Foundations of Information and Inference

Lecture Notes, Spring 2014

Colin Reimer Dawson, Ph.D.

Last Revised May 5, 2014

Contents

Part I

Overview and Preliminaries

Chapter 1

Introduction

1.1 Deductive and Plausible Reasoning

Scenario A policeman on patrol hears a burglar alarm, and looks across the street to see a broken window in a jewelry store. A man in a ski mask crawls out with a bag. The policeman concludes that this man is a criminal.

What's the reasoning process?

1.1.1 Strong Syllogism

We can identify several different forms of inference, which can be classified into two groups, depending on how "fallible" they are. The "strong" variety are logical, deductive inferences, which are bulletproof, provided their premises are true, because the premises are absolute: if p is true, then q is *always* true.

The most direct form of inference is **modus ponens**.

Syllogism Type 1 (Modus Ponens)

(Premise) If A is true, then B is true (e.g., if it rained, my grass is wet)

(Data) A is true (it rained)

(Inference) B is true (my grass is wet)

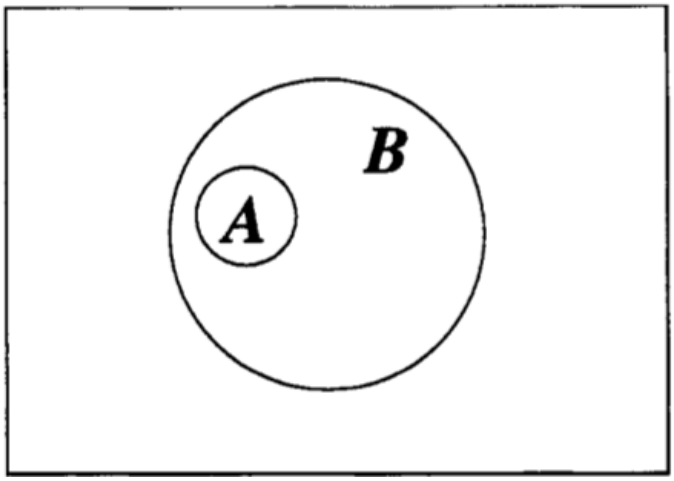

Figure 1.1: If A is only true in situations where B is true, then we can apply modus ponens (from A to B) or modus tollens (from the negation of B to the negation of A).

Given the same premise, namely that B is true whenever A is true, we can make a "negative inference" in the opposite direction. If B is false, then A cannot be true (since if it were, B would be true). This type of inference is called **modus tollens**.

Syllogism Type 2 (Modus Tollens)

(Premise) If A is true, then B is true (e.g., if it rained, my grass is wet)

(Data) B is false (my grass is not wet)

(Inference) A is not true (it must not have rained)

1.1.2 Weak Syllogism

Given the premise "A implies B", we can reason from A to B, or from NOT B to NOT A. A common mistake made by logic students (I've heard this called "modus morons") is to reason from B to A, or to reason from NOT A to NOT B. It should be clear from the Venn diagram that these inferences are not logically valid: A can be false even though B is true.

However, if we think of the diagram as containing possible configurations of the world, then learning that B holds rules out some of the ways that A can be false; similarly, learning that A fails rules out some of the ways that B can hold. So it seems reasonable to define the following "weak syllogisms", which deal not with absolutes, but with degrees of plausibility.

Syllogism Type 3

(Premise) If A is true, then B is true (e.g., if it rained, my grass is wet)

(Data) B is true (my grass is wet)

(Inference) A becomes more plausible (more "likely" that it rained)

Syllogism Type 4

(Premise) If A is true, then B is true

(Data) A is false (it didn't rain)

(Inference) B is less plausible (less likely my grass is wet)

But more or less plausible does not mean certain to be true or false. For example, what if there is a sprinkler system?

But what if we don't even have the nice clean implication that A always implies B, but we have a situation more like that in Figure 1.2, where any combination is possible? Suppose we learn that B is true. This rules out both some of the ways that A could be true *and* some of the ways that A could be false. It's hard to say anything at all about A.

But now imagine that most but not all of B overlaps with A. For example, A might indicate that there's rain and B might indicate that there's lightning. Sometimes there's lightning without rain, but usually there's rain, too. If we discover that there's lightning (B is true), then we rule out the possibility that there's lightning but no rain (A would be true), but that's relatively small as a proportion of all the ways A could happen; far more consequential is the fact that we've ruled out the possibility of neither lightning nor rain (this makes up most of the "possible worlds" that do not involve rain). It seems that in this case we are licensed to say that, having observed lightning rain is now *more likely* (but not guaranteed).

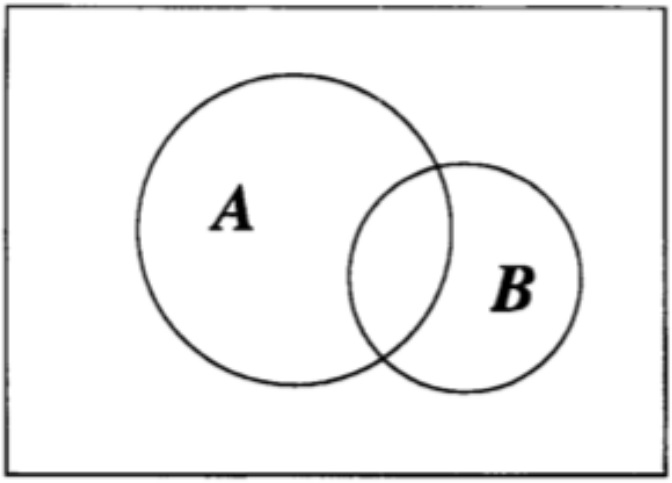

Figure 1.2: If neither event is contained inside the other, yet they overlap, then any combination is possible, and no strong syllogisms are available no matter the data.

This form of reasoning is far removed from modus ponens and modus tollens: not only are we reasoning in the wrong direction, but our premise isn't even guaranteed. Still, it seems perfectly reasonable, so long as we keep our uncertainty in mind.

Syllogism Type 5

(Premise) If A then B is more plausible (If there's rain, lightning becomes more likely.)

(Data) B is true (There's lightning.)

(Inference) A is more plausible (Rain is likely.)

Questions:

- Which of the above characterizes the policeman's reasoning process?
- What's missing from the description?
- What if this happened all the time, and the suspects turned out to be innocent?

1.1.3 Transitivity With Strong vs. Weak Inference

Weak and strong syllogism behave differently when we chain inferences together, using the conclusion of one inference as the data to the next. Strong syllogisms can be joined together without any loss of certainty: the conclusion of the first strong syllogism is every bit as valid as an observation would be.

For example, suppose A, B and C are propositions about a particular polygon that may or may not be true. A says the shape is a square, B says the shape is a rectangle, and C says the shape has four sides. Then we have the following transitive relationship.

Strong Transitivity

(Premise) If A is true then B is true (all squares are rectangles).

(Premise) If B is true then C is true (all rectangles have four sides)

(Data) A is true (we are given a square)

(Inference) C is true (our shape has four sides)

That is, since we have a square, we have a rectangle, by the first premise. Since we have a rectangle, the condition for the second premise is satisfied, which allows us to conclude that the shape has four sides. If we want to, we are licensed to chain the premises together like straws to get a new premise that says that all squares have four sides.

But what if we're using a weak syllogism, such as the following?

Slipping Confidence

(Premise) If it's cloudy, rain is more likely.

(Premise) If it rains, cancelling the baseball game is more likely.

(Data) It is cloudy

(Inference) (?) It is more likely that cancelling the baseball game is more likely.

As we chain these weaker premises together, our certainty slips. How about this case?

Weak Syllogism is not Modular

(Premise) If my grass is wet, it likely rained last night.

(Premise) If the sprinkler was on last night, my grass is wet.

(Data) The sprinkler was on last night.

(Inference) (?) It likely rained last night.

Finally, compare and contrast the following intuitive, yet formally puzzling, reasoning processes.

Explaining Away

(Premise) If my fuel tank is empty, the gauge will read "E"

(Premise) If the gauge battery is dead, the gauge will read "E"

(Datum 1) The gauge reads "E"

(Inference 1) Plausibility of empty tank goes up

(Inference 2) Plausibility of dead battery also goes up

(Datum 2) Battery is dead

(Inference 3) Plausibility of empty tank goes back down!

Here, finding a dead battery has an impact our beliefs about the fuel tank, even though these are not actually connected in any way, as can be seen in the second example.

Contrast With This

(Premise) If my fuel tank is empty, the gauge will read "E"

(Premise) If the gauge battery is dead, the gauge will read "E"

(Datum 1) Battery is dead

(Inference) The gauge will read "E"

(Inference) No effect on plausibility of empty tank!

Now, when we are ignorant about the gauge, the dead battery has no impact on our beliefs about the fuel tank. There seems to be some "memory" about how our beliefs were formed.

1.1.4 Building an Idealized "Common Sense Machine"

Imagine building an artificial system to do every day reasoning. Some are skeptical that machines could every really "think". The mathematician von Neumann (after whom the computing architecture underlying your personal computer, the "von Neumann machine" is named) responded to this skepticism by saying "If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that!". The point was that if the human brain does something, it is physically possible, so if we know how it is done, we can make a machine do it.

The modern field of Artificial Intelligence (AI) dates back to the 1950s, when scientists and engineers tried to build systems based purely on deductive reasoning. Though they had great success for "artificial" and highly constrained problems like chess, it was not until they started to incorporate weak inference that substantial progress began to be made in more realistic domains. Even something as seemingly simple as vision requires massive amounts of weak inference!

If we want to program a computer, or build a robot, to engage in everyday reasoning of the sort described in the examples above, then it's not enough to rely on our intuitions; nor do we want to settle for fuzzy statements like "rain is likely". We need to know exactly when, for example, finding a dead battery should influence our beliefs about a gas tank; and we need to know *how likely* it is to rain. When we are outside the realm of formal mathematics and logic, every conclusion has the possibility of being wrong, which has a cost. Even gathering more information to improve our conclusion can be costly. We need our system to be able to make intelligent decisions.

Conclusion: We need a formal theory of plausible reasoning!

- Limitation: We will confine machine to reasoning about "objective" propositions, not "opinions".

Chapter 2

Mathematical Reasoning

2.1 Mathematical Models

If we want to build a system that reasons about domains like weather, batteries and gas tanks, or burglaries, it needs to be able to *represent* the relevant quantities and propositions in its "mind" in such a way that it can do computation with them. That is, we need a **mathematical model**. What does that mean?

Mathematical models are an **abstraction** of the phenomenon they represent. Okay, what is abstraction?

We can make an analogy to an architectural blueprint. If we're building a house, we represent structures as perfect rectangles and 90 degree corners, but this is an idealization. For another example, what we call temperature is really the aggregate motion of billions of molecules. In both cases it seems like the abstraction is *simpler* than the thing it represents, in the sense that we have stripped away a lot of the detail to focus on the properties that matter for the task at hand.

Once we have this stripped down abstraction we can apply **mathematical theory** to the model to figure out how it behaves and what kinds of properties it has. Within the model, we can do calculations over the abstracted representations in order to make predictions about what will happen under different circumstances. We can then go back and apply these predictions to the real world.

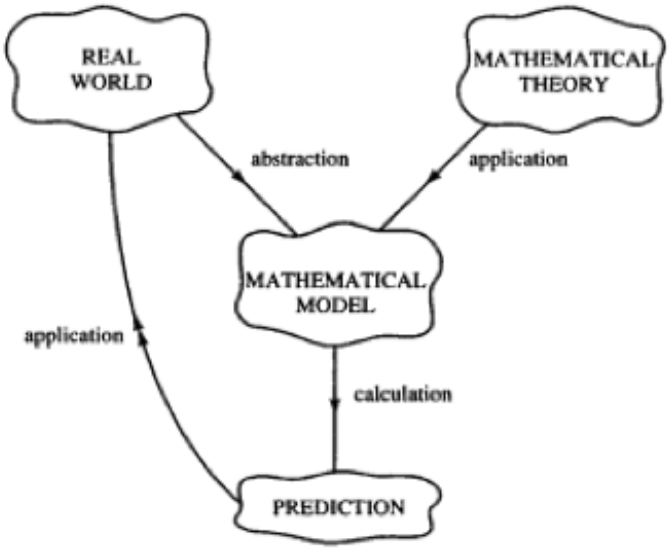

Figure 2.1: The relationship between models, theory and the world

2.2 Mathematical Proof

We want to develop mathematical models that *do* "common sense" reasoning via "weak syllogisms". However *to establish the properties of the models themselves*, we need to confine ourselves to logic and strong syllogism. Since the models are mathematical in nature (involving quantitative "degrees of belief"), that strong reasoning takes the form of mathematical proof. A background in doing proofs is not essential for this course, but some prior exposure to mathematical reasoning will ease the process.

2.2.1 Structure of Proofs

Most mathematical proofs we'll run into take one of four or so forms. To wit:

1. The **direct method** involves taking advantage of transitivity to string implications (or equalities or inequalities, which also obey transitivity) together until we can collapse them into the implication (or equality or inequality) that we set out to prove.

2. The **contrapositive method** also involves stringing together implications,

but instead of arriving at the original statement (of the form "if A then B"), we prove the logically equivalent *contrapositive*, "if NOT B then NOT A".

3. **Proof by contradiction** uses the "process of elimination" to show that if the statement we want to prove is false, then we arrive at a conclusion that we know is wrong (for example, that $0 = 1$, that a number is both positive and negative, etc.)

4. **Mathematical induction** is a technique used to show that an *infinite sequence* of propositions is true. This one is a bit different than the others, because of the nature of what we're trying to show.

Mostly we will need the direct method and proof by mathematical induction in this course, but let's consider all four of these in turn.

Before we start, let's give formal definitions of concepts that you're already familiar with so that we can be rigorous while focusing on the proofs techniques themselves and not on any complicated mathematical objects.

Definition 2.1. *An* **integer** *is any number that can be obtained from 0 by repeatedly adding 1 or -1.*

We will make use of the fact that when we add or multiply two integers together, the result is an integer. Ideally we should prove these facts as well, but we will just assume them for the sake of exposition.

Definition 2.2. *An* **even** *number is an integer that is equal to* $2y$ *for some integer* y*.*

Definition 2.3. *An* **odd** *number is an integer that is equal to* $2y+1$ *for some integer* y*.*

2.2.2 Direct Method

The goal is to prove a statement of the form "if A is true, then B is true", where A and B are propositions that are true or false. Or, sometimes, A and B are **predicates** that can be true or false of a particular object (such as a number), and we want to prove that "if A is true *of a particular object,* x, then B is true *of that same object,* x".

For example, in a moment we will prove that if the statement, A: "x is an odd number" holds for a particular x, then the statement B: "$x + 1$ is an even number"

holds for the same value of x. We can state this as, "for all integers x, if x is an odd number, then $x + 1$ is an even number".

Before we get to the example, let's look at the steps that are involved in proving a statement like this.

(i) We want to prove that whenever A is true, B must also be true. (That is, for any x that makes A true, it also makes B true)

(ii) For a particular x, either A is true or it is false. If it is false, then there is nothing to prove. So, **suppose** we have an x where A is true.

(iii) Use definitions and previously proved theorems to **construct a deductive chain** (e.g., "by definition, if A is true then C is true, and by Theorem 7.16, if C is true, then B is true")

(iv) **Conclude** that B is true via repeated application of **strong syllogism** (in this case, modus ponens).

(v) Since our only premise was that A holds (everything else was definition and fact), we have the general fact that $A \Rightarrow B$.

Note an important property of this process: if we want to prove that some if-then statement holds *for all* x, then we can't make any additional assumptions about the nature of x. So if A says that x is an even number, we can suppose that we have an even number, but we don't get to pick a *specific* even number, because we might be lucky and pick one for which B just happens to be true, even though there are others for which it doesn't.

We start off by stating what it is we're trying to prove. For something simple, we might call this a "Proposition"; for more "important" results, we might call it a "Theorem". Sometimes we have a "Lemma", which is usually something we need to prove in the service of a larger theorem. But the distinctions are subjective.

Proposition 2.1. *If x is an odd number, then $x + 1$ is an even number.*

Once we have stated our proposition, we indicate that we are beginning the proof. Remember the general strategy here: We will first **suppose** that the "if part" holds. Then we will try to arrive at a demonstrate the "then part" via a sequence of legal "moves".

Proof. Let x be an integer. If x is not odd, then there is nothing to prove (this part is usually omitted in practice). Suppose x is odd (the "if part"). Then, by definition

of an odd number, there is some integer y such that $x = 2y+1$ (we use the definition to give ourselves something concrete to work with).

Let y be such an integer (the definition says that one exists). Our goal is to show that $x + 1$ is even. In order to do this, we will want to use the definition of even: there is some other integer — let's call it w this time, so as not to confuse it with the y that we've already defined — such that $x+1 = 2w$. We just need to show that such a number w exists (and is an integer).

Let's see what we can do with $x + 1$. Using the equation from the definition of odd ($x = 2y + 1$), we can easily get to

$$x + 1 = (2y + 1) + 1$$

The rest is just simple algebraic manipulation.

$$\begin{aligned} x + 1 &= 2y + 2 \\ &= 2(y + 1) \end{aligned}$$

Since y is an integer and 1 is an integer, so is $y + 1$ (we're using the property of integers we stated above). Define $w = y + 1$. Then we have $x + 1 = 2w$ where w is an integer. That's what we needed, and so we have shown that $x + 1$ is even. □

The little square is a common notational convention to indicate that a proof is complete. Note that in the above proof, we didn't make any assumptions about the value of x except what was required by the statement: namely, that it is odd.

Let's try another one.

Proposition 2.2. *For all integers, w and x, if both w and x are odd, then $w + x$ is even.*

Proof. Suppose w and x are odd numbers. (Again, if this is not true, then there is nothing to prove.)

By definition of an odd number, there are some other integers, call them y and z, such that $w = 2y + 1$ and $x = 2z + 1$.

This time our goal is to say something about $w + x$ — in particular that it can be written as twice some integer. As before, we can rewrite the quantity of interest

$w+x$ by substituting equivalent quantities. Then, as before, we just do some algebraic manipulation to massage the right-hand side into the form 2 times something.

$$\begin{aligned} w + x &= (2y + 1) + (2z + 1) \\ &= 2y + 2z + 2 \\ &= 2(y + z + 1) \end{aligned}$$

Since y and z are integers, so is $y + z + 1$, and since $w + x$ is twice an integer, it is even. □

Again, we did not assume anything about w and x except that they are odd numbers. If we said anything more specific than this, our proof would no longer apply to every single pair of odd numbers.

Let's try one more.

Proposition 2.3. *If x is an even number, then x^2 is an even number.*

Proof. Suppose x is an even number. Then, *by definition of even*, there is an integer, y, so that $x = 2y$. Then we can rewrite x^2 by substituting for x:

$$\begin{aligned} x^2 &= (2y)^2 \\ &= 4y^2 \\ &= 2(2y^2) \end{aligned}$$

Since y is an integer, $y^2 = y \times y$ is an integer, and since *that's* an integer, so is $2y^2$. We have successfully written x^2 in the form required to be an even number. □

Note that in the above proofs I'm being super-extra-pedantic, in order to focus on the process. In practice this much detail is usually not necessary, but it's good to start out using a lot of detail to force you to examine your steps closely.

2.2.3 Proof by Contrapositive

The second proof method takes advantage of the fact that a statement of the form "If A is true, then B is true" is **logically equivalent** to the corresponding statement,

"If B is false, then A is false". With a bit of reflection, you should be able to convince yourself that these two are actually saying exactly the same thing.

Informally, two statements are logically equivalent if they yield the same Venn diagram over "possible worlds". In this case, "If A then B" implies a Venn diagram in which the set of worlds where A is true are contained entirely in the set of worlds where B is true. "If NOT B then NOT A" means that which A cannot be true when B is not true — in other words, all cases where A is true are also cases where B is true.

In some cases, it is difficult or cumbersome to prove a statement directly, but its contrapositive is much simpler. Because of logical equivalence, if we prove the contrapositive, then the original statement holds as well.

The basic structure of a contrapositive proof is as follows

(i) We want to prove that $A \implies B$ — in other words, that whenever A is true, B is true.

(ii) The **contrapositive**, $\overline{B} \Rightarrow \overline{A}$ (Not-B implies Not-A) is logically equivalent.

(iii) Prove the contrapositive by any other method (e.g., the direct method).

(iv) Because of logical equivalence, conclude that the original statement holds.

Let's look at an example.

Proposition 2.4. *If the product of two integers is even, at least one of the integers is even.*

First let's think about what we would have to do to prove this directly. We would have to suppose that we had two integers, say w and x, whose product, wx is even. We can rewrite wx as $2y$ where y is an integer, but then we're a little bit stuck. We could probably go through some fancy footwork and get where we need to go, but it's far simpler to prove the contrapositive.

Proof. Suppose w and x are integers. We can restate the original proposition as its contrapositive. The negation of the "then part" is that *it is not the case* that at least one of the integers is even. In other words, *neither* w nor x is even. The negation of the "if part" is that the product is *not* even.

So the contrapositive statement we are trying to prove is "If neither w nor x is even, then wx is not even". Now we just proceed using the direct method. We will use a

fact that we have not yet proved, but which you will prove for homework. And that is that every integer is either even or odd, but not both.

Suppose neither w nor x is even. Then, since all integers are either even or odd, and w and x are not even, they must both be odd.

Again, by definition of an odd number, there are some other integers, call them y and z, such that $w = 2y + 1$ and $x = 2z + 1$.

We are trying to prove something about the product, wx. So, we rewrite it, and then do some manipulations:

$$\begin{aligned} wx &= (2y+1)(2z+1) \\ &= 4yz + 2y + 2z + 1 \\ &= 2(2yz + y + z) + 1 \end{aligned}$$

Since $2yz + y + z$ is an integer, this means that wx is odd. So we have proved the contrapositive, and by logical equivalence, this establishes the original statement. □

2.2.4 Proof by Contradiction

Proof by contradiction is often confused with proof by contrapositive, because the first step is to assume that something is false. The difference is that, here, we assume that the statement we are proving is itself false, and show that it leads to something impossible (usually that some statement is both true and false, or that some obviously false equality, like $1 = 0$, holds).

The basic structure:

(i) We want to prove some statement, B (sometimes as above, we want to prove an implication, "if A then B").

(ii) Suppose the statement is false. (If the statement is "if A then B", what would make this false?)

(iii) Show that this leads to a contradiction, for example that some statement is both true and false, or that it leads to something obviously false, like $1 = 0$.

Let's look at an example. It is known that there is no largest prime number, as we will show now.

Proposition 2.5. *There are infinitely many prime numbers.*

Proof. Since this is a proof by contradiction, our first step is to suppose that the statement doesn't hold; that is to say, that there are *finitely* many primes. Suppose this is the case, and denote this finite number by n.

Then we can list all the primes in increasing order: $p_1, p_2, \ldots, p_n$. p_n is then the largest prime.

Since there is a fixed number of primes, we can multiply them all together and get a finite number, which is an integer larger than any prime (primes by definition are integers larger than one, so every time we multiply by a prime, we get something bigger than we had before).

We can call this big product q, which is equal to $p_1 \cdot p_2 \cdots p_n$. Now consider $q + 1$, which is also a finite integer.

If $q + 1$ has a factor besides 1 and itself, then it has a prime factor (if the first factor we find isn't prime, we can factor *it*; and so on until we get a prime).

But for every p (say p_2), we have

$$(q + 1)/p_2 = p_1 \cdot p_3 \cdot p_4 \cdots p_n + 1/p_2$$

But this is not an integer (p_2 is larger than 1, and 1 over any integer larger than one is not an integer), so q does not have a prime factor. Therefore q is prime.

But since q is bigger than the biggest prime, this contradicts the assumption that we listed all the primes. Therefore our assumption that there were finitely many primes must have been incorrect. □

See also the proof in Applebaum that $\sqrt{2}$ is irrational.

2.2.5 Proof by Mathematical Induction

Proof by mathematical induction is often likened to knocking down a series of dominos. We use it when we have a sequence of statements that we are trying to prove all at once. Usually the statements are numbered, with the numbers playing a role in defining the individual statements.

For example, consider the following generalization: "The sum of the first n positive integers is equal to $\frac{n(n+1)}{2}$". You can check this easily for any given value of n, but we want to prove that it holds for every positive integer and we can't check all of them individually. What we can do is to use induction to prove it for every positive integer at the same time.

Intuitively, imagine that each statement is a domino standing upright on a table, and we prove it by knocking it down. We don't have to knock each one down individually, though; we just have to ensure that they're set up in such a way that each one knocks down the next one as it falls, and then we just need to push over the first one. Voila! All the dominos are knocked down!

The basic structure of an induction proof:

(i) We want to prove that some infinite sequence of statements, call them $P(1), P(2), \ldots$, are all true.

(ii) Show that the first statement is true (we can knock down the first domino). This is called the **base case**.

(iii) Show that *if* the n^{th} statement is true (for general n), then the $n+1^{\text{th}}$ statement must be true (each domino knocks down the next one). The "if part" is called the **inductive hypothesis**. This whole part of the proof is called the **inductive step**.

(iv) Since the inductive step works for arbitrary n, the first statement implies the second, the second implies the third, and so on. Since this is deductive inference, the transitive property applies as many times as we want, and so by modus ponens, it works for any n.

Let's try to prove the generalization above.

Proposition 2.6. *The sum of the first n integers is equal to $\frac{n(n+1)}{2}$.*

Note that although this is written as one statement, it is really a sequence:

$$P(1) : \text{The sum of the first integer,} 1 = \frac{1(2)}{2}$$
$$P(2) : \text{The sum of the first two integers,} 1 + 2 = \frac{2(3)}{2}$$
$$P(3) : \text{The sum of the first three integers,} 1 + 2 + 3 = \frac{3(4)}{2}$$
$$\ldots$$

We can't check every single one, but we can use induction.

Proof. First, check the case where $n = 1$. Obviously $1 = \frac{1 \cdot 2}{2}$, and so we've proved the base case.

Now, make the inductive hypothesis: suppose the statement is true for some arbitrary $n = k$. We can't say anything more specific about this k beyond that it is a positive integer, or we lose the generality we need. We have to leave it as a variable. So what we are supposing is that

$$\sum_{j=1}^{k} j \quad = \quad \frac{k(k+1)}{2} \tag{2.1}$$

We want to use this fact to show that the statement holds when $n = k + 1$. That is, we need to show that $\sum_{j=1}^{k+1} j = \frac{(k+1)((k+1)+1)}{2}$. As we did in many of our proofs above, we can proceed by trying to rewrite the left-hand side of the equation we're trying to prove.

Note that the left-hand side in this $k + 1$ version is the same as the left-hand side of our inductive hypothesis (2.1) (which we get to assume is true), except that it has one additional term. So we can write it as

$$\sum_{j=1}^{k+1} j = (\sum_{j=1}^{k} j) + (k + 1)$$

But by the inductive hypothesis we can substitute for the sum, to get

$$\begin{aligned}\sum_{j=1}^{k+1} j &= \frac{k(k+1)}{2} + (k+1) \\ &= \frac{k(k+1)}{2} + \frac{2(k+1)}{2} \\ &= \frac{k(k+1) + 2(k+1)}{2} \\ &= \frac{(k+1)(k+2)}{2}\end{aligned}$$

This is just what we needed to show. Since we did not assume anything special about the value of k, this works for every choice.

Now, since we know the statement holds for $k = 1$, we can use the inductive step ("If $P(k)$ then $P(k+1)$") with $k = 1$ to get "If $P(1)$ then $P(2)$". Together with $P(1)$ we can just apply modus ponens to establish that $P(2)$ holds.

For $k = 2$, the inductive step becomes "If $P(2)$ then $P(3)$". Since we have established $P(2)$, this entails that $P(3)$ is true. We can keep doing this as many times as we want to establish $P(n)$ for *any* positive integer n.

In other words, the inductive step ensures that each domino will knock down the next one; so once we knock down the first domino, it knocks down the second, which in turn knocks down the third, and so on. □

Let's try another one.

Proposition 2.7. *If the product of n numbers is zero, then at least one of the individual numbers is zero.*

Here's a statement that we couldn't even check one n at a time if we wanted to. However, we can prove it easily by induction.

Proof. As before, we really have a sequence of statements.

$P(1)$: If the product of 1 number is 0, then the number is 0.

$P(2)$: If the product of 2 numbers is 0, then at least one of the numbers is 0.

$P(3)$: If the product of 3 numbers is 0, then at least one of the numbers is 0.

...

Start with the base case, where $n = 1$. If there is only one number being multiplied, then the product is just that number. Therefore, if the product is zero, the number is zero.

Now let's state the inductive hypothesis with $n = k$ (for arbitrary k). This says that if the product of k numbers is zero, then one of the k numbers is zero. We want to show that *if* this is true, then the statement holds for $n = k + 1$.

As above, we will start by taking $k + 1$ numbers, and trying to rewrite their product in such a way that we can take advantage of the inductive hypothesis. We don't get to use specific numbers, so we will write the numbers as $x_1, x_2, \ldots, x_k, x_{k+1}$. Now consider their product, which we assume to be zero. We have

$$x_1 \cdot x_2 \cdot \cdots \cdot x_k \cdot x_{k+1} = 0$$

If we define y to be the product of the first k terms, i.e., $y = x_1 x_2 \cdots x_k$, then we have

$$y \times x_{k+1} = 0 \tag{2.2}$$

Notice that we're using a similar strategy to the one we used above: break the $k + 1$ case into two pieces, one of which looks like the kth case. Then we can rely on our inductive hypothesis to do most of the work for us.

If $x_{k+1} = 0$, then the conclusion holds, since it said that at least one term had to equal zero. If $x_{k+1} \neq 0$, then we can divide both sides of 2.2 by x_{k+1}, which establishes that $y = 0$. But since y is the product of k numbers, the inductive hypothesis applies, and we can conclude that that one of the first k terms is zero. So, either $x_{k+1} = 0$ or one of the first k terms is zero. That means that one of the $k + 1$ terms is zero, and we have established that $P(k)$ implies $P(k + 1)$.

Finally, since $P(1)$ holds, and since we have shown that $P(k) \implies P(k + 1)$ for arbitrary k, then we can set $k = 1$ to conclude that $P(2)$ holds; then set $k = 2$ to conclude that $P(3)$ holds, and so on as many times as we want to prove that $P(n)$ holds for an arbitrary value of n. □

In a sense, our ability to count is based on this sort of inductive process: We know to start at 1, and we know how to get from one number to the next, therefore we can (in principle) count as high as we want.

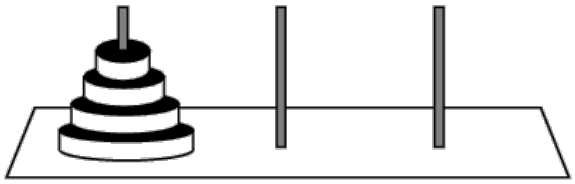

Figure 2.2: The Tower of Hanoi game (image courtesy `http://mathworld.wolfram.com/TowerofHanoi.html`)

The Tower of Hanoi Here's one more example that isn't about numbers.

A popular toy consists of three pegs and a set of circular discs of decreasing diameters, each with a hole in the center (see Fig 2.2). The game starts with the discs stacked in decreasing order of size on the first peg. The goal is to move one disc at a time to another peg, without ever placing a bigger disc on top of a smaller one, until all the discs are again arranged in order on the third peg. Prove that if there are n discs, then the puzzle can be solved in $2^n - 1$ moves.

As usual, the base case is easy. With only one disc, we can obviously complete the puzzle in one move. It happens that $1 = 2^1 - 1$, which what we need for our base case.

The key insight for the inductive step is that the only way we will ever be able to move the bottom disc to the third peg is if all the other discs are first moved (in order) to the middle peg. In other words, to successfully move $k + 1$ discs, we first have to solve the problem of moving a stack of k discs to another peg. Once we've done this, then it takes one move to put the bottom disc where it belongs, then we have to again move the other k discs. So if m_k is the number of moves needed to transfer k discs from one peg to another, then we will need $m_k + 1 + m_k = 2m_k + 1$ moves to transfer $k + 1$ discs.

If the inductive hypothesis holds, and it takes $m_k = 2^k - 1$ moves to transfer k discs, then by the logic above, it will take

$$2(2^k - 1) + 1 = 2^{k+1} - 2 + 1 = 2^{k+1} - 1$$

moves to transfer $k + 1$ discs. This completes the inductive step.

Once again, we have established the claim for $n = 1$, and we have established that *if* it holds for $n = k$, then it holds for $n = k + 1$, no matter the value of k. So, since it works for $n = k = 1$, it works for $n = k + 1 = 2$. And since it works for $n = 2$, it works for $n = 3$. And so on.

2.3 Exercises

Recall the following definitions.

- An integer x is **even** if there is some integer y such that $x = 2y$.
- An integer x is **odd** if there is some integer y such that $x = 2y + 1$.

You may assume basic properties of integers without proving them, unless otherwise noted. For example, the sum of two integers is an integer; the product of two integers is an integer; etc.

Prove the following statements by the direct method.

1. For every integer x, if x is even then $x + 1$ is odd.
2. For every integer x, if $x + 1$ is odd, then x is even.
3. For every integer x, if x is odd then x^2 is odd.
4. For every integer x, if x is odd then x^3 is odd.
5. For every pair of integers, x and y, if x and y are both odd, then xy is odd.
6. For every integer x, if x is even and y is any integer, then xy is even.

Let's define a new concept. We will call an integer **throdd** if it is one more than an integer multiple of 3. More formally, we will say that an integer x is **throdd** if there exists an integer y such that $x = 3y + 1$.

Prove the following properties of throdd integers.

7. If x is throdd then x^2 is throdd.
8. The product of any two throdd numbers is throdd.
9. The sum of two throdd numbers is never throdd (Hint: show that if the sum can be written in the form $3t + 1$ then t is not an integer)

Prove the following statements.

10. Let x, a and b be integers. Then $x + a = x + b$ *if and only if* $a = b$.
11. Let x, y and a be integers. Then $ax = ay$ *if and only if* $x = y$ OR $a = 0$.
12. Let x, a and b be integers. If $ax = 1$ and $bx = 1$, then $a = b$.

Prove the following generalizations by induction. Remember the two steps: (1) show that the statement works for the simplest case (usually $n = 1$), (2) show that *if* it works for some $n = k$, then it works for the case where $n = k + 1$. Therefore, since it works for $n = 1$, it works for $n = 1 + 1 = 2$. Therefore, it works for $n = 2 + 1$. And so on up to any positive integer. That is, (1) we can knock down the first domino, and (2) each domino will knock down the next one, therefore all the dominos will fall.

13. Let x be an arbitrary odd number. Then x^n is odd for every positive integer n.

14. The quantity $3^n - 1$ is even, for every positive integer n.

15. Let $x_1, x_2, x_3, \ldots, x_n$ all be odd numbers. Then the product $x_1 \cdot x_2 \cdot \cdots \cdot x_n$ is odd (Hint: use problem 5).

16. Every positive integer is either even or odd (Hint: Your inductive step will have two cases: one where n is even, and one where n is odd. For each case you will need to show that $n + 1$ is either even or odd.)

17. The sum of the first n odd integers is equal to n^2 (Hint: the k^{th} odd number can be written as $2k - 1$. For example, the 4^{th} odd number is 7, which can be written as $2 \times 4 - 1$.)

Here are some problems not involving numbers that make use of induction. They require a bit more creativity, but hopefully you'll find them fun.

18. If n people are in a room and every person shakes hands exactly once with every other person, then the total number of handshakes that take place is equal to $n(n - 1)/2$.

19. Imagine a square checkerboard with 2^n individual squares on a side, and one corner square cut out. This checkerboard can be tiled with "L-shaped" pieces consisting of three squares each. (Hint: Start with a $2^1 \times 2^1$ board as the base case. Now note that you can create a $2^{k+1} \times 2^{k+1}$ board by combining 4 $2^k \times 2^k$ boards. For example, we can get a 4×4 board by tiling four 2×2 boards.)

20. Consider a pile of n pebbles. For any number between 1 and $n - 1$, we can select that many pebbles and put them in a separate pile. For each of the remaining piles with more than one stone, we can further divide it into two (possibly unequal sized) piles. We can continue this until all piles have only one pebble. What is the minimum number of splits needed to do this (as an

expression involving n)? Prove that your result always holds. (Hint: try it out with some small values of n first to get the pattern; then see if you can prove that it holds for all n by induction.

Chapter 3

Counting and Sets

3.1 Counting

Most of us have, at some time or another, experienced an odd coincidence and wondered, "What are the chances of that?" For example, in a small group of people (say the 27 in this class) you might meet someone with the same birthday as you. If births are pretty evenly distributed throughout the year, you would expect only about 1 person in 365 to share *your* birthday. In a group of 27, it seems surprising to find someone that shares your birthday. It must be fate!

But consider that if someone else discovers a "birthday twin" they will likely be surprised as well. Should we infer that something fishy is going on? Winning the lottery is extremely unlikely for any given person, but *someone* wins most of the time, and to them it might seem like something mystical is happening.

There are lots of potential events that we might find surprising. The chances that *something* surprising happens will be much higher than the chances that any *particular* surprising thing occurs. Before we jump to conclusions, we should try to take a broader perspective.

For example, what's the probability that *some* two people in a group of 27 have the same birthday?

There are lots of different assumptions you might make about birthdays, but answering this question inevitably involves counting the *number of ways* different things can occur. For example:

- How many pairs of people are there?
- For each pair, how many different birthdays could be shared?
- We also want to count triples, but these include pairs, so how do we make sure we're not double counting?

The mathematical subject of **combinatorics** examines the properties of combinations, and gives us some useful "tricks" to count things efficiently.

3.1.1 Experiences and Outcomes

Before we start counting, we need to figure out what the "units" are that we're counting in the first place. We can be pretty general about this, and talk about counting "experiences".

Definition 3.1 (Experience)**.** *An **experience** is anything that takes on one of several values.*

Examples:

- Flipping a coin
- Rolling a die
- An election
- Observing someone's birthday

Definition 3.2 (Outcome)**.** *These values are called **outcomes**.*

Examples:

- Heads and tails
- 1-6
- Obama or Romney
- Jan. 1, Jan. 2, . . . , Dec. 31

Notation: Sequential Experiences The simplest way to consider combinations of experiences is in sequence. We'll write $A \circ B$ to mean "Experience A followed by Experience B".

This sequence of experiences is an experience in itself. Here, order is relevant!

3.1.2 Principles of Counting

Theorem 3.1 (The Basic Principle of Counting)**.** *If A and B are two experiences with n and m possible outcomes, respectively, then $A \circ B$ has nm outcomes.*

Proof. We can prove this by a slightly more complex version of induction with two indices instead of one. That is, our propositions can be denoted by $P(m, n)$ for integer choices of m and n.

Denote the individual outcomes of A by $a_1, \ldots, a_m$, and the individual outcomes of B by $b_1, \ldots, b_n$.

For the base case, if n and m are both 1, then the only possibility is the outcome given by (a_1, b_1), and so the theorem holds, since $1 = 1 \times 1$. For the inductive step, instead of establishing a chain of implications, we need to establish a "lattice". Take the analogy of a city whose streets form a grid. I can get from one intersection to another which is northeast of me by going a certain number of blocks to the north, and a certain number to the east.

For our inductive hypothesis, suppose that the theorem holds for some values of m and n, call them j and k. We need to do two things: first, show that it holds for $j + 1$ and k (that is, show that wherever we are we can go north), and second that it holds for j and $k + 1$ (show that we can go east).

Then, for any m and n we want, we can get there from the base case by a sequence of steps increasing one index (walking north) and then the other (walking east).

So, suppose the theorem holds for $m = j$ and $n = k$, and first suppose that $m = j+1$ and $n = k$. If we exclude outcomes where A has the outcome a_{j+1}, then by the inductive hypothesis there are jk outcomes. Now, outcome a_{j+1} can be followed by any of the outcomes from b_1 through b_k, for a total of k more outcomes. This gives us a total of $jk + k = (j + 1)k$ outcomes, which is what we wanted.

The exact same logic establishes that if $m = j$ and $n = k+1$, then there are $j(k+1)$ outcomes. So for any m and n we want, we can establish $P(m, n)$ by noting that

$P(1,1)$ is true, which implies (by "walking north" that $P(m,1)$ is true. Then this, by a series of steps "to the east", implies that $P(m,n)$ is true. □

Example: You have six shirts and four pairs of pants. How many distinct outfits can you create (assuming you don't care about style)?

We can think of choosing an outfit as first choosing a shirt (experience A) and then choosing a pair of pants (experience B). Overall, choosing an outfit is the experience $A \circ B$. Since A has 6 possible outcomes and B has 4 possible outcomes, the principle of counting tells us that $A \circ B$ has $6 \times 4 = 24$ outcomes.

What if the experience we are interested in has more than two "sub-experiences"?

Theorem 3.2 (The Generalized Principle of Counting)**.** *If $A_1, A_2, \ldots, A_n$ are n separate experiences with $r_1, r_2, \ldots, r_n$ outcomes, respectively, then the combined experience, $A_1 \circ A_2 \circ \cdots \circ A_n$ has $r_1 \times r_2 \times \cdots \times r_n$ possible outcomes.*

Proof. We know this works if $n = 2$ by the previous Theorem. So we can proceed by induction on n, with the base case already done.

The inductive step proceeds by assuming that the theorem holds for $n = k$ experiences (the inductive hypothesis), and trying to show that this *implies* that it holds for $n = k + 1$.

This has a "recursive" quality to it: if we have $k + 1$ separate experiences, then any sequence constructed from these individual experiences is itself an experience. In particular, the first k form the sequential experience $A_1 \circ \cdots \circ A_k$.

We can just think of the experience $A_1 \circ A_2 \circ \cdots \circ A_{k+1}$ as the experience $A_1 \circ A_2 \circ \cdots \circ A_k$ followed by the experience A_{k+1}. To emphasize the fact that sequences are themselves experiences, let's introduce the notation B_k to represent the sequence of the first k individual experiences. That is to say, $B_k = A_1 \circ A_2 \circ \cdots \circ A_k$.

So we can write $B_{k+1} = B_k \circ A_{k+1}$.

By our inductive hypothesis, B_k has $r_1 \times r_2 \times \cdots \times r_k$ outcomes. By definition, A_{k+1} has r_{k+1} outcomes.

Now we can apply Theorem 3.1 to $B_k \circ A_{k+1}$ to get that B_{k+1} has $(r_1 \times r_2 \times \cdots \times r_k) \times r_{k+1}$ outcomes.

This completes the inductive step, and so we are done. □

Example: A restaurant offers a four-course menu with 3 soups, 4 appetizers, 4 main courses, and 2 desserts. How many different meals could you have?

If A_1 is the experience of choosing a soup (3 outcomes), A_2 is the experience of choosing an appetizer (4 outcomes), A_3 is the experience of choosing a main course (4 outcomes), and A_4 is the experience of choosing dessert (2 outcomes), then the combined experience of selecting a meal is $A_1 \circ A_2 \circ A_3 \circ A_4$. By the generalized principle of counting, there are $3 \times 4 \times 4 \times 2$ unique meals.

3.1.3 Arrangements

Theorems 3.1 and 3.2 concern sequences of experiences where each position is a different *type* of experience, with (potentially) a different set of possible outcomes. What about the case when every outcome in a set occurs sometime, and only the order differs?

Example: The American League Central Division has five teams: The Chicago White Sox, the Cleveland Indians, the Detroit Tigers, the Minnesota Twins, and the Kansas City Royals. How many ways can they be ranked (excluding ties)?

We can think of selecting a ranking as a sequence of five experiences: choosing the first place team, choosing the second place team, third place, etc. We have five choices for first place, and then no matter who takes first, there will be four remaining choices for second place, etc. By Theorem 3.2, we therefore have $5 \times 4 \times 3 \times 2 \times 1 = 120$ possible rankings.

In general we can state the following theorem.

Theorem 3.3. *The number of distinct arrangements of n objects in a row (where an "arrangement" means that every object gets used exactly once) is*

$$n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$$

Proof. Notice that this in a sense just a special case of Theorem 3.2. There are n possible outcomes for the first choice; $n-1$ for the second (since one is used up); $n-2$ for the third; and so on, down to only 1 choice for the n^{th}.

If that's not convincing, we can prove it by (you guessed it) induction. Clearly if there is 1 object, there is only 1 ordering, and so the base case is done.

Now, suppose that k objects can be arranged in $k \times \cdots \times 2 \times 1$ ways, and suppose we have $k+1$ objects. Then we can reduce the problem to two experiences: first choose the first object, and then order the rest. There are $k+1$ ways to choose the first object, and by the inductive hypothesis, there are $k \times \cdots \times 1$ ways to order the rest. Then by the basic principle of counting we have $k+1 \times (k \times k-1 \times \cdots \times 1)$ ways to order all k. □

Notation: Factorials
This kind of "falling multiplication" comes up a lot in all sorts of areas of discrete math. So that we don't have to write it out all the time, we use the shorthand $n!$ (read "n factorial") to mean $n \times (n-1) \times \cdots \times 2 \times 1$.

It doesn't make a lot of sense to talk about 0! in terms of this falling multiplication, but it does make sense in a way to talk about the number of ways to arrange 0 objects. There's only one: the "empty list". Therefore, for consistency, we define 0! to be 1.

Question: What happens if you divide one factorial by another?

Example: If each of 28 students is assigned to one of two discussion times, and there are 14 seats in the discussion room, how many ways are there for everyone to choose a seat if (a) the times are already chosen, and (b) if the times need to be assigned as well?

Example: How does this compare to the number of ways 28 people could be assigned to one big table with 28 seats?

3.1.4 Combinations

A pizza place is offering a special on two-topping pizzas. They expect a lot of orders, so they want to make some of the pies ahead of time. If they offer six different

toppings, how many different pies would they need to prepare to guarantee that they have any combination ready that someone might ask for?

You might ask a couple of questions. First, can you get double of one topping? That obviously changes the answer. Second, does the order of the toppings matter? It would be odd if it did; if someone asks for a sausage and pepper pizza but instead they get a pepper and sausage pizza instead, is that really a different pizza?

In general, we can talk about the number of ways to choose r things (in our example, r is 2, because we are choosing two toppings) from a group of n (here n is 6). Obviously this only makes sense if $0 \leq r \leq n$). We want to make two distinctions: first, can we repeat a choice? If so, we are sampling **with replacement**. If not, we are sampling **without replacement**. Second, does the order matter? In some cases, it will; in our pizza example it doesn't.

Let's consider each case in turn.

3.1.5 Sampling With Replacement

Order Relevant

If the same outcome can repeat (equivalently, if you can choose the same thing more than once), then we are sampling **with replacement**.

- If A_1 is the experience corresponding to what happens first, A_2 is the experience corresponding to what happens second, etc., then for each $i = 1, 2, \ldots, r$, the experience A_i has n possible outcomes.
- So, by Theorem 3.2, there are $n \times n \times \cdots \times n = n^r$ total possible outcomes.

Let's record this fact as a theorem.

Theorem 3.4. *The number of ways to sample r things from a pool of n,* with replacement *with* order relevant *is given by n^r.*

Note that in this calculation we are implicitly assuming that order matters: if A_1 takes outcome 1 and A_2 takes outcome 2, we are counting that as a different overall outcome from the one where A_1 takes outcome 2 and A_2 takes outcome 1.

Order Irrelevant

This case turns out to be pretty complicated — so much so that there's not a nice formula for counting the number of possibilities! So we're going to skip it; but I encourage you to think about how you would go about counting in this case!

3.1.6 Sampling Without Replacement

Example: How many "administrations": president, VP, secretary and stenographer, are possible from a class of 27?

Example: How many distinct "councils" of four "members-at-large" are there?

The difference between these is whether the members of the sample are distinguished from each other (order relevant), or whether the sample is just a "pool" of interchangeable members.

Order Relevant

If order is relevant, then we just have $A_1 \circ A_2 \circ \cdots \circ A_r$, with n choices for A_1, $n-1$ for A_2, $n-2$ for A_3, and so on. In general, experience A_r occurs after $r-1$ items have already been removed from the pool, and so we have $n-(r-1) = n-r+1$ choices.

