2

Rules of Disorder

The theory of probability is a formal branch of mathematics with elegant theorems, complicated proofs, and its own book of jargon. Despite these potential obstacles, people use probability informally nearly every day. When we play games, decide what to wear by glancing at the morning sky, or pick the route we will take to get across town during rush hour, we often rely on crude perceptions of probability to make decisions in the face of uncertainty. Even the most math-phobic individuals occasionally use elementary aspects of probability theory to guide their actions. Unfortunately, such primitive applications of probability are often misguided and can lead to illogical decisions. Our intuition is not a viable substitute for the more formal theory of probability.

Examples of this are evident when we play games of chance. People differ substantially in skill, leading some players to win (and others to lose) far more frequently than they should by chance alone. Most often, what we think of as skill in such games is simply a measure of how accurately a player's actions are consistent with an understanding of probability theory—unless, of course, their success relies on cheating. By analogy, when chance events play an important role in the design, function, or behavior of organisms, our skill in interpreting patterns in nature depends on our understanding of the theory of probability.

In this chapter, we briefly develop a set of definitions, rules, and techniques that provides a theoretical framework for thinking about chance events. For the sake of brevity, we focus on the aspects of probability theory that are most critical to the issues raised in the rest of this book. For more in-depth coverage of probability we recommend Feller (1960), Isaac (1995), or Ross (1997).

2.1 *Events, Experiments, and Outcomes*

Every field of science and mathematics has its own vocabulary, and probability is no exception. Unfortunately, probability theory has the added feature of assigning technical definitions to words we commonly use to mean other things in everyday life. To avoid confusion, it is crucial that we speak the same language, and, to that end, some definitions are necessary.

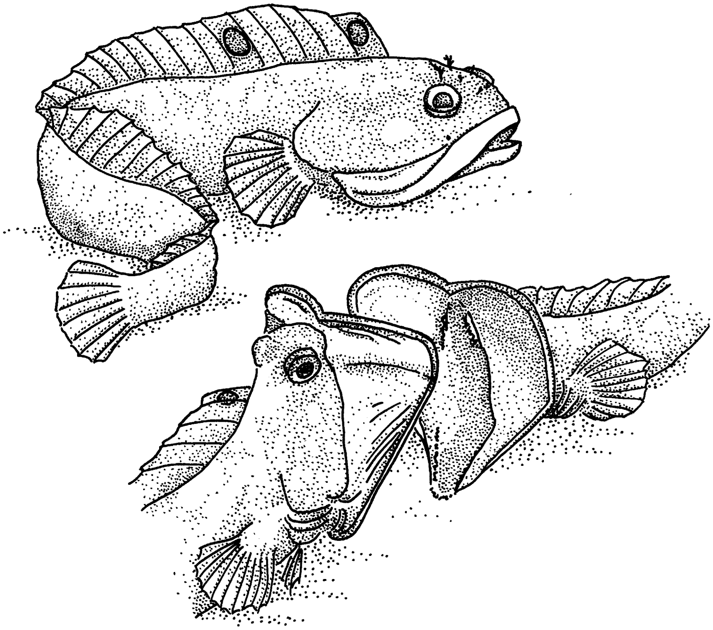

FIG. 2.1 The sarcastic fringehead. The upper panel shows the fish in repose. In the lower panel, two fringeheads engage in a ritual bout of mouth wrestling

In probability theory, the focus of our attention is an *event*. A more formal definition is given below, but put simply, an event is something that happens with some degree of uncertainty. Typically, books on probability theory use as examples events such as getting a one on the roll of a die, flipping a coin five times in a row without getting any heads, or sharing the same birthday with someone else at a party. These types of events are useful because they represent activities you can easily duplicate or imagine. As we have suggested, however, the uses of probability theory are far broader than playing games or matching birth dates; a large number of environmental and biological issues critically depend on the occurrence of uncertain events. Let's start with two biological examples where chance plays an important role.

2.1.1 Sarcastic Fish

One of the most ferocious fish found along the Pacific Coast of North America is the sarcastic fringehead (fig. 2.1). Although it rarely exceeds a foot in length, the fringehead has an enormous mouth, a pugnacious temperament, and the wary respect of fishermen, who have been known to do "amusing little dances while 6 in. long fish clamp sharp teeth around their thumbs" (Love 1991).

Fringeheads typically live in holes or crevices in the rocky substratum. They aggressively defend these shelters by lunging at anything that approaches, snapping open their capacious mouths. When the intruder is another fringehead looking for a new shelter, the two individuals often enter into a ritual match of "mouth wrestling" with their sharp teeth interlocked (Stokes 1998). As with many ritualized fights in animals, these matches are a relatively benign mechanism for establishing dominance, and the larger of the two individuals inevitably wins the battle and takes over the shelter. But sarcastic fringeheads seem to be poor judges of size. Due perhaps to poor eyesight, an inflated perception of their own bulk, or both, the fish appear incapable of accurately evaluating the size of another individual until they begin to wrestle.

Now suppose you are a fringehead guarding your shelter. Along comes another fringehead and that old, instinctive urge to dominate rises up within you. You lunge out and commence to wrestle. With your mouths pressed together it is quickly clear that you are substantially larger than your opponent (just as you thought!), and the intruder scurries away. Your shelter is safe. Later, a second fringehead arrives. Again, you rush to defend your shelter, but this time your luck runs out. You aren't quite the fish you thought you were, your mouth is smaller than his, and you end up homeless.

The stage is now set for a few basic definitions. In the vocabulary of probability theory, these wrestling matches are *experiments*.[1] Experiments are simply processes that produce some observation or measurement. Since you cannot predict the result of the wrestling experiments with complete certainty before you leave your shelter, we call these wrestling matches *random experiments*. Every time you repeat the experiment of defending your home, there is a single *outcome* (that is, one of the several possible results). The set of all possible outcomes for an experiment is called the *sample space* for that experiment. In the case of fringehead wrestling, there are only two possible *elementary outcomes*, success (s) or failure (f), and these together form the sample space.

Let's now turn our attention to another example of chance in the interaction among organisms.

2.1.2 Bipolar Smut

If you asked the average person on the street what he knew about smut and sex, he would probably feign ignorance and scurry away in search of a

[1] Note that the use of the word "experiment" in probability does not imply hypothesis testing as it might in inferential statistics or most fields of science.

policeman. If by chance you happened to ask a mycologist, however, you would evoke a *very* different response. We found this out when we naively inquired of a friend of ours if she knew anything interesting about reproduction in lower plants, and received in return an energetic lecture on the wonders of sex in the smuts.

Smuts, it turns out, are parasitic fungi in the order Ustilaginales. They commonly infect vascular plants, including a variety of economically important grains, and as a result have been studied in depth. Reproduction in smuts is bizarre by vertebrate standards. To be precise, smuts do not have separate sexes, but they nonetheless reproduce sexually.

