x}$$

$$\int x^a\, dx = \frac{x^{a+1}}{a+1}, \text{ for } a \neq -1 \qquad \int \frac{1}{x}\, dx = \log x \qquad \int e^{ax}\, dx = \frac{1}{a}\, e^{ax}$$

$$\int \log(x)\, dx = x \log(x) - x \qquad \int_t^{\infty} e^{-\lambda s}\, ds = \frac{1}{\lambda}\, e^{-\lambda t}$$

$$\int_0^t e^{-\lambda s}\, ds = \frac{1}{\lambda}\,(1 - e^{-\lambda t}) \qquad \int_t^{\infty} s\, e^{-\lambda s}\, ds = \frac{1 + \lambda t}{\lambda^2}\, e^{-\lambda t}$$

4. Variance

For any random variable X, $\mathsf{Var}\, X = \mathsf{E}(X - \mathsf{E}X)^2 = \mathsf{E}X^2 - (\mathsf{E}X)^2$. The standard deviation SD of X is the square-root of the variance, that is $\mathrm{SD}(X) = \sqrt{\mathsf{Var}\,(X)}$.

5. Discrete Distributions

- DiscreteUniform$\langle a_1, \ldots, a_n \rangle$ for integer $n \geq 1$ and the a_is distinct real numbers.

  A random variable X with a DiscreteUniform$\langle a_1, \ldots, a_n \rangle$ distribution has PMF
$$\mathsf{p}_X(u) = \begin{cases} \frac{1}{n} & \text{if } u \in \{a_1, \ldots, a_n\} \\ 0 & \text{otherwise.} \end{cases}$$

- Binomial$\langle n, p \rangle$ for integer $n > 0$ and $0 < p < 1$.

  A random variable X with a Binomial$\langle n, p \rangle$ distribution has PMF
$$\mathsf{p}_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \{0, \ldots, n\} \\ 0 & \text{otherwise.} \end{cases}$$

  In this case, $\mathsf{E}(X) = np$ and $\mathsf{Var}\,(X) = np(1-p)$.

  Special Case $n = 1$: We call the Binomial$\langle 1, p \rangle$ a Bernoulli$\langle p \rangle$ distribution.

- Poisson$\langle \lambda \rangle$ for $\lambda > 0$.

  A random variable X with a Poisson$\langle \lambda \rangle$ distribution has PMF
$$\mathsf{p}_X(k) = \begin{cases} e^{-\lambda} \dfrac{\lambda^k}{k!} & \text{if } k \in \mathbb{Z}_\oplus, \\ 0 & \text{otherwise.} \end{cases}$$

  In this case, $\mathsf{E}(X) = \lambda$ and $\mathsf{Var}\,(X) = \lambda$.

- NegativeBinomial$\langle k, p \rangle$ for integer $k \geq 1$ and $0 < p < 1$.

  A random variable X with a NegativeBinomial$\langle k, p \rangle$ distribution has PMF
$$\mathsf{p}_X(n) = \begin{cases} \binom{n-1}{k-1} p^k (1-p)^{n-k} & \text{if integer } n \geq k \\ 0 & \text{otherwise.} \end{cases}$$

  In this case, $\mathsf{E}(X) = k/p$ and $\mathsf{Var}\,(X) = k(1-p)/p^2$.

  Special Case $k = 1$: We call the NegativeBinomial$\langle 1, p \rangle$ the Geometric$\langle p \rangle$ distribution.

6. Continuous Distributions

- Uniform$\langle a,b\rangle$ distribution with $-\infty < a < b < \infty$.

  A random variable X with a Uniform$\langle a,b\rangle$ distribution has PDF

$$\mathsf{f}_X(u) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < u < b \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $\mathsf{E}(X) = (a+b)/2$ and $\mathsf{Var}\,(X) = (b-a)^2/12$.

- Normal$\langle \mu, \sigma^2\rangle$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

  A random variable X with a Normal$\langle \mu, \sigma^2\rangle$ distribution has PDF

$$\mathsf{f}_X(z) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}, \qquad -\infty < z < \infty.$$

In this case, $\mathsf{E}(X) = \mu$ and $\mathsf{Var}\,(X) = \sigma^2$.

  Useful facts:

  - If X has a Normal$\langle \mu, \sigma^2\rangle$ distribution, then $Z = (X-\mu)/\sigma$ has a Normal$\langle 0, 1\rangle$ distribution.
  - If Z has a Normal$\langle 0, 1\rangle$ distribution, then $X = \mu + \sigma Z$ has a Normal$\langle \mu, \sigma^2\rangle$ distribution.
  - The sum of two random variables with a Normal distribution also has a Normal distribution.

- Gamma$\langle \alpha, \lambda\rangle$ with $\alpha > 0$ and $\lambda > 0$.

  A random variable X with a Gamma$\langle \alpha, \lambda\rangle$ distribution has PDF

$$\mathsf{f}_X(z) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)}\, z^{\alpha-1}\, e^{-\lambda z} & \text{if } z > 0 \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $\mathsf{E}(X) = \alpha/\lambda$ and $\mathsf{Var}\,(X) = \alpha/\lambda^2$.

Special Case $\alpha = 1$: We call Gamma$\langle 1, \lambda\rangle$ the Exponential$\langle \lambda\rangle$ distribution. The PDF above reduces to $\mathsf{f}_X(z) = \lambda e^{-\lambda z}$ for $z > 0$.

- Beta$\langle a, b\rangle$ with $a, b > 0$.

  A random variable X with a Beta$\langle a, b\rangle$ distribution has PDF

$$\mathsf{f}_X(u) = \begin{cases} \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, u^{a-1}(1-u)^{b-1} & \text{if } 0 < u < 1, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $\mathsf{E}(X) = a/(a+b)$ and $\mathsf{Var}\,(X) = ab/(a+b)^2(a+b+1)$.

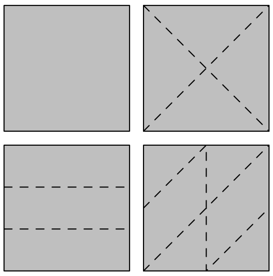
MEASURES

M

1. Fundamentals

A measure is a mathematical object that quantifies the size of sets. Each distinct measure embodies a different way to assess how big a set is. At first thought, it would seem that there is only one natural measure of a set's size—its cardinality. Indeed, the cardinality is a measure as we will define it below; given a set $\mathcal{A}$, $\#\mathcal{A}$ is a non-negative, possibly infinite, number. But if we examine the issue carefully, we realize that there are other reasonable ways to measure the size of a set. Consider the two sets $[1, 10]$ and $[10, 100]$. Both have the same cardinality ($= \#\mathbb{R}$), but our physical intuition suggests that the latter is "bigger" somehow. Indeed, we can compute the length of the two intervals, 9 and 90 respectively, which confirms our intuition. So, the length of a set provides a different measure of its size. We can also think about the lengths of the two intervals on a logarithmic scale (base 10): [1,10] spans the exponents 0 through 1 and [10,100] spans the exponents 1 through 2. Measured this way, the two sets have the same size, 1, which is finite.

Measures are based on a familiar idea—size. However, they may seem a bit abstract because mathematicians like to take a concept, whittle it down to its barest properties, and then explore the implications of those properties in contexts that are very different from the original idea. While this often turns out to be a profitable enterprise, it can seem daunting at first. Even though measures are likely unfamiliar to you as mathematics, try to keep in mind the strong intuition about size on which they rest as you proceed through the material. To build on this, we begin our exploration of measures with three examples that are familiar from everyday life: length, area, and volume.

1 

1.1. Length, Area, and Volume

We can measure the length of a curve, the area of a surface, the volume of some object. (Eureka!) From a mathematical perspective, the curve, surface, and object are just sets of points in space, so length, area, and volume give us a way to measure the extent of one, two, and three-dimensional sets respectively. This is the fundamental idea of a measure: a rule that associates to sets a number that quantifies the size of the set. To get an intuitive grasp of what a measure is, we will use length, area, and volume as our models. We will find the properties that these satisfy, and then ask which of these properties are essential for thinking about the size of sets. This will lead us to the general idea of a measure.

Property #1. Non-negativity. Length, area, and volume can in principle be any positive value; we just need to find a long enough curve, a big enough area, or a large enough object. They can be zero as well: a point is a curve that goes nowhere, a line is a rectangle with no width, a plane is a solid with no depth. If one is willing to posit an infinite space to the universe, we can imagine curves, surfaces, and volumes that are infinite as well. But negative length, negative area, negative volume have no meaning.

Property #2. Additivity. We can measure the length of a path on a map along roads from Pittsburgh, through Philadelphia and Trenton, New Jersey, to Newark, New Jersey. If we cut that path into three non-overlapping pieces – say from Pittsburgh to Philadelphia, Philadelphia to Trenton, and Trenton to Newark – then the lengths of the three pieces combined must be the same as the length of the original path. If I divide a rectangular region into non-overlapping pieces,[1] the total area of the pieces must equal the total area of the original region. Moreover, this same fact is true no matter how I partition the region. The same is true for volume as well.

We call this property *additivity.* To divide a set $\mathcal{A}$ – the points on a curve, in the rectangular region, or in the object – into non-overlapping pieces is to find a collection of pairwise disjoint sets $\mathcal{A}_1, \mathcal{A}_2, \ldots$ such that $\mathcal{A} = \bigcup_i \mathcal{A}_i$. The $\mathcal{A}_i$'s here represent the pieces into which we divide the original set. To say that the union of the pieces equals the original set is to say that we have included all the pieces that we made. To say that the sets are disjoint is to say that the pieces do not overlap. The additivity property states that the length (or area or volume, as appropriate) of $\mathcal{A}$ must equal the sum of the lengths (or areas or volumes) of the $\mathcal{A}_i$'s.

In our physical reality, we can imagine dividing a curve or surface or object into a finite (though perhaps large) number of pieces. If the additivity property holds for any finite number of pieces, we call it *finite additivity.* Mathematically it will be more convenient to require a stronger property: we require that additivity hold for any countable[2]number of pieces. We call this *countable additivity.*

2 Recall that countable means finite or countably infinite.

Any measure of size for which the additivity property holds must have another important property – monotonicity. To understand monotonicity, consider the path from Pittsburgh to Newark above. If we go along that path until Philadelphia and then stop, the length of the subpath cannot be bigger than the entire path. More generally, if $\mathcal{A} \subset \mathcal{B}$, then the size of $\mathcal{A}$ must be no greater than the size of $\mathcal{B}$ because of additivity: $\mathcal{B}$ can be divided into two disjoint pieces, one of which is $\mathcal{A}$, $\mathcal{B} = \mathcal{A} \cup \operatorname{compl}(\mathcal{A})$.

Property #3. Empty Set. In daily life, we rarely try to compute the length, area, or volume of the empty set. But to take these ideas into the mathematical realm, we need to come up with a sensible assignment that is consistent with our physical intuition. Since the magnitude of length, area, and volume derives from the extent of the curve, surface, or object, and since the empty set by definition has no points and thus no extent, it makes sense to take the length, area, and volume of the empty set as zero.

Property #4. Other Null Sets. The length of a point is 0. The area of a line, or indeed any curve, is 0. The volume of a plane, or any surface, is 0. The existence of these "null sets" tells us a lot about the nature of length, area, and volume. In particular, length is inherently one-dimensional, area two-dimensional, and volume three-dimensional; if we compute the size of a set of too low a dimension, we et 0.

Property #5. Translation Invariance. If I take a curve and shift it in space without changing its shape, its length does not change. Similarly for a region or object. Length, area, and volume are unchanging (invariant) under shifts in space (translations). If this weren't true, then walking around town could change your size.

3 An interval is a one-dimensional hyper-rectangle. Remember that $a < b$ is implicit in this notation.

Property #6. Hyper-Rectangles. An interval of the form $[a, b] \subset \mathbb{R}^3$ has length $b - a$. This is consistent with the fact that a point has length zero, because the point $\{a\}$ can be written as $[a, a]$. The additivity of length implies that the intervals of the form $]a, b[$, $]a, b]$, and $[a, b[$ all have the same length $b - a$. To see why, notice that the interval $[a, b]$ can be written as a union of pairwise disjoint intervals in three ways:

$$\begin{aligned}
[a, b] &= [a, a] \cup]a, b[\cup [b, b] \\
&= [a, a] \cup]a, b] \\
&= [a, b[\cup [b, b] ,
\end{aligned}$$

Additivity tells us that for any such disjoint partition of $[a, b]$, the sum of the lengths of the pieces must be $b - a$. Since the boundaries $[a, a]$ and $[b, b]$ have length 0 (they are zero-dimensional sets), all four types of intervals have the same length. This still holds if a or b are infinite; $]a, \infty[$, $]-\infty, b[$, and $]-\infty, \infty[$ all have length ∞.

Similarly, the area of a rectangle $[a_1, b_1] \times [a_2, b_2]$ is $(b_1 - a_1)(b_2 - a_2)$, and the volume of a rectangular solid $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$, $(b_1 - a_1)(b_2 - a_2)(b_3 - a_3)$. These hold for rectangles and rectangular solids with any part of the boundary missing because the boundary is a lower dimensional set.

These six properties determine length, area, and volume. In other words, any measure of set size that satisfies these properties must be length (if it is defined for one-dimensional sets), area (if it is defined for two-dimensional sets), or volume (if it is defined for three-dimensional sets). Our next task is to decide which of these six properties are quintessential properties for a measure of the size of a set and which are specific to length, area, and volume. Well, we only know one other candidate for a measure – cardinality – so finding which of these properties cardinality shares will help us determine the general features we wish a measure to satisfy.

For cardinality, only properties #1, #2, #3, and #5 hold. While these all seem quite general, it turns out that #5, translation invariance, is a bit too specialized. Here's a quick argument. If # is a measure of the size of sets, then the following should be as well: assign $\mathcal{A} \subset \mathbb{R}$ to be size $\#\mathcal{A} \cap [0, 1]$. This computes the cardinality of the part of $\mathcal{A}$

that lies in $[0, 1]$ and so has all of the features of cardinality except translation invariance. So we judge properties #1, #2, and #3 to be essential for measuring size.

This gives us a provisional definition of a measure: any function that maps a set to a non-negative (and possibly infinite) number, such that $\emptyset$ maps to 0 and the (countable) additivity property is satisfied. It turns out that there is nothing provisional here, as we will see below.

Before beginning with measures in the abstract, let's look a little more formally at the measures that correspond to length, area, and volume.

We begin in $\mathbb{R}$ and, since this is one-dimensional, with length. Suppose we have a subset $\mathcal{A} \subset]0, 1[$; how do we compute the length of $\mathcal{A}$? To make things interesting we will assume that $\mathcal{A}$ is a strict subset; we already know the length of $]0, 1[$.[4] What we are looking for is a mathematical algorithm that is guaranteed to produce the length of $\mathcal{A}$, although it may not be implementable in practice.

To start with, since $\mathcal{A} \subset]0, 1[$, we know that $\text{length}\big(\mathcal{A}\big) \leq 1$ by the monotonicity property of length mentioned above. Our approach will be to see how tight we can make this upper bound. Think about measuring the length of $\mathcal{A}$ by covering $\mathcal{A}$ with pieces of string. The subset $]0, 1[$ corresponds to a single piece of string of length 1; it covers $\mathcal{A}$ but if $\mathcal{A}$ has some gaps, our string is a bit too long. So, we can try to cover $\mathcal{A}$ by several pieces of string of total length < 1. Because $\mathcal{A}$ may have some very short parts, these pieces of string may be a little too long or may overlap each other a bit. That's ok for now. When we add up the length of all our pieces, we find that $\text{length}\big(\mathcal{A}\big) \leq$ the sum of the lengths of the pieces, which in turn is < 1. If we keep trying – cutting the pieces more and more carefully to get a tighter fit to $\mathcal{A}$ while still covering $\mathcal{A}$, then our upper bound on $\text{length}\big(\mathcal{A}\big)$ will get tighter and tighter. Thus, $\text{length}\big(\mathcal{A}\big)$ must be $\leq$ the total string length for every such covering of $\mathcal{A}$ by pieces of string, and we can get arbitrarily close to $\text{length}\big(\mathcal{A}\big)$ with some such covering.[5] It follows that $\text{length}\big(\mathcal{A}\big)$ must be the greatest lower bound of the lengths of all coverings of $\mathcal{A}$ by pieces of string.