So, again by Theorem 3.2, we get $n \times (n-1) \times \cdots \times (n-r+1)$.

This is just like counting the number of arrangements of n people, except that we stop choosing an item after the r^{th} position.

A shorthand for this "truncated factorial" is to write

$$\frac{n!}{(n-r)!}$$

since we want to leave off all the terms in the full $n!$ that are $n-r$ or smaller, which is accomplished by "cancelling them out" through division by $(n-r)!$.

We record this fact as follows.

Theorem 3.5. *The number of possibilities for sampling r items from a pool of n* without replacement *when* order is relevant *is*

$$n \times (n-1) \times \ldots (n-r+1)$$

or, equivalently

$$\frac{n!}{(n-r)!}$$

Order Irrelevant

If we don't care about order (such as in the pizza case, or the "members at large" case), then using the above results in overcounting, since we're tallying, say, "Adam, Liz, Carlos, Daniel" as different from "Carlos, Adam, Daniel, Liz", when we want them to be the same.

How bad is this overcounting?

Well, we're counting each combination exactly as many times as there are arrangements of the members. That is, $r!$ times. So if we start with the "order relevant" number, i.e., $\frac{n!}{(n-r)!}$, we can correct for the overcounting by dividing by $r!$. Then we are counting each combination once, for a total of $\frac{n!}{r!(n-r)!}$ combinations.

Example 2.5: An exam has ten questions on it. You have to answer seven out of the ten. How many choices do you have? What if you have to answer at least 3 out of the first five and 3 out of the last five?

3.1.7 Binomial Coefficients

Theorem 3.6 (Binomial Coefficients)**.** *There are*

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

ways to select r items from a set of n without replacement, with order irrelevant. These values are called ***binomial coefficients****.*

Proof. We just gave an intuitive justification for this formula in terms of correcting for overcounting. We could formally prove that this works by induction, but that turns out to be a bit more involved than usual, so we'll take a different approach.

One way to think about choosing r things out of a set of n is as one step on the way to getting a complete ordering of n things. Let B represent the experience of ordering all n things completely. We know from Theorem 3.3 that there will be a total of $n!$ ways of doing this.

If we want to put all n things in order, we could start by putting them into two piles: one representing the first r, and the other representing the remaining $n - r$. Call this division into piles A_1. The number of ways to do this is our binomial coefficient, $\binom{n}{r}$, which, for the moment, we suppose is unknown.

Once we've sorted our things into the two piles, then we need to order each pile in turn. Let A_2 represent ordering the first pile and A_3 represent ordering the second. Then A_2 has $r!$ outcomes (there are $r!$ ways to order the first pile), and A_3 has $(n-r)!$ outcomes.

Since $B = A_1 \circ A_2 \circ A_3$, the generalized principle gives us that

$$n! = \binom{n}{r} \times r! \times (n - r)! \tag{3.1}$$

Solving for the binomial coefficient, we get

$$\binom{n}{r} = \frac{n!}{r!(n - r)!} \tag{3.2}$$

□

So, revisiting our earlier examples, we can go back and calculate that there are

$$\begin{aligned} \binom{6}{2} &= \frac{6!}{2!4!} \\ &= \frac{6 \times 5}{2} = 15 \end{aligned}$$

different two-topping pizzas (assuming no double-toppings), and

$$\begin{aligned}\binom{27}{4} &= \frac{27!}{4!23!} \\ &= \frac{27 \times 26 \times 25 \times 24}{4 \times 3 \times 2 \times 1} \\ &= 9 \times 13 \times 25 \times 6 = 17550\end{aligned}$$

possible "councils" of "members at large".

Question: What's the relationship between $\binom{n}{r}$ and $\binom{n}{n-r}$?

There is an interesting recursive relationship between the binomial coefficients, which is represented by **Pascal's Triangle** (Fig. 3.1). If we arrange the binomial coefficients in rows, where each row represents a value of n, and the entries range from $r = 0$ to $r = n$, then the edges of the triangle consist all of 1s (since no matter what n is, if we're choosing nothing or choosing everything, there's only one way to do it), and any given entry is the sum of the two above it.

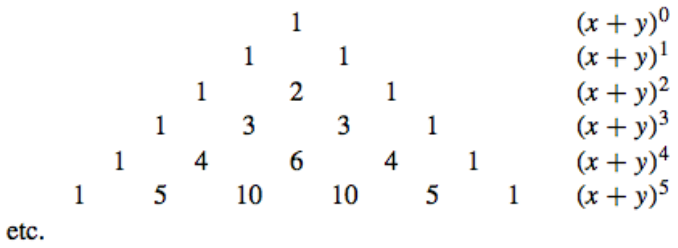

Figure 3.1: Pascal's Triangle

That is, for values of r other than 0 or n,

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r} \tag{3.3}$$

Intuitively, if we want to choose r things out of n, then from the first $n-1$ we either choose $r-1$ and then take the last one, or we choose r and omit the last one. These are mutually exclusive possibilities, with $\binom{n-1}{r-1}$ and $\binom{n-1}{r}$ options contained in each, respectively.

3.1.8 Multinomial Coefficients

Binomial coefficients can either be thought of as (a) the number of distinct subsets of size r from a larger set of n; or (b) the number of ways to divide n objects into two groups of sizes r and $n-r$, respectively (can you see why these are the same thing?)

If we view it the second way, it's easy to generalize the concept to an arbitrary number of "bins" (let's say k of them).

We can follow the same logic as before: in order to get a complete ordering of N objects, we can (1) group the objects into "bins", (2) decide how to order the objects in the first bin, (2) decide how to order the objects in the second bin, etc.

Let m represent the number we're trying to solve for — that is, the number of ways to divide the N objects into the bins. Then we have:

$$N! = m \times n_1! \times n_2! \times \cdots \times n_k! \tag{3.4}$$

Solving for m gives us

$$m = \frac{N!}{n_1! n_2! \cdots n_k!} \tag{3.5}$$

These are called **multinomial coefficients**. The notation for these is analogous to the one for the binomial coefficients.

Theorem 3.7 (Multinomial Coefficients)**.** *From an overall pool of N objects, there are*

$$\binom{N}{n_1, n_2, \ldots, n_k} = \frac{N!}{n_1! n_2! \ldots n_k!}$$

*ways to divide them into k groups, such that there are n_1 items in the first group, n_2 in the second, etc. order irrelevant. These values are called **multinomial coefficients**.*

Something to Think About: Randomly moving molecules tend to reach a "dynamic equilibrium", becoming evenly distributed in space. Can you explain why this is so using combinatorics?

Consider putting a bunch of particles in a cube which is divided into equal sized compartments, which the particles can freely travel between. A "macro level" description of the state of the system might consist of writing down how many particles there are in each compartment, whereas a "micro level" description would consist of identifying for each particle which compartment it's in. For any given macrostate, there are

multiple microstates that make it possible. Over time, as the particles move between compartments, the state of the system changes, such that in the long run, any given microstate is about as likely as any other. What does that mean for the likelihood of different macrostates?

3.2 The Algebra of Sets

The mathematics of probability is built on the mathematics of sets: sets are the basic object to which probabilities are assigned. So the last thing we need to do before we can officially start examining probability is to review some basic mechanics of sets and set operations.

3.2.1 Basic Terms and Notation

What is a set?

Modern mathematics has **set theory** at its foundation: a set is just a collection of "objects" (more precisely, symbols). I can have a set of numbers, a set of shoes, a set of people, a set of mathematical functions, anything that I might care to represent.

By definition, each member of a set is distinct (that is, there are no repetitions), and order doesn't matter. In terms of set equality, I can define a set called S_1, containing the integers from 1 to 3, but writing them in a different order doesn't change the set.

$$\{1, 2, 3\} = \{2, 1, 3\} = S_1$$

Sets don't have to be numeric in nature. $S_2 = \{Ana, Ben, Xiang\}$ is a perfectly valid mathematical set, for example.

Cardinality

The number of elements in a set is called its **cardinality**, which we can denote with the # symbol as follows:

$$\#(S_1) = \#(S_2) = 3$$

Many important sets have "infinite cardinality"; that is, they contain infinitely many things. Many commonly used sets of numbers have this property:

$$\begin{aligned}\mathbb{N} &= \{1, 2, 3, 4, \dots\} \\ \mathbb{Z} &= \{0, 1, -1, 2, -2, 3, -3, \dots\} \\ \mathbb{Q} &= \{0, 1, -1, 2, 1/2, -1/2, -2, 3, 1/3, -1/3, 2/3, -2/3, 3/2, -3/2, -3, 4, \dots\}\end{aligned}$$

The above are all infinite in "the same way": they can be put into a one-to-one correspondence with the "counting numbers" (i.e, $\mathbb{N}$). As such, they are called "countably infinite". This doesn't mean you can count them all — they're infinite, after all — it has to do with their correspondence to the counting numbers. To put it technically, each one can be put in a **bijection** (a one-to-one and onto function) with the set $\mathbb{N}$. In other words, we need to be able to come up with a scheme to put the elements in order so that we can figure out, for each counting number, which element is in that position, and for each element of the set, which position is it in.

In the case of the integers, we've used a scheme by which we start with zero, and then alternate between positive and negative integers, increasing the magnitude by 1 after each pair. So we know that any positive integer n occurs in position $2n$, and each nonpositive integer, $-n$, occurs in position $2n + 1$.

The scheme for the rational numbers is more complicated, but still well-defined: Proceed in the same way as for the integers, but before writing down a new integer, first list any fractions that can be constructed out of the integers listed so far, which do not reduce to something already in the list. In this way, we know that we will hit every rational number eventually.

The cardinality of these "countably infinite" sets is written as $\aleph_0$ (the Hebrew letter "aleph" with a subscript 0, known as "aleph nought"), though we won't really use that.

$$\#(\mathbb{N}) = \#(\mathbb{Z}) = \#(\mathbb{Q}) = \aleph_0$$

Other infinite sets have "too many" elements for this. For example, the real numbers, denoted by $\mathbb{R}$, has "uncountably infinite" elements: there is no way to map the set of reals onto the set of natural numbers in a one-to-one way, as the mathematician Cantor proved. Sets like this are said to have the "cardinality of the continuum", symbolized by $\mathfrak{c}$.

Intervals inside the reals are also uncountably infinite, which means that there are "more" numbers (in a well-defined way) between 0 and 1 than there are counting

numbers, even though both are infinite sets.

$$\#(\mathbb{R}) = \#([0,1]) = \mathfrak{c}$$

It seems strange that infinite sets that are subsets of each other can have the same cardinality. Two responses, in decreasing order of glibness, are that

(1) Infinity is wacky, and that

(2) The cardinality of two sets is said to be equal if and only if there exists an invertible function between them.

Elements and Subsets

A member of a set is called an **element**. We use the symbol $\in$ to say that the thing on the left is an element of the set on the right. For example,

$$1 \in \{1,2,3\}, \qquad Xiang \in \{Ana, Ben, Xiang\}$$

A set A is a **subset** of another set B if every element of A is also in B.

$$\begin{aligned} A &\subset B \\ \{2,3\} &\subset S_1 \\ \{Ana, Matt\} &\subset S_2 \end{aligned}$$

Sets with only one element are called **singleton sets**. They are different from elements.

$$1 \in S_1, \qquad \text{but} \qquad \{1\} \subset S_1$$

Note: I use $\subset$ to indicate general subsets, proper or improper. That is, a set is a subset of itself, and so I would write $A \subset A$, whereas some authors will write $A \subseteq A$ if they want to allow for the possibility that the two sets are equal, reserving $\subset$ for the case where the set on the right includes at least one element not in the set on the left. We will almost never care to make this distinction, and it's less writing. In the (rare) event that I want to denote a proper subset, I'll write $\subsetneq$.

If you are trying to prove that two sets are equal, you need to show that each one is a subset of the other. That is,

$$A = B \qquad \text{if and only if} \qquad A \subset B \quad \text{and} \quad B \subset A \tag{3.6}$$

Often times, subsets will be defined by some condition, such as "all numbers that are greater than 0", or "all students who have taken calculus". We can denote this using **set-builder notation**, which has the notation $\{x \in B; x \text{ satisfies some condition(s) }\}$, and is read as "all elements of B such that they meet the condition". For example, we can denote the set of even numbers as

$$\{x \in \mathbb{Z}; x/2 \in \mathbb{Z}\}$$

and we can represent intervals (which are subsets of the real numbers), as

$$\begin{aligned}
[a, b] &= \{x \in \mathbb{R}; a \leq x \leq b\} \\
(a, b) &= \{x \in \mathbb{R}; a \leq x \leq b\} \\
[a, b) &= \{x \in \mathbb{R}; a \leq x < b\} \\
(a, b] &= \{x \in \mathbb{R}; a < x \leq b\}
\end{aligned}$$

The **Venn Diagram** is a useful tool for visualizing subset relations.

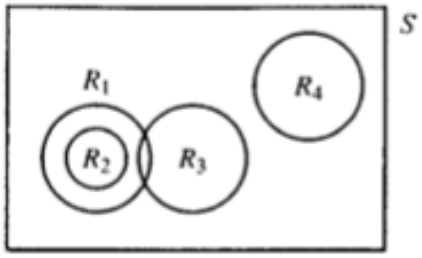

Figure 3.2: A Venn Diagram

3.2.2 Set Operations

In the same way that we apply operations like addition, negation, exponentiation, etc. to numbers, we can apply operations to sets. In probability we will always be working inside some "universal set", S. In this case we define the **complement** of a set A, which we write as $\overline{A}$, as all the elements in S that are not in A.

Definition 3.3. *Set Complement The* **complement** *of a set A in the context of a universal set S is written $\overline{A}$, and defined as*

$$\overline{A} = \{x \in S; x \notin A\} \tag{3.7}$$

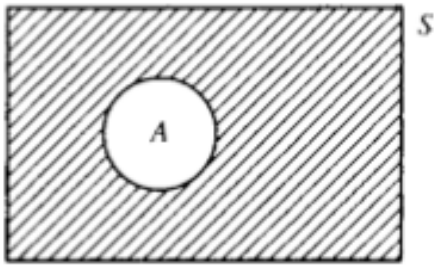

Figure 3.3: The Complement of a Set

The **union** of two sets A and B is the set that includes every element of S that is either in A OR B (it could be in both; that is, OR is inclusive OR).

Definition 3.4. *Set Union The* **union** *of two sets, A and B (in the context of a universal set S), is written $A \cup B$, and defined as*

$$A \cup B = \{x \in S; x \in A \ OR \ x \in B\} \tag{3.8}$$

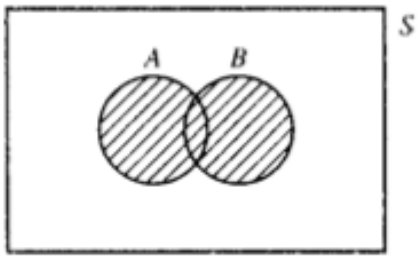

Figure 3.4: The Union of Two Sets

The **intersection** of two sets A and B is the set that includes only those elements of S that are in both A AND B.

Definition 3.5. *Set Intersection The* **intersection** *of two sets A and B (in the context of a universal set S) is written $A \cap B$ and defined as*

$$A \cap B = \{x \in S; x \in A \ AND \ x \in B\} \tag{3.9}$$

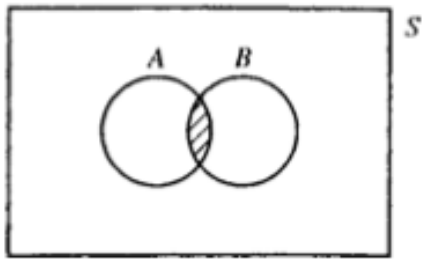

Figure 3.5: The Intersection of Two Sets

Note that intersection can only reduce the size of a set, whereas union can only increase it. For any two sets A and B, we have:

$$A \cap B \subset A \subset A \cup B \tag{3.10}$$
$$A \cap B \subset B \subset A \cup B \tag{3.11}$$

The Empty Set

Definition 3.6. *Empty Set The* ***empty set*** *is written* $\emptyset$*. It is the unique set that has no elements at all.*

We'll review some properties of the empty set shortly. It plays a particularly important role for us in defining "mutually exclusive", or **disjoint** sets.

Definition 3.7. *Disjoint Sets Two sets, A and B, are said to be* **disjoint** *if and only if they have no elements in common; in other words, if and only if*

$$A \cap B = \emptyset \tag{3.12}$$

Disjoint sets have special significance in probability, as we will soon see.

3.2.3 Boolean Laws

There is a fundamental relationship between set theory and logic. Logical predicates can be evaluated on objects of a certain kind; that is, in a "universal set" (our S). Each object in the universe takes on a value of TRUE or FALSE with respect to the predicate. Thus, predicates can be thought of as identifying *subsets*. We get the

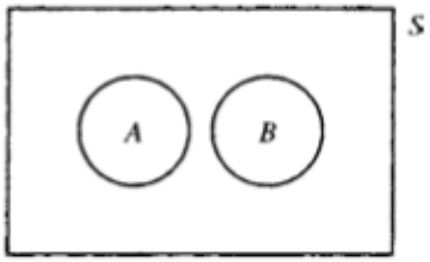

Figure 3.6: Two Disjoint Sets

following correspondence between logical objects and statements, and set theoretic objects and statements.

Logic	**Set Theory**
The predicate A	The set A
Trivially TRUE	The universal set, S
Trivially FALSE	The empty set, $\emptyset$
NOT A	$\overline{A}$
A OR B	$A \cup B$
A AND B	$A \cap B$
$A \implies B$	$A \subset B$
$A \iff B$	$A = B$

Table 3.1: Logic/Set Correspondences

There are some basic laws that apply to sets (and their propositional analogs) that are useful in simplifying expressions (see Table 3.2).

Many times when we want to find the probability of some proposition, it is easier to find the probability of a different but logically equivalent one. For example, the statement, "At least one of A, B, and C is TRUE" can be represented in set terms

Law	Set Formulation	Logical Formulation
Axioms of Two-Valued Logic	$\overline{\overline{A}} = A$	NOT (NOT A) $\iff$ A
	$A \cup \overline{A} = S$	A OR (NOT A) is TRUE
	$A \cap \overline{A} = \emptyset$	A AND (NOT A) is FALSE
	$\overline{S} = \emptyset$	NOT TRUE $\iff$ FALSE
	$\overline{\emptyset} = S$	NOT FALSE $\iff$ TRUE
Idempotence	$A \cup A = A$	A OR A $\iff$ A
	$A \cap A = A$	A AND A $\iff$ A
Commutativity	$A \cup B = B \cup A$	A OR B $\iff$ B OR A
	$A \cap B = B \cap A$	A AND B $\iff$ B AND A
Associativity	$(A \cup B) \cup C = A \cup (B \cup C)$	(A OR B) OR C $\iff$ A OR (B OR C)
	$(A \cap B) \cap C = A \cap (B \cap C)$	(A AND B) AND C $\iff$ A AND (B AND C)
Distributivity	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	A OR (B AND C) $\iff$ (A OR B) AND (A OR C)
	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	A OR (B AND C) $\iff$ (A OR B) AND (A OR C)
DeMorgan's Laws	$\overline{A \cup B} = \overline{A} \cap \overline{B}$	NOT(A OR B) $\iff$ (NOT A) AND (NOT B)
	$\overline{A \cap B} = \overline{A} \cup \overline{B}$	NOT(A AND B) $\iff$ (NOT A) OR (NOT B)
Conjunctive and Disjunctive Identities	$A \cup S = S$	A OR TRUE is TRUE
	$A \cap S = A$	A AND TRUE $\iff$ A
	$A \cup \emptyset = A$	A OR FALSE $\iff$ A
	$A \cap \emptyset = \emptyset$	A AND FALSE is FALSE
Trivial Containment	$\emptyset \subset A$	FALSE $\implies$ A

Table 3.2: Boolean Laws

as:

$$\begin{aligned}
A \cup B \cup C &= \overline{\overline{A \cup B \cup C}} && \text{(Two-Valued Logic)} \\
&= \overline{\overline{(A \cup B) \cup C}} && \text{(Associativity)} \\
&= \overline{\overline{(A \cup B)} \cap \overline{C}} && \text{(DeMorgan's Laws)} \\
&= \overline{(\overline{A} \cap \overline{B}) \cap \overline{C}} && \text{(DeMorgan's Laws)} \\
&= \overline{\overline{A} \cap \overline{B} \cap \overline{C}} && \text{(Associativity)}
\end{aligned}$$

We have used the laws to show the intuitive equivalence between "At least one is TRUE", and "It is not the case that all are FALSE". The latter can be more useful in many probabilistic contexts.

Another simple example is "Exclusive OR (XOR)". We want "A OR B but not both", i.e.:

$$\begin{aligned}
(A \cup B) \cap \overline{A \cap B} &= (A \cup B) \cap (\overline{A} \cup \overline{B}) && \text{(DeMorgan's Laws)} \\
&= (A \cap (\overline{A} \cup \overline{B})) \cup (B \cap (\overline{A} \cup \overline{B}) && \text{(Distributivity)} \\
&= (A \cap \overline{A}) \cup (A \cap \overline{B}) \cup (B \cap \overline{A}) \cup (B \cap \overline{B}) && \text{(Distributivity)} \\
&= \emptyset \cup (A \cap \overline{B}) \cup (B \cap \overline{A}) \cup \emptyset && \text{(Two-Valued Logic)} \\
&= (A \cap \overline{B}) \cup (\overline{A} \cap B) && \text{(D.I. Plus Commutativity)}
\end{aligned}$$

3.2.4 Set Difference

It is often useful to talk about the **difference** of two sets, which is the set consisting of members of the first set that are *not* also members of the second. We define the notation

$$A - B = A \cap \overline{B} \tag{3.13}$$

Notice that if A and B are *disjoint*, then $A - B = A$ and $B - A = B$. Also, A and $B - A$ are disjoint, regardless of the initial status of A and B, and $A \cup (B - A) = A \cup B$, because $B - A$ explicitly excludes anything in A. We can use this to turn a general

union of sets into a disjoint union of sets:

$$\begin{aligned} A_1 \cup A_2 \cup A_3 &= A_1 \cup (A_2 - A) \cup A_3 \\ &= A_1 \cup (A_2 - A_1) \cup (A_3 - (A_1 \cup (A_2 - A_1))) \\ &= A_1 \cup (A_2 - A_1) \cup (A_3 - (A_1 \cup A_2)) \end{aligned}$$

More generally, if we have n sets: $A_1, A_2, \ldots, A_n$, then:

$$\bigcup_{i=1}^{n} A_i = A_1 \cup (A_2 - A_1) \cup \cdots \cup \left[A_k - \bigcup_{i=1}^{k=1} A_i \right] \cup \cdots \cup \left[(A_n - \bigcup_{i=1}^{n-1} A_i) \right]$$

3.3 Exercises

1. A game is constructed in such a way that a small ball can travel down any one of three possible paths. At the bottom of each path there are four traps which can hold the ball for a short time before propelling it back into the game. In how many alternatives ways can the game evolve thus far?
2. How many ways can seven colored beads be arranged (a) on a straight wire, (b) on a circular necklace? (Hint: think about the kinds of manipulations that would or would not result in a "distinguishable" configuration of beads)
3. How many distinguishable five-card poker hands are possible? Note that re-ordering the cards in your hand does not change the nature of the hand.
4. Fifteen people enter a raffle, which has distinct first, second and third prizes. How many different combinations of winners are there?
5. You find six shirts at store A that you like, and eight more at store B, but you only want to buy four shirts in total. How many different selections can you make?
6. A firm has to choose seven people from its R and D team of ten to send to a conference on computer systems. How many ways are there of doing this
   (a) when there are no restrictions?
   (b) when two of the team are so indispensable that only one of them can be permitted to go?
   (c) when it is essential that a certain member of the team goes?

(d) when there are two subteams of five, each of which must send at least two people?

7. You register for an account at a website, and you are issued a temporary alphanumeric password which is six characters long and is *not* case-sensitive. Give an expression for the number of possible passwords under each of the following restrictions.

   (a) All characters are letters.

   (b) All characters are letters and the first one is 'q'.

   (c) All characters are *different* letters.

   (d) There are three letters and three numbers.

   (e) There is at least one number.

8. I'm having a party and we're ordering pizza. Some of my guests are vegetarian, and others insist on meat. My local pizza parlor offers three veggie toppings and three meat toppings. If I want to order two pies, each with two different toppings, such that one is vegetarian-friendly and the other has at least one meat topping, how many choices do I have?

9. (*) Suppose you have N objects, and you want to divide them into r groups so that the first group has n_1 objects, the second has n_2 objects, etc. Use induction on the number of groups to show that the number of ways to do this is

$$\frac{N!}{n_1!n_2!\ldots n_r!}$$

   Hint: The case where $r = 2$ is covered by the binomial coefficients. For the inductive step, break up the process into two parts. First, select which objects will go into the last group, and then divide the remaining objects.

10. Recall that we define the set theoretic difference, $A-B$ to be the set of elements that are in A but not in B. That is,

$$A - B = A \cap \overline{B}$$

    Use set algebra to show that the following equalities hold. You may want to draw a Venn diagram first, to convince yourself. I've done the first one for you.

    (a) $A \cap (B - C) = (A \cap B) - (A \cap C)$

Starting with the right-hand side:

$$\begin{aligned}
(A \cap B) - (A \cap C) &= (A \cap B) \cap \overline{(A \cap C)} && \text{(Def. of set diff.)} \\
&= (A \cap B) \cap (\overline{A} \cup \overline{C}) && \text{(DeMorgan's)} \\
&= (A \cap B \cap \overline{A}) \cup (A \cap B \cap \overline{C}) && \text{(Distributive)} \\
&= ((A \cap \overline{A}) \cap B) \cup (A \cap B \cap \overline{C}) && \text{(Associative)} \\
&= (\emptyset \cap B) \cup (A \cap B \cap \overline{C}) && \text{(Property of Complement)} \\
&= \emptyset \cup (A \cap B \cap \overline{C}) && \text{(Prop. of } \emptyset\text{)} \\
&= (A \cap B \cap \overline{C}) && \text{(Prop. of } \emptyset\text{)} \\
&= A \cap (B \cap \overline{C}) && \text{(Assoc.)} \\
&= A \cap (B - C) && \text{(Def. of Set Diff.)}
\end{aligned}$$

(b) $(A \cup B) - C = (A - C) \cup (B - C)$

(c) $A - (A - B) = A \cap B$

(d) $(A - B) - C = (A - C) - B = A - (B \cup C)$

(e) $A - (A \cap B) = A - B$

11. We can define the **symmetric difference** of two sets as the elements that are in one or the other, but not both (this is analogous to XOR in logic). Formally, we define

$$A \nabla B = (A \cup B) - (A \cap B)$$

(See the previous problem for the definition of the $-$ symbol.)

Use set algebra to establish the following properties of the symmetric difference.

(a) $A \nabla B = B \nabla A$

Starting with the left-hand side:

$$\begin{aligned}
A \nabla B &= (A \cup B) - (A \cap B) && \text{(Def. of symmetric diff.)} \\
&= (B \cup A) - (B \cap A) && \text{(Commutativity of } \cup \text{ and } \cap\text{)} \\
&= B \nabla A && \text{(Def. of symmetric diff.)}
\end{aligned}$$

(b) $A \nabla \emptyset = A$

(c) $A \nabla A = \emptyset$

(d) $A \nabla S = \overline{A}$

12. Define the set operation $\uparrow$ (analogous to NAND in logic) by setting

$$A \uparrow B = \overline{A \cap B}$$

Show that all other set operations can be built out of applications of $\uparrow$. In particular, show that, for any sets A and B:

(a) $A \uparrow A = \overline{A}$

(b) $(A \uparrow A) \uparrow (B \uparrow B) = A \cup B$

(c) $(A \uparrow B) \uparrow (A \uparrow B) = A \cap B$

Part II

Fundamentals of Conditional Probability

Chapter 4

Probability

4.1 Our Goal

Recall the first week of class when we discussed "plausible reasoning", and "weak inference". We engaged in reasoning of the form:

(Premise) If A is true then B becomes more plausible

(Data) B is true

(Inference) A becomes more plausible

But suppose we had to make a decision that depended on the status of A (for example, a jury hears evidence and needs to reach a guilty or not guilty verdict). It makes a big difference whether, by "more plausible" we mean

1. "Almost certain to be true", or
2. "Slightly less implausible than before, but still quite unlikely"

Recall our original goal: we want to develop a formal system that allows us to build models to carry out plausible reasoning. That is, we want a system that can carry out the intuitively reasonable "weak syllogisms", ultimately evaluating the **plausibility** of a proposition given some observations.

As we have seen, plausibility is not a binary quality: it is not the case that a statement is either plausible or implausible; there are degrees of plausibility. So it seems

reasonable that one criterion for our formal definition of plausibility is that it should be represented quantitatively — with real numbers.

Second, if we want a numeric scale to represent degrees of plausibility, it seems reasonable that greater numbers should be assigned to propositions which are more plausible, and smaller numbers to those which are less plausible.

Finally, we don't want our plausibility system to contradict itself. This has three implications.

First, if two propositions are logically equivalent in terms of what they say about the world (that is, if each logically implies the other), then they should be assigned the same plausibility, even if they contain different symbols. For example, if A, B and C are statements to be assigned plausibilities, then, the statement A AND (B OR C) must have the same plausibility as (A AND B) OR (A AND C), since they are logically equivalent.

Second, we don't want to ignore any evidence. For example, if we know that our car battery is dead, we need to take that into account in evaluating the plausibility that the fuel tank is empty; we mustn't reason from the gauge reading to the tank while ignoring the battery once we know about it.

Finally, if there are multiple reasoning paths from what we know to a conclusion, they should give the same plausibility assignment.

To make these criteria a bit more formal, we'd like a plausibility function, P, that takes a statement and returns a plausibility, and which has the following properties:

1. Plausibility is represented with a real number (i.e., $P(A) \in \mathbb{R}$ for any statement A).

2. Plausibility assignments should accord with common sense. That is, if A_1 is more plausible than A_2, then we should have $P(A_1) \geq P(A_2)$.

3. The reasoning mechanism must be consistent. Plausibilities should depend on all and only the available information, and not the specific syllogisms used, or the particular representation of a proposition.

The physicist Richard T. Cox proved that a plausibility scale that has all of these properties must obey the rules of probability. The proof is too involved for our purposes, but a version of it appears in Chapter 2 of the Jaynes book.

4.2 What Is Probability?

Before we can state the rules of probability, we need to say what kinds of things probabilities apply to. The reason we're doing this in the first place is to assign plausibilities to propositions; so what is the analog of a proposition in the world of probabilities?

4.2.1 Universe and Events

Typically we apply probabilities to *experiences* (in the general sense of the term we used earlier), which are associated with *outcomes*.

Definition 4.1 (Universe)**.** *For some experience, the associated* ***universe*** *is the* set*, U, consisting of all the outcomes associated with the experience. (This is more traditionally called the* ***sample space****.)*

Examples of Universes:

- Outcomes of a die roll: $\{1, 2, 3, 4, 5, 6\}$
- Outcomes of a pregnancy test: $\{+, -\}$
- Weather states: $\{\text{Rain}, \text{No Rain}\}$
- Winning candidates: $\{\text{Bush}, \text{Gore}, \text{Nader}\}$
- Jury verdicts: $\{\text{Guilty}, \text{Not Guilty}\}$

Definition 4.2 (Event)**.** *An* ***event****, E, is a proposition that becomes either TRUE or FALSE once we know the outcome of the experience. The proposition E can be equated with the set E of outcomes in U that make E TRUE.*

Example Events:

- "I roll an even number" $= \{2, 4, 6\}$
- "The test comes out positive" $= \{+\}$
- "A major party candidate wins" $= \{\text{Bush}, \text{Gore}\}$

An important concept that we will need later is that of a **partition** of the universe.

Definition 4.3 (Partition)**.** *The set of sets $\mathcal{E} = \{E_1, E_2, \ldots, E_n\}$ is called a* **partition** *of the universe U if every pair of events in $\mathcal{E}$ is disjoint, and the union of all the E_js is the whole universe. That is,*

(i) For every i, j from 1 to n such that $i \neq j$, we have

$$E_i \cap E_j = \emptyset$$

(ii)

$$E_1 \cup \cdots \cup E_n = U$$

Intuitively, a partition is a set of events that are mutually exclusive and exhaustive: every outcome in U is in exactly one (no more, no less) of the events in the partition.

For example, if the universe consists of the outcomes of a die roll ($U = \{1, 2, 3, 4, 5, 6\}$), then examples of partitions of U might be

$$\mathcal{E}_1 = \{\{1, 2, 3\}, \{4, 5, 6\}\}$$
$$\mathcal{E}_2 = \{\{1, 3, 5\}, \{2, 4, 6\}\}$$
$$\mathcal{E}_3 = \{\{1, 6\}, \{2, 5\}, \{3, 4\}\}$$

If the universe contains finitely many outcomes (i.e. U has finite cardinality), then an obvious partition consists of **atomic events**; that is, those with exactly one outcome. In the die roll example, this partition is

$$\mathcal{E}_4 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$$

4.2.2 Kolmogorov Axioms

We want to assign a real number, called a **probability**, to each event based on its plausibility. We adopt the following conventions:

1. Probabilities cannot be negative
2. The event corresponding to U itself has probability 1

   The requirements of common sense and consistency also suggest a third property, namely that of disjoint additivity:

3. If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

Collectively, these properties are known as **Kolmogorov's Axioms**, after the Russian mathematician, and represent the classical foundation of probability.

4.2.3 Probability Measures

Our goal is to define a function, P, whose domain is the set of events, whose range is the real numbers (or some subset thereof), and which obeys the constraints

(K1) Nonnegativity: $P(A) \geq 0$ for every event, A

(K2) Unity of the universe: $P(U) = 1$.

(K3) Disjoint additivity: $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$.

These defining properties entail several other properties

Properties of Probability Measures For all events A and B in the domain of P:

(P1) $P(\emptyset) = 0$

(P2) The complement rule: $P(\overline{A}) = 1 - P(A)$

(P3) General addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(P4) If $A \subset B$ then $P(B - A) = P(B) - P(A)$

(P5) Monotonicity: $P(A) \leq P(B)$ whenever $A \subset B$

(P6) Probability bounds: $0 \leq P(A) \leq P(U)$

(P7) If $\mathcal{E} = \{E_1, E_2, \ldots, E_n\}$ is a partition of U, then $\sum_{j=1}^{n} P(E_1) = 1$.

Some of the proofs are left for you to do as homework, but let's look at a couple of them.

Proof of (P1). Intuitively this must be the case: the chances that something in $\emptyset$ happens are 0, since there's nothing in there. But the point of proving this is that we want to show that it follows from the Kolmogorov axioms without having to state it as a separate principle.

The proof relies on the fact that the empty set is disjoint with any arbitrary set; that is, for any set A, $A \cap \emptyset = \emptyset$ (this was one of the set axioms). Therefore, we can take advantage of (K3), the disjoint additivity principle of probabilities, to write that

$$P(A \cup \emptyset) = P(A) + P(\emptyset)$$

Now, it is another set axiom that $A \cup \emptyset = A$, so substituting this in the left-hand side gives

$$P(A) = P(A) + P(\emptyset)$$

and now solving for $P(\emptyset)$ yields 0. □

Let's try another one.

Proof of (P2). The "complement rule" says that we can get the probability of a set's complement by subtracting the probability of the original set from 1. Intuitively, this says that the plausibility that something is false is 1 minus the probability that it's true.

Again we're going to make use of the disjoint additivity axiom (P3), this time noticing that for any set A, A and $\overline{A}$ are disjoint (they have no elements in common, by definition).

We also know that the union of A and $\overline{A}$ is the entire universe, U. That is, either A happens, or A doesn't happen; there's nothing else (this is another set axiom).

Combining these two facts, we can write

$$\begin{aligned} P(A \cup \overline{A}) &= P(A) + P(\overline{A}) \\ \implies P(U) &= P(A) + P(\overline{A}) \\ \implies 1 &= P(A) + P(\overline{A}) \\ \implies P(\overline{A}) &= 1 - P(A) \end{aligned}$$

where in the second-to-last line we've made use of (K2), the axiom that $P(U) = 1$. □

Let's try one more, this time one that involves an inequality.

Proof of (P5). This says that if A is a subset of B (in propositional terms, A implies B), then its probability cannot exceed the probability of B (in other words, a more specific event can't have a higher probability than a more general one).

This is proved by rewriting the larger set as the union of two disjoint sets: A and $B - A$ (remember, the latter is everything in B which is not in A, i.e., $B \cap \overline{A}$). Then we have

$$P(B) = P(A \cup (B - A)) = P(A) + P(B - A) \geq P(A) + 0$$

where we have used (K3), followed by (K1), which says that $P(A) \geq 0$. □

4.3 Interpreting Probability

So far, we've established that $P(U) = 1$, that $P(\emptyset) = 0$, and we can say certain things about probabilities of unions, intersections and complements. But how do we know what probabilities to assign to individual (non-empty) events in the first place?

Intuitively, it would seem that our system should resemble the way we talk about plausibility in natural language.

What do we mean when we say:

- "The probability a die will come up 2 is $1/6$"?
- "If a woman who is pregnant takes a pregnancy test, the probability that she will test positive is 0.8"?
- "The probability that a woman who gets a positive pregnancy test is pregnant is 0.7"?
- "There is a 40% chance of rain"?
- "Hillary Clinton has a 60% chance of winning the Democratic nomination for president in 2016"?
- "I'm 90% confident that the defendant is guilty"?

Do we mean the same thing by all of these?

4.3.1 The Principle of Exchangeability

For some simple experiences (flipping a coin, rolling a die), the elements of the universe are intuitively **exchangeable**: we could rearrange the labels without changing

the nature of the experience. That is to say, we have no reason to think that any one outcome is more likely than any other.

In these cases, provided the sample space is finite, we have

$$U = \{a_1, a_2, \ldots, a_n\}$$

and for all "atomic" events (those containing only one outcome):

$$P(\{a_1\}) = P(\{a_2\}) = \cdots = P(\{a_n\}) \tag{4.1}$$

Call this common value p. Since the atomic events form a partition of U, (K2) and (P7) imply

$$\begin{aligned} 1 = P(U) = \sum_{j=1}^{n} P(\{a_j\}) = \sum_{j=1}^{n} p = np \\ \implies p = \frac{1}{n} \end{aligned} \tag{4.2}$$

For an arbitrary set $A \in \mathcal{P}(S)$, we can write A as a disjoint union of r singleton sets (for some r). Therefore we can apply disjoint additivity (aka (K3), extended to more than 2 events using Exercise 9 at the end of this chapter) to get that

$$P(A) = P(\bigcup_{a_j \in A} \{a_j\}) = \sum_{a_j \in A} P(\{a_j\}) = \sum_{a_j \in A} \frac{1}{n} = \frac{r}{n} = \frac{\#(A)}{\#(U)} \tag{4.3}$$

In words, if the outcomes are exchangeable, then the probability of an event is just the number of ways it can be true divided by the total number of possibilities in the universe.

Random Samples In statistics, we are interested in making generalizations from samples to populations. When all possible samples are exchangeable, then we have a **simple random sample** (which can be with or without replacement, depending on what we mean by "all possible samples").

Example: Defective Items In a batch of 30 manufactured items, 5 are defective. If we choose 3 at random without replacement, what is the probability that all 3 are defective?

First, let's identify the universe. We are drawing 3 items from a batch of 30, so the U contains all possible subsets of 3 items from a batch of 30.

Now, what's the event of interest? The relevant condition is that all 3 items are defective, so A is the set of groups of 3 items, all defective.

Assuming the subsets in U are exchangeable (there's no bias in our selection procedure and no propensity for any items to be selected together), then we can use the equation

$$P(A) = \frac{\#(A)}{\#(U)}$$

and so all that's left is to do some counting.

How many elements are in U? We are choosing 3 items out of 30, without replacement, order irrelevant. There are

$$\#(U) = \binom{30}{3}$$

ways to do this. How many elements are in A? There are 5 defective items in the batch, so how many different groups of 3 are there out of these 5? Again,

$$\#(E) = \binom{5}{3}$$

If the principle of symmetry applies, the probability is

$$P(E) = \frac{\#(E)}{\#(S)} = \frac{\binom{5}{3}}{\binom{30}{3}}$$

Example: The Birthday Problem What is the probability that in a room of 27 people, there is a pair with the same birthday?

The universe is the set of all possible combinations of 27 birthdays. The event of interest is all combinations with at least one repetition. How many outcomes are in U?

U represents the set of outcomes for the experience $A_1 \circ A_2 \circ \ldots A_{27}$, where A_j represents the birthday of the j^{th}person. By Theorem 3.2, there are $365 \times 365 \times \cdots \times 365 = 365^{27}$ outcomes in S, and so if the principle of symmetry applies, every outcome has probability $1/365^{27}$.

How many individual outcomes are in E? This is pretty difficult to count directly because of all the different ways that a repetition could occur. It's much easier to count the outcomes in $\overline{E}$. So how many sequences of 27 *different* birthdays are there? This is like sampling without replacement, but where order matters (since each position is a different person). So, we have $\#(E) = \frac{365!}{(365-27)!} \approx 7.109 \times 10^{40}$. Then, assuming the combos are exchangeable, $P(\overline{E}) = \frac{365!}{(365-27)!}/365^{27} \approx 0.373$. Then, by the Complement Rule (aka (P5)), $P(E) = 1 - P(\overline{E}) \approx 0.627$.

4.3.2 Relative Frequency

The principle of symmetry is not always defensible. For example:

- What is the probability that you will win the lottery?
  ($U = \{\text{Win}, \overline{\text{Win}}\}$)
- What is the probability of rain?
  ($U = \{\text{Rain}, \text{Sun}, \text{Snow}, \text{Hail}\}$)
- What's the probability that a pregnant women gets a positive test result?
  ($U = \{+, -\}$)

In some cases, we can modify the sample space to create symmetric outcomes (as in the case of the die). But sometimes there's no sensible way to do this. Instead, we might appeal to **relative frequencies**: in all previous instances of this experience that we have access to, what proportion of the time did E occur?

- Of all the pregnant women who've taken the test before, 80% tested positive.
- Of all previous days with similar measurements of barometric pressure, wind speed, etc., it has rained in 40% of cases.

Relative Frequency and the Kolmogorov Axioms

Remember the three axioms we specified

(K1) Probabilities cannot be negative

(K2) The event corresponding to U itself has probability 1

(K3) If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

These all follow naturally from the relative frequency interpretation:

1. Fractions of counts over counts cannot be negative.
2. The proportion of the time that the outcome is in the set of possible outcomes is always 1.
3. If I have two disjoint sets of outcomes, then together their share of the whole is the sum of their individual shares.

Problems With Relative Frequency

Obtaining probabilities from relative frequencies is intuitive, but it is not always easy to do in a well-defined way. For one, proportions can come out differently on different occasions. Do we really want to say the probability is different based on our specific data if the conditions were the same?

The usual response is to define the probability not in terms of any specific data, but in terms of a **limit** of proportions.

But there is a generalization problem: how do we know what data to count? That is, what counts as an instance of "the same" experience? On the one hand, if we're very broad, then we may not convey useful information (e.g., if we just look at all previous days to compute the probability of rain, that's not very useful) If on the other hand we're very narrow, then we may have very little data to use (e.g., if I flip a coin once and it comes up heads, I wouldn't want to say that the probability that it will come up heads the next time is 1, just because I don't want to consider data from any coins that aren't exactly like this one)

4.3.3 Subjective Probability

The above issues are avoided by using **subjective probabilities**: assigning probabilities to basic outcomes based on one's impression of their relative plausibility / likelihood of occurrence. We can then apply the laws of probability to get a set of probabilities, which although *subjective*, are at least *consistent*.

Probability and Odds

It is in principle possible to elicit a person's subjective probabilities by asking them to place bets on an event, with particular **odds**.

Odds What is the maximum amount you would be willing to bet on a coin coming up heads if you would win \$1 if you are correct? How about on a die coming up 6? How about on the proposition that it rains tomorrow?

It makes sense to calibrate your bet based on your impression of the probability: if you repeatedly bet \$1 on "Heads", you will lose \$1 every time the coin comes up tails, and win \$1 every time it comes up heads. In the long run, if these happen equally often ($P(\text{Heads}) = \frac{1}{2}$), then you would break even; so if you have the opportunity to win \$1 by risking anything less than \$1, it's probably a good deal; but anything more than \$1 is a losing proposition in the long run.

With the die, you would presumably be more conservative: after all, you expect to lose five times as often as you win. But if you had the opportunity to put down \$0.20 for a chance to win \$1, then you would expect to break even in the long run. If the price is any greater, you'll lose money in the long run; any cheaper and you expect to win in the long run.

So, when we have a probability associated with an event, A, we can define the corresponding **odds** for A as the price-to-payout ratio associated with betting on A that we expect to result in our breaking even in the long run. The more likely A is to occur (i.e., the more likely it is that our bet will win), the more we might be willing to pay for the opportunity; and if we expect to win k times as often as we win, then we can afford to pay k times as much to play as we would win if we are successful. In other words,

Definition 4.4 (Odds)**.** *Suppose event A has probability $P(A)$. Then the* **odds** *of A are defined to be*

$$\text{Odds}(A) = \frac{P(A)}{P(\overline{A})}$$

So then, how do odds translate back to probability?

If our goal is to use people's subjective odds to discover the subjective probability they attach to an event, then we can just solve the equation above for $P(A)$, making use of the complement rule.

$$\begin{aligned} \text{Odds}(A) &= \frac{P(A)}{P(\overline{A})} = \frac{P(A)}{1 - P(A)} \\ \implies \frac{1}{\text{Odds}(A)} &= \frac{1}{P(A)} - 1 \\ \implies \frac{1}{P(A)} &= \frac{1}{\text{Odds}(A)} + 1 \\ \implies P(A) &= \frac{1}{\frac{1}{\text{Odds}(A)} + 1} \end{aligned}$$

Notice that if we express the odds as a fraction, $\frac{x}{y}$, then we can simplify the above:

$$P(A) = \frac{1}{\frac{y}{x} + 1} = \frac{\frac{x}{x}}{\frac{y}{x} + \frac{x}{x}} = \frac{x}{y + x}$$

Subjective Probability and the Kolmogorov Axioms

A set of odds is called **incoherent** if you can place multiple bets in such a way that you are guaranteed a profit regardless of the outcome (assuming that a bet on $\overline{A}$ is available at odds of $1/\mathsf{Odds}(A)$).

It can be shown that the probabilities induced by coherent odds must obey the Kolmogorov axioms. For example, the odds of any given event must be positive, since otherwise you can guarantee a profit by betting for negative cost and positive payout. This implies that probabilities are bounded between 0 and 1 (as you should verify). Moreover, if two events, A and B, are disjoint, then the odds will be incoherent if the implied $P(A \cup B)$ is not the implied $P(A)$ plus the implied $P(B)$.

The Scope of Subjective Probabilities

Framing probabilities in terms of degrees of confidence or belief allows us to assign probabilities not only to random phenomena, but to any proposition whose truth value we don't know for certain.

This includes events which have already happened.

- A woman takes a pregnancy test. By the time she takes the test, she's either pregnant or not. But if she applies subjective probability, she could use the test results to assign a measure of plausibility to the proposition that she is pregnant.
- In a criminal trial setting, the defendant either did what she's been accused of, or she didn't. But the jury must use the evidence to assign a measure of plausibility to the proposition that the defendant is guilty.

Some scientists find this *too* flexible, with too much room for opinion. However, if our goal is simply to make the best decisions possible given the available information, not to convince the scientific community that our decisions are correct, then we should be okay with using our background knowledge.

4.4 Exercises

1. Three fair coins are tossed.
   (a) List the outcomes in the universe, U, corresponding to this experience.
   (b) Identify the event sets corresponding to
   A: The first coin is a head.
   B: The second coin is a tail.
   C: There are at least two heads.
   (c) Calculate $P(A \cup B)$.
   (d) Calculate $P(A \cap B)$.
   (e) Calculate $P(\overline{A})$.
   (f) Calculate $P(\overline{C})$.
2. Two fair dice, one red and one green, are rolled. Let U be the set of all possible outcomes of this experience. Let A be the event corresponding to "the sum of the values is even", and let B be the event corresponding to "the sum of the values is divisible by 3".
   (a) List the outcomes in U.
   (b) List the outcomes in A and find $P(A)$.