This poses some potential problems. One of the advantages of separate sexes is that gametes from one individual cannot fuse with another gamete from the same individual. This eliminates the most extreme form of *inbreeding*—mating with yourself. In the smuts, individuals produce haploid spores,[2] each of which fuses with another spore to create a new generation of smut. But instead of having male and female individuals that produce gametes of distinctly different sizes (as you find in most animals and plants), smuts typically produce spores that are morphologically indistinguishable from one another.

Lacking discrete sexes, how do smuts avoid inbreeding? As with many other fungi, the smuts promote mating with other individuals (*outcrossing*) through the use of *compatibility genes*. In a simple case, a single gene locus has two or more alleles that determine whether spores are compatible for fusing. If the allele present at the compatibility locus differs between two spores, they can fuse; if the alleles are the same, they cannot.

For example, let's identify the different alleles at the compatibility locus by letters (e.g., *a*, *b*, *c*, etc.). Because spores are haploid, each has a single compatibility allele. If one spore has allele *a* and another spore has allele *b* (or *c*, or *d*, or anything but *a*), they can fuse. Smuts with this mating system are termed *bipolar* because two alleles determine mating compatibility. Other fungi (the tetrapolar fungi) take their sex to an even higher level by having a mating system with two separate compatibility loci. Here, spores must differ at both loci before they can fuse.

Note that in the bipolar mating system all adult smuts must be heterozygous at the compatibility locus. The only way they could be homozygous is if both of the spores that fused to form the adult had the same allele, and if they had the same allele, they could not fuse.

[2] In sexually reproducing organisms, each adult has two sets of chromosomes, one from each parent, and the organisms are therefore in a *diploid* state. In the process of manufacturing gametes by meiosis, the number of chromosomes is cut in half, and these special reproductive cells (spores in this case) are in a *haploid* state.

Now, imagine two smuts on wheat (or rye). Smut 1 has compatibility alleles a and b. The other (smut 2) has compatibility alleles c and d. You could not find a more perfectly matched couple! Let's use these fertile fungi to perform an experiment.

It's mating time, and each of the smuts produces a multitude of spores, which you carefully collect in a bag. After mixing the spores thoroughly to randomize which spores are in contact, you empty the bag onto a microscope slide. There are thousands of pairs of contiguous spores, and as you watch, some fuse successfully and begin to grow. Others remain unfused, never to know the life of a smut. What can we predict as to which spores will fuse and which will not?

We begin by noting that each pair of spores can be viewed as another example of a random experiment. The outcomes of the experiment are successful fusion or the failure to fuse, and success and failure depend on chance. But in this case, we have additional information that can be applied to the problem. We already know that the results of these fusion experiments depend on the underlying genetics. Therefore, let's examine the outcomes in terms of the genotypes of the spores at the compatibility locus.

For any given spore in the experiment described above, there are four possible results corresponding to the four compatibility alleles found in the parents (a, b, c, and d). Since there are four possible results for each spore in the pair, there is a combined total of sixteen possible outcomes (see box 2.1).

Box 2.1. Why sixteen and not eight?

If there are four possibilities for the first spore and another four for the second spore, why do we multiply instead of add to get the total number of possibilities? The answer can be seen by stepping through the problem. Assume the first spore has allele *a*. How many possibilities are there for the second spore? The answer is four, since the second spore could have any of the four alleles. Similarly, if the first spore has allele *b*, there are still four possible values for the second spore. Each possibility for spore 1 has four possibilities for spore 2. Therefore, there are four sets of four, which is sixteen total outcomes. Notice that some of the sixteen possible outcomes are functional duplicates because the order of the alleles doesn't matter to the genotype of the offspring (e.g., *a* fusing with *b* and *b* fusing with *a*). Sometimes, order does matter, and we will deal with the consequences of duplicate outcomes a little later.

Furthermore, the outcomes of this smut fusion experiment are more complex than those in our fringehead wrestling experiment. Recall that when fringeheads wrestle, the outcome is *elementary* (also known as simple or indecomposable):

either success or failure. In contrast, each genetic outcome of the smut experiment has two parts, each part corresponding to the allele from a potential parent. If we view the results of randomly picking a single spore as yielding an elementary outcome (= the spore's allele), the outcome of a fusion experiment is *complex*; it includes two elementary outcomes.

Complex outcomes can be represented by an ordered set. In this case, our outcome is an ordered pair of elementary outcomes (x, y), where x is the allele of one spore and y is the allele of the other spore. This fusion experiment is analogous to grabbing two socks blindly from a drawer. Spores with different incompatibility alleles are like socks with different colors. Using this sock analogy, successful sex in smuts is like picking two socks from the drawer that do not match.

Using these definitions and examples, we can now define an event more formally. An *event* is a *set of outcomes* from an experiment that satisfies some specified criterion. We use these two terms (event, set of outcomes) interchangeably. The most basic events associated with an experiment are those that correspond to a single outcome[3]—for example, winning a mouth wrestle, which we call event *Win*; or getting a pair of spores, both with allele a, which we call event AA. Each of these basic events includes only one of the following possible outcomes:

$$\mathit{Win} = (s)$$

$$AA = (a, a).$$

Events can also include several outcomes. For example, suppose you do not care which alleles a pair of spores has as long as they match. Given the mating strategy of smuts, the event *Match* could also be described as "a sexually incompatible pair of spores." The set associated with *Match* includes four of the possible sixteen outcomes of our experiment:

$$\mathit{Match} = \big\{(a, a), (b, b), (c, c), (d, d)\big\}. \tag{2.1}$$

In this fashion, we can define a wide variety of events related to the same fusion experiment.

2.1.3 Discrete versus Continuous

So far, the sample spaces we have discussed include a small number of possible outcomes. These are examples of *discrete sample spaces*. Discrete sample

[3] To help keep things as clear as possible, we'll use the convention of writing the names of *events* beginning with a capital letter and the names of *outcomes* in all lower-case letters.

spaces have a *countable* number of outcomes, by which we mean that we can assign an integer to each possible outcome. Any sample space with a finite number of outcomes is discrete, but some discrete sample spaces include an infinite number of outcomes. For example, we could ask how many wrestling matches a fringehead will enter before it loses for the first time. If we assume the fringehead is immortal and that there is an inexhaustible supply of challengers coming by, it is possible (although highly unlikely) that the fringehead could continue winning forever. The sample space for possible outcomes in this experiment is thus infinite, but countable.

Other experiments have an infinite number of possible outcomes that are *uncountable*. This commonly occurs in experiments where, by necessity, each outcome is measured using real numbers rather than integers. For example, the time it takes a predator to capture its prey can be measured to the fraction of a second, the average annual rainfall in Cincinnati can be calculated to a minute portion of an inch, and a compass heading can be determined to a minuscule part of a radian. In such situations, there are theoretically an infinite number of possible outcomes within any measurement interval, no matter how small the interval. Such experiments produce a *continuous sample space* with an uncountably infinite number of possible outcomes. We will return to continuous sample spaces in chapter 4.

