A PHILOSOPHICAL INTRODUCTION

TO

FORMAL LOGIC

By Jonathan Livengood
Freely you have received; freely give.
— Matthew 10:8
A Philosophical Introduction to Formal Logic

Brief Table of Contents

Chapter 0: Introduction
Chapter 1: Zeroth-Order Logic
Chapter 2: Natural Deduction for Zeroth-Order Logic
Chapter 3: First-Order Logic
Chapter 4: Natural Deduction for First-Order Logic
Chapter 5: Set Theory
Chapter 6: Probability Theory
Chapter 7: Causal and Statistical Reasoning
A Philosophical Introduction to Formal Logic

Chapter 0: What is Logic?

Logic is the normative study of reasoning. In other words, logic is the study of what makes reasoning good or bad. Since it is a normative study of reasoning, logic is different from psychology. The psychologist wants to know (among other things) how we actually reason. By contrast, the logician wants to know how we ought to reason, even if we never actually reason as we should. And yet, logic has a psychological subject insofar as reasoning is a mental process.

Now, the process of reasoning is a kind of movement from some collection of beliefs that the reasoner accepts at some time to some possibly different collection of beliefs that the reasoner accepts at a later time. But in order to facilitate the critical evaluation of reasoning, the logician abstracts away from the messy details of how the reasoning process is implemented by humans and represents instances of reasoning as static, linguistic objects, called arguments.¹

¹ Not all logicians have taken the view of logic taken here; however, the view here is a traditional one. For example, Whately (1811) writes: “Logic, in the most extensive sense which the name can with propriety be made to bear, may be considered as the Science, and also as the Art, of Reasoning. It investigates the principles on which argumentation is conducted, and furnishes rules to secure the mind from error in its deductions. Its most appropriate office, however, is that of instituting an analysis of the process of the mind in Reasoning; and in this point of view it is, as I have said, strictly a Science: while, considered in reference to the practical rules above mentioned, it may be called the Art of Reasoning.” Mill (1858) writes: “Logic is not the science of Belief, but the science of Proof, or Evidence. So far forth as belief professes to be founded upon proof, the office of logic is to supply a test for ascertaining whether or not the belief is well grounded.” Venn (1876) writes: “It is impossible to direct attention too prominently to the fact that logic (and therefore Probability as a branch of logic) is not concerned with what men do believe, but with what they ought to believe, if they are to believe correctly.” Peirce (1877) writes: “The object of reasoning is to find out, from the consideration of what we already know, something else which we do not know. Consequently, reasoning is good if it be such as to give a true conclusion from true premises, and not otherwise. Thus, the question of its validity is purely one of fact and not of thinking. A being the premises and B the conclusion, the question is, whether these facts are really so related that if A is B is. If so, the inference is valid; if not, not. It is not in the least the question whether, when the premises are accepted by the mind, we feel an impulse to accept the conclusion also. It is true that we do generally reason correctly by nature. But that is an accident; the true conclusion would remain true if we had no impulse to accept it; and the false one would remain false, though we could not resist the tendency to believe in it.” Moreover, this tradition is still very much alive today, for an example of which, see Priest (2000), who writes: “Logic is the study of what counts as a good reason for what, and why.”
An argument is a collection of sentences that are related to one another in a special way: some of the sentences, called the premisses of the argument, are intended to evidentially support another one of the sentences, called the conclusion of the argument. When an argument is good, the premisses succeed in supporting the conclusion. If the premisses of an argument fail to support its conclusion, then the argument is bad. In many cases, we have an intuitive, pre-theoretical sense for when an argument is good or bad. For example, consider the following argument:

\[
\begin{align*}
\text{Every politician is a liar.} \\
\text{Frank is a politician.} \\
\text{Frank is a liar.}
\end{align*}
\]

Even if you have never taken logic before, you probably recognize argument (1) as a good argument. And even if you cannot articulate exactly why argument (1) is a good argument, you still probably recognize that it is a good argument. If it is not obvious to you that argument (1) is a good argument, that’s okay for now. Regardless of whether or not you find the argument obviously good, keep the following question in mind: What is it about a good argument that makes it a good argument? Please stop and reflect on this question for a few minutes before continuing on.

When asking yourself what it is about a good argument that makes it a good argument, you might find it helpful to think about some examples of bad arguments as well. Compare argument (1) above with the following argument:

\[
\begin{align*}
\text{Every politician is a liar.} \\
\text{Suzy is not a politician.} \\
\text{Suzy is a liar.}
\end{align*}
\]

In writing out arguments, I will follow this stylistic convention: I will write the premisses of the argument above a solid line, and I will write the conclusion of the argument below the line. Sometimes, I will give a separate name to each line in an argument. For example, in an argument with two premisses and a conclusion, I might label the first premiss [A1], the second premiss [A2], and the conclusion [A3].
I claim that argument (2) is not a good argument. The premisses of argument (2) do not
evidentially support the conclusion of argument (2). Are you convinced? If so, ask yourself why
argument (2) is a bad argument. If you are not immediately convinced that argument (2) is a bad
argument, that’s okay for now. Learning about what makes an argument good or bad is the main
point of this text, after all!

We are now in a position to restate what it is that logic is all about. Logic is the study of
the relation of *evidential support*. The logician wants to know when some premisses are evidence
for the truth of some conclusion. But knowing that some premisses evidentially support some
conclusion does not necessarily tell you that the conclusion is true. The premisses may be false,
after all! Moreover, in an important sense, the logician *does not care* whether the conclusion of
an argument is true or false. Rather, the logician wants to know something conditional: if the
premisses were true, would they be evidence for the conclusion?

Most logicians today are interested in a special case of the relation of evidential support:
the relation of *logical consequence*.\(^3\) The relation of logical consequence is a *deductive* relation
that holds between the premisses of an argument and the conclusion of that argument when the
truth of the premisses *guarantees* the truth of the conclusion. Argument (1) above has this
feature. If the premisses of the argument are both true, then the conclusion *cannot* be false. If the
premisses and the conclusion of an argument stand in the relation of logical consequence, then
we say that the conclusion *follows logically* from the premisses. And if the conclusion of an
argument follows logically from its premisses, then the argument is a good one: the premisses
evidentially support the conclusion.

\(^3\) See Etchemendy (1990) for a very detailed discussion of the relation of logical consequence.
The premisses of an argument might evidentially support some conclusion despite the fact that the conclusion does not follow logically from the premisses. In fact, many arguments that we care about are such that even if the premisses of the argument are true, the conclusion is not guaranteed to be true. For example, we are often interested in arguments like the following:

Only one out of several million tickets in lottery \( \mathcal{L} \) is a winning ticket.

\( \text{(3)} \) Alistair has exactly one ticket in lottery \( \mathcal{L} \).

Alistair does not have a winning ticket.

In argument (3), if the premisses are both true, then the conclusion is very, very likely to be true. And so, it seems that the premisses evidentially support the conclusion. But the truth of the premisses does not guarantee the truth of the conclusion. Hence, the relation of logical consequence—understood as a deductive relation—only represents one species of evidential support. Sometimes, the conclusion of an argument stands in an inductive relation to the premisses. Therefore, even after we have a complete account of logical consequence, we will not have a complete account of evidential support. Later in this text, we will spend some time thinking about probability in order to better understand what makes arguments like (3) examples of good reasoning.

We have already idealized reasoning in an important way by replacing beliefs with sentences and processes of thought with static arguments. In order to better study what makes reasoning good or bad, we will use special symbols to construct a formal language within which we may describe various kinds of evidential relation with mathematical precision. We will think of the sentences of our formal language as translations of sentences that occur in ordinary language, and we will construct a well-defined relation within our formal language to represent the relation of evidential support—or some species of that relation—that holds for sentences in ordinary language. Hence, we are studying reasoning at two removes. In place of the messy
details about how brains represent the world and modify their representations over time, we substitute details about the structure of some linguistic objects. And instead of studying the linguistic objects directly, we study mathematical representations of them.

As Whately observed, logic is both a science and an art. We are mostly interested in the science of logic: our aim is to understand what makes reasoning good or bad. But we do not want to lose sight of the art of logic. We do not just want a formal account of good and bad reasoning. Ultimately, we want to be able to reason better as a result of studying logic. How can we learn to reason better? In this book, we will learn to reason better by studying formal models of good reasoning and by thinking about what makes reasoning good. We then try to make our own reasoning good by emulating our formal models, and in some cases, translating our own reasoning into a formal model to check it.

In order to improve our reasoning, we need to imagine that our reasoning really can be improved. And that brings us to the first rule of practical logic: In order to learn, we must desire to learn and not rest content with what we already think. So, ask questions. Challenge your own beliefs and the beliefs of the people around you. When you find that you do not understand something, inquire harder. Talk to your friends and classmates. Interrogate your teachers. Consult books and websites. Make careful observations of the world around you. Make guesses and test them. Run experiments. Never give up until Nature has relinquished her secrets to you. Nothing else you learn will ever be as important as having an unquenchable thirst to know.

References


Mill, J. (1858) *A System of Logic*. New York: Harper & Brothers. Available at [http://books.google.com/books?id=sh-e97TwKwkC&printsec=frontcover&dq=%22system+of+logic%22+in:mill&source=bl&ots=rFh05BWP_N&sig=zyOlo2g-_Wz9LqMwbVPLYBJeDtM&hl=en&sa=X&ei=ImwgUPfWDI6xqAHmoICADQ&ved=0CDQQ6AEwAA#v=snippet&q=%22logic%20is%20not%20the%20science%20of%20belief%22&f=false]


Chapter 1: Zeroth-Order Logic

In this chapter, we will begin building up our formal language. Our formal language will make use of the following collection of symbols. We will use capital letters, such as $P, Q,$ and $R$ (possibly with numerical subscripts, like $P_1$), called *sentence letters*. And we will use four special symbols, called *operators* or *sentential connectives*: $\sim, \land, \lor,$ and $\rightarrow$. In order to make our symbols easier to talk about, we will give them names. We will call the symbol $\sim$ a *tilde*. We will call the symbol $\land$ a *caret*. We will call the symbol $\lor$ a *wedge*. And we will call the symbol $\rightarrow$ an *arrow*. We will also use left and right parentheses, but they don’t have special names.

Now that we have the symbols of our language, we need to know how the *grammar* or the *syntax* of the language works. A sequence of symbols in our formal language is called a *formula*. In ordinary languages, like English, some sequences of words are grammatical, while others are not grammatical. For example, the sequence, “Bears like to eat honey,” is grammatical; whereas, “Like eat to bears honey,” is not grammatical. $^5$ Similarly, in our formal language, some formulas are grammatically correct, while others are not. The rules that tell us when a formula counts as grammatical are called *formation rules*. When a formula in our formal language follows the formation rules, it is called a *(syntactically) well-formed formula.*

1.1 *Formation Rules*

A single sentence letter standing alone is a well-formed formula. Hence, $P$ is a well-formed formula in our formal language. On the other hand, $PQ$ is not well-formed. We now identify the

$^5$ Notice that we are not talking about *meaning* here. Arguably, a sequence of words (in English or some other language) might be grammatical without being meaningful. The classic example, due to Chomsky (1957), is “Colorless green ideas sleep furiously.”
rest of the well-formed formulas recursively from the basis of the single sentence letters. First, suppose that $\phi$ and $\psi$ are well-formed formulas. Given that $\phi$ and $\psi$ are well-formed formulas, the following are also well-formed formulas:

$$\sim \phi$$
$$(\phi \land \psi)$$
$$(\phi \lor \psi)$$
$$(\phi \rightarrow \psi)$$

Nothing else counts as a well-formed formula. If a given formula cannot be constructed using the above rules, then it is not well-formed.

In order to make formulas easier to talk about when thought of as purely syntactic, we will adopt the following conventions for reading them. The formula $\sim P$ is to be read, “tilde-$P$.” The formula $(P \land Q)$ is to be read, “$P$ caret $Q$.” The formula $(P \lor Q)$ is to be read, “$P$ wedge $Q$.” And the formula $(P \rightarrow Q)$ is to be read, “$P$ arrow $Q$.” According to our formation rules, the following are examples of well-formed formulas:

$$Q$$
$$\sim \sim Q$$
$$(Q \land Q)$$
$$(R \lor (Q \land \sim P))$$
$$((P \land R) \rightarrow (Q \lor \sim S))$$

And the formulas below are examples of formulas that are not well-formed:

$$PQR$$
$$\sim \land ((P$$
$$P \land \rightarrow \lor S$$
$$(Q \sim S)$$
$$\land P$$

Our formation rules leave no room for anyone to misread a formula, but the rules do have the drawback of leading to formulas with lots of parentheses. Typically, we can improve readability without creating confusion by dropping the outermost parentheses. For example, instead of writing $(R \lor (Q \land \sim P))$, we will write $R \lor (Q \land \sim P)$. 
1.2 Translations

If our formal language is to be useful for evaluating arguments in ordinary language, then we need to know how to do two different translation steps. First, we need to be able to translate sentences from ordinary language into sentences of our formal language. And second, we need to be able to translate sentences from our formal language into sentences of ordinary language.

Translation is important, but there are two good reasons not to pay too much attention to translation in an introductory logic course. First, the formal languages developed in introductory logic courses are not powerful enough to adequately express very much of ordinary language. That will be true in this course. Second, formal languages in logic play another role not connected with their use in formalizing ordinary language: formalizing the foundations of mathematics. As a result, some of the usual formal language does not map very neatly onto ordinary language, as we will see. Nonetheless, some treatment of translation is necessary.

How should we start with translating sentences of ordinary language into our formal language? Conveniently, we have sentence letters. So, let’s just assign a unique sentence letter to each unique sentence of ordinary language. For example, we might use $P$, $Q$, $R$, and $S$ to represent the following sentences, respectively:

- Every child is innocent.
- Susan doesn’t believe that President Obama is secretly a Muslim.
- Jesus wept.
- Talan is my son, and his birthday is in October.

Now, we have a very generic way of translating into and out of our formal language. However, using a single sentence letter for each sentence of ordinary language ignores a lot of internal structure in the sentences of ordinary language. And we want our formal language to capture at least some of that internal structure, since the internal structure of sentences in ordinary language has a lot to do with the evidential relationships that they stand in.
Consider sentence $S$ above: Talan is my son, and his birthday is in October. Notice that sentence $S$ is composed of two distinct sentences joined by the conjunction and. Let $T$ stand for the sentence, “Talan is my son,” and let $W$ stand for the sentence, “[Talan’s] birthday is in October.” The three sentences – $S$, $T$, and $W$ – have some structural relationships that we want to capture. For example, if sentence $S$ is true, then it is the case both that sentence $T$ is true and that sentence $W$ is true. In order to begin to capture these sorts of relationships, we will use the operators in our formal language to represent some logically important terms and phrases, like “and,” “not,” and “if …, then ---.”

The first operator we will consider is the caret, $\land$, which we will use to stand for the conjunction and. Using $\land$ to stand for and, we can represent sentence $S$ by the formula $(T \land W)$. We will also use $\land$ to stand for several other conjunctions, like “also,” “but,” and “however.” At first glance, using the caret operator to represent such diverse conjunctions, which admittedly have different uses in ordinary English speech and writing, may look like a mistake. The reason for using the operator this way is that the differences in English are stylistic differences, rather than structural differences. Nothing is lost with respect to the evidential relationships that the sentence stands in if, for example, we substitute, “Hal wanted to go to the beach, and Mary wanted to go to the museum,” for “Hal wanted to go to the beach, but Mary wanted to go to the museum.” And it is the evidential relationships that we care about in logic. We will return to this issue briefly at the end of the next section.

In the same way as with the caret operator, we are going to use our other operator symbols to capture some of the internal structure of sentences in ordinary language. We will use the tilde $\sim$ to stand for not. For example, if we let $M$ stand for the sentence, “President Obama is secretly a Muslim,” then the formula $\sim M$ stands for the sentence, “President Obama is not
secretly a Muslim.” Notice that the *not* that the tilde translates might show up in many different ways and even in different positions in a sentence of ordinary language. Sometimes for clarity’s sake, the tilde is translated as “It is not the case that …,” which would lead us to translate the formula \( \sim M \) as the sentence, “It is not the case that President Obama is secretly a Muslim.”

Next up is the wedge. Logicians are wont to say that they are using the wedge symbol, \( \lor \), to stand for *or*. However, the *or* here differs a little bit from the typical use of *or* in English. Typically, sentences like, “Either Ira likes to draw or Carol is funny,” implicate that *exactly one* of the two clauses is true. If Ira likes to draw, then Carol is *not* funny. And if Carol is funny, then Ira *does not* like to draw. In other words, the ordinary English use of the word “*or*” is *exclusive*. If one of the two clauses is true, then the truth of the other one is excluded.

By contrast, the wedge stands for an *inclusive or*, which is used when both clauses joined by *or* might be true at the same time. For example, if \( D \) stands for the sentence, “Ira likes to draw,” and \( C \) stands for the sentence, “Carol is funny,” then the formula \( (D \lor C) \) stands for the sentence, “Ira likes to draw or Carol is funny *or both*.” Alternatively, one might say, “Ira likes to draw and/or Carol is funny.”

Logicians, especially those concerned with mathematics, have lots of good reasons to care about *inclusive-or*. However, translating the English *or* is not one of them. A better choice for translating the English *or* is the exclusive-or, sometimes denoted with the symbol \( \oplus \). We will look at the \( \oplus \) operator a bit more in the next section. In this text, we will follow ordinary convention in talking about the wedge as the *or* operator, but in translation exercises, we will make the intended translation obvious by using either the “or-both” construction or the “and/or” construction (but not both).
Finally, we will use the arrow $\rightarrow$ to stand for the structure, “If …, then ---,” which is called a *conditional*. For example, if $B$ stands for the sentence, “Bucky swings at the first pitch,” and $H$ stands for the sentence, “Bucky gets a hit,” then the formula $(B \rightarrow H)$ stands for the sentence, “If Bucky swings at the first pitch, then Bucky gets a hit.” We use the arrow to translate two other (equivalent) if-then expressions, as well: (1) “Bucky gets a hit if Bucky swings at the first pitch,” and (2) “Bucky swings at the first pitch only if Bucky gets a hit.” The conditional in ordinary English is notoriously difficult to translate into formal language, and as we will see in more detail in the next section, the arrow is an imperfect choice. However, for our purposes, the arrow is the best choice available. I will choose translation items that do not stretch the usefulness of our operator too much.

With English translations in hand, we now have two ways of referring to the operators: by their names (tilde, caret, wedge, and arrow) or by the English words they translate (not, and, or, and if-then). And we can do the same thing with formulas. The formula $\sim P$ can be read, “not-$P$.” The formula $(P \land Q)$ can be read, “$P$ and $Q$.” The formula $(P \lor Q)$ can be read, “$P$ or $Q$.” And the formula $(P \rightarrow Q)$ can be read, “If $P$, then $Q$,” or it can be read, “$P$ only if $Q$,” or it can even be read “$Q$ if $P$.”

Typically, when we use a single sentence letter to represent a sentence, it signals that we want to ignore the internal structure of that sentence for some reason. For example, we might not know how to formalize the internal structure of the sentence or perhaps we just don’t care about its internal structure. As an example of the first kind of reason, consider the sentence, “Susan doesn’t believe that President Obama is secretly a Muslim,” which we represented above with

---

6 When translating ordinary language sentences, the first reading is the most common. But in order to make the formula easier to read in-line and in order to preserve similarity to the other connectives, I will most often read the formula $(P \rightarrow Q)$ as “$P$ only if $Q$.” Either that, or I will stick to saying, “$P$ arrow $Q$.”
the sentence letter $Q$. Notice that this sentence has another sentence as a proper part: the sentence, “President Obama is secretly a Muslim,” which we represented above with the sentence letter $M$. The problem here is that we do not have tools to formalize the way in which sentence $M$ interacts with the rest of sentence $Q$ in order to produce sentence $Q$. Such tools have been developed, but they are well beyond the scope of the present course.

1.3 Truth Functions and Truth Tables

Anyone who has taken a bit of high school algebra should be familiar with the notion of a mathematical function. In simple, non-technical terms, a function is a kind of input-output machine for abstract objects, like numbers. A function has some number of places or arguments, which is the number of inputs that the function takes. For example, consider the plus-three function, which we might write as: $f(x) = x + 3$. The plus-three function is a one-place function. It takes a single argument as input. In the case of the plus-three function, the inputs are numbers. If you input the number 7, the plus-three function outputs the number 10. If you input the number 12, the function outputs the number 15. And so on. An example of a two-place function is the addition function, which we might write as: $g(x, y) = x + y$. If you input the numbers 3 and 8, the addition function outputs the number 11. If you input the numbers 25 and 103, the function outputs the number 128. And so on.

A truth function is just a special kind of input-output machine. Instead of taking in and returning numbers, a truth function takes in truth-values and outputs truth-values. In our formal language, we have exactly two truth-values: true and false. In this text, we will use T to stand for true, and we will use F to stand for false. In some textbooks, especially in computer science and related fields, writers use the numeral 1 to stand for true and the numeral 0 to stand for false.
At this point, we will begin to think of our operators as truth functions. Each operator symbol in our formal language will be associated with a unique truth function. I will first describe the truth functions associated with each operator in words, and then I will introduce a diagrammatic tool, called a truth table, in order to make the ideas clearer. First, the tilde, ~, will be associated with the truth function called *negation*. The negation is a one-place truth function. If the input to the negation is true, then the negation outputs false. And if the input to the negation is false, then the negation outputs true. Second, the caret, ∧, will be associated with the truth function called *conjunction*. The conjunction is a two-place truth function, and its inputs are called *conjuncts*. If the inputs to the conjunction are both true, then the conjunction outputs true. Otherwise, the conjunction outputs false. Third, the wedge, ∨, will be associated with the truth function called *disjunction*. The disjunction is a two-place truth function, and its inputs are called *disjuncts*. If the inputs to the disjunction are both false, then the disjunction outputs false. Otherwise, the disjunction outputs true. Finally, the arrow, →, will be associated with the truth function called the *material conditional*. The material conditional is also a two-place truth function. Unlike the conjunction and the disjunction, the order of the inputs matters for the material conditional. The first input to the material conditional is called the *antecedent*; the second input is called the *consequent*. Hence, in the formula \((P \to Q)\), the sentence letter \(P\) is the antecedent and the sentence letter \(Q\) is the consequent. If the antecedent of a material conditional is true and the consequent of the material conditional is false, then the material conditional outputs false. Otherwise, the material conditional outputs true.

Going forward, I will drop the input-output language and say that if the arguments of some truth function are such and so, then the truth function *is so and such or has so and such value*. In order to help you learn the material, I will repeat the previous paragraph in slightly
different language. We represent *negation* with the tilde, \( \sim \). The negation is false when its argument is true, and it is true when its argument is false. We represent *conjunction* with the caret, \( \wedge \). The conjunction is true if both of its arguments (its *conjuncts*) are true, and it is false otherwise. We represent *disjunction* with the wedge, \( \lor \). The disjunction is false if both of its arguments (its *disjuncts*) are false, and it is true otherwise. And we represent *material conditional* with the arrow, \( \rightarrow \). The material conditional is false if its first argument (its *antecedent*) is true and its second argument (its *consequent*) is false. Otherwise, the material conditional is true.

Each sentence letter in our formal language is assigned a single truth-value. In other words, we regard each sentence in our formal language as being either true or false, but not both. Remember that each sentence letter represents a sentence in ordinary language. So what we are saying here is that every sentence in ordinary language is either true or false but not both. For example, if we let the sentence letter \( R \) stand for the sentence, “Jesus wept,” then we are saying that it is either the case that Jesus wept or it is not the case that Jesus wept. We are denying that it could be the case that Jesus both wept and did not weep. Let us suppose that Jesus did, in fact, weep. Then the truth-value that we assign to \( R \) is T, for true. An assignment of a truth-value to a sentence letter is sometimes called an *interpretation*.\(^7\) Thinking of our sentence letters as interpreted—as having truth-values—we can think of a truth function as taking in sentence letters as inputs and returning a truth-value as output. For example, we could ask for the truth-value of the conjunction of \( P \) and \( R \) under the interpretation that says both that \( P \) is true and that \( R \) is true. What would the value of the conjunction be in that case?

\(^7\) An interpretation is itself a function that takes a sentence letter as its input and returns a truth-value as its output.
Recall that in some sense in logic we don’t care about the actual truth values of the sentence letters in our formal language. We only care about evidential relationships. We do not presume to know whether the sentences that we are representing with various sentence letters are true or false. Hence, we do not presume to know which interpretation to give to the sentence letters in our formal language. Consequently, we want to write down all of the possible truth-values that a collection of sentences might have and see how those are related to one another. A convenient way to write down all of the possible truth-values that a collection of sentences might have is to use something called a truth table.

The simplest truth table is just a list of the possible truth-values for a single sentence letter. For concreteness, consider the sentence letter $P$, which stands for the sentence, “Every child is innocent.” In our formal language, we allow only two possibilities: (1) the sentence represented by $P$ is true, or (2) the sentence represented by $P$ is false. We can write this in a truth table as follows:

<table>
<thead>
<tr>
<th>$P$</th>
<th>T</th>
<th>F</th>
</tr>
</thead>
</table>

Very simple! But suppose we have three sentence letters instead of just one:

$P = \text{“Every child is innocent.”}$
$Q = \text{“Susan doesn’t believe that President Obama is secretly a Muslim.”}$
$R = \text{“Jesus wept.”}$

None of these sentences appears as part of any of the others. Therefore, we assume that the truth-value that any one of the sentences takes is independent of the truth-values assigned to the other two sentences. The interpretation under which $P$ is true, $Q$ is false, and $R$ is false is just as admissible as the interpretation under which $P$ is false, $Q$ is true, and $R$ is false. Since our formal language has exactly two truth-values, the number of possible interpretations for $n$ sentence
letters (assuming the letters may be assigned truth-values independently) is $2^n$. In the case of three sentence letters, there are $2^3 = 8$ possible interpretations, which we can write in a truth table as follows:

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Each row in a truth table represents a way that the world could be. Hence, we can think of each line in a truth table as representing a *possible world*.

The truth-value of a truth function—negation, conjunction, disjunction, or material conditional—is completely determined by the truth-values of the arguments of the truth function. Consider again the three sentences:

$S =$ “Talan is my son, and his birthday is in October.”
$T =$ “Talan is my son.”
$W =$ “Talan’s birthday is in October.”

Sentence $S$ can be written as a conjunction of sentences $T$ and $W$. Since the truth-value of a truth function is completely determined by the truth-values of its arguments, if we know the truth-values of $T$ and $W$, then we can determine the truth-value of $S$. Consequently, the truth-values for $S$, $T$, and $W$ cannot be assigned independently, in the way that the truth-values for $P$, $Q$, and $R$ could be assigned independently. In other words, not all of the rows in the table at the top of the next page represent possible worlds. Some of them—the ones in which the letters are colored red—are *impossible worlds*. 
When we consider only the possible worlds, we get the truth table for the conjunction, which we can write as follows:

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>S = (T ∧ W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

As indicated earlier, the conjunction \((T ∧ W)\) is true only in the row (world) where \(T\) and \(W\) are both true. In every other row (world), the conjunction \((T ∧ W)\) is false.

In the same way, we can write out truth tables for our other truth functions. Let’s use the sentence letters \(X\) and \(Y\) to stand for two sentences that can be assigned truth-values independently. The following is the truth table for negation:

<table>
<thead>
<tr>
<th>X</th>
<th>~X</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

The table gives us what we expected: \(\sim X\) is false when \(X\) is true, and \(\sim X\) is true when \(X\) is false. Since negation is a one-place operator, we only need a table with two rows. For our other truth functions, we need tables with four rows, since the other truth functions have two arguments.

The truth table for conjunction is given at the top of the next page:
The conjunction is true only when both of its conjuncts are true. By contrast, the disjunction is only false when both of its disjuncts are false, as can be seen in the truth table for disjunction:

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>( (X \land Y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Finally, we can write the truth table for the material conditional as follows:

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>( (X \to Y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

Like the disjunction, the material conditional is only false in one row of the table. But unlike the disjunction, the material conditional is not symmetric with respect to its arguments: order matters. If the antecedent is true and the consequent is false, then the material conditional is false. Otherwise, it is true.

In Section 1.2, we noted that our disjunction does not map nicely onto the English word “or,” because the English word “or” is typically used as an exclusive-or. In fact, we can write out a truth table for the exclusive-or, which we symbolized with \( \oplus \).

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>( (X \oplus Y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>
The negation of the exclusive-or gives us another interesting operator, called the material bi-conditional. The material bi-conditional is symbolized with a double-headed arrow, $\leftrightarrow$, and its truth table can be written as follows:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$(X \leftrightarrow Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

A simple counting exercise shows that there are four possible one-place and 16 possible two-place truth functions. All of the one- and two-place truth functions are described in an appendix to this chapter. In addition to these, there are three-place truth functions, four-place truth functions, and so on. But we will not say much about them. We don’t need to. The reason we don’t need to say much about all of the possible truth functions is that using only our four basic truth functions, we can represent all of the other possible truth functions. All of them. We actually don’t even need all of the truth functions we have described, but counter-intuitively, having a few extra operators turns out to make the formal language much easier to work with.

1.4 Evaluating, Classifying, and Comparing Sentences

The truth tables for our truth functions let us compute the truth-values of compound sentences—sentences composed of simple sentences connected together by one or more operators—given that we know the truth-values of their component clauses. For example, suppose we want to evaluate the truth of the sentence, “Mary helps with dinner, but if she doesn’t help, then Martha will be angry,” under the assumption that we know the truth-values of the following two components, represented by $P$ and $Q$:

$P =$ “Mary helps with dinner.”
$Q =$ “Martha will be angry.”
Specifically, let’s suppose that $P$ and $Q$ are both true. We will often use the same few letters, like $P$, $Q$, and $R$, to represent different sentences in different arguments, as long as the context makes clear what the sentence letters are translating. So, don’t be confused by the fact that we used $P$ and $Q$ to stand for different sentences earlier on.

Given the translations above, the sentence, “Mary helps with dinner, but if she doesn’t help, then Martha will be angry,” may be translated into our formal language as follows:

$$(P \land (\sim P \rightarrow Q))$$

Let’s call this sentence $S$, just so that we have a simple way to refer back it. We could drop the outermost parentheses in order to make $S$ more readable:

$$P \land (\sim P \rightarrow Q)$$

But for now, let’s stick with the formally correct version. Now, we are supposing that we know the truth-values of the component clauses, $P$ and $Q$. In general we will neither know these values nor care what they are, but suppose that in this case we know that $P$ and $Q$ are both true. What is the truth-value of the sentence $(P \land (\sim P \rightarrow Q))$?

In order to compute the truth-value of $S$, we work from the inside out. That is to say, we will compute the truth-values of the parts of $S$ that are most enclosed by parentheses. In this case, the most-enclosed part of $S$ is $\sim P \rightarrow Q$. In cases like this, we first need to evaluate the negation. We know that the negation of a sentence letter has a truth-value opposite to the truth-value of the sentence itself. And we supposed that $P$ is in fact true. So, $\sim P$ is false. Let’s replace $\sim P$ with its truth-value, $F$, in the most enclosed part of $S$ to get $F \rightarrow Q$. We are supposing that we know the truth-value of $Q$: it is true. So, we want to know what the truth-value of a material conditional is when its antecedent is false and its consequent is true. From the table for the material conditional in Section 1.3, we see that when the antecedent is false and the consequent is true, the material
conditional is true. (In fact, whenever the antecedent of a material conditional is false, the
material conditional itself is true. So, we didn’t even need to know the truth-value of \( Q! \))

Just like we did for \( \sim P \), let’s replace \((F \rightarrow Q)\) with its truth-value, T, to get another new
sentence to evaluate: \((P \land T)\). At this point, we can just evaluate the sentence, since it is one of
our basic truth functions. We are supposing that the truth-value of \( P \) is true, and we know that a
conjunction is true just in case both of its conjuncts are true. So, the truth-value of our newest
sentence is true. Our procedure guarantees that the newest sentence has the same truth-value as
our original sentence \( S \). Hence, \( S \) is true. We noted earlier that the truth-value of \( S \) does not
depend on the truth-value of \( Q \). If \( Q \) had been false, \( S \) would still have been true. Would sentence
\( S \) have been true if \( P \) had been false?

In order to see how the truth-value of a compound sentence depends or does not depend
on the truth-values of its components, we will use truth tables, together with our method for
computing the truth-value of a sentence given the truth-values of its components. Let’s use the
method of truth tables for the sentence \( S \). The first thing we want to do is to write out all of the
possible combinations of truth-values for the basic components of \( S \): namely, \( P \) and \( Q \).

<table>
<thead>
<tr>
<th>( P )</th>
<th>( Q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Now that we have written down all of the possible combinations of the simplest components of
\( S \), we want to think about which more complicated parts of \( S \) we want to evaluate on our way to
being able to evaluate \( S \) as a whole. What we will do is repeat the basic procedure from earlier,
finding the truth-values of the parts of \( S \) based on how deeply enclosed they are and evaluating
negations first in the most deeply enclosed parts. Hence, as before, we first want to evaluate \( \sim P \),
and so, we add \( \sim P \) to our truth table. In order to be clear about which truth-values represent the possible combinations of truth-values for the simple sentences, let’s draw a double line separating the columns for the simple sentences from the columns for all the derived formulas.

<table>
<thead>
<tr>
<th>( P )</th>
<th>( Q )</th>
<th>( \sim P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>

As before, if the truth-value of \( P \) is true, then the truth-value of \( \sim P \) is false, and vice versa. Hence, we can fill in the truth-values for \( \sim P \) by writing down the opposite value from whatever we see in the column for \( P \).

<table>
<thead>
<tr>
<th>( P )</th>
<th>( Q )</th>
<th>( \sim P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

Don’t be thrown by the fact that there are more rows for \( P \) and \( \sim P \) in this table than there were in the table for negation by itself. We aren’t really doing anything different from what we did in Section 1.3, we are just applying the same idea with some redundancy.

Now, instead of replacing \( \sim P \) in \( S \) by its truth-value, we are simply going to write down the next most enclosed part and use what we have written down in the table for \( \sim P \) in order to calculate the truth-values for the formula related to what is next most enclosed. The reason for this difference from our earlier procedure is that there is now no unique truth-value that we could write down as a replacement for \( \sim P \). Therefore, we want to keep track of all the possible values that \( \sim P \) could take on. The next most enclosed part of \( S \) is \( \sim P \rightarrow Q \). So, we add the formula \( (\sim P \rightarrow Q) \) to our truth table:
In order to fill in the column for \((\sim P \rightarrow Q)\), we use the truth-values in the column for \(\sim P\) and the truth-values in the column for \(Q\), respectively, as the truth-values for the antecedent and the consequent of a material conditional. (If it helps you to think through the problem, feel free to simply repeat the column for \(Q\) to the right of the column for \(\sim P\).) Hence, the filled-in table looks like this:

<table>
<thead>
<tr>
<th>(P)</th>
<th>(Q)</th>
<th>(\sim P)</th>
<th>((\sim P \rightarrow Q))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T)</td>
<td>(T)</td>
<td>(F)</td>
<td>(T)</td>
</tr>
<tr>
<td>(T)</td>
<td>(F)</td>
<td>(F)</td>
<td>(T)</td>
</tr>
<tr>
<td>(F)</td>
<td>(T)</td>
<td>(T)</td>
<td>(T)</td>
</tr>
<tr>
<td>(F)</td>
<td>(F)</td>
<td>(T)</td>
<td>(T)</td>
</tr>
</tbody>
</table>

As in Section 1.3, the material conditional \((\sim P \rightarrow Q)\) will be false just in case its antecedent, \(\sim P\), is true and its consequent, \(Q\), is false. That only happens in the last line of the truth table, so only in the last line of the table is the conditional false.

Now, we are ready to add all of sentence \(S\) to our truth table. As long as we have made a note of what sentence \(S\) represents, we could simply write \(S\) in the last column of the truth table, and we will often make similar notes so that we can use single letters for complicated formulas in order to save space in our tables. But here, we’ll write out the whole sentence:

<table>
<thead>
<tr>
<th>(P)</th>
<th>(Q)</th>
<th>(\sim P)</th>
<th>((\sim P \rightarrow Q))</th>
<th>((P \wedge (\sim P \rightarrow Q)))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T)</td>
<td>(T)</td>
<td>(F)</td>
<td>(T)</td>
<td>(T)</td>
</tr>
<tr>
<td>(T)</td>
<td>(F)</td>
<td>(F)</td>
<td>(T)</td>
<td>(F)</td>
</tr>
<tr>
<td>(F)</td>
<td>(T)</td>
<td>(T)</td>
<td>(T)</td>
<td>(T)</td>
</tr>
<tr>
<td>(F)</td>
<td>(F)</td>
<td>(T)</td>
<td>(F)</td>
<td>(F)</td>
</tr>
</tbody>
</table>

In order to fill in the column for \((P \wedge (\sim P \rightarrow Q))\), we use the truth-values in the column for \(P\) and the truth-values in the column for \((\sim P \rightarrow Q)\) as the truth-values for the conjuncts of a
conjunction. As in Section 1.3, a conjunction is true just in case both of its conjuncts are true. Hence, the filled-in table looks like this:

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$\neg P$</th>
<th>$(\neg P \to Q)$</th>
<th>$(P \land (\neg P \to Q))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>$T$</td>
<td>$F$</td>
<td>$T$</td>
<td>$T$</td>
</tr>
<tr>
<td>$T$</td>
<td>$F$</td>
<td>$F$</td>
<td>$T$</td>
<td>$T$</td>
</tr>
<tr>
<td>$F$</td>
<td>$T$</td>
<td>$T$</td>
<td>$T$</td>
<td>$F$</td>
</tr>
<tr>
<td>$F$</td>
<td>$F$</td>
<td>$T$</td>
<td>$F$</td>
<td>$F$</td>
</tr>
</tbody>
</table>

We can now see, somewhat surprisingly, that the truth-value of sentence $S$ depends entirely on the truth-value for $P$. In fact, the two sentences always have the same truth-value!

