CHAPTER 2: PROBABILITY THEORY

Everyone has an intuitive idea of what probability is: it has to do with the likelihood that something you're interested in will happen. If you flip a coin, you may be interested in the likelihood of its coming up "heads"; if you throw a die, you might be interested in the likelihood of its coming up "6", if you plan a picnic tomorrow, you might be interested in the likelihood of it not raining tomorrow, and so on.

Probability theory is a major branch of mathematics, about which entire classes are taught and lengthy books written. In this chapter, we will cover the fundamentals of probability theory—enough so that you'll be able to use it to understand the fundamentals of data analysis. Before going into probability theory, however, it will be useful to provide a very brief synopsis of *set theory*, because some of the concepts and notations of set theory conveniently carry over into probability theory. So first a brief side trip to the world of set theory.

Set Theory

A set, very simply, is a well defined collection of things. A "thing" can be—well anything. "Well defined" simply means that you can unambiguously specify whether any given thing is or is not a member of the set. So for instance, the integers between 1 and 10 inclusive form a set; all integers (not just those between 1 and 10) form a set; all real numbers form a set, people who are members of the British House of Commons form a set; and so on.

Typically a set is referred to like this:

A = {members of a set}

where "A" is the name of the set (sets are usually but not always designated by a capital letter) and within the curly brackets is a specification of the Set A's members in the form of either a definition or a listing of the set's members. So for instance, we might designate,

A = {all positive integers smaller than 11}

or equivalently,

$$A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

Similarly we might designate,

B = {all integers}

and,

C = {all real numbers}

A Set's Cardinality

You may have noticed a distinction between, for example, Set A and Set B above. Set A contains a finite number of members (10, actually) while Set B contains an infinite number of members—you could go on forever, for example, listing integers. Those of you who are *really* paying attention may have even noticed, or at least intuited, that there's some kind of difference between Set B and Set C: even though the obviously both contain an infinite number of members, there's something different about the infinite number of integers and the infinite number of real numbers.

Let's be a little more definitive about what we mean here. A set's *cardinality* refers to the number of members that the set contains. Cardinality, it turns out, comes in three flavors: one finite flavor and two infinite flavors.

Finite sets

Finite sets are simple: they contain a finite number of members. Set A above is a finite set containing, as it does, 10 members. Some sets can be finite but very big, for example the set of all stars in the Milky Way.

Infinite sets

It probably won't surprise you to hear that an infinite set is one that contains an infinite number of members. However, as we hinted above, not all infinities are the same. Infinite sets are divided into *countably infinite and uncountably infinite* sets. The difference between countably and uncountably infinite is a bit hard to describe but is easiest to imagine like this.

Consider the set of all positive integers. Suppose you were asked to list all members of this set. You couldn't, of course, because there're an infinite number of members. However, you could at least start. You could, for instance, begin at 1, then list 2, 3, 4, and so on until you got tired. The key here is that, because each member of the set is discrete, beginning to list them is an option. "Countable" is a misnomer, because you can't count all of them, but you can at least point to the members individually.

Now consider the set of all positive real numbers. Again it's an infinite set. But suppose you were asked to begin listing them. This would present a problem. Even deciding on a starting point would be an issue. Suppose you decide to begin with the smallest number like did with the integers. What *is* the smallest member of the set? It's not zero, because zero, not being positive, isn't a member of the set. How about, say, 0.0001. Well, OK, but it isn't the smallest number in the set, because there's an infinite number of set members that are between zero and 0.0001. And so on. So you couldn't start with the smallest member. You could, of course, start anywhere, say 1.0. But then what's next? The point here is that members of this set aren't discrete—no matter what you pick, the next one isn't clear. A rough way of characterizing the distinction between countably and uncountably infinite sets is that each member of a countably infinite set discretely sits apart from its fellow set members, while the members of an uncountably infinite set are all sort of "infinitely squished together." We realize that this is less than ideal as a formal definition, but things will become clearer in a bit when we begin applying set theory to probability theory.

Subsets

A set, B is defined to be a *subset* of Set A if all members of B are also members of A. So for instance if,

A = {All automobiles}

and

B = {All Saabs}

then B is a subset of A because all Saabs are also automobiles. Notationally, we denote a subset relation as,

$A \subseteq B$

Notice that the "$\subseteq$" sign is sort of like the standard "$\leq$" sign. Just as, for example, $x \leq y$ means "x is less than *or* equal to y," $B \subseteq A$ means "Set B is a *proper subset* of A" *or* "Set B and Set A are the same set." By "proper subset" is meant that there is at least one member of A that isn't a member of B. Usually when we're concerned with subsets, we're concerned with proper subsets.

The Universal Set (W)

When using set theory in some real-life setting, there is typically some overall set of things that we're interested in. We saw this in Chapter 1 in which we were measuring soup volume in cans of Acme soup. Here, the overall set of things in which we were interested is what we

referred to as the population of all Acme soup cans. Within the context of set theory, this is referred to as a *universal set*, and the letter W is reserved to denote it.

Complements

Having defined a universal set, we are now in a position to define a complement. So consider a universal set, say

W = {All Oxford University undergraduates}

Now suppose that

A = {All female Oxford undergraduates}

Then the complement of A, designated $\overline{A}$ is all Oxford undergraduates who aren't females, i.e.,

$\overline{A}$ = {All male Oxford undergraduates}.

Unions

Suppose we have two sets, A and B. As an example,

A = {All left-handed humans}

and

B = {All male humans}

Then the *union* of A and B, designated (A ∪ B) is the set of all objects that are members of Set A *or* Set B or both. So in this case, we could, if we like, define a new set C as,

C = (A ∪ B) = {All humans who are left-handed or male or both)

Notice that (A ∪ B) usually (although not always) contains more members than either A or B individually.

Intersections

Let's consider the same two sets, A and B, i.e.,

A = {All left-handed humans}

and

B = {All male humans}

Then the *intersection* of A and B, designated (A ∩ B) is the set of all objects that are members of *both* Set A and Set B. So in this case, we could, if we like, define a new set D as,

D = (A∩B) = {All left-handed male humans)

Notice that (A ∩ B) usually (although not always) contains fewer members than either A or B individually.

