7

Probability Theory

7.1 Objectives

We study probability for a variety of reasons. First, probability provides a way of systematically and rigorously treating uncertainty. This is an important idea that actually developed rather late in human history. Despite major contributions from ancient and medieval scholars, the core of what we use today was developed in the seventeenth and eighteenth centuries in continental Europe due to an intense interest in gambling by various nobles and the mathematicians they employed. Key scholars of this period included Pascal, Fermat, Jacob Bernoulli, Johann Bernoulli, de Moivre, and later on Euler, Gauss, Lagrange, Poisson, Laplace, and Legendre. See Stigler (1986, 1999) or Dale (1991) for fascinating accounts of this period. In addition, much of the axiomatic rigor and notation we use today is due to Keynes (1921) and Kolmogorov (1933).

Interestingly, humans often think in probabilistic terms (even when not gambling), whether we are conscious of it or not. That is, we decide to cross the street when the probability of being run over by a car is sufficiently low, we go fishing at the lakes where the probability of catching something is sufficiently high, and so on. So, even when people are wholly unfamiliar with the mathematical formalization of probability, there is an inclination to frame uncertain

future events in such terms.

Third, probability theory is a precursor to understanding statistics and various fields of applied mathematics. In fact, probability theory could be described as "mathematical models of uncertain reality" because it supports the use of uncertainty in these fields. So to study quantitative political methodology, game theory, mathematical sociology, and other related social science subfields, it is important to understand probability theory in rigorous notation.

There are actually two interpretations of probability. The idea of **subjective probability** is *individually* defined by the conditions under which a person would make a bet or assume a risk in pursuit of some reward. In other words, probability differs by person but becomes apparent in the terms under which a person is willing to wager. Conversely, **objective probability** is defined as a limiting relative frequency: the long-run behavior of a nondeterministic outcome or just an observed proportion in a population. So objectivity is a function of physical observations over some period of time. In either case, the ideas discussed in this chapter apply equally well to both interpretations of probability.

7.2 Counting Rules and Permutations

It seems strange that there could be different and even complicated ways of counting events or contingencies. Minor complexities occur because there are two different features of counting: whether or not the order of occurrence matters, and whether or not events are counted more than once. Thus, in combining these different considerations there are four basic versions of counting rules that are commonly used in mathematical and statistical problems.

To begin, observe that the number of ways in which n individual units can be ordered is governed by the use of the factorial function from Chapter 1 (page 37):

$$n(n-1)(n-2)\cdots(2)(1) = n!.$$

This makes sense: There are n ways to select the first object in an ordered list, $n-1$ ways to pick the second, and so on, until we have one item left and there is only one way to pick that one item. For example, consider the set $\{A, B, C\}$. There are three (n) ways to pick the first item: A, B, or C. Once we have done this, say we picked C to go first, then there are two ways ($n-1$) to pick the second item: either A or B. After that pick, assume A, then there is only one way to pick the last item ($n-2$): B.

To continue, how do we organize and consider a range of possible choices given a set of characteristics? That is, if we are selecting from a group of people, we can pick male vs. female, young vs. old, college educated vs. non-college educated, and so on. Notice that we are now thinking about ***counting*** objects rather than just ***ordering*** objects as done above. So, given a list of known features, we would like a method for enumerating the possibilities when picking from such a population. Fortunately there is a basic and intuitive theorem that guides such counting possibilities. Intuitively, we want to "cross" each possibility from each characteristic to obtain every possible combination.

The Fundamental Theorem of Counting:

- If there are k distinct decision stages to an operation or process,
- each with its own n_k number of alternatives,
- then there are $\prod_{i=1}^{k} n_k$ possible outcomes.

What this formal language says is that if we have a specific number of individual steps, each of which has some set of alternatives, then the total number of alternatives is the product of those at each step. So for $1, 2, \ldots, k$ different characteristics we multiply the corresponding $n_1, n_2, \ldots, n_k$ number of features.

As a simple example, suppose we consider cards in a deck in terms of suit ($n_1 = 4$) and whether they are face cards ($n_2 = 2$). Thus there are 8 possible countable outcomes defined by crossing [Diamonds, Hearts, Spades, Clubs]

with [Face, NotFace]:

$$\begin{array}{c} \\ \mathbf{F} \\ \mathbf{NF} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{D} & \mathbf{H} & \mathbf{S} & \mathbf{C} \end{array} \\ \begin{pmatrix} F,D & F,H & F,S & F,C \\ NF,D & NF,H & NF,S & NF,C \end{pmatrix} \end{array}.$$

In general, though, we are interested in the number of ways to draw a subset from a larger set. So how many five-card poker hands can be drawn from a 52-card deck? How many ways can we configure a committee out of a larger legislature? And so on. As noted, this counting is done along two criteria: with or without tracking the order of selection, and with or without replacing chosen units back into the pool for future selection. In this way, the general forms of choice rules combine ***ordering*** with ***counting***.

The first, and easiest method, to consider is **ordered, with replacement**. If we have n objects and we want to pick $k < n$ from them, and replace the choice back into the available set each time, then it should be clear that on each iteration there are always n choices. So by the Fundamental Theorem of Counting, the number of choices is the product of k values of n alternatives:

$$n \times n \times \cdots n = n^k,$$

(just as if the factorial ordering rule above did not decrement).

The second most basic approach is **ordered, without replacement**. This is where the ordering principle discussed above comes in more obviously. Suppose again we have n objects and we want to pick $k < n$ from them. There are n ways to pick the first object, $n-1$ ways to pick the second object, $n-2$ ways to pick the third object, and so on until we have k choices. This decrementing of choices differs from the last case because we are not replacing items on each iteration. So the general form of ordered counting, without replacement using the two principles is

$$n \times (n-1) \times (n-2) \times \cdots \times (k+1) \times k = \frac{n!}{(n-k)!},$$

Here the factorial notation saves us a lot of trouble because we can express this list as the difference between $n!$ and the factorial series that starts with $k-1$. So the denominator, $(n-k)!$, strips off terms lower than k in the product.

A slightly more complicated, but very common, form is **unordered, without replacement**. The best way to think of this form is that it is just like ordered without replacement, except that we cannot see the order of picking. For example, if we were picking colored balls out of an urn, then *red,white,red* is equivalent to *red,red,white* and *white,red,red*. Therefore, there are $k!$ fewer choices than with ordered, without replacement since there are $k!$ ways to express this redundancy. So we need only to modify the previous form according to

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

Recall that this is the "choose" notation introduced on page 31 in Chapter 1. The abbreviated notation is handy because unordered, without replacement is an extremely common sampling procedure. We can derive a useful generalization of this idea by first observing that

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

(the proof of this property is a chapter exercise). This form suggests successively peeling off $k-1$ iterates to form a sum:

$$\binom{n}{k} = \sum_{i=0}^{k} \binom{n-1-i}{k-i}.$$

Another generalization of the choose notation is found by observing that we have so far restricted ourselves to only two subgroups: those chosen and those not chosen. If we instead consider J subgroups labeled $k_1, k_2, \ldots, k_J$ with the property that $\sum_{j=1}^{J} k_j = n$, then we get the more general form

$$\frac{n!}{\prod_{j=1}^{J} k_j!} = \binom{n}{k_1}\binom{n-k_1}{k_2}\binom{n-k_1-k_2}{k_3} \cdots \binom{n-k_1-k_2-\cdots-k_{J-2}}{k_{J-1}}\binom{k_J}{k_J},$$

which can be denoted $\binom{n}{k_1,k_2,\ldots,k_J}$.

The final counting method, **unordered, with replacement** is terribly unintuitive. The best way to think of this is that unordered, without replacement needs to be adjusted upward to reflect the increased number of choices. This form is best expressed again using choose notation:

$$\frac{(n+k-1)!}{(n-1)!k!} = \binom{n+k-1}{k}.$$

★ **Example 7.1: Survey Sampling**. Suppose we want to perform a small survey with 15 respondents from a population of 150. How different are our choices with each counting rule? The answer is, quite different:

Ordered, with replacement: $n^k = 150^{15} = 4.378939 \times 10^{32}$

Ordered, without replacement: $\frac{n!}{(n-k)!} = \frac{150!}{135!} = 2.123561 \times 10^{32}$

Unordered, without replacement: $\binom{n}{k} = \binom{150}{15} = 1.623922 \times 10^{20}$.

Unordered, with replacement: $\binom{n+k-1}{k} = \binom{164}{15} = 6.59974 \times 10^{20}$.

So, even though this seems like quite a small survey, there is a wide range of sampling outcomes which can be obtained.

7.2.1 The Binomial Theorem and Pascal's Triangle

The most common mathematical use for the choose notation is in the following theorem, which relates exponentiation with counting.

Binomial Theorem:

- Given any real numbers X and Y and a nonnegative integer n,

$$(X+Y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

An interesting special case occurs when $X = 1$ and $Y = 1$:

$$2^n = \sum_{k=0}^{n} \binom{n}{k},$$

which relates the exponent function to the summed binomial. Euclid (around 300 BC to 260 BC) apparently knew about this theorem for the case $n = 2$ only. The first recorded version of the full Binomial Theorem is found in the 1303 book by the Chinese mathematician Chu Shï-kié, and he speaks of it as being quite well known at the time. The first European appearance of the more general form here was due to Pascal in 1654.