Let’s consider a slightly more complicated example: “Jim rides a bicycle, and it’s not true that if Jim doesn’t ride a bicycle, then either he goes camping or he goes to the store (or both).” We’ll call this sentence $W$. First, we need to translate $W$ into our formal language. Let’s use the sentence letters $P$, $Q$, and $R$ to represent clauses in $W$ as follows:

$P = “Jim rides a bicycle.”$
$Q = “Jim goes camping.”$
$R = “Jim goes to the store.”$

Using our operator symbols and what we know about translations from English into our formal language, we can represent the sentence as follows:

$$(P \land \neg(\neg P \to (Q \lor R)))$$

Now that we have a representation in our formal language, we want to determine how the truth of the whole sentence is related to the truth of its component clauses. So, we begin to write out a truth table like we did earlier for sentence $S$. Again, we want to write out all of the possible combinations of truth-values for the simplest parts of our sentence, which in this case are $P$, $Q$, and $R$. We use the same pattern here that we used in Section 1.3:
Now that we have written down all of the possible combinations of the simplest components of $W$, we want to think about which more complicated parts of $W$ we want to evaluate on our way to being able to evaluate $W$ as a whole. Again, we will repeat the basic procedure from earlier, finding the truth-values of the parts of $W$ based on how deeply enclosed they are. The most deeply enclosed part of $W$ is $Q \lor R$. Hence, we add $(Q \lor R)$ to our truth table as follows:

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
<th>$(Q \lor R)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

We know how to evaluate $(Q \lor R)$ because it is a basic truth function that we introduced in Section 1.3: a disjunction is only false when both of its disjuncts are false; otherwise, it is true. Hence, the filled-in table looks like this:

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
<th>$(Q \lor R)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>
The next most enclosed part of \( W \) is \( \sim P \rightarrow (Q \lor R) \), but we need to evaluate \( \sim P \) before evaluating the whole expression. So let’s add \( \sim P \) and \( \sim P \rightarrow (Q \lor R) \) to our truth table.

<table>
<thead>
<tr>
<th>( P )</th>
<th>( Q )</th>
<th>( R )</th>
<th>( (Q \lor R) )</th>
<th>( \sim P )</th>
<th>( (\sim P \rightarrow (Q \lor R)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
</tbody>
</table>

After filling in the values for \( \sim P \), we are able to fill in the values for \( (\sim P \rightarrow (Q \lor R)) \). The completed truth table looks like this:

<table>
<thead>
<tr>
<th>( P )</th>
<th>( Q )</th>
<th>( R )</th>
<th>( (Q \lor R) )</th>
<th>( \sim P )</th>
<th>( (\sim P \rightarrow (Q \lor R)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
</tbody>
</table>

In order to save some space in our truth table, let’s give a new name to \( (\sim P \rightarrow (Q \lor R)) \). Let’s call it \( A \). The sentence letter \( A \) then stands for the sentence, “If Jim doesn’t ride a bicycle, then either he goes camping or he goes to the store (or both).”

Substituting into \( W \), we get the formula \( (P \land \sim A) \). In order to evaluate this, we need to first evaluate \( \sim A \). So, let’s add the formula \( \sim A \), and since we have seen the negation several times already, let’s go ahead and fill in the truth-values. After filling in the values for \( \sim A \), the only thing left is to evaluate the conjunction \( (P \land \sim A) \), which is the same sentence as \( W \). The result is given at the top of the next page.
|| P | Q | R | (Q ∨ R) | ~P | A | ~A | (P ∧ ~A) |
|---|---|---|---------|----|---|----|---------|
| T | T | T | T       | F  | T | F  | F       |
| T | T | F | T       | F  | T | F  | F       |
| T | F | T | T       | F  | T | F  | F       |
| T | F | F | F       | F  | T | F  | F       |
| F | T | T | T       | T  | T | F  | F       |
| F | T | F | T       | T  | T | F  | F       |
| F | F | T | T       | T  | T | F  | F       |
| F | F | F | F       | T  | F | T  | F       |

Did you recognize right away that the sentence W is never true? Surprisingly (I think), there is no possible world in which W is the case. Sentence W cannot possibly be true!

Now that we have some experience working with truth tables, I want to introduce some useful vocabulary. Every sentence in our formal language belongs to exactly one of three different categories, depending on what its truth table looks like. A sentence can be either **contradictory** or **tautologous** or **contingent**. (Sometimes, one sees the term “self-contradictory” used instead of “contradictory.” The terms are synonyms.)

A **contradictory sentence** is false in every row of its truth table. In other words, a contradictory sentence is false in every possible world. The sentence W above is an example of a contradictory sentence. The sentence W cannot possibly be true. No matter what the world is like, sentence W is false. Therefore, the truth-value of a contradictory sentence does not depend on the truth-values of its components.

On the other extreme are tautologous sentences. A **tautologous sentence** (sometimes called a **tautology** or a **logical truth**) is true in every row of its truth table. In other words, a tautologous sentence is true in every possible world. Hence, like a contradictory sentence, the truth-value of a tautologous sentence does not depend on the truth-values of its components.

The negation of the contradictory sentence W is a tautologous sentence. In fact, the negation of a contradictory sentence is always a tautologous sentence, regardless of which
contradictory sentence you pick. Likewise, the negation of a tautologous sentence is a
ccontradictory sentence.

Finally, a contingent sentence is true in at least one row of its truth table and false in at
least one row of its truth table. In other words, a contingent sentence is true in some, but not all,
possible worlds. Every sentence that we treat as simple we also treat as contingent. Compound
sentences may also be contingent. For example, our first example sentence in this section,
sentence $S$, is contingent. Sentence $S$ is true in the first two rows of its truth table and false in the
last two rows of its truth table.

Along with classifying individual sentences, truth tables let us make pairwise
comparisons of sentences. We care especially about four different ways that two sentences may
be related to each other. Two sentences may be equivalent, contradictory, consistent, or
inconsistent. Unlike in the case of classification, a single pair of sentences might stand in more
than one of these relations.

Two sentences are equivalent (or logically equivalent) if they have exactly the same truth
table with respect to the same set of simple sentences. In other words, two sentences are
equivalent if their truth-values are the same in every possible world. Another way of thinking
about equivalent sentences is that two sentences $X$ and $Y$ are equivalent just in case the bi-
conditional ($X \leftrightarrow Y$) is a tautology.

Two sentences are contradictory (or logically contradictory) if they always have the
opposite truth-values to each other. In other words, two sentences are contradictory if in each
possible world, one has the same truth-value as the negation of the other. Another way of
thinking about contradictory sentences is that two sentences $X$ and $Y$ are contradictory just in
case the exclusive-or ($X \oplus Y$) is a tautology.
Two sentences are *consistent* if there is at least one row of their truth table for which they are both true. In other words, two sentences are consistent if there is at least one possible world in which they are both true. By contrast, two sentences are *inconsistent* if there is no row of their truth table for which they are both true. In other words, two sentences are inconsistent if there is no possible world in which they are both true.

If two sentences are contradictory, then they are inconsistent. However, it is not the case that if two sentences are inconsistent, then they are contradictory. In fact, two inconsistent sentences might be equivalent! Below is a truth table for several illustrative sentences.

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$(P \to Q)$</th>
<th>$\neg(P \to Q)$</th>
<th>$\neg(Q \to P)$</th>
<th>$\neg P \lor Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

All of the sentences in the table are contingent sentences. The first and last sentences are equivalent. They are also consistent, since there is a row in the table for which they are both true (actually, three of them). The first and second sentences are contradictory, since they have a different truth-value on each line of the table. Hence, the first and second sentences are also inconsistent: there is no line on which they are both true. The first and third sentences are consistent because they are both true on the third line of the table; however, the two sentences are not equivalent. The second and third sentences are inconsistent, but they are not contradictory. There is no row of the table on which they are both true, but they do not always have opposite truth-values.

1.5 *Validity*

In the previous section, we used truth tables to evaluate, classify, and compare sentences. In this section, we will use truth tables to classify arguments. The basic idea, which goes back at least to
Peirce (1877) and was first made rigorous by Tarski (1936) in his analysis of logical consequence, is that an argument is good if it preserves truth. This will be our first—and most fundamental—formal conception of evidential support: an argument is good just in case it gives a true conclusion from true premisses.

Of course, an argument might do a better or worse job of preserving truth. As we said in Chapter 0, good non-ampliative (deductive) arguments preserve truth in all cases. They provide a guarantee that if the premisses are true, the conclusion is true. By contrast, a good ampliative (inductive) argument might *typically* preserve truth without giving any guarantee, but in this section, we will restrict attention to non-ampliative arguments. We will say that a good non-ampliative argument is *(deductively)* valid. And we will simply define validity in terms of truth-preservation. An argument is *valid* if whenever its premisses are true, its conclusion is true. Otherwise, it is *invalid*. We will reserve the term *cogent* for ampliative arguments whose conclusions are true at least half the time given that their premisses are true.

We will use truth tables to classify arguments as either valid or invalid. We will use the following procedure. First, write out a truth table that includes entries for the premisses and conclusion of the argument to be classified. Second, check every row in the truth table where all of the premisses are true: if in every row where the premisses are all true the conclusion is also true, then the argument is valid; otherwise, the argument is invalid. If there is no row in which all of the premisses are true, then the argument is valid. (In the case where there is no row in which all of the premisses are true, we often say that the argument is *vacuously* valid to distinguish such arguments from other valid arguments.)
Let’s work out the procedure for classifying arguments with some simple examples of valid and invalid arguments. The following is an instance of one of the simplest valid argument forms, called *modus ponens*:

If Jack is a fireman, then Jack gets to ride on a fire truck.
Jack is a fireman.

Therefore, Jack gets to ride on a fire truck.

First, let’s translate the argument into our formal language. Let $P = \text{“Jack is a fireman,”}$ and let $Q = \text{“Jack gets to ride on a fire truck.”}$ Then the argument may be re-written in our formal language as follows:

\[
\begin{array}{c}
(P \rightarrow Q) \\
\hline
P \\
\hline
Q
\end{array}
\]

In order to evaluate the argument, we need to write out a truth table that includes each line in the argument. Just to be especially careful, let’s write out the possible combinations of truth-values for the simple sentences in the argument and then write down all of the premisses and the conclusion, even though that will mean repeating some of the simple sentences in the truth table.

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$(P \rightarrow Q)$</th>
<th>$P$</th>
<th>$Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

From our definition, the argument is valid if whenever the premisses are both true, the conclusion is also true. The premisses are both true in exactly one row of the table, and the conclusion is also true in that row. The truth-values in the relevant row are highlighted in red.

Hence, the argument is valid.
Now let’s consider an invalid argument. The following is an instance of one of the simplest and most common invalid argument forms. It is so common that it gets its own name: the fallacy of *affirming the consequent*:

If Jack is a fireman, then Jack gets to ride on a fire truck.
Jack gets to ride on a fire truck.
Therefore, Jack is a fireman.

Using the same translation of simple sentences as before, the argument may be re-written in our formal language as follows:

\[
\frac{(P \rightarrow Q) \\
Q}{P}
\]

In order to evaluate the argument, we proceed as before. Write out a truth table that includes each line in the argument. Again, we will be extra careful in writing out the table so that we don’t make any mistakes. The only difference between this table and the previous one is that the last two columns are reversed.

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$(P \rightarrow Q)$</th>
<th>$Q$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

It is still true with respect to the new argument that the premisses are both true in the first row of the table. And in that row, the conclusion is also true. *However*, in our new argument, both of the premisses are also true in the third row. But in the third row, the conclusion is false. Again, the relevant rows are highlighted in red. Hence, by our definition, the argument is invalid.

Our verdict here makes sense because there are possible worlds in which Jack is not a fireman but does get to ride on a fire truck. For example, if Jack were the mayor of a town that owns fire trucks, then presumably, he could ride on a fire truck as part of a promotional stunt or...
in a town parade or maybe just because he felt like riding on a fire truck. And yet, Jack need not be a fireman in order to be mayor. Or if Jack were a firehouse dog, Jack might get to ride on the fire truck. And yet, Jack would not be a fireman because no dog is a fireman.

Another simple and common valid argument form is called modus tollens. Unlike modus ponens, modus tollens concludes with a negated sentence. Here is an instance:

If Jack is a fireman, then Jack gets to ride on a fire truck.
Jack does not get to ride on a fire truck.
Therefore, Jack is not a fireman.

Using the same translation of simple sentences as before, the argument may be re-written in our formal language as follows:

\[(P \rightarrow Q)\]
\[\sim Q\]
\[\sim P\]

In order to evaluate the argument, we proceed as before. We write out a truth table that includes each line in the argument. Again, we will be extra careful in writing out the table so that we don’t make any mistakes.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>(P \rightarrow Q)</th>
<th>(\sim Q)</th>
<th>(\sim P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

In this table, there is only one row (highlighted in red) in which both premisses are true, and in that row, the conclusion is true also. Hence, by our definition, the argument is valid.

Just as there is a common invalid argument form (affirming the consequent) that looks suspiciously like the valid argument form modus ponens, there is a common invalid argument form, called denying the antecedent, that looks suspiciously like the valid argument form modus tollens. The following argument is an instance of the fallacy of denying the antecedent:
If Jack is a fireman, then Jack gets to ride on a fire truck.  
Jack is not a fireman.  
Therefore, Jack does not get to ride on a fire truck.

Using the same translation of simple sentences as before, the argument may be re-written in our formal language as follows:

\[
(P \rightarrow Q) \\
\sim P \\
\sim Q
\]

In order to evaluate the argument, we proceed as before. We write out a truth table that includes each line in the argument. Again, we will be extra careful in writing out the table so that we don’t make any mistakes.

<table>
<thead>
<tr>
<th>P</th>
<th>Q</th>
<th>(P → Q)</th>
<th>~P</th>
<th>~Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

In this table, there are two rows in which both premisses of the argument are true: the third and fourth rows. These rows are highlighted in red. In the fourth row, the conclusion is also true, but in the third row, the conclusion is false. Hence, by our definition, the argument is invalid.

The notion of evidential support sketched in this section is curiously related to the material conditional. In a valid argument, if the premisses of the argument are all true, then the conclusion of the argument is true. That claim is correctly translated using the material conditional. Hence, we have another way of testing the validity of an argument using truth tables. Take all of the premisses in an argument that you want to test and put them all in a big conjunction. It doesn’t matter which premisses are most enclosed, but for consistency, let the last two premisses be the most enclosed. For example, if we had four premisses \(P, Q, R, \) and \(S\), we would write down \((P \land (Q \land (R \land S)))\). Now, make that conjunction the antecedent of a material
conditional whose consequent is the conclusion of the argument. For example, if our premisses
were $P, Q, R,$ and $S,$ and our conclusion were $W,$ then we would write down the following
material conditional: $((P \land (Q \land (R \land S))) \rightarrow W).$ The argument is valid if and only if the material
conditional is a tautology.
Appendix

Below are the four possible one-place truth functions:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$F(X)$</th>
<th>$\sim X$</th>
<th>$I(X)$</th>
<th>$T(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

The first one-place truth function, $F$, is false regardless of its argument. The negation we have seen before. The third one-place truth function is the identity. It is just the same as its argument. The fourth one-place truth function is true regardless of its argument.

Below are the 16 possible two-place truth functions. I have written them in four separate tables so that they fit nicely on the page:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$F(X, Y)$</th>
<th>$X \downarrow Y$</th>
<th>$\sim (X \leftrightarrow Y)$</th>
<th>$\sim P_X(X, Y)$</th>
<th>$X \oplus Y$</th>
<th>$X \uparrow Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$\sim (X \rightarrow Y)$</th>
<th>$\sim P_X(X, Y)$</th>
<th>$X \oplus Y$</th>
<th>$X \uparrow Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$P_X(X, Y)$</th>
<th>$X \leftrightarrow Y$</th>
<th>$P_Y(X, Y)$</th>
<th>$X \rightarrow Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$P_X(X, Y)$</th>
<th>$X \leftrightarrow Y$</th>
<th>$X \lor Y$</th>
<th>$T(X, Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>
References


In Chapter 1, we began building our formal language, we introduced truth functions and truth tables, and we described a special notion of evidential support, called validity, in which the conclusion of an argument is guaranteed to be true if its premisses are all true. We then showed how to use truth tables to test whether an argument in our formal language is valid. In this chapter, we provide an alternative account of evidential support in our formal language: proof.

What is the relationship between evidential support and proof? Recall that an argument is a collection of sentences, which is supposed to evidentially support some other sentence, called the conclusion of the argument. In Chapter 1, we said that an argument is good if it preserves truth, which we can check using a truth table. In this chapter, we will say that an argument is good if the conclusion of the argument may be proved from its premisses. We already know how to translate sentences from ordinary language into our formal language. The well-formed formulas (sentences) in our formal language represent sentences in ordinary language. What we need now is an account of proof in our formal language.

A proof is a finite sequence of well-formed formulas such that each formula in the sequence is either a premiss or derived from previous formulas in the sequence according to allowed transformation rules, also called rules of inference. The last formula in a proof is called the conclusion of the proof and should represent the conclusion of the argument that we want to evaluate. What justifies the rules? Some philosophers claim that the rules of inference are justified by intuition. According to these philosophers, the claim that this or that rule of inference is logically good is self-evident, perhaps because it is clear and distinct. The rules seem to be correct, so they must be correct. More recently, philosophers like Gentzen, Prawitz, and
Dummett have argued that the rules of inference themselves give meaning to the connectives, and hence, the rules of inference are analytic truths.¹

The rules of inference are also valid. Repeated application of the rules preserves truth, and hence, any argument that is proof-theoretically good is also valid. The converse also holds in zeroth-order logic, though we will not prove that it does so: if an argument is valid, then there is a proof of its conclusion from its premisses.² Disconcertingly, the two standards of goodness for arguments come apart when the formal language becomes too powerful, but we will not prove this claim either.³

Let \( \Gamma \) be a collection of well-formed formulas, and let \( \phi \) be a well-formed formula. We will use the symbol \( \models \) in order to represent the relation of entailment. If \( \Gamma \) consists of the premisses of an argument, then \( \Gamma \models \phi \) says that \( \Gamma \) entails \( \phi \). A collection \( \Gamma \) of formulas entails some formula \( \phi \) if and only if the formulas in \( \Gamma \) are all true only if the formula \( \phi \) is true. Entailment is the semantic notion of evidential support that we developed in Chapter 1. By contrast, we will use the symbol \( \vdash \) in order to represent the relation of implication. If \( \Gamma \) consists of the premisses of an argument, then \( \Gamma \vdash \phi \) says that \( \Gamma \) implies \( \phi \). A collection \( \Gamma \) of formulas implies some formula \( \phi \) if and only if the formula \( \phi \) is derivable from the premisses in \( \Gamma \). Implication is the syntactic notion of evidential support that we are considering in this chapter.⁴

The proof system that we will begin to develop in this chapter is called a system of natural deduction. Natural deduction systems were simultaneously introduced by Gentzen and

---

¹ For an introduction to proof theoretic semantics, see Schroder-Heister (2012).
² See Chapter 8 of Mates (1972) for a very clear exposition.
³ The basic point here is due principally to the work of Löwenheim, Gödel, Church, and Turing. For discussion of Gödel’s contribution, see Nagel and Newman (2001).
⁴ A formal language is sound iff \( \Gamma \vdash \phi \) implies that \( \Gamma \models \phi \). In other words, a formal language is sound just in case the proofs are valid arguments. A formal language is complete iff \( \Gamma \models \phi \) implies that \( \Gamma \vdash \phi \). In other words, a formal language is complete just in case the conclusions of valid arguments can be proved from the premisses of those arguments. Although we will not prove it, our formal language—up through first-order—is both sound and complete.
Jaśkowski in separate papers in 1934. The main motivation for the development of natural deduction systems was to capture a kind of reasoning common in mathematics where one may freely make assumptions or suppositions, so long as these assumptions are eventually discharged in the course of the proof. In this respect, natural deduction systems are different from axiomatic systems, in which only certain pre-specified formulas may be written down in a proof without justification.

The version of natural deduction developed in this text is similar to that of Suppes (1957). Since our natural deduction system allows us to make arbitrary assumptions, we need to keep careful track of what assumptions we make and whether they are adequately discharged in our proofs. To that end, we will adopt a very specific style when writing down proofs. I will now illustrate the proof style with an example. For now, it is not important that you understand how the proof works. What is important is that you have a model for our style of writing down proofs.

Example 2.1: \{ (\neg P \lor Q ), (R \rightarrow P ), R \} \vdash Q

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>(\neg P \lor Q )</td>
<td>A (premiss)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>(R \rightarrow P )</td>
<td>A (premiss)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(3)</td>
<td>R</td>
<td>A (premiss)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,3</td>
<td>(4)</td>
<td>P</td>
<td>2,3 \rightarrow E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(5)</td>
<td>\neg P</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,3</td>
<td>(6)</td>
<td>(\neg Q \rightarrow P )</td>
<td>4 \rightarrow I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(7)</td>
<td>(\neg Q \rightarrow \neg P )</td>
<td>5 \rightarrow I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,3,5</td>
<td>(8)</td>
<td>\neg\neg Q</td>
<td>6,7 \neg \neg I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,3,5</td>
<td>(9)</td>
<td>Q</td>
<td>8 \neg \neg E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,3</td>
<td>(10)</td>
<td>(\neg P \rightarrow Q )</td>
<td>5,9 \text{ CP}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>(11)</td>
<td>Q</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>(12)</td>
<td>(Q \rightarrow Q )</td>
<td>11 \text{ CP}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,2,3</td>
<td>(13)</td>
<td>Q</td>
<td>1,10,12 \lor \neg E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What we have in Example 2.1 is a proof of the conclusion that Q from the three premisses listed to the left of the turnstile symbol, \( \vdash \), in curly braces: namely, \((\neg P \lor Q )\), \((R \rightarrow P )\), and R. The proof itself has four columns. The first column, reading from left to right, keeps track of the
assumptions that were used to derive that line of the proof. Hence, we will call it the assumption column. The second column from the left gives the line number of the proof. In the third column, called the formula column, we write a well-formed formula that could be derived from earlier formulas using allowed rules of inference. Finally, the fourth column describes the derivation by identifying the inference rule and line numbers from which the formula was immediately derived. Call the last column the justification column.

The main difference between the assumption column and the justification column is that the assumption column keeps track of ultimate justification; whereas, the justification column keeps track of immediate justification. The purpose of the first column is to keep track of assumptions that we make and then discharge. For example, we assumed \( \sim P \) on line (5) and then discharged that assumption later on line (10). That is why the numeral 5 appears in the assumption column for lines (5), (7), (8), and (9) but not for line (10) or for any subsequent lines. Similarly, we assumed \( Q \) on line (11) and immediately discharged that assumption on line (12). The formula in the formula column on line (12) is a tautology, which you can see from our bookkeeping strategy by noticing that the assumption column for line (12) is empty. The formula on line (12) does not ultimately depend on any assumptions!

In our natural deduction system, we are allowed to write down any well-formed formula on any line in a proof. For example, consider the following line from the proof in Example 2.1:

\[
5 \quad (5) \quad \sim P \quad \text{A}
\]

Whenever we write down a formula in this way, we are using our first inference rule: the rule of assumption. The rule of assumption requires that we write down the line number as a new premiss number in the assumption column on that line of the proof. And we should write “A” in the justification column in order to indicate that we have used the rule of assumption. In many
cases, the formula being assumed is explicitly a premiss of an argument being formalized, as in lines (1), (2), and (3) of the proof in Example 2.1. In such cases, it is appropriate (though not required), to include a parenthetical remark indicating that the assumed formula is a premiss in the argument.

2.1 Inference Rules for Conjunction and Disjunction

In this section, I describe inference rules involving the conjunction and disjunction operators. In our proof system (as is typical for natural deduction systems), we will have an introduction rule and an elimination rule for each (basic) operator of our formal language. In this section, we will describe two introduction rules and two elimination rules. As you might guess from the names, an introduction rule tells you how to go from some formula or formulas that need not include the operator to a new formula that does include that operator. Similarly, an elimination rule tells you how to go from some formula that includes an operator to some new formula that does not include that operator.

Our first pair of rules will let us work with the conjunction operator. The conjunction introduction rule, which we denote by \( \land I \), essentially lets us add a conjunction. And the conjunction elimination rule, which we denote by \( \land E \), lets us remove a conjunction.

**Conjunction Introduction (\( \land I \))**

Suppose that \( \phi \) and \( \psi \) are well-formed formulas in our formal language. If \( \phi \) and \( \psi \) appear on separate lines in a proof, then the formula \( (\phi \land \psi) \) may be written on any later line. Schematically:

\[
\frac{\phi \quad \psi}{(\phi \land \psi)}
\]
When we use the conjunction introduction rule, we write down in the assumption column for \((\phi \land \psi)\) all of the numbers that appear in the assumption columns for \(\phi\) and \(\psi\) separately.

*Example 2.2*: \(\{P, Q, R\} \vdash ((P \land Q) \land R)\)

\[
\begin{array}{ccc}
1 & (1) & P & A \text{ (premiss)} \\
2 & (2) & Q & A \text{ (premiss)} \\
3 & (3) & R & A \text{ (premiss)} \\
1,2 & (4) & (P \land Q) & 1,2 \land I \\
1,2,3 & (5) & ((P \land Q) \land R) & 3,4 \land I \\
\end{array}
\]

**Conjunction Elimination (\(\land E\))**

Suppose that \(\phi\) and \(\psi\) are well-formed formulas in our formal language. If the formula \((\phi \land \psi)\) appears on some line in a proof, then \(\phi\) and \(\psi\) may each be written alone on a later line in the proof. Schematically, we may have either

\[
\begin{array}{c}
(\phi \land \psi) \\
\hline
\phi \\
\psi
\end{array}
\]

When we use the conjunction introduction rule, we have to write down in the assumption column for \(\phi\) (or for \(\psi\)) all of the numbers in the assumption column for \((\phi \land \psi)\).

*Example 2.3*: \(\{((P \land Q) \land R)\} \vdash Q\)

\[
\begin{array}{ccc}
1 & (1) & ((P \land Q) \land R) & A \text{ (premiss)} \\
1 & (2) & (P \land Q) & 1 \land E \\
1 & (3) & Q & 2 \land E \\
\end{array}
\]

44
Our next pair of rules will let us work with the disjunction operator. The disjunction introduction rule, which we denote by \( \lor I \), essentially lets us add a disjunction. And the disjunction elimination rule, which we denote by \( \lor E \), lets us remove a disjunction.

**Disjunction Introduction (\( \lor I \))**

Suppose that \( \phi \) and \( \psi \) are well-formed formulas in our formal language. If \( (\neg \phi \rightarrow \psi) \) appears on some line in a proof, then either the formula \( (\phi \lor \psi) \) or the formula \( (\psi \lor \phi) \) may be written on any later line. Schematically, we may have either

\[
\frac{(\neg \phi \rightarrow \psi)}{(\phi \lor \psi)} \quad \text{or} \quad \frac{(\neg \phi \rightarrow \psi)}{(\psi \lor \phi)}
\]

When we use the disjunction introduction rule, we write down in the assumption column for \( (\phi \lor \psi) \) the numbers that appear in the assumption column for \( (\neg \phi \rightarrow \psi) \).

**Example 2.4**: \( \{ (\neg P \rightarrow Q), (\neg R \rightarrow Q) \} \vdash ((P \lor Q) \land (Q \lor R)) \)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>((\neg P \rightarrow Q)) A (premiss)</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>((\neg R \rightarrow Q)) A (premiss)</td>
</tr>
<tr>
<td>1</td>
<td>(3)</td>
<td>((P \lor Q)) 1 (\lor I)</td>
</tr>
<tr>
<td>2</td>
<td>(4)</td>
<td>((Q \lor R)) 2 (\lor I)</td>
</tr>
<tr>
<td>1,2</td>
<td>(5)</td>
<td>(((P \lor Q) \land (Q \lor R))) 3,4 (\land I)</td>
</tr>
</tbody>
</table>

**Disjunction Elimination (\( \lor E \))**

Suppose that \( \phi \) and \( \psi \) are well-formed formulas in our formal language. If the formulas \( (\phi \lor \psi) \), \( (\phi \rightarrow \theta) \), and \( (\psi \rightarrow \theta) \) appear on separate lines in a proof, then \( \theta \) may be written on any later line in the proof. Schematically, we have:

\[
\frac{(\phi \lor \psi)}{(\phi \rightarrow \theta)} \frac{(\psi \rightarrow \theta)}{\theta}
\]
When we use the disjunction elimination rule, we have to write down in the assumption
column all of the numbers that appeared in the assumption column for the formulas
\((\phi \lor \psi), (\phi \rightarrow \theta), \) and \((\psi \rightarrow \theta)\).

**Example 2.5**: \{\((P \lor Q), (P \rightarrow (R \land S)), (Q \rightarrow (R \land S))\)\} \vdash R

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>((P \lor Q))</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>((P \rightarrow (R \land S)))</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>3</td>
<td>(3)</td>
<td>((Q \rightarrow (R \land S)))</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>1,2,3</td>
<td>(4)</td>
<td>((R \land S))</td>
<td>1,2,3 \lor E</td>
</tr>
<tr>
<td>1,2,3</td>
<td>(5)</td>
<td>(R)</td>
<td>4 \land E</td>
</tr>
</tbody>
</table>

**Example 2.6**: \{\((P \lor P), (P \rightarrow P)\)\} \vdash P

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>((P \lor P))</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>((P \rightarrow P))</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>1,2</td>
<td>(3)</td>
<td>(P)</td>
<td>1,2 \lor E</td>
</tr>
</tbody>
</table>

Notice that in Example 2.6, the formula in line (2) is implicitly used twice in the disjunction
elimination rule. Hence, when we add up all the assumptions used to derive our conclusion on
line (3), we only need to write down two numbers.

We now have a rule for making arbitrary assumptions and rules for working with
congruences and disjunctions. In the next section, I describe rules for working with negations
and material conditionals.

2.2 *Inference Rules for Negation and Material Conditional*

In this section, I describe inference rules involving negations and material conditionals. As with
conjunction and disjunction, we will have an introduction rule and an elimination rule for each
operator.
**Arrow Introduction (→I)**

Suppose that \( \phi \) and \( \psi \) are well-formed formulas in our formal language. If \( \psi \) appears on some line in a proof, then the formula \( (\phi \rightarrow \psi) \) may be written on any later line.

Schematically:

\[
\frac{\psi}{(\phi \rightarrow \psi)}
\]

When we use the arrow introduction rule, we write down in the assumption column for \( (\phi \rightarrow \psi) \) all the numbers that appear in the assumption column for \( \psi \).

**Example 2.7**: \( \{P, Q\} \vdash ((R \rightarrow P) \land (R \rightarrow Q)) \)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>P</th>
<th>A (premiss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(2)</td>
<td>Q</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>2</td>
<td>(3)</td>
<td>((R \rightarrow P))</td>
<td>1 →I</td>
</tr>
<tr>
<td>1</td>
<td>(4)</td>
<td>((R \rightarrow Q))</td>
<td>2 →I</td>
</tr>
<tr>
<td>1,2</td>
<td>(5)</td>
<td>((R \rightarrow P) \land (R \rightarrow Q))</td>
<td>3,4 ∧I</td>
</tr>
</tbody>
</table>

**Arrow Elimination (→E)**

Suppose that \( \phi \) and \( \psi \) are well-formed formulas in our formal language. If the formulas \( (\phi \rightarrow \psi) \) and \( \phi \) appear on separate lines in a proof, then \( \psi \) may be written on any later line in the proof. Schematically, we have:

\[
\frac{(\phi \rightarrow \psi)}{\phi} \quad \psi
\]

When we use the arrow elimination rule, we have to write down in the assumption column for \( \psi \) all of the numbers in the assumption column for \( (\phi \rightarrow \psi) \) and \( \phi \).
Example 2.8: \{((P \rightarrow P) \rightarrow (P \rightarrow R)), P \} \vdash R

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>((P \rightarrow P) \rightarrow (P \rightarrow R))</th>
<th>A (premiss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>(2)</td>
<td>P</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>2</td>
<td>(3)</td>
<td>(P \rightarrow P)</td>
<td>2 \rightarrow I</td>
</tr>
<tr>
<td>1,2</td>
<td>(4)</td>
<td>(P \rightarrow R)</td>
<td>1,3 \rightarrow E</td>
</tr>
<tr>
<td>1,2</td>
<td>(5)</td>
<td>R</td>
<td>2,4 \rightarrow E</td>
</tr>
</tbody>
</table>

Our next pair of rules will let us work with the negation operator. After describing rules for introducing and eliminating negations, we will have all of our basic rules of inference. All that will be left to do is describe conditional and indirect proof.

**Negation Introduction (\sim I)**

Suppose that \( \phi \) and \( \psi \) are well-formed formulas in our formal language. If the formulas \((\phi \rightarrow \psi)\) and \((\phi \rightarrow \sim \psi)\) appear on separate lines in a proof, then the formula \(\sim \phi\) may be written on any later line. Schematically, we have:

\[
\frac{(\phi \rightarrow \psi) \\ (\phi \rightarrow \sim \psi)}{\sim \phi}
\]

When we use the negation introduction rule, we write down in the assumption column for \(\sim \phi\) the numbers that appear in the assumption columns for the formulas \((\phi \rightarrow \psi)\) and \((\phi \rightarrow \sim \psi)\).

Example 2.9: \{P, \sim P\} \vdash \sim Q

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>P</th>
<th>A (premiss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>(2)</td>
<td>\sim P</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>1</td>
<td>(3)</td>
<td>(Q \rightarrow P)</td>
<td>1 \rightarrow I</td>
</tr>
<tr>
<td>2</td>
<td>(4)</td>
<td>(Q \rightarrow \sim P)</td>
<td>2 \rightarrow I</td>
</tr>
<tr>
<td>1,2</td>
<td>(5)</td>
<td>\sim Q</td>
<td>3,4 \rightarrow I</td>
</tr>
</tbody>
</table>
Example 2.9 gives us most of a technique for showing that anything follows from a contradiction. However, we need a way of removing negations in order to give a completely generic proof.

**Negation Elimination (¬E)**

Suppose that \( \phi \) is a well-formed formula in our formal language. If \( \sim \sim \phi \) appears on some line in a proof, then \( \phi \) may be written on any later line in the proof. Schematically, we have:

\[
\sim \sim \phi \quad \rightarrow \quad \phi
\]

When we use the negation elimination rule, we have to write down in the assumption column for \( \phi \) the numbers in the assumption column for the formula \( \sim \sim \phi \).

*Example 2.10: \{P, \sim P\} \vdash Q*

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td><em>P</em> &gt; A (premiss)</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td><em>\sim P</em> &gt; A (premiss)</td>
</tr>
<tr>
<td>1</td>
<td>(3)</td>
<td>(<em>Q \rightarrow P</em>) &gt; 1 \rightarrow I</td>
</tr>
<tr>
<td>2</td>
<td>(4)</td>
<td>(<em>Q \rightarrow \sim P</em>) &gt; 2 \rightarrow I</td>
</tr>
<tr>
<td>1,2</td>
<td>(5)</td>
<td><em>\sim Q</em> &gt; 3,4 \sim I</td>
</tr>
<tr>
<td>1,2</td>
<td>(6)</td>
<td><em>Q</em> &gt; 5 \sim E</td>
</tr>
</tbody>
</table>

Example 2.10 shows how to derive any arbitrary sentence from a contradiction. Think about the differences between Example 2.9 and Example 2.10.

We now have a rule for making arbitrary assumptions and rules for working with conjunctions, disjunctions, material conditionals, and negations. In the next section, I describe a rule for discharging assumptions.
2.3 The Rule of Conditional Proof

In a natural deduction proof system, we are allowed to freely assume any formula at any stage of a proof. However, when we make an assumption, we have to keep track of the fact that we have done so. We add a number for the supposition in the assumption column so that when we get to the end of a proof, we can be sure that we have proved our conclusion using only the allowed resources: namely, the premisses stated in the argument. Let me illustrate the need for our bookkeeping mechanism by contrasting a correct proof with two faulty ones.

Example 2.11: In this example, we will see a correct proof of the conclusion that \( \sim R \) from the premisses: \((P \lor Q)\), \((P \rightarrow \sim R)\), and \((Q \rightarrow \sim R)\). Then we will see two faulty attempts at proving \( \sim R \) from the same premisses.

This Proof is Correct:

1. (1) \((P \lor Q)\) A (premiss)
2. (2) \((P \rightarrow \sim R)\) A (premiss)
3. (3) \((Q \rightarrow \sim R)\) A (premiss)
4. (4) \(\sim R\) 1,2,3 ∨E

This Proof is NOT Correct:

1. (1) \((P \lor Q)\) A (premiss)
2. (2) \((P \rightarrow \sim R)\) A (premiss)
3. (3) \((Q \rightarrow \sim R)\) A (premiss)
4. (4) \(P\) A
5. (5) \(\sim R\) 2,4 →E

The faulty proof above correctly applies the rules of inference. However, we can see from our bookkeeping mechanism that the conclusion depends on an assumption that was not given in our original argument—namely, the formula assumed on line (4). But in order for the proof to be correct, the conclusion must ultimately depend only on the premisses of the argument. That is why we cannot just use the rule of assumption to
derive the conclusion immediately: such a proof would depend on an assumption not given in the premisses.

This Proof is NOT Correct:

1 (1) \(\sim R\) A

The proof above is not correct because the formula \(\sim R\) is not one of the premisses in the argument.

Sometimes it happens that we need (or want) to make some temporary assumptions in a proof and then discharge those assumptions before reaching the conclusion. Look back at the first example in this chapter, and you will see that the proof uses two such temporary assumptions.

We now formally introduce the rule of inference that allows us to discharge assumptions. The rule of conditional proof is the most complicated of our rules, and it typically takes students some time to get comfortable with using it. So don’t worry if the rule seems difficult initially. First, I will state the rule. Then I will try to explain how it works and give a couple of examples.