Ok, enough with string. The pieces of string in the above argument are just analogies for intervals. Consider a countable collection of in-

4 Nonetheless, we need our algorithm to work for $]0, 1[$, so we will have to revisit this case.

5 In other words, there can be no number ℓ such that such that $\text{length}\big(\mathcal{A}\big) < \ell$ and $\ell <$ the string length of every covering of $\mathcal{A}$.

tervals $]a_i, b_i[$ that "cover" $\mathcal{A}$, that is

$$\mathcal{A} \subset \bigcup_i \,]a_i, b_i[\, .$$

We do not require the intervals to be disjoint.[6] We call such a collection of intervals a *covering of* $\mathcal{A}$. Since any overlaps will be counted multiple times when adding up the lengths of the intervals, the additivity property of length implies that

$$\text{length}\Big(\mathcal{A}\Big) \leq \sum_i \text{length}\Big(]a_i, b_i[\Big) = \sum_i (b_i - a_i).$$

The $\sum_i (b_i - a_i)$ is the "length of the covering".

We can then consider a series of tighter and tighter coverings, trimming a little overlap here, cutting a little excess there, making the length of the covering closer and closer to the $\text{length}\Big(\mathcal{A}\Big)$. We know two facts:

1. $\text{length}\Big(\mathcal{A}\Big) \leq$ the length of every covering,
2. We can find some covering whose length is arbitrarily close to $\text{length}\Big(\mathcal{A}\Big)$.

It follows that $\text{length}\Big(\mathcal{A}\Big)$ is the greatest lower bound of all the lengths of coverings of $\mathcal{A}$. We are done.

This algorithm defines a measure $\text{length}\Big(\Big)$ which satisfies all the properties alluded to above. We can check that it makes sense with two simple examples, $]0, 1[$ and a point $\{1/2\}$. In the former case, the simple covering $]0, 1[$ is perfect and gives us a length of 1. In the latter case, the intervals $]1/2 - 1/n, 1/2 + 1/n[$ for positive integer n each cover $\{1/2\}$, but since their length goes to 0 as n goes to ∞, $\text{length}\Big(\{1/2\}\Big)$ must be 0, as we expect.

The same algorithm works to define area in $\mathbb{R}^2$, except we use two-dimensional intervals (i.e., rectangles), and to define volume in $\mathbb{R}^3$, except we use three-dimensional intervals (i.e., rectangular solids). In fact, we can use the same argument in any $\mathbb{R}^k$ using k-dimensional hyper-rectangles. This means that length, area, and volume are intimately related (as we knew from the start) and differ only in dimensionality. We call this family of measures on $\mathbb{R}^k$ *Lebesgue measures*[7] after the French mathematician Henri Lebesgue.

6 If the intervals were disjoint, then the sum of their lengths would equal the length of $\mathcal{A}$ by countable additivity. The point of the algorithm is that it is harder to find disjoint intervals that cover $\mathcal{A}$ than it is to find intervals that cover $\mathcal{A}$ without this constraint.

7 Lebesgue is pronounced Leb-ache.

Lebesgue measure on $\mathbb{R}$ is length. Lebesgue measure on $\mathbb{R}^2$ is area. Lebesgue measure on $\mathbb{R}^2$ is volume. Lebesgue measure on $\mathbb{R}^k$ for $k > 2$ is a "hyper-volume", a direct generalization of length, area, and volume which satisfies properties #1–#6 above (assigning measure $(b_1 - a_1) \cdots (b_k - a_k)$ to the hyper-rectangle $[a_1, b_1] \times \cdots \times [a_k, b_k]$). Because these measures have all the same properties except dimensionality, we use the same symbol, μ_{Leb}, to refer to all of them, and let the domain ($\mathbb{R}$, $\mathbb{R}^2$, etc.) be clear from context.

1.2. Measures Defined

Let $\mathcal{X}$ be any set. A measure on $\mathcal{X}$ is a function μ that maps the set of subsets of $\mathcal{X}$ to $[0, \infty]$ that satisfies (i) $\mu(\emptyset) = 0$ and (ii) the following *countable additivity* property:

Definition. For any countable and pairwise disjoint collection of subsets of $\mathcal{X}$ $\mathcal{A}_1, \mathcal{A}_2, \ldots$,

$$\mu\left(\bigcup_i \mathcal{A}_i\right) = \sum_i \mu(\mathcal{A}_i).$$

Remember that μ is just a function,[9] albeit on a strange domain, so if $\mathcal{A} \subset \mathcal{X}$, $\mu(\mathcal{A})$ is the image of the "point" $\mathcal{A}$ in $2^{\mathcal{X}}$ under the function μ, which is just a number. We call $\mu(\mathcal{A})$ the "measure of the set $\mathcal{A}$" or "measure of $\mathcal{A}$" for short. Notice that according to the definition, the measure of $\mathcal{A}$ may be ∞.

The countable additivity property seems rather abstract as it is stated above, but it comes with a definite physical intuitition that you should keep in mind: If we take a set and divide it into non-overlapping pieces, in any way, then the measure (think "size") of all the pieces adds up to the total measure (think "size") of the original set. See figure 1.

We have seen that the familiar ideas of length, area, and volume are measures, corresponding to Lebesgue measure on $\mathbb{R}$, $\mathbb{R}^2$, and $\mathbb{R}^3$, but these are so familiar that they can obscure some of the power of the idea. So, let's look at some very different examples of measures.

Example 1. How many elements? A natural measure associates to each set its cardinality. But this is exactly $\#(\mathcal{A})$ as we defined it earlier. Let us check that this actually defines a measure. First, $\#(\emptyset) = 0$ by definition. Second, if we take a collection of disjoint sets, then the cardinality of the union of these sets is the sum of the cardinalities

8 In a completely rigorous definition of a measure, we would have to be careful about specifying the set of subsets on which the measure is defined (i.e., the domain of μ). The reason is that if countable additivity holds it is possible in some cases to construct anomalous sets that cannot be measured at all. Thus, the domain of μ would sometimes need to be a strict subset of $2^{\mathcal{X}}$ that itself satisfies certain requirements. This is an important but very technical point that is largely irrelevant to our efforts here, so we ignore it in favor of the "practical" definition given.

9 As we have defined it, μ: $2^{\mathcal{X}} \to [0, \infty]$

because the sets do not overlap. Cardinality gives one way to quantify the size of a set; notice how it differs from length for subsets. For example, all intervals in $\mathbb{R}$ have the same cardinality.

Example 2. A point mass at 0. Consider a measure $\delta_{\{0\}}$ on $\mathbb{R}$ defined to give measure 1 to any set that contains 0 and measure 0 to any set that does not. Mathematically, we write the definition

$$\delta_{\{0\}}(\mathcal{A}) = \#(\mathcal{A} \cap \{0\}) = \begin{cases} 1 & \text{if } 0 \in \mathcal{A} \\ 0 & \text{otherwise,} \end{cases}$$

for $\mathcal{A} \subset \mathbb{R}$. We can similarly define a point mass at any point x by $\delta_{\{x\}}(\mathcal{A}) = \#(\mathcal{A} \cap \{x\})$.

Example 3. Counting measure on the integers. Consider a measure $\delta_{\mathbb{Z}}$ that assigns to each set $\mathcal{A}$ the number of integers contained in $\mathcal{A}$. We define this by

$$\delta_{\mathbb{Z}}(\mathcal{A}) = \#(\mathcal{A} \cap \mathbb{Z}).$$

Counting measure can be defined for sets other than $\mathbb{Z}$ in an analogous way.

Example 4. A Geometric Measure. Suppose that $0 < r < 1$. Define a measure on $\mathbb{R}$ that assigns to a set $\mathcal{A}$ a geometrically weighted sum over non-negative integers in $\mathcal{A}$. Specifically,

$$\mu(\mathcal{A}) = \sum_{i \in \mathcal{A} \cap \mathbb{Z}_{\oplus}} r^i.$$

For example, it follows that $\mu(\mathbb{R}) = \mu(\mathbb{Z}) = \mu(\mathbb{Z}_{\oplus}) = 1/(1-r)$ and that $\mu(\{0\}) = 1$.

Example 5. A Binomial Measure. Let n be a positive integer and let $0 < p < 1$. Define μ as follows:

$$\mu(\mathcal{A}) = \sum_{k \in \mathcal{A} \cap \{0,1,\ldots,n\}} \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k},$$

where ! denotes the factorial function, $j! = j(j-1)\cdots 1$.

Example 6. Bivariate Gaussian. Define a measure on $\mathbb{R}^2$ by

$$\mu(\mathcal{A}) = \int_{\mathcal{A}} \frac{1}{2\pi}\, e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy.$$

Example 7. Uniform on a Ball in $\mathbb{R}^3$. Let $\mathcal{B}$ be the set of points in $\mathbb{R}^3$ that are within a distance 1 from the origin; this is called the unit ball in $\mathbb{R}^3$. (It looks like a ball centered on $(0, 0, 0)$.) Define a measure on $\mathbb{R}^3$ as follows:

$$\mu(\mathcal{A}) = \frac{3}{4\pi}\,\mu_{\mathsf{Leb}}(\mathcal{A} \cap \mathcal{B}).$$

This measure yields $\mu(\mathbb{R}^3) = \mu(\mathcal{B}) = 1$. It looks like Lebesgue measure inside the unit ball but assigns measure 0 to any part of a set outside the ball.

Example 8. Not a measure. Suppose we define a set function that maps a set $\mathcal{A}$ to the smallest number (including ∞) that is $\geq$ every element of $\mathcal{A}$.[10] This arguably gives one way to quantify the size of the set but it is not a measure.

10 This is called the least upper bound, or supremum, of $\mathcal{A}$ and is denoted $\sup \mathcal{A}$. If $\mathcal{A}$ is finite, then $\sup \mathcal{A} = \max \mathcal{A}$, but some infinite sets like $]0, 1[$ have no maximum. What is $\sup\,[0, 1]$? $\sup\,]0, 1[$? $\sup \mathbb{R}$?

The definition of a measure implies two useful properties that it is worthwhile to see. Suppose that μ is a measure. Then the following hold:

1. Monotonicity. If $\mathcal{A} \subset \mathcal{B}$, then $\mu(\mathcal{A}) \leq \mu(\mathcal{B})$.
2. Subadditivity. If $\mathcal{A}_1, \mathcal{A}_2, \ldots$ is a countable collection of sets – not necessarily disjoint – then

$$\mu\left(\bigcup_i \mathcal{A}_i\right) \leq \sum_i \mu(\mathcal{A}_i).$$

The monotonicity property was discussed in Section 1.1. It follows from additivity because $\mathcal{B} = \mathcal{A} \cup \operatorname{compl}(\mathcal{A})$ and the two pieces are disjoint. The subadditivity property can be understood intuitively: if we divide up a set into *overlapping* pieces, then the sum of the measure of the pieces must be greater than if we made the pieces disjoint since any overlap is counted more than once.

1.3. A Brief Digression on Infinitesimals

Although the calculus was developed independently by Newton and Leibniz, much of the notation we use today comes from Leibniz. Of particular relevance for our purposes are the "infinitesimals" like dx and dy that appear in the expressions for derivatives and integrals: df/dx and $\int f(x)\,dx$. They act as semantic placeholders that tell us which variables we are operating with respect to, but they also have a deeper meaning that can be quite enlightening.

Let's start with $f(x)\,dx$. The definition of the integral that you used in calculus approximates the area under the curve $f(x)$ by successively

11 Actually, the integral with respect to Lebesgue measure is somewhat more general in that one can integrate a wider class of functions.

smaller rectangles. The expresion $f(x)\,dx$ is the area of a rectangle with height $f(x)$ and "infinitesimal" width dx that extends from x to $x + dx$. Thus, $f(x)\,dx$ is an infinitesimal area under the curve which we sum up during the integration. In a multivariate integral, the term $f(x, y)\,dx\,dy$ is the infinitesimal volume of a rectangular solid of height $f(x, y)$, horizontal width dx from x to $x + dx$, and vertical width dy from y to $y + dy$. The integral again adds up these infinitesimal values, yielding the volume under the two dimensional surface $f(x, y)$.

An infinitesimal quantity is one that is not zero but is arbitrarily – infinitely – small. From a rigorous mathematical viewpoint, this is nonsensical. A real number is either small but non-zero or is zero; there is nothing in between. What is actually meant is a more complicated limiting process that defines the shrinking of the rectangles. However, the idea of the infinitesimal is conceptually powerful. It explains an integral as a sum of quantities in the right units (areas for a one-dimensional integral, volumes for a two-dimensional integral, and so forth). It provides mnemonics for changes of variables and other transformations. It allows us to treat curves that change on arbitrarily small scales as a sequence of instantaneous values.

So what does something like dx mean? It is the length of an infinitesimal interval starting at x. When we see dx, it tells us three things: which variable we are dealing with, what value of that variable we are considering, and the length of an infinitesimal at that value. When we integrate with respect to measures in the subsections below, you will see an expression of the form $\mu(dx)$ where μ is a measure and dx is an infinitesimal. This is a short-hand notation for $\mu\Big([x, x + dx[\Big)$, that is, the measure of an infinitesimal interval starting at x. While more precise, the longer version is rather cumbersome, especially when repeated many many times. The beauty of infinitesimals is that this shorthand holds together nicely. For example, for Lebesgue measure, $\mu_{\mathsf{Leb}}\Big([x, x + dx[\Big) = dx$, so we write $\mathit{Lebesgue}(dx) = dx$, and the integral with respect to Lebesgue measure, defined below, reduces to the ordinary integral.[11] In two and three dimensions, we use expressions of the form $\mu(dx\,dy)$ and $\mu(dx\,dy\,dz)$, which mean respectively $\mu\Big([x, x + dx[\ \times\ [y, y + dy[\Big)$ and $\mu\Big([x, x + dx[\times [y, y + dy[\times [z, z + dz[\Big)$. You can see why we prefer the shorthand. But again things work out because for Lebesgue measure on $\mathbb{R}^2$ and $\mathbb{R}^3$ we have $\mu_{\mathsf{Leb}}(dx\,dy) \equiv \mu_{\mathsf{Leb}}\Big([x, x+dx[\ \times\ [y, y+dy[\Big) = dx\,dy$ and $\mu_{\mathsf{Leb}}(dx\,dy\,dz) \equiv \mu_{\mathsf{Leb}}\Big([x, x+dx[\times[y, y+dy[\times[z, z+dz[\Big) = dx\,dy\,dz$.

12 I say signed area because, as you may recall, when f dips below the x-axis the corresponding area counts negatively towards the integral.

1.4. Integration with respect to a Measure I: The Idea

In calculus, you learned that for a real-valued function f defined on a subset of $\mathbb{R}$, the *integral* of f, $\int f(x)\,dx$, is the (signed) [12] area between the curve $f(x)$ and the x-axis. Using measures, we can define a more powerful integral that reduces exactly to the Reimann integral where appropriate. Here, we give a conceptual overview of the integral with examples that show you what it means and how to compute it. In section 1.5, we describe the properties of the integral These two subsections should be sufficient for a strong working understanding of how to integrate. Section 1.6 gives the details of how this integral is constructed, which can help one understand exactly what is being computed.

We will consider functions $f\colon \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is any set and a measure μ on $\mathcal{X}$ and compute the integral of f with respect to μ, denoted $\int f(x)\,\mu(dx)$. This is a real number associated with the function. The "x" in $\int f(x)\,\mu(dx)$ is just a dummy variable and can be replaced by anything else. One way to think about the integral is as a sum of infinitesimal areas $f(x)\,\mu(dx)$ over all x. Here, $f(x)$ is a height and $\mu(dx)$ is a corresponding width; the measure enters only in the latter.

We begin with four examples which show that there is little new here.

1. For any function f,

$$\int g(x)\,\mu_{\mathsf{Leb}}(dx) = \int g(x)\,dx.$$

   In other words, integrals with respect to Lebesgue measure are computed exactly the same way that you learned in calculus. In terms of infinitesimals, $\mu_{\mathsf{Leb}}(dx) = dx$, so the "sum of infinitesimals" we get is the same as you are used to.