2.1.4 Drawing Pictures

To analyze chance events, it is often useful to view the sample space (the set of all possible outcomes) in a diagram. Traditionally, the entire sample space is represented by a box. In a discrete sample space, each possible outcome is then represented by a point within the box. For our sarcastic fringehead wrestling experiment, for instance, there would be only two dots in the box (one for a success, s, and one for a failure, f; fig. 2.2A). For our experiment in smut reproduction, the box has sixteen dots corresponding to the sixteen ordered pairs of alleles (fig. 2.2B).

Once the possible outcomes are drawn in the box, events can be represented as “disks,” where each disk is a closed curve that includes the set of points, if any, that satisfy the criteria of the event. This type of diagram is called a *Venn diagram*, after its originator John Venn.[4] A Venn diagram for our smut experiment is shown in figure 2.2B. Here, the disk labeled *Match* depicts the event of a pair of spores that have matched compatibility alleles.

[4] *Venn in doubt, draw a diagram*. Upon graduating from college, John Venn became a priest for five years, after which he returned to Cambridge University as a lecturer in Moral Science. Venn later grew tired of logic and devoted his time to writing history books and designing new machines. His most intriguing invention was a device to bowl cricket balls. The machine was so good it clean bowled one of the top stars of the Australian cricket team four times.

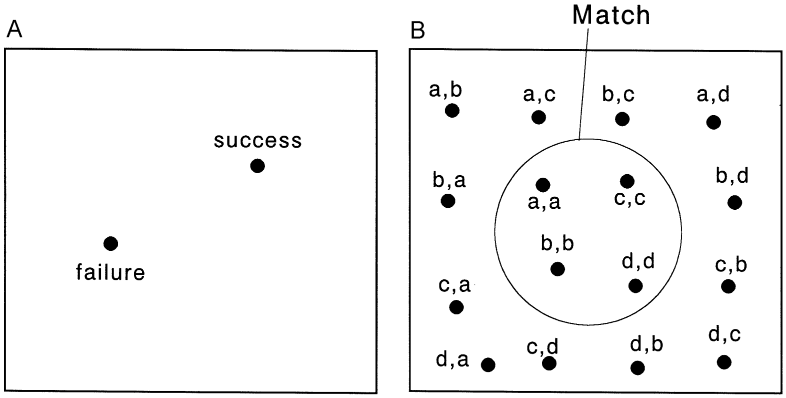

FIG. 2.2 Examples of Venn diagrams. Panel A depicts the complete sample space for a fringehead wrestling match. In this simple experiment, there are only two possible outcomes. Panel B shows the sixteen possible outcomes of the mating between smut 1 and 2. The event labeled "*Match*" includes the outcomes in which compatibility alleles are the same.

2.2 *Probability*

So far, we have thrown out a slew of definitions and drawn some potentially useful diagrams, but we have yet to touch on the central concept of our discussion, that of probability. In other words, we have characterized the possible outcomes of an experiment, but we currently have no means of estimating or predicting how frequently different outcomes might occur. In this section, we will develop the notion of the probability of an event.

For a particular manifestation of a random experiment, each outcome in the sample space has some possibility of occurring, but only one outcome can actually occur. The *probability* of a specific outcome is defined as the fraction of a large number of experiments that will yield this particular outcome.

DEFINITION 1 *Provided the number of experiments is very large:*

$$P(x) = \frac{\textit{number of occurrences of outcome } x}{\textit{total number of repeated random experiments}}.$$

This probability (denoted $P(x)$ for outcome x) must lie between 0 (the outcome never occurs) and 1 (the outcome always occurs).

In practice, we never know the precise probability of any uncertain event, but there are two general ways by which we can estimate its value: we can make an empirical estimate using repeated random experiments, or we can make a theoretical estimate using an idealized model of the random process.

Our experiment with smut sex provides an example. If the spores of smut 1 (with compatability alleles *a* and *b*) are mixed with spores of smut 2 (alleles *c* and *d*) and we repeatedly draw out pairs of spores at random, we find that each pairing occurs with equal frequency. Thus, each of the sixteen outcomes shown in figure 2.2B has a probability of 1/16.

A philosophical note is in order here. Given our definition, it is meaningless to talk about the probability of an experiment that cannot be repeated. If you carry out an experiment, an outcome results. But unless you can *repeat* the experiment, the number of occurrences of that outcome (= 1) must equal the number of trials (= 1). Thus, by our definition, the probability of an unrepeated experiment is exactly 1, in this context an uninformative number.

In contrast, as the number of repeated experiments grows, the frequency of occurrence becomes a better and better estimate of the actual probability of a given outcome for any single experiment. Formally, this rule is called the *Law of Large Numbers*. Fortunately, it expresses how most people intuitively think of the probability of a chance event.

2.3 *Rules and Tools*

Although estimating probabilities through the use of repeated random experiments is a common tool, there are many situations where this approach may be inaccurate, unacceptably expensive, unethical, or even impossible. For example, the number of experiments needed to get an accurate estimate of the probability may be inconveniently large, especially if you are trying to estimate the probability of rare events. In other cases, experimentation may be impractical since the event you are interested in may be something you are actively trying to avoid (e.g., an oil spill). You would not want to cause such events just to estimate their probability of occurrence. Finally, some questions may require experiments that society deems unethical. Examples include human trials of new drugs or surgical procedures where the expected risks are potentially large. To handle these cases, we need to develop models of probabilistic events. These models will be simplified abstractions of the real world, but they may help us to evaluate stochastic phenomena that we cannot study experimentally. Let's begin by considering how the probability of events builds on the probability of outcomes from an experiment.

2.3.1 Events are the Sum of Their Parts

Recall that an event is a set of outcomes that meets some criteria. If any outcome in the set occurs, the event occurs. As a result, the probability that an event will occur can be derived from the sum of the probabilities of all outcomes in its set:

Rule 1 $$P(\text{event } A) = \sum P(\text{outcomes in } A).$$

(Translation: If each of several different outcomes satisfies the criterion for an event, as in the event of getting a pair of smut spores with matching compatibility alleles, the probability of the event is simply the sum of the probabilities of the individual outcomes.)

To give some tangibility to this rule, let's return to the Venn diagram for our experiment in smut reproduction (fig. 2.2B). Here each of the points in the disk labeled *Match* satisfies the criterion for the event in which a pair of spores has matching alleles. Getting two a alleles *or* two b alleles *or* two c alleles *or* two d alleles are all satisfactory. Since each of these is a distinct outcome, we can simply add their respective probabilities (1/16) to obtain the overall probability of getting matching alleles: $4 \times 1/16 = 1/4$. In other words, the probability of the compound event *Match* is the sum of the probabilities of the individual outcomes enclosed by the disk in the Venn diagram. In fact, the probability of *any* event (no matter how complicated it may be) can be estimated if you can (1) identify the individual outcomes in the event and (2) if you know the probability of each of these outcomes.