**Conditional Proof (CP)**

Suppose that \(\phi\) and \(\psi\) are well-formed formulas in our formal language. Moreover, suppose that we introduced \(\phi\) into our proof on line \(j\) by using the rule of assumption. Since we used the rule of assumption to introduce \(\phi\) on line \(j\), we know that the ultimate justification for \(\phi\) — the number that goes in the assumption column on line \(j\) — is the number \(j\) itself. Now, suppose that the formula \(\psi\) appears on some line after line \(j\). And finally, suppose that the assumption number \(j\) appears in the assumption column for \(\psi\). Then on a later line, we may write the formula \((\phi \rightarrow \psi)\) on a later line, but unlike our other rules of inference, we do not write the number \(j\) in the assumption column for
We must write down all of the other numbers that appear in the assumption column for \( \psi \), but we do not write down the number associated with the formula \( \phi \).

Schematically, the rule of conditional proof looks like this:

```
: : 
: : : 
: (j) \( \phi \) A (for CP)
: : : 
: 1,...,i,j (m) \( \psi \) \( \text{<Rule>} \)
: : : 
: 1,...,i (n) (\( \phi \to \psi \)) j,m CP
```

The rule of conditional proof is similar to the arrow introduction rule, since in both cases the rule lets us derive (or infer) a material conditional. However, the rule of conditional proof is more powerful, since it allows us to discharge assumptions that we make. The idea is that we assume some formula and use it to derive another formula (possibly the very same formula). We then put the assumed formula into the antecedent of a material conditional with the derived formula as its consequent, and we remove the assumption number for the assumed formula. Think of it as pulling the assumption out of the assumption column and into the antecedent of a material conditional. Pictorially:

```
: : 
: (j) \( \phi \) A (for CP)
: : : 
: 1,...,i,j (m) \( \psi \) \( \text{<Rule>} \)
: : : 
: 1,...,i (n) (\( \phi \to \psi \)) j,m CP
```

As you might expect, the rule of conditional proof is especially valuable for proving material conditionals.
Example 2.12: \{ (P \rightarrow (Q \land R)) \} \vdash (P \rightarrow (Q \lor R))

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>( (P \rightarrow (Q \land R)) )</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>( P )</td>
</tr>
<tr>
<td>1,2</td>
<td>(3)</td>
<td>( (Q \land R) )</td>
</tr>
<tr>
<td>1,2</td>
<td>(4)</td>
<td>( R )</td>
</tr>
<tr>
<td>1,2</td>
<td>(5)</td>
<td>( (\sim Q \rightarrow R) )</td>
</tr>
<tr>
<td>1,2</td>
<td>(6)</td>
<td>( (Q \lor R) )</td>
</tr>
<tr>
<td>1</td>
<td>(7)</td>
<td>( (P \rightarrow (Q \lor R)) )</td>
</tr>
</tbody>
</table>

Example 2.12 illustrates a general strategy for using the rule of conditional proof. If the least enclosed operator in the conclusion to be derived is a material conditional, then use the rule of assumption to get the antecedent of that conditional. Proceed to derive the consequent of the conditional, and then use the rule of conditional proof to discharge your assumption. An even simpler example of this strategy is given in Example 2.13:

Example 2.13: \{ \} \vdash (P \rightarrow P)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>( P )</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>( (P \rightarrow P) )</td>
</tr>
</tbody>
</table>

Example 2.13 is an extremely simple proof, and yet, it illustrates our most conceptually difficult rule of inference. Moreover, you will notice something odd about Example 2.13: it is a proof of a conclusion from no premisses at all! Whenever a formula may be derived from the empty set of premisses—i.e. from no premisses at all—we call that formula a *theorem* of our proof system. Hence, the formula \( (P \rightarrow P) \) is a theorem of our proof system.

The derivation technique in Example 2.13 gets used over and over again in our natural deduction proof system. For example, we can use the derivation technique from Example 2.13 as follows:
Example 2.14: \( \{(P \vee P)\} \vdash P \)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>( (P \vee P) )</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>( P )</td>
<td>A (for CP)</td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td>( (P \to P) )</td>
<td>2 CP</td>
</tr>
<tr>
<td>1</td>
<td>(4)</td>
<td>( P )</td>
<td>1,3 ( \vee )E</td>
</tr>
</tbody>
</table>

The proof in Example 2.14 might look a little odd. Why didn’t we just stop when we wrote down the conclusion in line (2)? The answer is that when we wrote down \( P \) in line (2), the formula \( P \) depended on an assumption different from the premiss in the argument being formalized. Hence, we needed to discharge the assumption in that line before concluding the proof.

We can (and will) use the rule of conditional proof to show that many more complicated formulas are theorems of our proof system. In fact, every time we give a proof that has premisses, we could easily transform it into a proof of a theorem. Here is how. Take all of the premisses used in the proof and form a large conjunction. For example, if you had the set of premisses \( \{P, Q, R\} \), you would form the conjunction \( ((P \land Q) \land R) \). Now form a material conditional with the conjunction of the premisses as the antecedent of the conditional and the conclusion of the argument as the consequent of the conditional. That material conditional will be a theorem in our proof system. And we can prove the theorem using the old proof with a few small modifications.

The old proof begins by assuming the premisses on separate lines by the rule of assumption. In the new proof, we assume the antecedent of the material conditional for the purpose of conditional proof. We then use conjunction elimination to recover all of the premisses used in the old proof. Next, we use the same steps from the old proof to derive the conclusion of the old proof. Finally, we discharge the assumption into the antecedent of a material conditional
that has the old conclusion as its consequent. At this point, we have proved a theorem in our proof system.

**Example 2.15:** \{ \} \vdash (((\neg P \rightarrow Q) \land (\neg R \rightarrow Q)) \rightarrow ((P \lor Q) \land (Q \lor R)))

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>((\neg P \rightarrow Q) \land (\neg R \rightarrow Q))</td>
<td>A (for CP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(2)</td>
<td>((\neg P \rightarrow Q))</td>
<td>1 \land E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(3)</td>
<td>((\neg R \rightarrow Q))</td>
<td>1 \land E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(4)</td>
<td>((P \lor Q))</td>
<td>2 \lor I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(5)</td>
<td>((Q \lor R))</td>
<td>3 \lor I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(6)</td>
<td>(((P \lor Q) \land (Q \lor R)))</td>
<td>4,5 \land I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(7)</td>
<td>((1) \rightarrow (6))</td>
<td>1,6 \ CP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Example 2.15, we applied the procedure sketched above for converting a proof that has premisses into a proof of a theorem. (We converted the proof in Example 2.4.) Also, we introduced a convention to save space and make it easier to write down the proof, namely in line (7), we used line numbers to represent formulas.

At this point, our collection of inference rules for zeroth-order logic is complete. Our collection of rules has a lot of symmetry. For every operator, we have an introduction rule and an elimination rule. Moreover, we have a rule for introducing arbitrary assumptions, and we have a rule for discharging those assumptions. In the next section, we consider some common strategies for using our rules of inference in constructing proofs.

**2.4 Indirect Proof**

The rule of conditional proof naturally lends itself to the following proof strategy. Suppose you want to derive the formula \((\phi \rightarrow \psi)\), where \(\phi\) and \(\psi\) are well-formed formulas. Then assume the formula \(\phi\), derive the formula \(\psi\), and apply the rule of conditional proof to discharge the assumption. The foregoing proof strategy is sometimes called *direct proof*. 

55
By contrast, the rule of conditional proof may also be deployed in what is called *indirect proof*. Indirect proofs work by assuming a formula with either one more negation or one fewer negation than is in the conclusion to be proved and then deriving a pair of contradictory formulas. The initial assumption is then discharged and the rule of negation introduction (plus maybe the rule of negation elimination) is used to obtain the desired conclusion. Schematically, indirect proof looks like this:

\[
\begin{array}{cccc}
\vdots & \vdots & \vdots & \\
j & (j) & \phi & A^* \text{ (for reductio)} \\
\vdots & \vdots & \vdots & \\
1,...,i,j & (m) & \psi & \text{<Rule>}
\end{array}
\]

\[
\begin{array}{cccc}
\vdots & \vdots & \vdots & \\
1,...,i,k & (n) & \sim \psi & \text{<Rule>}
\end{array}
\]

\[
\begin{array}{cccc}
\vdots & \vdots & \vdots & \\
1,...,i & (o) & (\phi \to \psi) & j,m \text{ CP}
\end{array}
\]

\[
\begin{array}{cccc}
\vdots & \vdots & \vdots & \\
1,...,k & (p) & (\phi \to \sim \psi) & n \rightarrow I
\end{array}
\]

\[
\begin{array}{cccc}
\vdots & \vdots & \vdots & \\
1,...,i,k & (q) & \sim \phi & o,p \sim I
\end{array}
\]

In order to flag that we intend to reject the assumption that \( \phi \), we either add an asterisk to the justification or we add the remark that the assumption is made for *reductio* or both. Since the proof technique works by deriving a contradiction, indirect proof is sometimes called *proof by contradiction* or *reductio ad absurdum* (literally, reduction to absurdity).

**Example 2.16:** \{ \( \sim (P \to Q) \) \} \vdash \sim Q

\[
\begin{array}{cccc}
1 & (1) & \sim (P \to Q) & A \text{ (premiss)} \\
2 & (2) & Q & A^* \text{ (for reductio)} \\
2 & (3) & (P \to Q) & 2 \rightarrow I \\
 & (4) & (Q \to (P \to Q)) & 2,3 \text{ CP} \\
1 & (5) & (Q \to \sim (P \to Q)) & 1 \rightarrow I \\
1 & (6) & \sim Q & 4,5 \sim I
\end{array}
\]

**Example 2.17:** \{P\} \vdash \sim \sim P

\[
\begin{array}{cccc}
1 & (1) & P & A \text{ (premiss)} \\
2 & (2) & \sim P & A^* \text{ (for reductio)} \\
1 & (3) & (\sim P \to P) & 1 \rightarrow I \\
 & (4) & (\sim P \to \sim P) & 2 \text{ CP} \\
1 & (5) & \sim \sim P & 3,4 \sim I
\end{array}
\]
Example 2.18: \( \{P, Q\} \vdash \neg(P \to \neg Q) \)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th></th>
<th>(2)</th>
<th></th>
<th>(3)</th>
<th></th>
<th>(4)</th>
<th></th>
<th>(5)</th>
<th></th>
<th>(6)</th>
<th></th>
<th>(7)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P</td>
<td></td>
<td>2</td>
<td>Q</td>
<td></td>
<td>3</td>
<td>(P → \neg Q)</td>
<td>A* (for reductio)</td>
<td>4</td>
<td>\neg Q</td>
<td>1,3 →E</td>
<td>5</td>
<td>((P → \neg Q) → Q)</td>
<td>2 →I</td>
</tr>
<tr>
<td>1,2</td>
<td>\neg(P → \neg Q)</td>
<td>1,3 CP</td>
<td>6</td>
<td>((P → \neg Q) → \neg Q)</td>
<td>3,4 CP</td>
<td>7</td>
<td>\neg(P → \neg Q)</td>
<td>5,6 \neg I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Indirect proof can be used in two closely related but distinct ways: to prove a formula \( \phi \) or to prove a formula \( \neg \phi \). In the first case, one derives both the formula \((\neg \phi \to \psi)\) and the formula \((\neg \phi \to \neg \psi)\). Then one uses negation introduction \((\neg I)\) to obtain the formula \(\neg \neg \phi\), from which one may derive \(\phi\) by negation elimination. In the second case, one derives the formula \((\phi \to \psi)\) and the formula \((\phi \to \neg \psi)\). Then one uses negation introduction \((\neg I)\) to obtain the formula \(\neg \phi\) directly.

When one is trying to prove a material conditional, direct and indirect proof techniques may be combined by assuming both the antecedent of the conditional and the negation of the consequent of the conditional (or if the consequent is itself a negation, by assuming the formula being negated in the consequent). A proof is often easier to see when one has lots of assumptions to work with.
References


Chapter 3: First-Order Logic

In Chapter 1, we began building our formal language, we introduced truth functions and truth tables, and we described a special notion of evidential support in which the conclusion of an argument is guaranteed to be true if its premisses are all true. In Chapter 2, we described a proof system for our formal language. In this chapter, we add more complexity to our formal language. The more complex formal language that we are building in this chapter is the formal language of first-order logic or predicate logic.

As you might expect, our expanded formal language adds new symbols to the ones from Chapter 1. First, we will use lower-case letters at the beginning of the alphabet, like a, b, and c, which we will call constants, and we will use lower-case letters at the end of the alphabet, like x, y, and z, which we will call variables. Constants and variables are called terms. Second, we will use capital letters, like A, B, and C, which we will call predicates. And finally, we will use two special symbols, ∀ and ∃, which we will call quantifiers. Each quantifier gets its own name. The first quantifier, ∀, is called the universal quantifier. The second quantifier, ∃, is called the existential quantifier. Just as with zeroth-order logic, now that we have the symbols for our formal language, we need to provide a grammar.

3.1 Formation Rules

As with zeroth-order logic, we will recursively define the well-formed formulas of our language. The formation rules from zeroth-order logic are the first two formation rules for first-order logic. Thus, a single sentence letter is a well-formed formula, and if ϕ and ψ are well-formed formulas, then ~ϕ, (ϕ ∧ ψ), (ϕ ∨ ψ), and (ϕ → ψ) are well-formed formulas. In addition to the formation
rules from zeroth-order logic, we have the following two new rules. First, if $\Phi$ is a predicate symbol and $\alpha_1, \alpha_2, \ldots, \alpha_n$ are terms (either all constants or all variables or some combination of constants and variables), then $\Phi\alpha_1\alpha_2\ldots\alpha_n$ is a well-formed formula. Hence, the following are well-formed formulas:

- Aaa
- Bazd
- Cyz
- $(Aaz \rightarrow Byzd)$
- $\sim Cab$

And the following formulas are not well-formed:

- AB
- AaB
- aBy

Second, let $\phi$ be a well-formed formula and let $\alpha$ be a variable, then $(\forall\alpha)\phi$ and $(\exists\alpha)\phi$ are well-formed formulas. Notice that $\alpha$ must be a variable, not just an arbitrary term. In our formal language, the following are all well-formed formulas:

- $(\forall x)P$
- $(\forall y)Aay$
- $(\exists x)(\neg Bxzd \land Aay)$
- $(\forall x)(Bx \rightarrow P)$
- $(\exists x)(Bx \land Cx)$

However, the following are not well-formed:

- $(\forall a)P$
- $(\exists x \rightarrow \neg Bxzd)$
- $(\forall B)Baz$
- $(\exists x \lor \forall a)Cxa$
- $(\exists x)(Bx \land Cx \land Dx)$

According to our second new formation rule, both $(\forall x)(Bx \rightarrow Cx)$ and $((\forall x)Bx \rightarrow Cx)$ are well-formed, but as we will see, they say very different things. (In fact, the second one is well-formed
but does not get a truth-value. The moral is that you need to pay close attention to the parentheses!

For any variable $\alpha$, the expressions $(\forall \alpha)$ and $(\exists \alpha)$ are called quantifiers over $\alpha$. Consider the well-formed formulas $(\forall \alpha) \phi$ and $(\exists \alpha) \phi$. In each case, we will say that the scope of the quantifier over $\alpha$ is the formula $\phi$, and any term that appears in the formula $\phi$ is said to be in the scope of the quantifier. In a well-formed formula, the variable $\alpha$ is said to be bound by a quantifier if it appears in the scope of a quantifier over $\alpha$ or if it appears in either of the quantifier expressions $(\forall \alpha)$ or $(\exists \alpha)$. If a variable is not bound by a quantifier, then it is said to be a free variable. In the formulas below, the variables $x$ and $y$ are bound in some cases and free in other cases. Bound variables are bold. Free variables are red.

$$
\begin{align*}
Bx \\
(\forall y)Bx \\
(\sim Bx \lor Cy) \\
(\forall x)(\sim Bx \lor Cx) \\
(\forall x)(\sim Bx \lor Cx) \\
(\exists x)(\forall y)(\sim Bx \lor Cy) \\
(\exists x)(\forall y)(\sim Bx \lor Cy)
\end{align*}
$$

Any well-formed formula that contains no free variables is called a sentence. (Hence, only two of the formulas in the list above are sentences.) In our zeroth-order language, we did not need to distinguish between well-formed formulas and sentences: every well-formed formula in our zeroth-order language is a sentence. However, in our first-order language, the distinction is important for reasons having to do with translation and semantics.

3.2 Translations

In Chapter 1, we began with a picture of sentences in ordinary language as black boxes having lots of internal structure that is basically invisible. We began the process of opening up the black box and representing that internal structure, but we didn’t get very far. Our zeroth-order formal
language barely scratches the surface of the logical complexity of ordinary language. In this chapter, we gain considerably more power for formally representing ordinary language sentences. Consider the four English sentences that we first saw in Section 1.2:

- Every child is innocent.
- Susan doesn’t believe that President Obama is secretly a Muslim.
- Jesus wept.
- Talan is my son, and his birthday is in October.

With terms, predicates, and quantifiers, we can represent much more of the internal structure of these sentences than we could with the tools from Chapter 1. In order to do that, we need to think a little bit about what terms, predicates, and quantifiers may be used to represent.

Let’s start with terms. Terms are divided into two kinds: constants and variables. The constants of our formal language—a, b, c, and so on—represent the logical subjects of sentences. They may be used to stand for people, like Susan, Jesus, and Talan. Constants may also be used to stand for specific places, times, objects, events, or abstracta (like hope or justice). A constant on its own may represent an individual. By contrast, a variable is just a placeholder for a bunch of constants. A variable does not stand for anything on its own. If a variable stands for something, it is only because it is playing a specific role in a sentence in which it appears.

Predicates are truth functions that take terms as inputs. A predicate represents a property, like the property of being innocent or the property of having wept. When we join a constant term to a predicate, we are saying that the individual represented by the constant has or exhibits the property represented by the predicate. For example, if we let j stand for Jesus and W stand for the property of having wept, then we can translate the sentence, “Jesus wept,” into our formal language as Wj. (Notice that we are cheating a bit here with respect to tense.) Or if we let t stand for Talan and S stand for the property of being the son of the author of this text, then we can translate the sentence, “Talan is my son,” into our formal language as St. When we feed a
constant term into a predicate, the result is a sentence, which is either true or false. Talan either really is my son, making the sentence true, or he is not, making the sentence false.

Both examples of predicates in the previous paragraph are one-place functions: each predicate takes only a single argument. Predicates that take a single argument are sometimes called *monadic*. Not all predicates are monadic. A predicate could have any finite number of places. When a predicate takes more than one argument, it is called *polyadic*. But more commonly, a polyadic predicate is called a *relation*. An example of a two-place relation is the phrase, “... is taller than ---,” which we might represent with the letter H. In order to turn H into a sentence that gets a definite truth-value, we need to attach it to two constant terms. Using the constants j and t to represent Jesus and Talan, respectively, we could translate the sentence, “Jesus is taller than Talan,” into our formal language as Hjt. As we will see in more detail in the next section, order (typically) matters for relations. Suppose Jesus is more than one meter tall, and suppose Talan is less than one meter tall. In that case, Hjt is true, but Htj is false.

The real power of predicates and individuals comes in when we add the idea of quantification. The quantifiers let us talk about many individuals. The universal quantifier, ∀, is used to translate “for all” and related expressions. For example, the sentence (∀x)Gx says, “For all x, the predicate G is true of x,” or in other words, “Everything has the property G.”

Intuitively, the universal quantifier functions like an infinite conjunction of instances of the formula in the scope of the quantifier where the variable has been replaced by all the constant terms. From this perspective, (∀x)Gx is equivalent to the infinite sentence (Ga ∧ Gb ∧ Gc ∧ … and so on for all the constants. On another intuitive way of thinking about the sentence (∀x)Gx, the sentence says that any object you pick will have the property G. Hence, one might think of assertions involving the universal quantifier as part of a sort of game played by two people. If
one player asserts a universally quantified sentence, then he or she is saying that the other player cannot find an individual that fails to satisfy the formula in the scope of the quantifier. To say that everything has the property G is to say that no one can show me an instance of a non-G.

The existential quantifier, $\exists$, is used to translate “there exists” and related expressions. For example, the sentence $(\exists x)Gx$ says, “There exists an x such that the predicate G is true of x,” or in other words, “Something has the property G.” The existential quantifier only commits to there being at least one thing that satisfies the formula in the scope of the quantifier. Hence, intuitively, the existential quantifier functions like an infinite disjunction. From this perspective, $(\exists x)Gx$ is equivalent to the infinite sentence $(Ga \lor Gb \lor Gc \ldots$ and so on for all the constants. On an alternative intuitive reading, the sentence $(\exists x)Gx$ says that its utterer can pick an object that has the property G. Hence, one might think of assertions involving the existential quantifier as playing a slightly different game. If one player asserts an existentially quantified sentence, then he or she is saying that he or she can find at least one individual that satisfies the formula in the scope of the quantifier.

Using variables, predicates, and quantifiers, we can represent sentences like, “Everyone loves something or other,” and “Nothing is better than this piece of cake.” If we let P stand for the predicate “… is a person,” and we let L stand for the relation “… loves ---,” then we can translate the first sentence with $(\forall x)(Px \to (\exists y)Lxy)$, which says, “For anything you pick, if that thing is a person, then you can find at least one thing such that the first thing you picked loves the second thing.” (Notice how important the variable letters are for making the sentences clear! Using letters to name the possible objects under consideration actually makes the whole sentence easier to understand than using phrases like “that thing” and “the second thing” and so forth.) If we let the constant term c stand for “this piece of cake,” and we let the predicate letter B stand
for the relation “… is better than ---,” then we can translate the second sentence above with
~(∃x)Bxc, which says, “It is not the case that there is at least one thing that is better than this
piece of cake.” (What would change if you wanted to translate the sentence “Nothing is better
than cake”?)

Translations into our first-order formal language are very tricky. Let us consider a class
of sentences that have standard translations: the categorical sentences. A categorical sentence is
a sentence that has the same logical form as one of the following four example sentences:

All students are virtuous.
No politician is honest.
Some cats are good pets.
Some clarinets are not made of wood.

Not only is translating the categorical sentences into our first-order language a useful exercise,
but also, the categorical sentences have a venerable place in the history of logic. Aristotle’s logic
of the syllogism is essentially a logic for categorical sentences. Aristotle and the medieval
scholars who studied his writings noticed many properties of categorical sentences and built an
impressive account of their logical relations. We will not delve into the medieval logic of
categorical sentences but simply translate the general forms into our first-order formal language:

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-type</td>
<td>All Ss are Ps.</td>
<td>(∀x)(Sx → Px)</td>
</tr>
<tr>
<td>E-type</td>
<td>No Ss are Ps.</td>
<td>~(∃x)(Sx ∧ Px)</td>
</tr>
<tr>
<td>I-type</td>
<td>Some Ss are Ps.</td>
<td>(∃x)(Sx ∧ Px)</td>
</tr>
<tr>
<td>O-type</td>
<td>Some Ss are not Ps.</td>
<td>(∃x)(Sx ∧ ~Px)</td>
</tr>
</tbody>
</table>

It should be obvious from the translations that E-type sentences and I-type sentences are
contradictory sentences. What might not be immediately apparent is that A-type sentences and
O-type sentences are also contradictory sentences. That is, we could have translated the O-type
sentences as ~(∀x)(Sx → Px) instead. Hence, if an A-type sentence is true, then its
corresponding O-type sentence must be false, and vice versa. Later on, we will see that this fact has an interesting application to simple problems in probability theory.

With the standard translations of categorical sentences in hand, many sentences become simple to translate: just follow the recipe. For example, the first example sentence from Chapter 1, “Every child is innocent,” is just an A-type categorical sentence. Let’s translate it into our formal language. Let C represent the predicate “… is a child,” and let I represent the predicate “… is innocent.” Then the sentence, “Every child is innocent,” may be translated into our formal language as \((\forall x)(Cx \rightarrow Ix)\). What translation would have been appropriate if the target sentence had been “No child is innocent”?

3.3 Relations

Categorical sentences are interesting and useful. But the real power of first-order logic comes through relations (in combination with the quantifiers). We saw some examples of relations in the previous section: “… is taller than ---,” “… is better than ---,” and “… loves ---.” All of those relations have two places, and we will spend most of our time thinking about two-place relations. However, a relation may have arbitrarily many places. For example, we might talk about the relation “__ gives __ to __” or the relation “__ makes __ for __ on __.”

Since relations are just multiple-place predicates, relations are a kind of truth function. When one inputs a number of constant terms equal to the number of places in the relation, the relation returns a truth-value. Suppose that on Harry’s birthday, Sarah makes a cake to give to him. Let h stand for Harry, b stand for Harry’s birthday, a stand for Sarah, and c stand for the specific cake that Sarah makes. Moreover, let G stand for the relation “__ gives __ to __,” and let M stand for the relation “__ makes __ for __ on __.” Then \(G\)ach is true and \(M\)achb is true. But
Ghca is false, Mchab is false, and so are most of the possible combinations of these constants with those relations.

As I said above, we will spend most of our time thinking about two-place relations. For two-place relations, when we have Rab for the relation R with the constants a and b, we will say that a is R-related to b. The inputs to a relation are called the relata of the relation. Now, let’s think about special properties that hold for some (but not other) two-place relations. Some relations are reflexive. When a relation R is reflexive, every constant is R-related to itself. We can formally define the reflexive property as follows: A relation R is reflexive if and only if the sentence \((\forall x)Rxx\) is true. Examples of reflexive relations include “… is the same color as ---,” “… is at least as complicated as ---,” and “… weighs no more than ---.” By contrast, a relation is (strongly) irreflexive if no constant is R-related to itself. Formally, a relation R is irreflexive if and only if the sentence \((\forall x)\neg Rxx\) is true. Examples of irreflexive relations include “… is strictly greater than ---,” “… is significantly uglier than ---,” and “… is different than ---.” Some relations are not reflexive but they are also not irreflexive. For example, consider the relation “… is two times ---.” The sentence “Zero is two times zero,” is true; but the sentence “Four is two times four,” is false. Hence, for some constants, the relation “… is two times ---” is true when that constant is put into both places in the relation, but not so for just any constant you pick.

A relation is symmetric if it has the same truth-value when the same two constants are put into the relation, regardless of the order in which they are put into the relation. In other words, the relation holds in both directions: from the first relatum (the singular of relata) to the second and from the second relatum back to the first. Formally, a relation R is symmetric if and only if the sentence \((\forall x)(\forall y)(Rxy \rightarrow Ryx)\) is true. Examples of symmetric relations include: “… is the sibling of ---,” “… went on a date with ---,” and “… is located within five meters of ---.” If these
relations hold for constants a and b in that order, then the same relations hold for constants b and a in that order. For example, suppose that Mary and George, who are not siblings, are holding hands on their second date. Let m stand for Mary and g stand for George. Moreover, let S stand for “… is a sibling of ---,” let D stand for “… went on a date with ---,” and let W stand for “… is located within five meters of ---.” Then the following six material conditionals must all be true:
(Smg → Sgm), (Sgm → Smg), (Dmg → Dgm), (Dgm → Dmg), (Wmg → Wgm), and (Wgm → Wmg). Take a few seconds and make sure you are convinced that the conditionals are all true.

Why did we need to check six conditionals rather than just three?

One might think that if a relation is not symmetric, then it is asymmetric. But the standard use of “asymmetric” in logic and mathematics, which we will follow here, is a bit stronger. A relation R is asymmetric if and only if the sentence $\forall x)(\forall y)(Rxy \to \sim Ryx)$ is true. Examples of asymmetric relations include “… is the mother of ---,” “… is much furrier than ---,” and “… is smaller than ---.” Some relations are neither symmetric nor asymmetric. For example, while it is a sad fact that the relation “… loves ---” is not symmetric, since it sometimes happens that one’s love is not returned, we are spared the nightmare scenario in which “… loves ---” is asymmetric: sometimes a person’s love is returned.

A relation R is transitive just in case for any three things you pick, if the first is R-related to the second and the second is R-related to the third, then the first is also R-related to the third. Formally, R is transitive if and only if the sentence $(\forall x)(\forall y)(\forall z)((Rxy \land Ryz) \to Rxz)$ is true. Examples of transitive relations include “… is older than ---,” “… is poorer than ---,” and “… arrived later than ---.” If a relation is not transitive, then we say that it is non-transitive. The relation “… has met ---” is non-transitive. I have met people who have met the President of the United States. But I have not met the President of the United States. However, for many
collections of three people, it is true that the first has met the second, the second has met the third, and the first has met the third. Some relations are such that if a thing is related to a second thing and the second thing is related to a third thing, then the first thing cannot be related to the third thing. Such relations are called intransitive. The relation “… is the father of …” is intransitive. If George is the father of Bucky and Bucky is the father of Douglas, then George cannot be the father of Douglas. Formally, a relation \( R \) is intransitive if and only if the sentence

\[
(\forall x)(\forall y)(\forall z)((Rxy \land Ryz) \rightarrow \neg Rxz)
\]

is true.

A two-place relation over some collection of individuals is nicely represented by a directed graph. A directed graph is a collection of points, called vertices or nodes, possibly connected by arrows, called directed edges. An example of a directed graph is given in Figure 3.1 below:

**Figure 3.1: Directed Graph Over Letters a – e**

When representing a relation using a directed graph, nodes stand for concrete individuals, just like constants. In fact, we will often label the nodes (as in the graph above) with constant terms in order to make the graph easier to interpret. The directed edges in the graph represent the relation itself. Whenever an edge (arrow) in the graph goes from one node to a second node, the individual represented by the first node is related to the individual represented by the second
node. In the graph above, b is related to c by the represented relation. The individual b is also related to e, and the individual b is related to itself.

Using directed graphs to represent two-place relations makes the special properties we discussed above easy to see. For example, when a relation is reflexive, every node in the graph has a loop (an arrow from the node to itself). The relation represented in Figure 3.1 is not reflexive, since a, c, and d do not have loops. But the relation is not (strong) irreflexive, since both b and e do have loops. Similarly, when a relation is symmetric, every arrow in the graph must have a corresponding arrow pointing back in the opposite direction. In other words, a relation is symmetric if and only if between every pair of nodes there are either two arrows (one in each direction) or no arrows at all. The relation in Figure 3.1 is not symmetric, since a is related to c but c is not related to a. (The relation fails to be symmetric for other reasons as well. Can you see what they are?) However, the relation is not asymmetric, since b is related to e and also e is related to b.

Is the relation in Figure 3.1 transitive? Stop and think about the question for a few seconds before reading the next sentence. The answer is no, the relation represented in Figure 3.1 is not transitive. Why not? One way the relation fails to be transitive is this: d is related to e and e is related to b, but d is not related to b. Can you find some other ways in which the relation represented in Figure 3.1 fails to be transitive? Is the relation in Figure 3.1 intransitive? No, the relation is not intransitive either. One reason it is not intransitive is that d is related to e and e is related to itself. Why does that suffice to show that the relation is not intransitive? Another reason the relation in Figure 3.1 is not intransitive is that b is related to e, e is related to b, and b is related to itself.
Now, consider the relation “… comes earlier than --- in the alphabet.” For simplicity, we will restrict attention to the first five letters in the alphabet. The graph of the relation is given in Figure 3.2 below:

![Figure 3.2: Graph of the “… comes earlier than --- in the alphabet” Relation](image)

Take a few moments and think about which special properties the relation in Figure 3.2 exhibits. Is it reflexive? Symmetric? Transitive? The relation pictured in Figure 3.2 is irreflexive, since no node has an arrow pointing into itself. The relation is also asymmetric, since no pair of nodes is unrelated and no pair of nodes has two arrows connecting them. Finally, the relation is transitive.

If a relation is reflexive, symmetric, and transitive, then we say that it is an equivalence relation. The smallest and most important equivalence relation is the identity relation. The identity relation relates everything to itself and to nothing else. The graph of the identity relation over the first five letters of the alphabet is pictured in Figure 3.3:

![Figure 3.3: Graph of the Identity Relation](image)
The identity relation is on the one hand blindingly obvious and simple and on the other hand somewhat mysterious and puzzling. We know that everything is identical to itself and to nothing else. But what does it mean for something to be identical to itself? To get some sense of puzzlement concerning identity, consider the following problem having to do with the special case of personal identity.

Pick an event in what you think of as your past. Maybe it’s your eighth birthday or your graduation from high school. Are you the same person as that little boy or girl on his or her eighth birthday? Are you the same person as that high school graduate? Suppose we had a duplicator – something like a Star Trek transporter except that it makes two copies of anyone standing on the transport pad instead of just one. Now suppose that what you usually think of as your eight-year-old self had gotten into such a duplicator several years ago so that you have an exact duplicate. Are you identical to that eight-year-old? What about the other duplicate? If you are identical to that eight-year-old, shouldn’t the other duplicate also be identical to that eight-year-old? If so, then since identity is transitive, it seems that you are identical to the other duplicate! So, maybe you aren’t identical to the eight-year-old in the duplication scenario after all. But if you are not identical to the eight-year-old in the duplication scenario, why aren’t you? And if you are not identical to the eight-year-old in the duplication scenario, why think you are identical to the eight-year-old in the real-world case where there is no duplication?\(^5\)

For all its mystery, the identity relation is extremely important. Consequently, it gets its own special symbol. In our formal language, the identity relation will be represented by the equal sign, =. And rather than writing =ab, as we do for other relations, we will write (a = b). Identity,

\(^5\) For much more on the identity relation in general, see Noonan (2009). For more on personal identity, see Olson (2010).
or equality, is just a relation. As with other two-place relations, when we input two constant terms, the identity relation outputs a truth-value.

3.4 Small Worlds, Models, and Validity

In zeroth-order logic, we started with the idea of interpreting our sentence letters by assigning each sentence a truth-value. But we said that in some sense logicians do not care about what the actual truth-values are for any sentences that we consider. Rather, we care about whether specified collections of sentences (premises) stand in a relation of evidential support to some other sentence (conclusion). Consequently, we did not spend very long thinking about particular interpretations but spent our time thinking about all the possible interpretations that some collection of simple sentences could be given. In other words, we started thinking about possible worlds.

In zeroth-order logic, the possible worlds are easy to think about. A possible world is just a specification of the truth-values for a collection of sentences. This led us to truth tables. In effect, each row in a truth table picks out a different possible world. And then we used truth tables to test arguments for validity. We would like to do something similar in first-order logic, but unfortunately, the method of truth tables does not work nicely in first-order logic.

Whereas, in zeroth-order logic, our basic elements are complete sentences, in first-order logic, our basic elements are constants – concrete individuals that could be given names – and predicates – either properties that the constants might have or relations that the constants might stand in to one another. A sentence takes a truth-value. In our formal language, a sentence is either true or false. But a constant does not take a truth-value. It doesn’t make much sense to say that this or that desk is true or that the cupcake in my hand is false. (At least, we will suppose for this course that it doesn’t make any sense to say of a constant that it is true or false. A poet might
say that a cupcake is false as a metaphor, perhaps.) Similarly, predicates and relations are not the sorts of things that take truth-values either. Predicates like “… is sweet” are neither true nor false. And relations like “… jumps over ---” are neither true nor false. Only when combined together do we get things that take truth-values, e.g. “The cupcake in my hand is sweet.”

Instead of assigning truth-values directly to sentences in first-order logic, we will directly construct what we will call *small worlds*, and then the truth-values of sentences that we are interested in will be evaluated with respect to these small worlds. As the name suggests, small worlds are small! A small world specifies some constants that we will care about, and it specifies which predicates and relations hold for those constants. A small world is a toy version of a full-blown possible world. But whereas a possible world is usually thought of as a very rich sort of thing, a small world will be much more impoverished.

We are free to construct small worlds however we see fit, as long as each world contains at least one individual. For example, we could construct a small world having four constants (labeled a, b, c, and d), two monadic predicates (labeled G and H), and one two-place relation (labeled R). That gives the basic building blocks of the small world. In addition to the building blocks, themselves, a small world has to specify the way the predicates and relations apply (or fail to apply) to the specified constants. For example, we will construct our small world as follows. Let the predicate G apply to a and b, but not to c or d. Let the predicate H apply to b and c, but not to a or d. And let the following relations be the only ones that hold in our small world: a is R-related to b, a is R-related to d, b is R-related to b, c is R-related to b, and d is R-related to

---

6 We will not get bogged down in a rigorous, mathematical account of small worlds. The interested reader may consult Enderton (2001), Marker (2002), or Sider (2010). These writers use the term *structure* rather than our term *small world*.
a. Let’s call our small world $\mathcal{W}$. Drawing a picture is sometimes helpful. The drawing in Figure 3.4 represents the small world $\mathcal{W}$.

**Figure 3.4: A Small World**

The drawing in Figure 3.4 specifies the constants and the $R$-relations over those constants using a directed graph. In addition to the graph structure, the drawing uses closed curves (colored blue and red) to tell us that the constants $a$ and $b$ have the monadic predicate $G$ and the constants $b$ and $c$ have the monadic predicate $H$. For monadic predicates, an individual has the predicate if and only if the constant corresponding to that individual is enclosed by a curve representing the relevant predicate.

We could accomplish the same thing as our picture by completely describing the five features of our small world that characterize it, namely the individuals it contains, the monadic predicates that we care about, how those predicates apply or fail to apply to the individuals, the relations that we care about, and how those relations apply or fail to apply to the individuals. In the case of the small world pictured in Figure 3.4, the description we want could be given as follows:

- **Individuals:** $a, b, c, d$
- **Monadic Predicates:** $G, H$  
  The following hold: $Ga, Gb, Hb, Hc$
- **Relations:** $R$  
  The following hold: $Rab, Rad, Rbb, Rcb, Rda$
By convention, if a predication or relation is not listed, then its negation is understood to be true. For example, \( \sim \text{Gc} \) is true and \( \sim \text{Raa} \) is true.