(c) List the outcomes in B and find $P(B)$.

(d) List the outcomes in $A \cap B$ and find $P(A \cap B)$.

The next three problems are about a standard deck of playing cards, containing 52 cards, 13 each of 4 suits (clubs, diamonds, hearts, spades). Within each suit, the cards have "ranks" from 2 to 10, followed by Jack, Queen, King and Ace. In each case, you may assume that a simple random sampling procedure, without replacement, is used to draw from the deck.

3. Imagine randomly choosing five cards (without replacement) from the deck. What is the probability that you will have

   (a) No aces

   (b) At least one ace

   (c) Exactly one ace

4. Drawing four cards, what is the probability that you end up with one of each suit?

5. In poker, a "flush" is a five-card hand where all five cards have the same suit. A hand has "three of a kind" when any three cards have the same rank. "Four of a kind" is defined analogously.

   A flush beats a three of a kind, but is beaten by four of a kind. Demonstrate that this ranking of hands corresponds to how unlikely each hand type is to occur when drawing five cards at random (Hint: just count the number of different hands of each type, and then rely on the principle of exchangeability).

6. Prove ($P3$): For any events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. (Hint: Use set difference to write $A \cup B$ as a disjoint union of A and another set. Then, in a similar way, write A as a disjoint union of two other sets. Now apply (K3) and rearrange terms.)

7. Prove ($P4$): If $A \subset B$, then $P(B - A) = P(B) - P(A)$. (Hint: Write B as a disjoint union as you did for A in problem 6).

8. Prove ($P6$): For any set A, $0 \leq P(A) \leq P(U)$. (Hint: this can be done one or two lines if you use one of the other properties we proved in class)

9. Prove the following generalization of (K3) using induction on n: If $A_1, \ldots, A_n$ are "pairwise disjoint" events (that is, every pair is disjoint: for every pair of

integers $0 \leq i, j \leq n$ with $i \neq j$, $A_i \cap A_j = \emptyset$), then

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

10. Show that (P7) is a special case of the result in Exercise 9. (Note that for this problem you can assume the above result holds, even if you were not successful in proving it)

11. (*) One way to define a metric on the space of events is as the probability of their symmetric difference. That is, for any pair of events, A and B, define

$$d(A, B) = P(A \nabla B)$$

where, recall from the last chapter that we define $A \nabla B$ as $(A \cap \overline{B}) \cup (B \cap \overline{A})$.

Show that this distance function is a true metric; that is, show that for any events A, B and C,

(a) $d(A, A) = 0$

(b) $d(A, B) = d(B, A)$

(c) $d(A, C) \leq d(A, B) + d(B, C)$

Chapter 5

Conditional Probability

Consider the following scenario, which is similar to one we discussed in the first week:

I'm getting ready to leave the house in the morning, and I'm trying to decide whether or not to bring an umbrella. It's bulky, so I don't want to bring it if it's unlikely to rain, but I don't want to be caught in the rain with no protection. My decision will presumably depend on just what degree of plausibility I assign to the proposition that it's going to rain today (as well as the degree of discomfort I'll feel in each possible scenario).

Based on the fact that it's Tucson, I assign an initial probability of about 5% to rain.

But then I actually look out the window. It's not raining, but the sky is full of dark clouds. I reach into my bag of weak syllogisms, pull out number 5, and reason:

1. When it is about to rain, there is a greater tendency for the sky to be cloudy.
2. It is cloudy.
3. Therefore, it's more plausible that rain is coming.

Remember, we want our probability theory to parallel common sense reasoning, so we need a system that revises probabilities in the face of new evidence.

That is, we want to *update* our probability to reflect the *conditions* we are in. More formally, for some event A that we are interested in, and some event B that describes

our known conditions, we want to compute the **conditional probability** of A, *given* B.

5.1 Conditional Probability Motivation

Originally, we found ourselves in the universe, U, whose "outcomes" can be thought of as "possible worlds". We presumably are (or will be) in one of these, but we don't know which one, so we want to assign some degrees of plausibility to them.

In our example, each possible world has its own weather; one is cloudy and rainy, one is cloudy but not rainy, one is clear and sunny, etc.). We can represent the set of rainy worlds as the event A, the set of cloudy worlds as B, the set of cloudy and rainy worlds as $A \cup B$, and so on. We can then construct a probability measure, P, on events which are subsets of U, and we could assign some plausibility to the proposition that we are in a world in which it is raining, which we would write as $P(A)$.

However, once we learn that our world is cloudy (i.e., we observe the event B), none of the possible worlds that fail to be cloudy are plausible any more, and so we need to revise our probability measure to give these outcomes plausibilities of zero. In a sense, the universe has shrunk to include only cloudy worlds; it has become the set B.

In addition to being a valid probability measure (and hence obeying the Kolmogorov axioms), we want our revised probability measure on U to have a few intuitive properties.

First, it should assign a plausibility of 0 to $\overline{B}$, since by observing B we have logically ruled out its complement. Another property we want our revised probability measure to have is that for any events, say E_1 and E_2, which are *subsets* of B, the new measure should respect the previous relative plausibilities of these events, since neither one has had any possibilities ruled out.

Put another way, if E_1 was more plausible than E_2 *before* we observed B, it should continue to be more plausible afterwards; and if they were equally plausible before, they should remain equally plausible afterwards. In other words, when it comes to subsets of B, the value that our updated probability measure assigns to an event should be a **strictly increasing** function of the original probability assigned to that event.

Let's call the new probability measure P_B, to reflect our having observed B. We can already write down a few defining properties of P_B. The first one just enforces the Kolmogorov axioms, the second one reflects our having ruled out worlds not in B, and the last one reflects the strictly increasing property discussed above.

Definition 5.1 (Conditional Probability Measure)**.** *A* **conditional probability measure***, P_B, with respect to the universe U and the* **conditioning event** *B, is a function whose domain consists of events within U, and which satisfies*

1. *P_B is a probability measure, and so satisfies the three Kolmogorov axioms.*
2. *$P_B(\overline{B}) = 0$*
3. *If E_1 and E_2 are subsets of B, then*

$$P_B(E_2) > P_B(E_1), \qquad \textit{if and only if } P(E_2) > P(E_1)$$

These properties immediately imply the following proposition

Proposition 5.1. *For any event, E, $P_B(E) = P_B(E \cap B)$.*

Proof. We can write $E = (E \cap B) \cup (E \cap \overline{B})$, which is a disjoint union, so by $(K3)$, we have

$$P_B(E) = P_B(E \cap B) + P_B(E \cap \overline{B})$$

Note also that $E \cap \overline{B}$ is a subset of $\overline{B}$, and so by $(K1)$, monotonicity $(P5)$, and property 2 of a conditional probability measure, we have

$$0 \leq P_B(E \cap \overline{B}) \leq P(\overline{B}) = 0$$

which means that $P_B(E \cap \overline{B}) = 0$. Plugging this in above we get the result:

$$P_B(E) = P_B(E \cap B) + 0 = P_B(E \cap B)$$

□

The above proposition tells us that if we know what probabilities P_B assigns to all subsets of B, then we will have defined P_B on all events, since the conditional probability of any event is determined by the probability of its intersection with B, which is a subset of B.

We also get that

Proposition 5.2. $P_B(B) = 1$

This should be the case, since B is acting like the new "universe", and we'd like P_B to treat it as such.

Proof. Since P_B obeys the Kolmogorov axioms, it must also obey the complement rule, which follows from them. Therefore,

$$P_B(\overline{B}) = 1 - P_B(B)$$

But since $P_B(\overline{B}) = 0$ by definition of a conditional probability measure, we have

$$0 = 1 - P_B(B)$$

and so $P_B(B) = 1$ as claimed. □

Finally, property 3 says that for events which are subsets of B, $P_B(E)$ is a strictly increasing function of $P(E)$. That is, any time $P(E)$ increases, so does $P_B(E)$, *so long as* $E \subset B$. What that means is that, so long as we know that we are working with subsets of B, the only thing we need to know to compute the new probability, $P_B(E)$ is the old probability, $P(E)$. In other words, $P_B(E)$ is a function *only* of $P(E)$, and we can write $P_B(E) = u(P(E))$, where u is our strictly increasing update function. It turns out that u is a particularly simple sort of function; namely it multiplies probabilities by a constant. This follows from disjoint additivity.

Proposition 5.3. *Let P be a probability measure on a universe U and P_B be a conditional probability measure with respect to an event B. Let u be the update function mapping unconditional probabilities to conditional probabilities, such that for all $E \subset B$, $P_B(E) = u(P(E))$. Then*

$$u(p) = ap, \qquad \textit{where } a \textit{ is a constant that does not depend on } p.$$

Proof. Let E and F be disjoint subsets of B with $P(E) = p$ and $P(F) = q$. We have that $P_B(E) = u(P(E) = u(p)$ and $P_B(F) = u(P(F)) = u(q)$. Note that $E \cup F$ is also a subset of B, so that $P_B(E \cup F) = u(P(E \cup F))$. Since E and F are disjoint, we have $P(E \cup F) = p + q$. By $(K3)$, we have

$$u(p + q) = u(P(E \cup F)) = P_B(E \cup F) = P_B(E) + P_B(F) = u(p) + u(q)$$

That is to say, u is an **additive** function (an additive function has the property that $f(x + y) = f(x) + f(y)$). A function whose domain and range are real which

is monotonic (e.g. increasing) and additive is also linear, of the form $f(x) = ax$ for some constant a. The proof of this lemma is a straightforward application of elementary real analysis, but is beyond the scope of this course. □

Combining the above propositions, the following statements hold for any event $E \subset U$.

1. By Prop. 5.1, $P_B(E) = P_B(E \cap B)$.
2. By Prop. 5.2, $P_B(B) = 1$.
3. By Prop. 5.3, $P_B(E) = aP(E)$ whenever $E \subset B$. In particular $P_B(B) = aP(B)$.
4. Therefore $aP(B) = 1$, and so $a = \frac{1}{P(B)}$.
5. Since $E \cap B \subset B$, we have $P_B(E \cap B) = aP(E \cap B) = \frac{P(E \cap B)}{P(B)}$.
6. Putting this all together,

$$P_B(E) = P_B(E \cap B) = \frac{P(E \cap B)}{P(B)}$$

This means that a conditional probability measure is *unique*, and must be defined as follows.

Definition 5.2 (Conditional Probability Measure)**.** *The* unique *conditional probability measure P_B, with respect to the universe U and the conditioning set B, is defined on all events E in U as*

$$P_B(E) = \frac{P(E \cap B)}{P(B)}$$

In keeping with more standard notation for conditional probability, we will usually write $P(E|B)$ rather than $P_B(E)$, but it is important to keep in mind that for each fixed B, $P(\cdot|B)$ is a probability measure.

Technical Side Note We have shown that a conditional probability measure satisfying the properties laid out in definition 5.1 is unique. We have not actually shown that such a thing exists. In fact, it may not: in particular, if $P(B) = 0$, then our definition requires division by zero, which does not produce a well-defined function. When the universe can be partitioned into countably many atomic events, the union of all atomic events with individual probability zero also has zero probability, and

so we can restrict the universe to exclude this set without violating (K2). The only zero-probability event that remains is the empty set. Since it doesn't make sense to condition on the empty set ("given that the outcome is in the empty set" is a contradiction), we can assume that the conditioning set has positive probability.

When the universe cannot be partitioned into countably many atomic events, there are many cases when *every* atomic event has probability zero, and so restricting the universe by excluding these leaves us with nothing, which is not useful.

5.1.1 Alternative Derivations of Conditional Probability

We have given a formal derivation of conditional probability motivated only by "plausibility" considerations. Is this compatible with more standard notions of probability measures that are derived from the principle of exchangeability, or from limits of relative frequencies?

We can make sense of both of these questions by thinking of conditioning on an event B as restricting the universe U to the set B.

If the outcomes in U are exchangeable, then they should remain so upon restriction of the universe to B. If $U = \{a_1, \ldots, a_n\}$, then prior to conditioning on B, we have already seen that for any event, A, $P(A) = \frac{\#(A)}{\#(U)}$.

If the universe is B, then any event A restricted to the new universe B becomes $A \cap B$, and by the same exchangeability argument, we have

$$P_B(A) = P_B(A \cap B) = \frac{\#(A \cap B)}{\#(B)} = \frac{P(A \cap B)}{P(B)}$$

Similarly, if we take a relative frequency interpretation of probability, then prior to conditioning, the probability of the event A is the long-run ratio between the number of times the experience in question is realized with an outcome in A, out of all the times the experience occurs (that is, out of the number of times the experience takes an outcome in U). If we write $C(E)$ to be the count of how often the outcome is in E, then we have

$$P(A) = \frac{C(A)}{C(U)}$$

Conditioning on B simply restricts our attention to those times when the outcome of the experience is in B. The numerator of $P_B(A)$ then becomes the number of times

the outcome is in A out of this restricted set (that is, the number of times that the outcome is in both A and B; aka the number of times the outcome is in $A \cap B$) out of the total number of times that the outcome is in B. We then get

$$P_B(A) = \frac{C(A \cap B)}{C(B)} = \frac{C(A \cap B)}{C(U)} \times \frac{C(U)}{C(B)} = \frac{C(A \cap B)/C(U)}{C(B)/C(U)} = \frac{P(A \cap B)}{P(B)}$$

5.2 The Chain Rule

Notice that it follows directly from the definition of conditional probability that

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

The intuition behind this statement is very clear under a relative frequency interpretation: If I want to know the fraction of the time that the outcome of an experience is in both A and B, then I can first ask "what fraction of the time is the outcome in B?" (represented by $P(B)$), and then "out of those times, what fraction of the time is the outcome also in A?" (represented by $P(A|B)$). Multiplying these fractions together gives me the total fraction of the time that the outcome is in $A \cap B$.

We can apply this repeatedly (using induction and the associative property of intersections) to get

Definition 5.3 (The Chain Rule)**.** *For any number of arbitrary events $A_1, \ldots A_n$:*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots (A_n|A_1 \cap \cdots \cap A_{n-1})$$

The above is sometimes called the **chain rule**, because we are constructing a chain of conditional probabilities in lieu of one complicated joint probability. If we are interested in the probability of some complicated intersection, it may be easier to "decompose" the probability in this way (and perhaps avoid messy combinatorics).

Example 1. *Use the above type of decomposition to find the probability of selecting a batch of 3 defective items from a batch of 30 that includes a total of 5 defectives.*

Solution The sample space is all sequences of 3 draws. We can let $D_j, j = 1, 2, 3$ be the event that we select a defective item on the j^{th} draw. Then the ultimate event of interest is $D_1 \cap D_2 \cap D_3$.

We can decompose the probability as

$$P(D_1 \cap D_2 \cap D_3) = P(D_1) \times P(D_2|D_1) \times P(D_3|D_1 \cap D_2)$$

Now each of these is quite straightforward:

$$\begin{aligned} P(D_1) &= 5/30 \\ P(D_2|D_1) &= 4/29 \qquad \text{(4 remain out of 29 given that we already drew one)} \\ P(D_3|D_1 \cap D_2) &= 3/28 \qquad \text{(3 remain out of 28 given that we already drew 2)} \end{aligned}$$

So

$$P(D_1 \cap D_2 \cap D_3) = \frac{5}{30} \times \frac{4}{29} \times \frac{3}{28}$$

Example 2. *What's the probability of selecting at least 1 defective item in the first two draws?*

Solution Using the same notation as above, we're interested in $P(D_1 \cup D_2)$. It's easier to rewrite this as a disjoint union:

$$\begin{aligned} P(D_1 \cup D_2) &= P(D_1 \cup (D_2 - D_1)) \\ &= P(D_1) + P(D_2 \cap \overline{D_1}) \\ &= P(D_1) + P(\overline{D_1})P(D_2|\overline{D_1}) \\ &= \frac{5}{30} + (\frac{25}{30} \times \frac{5}{29}) \end{aligned}$$

(Note that we could also have computed this via the complement rule)

5.3 The Law of Total Probability

Recall from our discussion of probability measures that, if $\mathcal{E} = \{E_1, E_2, \ldots, E_n\}$ is a **partition** of U, then

$$1 = P(U) = P(E_1 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i) \tag{5.1}$$

We can construct partitions of other sets, too. For example, with $\mathcal{E}$ defined as above, we get a partition of the set A by taking

$$\mathcal{E} \cap A = \{E_1 \cap A, E_2 \cap A, \ldots, E_n \cap A\}$$

We will show that this works when $n = 2$. The inductive proof that this is a partition for arbitrary values of n is left as a homework problem.

Proof of case $n = 2$. Suppose $\{E_1, E_2\}$ is a partition of U, so that $E_1 \cap E_2 = \emptyset$ and $E_1 \cup E_2 = U$ (you should check for yourself that this means that $E_2 = \overline{E_1}$). We want to show that $\{E_1 \cap A\}$ and $\{E_2 \cap A\}$ together form a partition of A. We need to check the two properties of a partition: first that the sets are disjoint, and second that their union is A.

First, note that we have

$$\begin{aligned}(E_1 \cap A) \cap (E_2 \cap A) &= (E_1 \cap E_2) \cap (A \cap A) \\ &= \emptyset \cap A = \emptyset\end{aligned}$$

Next, we get

$$\begin{aligned}(E_1 \cap A) \cup (E_2 \cap A) &= (E_1 \cup E_2) \cap A \\ &= U \cap A = A\end{aligned}$$

where in the first line we have applied the distributive property in reverse (we have essentially "factored out" an A).

Since we have verified both conditions of being a partition, we have proved the claim. □

Then, by the same disjoint additivity argument we used in (9.2.3), we have

$$P(A) = P((E_1 \cap A) \cup \cdots \cup (E_n \cap A) = \sum_{i=1}^{n} P(E_i \cap A) \tag{5.2}$$

In the particular case where the partition consists of E and $\overline{E}$, (9.13) simplifies to

$$P(A) = P(A \cap E) + P(A \cap \overline{E}) \tag{5.3}$$

Combining the above with the chain rule applied to each term in the sum, we get the Law of Total Probability.

Theorem 5.1 (Law of Total Probability)**.** *Let A be an arbitrary event and let $\{E_1, \ldots, E_n\}$ be a partition of U. Then*

$$P(A) = \sum_{i=1}^{n} P(E_i)P(A|E_i)$$

In the special case where the partition consists of E and $\overline{E}$, (9.19) simplifies to

$$P(A) = P(E)P(A|E) + P(\overline{E})P(A|\overline{E}) \tag{5.4}$$

In many cases, the probability of an event A depends on some hidden state of the world. If we knew the hidden state, we could compute the probability in each case. If we have some beliefs about how likely each state is, then the Law of Total Probability tells us that we can compute the **marginal** probability of A by taking a weighted average of conditional probabilities, where we give a higher weight to states that we believe are more likely.

Example 3. *A baseball game is tied 2-2 going into the bottom of the ninth inning. The probability of scoring a run in the inning changes depending on what the first batter does. Let R be the event corresponding to scoring a run in the inning, and let B_0, B_1, B_2, B_3 and B_4 be events corresponding to the first batter (0) making an out, (1) getting on first, (2) getting on second, (3) getting on third, or (4) hitting a home run. These form a partition of the universe. Suppose we have the following probabilities:*

i	$P(B_i)$	$P(R\|B_i)$
0	0.680	0.167
1	0.240	0.400
2	0.050	0.615
3	0.005	0.850
4	0.025	1.000

What's the overall (marginal) probability of scoring a run in the inning?

Solution If we know that a team has a 16.7% chance of scoring given that the first batter makes an out, a 40% chance of scoring given that the first player reaches first base, etc., then intuitively the overall chance of scoring is an average of these. But it's not a simple average: an out is much more likely than, say, a home run, and so

the scoring probability conditioned on an out will influence the overall probability more. Using the law of total probability, we have

$$\begin{aligned} P(R) &= \sum_{i=0}^{4} P(B_i)P(R|B_i) \\ &= 0.680 \cdot 0.167 + 0.240 \cdot 0.400 + 0.050 \cdot 0.615 + 0.005 \cdot 0.850 + 0.025 \cdot 1.000 \\ &\approx 0.270 \end{aligned}$$

5.4 Bayes' Theorem

5.4.1 Motivation

Many quantities are difficult to observe directly (e.g., presence of a disease, pregnancy, guilt or innocence in a trial, truth or falsity of a scientific hypothesis, the nature and locations of physical objects) and must be inferred using tests (e.g. a medical test, data collected in an experiment) or other forms of evidence (e.g. testimony in a trial).

However, most of the time, probability models are framed in the opposite direction: we may have a reasonable model for the probabilities of different kinds of evidence, *conditioned on* a hypothesis being true (e.g., the probability of a positive test result *conditioned* on actually being pregnant; the probability that the burglar alarm is set off *conditioned on* no burglary taking place). But we have the evidence, and we want to assess a probability of our hypothesis being correct.

Unfortunately there is no unique, objective way to do this; however, if we are willing to assign subjective degrees of **prior plausibility** (a.k.a. **prior probabilities**) to each hypothesis (prior to looking at the evidence, that is), then probability theory tells us how to *revise* our initial beliefs in the face of evidence.

5.4.2 Derivation

Binary Hypotheses Suppose we have a hypothesis about the world (e.g., a burglary took place). We can represent this as a set H, and assign it an initial (**prior**)

probability, $P(H)$. Then, by the rules of probability, $P(\overline{H}) = 1 - P(H)$. Now, suppose we observe some **evidence** (e.g., the buglar alarm went off). Call this event E.

Suppose also that we have a **likelihood model**, which gives us $P(E|H)$, as well as $P(E|\overline{H})$. How do we use this to *update* the plausibility of H, starting with the **prior plausibility**, $P(H)$, and ending up with the **posterior plausibility**, $P(H|E)$?

If we had access to $P(H \cap E)$ and $P(E)$, we could just divide, using the definition of conditional probability.

$$P(H|E) = \frac{P(H \cap E)}{P(E)} \tag{5.5}$$

Unfortunately, we don't know these directly. However, we can compute $P(H \cap E)$ using the chain rule:

$$P(H \cap E) = P(H) \times P(E|H) \tag{5.6}$$

Substituting (5.6) into (5.5), we get

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \tag{5.7}$$

Now, we can apply the Law of Total Probability (Theorem 5.1) to the denominator to get

$$P(H|E) = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(\overline{H})P(E|\overline{H})} \tag{5.8}$$

which is now expressed entirely in terms of quantities that we have access to.

Multiple Hypotheses We may of course have more than one hypothesis of interest. In general, we might *partition* the space of possibilities into a set of n disjoint and exhaustive hypotheses, $H_1, H_2, \ldots, H_n$. Provided again that we can assign initial plausibilities $P(H_1), P(H_2), \ldots, P(H_n)$ (which must sum to 1 by property (P7)), as well as likelihood models associated with each hypothesis, which give us $P(E|H_1), P(E|H_2), \ldots, P(E|H_n)$ (which, note, need *not* sum to 1, as they are values drawn from different conditional probability measures), then following the same logic, we get for any $j \in \{1, 2, \ldots, n\}$:

$$P(H_j|E) = \frac{P(H_j)P(E|H_j)}{\sum_{i=1}^{n} P(H_i)P(E|H_j)} \tag{5.9}$$

Equations (9.20) and (9.32) are different forms of **Bayes' Theorem**:

Theorem 5.2 (Bayes Theorem (Two Versions))**.**
(a) Let E and H be events with $P(E), P(H) \neq 0$. Then

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

(b) Let E be an event in U, and let $\{H_1, H_2, \ldots, H_n\}$ be a partition of U with each $P(H_i) \neq 0$. Then for each $j \in \{1, 2, \ldots, n\}$

$$P(H_j|E) = \frac{P(H_j)P(E|H_j)}{\sum_{i=1}^{n} P(H_i)P(E|H_i)}$$

We derived Bayes Theorem using the definitions of conditional probability, and so $P(\cdot|E)$ must behave like a probability measure.

In particular, since $\{H_1, H_2, \ldots, H_n\}$ is a partition, we must have

$$\sum_{j=1}^{n} P(H_j|E) = 1 \tag{5.10}$$

This is easily verified from statement (b) of Bayes' Theorem:

$$\begin{aligned}\sum_{j=1}^{n} P(H_j|E) &= \sum_{j=1}^{n} \frac{P(H_j)P(E|H_j)}{\sum_{i=1}^{n} P(H_i)P(E|H_i)} \\ &= \frac{\sum_{j=1}^{n} P(H_j)P(E|H_j)}{\sum_{i=1}^{n} P(H_i)P(E|H_i)} \\ &= 1\end{aligned}$$

5.4.3 Example: Detectors

(The following is a generalization of the "binary symmetric channel" example from Ch. 4 in Applebaum)

Suppose we are interested in whether the environment is in some particular state H, or not (if not, it's in state $\overline{H}$). We have a detector that we can query, that tends to respond D when H is present, and $\overline{D}$ otherwise, though it is imperfect.

This scenario describes the pregnancy test, or any other medical test, a lie detector, a burglar alarm, etc.

The properties of the detector can be described by two numbers:

(1) its **sensitivity**, p (the conditional probability of a "present" result, given a true presence, also called the **true positive rate**), and

(2) its **specificity**, q (the conditional probability of an "absent" result, given a true absence, also called the **true negative rate**)

Ideally both p and q are close to 1. (Equivalently, we could be given the **false positive** and **false negative** rates, $1-q$ and $1-p$, respectively.

What's the probability that state H is present given that the detector says it is?

To apply Bayes theorem, we need one more piece of information: namely, the **prior probability** of H. Call this probability r (in the case of medical tests, etc., r is called the **prevalence**, which is how common the condition is in the relevant population).

Then

$$P(H|D) = \frac{P(H)P(D|H)}{P(H)P(D|H) + P(\overline{H})P(D|\overline{H})} = \frac{rp}{rp + (1-r)(1-q)}$$

When is this more than $1/2$? That is, when is it more likely than not that H is present?

Clearly we need $rp > (1-r)(1-q)$, which occurs if

$$\begin{aligned} rp &> 1 - r - q + rq \\ r(p+1-q) &> 1-q \\ r &> \frac{1-q}{p+1-q} \\ \text{Prevalence} &> \frac{\text{False Positive Rate}}{\text{Sensitivity} + \text{False Positive Rate}} \end{aligned}$$

In the special case where our detector is **symmetric** — that is, where $p = q$, H is more likely than not to be present given D iff

$$r > 1 - q$$

That is, the prevalence must be above the false positive (and false negative) rates.

5.4.4 The Update Factor

Notice that when we make an observation E (i.e., we restrict the set of possibilities to the set E), the plausibility of any other proposition H_j potentially changes. It goes from the prior plausibility $P(H_j)$ before the observation, to the posterior plausibility $P(H_j|E)$, afterwards. This *update* involves multiplying the prior by an update factor

$$\frac{P(E|H_j)}{P(E)}$$

(To see this, write the first version of Bayes' Theorem, and just bring the prior plausibility, $P(H)$ out of the fraction.)

If $\{H_1, H_2, \ldots, H_n\}$ is a partition, the update factor is

$$\frac{P(E|H_j)}{\sum_{i=1}^{n} P(H_i)P(E|H_i)}$$

As we noted in our discussion of the Law of Total Probability, the denominator can be regarded as a weighted average of the $P(E|H_i)$s, where the weights are the prior plausibilities.

In the simplest case where there are only two possible hypotheses, the update factor is

$$\frac{P(E|H)}{P(H)P(E|H) + P(\overline{H})P(E|\overline{H})}$$

Question: What must be true for an observation E to *increase* the plausibility of H?

Answer: The plausibility of H will go up iff $P(E|H)$ is greater than the weighted average (marginal) probability $P(E)$. This happens iff $P(E|H)$ is greater than $P(E|\overline{H})$ — that is, if H being true makes E *more likely* than it was if H were not true. Notice that this holds *regardless of the prior plausbility of* H. Bayes' Theorem is a precise expression of weak syllogism 5!

(Premise) If H is true then E becomes more plausible [than it would be otherwise]

(Data) E is true

(Inference) H becomes more plausible

5.5 Independence

Bayesian inference gives us a way to update the plausibility of any proposition after observing any another. In many cases (e.g., the burglar alarm, the pregnancy test, etc.), this change in plausibility can be large. Clearly, however, many pairs of propositions are unrelated. For example, observing that it's cloudy outside presumably has no bearing on whether a defendant is guilty.

5.5.1 Necessary and Sufficient Conditions

Definition 5.4 (Independence)**.** *We say that an event A is **independent** of another event B if the occurrence (observation) of B has no effect on the plausibility of A. That is to say, A is independent of B if and only if*

$$P(A|B) = P(A) \tag{5.11}$$

In Bayesian terms, for A to be independent of B, the prior plausibility of A ($P(A)$) and the posterior plausibility of A given B (i.e., $P(A|B)$) must be equal, which occurs iff the update factor is 1. Assuming neither $P(A)$ nor $P(B)$ is 0,

$$P(A|B) = P(A) \iff \frac{P(B|A)}{P(B)} = 1 \iff P(B|A) = P(B)$$

That is to say, A is independent of B iff B is independent of A. Independence is a symmetric (or commutative) relation, and we can simply say that "A and B are independent".

We can give a symmetric criterion for independence which is equivalent to (5.11) and avoids having to worry about potential division by zero:

$$\begin{aligned} & P(B|A) = P(B) \\ \iff & \frac{P(A \cap B)}{P(A)} = P(B) \\ \iff & P(A \cap B) = P(A)P(B) \end{aligned}$$

Definition 5.5 (Independence)**.** *A and B are independent if*

$$P(A \cap B) = P(A)P(B) \tag{5.12}$$

It is easy to compute the probability of the union of independent events, by plugging (9.36) into (P3):

If A and B are independent, then $P(A \cup B) = P(A) + P(B) - P(A)P(B)$ (P7*)

Intuitively, if the plausibility of B doesn't change upon observing A, then the plausibility of $\overline{B}$ (which is just $1 - P(B)$) will not change either.

Theorem 5.3. *If $P(A) \neq 0$ and A and B are independent, then so are A and $\overline{B}$.*

Proof.

$$\begin{aligned} P(A \cap \overline{B}) &= P(A)P(\overline{B}|A) && \text{(by the chain rule)} \\ &= P(A)(1 - P(B|A)) && \text{(by the complement rule)} \\ &= P(A)(1 - P(B)) && \text{(by independence)} \\ &= P(A)P(\overline{B}) && \text{(by the complement rule)} \end{aligned}$$

□

5.6 Independence of More than Two Events

We say that n events, $A_1, A_2, \ldots, A_n$ are mutually independent if knowledge about any subset of them does not change the plausibility of any non-overlapping subset.

What are the quantitative conditions for this mutual independence? It is tempting to suppose that the n events are independent if each one is independent of any other (this works for mutual disjointness, for example). However, it turns out that this condition (called **pairwise independence**) is insufficient.

For example, let U be the outcomes of two coin flips ($U = \{HH, HT, TH, TT\}$. Let A, be the event that the first coin comes up heads ($A = \{HH, HT\}$), let B be the event that the second coin comes up heads ($B = \{HH, TH\}$), and let C be the event where the two flips have the same outcome ($C = \{HH, TT\}$. Then, applying the principle of symmetry:

$$P(A) = P(B) = P(C) = 1/2$$

and

$$P(A \cap B) = P(\{HH\}) = \frac{1}{4} = P(A)P(B)$$
$$P(A \cap C) = P(\{HH\}) = \frac{1}{4} = P(A)P(C)$$
$$P(B \cap C) = P(\{HH\}) = \frac{1}{4} = P(B)P(C)$$

and so A, B and C are pairwise independent, but

$$P(C|A \cap B) = \frac{P(A \cap B \cap C)}{P(A \cap B)} = \frac{P(\{HH\})}{P(\{HH\})} = 1 \neq P(C)$$

See Exercise 4.20 in Applebaum for another example like this.

It is also possible to construct examples where the probability of n events occurring together is the product of their marginal probabilities, but where the events are not pairwise independent. For example, suppose that A, B and C are events such that $P(A) = 1/2$, $P(B|A) = \frac{1}{4}$, $P(B|\overline{A}) = \frac{1}{4}$, $P(C|A \cap B) = P(C|B) = 1$, and $P(C|\overline{B}) = 0$. To find the marginal probability of B we can use the law of total probability:

$$\begin{aligned} P(B) &= P(A)P(B|A) + P(\overline{A})P(B|\overline{A}) \\ &= \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{3}{4} \\ &= \frac{1}{2} \end{aligned}$$

Similarly, $P(C)$ is

$$\begin{aligned} P(C) &= P(B)P(C|B) + P(\overline{B})P(C|\overline{B}) \\ &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 \\ &= \frac{1}{2} \end{aligned}$$

We can find the joint probability of A, B and C using the chain rule:

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$$
$$\frac{1}{2} \times \frac{1}{4} \times 1 = \frac{1}{8}$$

and so we have $P(A \cap B \cap C) = P(A)P(B)P(C)$. But no pair of events is independent (most obviously, C depends deterministically on B).

Unfortunately, independence of n events can't be simplified much.

Definition 5.6 (Mutual Independence)**.** *We say that $A_1, A_2, \ldots, A_n$ are* **mutually independent** *iff for any choice of r events, $A_{i_1}, A_{i_2}, \ldots, A_{i_r}$ (for example, we could have $A_{i_1} = A_2$, $A_{i_2} = A_4$, $A_{i_3} = A_9$, etc.), we have*

$$P(A_{i_1} \cap A_{i_2} \cdots \cap A_{i_r}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_r}) \tag{5.13}$$

5.7 Exercises

1. Recall that a set of odds is called *coherent* if the associated probabilities obey the Kolmogorov axioms. Suppose three friends, A, B and C, plan to run a race to see who is fastest. A fourth friend decides to take bets on the outcome of the race. She offers odds of 3:1 against A (i.e., if someone bets $\$M$ on A, she will pay $\$3 \cdot M$ if A wins and collect $\$M$ if A loses), odds of 7:1 against B, and odds of 1:1 against C. Are these odds coherent? If so, prove it. If not, find a combination of bets you could place that would guarantee a profit for yourself no matter the outcome. Assume that you can bet on any runner at the stated odds, or against any runner at the inverse of the stated odds. The stakes need not be the same across bets.

2. Find the following probabilities.

   (a) Suppose 75% of women live past age 65 and 55% of women live past age 80. What is the probability that a woman who is now 65 will live past 80?

   (b) What's the probability that a roll of two fair six-sided dice will sum to 6? What's the probability of this once you observe that one of the dice came up 2?

   (c) A fair six-sided dice is tossed twice. The second toss is higher than the first. What's the probability that the first toss was at least 4?

3. The "multiplication rule" says that if A and B are events in a universe U equipped with a probability measure P, then the joint probability $P(A \cap B)$ can be written as $P(B)P(A|B)$. Using this, prove the chain rule by induction. Recall that the chain rule says that if $A_1, \ldots, A_n$ are events in U then we can

decompose their joint probability as a "chain of conditional probabilities":

$$P(A_1 \cap \ldots, \cap A_n) = P(A_1)P(A_2|A_1)\ldots P(A_n|A_1 \cap \cdots \cap A_{n-1})$$

4. A bowl contains seven red balls and five blue ones. Suppose three balls are drawn without replacement from the bowl. Use the chain rule to find

   (a) the probability that all three balls are red.

   (b) the probability that all three balls are blue.

   (c) the probability that the sequence is "red, blue, red".

5. Let H and D be events in a universe U.

   (a) Show that the unconditional plausibility of H is a weighted average of $P(H|D)$ and $P(H|\overline{D})$. In particular, show that

   $$P(H) = P(H|D)P(D) + P(H|\overline{D})P(\overline{D})$$

   (b) Explain why this means that observing D can only make the plausibility of H go up if observing $\overline{D}$ would make it go down.

In the game of blackjack, players are initially dealt two cards each, one that only they can see; the other that everyone can see. They can continue to ask for cards until the combined value of their cards exceeds 21 (the number cards, 2 through 10, are worth face value; face cards, Jack, Queen and King, are worth 10 each; and Aces can count as either 1 or 11). The goal is to get a score as close as possible to 21 without going over (a.k.a. going "bust"). If a player's first two cards have a combined value of 21, this is called "blackjack" (whether or not there is a black jack involved!)

6. Use the law of total probability to calculate the probability of getting "blackjack" (you can condition on different possibilities for the first card, for instance).

7. Suppose four cards from the deck were used up in the previous hand, and the new hand will be dealt from the cards that remain.

   (a) Suppose none of the four cards were aces, face cards or tens. What is the conditional probability of getting blackjack now?

   (b) Suppose all four cards were face cards or tens. What is the conditional probability of getting blackjack in this situation?

8. Recompute the probabilities in exercises 6 and 7, assuming that the initial deck contains four copies of each card (that is, it is four decks shuffled together). How has the use of additional decks changed the value of "counting cards" (i.e., keeping track of what cards have been used up) for predicting whether blackjack will occur?

9. Let $U = \{a_1, \ldots, a_n\}$ be a finite universe, and let $f : U \to [0, \infty)$ be an arbitrary function that assigns nonnegative real numbers (possibly greater than 1) to members of the universe. We will show that, for a fixed event B, we obtain a valid probability measure, P_B by defining, for any A,

$$P_B(A) = \frac{\sum_{a_i \in A \cap B} f(a_i)}{\sum_{a_j \in B} f(a_j)}$$

(where, by choosing $B = U$, we can get an "unconditional" probability, $P_B(A) = P(A)$). This requires three parts:

   (a) Show that P_B obeys (K1)

   (b) Show that P_B obeys (K2)

   (c) Show that P_B obeys (K3)

10. Let A and B be disjoint events, each with non-zero probability, and C be an arbitrary event. Define $\alpha = \frac{P(A)}{P(A)+P(B)}$.

   (a) Show that $P(C|A \cup B) = \alpha P(C|A) + (1 - \alpha)P(C|B)$.

   (b) Define a function $Q_C(\cdot)$ on events by setting $Q_C(A) = P(C|A)$. Use the result of part (a) to show that $Q_C(\cdot)$ is not a probability measure. (Hint: show that it does not obey (K3))

11. In each of the following scenarios there are two events, A and B. For each scenario, determine whether A and B are independent, and calculate $P(A \cup B)$.

   (a) $P(A) = 0.4$, $P(B) = 0.5$ and $P(A \cap B) = 0.20$.

   (b) $P(A) = 0.4$, $P(B) = 0.4$ and $P(\overline{A} \cap B) = 0.24$.

   (c) $P(A) = 0.7$, $P(B) = 0.8$ and $P(\overline{A} \cap \overline{B}) = 0.1$.

12. There are two dice, one of which is fair, and the other of which is loaded so that $P(1) = P(2) = P(3) = P(4) = 1/5$ and $P(5) = P(6) = 1/10$. Unfortunately you have lost track of which is which.

(a) You pick one die at random, and roll it. It comes up 6. What's the probability that it's the loaded die?

(b) You roll your chosen die again and it comes up 2. What's the probability that it's loaded, in light of all the evidence so far?

13. Bayes rule has an iterative property. Consider assessing the plausibility of a hypothesis, H. After a first piece of evidence, E_1 comes in, we update the plausibility of H from the prior plausibility, $P(H)$ to the posterior plausibility, $P(E|H)$. If a second piece of evidence, E_2, comes in, we can could start afresh, using $P(H)$ as the prior, and treating $F = E_1 \cap E_2$ as the evidence, applying Bayes' rule in the usual way:

$$P(H|F) = \frac{P(H)P(F|H)}{P(F)}$$

Alternatively, we can let the old posterior, $P(H|E_1)$, serve as the new prior, and compute the new posterior as

$$P(H|E_1 \cap E_2) = \frac{P(H|E_1)P(E_2|H \cap E_1)}{P(E_2|E_1)}$$

(a) Show that these two formulations give the same answer.

(b) Show that if, in addition, E_2 is *conditionally independent* of E_1 given H, meaning that

$$P(E_1 \cap E_2|H) = P(E_1|H)P(E_2|H)$$
$$P(E_1 \cap E_2|\overline{H}) = P(E_1|\overline{H})P(E_2|\overline{H})$$

then the *posterior odds* of H simplifies to

$$\frac{P(H|E_1 \cap E_2)}{P(\overline{H}|E_1 \cap E_2)} = \frac{P(H|E_1)}{P(\overline{H}|E_1)} \cdot \frac{P(E_2|H)}{P(E_2|\overline{H})}$$

14. (**The Three-Card Problem**) I have three double-sided cards. One is red on both sides; one is green on both sides; and one is red on one side and green on the other. I shuffle the cards behind my back, randomly select one, and then randomly select a side to show you. It is red. What's the probability that the *other* side of that card is also red?

15. (**The Monty Hall Problem**) In the gameshow *Let's Make a Deal*, a contestant has the opportunity to win a car by correctly guessing which of three doors it is behind (the other two doors contain gag prizes). Once the contestant selects her door, but before the door is opened, the host (Monty Hall, in the original version of the show) always opens one of the *other* two doors (not the one the contestant picked) to reveal a gag prize (this is always possible, since even if there is a gag prize behind the contestant's door, there will always be another gag prize behind one of the other doors). After this occurs, the contestant has the opportunity to switch her choice to the *other* remaining door. Whether this is a good thing to do depends on the conditional probability that the car is behind her original door, given the new evidence: if it is less than 0.5, she should switch; if it is greater she should stay; and if it is exactly 0.5 it doesn't matter what she does. Compute the optimal strategy by finding the conditional probability in question.

Chapter 6

Discrete Random Variables

For the rest of the chapter unless otherwise noted, we will let U denote a universe with finitely or countably infinitely many elements, and P denote a probability measure defined on all events in U. The domain of P is the **power set** of U, that is, the set of all possible subsets (including the empty set as well as U itself). The power set of U is denoted $\mathcal{P}(U)$.

Together, the triple $(U, \mathcal{P}(U), P)$ are called a **probability space**.

6.1 Random Variables

Most universes in real problems are large.

- $U = \{\text{All lists of 16 Birthdays}\}$ $\quad \#(U) = 365^{16}$
- $U = \{\text{All subsets of 10 items drawn from a lot of 30}\}$ $\quad \#(U) = \binom{30}{10}$

If the universe itself is large, the power set of the universe is extraordinarily huge: We have $\#(\mathcal{P}(U)) = 2^{\#(U)}$ (to count the number of subsets, consider choosing for every single element whether it gets included or excluded from the subset).

Typically we aren't interested in every possible event; but only those that pick out outcomes with some common "feature":

- $E =$ Lists with a repeated birthday
- $E =$ Subsets with no defective items

In principle, we could calculate the probability of these events by breaking them down into smaller disjoint sets whose probabilities we can compute. But it is simpler to introduce **random variables**, which let us work with these features directly.

Definition 6.1 (Random Variable)**.** *A **random variable**, X, is a* function*, whose domain is a universe, U, and whose range is a set R which is a subset of the real numbers $\mathbb{R}$. That is, for every $a \in U$, $X(a) \in R \subset \mathbb{R}$.*

Example If U is the set of all subsets of 3 items drawn from a lot of 30, we can define a random variable X which returns the number of defective items.

The range of X would then be the set $\{0, 1, 2, 3\}$.

Definition 6.2 (Discrete Random Variable)**.** *A random variable is called **discrete** if its range is either a finite set or a countably infinite set (e.g. $\mathbb{N}$, $\mathbb{Z}$, etc.).*

The random variable X defined above (counting defective items) is discrete because its range is a finite set.

In contrast, suppose U is the set of all possible Tucson days. We can define a random variable Y which, for each day gives its temperature. But since temperatures are real numbers, the range of Y is uncountably infinite, and so Y does not qualify as a discrete random variable.

We are often interested in events of this form: $E = \{a \in U : X(a) \in A\}$. In words, these are events that occur whenever X takes on one of a particular set of values.

- As a shorthand, we can write $E = \{X \in A\}$.
- If A is a singleton set (i.e., $A = \{x\}$ for some $x \in R$), then we can write E as $\{X = a\}$.
- If A is an interval, for example, $A = [a, b]$, we can write $E = \{a \leq X \leq x\}$. If the interval is unbounded on one end, such as $A = (-\infty, b]$, then we have $E = \{X \leq b\}$.

The probability of such events follows from the probability measure we have defined on events in U:

- $P(X \in A) = P(\{a \in U; X(a) \in A\})$
- $P(X = x) = P(\{a \in U; X(a) = x\})$

- $P(X \leq x) = P(\{a \in U; X(a) \leq x\})$

As such, everything we have established so far about the probability of events, their complements, unions, etc., still applies. All we are doing is providing a bit more structure to the universe (which, after all, is what all mathematics is about, isn't it?).

Example 4. *Two fair dice are thrown. What is the probability that the sum of the numbers appearing on the two upturned faces equals 7?*

Solution Note that the sample space consists of all possible ordered pairs of the numbers 1 through 6:

$$\begin{aligned} U &= \{(1,1),(1,2),\ldots,(6,5),(6,6)\} \\ \#(U) &= 6 \times 6 = 36 \qquad \text{(by the generalized counting theorem)} \end{aligned}$$

We can define a random variable X whose value is the sum of the scores of the two dice. In terms of X, the event of interest is

$$\begin{aligned} E &= \{X = 7\} = \{a \in U; X(a) = 7\} \\ &= \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \\ \implies \#(E) &= 6 \end{aligned}$$

It seems reasonable to assume that the ordered pairs are exchangeable in this case, as we would not expect the probabilities to change if we relabeled the dice, so we can compute

$$P(E) = \frac{\#(E)}{\#(U)} = \frac{6}{36} = \frac{1}{6}$$

If we repeat this process for other possible values of X, we will find that $P(X = x) = 0$ unless $x \in \{2,3,4,\ldots,11,12\}$, and that for these values, we can express the probability that $X = x$ with the following algebraic expression:

$$P(X = x) = \frac{6 - |x - 7|}{36}$$

Note that by restricting our attention to the sum of the dice, we have reduced the number of fundamental events from 36 to 11.

6.2 The PMF and CDF

Notice that a random variable X defines a **partition** of the universe, whose 'compartments' are the events $\{X = x\}$ for each x in the range of X. For our dice example, the events $\{X = 2\}$ through $\{X = 12\}$ partition all the possible outcomes of the two dice.

6.2.1 The Probability Mass Function of a Discrete Random Variable

Definition 6.3 (Probability Mass Function)**.** *Let X be a discrete random variable with range R. We define the **probability mass function**, f, associated with X to be a function that takes in a valid value for X, and returns the probability that X takes that value. Formally, the domain of f is the set R (the range of X), and the range of f is the interval $[0, 1]$, and $f(x)$ is defined for each $x \in R$ to be*

$$f(x) = P(X = x)$$

Theorem 6.1. *Let U be a universe, P a probability measure defined on events in U, let X be a random variable on U, and let f be the PMF of X as determined by the probability measure P.*
Then, treating R itself as a universe, there is a unique *probability measure, P', defined on* all *subsets of R, such that $P'(\{x\}) = f(x)$ for each $x \in R$. For any set $S \subset R$, we have $P'(S) = \sum_{x \in S} f(x)$.*

In other words, if we know $f(x)$ for every $x \in R$, then there is exactly one way to extend these values to a probability measure which is defined on all subsets of R. The way to do this is to take any subset S of R and set its probability to be the sum of the atomic probabilities of the elements of S.

One reason this matters is that it says that if we know the PMF of a discrete random variable, X, then we know everything that X is capable of telling us about.

Proof. There are two parts to this claim; an existence claim and a uniqueness claim. We need to check each part: first that the probability measure defined as stated obeys the Kolmogorov axioms, and second that no other probability measure on $\mathcal{P}(R)$ agrees with P' on all elements of R.

Existence. The first Kolmogorov axiom says that all probabilities are nonnegative. Since $f(x)$ is a probability, we know that $P'(\{x\})$ is nonnegative for every $x \in R$. Since the value of P' on any other set is obtained by summing these nonnegative numbers, P' is always nonnegative.