To show how rapidly the binomial expansion increases in polynomial terms, consider the first six values of n:

$$(X+Y)^0 = 1$$
$$(X+Y)^1 = X+Y$$
$$(X+Y)^2 = X^2 + 2XY + Y^2$$
$$(X+Y)^3 = X^3 + 3X^2Y + 3XY^2 + Y^3$$
$$(X+Y)^4 = X^4 + 4X^3Y + 6X^2Y^2 + 4XY^3 + Y^4$$
$$(X+Y)^5 = X^5 + 5X^4Y + 10X^3Y^2 + 10X^2Y^3 + 5XY^4 + Y^5.$$

Note the symmetry of these forms. In fact, if we just display the coefficient values and leave out exponents and variables for the moment, we get **Pascal's Triangle**:

```
                    1
                1       1
            1       2       1
        1       3       3       1
    1       4       6       4       1
1       5       10      10      5       1
```

which gives a handy form for summarizing binomial expansions (it can obviously go on further than shown here). There are many interesting features of

Pascal's Triangle. Any value in the table is the sum of the two values diagonally above. For instance, 10 in the third cell of the bottom row is the sum of the 4 and 6 diagonally above. The sum of the kth row (counting the first row as the zero row) can be calculated by $\sum_{j=0}^{k} \binom{k}{j} = 2^k$. The sum of the diagonals from left to right: $\{1\}, \{1\}, \{1,1\}, \{1,2\}, \{1,3,1\}, \{1,4,3\}, \ldots$, give the Fibonacci numbers (1,2,3,5,8,13,...). If the first element in a row after the 1 is a prime number, then every number in that row is divisible by it (except the leading and trailing 1's). If a row is treated as consecutive digits in a larger number (carrying multidigit numbers over to the left), then each row is a power of 11:

$$1 = 11^0$$
$$11 = 11^1$$
$$121 = 11^2$$
$$1331 = 11^3$$
$$14641 = 11^4$$
$$161051 = 11^5,$$

and these are called the "magic 11's." There are actually many more mathematical properties lurking in Pascal's Triangle, but these are some of the more famous.

7.3 Sets and Operations on Sets

Sets are holding places. A **set** is a bounded collection defined by its contents (or even by its lack thereof) and is usually denoted with curly braces. So the set of even positive integers less than 10 is

$$\{2, 4, 6, 8\}.$$

We can also define sets without necessarily listing all the contents if there is some criteria that defines the contents. For example,

$$\{X : 0 \leq X \leq 10, X \in \Re\}$$

defines the set of all the real numbers between zero and 10 inclusive. We can read this statement as "the set that contains all values labeled X such that X is greater than or equal to zero, less than or equal to 10, and part of the real numbers." Clearly sets with an infinite number of members need to be described in this fashion rather than listed out as above.

The "things" that are contained within a set are called **elements**, and these can be individual units or multiple units. An **event** is any collection of possible outcomes of an experiment, that is, any subset of the full set of possibilities, including the full set itself (actually "event" and "outcome" are used synonymously). So $\{H\}$ and $\{T\}$ are outcomes for a coin flipping experiment, as is $\{H, T\}$. Events and sets are typically, but not necessarily, labeled with capital Roman letters: A, B, T, etc.

Events can be abstract in the sense that they may have not yet happened but are imagined, or outcomes can be concrete in that they are observed: "A occurs." Events are also defined for more than one individual subelement (odd numbers on a die, hearts out of a deck of cards, etc.). Such defined groupings of individual elements constitute an event in the most general sense.

★ **Example 7.2:** **A Single Die**. Throw a single die. The event that an even number appears is the set $A = \{2, 4, 6\}$.

Events can also be referred to when they do *not* happen. For the example above we can say "if the outcome of the die is a 3, then A did not occur."

7.3.1 General Characteristics of Sets

Suppose we conduct some experiment, not in the stereotypical laboratory sense, but in the sense that we roll a die, toss a coin, or spin a pointer. It is useful to have some way of describing not only a single observed outcome, but also the full list of possible outcomes. This motivates the following set definition. The **sample space** $\mathcal{S}$ of a given experiment is the set that consists of all possible outcomes (events) from this experiment. Thus the sample space from flipping

a coin is $\{H, T\}$ (provided that we preclude the possibility that the coin lands on its edge as in the well-known *Twilight Zone* episode).

Sets have different characteristics such as countability and finiteness. A **countable set** is one whose elements can be placed in one-to-one correspondence with the positive integers. A **finite set** has a noninfinite number of contained events. Countability and finiteness (or their opposites) are not contradictory characteristics, as the following examples show.

★ **Example 7.3: Countably Finite Set**. A single throw of a die is a countably finite set,

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

★ **Example 7.4: Multiple Views of Countably Finite**. Tossing a pair of dice is also a countably finite set, but we can consider the sample space in three different ways. If we are just concerned with the sum on the dice (say for a game like craps), the sample space is

$$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

If the individual values matter, then the sample space is extended to a large set:

$$\begin{array}{llllll} \{1,1\}, & \{1,2\}, & \{1,3\}, & \{1,4\}, & \{1,5\}, & \{1,6\}, \\ & \{2,2\}, & \{2,3\}, & \{2,4\}, & \{2,5\}, & \{2,6\}, \\ & & \{3,3\}, & \{3,4\}, & \{3,5\}, & \{3,6\}, \\ & & & \{4,4\}, & \{4,5\}, & \{4,6\}, \\ & & & & \{5,5\}, & \{5,6\}, \\ & & & & & \{6,6\}. \end{array}$$

Also, if we have some way of distinguishing between the two dice, such as using different colors, then the sample space is even larger because it is now

possible to distinguish order:

$$\begin{array}{llllll}
\{1,1\}, & \{1,2\}, & \{1,3\}, & \{1,4\}, & \{1,5\}, & \{1,6\}, \\
\{2,1\}, & \{2,2\}, & \{2,3\}, & \{2,4\}, & \{2,5\}, & \{2,6\}, \\
\{3,1\}, & \{3,2\}, & \{3,3\}, & \{3,4\}, & \{3,5\}, & \{3,6\}, \\
\{4,1\}, & \{4,2\}, & \{4,3\}, & \{4,4\}, & \{4,5\}, & \{4,6\}, \\
\{5,1\}, & \{5,2\}, & \{5,3\}, & \{5,4\}, & \{5,5\}, & \{5,6\}, \\
\{6,1\}, & \{6,2\}, & \{6,3\}, & \{6,4\}, & \{6,5\}, & \{6,6\}.
\end{array}$$

Note that however we define our sample space here, that definition does not affect the probabilistic behavior of the dice. That is, they are not responsive in that they do not change physical behavior due to the game being played.

★ **Example 7.5: Countably Infinite Set**. The number of coin flips until two heads in a row appear is a countably infinite set:

$$\mathcal{S} = \{1, 2, 3, \ldots\}.$$

★ **Example 7.6: Uncountably Infinite Set**. Spin a pointer and look at the angle in radians. Given a hypothetically infinite precision measuring instrument, this is an uncountably infinite set:

$$\mathcal{S} = [0{:}2\pi).$$

We can also define the **cardinality** of a set, which is just the number of elements in the set. The finite set A has cardinality given by $n(A)$, $\bar{A}$, or $\|A\|$, where the first form is preferred. Obviously for finite sets the cardinality is an integer value denoting the quantity of events (exclusive of the null set). There are, unfortunately, several ways that the cardinality of a nonfinite set is denoted. The cardinality of a countably infinite set is denoted by $\aleph_0$ (the Hebrew aleph character with subscript zero), and the cardinality of an uncountably infinite set is denoted similarly by $\aleph_1$.

7.3.2 A Special Set: The Empty Set

One particular kind of set is worth discussing at length because it can seem confusing when encountered for the first time. The **empty set**, or **null set**, is a set with no elements, as the names imply. This seems a little paradoxical since if there is nothing in the set, should not the set simply go away? Actually, we *need* the idea of an empty set to describe certain events that do not exist and, therefore the empty set is a convenient thing to have around. Usually the empty set is denoted with the Greek letter phi: ϕ.

An analogy is helpful here. We can think of a set as a suitcase and the elements in the set are contents like clothes and books. Therefore we can define various events for this set, such as the suitcase has all shirts in it, or some similar statement. Now we take these items out of the suitcase one at a time. When there is only one item left in the set, the set is called a **singleton**. When this last item is removed the suitcase still exists, despite being empty, and it is also available to be filled up again. Thus the suitcase is much like a set and can contain some number of items or simply be empty but still defined. It should be clear, however, that this analogy breaks down in the presence of infinite sets.