Unfortunately, it is not always easy to both identify the individual outcomes in an event and know the probability of each outcome. As a result, we need to develop tools that allow us to estimate probabilities of sets we cannot measure directly.

Before leaving the subject of additive probabilities, we consider one corollary of rule 1:

Rule 2 $$\sum_{\text{all } i} P(x_i) = 1.$$

(Translation: The sum of the probabilities of all outcomes from an experiment equals one.)

This feature should be intuitive. As long as our sample space includes all possible outcomes, their probabilities must sum to one. In each experiment, *something* will happen. As we will see, rule 2 is used extensively for calculating the probability of complex events.

2.3.2 The Union of Sets

As we have seen in the case of matching alleles, complex events can be formed from combinations of individual *outcomes*. Similarly, we can build even more complex events by combining *events*. Let's explore two examples that will be useful later.

Recall that when a pair of smut spores share the same compatibility allele, fusion cannot occur. As a result, the action of compatibility genes is to reduce inbreeding. Reduce, yes, but it does not totally preclude the fusion between two spores produced by the same parent. Since all smuts are heterozygous at the compatibility locus, all smuts produce two spore genotypes, and two spores of different types can successfully fuse even though they have the same parent. Thus, inbreeding. Now, suppose that you are interested in the event of getting an inbred offspring from one or the other of the parents in our smut mating experiment. How would you describe this event?

To answer this question, we focus on simpler (although still complex) events. There are two parents in our experiment, smut 1 (with compatibility alleles a and b) and smut 2 (with c and d). Let's let I_1 denote the event of getting an inbred offspring from smut 1. There are two outcomes in this event [(a, b) and (b, a)]. Similarly, I_2 is the event of getting an inbred offspring from smut 2 [(c, d), (d, c)], and our overall event (let's call it I for inbred offspring in general) includes all of the outcomes in *either* I_1 *or* I_2. I is therefore a combination of two simpler sets of outcomes. We call this combination a *union*. The union of two sets is typically denoted by the union operator, ∪. Therefore, $I = I_1 \cup I_2$. This union is shown as a Venn diagram in figure 2.3.

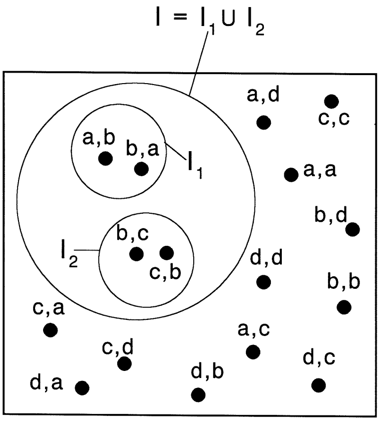

FIG. 2.3 A Venn diagram of the sample space for the smut sex experiment. The events labeled "I_1" and "I_2" include the outcomes corresponding to the inbred offspring of smuts 1 and 2, respectively. The event labeled "I" is the union of these two events. Note that events I_1 and I_2 do not share any outcomes.

For reasons that will become clear in a moment, a second example of the union of sets will be useful. Consider S_1, the event of getting an offspring of smut 1. S_1 is different from I_1 because in this case we do not care about the progeny's genotype. As long as one of its parents is smut 1, an offspring

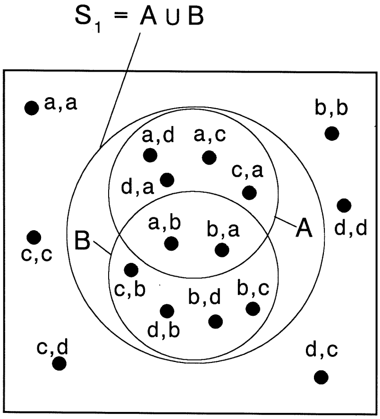

FIG. 2.4 A Venn diagram of the event S_1, the union between event A (outcomes containing allele a) and event B (outcomes containing allele b). Note that two outcomes are shared between events A and B.

qualifies. One simple way to describe this event follows from the realization that offspring of smut 1 must have either allele *a* or allele *b* (or both). Since smut 2 had neither of these two alleles, any new smut in our experiment with allele *a* (event A) and/or *b* (event B) must be an offspring of smut 1. Therefore, we can generate the event S_1 as the union between two events: $S_1 = A \cup B$. If you examine the Venn diagram in figure 2.4, you can see that six outcomes are found in each of these events. (Remember, *aa* and *bb* fusions do not produce offspring.) Unlike the previous example, these component events A and B do not have completely distinct outcomes. Their disks overlap because they share two offspring genotypes, (a, b) and (b, a). This overlap will have important consequences, which we discuss below.

2.3.3 The Probability of a Union

Now, our interest in this exercise is to estimate the probability of the complex events described by the union of simpler events. Consider first $P(I)$, the probability of obtaining an inbred smut. Recall from rule 1 that the probability of an event is the sum of the probabilities of its individual outcomes. By analogy, perhaps we can use the sum of the probabilities of the two events, I_1 and I_2, to estimate the probability of I. Can we really add the probabilities of events the way we can add the probabilities of individual outcomes?

Our logic in formulating rule 1 was that because the outcomes in an event are distinct, the probability that the event occurs is the sum of the probabilities of the individual outcomes. If you examine the events I_1 and I_2 (fig. 2.3), you will find that they indeed do not overlap. Therefore, by the same logic as in rule 1, we can use the sum of the probabilities of I_1 (= 1/8) and I_2 (= 1/8) to calculate the probability of I:

$$P(I) = P(I_1) + P(I_2) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}. \tag{2.2}$$

So far, so good, but what happens when the events in a union *do* overlap? Here we can use our second example to see if the simple summation of probabilities still holds. As we noted above, the event of getting an offspring from smut 1 is the union of events A and B. But these events overlap (fig. 2.4). If we simply summed the probabilities of A and B, two outcomes in S_1 $[(a, b), (b, a)]$ would get counted twice, once in A and once in B. By counting these outcomes twice, we would be overestimating the number of outcomes that qualify for event S_1. When this inflated number is inserted into our definition of probability, we would as a result overestimate the probability of S_1. Therefore, to calculate accurately the probability of S_1, we need to account for how shared outcomes affect the probability of the union of two or more sets. To do this, we use the concept of the intersection.

2.3.4 Probability and the Intersection of Sets

Outcomes that are shared by two sets are called the *intersection* between the sets. We denote the intersection by the operator, ∩. The intersection between the sets A and B includes two outcomes (see fig. 2.5):

$$A \cap B = \{(a, b), (b, a)\}.$$

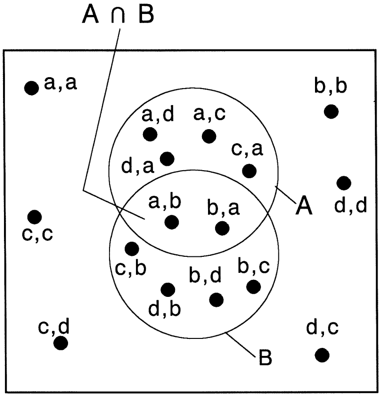

Fig. 2.5 A Venn diagram showing the intersection between event A (outcomes containing allele a) and event B (outcomes containing allele b). The interaction contains only those outcomes shared between events A and B.