Next, we need to check that $P'(R) = 1$. By definition, we have

$$P'(R) = \sum_{x \in R} f(x) = \sum_{x \in R} P(X = x)$$

Since the sets $\{X = x\}$ form a partition of the universe, property (P7) of probability measures implies that $\sum_{x \in r} P(X = x) = 1$.

Finally we need to check disjoint additivity. Suppose S and T are disjoint subsets of R. Then

$$P'(S \cup T) = \sum_{x \in (S \cup T)} f(x) = \sum_{x \in S} f(x) + \sum_{x \in T} f(x) = P'(S) + P'(T)$$

where the second equality follows from the fact that S and T are disjoint, and therefore do not share any elements whose probabilities might be counted twice if we split up the sum.

This completes the part of the proof that says that P' as defined is a probability measure, since we have checked all of the Kolmogorov axioms.

Uniqueness. It remains only to check that no other probability measure exists (say, $\tilde{P}$) that satisfies $\tilde{P}(\{x\}) = f(x)$. The key here is disjoint additivity. Suppose we define $\tilde{P}(\{x\}) = f(x)$ on every $x \in R$. Then for any set $S = \{x_1, \ldots, x_M\}$, which is a subset of R, we can write

$$S = \{x_1\} \cup \{x_2\} \cup \cdots \cup \{x_M\}$$

which is a disjoint union. Hence by the inductive extension of the disjoint additivity axiom, we must have

$$\tilde{P}(S) = \sum_{i=1}^{M} \tilde{P}(\{x_i\}) = \sum_{i=1}^{M} f(x_i) = P'(S)$$

and so $\tilde{P}$ is indistinguishable from P'. □

Note the following important fact about probability mass functions which has come out in the course of the above proof:

$$\sum_{x \in R} f(x) = 1 \tag{6.1}$$

In the case where the range of the random variable is a finite set, we can visualize the probability mass function p with a probability histogram (or "spike plot"), with a bar centered at each x_j, whose height is $p(x_j)$. If we again consider the random variable whose value is equal to the sum of the faces of two six-sided dice, then the probability histogram looks like Fig. 6.1

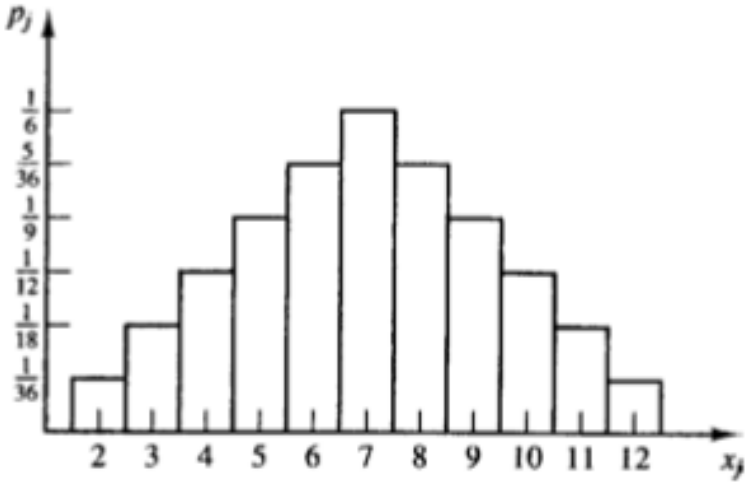

Figure 6.1: A probability histogram for the random variable representing the sum of two fair die rolls

6.2.2 The Cumulative Distribution Function

We often are interested in events of the form $\{X \leq x\}$ for some $x \in R$. For example, we might ask "What's the probability that there is at most one defective item in our draw?"

We can record probabilities of all events of this form using the **cumulative distribution function**.

Definition 6.4 (Cumulative Distribution Function)**.** *Let X be a discrete random variable with range R.*
*The **cumulative distribution function** for X is a function, F, from R to $[0,1]$, and defined for each $x \in R$ as*

$$F(x) = P(X \leq x)$$

Note that if R has a smallest element, x_1, so that we can list the elements of R in increasing order as $x_1, x_2, \ldots$, then

$$F(x_r) = \sum_{j=1}^{r} f(x_j)$$

where f is the probability mass function associated with X. This follows from the fact that the event $\{X \leq x_r\}$ is a disjoint union of the events $\{X = x_j\}$ for all j ranging from 1 to r, and so the probability of the union is the sum of the probabilities of these events, which are the $f(x)$s.

Theorem 6.2 (Basic Properties of the CDF)**.** *Let F be the CDF of a random variable X, with range R indexed by integers, so that $x_i < x_j$ whenever $i < j$ (technically this rules out certain discrete ranges, like the rational numbers, which cannot be indexed in this way, but it covers all the cases that we will be interested in, which are usually subsets of the integers). Then*
(i) $F(x_r) - F(x_{r-1}) = f(x_r)$ for all $r \geq 2$
(ii) F is a nondecreasing function. That is, whenever $x_t > x_s$

$$F(x_t) \geq F(x_s)$$

(iii) If R has a largest element, x_n, then $F(x_n) = 1$. (Otherwise $\lim_{r\to\infty} F(x_r) = 1$).

Proof.

(i) We can write

$$\begin{aligned} F(x_r) &= P(\{X \leq x_r\}) \\ &= P(\{X \leq x_{r-1}\} \cup \{X = x_r\}) \\ &= P(\{X \leq x_{r-1}\}) + P(\{X = x_r\}) && \text{(Disjoint additivity)} \\ &= F(x_{r-1}) + f(x_r) && \text{(Defs of } F \text{ and } f\text{)} \end{aligned}$$

Subtracting $F(x_{r-1})$ from the first and last expression gives us (i).

(ii) We have, using (i) and the fact that $f(x_r) \geq 0$ for all r:

$$F(x_r) = F(x_{r-1}) + f(x_r) \geq F(x_{r-1})$$

(iii)

$$\begin{aligned} F(x_n) &= \sum_{r=1}^{n} f(x_r) && \text{(by (6.2.2))} \\ &= 1 && \text{(by (6.1))} \end{aligned}$$

where the last line follows from the fact that we assume x_n to be the largest element of R, and so the sum is over every element of R.

□

The CDF tells us directly how to compute probabilities of events of the form $\{X \leq x\}$ for some $x \in R$. With barely any additional calculation, we can also use the CDF to compute probabilities for events of the form $\{X < x\}$ and $\{x_r \leq X \leq x_s\}$, as the next theorem illustrates.

Theorem 6.3 (More Properties of the CDF)**.** *Let X and F be defined as in the previous two theorems. Then*
(i) $P(X > x_r) = 1 - F(x_r)$
(ii) For $r \leq s$, we have

$$P(x_r \leq X \leq x_s) = F(x_s) - F(x_{r-1})$$

Proof.

(i) We can write

$$\begin{aligned} P(X > x_r) &= P(\overline{\{X \leq x_r\}}) \\ &= 1 - P(X \leq x_r) && \text{(by the complement rule)} \\ &= 1 - F(x_r) && \text{(by the Def. of F)} \end{aligned}$$

(ii) We have

$$\begin{aligned} P(x_r \leq X \leq x_s) &= P(\{X \leq x_s\} - \{X < x_r\}) \\ &= P(\{X \leq x_s\}) - P(\{X < x_r\}) \\ &\qquad ((\text{P5}), \text{ since } \{X < x_r\} \subset \{X \leq x_s\}) \\ &= P(\{X \leq x_s\}) - P(\{X \leq x_{r-1}\}) \\ &\qquad (\{X < x_r\} = \{X \leq x_{r-1}\}) \\ &= F(x_s) - F(x_{r-1}) \end{aligned}$$

□

6.3 Examples of Discrete Random Variables

Every situation has a different sample space. Random variables, however, take these sample spaces and create a partition, which, as we've seen, induces a simplified domain of events, and a simplified probability measure. As a result, problems that are very different at the level of sample space can give rise to random variables with very similar structures. This similarity allows us to generalize several probabilistic properties from one situation to another which is seemingly very different.

6.3.1 The Bernoulli Distribution

The simplest possible random variable takes the universe and divides it in two pieces. On one piece, the random variable takes the value 1. On the other, it takes the value 0. Any proposition which, for each outcome in U, is either true or false immediately defines such a random variable. This random variable can be set to 1 for all outcomes that make the proposition true, and 0 elsewhere. A random variable with this structure is called a **Bernoulli random variable**.

- Question: What's the range of a Bernoulli random variable?

Since we must have $\sum_{x=0}^{1} f(x) = 1$, the probability mass function of a Bernoulli random variable is determined entirely by setting one of its values, since the other value is obtained from the complement rule.

We define $p, 0 \leq p \leq 1$ to be equal to $f(1)$, which implies that $f(0) = 1 - p$.

This value p is called a **parameter** of the random variable. If X is a Bernoulli random variable with parameter p, you may see the notation

$$X \sim \mathcal{B}\text{ern}(p)$$

where the $\sim$ symbol is read "is distributed as".

We can write the probability mass function concisely as

$$f(x) = p^x(1-p)^{1-x}.$$

The use of exponents here is a bit "clever": since x is always 0 or 1, one of the exponents is always 1, and the other is always 0. Whichever term has an exponent of zero disappears since its value is just 1.

Categorical and Discrete Uniform Random Variables

We can generalize the notion of a Bernoulli random variable to one that takes more than two (but still a finite number of) values.

A random variable whose range is a finite set, $R = \{x_1, x_2, \ldots, x_n\}$, and which has no restrictions on its PMF, is called a **categorical** random variable.

In the most general case, we can't simplify the probability mass function beyond just stating all its values. We can denote

$$p_1 = f(x_1), \quad p_2 = f(x_2), \quad \ldots, \quad p_n = f(x_n)$$

and say

$$X \sim \mathcal{C}\text{at}(p_1, p_2, \ldots, p_n)$$

(Of course we do not have complete freedom in choosing the p_rs, as they must sum to 1)

If we're in a situation where the values of the random variable are exchangeable (that is, the numbers assigned are basically arbitrary labels), then we can set

$$p_1 = p_2 = \cdots = p_n$$

which, since $p(\cdot)$ sums to 1, implies as we've seen before that

$$p(x_r) = \frac{1}{n} \quad \text{for all } r \in \{1, 2, \ldots, n\}$$

Then, provided $x_1 < x_2 < \cdots < x_n$, by (6.2.2), we get

$$F(x_r) = \frac{r}{n}$$

In this case, the random variable has a **discrete uniform** distribution and we might write

$$X \sim \mathcal{DU}(n)$$

6.4 The Algebra of Random Variables

Remember that random variables are functions from U to $\mathbb{R}$. As such, we can manipulate random variables in the same way that we manipulate arbitrary real-valued functions.

If X and Y are random variables, α is a real number, and g is a function from $\mathbb{R}$ to $\mathbb{R}$ (for example $g(x) = x^2$ or $g(x) = \sin(x)$), then the following are all random variables:

(i) $X + Y$

(ii) αX

(iii) $X + \alpha$

(iv) XY

(v) $g(X)$

They are defined in the natural way. For any $a \in U$:

(i) $(X + Y)(a) = X(a) + Y(a)$

(ii) $(\alpha X)(a) = \alpha X(a)$

(iii) $(X + \alpha)(a) = X(a) + \alpha$

(iv) $(XY)(a) = X(a)Y(a)$

(v) $(g(X))(a) = g(X(a))$

For example:

(i) If X is the number of boys born in Tucson in a given year, and Y is the number of girls, then $X + Y$ is the total number of babies.

(ii) If X is the number of items of a certain kind that a store sells in a given day, and β is the profit margin per unit, then βX is a random variable representing the gross profit in a given day.

(iii) If α is the fixed operating cost of running the store for a day, then $\beta X - \alpha$ is the daily profit.

(iv) A machine produces window panes to target dimensions, but with some error. If X is the width of a given window, Y is its height, then XY describes the area of the windows.

(v) If X is the amplitude of a sound wave, then (roughly) $\log(X)$ is the decibel level.

6.4.1 Joint Distributions

If X and Y are two random variables on the same probability space $(U, \mathcal{P}(U), P)$, then any given sample element will have an X value and a Y value. For example, we might take U to be the set of students in this class, and let X be major (for simplicity, say ISTA vs. not ISTA) and Y be handedness. Technically we should assign numbers to each category — say 0 and 1. Using R_X to denote the range of X and R_Y to denote the range of Y, we have

$$R_X = \{0, 1\} (= \{\text{Non-ISTA}, \text{ISTA}\})$$
$$R_Y = \{0, 1\} (= \{\text{Left}, \text{Right}\})$$

Then we can talk about the probability of a particular *combination* of values (say, a left-handed ISTA major).

Definition 6.5 (Joint Distribution)**.** *Let X and Y be discrete random variables with ranges $R_X = \{x_1, x_2, \dots\}$ and $R_Y = \{y_1, y_2, \dots\}$, and probability mass functions f_X and f_Y, respectively. We define the **joint distribution** (or **joint mass function**) to be a function $f_{X,Y}$ from $R_X \times R_Y$ (the set of all ordered pairs (x, y), where $x \in R_X$ and $y \in R_Y$) to $[0, 1]$, which sets*

$$f_{X,Y}(x, y) = P(\{X = x\} \cap \{Y = y\})$$

In a sense we can think of a "two-dimensional" random variable, or **random vector**, (X, Y), whose range is $R_X \times R_Y$ (i.e., all ordered pairs whose first element is in

R_X and whose second element is in R_Y), and whose probability mass function is $f_{X,Y}$.

In our example, we can think of the joint mass function as applying to the cells of a 2×2 table (see Table 6.1)

		x (Major)	
		0 (non-ISTA)	1 (ISTA)
y (Hand)	0 (Left)	$f_{X,Y}(0,0)$	$f_{X,Y}(1,0)$
	1 (Right)	$f_{X,Y}(0,1)$	$f_{X,Y}(1,1)$

Table 6.1: A Simple Joint Mass Function

Then, for example, the probability of selecting a left-handed ISTA major is given by $f_{X,Y}(1,0)$.

It should be easy to see that the joint distribution function has the following properties:

Lemma 6.4 (Marginalization)**.**

(i) *For each* $y \in R_Y$, $\sum_{x \in R_X} f_{X,Y}(x,y) = f_Y(y)$

(ii) *For each* $x \in R_X$, $\sum_{y \in R_Y} f_{X,Y}(x,y) = f_X(x)$

(iii) $\sum_{x \in R_X} \sum_{y \in R_Y} f_{X,Y}(x,y) = 1$

Proof. Every element of the sample space gets mapped to some cell of the distribution table. The table therefore gives us three different ways of constructing a partition: one by taking each row as a block, one by taking each column as a block, and the third by taking each cell as a block.

Part (i) says that when we sum over the cells in a particular column, we get the total ("marginal") probability assigned to that column.

Part (ii) says that when we sum over the cells in a particular row, we get the total ("marginal") probability assigned to that row.

Part (iii) says that when we sum over all the cells, we get the probability of the entire sample space, which is 1.

More formally:

(i) For any $y \in R_Y$, we have

$$\begin{aligned}
\sum_{x \in R_X} f_{X,Y}(x,y) &= \sum_{x \in R_X} P(\{X = x\} \cap \{Y = y\}) \\
&= P(\bigcup_{x \in R_X} \{X = x\} \cap \{Y = y\}) && \text{(Disjoint Additivity)} \\
&= P((\bigcup_{x \in R_X} \{X = x\}) \cap \{Y = y\}) && \text{(Distributive)} \\
&= P(U \cap \{Y = y\}) && = P(Y = y) \\
& && (\{\text{X} = \text{x}\}\text{s form a partition}) \\
&= f_Y(y) && \text{(Def. of PMF)}
\end{aligned}$$

(ii) Exactly the same as (i), with roles reversed

(iii) We can either use (6.1), applied to (i) or (ii), or we can note that the sets $\{X = x_r\} \cap \{Y = y_s\}$, taking every combination of r and s, form a partition of S, and hence (P7) applies.

□

6.5 Expectation

So far, we have characterized random variables mainly in terms of their PMF and CDF. Both of these give us a list of probabilities that we can use to calculate the probability of any event we want.

Often we want more concise information, telling us

- What's a central, "typical" value of the feature represented by the random variable?
- How much variability is there in this feature?

Both of these questions (along with lots of others) can be addressed by taking **expectations** of functions of the random variable.

Analogy Suppose I own an electronics store. On 20% of the days my shop is open, I don't sell any laptops. 40% of the time I sell one, 30% of the time I sell 2, and 10% of the time I sell 3. How many laptops do I *expect* to sell tomorrow?

Our expectations are a weighted average of all the possible things that can happen, where I weigh more likely outcomes more heavily.

6.5.1 Expected Value

Definition 6.6 (Expected Value/Mean of a Random Variable)**.** *Let X be a random variable with range $R = \{x_1, x_2, \dots\}$ and probability mass function p. We define the* ***expectation*** *(or mean) of X, written $\mathbb{E}[X]$, to be*

$$\mathbb{E}[X] = \sum_{x \in R} x \cdot f(x)$$

We will sometimes use the Greek letter μ (pronounced "mew", like a cat) to stand for $\mathbb{E}[X]$. This is the Greek version of "m" for "mean".

In the laptop example, we have $R = \{0, 1, 2, 3\}$, so I have:

$$\begin{aligned}
\mu = \mathbb{E}[X] &= \sum_{x=0}^{3} x \cdot f(x_j) \\
&= 0 \cdot 0.2 + 1 \cdot 0.4 + 2 \cdot 0.3 + 3 \cdot 0.1 \\
&= 0 + 0.4 + 0.6 + 0.3 \\
&= 1.3
\end{aligned}$$

On average, I sell 1.3 laptops a day. Notice that the average is not a possible value for X. This is often the case for discrete random variables. One reason to define expected value this way is that, under a relative frequency interpretation of probability, it represents the "long run average" if we repeatedly observe the random variable. Consider repeating a random experiment N times, and observing the value of the random variable X each time. Letting $x^{(i)}$ denote the value of X on the ith run of the experiment, the sample mean of the runs will be

$$\frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

We can simplify this sum by grouping together identical terms, to get

$$\frac{1}{N} \sum_{x \in R_X} \text{Count}(X = x) \cdot x$$

In a sample of size N, we expect x_1 to occur about $N \cdot f(x_1)$ times (on average), x_2 to occur $N \cdot f(x_2)$ times, and so on, so that if we compute the sample mean, it will tend toward

$$\frac{1}{N} \sum_{x \in R_X} N \cdot f(x) \cdot x = \sum_{x \in R_X} f(x) \cdot x$$

which is exactly how we've defined expected value.

6.5.2 Expectation of Functions of a Random Variable

Suppose that, in the laptop example above, I charge \$1000 for each laptop I sell. How much do I expect to bring in tomorrow?

The logic is the same as when I determine the expected number of laptops sold; I just have to apply the income function first.

Let $h(x) = 1000x$ represent my gross income in a day. Then I can define

$$\begin{aligned}
\mathbb{E}\left[h(X)\right] &= \sum_{x=0}^{3} h(x) f(x) \\
&= 1000 \cdot 0 \cdot 0.2 + 1000 \cdot 1 \cdot 0.4 + 1000 \cdot 2 \cdot 0.3 + 1000 \cdot 2 \cdot 0.3 + 1000 \cdot 3 \cdot 0.1 \\
&= 0 + 400 + 600 + 300 \\
&= 1300
\end{aligned}$$

So I expect to bring in \$1300 on average in a given day.

Suppose I buy my laptops for \$800 each. If I want to compute my profit, as opposed to my raw income, I can set $g(x) = 1000x - 800x = 200x$

and compute my expected profit the same way.

Or I can take into account my fixed cost of operating the store; say, \$100 a day.

Then I have $f(x) = 200x - 100$, and I get

$$\begin{aligned}
\mathbb{E}[f(x)] &= \sum_{x=0}^{3}(200x - 100)f(x) \\
&= \sum_{x=0}^{3} 200xf(x) - \sum_{x=0}^{3} 100f(x) \\
&= 200\sum_{x=0}^{3} xf(x) - 100\sum_{x=0}^{3} f(x) \\
&= 200 \cdot \mathbb{E}[X] - 100 \cdot 1 \\
&= 260 - 100 \\
&= 160
\end{aligned}$$

So, I turn a profit of $160 a day, on average.

Notice the general pattern:

Definition 6.7 (Expectation of a Function of a Random Variable)**.** *Let X be a random variable with range R, and let h be any function which is defined on R, and which returns real numbers. Remember, we showed that $h(X)$ is a random variable in its own right. Then we have*

$$\mathbb{E}[h(X)] = \sum_{x \in R} h(x)f(x)$$

6.5.3 Properties of Expectation

Recall from our discussion of the algebra of random variables that if X and Y are both random variables, then $X + Y$ is also a random variable. Therefore, the expectation of $X + Y$ is a well-defined quantity.

We could compute it by looking at the range of $X+Y$ and computing a PMF for this random variable, which we would use to find the expectation. However, it's often easier to stick with pairs (x, y), and compute the weighted average of the numbers $x+y$, where the weights come from the *joint* distribution of X and Y. This will give us the same result; just with potentially more terms in the sum.

Definition 6.8. *Let X have range R_X and PMF f_X, let Y have range R_Y and PMF f_Y, and let $f_{X,Y}$ be the joint PMF of X and Y. Then*

$$\mathbb{E}[X+Y] = \sum_{(x,y)\in(R_X\times R_Y)} (x+y) f_{X,Y}(x,y)$$

It turns out that we never have to use this definition directly, due to the following Theorem.

Theorem 6.5 (a)**.** *For any two random variables, X and Y, on a common probability space,*

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Proof. This is a matter of some algebra, and using the Marginalization Lemma.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{(x,y)\in R_X\times R_Y} (x+y) f_{X,Y}(x,y) && \text{(Def. 6.8)}\\
&= \sum_{x\in R_X}\sum_{y\in R_Y} (x+y) f_{X,Y}(x,y)\\
&= \sum_{x\in R_X}\sum_{y\in R_Y} (x f_{X,Y}(x,y) + y f_{X,Y}(x,y))\\
&= \sum_{x\in R_X}\sum_{y\in R_Y} x f_{X,Y}(x,y) + \sum_{x\in R_X}\sum_{y\in R_Y} y f_{X,Y}(x,y)\\
&= \sum_{x\in R_X} x \sum_{y\in R_Y} f_{X,Y}(x,y) + \sum_{y\in R_Y} y \sum_{x\in R_X} f_{X,Y}(x,y)\\
&= \sum_{x\in R_X} x f_X(x) + \sum_{y\in R_Y} y f_Y(y) && \text{(Marginalization Lemma)}\\
&= \mathbb{E}[X] + \mathbb{E}[Y] && \text{(Def. of } \mathbb{E}\text{)}
\end{aligned}$$

□

Theorem 6.5 (b)**.** *If the function h is a* linear *function (that is, $h(x) = ax + b$ for some choices of constants a and b), then we can exchange the order of $\mathbb{E}$ and h. That is:*

$$\begin{aligned}
\mathbb{E}[h(X)] &= h(\mathbb{E}[X]), \text{ that is to say,}\\
\mathbb{E}[aX+b] &= a\mathbb{E}[X] + b
\end{aligned}$$

Proof. This comes straight from the definition. Suppose X has range R and PMF f. Then

$$\begin{aligned}
\mathbb{E}[aX+b] &= \sum_{x\in R}(ax+b)f(x) \\
&= \sum_{x\in R}(axf(x)+bf(x)) \\
&= \sum_{x\in R} axf(x) + \sum_{x\in R} bf(x) \\
&= a\sum_{x\in R} xf(x) + b\sum_{x\in R} f(x) \\
&= a\mathbb{E}[X] + b
\end{aligned}$$

where in the last line we've used the definition of Expectation in the first term, and the fact that the PMF sums to 1 in the second term. □

Note that we're free to set $b = 0$ or $a = 1$ in the above, in which case we get the following two statements:

$$\mathbb{E}[aX] = a\mathbb{E}[X] \tag{6.2}$$
$$\mathbb{E}[X+b] = \mathbb{E}[X] + b \tag{6.3}$$

6.5.4 Variance

Suppose we have already calculated $\mu = \mathbb{E}[X]$. Then this is just a number. If we now set $h(x) = (x-\mu)^2$, then $\mathbb{E}[h(X)]$ is called the **variance** of X. It gives us a way of measuring how far away X tends to be from its mean. It's also a measure of how much uncertainty is associated with our expectation of the value of X itself: if the variance is high, we "expect" to be off by a greater amount.

Definition 6.9 (Variance of a Random Variable)**.** *Let X be defined as above, and let $\mu = \mathbb{E}[X]$. We define the **variance** of X to be*

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}\left[(X-\mu)^2\right] = \sum_{x\in R}(x-\mu)^2 f(x)$$

Example In the case of the laptops, I have

$$\begin{aligned}
\mathbb{V}\text{ar}\left[X\right] &= \mathbb{E}\left[(X-\mu)^2\right] \\
&\quad \mathbb{E}\left[(X-1.3)^2\right] \\
&= \sum_{x=0}^{3}(x-1.3)^2 f(x) \\
&= (0-1.3)^2 \cdot 0.2 + (1-1.3)^2 \cdot 0.4 + (2-1.3)^2 \cdot 0.3 + (3-1.3)^2 \cdot 0.1 \\
&= 1.69 \cdot 0.2 + 0.09 \cdot 0.4 + 0.49 \cdot 0.3 + 2.89 \cdot 0.1 \\
&= 0.338 + 0.036 + 0.147 + 0.289 \\
&= 0.810
\end{aligned}$$

The variance has many useful and elegant mathematical properties. One of its limitations, however, is the fact that it's expressed in different units from those of X itself: namely, squared units. This fact is captured in the alternative symbol for variance, σ^2, where σ is the Greek letter "sigma".

We can return to the original units by taking the square root, leaving us with σ. This is called the **standard deviation** of X.

Definition 6.10 (Standard Deviation of a Random Variable)**.** *Let X be defined as above, with mean μ and variance $\mathbb{V}\text{ar}\left[X\right]$. We define the **standard deviation** of X to be the square root of the variance.*

$$\sigma = \sqrt{\mathbb{V}\text{ar}\left[X\right]}$$

Example The standard deviation in the daily number of laptops is $\sqrt{\sigma^2} = \sqrt{0.81} = 0.9$.

Notice that there's no function h that we can pick (in general) so that $\sigma = \mathbb{E}\left[h(X)\right]$. We need to compute the variance first in order to get the standard deviation.

6.5.5 Properties of the Variance

The following theorem gives us an easier way to compute variance than using the direct definition.

Theorem 6.6 (a)**.**

$$\mathbb{V}\mathrm{ar}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$$

Proof. Just follow the algebra, and apply Thm 6.5. Denote $\mu = \mathbb{E}\left[X\right]$.

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[X\right] &= \mathbb{E}\left[(X-\mu)^2\right] \\
&= \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mu\mathbb{E}\left[X\right] + \mu^2 && \text{(Thm. 6.5(b))} \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}\left[X\right]\mathbb{E}\left[X\right] + \mathbb{E}\left[X\right]^2 && \text{(Def. of } \mu\text{)} \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}\left[X\right]^2 + \mathbb{E}\left[X\right]^2 && (\mathbb{E}\left[X\right] \text{ is a number)} \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2
\end{aligned}$$

□

If we rescale X by multiplying by a constant, the variance changes in a simple way; but not just by the constant the way the mean does. Remember, variance is in squared units.

Theorem 6.6 (b)**.** *Let $a \in \mathbb{R}$ be a constant. We already know that aX is then a random variable, and so it has a variance. In particular,*

$$\mathbb{V}\mathrm{ar}\left[aX\right] = a^2\mathbb{V}\mathrm{ar}\left[X\right]$$

Proof. This is just algebra, and we could get the result directly from the definition of Variance, but it's easier to use the alternative expression for Variance from Thm 6.6(a), and then apply Thm 6.5(b).

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[aX\right] &= \mathbb{E}\left[(aX)^2\right] - \mathbb{E}\left[aX\right]^2 && \text{(Thm. 6.6(a))} \\
&= \mathbb{E}\left[a^2X^2\right] - \mathbb{E}\left[aX\right]^2 \\
&= a^2\mathbb{E}\left[X^2\right] - (a\mathbb{E}\left[X\right])^2 && \text{(Thm. 6.5(b))} \\
&= a^2\mathbb{E}\left[X^2\right] - a^2\mathbb{E}\left[X\right]^2 \\
&= a^2\left(\mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2\right) \\
&= a^2\mathbb{V}\mathrm{ar}\left[X\right] && \text{(Thm. 6.6(a))}
\end{aligned}$$

□

What should happen to the variance if we add a constant to a random variable?

We saw that when we add (respectively, subtract) a constant, the *mean* shifts up (respectively, down) by that exact amount. But the variance is a measure of how far away the variable tends to be from its mean. Since both the mean and the individual values are moving the same way, the variance should not change. Indeed this is the case:

Theorem 6.7. *Let X be a random variable with a defined variance, and let c be a real constant. Then*

$$\mathbb{V}\text{ar}\left[X + c\right] = \mathbb{V}\text{ar}\left[X\right]$$

Proof. Homework □

6.6 Exercises

1. (Adapted from Applebaum) A discrete random variable has range $R = \{0, 1, 2, 3, 4, 5\}$ and cumulative distribution whose first five values are

$$F(0) = 0, \quad F(1) = 1/9, \quad F(2) = 1/6$$
$$F(3) = 1/3, \quad F(4) = 1/2$$

   (a) Find $f(x)$, the value of the PMF, for each $x \in R$.

   (b) Find $\mathbb{E}\left[X\right]$ and $\mathbb{V}\text{ar}\left[X\right]$.

2. (Adapted from Applebaum) Three fair coins are tossed. Find the PMF of the random variable Y whose value is given by the total number of heads.

3. A bucket has a red balls and b blue ones. You draw k balls at random ($1 \leq k \leq a + b$) without replacement from the bucket. Let Y be a random variable representing the number of red balls in your sample. Find an expression representing the probability that $Y = y$.

4. Let X be a discrete random variable with range $R = \{x_1, \ldots, x_n\}$, where the elements are listed in increasing order. Show that the following probabilities can be computed from the CDF as follows:

   (a) $P(x_r < X \leq x_s) = F(x_s) - F(x_r)$

   (b) $P(x_r < X < x_s) = F(x_{s-1}) - F(x_r)$

   (c) $P(x_r \leq X < x_s) = F(x_{s-1}) - F(x_{r-1})$

   (d) $P(X \geq x_r) = 1 - F(x_{r-1})$

5. A game show allows contestants to win money by spinning a wheel. The wheel is divided into four equally sized spaces, three of which contain dollar amounts and one of which says BUST. The contestant can spin as many times as she wants, adding any money won to her total, but if she lands on BUST, the game ends and she loses everything won so far.

   (a) Suppose for the moment that the contestant keeps going until she goes bust. Let W be the random variable that represents the number of successful spins before going bust. Indicate the valid range for W, and find an expression for its CDF.

   (b) Use the CDF you found above to derive an expression for the PMF of W.

   (c) The dollar amounts on the non-BUST spaces are \$100, \$500, and \$5000. Let X_n be the random variable denoting the amount of money won or lost on the n^{th} spin. Find $\mathbb{E}[X_1]$ (note that there is nothing to lose on the first spin).

   (d) Let S_n be the cumulative amount won after n spins. Find $\mathbb{E}[X_n | S_{n-1}]$, as a function of S_n. How large does S_n have to get before the expected gain by spinning again is negative?

6. Use induction to extend Theorem 6.5(a), showing that if $X_1, \ldots, X_n$ are arbitrary random variables, then

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]$$

7. Let X be a random variable and $c \in \mathbb{R}$ a constant. Show that

$$\mathbb{V}\text{ar}[X + c] = \mathbb{V}\text{ar}[X]$$

8. (Adapted from Bolstad) A discrete random variable, Y, has a probability mass function given by the following table.

y	$f(y)$
0	0.2
1	0.3
2	0.1
3	0.1
4	0.1

   (a) Calculate $P(1 < Y \leq 3)$.

   (b) Calculate $\mathbb{E}[Y]$.

   (c) Calculate $\mathbb{V}\text{ar}[Y]$.

   (d) Let $W = 2Y + 3$. Calculate $\mathbb{E}[W]$.

   (e) Calculate $\mathbb{V}\text{ar}[W]$.

9. (Adapted from Bolstad) A discrete random variable, Y, has a probability mass function given by the following table.

y	$f(y)$
0	0.1
1	0.2
2	0.3
5	0.4

   (a) Calculate $P(0 < Y < 5)$.

   (b) Calculate $\mathbb{E}[Y]$.

   (c) Calculate $\mathbb{V}\text{ar}[Y]$.

   (d) Let $W = 3Y - 1$. Calculate $\mathbb{E}[W]$.

   (e) Calculate $\mathbb{V}\text{ar}[W]$.

10. (Adapted from Bolstad) Let X and Y be jointly distributed discrete random variables. Their joint probability distribution is given in the following table:

x	y					$f(x)$
	1	2	3	4	5	
1	0.02	0.04	0.06	0.08	0.05	
2	0.08	0.02	0.10	0.02	0.03	
3	0.05	0.05	0.03	0.02	0.10	
4	0.10	0.04	0.05	0.03	0.03	
$f(y)$						

(a) Find the marginal PMF of X.

(b) Find the marginal PMF of Y.

(c) Calculate the conditional probability $P(X = 3|Y = 1)$.

Chapter 7

Relationships Between Random Variables

We are often interested in making an *inference* about one random variable after having observed the value of another. In order to do this sensibly, we need to know how they relate: if it turns out that a random variable X takes a value near the top of its range (perhaps a lab test reveals high levels of a particular protein in a blood sample), how does that help me say what Y (e.g., the future growth rate of a tumor) is likely to do?

We can assess the relationship between random variables in several different ways. In this chapter we will start off by discussing measures of a *linear* relationship between two random variables, and then we'll move on to a more detailed assessment of the relationship, via **conditional distributions**. Using the latter, we will be able to carry out principled inferences from one variable to another (e.g., *given* that X is 300, what is the *expected value* of Y?)

7.1 Linear Relationships: Covariance and Correlation

If we have two random variables, X and Y defined on the same universe, U, it's often the case that, when one is high, the other is high, and when one is low, the other is low (or vice versa). For example, if X is height and Y is weight, I expect that taller

people will *tend* to weigh more (I may well have a short person who weighs more than someone taller, but *on average* height and weight go in the same direction).

7.1.1 Definitions

When two random variables are related in this way, we might say that they tend to "vary together", or that they "covary". We define the **covariance** between X and Y as follows:

Definition 7.1 (Covariance)**.** *Let X be a random variable with range R_X and mean μ_X, and let Y be another random variable on the same universe, with range R_Y and mean μ_Y. The **covariance** between X and Y is defined as*

$$\begin{aligned}\mathbb{C}\text{ov}\left(X,Y\right) &= \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \sum_{(x,y)\in R_X\times R_Y} (x-\mu_x)(y-\mu_Y)f_{X,Y}(x,y) \\ &= \sum_{x\in R_x}\sum_{y\in R_Y} (x-\mu_x)(y-\mu_Y)f_{X,Y}(x,y)\end{aligned}$$

Notice what we're doing here. We're looking at all possible *combinations* of x and y values, and getting a weighted average of the quantity $(x-\mu_X)(y-\mu_Y)$, where the weights come from the *joint distribution* of X and Y. In other words, the weights are the probability of *that combination*.

Now consider the quantity $(x-\mu_X)(y-\mu_Y)$ itself. When is this positive? When is it negative? When is it large/small?

We get a positive value for a particular (x,y) pair if both "coordinates" are on the same side of their respective mean. We get a negative value when they're on opposite sides. The absolute value is large when the distances from the mean are large.

Notice that

$$\mathbb{C}\text{ov}\left(X,X\right) = \mathbb{E}\left[(X-\mu_X)(X-\mu_X)\right] = \mathbb{E}\left[(X-\mu_X)^2\right] = \mathbb{V}\text{ar}\left[X\right]$$

This tells us that the magnitude of the covariance is affected by the scale of the random variables themselves: if we change the units of X (say, from inches to centimeters), the covariance between X and Y will change.

An alternative measure of the tendency of two random variables to increase or decrease together takes covariance and *standardizes* it, by removing the units.

Definition 7.2 (Correlation)**.** *Let X and Y be random variables with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2 (respectively). The **correlation** between X and Y is defined as*

$$\rho(X,Y) = \frac{\mathbb{C}\text{ov}\,(X,Y)}{\sigma_X \sigma_Y}$$

Remember that covariance is the expected value of the product of deviation scores. When I take the product of a deviation score in the X-direction and a deviation score in the Y-direction, the resulting quantity is like the area of a rectangle, whose width is on the X-scale and whose height is on the Y-scale. So if I want to create a "normalized rectangle", I should set its width to be a "standard unit" in the X-direction, and its height to be a "standard unit" in the Y-direction: that is, it should be σ_X wide and σ_Y high. Now I can measure any area using this standard rectangle as my unit.

So covariance measures the expected product of deviations in the original units of X and Y, whereas correlation measures the expected product of deviations in terms of a standard rectangle which is σ_X by σ_Y.

7.1.2 Basic Properties

Just as we had for expectation and variance, there are some important properties associated with the covariance, which are collected in the following theorem:

Theorem 7.1. *Let X, Y and Z be random variables and let a and b be any real-valued constants. Then*
(a) $\mathbb{C}\text{ov}\,(X,Y) = \mathbb{E}\,[XY] - \mathbb{E}\,[X]\,\mathbb{E}\,[Y]$
(b) $\mathbb{C}\text{ov}\,(X, aY + bZ) = a \cdot \mathbb{C}\text{ov}\,(X,Y) + b \cdot \mathbb{C}\text{ov}\,(X,Z)$
(c) $\mathbb{V}\text{ar}\,[X+Y] = \mathbb{V}\text{ar}\,[X] + 2\mathbb{C}\text{ov}\,(X,Y) + \mathbb{V}\text{ar}\,[Y]$

Proof. Homework. □

Notice the similarity between part (a) of this theorem and part (a) of Theorem 6.6. In fact, since $\mathbb{C}\text{ov}\,(X,X) = \mathbb{V}\text{ar}\,[X]$, we can derive that theorem as a special case of this one.

In part (c), notice that we *cannot* add variances the way we add expectations, in general. Only if the covariance between two variables is zero will the variance of the sum be the sum of the variances.

7.1.3 (*) Bounds on covariance and correlation

The next theorem gives us a range for the covariance between two variables, in terms of the variances of the individual variables. As a result, we will prove that the correlation must always be between -1 and 1.

Theorem 7.3. *Let X and Y be random variables with variances σ_X^2 and σ_Y^2, respectively. Then*
(a) $-\sqrt{\sigma_X^2\sigma_Y^2} \leq \mathbb{C}\text{ov}\,(X,Y) \leq \sqrt{\sigma_X^2\sigma_Y^2}$
(b) $-1 \leq \rho(X,Y) \leq 1$

To prove (a), we need a version of a powerful theorem, relevant to just about every branch of mathematics (including the proof of the "triangle inequality" which, among other things tells us that the shortest distance between two points is a straight line!). This theorem is called the *Cauchy-Schwartz inequality*. The version of it that we need is stated as follows.

Lemma 7.4 (Cauchy-Schwarz Inequality)**.** *Let X and Y be random variables with finite variances. Then*

$$\mathbb{E}\,[XY]^2 \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]$$

Proof of Lemma. Let t be any real number, and define a random variable given by $(X-tY)^2$. Notice that, since it is always the square of a real number, it is always nonnegative, and therefore its expectation is always nonnegative, no matter what we choose for t. We can write

$$\begin{aligned}\mathbb{E}\left[(X-tY)^2\right] &= \mathbb{E}\left[X^2 - 2tXY + t^2Y^2\right] \\ &= t^2\cdot\mathbb{E}\left[Y^2\right] - 2t\cdot\mathbb{E}\,[XY] + \mathbb{E}\left[X^2\right] && \text{(Thm 6.5)} \\ &\geq 0 && \text{(as } \mathbb{E} \text{ of a nonnegative variable)}\end{aligned}$$

All the expectations in this expression are just constant numbers. Consider the expression as a function of t. You can see that it's of the form at^2+bt+c, where, in this case $a=\mathbb{E}\,[Y^2]$, $b=-2\mathbb{E}\,[XY]$ and $c=\mathbb{E}\,[X^2]$.

Recall that, in the quadratic formula for finding roots of a quadratic polynomial, the term $\sqrt{b^2-4ac}$ determines whether there are 0, 1 or 2 real-valued roots. We know that our quadratic can have at most 1 root (because it is strictly nonnegative), and so $b^2-4ac\leq 0$. That is

$$4\mathbb{E}\,[XY]^2 \leq 4\mathbb{E}\left[Y^2\right]\mathbb{E}\left[X^2\right]$$

Dividing both sides by 4 gives the result. □

Proof of Theorem. The Lemma holds for *any* random variables X and Y; in particular we can replace X with $(X - \mu_X)$ and Y with $(Y - \mu_Y)$ (where μ_X and μ_Y are the respective means) to get

$$\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]^2 \leq \mathbb{E}\left[(X-\mu_X)^2\right]\mathbb{E}\left[(Y-\mu_Y)^2\right]$$

In other words

$$\mathbb{C}\mathrm{ov}\left(X,Y\right)^2 \leq \mathbb{V}\mathrm{ar}\left[X\right]\mathbb{V}\mathrm{ar}\left[Y\right]$$

Taking (positive) square roots:

$$\left|\mathbb{C}\mathrm{ov}\left(X,Y\right)\right| \leq \sqrt{\sigma_X^2\sigma_Y^2}$$

which is equivalent to

$$-\sqrt{\sigma_X^2\sigma_Y^2} \leq \mathbb{C}\mathrm{ov}\left(X,Y\right) \leq \sqrt{\sigma_X^2\sigma_Y^2}$$

proving (a). Now divide by $\sqrt{\sigma_X^2\sigma_Y^2}$ to get

$$-\frac{\sqrt{\sigma_X^2\sigma_Y^2}}{\sqrt{\sigma_X^2\sigma_Y^2}} \leq \frac{\mathbb{C}\mathrm{ov}\left(X,Y\right)}{\sqrt{\sigma_X^2\sigma_Y^2}} \leq \frac{\sqrt{\sigma_X^2\sigma_Y^2}}{\sqrt{\sigma_X^2\sigma_Y^2}}$$

In other words

$$-1 \leq \rho(X,Y) \leq 1$$

proving (b). □

7.2 Conditional Distributions and Independence

7.2.1 Conditional Distributions of a Random Variable

If we "observe" the value (say x_0) of one random variable (say X), then the possible outcmoes are limited to the set $\{X = x_0\}$. We will often be interested in examining how another random variable (say, Y) behaves, conditioned on that set.

Suppose Y has range R_Y, and consider the conditional probabilities

$$\begin{aligned} P(\{Y=y\}|\{X=x_0\}) &= \frac{P(\{X=x_0\}\cap\{Y=y\})}{P(\{X=x_0\})} \qquad y \in R_Y \\ &= \frac{f_{X,Y}(x_0,y)}{f_X(x_0)} \end{aligned}$$

There's nothing new here; we're just applying the definition of conditional probability, and examining the probabilities of the events $\{Y = y\}, y \in R_Y$, conditioned on the fixed event $\{X = x_0\}$.

Definition 7.3 (Conditional Distribution)**.** *We define the* ***conditional distribution*** *(or* ***conditional distribution function****) of Y given that $X = x_0$ to be a function $f_{Y|x_0} : R_Y \to [0, 1]$, defined as*

$$f_{Y|x_0}(y) = \frac{f_{X,Y}(x_0, y)}{f_X(x_0)}$$

As we did for conditional probability generally, we will usually write $f_{Y|X}(y|x_0)$ to mean $f_{Y|x_0}(y)$.

From this, we get a multiplication rule for random variables

Theorem 7.5 (Multiplication Rule for Random Variables)**.** *for any x and y, we can write the decomposition*

$$f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y|x)$$

We also have a version of Bayes Theorem for random variables. It's really just the version for a partition, where we use the particular partition induced by the random variable in question.

Theorem 7.6 (Bayes' Rule for Random Variables)**.**

$$f_{X|Y}(x|y_0) = \frac{f_X(x) f_{Y|X}(y_0|x)}{f_Y(y_0)} \tag{7.1}$$

$$= \frac{f_X(x) f_{Y|X}(y_0)}{\sum_{x' \in R_X} f_X(x') f_{Y|X}(y_0|x)} \tag{7.2}$$

Both of these theorems follow directly from their counterparts with individual events, since once we fix x and y, the PMFs are just giving us event probabilities.

Example: Inferring Whether a Die is Unbalanced You are the proud owner of a die factory. You have ten machines, nine of which produce perfectly balanced dice; but one of which has been discovered to produce off-balance dice. Unfortunately, before this was discovered, the boxes of dice from the different machines got all mixed together. You want to stop the faulty dice from hitting the stores, so you conduct an experiment.

Let Y be a random variable indicating whether or not a die came from the faulty machine. Since we don't care about distinguishing among the nine good machines, we'll just set $R_Y = \{0, 1\}$, with 1 representing the bad machine. If the machines produce equal numbers and we randomly choose a die, then

$$Y \sim \mathcal{B}\text{ern}(0.1)$$

Now let X denote the value of a die roll. We have the following conditional probabilities:

x	$f_{X\|Y}(x\|0)$	$f_{X\|Y}(x\|1)$
1	$1/6$	$3/12$
2	$1/6$	$2/12$
3	$1/6$	$2/12$
4	$1/6$	$2/12$
5	$1/6$	$2/12$
6	$1/6$	$1/12$

Question: If we roll a 1, what's the probability that the die is off-balance?

Solution We use Bayes' rule to compute

$$\begin{aligned} f_{Y|X}(1|1) &= \frac{f_Y(1)f_{X|Y}(1|1)}{f_X(1)} \\ &= \frac{f_Y(1)f_{X|Y}(1|1)}{\sum_{y=0}^{1} f_Y(y)f_{X|Y}(1|y)} \\ &= \frac{0.1 \cdot 3/12}{0.1 \cdot 3/12 + 0.9 \cdot 2/12} \\ &= \frac{3/120}{3/120 + 18/120} \\ &= 1/7 \end{aligned}$$

Note that, unlike in the homework problem where the prior distribution over the dice was uniform, here only 1/10 of the dice are faulty, and so, having observed an outcome favored by the faulty die, while the posterior probability of a faulty die is higher than the prior probability, the prior is still strong enough that it is still more likely than not that the die is a fair one.

Now suppose there are two faulty machines, but one weights the dice in the opposite direction. Let's let $R_Y = \{0, 1, 2\}$, with

y	$f_Y(y)$
0	0.8
1	0.1
2	0.1

and

x	$f_{X\|Y}(x\|0)$	$f_{X\|Y}(x\|1)$	$f_{X\|Y}(x\|2)$
1	$2/12$	$3/12$	$1/12$
2	$2/12$	$2/12$	$2/12$
3	$2/12$	$2/12$	$2/12$
4	$2/12$	$2/12$	$2/12$
5	$2/12$	$2/12$	$2/12$
6	$2/12$	$1/12$	$3/12$

Now, in order to compute the conditional distribution of Y given that $X = 1$ (i.e., given that we roll a 1), we're going to need to sum 3 terms in the denominator. But once we compute these three terms, we have our numerators for each of the three probabilities we're after:

$$f_{Y|X}(y|1) = \frac{f_{X,Y}(1,y)}{f_X(1)} = \frac{f_{X,Y}(1,y)}{\sum_{y'=0}^{2} f_{X,Y}(1,y')}$$

where

$$f_{X,Y}(1,y') = f_Y(y') f_{X|Y}(1|y') \qquad \text{for each } y' \in \{0,1,2\}$$

When we put this all together, we get Bayes Theorem. For each $y \in \{0,1,2\}$:

$$f_{Y|X}(y|1) = \frac{f_Y(y) f_{X|Y}(1|y)}{\sum_{y'=0}^{2} f_Y(y') f_{X|Y}(1|y')}$$

If we compute each term in the sum in the denominator, then we can quickly get $f_{Y|X}(y|1)$ for *each* y, just by plugging each one in to the numerator:

$$\begin{aligned} f_Y(0) f_{X|Y}(1|0) &= 0.8 \times \frac{2}{12} = \frac{16}{120} \\ f_Y(1) f_{X|Y}(1|1) &= 0.1 \times \frac{3}{12} = \frac{3}{120} \\ f_Y(2) f_{X|Y}(1|2) &= 0.1 \times \frac{1}{12} = \frac{1}{120} \end{aligned}$$

Then we get $f_X(1) = \sum_{y'=0}^{2} f_Y(y') f_{X|Y}(1) = \frac{20}{120}$, and

$$f_{Y|X}(0|1) = \frac{16/120}{20/120} = 0.80$$
$$f_{Y|X}(1|1) = \frac{3/120}{20/120} = 0.15$$
$$f_{Y|X}(2|1) = \frac{1/120}{20/120} = 0.05$$

Notice that $f_{Y|X}(0|1) = f_Y(0)$. In other words, the events $\{X = 1\}$ and $\{Y = 0\}$ are **independent**: rolling a 1 does not affect the probability that the die is fair. But does that mean that the result of the roll tells us nothing about which die was rolled?