2. For any function f,

$$\int g(x)\,\delta_{\{\alpha\}}(dx) = g(\alpha).$$

Consider the infinitesimal $\delta_{\{\alpha\}}(dx)$ as x ranges over $\mathbb{R}$. If $x \neq \alpha$, then the infinitesimal interval $[x, x + dx[$ does not contain α, so

$$\delta_{\{\alpha\}}(dx) \equiv \delta_{\{\alpha\}}\Big([x, x + dx[\Big) = 0.$$

If $x = \alpha$, $\delta_{\{\alpha\}}(dx) \equiv \delta_{\{\alpha\}}\Big([x, x + dx[\Big) = 1$. Thus, when we add up all of the infinitesimals, we get $g(\alpha) \cdot 1$.

3. For any function f,

$$\int g(x)\, \delta_{\mathbb{Z}}(dx) = \sum_{i \in \mathbb{Z}} g(i).$$

Again, consider the infinitesimal $\delta_{\mathbb{Z}}(dx)$ as x ranges over $\mathbb{R}$. If $x \notin \mathbb{Z}$, then the infinitesimal interval $[x, x + dx)$ does not intersect $\mathbb{Z}$ so $\delta_{\mathbb{Z}}(dx) \equiv \delta_{\mathbb{Z}}\Big([x, x + dx[\Big) = 0$. If $x \in \mathbb{Z}$, $\delta_{\mathbb{Z}}(dx) \equiv \delta_{\mathbb{Z}}\Big([x, x + dx[\Big) = 1$. (It is not more than one because an infinitesimal interval can contain at most one integer.) The term $g(x)\, \delta_{\mathbb{Z}}(dx)$ is thus $g(x)$ if $x \in \mathbb{Z}$ and 0 otherwise. When we add up all of the infinitesimals over x, we get the sum above.

4. Suppose $\mathcal{C}$ is a countable set. We can define counting measure on $\mathcal{C}$ to map $\mathcal{A}$ to $\#\mathcal{A} \cap \mathcal{C}$. For any function f,

$$\int g(x)\, \delta_{\mathcal{C}}(dx) = \sum_{v \in \mathcal{C}} g(v),$$

using the same basic argument as in the above example.

What we have just learned is that integrals with respect to Lebesgue measure are just ordinary integrals and that integrals with respect to Counting measure is just ordinary summation. So, in essence, we have a syntactic device for unifying sums and integrals to the same notation.

The next step is to consider measures built from Lebesgue and Counting measure.

1. Suppose μ is a measure that satisfies $\mu(dx) = f(x)\, \mu_{\mathsf{Leb}}(dx)$, then for any function g,

$$\int g(x)\, \mu(dx) = \int g(x)\, f(x)\, \mu_{\mathsf{Leb}}(dx) = \int g(x)\, f(x)\, dx.$$

As described in section 3, we say that f is the density of μ with respect to Lebesgue measure in this case.

2. Suppose μ is a measure that satisfies $\mu(dx) = p(x)\,\delta_{\mathcal{C}}(dx)$ for a countable set $\mathcal{C}$, then for any function g,

$$\int g(x)\,\mu(dx) = \int g(x)\,p(x)\,\delta_{\mathbb{C}}(dx) = \sum_{v \in \mathcal{C}} g(v)\,f(v).$$

Here, we say that p is the density of μ with respect to Counting measure on $\mathcal{C}$.

Except for #, all of the measures we will use in this course will have a density with respect to Lebesgue measure or with respect to Counting measure on some set. Therefore, integration with respect to measures is nothing new except for a nicely unified way to express summation.

1.5. Properties of the Integral

A function is said to be *integrable* with respect to μ if $\int |f(x)|\,\mu(dx) < \infty$. An integrable function has a well-defined and finite integral. If $f \geq 0$, the integral is always well-defined but may be ∞. (Consider 1 integrated with respect to Lebesgue measure on $\mathbb{R}$.)

As in calculus, we sometimes wish to integrate over a subset of the domain. Suppose μ is a measure on $\mathcal{X}$, $\mathcal{A} \subset \mathcal{X}$, and g is a real-valued function on $\mathcal{X}$. We define the integral of g over the set $\mathcal{A}$, denoted by $\int_{\mathcal{A}} g(x)\,\mu(dx)$, to be simply the integral of g using only those x that are elements of $\mathcal{A}$. This is given by

$$\int_{\mathcal{A}} g(x)\,\mu(dx) = \int g(x)\,1_{\mathcal{A}}(x)\,\mu(dx).$$

Notice that the points outside of $\mathcal{A}$ contribute nothing to the integral. Because $1_{\mathcal{X}} = 1$, we have that $\int_{\mathcal{X}} f(x)\,\mu(dx) = \int f(x)\,\mu(dx)$. Again, this is syntactic only; there is nothing new here.

For a practical working mastery of the integral, what you really need to know is how to manipulate it. Integrals of any kind are just sums, and they satisfy all the properties that a sum does. For example, $|2-3+4-5| \leq |2|+|-3|+|4|+|-5|$ because the absolute values in the latter prevent the negatives from reducing the total sum. This same idea translates to the property $|\int f(x)\,\mu(dx)| \leq \int |f(x)|\,\mu(dx)$; again, the absolute values inside the integral prevent any cancelation that would reduce the magnitude of the result. Similarly, $1+4+7+9 \leq 2+6+8+12$ because each term on the left is $\leq$ the corresponding term on the right. This translates to the property $f \le g$ implies $\int f(x)\,\mu(dx) \le \int g(x)\,\mu(dx)$.

When the integral is defined as in the next subsection, the following properties hold for every measure μ. In the list below, μ is a measure on $\mathcal{X}$, $\mathcal{A}, \mathcal{B} \subset \mathcal{X}$, $c \in \mathbb{R}$ and f and g are integrable functions.

1. *Constant Functions.* $\int_{\mathcal{A}} c\,\mu(dx) = c \cdot \mu(\mathcal{A})$.
2. *Linearity.*

$$\int_{\mathcal{A}} cf(x)\,\mu(dx) = c\int_{\mathcal{A}} f(x)\,\mu(dx)$$
$$\int_{\mathcal{A}} (f(x) + g(x))\,\mu(dx) = \int_{\mathcal{A}} f(x)\,\mu(dx) + \int_{\mathcal{A}} g(x)\,\mu(dx).$$

3. *Monotonicity.* If $f \le g$, then $\int_{\mathcal{A}} f(x)\,\mu(dx) \le \int_{\mathcal{A}} g(x)\,\mu(dx)$ for every set $\mathcal{A}$. This implies the following common special cases.
   - If $f \ge 0$, $\int f(x)\,\mu(dx) \ge 0$.
   - If $f \ge 0$ and $\mathcal{A} \subset \mathcal{B}$, $\int_{\mathcal{A}} f(x)\,\mu(dx) \le \int_{\mathcal{B}} f(x)\,\mu(dx)$.
4. *Null Sets.* If $\mu(\mathcal{A}) = 0$, then $\int_{\mathcal{A}} f(x)\,\mu(dx) = 0$.
5. *Absolute Values.* $|\int f(x)\,\mu(dx)| \le \int |f(x)|\,\mu(dx)$.
6. *Monotone Convergence.* If $0 \le f_1 \le f_2 \le \cdots$ is an increasing sequence of integrable functions that converge to f, then

$$\lim_{k\to\infty} \int f_k(x)\,\mu(dx) = \int f(x)\,\mu(dx).$$

7. *Linearity in region of integration.* If $\mathcal{A} \cap \mathcal{B} = \emptyset$,

$$\int_{\mathcal{A}\cup\mathcal{B}} f(x)\,\mu(dx) = \int_{\mathcal{A}} f(x)\,\mu(dx) + \int_{\mathcal{B}} f(x)\,\mu(dx).$$

To keep all of the above properties straight in your mind, it may help to think about the signed area under a curve. Each property translates into a physical statement about areas. For example, Monotonicity says that if one area contains another, it is bigger. The Absolute Values property results from the fact that any negative values subtract from the total area; the absolute values eliminate this. The Monotone Convergence property says that if we fill up the area under a curve f, building up in a series of steps, then this area gets closer and closer to the total area under f.

1.6. Integration with respect to a Measure II: The Details

To define an integral means to create a consistent and specific algorithm for computing the integral of any appropriate function.[13] As you know from experience, some functions are easier to integrate than others. For example, to integrate a multiple of the indicator function $1_{[0,1]}$ requires nothing special—we simply multiply height by width to get the area under the curve. For functions like this, we can define the integral to our complete satisfaction, so we start out defining the integral of simple functions like this. The next trick is to extend the definition of the integral to more complicated functions in a way that the integral behaves like we expect it to. Below, we define the integral with respect to a measure in three steps.

The integral introduced in calculus is called the Reimann integral. It is defined by approximating the area under the curve $f(x)$ with smaller and smaller rectangles. To define $\int f(x)\,\mu(x)$, we will do almost the same thing, but our method for approximating a function f is a bit different than filling in under the curve with rectangles. [14] On your first reading, you may want to skip this subsection. Come back to it later for deeper insights when you feel more comfortable with the integrals.

Step 1. Define the integral for simple functions.

A *simple function*, is any function that takes only a finite number of different values. All constant functions are simple functions because they take only one value. The indicator function of a set $\mathcal{A} \subset \mathcal{X}$ is a simple function because $1_{\mathcal{A}}$ takes only two values: 0 and 1. Any constant c times an indicator function, that is $c \cdot 1_{\mathcal{A}}$, is also a simple function because it takes at most the values 0 and c. Similarly, given disjoint sets $\mathcal{A}_1$ and $\mathcal{A}_2$, the linear combination $c_1 \cdot 1_{\mathcal{A}_1} + c_2 \cdot 1_{\mathcal{A}_2}$ is a simple function which can take the values 0, c_1, c_2.[15] In fact, any simple function f can be expressed as a linear combination of a finite number of indicator functions. That is, if f is *any* simple function on $\mathcal{X}$, then there is some finite integer n, there are non-zero constants $c_1, \ldots, c_n$, and there are *disjoint* sets $\mathcal{A}_1, \ldots, \mathcal{A}_n \subset \mathcal{X}$ such that

$$f = c_1 \cdot 1_{\mathcal{A}_1} + c_2 \cdot 1_{\mathcal{A}_2} + \cdots + c_n \cdot 1_{\mathcal{A}_n}.$$

Because the sets are disjoint, this function takes the $n+1$ values $0, c_1, \ldots, c_n$.

13 For mathematicians to be satisfied, the algorithm need only work in principle; we may not be able to implement it in practice.

14 The construction of the Reimann integral is usually rushed in calculus courses, and it is hardly the most exciting part of the material in any case. If you do not remember how it goes, do not worry. The method we will use is actually much simpler.

15 If $\mathcal{A}_1$ and $\mathcal{A}_2$ were not disjoint, we could define $\mathcal{B}_1 = \mathcal{A}_1 - \mathcal{A}_2$, $\mathcal{B}_2 = \mathcal{A}_2 - \mathcal{A}_1$, and $\mathcal{B}_3 = \mathcal{A}_1 \cap \mathcal{A}_2$ and then the function is equal to $c_1 \cdot 1_{\mathcal{B}_1} + c_2 \cdot 1_{\mathcal{B}_2} + (c_1 + c_2) \cdot 1_{\mathcal{B}_3}$. Make sure you understand why this is true. (Draw a picture in one example.)

16 ATTN Metapost figure with simple function and areas illustrated.

17 Remember: The variable of integration is a dummy variable.

So, if $f\colon \mathcal{X} \to \mathbb{R}$ is a simple function as just defined, what is the integral of f? Keeping in mind that the integral represents the (signed) area under the curve, all we have to do for each segment is to multiply the height c by the width $\mu(\mathcal{A})$ and add them all up. Specifically, if

$$f = c_1 \cdot 1_{\mathcal{A}_1} + c_2 \cdot 1_{\mathcal{A}_2} + \cdots + c_n \cdot 1_{\mathcal{A}_n}. \tag{M.1}$$

then

$$\int f\, \mu(dx) = c_1 \cdot \mu(\mathcal{A}_1) + c_2 \cdot \mu(\mathcal{A}_2) + \cdots + c_n \cdot \mu(1_{\mathcal{A}_n}). \tag{M.2}$$

This integral can be a real number, ∞, or $-\infty$ depending on the sets and the constants defining f. Take a look at the figure 16 and the following examples to get a feel for the argument,

Example 1. Suppose that μ is Lebesgue measure and that f is defined as follows:

$$f(x) = \begin{cases} 1 & \text{if } -1 < x < 1 \\ 3 & \text{if } 1 \le x < 4 \\ 5 & \text{if } 4 \le x < 10 \\ 2 & \text{if } -4 < x \le -1 \\ 4 & \text{if } -10 < x \le -4 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\int f(t)\, \mu(dt)$.[17]

The definition of f above makes clear that it is a simple function because it takes only 6 possible values. Let's express it in the form of equation (M.1):

$$f = 1_{]-1,1[} + 3 \cdot 1_{[1,4[} + 5 \cdot 1_{[4,10[} + 2 \cdot 1_{]-4,-1]} + 4 \cdot 1_{]-10,4]}. \tag{M.3}$$

Because μ is Lebesgue measure on $\mathbb{R}$, $\mu(dx) = dx$; so, $\int f\, \mu(dx) = \int f\, dx$, the familiar integral from calculus. The area under f can now be computed (draw a picture) by calculating the area under each separate piece and adding all the areas together. As above, we have

$$\begin{aligned} \int f\, \mu(dx) &= \mu\Big(]-1,1[\Big) + 3 \cdot \mu\Big([1,4[\Big) + 5 \cdot \mu\Big([4,10[\Big) + \\ &\qquad 2 \cdot \mu\Big(]-4,-1]\Big) + 4 \cdot \mu\Big(]-10,4]\Big) \\ &= 2 + 3 \cdot 3 + 5 \cdot 6 + 2 \cdot 3 + 4 \cdot 6 \\ &= 71. \end{aligned}$$

Example 2. Suppose f is defined as above and that $\mu = \delta_{\mathbb{Z}}$, counting measure on $\mathbb{Z}$. Find $\int f(x)\,\mu(dx)$.

The representation of f in terms of indicators is given in equation (M.3), and we need only compute the measures of the sets as in equation (M.2); see also the first line of the previous display. Since $\mu(\mathcal{A})$ is the number of integers in $\mathcal{A}$, we have that $\mu\Big(]-1,1[\Big) = 1$, $\mu\Big([1,4[\Big) = 3$, $\mu\Big([4,10[\Big) = 6$, and so forth. Hence, $\int f\,\mu(dx) = 1\cdot 1 + 3\cdot 3 + 5\cdot 6 + 2\cdot 3 + 4\cdot 6 = 70$.

Example 3. Suppose that f is any simple function and that $\mu = \delta_{\{0\}}$, a point mass at 0. Find $\int f(y)\,\mu(dy)$.

Here, μ of any set is 0 unless that set contains 0. So in the representation of f of equation (M.1), only one of the sets $\mathcal{A}_i$ will contain 0. The measure of this set will be 1, and the constant multiplying the corresonding indicator is just $f(0)$. It follows that $\int f(y)\,\delta_{\{0\}}(dy) = f(0)$ for any f.

Step 2. Define the integral for general non-negative functions, approximating the general function by simple functions.

Having learned how to compute the integral of any simple function, we turn next to a general non-negative function $f\colon \mathcal{X} \to [0,\infty[$. The idea is that we can approximate[18] such an f arbitrarily well by some non-negative simple function that is $\le f$, as illustrated in the marginal figure. The argument[19] that this claim is true is enlightening, but you can skip it without worry on the first reading.

If f and g are non-negative functions on the same domain then the integrals of f and g include no "negative areas"; hence, if $g \le f$, we should have $\int g(x)\,\mu(dx) \le \int f(x)\,\mu(dx)$. Indeed, this is true as can be seen from the definition above if f and g are both simple functions. We define the integral to make this true statement true. Thus, if $f\colon \mathcal{X} \to [0,\infty[$ is a general functions and $0 \le s \le f$ is a simple function, we require that $\int s(x)\,\mu(dx) \le \int f(x)\,\mu(dx)$. The closer that s approximates f, the closer we expect $\int s(x)\,\mu(dx)$ and $\int f(x)\,\mu(dx)$ to be.