The empty set

The set with nothing in it is called the *empty set* and is designated ∅.

Complements

Consider an event, A, that is part of a universal set, W. The *complement* of Set A, designated $\overline{A}$, consists of all members of A that are not members of W. As a simple example suppose,

W = {All Italian citizens}

A = {All female Italian citizens}

then,

$\overline{A}$ = {All male Italian citizens}

Mutually exclusive and exhaustive sets

Any number of sets, A, B, C,… are *mutually exhaustive* if the union of these sets is W, i.e., if among them they exhaust the universal set. So consider a universal set,

W = {All people who are living or have ever lived}

along with three subsets of W,

A = {All people born before 1990}

B = {All people born after 2000}

C = {All people born between 1975 and 2008}

Then (A∪B∪C) includes all of W; thus A, B, and C are mutually exhaustive.

Any number of sets, A, B, C,… are *mutually exclusive* if the intersection of any two of the sets is ∅. Again considering the universal set,

W = {All people who are living or have ever lived}

along with three different subsets of W,

A = {All people born before 1990}

B = {All people born after 2000}

C = {All people born between 1992 and 1995}

Because any two of these sets have an intersection of ∅, they the three are mutually exclusive.

Partitions

Finally, suppose some number of sets, sets, A, B, C,… are both mutually exclusive and mutually exhaustive. These sets are then said to *partition* the sample space. Again using the same example, three sets that would partition W would be,

A = {All people born before 1990}

B = {All people born after 2000}

C = {All people born between and including 1990 and 2000}

Basic Probability Theory

Equipped with this foundation in set theory, we are now ready to proceed to a discussion of probability theory. Everyone has a general idea of what probability is: it has to do with the likelihood that something you're interested in will happen. If you plan a picnic tomorrow, you're interested in whether it will rain. If you roll two dice you might be interested in whether the dice sum to 7. And so on.

Situations and Outcomes

To begin to formalize probability theory, we consider a *situation* along with some *outcome* of that situation in which we are interested. Here are some examples.

Situation	Outcome
Throw a die	Die comes up a "4"
Toss a coin	Coin comes up "heads"
Plan a picnic for tomorrow	It rains tomorrow
Shoot a free throw in basketball	You make the basket

A Basic Equation for Probability

Let's consider one of these situations: tossing a die with an interest in the die coming up a "4". At this point, everyone reading this book probably knows that the probability that a fair die turns up "4" is 1/6. Fine. Let's use this intuitive knowledge as a basis for understanding the logic behind a basic equation for probability. To begin with, let's consider a universal set consisting of

all possible outcomes of the situation. In probability theory, we refer to this universal set as "S" and we call it a *sample space*. In this simple example, it's easy to specify S, which is,

$S = \{1, 2, 3, 4, 5, 6\}$

Each member of S is referred to as an *elementary event*. Now let's define f(S) as the number ("f" for "frequency") of elementary events in S. In more complex examples, this won't be so easy to do, but in this example, it's obvious that f(S) = 6.

Now define another set—we'll call it A—as that subset of S that contains those elementary events corresponding to *Outcome* A, i.e., our outcome of interest. Again in this example, it's simple. If Outcome A is "die comes up a '4'" then,

$A = \{4\}$

and again using our "f" notation, f(A) = 1. The basic equation for probability is then:

$$p(A) = \frac{f(A)}{f(S)} \qquad \text{Equation 1.1}$$

or in this example,

$$p(A) = \frac{1}{6} = .167$$

Another Example and an important Caveat

The preceding example was pretty simple: intuitively and formally, the answer turned out to be 1/6. Let's consider another example that's a bit more complicated. Let's suppose, using census data, that you manage to track down all U.S. families who have exactly four children. Suppose further that we randomly select one such family. Define Outcome A to be: the selected family has 2 boys and 2 girls. What is the probability of this outcome?

To answer this question, we first need to establish S, the sample space of all possible outcomes, i.e., elementary events corresponding to this situation. To do so it's useful to start with *one* elementary event. An obvious candidate might be: all boys. Let's' designate this elementary event as "BBBB" indicating that the family had a boy four times in a row. Similarly, for example, "GBGB" would indicate that the family had a girl, then a boy, then another girl then finally another boy. With this notational scheme, we can, with some effort, systematically list all possible outcomes: the members of S, the sample space are,

Sequences of boys and girls in a 4-child family			
BBBB	BBBG	BBGB	BBGG
BGBB	BGBG	BGGB	BGGG
GBBB	GBBG	GBGB	GBGG
GGBB	GGBG	GGGB	GGGG

and it is easy to see that f(S) = 16. Our next step is to count up all members that correspond to our outcome of interest, i.e., 2 boys and 2 girls. This set is,

$A = \{ BBGG, BGGB, GBGB, GGBB, GBBG, BBGG\}$

So f(A) = 6 and p(A) = f(A)/f(S) = 6/16 = 0.0375

Notice that in this problems, we used the term "randomly select". This term has an important meaning in statistics and experimental design. Specifically, it means that we select our family

such that of all the 4-kid families we've identified (let's say, for the sake of simplicity, that there are 1,000,000 of them) each one has an equal chance of being chosen.

The equiprobability assumption

To use Equation 1.1, we must make an important assumption called the *equiprobability assumption*. This is that *all elementary events in the sample space occur with equal probability*. Within the context of our die example it means that the die is a *fair die*, i.e., that it is manufactured such that each face has an equal probability of coming up. This is a reasonable assumption because dice factories are very careful to manufacture dice this way.

With our 4-kid example, the equiprobability assumption is more suspect. For example, it appears to be true that slightly more than 50% of children born are boys. This would mean, for instance that the probability of a BBBB sequence would likely be a bit higher than the probability of a GGGG sequence.

To the degree that the equiprobability assumption is incorrect, any probability that we compute using Equation 1.1 is similarly incorrect. But if the equiprobability assumption is approximately correct—as it is for the 4-kid example—then Equation 1.1's answer is, likewise, also approximately correct.