What would happen if, instead of rolling a 1, we had rolled a 2? Now

$$f_Y(0) f_{X|Y}(2|0) = 0.8 \times \frac{2}{12} = \frac{16}{120}$$
$$f_Y(1) f_{X|Y}(2|1) = 0.1 \times \frac{2}{12} = \frac{2}{120}$$
$$f_Y(2) f_{X|Y}(2|2) = 0.1 \times \frac{2}{12} = \frac{2}{120}$$

Then we get

$$\begin{aligned} f_X(2) &= \sum_{y'=0}^{2} f_Y(y') f_{X|Y}(2|y') \\ &= 0.8 \cdot \frac{2}{12} + 0.1 \cdot \frac{2}{12} + 0.1 \cdot \frac{2}{12} \\ &= \frac{2}{12}(0.8 + 0.1 + 0.1) \\ &= \frac{2}{12} \cdot 1 = \frac{2}{12} \end{aligned}$$

and

$$f_{Y|X}(0|1) = \frac{16/120}{20/120} = 0.80$$
$$f_{Y|X}(1|1) = \frac{2/120}{20/120} = 0.10$$
$$f_{Y|X}(2|1) = \frac{2/120}{20/120} = 0.10$$

What do you notice?

Now, the *entire distribution* of Y is unchanged after conditioning on $\{X = 2\}$. In other words, the *entire random variable* Y is independent of the *event* $\{X = 2\}$. Learning that $\{X = 2\}$ provided (provode?) *no information* about Y, precisely because $\{X = 2\}$ had the same probability regardless of the value of Y (that's what let us factor out the $2/12$, leaving a term that had to sum to 1).

But, that doesn't mean our *whole experiment* was completely useless, since as we saw above, if we had rolled 1, we would get some information about Y.

7.2.2 Independence of Random Variables

Suppose none of the machines are faulty, and we have

y	$f_Y(y)$
0	0.8
1	0.1
2	0.1

and

x	$f_{X\|Y}(x\|0)$	$f_{X\|Y}(x\|1)$	$f_{X\|Y}(x\|2)$
1	$2/12$	$2/12$	$2/12$
2	$2/12$	$2/12$	$2/12$
3	$2/12$	$2/12$	$2/12$
4	$2/12$	$2/12$	$2/12$
5	$2/12$	$2/12$	$2/12$
6	$2/12$	$2/12$	$2/12$

Now what's going to happen if we roll the die and examine the conditional distribution of Y given the outcome of the roll?

Answer: Now, *every* outcome of X will leave the distribution of Y unchanged, since the conditional probability of *every* value of X given a value of Y does not depend on what value of Y that is. In other words, our experiment is completely useless as far as information about Y is concerned — no matter what we rolled, we would not learn anything about Y.

In situations like this, we say that X and Y are **independent**. This is a pretty strong statement: we need the following to be true.

Definition 7.4 (Independence of RVs)**.** *Let X and Y be discrete random variables with ranges R_X and R_Y, respectively. We say that X is **independent** of Y iff for* every $(x, y) \in R_X \times R_Y$,

$$f_{X|Y}(x|y) = f_X(x)$$

that is, iff

$$P(X = x|Y = y) = P(X = x)$$

In other words, every event of the form $\{X = x\}$ is independent of every event of the form $\{Y = y\}$.

Just as with independence of events, the definition is reversible from the definition of conditional probability, and we also have

Theorem 7.A (Independence of RVs, version 2)**.** *X and Y are independent iff for every $(x, y) \in R_X \times R_Y$,*

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

or, rewritten,

$$P(\{X = x\} \cap \{Y = y\}) = P(\{X = x\})P(\{Y = y\})$$

Proof. Homework. □

7.2.3 Properties of Independent Random Variables

Although independence of two random variables X and Y can be a bad thing, in that we can't learn anything about Y by observing X and vice versa, it does make a lot of calculations easier. In particular, if we want to investigate a random variable that is a function of X and Y, this is a lot simpler when X and Y are independent. Some useful facts are collected in the next theorem.

Theorem 7.7 (Properties of Independent Random Variables)**.** *If X and Y are independent random variables, then*
(a) $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$
(b) $\mathbb{C}\mathrm{ov}(X, Y) = 0$*, and if X and Y have nonzero variances, then* $\rho(X, Y) = 0$
(c) $\mathbb{V}\mathrm{ar}[X + Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]$

Proof of (a). We will assume in the proof that both ranges are finite, but nothing in the proof will require that to be true: the statement is just as true for RVs with infinite ranges, and even continuous ranges. As usual, denote the ranges of X and Y by R_X and R_Y, respectively. Then we have

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x\in R_X}\sum_{y\in R_Y} xyf_{X,Y}(x,y) && \text{(Def. of } \mathbb{E}[XY]\text{)}\\
&= \sum_x\sum_y xyf_X(x)f_Y(y) && \text{(By independence — Thm. 7.A)}\\
&= \sum_x xf_X(x)\sum_y yf_Y(y) && (x \text{ and } f_X(x) \text{ do not depend on } y)\\
&= \mathbb{E}[X]\,\mathbb{E}[Y] && \text{(Def. of } \mathbb{E}\text{)}
\end{aligned}$$

□

Proof of (b). This follows directly from Thm. 7.1, and part (a) of this theorem. We have

$$\begin{aligned}
\mathbb{C}\text{ov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\\
&= \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] && \text{(by (a) of this Thm)}\\
&= 0
\end{aligned}$$

Now, since $\rho(X,Y) = \mathbb{C}\text{ov}(X,Y)/\sqrt{\sigma_X^2\sigma_Y^2}$, by the fact that $\mathbb{C}\text{ov}(X,Y) = 0$ and the assumption that σ_X^2 and σ_Y^2 are both nonzero, $\rho(X,Y) = 0$. □

Proof of (c). This follows directly from Ex. 5.8(c) and part (b) of this Theorem.

$$\begin{aligned}
\mathbb{V}\text{ar}[X+Y] &= \mathbb{V}\text{ar}[X] + 2\mathbb{C}\text{ov}(X,Y) + \mathbb{V}\text{ar}[Y] && \text{(by Ex. 5.8(c))}\\
&= \mathbb{V}\text{ar}[X] + 2\cdot 0 + \mathbb{V}\text{ar}[Y] && \text{(by part (b) of this Thm)}\\
&= \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y]
\end{aligned}$$

□

Important Note!: We have proved that the above statements are *necessary* conditions for independence; that is, that *if* X and Y are independent, then statements (a), (b) and (c) are true. However, they are not *sufficient* conditions. That is, it may be that statements (a), (b) and (c) are true *without* X and Y being independent. You'll prove this for yourself by examining a counterexample in homework.

7.2.4 (*) Convolution

When X and Y are independent, we also have a relatively simple way to compute the probability law of $X + Y$. (We already know the mean and variance of $X + Y$ from previous theorems, but up to now we do not have the exact probabilities)

We define the **convolution** of the probability laws f_X and f_Y to be f_{X+Y}, the probability law of $X + Y$, when X and Y are independent.

Definition 7.5. *Let X and Y be independent random variables with probability laws p_X and p_Y, respectively, and let $W = X + Y$. Then the **convolution** of f_X and f_Y is a function defined as*

$$(f_X \star f_Y) = f_W = f_{X+Y}$$

where, for any $w \in \mathbb{R}_W$ we have

$$(f_X \star f_Y)(w) = P(W = w) = P(X + Y = w)$$

Theorem 7.8 (Convolution formula)**.** *Let X and Y be independent random variables with ranges R_X and R_Y, respectively. Then*

$$(f_X \star f_Y)(w) = \sum_{x \in R_X} f_X(x) f_Y(w - x)$$

Proof. For a particular value w, how can we have $X + Y = w$? Either $X = x_1$ and $Y = w - x_1$, or $X = x_2$ and $Y = w - x_2$, or, ... These are all disjoint possibilities (some of them may have zero probability, but that's okay — the empty set is disjoint with everything). So

$$\begin{aligned}
(f_X \star f_Y)(w) &= P(\{X = x_1\} \cap \{Y = w - x_1\}) + P(\{X = x_2\} \cap \{Y = w - x_2\}) + \dots \\
&\qquad\qquad \text{(Disjoint additivity)} \\
&= f_{X,Y}(x_1, w - x_1) + f_{X,Y}(x_2, w - x_2) \qquad \text{(Def. of } f_{X,Y}\text{)} \\
&= \sum_x f_{X,Y}(x, w - x) \\
&= \sum_x f_X(x) f_Y(w - x) \qquad \text{(by independence — Thm. 7.A)}
\end{aligned}$$

□

7.3 Independent and Identically Distributed (IID) Random Variables

7.3.1 I.I.D. Random Variables

We have seen cases where two or more random variables are just "copies" of each other (e.g., the numbers on two identical dice; the birthdays of two people sampled from a population, etc.). If X_1 and X_2 have the same range and the same PMF; that is, if we have

$$\begin{aligned} R_1 = R_2 =: R \\ f_{X_1}(x) = f_{X_2}(x) =: f(x) \qquad \text{for all } x \in R \end{aligned}$$

Then we say that X_1 and X_2 are **identically distributed**.

If, in addition, X_1 and X_2 are *independent*, then we say that X_1 and X_2 are **independent and identically distributed**, or I.I.D.

We can generalize this to an arbitrary number of random variables (say, n of them), $X_1, \ldots, X_n$. If all of these have the same range, R and the same PMF f, then they are identically distributed. If they are independent, then they are I.I.D.

Technically we haven't said what it means for a set of more than 2 random variables to be independent.

Definition 7.6 (Independence of Multiple Random Variables)**.** *Discrete random variables $X_1, X_2, \ldots, X_n$ are* **mutually independent** *if their joint PMF factors as a product of the individual PMFs:*

$$f_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

for every $(x_1, \ldots, x_n) \in R_1 \times \ldots R_n$ (that is, for every combination of values of the variables).

7.3.2 Sums and Means of I.I.D. Random Variables

In many cases we are interested in functions of the *sum* of I.I.D. random variables. (A particular case is the *mean* of I.I.D. variables, which, notice, is a deterministic (non-random) function of their sum)

Suppose $X_1, X_2, \ldots, X_n$ are I.I.D. with mean μ and variance σ^2. Notice that we don't need subscripts on the mean and variance: Since the random variables are identically distributed, they have the same mean and variance by definition. Since the X_is are random variables, so is their sum. Denote this new random variable by S:

$$S = X_1 + X_2 + \cdots + X_n$$

Question: What are the mean and variance of S?

Answer: We know (from an inductive extension of Thm. 6.5) that the expectation of a sum is the sum of the expectations; so we have

$$\mathbb{E}[S] = \mathbb{E}[X_1 + X_2 + \cdots + X_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \mu = n\mu$$

From a HW exercise, we know that *when two random variables are independent*, the variance of their sum is the sum of their variances. Using another homework exercise that says that if X, Y and Z are mutually independent, then $X + Y$ is independent of Z, we can extend this to n independent random variables: the variance of the sum is the sum of the variances. So we have

$$\mathbb{V}\mathrm{ar}[S] = \mathbb{V}\mathrm{ar}[X_1 + X_2 + \cdots + X_n] = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[X_i] = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

If we take S/n, we get the "sample mean" of the X_is. This is still a random variable: for one possible "sample", we get one sample mean; for another we get a different sample mean. We denote this random variable $\overline{X}$. We have

$$\begin{aligned}
\mathbb{E}\left[\overline{X}\right] &= \mathbb{E}\left[\frac{S}{n}\right] \\
&= \frac{1}{n}\mathbb{E}[S] && \text{(by Thm. 6.5)} \\
&= \frac{1}{n} n\mu \\
&= \mu
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\overline{X}\right] &= \mathbb{V}\text{ar}\left[\frac{S}{n}\right] \\
&= \frac{1}{n^2}\mathbb{V}\text{ar}\left[S\right] && \text{(by Thm 6.6)} \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

This captures the intuition that the more observations we have, the more "stable" the sample mean, $\overline{X}$ will be from one sample to the next; that is the closer $\overline{X}$ will tend to be to the true mean μ of the random variable; and hence the more reliable $\overline{X}$ becomes as an *estimator* of μ.

7.3.3 The Simple Random Walk

Imagine a really drunk guy staggering around on the sidewalk: he takes one step to the left, then one to the right, then two more to the left, maybe one to the right, etc. We might suppose that at each step, he's just as likely to move left as he is to move right, and what he does next doesn't depend on what he did before.

We can describe this kind of behavior with the **simple random walk**. This is one of the simplest cases of a **stochastic process**.

Each step is a random variable, which takes the value -1 (a step to the left) half the time, and $+1$ (a step to the right) the other half the time. The j^{th} step is captured by X_j, and $X_1, \ldots, X_n$ are I.I.D. We have

$$\begin{aligned}
f_X(-1) = f_X(1) &= 1/2 \\
f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) &= f_X(x_1) \cdots f_X(x_n)
\end{aligned}$$

The position of the drunkard after n steps is a random variable, which is described by S_n. If he takes the same number of steps left as right, then $S_n = X_1 + \ldots X_n = 0$, and he's right where he started. If he's taken more steps right, then S_n will be positive; if he's taken more steps left, then S_n is negative.

This same model can be used to describe a gambling game: Flip a coin. If it comes up heads, you win \$1. If it comes up tails, you lose \$1. Now S_n describes your total winnings after n flips.

In each case, the average position (or average winnings) is just the average result of a given step, which is 0. But the variance increases linearly with the number of steps: $\mathbb{V}\text{ar}[S_n] = n\sigma^2$ (here σ^2 turns out to be 1, as you should verify). This makes sense as well: the longer the guy staggers around, the harder it is to predict where he'll be.

7.3.4 The Biased Random Walk

We can generalize the simple random walk by saying that the probability that $X_j = 1$ can be any number $p \in [0, 1]$. Then $P(X_j = -1) = (1-p)$. If the X_js are still I.I.D., then we have a **biased random walk**. This can be used as a model for, say, roulette, where you bet on black every time. There, $p = 18/38$.

Exercise: Calculate the mean and variance of the position of the biased random walker (as a function of p) after n steps.

7.4 Binomial and Poisson Random Variables

7.4.1 The Binomial Distribution

Notice that in the biased random walk, the random variables took on one of two values: either -1 or 1, taking 1 with probability p. If instead of -1 and 1, they take on 0 or 1, then we just have $\{X_i\}_{i=1}^n \overset{i.i.d}{\sim} \mathcal{B}\text{ern}(p)$. In this case $Y = \sum_{i=1}^n X_i$ counts the number of 1s that occur. We say that Y has a **binomial distribution** with parameters n and p.

This is a model for any situation where we have a series of independent "trials", each of which can be a "success" or a "failure". The binomial distribution counts the number of successes in n trials.

We write

$$Y \sim \mathcal{B}\text{inom}((, n), p)$$

The PMF of a Binomial Random Variable

Theorem 7.9. *Let Y have a $\mathcal{B}\text{inom}((,n),p)$ distribution. The range of Y is then $\{0,1,\ldots,n\}$, and the PMF is*

$$f_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad \text{for each } y \in \{0,1,\ldots,n\}$$

Proof. Clearly if $X_1 = X_2 = \cdots = X_n = 0$, then $S = 0$. If we change a single X_j from 0 to 1 then S increases by 1. We can do this until $X_1 = X_2 = \cdots = X_n = 1$, at which point $Y = n$, and so the range of Y is $\{0,1,\ldots,n\}$ as stated.

$Y = y$ iff exactly y of the X_js are equal to 1. Then the remaining X_js are equal to 0. Let $i_1, i_2, \ldots, i_y$ be the indices of the random variables which are equal to 1, and $i_{y+1}, i_{y+2}, \ldots, i_n$ be the remaining indices. By independence of the X_js, we have

$$\begin{aligned} &P(\{X_{i_1} = 1\} \cap \cdots \cap \{X_{i_y} = 1\} \cap \{X_{i_{y+1}} = 0\} \cap \ldots \{X_{i_n} = 0\}) \\ &\quad = P(X_{i_1} = 1) \times \cdots \times P(X_{i_y} = 1) \times P(X_{i_{y+1}} = 0) \times \cdots \times P(X_{i_n} = 0) \\ &\quad = p^y (1-p)^y \end{aligned}$$

This is the probability that a *particular* choice of y trials are successes. If we choose a different subset of the random variables to be successes, this is a different, and disjoint event from the first, but has the same probability. So we have

$$P(Y = y) = p^y (1-p)^{n-y} + p^y (1-p)^{n-y} + \cdots + p^y (1-p)^{n-y}$$

with a term for each choice of exactly which X_js are equal to 1. How many terms are there? We have to choose y random variables out of a set of n, without replacement, to be equal to 1, with order irrelevant. So there are $\binom{n}{y}$ options, which means that

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

as stated in the Theorem. □

Perhaps you can see where the Binomial distribution got its name. Remember the Binomial Theorem?

$$(a+b)^n = \sum_{y=0}^{n} \binom{n}{r} a^y b^{n-y}$$

If we set $a = p$ and $b = 1 - p$, we get

$$1 = (p + 1 - p) = \sum_{y=0}^{n} \binom{n}{y} p^y (1-p)^{n-y}$$

which shows that the Binomial PMF is a valid PMF, as it sums to 1 over its range.

Mean and Variance of a Binomial Random Variable

Remember that we derived the Binomial random variable as the sum of n I.I.D. Bernoulli random variables, each with mean p and variance $p(1-p)$. We then get

$$\mathbb{E}[Y] = \mathbb{E}[X_1 + \cdots + X_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np$$

and, by independence,

$$\mathbb{V}\text{ar}[Y] = \mathbb{V}\text{ar}[X_1 + \cdots + X_n] = \sum_{i=1}^{n} \mathbb{V}\text{ar}[X_i] = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

So, if we conduct n trials and have a probability of p in succeeding with each one, then we *expect* about np successes in total. For example, if I try 1000 times and have a probability of 30% of succeeding each time, then I expect to have about $1000 \times (0.3) = 300$ successes.

7.4.2 The Poisson Distribution

Suppose you're fishing for carp in one of the many Tucson area lakes. Suppose in this particular lake, the expected number of fish caught is 3 every 5 hours. Of course, if we fish for 5 hours, we won't always catch exactly 3 fish. How could we model the distribution of fish caught in a 5 hour fishing expedition?

If we let X_j represent whether or not you catch a fish in the j^{th} hour, then we could let $X_1, \ldots, X_5$ be I.I.D. Bernoulli random variables with $p = 0.6$. Then $\mathbb{E}[X_j] = 0.6$, and $S_5 = \sum_{j=1}^{5} X_j$ has a $\mathcal{B}\text{inom}((,5), 0.2)$ distribution. So, $\mathbb{E}[S_5] = 5 \cdot 0.6 = 3$, as stated.

Is there anything strange about this choice of model?

One obvious objection is that we're saying that in any given hour, I can catch either 0 or 1 fish. As a result, in the 5 hour period, we're saying that it's impossible that I will catch more than 5 fish.

Instead, I could use 10 Bernoulli variables; one for each *half*-hour. Then S_{10} is the total number of fish in 10 half-hour periods (or, a total of 5 hours). But to come out with the right expected value, now I need to set $p = 0.3$:

$$\begin{aligned} & \{X_j\}_{j=1}^{10} \overset{i.i.d.}{\mathcal{B}\text{ern}}(0.3) \\ \implies & S_{10} \sim \mathcal{B}\text{inom}((,1)0, 0.3) \\ \implies & \mathbb{E}[S_{10}] \sim 10 \cdot 0.3 = 3 \end{aligned}$$

This is a bit better, as now I'm allowing that I might catch up to 10 fish. But this is still rather arbitrary: why should my choice of time-window matter for the distribution of the number of fish I expect to catch?

A typical solution to this type of problem is to let the time window become infinitesimal, so that it really doesn't make sense to talk about more than one "event" occurring in a single infinitesimal interval. Then, to model the number of events occurring in a large window (say, 5 hours), we can think about the number of windows (n) running off to infinity, while the probability of an event in any particular window (p) adjusts downward to make np stay constant.

So in our example, we can think about chopping up our 5 hour block into 5 pieces ($n = 5$ and $p = 0.6$), then 10 pieces ($n = 10$ and $p = 0.3$), then 20 ($n = 20, p = 0.15$), 100 ($n = 100, p = 0.03$), a million ($n = 10^6, p = 3 \times 10^{-6}$), etc.

Obviously the binomial probabilites get pretty difficult to deal with after awhile: imagine having to compute 1 million factorial! Fortunately, as $n \to \infty$ everything stabilizes, and we get an "asymptotic" distribution called the **Poisson distribution**.

7.4.3 Poisson PMF

How can we calculate the probability of y events occurring in an interval under the Poisson distribution? Remember, we get the Poisson by starting with a Binomial, and letting $n \to \infty$, while p compensates to keep np constant. Let's let $\lambda = np$,

representing the idea that this is a constant. Then we can write the $\mathcal{B}\text{inom}((, n), p)$ PMF as

$$\begin{aligned}
f_{Y_n}(y) &= \binom{n}{y} p^y (1-p)^{n-y} \\
&= \binom{n}{y} \left(\frac{np}{n}\right)^y \left(1 - \frac{np}{n}\right)^{n-y} \\
&= \binom{n}{y} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \\
&= \frac{n(n-1)\cdots(n-y+1)}{y!} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \\
&= \left(\frac{n(n-1)\cdots(n-y+1)}{n^y}\right) \left(\frac{\lambda^y}{y!}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-y}
\end{aligned}$$

Now consider what happens as n increases to ∞. The first term in parentheses has a polynomial in n of order y in both the numerator and denominator. As n goes to infinity, the lower order terms (i.e., those with exponents less than y) become much, much smaller than the y^{th} power terms. Hence the whole ratio will go to 1.

The last term in parentheses has a fixed exponent, and so as n goes to ∞, $\frac{\lambda}{n}$ goes to 0, leaving a 1 in the parentheses. So this term goes away as well.

That leaves

$$\begin{aligned}
\lim_{n\to\infty} f_{Y_n}(y) &= \lim_{n\to\infty} \left(\frac{\lambda^r}{y!}\right) \left(1 - \frac{\lambda}{n}\right)^n \\
&= \left(\frac{\lambda^y}{y!}\right) \lim_{n\to\infty} \left(1 + \frac{-\lambda}{n}\right)^n
\end{aligned}$$

It turns out that the limit of the last term is $e^{-\lambda}$. So, we have

$$\lim_{n\to\infty} p_{Y_n}(y) = \frac{\lambda^y}{y!} e^{-\lambda}$$

The end result is the PMF of the Poisson distribution, whose only parameter is the "rate", λ. If a random variable Y has this distribution, we will write

$$Y \sim \mathcal{P}\text{ois}(()\lambda)$$

7.4.4 Poission Properties

Validity as a PMF

How do we know that the Poisson pmf is a valid PMF? Clearly it takes nonnegative values, so that much is okay, but we should check that it sums to 1 over its range.

What is its range? Remember, we got it as the limit of Binomials with n going to infinity. The $\mathcal{B}\text{inom}((, n), p)$ distribution had a range of $\{0, 1, \ldots, n\}$. As n goes to ∞, this becomes the set of all nonnegative integers.

So if we sum from 0 to infinity, we get

$$\begin{aligned}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} &= e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= e^{\lambda} e^{\lambda} \\ &= e^{\lambda - \lambda} \\ &= e^{0} \\ &= 1\end{aligned}$$

where we have used the fact that $\sum_{y=0}^{\infty} \frac{\lambda^y}{y!}$ is the Taylor series for e^{λ} (don't worry if you don't know what that means — it's usually something you learn in Calc II or so).

Poisson Mean

What about the mean and variance? Again, we got the Poisson distribution as the limit of Binomials, each of which had mean $np = \lambda$. This was held constant as we took n to ∞, so we should expect that the Poisson distribution has a mean of λ as well.

Indeed, this is the case:

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y=0}^{\infty} y \frac{\lambda^r}{y!} e^{-\lambda} \\
&= e^{-\lambda}\left(0\frac{\lambda^0}{0!} + \sum_{y=1}^{\infty} y\frac{\lambda^r}{y!}\right) && \text{(Split off first term)} \\
&= e^{-\lambda}\sum_{y=1}^{\infty} y\frac{\lambda^r}{y!} \\
&= e^{-\lambda}\sum_{y=1}^{\infty} \frac{\lambda^y}{(y-1)!} && \text{(Properties of factorial)} \\
&= e^{-\lambda}\lambda\sum_{y=1}^{\infty} \frac{\lambda^{y-1}}{(y-1)!} && \text{(Distributive property)} \\
&= e^{-\lambda}\lambda\sum_{s=0}^{\infty} \frac{\lambda^s}{s!} && \text{(Set } s = y-1\text{)} \\
&= e^{-\lambda}\lambda e^{\lambda} && \text{(Taylor series again)} \\
&= \lambda
\end{aligned}$$

Poisson Variance

The $\mathcal{B}\text{inom}((, n), p)$ distribution has variance $np(1-p) = \lambda\left(1 - \frac{\lambda}{n}\right)$. We would expect the Poisson variance to be the limit of this quantity. Clearly as we take $n \to \infty$, the term in parentheses just goes to 1, and so we are left with λ. So the Poisson variance should be the same as its mean: namely λ. We can show this with a similar set of tricks as we used to prove that the mean was λ.

Let $Y \sim \mathcal{P}\text{ois}(()\lambda)$. Then

$$\begin{aligned}
\mathbb{V}\text{ar}[Y] &= \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2 && \text{(Thm 5.6(a))} \\
&= \mathbb{E}\left[Y^2\right] - \lambda^2
\end{aligned}$$

It's difficult to find $\mathbb{E}[Y^2]$ directly, but it turns out to be relatively easy to find

$\mathbb{E}[Y(Y-1)]$, which is equal to $\mathbb{E}[Y^2 - Y] = \mathbb{E}[Y^2] - \lambda$.

$$\begin{aligned}
\mathbb{E}[Y(Y-1)] &= \sum_{y=0}^{\infty} y(y-1)\frac{\lambda^y}{y!}e^{-\lambda} \\
&= e^{-\lambda}\left(0(0-1)\frac{\lambda^0}{0!} + 1(1-1)\frac{\lambda^1}{1!} + \sum_{y=2}^{\infty} y(y-1)\frac{\lambda^y}{y!}\right) \\
& \qquad \text{(Split off first two terms)} \\
&= e^{-\lambda}\sum_{y=2}^{\infty} y(y-1)\frac{\lambda^y}{y!} \\
&= e^{-\lambda}\sum_{y=1}^{\infty} \frac{\lambda^y}{(y-2)!} \qquad \text{(Properties of factorial)} \\
&= e^{-\lambda}\lambda^2\sum_{y=1}^{\infty} \frac{\lambda^{y-2}}{(y-2)!} \qquad \text{(Distributive property)} \\
&= e^{-\lambda}\lambda^2\sum_{s=0}^{\infty} \frac{\lambda^s}{s!} \qquad \text{(Set } s = y-2\text{)} \\
&= e^{-\lambda}\lambda^2 e^{\lambda} \qquad \text{(Taylor series again)} \\
&= \lambda^2
\end{aligned}$$

So we have

$$\mathbb{E}[Y(Y-1)] = \mathbb{E}\left[Y^2\right] - \lambda = \lambda^2$$

which means

$$\mathbb{E}\left[Y^2\right] = \lambda^2 + \lambda$$

which means

$$\begin{aligned}
\mathbb{V}\text{ar}[Y] &= \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2 \\
&= (\lambda^2 + \lambda) - \lambda^2 \\
&= \lambda
\end{aligned}$$

7.5 Exercises

1. Let X, Y and Z be random variables and a and b be real numbers. Show

(a) $\mathbb{C}\text{ov}(X, Y) = \mathbb{C}\text{ov}(Y, X)$

(b) $\mathbb{C}\text{ov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

(c) $\mathbb{C}\text{ov}(X, aY + bZ) = a\mathbb{C}\text{ov}(X, Y) + b\mathbb{C}\text{ov}(X, Z)$ (it may help you to break this into two parts, first showing that this holds when $b = 0$, and then when $a = b = 1$. The general result follows by combining these two cases.)

(d) $\mathbb{V}\text{ar}[X + Y] = \mathbb{V}\text{ar}[X] + 2\mathbb{C}\text{ov}(X, Y) + \mathbb{V}\text{ar}[Y]$ (Hint: write the left-hand side as a covariance between a variable and itself, and then apply a previous result from this problem. You do not need to write out any summations.)

2. Let X and Y be random variables. Show that $\rho(X, Y) = 0$ if and only if $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

3. (*/**) If $X_1, \ldots, X_n$ are random variables, then we can define a **random vector**, $X = (X_1, \ldots, X_n)$, whose range is the vector space $\mathbb{R}^n$. We define the variance (aka the **covariance matrix**), $\mathbf{\Sigma}$, of the random vector to be the $n \times n$ matrix whose i, j element is the covariance between X_i and X_j (and so the diagonal elements are the variances). This is of course a symmetric matrix. Consider an arbitrary linear combination of the X_is, $Y = a_1X_1 + \cdots + a_nX_n$. This is of course a random variable. If we define the constant vector $\mathbf{a} = (a_1, \ldots, a_n)$, then $Y = \mathbf{a}^T X$ can be thought of as the orthogonal projection of X onto the direction determined by $\mathbf{a}$.

   (a) It turns out that finding the variance of Y (aka the variance of X in the direction determined by $\mathbf{a}$) is easy: it is given by $\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}$ (which is equal to $\sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathbb{C}\text{ov}(X_i, X_j)$). Prove this using the results of Problem 1 (part (c) is especially useful).

   (b) Suppose $\mathbf{b}$ is another constant vector. Then

$$\mathbb{C}\text{ov}\left(\mathbf{a}^T X, \mathbf{b}^T X\right) = \mathbf{a}^T \Sigma \mathbf{b} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j \mathbb{C}\text{ov}(X_i, X_j)$$

   (c) Using these results show that if $\mathbf{A}$ is a matrix of constants, then the covariance matrix of the random vector given by $\mathbf{A}X$ is $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$.

4. Let X and Y be random variables whose joint distribution is given by the following table.

   (a) Calculate the mean and variance of Y.

		y				
		0	1	2	3	4
x	0	0.02	0.04	0.20	0.04	0.02
	1	0.02	0.12	0.10	0.12	0.02
	2	0.06	0.04	0.10	0.04	0.06

Table 7.1: Joint Distribution of X and Y. Values in each cell represent $f_{X,Y}(x, y)$.

(b) Calculate the conditional PMF, $f_{Y|x}$, for each value of X.

(c) Find the **conditional expectation** of Y given that $X = x$, for each value of X. We can write $\mathbb{E}[Y|0], \mathbb{E}[Y|1]$ and $\mathbb{E}[Y|2]$ to represent these values.

(d) Calculate the **conditional variance** of Y given that $X = x$, for each value of X. We can write $\mathbb{V}\text{ar}[Y|0], \mathbb{V}\text{ar}[Y|1]$ and $\mathbb{V}\text{ar}[Y|2]$ to represent these values.

(e) Notice that since X is a random variable, if we do *not* fix X, $\mathbb{E}[Y|X]$ is also a random variable (whose value depends on X), with range

$$\{\mathbb{E}[Y|0], \mathbb{E}[Y|1], \mathbb{E}[Y|2]\}$$

Similarly, $\mathbb{V}\text{ar}[Y|X]$ is a random variable (whose value depends on X) with range

$$\{\mathbb{V}\text{ar}[Y|0], \mathbb{V}\text{ar}[Y|1], \mathbb{V}\text{ar}[Y|2]\}$$

Find $\mathbb{E}[\mathbb{E}[Y|X]]$ and $\mathbb{E}[\mathbb{V}\text{ar}[Y|X]]$.

(f) Show that for any random variables, X and Y, we have $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$.

(g) (*) Show that for any random variables, X and Y

$$\mathbb{E}[\mathbb{V}\text{ar}[Y|X]] \leq \mathbb{V}\text{ar}[Y]$$

5. Prove: If X and Y are independent random variables, then

$$\mathbb{V}\text{ar}[X + Y] = \mathbb{V}\text{ar}[X - Y]$$

(Hint: Some of the results of problem 1 are useful here — remember that variance is equivalent to covariance between a variable and itself)

6. Let X and Y be random variables with ranges R_X and R_Y respectively. Let E and F be arbitrary sets such that $E \subset R_X$ and $F \subset R_Y$ (i.e., E represents a set of possible X values and F represents a set of possible Y values).

Prove: If X and Y are independent, then

$$P(\{X \in E\} \cap \{Y \in F\}) = P(X \in E)P(Y \in F)$$

(Hint: Remember that $\{X \in E\}$ is shorthand for $\{a \in U : X(a) \in E\}$. Use disjoint additivity to write each probability as a sum of individual PMF values)

7. Let S_n represent the position of a simple random walk after n steps. Find $P(S_3 = 0)$, $P(S_3 = -1)$ and $P(S_4 = 0)$.

8. Let X have a discrete uniform distribution with range $R_X = \{-2, -1, 1, 2\}$ and let $Y = X^2$, so that Y has range $R_Y = \{1, 4\}$.

   (a) Find the joint distribution of X and Y.

   (b) Show that X and Y are uncorrelated but not independent (Hint: to show that they are uncorrelated, use the equation for the covariance that relies on $\mathbb{E}[XY]$).

9. Let X have a Bernoulli distribution with parameter p and Y have a Bernoulli distribution with parameter q. Show that X and Y are independent if and only if they are uncorrelated (we have already shown that for *any* random variables, independence implies zero correlation, so the content of this problem is to show that *in the special case of two Bernoullis*, the implication runs the other way as well). (Hint: Write the covariance in terms of p and q using problem 1, and then use that to fill out the joint distribution table. Compare the result to the joint distribution table that assumes independence.)

10. Let $X_1, X_2, \ldots, X_n$ be i.i.d. steps in a random walk, such that each step has common PMF, f_X, with $f_X(1) = p$ and $f_X(-1) = 1 - p$. Define $S_n = X_1 + \cdots + X_n$, representing the position of the walk after n steps.

    (a) Find $\mathbb{E}[S_n]$.

    (b) Find $\mathbb{V}\text{ar}[S_n]$.

11. Fifteen items are selected from a production line, each of which has a probability of 0.3 of being defective. We can represent the state (defective of not) of the jth item using a Bernoulli random variable. Assuming these variables are independent, find the probability that

    (a) Exactly three items are defective

    (b) Strictly fewer than five items are defective

(c) Between two and six (inclusive) are defective

(d) Strictly more than four items are defective

(e) Strictly more than eleven items are good

12. Let $X \sim \mathcal{B}\text{inom}(n, 0.5)$, and let f_X be the PMF of X. Show that, for any x between 0 and n inclusive, we have

$$f_X(x) = f_X(n - x)$$

13. A sample of a radioactive material emits a certain type of particle at random intervals. Assume that the number of particles emitted in one second can be modeled by a Poisson distribution, such that the expected number emitted in one second is 0.7. In one second, what is the probability that

(a) Exactly one particle is emitted?

(b) More than three particles are emitted?

(c) Between one and four (inclusive) are emitted?

14. We know how to find $\mathbb{E}[X + Y]$ and and $\mathbb{V}\text{ar}[X + Y]$, but we have not yet seen how to find the PMF itself of the random variable $X + Y$. Let X and Y be arbitrary random variables, and define $Z = X + Y$. Consider what must happen for Z to take a particular value, z. If X has the value 0 (say), then $Z = z$ if and only if $Y = z - 0$. If $X = 1$ then $Z = z$ if $Y = z - 1$. And so on. In general,

$$f_{X,Z}(x, z) = f_X(x) f_{Z|X}(z|x) = f_X(x) f_{Y|X}(z - x|x)$$

Thus by the marginalization lemma,

$$f_Z(z) = \sum_{x \in R_X} f_X(x) f_{Z|X}(z|x) = \sum_{x \in R_X} f_X(x) f_{Y|X}(z - x|x)$$

(a) Suppose X_1 and X_2 are i.i.d. discrete uniform random variables with range $R = \{1, 2, 3, 4\}$. Derive the PMF of $X_1 + X_2$.

(b) (*) Suppose $X \sim \mathcal{B}\text{inom}(m, p)$ and $Y \sim \mathcal{B}\text{inom}(n, p)$ are independent and let $Z = X + Y$. Show that $Z \sim \mathcal{B}\text{inom}(m + n, p)$. (Hint: You will need to derive the relationship $\binom{m+n}{z} = \sum_{x=0}^{m} \binom{m}{x}\binom{n}{z-x}$ using a similar argument to the one we used to obtain the equation for $f_Z(z)$).

(c) Let $X_1 \sim \mathcal{P}\text{ois}(\lambda_1)$ and $X_2 \sim \mathcal{P}\text{ois}(\lambda_2)$ be independent. Show that if we define $Z = X_1 + X_2$, then $Z \sim \mathcal{P}\text{ois}(\lambda_1 + \lambda_2)$.

(d) Use induction and 14c to show that the sum of n independent Poisson random variables is Poisson with parameter equal to the sum of the original parameters.

Part III

Elementary Bayesian Inference

Chapter 8

Bayesian Inference About a Discrete Parameter

8.1 Inference in the Bayesian Universe

It is not uncommon to be in a situation where we have a good idea what *family* a random variable belongs to (e.g., binomial, Poisson), but we are not very sure what *parameter* the distribution has. For example, imagine you have developed a new drug to treat high cholesterol, but it has been observed that one of its side effects can be to cause Type II Diabetes. This is a severe side-effect, but is probably rare. Before your drug goes to market, you need some idea of how frequently it occurs, which is one of the purposes of a clinical trial.

For any given person using the drug, there is some (hopefully very small) chance, p, that they will develop diabetes. If n people use the drug, some number, call it Y, of them ($0 \leq Y \leq n$) will be afflicted. Assuming the individuals are independently sampled, the value Y can be modeled as the sum of i.i.d. Bernoulli trials, that is, as a Binomial distribution with parameters n and p. In this context we can interpret p as the "long run proportion" who experience the side effect out of all those who might use the drug.

In a clinical trial, we control the value of n, but p is unknown to us, and beyond our control; so we want to make an inference about it based on the data we collect from the trial. If we knew the value of p, then we could calculate the probability of any given y value, according to the Binomial PMF. But since we don't know p, we have

a set of *conditional* PMFs of the form $f(y|p)$, where there is a different one of these distributions for Y for each possible value of p.

In order to apply Bayes' theorem, we need to establish some prior beliefs about the unknown parameter, p. Even though p presumably has some fixed "true" value, we do not know what it is, and so the Bayesian approach is to treat it as a random variable whose distribution reflects the subjective **prior plausibility** we attach to any particular value. To avoid the inevitable confusion associated with naming a random variable P, let's use X to represent the unknown rate of diabetes among those taking the drug. We'll represent the **prior distribution** of X by its (unconditional) probability mass function, f_X, so that, as usual, for any hypothetical "rate of diabetes in drug users", x, $f_X(x)$ represents the prior plausibility that that's the right rate. Then, $f_{Y|X}(y|x)$ represents the conditional probability that y people get diabetes, *assuming* that the true long run rate is x.

8.1.1 The Bayesian Universe

The **Bayesian universe** for our clinical trial has two dimensions: the true rate of diabetes (represented by X) and the observed number of cases among those in the trial (represented by Y). We can represent this universe as a table, whose rows correspond to values of X (for now we will assume there are finitely many possibilities) and whose columns correspond to values of Y. Each cell in the table represents a pair (x, y) and has some probability, $f_{X,Y}(x, y)$, which we can find using the chain rule as

$$f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y|x)$$

Remember that f_X is our prior PMF and $f_{Y|X}$ is our Binomial PMF with p set to x.

Once the trial is complete, we get to observe Y; say that y_0 people out of the n we tested had the side-effect. However, we do not have direct access to X, and so it remains unknown. In the table representation of the universe, we know which *column* of the universe we are in, but not which row. Our task then is to compute degrees of **posterior plausibility** for each cell in the column (each corresponding to a different value of x). This is just a matter of applying Bayes' rule:

$$f_{X|Y}(x|y_0) = \frac{f_X(x) f_{Y|X}(y_0|x)}{f_Y(y_0)} = \frac{f_X(x) f_{Y|X}(y_0|x)}{\sum_{x' \in R_X} f_X(x') f_{Y|X}(y_0|x')}$$

(x_1, y_1)	(x_1, y_2)	...	(x_1, y_N)
(x_2, y_1)	(x_2, y_2)	...	(x_2, y_N)
...	...	...	...
(x_M, y_1)	(x_M, y_2)	...	(x_M, y_N)

$f_X(x_1)f_{Y\|X}(y_1\|x_1)$	$f_X(x_1)f_{Y\|X}(y_2\|x_1)$	...	$f_X(x_1)f_{Y\|X}(y_N\|x_1)$
$f_X(x_2)f_{Y\|X}(y_1\|x_2)$	$f_X(x_2)f_{Y\|X}(y_2\|x_2)$	...	$f_X(x_2)f_{Y\|X}(y_N\|x_2)$
...	...	...	...
$f_X(x_M)f_{Y\|X}(y_1\|x_M)$	$f_X(x_M)f_{Y\|X}(y_2\|x_M)$	...	$f_X(x_M)f_{Y\|X}(y_N\|x_M)$
$f_Y(y_1)$	$f_Y(y_2)$	...	$f_Y(y_N)$

Table 8.1: Top: A table representation of a Bayesian Universe involving unobservable variable X with range $\{x_1, x_2, \ldots, x_M\}$ and observable variable Y with range $\{y_1, y_2, \ldots, y_N\}$. Bottom: Representation of joint probabilities of elements of the Bayesian universe obtained from prior over X and conditional distribution of Y given X. The sums down each column represent marginal likelihoods of each y value.

8.2 Likelihood Functions

Notice the components of this expression. We have a **prior PMF**, $f_X(x)$, which is a probability distribution over x values. We have a conditional probability function $f_{Y|X}(y_0|x)$, giving us for each x the probability of the observed data, y_0. And we have a *marginal probability*, $f_Y(y_0)$, which gives us the overall (marginal) probability of the observed data, and is computed using the Law of Total Probability.

What is changing as we vary x? Well, f_X gives us a probability distribution over x values. The marginal f_Y, however, doesn't depend on x at all, and since we only care about its value at y_0, it acts like a constant for the purposes of our posterior distribution $f_{X|Y}$. Finally, consider $f_{Y|X}$. For each fixed x, $f_{Y|x}$ is a PMF over y values. However, we aren't varying y — it is fixed at y_0, since that's what we observed — we are varying x. So every time we change x, $f_{Y|X}(y_0|x)$ comes from a different PMF.

If instead of thinking about a collection of separate functions, $f_{Y|x_1}, f_{Y|x_2}$, and so on, we think about $f_{Y|X}$ as a single function of both x and y, then by fixing the value of y at y_0, we end up with a function of just x. This resulting function is no longer a PMF (it will not sum to 1, for example), but it is still a function that returns probabilities. We call this object a **likelihood function**, as it tells us how likely our data would be, under each of a set of hypotheses.

Definition 8.1 (Likelihood Function)**.** *Let X and Y be random variables with ranges R_X and R_Y, and let $f_{Y|X}(y|x) = P(Y = y|X = x)$ for each $(x, y) \in R_X \times R_Y$. Suppose Y is observed and has the value y_0. Then the function $f_{y_0|X} : R_X \to [0, 1]$, defined as*

$$f_{y_0|X}(x) = f_{Y|X}(y_0|x) = P(Y = y_0|X = x)$$

is called the likelihood function*, and is sometimes written $\mathcal{L}(x|y_0)$. We will stick to the notation $f_{Y|X}(y_0|x)$, but remember that when this expression appears in the context of Bayes' Theorem, it is used as a function of the conditioning variable X, not of the conditioned variable Y.*

Notice that the observed data influences the posterior *only* through the likelihood function: the only property of a hypothesis that affects its change in plausibility from prior to posterior is how likely it makes the data.

In our clinical trial example, suppose there were $n = 1000$ people who took the drug, and $Y = 2$ who developed Type II diabetes. Suppose we have four hypotheses about the "long-run" diabetes rate: (1) $X = 0.0001$, (2) $X = 0.001$, (3) $X = 0.01$ and (4) $X = 0.1$, which we assign equal prior probabilities. Then our likelihood function has the following values (obtained from the binomial PMF):

$$f_{Y|X}(2|0.0001) = \binom{1000}{2} 0.0001^2 0.9999^{998} \approx 0.0045$$

$$f_{Y|X}(2|0.001) = \binom{1000}{2} 0.001^2 0.999^{998} \approx 0.1840$$

$$f_{Y|X}(2|0.01) = \binom{1000}{2} 0.01^2 0.99^{998} \approx 0.0022$$

$$f_{Y|X}(2|0.1) = \binom{1000}{2} 0.1^2 0.9^{998} \approx 0$$

This obviously does not sum to 1, but remember, it is not supposed to, since it consists of values from different conditional PMFs and is not itself a PMF.

8.3 The Marginal Likelihood

Having defined the prior distribution and the likelihood function, the only component that remains in determining the posterior distribution is the **marginal likelihood**,

$f_Y(y_0)$, or the overall probability of the data, averaging over hypotheses. Since its computation requires summing over all possible hypotheses, it appears to be the most cumbersome component of calculating the posterior distribution. However, since it does not depend on x, it only needs to be computed once; and since the terms in the sum are the same ones that appear in the numerator of each posterior probability, we are doing most of the computation already anyway.

Recall that in the "Bayesian Universe" table layout, the data tells us which column of the universe we are in, and Bayes' theorem is used to find the conditional probabilities associated with each cell in that column. But since the numerator of Bayes' Theorem is just a joint probability, $f_{X,Y}(x, y)$, we obtain the posterior probability of the cell (i.e., conditioned on being in the column) by simply *renormalizing* the set of joint probabilities so that they sum to 1; that is, by dividing by their sum. This does not change the *relative* magnitudes of the probabilities; it just ensures that collectively they form a valid probability distribution.

Notice also that if we do not need the full posterior distribution but only need to know which hypothesis is the most likely given the data, then we do not need to calculate the marginal likelihood at all, since the highest posterior probability is also the highest joint probability (within the restricted universe represented by the column corresponding to the observed y value).