7.3.3 Operations on Sets

We can perform basic operations on sets that define new sets or provide arithmetic and boolean (true/false) results. The first idea here is the notion of containment, which specifies that a set is composed entirely of elements of another set. Set A is a **subset** of set B if every element of A is also an element of B. We also say that A is contained in B and denote this as $A \subset B$ or $B \supset A$. Formally,

$$A \subset B \iff \forall X \in A, X \in B,$$

which reads "A is a subset of B if and only if all values X that are in A are also in B." The set A here is a **proper subset** of B if it meets this criteria *and* $A \neq B$. Some authors distinguish proper subsets from the more general kind

where equality is allowed by using $\subset$ to denote only proper subsets and $\subseteq$ to denote the more general kind. Unfortunately this notation is not universal.

Subset notation is handy in many ways. We just talked about two sets being equal, which intuitively means that they must contain exactly the same elements. To formally assert that two sets are equal we need to claim, however, that both $A \subset B$ and $B \subset A$ are true so that the contents of A exactly match the contents of B:

$$A = B \iff A \subset B \text{ and } B \subset A.$$

Sets can be "unioned," meaning that they can be combined to create a set that is the same size or larger. Specifically, the **union** of the sets A and B, $A \cup B$, is the new set that contains all of the elements that belong to either A or B. The key word in this definition is "or," indicating that the new set is inclusive. The union of A and B is the set of elements X whereby

$$A \cup B = \{X : X \in A \text{ or } X \in B\}.$$

The union operator is certainly not confined to two sets, and we can use a modification of the "$\cup$" operator that resembles a summation operator in its application:

$$A_1 \cup A_2 \cup \ldots \cup A_n = \bigcup_{i=1}^{n} A_i.$$

It is sometimes convenient to specify ranges, say for $m < n$, with the union operator:

$$A_1 \cup A_2 \cup \ldots \cup A_m = \bigcup_{i \leq m} A_i.$$

There is an obvious relationship between unions and subsets: An individual set is always a subset of the new set defined by a union with other sets:

$$A_1 \subset A \iff A = \bigcup_{i=1}^{n} A_i,$$

and this clearly works for other constituent sets besides A_1. We can also talk about nested subsets:

$$A_n \uparrow A \Longrightarrow A_1 \subset A_2 \subset \ldots A_n, \text{ where } A = \bigcup_{i=1}^{n} A_i$$

$$A_n \downarrow A \Longrightarrow A_n \subset A_{n-1} \subset \ldots A_1, \text{ where } A = \bigcup_{i=1}^{n} A_i.$$

So, for example, if A_1 is the ranking minority member on the House appropriations committee, A_2 is the minority party membership on the House appropriations committee, A_3 is the minority party membership in the House, A_4 is the full House of Representatives, A_5 is Congress, and A is the government, then we can say $A_n \uparrow A$.

We can also define the **intersection** of sets, which contains only those elements found in both (or all for more than two sets of interest). So $A \cap B$ is the new set that contains all of the elements that belong to A and B. Now the key word in this definition is "and," indicating that the new set is exclusive. So the elements of the intersection do not have the luxury of belonging to one set or the other but must now be a member of both. The intersection of A and B is the set elements X whereby

$$A \cap B = \{X : X \in A \text{ and } X \in B\}.$$

Like the union operator, the intersection operator is not confined to just two sets:

$$A_1 \cap A_2 \cap \ldots \cap A_n = \bigcap_{i=1}^{n} A_i.$$

Again, it is convenient to specify ranges, say for $m < n$, with the intersection operator:

$$A_1 \cap A_2 \cap \ldots \cap A_m = \bigcap_{i \leq m} A_i.$$

Sets also define complementary sets by the definition of their existence. The **complement** of a given set is the set that contains all elements not in the original set. More formally, the complement of A is the set $A^{\complement}$ (sometimes denoted A' or $\bar{A}$) defined by

$$A^{\complement} = \{X : X \notin A\}.$$

A special feature of complementation is the fact that the complement of the null set is the sample space, and vice versa:

$$\phi^{\complement} = \mathcal{S} \quad \text{and} \quad \mathcal{S}^{\complement} = \phi.$$

This is interesting because it highlights the roles that these special sets play: The complement of the set with everything has nothing, and the complement of the set with nothing has everything.

Another common operator is the **difference operator**, which defines which portion of a given set is *not* a member of the other. The difference of A relative to B is the set of elements X whereby

$$A \setminus B = \{X : X \in A \text{ and } X \notin B\}.$$

The difference operator can also be expressed with intersection and complement notation:

$$A \setminus B = A \cap B^{\complement}.$$

Note that the difference operator as defined here is not symmetric: It is not necessarily true that $A \setminus B = B \setminus A$. There is, however, another version called the **symmetric difference** that further restricts the resulting set, requiring the operator to apply in both directions. The symmetric difference of A relative to B and B relative to A is the set

$$A \triangle B = \{X : X \in A \text{ and } X \notin B \text{ or } X \in B \text{ and } B \notin A\}.$$

Because of this symmetry we can also denote the symmetric difference as the union of two "regular" differences:

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cap B^{\complement}) \cup (B \cap A^{\complement}).$$

★ **Example 7.7: Single Die Experiment**. Throw a single die. For this experiment, define the following sample space and accompanying sets:

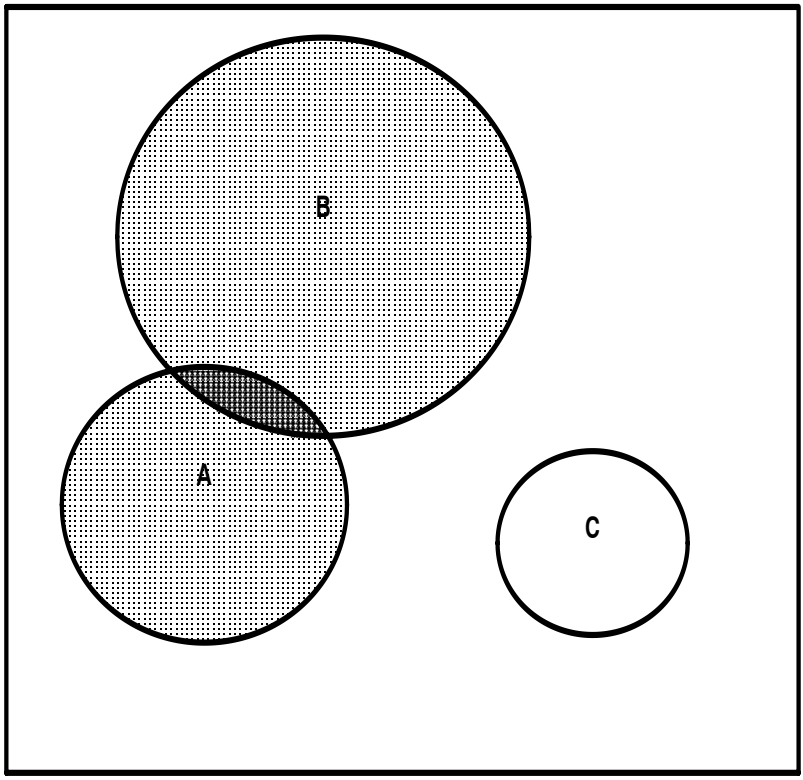
$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$
$$A = \{2, 4, 6\}$$
$$B = \{4, 5, 6\}$$
$$C = \{1\}.$$

So A is the set of even numbers, B is the set of numbers greater than 3, and C has just a single element. Using the described operators, we find that

$$A^{\complement} = \{1, 3, 5\} \quad A \cup B = \{2, 4, 5, 6\} \quad A \cap B = \{4, 6\}$$
$$B \cap C = \phi \quad (A \cap B)^{\complement} = \{1, 2, 3, 5\} \quad A \setminus B = \{2\}$$
$$B \setminus A = \{5\} \quad A \triangle B = \{2, 5\} \quad (A \cap B) \cup C = \{1, 4, 6\}.$$

Figure 7.1 illustrates set operators using a **Venn diagram** of three sets where the "universe" of possible outcomes ($\mathcal{S}$) is given by the surrounding box. Venn diagrams are useful tools for describing sets in a two-dimensional graph. The intersection of A and B is the dark region that belongs to both sets, whereas the union of A and B is the lightly shaded region that indicates elements in A or B (including the intersection region). Note that the intersection of A or B with C is ϕ, since there is no overlap. We could, however, consider the nonempty sets $A \cup C$ and $B \cup C$. The complement of $A \cup B$ is all of the nonshaded region, including C. Consider the more interesting region $(A \cap B)^{\complement}$. This would be every part of $\mathcal{S}$ *except* the intersection, which could also be expressed as those elements that are in the complement of A *or* the complement of B, thus ruling out the intersection (one of de Morgan's Laws; see below). The portion of A

Fig. 7.1. Three Sets



that does not overlap with B is denoted $A \backslash B$, and we can also identify $A \triangle B$ in the figure, which is either A or B (the full circles) but not both.

There are formal properties for unions, intersections, and complement operators to consider.