More General Laws of Probability

Equation 1.1 is widely used in everyday applications of probability theory. However, as we have just seen, it does have the limitation that the equiprobability assumption is required for it to be accurate. In this section we describe some important laws of probability that are more *general*—they do not require the validity of the equiprobability assumption to be true.

The Addition Rule for Mutually Exclusive Events

Suppose you throw a die. To make life more interesting, suppose that the die is not fair, but that the probabilities of the various faces coming up are:

p(1) = .10

p(2) = .10

p(3) = .15

p(4) = .15

p(5) = .20

p(6) = .30

which means that the equiprobability assumption is violated and Equation 1.1 can't be used. What is the probability that the die comes up an even number?

Note first that the probability we seek is of the *union* of three events, i.e., letting "E" be the event, "die comes up an even number,"

$p(E) = p(2 \cup 4 \cup 6)$

Note that the three events, die comes up 2, 4, or 6 are *mutually exclusive*, i.e., if one of them (e.g., die comes up 4) occurs then neither of the others can occur.

To compute the probability of the union of mutually exclusive events, we simply add the individual probabilities, i.e., in general, for N mutually exclusive events,

$$p(A \cup B \cup C \cup \ldots \cup N) = p(A) + p(B) + p(C) + \ldots + p(N) \qquad \text{Equation 1.2}$$

which makes the answer to our problem easy to compute:

$$p(E) = p(2 \cup 4 \cup 6) = p(2) + p(4) + p(6) = .10 + .15 + .30 = \mathbf{.55}$$

To recapitulate: Equation 1.2 can be used whenever one is trying to compute of an event that is itself the union of mutually exclusive events.

And if the Events are *not* Mutually Exclusive?

Then things can become complicated particularly when there are more than two events. However, in the common situation where there are only two events to contend with, it's still relatively simple. Suppose, for example, we pick a playing card from an elderly deck in which 7 of the cards are missing leaving only 45 rather than the standard 52. In particular, for the four suits,

Clubs:	All 13 cards are there; thus,	p(C) = 13/45 = .289
Diamonds:	Missing, the 2, 3, 4, and 10; thus,	p(D) = 9/45 = .200
Hearts:	Missing the Jack and King; thus,	p(H) = 11/45 = .244
Spades:	Missing the Ace; thus,	p(S) = 12/45 = .267

and other than that, the deck is normal and well-shuffled. Now we ask the question: what is the probability that the card we select is a Heart, or a Face card (or both)? Thus, we wish to compute p(H ∪ F) where H is the event, "Heart picked" and F is the event, "Face card picked". The relevant equation is, in general for events A and B,

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \qquad \text{Equation 1.3}$$

so in this example,

$$p(H \cup F) = p(H) + p(F) - p(H \cap F)$$

We already know that p(H) = 11/45. How about p(F)? Among the 45 cards, only the two face cards from the Heart suit are missing, so there are 10 face cards remaining and p(F) = 10/45. Finally, p(H ∩ F) = 2/45—the probability of selecting the one remaining Heart face card, the Queen. So,

$$p(H \cup F) = 11/45 + 10/45 - 1/45 = 20/45 = .444$$

It is useful to understand how Equation 1.3 works from a slightly different perspective. There are f(H ∪ F) = 20 cards in the deck that fall into the union, (H ∪ F)—the 11 hearts plus the 9 face cards from the other three suits. So p(H ∪ F) can be computed as,

$$p(H \cup F) = \frac{f(H \cup F)}{f(S)} = \frac{20}{45} = .444$$

Now notice that when you add p(H) = 11/45 to p(F) = 10/45, you've counted the probability of the intersection, i.e., the probability of selecting the Queen of Hearts, twice: once as part of p(H) and again as part of p(F). So you have to subtract it off to get the correct answer.

Before leaving this topic, notice one more thing about Equation 1.3: if A and B are mutually exclusive, then f(A ∩ B) = 0 and p(A ∩ B) = 0 as well, which means that Equation 1.3 would reduce to,

$$p(A \cup B) = p(A) + p(B),$$

as, by Equation 1.2, it is supposed to for two mutually exclusive events. So Equation 1.3 is a general equation, that applies to any two events, no matter how they relate to one another.

The Complement Rule

Let's go back to our 4-kid example. Suppose we wish to compute the probability that a randomly-selected family has at least one boy, i.e., p(1 boy ∪ 2 boys ∪ 3 boys ∪ 4 boys). We could use Equation 1.2, along with our sample space of all f(S) = 16 possible outcomes back on p. xx to add up the probabilities of the four mutually exclusive events whose union we're seeking i.e.,

$$p(\text{at least one boy}) = p(1\text{ boy}) + p(2\text{ boys}) + p(3\text{ boys}) + p(4\text{ boys})$$

which, using the sample space on p. xx, we can compute to be,

$$p(\text{at least one boy}) = 4/16 + 6/16 + 4/16 + 1/16 = 15/16 = .937$$

However, a simpler way of doing this would be to consider the *complement* of the event we're seeking. Note that,

$$p(A) = \frac{f(A)}{f(S)}$$

and that

$$p(\overline{A}) = \frac{f(\overline{A})}{f(S)}$$

Now because $f(\overline{A}) = f(S) - f(A)$,

$$p(\overline{A}) = \frac{f(S-A)}{f(S)} = \frac{f(S)-f(A)}{f(S)} = 1 - \frac{f(A)}{f(S)} = 1 - p(A)$$

or rearranging terms,

$$p(A) = 1 - p(\overline{A}) \qquad \text{Equation 1.4}$$

Equation 1.4 is quite useful because it's often easier to compute the probability of some event's complement than to compute the probability of the event itself. Returning to our example problems, if Event A that we're seeking is "at least one boy" then its complement, Event $\overline{A}$, would be "all girls". It's easy to see that p(all girls) = 1/16, so p(at least one boy) would be, (1 – 1/16) = 15/16 = .937, as we already worked out via the more complicated sum-of-mutually-exclusive-events route.