In our clinical trial example, we can find the marginal likelihood by taking a weighted average of the values of the likelihood function, where the weights come from the prior. We have

$$
\begin{aligned}
f_Y(y_0) &= \sum_{x \in R_X} f_X(x) f_{Y|X}(y_0|x) \\
f_Y(2) &= f_X(0.0001) f_{Y|X}(2|0.0001) + f_X(0.001) f_{Y|X}(2|0.001) \\
&\qquad + f_X(0.01) f_{Y|X}(2|0.01) + f_X(0.1) f_{Y|X}(2|0.1) \\
&\approx 0.25 \cdot 0.0045 + 0.25 \cdot 0.1840 + 0.25 \cdot 0.0022 + 0.25 \cdot 0 \\
&\approx 0.0011 + 0.0460 + 0.0006 + 0 \qquad\qquad \approx 0.0477
\end{aligned}
$$

We can then find the posterior probabilities by dividing each joint probability by the

x	Prior, $f_X(x)$	Likelihood, $f_{Y\|X}(2\|x)$	Joint, $f_{X,Y}(x,2)$	Posterior, $f_{X\|Y}(x\|2)$
0.0001	0.25	0.0045	0.0011	0.024
0.001	0.25	0.1840	0.0460	0.965
0.01	0.25	0.0022	0.0006	0.024
0.1	0.25	0	0	0
Sum	1.00	—	0.0477	1

Table 8.2: Posterior calculation for the drug side-effect example

marginal likelihood

$$f_{X|Y}(0.0001|2) = \frac{f_X(0.0001)f_{Y|X}(2|0.0001)}{f_Y(2)} \approx \frac{0.0011}{0.0477} = 0.024$$
$$f_{X|Y}(0.001|2) = \frac{f_X(0.001)f_{Y|X}(2|0.001)}{f_Y(2)} \approx \frac{0.0460}{0.0477} = 0.965$$
$$f_{X|Y}(0.01|2) = \frac{f_X(0.01)f_{Y|X}(2|0.01)}{f_Y(2)} \approx \frac{0.0011}{0.0477} = 0.024$$
$$f_{X|Y}(0.1|2) = \frac{f_X(0.1)f_{Y|X}(2|0.1)}{f_Y(2)} \approx \frac{0}{0.0477} = 0$$

8.4 Example: Hypergeometric Distribution

Let's consider another simple example. Suppose we have a bag of candy containing 6 mini chocolate bars, some of which are dark chocolate and some of which are milk chocolate, but we don't know how many there are of each variety. We gain information by pulling a piece of candy out of the bag and observing its color (obviously if we just did this six times without replacement we would know exactly what was in the bag, but you should have in mind a situation with a very large, sometimes even infinite, population of potential observations, where it is only practical to observe a small proportion).

We can set X to be the number of dark chocolates in the bag (with six total, we have $R_X = \{0, 1, 2, 3, 4, 5, 6\}$), and Y to be the number of dark chocolates in our sample (for now, let's suppose our sample consists of only one candy so that $R_Y = \{0, 1\}$). If we don't have much background information about where the bag came from, a reasonable thing to do might be to put a discrete uniform prior on X, which amounts

to saying that we think that the number of dark chocolates is equally likely to be 0, 1, 2, and so on, *a priori*. So, since there are seven possibilities, the prior distribution over x is

$$f_X(0) = f_X(1) = f_X(2) = f_X(3) = f_X(4) = f_X(5) = f_X(6) = \frac{1}{7}$$

What does the likelihood function, $f_{Y|X}(y|x)$, look like? Well, if there are x dark chocolates in the whole bag, then there are $6 - x$ milk chocolates, so assuming that every candy in the bag has an equal chance of being picked, the probability of choosing a dark chocolate (i.e., $f_{Y|X}(1|x)$) is $\frac{x}{6}$. Similarly, the probability of choosing a milk chocolate is $f_{Y|X}(0|x) = 1 - \frac{x}{6}$.

If the candy we draw is dark chocolate, then writing out our likelihood function, we get

$$f_{Y|X}(1|0) = 0 \qquad f_{Y|X}(1|1) = \frac{1}{6} \qquad f_{Y|X}(1|2) = \frac{2}{6} \qquad f_{Y|X}(1|3) = \frac{3}{6}$$
$$f_{Y|X}(1|4) = \frac{4}{6} \qquad f_{Y|X}(1|5) = \frac{5}{6} \qquad f_{Y|X}(1|6) = 1$$

Multiplying prior by likelihood for each x yields joint probabilities:

$$f_{X,Y}(0,1) = 0 \qquad f_{X,Y}(1,1) = \frac{1}{42} \qquad f_{X,Y}(2,1) = \frac{2}{42} \qquad f_{X,Y}(3,1) = \frac{3}{42}$$
$$f_{X,Y}(4,1) = \frac{4}{42} \qquad f_{X,Y}(5,1) = \frac{5}{42} \qquad f_{X,Y}(6,1) = \frac{6}{42}$$

That is, we have $f_{X,Y}(x,1) = \frac{x}{42}$. Finally, normalizing these joint probabilities by their sum yields the posterior distribution $f_{X|Y}(x|1) = \frac{x}{42}/\frac{21}{42} = \frac{x}{21}$. So the most likely situation is that all the candy in the bag is dark chocolate, which has posterior probability $\frac{6}{21}$; but a bag with 5 dark and 1 milk is almost as likely, with posterior probability $\frac{5}{21}$, etc.

Now suppose we draw a second candy, and it is also dark chocolate. We want to update our distribution over the number of dark chocolates in the original bag to reflect our new observation. You may recall from Exercise 5.13 that there are two equivalent ways of doing this: first, we could start over from our original prior, and compute a likelihood function for our complete sample. Alternatively, since we already have a posterior distribution given the first observation, we can just turn this into our new prior and just worry about the likelihood for the second observation.

x	Prior, $f_X(x)$	Likelihood, $f_{Y_1\|X}(1\|x)$	Joint, $f_{X,Y}(x,1)$	Posterior, $f_{X\|Y_1}(x\|1)$
0	1/7	0/6	0/42	0/21
1	1/7	1/6	1/42	1/21
2	1/7	2/6	2/42	2/21
3	1/7	3/6	3/42	3/21
4	1/7	4/6	4/42	4/21
5	1/7	5/6	5/42	5/21
6	1/7	6/6	6/42	6/21
Sum	1.00	—	21/42	1

Table 8.3: Posterior calculation for the chocolates example (first draw)

If we take the first approach, then the prior probability for each x value from 0 to 6 is $\frac{1}{7}$. The likelihood of drawing two dark chocolates without replacement if there were x in the original bag of six is then the probability that the first draw is dark, multiplied by the conditional probability that the second candy is dark chocolate, *given* that the first is dark. Let Y now represent the combined data in our two draws from the bag. If we draw two successive dark chocolates, then $Y = 2$. We have

$$f_{Y|X}(2|x) = \frac{x}{6}\frac{x-1}{5}$$

So then the joint probabilities of each hypothesis with the evidence become

$$f_{X,Y}(x,2) = \frac{1}{7} \cdot \frac{x(x-1)}{30} = \frac{x(x-1)}{210}$$

Summing these we get $f_Y(2) = \frac{0+0+2+6+12+20+30}{210} = \frac{70}{210}$, and so the posterior probabilities become

$$f_{X|Y}(x|2) = \frac{f_{X,Y}(x,2)}{f_Y(2)} = \frac{x(x-1)}{70}.$$

But this is a bit easier using the second (iterative) approach, provided we had already done the work to update the distribution after the first observation.

To clearly distinguish between the two different observations, let's define Y_1 to be 1 if the first observation is dark chocolate (0 otherwise) and Y_2 to be 1 if the second observation is dark (side question: are Y_1 and Y_2 i.i.d.?). Then as we calculated above, we have

$$f_{X|Y_1}(x|1) = \frac{x}{21}$$

The posterior over X given the first two observations together can be written as

$$f_{X|Y_1,Y2}(x|y_1,y_2) = \frac{f_{X|Y_1}(x|y_1)f_{Y_2|X,Y_1}(y_2|x,y_1)}{f_{Y_2|Y_1}(y_2|y_1)} = \frac{f_{X|Y_1}(x|y_1)f_{Y_2|X,Y_1}(y_2|x,y_1)}{\sum_{x'} f_{X|Y_1}(x'|y_1)f_{Y_2|X,Y_1}(y_2|x',y_1)}$$

where we have just conditioned everything on Y_1 and then otherwise written out Bayes' rule as normal (you can think of the set where $Y_1 = y_1$ as playing the role of the universe). Given that there were x dark chocolates in the bag to start and given that we have drawn a dark chocolate on our first pick (that is, that $Y_1 = 1$), the probability of drawing another one is

$$f_{Y_2|X,Y_1}(1|x,1) = \frac{x-1}{5}$$

Then the joint probability (of X and Y_2, given Y_1), $f_{X,Y_1,Y_2}(x,1,1)$ is

$$f_{X,Y_2|Y_1}(x,1,1) = f_{X|Y_1}(x|1)f_{Y_2|X,Y_1}(1|x,1) = \frac{x}{21}\frac{x-1}{5} = \frac{x(x-1)}{105}$$

We can find $f_{Y_2|Y_1}(1|1)$ by summing these joints over $x = 0,1,2,3,4,5,6$, to get

$$f_{Y_2|Y_1}(1|1) = \frac{0+0+2+6+12+20+30}{105} = \frac{70}{105}$$

and so dividing the joint by the marginal, we get the posterior,

$$f_{X|Y_1,Y_2}(x|1,1) = \frac{x(x-1)}{70}, x \in \{0,1,2,3,4,5,6\}$$

as before.

x	Prior, $f_{X\|Y_1}(x\|1)$	Likelihood, $f_{Y_2\|X,Y_1}(1\|x,1)$	Joint, $f_{X,Y_1,Y_2}(x,1,1)$	Posterior, $f_{X\|Y_1,Y_2}(x\|1,1)$
1	1/21	0/5	0/105	0/70
2	2/21	1/5	2/105	2/70
3	3/21	2/5	6/105	6/70
4	4/21	3/5	12/105	12/70
5	5/21	4/5	20/105	20/70
6	6/21	5/5	30/105	30/70
Sum	1.00	—	70/105	1

Table 8.4: Posterior calculation for the chocolates example (second draw)

8.5 Exercises

1. There is an urn containing 9 balls, which can be either green or red. The number of red balls in the urn is not known. One ball is drawn at random from the urn, and its color is observed.

   (a) What is the Bayesian universe of the experiment?

   (b) Let X be the number of red balls in the urn. Assume that all possible values of X from 0 to 9 are equally likely. Let $Y_1 = 1$ if the first ball drawn is red, and $Y_1 = 0$ otherwise. Create a table containing the joint distribution of X and Y_1.

   (c) Find the marginal distribution of Y_1, and put it in the table.

   (d) Suppose a red ball was drawn. Identify the reduced Bayesian universe.

   (e) Find the posterior PMF for X, given that $Y_1 = 1$, by filling in a table like the one we used for the example with the chocolates:

x	$f_X(x)$	$f_{Y\|X}(1\|x)$	$f_{X,Y}(1,x)$	$f_{X\|Y}(x\|1)$
0				
...				
9				
Sum		—		

2. Suppose that a second ball is drawn from the urn in the previous problem, without replacing the first. Let $Y_2 = 1$ if the second ball is red, and $Y_2 = 0$ if it is green. Use the posterior distribution of X from the previous question as the prior distribution for X. Suppose the second ball is green. Find the posterior distribution of X.

3. Suppose we look at the two draws from the urn (without replacement) as a single experiment. The results were first draw red, second draw green. Find the posterior distribution of X in one step, and verify that it is the same as the one you got in the previous problem.

4. You grab a coin from a jar in a magic shop, and flip it 20 times. You get 17 heads.

   (a) Naturally, you suspect the coin may not be a fair one. Graph the likelihood function for this data, varying the heads probability continuously from 0 to 1. (You can use a computer or a graphing calculator for this)

(b) You learn that 90% of the coins in the jar were fair, 5% were "trick coins" that come up heads 90% of the time, and 5% were trick coins that come up tails 90% of the time, and the rest are fair. Calculate the posterior probability that your coin is fair.

5. Three competing economic models use different equations to predict how often recessions occur. All three models agree that Y, the number of recessions per century, has a Poisson distribution, but they disagree on the expected number. Model A says that there will be 15 recessions on average in a century. Model B says the average number is 20. Model C says the average is 25. Assume that, *a priori*, you find all three theories equally plausible.

   (a) There were actually 20 recessions in the century in question. Graph the likelihood function for thie data, varying the Poisson parameter continuously from 0 to 100.

   (b) Compute the posterior plausibilities of the three models.

6. Imagine someone rolls a fair die with an unknown number of sides several times, and writes down the outcomes. Let X represent the number of sides on the die. Let Y_1 be the outcome of the first roll, Y_2 be the outcome of the second roll, etc.

   (a) Suppose the first roll is a 1 ($Y_1 = 1$). Assuming the rolls are independent, find $f_{Y_1|X}(1|4)$ and $f_{Y_1|X}(1|6)$. Now find the likelihood function, $f_{Y_1|X}(1|x)$, in general (as a function of x).

   (b) Now suppose the die is rolled n times, and the result is never above 4, so that it could have come either from the 4-sided or the 6-sided die. Let D represent the event that has occurred (for example, D might be $\{Y_1 = 1 \cap Y_2 = 1 \cap Y_3 = 4\}$ if $n = 3$). Calculate $P(D|X = x)$, as an expression containing n and x. (Note that, when the die is fair, it is not necessary to know the exact sequence of rolls to find this probability; only that no roll is above 4.)

   (c) Suppose you know that the die either has 4 or 6 sides, and you thought these were equally likely before you saw the data. Find the posterior distribution of X given that $Y_1 = 1$.

   (d) Find the posterior ratio, $P(X = 6|D)/P(X = 4|D)$ (where D is defined as in (b)) as an expression containing n (note that you can calculate this ratio without calculating either of the probabilities individually, since some

things will cancel out). Sketch a graph with n on the horizontal axis and the natural log of this ratio on the vertical axis (use a computer or graphing calculator).

(e) Intuitively, under the assumption that the die is six-sided, it becomes more and more of a "suspicious coincidence" the more times you roll if you never get a 5 or 6. Explain how the your graph supports this intuition.

Chapter 9

Parameter Estimation With a Continuous Prior

9.1 Motivation

In the last chapter we looked at the example of a clinical trial in which we were trying to estimate the rate of a particular side effect. We modeled the number of instances of this side effect in a sample of n people using a binomial distribution with known parameter n (the number of people), and unknown parameter p (the long-run rate of the side-effect in the broader population). Recall that our approach was to treat the parameter as a random variable, on which we placed a prior distribution, enabling us to condition on our observations to obtain a posterior distribution. We could then use the posterior to answer a variety of questions, such as "How likely is it that more than 1 person in 1000 who take the drug will develop the side-effect?".

However, at the time, we were forced into an unnatural simplification: namely that we pick a finite number of possibilities for p, so that we could assign prior probabilities to each one. This is unnatural because it means that we are asserting before we even begin that it is impossible for the true probability to be anything other than these few values. In practice, we may be able to live with this artificiality if we allow for enough different values that are close enough together (e.g., perhaps we only care to estimate the probability to the nearest 0.01), but there is a more principled solution: represent the probability with a **continuous** random variable, that can take *any* real number between 0 and 1.

But wait; if any real number is possible, how do we assign a probability to any one of them? And how do we take a sum over all real numbers in an interval?

We will see that these two concerns need to be addressed by treating continuous distributions a bit differently than discrete ones. First, we will need to replace the concept of a Probability Mass Function with a Probability Density Function (which will mean that we are no longer assigning probabilities to individual values, but instead will say how "dense" the probability mass is near a value), and we will need to replace sums with integrals. Other than that, though, most of what we know about discrete distributions (how conditional probability works, for example) will still hold.

9.2 Essentials of Continuous Probability

Unfortunately we can't cover continuous probability in the same depth that we covered discrete probability (partly because of time, partly because of technical obstacles), but since most of the concepts are analogous, hopefully a whirlwind "crash course" (i.e., a quick definition of important concepts without formal proofs) will be enough for our present purposes.

9.2.1 Motivation

We might define a continuous random variable as one whose range is uncountably infinite; that is, where there is no way to map the elements of the range to the counting numbers ($\mathbb{N}$). Or, we could be somewhat less flexible and say that a continuous random variable is one whose range is an interval in the real numbers (or perhaps $\mathbb{R}$ itself).

It turns out, however, that we want a definition which is a bit more restrictive still.

Definition 9.1 (Continuous Random Variable)**.** *A* ***continuous random variable*** *is a random variable whose cumulative distribution function (CDF) is continuous (in the usual sense of a continuous function).*

Remember that we originally defined a random variable as a function that takes an element of the universe and assigns to it a real number; so any random variable has a CDF. If X is a random variable with range R, then for any $x \in \mathbb{R}$, the CDF of X

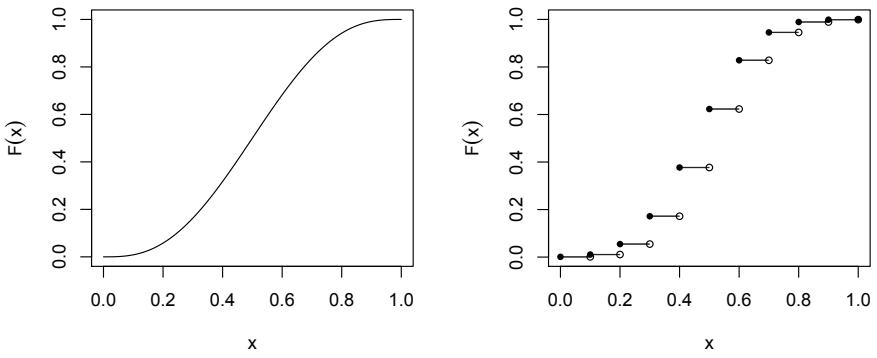

Figure 9.1: Examples of a continuous CDF (left) and a discrete CDF (right)

evaluated at x is given by $F(x) = P(X \leq x)$. Nothing in this definition requires any restriction on the nature of R, so it applies equally well to discrete and non-discrete random variables.

An example of a continuous CDF is in (e.g Fig. 9.1 (left)). Contrast this with the CDF of a discrete random variable, which has jumps at the points that have positive probability (e.g. Fig. 9.1 (right)).

A motivation for defining things this way is the following: Consider two half-open intervals, $(a, b]$ and $(a, b - \varepsilon]$, where ε is a small positive number. If I make ε small enough, then these two intervals are almost identical. The probability that X falls in the first interval these will be bigger, since it includes everything in the second interval, plus the outcomes where X is between $b - \varepsilon$ and b.

Notice that the difference between these probabilities is $P(b-\varepsilon < X \leq b)$. Intuitively, for a continuous random variable, this difference should be small, and ought to vanish completely as we make ε smaller and smaller. By the properties of the CDF, we can write this probability as $F(b) - F(b - \varepsilon)$:

$$\begin{aligned} P(b - \varepsilon < X \leq b) &= P((X \leq b) \cap \overline{X \leq b - \varepsilon}) \\ &= P(X \leq b) - P(X \leq b - \varepsilon) \qquad \text{(property } (P4)) \\ &= F(b) - F(b - \varepsilon) \end{aligned}$$

If the CDF is continuous, this difference will disappear as we shrink ε down to nothing, which is intuitively what we want from a continuous random variable.

Note that when the CDF is continuous, the random variable can never have positive probability at a single point. To see why, suppose $X = b$ with probability greater than zero. Then no matter how small we make ε, the event $a < X \leq b - \varepsilon$ does not contain b, and so we do not get any of the probability associated with the event $X = b$. Similarly, if $X = a$ with positive probability, then $a < X \leq b$ excludes $X = a$, but $a - \varepsilon < X \leq b$ includes it, no matter how small ε is, and so the probabilities of these two events will always differ by at least $P(X = a)$. And so, as we saw above, the CDF will not be continuous.

Lemma 9.1. *If X has a continuous CDF, then $P(X = b) = 0$, for every choice of b.*

Proof. By property (P4) of a probability measure, $P(X = b) := P(X \in \{b\}$ must be smaller than $P(X \in (b-\varepsilon, b]$, no matter what b and ε I choose, since $\{b\} \subset (b-\varepsilon, b]$. Also, by the properties of the CDF,

$$P(X \in (b - \varepsilon, b]) = F(b) - F(b - \varepsilon)$$

But since F is continuous, I can make this difference as close to zero as I want by making ε small enough. So, since $P(X = b)$ must be smaller than any arbitrarily small positive number, and it cannot be negative, as a probability, the only possibility is that $P(X = b) = 0$. □

An implication of this is that closed, open and half-open intervals with the same endpoints always have the same probability.

Lemma 9.2. *If X is continuous, then*

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) \tag{9.1}$$

Proof. Left for homework. Hint: represent intervals with closed ends as the disjoint union of an interval with an open end, and a singleton set consisting of the endpoint. Now apply the basic properties of probability measures and the fact that point probabilities are zero. □

9.2.2 Probability Density Functions

Motivation and Definition

If the probability of being equal to any given point is zero, the whole notion of a probability mass function is rendered completely useless, since it would have to be zero for every single value of the random variable. And yet, I can define two completely different continuous CDFs, which will give me two completely different continuous random variables. This means that, unlike discrete random variables, continuous random variables are not defined by the probabilities of single values.

We could get something like a PMF by chopping up the range of our random variable into intervals, and measuring the probability of falling in that interval. (This is a bit like the fishing example from Chapter 7, where we chopped up time into smaller and smaller intervals). If the intervals are small enough, then this would capture most of the information contained in the CDF; but not all of it, since I wouldn't know what the probabilities would be if I made my intervals even smaller.

If I keep making the intervals smaller and smaller, the probability of falling in that interval will shrink to zero. But if I balance out the shrinking interval by dividing the probability in the interval by the width of the interval, then I have some hope of converging to a meaningful number (think back to the derivation of the Poisson distribution, when we started with a binomial distribution, took n to infinity, and compensated by taking p to zero so that np stayed constant).

In physical settings, we define "density" as mass divided by volume. Here we do the same thing: start with the probability mass in an interval, divide by the "volume" of the interval (in this case its width), and that gives the average density in the interval. To get the density at a point, shrink the interval to zero and see what happens to the density.

What I want is a density function, f, which for each x in the range of X gives me the limiting density that I get by shrinking an interval down to that point. In other words:

$$\begin{aligned} f(x) &= \lim_{\varepsilon \downarrow 0} \frac{P(x-\varepsilon \leq X \leq x+\varepsilon)}{(x+\varepsilon)-(x-\varepsilon)} \\ &= \lim_{\varepsilon \downarrow 0} \frac{F(x+\varepsilon)-F(x-\varepsilon)}{2\varepsilon} \end{aligned} \tag{9.2}$$

This is exactly the derivative of F evaluated at the point x. One way to interpret

this quantity is that it captures how quickly F is increasing as it passes the point x; that is, as the slope of the line tangent to F at the value x.

We call this limit the **probability density** at x.

Definition 9.2 (Probability Density Function)**.** *If X is a continuous random variable with a differentiable CDF, its **probability density function** (or PDF), f, is the derivative of the CDF.*

Interpreting the density is similar to interpreting the PMF of a discrete random variable: the random variable is likely to take values where the density is high, and unlikely to take values where the density is low.

Properties of the PDF

Notice that since the the expression in (9.2) whose limit we're taking will always be nonnegative (it's the ratio of a probability and a positive number). Therefore, density is always nonnegative.

If we want to calculate the probability in some interval, we have already seen that we can just take the difference of the CDF values at the endpoints. But in many cases, we do not have an expression for the CDF; we only have an expression for the PDF. How can we calculate probabilities in that case?

Recall that with a discrete random variable, if we wanted the probability that it falls in some range, we just add up the probabilities of all the points in that range. We clearly can't add point probabilities in the continuous case, since they're all zero.

The idea is that, since the PDF is obtained by taking the derivative of the CDF, we can go backwards by taking the antiderivative. The difference between the CDF value at two points is just the integral of the PDF over that interval.

Lemma 9.3. *If X is a continuous random variable with PDF f, we have*

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{9.3}$$

Proof. This is a direct consequence of the Fundamental Theorem of Calculus, and the fact that interval probabilities are given by differences of the CDF. The antiderivative

of f is F, and so

$$\int_a^b f(x)dx = F(b) - F(a) = P(a \leq X \leq b)$$

□

Corollary 9.4. *For any continuous random variable X with range R and density f*

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Proof. By the previous lemma,

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \lim_{N\to\infty} \int_{-N}^{N} f(x)dx \\ &= \lim_{N\to\infty} P(-N \leq X \leq N) \\ &= P(-\infty < X < \infty) \\ &= 1\end{aligned}$$

□

A take-home message is that if we graph the density function, the probability associated with an interval is the area under the curve in that interval. The area under the *entire* curve is always 1.

An important implication of the fact that PDFs must always integrate to 1 is that if we know a PDF up to a constant multiple, then we can work out the exact PDF (that is to say, the unique value of the multiplicative constant) by enforcing that the resulting expression must integrate to 1.

Lemma 9.5. *For every nonnegative function g defined on the real line whose integral (over all of $\mathbb{R}$) is finite, there is a* unique *constant k (which does not depend on x), such that*

$$f(x) = k \cdot g(x)$$

is a valid PDF.

Proof. Suppose g is such a function, and define

$$c = \int_{-\infty}^{\infty} g(x)dx$$

Then by setting $f(x) = \frac{1}{c}g(x)$, we get

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} \frac{1}{c}g(x)dx \\ &= \frac{1}{c}\int_{-\infty}^{\infty} g(x)dx \\ &= \frac{1}{c} \cdot c \\ &= 1\end{aligned}$$

So f is a valid PDF. Clearly, multiplying g by any other constant will result in a function which does not integrate to 1, and so the resulting PDF is unique. □

9.2.3 Analogies With Discrete Distributions

There are basically two key correspondences between the discrete case and the continuous case. By making these two substitutions, almost everything we've learned about discrete distributions carries over to continuous distributions as well.

The main analogies are as follows:

Discrete Version	**Continuous Counterpart**
Probability Mass Function	Probablility Density Function
Summation	Integration

We've already seen one example of this: whereas a PMF must sum to one (over the entire range), a PDF must integrate to 1 (over the entire range).

The same analogy holds for expectation, variance, etc.

Expectation and Variance

Recall that for a discrete random variable, X, with range R and PMF f, we defined

$$\mathbb{E}[X] = \sum_{x \in R} xf(x)$$

For a continuous random variable with density f, we have

$$\mathbb{E}[X] = \int_R xf(x)dx \tag{9.4}$$

Similarly, if h is a real-valued function defined on R, in the discrete case we had

$$\mathbb{E}\left[h(X)\right] = \sum_{x \in R} h(x) f(x)$$

The continuous counterpart is

$$\mathbb{E}\left[h(X)\right] = \int_R h(x) f(x) dx \tag{9.5}$$

For the particular case where $h(x) = (x - \mathbb{E}\left[X\right])^2$, we get the variance. In the discrete case:

$$\mathbb{V}\mathrm{ar}\left[X\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \sum_{x \in R} (x - \mathbb{E}\left[X\right])^2 f(x)$$

In the continuous case:

$$\mathbb{V}\mathrm{ar}\left[X\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \int_R (x - \mathbb{E}\left[X\right])^2 f(x) dx \tag{9.6}$$

All the properties we proved about expectation and variance with discrete random variables hold for continuous random variables as well, for basically the same reasons (most of the operations we did to prove these in the discrete case involved pulling constants out of sums, or breaking one sum into multiple separate sums — both of these manipulations work with integrals too). For example:

$$\mathbb{E}\left[X + Y\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right] \tag{9.7}$$

$$\mathbb{E}\left[aX + b\right] = a\mathbb{E}\left[X\right] + b \tag{9.8}$$

$$\mathbb{V}\mathrm{ar}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 \tag{9.9}$$

$$\mathbb{V}\mathrm{ar}\left[X + a\right] = \mathbb{V}\mathrm{ar}\left[X\right] \tag{9.10}$$

$$\mathbb{V}\mathrm{ar}\left[aX\right] = a^2 \mathbb{V}\mathrm{ar}\left[X\right] \tag{9.11}$$

9.2.4 Examples of Continuous Random Variables

Uniform Distribution

The simplest continuous random variable is the one with a constant density within some interval, $[a, b]$, and zero elsewhere. This is the **uniform distribution** (Fig. 9.2).

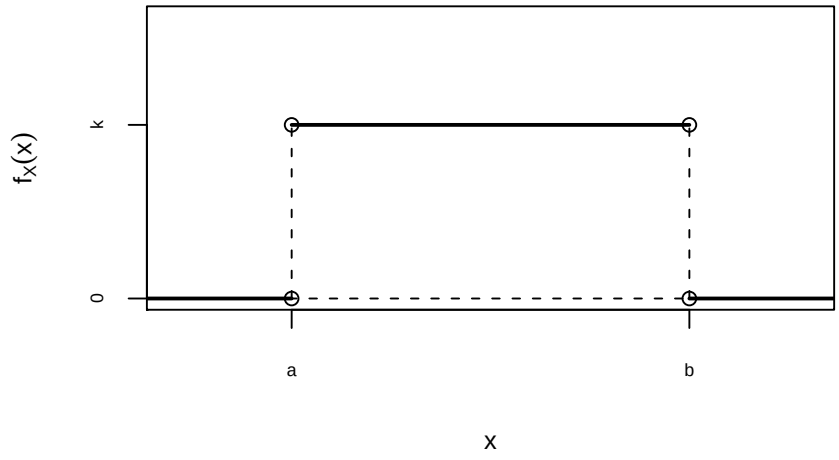

Figure 9.2: Density of a $\mathcal{U}(a, b)$ random variable. Notice that the density is technically undefined at the endpoints, since the CDF is not differentiable there. Notice also that for the area under the curve to be 1, we must have $k = \frac{1}{b-a}$.

Definition 9.3 ((Continuous) Uniform Distribution)**.** *A random variable X has a* ***continuous uniform distribution*** *with parameters a and b if its density is given by*

$$f(x) = \begin{cases} k & \text{for } x \in (a, b) \\ 0 & \text{for } x \notin [a, b] \end{cases} \tag{9.12}$$

for a constant k, chosen to make $\int_a^b f(x) = 1$.

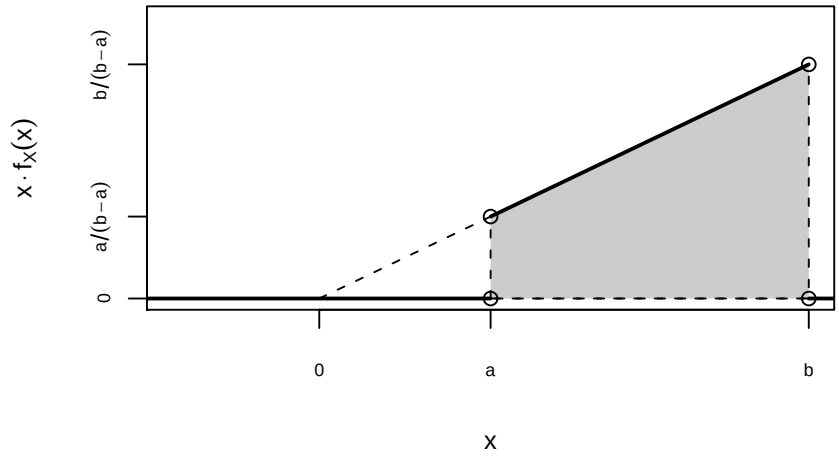

Figure 9.3: The function $x \cdot f(x)$ for a $\mathcal{U}(a, b)$ random variable. The area of the shaded region is the mean of the random variable.

If X has a uniform distribution on $[a, b]$, we can write $X \sim \mathcal{U}(a, b)$. The expected value is

$$\begin{aligned}\mathbb{E}[X] &= \int_a^b x \cdot f(x)\, dx = \int_a^b \frac{x}{b-a} dx \\ &= \frac{x^2}{2(b-a)}\Big|_{x=a}^{b} = \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}\end{aligned}$$

We could also have found this by noting that it is the area under the function $\frac{x}{b-a}$, from a to b. This function is drawn in Fig. 9.3.

Another way to calculate this is to notice that the mean is the area under the function $x \cdot f(x)$, which in this case is a simple polygon. In particular it is the difference between the areas of two triangles, which can be found in the standard way ($\frac{1}{2}$ Base $\times$ Height). You should verify that we get the same answer

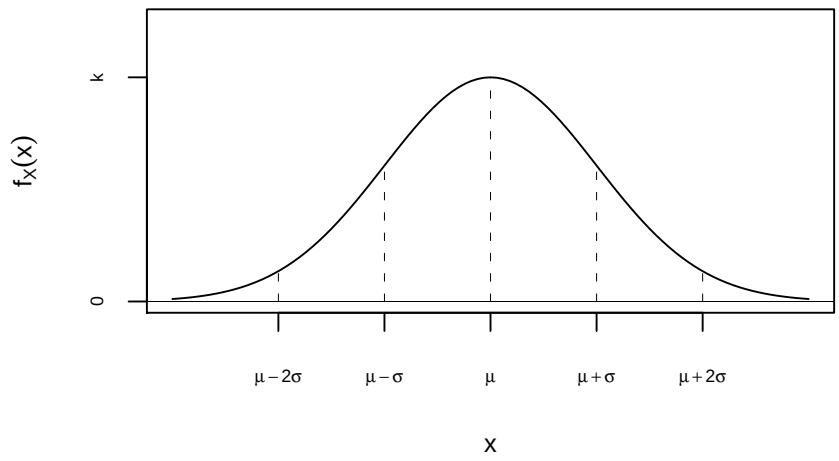

Figure 9.4: Density of the Normal distribution. The value of k turns out to be $\frac{1}{\sqrt{2\pi\sigma^2}}$.

Normal Distribution

The **normal distribution**, parameterized by its mean μ and its variance σ^2, has density defined on the whole real line, given by

$$f(x) = k \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{9.13}$$

It turns out that the required value of k is $\frac{1}{\sqrt{2\pi\sigma^2}}$. For a random variable with the above density, we write $X \sim \mathcal{N}(\mu, \sigma^2)$.

There is no algebraic expression for the CDF; however there's a multitude of fast algorithms out there to compute Normal probabilities numerically. These algorithms take advantage of the fact that we can transform X into a "standardized" version (known as a "standard normal" random variable with mean 0 and standard deviation 1, by setting $Z = (X - \mu)/\sigma$. Since this is a one-to-one and strictly increasing transformation, the CDF of X at the value x is equivalent to the CDF of Z at the corresponding value $z = (x - \mu)/\sigma$.

9.2.5 Joint and Conditional Densities

Joint Density

If I have two continuous random variables, X and Y, I can draw x and y axes, and ask about the probability of being in any particular region in the plane. In particular, if I draw a rectangle that goes from a to b in the x direction and from c to d in the y direction, then the probability of being in that rectangle is

$$P(\{a \leq X \leq b\} \cap \{c \leq Y \leq d\})$$

By the same logic as above, the probability at any single point will be zero. But we can do the same limit trick we did before to get the **joint density** at a point (x, y): take a sequence of rectangles that shrink down to the point, and compute the probability density in the rectangle, given by the probability mass divided by the "volume" (really area, this time) of the rectangle. The density $f_{X,Y}$ at the point (x, y) is the limit of these rectangle densities.

Equipped with the concept of a joint density, we get a bunch more analogies with discrete distributions:

$$\int_{R_X} f_{X,Y}(x, y)dx = f_Y(y) \tag{9.14}$$

$$\int_{R_Y} f_{X,Y}(x, y)dy = f_X(x) \tag{9.15}$$

$$\int_{R_X} \int_{R_Y} f_{X,Y}(x, y)dy\ dx = 1 \tag{9.16}$$

$$\mathbb{E}\left[h(X, Y)\right] = \int_{R_X} \int_{R_Y} h(x, y) f_{X,Y}(x, y)dy\ dx \tag{9.17}$$

$$\mathbb{C}\text{ov}\left(X, Y\right) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \tag{9.18}$$

Notice that in every case, the equation is the same as the counterpart for discrete random variables, except that we've replaced the probability mass $f_{X,Y}(x, y)$ with the density $f_{X,Y}(x, y)$, and we've replaced sums by integrals.

Conditional Density

We also have conditional density, which is defined in the same way as in the discrete case.

Definition 9.4 (Conditional Density)**.** *For continuous random variables X and Y, the conditional density at $Y = y$ given that $X = x$ is defined as*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{9.19}$$

We also get the multiplication rule (which we can easily extend by induction to the general chain rule)

$$f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y) \tag{9.20}$$

Joint Mass-Density Functions

What happens if, say, X is continuous (with density f_X) and Y is discrete (with PMF f_Y)? Conditioned on any particular value of Y, the distribution of X will still be continuous with density $f_{X|Y}$ that integrates to 1; similarly, conditioned on any particular value of X, the distribution of Y will be discrete with PMF $f_{Y|X}$, which sums to 1. It stands to reason that we should have a joint "density-mass" function, $f_{X,Y}$, defined as

$$f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y) \tag{9.21}$$

and that

$$\int_{R_X} f_{X,Y}(x,y)dx = f_Y(y) \tag{9.22}$$

$$\sum_{y \in R_Y} f_{X,Y}(x,y) = f_X(x) \tag{9.23}$$

$$\sum_{y \in R_Y} \left(\int_{R_X} f_{X,Y}(x,y)dx \right) = \int_{R_X} \left(\sum_{y \in R_Y} f_{X,Y}(x,y) \right) dx = 1 \tag{9.24}$$

Independence

If X and Y are independent then for every (x, y),

$$f_{Y|X}(y|x) = f_Y(y) \tag{9.25}$$

$$f_{X|Y}(x|y) = f_x(x) \tag{9.26}$$

$$f_{x,y}(x,y) = f_X(x)f_Y(y) \tag{9.27}$$

(where f is a density function when the variable in question is continuous, and a mass function when it is discrete)

Bayes Rule

Bayes rule is the same for continuous random variables as well

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)} \tag{9.28}$$

$$= \frac{f_X(x)f_{Y|X}(y|x)}{\int_{R_X} f_{X,Y}(x',y)dx'} \tag{9.29}$$

$$= \frac{f_X(x)f_{Y|X}(y|x)}{\int_{R_X} f_X(x')f_{Y|X}(y|x')dx'} \tag{9.30}$$

$$\tag{9.31}$$

(where, again, if one variable is discrete, we interpret its f as a PMF)

Notice, though, that if we're interested in the conditional distribution of X given a particular value of Y, the function on the left-hand side of (9.28) is a function of x only; and so the denominator on the right is a constant (since we have averaged over *all possible* x-values). So we can write

$$f_{X|Y}(x|y) = k_y \cdot f_X(x)f_{Y|X}(y|x) \tag{9.32}$$

By Lemma 9.5, if we know $f_X(x)f_{Y|X}(y|x)$, then there is a unique value of k_y that makes $f_{X|Y}(x|y)$ a valid pdf. Note that since $f_{X|Y}$ is a density of X alone, k_y only needs to be a constant *with respect to* x. It does depend on y, and in fact is equal to $1/f(y)$.

If we leave off k_y and work with the non-constant part of the posterior density, we might get something that looks like the PDF for a known family of random variables (e.g., Normal), apart from a constant multiple term. When this happens, we know that there will be *some* constant, say, c_y, that we can put in to make the result an actual PDF in that family (Normal, say). Since any other constant would result in a function that did not integrate to 1 (and hence is not a valid PDF), we can conclude that the value of k_y that we had discarded *must be* equal to c_y, and the PDF that we have is in fact a member of the family in question.

By arguing this way, we avoid having to do any actual integration to find k_y. This happens when the prior distribution of X and the conditional distribution of $Y|X$ are "conjugate", which we define next.

9.3 Parameter Estimation With Conjugate Priors

9.3.1 Motivation

Recall our clinical trial example. A drug produces a certain side effect in some proportion, X, of people who take it. We don't know what that proportion is, so we want to estimate it by doing a clinical trial. In our trial, if n randomly selected individuals take the drug, then some number, Y, of them will experience the side-effect, where Y ranges from 0 to n.

The Bayesian approach we took to estimating X was to treat it as a random variable, so that the conditional distribution $Y|X = x$ has a $\mathcal{B}\text{inom}(n, x)$ distribution. Here, x represents the probability that any given person has the side-effect (or, equivalently, the "long run proportion" of people who will experience the side-effect).

To find a plausibility distribution over X values given the data from the trial, we first need to assign a prior to X. When we first introduced this example, we used a discrete prior over a few values that X might take. But this was artificial and unrealistic, since we were categorically ruling out in advance all the other potential values for X, which meant that if the true value were not in our discrete set, we could never discover the truth. We'd like a prior on X whose range is the full interval $[0, 1]$, so that nothing which is possible in principle is ruled out completely.

9.3.2 The Beta-Binomial Model

Let's start out by supposing that the prior density of X is the same everywhere on $[0, 1]$. What that means conceptually is that we're saying it's equally likely that the true value is between 0 and 0.001, as it is to be between 0.001 and 0.002, which in turn has the same prior probability of being between 0.002 and 0.003, etc. Moreover, *if* the true value is between 0 and 0.001, it is just as likely to be between 0 and 0.0001 as it is to be between 0.0001 and 0.0002, and so on.

In other words, X has a continuous uniform prior, and its density is just 1 on that interval. Conditioned on X, Y has a $\mathcal{B}\text{inom}(n, x)$ distribution. So, we have

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)} \\ &= \frac{1 \cdot \binom{n}{y} x^y (1-x)^{n-y}}{f_Y(y)} \end{aligned}$$

Remember, the result is a function of x, not a function of y, so we can absorb everything that doesn't have an x in it — here, the binomial coefficient, and the denominator — into a constant, k_y, leaving us with

$$f_{X|Y}(x|y) = k_y \cdot x^y (1-x)^{n-y}$$

Since this is a function of x and x is a continuous random variable, this is a probability *density* function.

If we could integrate this expression, we could find the value of k_y that made $f_{X|Y}$ a valid PDF. It's just a polynomial in x, so we could in principle multiply everything out and get a value analytically; but it's an ugly polynomial, and n and y might be quite big. Fortunately, this has the form of a known distribution, called the **Beta distribution**.

The Beta Family

Definition 9.5 (Beta Distribution)**.** *A random variable X with range $[0, 1]$ has a **beta distribution** with parameters a and b if its density is*

$$f(x) = k_{a,b} \cdot x^{a-1} (1-x)^{b-1} \tag{9.33}$$

where a and b are positive real numbers, and we've given k some subscripts to emphasize the fact that it's a constant as far as x is concerned, but it does depend on a and b.

The constant $k_{a,b}$ doesn't always correspond to any algebraic expression, but if a and b are both integers, then

$$k_{a,b} = \frac{(a+b-1)!}{(a-1)!(b-1)!} = (a+b-1)\binom{a+b-2}{a-1}$$

In general, the constant is given in terms of the **gamma function**, which extends the idea of the factorial to non-integer values.

For a positive integer α, we have $\Gamma(\alpha) = (\alpha - 1)!$. For every positive real value, we have

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \tag{9.34}$$

In the case of the Beta family, We have

$$k_{a,b} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{9.35}$$

So, if we write out the constant, $k_{a,b}$, the Beta density is given by

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot x^{a-1}(1-x)^{b-1}. \tag{9.36}$$

Notice that if $a = b = 1$, then we get the $\mathcal{U}(0,1)$ distribution, since the exponents are zero, leaving a constant density. We know this without even working out the value of the constant, since there is only one possible constant density function on any given interval. (If we did work out the constant for $a = b = 1$ using the fact that $\Gamma(1) = (1-1)! = 1$ and $\Gamma(2) = (2-1)! = 1$, we would see that it is 1, as it must be).

Calculating the mean of a $\mathcal{B}$eta(a, b) distribution requires integrating a complicated expression; but there's a short-cut, which is similar to the one we used to find the mean and variance of the Poisson. We have

$$\begin{aligned}\mathbb{E}[X] &= \int_0^1 x \cdot k_{a,b} \cdot x^{a-1}(1-x)^{b-1}dx \\ &= k_{a,b}\int_0^1 x^{a+1-1}(1-x)^{b-1}dx\end{aligned}$$

Notice that the new expression in the integral is the non-constant part of a Beta distribution, if we set the first parameter to $a + 1$ and the second one to b. And so if we multiply what's inside the integral by the corresponding constant for the $Beta(a + 1, b)$ distribution, which we can call $k_{a+1,b}$, then the result will integrate to 1, and the integral will disappear. But in order to do this without changing the end result, we also need to multipy by the inverse of this constant, which we put outside the integral. So, we are left with an answer of $k_{a,b}/k_{a+1,b}$. If we write out

these constants, and do some algebra, we find that the result simplifies dramatically to $a/(a+b)$, as seen below:

$$\begin{aligned}
\mathbb{E}[X] &= k_{a,b}\int_0^1 x^{a+1-1}(1-x)^{b-1}dx \\
&= k_{a,b}\frac{1}{k_{a+1,b}}\int_0^1 k_{a+1,b}x^{a+1-1}(1-x)^{b-1}dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} \quad \text{(since the integral is 1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a+b+1)}\frac{\Gamma(a+1)}{\Gamma(a)} \\
&= \frac{\Gamma(a+b)}{(a+b)\Gamma(a+b)}\frac{a\Gamma(a)}{\Gamma(a)} \qquad \text{(by (9.34))} \\
&= \frac{a}{a+b}
\end{aligned}$$

So, back to our clinical trial example, we had found that

$$f_{X|Y}(x|y) = k_y \cdot x^y(1-x)^{n-y}$$

But notice that this has the form of a Beta distribution, with parameters $y+1$ and $n-y+1$. So, to sum up, if $Y|(X=x)$ has a $\mathcal{B}\text{inom}((,n),x)$ distribution, and we use a $\mathcal{U}01$ prior for X, then upon observing $Y=y$, the posterior distribution for X becomes $\mathcal{B}\text{eta}(y+1, n-y+1)$. In the next chapter we will discuss how we can take this estimate and come up with a principled way to find a single value to estimate X, but for now suppose we use the posterior expected value. Then our "best estimate" for the rate of the side effect in the population becomes

$$\hat{x} = \frac{y+1}{n-y+1}$$

What happened to the distribution of X when we did our Bayesian update? We started with a $\mathcal{U}01$ distribution, and ended up with a $\mathcal{B}\text{eta}(y+1, n-y+1)$ distribution. But notice that actually, the prior we started with was a special case of a Beta distribution, too: namely a $\mathcal{B}\text{eta}(1,1)$ distribution. If X has a $\mathcal{B}\text{eta}(1,1)$ prior, then its density is

$$f_X(x) = k_{1,1}\cdot x^{1-1}(1-x)^{1-1} = k\cdot 1$$

This is proportional to, and hence equal to, the $\mathcal{U}01$ distribution, which has constant density 1.

What happens if we start with a different Beta prior on X? In general, suppose $X \sim \mathcal{B}\text{eta}(a, b)$. Then if $Y|(X = x) \sim \mathcal{B}\text{inom}(n, x)$, the posterior density of X becomes

$$\begin{aligned} f_{X|Y}(x|y) &= k_y \cdot f_X(x) f_{Y|X}(y|x) \\ &= k_y \cdot x^{a-1}(1-x)^{b-1} \cdot \binom{n}{y} x^y (1-x)^{n-y} \\ &= k_{a,b,y} x^{a+y-1}(1-x)^{b+n-y-1}. \end{aligned}$$

Notice that the non-constant part corresponds to that of a Beta density, with parameters $a+y$ and $b+n-y$. So, since there is only one valid PDF which is proportional to a $\mathcal{B}\text{eta}(a+y, b+n-y)$ density, and that's the $\mathcal{B}\text{eta}(a+y, b+n-y)$ itself, we must have $X|(Y=y) \sim \mathcal{B}\text{eta}(a+y, b+n-y)$.

So we have just shown that if we start with *any* Beta prior on X, and update using data from a Binomial distribution (with parameters n and X), the resulting posterior is a Beta distribution. In other words, the Beta family is "closed under binomial updates". This important closure property is called **conjugacy**.

9.3.3 Conjugacy

In the example above, we started with a Beta prior and ended up with a Beta posterior after observing Binomially distributed data. This is because the Beta family is the **conjugate prior** family for the Binomial distribution.

Definition 9.6 (Conjugate Priors)**.** *Suppose X and Y are such that the distribution of Y comes from a family that has X as a parameter. A prior on X is said to be* **conjugate** *to that family if the resulting posterior distribution of $X|Y$, always comes from the same parametric family as the prior.*

So, for example a Beta prior is *conjugate* to a binomial likelihood, since the resulting posterior distribution is also a beta distribution.