Properties For Any Three Sets A, B, and C, in $\mathcal{S}$

$\rightarrowtail$	Commutative Property	$A \cup B = B \cup A$
		$A \cap B = B \cap A$
$\rightarrowtail$	Associative Property	$A \cup (B \cup C) = (A \cup B) \cup C$
		$A \cap (B \cap C) = (A \cap B) \cap C$
$\rightarrowtail$	Distributive Property	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
		$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
$\rightarrowtail$	de Morgan's Laws	$(A \cup B)^{\complement} = A^{\complement} \cap B^{\complement}$
		$(A \cap B)^{\complement} = A^{\complement} \cup B^{\complement}$

As an illustration we will prove $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ by demonstrating inclusion (subsetting) in both directions to establish the equality.

- First show $A \cup (B \cap C) \subset (A \cup B) \cap (A \cup C)$ by demonstrating that some element in the first set is also in the second set:

  Suppose $X \in A \cup (B \cap C)$, so $X \in A$ or $X \in (B \cap C)$

  If $X \in A$, then $X \in (A \cup B)$ and $X \in (A \cup C)$

  $\therefore X \in (A \cup B) \cap (A \cup C)$

  Or if $X \notin A$, then $X \in (B \cap C)$ and $X \in B$ and $X \in C$

  $\therefore X \in (A \cup B) \cap (A \cup C)$

- Now show $A \cup (B \cap C) \supset (A \cup B) \cap (A \cup C)$ by demonstrating that some

element in the second set is also in the first set:

Suppose $X \in (A \cup B) \cap (A \cup C)$ so $X \in (A \cup B)$ and $X \in (A \cup C)$

If $X \in A$, then $X \in A \cup (B \cap C)$

Or if $X \notin A$, then $X \in B$ and $X \in C$,

since it is in the two unions but not in A.

$\therefore X \in A \cup (B \cap C)$

- Since $A \cup (B \cap C) \subset (A \cup B) \cap (A \cup C)$ *and* $A \cup (B \cap C) \supset (A \cup B) \cap (A \cup C)$, then every element in the first set is in the second set, and every element in the second set is in the first set. So the sets must be equal.

The case where two sets do not have an intersection (say A and C in Figure 7.1) is important enough that it has a special name. Two sets A and B are **disjoint** when their intersection is empty: $A \cap B = \phi$. This is generalizable as well. The k sets $A_1, A_2, \ldots, A_k$ are **pairwise disjoint**, also called **mutually exclusive**, if $A_i \cap A_j = \phi \ \forall i \neq j$. In addition, if $A_1, A_2, \ldots, A_k$ are pairwise disjoint and we add the condition that $\bigcup_{i=1}^{k} A_i = \mathcal{S}$ (i.e., that they cover the sample space completely), then we say that the $A_1, A_2, \ldots, A_k$ are a **partition** of the sample space. For instance, the outcomes $\{1, 2, 3, 4, 5, 6\}$ form a partition of $\mathcal{S}$ for throwing a single die because they are pairwise distinct and nothing else can occur. More formally, $A_1, A_2, \ldots, A_k$ are a partition of $\mathcal{S}$ iff

- $A_i \cap A_j = \phi \ \forall i \neq j$.
- $\bigcup_{i=1}^{k} A_i = \mathcal{S}$.

★ **Example 7.8: Overlapping Group Memberships**. Sociologists are often interested in determining connections between distinct social networks. Bonacich (1978) used set theory to explore overlapping group memberships such as sports teams, clubs, and social groups in a high school. The data are given by the following cross-listing of 18 community members and 14 social events they could possibly have attended. An "X" indicated that the

individual on that row attended the social event indexed by the column. The idea is that the more social events two selected individuals have in common, the closer they are in a social network.

	Event													
Ind.	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*
1	X	X	X	X	X	X		X	X					
2	X	X	X		X		X	X						
3		X	X	X	X	X	X	X	X					
4	X		X	X	X	X	X	X						
5			X	X	X		X							
6			X		X	X		X						
7					X	X	X	X						
8						X		X	X					
9					X		X	X	X					
10							X	X	X			X		
11								X	X	X		X		
12								X	X	X		X	X	X
13							X	X	X	X		X	X	X
14						X	X		X	X	X	X	X	X
15							X	X		X	X	X		
16								X	X					
17									X		X			
18									X		X			

We can see that there are two relatively distinct groups here, with reasonable overlap to complicate things. Counting the social events as sets, we can ask some specific questions and make some observations. First, observe that only $M = N$; they have the same members: $\{12, 13, 14\}$. A number of sets are disjoint, such as (A, J), (B, L), (D, M), and others. Yet, the full group of sets, $A{:}N$, clearly does not form a partition due to the many nonempty intersections. In fact, there is no subset of the social events that forms a partition. How do we know this? Consider that either I or K would have to be included in the formed partition because they are the only two that include individuals 17 and 18. The set K lacks individual 16, necessitating inclusion of H or I, but these overlap with K elsewhere. Similarly, I lacks individual 15, but each of the five sets that include this individual overlap somewhere

with I. Thus no subset can be configured to form a partition. Let's now test out the first de Morgan's Law for the sets E and G:

$$\begin{aligned}(E \cup G)^{\complement} &= (\{1,2,3,4,5,6,7,9\} \cup \{2,3,4,5,7,9,10,13,14,15\})^{\complement} \\ &= (\{1,2,3,4,5,6,7,9,10,13,14,15\})^{\complement} \\ &= \{8,11,12,16,17,18\} \\ E^{\complement} \cap G^{\complement} &= \{1,2,3,4,5,6,7,9\}^{\complement} \cap \{2,3,4,5,7,9,10,13,14,15\}^{\complement} \\ &= \{8,10,11,12,13,14,15,16,17,18\} \\ &\qquad \cap \{1,6,8,11,12,16,17,18\} \\ &= \{8,11,12,,16,17,18\},\end{aligned}$$

which demonstrates the property.

★ **Example 7.9: Approval Voting**. This example derives from the review of formal models of voting in Gill and Gainous (2002). In approval voting, voters are allowed to vote for (approve of) as many candidates as they want but cannot cast more than one vote for each candidate. Then the candidate with the most of these approval votes wins the election. Obviously this system gives voters a wide range of strategies, and these can be analyzed with a formal (mathematical) model.

Given $K \geq 3$ candidates, it appears that there are 2^K possible strategies, from the counting rules in Section 7.2, but because an abstention has the same net effect as voting for every candidate on the ballot, the actual number of different choices is $2^K - 1$. We can formalize approval voting as follows:

- Let w, x, y, z be the individual candidates from which a group of voters can choose, and let wPx represent a given voter's strict preference for w over x. A multicandidate strict preference order is denoted $wPxPyPz$.

- No preference between w and x is denoted by wIx, meaning that the voter is indifferent or ambivalent between the two.
- Strict preference and indifference have transitive relations, $wPx, xPy \rightarrow wPz$ and $wIx, xIy \rightarrow wIy$.
- Every possible grouped ordering can be separated into ℓ nonempty subsets where

$$[w_1, w_2, \ldots], [x_1, x_2, \ldots], [y_1, y_2, \ldots], [z_1, z_2, \ldots], \ldots, [\ell_1, \ell_2, \ldots]$$

, denoted as $W, X, Y, Z, \ldots, L$, and the voter is indifferent among the candidates within any single such subset while still strictly preferring every member of that subset to any of the other candidate subsets lower in the preference ordering.

When $\ell = 1$, the voter is called *unconcerned* and has no strict preference between any candidates. If $\ell = 2$, then the voter is called *dichotomous*, *trichotomous* if $\ell = 3$, and finally *multichotomous* if $\ell \geq 4$. If all voters have a dichotomous preference, then an approval voting system always produces an election result that is majority preferred, but when all preferences are not dichotomous, the result can be different. In such cases there are multiple *admissible voter strategies*, meaning a strategy that conforms to the available options among k alternatives and is not uniformly dominated (preferred in all aspects by the voter) by another alternative.

As an example, the preference order wPx with xPy has two admissible sincere strategies where the voter may have given an approval vote for only the top alternative w or for the two top alternatives w, x. Also, with multiple alternatives it is possible for voters to cast *insincere* (strategic) votes: With $wPxPy$ she prefers candidate w but might select only candidate x to make x close to w without helping y.

For two given subsets A and B, define the union $A \cup B = \{a : a \in A \text{ or } a \in B\}$. A subset that contains only candidate w is denoted as $\{w\}$, the subset that contains only candidate x is denoted as $\{x\}$, the subset containing only

candidates w and x is denoted as $\{w, x\}$, and so on. A *strategy*, denoted by S, is defined as voting for some specified set of candidates regardless of actual approval or disapproval. Now consider the following set-based assumptions for a hypothetical voter:

- P: If wPx, then $\{w\}P\{w, x\}P\{x\}$.
- I: If $A \cup B$ and $B \cup C$ are nonempty, and if wIx, xIy, and wIy for all $w \in A$, $x \in B$, $y \in C$, then $(A \cup B)I(B \cup C)$.
- $M(P) = A_1$ is the subset of the most-preferred candidates under P, and $L(P) = A_n$, the subset of the least-preferred candidates under P.