This is how we compute the integral of f. Loosely speaking, we define $\int f(x)\,\mu(dx)$ to be the largest value $\int s(x)\,\mu(dx)$ for a simple function $0 \le s \le f$. To be more precise, we define the integral $\int f(x)\,\mu(dx)$ to be the smallest value I such that $\int s(x)\,\mu(dx) \le I$ for all simple functions

18 ATTN: Metapost figure of successive approximations.

19 To see this, choose some large integers k, and define $\mathcal{A}_i = f^{-1}\Big([i/2^k, (i+1[\,/2^k)\Big)$ for $0 \le i < k2^k$ and $\mathcal{A}_{k2^k} = f^{-1}\Big([k,\infty[\Big)$, where as usual $f^{-1}(\mathcal{B})$ refers to the inverse image of the set $\mathcal{B}$. Define a simple function s_k by

$$s_k(x) = \sum_{i=0}^{k2^k} \left(\frac{i}{2^k}\right) 1_{\mathcal{A}_i}.$$

Since $s(x) = i/2^k \le f(x)$ for all $x \in \mathcal{A}_i$, we have $s \le f$. Also, $|s_k(x) - f(x)| \le 2^{-k}$ except possibly when $x \in \mathcal{A}_{k2^k}$. and as k gets larger and larger, each x is eventually outside $\mathcal{A}_{k2^k}$.

$0 \le s \le f$.[20]

20 This is the least upper bound (supremum) as defined in note 10. In other words, $\int f(x)\,\mu(dx)$ is the least upper bound of the set of numbers $\int s(x)\,\mu(dx)$ such that s is simple and $0 \le s \le f$. We need the least upper bound because there might not be a largest $\int s(x)\,\mu(dx)$.

Step 3. Define the integral for general real-valued functions by separately integrating the positive and negative parts of the function.

Finally, if $f \colon \mathcal{X} \to \mathbb{R}$ is a general function, we can define its *positive part* f^+ and its *negative part* f^- by[21]

21 ATTN: Metapost Figure. Function with positive and negative parts

$$
\begin{aligned}
f^+(x) &= \max(f(x), 0) \\
f^-(x) &= \max(-f(x), 0).
\end{aligned}
$$

Notice that both f^+ and f^- are non-negative functions (so we know how to integrate them) and that $f = f^+ - f^-$. Consequently, we define

$$\int f(x)\,\mu(dx) = \int f^+(x)\,\mu(dx) - \int f^-(x)\,\mu(dx).$$

This is a well-defined number (possibly infinite) if and only if at least one of f^+ and f^- has a finite integral.

This defines the integral of a function f with respect to a measure μ. The advantage of seeing the construction is that we understand exactly what the integral means, but we can operate just fine using only the examples and properties described in the previous two subsections. Fortunately, one can show that this construction implies that all of those properties are true.

2. Some Important Examples

We saw a variety of examples of measures above. However, every measure that we will encounter in this course will be built from one of the three most important examples, which we summarize in the following:

1. Cardinality. This is denoted by $\#(\mathcal{A})$ for any set $\mathcal{A}$ and can be defined as a measure on any domain.
2. Counting measure.

   Suppose $\mathcal{C}$ is a countable set. Then, the counting measure with respect to $\mathcal{C}$, $\delta_{\mathcal{C}}$, is defined by $\delta_{\mathcal{C}}(\mathcal{A}) = \#\mathcal{A} \cap \mathcal{C}$. Counting measure on the integers, $\delta_{\mathbb{Z}}$, and the point mass at a point x, $\delta_{\{x\}}$, are special cases.

   Integrals with respect to counting measure are just sums. That is,

   $$\int f(x)\,\delta_{\mathcal{C}}(dx) = \sum_{z \in \mathcal{C}} f(z).$$

Or, put another way, sums and ordinary integrals are both just integrals, but with respect to different measures.

3. Lebesgue measure. This can be defined on any $\mathbb{R}^k$. Lebesgue measure on $\mathbb{R}$ gives length, Lebesgue measure on $\mathbb{R}^2$ gives area, and Lebesgue measure on $\mathbb{R}^3$ gives volume, all three very familiar. We always denote Lebesgue measure by μ_{Leb} and let context indicate the appropriate domain, but it will be rare that we have to refer to it explicitly.

   Lebesgue measure on any $\mathbb{R}^k$ satisfies the following properties:

   A. A hyper-rectangle[22] $]a_1, b_1[\, \times \cdots \times \,]a_k, b_k[$ in $\mathbb{R}^k$ has Lebesgue measure $(b_1 - a_1) \cdots (b_k - a_k)$.
   B. If $\mathcal{C}$ is any countable set, $\mu_{\mathsf{Leb}}(\mathcal{C}) = 0$.
   C. If $\mathcal{B}$ is the result of shifting the set $\mathcal{A}$ by some amount (e.g., $[4, 5]$ is obtained by shifting $[0, 1]$ by 4 in $\mathbb{R}$), then $\mu_{\mathsf{Leb}}(\mathcal{B}) = \mu_{\mathsf{Leb}}(\mathcal{A})$.

   Integrals with respect to Lebesge measure are just the ordinary integrals you are familiar with from calculus. In particular, $\mu_{\mathsf{Leb}}(dx) = dx$ on $\mathbb{R}$, $\mu_{\mathsf{Leb}}(dx\, dy) = dx\, dy$ on $\mathbb{R}^2$, $\mu_{\mathsf{Leb}}(dx\, dy\, dz) = dx\, dy\, dz$ on $\mathbb{R}^3$, and $\mu_{\mathsf{Leb}}(dx_1 \cdots dx_k) = dx_1 \cdots dx_k$ on $\mathbb{R}^k$.

22 A hyper-rectangle in $\mathbb{R}$ is just an interval; in $\mathbb{R}^2$ it is a rectangle; and in $\mathbb{R}^3$ it is a rectangular solid.

3. How to make new measures from old ones

Sums and Multiples. Consider the point mass measures at 0 and 1, $\delta_{\{0\}}$ and $\delta_{\{1\}}$, and construct a two new measures on $\mathbb{R}$ $\mu = \delta_{\{0\}} + \delta_{\{1\}}$ defined by

$$\mu(\mathcal{A}) = \delta_{\{0\}}(\mathcal{A}) + \delta_{\{1\}}(\mathcal{A}),$$

and $\nu = 4\delta_{\{0\}}$ defined by

$$\nu(\mathcal{A}) = 4 \cdot \delta_{\{0\}}(\mathcal{A}),$$

for all $\mathcal{A} \subset \mathcal{X}$. We say that μ is the sum of $\delta_{\{0\}}$ and $\delta_{\{1\}}$ and that ν is a multiple of $\delta_{\{0\}}$ by a factor or 4. The measure μ counts how many elements of $\{0, 1\}$ are in its argument. (We will denote counting measures like this by a common convention; specifically, $\mu = \delta_{\{0,1\}}$ in this case.) Notice that $\delta_{\mathbb{Z}}$, the counting measure of the integers can be re-expressed as a sum of all pointmasses $\delta_{\{i\}}$ for integers i:

$$\delta_{\mathbb{Z}} = \sum_{i=-\infty}^{\infty} \delta_{\{i\}}.$$

Similarly, we can also write ν as a sum

$$\nu = \delta_{\{0\}} + \delta_{\{0\}} + \delta_{\{0\}} + \delta_{\{0\}}.$$

Because we have defined measures to be non-negative, a multiple of a measure is only a measure if the constant factor is non-negative.

By combining the operations of summation and multiplication, we can construct measures like $4\delta_{\{0\}} + 3\delta_{\{1\}}$ that are linear combinations of other measures (with positive coefficients). For example, the geometric measure in example ATTN above can be written as $\sum_{i=0}^{\infty} r^i \, \delta_{\{i\}}$.

Restriction to a Subset. Consider Lebesgue measure on $\mathbb{R}$, μ_{Leb}. We can define a new measure on $]0,1[$ which maps a set $\mathcal{A} \subset\ ]0,1[$ to $\mu_{\mathsf{Leb}}(\mathcal{A})$. This is effectively the same measure but it is restricted to subsets of $]0,1[$ only. In general, suppose μ is a measure on $\mathcal{X}$ and $\mathcal{B} \subset \mathcal{X}$. We can define a new measure on $\mathcal{B}$ which maps $\mathcal{A} \subset \mathcal{B}$ to $\mu(\mathcal{A})$. This is called the restriction of μ to the set $\mathcal{B}$.

The Measure Induced by a Function. Consider Lebesgue measure on $\mathbb{R}$ and the function $g\colon \mathbb{R}to\ ]0,\infty[$ defined by $g(x) = 10^x$. The function g connects subsets of $(0,\infty)$ to subsets of $\mathbb{R}$ via the inverse image. For example, the inverse image of $]1,100[\ \subset\ ]0,\infty[$ is $]0,2[\ \subset \mathbb{R}$ because these are the exponents of 10 that produce the values 1 through 100. We can use this relationship to construct a new measure on $(0,\infty)$: $\nu(\mathcal{A}) = \mu_{\mathsf{Leb}}(g^{-1}(\mathcal{A}))$ for $\mathcal{A} \subset\ ]0,\infty[$.

Thought Question Find $\nu\big(]1,2[\big)$, $\nu\big(]1,100[\big)$, $\nu\big(]1,10000[\big)$, and $\nu\big(]100,10000[\big)$.

In general, suppose μ is a measure on $\mathcal{X}$ and $g\colon \mathcal{X} \to \mathcal{Y}$ is a function. We can use μ and g to define a new measure ν on $\mathcal{Y}$ by

$$\nu(\mathcal{A}) = \mu(g^{-1}(\mathcal{A}))$$

for $\mathcal{A} \subset \mathcal{Y}$. We call this the measure *induced from μ by the function g*. Notice that we can think of the inverse image g^{-1} as mapping subsets of $\mathcal{Y}$ to subsets of $\mathcal{X}$, that is as a function $2^{\mathcal{Y}} \to 2^{\mathcal{X}}$. In that sense, the measure induced from μ by g can be written as $\mu \circ g^{-1}$. This technique allows us to carry over a measure defined on one space to another space with no extra work. It is very important in probability theory as we will see in Chapter 3.

Thought Question Suppose $h(x) = x^2$. Describe $\mu_{\text{Leb}} \circ h^{-1}$.

If $\nu = \mu \circ g^{-1}$ for some function, g, then we can make a useful substitution in integrals that illuminates the relationship between μ and ν. Specifically, $f(y)\,\nu(dy) = f(g(x))\,\mu(dx)$, for $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, so wherever we see one, we can substitute the other. In other words, for any function $f\colon \mathcal{Y} \to \mathbb{R}$,

$$\int f(y)\,\nu(dy) = \int f(g(x))\,\mu(dx).$$

Integrating a Density. Suppose μ is a measure on $\mathcal{X}$ and $f\colon \mathcal{X} \to \mathbb{R}$. We can define a new measure ν on $\mathcal{X}$ as follows:

$$\nu(\mathcal{A}) = \int_{\mathcal{A}} f(x)\,\mu(dx). \tag{M.4}$$

(Recall the definition of the integral over a set above.) Because of the linearity and convergence properties of the integral, ν satisfies all the properties of a measure including countable additivity.

We say that f is the *density* of the measure ν with respect to the measure μ. The name "density" seems an odd choice, but it does have a strong physical intuition which we will explore fully in Chapter 3.

We often turn this phrase around. If ν and μ are two measures for which the equation (M.4) holds for every $\mathcal{A} \subset \mathcal{X}$, we say that ν has a density f with respect to μ. This implies two useful results:

- $\mu(\mathcal{A}) = 0$ implies $\nu(\mathcal{A}) = 0$.
- $\nu(dx) = f(x)\,\mu(dx)$. In other words, we can substitute for $\nu(dx)$ in integrals: for real-valued g on $\mathcal{X}$, $\int g(x)\,\nu(dx) = \int g(x)\,f(x)\,\mu(dx)$.

Densities play an important role in probability theory.

Example 1. Suppose, given μ, we define ν by $\nu(\mathcal{A}) = \mu(\mathcal{A} \cap \mathcal{B})$ for some fixed set $\mathcal{B} \subset \mathcal{X}$. Then, ν assigns 0 measure to anything outside of $\mathcal{B}$; for any subset of $\mathcal{B}$, μ and ν are the same. We can rewrite ν as follows

$$\nu(\mathcal{A}) = \int_{\mathcal{A}} 1_{\mathcal{B}}(x)\,\mu(dx).$$

Thus, $1_{\mathcal{B}}(x)$ is the density of ν with respect to μ. That the density is 0 outside $\mathcal{B}$ reflects the fact that ν assigns 0 measure to anything outside $\mathcal{B}$. That the density is 1 inside $\mathcal{B}$ reflects that fact that μ and ν are the same for any subset of $\mathcal{B}$. We could write this relationship in terms of infinitesimals as follows:

$$\nu(dx) = \begin{cases} \mu(dx) & \text{if } x \in \mathcal{B} \\ 0 & \text{if } x \notin \mathcal{B}. \end{cases}$$

Example 2. The geometric measure defined above ATTN(xref) has a density with respect to $\delta_{\mathbb{Z}}$, and the density is the function p given by

$$p(x) = \begin{cases} r^x & \text{if } x \in \mathbb{Z} \\ 0 & \text{otherwise.} \end{cases}$$

Example 3. The point mass $\delta_{\{0\}}$ also has a density with respect to $\delta_{\mathbb{Z}}$. To see this, notice that

$$\int g(x)\, \delta_{\mathbb{Z}}(dx) = \sum_{i=-\infty}^{\infty} g(i).$$

If p is the function given by

$$p(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise,} \end{cases}$$

then

$$\int g(x)\, p(x)\, \delta_{\mathbb{Z}}(dx) = g(0).$$

By taking $g = 1_{\mathcal{A}}$ for a subset $\mathcal{A}$ of $\mathcal{X}$, we get

$$\int_{\mathcal{A}} p(x)\, \delta_{\mathbb{Z}}(dx) = \delta_{\{0\}}(\mathcal{A}).$$

Thus, p is the density of $\delta_{\{0\}}$ with respect to $\delta_{\mathbb{Z}}$.

Example 4. Define a measure μ_G on $\mathbb{R}$ by

$$\mu_G(\mathcal{A}) = \int_{\mathcal{A}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\, dx.$$

This is called standard Gaussian measure. It has a density $f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}$ with respect to Lebesgue measure on $\mathbb{R}$. Notice that $\mu_G(\mathbb{R}) = 1$ because the density decays so rapidly for large values of $|x|$.

4. Other Types of Measures

Suppose that μ is a measure on $\mathcal{X}$.

If $\mu(\mathcal{X}) = \infty$, we say that μ is an infinite measure. If $\mu(\mathcal{X}) < \infty$, we say that μ is a finite measure. If $\mu(\mathcal{X}) = 1$, we say that μ is a probability measure. We will talk much more about probability measures throughout the course.

If there is a countable set $\mathcal{S}$ such that $\mu(\mathcal{X} - \mathcal{S}) = 0$, we say that μ is a *discrete* measure. Equivalently, μ has a density with respect to counting measure on $\mathcal{S}$. If μ has a density with respect to Lebesgue measure, we say that μ is a *continuous* measure. If μ is neither continuous nor discrete, we say that μ is a mixed measure.

NOTATION

N

1. Conventions

1.1. Generic Objects

Each class of objects (sets, functions, measures, numbers) has certain symbols that are used to represent them this will be consistent. Similarly delimiters such as [] and {} are only used in one way, for intervals and sets respectively.

1.2. Subscripts

Subscripts are used to specify objects given by generic symbols. The subscript can be an index in a collection, the name of a random variable, or an arbitrary letter or word. The subscript is part of the name

1.3. Operators and Functions

There is a difference between operators and functions. Expected values, variances, and entropies, for instance, are denoted by operators which act on objects of some definite type. Functions are defined precisely in Appendix F.