Why are we bothering to construct small worlds? The fundamental motivation for constructing small worlds is to give us a way of evaluating the sentences in our formal language. Small worlds let us evaluate sentences because we can say what the truth-value of a sentence is with respect to a specific small world. For example, with respect to the small world \( \mathcal{W} \) described above (and represented in Figure 3.4), the sentence \( \text{Ga} \) is true and the sentence \( \text{Ha} \) is false. Here are some examples of sentences that are true with respect to \( \mathcal{W} \):

\[
\begin{align*}
\sim \text{Gd} & \\
(\text{Ga} \rightarrow \text{Hc}) & \\
(\text{Gb} \land \text{Hb}) & \\
(\text{Rab} \lor \text{Gd}) & \\
(\text{Rdc} \rightarrow \text{Hd}) & \\
\sim (\text{Rad} \land \text{Ha}) & \\
(\exists x)(\text{Gx} \land \text{Hx}) & \\
(\forall x)(\exists y) \text{Rxy} & \\
(\exists x)(\exists y)((\text{Gx} \land \text{Hy}) \land \text{Rxy}) & \\
(\forall x)(\text{Rx} \rightarrow (\text{Gx} \land \text{Hx})) &
\end{align*}
\]

And here are some sentences that are false with respect to \( \mathcal{W} \):

\[
\begin{align*}
\sim \text{Ga} & \\
(\text{Ga} \rightarrow \text{Hd}) & \\
(\text{Gb} \land \text{Gc}) & \\
(\text{Rac} \lor \text{Ha}) & \\
(\text{Rbb} \rightarrow \text{Gd}) & \\
\sim (\text{Rbb} \land \text{Ga}) & \\
(\forall x)(\text{Gx} \land \text{Hx}) & \\
(\forall x)(\exists y) \text{Ryx} & \\
(\exists x)(\exists y)((\text{Gx} \land \text{Hy}) \land (x = y)) & \\
(\forall x)(\text{Rdx} \rightarrow (\text{Gx} \land \text{Hx})) &
\end{align*}
\]

Whenever a sentence \( \phi \) is true with respect to some small world \( \mathcal{M} \), we will say that \( \mathcal{M} \) is a model for \( \phi \). We will write \( \{ \} \models_{\mathcal{M}} \phi \) in order to indicate that \( \mathcal{M} \) is a model for \( \phi \). In other words, if we write \( \{ \} \models_{\mathcal{M}} \phi \), then the sentence \( \phi \) is true with respect to the small world \( \mathcal{M} \). A small world \( \mathcal{M} \) is a model for a collection of sentences if and only if \( \mathcal{M} \) is a model for every sentence in the collection.

As you might expect, if we write \( \{ \psi_1, \psi_2, \ldots, \psi_n \} \models_{\mathcal{M}} \phi \) for some small world \( \mathcal{W} \) and some sentences \( \psi_1, \psi_2, \ldots, \psi_n \), and \( \phi \), then we are saying that if \( \psi_1, \psi_2, \ldots, \psi_n \) are all true with respect to small world \( \mathcal{W} \), then \( \phi \) is also true with respect to small world \( \mathcal{W} \). Notice that the small
world $\mathcal{W}$ need not be a model for any of $\psi_1, \psi_2, \ldots, \psi_n, \phi$ in order for \{ $\psi_1, \psi_2, \ldots, \psi_n$ \} $\vDash_{\mathcal{W}} \phi$ to be the case. In fact, if $\mathcal{W}$ is not a model for each of $\psi_1, \psi_2, \ldots, \psi_n$, then \{ $\psi_1, \psi_2, \ldots, \psi_n$ \} $\vDash_{\mathcal{W}} \phi$ is vacuously true. That is, it is true regardless of the relationship between $\mathcal{W}$ and $\phi$.

Logic is the normative study of reasoning. Logicians want to say what makes reasoning good or bad. Consequently, we have been trying to give an account of the evidential support relation. We said that if an argument preserves truth, then it is good. And we gave this kind of argumentative success a name: if the conclusion of an argument is guaranteed to be true when the premisses of the argument are all true, then the argument is valid. The same basic idea extends to first-order logic, but in first-order logic, truth is relative to a model. In first-order logic, an argument is said to be valid just in case every model for the premisses of the argument is also a model for the conclusion of the argument. In other words, an argument is valid just in case for every small world $\mathcal{M}$, if $\mathcal{M}$ is a model for the premisses, then $\mathcal{M}$ is a model for the conclusion. We will write \{ $\psi_1, \psi_2, \ldots, \psi_n$ \} $\vDash \phi$ in order to indicate that the argument having $\psi_1, \psi_2, \ldots, \psi_n$ as premisses and $\phi$ as conclusion is valid. If \{ \} $\vDash_{\mathcal{M}} \phi$ for every small world $\mathcal{M}$, then we write \{ \} $\vDash \phi$, and we say that $\phi$ is a logical truth.

Unlike with truth tables in zeroth-order logic, we do not have a mechanical procedure in first-order logic for determining whether an argument is valid or invalid. Instead, we will have to be clever in finding counter-examples to establish when arguments are invalid. Since our definition of validity is conditional, in order to establish that an argument is invalid, we need to produce a small world that is a model for the premisses but is not a model for the conclusion. If we can do that, then the argument is invalid.

Let’s see how counter-examples work in practice. First, consider the claim that
\{ (\exists x) Gx \} \models (\forall x) Gx. Let G be the predicate “… is green.” Then the argument is that since there is at least one thing that is green (premiss), everything is green (conclusion). The argument should look bad or at least funny enough to make you worried about its validity. In fact, the argument is not valid. In order to show that the argument is invalid, we need to construct a small world in which it is true that (\exists x) Gx but false that (\forall x) Gx. Using the translation of the predicate given above, we need a small world in which at least one thing is green and at least one thing is not green. Consider the small world consisting of the two constants a and b along with a single one-place predicate G. In our small world, we let G apply to the constant a but not to the constant b. Hence, with respect to our small world, Ga is true but Gb is false. Since Ga is true, it is true that there is at least one thing that has the property G. And so we have shown that the argument is invalid.

Second, consider the claim that \{ (\forall x) (Gx \rightarrow Hx), \neg (\exists x) Gx \} \models \neg (\exists x) Hx. Let G be the predicate “… is granite,” and let H be the predicate “… is heavy.” Then the argument says that since everything granite is heavy (premiss) and nothing is granite (premiss), nothing is heavy (conclusion). Remember that we are not concerned with whether the premisses are true or false. Rather we are concerned with whether the premisses would evidentially support the conclusion if the premisses were true. Again, the argument should look bad. In order to see that it definitely is bad, consider the small world consisting of a single constant, say d. Then suppose that the predicate H applies to d but the predicate G does not apply to d. In that case, both of the premisses are satisfied. Since we only have one constant, we only have to check that (Gd \rightarrow Hd) is true and that d does not manifest the property G. Both of those hold, since Hd is true and Gd is false. And yet, since Hd is true, the conclusion of the argument, namely that \neg (\exists x) Hx, is false. Contrary to the conclusion of the argument, there is at least one thing that has property H,
namely whatever is represented by d. What if we replace the conclusion with its negation? Is it true that \[ \{ (\forall x)(Gx \rightarrow Hx), \neg(\exists x)Gx \} \models (\exists x)Hx? \]
References


A Philosophical Introduction to Formal Logic

Chapter 4: Natural Deduction for First-Order Logic

In Chapter 3, we used small worlds to give a formal account of validity in first-order logic. An argument in first-order logic is valid if and only if in every small world in which the premisses are all true, the conclusion is true as well. Thus, Chapter 3 did the same sort of work for first-order logic that Chapter 1 did for zeroth-order logic. In Chapter 2, we developed an alternative account of evidential support in our formal language: proof. The idea in Chapter 2 was to develop a proof system and then say that an argument is good if its conclusion may be derived from its premisses using that proof system. In this chapter, we extend the proof system developed in Chapter 2. The goal is the same: we will say that an argument is good if its conclusion may be derived from its premisses using our proof system.

As in Chapter 2, our rules of inference come in pairs. In this chapter, we will see six new rules of inference: two for the universal quantifier, two for the existential quantifier, and two for the identity relation. But before we get to those rules, we need to discuss two operations: substitution and partial substitution.

4.1 Substitution and Partial Substitution

In this section, I describe two operations on well-formed formulas in our formal language: substitution and partial substitution. The formal machinery here can look a little intimidating, but the idea is very simple: we are replacing some terms in a formula with other terms.

Let $\phi$ be a well-formed formula in our language. Let $x$ be an arbitrary variable, and let $c$ be an arbitrary constant. The expression $\phi[x/c]$ represents the substitution operation applied to the formula $\phi$. Specifically, the expression $\phi[x/c]$ denotes the formula obtained from $\phi$ by
replacing all *free* occurrences of *x* in *φ* with the constant *c*. Below are some examples of the substitution operation:

<table>
<thead>
<tr>
<th><em>φ</em></th>
<th><em>φ</em>[x/c]</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Fa</em></td>
<td><em>Fa</em></td>
</tr>
<tr>
<td><em>Fx</em></td>
<td><em>Fc</em></td>
</tr>
<tr>
<td>(<em>Gx ∧ Hxy</em>)</td>
<td>(<em>Gc ∧ Hcy</em>)</td>
</tr>
<tr>
<td>(∀<em>x</em>)Rxx</td>
<td>(∀<em>x</em>)Rxx</td>
</tr>
<tr>
<td>((∀<em>x</em>)Rxy → Hzx)</td>
<td>((∀<em>x</em>)Rxy → Hzc)</td>
</tr>
</tbody>
</table>

In all of these cases, the formula on the right is obtained from the formula on the left by replacing free occurrences of the variable *x* in *φ* with the constant *c*. If the variable *x* does not occur free, then it is not replaced.

We also define a similar operation, which we will call *partial substitution*. Let *φ* be a well-formed formula in our language, and let *b* and *c* be arbitrary constants. The expression *φ*[b/c] represents the *partial substitution operation* applied to the formula *φ*. Specifically, the expression *φ*[b/c] denotes the formula obtained from *φ* by replacing some occurrences of the constant *b* in *φ* with the constant *c*. Unlike the substitution operation, the partial substitution operation does not always have a unique result. Below are some examples of the partial substitution operation:

<table>
<thead>
<tr>
<th><em>φ</em></th>
<th><em>φ</em>[b/c]</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Fb</em></td>
<td><em>Fc</em></td>
</tr>
<tr>
<td>(<em>a = b</em>)</td>
<td>(<em>a = c</em>)</td>
</tr>
<tr>
<td><em>Rbb</em></td>
<td><em>Rbc, Rcb, Rcc</em></td>
</tr>
<tr>
<td>(∀<em>x</em>)Rxx</td>
<td>(∀<em>x</em>)Rxx</td>
</tr>
<tr>
<td>((∃<em>x</em>)Rx → <em>(a = b</em>))</td>
<td>(((∃<em>x</em>)Rx → *(a = c))),</td>
</tr>
<tr>
<td></td>
<td>(((∃<em>x</em>)Rc → *(a = b))),</td>
</tr>
<tr>
<td></td>
<td>(((∃<em>x</em>)Rx → *(a = c)))</td>
</tr>
</tbody>
</table>
For the third and fifth entries on the left-hand side, the result of applying the partial substitution operation is ambiguous. In either case, any of the three formulas on the right would represent an appropriate application of partial substitution. The formulas on the right are obtained from the formulas on the left by replacing some occurrences of the constant \( b \) in \( \phi \) with the constant \( c \).

### 4.2 Universal Introduction and Elimination

Our first new rule is universal elimination. Universal elimination takes us from a formula that begins with a universal quantifier expression to a formula without that specific universal quantifier expression.

**Universal Elimination (\( \forall E \))**

Suppose that \((\forall x)\phi\) is a well-formed formula in our formal language. Suppose that \( x \)—the same \( x \) that appears in \((\forall x)\phi\)—is an arbitrary variable, and suppose that \( c \) is an arbitrary constant. If \((\forall x)\phi\) appears on some line in a proof, then the formula \( \phi[x/c] \) may be written on any later line. Schematically:

\[
\frac{(\forall x)\phi}{\phi[x/c]}
\]

When we use the universal elimination rule, we write down in the assumption column for \( \phi[x/c] \) all of the numbers that appear in the assumption column for \((\forall x)\phi\).

**Example 4.1:** \{ (\(\forall x\)Fx) \( \vdash \) (Fa \& (Fb \lor Fc)) \}

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>((\forall x)Fx)</th>
<th>(2)</th>
<th>Fa</th>
<th>(3)</th>
<th>Fc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(4)</td>
<td>(~Fb ( \rightarrow ) Fc)</td>
<td>(5)</td>
<td>(Fb \lor Fc)</td>
<td>(6)</td>
<td>(Fa &amp; (Fb \lor Fc))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A (premiss)</td>
<td></td>
<td>1 ( \forall E )</td>
<td></td>
<td>1 ( \forall E )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 ( \rightarrow I )</td>
<td></td>
<td>4 ( \lor I )</td>
<td></td>
<td>2,5 ( \land I )</td>
</tr>
</tbody>
</table>
Example 4.2: \{ (\forall x)(\forall y)(\forall z)((Rxy \land Ryz) \rightarrow Rxz), Rab \} \vdash (Rbc \rightarrow Rac)

1 (1) (\forall x)(\forall y)(\forall z)((Rxy \land Ryz) \rightarrow Rxz) A (premiss)
2 (2) Rab A (premiss)
3 (3) Rbc A (for CP)
1 (4) (\forall y)(\forall z)((Ray \land Ryz) \rightarrow Raz) 1 \forall E
1 (5) (\forall z)((Rab \land Rbz) \rightarrow Raz) 4 \forall E
1 (6) ((Rab \land Rbc) \rightarrow Rac) 5 \forall E
2,3 (7) (Rab \land Rbc) 2,3 \land I
1,2,3 (8) Rac 6,7 \rightarrow E
1,2 (9) (Rbc \rightarrow Rac) 3,8 \rightarrow E

Example 4.3: \{ (\forall x)(Gx \lor Hx), (\forall x)(Gx \rightarrow Bx), (\forall x)(Hx \rightarrow Bx) \} \vdash Ba

1 (1) (\forall x)(Gx \lor Hx) A (premiss)
2 (2) (\forall x)(Gx \rightarrow Bx) A (premiss)
3 (3) (\forall x)(Hx \rightarrow Bx) A (premiss)
1 (4) (Ga \lor Ha) 1 \forall E
2 (5) (Ga \rightarrow Ba) 2 \forall E
3 (6) (Ha \rightarrow Ba) 3 \forall E
1,2,3 (7) Ba 4,5,6 \lor E

Universal elimination has no constraints on its application: if we have (\forall x)(\phi), then we may write \phi[x/c] for any constant c on any later line. No special conditions need to be satisfied in order to apply the rule of universal elimination. By contrast, our second new rule, universal introduction, does have some constraints on its application.

**Universal Introduction (\forall I)**

Suppose that \phi is a well-formed formula in our formal language. Let x be an arbitrary variable, and let c be any constant that does not appear in the formula \phi. If the formula \phi[x/c] appears on some line in a proof, then (\forall x)\phi may be written on any later line provided the formula \phi[x/c] does not depend on any formula containing the constant c.

The first constraint requires that we replace all of the tokens of whatever constant we are replacing with a variable. The second constraint requires that the constant we replace
when we add the quantifier is essentially arbitrary since it does not depend on any
previous occurrences of that specific constant. Schematically, we have:

$$\phi[x/c]$$

$$\frac{}{(\forall x)\phi}$$

When we use the universal introduction rule, we have to write down in the assumption
column for $$(\forall x)\phi$$ all of the numbers in the assumption column for $$\phi[x/c]$$.

Example 4.4: \{ $$(\forall x)\text{Fxx}$$ \} $\vdash (\forall y)\text{Fyy}$$

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>(\forall x)\text{Fxx}</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>1</td>
<td>(2)</td>
<td>\text{Faa}</td>
<td>1 \forall E</td>
</tr>
<tr>
<td>1</td>
<td>(3)</td>
<td>(\forall y)\text{Fyy}</td>
<td>2 \forall I</td>
</tr>
</tbody>
</table>

Example 4.4 shows that one’s choice of variable in the universal quantifier expression does not
matter. Example 4.4 allows us to illustrate the importance of the constraint that we replace every
token of whatever constant we are replacing with a variable. To see the importance of the
constraint, consider a case in which the constraint is violated. Suppose we could (and did)
replace only one of the occurrences of a in Faa with a variable bound by the quantifier
expression $$\forall y$$. For concreteness, suppose we replaced only the first occurrence. Then, we
would have concluded $$\forall y\text{Fya}$$. But that conclusion is not supported evidentially by the premiss.
The sentence $$\forall y\text{Fyy}$$ says that everything is F-related to itself. But the sentence $$\forall y\text{Fya}$$ says
that everything is F-related to a. To see why this kind of inference should be barred, consider an
example that ought to satisfy the premiss, “Everything is exactly as fast as itself.” Now, consider
the sentence, “Everything is exactly as fast as Usain Bolt.” If true, then no one is faster or slower
than anyone else, which is not only incorrect, it is clearly distinct from the claim that everything
is exactly as fast as itself.
Example 4.5: \{ (\forall x)(\forall y)(Gxy \rightarrow Hx), (\forall x)Gxx \} \vdash (\forall x)Hx

1 \hfill (1) \hfill (\forall x)(\forall y)(Gxy \rightarrow Hx) \hfill A \text{ (premiss)}
2 \hfill (2) \hfill (\forall x)Gxx \hfill A \text{ (premiss)}
1 \hfill (3) \hfill (\forall y)(Gay \rightarrow Ha) \hfill 1 \ \forall E
1 \hfill (4) \hfill (Gaa \rightarrow Ha) \hfill 3 \ \forall E
2 \hfill (5) \hfill Gaa \hfill 2 \ \forall E
1,2 \hfill (6) \hfill Ha \hfill 4,5 \ \rightarrow E
1,2 \hfill (7) \hfill (\forall x)Hx \hfill 6 \ \forall I

In example 4.5, the constraints on universal introduction are satisfied, since the resulting formula, \((\forall x)Hx\), contains no constants, and no constants appear in either line (1) or line (2), which are the ultimate justifications for the conclusion in line (7).

4.3 Existential Introduction and Elimination

Our next pair of rules will let us work with the existential quantifier. In a sense, the rules for existential quantification are flipped compared with the rules for universal quantification. Whereas universal elimination has no constraints, existential elimination does; and whereas universal introduction has constraints, existential introduction does not. Hence, we begin with existential introduction as the simpler case.

Existential Introduction (\(\exists I\))

Suppose that \((\exists x)\phi\) is a well-formed formula, where x is an arbitrary variable. Further suppose that c is an arbitrary constant. If the formula \(\phi[x/c]\) appears on some line in a proof, then the formula \((\exists x)\phi\) may be written on any later line. Schematically:

\[
\frac{\phi[x/c]}{(\exists x)\phi}
\]

When we use the existential introduction rule, we write down in the assumption column for \((\exists x)\phi\) the numbers that appear in the assumption column for \(\phi[x/c]\). The basic
idea here is that whenever some formula containing a constant appears in a proof, we
know that there exists something satisfying the formula: namely, the constant that
appears in it. Hence, whenever a formula containing some constants appears in a proof,
we may replace as many of the tokens of a single type of constant—for example,
occurrences of the constant c—with a single variable that is then bound by an existential
quantifier.

Example 4.6: \{ Fa \} \vdash (\exists x)Fx

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>Fa</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>1</td>
<td>(2)</td>
<td>(\exists x)Fx</td>
<td>1 \exists</td>
</tr>
</tbody>
</table>

Example 4.7: \{ Gaa \} \vdash (\exists x)(\exists y)Gxy

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>Gaa</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>1</td>
<td>(2)</td>
<td>(\exists y)Gay</td>
<td>1 \exists</td>
</tr>
<tr>
<td>1</td>
<td>(3)</td>
<td>(\exists x)(\exists y)Gxy</td>
<td>2 \exists</td>
</tr>
</tbody>
</table>

Notice that in Example 4.7 we could have directly proved that (\exists x)Gxx from the premiss that Gaa using existential introduction. However, unlike with universal introduction, we are not
required to bind up all of the tokens of whatever type of constant we are replacing. We may
replace as many or as few occurrences of the relevant type of constant as we like. The following
example illustrates this feature of existential introduction once more:

Example 4.8: \{ Laaaaa \} \vdash (\exists x)(\exists y)(\exists z)Laxxyz

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>Laaaaa</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>1</td>
<td>(2)</td>
<td>(\exists z)Laaaaz</td>
<td>1 \exists</td>
</tr>
<tr>
<td>1</td>
<td>(3)</td>
<td>(\exists y)(\exists z)Laaayz</td>
<td>2 \exists</td>
</tr>
<tr>
<td>1</td>
<td>(4)</td>
<td>(\exists x)(\exists y)(\exists z)Laxxyz</td>
<td>3 \exists</td>
</tr>
</tbody>
</table>
Again, existential introduction has no special constraints on its application. We may replace
some or all of the constants in a formula with a variable that we then bind up with an existential
quantifier. By contrast, existential elimination is similar to universal introduction in that it has
two constraints on its application.

**Existential Elimination (∃E)**

Suppose that \( \phi \) and \( \psi \) are well-formed formulas in our formal language and that the
constant \( c \) does not appear in either formula. If the formulas \( (\exists x) \phi \) and \( (\phi[x/c] \rightarrow \psi) \)
appear on separate lines in a proof, then \( \psi \) may be written on any later line in the proof,
provided the formula \( (\phi[x/c] \rightarrow \psi) \) does not depend for its ultimate justification on any
formula in which the constant \( c \) appears. Schematically:

\[
\frac{(\exists x) \phi}{(\phi[x/c] \rightarrow \psi)} \psi
\]

When we use the existential elimination rule, we have to write down in the assumption
column all of the numbers that appear in the assumption columns for the formulas
\((\exists x) \phi \) and \( (\phi[x/c] \rightarrow \psi) \).

*Example 4.9*: \{ (\exists x)(Fx \land Gx) \} ⊢ (\exists x)Fx \land (\exists x)Gx

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>(\exists x)(Fx \land Gx)</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>(Fa \land Ga)</td>
</tr>
<tr>
<td>2</td>
<td>(3)</td>
<td>Fa</td>
</tr>
<tr>
<td>2</td>
<td>(4)</td>
<td>Ga</td>
</tr>
<tr>
<td>2</td>
<td>(5)</td>
<td>(\exists x)Fx</td>
</tr>
<tr>
<td>2</td>
<td>(6)</td>
<td>(\exists x)Gx</td>
</tr>
<tr>
<td>2</td>
<td>(7)</td>
<td>((\exists x)Fx \land (\exists x)Gx)</td>
</tr>
<tr>
<td>(8)</td>
<td>(2) \rightarrow (7)</td>
<td>2,7 \ CP</td>
</tr>
<tr>
<td>1</td>
<td>(9)</td>
<td>((\exists x)Fx \land (\exists x)Gx)</td>
</tr>
</tbody>
</table>
In Example 4.9, \( \phi = (F \land G) \), \( \phi[x/a] = (Fa \land Ga) \), and \( \psi = ((\exists x)Fx \land (\exists x)Gx) \). Notice that we used the constant \( a \) in the proof, so we have \( \phi[x/a] \) rather than \( \phi[x/c] \). The constraints on existential elimination are satisfied, since no constants appear in \( \phi \) or \( \psi \), and the conditional \( (\phi[x/c] \rightarrow \psi) \), which appears in line (8) of the proof, does not ultimately depend on anything (so, in particular, it does not depend on any formula in which the constant \( a \) occurs).

**Example 4.10**: \( \{ (\exists x)(\forall y)Fxy \} \vdash (\forall y)(\exists x)Fxy \)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>(\exists x)(\forall y)Fxy</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>(\forall y)Fay</td>
</tr>
<tr>
<td>2</td>
<td>(3)</td>
<td>Fab</td>
</tr>
<tr>
<td>2</td>
<td>(4)</td>
<td>(\exists x)Fxb</td>
</tr>
<tr>
<td>2</td>
<td>(5)</td>
<td>(\forall y)(\exists x)Fxy</td>
</tr>
<tr>
<td></td>
<td>(6)</td>
<td>(\forall y)Fay \rightarrow (\forall y)(\exists x)Fxy</td>
</tr>
<tr>
<td>1</td>
<td>(7)</td>
<td>(\forall y)(\exists x)Fxy</td>
</tr>
</tbody>
</table>

In Example 4.10, \( \phi = (\forall y)Fxy \), \( \phi[x/a] = (\forall y)Fay \), and \( \psi = (\forall y)(\exists x)Fxy \). Again, we used the constant \( a \) in the proof, so we have \( \phi[x/a] \) rather than \( \phi[x/c] \). The constraints on existential elimination are satisfied, since no constants appear in \( \phi \) or \( \psi \), and the conditional \( (\phi[x/c] \rightarrow \psi) \), which appears in line (6) of the proof, does not ultimately depend on anything (so, in particular, it does not depend on any formula in which the constant \( a \) occurs).

In examples 4.9 and 4.10, we see a general strategy for using the rule of existential elimination. Whenever you have an existentially quantified formula, assume a formula that instantiates the existential for some constant that has not yet appeared in the proof. Then use that assumption (possibly along with other assumptions) to derive the conclusion (or some important step toward the conclusion). Discharge the assumption using the rule of conditional proof and then apply existential elimination. We now have rules for working with universal and existential quantifiers. In the next section, I describe rules for working with the identity relation.
4.4 Identity Introduction and Elimination

In this section, I describe inference rules for the identity relation. Our first rule in this section, identity introduction, looks a little funny when compared to our other rules of inference in that it allows us to write something down on the very first line of a proof. Specifically, identity introduction lets us write down the formula saying that everything is identical to itself.

**Identity Introduction (≡I)**

On any line of a proof, the formula \((∀x)(x = x)\) may be written without writing anything in the assumption column. Schematically:

\[
(∀x)(x = x)
\]

Again, when we use the identity introduction rule, we do not need to write anything in the assumption column.

*Example 4.11*: \{ \} \vdash (P ∨ (a = a))

1. \((∀x)(x = x)\) \hspace{1cm} ≡I
2. \((a = a)\) \hspace{1cm} 1 ∀E
3. \((¬P → (a = a))\) \hspace{1cm} 2 →I
4. \((P ∨ (a = a))\) \hspace{1cm} 3 ∨I

*Example 4.12*: \{ (a ≠ a) \} ⊢ P

1. (a ≠ a) \hspace{1cm} \text{A (premiss)}
2. \((∀x)(x = x)\) \hspace{1cm} ≡I
3. \((a = a)\) \hspace{1cm} 2 ∀E
1. (4) \((¬P → (a ≠ a))\) \hspace{1cm} 1 →I
1. (5) \((¬P → (a = a))\) \hspace{1cm} 1 →I
1. (6) \(¬¬P\) \hspace{1cm} 4,5 ¬I
1. (7) \(P\) \hspace{1cm} 6 ¬E
In Example 4.12, we make use of the fact that the formula \((a \neq a)\) is just a way of writing the formula \(\sim(a = a)\). Whenever we have a formula of the form \((a \neq a)\), we can make use of our usual reductio strategy.

Identity introduction lets us write down the formula \((\forall x)(x = x)\) without writing anything in the assumption column. We have a special name for formulas that may be written down without incurring any assumption debts: these formulas are *axioms*. Hence, the rule of identity introduction amounts to the statement that the formula \((\forall x)(x = x)\) is an axiom.

Our next rule lets us replace occurrences of a constant in a formula with any constant that is related to the first by the identity relation. For example, if we have a formula like \(Ha\), and we know (or assume) that \((a = b)\), then we can replace the \(a\) in \(Ha\) with \(b\) to obtain \(Hb\).

**Identity Elimination \((=E)\)**

Suppose that \(\phi\) is a well-formed formula in our formal language. If the formula \(\phi\) appears on some line in a proof and the identity of \(b\) and \(c\), expressed either as \((b = c)\) or as \((c = b)\), appears on some other line in a proof, then \(\phi[[b/c]]\) may be written on any later line in the proof. Schematically:

\[
\frac{\phi}{\phi[[b/c]]} \quad \text{or} \quad \frac{\phi}{\phi[[b/c]]}
\]

\((b = c)\) or \((c = b)\), respectively. When we use the identity elimination rule, we have to write down in the assumption column for \(\phi[[b/c]]\) all of the numbers in the assumption columns for \(\phi\) and for \((b = c)\) or \((c = b)\), depending on form.

Just to be completely clear, the identity elimination rule makes use of the *partial* substitution operation, \([[b/c]]\), not the ordinary substitution operation \([x/c]\). When using identity elimination, we are free to replace as many or as few instances of a constant as we like.
Example 4.13: \{ H_a, (\forall x)(x = b) \} \vdash H_b

1  (1)  H_a  \hspace{1cm} \text{A (premiss)}
2  (2)  (\forall x)(x = b)  \hspace{1cm} \text{A (premiss)}
2  (3)  (a = b)  \hspace{1cm} \text{2 \ \exists E}
1,2  (4)  H_b  \hspace{1cm} \text{1,3 \ = E}

Example 4.14: \{ H_a, \neg H_b \} \vdash (a \neq b)

1  (1)  H_a  \hspace{1cm} \text{A (premiss)}
2  (2)  \neg H_b  \hspace{1cm} \text{A (premiss)}
3  (3)  (a = b)  \hspace{1cm} \text{A* (for reductio)}
1,3  (4)  H_b  \hspace{1cm} \text{1,3 \ = E}
1  (5)  ((a = b) \rightarrow H_b)  \hspace{1cm} \text{3,4 \ CP}
2  (6)  ((a = b) \rightarrow \neg H_b)  \hspace{1cm} \text{2 \rightarrow I}
1,2  (7)  (a \neq b)  \hspace{1cm} \text{5,6 \ \neg I}

And that gives us all of the rules that we need in order to prove all the truths in first-order logic with identity!
In previous chapters we have developed a formal language adequate for translating many ordinary English sentences, and we have developed techniques, including a formal proof theory, that let us establish the validity or invalidity of arguments in our formal language. Up to now, we have been paying attention to deductive arguments, which have conclusions that are guaranteed to be true if their premisses are all true. The conclusion of a deductive argument is, in some sense, already asserted in the premisses. In later chapters, we will be paying attention to ampliative arguments, which have conclusions that are not guaranteed to be true even if their premisses are all true. The conclusion of an ampliative argument goes beyond what is asserted in the premisses and takes on some additional risk.

In this book, we are going to use probability theory to think about ampliative inferences. But before we can talk sensibly about probability theory, we need an important piece of machinery: sets. Hence, this chapter is essentially an interlude between the part of the book on deductive logic and the part of the book on ampliative logic. However, despite the fact that set theory is only serving as a bridge for us, it is (in fact) a very rich and important topic both for mathematicians and for philosophers. For mathematicians, sets have long been a sort of lingua franca for its diverse sub-fields. For philosophers, sets are both an important object of investigation in themselves (especially in the philosophy of mathematics) and a formal tool that has been used to give accounts of properties and relations, possible worlds, and propositions, just to name a few important topics.
5.1 Object Language and Meta-Language

In previous chapters, we introduced a formal language, which contains strings of symbols like, $(P \rightarrow Q)$, $(\sim S \land P)$, and $(\exists x)(\forall y)Ryx$. We use the machinery of our formal language to represent sentences in some natural language, like English, in order to understand how the represented sentences are evidentially related to one another. Sometimes, we make use of a formal language. For example, we might use a formal language in order to assure ourselves that an inference we are making in natural language is a good one. But sometimes, we want to talk about our logical language. For example, we might want to know whether the conclusion of a valid argument can always be proved from its premisses. When we are talking about a language, the language we are talking about is called the object language. The language we use to talk about an object language is called a meta-language. Our meta-language is predominantly English, but there are a few special symbols that we have used that are part of our meta-language. For example, when we write down a rule of inference, like:

**Conjunction Introduction (\(\wedge I\))**

Suppose that \(\phi\) and \(\psi\) are well-formed formulas in our formal language. If \(\phi\) and \(\psi\) appear on separate lines in a proof, then the formula \((\phi \land \psi)\) may be written on any later line. Schematically:

\[
\begin{array}{c}
\phi \\
\psi
\end{array}
\]

\((\phi \land \psi)\)

We are saying something in our meta-language about how various expressions in our formal language, taken as an object of conversation, may be transformed into other expressions in our formal language. The English words, the Greek letters, and the horizontal line are part of our meta-language, not part of our object language. Our formal language doesn’t contain any Greek
letters, but the Greek letters allow us to talk about formulas that are contained by the formal language.

When we write \{ (R \land S), (S \rightarrow Q) \} \vdash Q, we are making a claim in our meta-language about our formal language as an object. Specifically, we are saying that within the formal system we are considering, the formula Q may be proved from the formulas appearing in the collection \{ (R \land S), (S \rightarrow Q) \}, which from now on, we will call a set. Our formal language does not contain symbols like {, }, or \vdash. And so far, our formal language does not contain any sets. In Section 5.6, we will show how one could introduce sets and notation for working with sets into our formal language. But for now, we are going to think about sets as belonging to the meta-language. Having a clear distinction between object language and meta-language will be helpful when we start thinking about probability in the next chapter.

5.2 Sets and Boxes: An Informal Introduction to Sets

Naïvely, one might say that a set is a collection or grouping of things. The things collected together into a set are called the elements or members of the set. A set might be a collection of almost anything: a set of hats, a set of people, a set of dishes, or a set of numbers. Whenever something is an element of a set, we say that the thing is in or belongs to that set. A set might be very small, having no members or having only a few members, or a set might be very large, having an infinite number of members. A set might have other sets as members, and (in some set theories) a set might even have a copy of itself as a member.

However, sets are not ordinary, common-sense collections of objects.\(^1\) In the first place, a set is distinct from its elements. The set, or collection, of all of Gene Hackman’s hats is different

\(^1\) For more discussion of how sets differ from ordinary, common-sense collections, with special attention to plural reference, see Chapter Three of Stephen Pollard’s *Philosophical Introduction to Set Theory*. 

95
from (the plurality of) those hats. In this way, a set is analogous to a container. A box with several objects in it is different from the objects themselves. In the second place, a set might be empty or it might contain only a single element, even though we ordinarily think of collections as including several different things. My stamp collection—regarded as a set—is empty. But normally, I would just say that I do not have a stamp collection. The set that has no members at all is special enough to get its own name. The set that contains nothing is called the empty set or the null set. Any set that contains exactly one element is called a singleton set.

A third way in which sets are different from ordinary, common-sense collections is that set membership is not transitive. If the United States were regarded as a set containing the fifty states (plus miscellaneous territories, like Guam and the District of Columbia) and if each state were regarded as a set containing the counties in that state, then it would be true to say that Champaign County is in Illinois, and it would be true to say that Illinois is in the United States. But it would be false to say that Champaign County is in the United States! Thinking about sets as boxes does not seem to help at first. Suppose I have a box of eight crayons of different colors, including a green crayon. I decide to put my box of crayons into a larger box with some other art supplies. Ordinarily, I might say that there is a green crayon in the box of art supplies. But if the boxes are sets, then that would not be correct. The green crayon is in the box of crayons, but it is not in the box of art supplies. One way to preserve the analogy with boxes is to treat “in” as equivalent to “immediately inside.” That way, when you say that something is in a given box, you mean that one can find the indicated thing without opening any additional boxes.

A fourth way in which sets are different from ordinary, common-sense collections is that sets do not have duplicate members. Ordinarily, we would treat a box containing one green crayon as different from a box containing three green crayons, but if the boxes are thought of as
sets of colors, then the two are identical. The elements of a set do not take up space in a set in the same way that ordinary objects take up space in a box. At this point, the analogy with boxes becomes very strained. In order to preserve the analogy, we have to imagine that we know nothing about the spatial arrangement of the things in the box. We cannot even know what places or relative positions the various objects occupy.

5.3 Set Notation and Set Definition
A set is a mathematical object. We use curly braces to denote sets. For example, \{1, 2, 3\} is the set with elements 1, 2, and 3. Sets are unordered. The arrangement of the elements in a set is not important. Hence, the set \{1, 2, 3\} is identical to the set \{2, 3, 1\}. As mentioned earlier, sets do not have duplicate members. Hence, the set \{1, 2, 3\} is identical to the set \{1, 1, 2, 3\}, it is identical to the set \{1, 2, 2, 3, 3, 3, 1\}, and so on. These properties are not exclusive to sets of numbers. For example, consider the set of trucks \{pickup truck, dump truck, garbage truck, grain truck, semi truck\}. That set of trucks is identical to the set \{dump truck, garbage truck, pickup truck, semi truck, grain truck\}, and it is also identical to the set \{pickup truck, pickup truck, dump truck, semi truck, garbage truck, grain truck, garbage truck\}.

A set is not merely the items that are its elements. And (in our set theory) a thing is not the same as a set containing that thing. For example, suppose Billy, Suzy, and Tom go out for drinks one night. Billy, Suzy, and Tom are a group, but the three of them are not the same as the set whose elements are Billy, Suzy, and Tom. Moreover, Suzy is not the same as the singleton set whose only element is Suzy. In general, \(a \neq \{a\}\). Our analogy with boxes is helpful for understanding the point here: a box containing a green crayon is not the same thing as a green crayon.
One might try to define a set in any of three different ways. First, we might define a set by *listing* all of its elements. The following define sets by listing the elements of the set:

\{\text{plastic, paper}\}
\{\text{Harry, Joe, yellow, justice, bologna, ketchup}\}
\{\text{mailbox, 2, Earth}\}
\{\{\}, \{1\}, \{2\}, \{1, 2\}\}

Second, we may describe a set by predication. We write \{x \mid Fx\}, where F is a predicate, to denote the set of all things that satisfy the predicate F. For example, \{x \mid x \text{ is a fireman}\} denotes the set of all firemen. We will use bold-faced capital letters and some special symbols to represent sets. For example, we might use \textbf{F} to denote the set \{x \mid x \text{ is a fireman}\} or we might use \mathbb{Z}^+ to denote the set \{x \mid x \text{ is an integer greater than zero}\} picks out the set of positive integers.