Suppose we look once again at the voter who has the preference order $wPxPyPz$, while all other voters have dichotomous preferences, with some being sequentially indifferent (such as wIx and yIz), and some strictly prefer w and x to y and z, while the rest prefer y and z to w and x. Each of the other voters uses their unique admissible strategy, so that the aggregated preference for w is equal to that of x ,$f(w) = f(x)$, and the aggregated preference for y is equal to that of z, $f(y) = f(z)$. Now assume that the voter with preference $wPxPyPz$ is convinced that there is at least a one-vote difference between w and y, $f(w) \geq f(y) + 1$; therefore, $\{w, y\}$ is a good strategy for this voter because a vote for w ensures that w will receive at least one more vote than x, and a vote for y ensures that y will receive at least one more vote than z. Therefore, $\{w, y\}$ ensures that the $wPxPyPz$ voter's most-preferred candidate gets the greatest votes and $wPxPyPz$ voter's least-preferred candidate gets the fewest votes.

7.4 The Probability Function

The idea of a probability function is very basic and very important. It is a mapping from a defined event (or events) onto a metric bounded by zero (it cannot happen) and one (it will happen with absolute certainty). Thus a probability function enables us to discuss various *degrees of likelihood of occurrence* in a

systematic and practical way. Some of the language here is a bit formal, but it is important to discuss probability using the terminology in which it was codified so that we can be precise about specific meanings.

A collection of subsets of the sample space $\mathcal{S}$ is called a **sigma-algebra** (also called a **sigma-field**), and denoted $\mathfrak{F}$ (a fancy looking "F"), if it satisfies the following three properties:

(i) **Null Set Inclusion.** It contains the null set: $\phi \in \mathfrak{F}$.

(ii) **Closed Under Complementation.** If $A \in \mathfrak{F}$, then $A^{\complement} \in \mathfrak{F}$.

(iii) **Closed Under Countable Unions.** If $A_1, A_2, \ldots \in \mathfrak{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathfrak{F}$.

So if A is any identified subset of $\mathcal{S}$, then an associated (minimal size) sigma-algebra is $\mathfrak{F} = \{\phi, A, A^{\complement}, \mathcal{S}\}$. Why do we have these particular elements? We need ϕ in there due to the first condition, and we have identified A as a subset. So by the second condition we need $\mathcal{S}$ and $A^{\complement}$. Finally, does taking unions of any of these events ever take us out of $\mathcal{S}$. Clearly not, so this is a sigma-algebra. Interesting enough, so is $\mathfrak{F}' = \{\phi, A, A^{\complement}, A, A, \mathcal{S}, A^{\complement}\}$ because there is no requirement that we not repeat events in a sigma-algebra. But this is not terribly useful, so it is common to specify the minimal size sigma-algebra as we have originally done. In fact such a sigma-algebra has a particular name: a **Borel-field**. These definitions are of course inherently discrete in measure. They do have corresponding versions over continuous intervals, although the associated mathematics get much more involved [see Billingsley (1995) or Chung (2000) for definitive introductions].

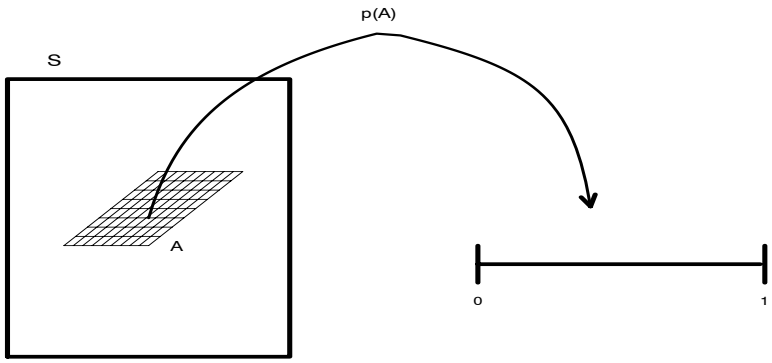
★ **Example 7.10: Single Coin Flip**. For this experiment, flip a coin once.

This produces

$$\mathcal{S} = \{H, T\}$$
$$\mathfrak{F} = \{\phi, H, T, (H, T)\}.$$

Given a sample space $\mathcal{S}$ and an associated sigma-algebra $\mathfrak{F}$, a **probability function** is a mapping, p, from the domain defined by $\mathfrak{F}$ to the interval $[0:1]$. This is shown in Figure 7.2 for an event labeled A in the sample space $\mathcal{S}$.

Fig. 7.2. The Mapping of a Probability Function



The **Kolmogorov probability axioms** specify the conditions for a proper probability function:

- The probability of any realizable event is between zero and one: $p(A_i) \in [0{:}1] \ \ \forall a_i \in \mathfrak{F}$.
- Something happens with probability one: $p(\mathcal{S}) = 1$.
- The probability of unions of n pairwise disjoint events is the sum of their individual probabilities: $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} p(A_i)$ (even if $n = \infty$).

It is common to identify an experiment or other probabilistic setup with the **triple** (also called a probability space or a probability measure space) consisting of $(\mathcal{S}, \mathfrak{F}, P)$, to fully specify the sample space, sigma-algebra, and probability function applied.

7.5 Calculations with Probabilities

The manipulation of probability functions follows logical and predictable rules. The probability of a union of two sets is no smaller than the probability of an intersection of two sets. These two probabilities are equal if one set is a subset of another. It also makes intuitive sense that subsets have no greater probability

than the enclosing set:

$$\text{If } A \subset B, \text{ then } p(A) \leq p(B).$$

The general rules for probability calculation are straightforward:

Calculations with Probabilities for A, B, and C, in $\mathcal{S}$

$\rightarrowtail$	Probability of Unions	$p(A \cup B)$ $= p(A) + p(B) - p(A \cap B)$
$\rightarrowtail$	Probability of Intersections	$p(A \cap B)$ $= p(A) + p(B) - p(A \cup B)$ (also denoted $p(A, B)$)
$\rightarrowtail$	Probability of Complements	$p(A^{\complement}) = 1 - p(A)$, $p(A) = 1 - p(A^{\complement})$
$\rightarrowtail$	Probability of the Null Set	$p(\phi) = 0$
$\rightarrowtail$	Probability of the Sample Space	$p(\mathcal{S}) = 1$
$\rightarrowtail$	Boole's Inequality	$p(\bigcup_j A_j) \leq \sum_j p(A_j)$

Either of the first two rules can also be restated as $p(A \cup B) + p(A \cap B) = p(A) + p(B)$, which shows that the intersection is "double-counted" with naive addition. Note also that the probability of the intersection of A and B is also called the joint probability of the two events and denoted $p(A, B)$.

We can also now state a key result that is quite useful in these types of calculations.

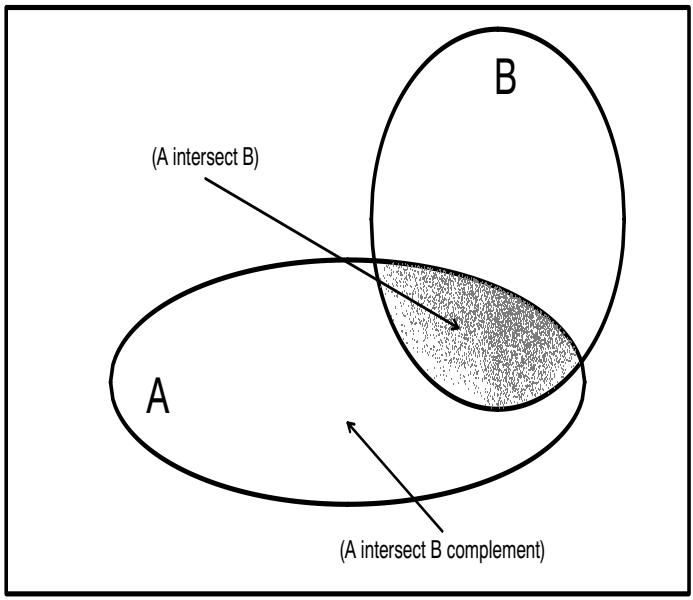
The Theorem of Total Probability:

- Given any events A and B,

- $p(A) = p(A \cap B) + p(A \cap B^{\complement})$.

This intuitively says that the probability of an event A can be decomposed into to parts: one that intersects with another set B and the other that intersects with the complement of B, as shown in Figure 7.3. If there is no intersection or if B is a subset of A, then one of the two parts has probability zero.

Fig. 7.3. Theorem of Total Probability Illustrated



More generally, if $B_1, B_2, \ldots, B_n$ is a partition of the sample space, then

$$p(A) = p(A \cap B_1) + p(A \cap B_2) + \ldots + p(A \cap B_n).$$

★ **Example 7.11: Probabilistic Analysis of Supreme Court Decisions**. Probability statements can be enormously useful in political science research. Since political actors are rarely deterministic enough to predict with certainty, using probabilities to describe potential events or actions provides a means of making claims that include uncertainty.