1.4. Naming

While it is often possible to give explicit expressions for functions or sets or distributions, we will often give them names to shorten our reference. Thus, we name random variables and sets and give names to families of distributions that arise frequently.

2. Common Errors

Some of the most common errors involve the following:

- Having objects of different type on opposite sides of the equal sign. For instance, $X = \mathcal{A}$ where X is a real-valued random variable and $\mathcal{A}$ is an event or $X(\omega) = Y$ where X and Y are random variables but $X(\omega)$ is a number.
- Treating local (dummy) variables as though they have meaning outside their scope. For example, indices of summation and variables of integration are only defined within the sum or integral. Similarly, if we define a function say by "Let $h(x) = x^2$", the variable used in that definition is local to the definition. It is analogous to the definition of a function in a programming language.
- Using algebraic variables where random variables are required. If x is a number then $\{ x = 2 \}$ say is either all of Ω or the empty set.
- Taking the expected value of an event $\mathsf{E}\mathcal{A}$ or the probability of a random variable $\mathsf{P}(X)$. These operators act on specific types of objects, putting the wrong type gives an expression that makes no sense.
- Having probabilities < 0 or > 1. Can't happen.
- Treating the different possible values of a random variable as different random variable. If X is a random variable that takes values 3, 7, or 9. Then, we can consider events $\{ X = 3 \}$, $\{ X = 7 \}$, $\{ X = 9 \}$, and related events or indicators derived thereof. These are different values of the same random variable.

A useful motto to keep in mind when reviewing your own work is "Think like a compiler". Make sure your types match and that everything you use has been properly defined. Watch out for local variables that are used outside their scope.

3. Vectors and All That

We often use vector valued random variables, but we tend to treat them implicitly as a collection of real valued random variables. The definitions in Chapter 3 all work in either case, but it can sometimes be confusing keeping the distinctions straight. This is worth reviewing carefully. Remember if X is $\mathbb{R}^n$ valued, we can always write it as $(X_1, \ldots, X_n)$ for real-valued component variables $X_1, \ldots, X_n$. Describing the joint distribution of the components is equivalent to describing the distribution of the vector.

4. Mathematical Preliminaries

4.1. Sets

See Appendix S for details and definitions.

Basics	
Mnemonic	When you see $\{\}$s, you know you are dealing with a set
Names	Scripted, capital roman letter letters e.g., $\mathcal{A}, \mathcal{E}, \mathcal{F}, \mathcal{S}$
Defining	$\{$`<candidate points>` such that `<condition>`$\}$ e.g., $\{x \in \mathbb{R}$ such that $0 < x < 2\}$
Operations	
Union	$\bigcup$ e.g, $\mathcal{A} \cup \mathcal{B}$, $\bigcup\limits_{i=1}^{\infty} \mathcal{A}_i$
Intersection	$\bigcap$ e.g, $\mathcal{A} \cap \mathcal{B}$, $\bigcap\limits_{i=1}^{\infty} \mathcal{A}_i$
Complement	$\operatorname{compl}()$ e.g, $\operatorname{compl}(\mathcal{A})$, $\operatorname{compl}(\mathcal{B} \cap [0,1])$
Difference	$-$ (e.g, $\mathcal{A} - \mathcal{B}$)
Intervals	
Closed	$[a, b] = \{x \in \mathbb{R}$ such that $a \leq x \leq b\}$
Open	$]a, b[= \{x \in \mathbb{R}$ such that $a < x < b\}$
Half-Closed	$[a, b[= \{x \in \mathbb{R}$ such that $a \leq x < b\}$
Half-Open	$]a, b] = \{x \in \mathbb{R}$ such that $a < x \leq b\}$

Named Spaces

Integers	$\mathbb{Z}$
Positive Integers	$\mathbb{Z}_+$
Non-negative Integers	$\mathbb{Z}_\oplus$
Rational Numbers	$\mathbb{Q}$
Real Numbers	$\mathbb{R}$
Positive Reals	$\mathbb{R}_+$
Non-negative Reals	$\mathbb{R}_\oplus$
Complex Numbers	$\mathbb{C}$
Numeric Variables	Lower case roman letters, e.g., $x, y, z, a, b, r, s, t, u, v, w$ used for reals and i, j, k, l, m, n for integers
Numeric Parameters	Lower case greek or roman letters (e.g., λ, γ)

4.2. Functions

Generic	Lower case, roman letters (e.g., f, g, h)
Random Variables	Upper case, roman letters (e.g., X, Y, Z)
Evaluated	Paretheses contains argument list (e.g., $g(x)$)
Declaring	`<name>`: `<domain>` $\to$ `<codomain>`

5. Foundations of Probability Models

Outcomes

Outcome space	Ω
Elementary outcome	ω
Events	{`<Defining Condition>`}, e.g., $\{\, X \in \mathcal{A} \,\}$ or $\{\, Y = 4 \text{ and } Z > 10 \,\}$

Random Variables

Names	Upper case roman letters (e.g., X, Y, Z)
Evaluated	$X(\omega)$ for arbitrary $\omega \in \Omega$, e.g., $X(\omega)$
Observed	Pass selected elementary outcome ω^*
Functions	$g(X)$ is the random variable $\omega \mapsto g(X(\omega))$, e.g., X^2, e^X, $g(X)$
Indicators	$1_{\texttt{<event>}}$, e.g., $1_{\{\, X>2 \,\}}$ or $1_{\mathcal{A}}$
Abbreviation	RV or rv

6. Expected Values and Probabilities

Expected Value	Operates on random variables, e.g., $\mathsf{E}X$ Parenthese needed only for grouping, e.g., $\mathsf{E}(X+Y)$ or $\mathsf{E}(X^2)$
Probabilities	Operates on events, e.g., $\mathsf{P}(\mathcal{A})$ or $\mathsf{P}\{\,X > 10\,\}$ Parenthese needed only for grouping, e.g., $\mathsf{P}(\{\,X > 10\,\} \cup \{\,Y < 10\,\})$
Variance	Operates on random variables, e.g., $\mathsf{Var}\,X$ or $\mathsf{Var}\,(Z+U)$
Entropy	Operates on random variables, e.g., $\mathsf{H}X$ or $\mathsf{H}(Z+U)$
Conditional on RV	A random variable $\mathsf{E}(\texttt{<random variable>} \mid \texttt{<random variable>})$, e.g., $\mathsf{E}(X \mid Y)$
Conditional on Event	A number $\mathsf{E}(\texttt{<random variable>} \mid \texttt{<event>})$, e.g., $\mathsf{E}(X \mid \{\,Y > 2\,\})$
Conditional Probability	$\mathsf{P}(\texttt{<event>} \mid \texttt{<event>})$, e.g., $\mathsf{P}(\{\,X > 10\,\} \mid \{\,Y > 2\,\})$

6.1. Conditional Expected Values

Suppose that X and Y are random variables and $\mathcal{A}$ and $\mathcal{B}$ are events.

$\mathsf{E}(Y \mid X)$ denotes the *random variable* computed as a function of X that best predicts Y.

We define $\mathsf{E}(Y \mid X \ \mathsf{near}\ x)$ to be the value that this random variable takes when X is near x. If X is discrete, we can write this as $\mathsf{E}(Y \mid X = x)$.

$\mathsf{E}(X \mid \mathcal{A})$ is the conditional expected value of X given that the event $\mathcal{A}$ occured. $\mathsf{E}(X \mid \mathcal{A}) = \mathsf{E}(X \mid 1_{\mathcal{A}} = 1) = \mathsf{E}(X\,1_{\mathcal{A}})/\mathsf{P}(\mathcal{A})$.

$\mathsf{P}(\mathcal{B} \mid \mathcal{A})$ is the conditional probability of $\mathcal{B}$ given that the event $\mathcal{A}$ occured. $\mathsf{P}(\mathcal{B} \mid \mathcal{A}) = \mathsf{E}(1_{\mathcal{B}} \mid \mathcal{A})$.

7. Distributions

Generic	D
of a RV	D_X is the distribution of X
PMF	p_X is the PMF of discrete rv X
PDF	f_X is the PDF of continuous rv X
CDF	F_X is the CDF of a rv X
SF	S_X is the SF of a rv X
Joint	If $Z = (X_1, \ldots, X_n)$ is a vector-valued rv

	the distribution of Z is the same as the *joint* distribution of $X_1, \ldots, X_n$
PMF	$\mathsf{p}_{X_1,\ldots,X_n}(u_1, \ldots, u_n) = \mathsf{P}\{\, X_1 = u_1 \text{ and } \cdots \text{ and } X_n = u_n \,\}$
PDF	$\mathsf{f}_{X_1,\ldots,X_n}(u_1, \ldots, u_n) du_1 \cdots du_n = \mathsf{P}\{\, X_1 \text{ near } u_1 \text{ and } \cdots \text{ and } X_n \text{ near } u_n \,\}$
CDF	$\mathsf{F}_{X_1,\ldots,X_n}(u_1, \ldots, u_n) = \mathsf{P}\{\, X_1 \le u_1 \text{ and } \cdots \text{ and } X_n \le u_n \,\}$
Near	The near operator used like $=$, a near b, but represents infinitesimal not exact equality

7.1. The near Operator and Infinitesimals

The near operator is the key to seeing the common structure in the calculations we will do across the discrete-continuous divide. Review carefully the practical definitions of the PDF and conditional PDF given in the text (Chapters 3 and 4) in terms of the near operator. Remember that for discrete variables, near and $=$ are interchangeable. Study the algorithm for translating discrete expressions to the continuous case and vice versa in Section 3.5.

7.2. Conditional Distributions

PMF	$\mathsf{p}_{X\mid Y}(u \mid v) = \mathsf{P}\{\, X = u \mid Y \text{ near } v \,\}$
PDF	$\mathsf{f}_{X\mid Y}(u \mid v)\, du = \mathsf{P}\{\, X \text{ near } u \mid Y \text{ near } v \,\}$
CDF	$\mathsf{F}_{X\mid Y}(u \mid v) = \mathsf{P}\{\, X \le u \mid Y \text{ near } v \,\}$
SF	$\mathsf{S}_{X\mid Y}(u \mid v) = \mathsf{P}\{\, X > u \mid Y \text{ near } v \,\}$

To interpret the above, remember that for any discrete random variable D, $\{\, D \text{ near } v \,\} = \{\, D = v \,\}$. Both the above definitions thus work whether Y is discrete or continuous. If X and Y are discrete, for instance,

$$\mathsf{p}_{X\mid Y}(u \mid v) = \mathsf{P}\{\, X = u \mid Y = v \,\}.$$

Conditional distributions are similarly defined for vector-valued random variables as well Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ be vector-valued rvs.

If X is discrete, we have the conditional PMF of X given Y:

$$\mathsf{p}_{X_1 \ldots X_n \mid Y_1 \ldots Y_m}(u_1, \ldots, u_n \mid v_1, \ldots, v_m) = \mathsf{P}\{\, X_1 = u_1 \text{ and } \cdots \text{ and } X_n = u_n \mid Y_1 \text{ near } v_1 \text{ and } \cdots \text{ and } Y_m \text{ near } v_m \,\}.$$

If X is continuous, we have the conditional PDF of X given Y:

$$\mathsf{f}_{X_1 \dots X_n \mid Y_1 \dots Y_m}(u_1, \dots, u_n \mid v_1, \dots, v_m)\, du_1 \cdots du_n =$$
$$\mathsf{P}\{\, X_1 \text{ near } u_1 \text{ and } \cdots \text{ and } X_n \text{ near } u_n \mid Y_1 \text{ near } v_1 \text{ and } \cdots \text{ and } Y_m \text{ near } v_m \,\}.$$

7.3. Canonical Distributions

Certain distributions arise frequently in applications, in various guises. These distributions can be arranged in families where different distributions in the same family have the same mathematical form up to differences in certain *parameters.* One specifies a canonical distribution by giving the family name (in roman type with the first character capitalized) and the values of one or more parameters (listed within angle brackets $\langle\,\rangle$).

Examples of the most frequently used families are as follows:

- Bernoulli$\langle p \rangle$
- Binomial$\langle n, p \rangle$
- Geometric$\langle p \rangle$
- NegativeBinomial$\langle k, p \rangle$
- Poisson$\langle \lambda \rangle$
- DiscreteUniform$\langle x_1, \dots, x_n \rangle$
- Exponential$\langle \lambda \rangle$
- Gamma$\langle \alpha, \lambda \rangle$
- Uniform$\langle a, b \rangle$
- Normal$\langle \mu, \sigma^2 \rangle$
- Beta$\langle \alpha, \beta \rangle$
- PoissonArrival$\langle \lambda \rangle$

SETS

1. Fundamentals

A *set* is a collection of distinct objects, called *elements* of the set. The elements of a set can be anything whatsoever—thus, {`emu`, `frog`, `newt`} and {`?`,`!`,`'`} are valid sets, the first consisting of the English words for specific animals and the second of punctuation marks—as long as it is understood that no two elements of a set can be exactly the same. It is well-known to mathematicians that this informal definition of sets can lead to some bizarre logical contradictions (consider the set of all sets ?!@*#?), but in this course, we will never need to venture near such strange beasts.[1]

Sets are the most fundamental objects in mathematics in the sense that all other mathematical ideas are either built from sets or use them.[2] Indeed, many entire branches of mathematics, including probability theory, involve studying the properties of a set adorned with some additional structure. For example, we can define two operations $+$ and $\times$ on the set of real numbers[3] that each take a pair of numbers to a new number via certain rules (addition and multiplication). The set of real numbers adorned with these operations forms what mathematicians call a *field*. In probability theory, we consider a set of outcomes for some random experiment and adorn this set with a measure of how likely the different outcomes are to occur.

Given the idea of a set, we would like to be able to do something with it. At minimum we must be able to define new sets, determine if two sets are equal or if one lies "inside" the other, decide whether or not an object is an element of a particular set, and describe the number of elements a set contains.

1 Mathematicians have developed a complete axiomatic theory of sets that circumvents these problems.

2 What about numbers, you might ask? It turns out that numbers themselves can be defined as sets. For example, the integer zero is defined as the empty set; the integer one is the set whose single member is the empty set; the integer k is the set whose k members are the sets representing $0, 1, \ldots, k-1$. This same construction can be extended to all integers, then to the rational and real numbers, and beyond.

3 For now, I will assume that the set of real numbers and the set of integers are passingly familiar; the former contains all "numbers" between $> -\infty$ and $< \infty$ and the latter contains those real numbers with no fractional part. See Section 2 below for a more complete discussion.

In mathematical text, we will denote sets by enclosing the collection of elements in braces $\{\}$'s. Throughout this course, we will use braces only to represent sets, so whenever you see something so enclosed you will know what type of object it is. If the set has only a few elements, it is easy to specify it completely by listing the elements explicitly. For example, $\{2,4,6,8\}$ is the set of positive, even integers less than 10. There is no preferred order in the listing of a set since the set embodies the *collection* of elements not an ordered list of elements. Hence, $\{8,2,4,6\}$, $\{4,8,2,6\}$, $\{6,8,2,4\}$, and any other permuted listing all represent the same set. Most sets have too many elements to list explicitly, so we describe them by defining the conditions that an object must satisfy to be an element of the set. This description is enclosed in $\{\}$'s as usual. For example, $\{$real numbers x such that $x < 2$ and $x > -16\}$ is the set of all real numbers that are strictly smaller than 2 and strictly bigger than -16, and $\{$integers p such that p is prime$\}$ is the set of prime numbers. The variables x and p in these descriptions are *dummy variables*; they have meaning only within the definition of the set as labels for an arbitrary candidate (in this case, a real number or an integer) that make it easier to specify the defining conditions. Thus, the set $\{$real numbers z such that $-16 < z < 2\}$ is exactly the same set as the former because it defines the same set of numbers. It is irrelevant that in the latter we use a different dummy variable to specify the defining conditions or that we state those conditions differently.