Conditional Probability

Remember Estatia? Suppose that the Estatian department of labor statistics (EDLS) is investigating employment patterns in the small Estatian town of Eastwich. The Eastwich labor force, like Eastwich itself, is small, consisting only of 250 people. The EDLS initially categorizes these 250 workers into Set E, those who are employed versus Set $\overline{E}$, those who are unemployed. Additionally, the EDLS categorizes the 250 workers into Set B, those who are blue-collar workers versus Set $\overline{B}$, those who are white-collar workers. It turns out that f(E) = 175 employed people and f(B) = 100 blue-collar workers.

A Contingency Table

Table 2.1 is the beginning of a useful representation of many things including the kind of set interactions that we've just described. In it the sample space, S, is represented as a rectangle, the one with the double-line border. At the bottom-right, "f(S) = 250" indicates that there are 250 elementary events in this sample space.

The sample space is then divided in two ways. The two columns partition S into the two sets, B and $\overline{B}$, while the two rows partition S into E and $\overline{E}$. The *marginal* frequencies shown in the

Table 2.1

	B (Blue-collar)	$\overline{B}$ (White-collar)	
E (Employed)	$f(E \cap B) = 50$		$f(E) = 175$
$\overline{E}$ (Unemployed)			$f(\overline{E}) = 75$
	$f(B) = 100$	$f(\overline{B}) = 150$	$f(S) = 250$

far-right column and the bottom row represent how the f(S) = 250 people are divided up according to employment versus non-employment (f(E) and f($\overline{E}$) and according to blue-collar versus white-collar (f(B) and f($\overline{B}$). Note how both the column and row marginal frequencies add to the total frequency, f(S) = 250. Note also that this information is represented *spatially* (which is to say *intuitively*) in that the employed people can be thought of as occupying the top row, the blue-collar people can be thought of as occupying the left column, and so on.

In Table 2.1, we have also provided one new piece of information—the number of employed blue-collar people, i.e., f(E ∩ B) = 50. Again we can think spatially; these employed, blue-collar people occupy the top-right *cell* of the contingency table. Note also that the term *intersection* also has a spatial meaning: f(E ∩ B) is literally the cell comprising the intersection of employed people and blue-collar people.

Having provided f(E ∩ B), we can now compute the frequencies of the remaining cells in the table as shown in the more complete Table 2.2. So, for example, if there are 175 employed people and 50 of them are blue-collar, then the remaining 125 must be white-collar, which allows us to fill in the intersection of Employed and Blue-collar with f(E ∩ $\overline{B}$) = 125. And so on for the remaining two cells.

Table 2.2

	B (Blue-collar)	$\overline{B}$ (White-collar)	
E (Employed)	f(E∩B) = 50	f(E∩ $\overline{B}$) = 125	f(E) = 175
$\overline{E}$ (Unemployed)	f($\overline{E}$∩B) = 50	f($\overline{E}$∩ $\overline{B}$) = 25	f($\overline{E}$) = 75
	f(B) = 100	f($\overline{B}$) = 150	f(S) = 250

Inspection of Table 2.2 quickly shows that everything adds up very neatly: the two cell frequencies in each column add to the column frequency; the two cell frequencies in each row add to the row frequency, and all four cell frequencies add to f(S) = 250. Note that providing just one cell frequency was sufficient to fill in to total table: as it happened we provided f(E ∩ B) = 50, but we *could* have provided any one of the four cell frequencies.

Frequencies to Probabilities

As is evident Table 2.2 describes the situation we're interested in with respect to *frequencies*: it is descriptive in that it tells us how many people are in which of the various sets. But we can think of the same information in a somewhat different way. Suppose that we were to randomly select a person from the workforce—and by "randomly select" we mean that each of the f(S) = 250 people has an equal chance of being selected—and ask what is the *probability* that the selected person is in this or that set?

Moving from the Table 2.2 frequencies to such probabilities is simple. Given the assumption of random selection, we can use Equation 2.1 for computing any probability of interest. For instance, the probability that a randomly-selected person is employed is,

$$p(E) = \frac{f(E)}{f(S)} = \frac{100}{250} = .600$$

or, the probability that a randomly-selected person is an unemployed blue-collar worker is,

$$p(\overline{E} \cap B) = \frac{f(\overline{E} \cap B)}{f(S)} = \frac{50}{250} = .200$$

Using similar logic, we can reproduce Table 2.2, expressing everything in terms of probabilities, which is done in Table 2.3. Note that, as in Table 2.2, things sum in an organized way: cell probabilities in each column sum to the column probability, cell probabilities in each row sum to the row probability. Also, the two column marginal probabilities, the two row marginal probabilities, and the four cell probabilities add to 1.0.

Table 2.3

	B (Blue-collar)	$\overline{B}$ (White-collar)	
E (Employed)	$p(E \cap B) = \frac{50}{250} = .20$	$p(E \cap \overline{B}) = \frac{125}{250} = .50$	$p(E) = \frac{125}{250} = .70$
$\overline{E}$ $\overline{E}$ (Unemployed)	$p(\overline{E} \cap B) = \frac{50}{250} = .20$	$p(\overline{E} \cap \overline{B}) = \frac{25}{250} = .10$	$p(\overline{E}) = \frac{75}{250} = .30$
	$p(B) = \frac{100}{250} = .40$	$p(\overline{B}) = \frac{150}{250} = .60$	$p(S) = \frac{250}{250} = 1.00$

Joint Probabilities

Each cell of Table 2.3 contains the probability of an *intersection* of two events—$p(E \cap B)$, $p(E \cap \overline{B})$, etc. The probability of an intersection is known as a *joint probability*, referring to the fact that it is the probability of two things jointly happening.

Conditional versus Unconditional Probability

As we've seen, using Table 2.2, it is easy to compute a probability such as p(E), which is f(E)/f(S) = 175/250 = .70. This kind of probability is referred to as *unconditional probability*—it is simply the overall probability that something occurs.

Now suppose we were to do something a bit different. We again randomly select a person from the Eastwich workforce. Suppose that we learn from this person—say her name is Zelda—that she's a blue-collar worker. Knowing this piece of information about Zelda, we can now ask: what is the probability that *she* is employed?