Jeffrey Segal (1984) looked at Supreme Court decisions to review search and seizure cases from lower courts. He constructed a model using data from all 123 Fourth Amendment cases from 1962 to 1981 to explain why the Court upheld the lower court ruling versus overturning it. The objective was to make probabilistic statements about Supreme Court decisions given specific aspects of the case and therefore to make predictive claims about future actions. Since his multivariate statistical model simultaneously incorporates all these variables, the probabilities described are the effects of individual variables holding the effects of all others constant.

One of his first findings was that a police search has a 0.85 probability of being upheld by the Court if it took place at the home of another person and only a 0.10 probability of being upheld in the detainee's own home. This is a dramatic difference in probability terms and reveals considerable information about the thinking of the Court. Another notable difference occurs when the search takes place with no property interest versus a search on the actual person: 0.85 compared to 0.41. Relatedly, a "stop and frisk" search case has a 0.70 probability of being upheld whereas a full personal search has a probability of 0.40 of being upheld. These probabilistic findings point to an underlying distinction that justices make in terms of the personal context of the search.

Segal also found differences with regard to police possession of a warrant or probable cause. A search sanctioned by a warrant had a 0.85 probability of being upheld but only a 0.50 probability in the absence of such prior authority. The probability that the Court would uphold probable cause searches (where the police notice some evidence of illegality) was 0.65, whereas those that were not probable cause searches were upheld by the Court with probability 0.53. This is not a great difference, and Segal pointed out that it is confounded with other criteria that affect the overall reasonableness of the search. One such criteria noted is the status of the arrest. If the search is performed subject to a lawful arrest, then there is a (quite impressive) 0.99 probability of being

upheld, but only a 0.50 probability if there is no arrest, and all the way down to 0.28 if there is an unlawful arrest.

What is impressive and useful about the approach taken in this work is that the author translates extensive case study into probability statements that are intuitive to readers. By making such statements, underlying patterns of judicial thought on Fourth Amendment issues are revealed.

7.6 Conditional Probability and Bayes Law

Conditional probability statements recognize that some prior information bears on the determination of subsequent probabilities. For instance, a candidate's probability of winning office are almost certain to change if the opponent suffers a major scandal or drops out of the race. We would not want to ignore information that alters probability statements and **conditional probability** provides a means of systematically including other information by changing "$p(A)$" to "$p(A|B)$" to mean the probability that A occurs given that B has occurred.

★ **Example 7.12: Updating Probability Statements**. Suppose a single die is rolled but it cannot be seen. The probability that the upward face is a four is obviously one-sixth, $p(x = 4) = \frac{1}{6}$. Further suppose that you are told that the value is greater than three. Would you revise your probability statement? Obviously it would be appropriate to update since there are now only three possible outcomes, one of which is a four. This gives $p(x = 4|x > 3) = \frac{1}{3}$, which is a substantially different statement.

There is a more formal means of determining conditional probabilities. Given two outcomes A and B in $\mathcal{S}$, the probability that A occurs given that B occurs is the probability that A and B both occur divided by the probability that B occurs:

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

provided that $p(B) \neq 0$.

★ **Example 7.13: Conditional Probability with Dice**. In rolling two dice labeled X and Y, we are interested in whether the sum of the up faces is four, given that the die labeled X shows a three. The unconditional probability is given by

$$p(X+Y=4) = p(\{1,3\},\{2,2\},\{3,1\}) = \frac{1}{12},$$

since there are 3 defined outcomes here out of 36 total. The conditional probability, however, is given by

$$\begin{aligned} p(X+Y=4|X=3) &= \frac{p(X+Y=4, X=3)}{p(X=3)} \\ &= \frac{p(\{3,1\})}{p(\{3,1\},\{3,2\},\{3,3\}\{3,4\},\{3,5\},\{3,6\})} \\ &= \frac{1}{6}. \end{aligned}$$

We can rearrange $p(A|B) = \frac{p(A\cap B)}{p(B)}$ to get $p(A|B)p(B) = p(A \cap B)$. Similarly, for the set $B^{\complement}$, we get $p(A|B^{\complement})p(B^{\complement}) = p(A \cap B^{\complement})$. For any set B we know that A has two components, one that intersects with B and one that does not (although either could be a null set). So the set A can be expressed as the sum of conditional probabilities:

$$p(A) = p(A|B)p(B) + p(A|B^{\complement})p(B^{\complement}).$$

Thus the Theorem of Total Probability can also be reexpressed in conditional notation, showing that the probability of any event can be decomposed into conditional statements about any other event. It is possible to further extend this with an additional conditional statement. Suppose now that we are interested in decomposing $p(A|C)$ with regard to another event, B and $B^{\complement}$. We start with the definition of conditional probability, expand via the most basic form of the

Theorem of Total Probability, and then simplify:

$$\begin{aligned} p(A|C) &= \frac{p(A \cap C)}{p(C)} \\ &= \frac{p(A \cap B \cap C) + p(A \cap B^{\complement} \cap C)}{p(C)} \\ &= \frac{p(A|B \cap C)p(B \cap C) + p(A|B^{\complement} \cap C)p(B^{\complement} \cap C)}{p(C)} \\ &= p(A|B \cap C)p(B|C) + p(A|B^{\complement} \cap C)p(B^{\complement}|C). \end{aligned}$$

It is important to note here that the conditional probability is order-dependent: $p(A|B) \neq p(B|A)$. As an illustration, apparently in California the probability that a highway motorist was in the left-most lane given they subsequently received a speeding ticket is about 0.93. However, it is certainly not true that the probability that one receives a speeding ticket given they are in the left lane is also 0.93 (or this lane would be quite empty!). But can these conditional probabilities be related somehow?

We can manipulate the conditional probability statements in parallel:

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

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

Wait a minute! We know that $p(A \cap B) = p(B \cap A)$, so we can equate

$$\begin{aligned} p(A|B)p(B) &= p(B|A)p(A) \\ p(A|B) &= \frac{p(A)}{p(B)}p(B|A) \\ &= \frac{p(A)p(B|A)}{p(A)p(B|A) + p(A^{\complement})p(B|A^{\complement})}, \end{aligned}$$

where the last step uses the Total Probability Theorem. This means that we have a way of relating the two conditional probability statements. In fact, this

is so useful that it has a name, **Bayes Law**, for its discoverer, the Reverend Thomas Bayes (published posthumously in 1763).

Any joint probability can be decomposed into a series of conditional probabilities followed by a final unconditional probability using the **multiplication rule**. This is a generalization of the definition of conditional probability. The joint distribution of k events can be reexpressed as

$$p(A_1, A_2, \ldots, A_k) = p(A_k|A_{k-1}, A_{k-2}, \ldots, A_2, A_1) \times p(A_{k-1}|A_{k-2}, \ldots, A_2, A_1) \cdots p(A_3|A_2, A_1)p(A_2|A_1)p(A_1).$$

So we can "reassemble" the joint distribution form starting from the right-hand side using the definition of conditional probability, giving

$$p(A_2|A_1)p(A_1) = p(A_1, A_2)$$
$$p(A_3|A_2, A_1)p(A_2, A_1) = p(A_1, A_2, A_3)$$
$$p(A_4|A_3, A_2, A_1)p(A_3, A_2, A_1) = p(A_1, A_2, A_3, A_4),$$

and so on.

7.6.1 Simpson's Paradox

Sometimes conditioning on another event actually provides *opposite* results from what would normally be expected. Suppose, for example, a state initiated a pilot job training program for welfare recipients with the goal of improving skill levels to presumably increase the chances of obtaining employment for these individuals. The investigators assign half of the group to the job placement program and leave the other half out as a control group. The results for the full group and a breakdown by sex are provided in Table 7.1.

Looking at the full group, those receiving the job training are somewhat more likely to land employment than those who did not. Yet when we look at these same people divided into men and women, the results are the opposite! Now

Table 7.1. ILLUSTRATION OF SIMPSON'S PARADOX: JOB TRAINING

		Job	No Job	Placement Rate
Full Group, $n = 400$	Training	100	100	50%
	No Training	80	120	40%
Men, $n = 200$	Training	90	60	60%
	No Training	35	15	70%
Women, $n = 200$	Training	10	40	20%
	No Training	45	105	30%

it appears that it is better not to participate in the job training program for both sexes. This surprising result is called **Simpson's Paradox**.

How can something that is good for the full group be bad for all of the component subgroups? A necessary condition for this paradox to arise is for Training and Job to be correlated with each other, *and* Male to be correlated with both Training and Job. So more men received job training and more men got jobs. In other words, treatment (placement in the program) is confounded with sex. Therefore the full group analysis "masks" the effect of the treatment by aggregating the confounding effect out. This is also called **aggregation bias** because the average of the group averages is not the average of the full population.