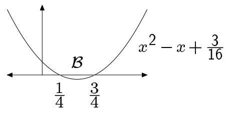
Once we start defining sets, it makes our lives easier if we can give them more succinct names by which to refer to them later. (Naming is an activity that both mathematicians and computer scientists enjoy.) We name sets by associating a symbol with the set using an $=$ sign. For example, if I write "let $\mathcal{A} = \{$real numbers u such that $(u+1)^2 < 1\}$" then I can use $\mathcal{A}$ or the complete specification of the set interchangeably. If I am going to refer to this set many times, it is much more convenient to use $\mathcal{A}$ than to repeat the definition. Notice that, on the face of it, the symbol $\mathcal{A}$ gives no indication that it refers to a set. This is a disadvantage of using names: we have to discern from context (or just remember) what a particular symbol means. The same phenomenon occurs in computer programs: the variables `x` and `A` can stand for objects of any type, only context and memory help us determine what they are. Programmers make this task easier for someone reading their program by using mnemonic names and by sticking to a naming convention that makes it easier to identify types. We will do the same thing; see Appendix N for a complete description of our conventions. In particularly, we will use script Roman capital letters to denote sets.

When defining a set, one must be precise and unambiguous. Pictures of sets, although not always precise, are nonetheless often helpful for understanding what a definition really means. For example, consider two sets $\mathcal{B} = \{\text{real numbers } x \text{ such that } x^2 - x < -3/16\}$[4] and $\mathcal{C} = \{\text{real ordered pairs } (x, y) \text{ such that } 0 < x \leq y < 1\}$.[5] With some practice, it will become easier to picture these sets in your mind, but for now, take a look at the corresponding figures. In the first case, $\mathcal{B}$ contains all real numbers bigger than 1/4 and smaller than 3/4 as can be seen by referring to the superimposed quadratic curve $x^2 - x + 3/16 = (x - 1/4)(x - 3/4)$. In the second case, the set consists of ordered pairs, which we can interpret as coordinates in the x-y plane. The set $\mathcal{C}$ then contains all those points inside the unit square ($\{(x, y) \text{ such that } 0 < x < 1 \text{ and } 0 < y < 1\}$) above the main diagonal (i.e., x-coordinate smaller than y-coordinate). The pictured triangle results. Pictures of sets can also be used more abstractly to help understand the relationship between different sets.[6] In such pictures, called Venn diagrams, the specific definition of the sets is not important; you look at Venn diagrams to gain an instant visual interpretation. Two sets which share no elements in common can be thought of as "non-overlapping"; a set whose elements are all elements of a second set can be thought of as "contained in" that second set. Both of these relationships are immediately clear from the Venn diagram and hold true quite generally.

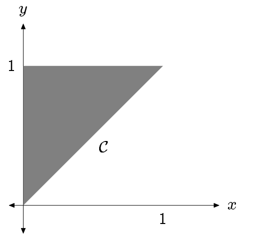
Thought Question Is $\{2, 2, 4, 6, 8\}$ a set according to the definition? Why or why not?

Thought Question Is $\{\{0\}, \{0, 1\}, \{0, 1, 2\}\}$ a set according to the definition? Why or why not?

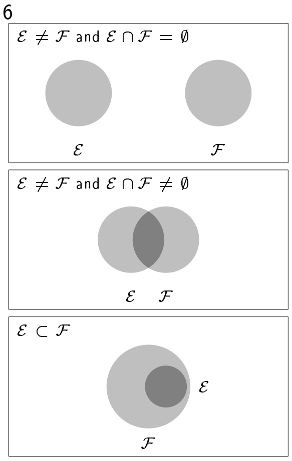
4



5



6



1.1. Comparing Sets

Two sets $\mathcal{E}$ and $\mathcal{F}$ are said to be equal,[7] denoted $\mathcal{E} = \mathcal{F}$, if and only if they have exactly the same elements. If $\mathcal{E}$ and $\mathcal{F}$ are not equal, we write $\mathcal{E} \neq \mathcal{F}$. Figure 6 illustrates three different ways that $\mathcal{E}$ and $\mathcal{F}$ can be unequal: non-overlapping, overlapping but containing distinct elements, or one contained in the other. Given two sets $\mathcal{E}$ and $\mathcal{F}$, we say that $\mathcal{E}$ is a *subset* of $\mathcal{F}$, denoted $\mathcal{E} \subset \mathcal{F}$, if every element of $\mathcal{E}$ is also an element of cF.[8] (For symmetry, we also say in this case that $\mathcal{F}$ is a *superset* of $\mathcal{E}$, denoted $\mathcal{F} \supset \mathcal{E}$.) As I have defined it above, the following two statements are true:

- Every set is a subset of itself.
- $\mathcal{E} = \mathcal{F}$ if and only if both $\mathcal{E} \subset \mathcal{F}$ and $\mathcal{F} \subset \mathcal{E}$ are true.

You should convince yourself that both these facts follow from the definition.[9]

1.2. Membership

To indicate that an object x is an element of a set $\mathcal{E}$, we write $x \in \mathcal{E}$. The operator $\in$ here is read in one of several equivalent ways "element of", "member of", "in", "belongs to", and so forth. Sometimes it is useful to indicate the opposite, that x is not an element of $\mathcal{E}$; we do this by writing $x \notin \mathcal{E}$.[10] Both operators $\in$ and $\notin$, take an object on the left and a set on the right. Thus, $2 \in \{2, 4, 6, 8\}$ and $0 \in \{\text{real numbers } z \text{ such that } -16 < z < 2\}$, $0 \notin \{2, 4, 6, 8\}$ and $2 \notin \{\text{real numbers } z \text{ such that } -16 < z < 2\}$.

A very important technique to master is the argument that logically proves two sets equal using the second bulleted result above. As an example, consider the two sets

$$\mathcal{E} = \Big\{\text{ordered pairs } (x, y) \text{ such that } x^2 + y^2 \leq 1\Big\}$$
$$\mathcal{F} = \{\text{ordered pairs } (r\cos\theta, r\sin\theta) \text{ such that } 0 \leq \theta < 2\pi, 0 \leq r \leq 1\}.$$

To show that $\mathcal{E} = \mathcal{F}$, I need to show that both $\mathcal{E} \subset \mathcal{F}$ and $\mathcal{F} \subset \mathcal{E}$ are true. To do this, I need to show that for any element $p \in \mathcal{E}$ that I choose, it is also true that $p \in \mathcal{F}$. Choose $p \in \mathcal{E}$. Because this is an arbitrary point,[11] all we know about it is that p satisfies the conditions defining $\mathcal{E}$; namely, $p = (x, y)$ with $x^2+y^2 \leq 1$. Let $r = \sqrt{x^2 + y^2}$ and let

7 This sentence is an example of mathematician-speak. When I say "two sets $\mathcal{E}$ and $\mathcal{F}$ ...", $\mathcal{E}$ and $\mathcal{F}$ are place holders for two arbitrary sets. They may even stand for the same set twice, as they do if the sets are equal.

8 The phrase "$\mathcal{E}$ is contained in $\mathcal{F}$" is often used as a loose synonym of "$\mathcal{E}$ is a subset of $\mathcal{F}$" because of the intuition offered by the Venn diagram.

9 If $\mathcal{E} \subset \mathcal{F}$ and $\mathcal{E} \neq \mathcal{F}$, we say that $\mathcal{E}$ is a *strict* subset of $\mathcal{F}$. This turns out to be a much less useful a relation in general than the subset relation as I defined it.

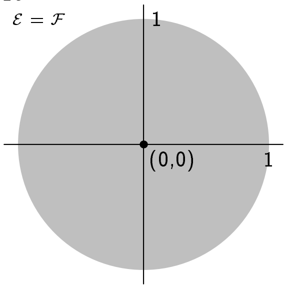
10 The slash through the $\in$ means "no" or "not", by analogy to the no left turn sign and its kin.

11 The word point here is synonymous with element; it is usually used when the set represents physical space.

$\theta = \arccos(x/r)$ (or 0 if $r = 0$). (In other words, θ is the angle such that $\cos\theta = x/r$ which we can find because x/r is between 0 and 1.) Then, $x = r\cos\theta$, and because $x^2+y^2 = r^2$ and $(\cos\theta)^2+(\sin\theta)^2 = 1$, we also have that $y = r\sin\theta$. Hence, $p = (x, y) = (r\cos\theta, r\sin\theta) \in \mathcal{F}$. Since the point p was a completely arbitrary element of $\mathcal{E}$, this argument[12] works for any point in $\mathcal{E}$. We have thus established that $\mathcal{E} \subset \mathcal{F}$. Now choose an arbitrary point $p \in \mathcal{F}$. Again because this is an arbitrary point, all we know about it is that p satisfies the conditions defining $\mathcal{F}$; namely, $p = (r\cos\theta, r\sin\theta)$ with $0 \le r \le 1$ and $0 \le \theta < 2\pi$. But as we saw above, the values $x = r\cos\theta$ and $y = r\sin\theta$ satisfy $x^2 + y^2 \le r^2$. Thus, $\mathcal{F} \subset \mathcal{E}$, and we have shown that $\mathcal{E} = \mathcal{F}$. (See the picture[13] of this set at the right.) It is quite common for two completely different descriptions of the same set to arise, and when this happens, we need to apply this argument to show that they are indeed the same. To summarize: given two sets $\mathcal{E}$ and $\mathcal{F}$. I have to show that (i) for any element $x \in \mathcal{E}$, it is also true that $x \in \mathcal{F}$, and (ii) for any element $x \in \mathcal{F}$, it is also true that $x \in \mathcal{E}$.

12 What we have done is constructed an algorithm that takes a point in $\mathcal{E}$ and shows that it is also in $\mathcal{F}$. Because the workings of the algorithm do not depend on any specific features of the point, we can apply it to every point in $\mathcal{E}$ with the same result, thus proving the claim.

13



Thought Question Is {real numbers x such that $0 < x < 1$} $\subset$ {real numbers x such that $0 \le x \le 1$}? Are the two sets equal?

1.3. The Size of a Set

The *cardinality* of a set is the number of items it contains. We denote the cardinality of a set $\mathcal{S}$ by $\#\mathcal{S}$. This can be finite or infinite but must be non-negative. We will see examples of both below, including an answer to the question of whether infinity plus 1 is bigger than infinity.

The size of a set can be defined in many ways. A *measure* is a mathematical object that quantifies the size of sets. Each distinct measure embodies a different way to assess how big a set is. Measures are the subject of Appendix M.

1.4. Upper and Lower Bounds

A set $\mathcal{A}$ of real numbers is said to be *bounded above* if there is a $u < \infty$ such that every $x \le u$ for every $x \in \mathcal{A}$. This means that if you move a certain finite distance to the right along the real line, then every element of $\mathcal{A}$ is to the left of you. A set $\mathcal{A}$ of real numbers is said to be *bounded below* if there is a $\ell > -\infty$ such that every $\ell \leq x$ for every $x \in \mathcal{A}$. This means that if you move a certain finite distance to the left along the real line, then every element of $\mathcal{A}$ is to the right of you. A set is *bounded* if it bounded above and bounded below, otherwise it is *unbounded*.

The number u in the definition of "bounded above" just given is called an upper bound of $\mathcal{A}$. If $\mathcal{A}$ is bounded above, an *upper bound* of $\mathcal{A}$ is any real number that is $\geq$ every element of $\mathcal{A}$. If $\mathcal{A}$ is not bounded above, then ∞ is the only upper bound for $\mathcal{A}$. The smallest possible upper bound for a set $\mathcal{A}$ of real numbers is called the *least upper bound* (or *supremum*) and is denoted $\sup \mathcal{A}$. The least upper bound of $\mathcal{A}$ satisfies the following:

- $\sup \mathcal{A}$ is an upper bound for $\mathcal{A}$,
- If u is any upper bound for $\mathcal{A}$, then $\sup \mathcal{A} \leq u$.

If $\mathcal{A}$ has a largest value then it equals $\sup \mathcal{A}$. For instance, any finite set of numbers has a maximum, which satisfies the two defining properties of the least upper bound. However, some sets do not have a largest value, including the set of all real numbers and $\{$real numbers x such that $0 < x < 1\}$.

Similarly, the number ℓ in the definition of "bounded below" just given is called a lower bound of $\mathcal{A}$. If $\mathcal{A}$ is bounded below, a *lower bound* of $\mathcal{A}$ is any real number that is $\leq$ every element of $\mathcal{A}$. If $\mathcal{A}$ is not bounded below, then $-\infty$ is the only lower bound for $\mathcal{A}$. The biggest possible lower bound for a set $\mathcal{A}$ of real numbers is called the *greatest lower bound* (or *infimum*) and is denoted $\inf \mathcal{A}$. The greatest lower bound of $\mathcal{A}$ satisfies the following:

- $\inf \mathcal{A}$ is a lower bound for $\mathcal{A}$,
- If ℓ is any lower bound for $\mathcal{A}$, then $\ell \leq \sup \mathcal{A}$.

If $\mathcal{A}$ has a smallest value then it equals $\inf \mathcal{A}$. For instance, any finite set of numbers has a minimum, which satisfies the two defining properties of the greatest lower bound. Analogously to the above, some sets do not have a smallest value, including the real numbers and $\{$real numbers x such that $0 < x < 1\}$.

Thought Question If $\mathcal{A}$ is a set of real numbers and $\#\mathcal{A} < \infty$, can $\mathcal{A}$ be unbounded? Why or why not?

Thought Question What is the least upper bound of the set of real numbers x such that $0 < x < 1$? What is the least upper bound of the set of real numbers x such that $0 \le x \le 1$? What is the least upper bound of the set of real numbers? What is $\inf \{0\}$? What is $\inf \{1/n \text{ such that } n \in \mathbb{Z}_+\}$?

Thought Question If $\mathcal{A}$ is a set of real numbers and $\mathcal{B} = \{-x \text{ such that } x \in \mathcal{A}\}$, how do $\inf \mathcal{A}$ and $\sup \mathcal{A}$ relate to $\sup \mathcal{B}$ and $\inf \mathcal{B}$?

1.5. Collections of Sets

We will frequently deal with a collection of many related sets at once, so it is useful to have a convention for naming them. Consider the following collections of sets:

- Sets $\mathcal{A}_k = \{2, \ldots, 2k\}$ for each integer $k \ge 1$,
- Sets $\mathcal{B}_x = \{\text{real } z \text{ such that } 0 \le z \le x\}$ for each real $x > 0$,
- Sets $\mathcal{C}_\aleph = \{a, \ldots, \aleph\}$ for each letter $\aleph$ in $\{a, \ldots, z\}$.

There are two ingredients that define each of these collections. First, we define an *index set*; the value of the index determines which set in the collection we are considering. In these cases, the integers, real numbers, and lowercase letters respectively are the index sets. Second, we define the one set in the collection for each element of the index set. In practice, we refer to the sets in the collection using a common letter symbol (e.g., $\mathcal{A}$, $\mathcal{B}$, etc.) with the index as a subscript. Typically, we will use the integers or non-negative integers as the index set, and in that case the collection is also called a sequence of sets. When we want to refer to the collection as a whole we will informally speak of the $\mathcal{A}_k$'s or the $\mathcal{B}_x$'s. If we want to be more formal in discussing an entire collection $\mathcal{A}_\gamma$ with $\gamma \in \mathcal{I}$, we will write $(\mathcal{A}_\gamma)_{\gamma \in \mathcal{I}}$.

2. Some Important Examples

2.1. The Empty Set

The simplest possible set is the set that contains no elements, called the *empty set*. Because it is so important, the empty set has its own symbol $\emptyset$ which we will use because it is so common. Naturally, $\#\emptyset = 0$. The empty set is a subset of every set.[14]

14 Why? Consider a set $\mathcal{A}$. Because $\emptyset$ contains no elements, it is true by default that every element of $\emptyset$ is also an element of $\mathcal{A}$. This argument works for any set $\mathcal{A}$.