In either case, we can use the identity relation to attach a plain name to a set described in list or predicate notation, like \textbf{F} = \{x \mid x \text{ is a fireman}\}. In addition to using bold-faced, capital letters to denote sets, we will sometimes use special symbols to denote specific sets. For example, we will use \emptyset to denote the empty set, and we will use \mathbb{Z} to denote the set of all the integers, i.e. the set of positive and negative counting numbers plus zero: \{…, -3, -2, -1, 0, 1, 2, 3, …\}. From time to time, we will also use the bold-faced capital letter \textbf{U} to denote the universe of discourse.

However, when we define sets by predication, we have to be very careful. Paradox lurks here! Some early set theorists thought that predication was a perfectly safe way of defining sets. After all, using predication to describe the members of a set is a very natural way to proceed. Set theories that allow you to define arbitrary sets by predication are sometimes called *naïve set theory* because they uncritically make this intuitively correct but paradox-inducing assumption.

---

2 The universe of discourse is not a context-invariant set, like the null set. That is, for us, \textbf{U} does not pick out the same set every time we write it down. Rather, the set denoted by \textbf{U} will change from one context to another. In some developments of set theory, like in Quine’s New Foundations, there is a unique, determinate universal set: a set of literally everything. But in our approach to set theory, the universe of discourse is *not* the set of literally everything. Given the way we talk about sets, a genuine universal set would lead to paradoxes.
The classical example of a paradox of naïve set theory comes from a letter that Bertrand Russell famously wrote to the logician Gottlob Frege. Use predication to define the set $S$ below:

$$S = \{ x \mid x \text{ is a set that is not a member of itself} \}$$

The set $S$ is the set of all sets that are not members of themselves. The set of cats is in the set $S$, since the set of cats is not a cat (and hence, not a member of the set of cats). But the set of all sets having more than three elements is not in the set $S$, since the set of all sets having more than three elements has more than three elements and so must be a member of itself. Now, ask yourself, “Is $S$ an element of the set $S$?” Really stop and think about this before reading on.

To see what is wrong with the set $S$, first suppose that $S$ is a member of itself. Then by its definition, $S$ is not a member of itself. So, we cannot consistently suppose that $S$ is a member of itself. Okay then, try the other possibility: suppose that $S$ is not a member of itself. In that case, the set $S$ satisfies the predicate “… is a set that is not a member of itself.” So, $S$ must be a member of $S$! So we cannot consistently assume that $S$ is not a member of itself either! But $S$ must either be a member of itself or not. And so, $S$ is paradoxical. There must be something wrong with the way we have defined the set $S$.

One way to avoid the paradoxes of naïve set theory is to define sets recursively. Begin with a list of allowed sets. Then provide a rule that constructs new sets using already constructed sets. For example, begin with the empty set and construct an infinite sequence of sets as follows. Let $T_0 = \{ \}$. For each $T_i$, where $i$ is a natural number greater than zero, let $T_i = \{ T_{i-1} \}$. The sequence just defined looks like this:

$$
\begin{align*}
T_0 & = \{ \} \\
T_1 & = \{ T_0 \} = \{ \{ \} \} \\
T_2 & = \{ T_1 \} = \{ \{ T_0 \} \} = \{ \{ \{ \} \} \} \\
T_3 & = \{ T_2 \} = \{ \{ T_1 \} \} = \{ \{ \{ T_0 \} \} \} = \{ \{ \{ \{ \} \} \} \}
\end{align*}
$$
Since the null set is a safe set and the rule for set construction outputs safe sets given safe sets as inputs, all of the T-sets are safe sets as well.

5.4 Some Special Relations and Operations for Sets

So, a set is a mathematical object having elements or members. The elements of a set stand in the set membership relation with respect to the set that contains them, and we introduce a special relation symbol, $\in$, to denote that relation. The set membership relation is a two-place relation: it has two blanks to be filled in. If $b$ is an element of a set $B$, we will write $b \in B$. And we will write $b \notin B$ in order to denote that $b$ is not an element of $B$, in the same way that we write $a \neq b$ to denote that $a$ is not identical to $b$. For any object you pick and any set you pick, the object is either an element of the set or not. If, for example, the object named by $b$ is a member of the set named by $B$, then the sentence $b \in B$ is true; otherwise, the sentence is false. And if the object named by the constant $b$ is not a member of the set named by $B$, then the sentence $b \notin B$ is true; otherwise, the sentence is false. The set membership relation is a primitive in our meta-language. We will not reduce it to anything or analyze it in terms of anything.

Using the set membership relation, we may now describe three useful operations on sets: intersection, union, and set difference. Each operation is a function that takes two sets as inputs and returns one set as output. In general, if $\alpha$ and $\beta$ are arbitrary sets, then the following are sets as well:

$$\alpha \cap \beta \quad \text{called the intersection of } \alpha \text{ and } \beta$$

$$\alpha \cup \beta \quad \text{called the union of } \alpha \text{ and } \beta$$

$$\alpha - \beta \quad \text{called the difference between } \alpha \text{ and } \beta$$
The specific set that results from the application of a set operation is defined in terms of the set membership relation. Let \( \alpha \) and \( \beta \) be arbitrary sets and let \( \delta \) be an arbitrary object. The object \( \delta \) itself might or might not be a set. We may now define the set operations of intersection, union, and set difference as follows:

\[
\delta \in (\alpha \cap \beta) =_{df} ((\delta \in \alpha) \land (\delta \in \beta))
\]

\[
\delta \in (\alpha \cup \beta) =_{df} ((\delta \in \alpha) \lor (\delta \in \beta))
\]

\[
\delta \in (\alpha - \beta) =_{df} ((\delta \in \alpha) \land (\delta \notin \beta))
\]

According to the definitions above, something is an element of the intersection of two sets just in case it is an element of both of the intersecting sets. Something is an element of the union of two sets just in case it is either an element of one or an element of the other (or both). And something is an element of the set difference between two sets just in case it is an element of the first and \textit{not} an element of the second.

Consider the sets \( A = \{\text{plastic, paper, canvass}\} \) and \( B = \{\text{plastic, justice, mailbox, 2}\} \). The intersection of \( A \) and \( B \) is the set \( A \cap B = \{\text{plastic}\} \), since plastic is the only thing that is both an element of the set \( A \) and an element of the set \( B \). The union of the set \( A \) and the set \( B \) is the set \( A \cup B = \{\text{plastic, paper, canvass, justice, mailbox, 2}\} \), since each of those items is either an element of the set \( A \) or an element of the set \( B \) (or both). And the difference between the set \( A \) and the set \( B \) is the set \( A - B = \{\text{paper, canvass}\} \), since paper and canvass are in the set \( A \) but not in the set \( B \). Notice that the set \( B \cap A \) is the same as the set \( A \cap B \), since the zeroth-order operator \( \land \) is commutative, and similarly, the set \( B \cup A \) is the same as the set \( A \cup B \), since the operator \( \lor \) is commutative. However, the set difference operator is \textit{not} commutative. Generally, the set \( B - A \) is different than the set \( A - B \).
We can draw some pictures, called *Venn diagrams*, to help us think about the basic set operations. In a Venn diagram, named for the nineteenth century logician John Venn, we represent the universe of discourse as a rectangle, and we represent particular sets with circles or ovals drawn inside that rectangle, as in Figure 5.1:

![Figure 5.1: Venn Diagram for One Set](image)

If we have more than one set to work with, we begin by drawing our circles so that there is a region of the diagram for every possible place that an element of the universe of discourse might appear. For example, with the two sets $A$ and $B$, we would draw the diagram in Figure 5.2:

![Figure 5.2: Venn Diagram for Two Sets](image)

In Figure 5.1, there are two distinct regions: one inside the set $A$ and one outside the set $A$. In Figure 5.2, there are four distinct regions: one outside both the set $A$ and the set $B$, one inside the
set \textbf{A} but outside the set \textbf{B}, one inside the set \textbf{B} but outside the set \textbf{A}, and one inside both the set \textbf{A} and the set \textbf{B}. Those four regions are labeled in Figure 5.3:

![Venn Diagram for Two Sets with Distinct Regions Numbered](image)

**Figure 5.3:** Venn Diagram for Two Sets with Distinct Regions Numbered

Venn represented the fact that an object was located in a specific region of a diagram by shading or blackening out all the regions of the diagram where the object was not located. For example, if we wanted to represent that everything in the universe of discourse is in the set \textbf{A}, we might draw the diagram in Figure 5.4:

![Venn Diagram Representing that Everything is in the Set A](image)

**Figure 5.4:** Venn Diagram Representing that Everything is in the Set \textbf{A}

With the interpretation of the diagrams firmly in mind, let’s see how the set operations look diagrammatically. Suppose that we want to say that some object is located in the intersection of the sets \textbf{A} and \textbf{B}. Then, the Venn diagram would look like the one in Figure 5.5:
The intersection of the sets $A$ and $B$ is the overlap, what we labeled as Region 4 in Figure 5.3. By contrast, the union of the sets $A$ and $B$ is all of the area covered by either one, as seen in Figure 5.6:

In the diagram, the union of the sets $A$ and $B$ is represented by all of the area covered by the circle for $A$ and all of the area covered by the circle for $B$, including the overlap of the two. Only the region outside of both—in what we called Region 1 in Figure 5.3—is shaded out. Finally, we can represent the set difference between $A$ and $B$ as follows in Figure 5.7:
The diagram for the set difference between \( B \) and \( A \) would be a mirror image of the diagram in Figure 5.7.

Suppose that every element of the set \( A \) is also an element of the set \( B \). Then we say that \( A \) is a subset of \( B \), which we denote by writing \( A \subseteq B \). For example, if \( A \) is the set \( \{1, 2, 3\} \) and \( B \) is the set \( \{1, 2, 3, 4\} \), then \( A \) is a subset of \( B \). We have lots of quite ordinary examples of the subset relation. Everything that weighs at least five pounds also weighs at least three pounds. Everything that has been colored green all over has been colored some color all over. Every cat is a mammal. And so on. In each of these cases, if we think of the groups as sets, then the subset relation holds between the groups. The set of things that weigh at least five pounds is a subset of the set of things that weight at least three pounds. The set of green things is a subset of the set of things with color. The set of cats is a subset of the set of mammals.

Most sets have many different subsets. In fact, every set other than the empty set has more than one subset. For example, the set \( A = \{1, 2, 3\} \) has eight subsets, including itself:

\[
\emptyset \quad \{1\} \quad \{1, 2\} \quad \{1, 3\} \quad \{1, 2, 3\}
\]

\[
\{2\} \quad \{1, 3\} \quad \{2, 3\}
\]

\[
\{3\}
\]
In general, a set having \( n \) elements has \( 2^n \) subsets. The definition of the subset relation (in terms of the set membership relation) contains a material conditional. The set \( A \) is a subset of the set \( B \) just in case for anything you pick if that thing is an element of \( A \), then it is an element of \( B \). Hence, the empty set, \( \emptyset \), is a subset of every set because the antecedent of the conditional is false.

We now define one more operation on sets. The set whose elements are all of the subsets of an arbitrary set \( \alpha \) is called the power set of \( \alpha \), denoted \( \mathcal{P}(\alpha) \). The power set operator is the function \( \mathcal{P}(\cdot) \) that takes a set as input and outputs its power set. We’ve already seen the power set of the set \( A = \{1, 2, 3\} \), which is \( \mathcal{P}(A) = \{ \emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, A \} \). If a set has \( n \) elements, then its power set has \( 2^n \) elements.

5.5 Extensionality and Some Philosophy of Sets

The most important idea in set theory is that the identity of a set is completely determined by the elements that it contains, collectively called the extension of the set. You might wonder how else the identity of a set might be determined. Good question! A natural alternative to determining the identity of a set by its extension is determining the identity of a set by the rule one used in forming or picking out the set, which rule is called the intension of the set. Rules are plentiful things, and the same extension—the same elements that are collected into a specific set—might be specified by any number of cleverly thought-up rules. The extension of the empty set offers an excellent example. The rule, “Pick all of Jonathan Livengood’s siblings,” picks out exactly the elements of the empty set, since Jonathan Livengood has no siblings. The rule, “Pick all of the Beatles albums recorded before 1900,” also picks out exactly the elements of the empty set.
So do the rules, “Pick all of the moons of Mercury,” and, “Pick all of the ocean trenches deeper than the Mariana Trench.” These all have the same extension, but the rules are clearly different.

Extensionality involves denying that sets formed in different ways are different sets. According to extensionality, the intension of a set is *not part* of its identity. For an example that does not involve the empty set, consider the universe of discourse consisting of the hats pictured in Figure 5.8 below:

![Figure 5.8: A Universe of Hats](image)

Now, consider the following two classification rules: (1) a hat is a member of the set $S$ if and only if it is labeled with a vowel; and (2) a hat is a member of the set $T$ if and only if it is red. The two sets have the same extension in our universe of hats: they both pick out hats a, e, i, o, and u.\(^3\) But obviously, they do not have the same intension.

\(^3\) My wife insists that hat a is pink, not red, but let us pretend!
According to extensionality, there is exactly one empty set. The set of unicorns is identical to the set of faeries, since neither unicorns nor faeries exist. Similarly, by extensionality, the sets $S$ and $T$ from our universe of hats are identical. The identity of the two sets holds even though we can imagine alternative universes where $S$ and $T$ are not identical. If we were to swap the labels $a$ and $p$, for example, then the sets $S$ and $T$ as we have defined them would no longer be identical.

Extensionality looks like a very simple and straightforward idea, but appearances can be deceiving. In some standard developments of set theory—for example in versions of Zermelo-Fraenkel (ZF) set theory like the one we develop in Section 5.6—the only things that exist are sets. In other set theories, like Quine’s New Foundations (NF), some things that exist—and that may be elements of sets—are not sets but are called *ur-elements*. At first glance, the most natural way to think about our universe of hats is as a set of ur-elements. After all, a fez is not a set, right? Well, not so fast! That a fez is not a set seems like a reasonable thing to say. But then, it also seems reasonable to say that a number is not a set. And yet, according to one standard analysis of numbers in contemporary mathematics, numbers are sets. For example, the sequence of $T$ sets above might be understood as a *definition* of the natural numbers. Or, we might define the natural numbers with the infinite sequence below, instead:

\[
0 = \emptyset \\
1 = \{0\} \\
2 = \{0, 1\} \\
3 = \{0, 1, 2\} \\
:\]

On this analysis, the number one is a member of every natural number greater than one. Weird! Weirder still is the possibility that something like a fez might be a set. But there might be a good motivation for thinking that a fez really is a set.
Suppose that no hat is a set. What should we say about how the membership relation applies to a thing that is not a set? That is, supposing the fez is not a set, what should we say about the truth-value of the sentence, “The fireman’s helmet is an element of the fez”? One natural thing to say is that the sentence is false. Anything that is not a set has no members. But in that case, things that we would have called “non-sets” really are sets: every non-set has the same extension as the empty set. By extensionality, every non-set is identical to the empty set. The construction hat, the fireman’s helmet, the crown, and all the straw hats in our universe of hats are identical to the empty set. And therefore, all of the elements in our universe of hats are identical to one another. We thought there were 21 things in our universe, and we thought that those things were hats, not sets. But it turns out that there is only one thing in our universe of discourse after all: the empty set!

In order to avoid these nasty consequences, we can make one of several moves. Here are three possibilities. We could qualify the axiom of extensionality so that it applies only to objects that we have previously identified as sets. That keeps ur-elements like fezzes (if they really are ur-elements) from turning out to be identical to the empty set. Alternatively, we could modify the membership relation so that it applies only to sets. But in order to carry out the modification, we will need to drop our commitment to classical, two-valued logic and introduce a truth-value gap. Our approach is to treat everything as a set, just as we did with the natural numbers. In that case, we get the odd consequence that fezzes really are sets, but not the odd consequence that every hat is really the empty set. In terms of our formalism, we will take the third option: everything

---

4 The reason that a truth-value gap might be called for here is that we can write down sentences like, “The construction hat is a member of the fez.” In classical, two-valued logic, such a sentence must be either true or false. But so long as extensionality applies to everything in the universe, assigning either truth-value to the target sentence implies that the fez is a set.
will be treated as if it is a set of some sort or other, and every distinct thing in our universe will be a distinct set.

Sets present several difficult philosophical challenges. Some of these are metaphysical and some are epistemological. For example, we want to know what the set membership relation is like. Is being in a set similar to being in a box? If not, what does it mean for something to be in a set? This kind of problem has been around nearly as long as philosophy itself. One can see the problem of set membership (or something very much like it) in Plato’s doctrine of the Forms. We also want to know how to identify sets. Boxes are things in the world that we can point at. What about sets? Maybe we can point to them as well, if things like fezzes are really sets. But if fezzes really are sets, how do we come to know which sets they are? Going forward, we will ignore the philosophical problems (which are mostly problems in the philosophy of mathematics) and focus on developing set theory at a purely formal level.

5.6 Adding Sets to Our Formal Language: A Simple Axiomatic Set Theory

So far, we have been thinking of sets as part of our meta-language—the language we use to talk about our formal, first-order language. In this section, we show how to add sets to our formal language. In order to avoid paradoxes, we will use a simplified version of the axiomatic approach to set theory first described by Ernst Zermelo (1908).

Rather than giving new rules of inference for working with sets, we will state some axioms, which may be written on a line of a proof at any time without incurring any assumption debts. We actually don’t have to introduce any new notation to our first-order language. Since sets are definite (though abstract) objects, they are the sorts of things that could be given proper names. Hence, we could use constant terms, like a or b, to represent sets, like {1, Bob, {3}} or
And that is exactly what we will do. However, in addition to using lower-case letters to denote sets, we will go ahead and use the special symbol $\emptyset$ to denote the empty set. As before, constants, like $a$, $b$, and $c$, and variables, like $x$, $y$, and $z$, are terms in our language. The only difference is that now, the constant symbols (and $\emptyset$) represent sets, and the variable symbols are placeholders for sets.

With respect to our formal language, the set membership relation is just like any other relation. If we fill up both blanks in the set membership relation with terms, the result is a well-formed formula. If the terms are both constants, the result is a sentence that gets a truth-value. And we can quantify over variable terms that fill up blanks in the set membership relation. For example, the formula $(\exists x)(x \in a)$ says that there is something that is an element of the thing named by $a$. As we have already said, we will assume that everything that gets a name is, in fact, a set. The assumption that everything is a set is not philosophically indefensible, but our motivation is mainly to make our formalism a bit simpler.

Although we are not introducing explicit rules for the set operations, we will make use of the definitions in formal proofs. The idea is that whenever a formula having the form on the left-hand side of the $\equiv$ symbol in one of our definitions appears on some line in a proof, we may write the corresponding formula having the form on the right-hand side of the $\equiv$ symbol on a later line by appealing to the relevant definition. Several examples will illustrate.

**Example 5.1:** $\{ a \in b \} \vdash (a \in (b \cap b))$

<table>
<thead>
<tr>
<th>Line</th>
<th>Formula</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1) $a \in b$</td>
<td>A (premiss)</td>
</tr>
<tr>
<td>1</td>
<td>(2) $(a \in b) \land (a \in b)$</td>
<td>1 $\land$</td>
</tr>
<tr>
<td>1</td>
<td>(3) $(a \in (b \cap b))$</td>
<td>2 Defn $\cap$</td>
</tr>
</tbody>
</table>
Example 5.2: \( \{ (d \in a) \} \vdash (d \in (a \cup b)) \)

1 (1) \( (d \in a) \) A (premiss)
1 (2) \( (d \notin b) \rightarrow (d \in a) \) 1 \( \rightarrow I \)
1 (3) \( (d \in a) \lor (d \in b) \) 2 \( \lor I \)
1 (4) \( (d \in (a \cup b)) \) 3 Defn \( \cup \)

Example 5.3: \( \{ (d \in (a \cup b)), (d \notin a) \} \vdash (d \in b) \)

1 (1) \( (d \in (a \cup b)) \) A (premiss)
2 (2) \( (d \notin a) \) A (premiss)
1 (3) \( (d \in a) \lor (d \in b) \) 1 Defn \( \cup \)
4 (4) \( (d \in a) \) A (for CP)
2 (5) \( (d \notin b) \rightarrow (d \notin a) \) 2 \( \rightarrow I \)
4 (6) \( (d \notin b) \rightarrow (d \in a) \) 4 \( \rightarrow I \)
2,4 (7) \( \neg (d \notin b) \) 5,6 \( \neg I \)
2,4 (8) \( (d \in b) \) 7 \( \neg E \)
2 (9) \( (d \in a) \rightarrow (d \in b) \) 4,8 CP
10 (10) \( (d \in b) \) A (for CP)
(11) \( (d \in b) \rightarrow (d \in b) \) 10 CP
1,2 (12) \( (d \in b) \) 3,9,11 \( \lor E \)

Example 5.4: \( \{ (\exists x)(x \in (a \setminus b)) \} \vdash (\exists x)(x \in (a \cup b)) \)

1 (1) \( (\exists x)(x \in (a \setminus b)) \) A (premiss)
2 (2) \( d \in (a \setminus b) \) A (for CP)
2 (3) \( (d \in a) \land (d \notin b) \) 2 Defn \( \setminus \)
2 (4) \( (d \in a) \) 3 \( \land E \)
2 (5) \( (d \notin b) \rightarrow (d \in a) \) 4 \( \rightarrow I \)
2 (6) \( (d \in a) \lor (d \in b) \) 5 \( \lor I \)
2 (7) \( d \in (a \cup b) \) 6 Defn \( \cup \)
2 (8) \( (\exists x)(x \in (a \cup b)) \) 7 \( \exists I \)
(9) \( (d \in (a \setminus b)) \rightarrow (\exists x)(x \in (a \cup b)) \) 2,8 CP
1 (10) \( (\exists x)(x \in (a \cup b)) \) 2,9 \( \exists E \)

Now, for the axioms. Our first axiom for set theory, the axiom of extensionality, says that the identity of a set is completely determined by its extension. We do not need to know how the elements of a set were collected together. We don’t need to know by what rule(s) two sets were formed to decide whether or not they are the same. According to the axiom of extensionality, two
sets are identical if and only if they have exactly the same elements. For clarity in the exposition, we will sometimes use square brackets interchangeably with parentheses. Bearing that in mind, we may state the axiom of extensionality as follows:

\[(\forall x)(\forall y)[(x = y) \leftrightarrow ((\forall z)((z \in x) \leftrightarrow (z \in y)))]\]

The axiom says (in plain language) that two sets are identical if and only if they have exactly the same elements.

We make use of the axiom of extensionality in proofs by eliminating the universal quantifiers over x and y. The variables x and y are replaced by the names of sets that we want to show to be identical. For example, if we wanted to show that the set a is identical to the set b, we would instantiate the axiom of extensionality to obtain the biconditional:

\[(a = b) \leftrightarrow (\forall z)((z \in a) \leftrightarrow (z \in b))\]

The biconditional in (AEI) says that the sets a and b have all and only the same elements if and only if they are identical. In order to complete a proof that the set a is identical to the set b, we would need to show that for anything you pick, it is an element of a if and only if it is an element of b. In using the axioms of set theory, it will be convenient to have rules of inference for working with the biconditional. As with our other rules of inference, we have an introduction rule and an elimination rule.

**Biconditional Introduction (↔I)**

Suppose that \(\phi\) and \(\psi\) are well-formed formulas in our formal language. If the formulas \((\phi \rightarrow \psi)\) and \((\psi \rightarrow \phi)\) appear on separate lines in a proof, then the formula \((\phi \leftrightarrow \psi)\) may be written on any later line. Schematically:

\[
\frac{(\phi \rightarrow \psi) \\
(\psi \rightarrow \phi)}{(\phi \leftrightarrow \psi)}
\]
When we use the biconditional introduction rule, we write down in the assumption column for $(\phi \leftrightarrow \psi)$ all the numbers that appear in the assumption columns for $(\phi \rightarrow \psi)$ and $(\psi \rightarrow \phi)$.

Example 5.5: $\{ \}$ ⊢ $(a \in b) \leftrightarrow (a \in b)$

1. $(a \in b)$ A (for CP)
2. $((a \in b) \rightarrow (a \in b))$ 1 CP
3. $((a \in b) \leftrightarrow (a \in b))$ 2 ↔I

Biconditional Elimination ($\leftrightarrow E$)

Suppose that $\phi$ and $\psi$ are well-formed formulas in our formal language. If the formulas $(\phi \leftrightarrow \psi)$ and $\phi$ appear on separate lines in a proof, then $\psi$ may be written on any later line in the proof. And if the formulas $(\phi \leftrightarrow \psi)$ and $\psi$ appear on separate lines in a proof, then $\phi$ may be written on any later line in the proof. Schematically, we have:

\[
\begin{array}{c}
\phi \\
\hline
\psi
\end{array}
\quad
\begin{array}{c}
\phi \\
\hline
\psi
\end{array}
\]

When we use the biconditional elimination rule, we have to write down in the assumption column for the conclusion all of the numbers in the assumption columns for the premisses.

Example 5.6: $\{ (a = a) \leftrightarrow ((d \in a) \leftrightarrow (b \in a)), (b \in a) \} \vdash (d \in a)$

1. $(a = a)$ A (premiss)
2. $(b \in a)$ A (premiss)
3. $(\forall x)(x = x)$
4. $(a = a)$
5. $((d \in a) \leftrightarrow (b \in a))$ 1,4 ↔E
6. $(d \in a)$ 2,5 ↔E
Now that we have some rules for working with biconditionals, let’s put it all together with the axiom of extensionality in a proof of identity.

**Example 5.7**: \( \left\{ \right\} \vdash (a = (a \cap a)) \\

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1) ((b \in a))</td>
<td>A (for CP)</td>
</tr>
<tr>
<td>1</td>
<td>(2) (((b \in a) \land (b \in a)))</td>
<td>1 (\land I)</td>
</tr>
<tr>
<td>1</td>
<td>(3) ((b \in (a \cap a)))</td>
<td>2 Defn (\cap)</td>
</tr>
<tr>
<td>5</td>
<td>(4) ((b \in a) \rightarrow (b \in (a \cap a)))</td>
<td>1,3 CP</td>
</tr>
<tr>
<td>5</td>
<td>(5) ((b \in (a \cap a)))</td>
<td>A (for CP)</td>
</tr>
<tr>
<td>5</td>
<td>(6) (((b \in a) \land (b \in a)))</td>
<td>5 Defn (\cap)</td>
</tr>
<tr>
<td>5</td>
<td>(7) ((b \in a))</td>
<td>6 (\land E)</td>
</tr>
<tr>
<td>(8)</td>
<td>((b \in (a \cap a)) \rightarrow (b \in a))</td>
<td>5,7 CP</td>
</tr>
<tr>
<td>(9)</td>
<td>((b \in a) \leftrightarrow (b \in (a \cap a)))</td>
<td>4,8 (\leftrightarrow I)</td>
</tr>
<tr>
<td>(10)</td>
<td>((\forall z)((z \in a) \leftrightarrow (z \in (a \cap a))))</td>
<td>9 (\forall I)</td>
</tr>
<tr>
<td>(11)</td>
<td>((\forall x)(\forall y)((\forall z)((z \in x) \leftrightarrow (z \in y)) \leftrightarrow (x = y)))</td>
<td>Ax. Ext.</td>
</tr>
<tr>
<td>(12)</td>
<td>((\forall y)((\forall z)((z \in a) \leftrightarrow (z \in y)) \leftrightarrow (a = y)))</td>
<td>11 (\forall E)</td>
</tr>
<tr>
<td>(13)</td>
<td>((\forall z)((z \in a) \leftrightarrow (z \in (a \cap a))) \leftrightarrow (a = (a \cap a)))</td>
<td>12 (\forall E)</td>
</tr>
<tr>
<td>(14)</td>
<td>((a = (a \cap a)))</td>
<td>10,13 (\leftrightarrow E)</td>
</tr>
</tbody>
</table>

The basic idea illustrated in Example 5.7 recurs over and over again in proofs of identity.

Suppose we want to prove that some set \(a\) is identical to some other set \(b\). We first assume that an arbitrary constant is an element of the set \(a\). We then show that that constant is an element of the set \(b\), and we discharge the assumption. Then we do the same thing in reverse. We assume that the same arbitrary constant is an element of the set \(b\). We then show that that constant is an element of the set \(a\), and we discharge the assumption. That lets us introduce a biconditional. Since the constant we used was arbitrary, we can then introduce a universal to quantify over the arbitrary constant that we picked. Finally, we strip down the axiom of extensionality by instantiating the variables \(x\) and \(y\) with the constants \(a\) and \(b\), and we use biconditional elimination in order to complete the proof that the set \(a\) is identical to the set \(b\).
When set theory was first being developed at the end of the nineteenth century, mathematicians and philosophers thought that sets could be defined by predication. This intuitive idea is sometimes captured in what is called the naïve comprehension axiom (schema):

\[(\exists x)(\forall y)((y \in x) \leftrightarrow Fy)\]

where \(F\) is an arbitrary predicate. (The fact that \(F\) is an arbitrary predicate is what makes naïve comprehension an axiom *schema* rather than a simple axiom. A schema encodes an infinite number of formulas.) According to the naïve comprehension axiom, there is a set whose elements are exactly those things that have the predicate \(F\). As we saw in Section 5.2, defining sets by arbitrary predication leads to absurdity. Hence, the axiom (schema) above—and the set theory that is based on it—is called naïve.

In the usual development of Zermelo-Fraenkel set theory, the naïve comprehension axiom is replaced with a less-naïve comprehension axiom, sometimes called the *axiom of separation*. The axiom of separation defines sets by qualifying existing sets. The axiom of separation says that for any set, the elements of the set that *also* satisfy some predicate form a set. In this way, contradiction is avoided, since every set specified by the axiom of separation is separated away from an existing set known to be non-absurd.

We simplify things by beginning with the empty set and then building up every other set that we need from the empty set. The motivation here is the idea that if any set is acceptable, the empty set is acceptable. Hence, we need an axiom asserting the existence of the empty set:

\[(Ax. \text{ Empty Set}) \quad (\exists x)(\forall y)(y \notin x)\]

The axiom of the empty set entails that there exists at least one set that has no elements, and as we saw earlier, the axiom of extensionality entails that there is at most one set that has no elements. Hence, we can talk about *the* empty set.
The axiom of the empty set is an existentially quantified sentence. Hence, we need the rule of existential elimination in order to make use of the axiom of the empty set. For example:

**Example 5.8:** \( \{ \} \vdash (\exists x) (a \notin x) \)

<table>
<thead>
<tr>
<th>Step</th>
<th>Line</th>
<th>Formula</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>( (\exists x)(\forall y)(y \notin x) )</td>
<td>Ax. Empty Set</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>( (\forall y)(y \notin b) )</td>
<td>A (for CP)</td>
</tr>
<tr>
<td>2</td>
<td>(3)</td>
<td>( b \notin b )</td>
<td>2 ( \forall E )</td>
</tr>
<tr>
<td>4</td>
<td>(4)</td>
<td>( (\exists x)(a \notin x) )</td>
<td>3 ( \exists I )</td>
</tr>
<tr>
<td>5</td>
<td>(5)</td>
<td>( (\forall y)(y \notin b) \rightarrow (\exists x)(a \notin x) )</td>
<td>2,4 ( \text{CP} )</td>
</tr>
<tr>
<td>6</td>
<td>(6)</td>
<td>( (\exists x)(a \notin x) )</td>
<td>1,5 ( \exists E )</td>
</tr>
</tbody>
</table>

The axiom of the empty set tells us that there exists at least one set that has no elements. But a set theory with only one set would be boring. So, we need some axioms that let us build new sets using the empty set. The result is a set theory that is literally built on nothing!

The first constructive axiom that we consider is called the *axiom of pairing*, which we will use to construct sets called *singletons*, which have exactly one element, and sets called *pair sets*, which have exactly two elements. Formally, the axiom says:

\[
(\text{Ax. Pairing}) \quad (\forall x)(\forall y)(\exists z)(\forall w)((w \in z) \leftrightarrow ((w = x) \vee (w = y)))
\]

The axiom of pairing says that something is in the set being constructed just in case it is identical to one or the other of two specified elements.

**Example 5.9:** \( \{ \} \vdash (\exists x)(\exists y)(\exists z)((x \in z) \land (y \in z)) \)

<table>
<thead>
<tr>
<th>Step</th>
<th>Line</th>
<th>Formula</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1)</td>
<td>( (\forall x)(\forall y)(\exists z)(\forall w)((w \in z) \leftrightarrow ((w = x) \vee (w = y))) )</td>
<td>Ax. Pair</td>
</tr>
<tr>
<td>2</td>
<td>(2)</td>
<td>( (\forall y)(\exists z)(\forall w)((w \in z) \leftrightarrow ((w = a) \vee (w = y))) )</td>
<td>1 ( \forall E )</td>
</tr>
<tr>
<td>3</td>
<td>(3)</td>
<td>( (\exists z)(\forall w)((w \in z) \leftrightarrow ((w = a) \vee (w = b))) )</td>
<td>2 ( \forall E )</td>
</tr>
<tr>
<td>4</td>
<td>(4)</td>
<td>( (\forall w)((w \in d) \leftrightarrow ((w = a) \vee (w = b))) )</td>
<td>A (for CP)</td>
</tr>
<tr>
<td>4</td>
<td>(5)</td>
<td>( d \in d \leftrightarrow ((a = a) \vee (a = b)) )</td>
<td>4 ( \forall E )</td>
</tr>
<tr>
<td>4</td>
<td>(6)</td>
<td>( b \in d \leftrightarrow ((b = a) \vee (b = b)) )</td>
<td>4 ( \forall E )</td>
</tr>
<tr>
<td>7</td>
<td>(7)</td>
<td>( (\forall x)(x = x) )</td>
<td>=I</td>
</tr>
<tr>
<td>8</td>
<td>(8)</td>
<td>( a = a )</td>
<td>7 ( \forall E )</td>
</tr>
<tr>
<td>9</td>
<td>(9)</td>
<td>( a \neq b \rightarrow (a = a) )</td>
<td>8 ( \rightarrow I )</td>
</tr>
<tr>
<td>10</td>
<td>(10)</td>
<td>( ((a = a) \vee (a = b)) )</td>
<td>9 ( \forall I )</td>
</tr>
</tbody>
</table>
Notice that the proof in Example 5.9 does not guarantee that there is a set with two distinct elements. For all we’ve shown, a and b could be identical! How would you show that there exists a set that has two distinct elements? When the elements that we specify are identical to one another, the constructed set will have exactly one element. We use this fact to get off the ground using only the empty set. By instantiating both x and y with the empty set, we get the singleton set whose only element is the empty set. Once we have such a set, we are off and running: we can construct indefinitely many further singleton and pair sets using the axiom of pairing.

The second constructive axiom that we consider is called the **axiom of union**. The axiom of union is designed to let us construct sets out of the elements of the elements of an existing set. That is not a typo! We use the axiom of union to construct sets whose elements are all and only the elements that belong to all of the elements of a given set. Formally, the axiom says:

\[ (Ax. \text{ Union}) \quad (\forall x)(\exists y)(\forall z)((z \in y) \leftrightarrow (\exists w)((w \in x) \land (z \in w)))) \]

Suppose that we use the axiom of pairing to construct the following four singleton sets: \( a = \{\emptyset\} \), \( b = \{a\} = \{\{\emptyset\}\} \), \( c = \{b\} = \{\{\{\emptyset\}\}\} \), and \( d = \{c\} = \{\{\{\{\emptyset\}\}\}\} \). Then use the pairing axiom again to construct the following two pair sets: \( \{a, b\} \) and \( \{c, d\} \). Use the pairing axiom one last time to construct the pair set: \( \{\{a, b\}, \{c, d\}\} \). Finally, we use the axiom of union to construct the four-element set \( \{\emptyset, a, b, c\} \), which is the set consisting of exactly the elements of the sets a, b, c, and
d. That is, the union set is the set consisting of the elements of the elements of the pair set that we constructed: \{\{a, b\}, \{c, d\}\}. A formal proof of this construction would be too long to conveniently include in this text.

The third constructive axiom that we consider is called the *axiom of power set*. When every element of a set \(\alpha\) is also an element of a second set \(\beta\), we say that \(\alpha\) is a *subset* of \(\beta\), denoted by \((\alpha \subseteq \beta)\). The axiom of power set is designed to let us construct a set whose elements are all of the subsets of an existing set. We give a formal definition of the subset relation, as we did for the set operations, as follows:

\[
(\alpha \subseteq \beta) \quad =_{df} \quad (\forall z)((z \in \alpha) \rightarrow (z \in \beta))
\]

The subset relation is just that: a *relation*, not merely a function. Hence, when \(\alpha\) and \(\beta\) are terms, \((\alpha \subseteq \beta)\) is a well-formed formula, and when \(\alpha\) and \(\beta\) are constants, \((\alpha \subseteq \beta)\) is a sentence. We can give proofs using the definition of subset, as we did with the set operations.

**Example 5.10:** \{\} \vdash (\forall x)(x \subseteq x)

\[
\begin{align*}
1 \quad (1) & \quad (d \in a) \quad \text{A (for CP)} \\
2 \quad (2) & \quad (d \in a) \rightarrow (d \in a) \quad 1 \text{ CP} \\
3 \quad (3) & \quad (\forall z)((z \in a) \rightarrow (z \in a)) \quad 2 \forall I \\
4 \quad (4) & \quad (a \subseteq a) \quad 3 \text{ Defn } \subseteq \\
5 \quad (5) & \quad (\forall x)(x \subseteq x) \quad 4 \forall I
\end{align*}
\]

**Example 5.11:** \{(a \subseteq b), (d \notin b)\} \vdash (d \notin a)

\[
\begin{align*}
1 \quad (1) & \quad (a \subseteq b) \quad \text{A (premiss)} \quad 1 \text{ Defn } \subseteq \\
2 \quad (2) & \quad (d \notin b) \quad \text{A (premiss)} \\
3 \quad (3) & \quad (\forall z)((z \in a) \rightarrow (z \in b)) \quad 1 \text{ Defn } \subseteq \\
4 \quad (4) & \quad (d \in a) \rightarrow (d \in b) \quad 3 \forall E \\
5 \quad (5) & \quad (d \in a) \rightarrow (d \notin b) \quad 2 \rightarrow I \\
6 \quad (6) & \quad (d \notin a) \quad 4,5 \sim I
\end{align*}
\]

With the subset relation in hand, we may now state the axiom of power set as follows:

\[
(\text{Ax. Power}) \quad (\forall x)(\exists y)(\forall z)((z \in y) \leftrightarrow (z \subseteq x))
\]
The set whose elements are all of the subsets of the set $\alpha$ is called the *power set* of $\alpha$, denoted by $\mathcal{P}(\alpha)$. For example, consider the set $a = \{1, 2, 3\}$. The power set of $a$, $\mathcal{P}(a)$, is the set:

$$\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, a\}$$

As noted previously, if a set has $n$ elements, its power set has $2^n$ elements.
References


The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man’s mind.