Look carefully at Table 2.2. Knowing that Zelda is blue-collar means that she is one of the f(B) = 100 people in the left column of Table 2.2. Furthermore, of these f(B) = 100 people, $f(E \cap B) = 50$ of them are employed. This means that the probability that Zelda is employed is 50/100 = .50. This is an example of what is called *conditional probability*: the probability of some event given that something else is true, in this case, the probability of a person's being employed given that the person is blue-collar. Generally, a conditional probability of some outcome A, given that something else, B is true, is designated by p(A|B), i.e., the vertical line, "|" is shorthand for "given." In our example, we are thus talking about the probability, p(E|B).

The Equation for Conditional Probability

In this example, it should be easy to see that an equation for the conditional probability that we sought was, implicitly, computed by the equation,

$$p(E \mid B) = \frac{f(E \cap B)}{f(B)}$$

which is fine. However the way that conditional probability is usually expressed is to divide numerator and denominator by f(S) to get,

$$p(E \mid B) = \frac{[f(E \cap B)/f(S)]}{[f(B)/f(S)]} = \frac{p(E \cap B)}{p(B)}$$

or,

$$p(E \mid B) = \frac{.20}{.40} = .50$$

just as we computed it to be before. So, to summarize, for any two events A and B, the conditional probability of A given B is,

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

Some Examples

Just to make sure that the use of Equation 1.6 is clear, here are some examples of how it is used based on our Eastwich workforce example.

$$p(B \mid E) = \frac{p(B \cap E)}{p(E)} = \frac{.20}{.70} = .29$$

$$p(E \mid \overline{B}) = \frac{p(E \cap \overline{B})}{p(\overline{B})} = \frac{.20}{.40} = .50$$

$$p(\overline{E} \mid \overline{B}) = \frac{p(\overline{E} \cap \overline{B})}{p(\overline{B})} = \frac{.10}{.60} = .17$$

Some Important Properties of Conditional Probability

Here are some important things to keep in mind about the relation between certain pairs of conditional probabilities.

The complementary relation between p(A\|B) and p($\overline{A}$ \|B)

Take a look at two more conditional probabilities that can be computed from Table 2.3. The first is,

$$p(B \mid E) = \frac{p(B \cap E)}{p(E)} = \frac{.20}{.70} = .29$$

and the second is,

$$p(\overline{B} \mid E) = \frac{p(\overline{B} \cap E)}{p(E)} = \frac{.50}{.70} = .71$$

Notice that these two probabilities add to 1.0. In general, the two probabilities, p(A|B) and p($\overline{A}$ |B) must add to 1.0. Intuitively, the reason is this. Both probabilities are restricted to the case when B is true. Given that B is true, either A or $\overline{A}$ must occur. That is, given that B is true, p(A|B) and p($\overline{A}$ |B) are complements of one another.

The non-relation between p(A|B) and p(B|A)

Above, we computed that p(E|B) = .50 and that p(B|E) = .29. These two probabilities don't seem to have much to do with one another—for instance, they're not equal; they don't add to 1.0, etc. This observation is correct: in general there's no necessary relation between the two conditional probabilities, p(A|B) and p(B|A).

There is a reason that this lack of relation is interesting beyond its value in understanding statistics: these two conditional probabilities are frequently confused in real life—which leads to no end of mischief in interpretation of various real-world phenomena. Here are two examples of such confusion.

Example 1: Marijuana as a starter drug. Many years ago, the U.S. Government embarked on quite a crusade to persuade the public that marijuana was a highly dangerous drug. Among the many charges that government agencies leveled against this weed was that it was a *starter drug*, i.e., that marijuana use led inexorably to the use of even more dangerous drugs such as heroin. In support of this assertion, the government pointed to the (correct) fact that a large majority of heroin users had smoked marijuana prior to using heroin. This, they claimed, was proof that marijuana use led to heroin.

Let's couch this situation within the framework of conditional probability. We'll designate event M "uses marijuana" and event H "uses heroin." The correct fact relied upon by the government—that a high percentage or heroin users also used heroin—can be expressed as

p(M|H) is high

However, the *conclusion* reached by the government—that marijuana use leads to heroin use—is expressed by,

p(H|M) is high

which does not necessarily follow because, as we've just seen, the two conditional probabilities p(H|M) and p(M|H) bear no necessary relation to each other.

To be more specific, suppose that in Estatia there are 4,000,000 marijuana users out of the country's 20,000,000 residents, i.e., p(M) = .20. Suppose further that there are 1,200,000 heroin users, i.e., p(H) = .06. Suppose further that of these 1,200,000 heroin users, 1,000,000 of them began by using marijuana, i.e., p(M∩H) = .05. Now we can compute the relevant conditional probabilities. The probability of being a marijuana user given that you're a heroin user is,

$$p(M \mid H) = \frac{p(M \cap H)}{p(H)} = \frac{.05}{.06} = .83$$

while the probability that you're a heroin user given that you're a marijuana user is,

$$p(H \mid M) = \frac{p(M \cap H)}{p(M)} = \frac{.05}{.20} = .25$$

So applied to the Estatian population it's true that heroin users tend to be marijuana users, but not true that marijuana users tend to be heroin users.

Example 2: The case against universal AIDS testing. In Estatia, a test is developed by the Estatian billionaire-philanthropist-medical researcher, Alex d'Irchesse whose purpose is to detect the presence of the AIDS virus. The d'Irchesse test is very good in the following ways. First the test's false-negative rate is zero; that is, the test *never* shows negative if given to a person who in fact has AIDS. Second, the test's false-positive rate is only 5%; that is, if the test is administered to a person who does *not* have AIDS, the test will only register positive 5% of the time. Furthermore, d'Irchesse is willing to finance the test's easy and universal availability to all Estatians. Should a grateful nation accept this offer?

"Why not" would probably be a pretty common answer. The test appears to be almost perfect, it's free, and it's universal. What's not to like? Well, here's the problem. Under reasonable assumptions, a person testing positive on the test would—counterintuitively—actually have a very small probability of having AIDS. Let's get a general idea of why this is, and then we'll proceed to a numerical example. The disconnect between intuition and reality, issues from, as you may have guessed, a confusion between p(A|B) and p(B|A).