More to the point, the intersection includes the same two outcomes that get counted twice if we add the probabilities of events A and B to arrive at the probability of S_1. The probability of S_1 should clearly include the probability of each of these two outcomes, but the probability of these shared outcomes should be included only once. As a result, we can estimate $P(S_1)$ from the sum of $P(A)$ and $P(B)$ by adjusting for the double counting of all outcomes shared by A and B (that is, by subtracting $P(A \cap B)$). Thus,

Rule 3 if $S_1 = A \cup B$, then $P(S_1) = P(A) + P(B) - P(A \cap B)$.

(Translation: If two events share outcomes, then the sum of the probabilities of the two events always exceeds the probability of the union of those two events. The difference is the probability of the shared outcomes, which erroneously gets counted twice.)

Note that this rule applies equally well to our initial example of inbred smuts. In this case, there is no overlap between I_1 and I_2, $P(I_1 \cap I_2) = 0$, and $P(I) = P(I_1) + P(I_2)$ as advertised.

2.3.5 The Complement of a Set

Our laboratory experiments with reproduction in smuts greatly simplifies the real-world phenomenon by focusing on only two individuals. In an actual field setting, spores from one adult smut could potentially contact spores from a large number of other individuals. Furthermore, we have assumed here that there are only four alleles at the compatibility locus, but in actual populations the number of alleles may exceed a hundred. Both of these factors (more potential parents and more alleles) make it far more complicated to estimate probabilities in real populations.

For example, suppose you were a smut trying to estimate the probability that a particular individual spore you have produced could fuse with other spores encountered in the field. Let's assume that within the local smuts there are a hundred alleles at the compatibility locus, only one of which is contained in this particular spore. One approach to estimating your chance of producing offspring would be to sample the relative frequency of each of the other ninety-nine alleles in the population. Each of these alleles is compatible with the individual spore in question, and if you knew the probability of encounter for each of these genotypes, you could estimate the overall probability that this individual spore will successfully fuse. This approach is indeed possible, but it is *very* laborious.

A far simpler approach would be to focus on the single allele with which your spores could *not* fuse. Since the sum of the probabilities of all outcomes of an experiment must equal 1 (see rule 2), we can estimate the probability of an event by going in the back door, so to speak. If you could estimate the probability

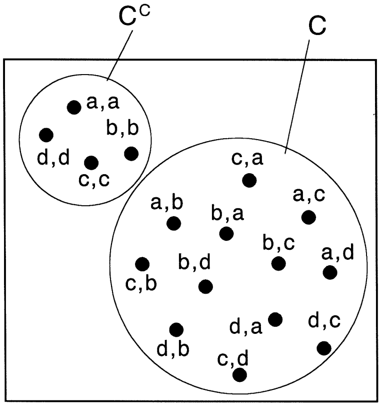

FIG. 2.6 A Venn diagram showing the relationship between event C and its complement, C^c. All outcomes not in an event are in its complement.

of incompatibility, you could subtract it from 1 to estimate the probability of fusion. By turning the question on its head, a far simpler answer emerges.

This process leads us to define yet one more term. In probability theory, the *complement* of X is all outcomes in the sample space that are not in the set X. The complement is symbolized as X^c. In the example just discussed, we are interested primarily in the set C, the alleles with which our individual spore is compatible. But we carry out our calculation using C^c (the complement of C), the alleles with which our spore is *in*compatible (see the Venn diagram in fig. 2.6). The classic example of an event that is much easier to address as a complement is the question of estimating the probability that at least two individuals in a crowd share the same birthday. This probability is difficult to calculate directly, but quite simple to estimate using the complementary event. In other words, it is far easier to address the problem by estimating the probability that no two individuals share a birthday than it is to estimate the probability that at least two do share a birthday. We will leave the proof of this assertion as an exercise for you (see question 6 at the end of the chapter).

We note for future use the following fact:[5]

Rule 4 $$P(X) + P(X^c) = 1.$$

This follows from rule 2. Because all events are either in X or X^c, the sum of $P(X)$ and $P(X^c)$ must be 1.

[5] A fact well known to country music fans. As Clay Walker laments to his departed sweetheart, "The only time I ever miss you, honey/ is when I'm alone and when I'm with somebody."

2.3.6 Additional Information and Conditional Probabilities

Let's now return to the event of getting an inbred offspring from smut 1 (I_1). Remember that this event includes two outcomes: $I_1 = \{(a, b), (b, a)\}$. Suppose there is one spore type whose genotype you can accurately identify. For example, suppose there is a rare mutation that, if present, causes spores with compatibility allele a to have a different color. As you scan a group of spores, you can thus identify with certainty the genotype of an occasional spore.

Suppose you randomly pick a pair of spores from those produced by smuts 1 and 2. You notice that one of the two spores has the color mutation. As a result, you know this spore has allele a. How does this additional information affect our estimate of the probability that this pair of fused spores will produce an inbred offspring of smut 1? It is clear that $P(I_1)$ must change based upon the additional information we now have, because only seven of the sixteen outcomes in our sample space are now possible—(a, a), (a, b), (b, a), (a, c), (c, a), (a, d), (d, a). At least one of the two alleles in each of these fusions is a.

We call the probability of an event based on the known occurrence of a separate event a *conditional probability*. It is denoted $P(X \mid Y)$, which is read "the probability of event X given that event Y occurs." Note that the conditional probability $P(X \mid Y)$ does not require that event Y happen first. For example, our task here is to find $P(I_1 \mid Color)$, the probability that a random pair of spores produces an inbred offspring of parent 1 given that (because of its color) at least one of the spores is known to have allele a. The number of inbred offspring produced (and therefore the process by which we calculate the probability of their production) is the same whether we observe the color of spores after or before they fuse. As long as we know for sure that event Y occurs, its temporal relationship to event X is irrelevant.

Let's examine a Venn diagram for our smut experiment (fig. 2.7) to see if we can figure out how the probability of I_1 will change given that we know that *Color* occurs. As we have seen, the event *Color* includes only seven outcomes from our original sample space of sixteen. In other words, if we see a spore with the color mutation, there is no chance for any of the nine outcomes that are not in *Color* to occur. In essence, the knowledge that the event *Color* has occurred shrinks our sample space.