– James Clerk Maxwell (1850)

In previous chapters, we constructed a formal language that we could use to translate a small fragment of ordinary language; we characterized a species of good arguments, the valid ones, which are such that if the premisses are all true, then the conclusion is guaranteed to be true; and we described a system of deductive rules of inference (also called transformation rules) that let us derive some sentence(s) on the basis of some sentences thought of as premisses. Deductive logic is a beautiful and important tool. Using the deductive logic that we have constructed, we could build up many more complicated mathematical systems by introducing axioms and definitions, as we did with set theory at the end of Chapter 5. However, as Maxwell suggests, we would be in a very sorry state if we were limited to deductive logic alone, for while deductive logic characterizes the very strong notion of evidential support that we have called validity, we rarely hold our arguments to such a high standard either in science or in ordinary life. Rather, we are usually contented to observe that if some premisses are true, then some conclusion is likely to be true or (what is an even weaker claim) more likely to be true given the evidence. In this chapter, we will expand our logic to characterize cases where the premisses of an argument
confirm its conclusion but (might) fall short of providing a deductive guarantee. We give an account of confirmation in terms of probability theory. Our approach will be to treat probability as a generalization of the valuation function that assigns truth-values to the sentences in our formal language. As Peirce (1878) puts it, “The theory of probabilities is simply the science of logic quantitatively treated.”

Probability is a peculiar kind of measure. For the moment, I want to set aside the question of what exactly probability is a measure of and focus on the formal properties of the measure itself. But before getting too deeply into a formal account of probability, let me say a few informal things about probability in order to make the ideas a little more intuitive. In ordinary English, probability is synonymous with likelihood or chance. Ordinarily, we talk about occurrences as being more or less probable – that is, as having more or less of the stuff that probability measures—whatever that stuff is. For example, I do not know whether it will rain tomorrow or not. There is some probability that it will rain tomorrow and some that it will not. In other words, there is some amount of the stuff that probability measures on the side of rain and some on the side of not-rain. In this vein, we say that it is more or less likely that it will rain tomorrow, or we say that the chance it will rain tomorrow is such and so much. Hence, the theory of probability is just that: a theory of a pre-existing thing or of a pre-theoretical notion. We want to use mathematical tools to develop a deeper, fuller understanding of that thing: probability. And in so doing, we hope to get a better theory of inductive arguments and what makes inductive arguments good or bad.

1 We will say more about confirmation in Section 6.4. For developments and discussions of confirmation, see Carnap (1950), Carnap (1952), Horwich (1982), Maher (1993), and Fitelson (2001). For criticisms, see Earman (1992) and Norton (2011).
6.1 Valuations, Probability Functions, and Universes of Discourse

In our construction of zeroth-order logic, we implicitly appealed to a valuation function that maps sentences in our formal language to truth-values. In other words, we appealed to a function that takes a sentence as input and gives either True or False as output. Moreover, we explicitly noted that the sentential connectives act like truth functions. If we know the truth-values for the simple sentences, then we can determine the truth-values of complex sentences built up from the simple sentences using the sentential connectives. For example, if \( v(\cdot) \) is the valuation function, then we could define the conjunction by saying that \( v(P \land Q) = T \) if and only if \( v(P) = T \) and \( v(Q) = T \). The probability function assigns to each sentence of our language an analogue of a truth-value: a probability value or a probability. However, unlike in zeroth-order logic, knowing the probability values for the simple sentences is not enough (in general) to determine the probability values for the complex sentences in which those simple sentences appear as parts. Rather, we need some additional constraints, which are typically imposed by some combination of an interpretation of probability and facts about what the world is like.

The syntax for probability theory is simple. If \( \varphi \) is a sentence in our formal language, then \( \Pr(\varphi) \) denotes the probability of \( \varphi \), and we write \( \Pr(\varphi) = p \) to indicate that the specific probability value assigned to \( \varphi \) is \( p \). To illustrate, let \( W = \ldots \text{is a planet,} \) let \( A = \ldots \text{supports life,} \) and let \( e \) denote the Earth. Then the sentence \( (\exists x)((Wx \land Ax) \land (x \neq e)) \) says that there exists a planet other than Earth that supports life, and the complicated-looking expression \( \Pr((\exists x)((Wx \land Ax) \land (x \neq e))) = 0.5 \) says that the probability that there exists a planet other than Earth that supports life is one half. Importantly, given the way we have built our formal language, expressions like \( \Pr((\exists x)((Wx \land Ax) \land (x \neq e))) = 0.5 \) are not formulas in first-order logic. They are part of a meta-language within which we can talk about sentences of first-order
logic. Given that $v(\cdot)$ denotes the valuation function in zeroth- and first-order logic, we should recognize that $v(\exists x)((Wx \land Ax) \land (x \neq e)) = T$ is not a sentence in first-order logic either. We are here doing something akin to *mentioning* a sentence in order to say something about it without actually *using* the sentence. That is, we are doing something similar to what we do when we observe that “goat” has four letters or that “smiles” has a “mile” between its two s’s.

The semantics for probability theory is not so simple. In fact, there is a lot of controversy about how best to understand what we mean when we talk about probability. For now, we will content ourselves with three mathematical constraints on the probability function, first stated by Kolmogorov (1933), given below in terms of arbitrary sentences $\varphi$ and $\psi$.

- **[Non-negativity]** The probability of each sentence is a non-negative real number.
- **[Normality]** If $\{\} \vDash \varphi$, then $Pr(\varphi) = 1$.
- **[Finite Additivity]** If $\{\} \vDash \neg(\varphi \land \psi)$, then $Pr(\varphi \lor \psi) = Pr(\varphi) + Pr(\psi)$

Although there seems to be no end to debate about the meaning of assertions about probability, there is considerably less debate about the mathematical development of the theory. The constraints above are sometimes called the *Kolmogorov axioms* for probability, and we can use the axioms to compute the probabilities of various sentences as long as we have initial probability assignments for the right contingent sentences.

*Example 6.1:* Suppose the probability of the sentence $A$ is $1/3$, the probability of the sentence $B$ is $1/4$, and the sentences $A$ and $B$ are inconsistent. We want to know the probability that either $A$ is true or $B$ is true. Since $A$ and $B$ are inconsistent, it follows that $\{\} \vDash \neg(A \land B)$. Hence, by Finite Additivity, the probability that either $A$ is true or $B$ is true is $1/3 + 1/4 = 7/12$.

*Example 6.2:* We want to know the probability that either $A$ is true or $\neg A$ is true. A simple truth table construction shows that $\{\} \vDash (A \lor \neg A)$. Hence, the probability that either $A$ is true or $\neg A$ is true is equal to one.

---

2 That is not to say that there is *no* debate, just that there is *less*. Two features of the mathematics do come in for disputes: (1) the treatment of additivity, and (2) the treatment of conditional probability.
Notice that while the Kolmogorov axioms tell us both that $\Pr(A \lor \neg A) = 1$ and also that $\Pr(A \lor \neg A) = \Pr(A) + \Pr(\neg A)$, the axioms allow the probability of $A$ to be any number between zero and one inclusive. In order to simplify things, we will often adopt the assumption that each of the sentences describing possible outcomes of an experiment or observation, like the flip of a coin or the spin of a roulette wheel, are all equally likely.

We will often find it convenient to think about sentences in terms of the set of worlds in which they are true. Thinking about sentences that way lets us reformulate the Kolmogorov axioms in terms of set theory. Let’s see how that works. Let $U$ denote a set, called the universe of discourse. Any element $u$ of the set $U$ is called a unit. In order to simplify the discussion somewhat, we will assume that the universe of discourse is a finite set. (If the universe is infinite, then serious mathematical complications arise, which are beyond the scope of this book.) The universe of discourse is a coarse-grained version of the space of all possible worlds, which specifies all of the possible outcomes of a proposed experiment.³

**Example 6.3:** Imagine that we want to observe the behavior of some dice. Specifically, we want to toss two six-sided dice and record the results. What does the universe of discourse look like? The universe should encode all of the possible outcomes of the toss of two dice. The 36 possible outcomes are listed in Table 1 below. Each outcome is an ordered pair, where the first entry is the number on the first die and the second entry is the number on the second die:

<table>
<thead>
<tr>
<th>&lt;1,1&gt;</th>
<th>&lt;1,2&gt;</th>
<th>&lt;1,3&gt;</th>
<th>&lt;1,4&gt;</th>
<th>&lt;1,5&gt;</th>
<th>&lt;1,6&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2,1&gt;</td>
<td>&lt;2,2&gt;</td>
<td>&lt;2,3&gt;</td>
<td>&lt;2,4&gt;</td>
<td>&lt;2,5&gt;</td>
<td>&lt;2,6&gt;</td>
</tr>
<tr>
<td>&lt;3,1&gt;</td>
<td>&lt;3,2&gt;</td>
<td>&lt;3,3&gt;</td>
<td>&lt;3,4&gt;</td>
<td>&lt;3,5&gt;</td>
<td>&lt;3,6&gt;</td>
</tr>
<tr>
<td>&lt;4,1&gt;</td>
<td>&lt;4,2&gt;</td>
<td>&lt;4,3&gt;</td>
<td>&lt;4,4&gt;</td>
<td>&lt;4,5&gt;</td>
<td>&lt;4,6&gt;</td>
</tr>
<tr>
<td>&lt;5,1&gt;</td>
<td>&lt;5,2&gt;</td>
<td>&lt;5,3&gt;</td>
<td>&lt;5,4&gt;</td>
<td>&lt;5,5&gt;</td>
<td>&lt;5,6&gt;</td>
</tr>
<tr>
<td>&lt;6,1&gt;</td>
<td>&lt;6,2&gt;</td>
<td>&lt;6,3&gt;</td>
<td>&lt;6,4&gt;</td>
<td>&lt;6,5&gt;</td>
<td>&lt;6,6&gt;</td>
</tr>
</tbody>
</table>

**Table 6.1: Possible Outcomes of the Toss of Two Six-Sided Dice**

³ In practice, the universe of discourse includes only the outcomes that we care to consider or that we count as live possibilities. If we’re considering an experiment in which we throw a pair of dice, we do not usually consider the possibility that the dice both land on a vertex and stay there or the possibility that the dice explode or the like.
The universe of discourse—or the way we carve up and think about the space of all possible worlds—is relative to our choice of problem.

Assuming that the events listed in Table 6.1 are, indeed, all of the possible outcomes of tossing two dice, there are 36 sentences of the form, “The dice landed \(<x, y>\),” such that each sentence is contingent, the disjunction of all of the sentences is a tautology, and each pair of sentences is inconsistent. The Kolmogorov axioms require that the disjunction be assigned probability one and that the disjunction of any collection of the 36 sentences be assigned a probability equal to the sum of the individual probabilities of the sentences in the collection. However, as far as the axioms go, any one of the 36 sentences may be assigned any probability value we like between zero and one.

**Example 6.4:** Imagine that we want to observe the age of a student. Specifically, we want to pick a student at random from the class and have the student tell us his or her age in years.⁴ What does the universe of discourse look like? The universe in this case is a set of numbers representing years of age. For every student in the class, if the student is \(n\) years old, then the number \(n\) is an element of the universe. Nothing else is in the universe. For concreteness, suppose that in the class there are 20 students who are 18 years old, 30 students who are 19 years old, 25 students who are 20 years old, 15 students who are 21 years old, 5 students who are 22 years old, 4 students who are 23 years old, and 1 student who is 47 years old. Then the universe of discourse is the set \(U = \{18, 19, 20, 21, 22, 23, 47\}\).

Again, in this book, we will only be concerned with finite universes of discourse. Each unit in the universe of discourse represents some possible occurrence relative to the contemplated experiment or observation, and each subset of the universe of discourse is called an *event*.

Now, let \(E\) be an arbitrary event—that is, an arbitrary subset of \(U\).⁵ We can think of the probability function \(Pr(\cdot)\) as mapping events into the real numbers. Hence, we denote the probability of the event \(E\) by \(Pr(E)\). We will sometimes talk about the probability that a unit

---

⁴ We will assume that no one lies about his or her age!
⁵ For the sake of readability, we will sometimes use capital letters to refer to sets in this chapter.
belongs to \( E \). These expressions are interchangeable. If the probability of the event \( E \) is equal to \( p \), for some real number \( p \), we write, \( \Pr(E) = p \). In an experiment, an event is said to occur if and only if the outcome of the experiment (a unit in the universe) is a member of the relevant event.

*Example 6.5*: Consider the event of picking a student in the class from Example 6.4 who is older than 21 years. The set \( E = \{22, 23, 47\} \) represents the event described. Now, suppose that I pick a student who is 23 years old. In that case, the event \( E \) is said to occur because \( 23 \in E \), and the relevant unit is 23.

We may now restate the Kolmogorov axioms in terms of set theory. Let \( U \) be the universe of discourse, and let \( A \) and \( B \) denote arbitrary subsets of \( U \), i.e. arbitrary events. Then a probability function satisfies the following constraints:

- [Non-negativity] \( \Pr(A) \geq 0 \).
- [Normality] \( \Pr(U) = 1 \).
- [Finite Additivity] If \( A \cap B = \emptyset \), then \( \Pr(A \cup B) = \Pr(A) + \Pr(B) \).

Say that an event containing exactly one unit is a *simple event*. A simple event is a singleton set, and the union over all of the simple events is identical to the universe. If the universe of discourse is finite, we can express the non-negativity and finite additivity constraints in terms of simple events.

- [Non-negativity] The probability of each simple event is non-negative.
- [Finite Additivity] The probability of an arbitrary event \( E \) is equal to the sum of the probabilities of the simple events that are subsets of \( E \).

Let’s think through some examples in terms of simple events.

*Example 6.6*: Consider the universe from Example 6.3, consisting of the 36 possible results of tossing a pair of six-sided dice. Suppose that each simple event in \( U \) is assigned the same probability. What value must be assigned to the simple events in order to satisfy the three requirements on probability assignments? The sum of all of the simple events must be equal to one in order to satisfy the second two constraints. Given that all of the values must be the same, to find the probability assigned to each simple event, we simply divide one by the number of simple events: 36. So, the probability assigned to each simple event is \( 1/36 \).
Example 6.7: Suppose that the simple events in Example 6.3 are all assigned the same probability value, as in Example 6.6. What probability is assigned to the event, “The sum of the two dice is strictly greater than 9”? Well, the event described is equal to the set \( E = \{<4,6>, <5,5>, <5,6>, <6,4>, <6,5>, <6,6>\} \), and the set \( E \) has six elements. So, the probability that a unit \( u \) is an element of \( E \) is equal to the sum of the probabilities of the six simple events \( E_1 = \{<4,6>\}, E_2 = \{<5,5>\}, E_3 = \{<5,6>\}, E_4 = \{<6,4>\}, E_5 = \{<6,5>\}, E_6 = \{<6,6>\} \). Since each of the simple events was assigned the same probability of 1/36, the probability of \( E \) is 6/36 = 1/6.

If the simple events are assigned the same probability values, we say that the simple events are *equally likely*. Classical probability theory assumes that every simple event is equally likely. In some cases, like the dice problems of Examples 6.3, 6.6, and 6.7, the assumption that every simple event is equally likely seems reasonable. That classical probability theory gives reasonable answers to problems about dice is not especially surprising, since classical probability theory was first developed in the 17th century to better understand exactly such problems.\(^6\)

When the simple events are all equally likely, the probability of an arbitrary sentence may be computed by taking the number of outcomes that make the sentence true divided by the total number of possible outcomes. Hence, in classical probability theory, probability problems reduce to counting problems. However, counting problems can be very difficult. For example, consider the following problem.

Example 6.8: The United States citizenship test consists of ten questions drawn from a set of 100 possible questions. Say that two tests are identical if and only if the same questions appear on each test (but possibly in different orders). How many distinct tests are there? You could try to write them all down and count them that way, but it would take you a while, since there are more than 17 trillion distinct tests constructed in this way!\(^7\)

---

\(^6\) The usual story of the development of probability theory is that the Chevalier de Méré (1607-1684) approached Fermat and Pascal with some questions about dice games that de Méré enjoyed playing. He wanted to know, for example, what *odds* to lay on the possibility of rolling at least one six in four throws of a fair, six-sided die.

\(^7\) Since we are ignoring the order in which the questions appear, the counting problem in Example 6.8 is to find how many ways one may choose ten objects from a collection of 100 objects. For natural numbers \( n \) and \( m \), where \( m > n \), the number of ways one may choose \( n \) objects from among \( m \) objects is equal to \( m! / (m-n)! n! \), where \( ! \) is the factorial operator. The number of ways to choose ten questions from 100 possible ones is equal to 100! / (90! 10!), which is (100 \cdot 99 \cdot 98 \cdot 97 \cdot 96 \cdot 95 \cdot 94 \cdot 93 \cdot 92 \cdot 91) / (10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2) \cong 17,310,000,000,000.
Classical probability theory gains a lot of power from the assumption of equal likelihood. However, in many cases, the assumption looks unreasonable.

*Example 6.9:* Suppose there are 100 students in the class. Further suppose that the ages of the students in years are 18, 19, 20, 21, and 22. What is the probability that a student chosen at random is 18 years old?

If we make an equal likelihood assumption, the answer to Example 6.9 is easy to give: 1/5. However, that answer supposes that we are drawing at random from the five ages represented in the class, which does not look especially plausible. How are probability assignments to be made in such cases?

Sometimes we can make the classical theory work by being more specific. In this spirit, we might think of our random draw as coming from the 100 students, as opposed to coming from the five ages.

*Example 6.10:* Suppose there are 30 students who are 18 years old, 25 who are 19 years old, 20 who are 20 years old, 15 who are 21 years old, and 10 who are 22 years old. The following is a natural assignment of probabilities for sentences describing the age of a student sampled from the class:

\[
\begin{align*}
\Pr(\text{The student is 18 years old.}) &= 30/100 = 3/10 \\
\Pr(\text{The student is 19 years old.}) &= 25/100 = 1/4 \\
\Pr(\text{The student is 20 years old.}) &= 20/100 = 1/5 \\
\Pr(\text{The student is 21 years old.}) &= 15/100 = 3/20 \\
\Pr(\text{The student is 22 years old.}) &= 10/100 = 1/10
\end{align*}
\]

Note that the probability assignments are all non-negative real numbers and that they sum to one as required by the axioms.

The probabilities assigned in Example 6.10 are related to the hypothetical counts of students in the class. The probability assignment effectively assumes that each *student* is as likely to be chosen as any other. However, the probability assignment of Example 6.10 is not required by our constraints on probabilities. An infinite number of other assignments is consistent with the Kolmogorov axioms!
Example 6.11: The following probability assignment works just as well as the assignment in Example 6.10 from a formal point of view:

- \( \Pr(\text{The student is 18 years old.}) = 1/16 \)
- \( \Pr(\text{The student is 19 years old.}) = 1/16 \)
- \( \Pr(\text{The student is 20 years old.}) = 1/8 \)
- \( \Pr(\text{The student is 21 years old.}) = 1/4 \)
- \( \Pr(\text{The student is 22 years old.}) = 1/2 \)

As in Example 6.10, the probability of each sentence is a non-negative real number, and the sum of the probabilities is equal to one.

One might object that the probability assignment in Example 6.11 looks strange. (Indeed, it does look strange given the composition of the class described in Example 6.10.) However, the probability assignment in Example 6.11 satisfies the axioms. What more could anyone want? The answer to that question depends in part on what it is that probability values are supposed to be representing, i.e. the answer depends on one’s interpretation of probability. We will return to interpretation of probability in Section 6.3.

6.2 Derived Rules and Conditional Probabilities

Using the Kolmogorov axioms, we can derive several additional rules (theorems, really) that make probability computations easier. We begin with a statement in terms of arbitrary sentences in our formal language and then consider restatements in terms of sets. Here are four especially useful derived rules for arbitrary sentences \( \varphi \) and \( \psi \):

- [Contradiction]: If \( \{ \} \models \neg \varphi \), then \( \Pr(\varphi) = 0 \).
- [Equivalence]: If \( \varphi \) and \( \psi \) are logically equivalent, then \( \Pr(\varphi) = \Pr(\psi) \).
- [Complement]: \( \Pr(\neg \varphi) = 1 - \Pr(\varphi) \).
- [General Add]: \( \Pr(\varphi \lor \psi) = \Pr(\varphi) + \Pr(\psi) - \Pr(\varphi \land \psi) \).

In Example 6.12 below, we derive the rule we are calling Equivalence. Derivations of the other rules are left as exercises for the reader.
Example 6.12: Suppose that $\varphi$ and $\psi$ are logically equivalent. A simple truth table construction shows that $\{\} \models \sim (\varphi \land \sim \varphi)$. Since $\varphi$ and $\psi$ are logically equivalent, it follows that $\{\} \models \sim (\varphi \land \sim \psi)$. By Finite Additivity, $\Pr(\varphi \lor \sim \psi) = \Pr(\varphi) + \Pr(\sim \psi)$. And by the derived rule Complement, $\Pr(\varphi) + \Pr(\sim \psi) = \Pr(\varphi) + 1 - \Pr(\psi)$. A simple truth table construction shows that $\{\} \models (\varphi \lor \sim \varphi)$. And again, since $\varphi$ and $\psi$ are logically equivalent, it follows that $\{\} \models (\varphi \lor \sim \psi)$. By Normality, $\Pr(\varphi \lor \sim \psi) = 1$. Therefore, $\Pr(\varphi \lor \sim \psi) = \Pr(\varphi) + 1 - \Pr(\psi) = 1$. Some simple arithmetic then shows that $\Pr(\varphi) = \Pr(\psi)$.

Probably the most useful of our derived rules is Complement. The most obvious application of Complement is in finding the probability that some sentence is false or that some event does not occur.

Example 6.13: Suppose we pull one card from a standard 52 card deck. Assume that each card is equally likely to be drawn. What is the probability that the card we pulled is not an ace? Since there are exactly four aces in a standard deck, the probability of drawing an ace is $4/52 = 1/13$. So, by our derived Complement rule, the probability of not drawing an ace is $1 - 1/13 = 12/13$.

However, the most useful application of Complement, which we will see again and again, is in finding the probability of existential sentences like, “At least one head turns up in three tosses of a fair coin.”

Example 6.14: We want to find the probability that at least one head turns up in three tosses of a fair coin. We could try to compute this directly using Finite Additivity by adding up the probabilities for the sentences, “Exactly one head turns up,” “Exactly two heads turn up,” and “Exactly three heads turn up.” However, we can get the answer more efficiently by directly calculating the probability that no heads turn up and then subtracting that probability from one.

Example 6.15: Suppose we pull one card from a standard 52 card deck. Assume that each card is equally likely to be drawn. We know that there are exactly four aces in the deck. What is the probability that the card we pulled is not an ace? Let the set of aces in the deck be denoted by $A = \{<\text{ace, heart}>, <\text{ace, spade}>, <\text{ace, diamond}>, <\text{ace, club}>\}$. The probability of drawing an ace is $\Pr(A) = 4/52 = 1/13$. So, by the complement rule, the probability of not drawing an ace is $\Pr(U - A) = 1 - \Pr(A) = 1 - 1/13 = 12/13$. Notice that this is the same problem and the same answer as in Example 6.13.
Our first set theoretic version of the Finite Additivity constraint is sometimes called the *Special Rule for Unions*, since it describes how to calculate the probability of the union of two events in the special case where those sets are non-overlapping.

**Special Rule for Unions**

Let $\alpha$ and $\beta$ be arbitrary subsets of the universe $U$.

If $(\alpha \cap \beta) = \emptyset$, then $\Pr(\alpha \cup \beta) = \Pr(\alpha) + \Pr(\beta)$.

*Example 6.16:* Suppose we make one toss of a fair six-sided die. We want to know the probability that the die shows either a number strictly greater than four or a number strictly less than three. That is, we want to know the probability of an observed unit being an element of either $E_L = \{1, 2\}$ or $E_G = \{5, 6\}$. Hence, we want to find $\Pr(E_L \cup E_G)$. Since $E_L \cap E_G = \emptyset$, we may apply the special rule for unions to obtain the probability of the event of interest. The probability of $E_L$ is $2/6 = 1/3$, and the probability of $E_G$ is $2/6 = 1/3$, so the probability of $E_L \cup E_G$ is equal to $1/3 + 1/3 = 2/3$.

Similarly, our initial version of Finite Additivity in terms of sentences tells us how to deal with disjunctions under the special condition that the disjuncts cannot both be true. But we often want a more general rule. The derived rule General Add drops the assumption that the disjuncts are inconsistent with each other. The set theoretic analogue of General Add, which we call the **General Rule for Unions**, drops the assumption that the events under consideration are non-overlapping.

**General Rule for Unions**

Let $\alpha$ and $\beta$ be arbitrary subsets of the universe $U$.

$$\Pr(\alpha \cup \beta) = \Pr(\alpha) + \Pr(\beta) - \Pr(\alpha \cap \beta)$$

*Example 6.17:* Suppose we make one toss of a fair six-sided die. We want to know the probability that the die shows either an even number or a number strictly greater than three. That is, we want to know the probability of an observed unit being an element of either $E_E = \{2, 4, 6\}$ or $E_3 = \{4, 5, 6\}$. We could calculate the probability directly by counting the simple events in the union $E_E \cup E_3 = \{2, 4, 5, 6\}$ and obtaining $4/6 = 2/3$ as the probability. However, in many cases, it is easier to use the formula. The simple case here illustrates how the formula works. In this case, $E_E \cap E_3 = \{4, 6\}$, which has probability $2/6 = 1/3$. The individual events $E_E$ and $E_3$ have probabilities $3/6 = 1/2$ and $3/6 = 1/2$. Applying the formula, we get $\Pr(E_E \cup E_3) = 3/6 + 3/6 - 2/6 = 4/6 = 2/3$. 

133
Remember that when the outcome of an experiment or observation is in the intersection of two events, the outcome is an element of both sets.

Up until now, we have been talking about what are called *unconditional probabilities*, which are probability statements of the form \( \Pr(H) \), where \( H \) is an arbitrary sentence in our formal language. Often, however, we want to consider the probability of \( H \) on the assumption that some other sentence \( E \) is true. Call \( E \) the *condition* or the *conditioning sentence*. Sometimes we know (or think that we know) that \( E \) is true, and we want to know the probability of \( H \) given *that knowledge*. And sometimes, we want to know what we should say about the probability of \( H \) if we learn some fact or other. In order to account for such cases, we define a new idea: *conditional probability*. The conditional probability of \( H \) given \( E \) is denoted \( \Pr(H \mid E) \), and it is usually analyzed as follows:

\[
\Pr(H \mid E) = \frac{\Pr(H \land E)}{\Pr(E)}
\]

Following Hájek (2003), we’ll call this the RATIO analysis of conditional probability. Since we are dividing by the probability of the sentence \( E \), the sentence \( E \) must not be a contradiction for the conditional probability to be well-defined. In other words, we are not allowed to divide by zero! We can translate our account of conditional probability into set theory by treating \( H \) and \( E \) as sets and replacing \( \land \) with \( \cap \).

*Example 6.18:* Suppose we pull one card from a standard 52 card deck. Assume that every card is equally likely to be drawn. I look at the card (but you do not), and I tell you that it is a face card. What is the probability that the card is a jack? We know that 12 of the 52 cards are face cards: four kings, four queens, and four jacks. And we know that 4 of the 52 cards are both face cards and jacks. Hence, the probability that the card is a jack given that it is a face card is \( \Pr(\text{Jack} \land \text{Face card}) \) / \( \Pr(\text{Face card}) = \frac{4/52}{12/52} = 4/12 = 1/3 \).
Following the RATIO analysis, we can calculate a conditional probability by calculating two unconditional probabilities and dividing. Alternatively, we can think of the conditioning sentence (or the corresponding set) as determining a new universe of discourse.

*Example 6.19:* Suppose we toss a pair of fair dice. One of the dice is red and one is white. After the toss, you see that the red die is showing three. What is the probability that the sum of the numbers showing on the two dice is greater than six? We can write out all the possible outcomes where the red die shows three and the sum is greater than six as follows (with the first element of each ordered pair representing the red die):

\[ <3, 4>, <3, 5>, <3, 6>. \]

Moreover, we know that there are exactly six tosses where the red die shows three. Hence, the probability that the dice sum to more than six given that the red die shows three is

\[ \frac{3}{36} / \frac{6}{36} = \frac{3}{6} = \frac{1}{2}. \]

We could also obtain the conditional probability value by finding the probability that the sum is greater than six in the universe of discourse determined by the condition that the red die shows three. The restricted universe of discourse has six total elements, exactly three of which represent a sum greater than six.

Notice that in Example 6.19, knowing that the red die came up three actually lowered the probability that the dice summed to a number strictly greater than six. Sometimes, conditioning raises the probability. Sometimes, conditioning lowers the probability. And sometimes, conditioning leaves the probability unchanged!

*Example 6.20:* Suppose we toss a pair of fair dice. Again, one of the dice is red and one is white. What is the probability that the sum of the dice is seven? The outcomes where the sum is seven are these:

\[ <1, 6>, <2, 5>, <3, 4>, <4, 3>, <5, 2>, <6, 1>. \]

And we know that there are 36 possible outcomes altogether. Hence, if each outcome is equally likely (as indicated by the claim that the dice are fair), the probability of a seven is \[ \frac{6}{36} = \frac{1}{6}. \]

Now suppose that after the toss, you see that the red die is showing four. What is the conditional probability that the sum of the dice is seven? There is exactly one outcome where the red die shows four and the sum of the two dice is seven, namely the toss represented by \( <4, 3> \). If each toss is equally likely, then the probability that the red die shows four and the sum of the two dice is seven is equal to \( \frac{1}{36} \). Hence, the conditional probability that the sum is seven given that the red die shows four is equal to \( \frac{1}{36} / (\frac{6}{36}) = \frac{1}{6} \). The unconditional probability that the sum is seven equals the conditional probability that the sum is seven given that the red die shows four.
If the probability of \( H \) given \( E \) is the same as the probability of \( H \), then we say that the two sentences are independent. Formally: Two sentences \( H \) and \( E \) are independent if and only if

\[
Pr(H \mid E) = Pr(H).
\]

Notice that this requirement could equally have been written as \( Pr(E \mid H) = Pr(E) \), although the specific probability values will generally be different in the two formulations. If \( H \) and \( E \) are independent, we write \( H \perp E \). If \( Pr(H \mid E) \neq Pr(H) \), then the sentences \( H \) and \( E \) are associated. If \( Pr(H \mid E) > Pr(H) \), then \( H \) and \( E \) are positively associated. If \( Pr(H \mid E) < Pr(H) \), then \( H \) and \( E \) are negatively associated.

When two sentences are independent, then the definition of conditional probability entails a result that we call the multiplication rule. Applying our notion of independence to our definition of conditional probability, you can see that if two sentences \( H \) and \( E \) are independent, the probability of their conjunction is equal to the product of their unconditional probabilities:

\[
\text{If } H \perp E, \text{ then } Pr(H \land E) = Pr(H) \cdot Pr(E).
\]

For games of chance, repeated experiments are usually assumed to be independent. For example, if I throw a die two times in succession, I usually assume that the outcomes of the two throws are independent of one another. Given the probabilities, if we learn what comes up on the first throw, we do not gain any useful information about what will come up on the second throw. Similarly, learning what comes up on the second throw does not give us any useful information about what came up on the first throw. Whenever we know that two sentences are independent, our calculations are greatly simplified by using the multiplication rule.

*Example 6.21:* Assume that each toss of a fair, six-sided die is an independent event. What is the probability of throwing three sixes in succession? The probability of throwing one six on a single toss of a fair die is \( 1/6 \). Each toss has the same probability, and the tosses are independent. So, the probability of three sixes in succession is \( (1/6) \cdot (1/6) \cdot (1/6) = 1/216 \).
By combining the multiplication rule with the rule for complements, we may solve one of the problems that the Chevalier de Méré (1607-1684) asked Fermat and Pascal to think about:

*Example 6.22*: Assume that each toss of a fair, six-sided die is independent of every other toss. What is the probability of throwing at least one six in four tosses? We will find this probability by first computing the probability that we see no sixes in four tosses and then subtracting that value from one. The probability that you throw something other than a six on any single toss is $\frac{5}{6}$. Since each toss is independent, the probability of throwing no six four times in succession is $\left(\frac{5}{6}\right) \cdot \left(\frac{5}{6}\right) \cdot \left(\frac{5}{6}\right) \cdot \left(\frac{5}{6}\right) = \frac{625}{1296}$. By Complement, the probability of throwing at least one six in four tosses is equal to one minus the probability of throwing no sixes in three tosses. So, the probability value that we want to compute is $1 - \frac{625}{1296} = \frac{671}{1296}$, which is just a bit more than $\frac{1}{2}$.

Remarkably, what we have seen so far—the three Kolmogorov constraints, the equal likelihood assumption, and the RATIO analysis of conditional probability—is a complete account of classical probability theory. With just these tools, we can solve an enormous and varied collection of probability problems. And yet, there is a good sense in which we do not really know what we mean by “probability.”

### 6.3 Interpretations of Probability

Having laid out the basic mathematical machinery of probability, we take a step back in this section to consider what probability claims mean. The problem of interpreting probability theory—and thereby giving meaning to the mathematical machinery—is a philosophical problem, and one might wonder what practical value there is in solving such a problem. In his excellent Stanford Encyclopedia of Philosophy article on “Interpretations of Probability,” Alan Hájek notes that the Kolmogorov axioms say very little about how we should make initial probability assignments, and he remarks that “for guidance with [assigning initial probabilities], we need to turn to the interpretations of probability.” Interpretations of probability help us to assign initial probability values by identifying what it is that probability measurements are supposed to be measuring. In the rest of this section, we will consider three standard
interpretations of probability: frequentist, evidentialist, and personalist. The interpretations we consider do not agree in every instance as to what probability value (if any) ought to be assigned to a given sentence or event. Insofar as this represents a genuine disagreement, at most one of the three interpretations can be correct. So we need to make a choice. In order to illustrate the differences, we will focus on how the various interpretations determine what are sometimes called prior probabilities, i.e. assignments of probability values in the absence of evidence. As we shall see, each interpretation has something to recommend it to us. At the same time, the different interpretations all have weaknesses. So the choice is not obvious. And it is possible that there are so-far unconceived alternatives preferable to the interpretations discussed here.

The frequentist interpretation of probability maintains that the probability of a sentence is a measure of the frequency with which that sentence is made true by some run of experiments (often called a reference class in the context of frequentism). Recall (from Section 6.1) that classical probability theory maintains that every possible outcome is equally likely, and the probability of a sentence is just the number of possible outcomes favorable to the sentence divided by the total number of possible outcomes. On the classical interpretation, the probability of rolling a one on a six-sided die is 1/6 because there are six possible ways for the die to turn up, only one of which is a six. The frequentist interpretation makes probability a matter of what actually happens, not what might possibly happen. The frequentist, when confronted with a six-sided die, will ask how often the number one has turned up in some sequence of throws. The

---

8 Hájek recognizes five main interpretations of probability, and there are other less common interpretations.  
9 Actually, this isn’t quite true in general. Some frequentists maintain that probability has to do with what actually happens. These frequentists are sometimes called actual frequentists or actualists. (And actualists come in two subtypes: finite and infinite.) But some other frequentists identify probability not with any actual sequence of experiments but with an indefinite (infinite) long run of possible experiments. These hypothetical frequentists thus disconnect probability from what actually happens and connect it to what might possibly happen.
relative frequency of ones (successes) out of throws (experiments) is the probability of the number one turning up (success).

Insofar as the prior probability is supposed to be the probability assigned before any evidence is collected, the frequentist will deny that any definite prior probability may be assigned. After all, the quantity that the frequentist would need to evaluate in such a case is 0/0. No number will be assigned to the probability of a sentence before evidence is available. In practice, when a frequentist needs a “prior probability” for a calculation (for example, in order to apply the Bayes’ Theorem that we discuss in the next section) the probability value is not really prior to all evidence, and the frequentist will take the prior (or base rate) from what is known about experiments similar to the one being considered. For example, suppose a frequentist is asked about the probability that a woman has breast cancer given a positive mammogram. The prior probability of breast cancer for the frequentist will be the relative frequency of breast cancer in the target population, and estimates of that frequency will be based on previous observations.

But herein lies a weakness of the frequentist interpretation of probability. What reference class should the frequentist consider when assigning probability values? Suppose Susan comes into the hospital for a mammogram. Susan is a 46-year-old, African-American woman who stands 5’4” tall, weighs 130 lbs. and smokes occasionally. Susan lives in Chicago but was born in New Orleans. She likes coffee with two creams and no sugar, listens to Thelonious Monk, and owns a cat. What reference class should the frequentist consider when assigning the prior probability of breast cancer for Susan? Should the frequentist consider only the fact that Susan is a woman? What about her age or taste in coffee and music? One might be tempted to say that the
most informative universe should be chosen, but then, it will likely be the case that the
frequentist cannot assign any prior probability to the claim that Susan has breast cancer.