2.2. Numbers

The next most important sets, and those perhaps most familiar to you, are sets of numbers. The first numbers you probably came across were the positive integers 1, 2, 3, ..., also called the natural numbers or counting numbers. Certain fundamental sets of numbers, like the positive integers, arise so frequently that we give them special symbols. We denote the set of positive integers by $\mathbb{Z}_+$, where the $\mathbb{Z}$ represents for the German word for integer "Zahl" and the '+' indicates positive numbers. Once you learned about zero, your horizons had expanded to the set of non-negative integers, or whole numbers, $\mathbb{Z}_\oplus = \{0, 1, 2, \ldots\}$, the $\oplus$ here gives a mnemonic for "positive and zero = non-negative". Having discovered zero, the next step is to realize that negative numbers make sense. (If you borrow 4 apples from me and eat them, how many apples do you have?) Combine the negative integers[15] $-1, -2, -3, \ldots$ with the non-negative integers and you get the *integers* $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. The sets $\mathbb{Z}$, $\mathbb{Z}_+$, and $\mathbb{Z}_\oplus$ are all infinite sets, and none of these three sets contains a largest element,[16] though $\mathbb{Z}_+$ and $\mathbb{Z}_\oplus$ do have a smallest element.

The set of *rational numbers* $\mathbb{Q}$ is defined by

$$\mathbb{Q} = \{p/q \text{ such that } p \in \mathbb{Z} \text{ and } q \in \mathbb{Z}_+\}.$$

Remember that the elements of a set are unique, so the ratios p/q and $2p/2q$, which are equal, appear only once in $\mathbb{Q}$. The set $\mathbb{Q}$ contains the numbers whose decimal representation has a fixed (or repeating) number of digits. All numbers representable on a computer are rational, for instance. Because there are infinitely many integers, $\mathbb{Q}$ is also an infinite set. What may be more surprising is the following property of $\mathbb{Q}$: between any two rational numbers $q_1 < q_2$, there are infinitely many more rational numbers.[17] Try to devise an argument to prove this.

However, most numbers, including such favorites as e, π, and $\sqrt{2}$, cannot be represented as rational numbers; they require an infinite, non-repeating sequence of digits in their decimal representation. These are called irrational numbers.[18] The set of *real numbers* $\mathbb{R}$ consists of all the rational and all the irrational numbers. We often refer to $\mathbb{R}$ colloquially as "the real line". As we will see below, the cardinality of $\mathbb{R}$ is much greater than the cardinality of $\mathbb{Z}$; that is, there are many

15 We could denote these by $\mathbb{Z}_-$ if we ever need them, although the negative integers alone do not come up too often.

16 Argument: Given an arbitrary integer k, I can always form $k+1$.

17 This is called the Archimedean property, after the Greek philosopher Archimedes.

18 These numbers made mathematicians just a little crazy.

more real numbers than integers. (Oddly enough, the sets $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{Z}_\oplus$, and $\mathbb{Q}$ all have the same cardinality; we will see why later on.) It is also useful to define $\mathbb{R}_+$ to be the set of positive real numbers and $\mathbb{R}_\oplus$ to be the set of non-negative real numbers. By definition, we have that $\mathbb{R}_+ \subset \mathbb{R}_\oplus \subset \mathbb{R}$.

The set of *complex numbers* $\mathbb{C}$ consists of all combinations $a + \imath b$ where $a, b \in \mathbb{R}$ and $\imath = \sqrt{-1}$ (electrical engineers use $\jmath$). By taking $b = 0$, we just get the real numbers back, so $\mathbb{R} \subset \mathbb{C}$. Complex numbers arise as the solutions to certain polynomial equations. For example, the equation $x^2 + 1 = 0$ has two complex solutions, $\pm\imath$, but no real solutions.

These sets that we have introduced are contained one within the other: $\mathbb{Z}_+ \subset \mathbb{Z}_\oplus \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$. Most of the time, we will deal with real numbers, and if a number's type is unspecified, you can assume that it is real by default. Hence, if we write $\{x \text{ such that } x > 30\}$ we mean that $x \in \mathbb{R}$. (Of course, it does not hurt to be precise, so feel free to write $\{x \in \mathbb{R} \text{ such that } x > 30\}$. And remember to indicate when you mean something specific, like $x \in \mathbb{Z}$.)

Thought Question Is $\sqrt{25}$ a rational number?

Thought Question Can you find a real number that is not a complex number? Can you find a complex number that is not a real number?

2.3. Intervals

An *interval* of real numbers is a subset of $\mathbb{R}$ containing all numbers in a certain range. There are four types of intervals distinguished only by which boundary points are part of the set. Closed intervals $[a, b]$ are defined for $a, b \in \mathbb{R}$ with $a < b$ by

$$[a, b] = \{x \in \mathbb{R} \text{ such that } a \le x \le b\}.$$

The orientation of the brackets is intended to help you remember that closed intervals contain both their end-points. A special case that arises often is the *unit interval* $[0, 1]$. Open intervals $]a, b[$ contain neither of their endpoints; they are defined for $a, b \in \mathbb{R}$ with $a < b$ by

$$]a, b[= \{x \in \mathbb{R} \text{ such that } a < x < b\}.$$

Again, note how the bracket orientation encodes the definition. (We can also define an open interval with $a = -\infty$ and $b = \infty$; this $]-\infty, \infty[$ is just the set of real numbers.) The two types of half-open intervals are $[a, b[$ and $]a, b]$. The former are defined for $a < b$ with $a \in R$ and b in $\mathbb{R}$ or equal to ∞; the latter are defined for $a < b$ with $a \in R$ or equal to ∞ and b in $\mathbb{R}$. These are given by

$$\begin{aligned}
[a, b[&= \{x \in \mathbb{R} \text{ such that } a \leq x < b\} \\
]a, b] &= \{x \in \mathbb{R} \text{ such that } a < x \leq b\} .
\end{aligned}$$

Thus, $[0, 1[$, $[0, 2[$, and $[0, \infty[$ are all valid intervals; the last of which contains all non-negative numbers in $\mathbb{R}$. Occasionally, we wish to allow for the possibility that a number is ∞ or $-\infty$, so we will use infinities in interval specifications as a convenient shorthand. For example, $[0, \infty]$ is the set of non-negative real numbers union the set $\{\infty\}$.

Thought Question If a is a real number, what is the set $[a, a]$?

Thought Question Under what situations is an interval of the form $]a, b[$ bounded?

Thought Question What are the least upper bound and greatest lower bound of the interval $[a, b[$?

2.4. Vectors

A *vector*[19] is an ordered sequence of objects. We denote a vector as a comma separated list surrounded by *parentheses*. For example, $(2, 4)$, $(1, 2, 5)$, and $(3, 4.2, 19, -20, 4.24, 10^{-10})$ are vectors of real numbers with length 2, 3, and 6. If a vector has length k, we call it a k-vector. In the example above, $(1, 2, 5)$ is a 3-vector and $(2, 4)$ is a 2-vector. Another name for a 2-vector is *ordered pair*; an ordered pair of real numbers is just the familiar object from your mathematics classes used to denote two-dimensional coordinates. A 1-vector is just the object itself (the parentheses are superfluous in this case). It is also possible to have vectors of infinite length such as $(1, e, e^2, e^3, \ldots)$ or $(x_1, x_2, x_3, \ldots)$; an infinite vector is also called a *sequence*. Unlike sets, the order in which a vector's elements are listed matters. While the sets $\{1, 2, 5\}$, $\{5, 2, 1\}$, and $\{2, 1, 5\}$ are all the same, the vectors $(1, 2, 5)$, $(5, 2, 1)$, and $(2, 1, 5)$

19 This construct is also called a *tuple*.

are all different. We often need distinguish between vectors and single numbers in isolation. The latter are called *scalars*.

The elements in the list given by a vector are called its *components*. If x is a vector then by convention x_i is the ith component. Conversely, given numbers (i.e., scalars) $x_1, x_2, \ldots, x_n$, we can construct the vector x by putting these components together: $x = (x_1, x_2, \ldots, x_n)$. A key fact about vectors is that we can always move interchangeably between using the vector as an object itself (e.g., as the symbol x) and working with the components (e.g.,the list $x_1, \ldots, x_n$).

It is often useful to consider sets of vectors. Consider three important examples. First, $\mathbb{R}^1$ is the set of 1-vectors of real numbers, but since a 1-vector (x) is just x, $\mathbb{R}^1 = \mathbb{R}$, the real line. Second, $\mathbb{R}^2$ is the set of 2-vectors – ordered pairs – of real numbers; each element is of the form (x, y) for some real numbers x and y. This yields the "x-y plane" of two-dimensional coordinates you have used in your mathematics classes. Each vector (x, y) corresponds to a point in the plane. Third, $\mathbb{R}^3$ is the set of 3-vectors – ordered triples – of real numbers; each element is of the form (x, y, z) for some real numbers x, y, z. Each such vector gives the coordinates of a point in three-dimensional space; $\mathbb{R}^3$ is the set of all such points. Of course, there is no need to stop at three. For any $k \in \mathbb{Z}_+$, we define $\mathbb{R}^k$ to be the set of k-vectors of real numbers. Each element of $\mathbb{R}^k$ is of the form $(x_1, \ldots, x_k)$, where the components are real numbers. We also write $\mathbb{R}^\infty$ to denote the set of *sequences* of real numbers. For example, $(1, 1/2, 1/3, 1/4, \ldots)$ is an ∞-vector whose nth component is the number $1/n$.

Vectors have the property that they can be added together componentwise and they can be scaled by a number (thus the name scalar). If we have to vectors of the same length $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, then the vector $x + y$ has components $x_i + y_i$. That is,

$$x + y = (x_1 + y_1, \ldots, x_n + y_n).$$

Similarly, the vector cx has components cx_i. That is,

$$cx = (cx_1, \ldots, cx_n).$$

for example, $(2, 1, 3) + (1, -1, 3) = (3, 0, 3)$ and $3(1, 2, 3) = (3, 6, 9)$. The two operations can also be combined. For instance, if x and y are two vectors and $0 \leq \lambda \leq 1$ is a scalar, then $\lambda x + (1-\lambda)y$ is the point λ of the way along the line between x and y. Similarly, $x - y$ is the vector $x + (-1)y$.

Vectors in $\mathbb{R}^k$ (for $k \in \mathbb{Z}_+$) are often represented by arrows from the origin of the coordinate system to the point represented by the vector. This can be an intuitively useful heuristic; it represents a vector as a direction and a magnitude. The vector $x+y$ is the arrow from the origin to the point obtained by following the x arrow to x and then following a copy of the y arrow shifted to x. Similarly, $x - y$ is the arrow from y to x. (You can remember this because $y + x - y = x$.)

We can also sometimes compare vectors in $\mathbb{R}^k$. If $a, b \in \mathbb{R}^k$, we say that $a \leq b$ if $a_i \leq b_i$ for all $i = 1, \ldots, k$. Analogous definitions work for $<$, $>$, and $\geq$. Notice that two vectors need not be comparable; that is, for $k > 1$, both $a \leq b$ and $a > b$ can be false.

If $x \in \mathbb{R}^n$, we can denote the *Euclidean length* of the vector by $|x|$, where

$$|x| = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{S.1}$$

In $\mathbb{R}$, this reduces to the ordinary absolute value. For a vector x, $|x|$ is the distance between the point (in n-dimensional space) represented by x and the origin of the coordinate system. Indeed, we can use this to measure the distance between any two points. If $x, y \in \mathbb{R}^n$, then

$$|x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{S.2}$$

In $\mathbb{R}$, this reduces to the ordinary absolute difference between two numbers.

2.5. Balls

Balls are sets in $\mathbb{R}^k$ consisting of points at most a fixed Euclidean distance away from a specified point. The open ball around x of radius r is the set $\mathcal{B}_x(r)$ which contains within r of x except for those on the spherical surface.

$$\mathcal{B}_x(r) = \{x \text{ such that } |x - u| < r\}. \tag{S.4}$$

The analogous closed ball $\overline{\mathcal{B}}_x(r)$ contains in addition the spherical surface.

$$\overline{\mathcal{B}}_x(r) = \{x \text{ such that } |x - u| \leq r\}. \tag{S.4}$$

In $\mathbb{R}$, the open ball is an interval of the form $[x, x+r[$ and the closed ball is an interval of the form $[x, x+r]$.

2.6. Product Sets

If $\mathcal{A}$ is a set and k is a positive integer, we use $\mathcal{A}^k$ to denote[20] the set of k-vectors of elements of $\mathcal{A}$. Each element of $\mathcal{A}^k$ can be written in the form $(a_1, \ldots, a_k)$ where the components a_i are all in $\mathcal{A}$.

20 Note that $\mathcal{A}^1 = \mathcal{A}$. We will see later that this exponent notation actually makes sense in other ways as well.

Thought Question Describe what the set $[0,1]^3$ looks like.

We can generalize this notion by considering vectors whose elements are obtained from an arbitrary sequence of sets. For example, consider the "rectangle" in the x-y plane consisting of points with $0 \le x \le 10$ and $0 \le y \le 5$. This rectangle set consists of ordered pairs (x, y) with $x \in [0, 10]$ and $y \in [0, 5]$. Similarly, the points (x, y, z) in $\mathbb{R}^3$ with $x \in\ ]0,1[$, $y \in\ ]0,2[$, and $z \in\ ]0,4[$ form an open rectangular solid. Such vector sets are called *product sets*. We denote the product by a $\times$ operator between the sets. For example, the rectangle above is the set $[0, 10] \times [0, 5]$. In general, if $\mathcal{A}_1, \ldots, \mathcal{A}_m$ are any sets, then the product set $\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_m$ is defined as follows:

$$\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_m = \{(a_1, \ldots, a_m) \text{ such that } a_i \in \mathcal{A}_i, i = 1, \ldots, m\}.$$

The vector sets we saw previously are also product sets. For instance, $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$. We get $\mathcal{A}^k$ back from this operation by

$$\mathcal{A}^k = \underbrace{\mathcal{A} \times \cdots \times \mathcal{A}}_{k \text{ times}}.$$

Hyper-Rectangles. A special case of product sets are products of intervals. A product of k intervals forms what is called a k-dimensional hyper-rectangle. For example, $[0,1] \times [0,2] \times [0,3] \times [0,4]$ is a four-dimensional subset of $\mathbb{R}^4$. A 1-dimensional hyper-rectangle is just an interval in $\mathbb{R}$; a 2-dimensional hyper-rectangle is just a rectangle in $\mathbb{R}^2$; a 3-dimensional hyper-rectangle is just a rectangular solid in $\mathbb{R}^3$. The higher dimensional cases serve as a useful generalization. In fact, we use interval notation to define hyper-rectangles in $\mathbb{R}^k$. For example, if $a, b \in \mathbb{R}^k$ with $b \ge a$, $[a, b]$ is the set of all points x such that $a \le x \le b$, meaning $a_i \le x_i \le b_i$ for $i = 1, \ldots, k$. The analogous definition works for $[a, b[$, $]a, b]$, and $ooI(a, b)$ replacing $\le$ with $<$ for open-ended cases.

3. How to make new sets from old ones

Unions. The word "union" implies a bringing together, and that is exactly what you should think when considering the union of sets. See the examples in the figure.[21] If $\mathcal{A}$ and $\mathcal{B}$ are sets, then their *union* $\mathcal{A} \cup \mathcal{B}$ is the set containing those elements that belong to either $\mathcal{A}$ *or* $\mathcal{B}$:

$$x \in \mathcal{A} \cup \mathcal{B} \text{ if and only if } x \in \mathcal{A} \text{ or } x \in \mathcal{B}.$$

The "or" here is a logical or; it is true if x is in either or both sets. As usual with the definition of sets, no element is repeated; as the logical formulation above shows, even if an object is an element of both $\mathcal{A}$ and $\mathcal{B}$ it occurs only once in the union. Hence, $[0,2] \cup [1,4] = [0,4]$ and $\{a,b,c\} \cup \{c,r,g\} = \{a,b,c,r,g\}$.

Suppose now that I have three sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. One can form the union of these three sets in three different ways:

1. $(\mathcal{A} \cup \mathcal{B}) \cup \mathcal{C}$,
2. $\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C})$,
3. $(\mathcal{A} \cup \mathcal{C}) \cup \mathcal{B}$.