Let's let A be the event, "person has AIDS" and P be the event, "person tests positive". The false-positive rate of 1% can now be couched as, $p(P|\overline{A}) = .05$ which, in keeping with the confusion we've been discussing, promotes the misbegotten intuition that, $p(\overline{A}|P)$—the probability that you don't have AIDS given that you test positive—is probably pretty small as well (which, of course, would imply that p(A|P), the probability that you *do* have AIDS given that you test positive must be pretty high). But, as we've been emphasizing, these two probabilities, $p(P|\overline{A})$ and $p(\overline{A}|P)$ bear no necessary relation to each other so this conclusion, intuitive though it might be, isn't valid.

Let's be more specific. As we know from the previous example, there are 20,000,000 people in Estatia. What's more, it's believed that 1%, or 200,000 of them actually have AIDS, i.e.,

$p(A) = .01$

Now what happens with universal testing? First let's calculate how many people will test positive. Such people will include the 200,000 people who actually *do* have AIDS (remember the false-negative rate is zero) plus, because of the 5% false-positive rate, 5% of the remaining 19,800,000 people who don't have AIDS, or 990,000 additional people. So

$f(P) = 200{,}000+990{,}000 = 1{,}190{,}000$ people in all

and so,

$p(P) = 1{,}190{,}000/20{,}000{,}000 = .0595.$

Meanwhile, again because of the zero false-negative rate, the only people in the set, $(P \cap A)$ are the 200,000 people actually afflicted with AIDS, i.e.,

$f(P \cap A) = 200{,}000,$

and so

$p(P \cap A) = 200{,}000/20{,}000{,}000 = .01.$

Armed with these fact, we can, finally, compute that,

$$p(A|P) = \frac{p(A \cap P)}{p(P)} = \frac{.0100}{.0595} = .17$$

which means, because $p(\overline{A}|P) = 1\text{-}p(A|P)$, that $p(\overline{A}|P) = .83$.

So let's just sum up: If you test positive on this very reliable d'Irchesse test, the probability is .17 that you actually have AIDS but *.83 that you are a false positive*. In other words, given that you test positive for AIDS, the probability is greater, by almost a factor of 5, that you don't have AIDS than that you do have AIDS. This is going to unnecessarily scare a lot of people and seems like a persuasive argument against a universal AIDS test.

Is this to say that AIDS tests (or similar tests for the presence of any other disease), unless perfect, should never be used? Well no. Here's why and what can be done with such tests. The basic reason for this problem is that the overall proportion of people who have AIDS—what's known as the *base rate*—is very low, only 1% in the Estatistan population. This is why the number of false positives is so high—5% of the remaining 99% of the people is still a lot of people compared to the number of people who actually have AIDS. So instead of administering the test universally, it can instead be administered only to those individuals who have a relatively

high probability of having AIDS to begin with. Let's suppose, for example, that we consider only the 300,000 Estatians who report both using intravenous drug and engaging in unprotected sex. Suppose that 60% of *these* individuals are estimated to have AIDS. If the d'Irchesse test were given to those individuals only, we could compute (and we urge you to do these computations yourself) that,

$p(A|P) = .97$

$p(\overline{A}|P) = .05$

which is considerably more in keeping with what we would like.

Conditional probability and joint probability are different!

Gabriella, an Estatian high-school student is taking Driver's Education. One day her teacher asserts to the class that, "most accidents occur within 10 kilometers of home" and went on to list the reasons that driving close to home is more dangerous than driving far from home—a driver is lulled by familiarity into not paying attention, is more likely to be driving with distracting friends, etc.

Gabriella was not so sure though that her teacher's conclusion—that driving close to home is more dangerous than driving far from home—necessarily followed from the fact that "most accidents occur within 10 kilometers of home." Most *driving* takes place within 10 kilometers of home," she reasoned, "so maybe that's why most accidents occur there." Enlisting the help of her mother who happened to be a statistician, Gabriella researched Estatian driving and accident records. She discovered, confirming her intuitions, that 85% of Estatian driving trips were to locations within 10 kilometers of home. Digging further, she found that on .8% of all Estatian trips, there is an accident of one sort or another (Estatians are somewhat crazy drivers). Focusing only on these .8% of the trips that resulted in accidents, she nailed down the last fact she needed: they divided themselves into .6% that occurred within 10 kilometers of home and .2% that occurred elsewhere.

Let's translate Gabriella's findings into joint and conditional probabilities. We'll let H be "a trip within 10 miles of home" and A be "have an accident on a trip". Gabriella's discovery that 85% of trips were close to home translates to,

$p(H) = .85$

which means that the probability of a trip far from home is,

$p(\overline{H}) = 1\text{-}p(H) = .15$

We can also translate Gabriella's finding about accidents near and far from home to relevant joint probabilities of…

…an accident close to home: $p(A \cap H) = .006$

and

…an accident far from home: $p(A \cap \overline{H}) = .002$

confirming the fact cited by Gabriella's Driver's Ed teacher: an accident is three times as likely to occur close to home than far from home. But to assess how *dangerous* is driving close to versus far from home, Gabriella calculated *conditional probabilities*, i.e., the probability of an accident *given that* you're close or far from home. That is, we can compare the relevant conditional probabilities of…

…an accident given that you're close to home: $p(A \mid H) = \frac{p(A \cap H)}{p(H)} = \frac{.006}{.850} = .00706$

and

…an accident given that you're far from home: $p(A \mid \overline{H}) = \frac{p(A \cap \overline{H})}{p(H)} = \frac{.002}{.150} = .01333$

In other words, the Driver's Ed instructor, although correct in his fact, was wrong in his conclusion: it's actually almost twice as dangerous driving far from home as driving close to home. The disconnect comes about because of the difference between joint probability (which underlies the fact) and conditional probability (which underlies the conclusion).

Independence

Let's return to Table 2.3 which describes some employment facts in the Estatian town of Eastwich. Suppose that we are interested in the probability that a person is employed given that he or she is white versus blue collar. We can compute these conditional probabilities,

$$p(E \mid B) = \frac{p(E \cap B)}{p(B)} = .2/.4 = 0.50$$

and

$$p(E \mid \overline{B}) = \frac{p(E \cap \overline{B})}{p(\overline{B})} = .5/.7 = 0.71$$

So there seems to be a *connection* between blue/what collar and employment status in that the probability is higher that a white-collar person is employed (0.714) than if a blue-collar person is employed (0.500). This fact can be summarized by saying that employment status *depends on* blue/white collar, i.e., they are *not independent*.