The frequentist interpretation focuses on actual outcomes of experiments. By contrast, the
evidentialist interpretation (sometimes called the logical interpretation) focuses on possible
outcomes of experiments and their evidential relationships. The goal of the evidential
interpretation of probability is to treat probability as an extension or generalization of ordinary
logic. Instead of two truth values, the evidentialist admits infinitely many truth values. And
instead of a relation of logical implication, by which the truth of some premisses guarantees the
truth of some conclusion, the evidentialist settles for a relation of confirmation, by which the
truth of some premisses makes the truth of some conclusion more likely. For the evidentialist,
then, probability is a generalized measure of evidential support. It is a measure of how much
truth value one should assign to a claim based on some evidence.

Evidentialists maintain that for any body of evidence, there is a unique probability that
ought to be assigned to any sentence one might consider. A natural reaction, then is to ask what
the evidentialist should say in cases where there is no evidence. Typically, evidentialists appeal
to a symmetry principle in assigning probabilities based on no evidence. The earliest such
symmetry principle is called the principle of indifference or sometimes the principle of
insufficient reason. The idea is that if you have no reason to think that one or another outcome of
an experiment is more likely than any other, then you should assign equal probability to all
possible outcomes. If your evidence is symmetric with respect to a collection of sentences, then
your probability assignments should be similarly symmetric with respect to those sentences.
According to the principle of indifference, if an experiment, like tossing a six-sided die, has six
possible outcomes, then the probability of each outcome is 1/6. However, appealing to symmetry
principles is widely regarded as unworkable in light of examples like the following. Suppose you
are the owner and operator of a ball bearing factory. You manufacture solid steel spheres having
a radius between 1 cm and 3 cm, inclusive. Suppose I visit your factory and observe a newly
made ball bearing. What is the probability that the bearing has a radius less than 2 cm? The
natural answer is ½, since the interval [1, 2] stands to the interval [1, 3] in the ratio 1:2. But now
consider. The surface area of a bearing made in your factory is between $4\pi \, \text{cm}^2$ and $36\pi \, \text{cm}^2$.
What is the probability that the surface of the newly made bearing that I observed is less than
$16\pi \, \text{cm}^2$? Since the interval $[4\pi, 16\pi]$ stands to the interval $[4\pi, 36\pi]$ in the ratio 3:8, by the
same reasoning as before, we should say that the probability is $\frac{3}{8}$ that the ball bearing I observed
has surface area less than $16\pi \, \text{cm}^2$. But notice that the surface area of a ball bearing is less than
$16\pi \, \text{cm}^2$ if and only if its radius is less than 2 cm. Hence, by perfectly natural reasoning, we end
up assigning different probabilities to two logically equivalent sentences. We have already seen
(in Example 6.12) that it follows from the Kolmogorov axioms that two logically equivalent
sentences must be assigned the same probability. So, our natural reasoning here violates the
Kolmogorov axioms.\(^{10}\)

One response to cases like the ball bearing factory is to reject the evidentialist assumption
that in the absence of evidence there is only one rationally permissible assignment of probability
for any given sentence. Rejecting the evidentialist requirement leads naturally to the personalist
interpretation of probability (sometimes called the subjectivist interpretation), which begins with
the idea that probability is a measure of one’s personal ignorance or a measure of one’s personal
degrees of belief with respect to a given sentence. In order to distinguish what they mean by

\(^{10}\)Evidentialists like E.T. Jaynes have made sophisticated attempts to avoid problems like the ball bearing factory by
using the mathematical machinery of transformation groups. Challenges have been raised against such attempts to
save the objectivity of assignments of prior probabilities. For considerably more detail on the philosophical issues,
“probability” from what a frequentist means by the same word, personalists often distinguish between \textit{credences} and \textit{chances}. The chance of some event (if there is any such thing at all) is an objective feature of the world. The credence that one assigns to a sentence is the degree of belief one has that the sentence is true. Personalists usually assume that one’s degrees of belief are closely connected with the actions that one would be willing to take in various circumstances. For example, unless you have a death wish, if you really believe that the building is on fire, then you will try to get out of the building! If there is a tight connection between beliefs and actions, personalists argue, then credences may be measured through behavior. Typically, personalists appeal specifically to \textit{betting} behavior in order to represent an epistemic agent’s credences.

The most minimalist of personalists think that the only thing that matters is the actual degree of belief that an agent has at a given time. However, as Hájek points out, most people fail to have degrees of belief that satisfy the axioms of probability. In response to this point, personalists usually say that they are interested in \textit{rational} degrees of belief and claim that an agent’s credences must satisfy some specific constraints in order to count as rational. The first constraint usually advanced by personalists is that credences must satisfy the probability axioms. The second constraint is that one must have a rule for updating degrees of belief given new evidence. Often, Bayes’ rule (which we will discuss in the next section) is endorsed as the correct rule for updating one’s degrees of belief. Some personalists add to these constraints a requirement that degrees of belief be \textit{calibrated} to relative frequencies, but there is considerable debate about how calibration is supposed to work for a personalist.

Going forward, let us assume that we are talking about a personalist who thinks that rational credences must satisfy the axioms of probability. When asked to assign prior probabilities to some sentence, the personalist may give whatever answers he or she wants to
give consistent with the axioms of probability. One might think that this by itself is a *reductio* of the personalist interpretation. After all, if the personalist may assign any probability at all to a sentence, what could the connection be between probabilities and occurrences in the world? What would be the point of probability? Personalists who endorse an update rule often reply by citing one or another theorem showing that if two people share an update rule, like Bayes’ rule, and never assign any contingent sentence a prior probability equal to zero or one, then given enough evidence, those people will come to have arbitrarily similar degrees of belief.

6.4 *Bayes’ Rule, Bayes’ Theorem, and the Law of Total Probability*

With the assumption of equal likelihood, the Kolmogorov axioms, the definition of conditional probability, and the multiplication rule, we can solve a large range of probability problems. However, what we can do with the probability theory we have so far developed is limited in an important way: the probability values we have calculated so far are static, representing what is the case at a specific time and for an unchanging evidentiary base. In order to solve dynamic problems, we need a new device: Bayes’ rule. But before describing Bayes’ rule, I would like to provide a little philosophical context. In 1739, the Scottish philosopher David Hume published *A Treatise of Human Nature*. In the first part of his *Treatise*, Hume raised skeptical worries about the possibility of causal knowledge. Hume argued that knowledge of cause and effect relations cannot be obtained by reasoning without experience. But he went even further, arguing that no amount of experience gives any rational support to a causal inference. Restating the problem in *An Enquiry Concerning Human Understanding* in 1748, Hume wrote:

> As to past Experience, it can be allowed to give direct and certain information of those precise objects only, and that precise period of time, which fell under its cognizance: but why this experience should be extended to future times, and to other objects, which for aught we know, may be only in appearance similar; this is the main question on which I would insist. The bread, which I formerly eat, nourished me; that is, a body of such
sensible qualities was, at the time, endued with such secret powers: but does it follow, that other bread must also nourish me at another time, and that like sensible qualities must always be attended with like secret powers? The consequence seems nowise necessary. At least, it must be acknowledged that there is here a consequence drawn by the mind; that there is a certain step taken; a process of thought, and an inference, which wants to be explained. These two propositions are far from being the same, I have found that such an object has always been attended with such an effect, and I foresee, that other objects, which are, in appearance, similar, will be attended with similar effects. I shall allow, if you please, that the one proposition may justly be inferred from the other: I know, in fact, that it always is inferred. But if you insist that the inference is made by a chain of reasoning, I desire you to produce that reasoning. (23-24)

On the very next page, Hume again calls for the reader to exhibit the reasoning that would provide rational justification for causal inferences from experience.

It is only after a long course of uniform experiments in any kind, that we attain a firm reliance and security with regard to a particular event. Now where is that process of reasoning which, from one instance, draws a conclusion, so different from that which it infers from a hundred instances that are nowise different from that single one? … I cannot find, I cannot imagine any such reasoning. (25)

Hume could not imagine reasoning that would solve his problem. But some of his contemporaries were cleverer. Thomas Bayes may have provided a solution to Hume’s problem as early as 1749.11

Bayes’ proposal comes in two parts. The first is a simple dynamical rule for updating probability assignments on the basis of observational evidence. According to Bayes’ rule, the probability that one ought to assign to a sentence \( H \) after learning that some sentence \( E \) is true is just the conditional probability of \( H \) given \( E \). That is, the new unconditional probability assignment for \( H \) should be equal to the conditional probability of \( H \) given \( E \) that one assigned before learning that \( E \) is true.

\[
\text{[Bayes’ Rule]} \quad \Pr_{\text{New}}(H) = \Pr_{\text{Old}}(H \mid E)
\]

11 Bayes never published his result during his lifetime. It was published posthumously in 1764 by his friend Richard Price. However, Zabell (1989) provides a compelling argument—based on a passage in David Hartley’s 1749 work, *Observations on Man*—that Bayes proved his result sometime in the 1740s.
Sometimes the rule is written as $\Pr_E(H) = \Pr(H \mid E)$ to denote the fact that the probability assignment on the left-hand side of the equation is a result of learning that $E$ is true. If you apply Bayes’ rule to update your probability assignment for some sentence $H$ on the basis of some evidence $E$, you are said to be *conditioning* or *conditioning* on your evidence $E$.

The application of Bayes’ rule to Hume’s problem is simple. Hume wants to know how we could come to believe that some piece of bread, which we have not yet eaten, has a secret power to nourish us given only some observation of its qualities: its color and shape, its smell, the way it feels to the touch, etc. Bayes’ rule tells us that the probability we should assign to the claim that the bread nourishes given that we observe some specified qualities of the bread is just the conditional probability that the bread nourishes given that the bread has those specified qualities. One might worry that Bayes’ rule isn’t much of a solution because while it says in a formal language that one ought to assign some probability to the claim that the bread nourishes given that some qualities are observed, it doesn’t tell us how to compute the required conditional probability. Hence, the *process of reasoning* that Hume wanted to understand is left mysterious.

The second part of Bayes’ solution to Hume’s problem, called *Bayes’ theorem*, provides a way of computing the conditional probability of $H$ given $E$ indirectly through probability assignments that are plausibly more accessible to the agent than is the conditional probability of $H$ given $E$ itself. In its simplest form, Bayes’ theorem may be stated as follows, where $H$ and $E$ are arbitrary sentences:

$$
\Pr(H \mid E) = \frac{\Pr(E \mid H) \cdot \Pr(H)}{\Pr(E)}
$$

The conditional probability of $H$ given $E$ in the formula is sometimes called the *posterior* probability of $H$. And the unconditional probability of $H$ is called the *prior* probability of $H$. To repeat a bit: We can think of $H$ as a hypothesis and $E$ as some evidence. In this way, we can see...
how Bayes’ rule and Bayes’ theorem might serve to answer Hume’s problem. Bayes’ rule—together with Bayes’ theorem—tells us how to update our probability assignment for a hypothesis (e.g. that a given object has an unobserved causal power) given some evidence (e.g. that the object has some observed qualities, which have been associated with such a causal power in the past).

Bayes’ theorem follows from two simple facts: the commutativity of conjunction and the definition of conditional probability. By the definition of conditional probability,

\[
Pr(H \mid E) = \frac{Pr(H \land E)}{Pr(E)}
\]

By the commutativity of conjunction (or of set intersection), \((H \land E) = (E \land H)\), which means that

\[
Pr(H \mid E) = \frac{Pr(E \land H)}{Pr(E)}
\]

Again, by the definition of conditional probability:

\[
Pr(E \mid H) = \frac{Pr(E \land H)}{Pr(H)}
\]

Multiplying by \(Pr(H)\), we see that \(Pr(E \land H) = Pr(E \mid H) \cdot Pr(H)\). Substituting into the equation for the conditional probability of \(H\) given \(E\), we have Bayes’ theorem:

\[
Pr(H \mid E) = \frac{Pr(E \mid H) \cdot Pr(H)}{Pr(E)}
\]

Now, let’s see how Bayes’ theorem is applied in a simple case.

*Example 6.23:* Suppose I have a big bucket of marbles. Some of the marbles are blue, and some of the marbles have a white swirl pattern. (And it is possible that some are blue and have a white swirl.) I pick out a marble at random and tell you that it has a white swirl. Suppose the probability of drawing a blue marble is 1/5, the probability of drawing a marble with a white swirl is 1/10, and the conditional probability that a marble has a white swirl *given* that it is blue is 1/2. What is the probability that the marble I drew is
blue? What does this tell you about the relationship between blue marbles and marbles with white swirls?

To get the answer to the first question, apply Bayes’ theorem. Let \( B \) stand for the sentence, “The marble I picked is blue,” and let \( W \) stand for the sentence, “The marble I picked has a white swirl.” Then, by Bayes’ theorem,

\[
\Pr(B \mid W) = \frac{\Pr(W \mid B) \cdot \Pr(B)}{\Pr(W)} = \frac{1/2 \cdot 1/5}{1/10} = 1.
\]

Assuming that I have a finite number of marbles in my bucket, every marble with a white swirl on it is blue!

In many applications, we do not directly know the unconditional probability of the conditioning sentence that appears in Bayes’ theorem. However, we often know enough to derive the unconditional probability of the evidence by using something called the law of total probability.

In order to understand the law of total probability, it will be helpful to have a bit of additional technical machinery in place. Specifically, we need to define a partition. A partition is a collection of sentences \( \varphi_1, \ldots, \varphi_n \) such that the following two conditions are satisfied:

1. For all \( 1 \leq i, j \leq n \), such that \( i \neq j \), \( \{ \} \models \neg (\varphi_i \land \varphi_j) \), and
2. \( \{ \} \models (\varphi_1 \lor \ldots \lor \varphi_n) \).

The sentences \( \varphi_1, \ldots, \varphi_n \) are said to partition logical space. We can also think about partitions in terms of set theory. Imagine all of logical space as being a single set. Then a partition of logical space is a collection of non-overlapping sets that completely cover it. In fact, we can partition any set. To illustrate, suppose we have a set \( S = \{1, 2, 3, 4, 5\} \). The collection of sets consisting of \( \{1, 2\} \), \( \{3\} \), and \( \{4, 5\} \) is a partition of \( S \). The pair of sets \( \{1, 2, 3, 4\} \) and \( \{5\} \) is also a partition of \( S \). However, the pair of sets \( \{1, 2\} \) and \( \{4, 5\} \) is not a partition of \( S \), since they do not completely cover \( S \). And the pair of sets \( \{1, 2, 3\} \) and \( \{3, 4, 5\} \) is not a partition of \( S \), since they overlap.

When we say that a set of sentences partitions logical space, we mean that for each world in logical space, exactly one of the sentences in the partition is true. Consequently, the set of all
the possible worlds may be divided up into a number of subsets equal to the number of sentences
in a partition in such a way that for every subset, there exists a sentence in the partition such that
every world in that subset is a model of the sentence. (As an exercise for the reader, think about
how that last claim could be translated into first-order logic.) The result might look something
like the picture in Figure 6.1 below, where the whole rectangle represents logical space.

Figure 6.1: A Partition of Logical Space

In Figure 6.1, logical space is partitioned by the collection of five sentences \{A, B, C, D, E\}. Each world in the space is a model of exactly one of those five sentences.

The law of total probability states that if the sentences \(\phi_1, \ldots, \phi_n\) form a partition of
logical space, then the following equation holds for any arbitrary sentence \(\psi\):

\[
\Pr(\psi) = \sum_{i=1}^{n} \Pr(\psi \mid \phi_i) \cdot \Pr(\phi_i) = \Pr(\psi \mid \phi_1) \cdot \Pr(\phi_1) + \ldots + \Pr(\psi \mid \phi_n) \cdot \Pr(\phi_n)
\]

In the special case that the partitioning sentences are a pair of the form \(\phi\) and \(\sim \phi\), the law of
total probability states that for any sentence \(\psi\):

\[
\Pr(\psi) = \Pr(\psi \mid \phi) \cdot \Pr(\phi) + \Pr(\psi \mid \sim \phi) \cdot \Pr(\sim \phi)
\]
Pr(ψ) = Pr(ψ | φ)·Pr(φ) + Pr(ψ | ~φ)·Pr(~φ)

In words: The law of total probability lets us calculate the unconditional probability of an event from (1) the probabilities of the event conditional on each of the sentences in a partition and (2) the unconditional probabilities of each of the sentences in the partition.

Example 6.24: Suppose I have a fair coin, two different spinner wheels, and a bell. The first spinner has three equally likely outcomes: red, blue, and green. The second spinner wheel has five equally likely outcomes: 1, 2, 3, 4, and 5. I have decided to ring the bell according to the following rule. First, I flip the coin. If it is heads, I spin the first spinner. If it is tails, I spin the second spinner. If the first spinner turns up red, then I ring the bell. If the second spinner turns up an odd number, then I ring the bell. Otherwise, I do not ring the bell. On any flip of the coin, what is the probability that I ring the bell?

Let A be the sentence, “I ring the bell.” Let B be the sentence, “I spin the first spinner,” and let ~B be the sentence, “I spin the second spinner.” By the law of total probability, Pr(A) = Pr(A | B)·Pr(B) + Pr(A | ~B)·Pr(~B). Since the coin that determines which spinner I spin is fair, Pr(B) = Pr(~B) = 1/2. Since for each spinner, all of the outcomes are equally likely, Pr(A | B) = 1/3 and Pr(A | ~B) = 3/5.

Hence, Pr(A) = (1/3)(1/2) + (3/5)(1/2) = 14/30 = 7/15.

Example 6.25: In example 6.24, the coin is fair. Suppose that I replace it with a coin biased towards heads so that heads has a probability of 4/5 on any toss. Everything else in the setup of Example 6.24 is left alone. On any flip of the biased coin, what is the probability that I ring the bell?

The only things that change with the biased coin are the unconditional probabilities that each spinner is chosen. That is, instead of Pr(B) = Pr(~B) = 1/2, we now have Pr(B) = 4/5 and Pr(~B) = 1/5. Hence, Pr(A) = (1/3)(4/5) + (3/5)(1/5) = 29/75.

The law of total probability is very powerful, but why does it hold? I will illustrate why the law of total probability works in the special case of a two-sentence partition using the picture in Figure 6.2 below.
In Figure 6.2, the black oval in the center of the diagram represents the collection of models of the sentence \( A \), the red-bordered area on the left half of the rectangle represents the collection of models of the sentence \( B \), and the blue-bordered area on the right half of the rectangle represents the collection of the models of the sentence \( \sim B \). The probability of the sentence \( A \) is equal to the probability of the models that it shares with \( B \) plus the probability of the models that it shares with \( \sim B \). We can see from the figure that \( \Pr(A) = \Pr(A \land B) + \Pr(A \land \sim B) \). From the definition of conditional probability, the probability of the conjunction of two sentences is equal to the product of the conditional probability of the first given the second and the unconditional probability of the second. In the case under consideration, \( \Pr(A \land B) = \Pr(A \mid B) \cdot \Pr(B) \).

Similarly, \( \Pr(A \land \sim B) = \Pr(A \mid \sim B) \cdot \Pr(\sim B) \). By replacing the probability terms on the right-hand side of the sum above with the expressions we just obtained from the definition of conditional probability, we get the law of total probability. The figure and discussion are not a proof of the law of total probability, but they do suggest a way of proving the law. I leave the proof as an exercise to the reader.
The law of total probability lets us fill in the denominator on the right-hand side of Bayes’ theorem without being able to evaluate the probability of our evidence directly. As with Bayes’ theorem itself, we will often have better (or at least more direct) epistemic access to the probabilities on the right-hand side of the equal sign than we will to the probability on the left-hand side.

Example 6.26: Acme Corporation makes catapults that it sells to people (or coyotes) hunting road runners. However, some of the catapults are faulty and inevitably cause harm to their users. Luckily, Acme has a simple test of the quality of its catapults that is 90% accurate. Suppose that Acme manufactures 10,000 catapults. Among the catapults, 1000 are faulty. Among the faulty catapults, 900 test as faulty and 100 do not. Among the properly built catapults, 900 test as faulty and 8100 do not. If Wile E. Coyote’s catapult tests out as faulty, what is the probability that it actually is faulty?

The unconditional probability that the catapult is faulty is 1000 / 10000. The conditional probability that the catapult tests as faulty, given that it really is faulty is 900 / 1000. The conditional probability that the catapult tests as faulty given that it is not faulty is 900 / 9000. Hence, the probability that Mr. Coyote’s catapult is faulty given the test result may be calculated using Bayes’ theorem as follows:

\[ \Pr(\text{faulty} \mid \text{test}) = \frac{(1000/10000) \cdot (900/1000)}{(1000/10000) \cdot (900/1000) + (9000/10000) \cdot (900/9000)} \]

\[ = 1/2 \]

Hence, we should expect that a catapult is faulty half the time when Acme’s test says that it is faulty.

As several psychological experiments have shown, humans tend to think that the probability of some event given a positive test for the occurrence of that event is equal to the reliability of the test. But Bayes’ theorem shows us what we should have known: that in order to correctly measure the value of some evidence, we have to take into consideration all of our evidence, not just the most recent evidence. Ignoring the prior probability (or base rate) is an error sometimes called base rate neglect or the base rate fallacy.

12 http://www.youtube.com/watch?v=5aCgSwmm5Ho In this case, it seems that Acme has sub-contracted the work with the Road-Runner Manufacturing Company, which is unsurprising, really.
13 See Kahneman and Tversky (1973) for an early example of this sort of research.
6.5 Confirmation and Evidential Favoring

At the beginning of this book, we described logic as the normative study of reasoning. Logic is the study of what makes an argument good or bad. Hence, a central concern in logic is describing how some sentences are evidentially related to one another. If we treat the premisses of an argument as being (or describing) our evidence with respect to some claim (the conclusion of the argument), then we could say that the general problem of logic is to characterize the quality of our evidence. In the deductive setting, the truth of the premisses of an argument guarantee the truth of the conclusion. But in the ampliative (inductive) setting, the conclusion of an argument having true premisses might be false. Hence, our strategy is to replace implication with a weaker evidential support relation: confirmation.

We will make the confirmation relation formally precise using conditional probability. In fact, for us conditional probability is all there is to the general notion of evidential support. The idea is to evaluate the goodness (or badness) of an argument by taking the probability of the conclusion conditional on the premisses of the argument.

Example 6.27: Suppose we think that each face of an ordinary six-sided die is equally likely to turn up on a single throw. We are told that after being thrown, the die turned up a number greater than three. We then argue as follows:

\[
\begin{align*}
\text{The die turned up a number greater than three.} \\
\hline
\text{The die turned up the number six.}
\end{align*}
\]

Let “\(x = 6\)” denote the sentence, “The die turned up the number six.” And let “\(x > 3\)” denote the sentence, “The die turned up a number greater than three.” Then the evidential support that the premiss gives to the conclusion may be described quantitatively by taking the conditional probability. Given the equal likelihood assumption and the ratio definition of conditional probability, we have \(\Pr(x = 6 \mid x > 3) = 1/3\).

Example 6.28: Suppose we know that 82\% of in-state students at UIUC are from Chicago. And we believe that Becky is randomly chosen from among the in-state students at UIUC. We then argue as follows:
82% of in-state students at UIUC are from Chicago.
Becky is an in-state student at UIUC.

Becky is from Chicago.

Since we think that Becky is chosen at random, we may apply the equal likelihood assumption to the population from which Becky was drawn. Hence, the evidential support conferred by the premisses here is 0.82.

Conditional probability gives us an interesting and useful account of evidential support. But there are some further wrinkles that need to be ironed out. To see what some of the issues are, let’s consider two distinctions that one may draw when talking about confirmation.

The first distinction is between absolute confirmation and incremental confirmation. In some cases, when we say that a claim has been confirmed (or disconfirmed) by some evidence, we mean that the evidence makes the claim more (or less) likely to be true. In such cases, we are using “confirmation” in its incremental sense. Formally, we say that some evidence $E$ incrementally confirms some hypothesis $H$ if and only if $\Pr(H | E) > \Pr(H)$. If $\Pr(H | E)$ is less than $\Pr(H)$, then we say that $E$ incrementally disconfirms $H$. And if the two are equivalent, then $E$ is confirmationally or evidentially irrelevant to $H$. (Notice that incremental confirmation tracks association, which we first mentioned in Section 6.3.)

In other cases, when we say that a claim has been confirmed by some evidence, we mean that the evidence makes the claim sufficiently likely that it is beyond reasonable dispute. In such cases, we are using the term “confirmation” in an absolute sense. Formally, we will say that $E$ absolutely confirms $H$ if and only if $\Pr(H | E) > \Pr(H)$ and $\Pr(H | E) > k$, for some suitably large value of $k$. In other words, some evidence absolutely confirms some hypothesis if and only if the evidence incrementally confirms the hypothesis and the posterior probability of the hypothesis

---

14 Assessing the precise, quantitative degree to which some premisses evidentially support some conclusion—an idea distinct from conditional probability—requires us to take into account the extent to which the conclusion is evidentially supported when one has no evidence. We will come back to this topic shortly.
conditional on the evidence is sufficiently large. Figuring out exactly where to draw the cutoff point is (I think) an intractable problem, like figuring out exactly how many grains of sand one has to have in order to have a heap. However, we might be able to make progress by sticking to easy cases: where $\Pr(H \mid E) > \Pr(H)$ and even for very large values of $k$, $\Pr(H \mid E) > k$.

While posterior probability might be a good measure of evidential support, both incremental confirmation and absolute confirmation appear to be distinct from posterior probability. For starters, some hypothesis might have a high posterior probability given a piece of evidence that incrementally disconfirms it, if its prior probability (or its probability given the rest of one’s evidence) is high enough. So, posterior probability is not a measure of incremental confirmation. Moreover, it would be a mistake to say that disconfirming evidence absolutely confirms any hypothesis that had a high enough prior probability. Hence, posterior probability is not a measure of absolute confirmation either. (One might try a trick of language here and say that the hypothesis is absolutely confirmed in light of one’s total evidence. But doing so does not help to clarify how individual pieces of one’s evidence are related to the hypothesis.) The problem here seems to be that when we say that a piece of evidence confirms (or disconfirms) some hypothesis, we are talking about how the probability of the hypothesis changes in light of the new evidence. Plausibly, posterior probability represents the rational credence one should have given one’s total evidence and one’s starting point. But posterior probability does not (on its own) indicate the degree to which a single piece of evidence supports or undermines a hypothesis.

The second distinction is between subjective confirmation and objective confirmation. Roughly, the distinction here is between the degree of support that some sentence would confer on some hypothesis for an agent—possibly an ideal agent—with given initial credences and the
degree of support that some sentence would confer for an ideal agent whose initial credences represent a state of total ignorance. Making sense of objective confirmation in this sense requires solving the problem of how to objectively assign genuine, uninformative prior probabilities for some collection of sentences, which brings us back to the ball bearing factory from Section 6.3 and similar challenges to the principle of indifference.

By contrast with objective confirmation, subjective confirmation depends on the background knowledge that an agent brings to a problem. When we are thinking about objective confirmation, we want the same piece of evidence to have the same value—or provide the same degree of support—with respect to a given claim for every agent. But when we are thinking about subjective confirmation, we (plausibly) do not want the same piece of evidence to have the same value for every agent. Suppose NASA scientists report careful measurements at many points on the surface of Mars between 1992 and 2012, indicating that the global mean temperature on Mars fell slightly during that time. Meanwhile, climate scientists tell us that the global mean temperature on Earth rose during the same period. For an agent who knows nothing else about climate science, this would be strong evidence for human-caused global warming. The impact on such an agent’s credence should be very large if she is being rational. But for an agent who knows a lot about climate science, the new evidence would only slightly confirm the hypothesis of anthropogenic global warming. Since the agent already knows a lot about climate science and hence has excellent reason to believe in anthropogenic global warming, the new piece of evidence won’t change her beliefs very much.

Can we say anything more definite about the degree to which some evidence confirms a hypothesis? One might plausibly want a quantitative measure of the degree of support that a given piece of evidence provides to a given hypothesis. We will say that a confirmation measure
is a function having two inputs and one output, where the inputs are a hypothesis and some
purported evidence, and the output is a number representing the degree to which the hypothesis
is confirmed (or disconfirmed) by the evidence. Let \( c \) be the confirmation measure (or function).
Plausibly, the measure \( c \) should satisfy the following basic requirements, which follow from our
qualitative remarks above:

1. If \( \Pr(H | E) > \Pr(H) \), then \( c(H, E) > 0 \);
2. If \( \Pr(H | E) < \Pr(H) \), then \( c(H, E) < 0 \); and
3. If \( \Pr(H | E) = \Pr(H) \), then \( c(H, E) = 0 \).

Perhaps the most obvious and natural proposal for the exact form of the confirmation measure is
to take \( c \) to be the difference between the posterior probability of \( H \) given the evidence and the
prior probability of \( H \). Formally, define \( c(H, E) = \Pr(H | E) - \Pr(H) \). The difference measure
satisfies requirements (1) through (3). However, Fitelson (2001) points out that there are many
functions that satisfy requirements (1) through (3). That much is, perhaps, unsurprising.
However, even prima facie plausible alternative confirmation measures do not agree (in general)
with respect to which of several hypotheses is confirmed the most by a given piece of evidence.
And that really is surprising.

In light of the various challenges to confirmation theory, only some of which we have
touched on in this chapter, one might try to formulate some alternative way of judging the value
of evidence with respect to various theories. Likelihoodism is one such alternative. Following
Edwards (1972, 9), we will say that the likelihood of a hypothesis \( H \) given some evidence \( E \),
denoted \( \mathcal{L}(H | E) \), is proportional to the conditional probability of \( E \) given \( H \), with the constant
of proportionality being arbitrary. Some writers, such as Sober (2008, 9), identify the likelihood
of \( H \) given \( E \) with the conditional probability of \( E \) given \( H \). However, it is best to distinguish the
two. The likelihood should be thought of as a function of the hypothesis \( H \); whereas, the probability is a function of the evidence \( E \). The Likelihood Principle maintains that the empirical significance of some evidence \( E \) with respect to some hypothesis \( H \) is exhausted by the conditional probability \( \Pr(E \mid H) \). So far, this is consistent with a Bayesian approach. Recall the basic form of Bayes’ theorem:

\[
\Pr(H \mid E) = \frac{\Pr(E \mid H) \cdot \Pr(H)}{\Pr(E)}
\]

Notice that the only way the evidence really matters to what a Bayesian ultimately believes is by way of the conditional probability \( \Pr(E \mid H) \). (From the likelihoodist perspective, the prior probability of the evidence is just a proportionality constant.)

But while likelihoodists accept the Likelihood Principle, they reject the full Bayesian framework in two ways. First, they put the focus on evidential favoring, which is a comparative idea, rather than on confirmation, which could be applied to a single hypothesis on its own. Hence, likelihoodists think that evidence figures in science by way of theory selection. Second, likelihoodists reject the claim that prior probabilities have an important role to play in theory selection. Instead, they argue that we can and should talk about evidential favoring without ever mentioning prior probabilities. Rather than trying to say to what degree some hypothesis is confirmed by some evidence, the likelihoodist aims simply to say whether some evidence favors one hypothesis \( H_1 \) over a rival \( H_2 \) and to what degree. Formally, the likelihoodist says that evidence \( E \) favors \( H_1 \) over a rival \( H_2 \) if and only if \( \Pr(E \mid H_1) > \Pr(E \mid H_2) \). The likelihoodist’s thesis about evidential favoring is sometimes called the law of likelihood. The likelihood ratio,

\[
\lambda = \frac{\Pr(E \mid H_1)}{\Pr(E \mid H_2)}
\]

provides a quantitative measure of the degree to which evidence \( E \) favors hypothesis \( H_1 \) over hypothesis \( H_2 \). If \( \lambda \) is greater than one, then the evidence favors \( H_1 \) over \( H_2 \) to
If \( \lambda \) is less than one, then the evidence favors \( H_2 \), and if \( \lambda \) is equal to one, then the evidence favors neither hypothesis.

6.6 Probability in the Object Language

So far, we have been treating probability as part of the meta-language. The probability function is analogous to the valuation function and applies to sentences of the object language. But we could also add probability to our logical language. In order to do so, we need to do two things. First, we need to introduce a new operator: the probability function. The probability function maps sets into real numbers. As before, a lower-case letter stands for an individual constant or an individual variable. Lower-case letters are terms in our logical language, and set operations on terms are themselves terms in our language. If \( \alpha \) is an arbitrary term, then \( \Pr(\alpha) \) is also a term, called the probability of \( \alpha \). The probability of a term is itself a term in our logical language. In order to get a sentence involving the probability function, we need to have some predicate or relation. We will use the standard arithmetical relations \(<, \leq, >, \geq\), and =. We also introduce ordinary arithmetical symbols, and although we will not formalize arithmetic (though we could), we will allow standard arithmetic operations in proofs with the justification “arithmetic.” Hence, \( \Pr(a) > \Pr(b), \Pr(a) = \Pr(b), \) and \( \Pr(a) + \Pr(b) = \Pr(c) \) are all examples of well-formed formulas in our language.\(^\text{15}\) And as usual, we may quantify over the constants in such formulas to get formulas like \((\forall x)(\exists y)(\Pr(x) = \Pr(y))\).

The second thing we need to do in order to add probability to our logical language is formalize our constraints on probability assignments. We will formalize the constraints using three axioms of probability. The three axioms are called non-negativity, normality, and finite

\(^{15}\) Since these formulas all involve constants, each of these formulas is also a sentence, which gets a truth-value.
additivity. Each axiom may be written on any line of a proof with no assumptions. The justification for writing it down is the name of the axiom (non-neg, norm, or fin-add). Below are some examples of the use of the axioms in proofs.

1. **Non-Negativity**: \((\forall x)(\text{Pr}(x) \geq 0)\)

   *Example 6.29*: Show that \(\{\} \vdash \text{Pr}(a) + \text{Pr}(b) \geq 0\)

   \[
   \begin{align*}
   (1) & \quad (\forall x)(\text{Pr}(x) \geq 0) \quad \text{non-neg} \\
   (2) & \quad \text{Pr}(a) \geq 0 \quad 1 \forall E \\
   (3) & \quad \text{Pr}(b) \geq 0 \quad 1 \forall E \\
   (4) & \quad \text{Pr}(a) + \text{Pr}(b) \geq 0 \quad 2, 3 \text{ arith}
   \end{align*}
   \]

2. **Normalization**: \(\text{Pr}(U) = 1\)

   *Example 6.30*: Show that \(\{ (U = (a \cup b)) \} \vdash \text{Pr}(a \cup b) = 1\).

   \[
   \begin{array}{c|c}
   \hline
   1 & \begin{align*}
   (1) & \quad U = (a \cup b) \quad \text{A (premiss)} \\
   (2) & \quad \text{Pr}(U) = 1 \quad \text{norm} \\
   (3) & \quad \text{Pr}(a \cup b) = 1 \quad 1, 2 =E
   \end{align*} \\
   \hline
   \end{array}
   \]

3. **Finite Additivity**: \((\forall x)(\forall y)(((x \cap y) = \emptyset) \to (\text{Pr}(x \cup y) = \text{Pr}(x) + \text{Pr}(y)))\)

   *Example 6.31*: Show that \(\{(U \cap \emptyset) = \emptyset), ((U \cup \emptyset) = U) \} \vdash \text{Pr}(\emptyset) = 0\).

   \[
   \begin{array}{c|c}
   \hline
   1 & \begin{align*}
   (1) & \quad (U \cap \emptyset) = \emptyset \quad \text{A} \\
   (2) & \quad (U \cup \emptyset) = U \quad \text{A} \\
   (3) & \quad (\forall x)(\forall y)(((x \cap y) = \emptyset) \to (\text{Pr}(x \cup y) = \text{Pr}(x) + \text{Pr}(y))) \quad \text{fin-add} \\
   (4) & \quad (\forall y)(((U \cap y) = \emptyset) \to (\text{Pr}(U \cup y) = \text{Pr}(U) + \text{Pr}(y))) \quad 3 \forall E \\
   (5) & \quad ((U \cap \emptyset) = \emptyset) \to (\text{Pr}(U \cup \emptyset) = \text{Pr}(U) + \text{Pr}(\emptyset)) \quad 4 \forall E \\
   (6) & \quad \text{Pr}(U \cup \emptyset) = \text{Pr}(U) + \text{Pr}(\emptyset) \quad 1, 5 \to E \\
   (7) & \quad \text{Pr}(U) = \text{Pr}(U) + \text{Pr}(\emptyset) \quad 1, 2 \text{ E} \\
   (8) & \quad \text{Pr}(U) = 1 \quad \text{norm} \\
   (9) & \quad 1 = 1 + \text{Pr}(\emptyset) \quad 7, 8 =E \\
   (10) & \quad \text{Pr}(\emptyset) = 0 \quad 9 \text{ arith}
   \end{align*} \\
   \hline
   \end{array}
   \]
In most cases, it would be a waste of time to try to rigorously solve probability problems by writing down a proof in our formal system. However, it is worth seeing that probability problems may indeed be solved in this way, at least in principle.
References


Hume, D. (1748) *An Enquiry Concerning Human Understanding*.


In the previous chapter, we discussed some basic ideas in probability theory. In this chapter, we expand our discussion of probability theory and apply it to make inferences about quantitative features of real-world populations. The line separating probability theory from statistics is fuzzy. One way to separate probability theory from statistics is to think of probability theory as an uninterpreted syntactic system and then think of statistics as beginning with a semantics for probability theory. From that perspective, we started thinking about statistics in the middle of the previous chapter, when we discussed possible interpretations of probability. A different way to separate probability theory from statistics is to think of probability as providing a machinery for solving a variety of learning problems and to think of statistics as applying that machinery in specific cases involving inference from a sample to a population. In this chapter, we are going to treat statistics as an application of probability theory to solve a class of learning problems. For us, a statistical problem begins by identifying a population of interest: a collection of things that we want to learn something about. In a statistical problem, we draw inferences about a population of interest on the basis of facts about a sample drawn from that population.