This is particularly nice for making point estimates, since we can take expectations and variances, etc. easily (i.e., without having to compute any integrals) by using known properties of the parametric family. If we can choose parameters so that the prior has the properties we want (e.g., with a particular mean and variance, or

mode and quantiles, etc.), then we are making use of this information in our decision process.

Equivalent Data Interpretation

With some kinds of distributions, conjugate priors have an interesting property. In the Beta-Binomial model, if we start with a $\mathcal{B}\text{eta}(a, b)$ prior on X and get an observation of y "successes" from a $\mathcal{B}\text{inom}(n, x)$ distribution, the posterior on X is a $\mathcal{B}\text{eta}(a + y, b + n - y)$ distribution: that is, the first posterior parameter is the first prior parameter plus the number of successes, and the second posterior parameter is the second prior parameter plus the number of failures observed.

This gives an intuitive interpretation of the Beta parameters as characteristics of an "equivalent sample". The a parameter is simulating a prior "number of successes", while the b parameter simulates a prior "number of failures". To see this, imagine that a were reduced by 1 and the number of observed successes was increased by 1 (so that n becomes $n + 1$ and y becomes $y + 1$). The resulting posterior would be $\mathcal{B}\text{eta}(a - 1 + y + 1, b + n + 1 - (y + 1)$, or, after simplification, $\mathcal{B}\text{eta}(a + y, b + n - y)$; exactly as before! All we have done is traded a prior observation for a real observation. Similarly, we can trade units of b for observed failures without changing the prior. This illustrates the effect that the prior has on the posterior: it works as though you were "hallucinating" extra data.

Another way to look at the parameters of the Beta distribution is as follows. Instead of a and b, we might parameterize the distribution in terms of $a/(a + b)$ (i.e., a mean) and $(a + b)$. Together these determine a and b, and vice-versa. But here, our parameters have another intuitive interpretation: $a/(a + b)$ acts as a prior "point estimate" (it's the prior mean), and $(a + b)$ acts as an "equivalent sample size", or a measure of how much weight the prior should have compared to the data. As $a + b$ increases relative to n (the actual sample size), the prior has a bigger influence on the resulting posterior; as it decreases, the data "speaks louder".

If we examine the posterior mean, we can see that it is a weighted average of the prior mean and the sample proportion of successes, where the weights are the "equivalent sample size" and the real sample size, respectively. Again, let X have a $\mathcal{B}\text{eta}(a, b)$ prior and $Y|X$ have a $\mathcal{B}\text{inom}(n, x)$ distribution. Then the posterior, $X|Y$ has a $\mathcal{B}\text{eta}(a + y, b + n - y)$ distribution, with mean

$$\mathbb{E}\left[X|Y = y\right] = \frac{a + y}{a + y + b + n - y} = \frac{a + y}{a + b + n}$$

which, with some algebra, can be rewritten as

$$\begin{aligned}\mathbb{E}\left[X|Y=y\right] &= \frac{a+y}{a+b+n} \\ &= \frac{\frac{a}{a+b}\cdot(a+b)+\frac{y}{n}\cdot n}{(a+b)+n} \\ &= \frac{a}{a+b}\cdot\frac{a+b}{a+b+n}+\frac{y}{n}\cdot\frac{n}{a+b+n}\end{aligned}$$

In other words, the posterior mean is a weighted average between the prior mean, $\frac{a}{a+b}$, and the sample mean, $\frac{y}{n}$, where the respective weights are the share of the combined (actual and "hallucinated") sample size, $a+b+n$ which is taken up by effective and real data.

9.3.4 The Gamma-Poisson Model

Another example of conjugacy is the following. The **gamma family** of distributions consists of random variables whose range is the nonnegative real numbers, and whose PDF has the form

$$f(x) = k_{a,b}\cdot x^{a-1}e^{-bx}, \tag{9.37}$$

where $a, b > 0$ are parameters and $k_{a,b}$ is the normalization constant. We will write $X \sim \mathcal{G}\text{amma}(a,b)$ to indicate that X has the range and density above. It will turn out that this family is the conjugate prior for the Poisson parameter.

Lemma 9.6. *If $Y|X$ has a $\mathcal{P}\text{ois}(X)$ distribution and X has a $\mathcal{G}\text{amma}(a,b)$ prior, then the posterior $X|(Y=y)$ has a $\mathcal{G}\text{amma}(a+y,b+1)$ distribution. In other words, the Gamma family is the conjugate prior for the Poisson parameter.*

Proof. Let $Y|X \sim \mathcal{P}\text{ois}(X)$ and $X \sim \mathcal{G}\text{amma}(a,b)$. Then

$$\begin{aligned}f_X(x) &= k_{a,b}\cdot x^{a-1}e^{-bx} \\ f_{Y|X}(y|x) &= \frac{e^{-x}x^y}{y!} \\ f_{X|Y}(x|y) &= k_{a,b}\cdot k_y\cdot x^{a+y-1}e^{-(b+1)x} \\ &= k_{a,b,y}\cdot x^{a+y-1}e^{-(b+1)x},\end{aligned}$$

where all factors that do not contain x are absorbed into the constant, k''. The result is the PDF of a $\mathcal{G}\text{amma}(a+y,b+1)$ distribution. □

As with the Beta-Binomial model, the Gamma-Poisson model has a similar "effective sample" property associated with the parameters of the prior. Recall that the mean of a Poisson distribution is the parameter itself. Given a single observation, y, a reasonable estimate for the mean might be y itself (on average, over all possible observations, this estimate is correct; this property is known in statistics as "unbiasedness"). The a parameter of the Gamma prior acts similarly: as a *prior* point estimate for the mean.

The b parameter acts as the "effective sample size": it determines how much weight the prior estimate, a, has, relative to the purely data-based estimate, y, which in this case has a weight of 1, since we have only one observation. As with the Beta-Binomial model, the resulting posterior mean is a weighted average of the prior estimate, a and the data estimate, y, where the weight on the former is determined by b, and the weight on the latter is determined by the number of observations, in this case 1.

The mean of a $\mathcal{G}\text{amma}(a, b)$ distribution is $\frac{a}{b}$, and so the posterior mean is

$$\begin{aligned}\mathbb{E}\left[X|Y=y\right] &= \frac{a+y}{b+1} = \frac{\frac{a}{b}\cdot b + y\cdot 1}{b+1} \\ &= \frac{a}{b}\left(\frac{b}{b+1}\right) + y\left(\frac{1}{b+1}\right)\end{aligned}$$

9.3.5 Other Conjugate Examples

- If $Y|\sigma^2$ has a $\mathcal{N}(\mu, \sigma^2)$ distribution (where μ is fixed), then the **Inverse Gamma family** is the conjugate prior for σ^2.
- If $Y|\mu$ has a $\mathcal{N}(\mu, \sigma^2)$ distribution (where σ^2 is fixed), then the **Normal family** is the conjugate prior family for μ.

9.4 Exercises

1. A "triangular density" on the interval $[-1, 1]$ is given by the piecewise linear expression

$$f(x) = \begin{cases} k(1-x), & x \geq 0 \\ k(1+x), & x < 0 \end{cases}$$

which yields an isosceles triangle with a peak at 0.

(a) Find the value of k that makes this a valid density (Note: you do not necessarily need to do any integration, although you can — a simple geometric argument is sufficient)

(b) Find $P(|X| > \frac{1}{2})$ (again, a geometric argument is enough).

2. Consider a circle with radius r and center $(0, 0)$ on a plane. The points on the circle satisfy, $x^2 + w^2 = r^2$ (where x is the horizontal coordinate and w is the vertical one). If we take just the "upper half" of the circle, so that $w > 0$, we can solve the circle equation for w in terms of x and r to get $w = \sqrt{r^2 - x^2}$. Treating r as a constant, we can write $g(x)$ in place of w to yield the function $g(x) = \sqrt{r^2 - x^2}$.

(a) What is the valid domain, R_X, on which $g(x)$ is defined (as a real number)?

(b) Suppose X is a random variable with range R_X (as defined in the previous part). By Lemma 9.5, there is a unique constant, k_r, such that $f(x) = k_r \cdot g(x)$ is a valid PDF (k_r depends on the radius r that we chose, but we are assuming this is itself constant). What is the value of k_r? (As usual, you can work this out with geometry — no calculus is necessary.)

3. Suppose Y is a random variable with minimum value 0 and maximum value x, and is uniformly distributed between those two extremes (that is, $Y \sim \mathcal{U}0x$). The value of x is unknown.

(a) The prior range of x is $[0, \infty)$. Given an observation, $Y = y_0$, what is the posterior range for x?

(b) What is the formula for the likelihood function, $f_{Y|X}(y_0|x)$? Where is it largest? (You do not need calculus to answer this — draw the graph if you aren't sure.) Why, intuitively, should it be largest there?

(c) Suppose X has a $\mathcal{G}$amma(a, b) prior. Find the non-constant component of the posterior PDF. Is the posterior a Gamma distribution? How do you know?

4. (*) We have seen a "short-cut" to find that the mean of a $\mathcal{B}$eta(a, b) distribution is $a/(a+b)$, by recognizing the integrand as being proportional to a $\mathcal{B}$eta$(a+1, b)$ distribution, finding the value it must have in terms of Gamma functions, and then using the properties of the Gamma function to simplify the result. Use a

similar trick to show that the variance of a $\mathcal{B}\text{eta}(a, b)$ distribution is

$$\frac{ab}{(a+b)^2(a+b+1)}$$

5. (Adapted from Bolstad 8.1) In order to determine how effective a magazine is at reaching its target audience, a market research company selects a random sample of n people from the target audience and interviews them. Let X represent the proportion of the target audience that has seen the latest issue and Y be the number in the interview group who has seen it.

   (a) Suppose $n = 150$. What is the distribution of $Y|X = x$? Use your favorite software (e.g., R) to graph the distribution for an x value of your choice.

   (b) Suppose $Y = 29$. Graph the **likelihood function** for X.

   (c) (*) By differentiating the likelihood function and setting to zero, find the **maximum likelihood estimate** of X.

   (d) Using a uniform prior on X, find the posterior distribution of $X|Y = 29$.

   (e) Find $\mathbb{E}[X|Y = 29]$.

6. (Adapted from Bolstad 8.4) You are going to take a random sample of voters in a city in order to estimate the proportion, X, who support stopping the fluoridation of the municipal water supply. Before you analyze the data, you need a prior distribution for X. You decide that your prior mean is 0.4 and your prior variance is 0.01.

   (a) Assuming you want to use a Beta prior, what values of the parameters a and b do you need to use to satisfy the constraints on the mean and variance? (Hint: use the result Problem 2, even if you did not prove it)

   (b) What is the equivalent sample size of your prior?

   (c) Out of 100 voters polled, $y = 21$ support the removal of fluoridation. Find the posterior distribution for X.

   (d) Find the posterior mean and variance.

7. The **exponential distribution** with **rate parameter** λ and range $[0, \infty)$ is often used to model the amount of time that passes between two events. Its density is given by

$$f(y|\lambda) = \lambda e^{-\lambda y}$$

(a) Show that this is a special case of a $\mathcal{G}$amma(a, b) distribution, by finding the values of a and b that make the densities equivalent.

(b) Show that the $\mathcal{G}$amma(a, b) family is also a conjugate prior for the rate parameter. That is, if the prior density on λ has the form

$$f(\lambda) = k_{a,b} \cdot \lambda^{a-1} e^{-b\lambda}$$

where $a, b > 0$ are parameters and $k_{a,b} > 0$ is a normalizing constant that depends on a and b but not on λ, then the posterior density $f(\lambda|y)$ has the same form, for different values of a, b and k.

(c) (*) Show that if $Y_1, \ldots, Y_n$ are i.i.d. exponential with rate λ, then using a $\mathcal{G}$amma(a, b) prior for λ results in a Gamma posterior density, $f(\lambda|y_1, \ldots, y_n)$, and find its parameters.

Chapter 10

Decision-Making

10.1 Optimal Decisions

10.1.1 Motivation

We have spent a lot of time trying to calculate probabilities of various events. In particular, we've wanted to know how probabilities change when we condition on observations. Of course, while knowing the distribution of a variable is interesting, at some point we usually want to make some kind of decision based on what we think the value is or will turn out to be.

If we have a random variable X, whose value we don't know (possibly because it hasn't been realized yet, but maybe because we just don't have all the information), we can make an inference about its (unknown or future) value if I can characterize its distribution.

If we observe some other random variable, Y, we can improve our inferences about X by using its posterior distribution given Y. But what do we do with that distribution? A distribution is kind of a complex thing, and sometimes we need to make a specific prediction. And sometimes we need to translate the distribution into *action*. To do this in a principled way, we need a way of evaluating how good or bad an action is in light of the beliefs we have formed about the situation.

Example 1 Suppose I engage in a friendly office pool to pick the winner of the NCAA men's basketball championship. At the Final Four stage, the possibilities are Connecticut, Florida, Kentucky and Wisconsin.

I can define a random variable, X, with range $R = \{1, 2, 3, 4\}$, corresponding to these four teams (in, say, alphabetical order). Based on extensive research, suppose I assign the following probability distribution to X:

x	1	2	3	4
$P(X = x)$	0.2	0.4	0.1	0.3

Which team should I bet on? Why?

In this case, I will eventually observe the value of X, and if I picked the correct value, I win. If I picked anything else, I lose. How much I win or lose depends on the way the game is set up, but most likely, there's no "partial credit" (if I get credit for picking one of the two teams in the final round, I'm going to need another random variable, to capture all possible outcomes).

This is an example where the best decision rule turns out to be "Pick the value with the highest probability".

Example 2 What if my decision involves the unknown variable corresponding to whether or not I have cancer? As we've seen, getting a positive test result is certainly no guarantee that I actually have the disease, and so proceeding with an expensive treatment based on a single test may not be sensible.

Suppose X represents whether or not I have cancer ($R_X = \{0, 1\}$), and Y represents the outcome of a medical test ($R_Y = \{0, 1\}$), where 1 indicates cancer or a positive test result, respectively.

If the test is positive, then I can compute the conditional distribution of X given that $Y = 1$. If the test has a sensitivity of 0.95 and a specificity of 0.99, and the disease has a base rate of 1 in 10000, then I would get

x	$f_X(x)$	$f_{Y\|X}(1\|x)$	$f_{X,Y}(x, 1)$	$f_{X\|Y}(x\|1)$
0	0.9999	0.01	0.009999	0.9906
1	0.0001	0.95	0.000095	0.0094
			$f_Y(1) = 0.00194$	

(You should check for yourself that this is indeed the correct posterior given the information provided)

So the probability that I actually have the disease is less than 1%. If I'm asked "Do you have cancer?", what should I answer?

Most likely, the answer is no... But this is potentially a life-or-death matter, and so my actual actions (do I get treatment?) depend on more than just the probability that I have cancer; I need to take into account things like the cost of treatment (both in terms of dollars and possible side-effects), my expected quality of life and prognosis under each of the four combinations of (1) treating real cancer, (2) unnecessary treatment, (3) untreated real cancer, (4) not treating non-cancer, etc. I want to make my decision based on some kind of cost-benefit analysis (even though the costs and benefits might be pretty intangible). I want to make the decision that minimizes the "expected badness" of what happens afterwards.

So far, the number of available decisions has matched the number of possible values of a variable. There's no reason this has to be the case, however: I might be able to choose (a) treatment, (b) further testing, or (c) doing nothing. I need to assess the goodness/badness of each decision for each possible state of the world.

10.1.2 Loss Functions

No amount of mathematical theory can tell us how exactly to compare the badness of spending $10,000 on an unnecessary treatment to the badness of letting cancer go untreated. This is subjective. But *if* we can assign numbers to the badness of all possible scenarios, then we can use decision theory to tell us which option minimizes our "expected badness".

To do this, we first need to define a **loss function**.

Definition 10.1 (Loss Function)**.** *Let X be a random variable with range R, and let $\mathcal{A}$ be a set of possible actions. A **loss function**, L, assigns to each "state-action pair", $(x, a) \in R \times \mathcal{A}$, a real number. The value $L(x, a)$ expresses the cost of taking action a when the true state of the world is x.*

Example 1 In our basketball example, we have $\mathcal{A} = \{1, 2, 3, 4\}$ (corresponding to possible bets). Our loss function might look like Table 10.1.

Example 2 In the binary-choice version of our medical example, we might have $\mathcal{A} = \{0, 1\}$, representing whether or not we seek treatment, and the loss function

		a			
		1	**2**	**3**	**4**
x	**1**	0	1	1	1
	2	1	0	1	1
	3	1	1	0	1
	4	1	1	1	0

Table 10.1: Entries in the table represent $L(x, a)$ for the example of picking a winning basketball team.

		a	
		0	**1**
x	**0**	0	1
	1	50	0

Table 10.2: Entries in the table represent $L(x, a)$ for the example of choosing whether to proceed with cancer treatment.

might look like Table 10.2 If we allow for the third option of further testing, we might have $\mathcal{A} = \{0, 1, 2\}$, for "Do nothing", "Get treatment", "Get further testing". Then we might get something like Table 10.3

10.1.3 Risk

We'd like a policy that minimizes our loss. If we knew that $X = x$, we could just pick the action with the smallest loss for that x. But if we knew $X = x$, we wouldn't be doing this; the whole point is that we want to make the best decision in the face of *uncertainty* about X.

One way to think about how to do this is to imagine repeatedly encountering the

		a		
		0	**1**	**2**
x	**0**	0	5	1
	1	50	0	1

Table 10.3: Entries in the table represent $L(x, a)$ for the example of choosing whether to proceed with cancer treatment, get further testing, or do nothing.

same scenario; making a decision each time; and then realizing some amount of loss. Over time, we accumulate "costs", and so we'd like to keep this running total as small as possible. We can do this by taking the action with the smallest *average loss*; that is, where the *expected value* of our loss is smallest.

Definition 10.2 (Risk)**.** *The **risk**, r associated with an action, a, is the expected value of the associated loss. We define*

$$r(a) = \mathbb{E}\left[L(X, a)\right] = \sum_{x \in R} L(x, a) f(x)$$

where f is the PMF of X (if X is continuous, replace the PMF with the PDF, and replace the sum with an integral).

In words, if we pick an action, a, the risk associated with that action is a weighted average of all possible losses, where the losses come from the relevant column in L, and the weights come from the PMF of X.

Notice that if we represent the distribution of X as a row vector, $\boldsymbol{\pi}$ and the loss function as a matrix $\mathbf{L}$, where rows correspond to x values and columns correspond to actions, then the risk function can be obtained (as another row vector) by matrix multiplication:

$$\mathbf{r} = \boldsymbol{\pi}\mathbf{L} \tag{10.1}$$

Example 1 In our basketball example, we had

$$\boldsymbol{\pi} = \begin{pmatrix} 0.2 & 0.4 & 0.1 & 0.3 \end{pmatrix}$$
$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

which yields

$$\mathbf{r} = \begin{pmatrix} 0.8 & 0.6 & 0.9 & 0.7 \end{pmatrix}$$

and so, as we determined, based on this distribution and loss function, our risk would have been minimized by betting on 1, Florida. Of course, if we had started with a different distribution, we could get a different answer.

Example 2 In the cancer example where we have three actions available: do nothing, get treatment, and get further testing, we have

$$\boldsymbol{\pi} = \begin{pmatrix} 0.9906 & 0.0094 \end{pmatrix}$$
$$\mathbf{L} = \begin{pmatrix} 0 & 5 & 1 \\ 250 & 0 & 1 \end{pmatrix}$$

which yields

$$\mathbf{r} = \begin{pmatrix} 2.350 & 4.953 & 1.000 \end{pmatrix}$$

Here the lowest-cost action is to get further testing. Notice that this isn't the best action within either row of the loss matrix, but by averaging over rows, it wins out as the "compromise" option.

Of course, if the probability of cancer were higher, then it might be better to just go ahead with treatment; similarly, when the probability dips below some threshold, doing nothing will win out.

10.1.4 Invariance Properties of Minimum Risk Action

In the examples above, you will notice that we put a zero in every row of the loss matrix, and a one somewhere in the matrix (though not necessarily in every row). Why do we do this? There's no special reason we need to, but it turns out that we always can do so without impacting our final decision. This is helpful in figuring out how to design a loss function.

Lemma 10.1. *Let X be a random variable with range R and PMF (or PDF) f, let $\mathcal{A}$ be a set of actions, and let L be a loss function on $R \times \mathcal{A}$. Denote by a_L^* the minimum risk action, $a_L^* = \arg\min_a r_L(a) = \arg\min_a \mathbb{E}\left[L(X, a)\right]$ associated with the loss function L.*

(a) Define $L'(x, a) = c \cdot L(x, a)$ for some $c > 0$. Then $a_{L'}^* = a_L^*$.

(b) Define $\tilde{L''}(x, a) = \cdot L(x, a) + c(x)$ where $c(x) \in \mathbb{R}$ for each x. Then $a_{L''}^* = a_L^*$.

What this lemma says is that (a), the entire loss matrix can be rescaled without affecting the final decision, and (b) each row in the matrix can be shifted (so that, for example, the smallest value is zero) without affecting the final decision.

Proof. (a) For a given action a, we have

$$\begin{aligned} r_{L'}(a) &= \sum_{x \in R} L'(x,a) f(x) \\ &= \sum_{x \in R} c \cdot L(x,a) f(x) \\ &= c \sum x \in R L(x,a) f(x) \\ &= c \cdot r_L(a) \end{aligned}$$

since c is positive and does not depend on a, rescaling the risk function does not change which a gives the smallest risk.

(b) Similarly, we have

$$\begin{aligned} r_{L''}(a) &= \sum_{x \in R} L''(x,a) f(x) \\ &= \sum_{x \in R} \left(L(x,a) + c(x) \right) f(x) \\ &= \sum_{x \in R} L(x,a) f(x) + \sum_{x \in R} c(x) f(x) \\ &= r_L(a) + \mathbb{E}\left[c(X) \right] \end{aligned}$$

Since the additive constant does not depend on a, the value of a that minimizes the risk is unchanged.

□

Here's a rough procedure that can be followed in light of the above lemma. For each possible state x, figure out which action would be best if that were the true state of the world. Assign that action a loss of zero (i.e., put a zero somewhere in each row).

Then, find a state-action combination which is less good than the best one in its row, but whose cost is easy to compare to others. Assign that one a loss of 1. Now, for every other combination, think about how many times worse (or better) it is relative to the best action in its row than the reference combination, and assign it that multiplier for a loss.

In the basketball betting example, we start by assigning a loss of zero to each correct bet. All other bets are equally bad (that is, each one is one times as bad as any other), so we assign them all a loss of 1.

In the simple cancer example with just two actions, pick the preferred action for each cancer status. If you have cancer you want to treat it, and if you don't you don't. Then, think about treatment when you don't have the disease. If that has a loss of 1, how much worse is foregoing treatment when you do have it? In the example, we said it was 50 times worse.

Of course, if there's a more intuitive way to design a loss function, you should do that instead.

10.1.5 Bayes Risk and Bayes Decision Rules

In the cancer-testing example, we computed our X probabilities based on an observation; namely that a test came out positive. If the test had produced different results, we would have a different distribution, and potentially a different optimal decision.

In general, if we observe Y, we want to make our decision based on the unknown but related variable X using its *conditional distribution*. That is, given an observation $Y = y_0$, we need to compute $f_{X|Y}(x|y_0)$ for each $x \in R_X$ so that we can select the optimal decision. We do this by minimizing the **conditional risk**

Definition 10.3 (Conditional Risk)**.** *Given an observation $Y = y_0$, a unobservable variable X, a set of possible actions $\mathcal{A}$, and a loss function defined for each (x, a) pair, the **conditional risk** is given by*

$$r(a|y) = \mathbb{E}\left[L(X, a)|Y = y_0\right] = \sum_{x \in R_X} L(x, a) f_{X|Y}(x|y_0)$$

If we were planning to make repeated observations, and make a decision for each one (for example, we might want to design a computer monitoring system that examines widgets for signs of defectiveness, and automatically decides whether to let them through or toss them; or a device to detect when a golden goose egg is likely to be spoiled), we would want a function that took in an observation and output a decision.

Definition 10.4 (Decision Rule)**.** *Given a random variable Y with range R_Y, whose value will observed, and a set of possible actions $\mathcal{A}$, a **decision rule** is a function that takes observations and returns actions directly. Formally, its domain is R_Y, and its range is $\mathcal{A}$.*

We can define a decision rule any way we want; but a natural thing to do will be to pick, for each $y \in R_Y$, the action that minimizes the conditional risk. The decision rule that does this is called the **Bayes Decision Rule**.

Definition 10.5 (Bayes Decision Rule)**.** *The **Bayes decision rule**, $d^\star$ is the decision rule which minimizes the conditional risk for each y:*

$$d^\star(y) = \arg\min_a r(a|y)$$

Example 2 We have already found the value of the Bayes decision rule when the test is positive for cancer. The optimal action is $a = 2$: get more testing, which had a risk of 1. In our new notation:

$$d^\star(1) = 2$$

What about when the test is negative; that is, when $Y = 0$? We have

x	$f_X(x)$	$f_{Y\|X}(1\|x)$	$f_{X,Y}(x,1)$	$f_{X\|Y}(x\|1)$
0	0.9999	0.99	0.989905	0.99999
1	0.0001	0.05	0.000005	0.00001
			$f_Y(1) = 0.98991$	

Repeating the loss function from before, we have

$$\boldsymbol{\pi}|(Y=0) = \begin{pmatrix} 0.99999495 & 0.00000505 \end{pmatrix}$$
$$\mathbf{L} = \begin{pmatrix} 0 & 5 & 1 \\ 250 & 0 & 1 \end{pmatrix}$$

which yields

$$\mathbf{r}|(Y=0) = \begin{pmatrix} 0.0012625 & 4.999975 & 1.000000 \end{pmatrix}$$

Now the risk is lowest if we do nothing. In other words

$$d^\star(0) = 0$$

Now, having worked out the optimal decision for each potential observation, we can, if we want, just keep d^*, and use it to convert observations into decisions directly.

10.2 Common Loss Functions for Estimation

Obviously the choice of loss function has a large impact on what decision is optimal, and so it is worth putting in some effort studying the nature of a problem in order to come up with a suitable loss function.

In many cases, the decision amounts to picking an x value. For example, X might represent some future event, which we are trying to predict; or X might represent the parameter of a distribution, which we are trying to estimate. In these situations, the action set is the same as the range: $\mathcal{A} = R$, where $\mathcal{A}$ is the set of possible actions and R is the range of the random variable X. This class of decision problem is called an **estimation problem**, since the result is an *estimate* of the value of X.

In this situation, there are several "standard" loss functions with some elegant properties. We'll look at a few of these now. In each case, we will use $f(x)$ to represent the PMF or PDF of X, but in the event that we have data, $Y = y_0$, everything stays the same if we just replace X with $X|(Y = y_0)$ and $f(x)$ with $f(x|y_0)$

10.2.1 Zero-One Loss

In the basketball example, there was no "partial credit"; our action was either correct or incorrect, with a fixed, constant cost for being incorrect (i.e., it didn't matter in what way we were wrong).

There, we used a **zero-one** loss function: there was no cost associated with being right, and a unit cost for all the different ways of being wrong. Notice that in this case it doesn't matter what specific numbers we choose, as long as the cost of being correct is lower than the cost of being wrong. For simplicity, we set the former to 0 and the latter to 1.

The resulting optimal decision was the value with the highest probability. This is always the case with zero-one loss.

Lemma 10.2. *For a decision problem where $\mathcal{A} = R$, the optimal decision under zero-one loss is the x value with the highest probability; that is, $a^* = \arg\max f(x)$.*

Proof. The optimal decision is the one that minimizes $r(a)$. We have

$$\begin{aligned} r(a) &= \sum_{x \in R} L(x, a) f(x) \\ &= \sum_{x \neq a} f(x) \\ &= 1 - f(a) \end{aligned}$$

where the second line follows because the term where $x = a$ has loss of zero.

Clearly we make this quantity smaller by making $f(a)$ bigger; and so we should choose the a where $f(a)$ is at a maximum. □

10.2.2 Distance-Based Loss

When the unknown variable is quantitative rather than qualitative, it is natural to define a loss function which is based on the difference between our guess and the true value. Two natural choices are **squared error loss** and **absolute loss**. Both of these make sense when

(1) $\mathcal{A} = R$, and

(2) R has a natural "distance" metric (that is, X is on an interval or ratio scale).

Squared Error Loss

Definition 10.6 (Squared Error Loss)**.** *Under the above conditions, the* ***squared error loss function*** *is given by*

$$L(x, a) = (a - x)^2$$

Lemma 10.3. *Under squared error loss, the optimal action is* $a^* = \mathbb{E}[X]$.

Proof. As usual, we want to minimize $r(a)$. We have

$$\begin{aligned} r(a) &= \mathbb{E}[L(X, a)] \\ &= \mathbb{E}\left[(a - X)^2\right] \\ &= \mathbb{E}\left[a^2 - 2aX + X^2\right] \\ &= a^2 - 2a\mathbb{E}[X] + \mathbb{E}\left[X^2\right] \end{aligned}$$

Since we want to minimize this expression by choosing the best a, we differentiate with respect to a and set the result to zero.

$$\begin{aligned} \frac{dr(a)}{da} &= 2a - 2\mathbb{E}[X] \equiv 0 \\ \implies 2a^* &= 2\mathbb{E}[X] \\ \implies a^* &= \mathbb{E}[X] \end{aligned}$$

□

The second derivative of $r(a)$ is 2, which is positive, and so $a^* = \mathbb{E}[X]$ really is a *minimum*.

Absolute Error Loss

Definition 10.7 (Absolute Error Loss)**.** *Under the same conditions as above, the* ***absolute error loss function*** *is given by*

$$L(x, a) = |a - x|$$

Lemma 10.4. *Under squared error loss, the optimal action is a* ***median*** *value of* X*.*

Let's first try to get an intuitive grasp of why this should be true. Consider choosing an action a. Each possible x value contributes $|x - a|$ to the expected loss, weighted by its probability. For those x values to the right of a (in the sense of a number line), we can make their contribution smaller by increasing the value of a, since then the distance from a to these points will be less. But this comes at the cost of increasing the distance from a to all the x values to its left. When do the benefits outweigh the costs?

A move to the right is "worth it" when the values on the right have greater weight in the average; that is, if their total probability is greater. So we can keep getting benefits from moving a right as long as the probability that $X > a$ is greater than the probability that $X < a$. Of course, the farther right we move, the smaller $P(X > a)$ is getting, and the bigger $P(X < a)$ is getting. At some point these will even out; and then any further benefit from moving to the right will be negated by the increased distance to the points on the left. But this equilibrium point is precisely the median, which is the point, a, where $P(X > a) = P(X < a)$.

In the proof below we will assume that X is continuous, because this will actually make life a little easier. The reason is that minimization is easier when you can take derivatives, which only apply in the continuous case. However, if X is discrete, the main idea of the proof still holds; we just end up with a *range* of a values that minimize the risk.

Proof. The risk is

$$\begin{aligned} r(a) &= \int_{x \in R} L(x,a) f(x)\ dx \\ &= \int_{x \in R} |a-x| f(x)\ dx \\ &= \int_{x \le a} (a-x) f(x)\ dx + \int_{x > a} (x-a) f(x)\ dx \end{aligned}$$

where we have gotten rid of the pesky absolute value by considering the part of the range where $a - x$ is positive separately from the part where it is negative. In the latter case, the absolute value is equal to $x - a$.

We can rewrite this as

$$\begin{aligned} r(a) &= \int_{x \le a} (a-x) f(x)\ dx - \int_{x > a} (a-x) f(x)\ dx \\ &= \int_{x \le a} a f(x)\ dx - \int_{x \le a} x f(x)\ dx - \int_{x > a} a f(x)\ dx + \int_{x > a} x f(x)\ dx \\ &= a \left[\int_{x \le a} f(x)\ dx - \int_{x > a} f(x)\ dx \right] - \left[\int_{x \le a} x f(x)\ dx - \int_{x > a} x f(x)\ dx \right] \\ &= a \left[F(a) - (1 - F(a)) \right] - \left[\int_{x \le a} x f(x)\ dx - \left(\int_{x \in R} x f(x) - \int_{x \le a} x f(x)\ dx \right) \right] \\ &= a(2F(a) - 1) - 2 \int_{x \le a} x f(x) + \mathbb{E}\left[X\right] \\ &= 2aF(a) - a - 2 \int_{x \le a} x f(x)\ dx \end{aligned}$$

If we take the derivative with respect to a, we get

$$\frac{dr(a)}{da} = 2a \frac{d}{da} F(a) + 2F(a) \frac{d}{da} a - 1 - 2af(a)$$

where the derivative of an integral with respect to the upper limit of integration is just the integrand (by the Fundamental Theorem of Calculus). Since the derivative of the CDF is the PDF, this is equal to

$$\frac{dr(a)}{da} = 2af(a) + 2F(a) - 1 - 2af(a) = 2F(a) - 1$$

Setting this to zero gives us $F(a) = \frac{1}{2}$. In other words, a must be the median! □

10.3 Exercises

You can use any computational tool you want to help you with the calculations, table-making, etc. in this assignment; just record what code/commands you used.

All questions deal with the following scenario:

You are developing an automated object recognizer. The world of objects consists of three different shapes (squares, circles and triangles), each of which can be red, green, or blue, for a total of nine distinct object types. Your recognizer will rely on data from imperfect color and shape sensors.

Forty percent of objects are circles. Squares and triangles occur equally often. We have the following conditional probabilities of color given shape:

		Color		
		Red	Green	Blue
Shape	Square	0.6	0.3	0.1
	Circle	0.2	0.6	0.2
	Triangle	0.1	0.3	0.6

Table 10.4: Conditional probabilities of color given shape

1. Compute the joint distribution over the nine object types.
2. Suppose the most important thing for recognition is getting the shape right, but that there is also value in getting the color right. Construct a loss function (for the nine true object types and nine possible classifications) that reflects these priorities.
3. Compute the prior risk associated with each choice.

4. Suppose you could observe the color with perfect accuracy. For each color, compute the posterior distribution over shapes, and calculate the conditional risk function associated with each choice of shape (using the same loss function as above).

5. When the true color is blue, your color sensor has an 0.8 probability of returning 'blue', and a 0.1 probability of returning each of the other colors. The analogous distribution holds for the other true colors. Assuming the output of the color sensor is independent of shape, compute a new posterior distribution over the nine object types for each color that the sensor might report.

6. Compute the posterior risk (using the same loss function) for each classification choice assuming the sensor says 'blue'. Do the same for 'green' and 'red' sensor readings.

7. What is the minimim-risk decision rule? I.e., for each sensor reading about color, what is the best shape to report?

8. Let's add a shape sensor to the mix. The shape sensor is more accurate than the color sensor: for a true square, it has a 0.9 probability of producing the correct label, and a 0.05 probability of returning each of the other shapes; analogously for the other true shapes. The sensor's output is independent of both the true color and the output of the color sensor. Suppose the color sensor says 'blue'. For each value that the shape sensor might return, compute the posterior distribution over the nine object types given the output of *both* sensors.

9. Using the results of the previous part, compute the posterior risk associated with each of the nine choices, under each value of the shape sensor (again assuming that the color sensor says "blue").

10. What is the minimum-risk decision rule mapping shape sensor outputs to classification decisions (again assuming that the color sensor says "blue")?

Part IV

Other Applications of Conditional Probability

Chapter 11

Information Theory

11.1 Information

How informative would you find the following (not necessarily true) statements?

1. It's hot in Tucson, AZ.
2. It's hot in Anchorage, Alaska.

How about these?

3. My car is yellow.
4. My license plate is 416 SPG.

What's the difference?

- What's the difference betwen 1 and 2?
- What about between **??** and **??**?

Intuitively, it seems that there is more information conveyed by statements which are more surprising. Surprise can result from learning that (a) something rare has happened, (b) something that deviated from your previous beliefs happened, (c) one of a large number of possibilities has happened. If you already knew something, there's not really any new information in hearing it again.

Colloquially, we tend not to think that a statement conveys much information if we don't care about it (e.g., learning my license plate is probably not important to you,

unless you're trying to report some illegal activity on my part or something).

Formally, though, it's difficult to quantify "relevance", so for now, we'll stick to the "surprise" component of information.

11.1.1 Criteria for a Measure of Information

Given an event space, U, and a probability measure on events in that space, we'd like to define an **information function**, which takes an event and tells us how much information it contains using a real number. Let's call this function I. How should it be defined?

Consider the simple example of drawing a card from a standard 52-card deck. Rank the following in terms of their information content:

(a) Learning the card is a spade (E_1)

(b) Learning the card is a nine (E_2)

(c) Learning the card is a nine of spades ($E_1 \cap E_2$)

The information content increases as the probability goes down. This is our first criterion:

1. $I(E)$ decreases as $P(E)$ increases; that is, **information is a decreasing function of probability**

Now, compare the following:

(a) Learning the card is a nine (E_2) and then learning that it is a spade (E_1)

(b) Learning the card is black (E_3) and then learning that it is a spade (E_2)

(c) Learning the card is a spade (E_2) and then learning that it is black (E_3)

In all three cases, we learn E_1: the card is a spade. Does it always contain the same information?

In a and c, it would seem that it contains the same information: we've narrowed the possibilities from four suits to one. In b, however, by the time we learn the card is a spade, we had already learned it was black, and so we were only going from two possible suits to one.

So, when we learn something, it matters what we knew to start with! That is, what counts is the *conditional* probability of the event, given everything we know so far! We add information when we learn something new, but only to the extent that we actually learned something new. If we learn E and then F, we get a total of $I(E) + I(F)$ only if $P(F) = P(F|E)$: that is, if E and F are independent.

2. If E and F are independent, then $I(E \cap F) = I(E) + I(F)$.

2*. In general, we can think of $I(E \cap F) = I(E) + I_E(F)$, where $I(F|E)$ is *conditional information* of F given E, which is equal to the information contained in an event with probability $P(F|E)$.

Finally, if we observe an event that had probability 1, this should have zero information. Then, by the first criterion, since no event can have probability greater than 1, no event can have information less than 0.

3. If $P(E) = 1$, then $I(E) = 0$.

To summarize

Definition 11.1 (Criteria for Information)**.** *Given an event space U with probability measure P, we say that I is an **information function** if it satisfies the following criteria (where we let E and F be arbitrary events in the domain of P):*

1. *$I(E)$ is a strictly decreasing function of $P(E)$. That is*
   (a) *$I(E) = I(F)$ iff $P(E) = P(F)$*
   (b) *$I(E) < I(F)$ iff $P(E) > P(F)$*
2. *If E and F are independent, then*

$$I(E \cap F) = I(E) + I(F)$$

3. *Certain events have no information: If $P(E) = 1$, then $I(E) = 0$.*

Theorem 11.1. *Any information function I must be of the form*

$$I(E) = -K \log(P(E)) \tag{11.1}$$

for some real number $K > 0$

It's straightforward to check that a function with this form does have the properties of an information function.

1. First, note that $\log(\cdot)$ is a strictly increasing function; so multiplying by a

negative number turns it into a strictly decreasing function (see Fig. 11.1).

2. If E and F are independent, then $P(E \cap F) = P(E)P(F)$. It is a general property of $\log(\cdot)$ that $\log(xy) = \log(x) + \log(y)$.

3. We have $\log(1) = 0$.

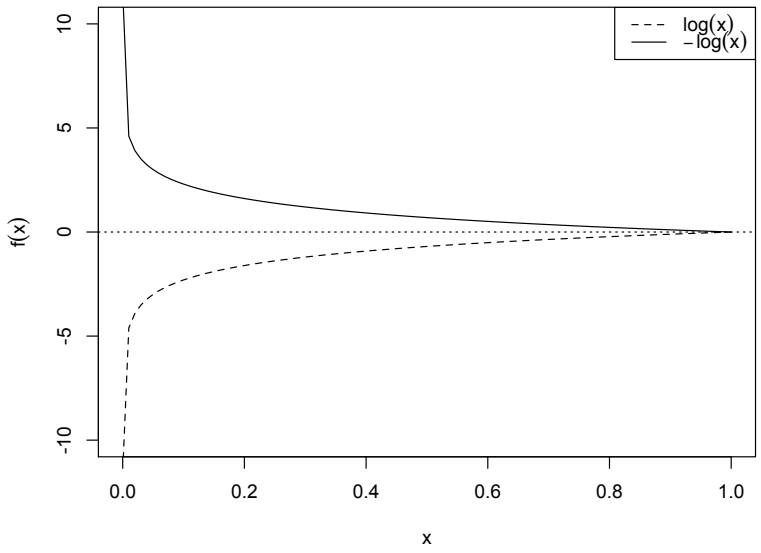

Figure 11.1: Graph of $\log(x)$ and $-\log(x)$

The proof that no other function satisfies these properties is more involved than usual, so we omit it, but you can read it as part of the proof of Theorem 6.8, in section 6.6 of the Applebaum book.

Taking the theorem for granted, what should we choose for K? The value which is most convenient depends on the application (it's just a matter of what unit of information we want to use), but one sensible convention is that discovering that one of two equally likely possibilities is true (e.g., observing the outcome of a coin flip) yields one unit of information. That is, if A is (say) the event "Coin comes up

heads", then $P(A) = \frac{1}{2}$, so we want

$$\begin{aligned} I(A) &= -K \log(\frac{1}{2}) = 1 \\ \implies K \log(2) &= 1 \\ \implies K &= \frac{1}{\log(2)} \end{aligned}$$

In other words we set

$$I(E) = -\frac{\log(P(E))}{\log(2)} = -\log_2(P(E)) \tag{11.2}$$

where the second equality follows from the general fact of logarithms that

$$\log_a(x) = \log_b(x) \log_a(b)$$

This yields the unit of information called the **bit** (short for "binary digit"): If we have a set of outcomes which are all equally likely, then then the number of digits that we need to record what happens depends on the number of possibilities.

- If there's only 1 outcome possible, we don't need to record anything — we need zero bits.
- If there are 2 possible outcomes, we can label them 0 and 1, and we need one binary digit (bit) to record what happens:

$$\text{Record } \begin{cases} 0 & \text{if outcome 0 happens} \\ 1 & \text{if outcome 1 happens} \end{cases}$$

- If there are 4 possibilities, we can label them $0, 1, 2$ and 3, and we need 2 binary digits to record the outcome:

$$\text{Record } \begin{cases} 00 & \text{if outcome 0 happens} \\ 01 & \text{if outcome 1 happens} \\ 10 & \text{if outcome 2 happens} \\ 11 & \text{if outcome 3 happens} \end{cases}$$

In general, with n bits we can record 2^n different outcomes. Or, reversing the relationship, if there are $k = 2^n$ possibilities, all equally likely, then the information conveyed by any one of them is

$$-\log_2(\frac{1}{2^n}) = \log_2(2^n) = n$$

bits

If instead we set $K = 1$ and use the natural log, then we get information measured in **nats**. There are about 1.44 (i.e. $1/\log(2)$) bits to one nat.

11.2 Entropy

Measuring information is valuable when trying to decide what inquiries to make, or what data to collect, what experiments to do, etc. A reasonable question to ask is "of the options available to me, what's the most efficient use of my resources, in terms of gaining information?"

Of course, if we knew exactly what the data would be, then it wouldn't give us any information (as its conditional probability would be 1, given the data we already had). In any interesting setting, we're not choosing *exactly what data* to get, but what *data source* to pursue; or what *type of data* to gather. That is to say, which *random variable* should we observe?

To answer this question, we might ask which random variable gives us the greatest *expected information*, given our present state of knowledge.

11.2.1 Information Associated With a Random Variable

Suppose we have a random variable X with range $R = \{x_1, x_2, \ldots, x_n\}$ and PMF f. If we observe the value of X, and discover that $X = x_r$, then we gain

$$I(X = x_r) = -\log_2(f(x_r))$$

bits of information. But since we don't know in advance what value X will take, the amount of information we get by observing X is itself a random variable!

Given the probability law of X, we can define a random variable, $I(X)$, whose range is

$$\begin{aligned} R_{I(X)} &= \{I(x_1), I(x_2), \ldots, I(x_n)\} \\ &= \{-\log(f(x_1)), -\log(f(x_2)), \ldots, -\log(f(x_n))\} \end{aligned}$$

That is, its values are the real numbers corresponding to the information contained in the events $\{X = x_1\}, \{X = x_2\}, \ldots \{X = x_n\}$.

11.2.2 Entropy Definition

Then, the **expected information** for X is

$$\mathbb{E}\left[I(X)\right] = \sum_{i=1}^{n} -\log(f(x_i))f(x_i)$$

This is called the **entropy** of X, due to its connection with entropy in physics as a measure of the "disorder" in a system. (Here, since information is associated with low probabilities, entropy is a measure of uncertainty associated with X. The higher the inherent uncertainty, the more information, and hence reduction in uncertainty, there is to be gained.)

Definition 11.2 (Entropy)**.** *The **entropy**, H, of a random variable X with range $\{x_1, \ldots, x_n\}$ is the expected value of the information gained by observing its value:*

$$H(X) = \mathbb{E}\left[I(X)\right] = -\sum_{i=1}^{n} \log(f(x_i))f(x_i) \tag{11.3}$$

where we adopt the convention that $0\log(0) = 0$.

Example: Suppose $X \sim \mathcal{B}\text{ern}(p)$. Find $H(X)$, as a function of p. For what value of p would you expect $H(X)$ to be greatest? Is your intuition correct?

Using (11.3) we get

$$\begin{aligned} H_p(X) &= -\sum_{x=0}^{1} \log(f(x))f(x) \\ &= -\log(1-p)\cdot(1-p) - \log(p)\cdot p \end{aligned} \tag{11.4}$$

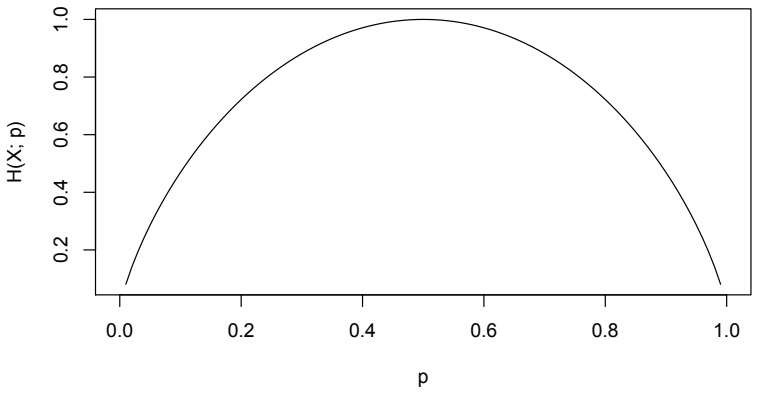

Figure 11.2: Entropy of a Bernoulli Distribution for Different Values of p

The graph of this function is in Fig. 11.2. To find the maximum, take the derivative with respect to p and set it to zero:

$$\begin{aligned}
0 = \frac{d}{dp} H_p(X) &= -\frac{1}{(1-p)\log(2)} \cdot (-1) \cdot (1-p) - \log_2(1-p) \cdot (-1) \\
&\quad - \frac{1}{p\log(2)} \cdot 1 \cdot p - \log_2(p) \cdot 1 \\
&= \frac{1}{\log(2)} + \log_2(1-p) - \frac{1}{\log(2)} - \log_2(p) \\
&= \log_2(1-p) - \log_2(p) \\
\implies log_2(1-p) &= \log_2(p) \\
\implies 1-p &= p \\
\implies p &= \frac{1}{2}
\end{aligned}$$

To check that this is a maximum and not a minimum, take the second derivative and

check its sign:

$$\begin{aligned}\frac{d^2}{dp^2}H(x;p) &= \frac{d}{dp}\left(\log_2(1-p) - \log_2(p)\right) \\ &= \frac{1}{(1-p)\log(2)} \cdot (-1) - \frac{1}{p\log(2)} \\ &= -\frac{1}{(1-p)\log(2)} - \frac{1}{p\log(2)}\end{aligned}$$

Since both p and $1-p$ are positive (as is $\log(2)$), the second derivative is negative everywhere, so $p = \frac{1}{2}$ gives the maximum.