We can use this information to calculate the modified probability of I_1 by using our existing techniques applied to this modified sample space. Our new, reduced sample space has only seven outcomes, and the event I_1 includes two of them $[(a, b), (b, a)]$. Thus, by our definition of probability, $P = 2/7$. In the long run it will be useful to express this conclusion in a more general fashion. According to rule 1, we should be able to estimate the probability of the event

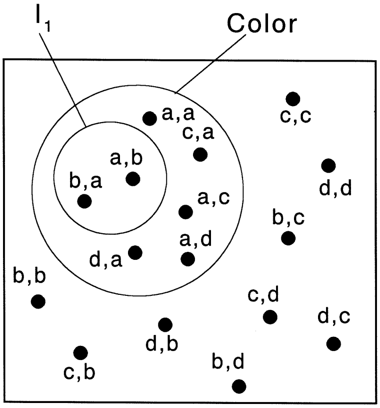

FIG. 2.7 An example of conditional probability. The event "*Color*" includes all outcomes that contain allele a. If we know that this event occurs (that is, we can see that a spore chosen at random is a different color than usual), this information affects the probability that event I_1 (an inbred offspring from smut 1) occurs.

I_1 by summing the probabilities of the two conditioned outcomes in I_1. This is indeed true, but we have a problem if our probabilities for the outcomes are based on our original sample space. In this case, each outcome has $P = 1/16$, and if we summed the probabilities of the seven possible outcomes in *Color*, they would not equal 1. In fact, they would sum to $P(Color)$. This suggests a solution. If we divide each outcome's probability by the total probability in our new sample space, $P(Color)$, the probabilities in the new sample space sum to 1.

Now all we need is an expression that defines those outcomes in our new sample space *Color* that also satisfy the event I_1, and our task will be complete. Solving this problem is easy, since the set of outcomes that satisfies both I_1 and *Color* is, by definition, the intersection of the two events. Thus, if we divide the probability of this intersection by $P(Color)$ to adjust the probabilities, we arrive at a formula for the probability of producing an inbred offspring of smut 1 given that we know one spore has allele a:

$$\begin{aligned} P(I_1 \mid Color) &= \frac{P(I_1 \cap Color)}{P(Color)} \\ &= \frac{2/16}{7/16} = \frac{2}{7}. \end{aligned} \tag{2.3}$$

Or, in the general terms of any two events X and Y,

Rule 5 $$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}.$$

(Translation: The probability that an event occurs, given that a second event occurs, is simply the probability that both events occur [that is, the probability of their intersection] divided by the probability of the event known to have occurred.)

One bonus from our effort with conditional probabilities is that we also gain a new formula to calculate the probability of the intersection between two events. If you rearrange the equation in rule 5, you get

$$P(X \cap Y) = P(X \mid Y) \cdot P(Y). \tag{2.4}$$

This formula provides an intuitively pleasing definition for the probability of the intersection of two events. Event Y occurs with probability $P(Y)$. Given that Y occurs, event X occurs with probability $P(X \mid Y)$. Therefore, the probability that both X and Y occur (that is, $P(X \cap Y)$) is the product of these two terms.

Now, in arriving at this conclusion, we have arbitrarily assumed that event Y is known to occur, but we could just as easily have assumed that X occurred. Thus, we can alternatively obtain the probability of the intersection of the two events using the probability that X occurs and the probability of Y given X. Therefore, it must also be true that

$$P(X \cap Y) = P(Y \mid X) \cdot P(X). \tag{2.5}$$

This equivalent form provides an important step to another useful rule in probability.

2.3.7 Bayes' Formula

Normally when we deal with equations, we try to simplify them as much as possible. This usually makes it easier to interpret what they mean. Sometimes, however, we can learn something by rearranging the equation into a more complex form. Let's return to our definition of a conditional probability (rule 5) for an important example. The probability that an event X occurs given that a second event Y occurs is equal to

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}. \tag{2.6}$$

Let's expand this simple formula and see where it gets us. As we just discovered, the probability of the intersection between two events (the numerator here) can be written as $P(X \cap Y) = P(Y \mid X) \cdot P(X)$. To expand the denominator, we can use a trick based on the fact that the combination of an event (e.g., X) and its complement (X^c) includes all possible outcomes (rule 4). Therefore, we can write

$$P(Y) = P(Y \cap X) + P(Y \cap X^c). \tag{2.7}$$

In other words, all outcomes in Y must either be in X or its complement. If we further expand this formula for $P(Y)$ using eq. (2.4), we obtain

$$P(Y) = P(Y \mid X) \cdot P(X) + P(Y \mid X^c) \cdot P(X^c). \tag{2.8}$$

Now we are ready to thoroughly complicate rule 5. Substituting eq. (2.5) for the numerator and eq. (2.8) for the denominator into the formula for a conditional probability in eq. (2.6), we obtain

$$P(X \mid Y) = \frac{P(Y \mid X) \cdot P(X)}{P(Y \mid X) \cdot P(X) + P(Y \mid X^c) \cdot P(X^c)}. \tag{2.9}$$

This result may not seem like much of an accomplishment given the simple formula with which we started, but in fact this formula proves to be a very powerful tool.

This equation was originally proposed by Thomas Bayes, another English theologian and part-time mathematician. It is known as *Bayes' formula*. Notice that the equality shown here contains on its left side the probability for X conditioned on the occurrence of Y. In contrast, on the right-hand side, the conditional probabilities are all for Y conditioned on the occurrence of either X or X^c. In other words, the probability of one event conditioned on a second can be used to calculate the probability of the second event conditioned on the first. Therein lies the utility of Bayes' formula.

2.3.8 AIDS and Bayes' Formula

Your head is probably spinning from all these conditions, so let's consider an example to show how useful Bayes' formula can be. Isaac (1995) provides an excellent analysis of issues related to testing for HIV that shows how useful Bayes' formula can be.

For several years, blood and saliva tests have existed that can very accurately assess whether an individual has been infected with the AIDS virus, HIV. Although these tests are quite accurate, they occasionally make mistakes. There are two types of mistakes: false positives (where the individual tests positive but has never been exposed to the AIDS virus) and false negatives (where infection has occurred, but the test does not detect it). Experimental estimates of the likelihood of these events suggest that the existing tests for HIV are extremely accurate. If an individual is infected with HIV (= event Inf), existing tests will be positive (= event Pos) roughly 99.5% of the time. In other words,

$$P(Pos \mid Inf) = 0.995. \tag{2.10}$$

From this conditional probability, we can immediately calculate the probability of one type of mistake, a false negative. If a positive blood test for an infected individual occurs 99.5% of the time, this means that 0.5% of blood tests from infected individuals are negative (= event Neg). In other words,

$$P(\mathit{FalseNegative}) = P(Neg \mid Inf) = 1 - P(Pos \mid Inf) = 0.005. \qquad (2.11)$$

Using a similar procedure, we can estimate the probability of a false positive. In this case, we start with $P(Neg \mid NInf)$, the probability that we get a negative test result from a person who is not infected. In practice, estimating $P(Neg \mid NInf)$ is difficult because we need individuals who we know with certainty have not been infected (which requires a separate, unequivocal means of testing for HIV). Reasonable estimates of $P(Neg \mid NInf)$ using control groups with no likely risk of exposure to HIV suggest that this probability is roughly 0.995. As a result, the probability of a false positive is

$$P(\mathit{FalsePositive}) = P(Pos \mid NInf) = 1 - P(Neg \mid NInf) = 0.005. \qquad (2.12)$$

Therefore, the probabilities of test errors (either positive or negative) are extremely small for an individual test.