The basic elements of statistical analysis are called statistical units, where a statistical unit is an entity (broadly construed) having properties that we can measure. For example, a virus, a boll of cotton, a can of Coke, a human being, a house, an acre of farm land, a planet, and a
galaxy are all possible statistical units. Each one has properties that we can measure. A population is a collection (technically, a set) of statistical units. For example, viruses on my hands, cotton bolls in a specific field in Texas, cans of Coke in the Chicago metropolitan area, humans living in China, houses in Denmark, acres of farm land in sub-Saharan Africa, planets orbiting stars in the Large Magellanic Cloud, and galaxies within 100 million light years of the Milky Way are all populations. Often, the populations we care about are too big for us to exhaustively observe. And this is where statistical inference comes in.

The aim of statistics – the reason for the whole enterprise—is to learn something about the big, messy populations that we cannot exhaustively observe or catalogue on the basis of samples that we draw from those populations. Hence, we may think of statistics as a sustained attempt to solve Hume’s problem of induction: to show how to make reasonable inferences from what one has observed to what one has not yet observed. In other words, statistics is a logic of induction.¹

The basic problem of statistics is estimation. In many cases, we want to know something about some feature of a population called a parameter. In order to estimate the value of the parameter, we take a sample from the population and determine the value of a corresponding statistic. That statistic, together with background knowledge (if we have any), lets us estimate the value of the parameter. For example, we want to know how much caffeine is in a typical can of Coke in the Chicago metropolitan area. We take a sample of cans of Coke from the Chicago metropolitan area, we measure how much caffeine is in each can, and we use those measurements to determine precisely how much caffeine is in a typical can in our sample. The

¹ I say that statistics it is a logic of induction because there are alternatives. See Norton (2003, 2010) for examples and discussion. See Romeijn (2011) for an extended discussion of the relationship between statistics and inductive logic.
amount of caffeine in a typical can in our sample is a statistic that corresponds to the parameter of interest—the amount of caffeine in a typical can in the population as a whole. Estimation problems are ubiquitous in science and everyday life. Here are some examples of questions that require solving an estimation problem:

1. What is the average salinity of a cubic meter of water in the Atlantic Ocean?
2. How many blades of grass are on the main quad of the University of Illinois, Urbana-Champaign campus?
3. What is the difference between the median incomes of fishermen and loggers in Maine?
4. How many muons enter Earth’s atmosphere every hour?
5. If a U.S. state has a two percent higher poverty rate than a neighboring state, what should we predict about the relationship between the science education attainment levels in the two states (e.g. as measured by tests like those administered as part of the NAEP)?
6. How many more people would die in automobile crashes each year for every five mile per hour increase to the current speed limit on interstate highways in Indiana?
7. What would be the effect on suicide rates of a handgun buy-back program in Chicago?
8. What would the global annual average temperature be in 2100 if we cut CO₂ emissions by 50% next year?
9. Would studying philosophy cause you to be significantly better at standardized graduate readiness examinations than you would have been had you not studied philosophy?

Questions [1] – [5] could be answered, in principle, by looking at all of the members of the relevant population. However, the populations are typically so large or otherwise difficult to access that we want to make inferences about them without exhaustively looking at them all. Hence, we collect samples, instead. Probability theory tells us how to move from facts about the
samples we collect to facts about the populations those samples come from. Questions [6] – [9] require further assumptions, which we discuss in Sections 7.5 and 7.6.

Summing up, the big picture looks like this. We identify a population. We take a sample from that population and measure the units in the sample. We then use our measurements to characterize the sample in terms of some statistics. And finally, we use probability theory to make inferences about the population on the basis of the statistics.

7.1 Random Variables and Probability Distributions

Recall that a population is a collection of statistical units, each of which may be regarded as a collection of measurable properties. A random variable (or simply a variable) is a function from a well-defined population into a collection of values. The collection of values (called the state space) is often just the real numbers, but it is sometimes convenient to use non-numeric values, like “green,” “Republican,” or “female.” Random variables typically represent property universals (or whatever your preferred metaphysical alternative to universals is), like height, mass, operating budget, and so on. The value of a variable $V$ with respect to a specific unit $u$, denoted $V(u) = v$, represents the instantiation of the property $V$ for that unit. And hence, expressions of the form $V(u) = v$ are sentences in our formal language. As a result, it makes sense to assign a probability to expressions like $V(u) = v$. In the special case where a variable $V$ has exactly two possible values—for example, 0 and 1—we can think of $V(u) = 0$ as the negation of $V(u) = 1$. In the (more) general case where a variable $V$ has some finite number $m$ of values, the disjunction $V(u) = v_1 \lor V(u) = v_2 \lor \ldots \lor V(u) = v_m$ is a logical truth and each pair $V(u) = v_i$, $V(u) = v_j$ for $i \neq j$ is inconsistent. Hence, by finite additivity, $Pr(V(u) = v_1) + \ldots + Pr(V(u) = v_m) = 1$.

We are often interested in the way the values of a random variable are distributed in a population. The simplest possible distribution that a variable could have is called the Bernoulli
distribution. And the standard example of a variable that has a Bernoulli distribution is a variable that records the result of a single coin flip. Formally, a Bernoulli distribution is defined as follows for a variable $X$:

$$Ber(x; p) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Where $p$ is the probability that the variable takes the value 1. We can imagine that we have tossed a thousand coins out onto a floor. Each one comes up either heads or tails. From a frequentist perspective, the probability of heads ($p$) is the proportion of units $u$ in the population that came up heads (i.e. for which $X(u) = 1$). From a personalist perspective, the probability of heads ($p$) is the credence one has that any given coin came up heads (i.e. for which $X(u) = 1$).

Now, suppose we flip a coin $n$ times in succession. Let $X$ be a variable that counts the number of heads that turn up in those $n$ tosses of a coin. The value of the variable $X$ could be 0, 1, 2, …, up to the total number of tosses $n$. And the probability that $X$ has one or another possible value is given by the binomial distribution, which is defined as follows:

$$Bin(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Where $p$ is the probability of heads on any single toss (sometimes called the bias of the coin), $n$ is the number of tosses, and the combination function is defined as

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

The binomial distribution gives the probability of seeing $x$ successes in $n$ trials where each trial is independent of all the others and has probability $p$ of success. Hence, the binomial distribution corresponds to the case of sampling with replacement, i.e. drawing a sample by taking one unit from the population, making a measurement on it, and then returning it to the population before
taking another. For ten flips of a *fair* coin—one that has an equal probability of coming up heads or tails—the probability of seeing \( x \) heads turn up is equal to:

\[
Pr(X = x) = \text{Bin}(x;10,0.5) = \binom{10}{x} 0.5^x \cdot 0.5^{10-x} = \binom{10}{x} 0.5^{10}
\]

Approximate probabilities for the number of heads in ten tosses are given in Table 7.1.

<table>
<thead>
<tr>
<th>Number of Heads</th>
<th>Approximate Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.001</td>
</tr>
<tr>
<td>1</td>
<td>0.010</td>
</tr>
<tr>
<td>2</td>
<td>0.044</td>
</tr>
<tr>
<td>3</td>
<td>0.117</td>
</tr>
<tr>
<td>4</td>
<td>0.205</td>
</tr>
<tr>
<td>5</td>
<td>0.246</td>
</tr>
<tr>
<td>6</td>
<td>0.205</td>
</tr>
<tr>
<td>7</td>
<td>0.117</td>
</tr>
<tr>
<td>8</td>
<td>0.044</td>
</tr>
<tr>
<td>9</td>
<td>0.010</td>
</tr>
<tr>
<td>10</td>
<td>0.001</td>
</tr>
</tbody>
</table>

**Table 7.1: Approximate Probabilities for a Binomial**

When a probability distribution, like the binomial, is discrete, its probability function is sometimes called a probability *mass* function. The reason is that at each value for a variable that has a discrete distribution, that value has a definite amount of the stuff measured by the probability function. The binomial distribution with parameters \( n = 10 \) and \( p = \frac{1}{2} \) is pictured in Figure 7.1:
The binomial distribution lets us model sampling with replacement. But in many real-world applications, we want to sample without replacement. For example, when we play card games, like poker or hearts, we are sampling without replacement. Since the binomial does not model such cases, we need a new probability distribution. Suppose we draw \( n \) marbles all at once from a large bucket. Let \( X \) be a variable that counts the number of green marbles in our sample of \( n \) marbles. The value of the variable \( X \) could be 0, 1, 2, \ldots, up to \( A \), where \( A \) is the total number of green marbles in the population. The probability that \( X \) has one or another possible value is given by the hypergeometric distribution, which is defined as follows:

\[
Hyp(x; A, B, n) = \binom{A}{x} \binom{B}{n-x} / \binom{A+B}{n}
\]

**Figure 7.1: Binomial distribution with parameters \( n = 10 \) and \( p = \frac{1}{2} \)**
Where $x$ is the number of green marbles in the sample, $A$ is the number of green marbles in the population, $B$ is the number of non-green marbles in the population, and $n$ is the sample size.

The hypergeometric distribution is similar to the binomial insofar as we are dividing the population into two categories, which we might think of as successes and failures. However, unlike the binomial distribution, the hypergeometric distribution assumes that the trials are not all independent of one another. Let’s consider an example. Suppose our bucket contains 30 green marbles and 40 non-green marbles. We pick 15 marbles out of the bucket without replacement.

The probability of seeing $x$ green marbles in our sample is equal to:

$$\Pr(X = x) = Hyp(x; 30, 40, 15) = \frac{\binom{30}{x} \binom{40}{15-x}}{\binom{70}{15}}$$

Approximate probabilities for the number of $A$ units in a sample of size 15 (where the number of $A$ and $B$ units in the population are 30 and 40, respectively) are given in Table 7.2.

<table>
<thead>
<tr>
<th>Number of Heads</th>
<th>Approximate Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.007</td>
</tr>
<tr>
<td>3</td>
<td>0.031</td>
</tr>
<tr>
<td>4</td>
<td>0.088</td>
</tr>
<tr>
<td>5</td>
<td>0.167</td>
</tr>
<tr>
<td>6</td>
<td>0.225</td>
</tr>
<tr>
<td>7</td>
<td>0.217</td>
</tr>
<tr>
<td>8</td>
<td>0.151</td>
</tr>
<tr>
<td>9</td>
<td>0.076</td>
</tr>
<tr>
<td>10</td>
<td>0.027</td>
</tr>
<tr>
<td>11</td>
<td>0.007</td>
</tr>
<tr>
<td>12</td>
<td>0.001</td>
</tr>
<tr>
<td>13</td>
<td>0.000</td>
</tr>
<tr>
<td>14</td>
<td>0.000</td>
</tr>
<tr>
<td>15</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 7.2: Approximate Probabilities for a Hypergeometric

---

2 Approximate probabilities for 0, 13, 14, and 15 to seven digits are 5.58e-5, 1.295e-4, 8.1e-6, and 2e-7.
We can draw a plot of the hypergeometric distribution, just as we drew a plot of the binomial distribution. The result is given in Figure 7.2:

![Hypergeometric distribution plot](image)

**Figure 7.2: Hypergeometric distribution with parameters $n = 15, A = 30,$ and $B = 40$**

The distribution is not symmetric because the values of $A$ and $B$ are not equal.

Both the binomial distribution and the hypergeometric distribution treat the units in the population as falling into one or the other of two categories. However, in many cases, we want to reason about populations whose units fall into one of some number of categories greater than two. The multinomial distribution and the multivariate hypergeometric distribution generalize the binomial and the hypergeometric, respectively.

Here is a simple example. Around Halloween, the philosophy department keeps a big bowl of M&M candies in the main office.³ M&M candies come in six different colors: brown, yellow, green, red, orange, and blue. Suppose the bowl contains 2,000 M&Ms distributed

---

³ This is, sadly, a fictional example. Do not rush the department looking for M&Ms.
roughly evenly across the six kinds such that there are 323 brown, 327 yellow, 318 green, 332 red, 344 orange, and 356 blue M&Ms in the bowl. You scoop out 251 M&Ms from the bowl.

What is the probability that in your 251 M&Ms you have 42 brown, 43 yellow, 41 green, 39 red, 42 orange, and 44 blue M&Ms? Since you are sampling without replacement, the answer is given by the multivariate hypergeometric, which is defined as:

$$\text{MultHyp}(\bar{x}; \tilde{K}, n) = \frac{\prod_{i=1}^{m} \binom{K_i}{x_i}}{\binom{N}{n}} = \frac{\left(\frac{K_1}{x_1}\right) \cdot \left(\frac{K_2}{x_2}\right) \cdot \ldots \cdot \left(\frac{K_m}{x_m}\right)}{\binom{N}{n}}$$

Where $N = \sum_{i=1}^{m} K_i = K_1 + K_2 + \ldots + K_m$ is the total number of units in the population,

$n = \sum_{i=1}^{m} x_i = x_1 + x_2 + \ldots + x_m$ is the total number of units in the sample, $K_i$ is the number of units in the population that fall into the $i^{th}$ category, and $x_i$ is the number of units in the sample that fall into the $i^{th}$ category. Hence, the answer to the question above is:

$$\binom{323}{42} \cdot \binom{327}{43} \cdot \binom{318}{41} \cdot \binom{332}{39} \cdot \binom{344}{42} \cdot \binom{356}{44} \cdot \binom{2000}{251}$$

which is too difficult for my computer to calculate. Now, if we were sampling with replacement from the same population—taking out one M&M at a time, recording its color, and then tossing it back in the bowl before drawing another—we would calculate the probability of seeing the exact same sample using the multinomial distribution, which is defined as:

$$\text{Mult}(\bar{x}; n, \tilde{p}) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdot \ldots \cdot p_k^{x_k}$$

Where the combination function from the binomial distribution has been replaced by:
$$\binom{n}{x_1,\ldots,x_k} = \frac{n!}{x_1!\cdots x_k!}$$

Each $p_i$ term in the multinomial distribution is the single-case probability of drawing a unit from the $i^{th}$ category, and each $x_i$ term is the number of units in the $i^{th}$ category that appear in the sample. Hence, the probability of seeing the same sample as before except drawn with replacement is given by:

$$\begin{vmatrix} 251 \\ 42, 43, 41, 39, 42, 44 \end{vmatrix} \begin{vmatrix} 323 \\ 2000 \end{vmatrix}^{42} \cdot \begin{vmatrix} 327 \\ 2000 \end{vmatrix}^{43} \cdot \begin{vmatrix} 318 \\ 2000 \end{vmatrix}^{41} \cdot \begin{vmatrix} 332 \\ 2000 \end{vmatrix}^{39} \cdot \begin{vmatrix} 344 \\ 2000 \end{vmatrix}^{42} \cdot \begin{vmatrix} 356 \\ 2000 \end{vmatrix}^{44}$$

Which is also too difficult for the computer. Statisticians have developed other distributions that are easier to use for calculation and that approximate the values of the multinomial and multivariate hypergeometric distributions.

### 7.2 Statistical Inference from a Bayesian Perspective

In this section, I want to consider the problem of inferring the value of a parameter—a feature of a distribution that characterizes a population—from a Bayesian point of view. Here is a sketch of how Bayesian inference goes. (Compare this to the Big Picture, as sketched above.) First, the Bayesian models the data-generating process (the population plus the sampling scheme) using a probability distribution. The probability distribution that the Bayesian uses to represent the data-generating process has some parameter(s) that we care about. Second, the Bayesian models her uncertainty about the parameter(s) of interest by using another probability distribution. For example, the Bayesian might model a sequence of coin flips with a binomial distribution, where she is interested in the bias of the coin. The bias of the coin is a parameter of the binomial distribution and a feature of the population. The Bayesian then represents her uncertainty about what specific bias the coin has before any new data come in. Third, the Bayesian collects data
generated from the population according to some appropriate sampling scheme. And fourth, the Bayesian updates her prior distribution over the parameters in light of the data. Let’s see how the basic idea works in a bit more detail for the specific example of inference about the binomial.

Suppose some epistemic agent wants to estimate the proportion of green marbles in a jar.\(^4\) The agent collects data by repeatedly pulling a marble from the jar, recording the color, and then returning it to the jar. She thinks that a good model for the data-generating process is a binomial distribution, and she is interested in estimating the value of the parameter \(p\). Now that she has a guess about what the data-generating process looks like, she needs another distribution to characterize her uncertainty about the parameter \(p\). The usual choice for modeling uncertainty about the parameter \(p\) in a binomial distribution is called the *beta* distribution. The beta distribution has two parameters, \(\alpha\) and \(\beta\), which are sometimes called *shape* parameters. Unlike the distributions we have considered so far, the beta distribution is a *continuous* function, which ranges over the real interval \((0, 1)\) and is given by the following equation:

\[
\text{Beta}(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1},
\]

And the function \(\Gamma(x)\) is defined as:

\[
\Gamma(x) = \int_0^x t^{x-1} e^{-t} dt,
\]

which is equal to \((n - 1)!\) whenever \(x\) is a natural number. If you find the above equations enlightening, great! If not, don’t worry too much about it: just know that you have more to learn in this area. While the details of the equation for the beta distribution are not important for our

\(^4\) This is a toy case, but there are lots of real-world analogues. Some examples include estimating the proportion of registered voters who will vote for either of two candidates in a large election, estimating the proportion of animals in a population having a specific genetic marker, and estimating the number of times some event occurs in a given number of opportunities.
purposes, we will be well-served to graph it for some values of $\alpha$ and $\beta$ to see how the shape of the distribution changes in response to changes in the values of the shape parameters.

Suppose our imaginary epistemic agent thinks that all possible values of the proportion of green marbles in the jar are equally likely. Then her prior distribution for the proportion $p$ will be given by a beta distribution with parameter values $\alpha = 1$ and $\beta = 1$. Our agent’s prior is pictured graphically in the upper right hand corner of Figure 7.3.

The measure theory (and calculus) required to really work with the Beta distribution (and other standard continuous distributions, like the normal and the exponential) is beyond the scope
of this book, but one thing is worth pointing out. For discrete distributions like the binomial, the value output by the function (sometimes called a probability *mass* function) for some specific input value(s) is a probability. By contrast, for continuous distributions like the beta, the value output by the function (sometimes called a probability *density* function) for some specific input value(s) is *not* a probability. In a continuous distribution, probabilities correspond to the area under the curve in some *interval*. So, suppose we have a beta distribution with shape parameters both equal to one. Such a distribution is flat, as you can see in Figure 7.3. The value of the beta distribution for $p = 0.5$ is 1, but that value is not a *probability*. (The probability that $p$ is exactly equal to 0.5 is actually zero!) By contrast, the area under the curve between 0.4 and 0.6 is 0.2, which is the probability that the value of the parameter $p$ is in the interval $[0.4, 0.6]$.

Our imaginary epistemic agent now draws 15 marbles with replacement from the bucket and observes 4 green marbles and 11 non-green marbles. In general, if an agent’s prior distribution is a beta with parameterization $\alpha = a$ and $\beta = b$, and we update it using Bayes’ rule on the basis of $s$ successes and $f$ failures, then the posterior distribution is also a beta distribution, and it has parameterization $\alpha = a + s$ and $\beta = b + f$. Hence, after seeing the data, our Bayesian agent has credences described by a beta distribution with $\alpha = 1 + 4$ and $\beta = 1 + 11$. The resulting distribution is pictured in Figure 7.4:
The Bayesian agent may use her posterior distribution to calculate something called a *credible interval*. A $100 \cdot (1 - \alpha)\%$ credible interval for the parameter $p$ is any interval $(a, b)$ such that $\Pr(a < p < b) = 1 - \alpha$ according to the posterior distribution. Hence, a Bayesian credible interval may be determined in any one of several different ways. Here are four: (1) take the interval such that the mean of the posterior distribution (if it exists) is exactly in the middle of the interval; (2) take the interval, called the *equal tails interval*, such that according to the posterior distribution, the probability that the parameter is smaller than the lower endpoint of the interval is the same as the probability that the parameter is greater than the upper endpoint of the interval; (3) take the narrowest continuous interval that has the required probability; (4) take the interval such that the endpoints of the interval have a greater value than any of the points outside the interval.

Suppose our agent uses the second method and calculates a 95% credible interval such that 2.5% of the total area under the distribution is to the right of the upper endpoint of the
interval and 2.5% of the total area under the distribution is to the left of the lower endpoint of the interval. The result is the interval (0.11, 0.52). After seeing the data, our Bayesian agent believes that there is about a 95% chance that the proportion of green marbles in the bucket is in the interval (0.11, 0.52). She also believes that there is a 2.5% chance that the proportion of green marbles in the bucket is larger than 0.52, and she believes that there is a 2.5% chance that the proportion of green marbles is smaller than 0.11.

Our epistemic agent began with a flat prior with respect to the possible values of $p$. The flat distribution is a natural choice for representing a lack of information about the value of $p$. However, the flat distribution is not invariant under transformations of scale. (It suffers from the same problem that we saw with the ball bearing factory in our discussion of evidentialism in Chapter 6.) Harold Jeffreys (1946) proposed a way of constructing priors that are invariant under transformations of scale. For proportions, the Jeffreys prior is a beta distribution with parameterization $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$. Jeffreys priors have their own problems, and serious engagement with the problem of non-informative priors is beyond the scope of this text. However, you should be aware that there is interesting work to be done here.

7.3 Observing Versus Doing

At the beginning of Section 7.1, we introduced the idea of a random variable. We said that the value of a variable $V$ with respect to a specific unit $u$, denoted $V(u) = v$, represents the instantiation of the property $V$ for that unit. In an ideal case, the value of a variable also represents the result of some measurement process applied to the unit. (In real cases, our measurements are often only indicators of the way a given unit actually instantiates some property universal.) Expressions assigning a value to a variable for a unit, like $V(u) = v$, are sentences of our formal language. Hence, it makes sense to talk about the probability that a unit
takes a specific value for a variable. By a slight extension, we will let \( \Pr(V = v) \) denote the probability that a randomly chosen unit has value \( v \) for variable \( V \). In other words, \( \Pr(V = v) \) is the probability that property \( V \) is measured to be \( v \) for a randomly chosen unit. But of course a given unit might have lots of properties that we can measure. For example, we might observe that some unit \( u \) has a \textit{Height} of 178 cm and a \textit{Weight} of 68 kg. Since expressions like \( V(u) = v \) are sentences in our formal language, so are conjunctions, like \( H(u) = h \land W(u) = w \). Hence, it also makes sense to talk about the probability of such conjunctions. And by extension, the probability of conjunctions of the form \( H = h \land W = w \).

If we have a collection of random variables, say \( V_1, \ldots, V_n \), then the \textit{joint probability distribution} over the collection \( V \) is the distribution of probability values with respect to the conjunction \( V_1 = v_1 \land V_2 = v_2 \land \ldots \land V_n = v_n \), for all of the possible combinations of values for the variables. Conditional probability for random variables is then given by the ratio account as before:

\[
\Pr(Y = y \mid X = x) = \frac{\Pr(Y = y \land X = x)}{\Pr(X = x)}
\]

for specified combinations of values \( x \) and \( y \). Similarly, we can extend the definition of association and independence from events to random variables. We say that two variables \( X \) and \( Y \) are \textit{independent} if and only if \( \Pr(Y = y \mid X = x) = \Pr(Y = y) \) for all values \( x \) of \( X \) and \( y \) of \( Y \). If there exists even a single pair of values \( <x, y> \) such that \( \Pr(Y = y \mid X = x) \neq \Pr(Y = y) \), then \( X \) and \( Y \) are (statistically) associated.

When two variables are associated, knowing the value that one of them takes is (at least sometimes) informative about what value the other variable takes. When measured associations reflect a genuine relation between property universals, as opposed to an artifact of sampling error (bad luck), then the value of one property may sometimes be inductively inferred from the value
of the other. Suppose that our statistical units are small coastal towns in the Eastern United States in a given month. We have two binary variables: $B$ and $S$. When we input a town (at a specific month) into the variable $B$, the output is whether the frequency of boating accidents is high or low. When we input a town into the variable $S$, the output is whether the volume of ice cream sales is high or low. Now, suppose that the two variables are positively associated. If we find that $B(u) = \text{high}$ for some unit $u$, then we should increase our credence that $S(u) = \text{high}$. (Our new credence that $S = \text{high}$ might still be low, but it should go up from baseline.)

Statistical or probabilistic associations provide a good basis for observational predictions. However, they do not always provide a good basis for actions. What would provide a good basis for action? The answer is causal structure. If some property $X$ causes some other property $Y$, then changing the value of $X$ is an effective strategy for changing the value of $Y$. Changing the value of $X$ might not be the best strategy or the most effective strategy. But it will be effective. However, the fact that two variables are associated does not provide reason to think that changing the value of one will be an effective way to change the value of the other. We think that in our example above, boating accidents neither cause nor are caused by ice cream sales. Rather, the two are related because some third factor, like temperature, is a common cause of both.

Despite the fact that $B$ and $S$ are statistically associated, imposing a heavy tax on ice cream in order to keep down sales would be an ineffective strategy for reducing the frequency of boating accidents. In short, there is an important difference between observing and doing.

In general, a statistical association between two variables $X$ and $Y$ may be explained in one of four different ways: by chance, by $X$ causing $Y$, by $Y$ causing $X$, or by a common cause (or collection of common causes) of both $X$ and $Y$. In some cases, several different causal stories are plausible, and good policy decisions depend on figuring out which story is correct. Let’s consider
a real-world example. There is a weak association between firearm availability rates and homicide rates across high-income nations. What explains it? The association could be a fluke. After all, the observed association is due almost entirely to the large homicide rate and the large gun ownership rate in the United States. Removing the United States from the dataset eliminates the association. If it is not a fluke, the explanation could be that gun availability causes homicides. The mechanism here is simple and obvious: guns make killing easier, so when people have easy access to guns, homicides increase. Or it could be that homicides cause gun availability. Perhaps people observe the high homicide rate and reason that they will be safer with a gun than without. Or the association could be due to a common cause. For example, vicious, murderous people might be more likely to commit homicide and be more likely to make a gun or to purchase a gun.

What does it mean for one variable to cause another variable? First, let me give an intuitive gloss: A variable $X$ causes a variable $Y$ just in case if we hold everything else fixed and wiggle $X$, $Y$ wiggles along. For example, suppose that you toggle a switch. Whenever the switch has been set to the on position, a light is on, and whenever the switch has been set to the off position, the light is off. Then (speaking a bit loosely) the switch is a cause of the light. The relationship between the switch and the light might depend on other variables taking specific values, e.g. that there is electricity going to the circuit. More formally, we say that the variable $X$ is a direct structural cause of the variable $Y$ relative to some collection of other variables, the $Z$s, if and only if for some valuation of the $Z$s, there exists a pair of distinct values for $X$ such that if $X$ is set to one value, $Y$ equals $y$, and if $X$ is set to the other value, $Y$ does not equal $y$. The pair of values for $X$ is called a test pair. Importantly, causation need not be deterministic. For example,
smoking causes lung cancer. But smoking does not guarantee that one will get lung cancer. In fact, most people who smoke do not get lung cancer. However, if everything else is equal, smoking radically increases the chances of getting lung cancer.\footnote{Lung cancer is very rare, but it is something like five times as likely among smokers than among non-smokers.}

We are going to use directed graphs to represent causal relationships. Our graphs will have two basic elements: a collection of vertices, $A, B$, and so forth, which we also think of as random variables; and a collection of directed edges between pairs of vertices. When we have a directed edge, e.g. $A \rightarrow B$, in a graph, we are saying that there is a direct (structural) causal relation that points in the direction of the edge, e.g. that $A$ is a direct (structural) cause of $B$. When we have no edge connecting a pair of variables, then we are saying that there is no direct (structural) causal relationship between the vertices.

7.4 Independence and Conditional Independence

Recall that in Section 6.2 we said that two sentences $\phi$ and $\psi$ are (probabilistically or statistically) independent when the conditional probability $\Pr(\phi | \psi)$ is equal to the unconditional probability $\Pr(\phi)$. In this section, we extend the notion of independence in two ways. First, we generalize the idea of independence from sentences to properties (as represented by random variables). And then we extend our account from cases of unconditional independence to cases of conditional independence.

Let $u$ be a statistical unit, let $V$ be a variable, and let $v$ be a value in the range of $V$. Then an expression of the form $V(u) = v$ may be translated by a sentence in our language that says a given individual has a specific instance of a given property. For example, if $\text{Shape}(u) = \text{square}$, then we are saying that unit $u$ has (or instantiates) a square shape. We have already generalized
this slightly by understanding expressions of the form $V = v$ to denote the set of all units for which $V$ has the value $v$. And then we have understood probability claims such as $\Pr(V = v) = p$ to be saying that the probability that an arbitrarily chosen unit has the value $v$ for $V$ is equal to $p$. Hence, a probability claim might be made in terms of sentences or in terms of sets of worlds, which we have called events. And a nice way of thinking about an event is as a way of having a property instance. Consequently, our earlier definition of independence for sentences applies straightforwardly to events understood as ways of having property instances. For example, we might talk about the probability of having a square shape or the probability of having a mass of ten kilograms. If $\Pr(\text{Shape} = \text{square} \mid \text{Mass} = 10 \text{ kg}) = \Pr(\text{Shape} = \text{square})$, then we say that having a square shape is independent of having a mass of ten kilograms. In general, we say that $X = x$ is independent of $Y = y$ if and only if $\Pr(X = x \mid Y = y) = \Pr(X = x)$. We are now in position to extend the notion of independence to properties (as represented by random variables). We say that the variable $X$ is independent of the variable $Y$ if and only if for all $x$ and all $y$, $X = x$ is independent of $Y = y$. For all of these definitions, if two things are not independent, they are associated.

So far, we have talked about independence as a relation that holds between two things without regard to the way anything else might be. However, just as we have defined a notion of conditional probability, we might define a notion of conditional independence. That is, we might write down a formal definition to try to cash out our ordinary sense that two things might not be informationally related if we know some third thing to be the case. Formally, we say that $A$ is independent of $B$ given (or conditional on) $C$ if and only if $\Pr(A \mid B \land C) = \Pr(A \mid C)$. We can think of what is going on here as restricting the universe of discourse to $C$ and then asking about that universe whether $A$ and $B$ are unconditionally independent. The other expressions involving
independence work similarly. Hence, we will say that \( X = x \) is independent of \( Y = y \) given (or conditional on) \( Z = z \) if and only if \( \Pr(X = x \mid Y = y \land Z = z) = \Pr(X = x \mid Z = z) \). When we are working with variables, we will sometimes write \( \Pr(X = x \mid Y = y, Z = z) \) instead of using the conjunction operator between \( Y = y \) and \( Z = z \). Finally, we generalize conditional independence to variables by saying that \( X \) is independent of \( Y \) given (or conditional on) \( Z \) if and only if for every choice of values \( x, y, \) and \( z \) for \( X, Y, \) and \( Z \), \( \Pr(X = x \mid Y = y \land Z = z) = \Pr(X = x \mid Z = z) \).

7.5 Causal Inference from Experimentation

Causal, but not statistical, relations are poised to guide our actions and inform our policy decisions. But how can we figure out what causes what? The most reliable way to figure out what causes what is to conduct controlled experiments. The central problem of causal inference is a problem of under-determination. A statistical association between two variables \( X \) and \( Y \) might be explained by any of three different causal structures. The variable \( X \) might be a direct cause of the variable \( Y \). The variable \( Y \) might be a direct cause of the variable \( X \). Or there might be a common cause or a family of common causes of the variables \( X \) and \( Y \). Hence, the true causal structure is under-determined by the data.

However, suppose we could conduct an experiment in which we set the value of the variable \( X \). Since we are determining the value of the variable \( X \), we know that it has no other direct causes. Therefore, if we observe an association between \( X \) and \( Y \), we know that the association cannot be explained by the fact that \( Y \) causes \( X \). Neither can it be explained by any unmeasured common cause (sometimes called a confounder). The variables \( X \) and \( Y \) might be associated in virtue of the fact that the intervention itself is a common cause of both \( X \) and \( Y \). To avoid such difficulties, we often attempt to make the assignment of values for \( X \) to units in an
experiment are made randomly. Since we do not expect a random process to be a cause of $Y$, if we randomly assign values to $X$ and observe an association between $X$ and $Y$, we conclude that $X$ is a direct structural cause of $Y$.

7.6 Causal Inference from Observation

Even if we can’t conduct experiments, we can sometimes learn something about the causal structure that gives rise to our observed data. We are going to use partially directed graphs to represent what we know about the causal relationships that hold in a given case. As before, our graphs will have vertices, $A$, $B$, and so forth, (which we also think of as random variables) and also directed edges between pairs of vertices. In addition, our graphs will include undirected edges between pairs of vertices. When we have an undirected edge, e.g. $C - D$, in a graph, we are saying that there is a direct (structural) causal relationship between the vertices. Either $C \rightarrow D$ or $D \rightarrow C$, but we don’t know (or are declining to say) which.

We are already making a big assumption at this point: there are no unrepresented common causes. That is, undirected edges are not simple statistical associations, which could be explained by the presence of a common cause not included in the graph. The name for this assumption is causal sufficiency.\(^7\) In addition to causal sufficiency, we are going to add another assumption: that the causal structure over a collection of variables is acyclic. The acyclicity assumption says that no causal graph that we draw will have a sequence of vertices $X_{(1)}$, ..., $X_{(n)}$ such that $X_{(i)} \rightarrow X_{(i+1)}$ for $i$ in 1 to $n$-1 and such that $X_{(1)} = X_{(n)}$.

\(^7\) Assuming causal sufficiency is not required for causal search. See Pearl (2000) and Spirtes et al. (2000) for procedures that do not assume sufficiency.
Given the assumption of causal sufficiency (that there are no unrepresented common causes) and the assumption of acyclicity, we want to draw inferences from patterns of statistical relations to a causal structure, represented by a partially directed graph.\(^8\) We begin with two basic ideas. The first basic idea is that if two variables are unconditionally associated with one another and continue to be associated with one another regardless of what other variables we condition on, then those two variables are adjacent in the graph, where two variables are said to be adjacent if there is either a directed edge or an undirected edge between them. Given our assumptions, two variables are adjacent just in case one of the variables is a direct (structural) cause of the other.

The second basic idea is that unshielded colliders have some peculiar properties that make causal inference work. An unshielded collider (also called a v-structure or an immorality) is a collection of three variables, e.g. \(A, B,\) and \(C,\) such that two variables are each direct (structural) causes of a third but are themselves unrelated, e.g. \(A \rightarrow B \leftarrow C.\) Unshielded colliders exhibit a unique pattern of conditional independences—a pattern that distinguishes them from all other causal graphs over three variables. In a completely connected graph, e.g. any of the graphs in Figure 7.5, everything is unconditionally associated with everything else, and each variable is conditionally associated with each other variable conditional on the third variable.

\(^8\) Of course, we might have common causes in the graph, e.g. \(A \leftarrow B \rightarrow C,\) but we won’t have any variable \textit{not in our graph} that stands in a direct (structural) causal relation to more than one vertex in our graph.
Four graphs over three variables are missing one edge. All four of these are pictured in Figure 7.6. Three of these graphs are statistically indistinguishable. The unshielded collider graph is statistically different from the other three.
The three graphs on the left all have the following properties. Each of the variables is unconditionally associated with each of the other variables. Moreover, in each graph, $A$ is conditionally independent of $C$ given $B$. That is, $B$ screens off $A$ from $C$ (and vice versa). By contrast, the graph on the right has the following properties. Variables $A$ and $C$ are unconditionally independent of one another, but both are unconditionally associated with $B$. However, $A$ and $C$ are conditionally *associated* given $B$. That last fact is often surprising at first, so here is an illustrative example. We have three variables: one representing whether or not Suzy smiles, one representing whether or not Suzy gets an ice cream cone, and one representing whether or not Suzy gets a balloon. Suppose that whether Suzy gets an ice cream cone is independent of whether she gets a balloon, and suppose that Suzy smiles if she gets a balloon, an ice cream cone, or both. Now, suppose that we know that Suzy smiled. That is like conditioning on variable $B$ in Figure 7.6. Having that bit of information induces a statistical association between the other two variables: in at least some cases, knowing how one of the variables comes out tells us how the other variable comes out. For example, if we know that Suzy smiled and we know that Suzy did not get a balloon, then we may infer that Suzy got an ice cream cone.

The basic ideas above are formalized by the *causal Markov condition*, which may be stated in various ways. The (local) causal Markov condition says that each variable in a graph is independent of its non-descendants (the vertices in the graph that cannot be reached from the target variable by following edges in the direction they point) conditional on its parents (the variables that are direct structural causes of the target variable). The causal Markov condition is a conditional, so it supports inferences from statistical *associations* to graphical structure. Note that for all the causal Markov condition says, some variables might be independent despite being causally related. In order to bar such a possibility, we sometimes make a further assumption—the
causal faithfulness condition—which says that the only independences in the statistical data are the ones entailed by the causal Markov condition.

Now, assume that from the patterns of statistical association that we have with respect to some collection of variables we can identify the adjacencies and the unshielded colliders in the graph representing the causal structure for those variables. Chris Meek showed how to get further orientation information out of the resulting partially directed graph.⁹ Specifically, Meek proposed the four rules pictured in Figure 7.7.

Each rule orients an undirected edge in a graph. The rules work by applying two assumptions that we made above. First, recall that we assumed our causal graphs are acyclic. Pick an undirected edge in a graph. If orienting the edge one way produces a cycle, then it must be oriented the other way. Otherwise, the acyclicity assumption is violated. Second, recall that we assumed (via the Markov condition) that we can identify all of the unshielded colliders from patterns of statistical associations and independences. Pick an undirected edge in a graph. If orienting the edge one way produces an unshielded collider, then it must be oriented the other way. Otherwise, contra hypothesis, we did not identify all of the unshielded colliders from the statistical data.

⁹ See Meek (1995).
All of Meek’s rules follow from the above considerations. For example, his first rule says that if we have $A \rightarrow B - C$ in a graph, then we may orient the undirected edge from $B$ into $C$.

Why? Because if the edge were oriented from $C$ into $B$, then we would have an unshielded collider. Proving the other rules is left as an exercise to the reader.
References