We can also analyze this using the conditional probability version of the Total Probability Theorem. Label the events J for a job, T for training, and M for male. Looking at the table it is easy to observe that the $p(J|T) = 0.5$ since a total of 200 individuals got the training and 100 acquired jobs. Does this comport with the conditioning variable?

$$\begin{aligned} p(J|T) &= p(J|M \cap T)p(M|T) + p(J|M^{\complement} \cap T)p(M^{\complement}|T) \\ &= (0.6)\left(\frac{90+60}{100+100}\right) + (0.2)\left(\frac{10+40}{100+100}\right) \\ &= 0.5. \end{aligned}$$

Thus the explanation for why $p(J|T) > p(J|T^{\complement})$ but $p(J|M \cap T) < p(J|M \cap T^{\complement})$ and $p(J|M^{\complement} \cap T) < p(J|M^{\complement} \cap T^{\complement})$ is the unequal weighting between men and women.

7.7 Independence

In the last section we found that certain events will change the probability of other events and that we are advised to use the conditional probability statement as a way of updating information about a probability of interest. Suppose that the first event does *not* change the probability of the second event. For example, if we observe that someone drives a blue car, it does not change the probability that they will vote for the Republican candidate in the next election. Conversely, if we knew that this person voted for the Republican candidate in the last election, we would certainly want to update our unconditional probability.

So how do we treat the first case when it does not change the subsequent probability of interest? If all we are interested in is the probability of voting Republican in the next election, then it is obviously reasonable to ignore the information about car color and continue to use whatever probability we had originally assigned. But suppose we are interested in the probability that an individual votes Republican (event A) *and* owns a blue car (event B)? This joint probability is just the product of the unconditional probabilities and we say that A and B are **independent** if

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

So the subject's probability of voting for the Republican and driving a blue car is just the probability that she votes for the Republican times the probability that she owns a blue car. Put another way, the intersection occurs by chance, not by some dependent process.

The idea of independence can be generalized to more than two events. A set of events $A_1, A_2, \ldots, A_k$ is **pairwise independent** if

$$p(A_i \cap A_j) = p(A_i)p(A_j) \ \ \forall i \neq j.$$

This means that if we pick any two events out of the set, they are independent of each other. Pairwise independence does *mean* the same thing as general independence, though it is a property now attached to the pairing operation. As an example, Romano and Siegel (1986) give the following three events for two tosses of a fair coin:

- **Event A:** Heads appears on the first toss.
- **Event B:** Heads appears on the second toss.
- **Event C:** Exactly one heads appears in the two tosses.

It is clear that each event here has probability of $\frac{1}{2}$. Also we can ascertain that they are each pairwise independent:

$$p(A \cap B) = \frac{1}{4} = p(A)p(B) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$$

$$p(B \cap C) = \frac{1}{4} = p(B)p(C) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$$

$$p(A \cap C) = \frac{1}{4} = p(A)p(C) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right),$$

but they are not independent as a group because

$$p(A \cap B \cap C) = 0 \neq p(A)p(B)p(C) = \frac{1}{8}.$$

So independence is a property that changes with group constituency. In addition, independence can be conditional on a third event.

Events A and B are **conditionally independent** on event C:

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

Returning to the example above, A and B are not conditionally independent either if the condition is C because

$$p(A \cap B|C) = 0,$$

but

$$p(A|C) = \frac{1}{2}, \qquad p(B|C) = \frac{1}{2},$$

and their product is clearly not zero.

An important theorem states that if A and B are independent, then functions of A and B operating on the same domain are also independent. As an example, suppose we can generate random (equally likely) integers from 1 to 20 (usually done on computers). Define A as the event that a prime number occurs except the prime number 2:

$$p(A) = p(x \in \{1, 3, 5, 7, 11, 13, 17, 19\}) = \frac{8}{20},$$

and B as the event that the number is greater than 10:

$$p(B) = p(x > 10) = \frac{10}{20}.$$

Since there are 4 odd primes above and below 10, A and B are independent:

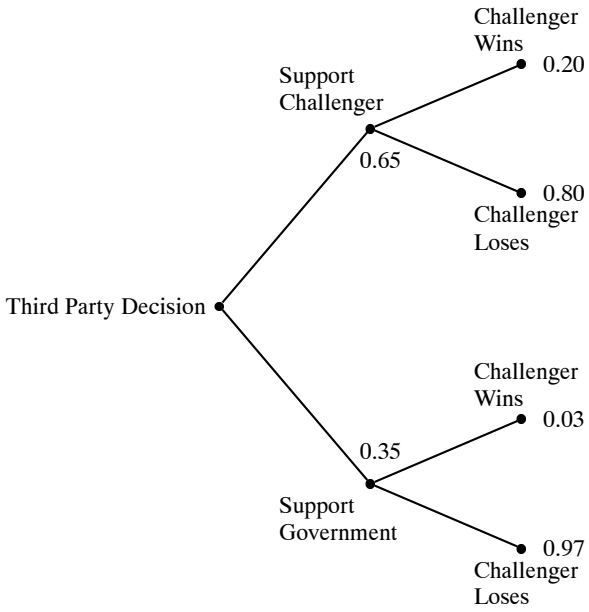
$$p(A \cap B) = \frac{4}{20} = p(A)p(B) = \frac{8}{20}\frac{1}{2}.$$

For example, use the simple functions $p(g(A)) = p(A^{\complement}) = \frac{12}{20}$ and $p(h(B)) = p(B^{\complement}) = \frac{1}{2}$. Then $\frac{12}{20} \times \frac{1}{2} = \frac{3}{10} = p(g(A) \cap (h(B))$.

★ **Example 7.14: Analyzing Politics with Decision Trees**. Another way of looking at conditional probabilities is with game or decision trees (depending on the specific application). Bueno de Mesquita, Newman, and Rabushka (1985) suggested this method for forecasting political events and applied it to the test case of Hong Kong's reversal to China (PRC). The decision tree in Figure 7.4 shows the possible decisions and results for a third party's decision to support either the challenger or the government in an election, given that they have ruled out doing nothing. Suppose we were trying to anticipate the behavior of this party and the resulting impact on the election (presumably the support of this party matters).

Hypothetical probabilities of each event at the nodes are given in the figure. For instance, the probability that the challenger wins is 0.03 and the probability that the challenger loses is 0.97, *after the third party has already*

Fig. 7.4. Marginal Contribution of Third Parties



thrown its support behind the government. Correspondingly, the probability that challenger wins is 0.20 and the probability that the challenger loses is 0.80, *after the third party has already thrown its support behind the challenger*. So these are conditional probabilities exactly as we have studied before but now show diagrammatically. In standard notation these are

$$p(C|SC) = 0.20 \qquad p(G|SC) = 0.80$$
$$p(C|SG) = 0.03 \qquad p(G|SG) = 0.97,$$

where we denote C as challenger wins, G as government wins, SG for the third party supports the government, and SC for the third party supports the challenger. As an analyst of future Hong Kong elections, one might be estimating the probability of either action on the part of the third party, and these are given here as $p(SC) = 0.65$ and $p(SG) = 0.35$. In other words, our study indicates that this party is somewhat more inclined to support the opposition. So what does this mean for predicting the eventual outcome?

We have to multiply the probability of getting to the first node of interest times the probability of getting to the outcome of interest, and we have to do this for the entire tree to get the full picture.

Looking at the first (top) path of the tree, the probability that the challenger wins when the third party supports them is $(0.20)(0.65) = 0.13$. Conceptually we can rewrite this in the form $p(C|SC)p(SC)$, and we know that this is really $p(C, SC) = p(C|SC)p(SC)$ from the definition of conditional probability on page 312. So the probabilities at the nodes in the tree figure are only "local" in the sense that they correspond to events that can happen at that geographic point only since they assume that the tree has been traversed already to that point. This makes a nice point about conditional probability. As we condition on events we are literally walking down a tree of possibilities, and it is easy to see that such trees can be much wider (more decisions at each node) and much deeper (more steps to the final event of interest).

7.8 Odds

Sometimes probability statements are reexpressed as **odds**. In some academic fields this is routine and researchers view these to be more intuitive than standard probability statements. To simplify for the moment, consider a sample space with only two outcomes: success and failure. These can be defined for any social event we like: wars, marriages, group formations, crimes, and so on. Defining the probability of success as $p = p(S)$ and the probability of failure as $q = 1 - p = p(F)$, the odds of success are

$$\text{odds}(S) = \frac{p}{q}.$$

Notice that while the probability metric is confined to $[0:1]$, odds are positive but unbounded. Often times odds are given as integer comparisons: "The odds of success are 3 to 2" and notated $3:2$, and if it is convenient, making the second number 1 is particularly intuitive. Converting probabilities to odds does not lose any information and probability information can be recovered. For instance,

with only two events, if the odds are $3{:}2$, then $p(S) = \frac{3}{5}$ and $p(F) = \frac{2}{5}$. More generally, if $odds(S) = \alpha{:}\beta$, then the two probabilities are

$$p(S) = \frac{\alpha}{\alpha + \beta} \qquad p(F) = \frac{\beta}{\alpha + \beta}.$$

These calculations are more involved but essentially the same for more than two possible outcomes.

★ **Example 7.15: Parental Involvement for Black Grandmothers**. Pearson et al. (1990) researched the notion that black grandparents, typically grandmothers, living in the same household are more active in parenting their grandchildren than their white counterparts. The authors were concerned with testing differences in extended family systems and the roles that members play in child rearing. They obtained data on 130 black families where the grandmother lived in the house and with reported levels of direct parenting for the grandchildren.