Now suppose that a misguided law is passed requiring all individuals to take a blood test for HIV infection, the intent being to quarantine infected individuals. If we select a random individual whose test is positive, what is the probability that this random individual is actually infected with HIV? If we translate this question into a conditional probability, we are asking what is

$$P(Inf \mid Pos)?$$

Notice that this conditional probability differs fundamentally from the conditional probabilities in our estimates of false positives and false negatives. Here we are trying to estimate the probability of actual infection conditioned on a test result. In eqs. (2.11) and (2.12), the reverse is true—we estimated the probability of a test result conditioned on a state of infection. This is a perfect opportunity to use Bayes' formula, which allows us to use probabilities conditioned on one event to estimate probabilities conditioned on another.

To simplify the interpretation, let's insert the events of this problem into Bayes' formula, eq. (2.9):

$$P(Inf \mid Pos) = \frac{P(Pos \mid Inf) \cdot P(Inf)}{P(Pos \mid Inf) \cdot P(Inf) + P(Pos \mid NInf) \cdot P(NInf)}. \qquad (2.13)$$

We have already estimated the conditional probabilities on the right side. All we need are estimates for $P(Inf)$, the fraction of the population that is infected.

The Centers for Disease Control estimated that there were 293,433 individuals reporting infection by HIV or AIDS in the United States in 1996. Given a population of approximately 270 million, this yields a rough estimate of $P(Inf) = 0.001$. Although this number surely underestimates the total number of infected individuals, it gives us a ballpark estimate of the probability of infection:

$$P(Inf) = 0.001, \text{ therefore } P(NInf) = 0.999. \tag{2.14}$$

Substituting these values into eq. (2.13), we obtain

$$P(Inf \mid Pos) = \frac{(0.995)(0.001)}{(0.995)(0.001) + (0.005)0.999} = 0.16. \tag{2.15}$$

This is an unexpected and disturbing result. Despite the fact that the blood test has only a minuscule chance of false positives (0.5%), a positive blood test implies only a 16% chance that an individual is actually infected. How can this be? Looking at Bayes' formula provides a clear explanation for this seeming paradox. Although individual tests have a low chance of error, most individuals who are tested are not infected with HIV. Therefore, we are multiplying a small probability of false positives by a large number of uninfected individuals. Even a minute probability of false positives for individual tests can in this circumstance produce many more false positives than true positives. As long as the disease is rare, even a very accurate test of infection will not be able to accurately identify infected individuals in a random test.

2.3.9 The Independence of Sets

In deriving Bayes' formula, we made repeated use of the information provided by conditional probabilities. That is, knowing that Y occurs gives us new insight into the probability that X occurs. There are cases, however, where the conditional probability of event X given event Y is the same as the unconditional probability of X. In other words, the added information of knowing that event Y occurs tells us nothing about the probability of event X occurring. Thus, if $P(X \mid Y) = P(X)$, the two events X and Y are said to be *independent* events. Independence of events turns out to be a very useful feature. Consider rule 5. If we use our definition of independence, we can substitute for the conditional probability on the left-hand side of the equation, $P(X \mid Y)$, to get

$$P(X) = \frac{P(X \cap Y)}{P(Y)}, \text{ if } X \text{ and } Y \text{ are independent.} \tag{2.16}$$

If we then multiply both sides of this equation by $P(Y)$, we obtain the famous *product rule* of independent events:

Rule 6 If X and Y are independent events, $P(X \cap Y) = P(X) \cdot P(Y)$.

(Translation: If two events do not influence each other's probability of occurring, the probability that both events will occur is simply the product of the probabilities that they individually occur.)

It turns out that this product rule can be extended to any number of independent events. If there are n independent events, the probability that all n events occur (i.e., the intersection of the n events) is the product of the n probabilities of the individual events.

As an example of the utility of the product rule, let's return once again to our sarcastic fringehead experiment. In this particular example, our fringehead won his first match and lost the second. But suppose we are interested in a more general question, the probability that the fringehead wins (and thereby retains his shelter) for the first time on the ith experiment. Let event $L_i =$ (retains shelter on match i). There are only two outcomes for each individual bout, success and failure, s and f. To keep track of the particular bout in which an outcome occurred, we use subscripts. Thus, s_{15} denotes a success in the fifteenth wrestling match.

Let's suppose that the probability of winning a match is p, and of losing is $q = 1 - p$. Assume also that the probability of winning is not affected by what happens during a previous wrestling match (i.e., bouts are independent—fringeheads do not become better wrestlers with more practice). Unless the fringehead retains its shelter in the first wrestling match, each L_i will be a series of $(i - 1)$ f's followed by a single s. For instance, if our fish first wins in the fifth match,

$$L_5 = (f_1, f_2, f_3, f_4, s_5). \tag{2.17}$$

Therefore, L_i is the same as

$$f_1 \cap f_2 \cap \cdots f_{i-1} \cap s_i. \tag{2.18}$$

Since the events associated with each match are independent, we can calculate the probability of this intersection using the product rule. The probability of s is p, and the probability of f is q. Therefore,

$$P(L_i) = q \cdot q \cdot q \cdot \ldots \cdot p \text{ where there are } (i-1) q\text{'}s. \tag{2.19}$$

Thus,

$$P(L_i) = q^{i-1}p. \tag{2.20}$$

For example, if $p = 0.25$, the probability that our fish will first win on match i is shown in figure 2.8. There is only about an 8% chance that the first win will be in the fifth bout.

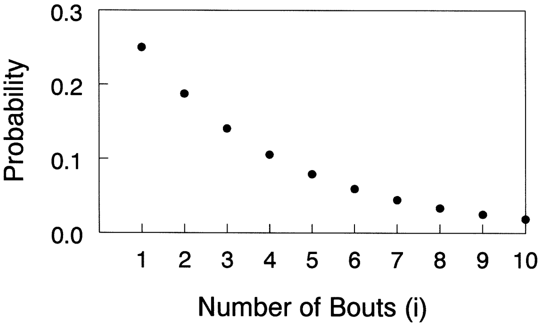

FIG. 2.8 The probability that a fringehead will first win on its ith match. Here we have assumed that in each bout the fish has a 25% chance of winning. This figure is a graphical representation of eq. (2.20).

2.4 *Probability Distributions*

In the last section we explored a variety of methods to calculate the probability for a particular outcome, and in the last example (how many bouts a fringehead will lose before winning) we even managed to derive an equation that describes the probability for each possible outcome. In other words, with a bit of diligent bookkeeping we can keep track of the frequency of occurrence of each and every outcome in the entire sample space, thereby associating each outcome with a probability. We can then lump outcomes into events and appropriately calculate the probability of each event.