Again, suppose that the town workforce of 250 people is partitioned according to the 175 employed people and the remaining 75 unemployed people. Suppose however, that the other way of dividing up the workforce is by gender and in particular, there are 100 females and 150 males, as shown in Table 2.4. So in terms of the marginal frequencies, Table 2.4 looks exactly like Table 2.2, where blue/white collar is substituted for gender.

Table 2.4

	F (Female)	$\overline{F}$ (Male)	
E (Employed)	$f(E \cap F) = 70$	$f(E \cap \overline{F}) = 105$	$f(E) = 175$
$\overline{E}$ $\overline{E}$ (Unemployed)	$f(U \cap \overline{B}) = 30$	$f(\overline{E} \cap \overline{F}) = 45$	$p(\overline{E}) = 75$
	$f(F) = 100$	$f(\overline{F}) = 150$	$f(S) = 250$

Note, however, that the joint frequencies in Table 2.4 are different from those of Table 2.2 and similarly the when the data are expressed as probabilities (Table 2.5) the four joint probabilities differ from the associated joint probabilities of Table 2.3.

Just as we computed the conditional probabilities of being employed given that a worker is blue- or white-collar, we can compute the probabilities of being employed given that a worker is female or male. These probabilities are,

$$p(E \mid F) = \frac{p(E \cap F)}{p(F)} = .28/.40 = 0.70$$

and

$$p(E \mid \overline{F}) = \frac{p(E \cap \overline{F})}{p(\overline{F})} = .42/.60 = 0.70$$

In other words, there is no connection between gender and employment status: the probability of being employed is the same, 0.7, whether a worker is a female or a male, or simply a randomly picked person from the town, i.e., the *unconditional* employment probability is also 0.7.

Table 2.5

	F (Female)	$\overline{F}$ (Male)	
E (Employed)	$p(E \cap F) = \frac{70}{250} = .28$	$p(E \cap \overline{F}) = \frac{105}{250} = .42$	$p(E) = \frac{125}{250} = .70$
$\overline{E}$ (Unemployed)	$p(U \cap \overline{B}) = \frac{30}{250} = .12$	$p(\overline{E} \cap \overline{F}) = \frac{45}{250} = .18$	$p(\overline{E}) = \frac{75}{250} = .30$
	$p(F) = \frac{100}{250} = .40$	$p(\overline{F}) = \frac{150}{250} = .60$	$p(S) = \frac{250}{250} = 1.00$

Independence Defined

These intuitive examples lead us to a formal defendant of independence: two events, A and B are independent if and only if,

$$p(A \mid B) = p(A \mid \overline{B}) = p(A)$$

The multiplication rule for independent events

This definition of independence allows derivation of another central law of probability: the multiplication rule for independent events. From the equation for independence, and from our basic definition of condition probability, we can deduce that when A and B are independent,

$$p(A \mid B) = p(A) = \frac{p(A \cap B)}{p(B)}$$

or, multiplying both sides of this equation by p(B),

$$p(A \cap B) = p(A)p(B)$$

That is: the probability of the joint event is the *product* of the two individual probabilities.

Empirical versus Theoretical Independence

In the examples we've been describing so far, independence (as in the gender example) or lack of it (as in the race example) have been determined by data. In other words, there was no way of knowing whether independence would hold until we actually went in and looked at the frequencies in the various cells of the contingency tables. These are examples of what we call *empirical independence*.

In contrast, one often begins with the proposition that two events are independent because of *a priori* knowledge of how the world works. If one can safely assume that two events are independent then one can use the multiplication rule to compute joint-event probability. As an example, suppose that a fair coin is flipped twice. What is the probability that both flips come up heads? This probability can be expressed as $p(H1 \cap H2)$ where H1 and H2 are the probabilities of a head coming up on the first and second toss respectively. Now here is the critical thing: we *can*

generally assume that H1 and H2 are independent of one another: in general, there is no reason to expect that the outcome of one toss will affect the outcome of the other. Therefore, under the assumption of independence, we can use the multiplication rule to calculate that,

$$p(H1 \cap H2) = p(H1) \times p(H2) = .5 \times .5 = .25$$

The multiplication rule can be applied to arbitrarily many *mutually independent* events. Suppose that, for example, one flips a coin twice and then throws two dice. What is the probability that both coin tosses will turn up heads *and* that first die throw will turn up a 5 *and* that the second die throw will turn up an odd number? Again we can assume *a priori* that each of these four events is independent of each of the other three, i.e., that none of these events affects any of the other 3, so they are all mutually independent. The multiplication rule can be used to compute the joint probability of all four events: denoting them as H1, H2, D5, and DE,

$$p(H1 \cap H2 \cap D5 \cap DE) = p(H1) \times p(H2) \times p(D5) \times p(DE) = (1/2) \times (1/2) \times (1/6) \times (3/6) = .024$$

Joint Probabilities of any Two Events

The multiplication rule is used pervasively because in many instances, the events of which the joint probability is sought can be construed as mutually independent. But what if events are *not* necessarily independent? In this case, joint probabilities can often be computed because relevant conditional and unconditional probabilities are known. That is, considering two events, A and B, if one knows the two probabilities, p(A) and p(A|B), one can reason as follows: we know that

$$p(A \mid B) = \frac{p(A \cap B)}{p(A)}$$

and therefore, by multiplying both sides of this equation by p(A), we arrive at,

$$p(A \cap B) = p(A)p(B|A)$$

Note, by the way, that if A and B *are* independent, then p(B|A) = p(B) and the equation becomes the multiplication rule for two independent events.

Here are two examples of how this formula can be used.