Fortunately, by the definition above, all of these lead to exactly the same set,[22] which we can confidently refer to as *the* union of $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$.

$$x \in \mathcal{A} \cup \mathcal{B} \cup \mathcal{C} \text{ if and only if } x \in \mathcal{A} \text{ or } x \in \mathcal{B} \text{ or } x \in \mathcal{C}.$$

What works for two sets or three sets, works for any number of sets. As an example, consider the collection of sets $\mathcal{A}_1, \ldots, \mathcal{A}_{10}$ defined by $\mathcal{A}_i = \{2^{i-2}, 2^{i-1}, 2^i\}$. The union of the sets in this collection is denoted $\bigcup_{i=1}^{10} \mathcal{A}_i$. Logically, a number is in the union if it is in at least one of the $\mathcal{A}_i$; hence, $\bigcup_{i=1}^{10} \mathcal{A}_i = \{1/2, 1, 2, 4, 8, \ldots, 1024\}$. Check this carefully yourself. The same logic also works if the collection of sets is infinite. Consider for instance the infinite collection of sets defined by $\mathcal{A}_i = [1/(i+1), 1[$ for $i \in Z_+$. The union of all the $\mathcal{A}_i$'s is $\bigcup_{i=1}^{\infty} \mathcal{A}_i$ which contains any point that is in at least one of the $\mathcal{A}_i$'s.

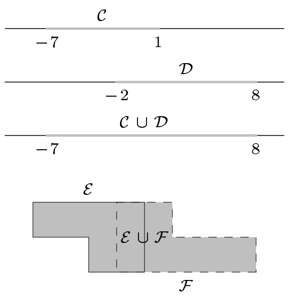
Thought Question Is $0 \in \bigcup_{i=1}^{\infty} \mathcal{A}_i$? Is 1? What is this union?

In general if $\mathcal{A}_1, \mathcal{A}_2, \ldots$ is a collection of sets, finite or infinite, we have

$$x \in \bigcup_i \mathcal{A}_i \text{ if and only if } x \in \mathcal{A}_i \text{ for at least one } i.$$

21 $\mathcal{A} = \{0,1,4\}$ $\quad \mathcal{B} = \{0,2,4,9\}$

$\mathcal{A} \cup \mathcal{B} = \{0,1,2,4,9\}$



22 It may be useful to work through the logic that proves these three sets equal.

If the index set is understood, we will write leave the range of indices implicit, as in $\bigcup_i$; when we need to be specific, we will include the indices explicitly as above.

Intersections. At the intersection of two roads is a patch of ground that is common to both of them; this is the image underlying the intersection of sets. If $\mathcal{A}$ and $\mathcal{B}$ are sets, then their *intersection* $\mathcal{A} \cap \mathcal{B}$ is the set containing those elements that belong to both $\mathcal{A}$ *and* $\mathcal{B}$:

$$x \in \mathcal{A} \cap \mathcal{B} \text{ if and only if } x \in \mathcal{A} \text{ and } x \in \mathcal{B}.$$

See the examples in the figure.[23] The "and" here is a logical and; it is true if x is in both sets.[24]

What is the intersection of two sets with no elements in common? (Think about this before proceeding.) If two sets have no elements in common, then their intersection must be the set with no elements, namely the empty set.

Thought Question What is the intersection of a set with itself? If $\mathcal{A} \subset \mathcal{B}$, what is $\mathcal{A} \cap \mathcal{B}$?

Just as unions can be defined for arbitrary collections of sets, so can intersections. In the example above, where $\mathcal{A}_i = [1/(i+1), 1[$ for $i \in Z_+$, the intersection $\bigcap_{i=1}^{\infty} \mathcal{A}_i$ is a set whose elements are those numbers that are in *every* $\mathcal{A}_i$. This is just the set $[1/2, 1[$. Check this for yourself. In general, for a collection $\mathcal{A}_1, \mathcal{A}_2, \ldots$, finite or infinite,

$$x \in \bigcap_i \mathcal{A}_i \text{ if and only if } x \in \mathcal{A}_i \text{ for every } i.$$

If the index set is understood, we will write leave the range of indices implicit, as in $\bigcap_i$; when we need to be specific, we will include the indices explicitly as above.

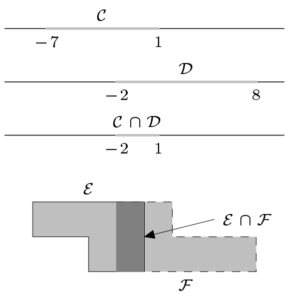
Differences and Complements. Given two sets $\mathcal{A}$ and $\mathcal{B}$, the *difference* $\mathcal{A} - \mathcal{B}$ is defined to be the set of elements of $\mathcal{A}$ that are not in $\mathcal{B}$:

$$x \in \mathcal{A} - \mathcal{B} \text{ if and only if } x \in \mathcal{A} \text{ and } x \notin \mathcal{B}.$$

See the figure[25] Notice that this definition treats $\mathcal{A}$ and $\mathcal{B}$ differently; just as in a numerical subtraction, the order matters.[26]

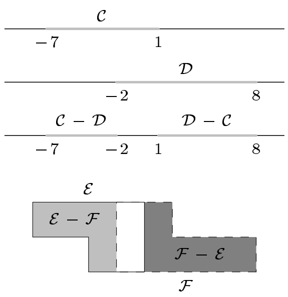
Very often, we work exclusively with subsets of some larger universe set $\mathcal{U}$, such as the set of real numbers or as we will see later the set of

23 $\mathcal{A} = \{0,1,4\}$ $\mathcal{B} = \{0,2,4,9\}$
$\mathcal{A} \cap \mathcal{B} = \{0,4\}$



24 In terms of logic or digital circuitry, you can think of the union operation as representing "Or" and the intersection operation as representing "And". If you see these words in the definition of a set, this is a clue that the set is built from a union or intersection.

25 $\mathcal{A} = \{0,1,4\}$ $\mathcal{B} = \{0,2,4,9\}$
$\mathcal{A} - \mathcal{B} = \{1\}$ $\mathcal{B} - \mathcal{A} = \{2,9\}$



26 A related operation for which order does not matter is called the symmetric difference. The symmetric difference $\mathcal{A}\Delta\mathcal{B}$ contains those points in either but not both of the sets: $\mathcal{A}\Delta\mathcal{B} = (\mathcal{A} - \mathcal{B}) \cup (\mathcal{B} - \mathcal{A})$. This is the set equivalent of the logical exclusive or.

outcomes of an experiment. When the universe set is clear from context, we speak of the *complement* of a set. If $\mathcal{A} \subset \mathcal{U}$, then the complement of $\mathcal{A}$, denoted by $\operatorname{compl}(\mathcal{A})$, is just $\mathcal{U} - \mathcal{A}$. This contains all the points in the superset that are not in $\mathcal{A}$. The complement is convenient when there is an overarching superset; it is the set equivalent of a logical not. Notice that $\mathcal{A} \cap \operatorname{compl}(\mathcal{A}) = \emptyset$; a set and its complement are always disjoint.[27] Also, $\mathcal{A} \cup \operatorname{compl}(\mathcal{A}) = \mathcal{U}$.

27 Analogy: A statement cannot logically be both true and false.

Mixing the Operations. We have already seen that we can take unions or intersections of collections of sets in any order and get the same result; in other words, no parentheses are needed. This is no longer true when we mix unions, intersections, and differences in the same expression. The following general relationships hold:

$$\mathcal{A} \bigcap \left(\cup_i \mathcal{B}_i \right) = \bigcup_i (\mathcal{A} \cap \mathcal{B}_i)$$

$$\mathcal{A} \bigcup \left(\cap_i \mathcal{B}_i \right) = \bigcap_i (\mathcal{A} \cup \mathcal{B}_i)$$

$$\operatorname{compl}\left(\bigcup_i \mathcal{B}_i \right) = \bigcap_i \operatorname{compl}(\mathcal{B}_i)$$

$$\operatorname{compl}\left(\bigcap_i \mathcal{B}_i \right) = \bigcup_i \operatorname{compl}(\mathcal{B}_i),$$

where the $\mathcal{B}_i$ are some collection of sets, finite or infinite. You should try to reason out verbally why these statements are true. For instance, if a point x is in $\mathcal{A}$ *and* at least one of the $\mathcal{B}$'s, then there is some i for which $x \in \mathcal{A} \cap \mathcal{B}_i$; hence x is in the union of these sets. The reverse argument works as well.

Thought Question Find a simpler expression for

$$\operatorname{compl}\left(\bigcap_{n=1}^{\infty} \left]-1/n, 1/n\right[\right).$$

Disjoint Sets. To say that two sets $\mathcal{A}$ and $\mathcal{B}$ are *disjoint* means that they have no elements in common,[28] that is $\mathcal{A} \cap \mathcal{B} = \emptyset$. Similarly, if $\mathcal{A}_i$ is a collection of sets, it is said to be a *pairwise disjoint* collection if for every index i and j with $i \neq j$, $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$; in other words, every pair of sets in the collection is disjoint. The collection of sets $\mathcal{A}_i = \left]i, i+1\right[$ for $i \in \mathbb{Z}$ is pairwise disjoint but the collection $\mathcal{B}_j = [0, j[$ for $j \in \mathbb{Z}_+$

28 Pictorially in Venn diagrams, sets are disjoint when they do not overlap.

is not. If $\mathcal{A}_i$ is a disjoint collection, the $\bigcup_i \mathcal{A}_i$ is called a *disjoint union*. The property of disjointness arises frequently in probability theory, and disjoint unions are special only because they make some computations easier.

Thought Question If $(\mathcal{A}_i)$ is a collection of sets indexed by $\mathbb{Z}_+$, and $\bigcap_{i \in \mathbb{Z}_+} \mathcal{A}_i = \emptyset$, does it imply that the collection is pairwise disjoint?

Thought Question (Eventually and Infinitely Often)

Suppose we have a sequence of sets $\mathcal{A}_1, \mathcal{A}_2, \ldots$. Describe the meaning of the following two sets:

$$\left\{\mathcal{A}_i \text{ eventually as } i \to \infty\right\} = \bigcup_{i=1}^{\infty} \bigcap_{j \geq i} \mathcal{A}_j$$

$$\left\{\mathcal{A}_i \text{ infinitely often as } i \to \infty\right\} = \bigcap_{i=1}^{\infty} \bigcup_{j \geq i} \mathcal{A}_j.$$

Hint: If $i = 10$, the set inside the union on the top line is $\bigcap_{j=10}^{\infty} \mathcal{A}_j$, and the set inside the intersection on the bottom line is $\bigcup_{j=10}^{\infty} \mathcal{A}_j$. What do these sets represent? Think about the expressions above from the inside out.

The Power Set of a Set. Given any set $\mathcal{S}$, we can form a new set that consists of subsets of $\mathcal{S}$. This is called the power set of $\mathcal{S}$ and is denoted by $2^{\mathcal{S}}$. For our purposes, this is mostly a curiosity, but it does provide a nice example of how mathematical notation internalizes meaning. Suppose that $\mathcal{S}$ has a finite cardinality, then it has $2^{\#\mathcal{S}}$ elements. (Why? Order the elements of $\mathcal{S}$ in some arbitrary way and for each subset of $\mathcal{S}$ construct a binary number with one bit per element which is 1 if that element is in the subset and 0 otherwise. There are exactly $2^{\#\mathcal{S}}$ such numbers.) If $\mathcal{S}$ is non-empty, then it can be shown that the cardinality of $2^{\mathcal{S}}$, called $2^{\#\mathcal{S}}$, is strictly larger than the cardinality of $\mathcal{S}$. This is not obvious for sets with infinite cardinality. Notice that this gives us a whole sequence of "infinities", each larger than the next.

4. Finite, Countable, and Uncountable Sets

A *finite* set is a set whose cardinality is a finite integer. We can express the same idea in a more complicated but equivalent way by constructing an explicit correspondence between a set and the first n integers for some integer n. Consider a set $\mathcal{S}$. List out the elements of $\mathcal{S}$ on one side of a (potentially very big) page. If for some $n \in \mathbb{Z}_+$ you can list the numbers $1, \ldots, n$ on the other side of the page and connect each of these numbers to exactly one element of $\mathcal{S}$ with a line, then $\mathcal{S}$ is finite and has cardinality n. Any set that is not finite is called an infinite set.

Why go through all this when you can simply count, you may ask. Two answers: (i) to a mathematician this process of listing and connecting *is* counting, and (ii) this formalized notion of cardinality generalizes better to infinite sets.

We have already seen that there is a whole sequence of infinities, but it is unclear what those numbers represent. To get a feel for this, let's start with the integers, which are an infinite set.[29] List the integers on the left side of the page (one per line), and list the *even* integers on the other side. Next, draw a line from integer k on the left to integer $2k$ on the right. As you look over your list, you will notice that every number on the left is connected to one and only one number on the right, and vice versa. We have constructed an explicit correspondence between the integers and even integers; by definition, the two sets have the same cardinality. This is true even though one is a subset of the other! (Infinite sets can be odd.) The same type of argument can be used to show that $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{Z}_\oplus$, and $\mathbb{Q}$ all have the same cardinality.

Any infinite set that can be put in explicit, one-to-one correspondence (every point in one set is connected to only one point in the other set and vice versa) with the integers is called *countably infinite*. A set that is either finite or countably infinite is called *countable*; even if it takes forever, we can count the elements of such a set. A set that is not countable is called *uncountable* or uncountably infinite. There are many more elements of such a set than there are integers.

Are there any uncountable sets? Yes, $\mathbb{R}$ is uncountable. ($\mathbb{C}$ and $\mathbb{R}^k$ for any $k \geq 1$ are also uncountable, as a result.) To see this, we will show that the interval $]0, 1[$ is uncountable, which implies that any interval including $\mathbb{R} =]-\infty, \infty[$ is itself uncountable.[30] The proof that this is

29 Suppose that $\mathbb{Z}$ were finite. Then we could find some finite $n \in \mathbb{Z}_+$ such that $\mathbb{Z}$ is in explicit correspondence to $\{1, \ldots, n\}$. But take the maximum of the integers connected to 1 through n by the lines we've drawn and add one. This cannot be connected to any number in $\{1, \ldots, n\}$ (because it is bigger than the maximum of such numbers) but is clearly an integer. Oops. Since we come to a contradiction if $\mathbb{Z}$ is finite, $\mathbb{Z}$ must be infinite.

30 If $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{A}$ is uncountable, then so is $\mathcal{B}$ because there is no way to put the subset $\mathcal{A}$ or therefore $\mathcal{B}$ into correspondence with $\mathbb{Z}$.

so is the famous Cantor's diagonal argument, and it is a hoot.

Consider numbers in $]0,1[$. Every such number can be written as $\sum_{k=1}^{\infty} a_k\, 2^{-k}$ for some sequence of a_k's that are all 0 or 1.[31] In other words, we can write each such number as a binary fraction of the form $0.a_1a_2a_3a_4\ldots$. If we throw a way all terminating sequences of a_k's, those with only finitely many 1's, then this binary fraction representation is unique. (There aren't really that many of these in some sense, so this is no loss), For instance, $1/2$ has a representation $0.01111111111111111\ldots$ since $0.10000000000000000\ldots$, with only one 1, has been thrown away.

31 Note: $\sum_{k=1}^{\infty} 2^{-k} = 1$.

That's all details; it can all be demonstrated but take my word for it for now. The real fun comes in the next step. Suppose that the interval $]0,1[$ is countable. Then, we can list out all these binary fractions and the integers and draw appropriate lines between the lists. Suppose we do this and suppose our list of the elements of $]0,1[$ looks as follows:

$$
\begin{array}{l}
0.a_{11}a_{12}a_{13}a_{14}a_{15}\ldots \\
0.a_{21}a_{22}a_{23}a_{24}a_{25}\ldots \\
0.a_{31}a_{32}a_{33}a_{34}a_{35}\ldots \\
0.a_{41}a_{42}a_{43}a_{44}a_{45}\ldots \\
0.a_{51}a_{52}a_{53}a_{54}a_{55}\ldots \\
\ldots \\
\ldots \\
\ldots
\end{array}
$$

Take this list and construct a correspondence with the integers, which we can do by the assumption that $]0,1[$ is countable. But now make a new number $0.b_1b_2b_3\ldots$ where $b_k = 1 - a_{kk}$. This new number is not in our original list because it differs from every number on that list in at least one digit. Ooops. What this means is that no matter how we attempt to construct a correspondence between $]0,1[$ and the integers we always miss some of the numbers in $]0,1[$. That is, $]0,1[$ is uncountable.