Looking at the graph in Fig. 11.2, you can see that when "successes" (i.e., $X = 1$) are either certain or impossible, the entropy is 0. There is no expected information from observing the outcome if we know it in advance. When success is almost certain or almost impossible, there's very little expected information, since most of the time we're just going to observe what we thought would happen. As the probabilities get closer to 50/50, we have less and less idea about what will happen, and so there is more information gained by making an observation.

We can think of entropy as a measure of our prior uncertainty about the value of a random variable.

11.2.3 Properties of Entropy

Some important properties of Entropy are collected in the following theorem.

Theorem 11.4 (Properties of Entropy)**.** *Let X be a discrete random variable with range R. Then*
(a) $H(X) \geq 0$ with $H(X) = 0$ iff $P(X = x) = 1$ for some $x \in R$.
(b) $H(X) \leq \log(\#(R))$ with $H(X) = \log(\#(R))$ iff X has a uniform distribution. Note that a uniform distribution is only possible if $\#(R) < \infty$, whereas if $\#(R) = \infty$, then the theorem isn't really saying anything at all.

Before we prove the theorem, let's look at what it's saying.

Part (a) says that no random variable has "negative uncertainty"; and that unless the random variable is certain to take a particular value, there is some uncertainty associated with it (and some information expected by observing it).

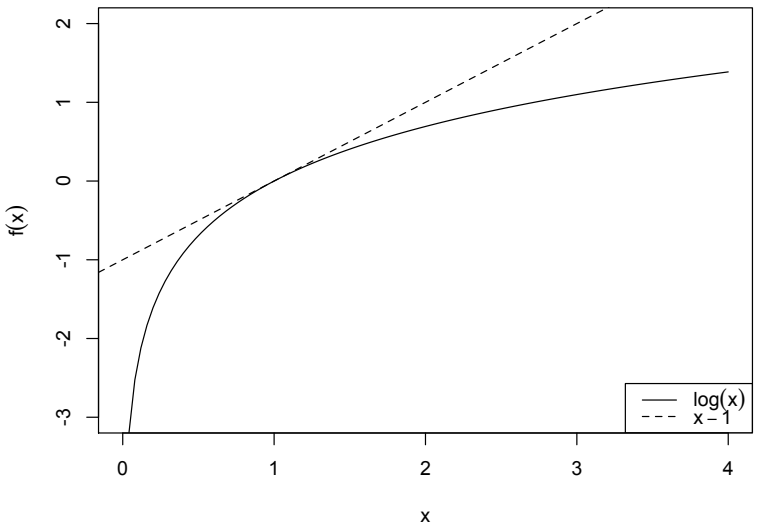

Figure 11.3: Plot of $\log(x)$ and $x - 1$

Part (b) gives us an upper bound on the entropy of a random variable, and says that for a particular range, the uniform distribution is the most uncertain. This generalizes what we showed about the Bernoulli distribution above.

Proof.

(a) To show that $H(X) \geq 0$, note that we have

$$H(X) = -\sum_{x \in R} f(x) \log(f(x))$$

Each term in the sum is the product of a nonnegative number ($f(x)$) and a negative number ($\log(f(x))$), giving a nonpositive number. Taking the negation of the entire sum thus yields a nonnegative number.

The sum will be zero if and only if all the terms are zero, which happens if either the probability is zero, or if the log probability is zero. The latter happens when the probability is one. But only one value can have probability one, and if it does, all the rest are zero. Which is exactly the case of the certain variable.

(b) To prove the upper bound on the entropy, we first need a Lemma which has nothing to do with information theory.
Lemma 11.3. $\log(x) \leq x - 1$ *with* $\log(x) = x - 1$ *iff* $x = 1$

Proof of Lemma. It is easy to see that this is true by inspecting a graph of the two functions (see Fig. 11.3). To prove the statement first note that when $x = 1$, $\log(1) = 0 = (1 - 1)$. Now, taking derivatives of both sides, we see that

$$\frac{d}{dx}(x - 1) = 1$$

whereas

$$\frac{d}{dx}\log(x) = \frac{1}{x}$$

which is less than 1 when $x > 1$, and greater than 1 when $x < 1$. This shows what we see in the graph: that to the right of the point where they touch, $\log(x)$ increases more slowly than $x - 1$, and hence remains below $x - 1$ forever. As we move left from 1, $\log(x)$ falls off more quickly than $x - 1$, and so again remains below it.

(Even this is a bit informal: what we really should do is note that, for $x > 1$:

$$\log(x) = \log(1) + \int_1^x \frac{d}{dx}\log(x)dx$$
$$x - 1 = (1 - 1) + \int_1^x \frac{d}{dx}(x - 1)dx$$

Since the first terms are equal in both cases, and since the boundaries of integration are equal, the relationship between the two expressions is determined by the integrand, which is greater in the first expression.

For $x < 1$, the argument is the same, except that we have a difference, rather than a sum, and the bounds on the integral are reversed.) □

To prove part (b) of the theorem, we will assume that X has $n < \infty$ values in its range with positive probability, since $if \#(R) = \infty$, there is nothing to prove. We will show that $H(X) - \log(n) \leq 0$, which is equivalent to the statement in

the theorem.

$$\begin{aligned}
H(X) - \log(n) &= -\left(\sum_{x\in R} f(x)\log(f(x))\right) - \log(n) \\
&= -\left(\sum_{x\in R} f(x)\log(f(x))\right) - \log(n)\sum_{x\in R} f(x) \\
&\qquad\qquad (\text{since } \textstyle\sum_{j=1} f(x) = 1) \\
&= -\left(\sum_{x\in R} f(x)\left[\log(f(x)) + \log(n)\right]\right) \qquad (\text{Distributive}) \\
&= -\left(\sum_{x\in R} f(x)\log(f(x)\cdot n)\right) \qquad (\text{Frop. of } \log(\cdot)) \\
&= \left(\sum_{x\in R} f(x)\log\left(\frac{1}{f(x)\cdot n}\right)\right) \qquad (\text{Frop. of } \log(\cdot)) \\
&\leq \left(\sum_{x\in R} f(x)\left(\frac{1}{f(x)\cdot n} - 1\right)\right) \qquad (\text{Lemma 11.3}) \\
&= \left(\sum_{x\in R}\left(\frac{1}{n} - f(x)\right)\right) \\
&= \left(\sum_{x\in R}\frac{1}{n} - \sum_{x\in R} f(x)\right) \\
&= (1-1) \\
&= 0
\end{aligned}$$

And so, $H(X) - \log(n) \leq 0$, and $H(X) \leq \log(n)$, as claimed.

From Lemma 11.3, equality holds if and only if $\frac{1}{f(x)\cdot n} = 1$ for all x; that is, iff $f(x)\cdot n = 1$ for all x, which holds iff $f(x) = \frac{1}{n}$ for all x. Namely, when X has a Uniform distribution. □

11.2.4 Entropy and the Principle of Symmetry

Theorem 11.4(b) tells us that, when a random variable can take n different values, we have maximal uncertainty when each value has the same probability. This is a

justification for the Principle of Symmetry: If we really don't have any particular knowledge about a distribution except that there are n possible outcomes, assigning each outcome the probability $\frac{1}{n}$ is a way of conveying maximal uncertainty.

11.3 Conditional Entropy and Mutual Information

11.3.1 Joint Entropy

The probability law of a random variable X gives the probability of each value $x \in \mathbb{R}_X$. The joint distribution function of random variables X and Y gives the probability of each combination of values, $(x, y) \in R_X \times R_Y$.

Given a random variable X, and its probability law, we created the random variable $I(X)$, with values $I(x) = -\log(f(x))$. We called the expected value of this random variable the *entropy* of X.

We can do something analogous with the joint distribution of X and Y. If we observe that $X = x_r$ and $Y = y_s$, we have observed the event $E = \{X = x_r\} \cap \{Y = y_s\}$, and the information gained is

$$I(E) = -\log(P(E)) = -\log(f(x_r, y_s))$$

What is our expected information gain by observing *both X and Y*? The idea is the same as always: take a weighted average of the information values, with the weights corresponding to the (joint) probabilities.

Definition 11.3 (Joint Entropy)**.** *If X and Y are random variables with ranges R_X and R_Y, respectively, then their **joint entropy** is given by*

$$H(X, Y) = -\sum_{x \in R_X} \sum_{y \in R_Y} p(x, y) \log(p(x, y)) \tag{11.5}$$

Notice that the definition is symmetric: $H(X, Y) = H(Y, X)$.

Example: Let X_1 and X_2 have Bernoulli distributions, representing the outcomes of two independent coin flips ($X_j = 1$ when the j^{th} coin comes up heads). Confirm that $H(X_1) = H(X_2) = 1$ bit.

(a) Compute $H(X_1, X_2)$.

(b) Now suppose $X_3 = 1$ when the first coin comes up tails. Note that the distribution of X_3 is the same as that of X_2. Compute $H(X_1, X_3)$.

(c) Let Y represent the total number of heads in two flips. Compute $H(Y)$ and $H(X_1, Y)$.

Notice that $H(X_1, X_2) > H(X_1, X_3)$, even though X_2 and X_3 have the same distribution. Meanwhile, $H(X_1, X_2) = H(X_1, Y)$, even though $H(Y) = 1.5 > 1 = H(X_2)$. Why is that?

Together X_1 and X_2 tell you everything about the outcome of the two flips. If you know both, you know exactly what happened. But the same is true for X_1 and Y together: if you know how the first coin came out, and you know how many heads there were in total, you can deduce the outcome of the second flip.

Meanwhile, X_3 doesn't tell you anything new once you know X_1, so learning both yields the same amount of information as learning X_1 alone.

11.3.2 Conditional Information

When two events (call them E and F) are independent, the information conveyed by their intersection is the sum of the information conveyed by each event individually: $I(E \cap F) = I(E) + I(F)$.

When they're not independent, the information in their intersection is the sum of (1) the information conveyed by E alone, and (2) the information contained in F *that we didn't already have* from E.

How can we determine the second piece? Recall that information in an event is detemined by its probability: $I(E) = -\log(P(E))$. Once we know E, what's the probability of F? That's exactly its *conditional probability.*

Definition 11.4. *The* ***conditional information*** *associated with F, having observed E is*

$$I(F|E) = -\log(P(F|E))$$

So then, we have

$$\begin{aligned} I(E \cap F) &= -\log(P(E \cap F)) \\ &= -\log(P(E)P(F|E)) \\ &= -\left[\log(P(E)) + \log(P(F|E))\right] \\ &= I(E) + I(F|E) \end{aligned}$$

Returning to our previous example, what is $I(Y = 2|X = 1)$? In words, it's the additional information gained by learning that there were two heads, once we knew the first coin was heads. Computing it:

$$\begin{aligned} I(Y = 2|X = 1) &= -\log_2\left(\frac{P(\{X = 1\} \cap \{Y = 2\})}{P(X = 1)}\right) \\ &= -\log\left(\frac{1/4}{1/2}\right) \\ &= -\log\left(\frac{1}{2}\right) = 1 \end{aligned}$$

So we get one additional bit of information by learning that there were two heads. This is exactly the information conveyed by the second flip coming up heads.

11.3.3 Conditional Entropy

Equipped with the concept of conditional information, we can now ask, how much additional information do we *expect* to get by observing Y (say), given that we've already observed $X = x_r$ (say)? We do the same thing we've done all along: take a weighted average of the conditional information associated with each value of Y, weighted by its probability *given all available data* (in this case, that $X = x_r$). This gives us the **conditional entropy** of Y, given that $X = x_r$.

Motivation and Definitions

Definition 11.5 (Conditional Entropy)**.** *Let X and Y be discrete random variables with ranges R_X and R_Y, respectively. The **conditional entropy** of Y given that $X = x_r$ is*

$$H(Y|X = x_r) = -\sum_{y \in R_Y} f(y|x_r)\log(f(y|x_r)) \tag{11.6}$$

As before, we adopt the convention that $0 \cdot \log(0) = 0$.

Conditional entropy is an instance of a **conditional expectation**, which you examined in Exercise 6.40. To restate:

Definition 11.6 (Conditional Expectation)**.** *Let X and Y be discrete random variables with ranges R_X and R_Y, respectively. The **conditional expectation** of Y given $X = x_r$ is defined as*

$$\mathbb{E}\left[Y|X = x_r\right] = \sum_{y \in R_Y} y \cdot f_(y|x_r)$$

More generally, we have the conditional expectation of a *function* of Y:

Definition 11.7 (Conditional Expectation Generalized)**.** *In addition, let h be a real-valued function defined on R_Y. Then the conditional expectation of $h(Y)$ given $X = x_r$ is*

$$\mathbb{E}\left[h(Y)|X = x_r\right] = \sum_{y \in R_Y} h(y) \cdot f_{Y|X}(y|x_r)$$

Since $I(Y = y|X = x_r)$ is a real-valued function of the value of Y, we can set $h(y) = I(Y = y|X = x_r)$ in Def. 11.7 to get

$$\begin{aligned}
\mathbb{E}\left[I(Y|X = x_r)\right] &= \sum_{y \in R_Y} I(Y = y|X = x_r) f(y|x_r) \\
&= \sum_{y \in R_Y} -\log(f(y|x_r)) f(y|x_r) \\
&= H(Y|X = x_r)
\end{aligned}$$

Notice that we are averaging over all possible values of Y when we compute conditional entropy, whereas, we are fixing a *particular* value of X. We are assuming that we have observed X and it has realized the value x_r. So, for each different value of r that we choose, we will get a different value for the conditional entropy of Y.

If we view the conditional expectation of Y given that $X = x_r$ *as a function of* x_r, then the result is actually a random variable, $\mathbb{E}\left[Y|X\right]$, with range

$$R_{\mathbb{E}[Y|X]} = \{\mathbb{E}\left[Y|X = x_1\right], \mathbb{E}\left[Y|X = x_2\right], \dots\}$$

and probabilities given by

$$P(\mathbb{E}\left[Y|X\right] = \mathbb{E}\left[Y|X = x_r\right]) = P(X = x_r)$$

The same logic applies if we treat $H(Y|X = x)$ as a function of x: we get a random variable $H(Y|X)$ with range

$$R_{H(Y|X)} = \{H(Y|X = x_1), H(Y|X = x_2), \dots\}$$

and probabilities

$$P(H(Y|X) = H(Y|X = x_r)) = P(X = x_r)$$

If we take the expected value of this random variable, we're taking an expectation of an expectation. We call this value $H(Y|X)$ (note that before we had $H(Y|X = x)$ for a *particular* x value, whereas now we're averaging over x values):

$$H(Y|X) = \mathbb{E}\left[H(Y|X = x)\right] = \mathbb{E}\left[\mathbb{E}\left[I(Y|X)\right]\right] = \sum_{x \in R_X} \mathbb{E}\left[I(Y|X = x)\right] f(x) \quad (11.7)$$

The first expectation tells us the expected information remaining to be gained by observing Y, once we have observed a particular value of X. Or, flipping things around, it tells us how much uncertainty about the value of Y we have left after we observe that $X = x_r$ (remember, uncertainty is the same as potential information gain).

The second expectation, this time over the distribution of X, tells us how much uncertainty about Y we *expect* to have left after we observe X (without saying what value X will take).

Working out the equation, we get

Lemma 11.6.

$$H(Y|X) = -\sum_{x \in R_X} \sum_{y \in R_Y} f(x, y) \log(f(y|x)) \quad (11.8)$$

Proof.

$$\begin{aligned}
H(Y|X) &= \mathbb{E}\left[\mathbb{E}\left[I(Y|X)\right]\right] \\
&= \mathbb{E}\left[-\sum_{y\in R_Y} f(y|x)\log(f(y|x))\right] \\
&= -\sum_{x\in R_X}\left(\sum_{y\in R_Y} f(y|x)\log(f(y|x))\right) f(x) \\
&= -\sum_{x\in R_X}\sum_{y\in R_Y} f(y|x)f(x)\log(f(y|x)) \\
&= -\sum_{x\in R_X}\sum_{y\in R_Y} f(x,y)\log(f(y|x))
\end{aligned}$$

□

Properties of Conditional Entropy

If X and Y are independent, then we do not expect any change in our uncertainty about Y after observing X. Hence, we should have

Lemma 11.7. *If X and Y are independent, then*

$$H(Y|X) = H(Y)$$

Proof. From Lemma 11.6, we have

$$\begin{aligned}
H(Y|X) &= -\sum_{x\in R_X}\sum_{y\in R_Y} f(x,y)\log(f(y|x)) \\
&= -\sum_{x\in R_X}\sum_{y\in R_Y} f(x)f(y)\log(f(y)) && \text{(by Independence of } X \text{ and } Y\text{)} \\
&= \sum_{x\in R_X} f(x)H(Y) \\
&= H(Y)\sum_{x\in R_X} f(x) \\
&= H(Y) && \text{(by properties of probability laws)}
\end{aligned}$$

□

Recall that we interpret the conditional entropy $H(Y|X)$ as the amount of (remaining) uncertainty we expect to have about Y after we observe X. Consider what happens if we observe X and then Y. First, upon observing $X = x$, we reduce our uncertainty by $I(X = x)$. Then, upon observing $Y = y$, we reduce our uncertainty by an additional $I_{X=x}(Y = y)$. So, overall, our uncertainty has been reduced by $I(\{X = x\} \cap \{Y = y\}) = I(X = x) + I_{X=x}(Y = y)$.

By how much do we *expect* to reduce our uncertainty if we plan to observe X and Y but haven't done so yet? We have already seen that this is $H(X, Y)$. Breaking this into stages, when we observe X, we expect a reduction in uncertainty of $H(X)$ bits. When we then observe Y, we expect (without knowing what X value we will have observed) a reduction of $H(Y|X)$ bits. Hence, we expect a total reduction in uncertainty of $H(X, Y) = H(X) + H(Y|X)$ bits.

This is analogous to the way we can break apart a set $A \cup B$ into two disjoint pieces: A and $B - A$. First take A in its entirety, then take the part of B which is left over.

Is the intuition correct that in fact the joint entropy is the sum of these two pieces? The next theorem shows that it is.

Theorem 11.8 (Decomposition of Joint Entropy)**.** *Let X and Y be discrete random variables with ranges R_X and R_Y. Then*

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Proof. Starting off by rewriting the left-hand-side by applying Def. 11.3, we get

$$\begin{aligned}
H(X, Y) &= -\sum_{x \in R_X} \sum_{y \in R_Y} f(x, y) \log(f(x, y)) \\
&= -\sum_{x \in R_X} \sum_{y \in R_Y} f(x, y) \log(f(x) f(y|x)) && \text{(Decomposition rule)} \\
&= -\sum_{x \in R_X} \sum_{y \in R_Y} f(x, y) \log(f(x)) - \sum_{x \in R_X} \sum_{y \in R_Y} f(x, y) \log(f(y|x)) \\
&= -\sum_{x \in R_X} \log(f(x)) \sum_{y \in R_Y} f(x, y) + H(Y|X) && \text{(Lemma 11.6)} \\
&= -\sum_{x \in R_X} \log(f(x)) f(x) + H(Y|X) && \text{(Marginalization Lemma)} \\
&= H(X) + H(Y|X) && \text{(Def. 11.2)}
\end{aligned}$$

This establishes the first equality. Since $H(X,Y) = H(Y,X)$ we can apply the exact same proof to get the remaining equality. □

Combining this Theorem with Lemma 11.7, we get the following useful Corollary:
Corollary 11.9. *If X and Y are independent, then*

$$H(X,Y) = H(X) + H(Y)$$

11.3.4 Mutual Information

Motivation and Definition

In Theorem 11.8, we decomposed the total information in X and Y together into two pieces: one due to X, and one unique to Y (or, if we want, one due to Y, and one unique to X). What about the information contained in Y itself?

Conceptually we can talk about the information in Y itself containing two parts. The first is *not* shared with X (this is the conditional entropy of Y given X, which we already defined). The second is the part which *is* shared with X. This is similar to the way we can break apart a set A into two disjoint pieces:

$$A = (A \cap B) \cup (A \cap \overline{B}) = (A \cap B) \cup (A - B)$$

The conditional entropy $H(Y|X)$ is analogous to the $A - B$ part above. The total entropy $H(Y)$ is analogous to A itself.

The intersection part is analogous to what is called the **mutual information** of X and Y.

Definition 11.8 (Mutual Information)**.** *The **mutual information** of X and Y is written $I(X,Y)$, and defined as*

$$I(X,Y) = H(Y) - H(Y|X) \tag{11.9}$$

Again, this is similar to the way in which we can write

$$A \cap B = A - (A - B)$$

Properties of Mutual Information

The next theorem (a) gives us a formula to compute mutual information which is easier to use than directly applying the defintion; (b) tells us that mutual information is symmetric (as we would expect from the set intersection analogy); that independent random variables do not share any information (as we should also expect by the conceptual meaning of independence).

Theorem 11.10 (Properties of Mutual Information)**.** *Let X and Y be discrete random variables with ranges R_X and R_Y, respectively. Then*
(a)

$$I(X,Y) = \sum_{x \in R_X} \sum_{y \in R_Y} f(x,y) \log \left(\frac{f(x,y)}{f(x)f(y)} \right)$$

(b)

$$I(X,Y) = I(Y,X)$$

(c) If X and Y are independent, then

$$I(X,Y) = 0$$

Proof. Guidelines for a proof were given in class. Fill out the details for yourself. □

11.4 Exercises

1. A random variable, X, takes three possible values, which occur with probabilities 0.1, 0.3 and 0.6. Find the information associated with each event, and then find $H(X)$.

2. A word in a code consists of five binary digits. Each digit is chosen independently of the others and the probability of any particular digit being a 1 is 0.6. Find the information associated with the following events

   (a) at least three 1s

   (b) at most four 1s

   (c) exactly two 0s

3. Let Y be an arbitrary discrete random variable with range $R = \{y_1, \ldots, y_n\}$ and PMF f. Show that

$$2^{-H(Y)} = f(y_1)^{f(y_1)} \cdot \ldots \cdot f(y_n)^{f(y_n)}$$

(where entropy is assumed to be measured in bits)

4. Show that the following information-theoretic version of Bayes' theorem holds

$$2^{H(X|Y)} = \frac{2^{H(X)}2^{H(Y|X)}}{2^{H(Y)}}$$

5. Recall that the "Bayesian update factor" from a prior on X to a posterior given the observation $Y = y$ is

$$u(x, y) = \frac{f(y|x)}{f(y)}$$

Show that $\mathbb{E}\left[\log(u(X, Y))\right]$ is equal to the mutual information $I(X; Y)$.

6. (*) Let X_1 and X_2 be discrete random variables with common range R, and respective PMFs p and q. Recall that the KL divergence of X_2 from X_1 is defined as

$$D(X\|Y) = \sum_{x \in R} p(x) \log(\frac{p(x)}{q(x)})$$

   (a) Show that $D(p\|q) \geq 0$, with equality if and only if $p(x) = q(x)$ for all $x \in R$. (Hint: Use the lemma proved in class that says that $\log(x) \leq x - 1$ with equality iff $x = 1$)

   (b) Show that if Y has a uniform distribution, then

$$D(X\|Y) = \log(|R|) - H(X)$$

   where $|R|$ denotes the cardinality of R.

   (c) Let $W = (X, Y)$; that is, let W be a random vector whose range is all combinations of X and Y. Let Z be analogous random vector obtained if X and Y are assumed to be independent. Show that

$$D(W\|Z) = I(X; Y)$$

7. Show that for any discrete random variables X and Y,

$$H(Y|X) \leq H(Y)$$

with equality iff X and Y are independent (Hint: use part (a) of the previous problem)

8. Show that $I(X;Y) \geq 0$ with equality iff X and Y are independent. (Hint: use a previous exercise)

Chapter 12

Markov Chains

Note: the material in this section corresponds to Chapter 10, not Chapter 7, of Applebaum.

12.1 Stochastic Processes

Consider an example we looked at earlier: the simple symmetric random walk. Recall the setup: we begin at a position labeled 0. At every time step, we flip a coin to decide whether to step left or right. We defined a sequence of $I.I.D.$ random variables, Y_1, Y_2, etc., which represented the result of our coin flip: each one had a probability of $1/2$ of taking the value -1, in which case we step left, and a probability of $1/2$ of taking the value 1, in which case we step right.

For the random walk, we can represent our position at time t as the sum of the first t random variables:

$$S_t = Y_1 + Y_2 + \cdots + Y_t$$

Notice that if we consider the sequence of random variables $X_0, X_1, X_2, \ldots$, these are neither independent nor identically distributed. X_0 is assumed to equal 0 with probability 1. Then we have

$$p_{X_1}(1) = p_{Y_1}(1) = \frac{1}{2}$$
$$p_{X_1}(-1) = p_{Y_1}(-1) = \frac{1}{2}$$

Marginally, X_2 will be equal to -2 with probability 1/4 (if $Y_1 = Y_2 = -1$), will be equal to 0 with probability 2/4, since we have

$$\{X_2 = 0\} = (\{Y_1 = -1\} \cap \{Y_2 = 1\}) \cup (\{Y_1 = 1\} \cap \{Y_2 = -1\})$$

and will be equal to 2 with probability 1/4 (when $Y_1 = Y_2 = 1$).

But consider the distribution of X_2 conditioned on X_1. We have

$$\begin{aligned} P_{X_1=-1}(X_2 = -2) = P(Y_2 = -1) = \frac{1}{2} \neq P(X_2 = -2) \\ P_{X_1=-1}(X_2 = 0) = P(Y_2 = 1) = \frac{1}{2} \\ P_{X_1=1}(X_2 = 0) = P(Y_2 = -1) = \frac{1}{2} \\ P_{X_1=1}(X_2 = 2) = P(Y_2 = 1) = \frac{1}{2} \neq P(X_2 = 2) \end{aligned}$$

So what we have in $X_1, X_2, \dots$ is a sequence of *dependent* random variables, which collectively describe the state of a system as it evolves over time. This is an example of a **stochastic process**: "stochastic" because it involves an element of randomness, and "process" because it unfolds over time.

12.1.1 Basic Definitions

Mathematically, we can define a stochastic process as follows.

Definition 12.1 (Stochastic Process)**.** *A **stochastic process** is a collection of random variables, $\{X_t; t \in I\}$, where I is an index set. We can think of I as a set of time points.*

In some cases $I = [0, \infty)$, that is, there is an initial state corresponding to X_0, and we can observe the system at arbitrary, continuous time points. In many cases (and for the purposes of this course), time is discretized, and we observe the process at intervals. Here we will typically have $I = \mathbb{Z}_+ = \{0, 1, 2, \dots\}$, which means that our stochastic process is $\{X_0, X_1, X_2, \dots\}$.

It is convenient to let all the X_ts have the same range, corresponding to all possible states of the system. In the case of the simple symmetric random walk, this wasn't the case: $R_{X_0} = \{0\}$, $R_{X_1} = \{-1, 1\}$, $R_{X_2} = \{-2, 0, 2\}$, etc. But we are free to

extend the ranges of each variable to be the union of the ranges of all the variables in the process, and then set as many probabilities to zero as we need to to preserve the original distributions. In this way, we can always find a single common range. We call this the **state space** of the stochastic process.

Definition 12.2 (State Space)**.** *The **state space**, S, of a stochastic process $\{X_t; t \in I\}$ is the common range of the random variables. If necessary, take unions to ensure that*

$$S = R_{X_1} = R_{X_2} = \dots$$

Question: What is the state space for the simple symmetric random walk?

12.1.2 Example: Random Walks

We have already looked at the simple symmetric random walk, where our steps are determined by a sequence of I.I.D. random variables, each of which is 1 with probability 1/2 and -1 with probability 1/2.

We can easily generalize this by dropping the symmetry requirement, and let $Y_1, Y_2, \dots$ be I.I.D. so that

$$P(Y_t = 1) = p, \qquad P(Y_t = -1) = 1 - p, \qquad \text{for arbitrary } p \in [0, 1]$$

The cumulative sums of the Y_ts define a **random walk**. When $p \neq 1/2$, our random walk is biased, and so we expect it to gradually drift off to the right (i.e., toward $+\infty$)when $p > 1/2$, and to the left (i.e., toward $-\infty$) if $p < 1/2$.

This biased random walk describes most luck-based casino games: if winning corresponds to a step to the right and losing corresponds to a step to the left, the number of wins minus the number of losses is described as a biased random walk, which gradually drifts off to $-\infty$ if we keep playing forever. The bias ensures that the house always wins in the long run.

12.1.3 Some Remarks on Notation

At time t, we can describe everything that's happened so far if we know $X_0, X_1, X_2, dots, X_t$. We can describe a complete "path" up to time t by recording the values of these variables in a vector, $(x_0, x_1, \dots, x_t)$, where each $x_j \in S$.

Joint Probabilities

From the perspective of "predicting the future", we may be interested in the distribution over paths up to a particular time t; that is, in the joint probabilities

$$P(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \cdots \cap \{X_t = x_t\})$$

Writing out everything with brackets and intersections gets pretty cumbersome, so as a shorthand we'll write the above as

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t)$$

Of course, we can use this same notation to denote the joint probabilities of any subset of X_js.

Conditional Probabilities

Because of the dependence of each random variable on the past, it can be difficult to calculate the marginal distribution of any given X_t. However, it is often simpler to look at the *conditional* distribution of X_t, given everything that has happened previously. As with joint probabilities, our usual notation for conditional probability gets very cumbersome when we're conditioning on lots of random variables at once, so we'll revert to the more standard conditional probability notation and write the conditional probability that $X_t = x_t$ given everything before as

$$P(X_t = x_t | X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1})$$

12.2 Markov Chains

Consider the simple random walk. How would we find the probability of being in state k at time t?

In order to be at k after t steps, we either need to be at $k-1$ after $t-1$ steps and go right, or we need to be at $k+1$ after $t-1$ steps and go left. That is

$$P(X_t = k | X_{t-1} = k-1) = \frac{1}{2}, \qquad P(X_t = k | X_{t-1} = k+1) = \frac{1}{2}$$

and for any other value of X_{t-1}, the conditional probability of ending up in state k is zero. Suppose we've observed the walk at every time up to time t, and we want to know the distribution of X_{t+1}. That is, we want to evaluate

$$P(X_{t+1} = k_{t+1} | X_0 = k_0, X_1 = k_1, \ldots, X_t = k_t)$$

How would we calculate this?

Notice that the random walk is such that if we know where we are at t, then our predictions about the future won't be affected if we forget the values of X_0 up to X_{t-1}. This is called the **Markov property**. In words, it says that predictions of the future are no better if we know the entire history of the process than they are if we just know the present state.

A stochastic process with a discrete index set (think time steps, rather than continuous time) that has the Markov property is called a **Markov chain**. If the index set is continuous, the process is called a **Markov process**. Formally

Definition 12.3 (Markov chain)**.** *A stochastic process* $\{X_t; t \in \mathbb{Z}_+\}$ *is called a* ***Markov chain*** *if for every time* t *and every path* $(k_0, k_1, \ldots, k_t, k_{t+1})$*,*

$$P(X_{t+1} = k_{t+1} | X_0 = k_0, \ldots X_t = k_t) = P(X_{t+1} = k_{t+1} | X_t = k_t)$$

This is an instance of what's called **conditional independence**. If we condition on X_t, then X_{t+1} is independent of X_0 through X_{t-1}: further conditioning on them doesn't change the distribution.

If we have a Markov chain, then we can characterize it completely if we know

(1) $P(X_0 = k$, for each k in the state space S

(2) $P(X_1 = k | X_0 = j)$ for each combination $j, k \in S$

(3) $P(X_2 = k | X_1 = j)$ for each $j, k \in S$

(4) etc.

As a shorthand, we can write

$$\pi_k \stackrel{\text{def.}}{=} P(X_0 = k) \tag{12.1}$$

and

$$P_{jk}^{t,t+1} \stackrel{\text{def.}}{=} P(X_{t+1} = k | X_t = j) \tag{12.2}$$

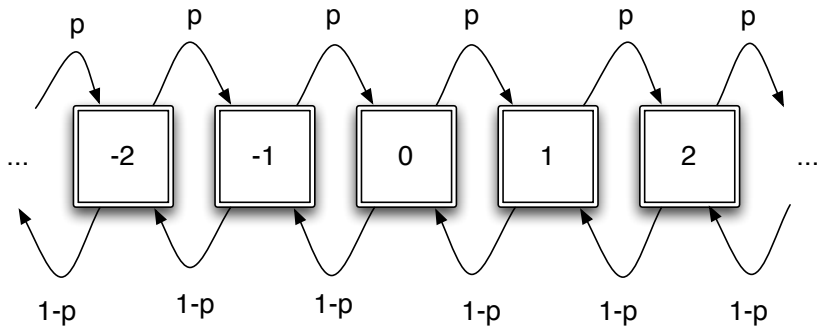

Figure 12.1: State Diagram for the Random Walk

These conditional probabilities are called the (one-step) **transition probabilities** of the Markov chain, as they give us the probability of *transitioning* to state k from state j in one time step.

Definition 12.4 (Stationarity)**.** *If the transition probabilities $P_{jk}^{t,t+1}$ for each ordered pair (j,k) of states do not change over time (that is, if they do not depend on the value of t), then we say that the Markov chain is **stationary**. More formally, a Markov chain is stationary if, for each $j,k \in S$, we have*

$$P_{jk}^{0,1} = P_{jk}^{1,2} = \cdots = P_{jk}^{t,t+1} = \cdots = P_{jk}$$

When we have a stationary Markov chain (with a discrete state space), we can visualize it using a **state diagram**, where we let states be boxes, and connect states to other states by arrows, labeled by the corresponding transition probability. A sample state diagram for the random walk is in Fig. 12.1

For a stationary Markov chain, if we know the π_k for each $k \in S$ and we know the transition probabilities P_{jk} for each $j,k \in S$, then we can compute any probability we want. If the state space is finite (say, $S = \{0,1,\ldots,n\}$), we can represent the intial probabilities as a vector

$$\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_n) \tag{12.3}$$

and the transition probabilities as a (square) matrix

$$\mathbf{P} = \begin{pmatrix} P_{00} & P_{01} & \ldots & P_{0n} \\ P_{10} & P_{11} & \ldots & P_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ P_{n0} & P_{n1} & \ldots & P_{nn} \end{pmatrix} \tag{12.4}$$

This **transition matrix** is the "soul" of the stochastic process, as it describes the actual "process" part (the initial probabilities just tell us how to get in the door). The interesting theoretical properties of stochastic processes depend on this transition matrix, as the influence of the initial probabilities disappears over time, provided the transition matrix is well-behaved ("well-behaved" is well-defined, but beyond the scope of this course, unfortunately).

Notice that transition matrices satisfy the following properties:

(i) $P_{jk} \geq 0$ for all $j, k \in S$

(ii) $\sum_k P_{jk} = 1$ for each $j \in S$

Proof. Property (i) is obvious from the definition of P_{jk} as a conditional probability. Property (ii) also follows from the properties of conditional distributions (e.g. from Exercise 5.40(i), which you did for HW). If fix the value I'm conditioning on, then the conditional distribution must sum to 1. Notice that (ii) is asymmetric. In terms of the matrix representation, it says that the *rows* must sum to 1. There is no reason, however, that the columns should sum to 1. A simple counterexample is the (rather silly) chain in which every state transitions to 0 with probability 1. The transition matrix here will have 1s in every entry in the first column, and 0s everywhere else.

Note that if the Markov chain is not stationary, then we need a sequence of transition matrices, $(\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_t, \ldots)$, where the entry in the j^{th}row and k^{th}column of the matrix $\mathbf{P}_t$ represents $P(X_t = k | X_{t-1} = j)$. □

12.2.1 Example: Gambling With Finite Resources

Suppose two players have a total of M silver dollars between them. They repeatedly place bets, in which Player 1 has a probability of p of winning \$1 from player 2, and a probability of $1 - p$ of losing \$1 to player 2. This continues until one player has all \$M.

We can model the evolving bankroll of Player 1 with a Markov chain. Since the probability of winning stays constant over time, the chain is stationary.

Question: What's its state space? What's the transition matrix? Try drawing a state diagram.

Answer: Since at any given time the gambler can have anywhere from 0 to M dollars, the state space is $S = \{0, 1, \ldots, M\}$.

The first row of the transition matrix corresponds to the conditional distribution of the gambler's stash after starting with \$0. Since this is a game-ending state, she will stay here with probability 1.

The second row is the distribution when she has exactly \$1. She has probability p of ending up with \$2, and probability $1 - p$ of ending up with \$0. Similarly, if she starts with \$2, she has probability p of ending up with \$3 and probability $1 - p$ of ending up with \$1. And so on. Finally, if she starts with \$M, the game is over, and she will keep all \$M with probability 1. Altogether, we get

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \ldots & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & \ldots & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \ldots & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

12.2.2 Calculating Joint Probabilities

Given a Markov chain, how can we calculate the probability of a path? Intuitively, if we want to find the probability of the path $(k_0, k_1, k_2, \ldots, k_t)$, it seems like we should be able to take the probability of starting at k_0, multiply by the conditional probability of going to k_1 from k_0, then by the conditional probability of going to k_2 from k_1, etc. In fact, this works, because of the Markov property, as well as the "chain rule" of conditional probability:

Lemma 12.1 (Chain Rule of Conditional Probability)**.** *Let $A_1, A_2, \ldots, A_n$ be events. Then*

$$P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\ldots P(A_n|A_1, \ldots, A_{n-1})$$

where the commas are shorthand for intersections.

Proof. This a straightforward inductive generalization of the "decomposition rule". For a base case, take $n = 2$. Then the result follows from that rule. Now suppose

the result holds for some $n = m$. Then we can write

$$P(A_1, \ldots, A_m, A_{m+1}) = P(A_1, \ldots, A_m)P(A_{m+1}|A_1, \ldots, A_m)$$

The inductive hypothesis gives us that

$$P(A_1, \ldots, A_m) = P(A_1)P(A_2|A_1) \cdots P(A_m|A_1, \ldots, A_{m-1})$$

Substituting this into the equation above gives us the result. □

Theorem 12.2 (Path Probabilities)**.** *Let $\{X_t; t \in \mathbb{Z}_+\}$ be a stationary Markov chain with transition matrix $\mathbf{P} = (P_{jk})$. Then we have*

$$P(X_0 = k_0, X_1 = k_1, \ldots, X_t = k_t) = \pi_{k_0} P_{k_0,k_1} P_{k_1,k_2} \cdots P_{k_{t-1},k_t}$$

Proof. By the Chain Rule, we can write

$$\begin{aligned} P(X_0 = k_0, \ldots, X_t = k_t) &= P(X_0 = k_0)P(X_1 = k_1|X_0 = k_0) \\ &\quad \times P(X_2 = k_2|X_0 = k_0, X_1 = k_1) \\ &\quad \times \cdots \times P(X_t = k_t|X_0 = x_0, \ldots, X_{t-1} = k_{t-1}) \\ &= P(X_0 = k_0)P(X_1 = k_1|X_0 = k_0) \\ &\quad \times P(X_2 = k_2|X_1 = k_1) \qquad \text{(Markov Property)} \\ &\quad \times \cdots \times P(X_t = k_t|X_{t-1} = k_{t-1}) \\ &= \pi_{k_0} P_{k_0,k_1} P_{k_1,k_2} \cdots P_{k_{t-1},k_t} \qquad \text{(Def. of } \pi \text{ and } \mathcal{P}_{jk}) \end{aligned}$$

□

Example 1

Suppose we have a Markov chain with state space $S = \{0, 1, 2\}$, initial distribution $\boldsymbol{\pi} = (0.25, 0.5, 0.25)$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0.3 & 0.2 & 0.5 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

Find the probability of the path $(0, 2, 1)$.

Solution By Theorem 12.2, we have

$$\begin{aligned} P(X_0 = 0, X_1 = 2, X_2 = 1) &= \pi_0 P_{02} P_{21} \\ &= 0.25 \times 0.5 \times 0.4 \\ &= 0.05 \end{aligned}$$

12.3 The Chapman-Kolmogorov Equations

In many cases we don't care about the probabilities of complete paths; we just want to know how likely we are to be in state b at a particular time (perhaps given that we are in state a to start). We could, of course, list all possible paths from a to b, compute their probabilities using Theorem 12.2, but this would be extremely laborious! Fortunately, there's a more elegant way to do this.

12.3.1 Two-Step Transition Matrix

To start simple, consider how you would compute the probability that $X_2 = k_2$ given that $X_0 = k_0$. By the Marginalization Lemma, and Theorem 12.2, we have

$$\begin{aligned} P(X_2 = k_2 | X_0 = k_0) &= \sum_{k_1 \in S} P(X_1 = k_1, X_2 = k_2 | X_0 = k_0) \\ &= \sum_{k_1 \in S} \frac{P(X_0 = k_0, X_1 = k_1, X_2 = k_2)}{P(X_0 = k_0)} \\ &= \sum_{k_1 \in S} \frac{\pi_{k_0} P_{k_0,k_1} P_{k_1,k_2}}{\pi_{k_0}} \\ &= \sum_{k_1 \in S} P_{k_0,k_1} P_{k_1,k_2} \end{aligned}$$

Using this expression, we can construct a matrix of two-step transition probabilities.

$$\mathbf{P}^{(2)} = \left(P_{ij}^{(2)} \right) \qquad \text{where } P_{ij}^{(2)} = \sum_{k \in S} P_{ik} P_{kj}$$

If you're familiar with some matrix algebra, you'll recognize this as the definition of $\mathbf{P}^2$. If you're not, here's what you need to know. If $A = (a_{ik})$ is an $n \times p$ matrix and

$B = (b_{kj})$ is a $p \times m$ matrix (that is, if B has the same number of rows that A has columns), then the product AB is an $n \times m$ matrix, and the entry in its i^{th} row and j^{th} column is

$$\sum_{k=1}^{p} a_{ik} b_{kj}$$

In the specific case where A is a square (say, $n \times n$) matrix, then A^2 is defined to be AA, and the entry in its i^{th} row and j^{th} column is

$$\sum_{k=1}^{n} a_{ik} a_{kj}$$

Hence from above, for a stationary Markov chain, its two-step transition matrix $\mathbf{P}^{(2)}$ is equal to $\mathbf{P}^2$.

12.3.2 N-step Transition Matrix

That covers the case of two steps; what about an arbitrary number of steps (say, n)?

Theorem 12.3 (Chapman-Kolmogorov)**.** *Let $\{X_t; t \in \mathbb{Z}_+\}$ be a stationary Markov chain with a finite state space S and transition matrix $\mathbf{P}$. Its n-step transition matrix is given by*

$$\mathbf{P}^{(n)} = \mathbf{P}^n$$

where $\mathbf{P}^n$ is the n-fold matrix product of $\mathbf{P}$.

Proof. First a side-note on matrix algebra: since $\mathbf{P}$ is a square matrix, $\mathbf{P}^2$ is also square with the same number of rows and columns, and so we can define $\mathbf{P}^3 = \mathbf{P} \times \mathbf{P}^2$, and so on to $\mathbf{P}^n$.

We proceed by induction. The case $n = 1$ is a tautology (actually, we proved the theorem for $n = 2$ above, but we didn't actually need to treat that case separately).

Now suppose the theorem holds for some $n = m$. Then consider the entry $P_{ij}^{(m+1)}$; that is,

$$P_{ij}^{(m+1)} = P(X_{m+1} = j | X_0 = i)$$

We have, by the Marginalization Lemma, the definition of conditional probability, the chain rule, and the Markov property

$$\begin{aligned}
P(X_{m+1} = j | X_0 = i) &= \sum_{k \in S} P(X_{m+1} = j, X_m = k | X_0 = i) \\
&= \sum_{k \in S} \frac{P(X_{m+1} = j, X_m = k, X_0 = i)}{P(X_0 = i)} \\
&= \sum_{k \in S} \frac{P(X_0 = i) P(X_m = k | X_0 = i) P(X_{m+1} = j | X_m = k, X_0 = i)}{P(X_0 = i)} \\
&= \sum_{k \in S} P(X_{m+1} = j | X_m = k, X_0 = i) P(X_m = k | X_0 = i) \\
&= \sum_{k \in S} P(X_m = k | X_0 = i) P(X_{m+1} = j | X_m = k) \\
&= \sum_{k \in S} P_{ik}^{(m)} P_{kj}
\end{aligned}$$

This is, by definition, the entry in the i^{th}row and j^{th}column of the matrix $\mathbf{P}^{(m)} \times \mathbf{P}$, and so

$$\mathbf{P}^{(m+1)} = \mathbf{P}^{(m)} \times \mathbf{P}$$

But by the induction hypothesis, $\mathbf{P}^{(m)} = \mathbf{P}^m$, and so we have

$$\mathbf{P}^{(m+1)} = \mathbf{P}^m \times \mathbf{P} = \mathbf{P}^{m+1}$$

This establishes the induction step, and so the theorem holds for all n. □

This is great — now, given a stationary Markov chain, if we know the chain is in state j at some time s (that is, we know $X_s = j$, we can compute the distribution of X_{s+t} simply by computing $\mathbf{P}^t$ (which is easily done on the computer, with R or MATLAB, for example) and reading out its j^{th} row!

Example 2

Suppose we have a two-state Markov chain with $S = \{0, 1\}$

$$\mathbf{P} = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$$

Find

(i) $\mathbf{P}^{(2)}$

(ii) $P(X_2 = 1|X_0 = 0)$

Solution By Theorem 12.3, we have

$$\begin{aligned}\mathbf{P}^{(2)} = \mathbf{P} \times \mathbf{P} &= \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix} \times \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix} \\ &= \begin{pmatrix} p^2 + (1-p)^2 & p(1-p) + (1-p)p \\ (1-p)p + p(1-p) & (1-p)^2 + p^2 \end{pmatrix}\end{aligned}$$

Then $P(X_2 = 1|X_0 = 0) = P_{01}^{(2)} = 2p(1-p)$

To confirm that what we did makes sense, let's check that the rows of the two-step transition matrix sum to 1.

$$p^2 + (1-p)^2 + 2p(1-p) = (p + (1-p))^2 = 1^2 = 1$$

12.3.3 Marginal Distributions Down the Chain

From here, computing the marginal distribution of any X_t is straightforward. We have

$$\begin{aligned}P(X_t = k) &= \sum_{i \in S} P(X_t = k, X_0 = i) \\ &= \sum_{i \in S} P(X_0 = i) P(X_t = k | X_0 = i) \\ &= \sum_{i \in S} \pi_i P_{ik}^{(t)}\end{aligned}$$

Notice, though, that if we treat $\boldsymbol{\pi}$ as a $1 \times p$ matrix (where $p = \#(S)$), this is just the k^{th} entry in the matrix that we get when we multiply $\boldsymbol{\pi} \times \mathbf{P}^{(t)}$. Therefore, we get

Theorem 12.4 (Marginal Distributions of Stationary Markov Chain)**.** *Let $\{X_t; t \in \mathbb{Z}_+\}$ be a stationary Markov chain with a finite state space S, initial distribution $\boldsymbol{\pi}^{(0)}$ and transition matrix $\mathbf{P}$. Denote the marginal distribution of X_t as a row vector $\boldsymbol{\pi}^{(t)}$. Then*

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^{(0)} \mathbf{P}^t$$

Example 3

Return to the chain from Example 2, and suppose $\boldsymbol{\pi}^{(0)} = (q,\ 1\ -\ q)$. Find the marginal distribution of X_2.

Solution By Theorem 12.4, we have

$$\boldsymbol{\pi}^{(2)} = \boldsymbol{\pi}^{(0)}\mathbf{P}^2$$

where from Example 2 we calculated that

$$\mathbf{P}^2 = \begin{pmatrix} p^2 + (1-p)^2 & 2p(1-p) \\ 2p(1-p) & (1-p)^2 + p^2 \end{pmatrix}$$

Putting these together, we get

$$\begin{aligned} P(X_2 = 0) &= q(p^2 + (1-p)^2) + 2(1-q)p(1-p) \\ P(X_2 = 1) &= 2qp(1-p) + (1-q)(p^2 + (1-p)^2) \end{aligned}$$

You should check that this sums to 1.