Three dichotomous (yes/no) effects were of direct interest here. ***Supportive behavior*** was defined as reading bedtime stories, playing games, or doing a pleasant outing with the child. This was the main variable of interest to the researchers. The first supporting variable was ***punishment behavior***, which was whether or not the grandmother punished the child on misbehavior. The second supporting variable was ***controlling behavior***, which meant that the grandmother established the rules of behavior for the child.

Pearson et al. looked at a wide range of explanations for differing levels of grandmother involvement, but the two most interesting findings related to these variables. Grandmothers who took the punishment behavior role versus not taking on this role had an odds ratio of $2.99:1$ for exhibiting supportive behavior. Furthermore, grandmothers who took the controlling behavior role versus not doing so had an odds ratio of $5.38:1$ for exhibiting supportive behavior. Therefore authoritarian behavior strongly predicts positive parenting interactions.

7.9 New Terminology

aggregation bias, 316
Bayes Law, 315
Binomial Theorem, 289
Borel-field, 307
cardinality, 294
complement, 298
conditional probability, 312
conditionally independent, 318
countable set, 293
difference operator, 298
disjoint, 302
elements, 292
empty set, 295
event, 292
finite set, 293
Fundamental Theorem of Counting, 286
independent, 317
intersection, 297
Kolmogorov probability axioms, 308
multiplication rule, 315
mutually exclusive, 302
null set, 295
odds, 321
objective probability, 285
ordered, with replacement, 287
ordered, without replacement, 287
pairwise disjoint, 302
pairwise independent, 317
partition, 302
Pascal's Triangle, 290
probability function, 307
proper subset, 295
sample space, 292
set, 291
sigma-algebra, 307
sigma-field, 307
Simpson's Paradox, 316
singleton, 295
subjective probability, 285
subset, 295
symmetric difference, 298
Theorem of Total Probability, 309
triple, 308
union, 296
unordered, with replacement, 289
unordered, without replacement, 288
Venn diagram, 299

Exercises

7.1 A fair coin is tossed 20 times and produces 20 heads. What is the probability that it will give a tails on the 21st try?

7.2 At the end of a legislative session there is time to vote on only three more bills. Pending there are 12 bills total: 6 on foreign policy, 4 on judicial affairs, and 2 on energy policy. Given equally likely selection, what is the probability that

- exactly one foreign policy bill will receive a vote?
- all three votes will be on foreign policy?
- one of each type will receive a vote?
- no judicial affairs bills will receive a vote?

7.3 Prove that $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$.

7.4 Develop two more rows to the Pascal's Triangle given on page 290 and show that the "magic 11" property holds.

7.5 Suppose you had a pair of four-sided dice (they exist), so the set of possible outcomes from summing the results from a single toss is $\{2, 3, 4, 5, 6, 7, 8\}$. Determine the probability of each of these outcomes.

7.6 For some set A, explain $A \cup A$ and $A \cap A$.

7.7 Prove de Morgan's Laws for two sets A and B.

7.8 The probability that marriage A lasts 20 years is 0.4, the probability that marriage B lasts 20 years is 0.25, and the probability that marriage C lasts 20 years is 0.8. Under the assumption of independence, calculate the probability that they all last 20 years, the probability that none of them last 20 years, and the probability that at least one lasts 20 years.

7.9 If $(D|H) = 0.5,\ p(D) = 1$, and $p(H) = 0.1$, what is the probability that H is true given D?

7.10 In rolling two dice labeled X and Y, what is the probability that the sum of the up faces is four, given that either X or Y shows a three.

7.11 Show that the Theorem of Total Probability also works when either of the two sets is the null set.

7.12 You are trying to form a coalition cabinet from the six major Italian political parties (given by just initials here). There are three senior members of DC, five senior members of the PCI, four senior members of PSI, two senior members of PSDI, five senior members of PRI, and three senior members of PLI, all vying for positions in the cabinet. How many ways could you choose a cabinet composed of two from each party?

7.13 If events A and B are independent, prove that $A^{\complement}$ and $B^{\complement}$ are also independent. Can you say that A and $A^{\complement}$ are independent? Show your logic.

7.14 Suppose we roll a single die three times. What is the probability of:

(a) three sixes?

(b) exactly one six?

(c) the sum of the three rolls is 4?

(d) the sum of the three rolls is a prime number?

7.15 Use this joint probability distribution

		X		
		0	1	2
	0	0.10	0.10	0.01
Y	1	0.02	0.10	0.20
	2	0.30	0.10	0.07

to compute the following:

(a) $p(X < 2)$.

(b) $p(X < 2|Y < 2)$.

(c) $p(Y = 2|X \leq 1)$.

(d) $p(X = 1|Y = 1)$.

(e) $p(Y > 0|X > 0)$.

7.16 Al and George want to have a "town hall" style debate. There are only 100 undecided voters in the entire country from which to choose an audience. If they want 90 of these people, how many different sets of 90 can be chosen (unordered, without replacement)?

7.17 Someone claims they can identify four different brands of beer by taste. An experiment is set up (off campus of course) to test her ability in which she is given each of the four beers one at a time without labels or any visual identification.

(a) How many different ways can the four beers be presented to her one at a time?

(b) What is the probability that she will correctly identify all four brands simply by guessing?

(c) What is the probability that she will incorrectly identify only one beer simply by guessing (assume she does not duplicate an answer)?

(d) Is the event that she correctly identifies the second beer disjoint with the event that she incorrectly identifies the fourth beer?

7.18 A company has just placed an order with a supplier for two different products. Let

E = the event that the first product is out of stock

F = the event that the second product is out of stock

Suppose that $p(E) = 0.3$, $p(F) = 0.2$, and the probability that at least one is out of stock is 0.4.

(a) What is the probability that both are out of stock?

(b) Are E and F independent events?

(c) Given that the first product is in stock, what is the probability that the second is also?

7.19 Suppose your professor of political theory put 17 books on reserve in the library. Of these, 9 were written by Greek philosophers and the rest were written by German philosophers. You have already read all of the Greeks, but none of the Germans, and you have to ask for the books one at a time. Assuming you left the syllabus at home, and you have to ask for the books at random (equally likely) by call letters:

(a) What is the probability that you have to ask for at least three books before getting a German philosopher?

(b) What is the highest possible number of times you would have to ask for a book before receiving a German philosopher?

7.20 Suppose you first flipped a quarter, then flipped a dime, and then flipped a nickel.

(a) What is the probability of getting a heads on the nickel *given* you get tails on the quarter and heads on the dime?

(b) Are the events getting a tails on the quarter and getting a tails on the nickel disjoint?

(c) Are the events getting a tails on the dime and a heads on the dime independent?

7.21 In a given town, 40% of the voters are Democrats and 60% are Republican. The president's budget is supported by 50% of the Democrats and 90% of the Republicans. If a randomly (equally likely) selected voter is found to support the president's budget, what is the probability that they are a Democrat?

7.22 At Cafe Med on Telegraph Avenue, 60% of the customers prefer regular coffee and 40% prefer decaffeinated.

(a) Among 10 randomly (equally likely) selected customers, what is the probability that at most 8 prefer regular coffee?

(b) Cafe Med is about to close and only has 7 cups of regular left but plenty of decaffeinated. What is the probability that all 10 remaining customers get their preference?

7.23 Assume that 2% of the population of the United States are members of some extremist militia group, ($p(M) = 0.02$), a fact that some members might not readily admit to an interviewer. We develop a survey that is 95% accurate on positive classification, $p(C|M) = 0.95$, and 97% accurate on negative classification, $p(C^{\complement}|M^{\complement}) = 0.97$. Using Bayes' Law, derive the probability that someone positively classified by the survey as being a militia member really is a militia member. (Hint: Draw a Venn diagram to get $p(C)$ and think about the Theorem of Total Probability).

7.24 Suppose we have two urns containing marbles. The first urn contains 6 red marbles and 4 green marbles, and the second urn contains 9 red marbles and 1 green marble. Take one marble from the first urn (without looking at it) and put it in the second urn. Then take one marble from the second urn (again without looking at it) and put it in the first urn. What is the probability of now drawing a red marble from the first urn?

7.25 Corsi (1981) examined political terrorism, responses to terrorist acts, and the counter-response of the terrorists for 1970 to 1974. For the type of events where a target is seized and held at an unknown site (like kidnapping) he found that 55.6% ($n = 35$) of the time the government involved capitulated. Given that this happened, 2.9% of the time the terrorists increased their demands, 91.4% of the time there was no change in these demands, and 5.7% of the time contact is lost. Of these three events, the number of times that there was known to be no damage or death was 1, 31, and 1, respectively. Construct a tree diagram that contains the conditional probabilities at each level.

7.26 Suppose there are three possible outcomes for an experiment: A, B, and C. If the odds of A over B are $9{:}1$ and the odds of B over C are $3{:}2$, what are the probabilities of the three events?