Example 1: Imagine that there is a seminar consisting of five people: three females and two males. Each week, one person is randomly chosen to lead the seminar, with the provision that different people must lead it in Week 1 and Week 2. Now what is the probability that the seminar leader is a female in both Weeks 1 and 2, a probability that we can denote as, p(F1∩F2)? To answer this question, we reason first that p(F1) = 3/5, as the first week's leader is chosen randomly from the five people. Now given F1—that is, given that a female is chosen to lead the seminar on Week 1—what is the probability that another female is chosen to lead on Week 2? This probability, p(F2|F1), can be calculated by noting that on Week 2, there are only four people under consideration—the two males and the two females who *weren't* the leader on Week 1. Therefore, this probability is p(F2|F1) = 2/4, and the joint probability we seek is,

$$p(F1 \cap F2) = p(F1) \times p(F2|F1) = (3/5) \times (2/4) = 0.30$$

Example 2: Imagine a population of people that is divided into three genetic types: 12% of the population are of Type A, 35% are of Type B, and the remaining 53% are of Type C. Now suppose that a particular disease is contracted by Type A people with a probability of 0.22, by Type B people with a probability of 0.38, and by Type C people with a probability of 0.16. What is the probability that a random person from the population will contract the disease?

To answer this question, we can define some terms and match them with the numbers from the example:

p(D) is the probability of a random person contracting the disease. This, of course, is what we're trying to compute.

The probabilities that a random person will be of one of the three genetic types are, p(A) = .12, p(B) = .35, and p(C) = .53.

The conditional probabilities that people of the three genetic types will contract the disease are, p(D|A) = 0.22, p(D|B) = 0.38, and p(D|C) = 0.16.

Now, we can construe the event, "contracting the disease" as being the union of three mutually exclusive joint events, i.e., a person can by Type-A and contract the disease or can be Type-B and contract the disease or can be Type-C and contract the disease, i.e.,

$$D = (A \cap D) \cup (B \cap D) \cup (C \cap D)$$

By using the addition rule for mutually exclusive events, we arrive at,

$$p(D) = p(A \cap D) + p(B \cap D) + p(C \cap D)$$

and from what we have just learned, each of these three joint probabilities is the product of an unconditional and a conditional probability, i.e.,

$$p(D) = p(A)p(D|A) + p(B)p(D|B) + p(C)p(D|C) = 0.12 \times 0.22 + 0.35 \times 0.38 + 0.53 \times 0.16 = 0.244$$

That is, not quite a quarter of the population is expected to contract the disease.

Bayes Theorem

Earlier in this chapter we pointed out that there is no immediately obvious relation between the two seemingly related probabilities, p(A|B) and p(B|A). However, these two probabilities can be related by an equation as follows. Recall that the equation for conditional probability is,

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

and, as we have just seen, $p(A \cap B) = p(A)p(B|A)$. Combining these two facts, we arrive at,

$$p(A \mid B) = \frac{p(A)p(B \mid A)}{p(B)}$$

which is referred to as "Bayes Theorem."

To see how Bayes Theorem, might be useful, suppose that there has been invented a free and easy-to-administer AIDS test. Suppose that the test is relatively error free in that the true-positive rate is 100%, and the false-positive rate is but 5%. This, of course, means that a person with AIDS has a100% chance of testing positive, while a person who does not have AIDS has but a 5% chance of testing positive. Finally suppose that the prevalence of AIDS in the population, i.e., the probability that a randomly-chosen person has AIDS is one in a hundred.

The question is: should the test be administered to everyone? At first glance, the answer is, "of course, why not?" The intuition is that because the test is relatively error-free, everyone with AIDS will know for sure that they have it and can embark on an appropriate course of medical action, while almost everyone without AIDS will be able to rest easy. The only downside is that a small number—5%—of AIDS-free people will undergo the emotional trauma and needless

medical hassles associated with falsely testing positive and thus thinking that they have AIDS when they don't.

But let's examine this seemingly minto downside in a different way. When we think about it, the 5% probability of an AIDS-free person testing positive is, while somewhat useful, not what we really would like to know. Of primary interest is the conditional probability, p(D|P). That is, if I am a random person and I test positive for AIDS, what's the probability that I actually do have the disease? Intuitively, it would seem that this probability must be pretty small since the test's error rate is so low. But is it? We can use Bayes Theorem to find out.

Define "D" as "has the AIDS disease" and "P" as "tests positive." Translating what we know so far into probability notation, gives us:

Probability that a random person has AIDS: $p(D)=.0100$

Probability of a true positive: $p(P|D)=1.000$

Probability that a random person doesn't have AIDS: $p(\overline{D}) = 1 - p(D) = 1 - .010 = 0.990$

Probability of a false positive: $p(P \mid \overline{D}) = 0.050$

Probability that a random person will test positive: $p(P)=0.595$

Where did this last probability, p(P)=0.595 come from? We reasoned as follows. The probability of testing positive can be partitioned into two joint probabilities that add together: the probability of having AIDS and testing positive and the probability of not having AIDS and testing positive, i.e., $p(P) = p(P \cap D) + p(P \cap \overline{D})$. We can calculate these probabilities separately using what we already know:

$p(P \cap D)=p(D)p(P|D)=0.010 \text{ x } 1.000 = 0.010$

and,

$p(P \cap \overline{D}) = p(\overline{D})p(P \mid \overline{D})=0.990 \text{ x } 0.050 = 0.0495$

So the sum of them is, p(P) = 0.0100+0.0495=0.0595 as indicated above.

Now we can plug all of this into Bayes Theorem:

$$p(D \mid P) = \frac{p(D)p(P \mid D)}{p(P)} = \frac{0.010 \text{ x } 1.000}{0.0595} = 0.168$$

This is a counterintuitive result: despite the test's low error rate, a random person in the population who tests positive for AIDS has only about a 17% chance of actually having AIDS!

Looking at the Bayes-Theorem equation applied to this AIDS-test question provides a bit more insight into what's going on. If we're interested on the probability of the disease given a positive test, then the only part of the population of interest is p(P)=0.0595, the proportion of the population in the equation's denominator that *does* test positive. Because the overall incidence of AIDS in the population is so small—the 1% that shows up in the equation's numerator—the portion of the tests-positive pool that comes from true-positive people (0.01) is small relative to the portion that comes from false-positive people (0.0495).