This brings us to an important juncture in our exploration. Imagine writing down all the possible events in an experiment in one column of a table, and generating a companion column with each event's corresponding probability. This ensemble of paired outcomes and probabilities is called the *probability distribution* of the experiment.

Probability distributions will be of central importance throughout the rest of this book, and it will be best to take some time here to make sure that the concept is abundantly clear. Consider the same two smuts we have dealt with before, one with compatibility alleles a and b, the other with alleles c and d. Each produces an abundance of spores, and the spores are mixed randomly, and a single pair is chosen. What is the probability distribution for this reproductive experiment?

TABLE 2.1 The Probability Distribution for Spores Produced by the Mating of Two Smuts

Outcome	*Probability*	*Outcome*	*Probability*
a, a	1/16	c, a	1/16
a, b	1/16	c, b	1/16
a, c	1/16	c, c	1/16
a, d	1/16	c, d	1/16
b, a	1/16	d, a	1/16
b, b	1/16	d, b	1/16
b, c	1/16	d, c	1/16
b, d	1/16	d, d	1/16

Note: One smut has compatibility alleles a and b, the other has alleles c and d.

TABLE 2.2 The Probability Distribution for the Spores of the Two Smuts of Table 2.1

Event	*Probability*	*Event*	*Probability*
a, a	1/16	b, c	1/8
a, b	1/8	b, d	1/8
a, c	1/8	c, c	1/16
a, d	1/8	c, d	1/8
b, b	1/16	d, d	1/16

Note: In this case, the order of alleles is *not* taken into account.

First, we list the sample space for the simplest outcomes and their associated probabilities in table 2.1. In this case, each outcome has equal probability. But, as we have noted before, several of these simple outcomes are functionally equivalent [(a, b) and (b, a), for instance]. Thus, if we define an event as having a distinct allelic type *independent of order*, we have the list in table 2.2. The probability of an event in which alleles match is half that of events in which alleles are different.

We could simplify matters even more by again redefining what we mean by an event in the pairing of spores. Suppose that instead of tabulating the genotypes of paired spores we keep track of whether they fuse or not. In this case, the two possible events are *Fusion* and *Nonfusion*, and the corresponding probability distribution is as follows:

Event	*Probability*
Fusion	0.75
Nonfusion	0.25

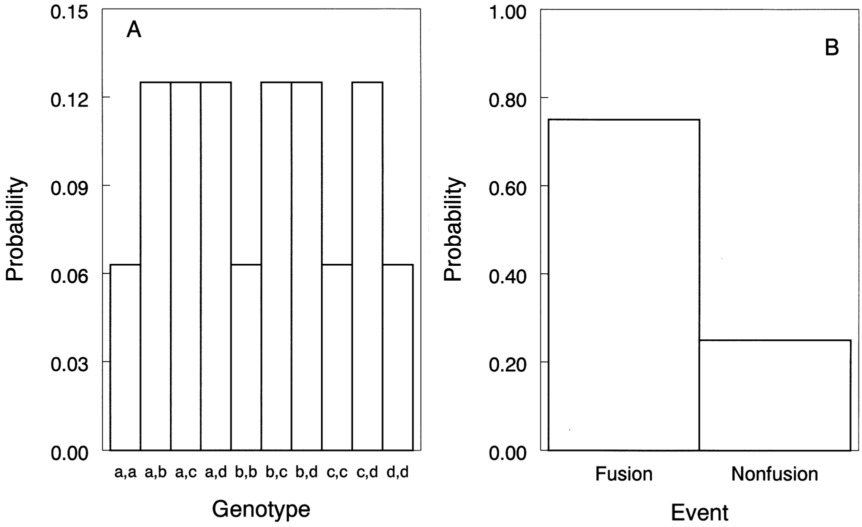

FIG. 2.9 Probability distributions for the mating between smuts 1 and 2. In panel A, events are defined in terms of the genotype of the paired spores. In panel B, events are defined with regard solely to whether the paired spores fuse. The same experiment can lead to different probability distributions depending on how events are defined.

Twelve of the sixteen possible ordered pairs are capable of fusion, so (given random pairing) fusion occurs three quarters of the time. The remaining pairs (those with matching compatibility alleles) do not fuse.

Tables of events and probabilities can be cumbersome when dealing with all but the smallest sample spaces. As a practical alternative, it is often handy to plot probability distributions graphically. This is traditionally done as a histogram. For example, the probability distribution of unordered allelic types is shown in figure 2.9A, and that for fusion/nonfusion in figure 2.9B. Alternatively, the probability distribution can be graphed as a scatter plot, as suggested by figure 2.8.

These simple probability distributions may seem obvious, and you are perhaps wondering why we are belaboring their existence. We have taken extra care in presenting the concept of probability distributions because these distributions are so fundamentally important. This importance lies in the fact that the probability distribution contains *all the information that can be known* about a given random experiment. Once you have established the criteria that define an event, have listed the sample space of possible outcomes, and have associated each event with a probability, you have in hand all the information that it is possible to obtain about the stochastic process in question.

This is not to say that this information cannot be processed further. For example, we will see in a moment how the probability distribution of simple events can be used to calculate the probability distribution of more complex events. In essence, this is what we did in working from the distribution of ordered allele pairs to the simple distribution of spore fusions. Knowledge of the probability distribution can also be used to calculate useful indices such as the average value of an experiment's outcomes and the typical variability of these outcomes. These uses of probability distributions will be covered in detail in chapters 3 and 4.

Note that it isn't necessary to present a probability distribution as a table. In many cases, a distribution can be described in more compact form by an equation. For example, eq. (2.20) describes the probability of event L (a fringehead winning) occurring first on trial i. Because this equation is in essence a shorthand notation for a list of L, $P(L)$, it describes the probability distribution for L.

2.5 *Summary*

In this chapter we have developed a set of tools to help deal with estimating the probability of relatively complex events. We can use Venn diagrams to answer complex problems with a somewhat brute-force, graphical approach: draw a diagram with all possible outcomes and then choose those outcomes that are included in your event. Alternatively, by decomposing the problem into a set of simpler events, we can break what seems like a difficult question into a series of manageable tasks. Our arsenal of tools now allows us to examine the probability of (1) outcomes that are shared by different events (=*intersections*), (2) outcomes that occur in any of two or more events (=*unions*), (3) outcomes that are not part of a particular event (=*complements*), and (4) outcomes whose chance of occurrence depends on the occurrence of other outcomes (=*conditional probabilities*). The combination of these tools (and some diligent bookkeeping) allows us to associate every event with its probability, and thereby to specify the *probability distribution* for an experiment. Through the use of the probability distribution we will be able to explore a number of biological phenomena in which chance plays a crucial role.

2.6 *Problems*

1. Hand gestures play an important role in many cultures. For example, a raised index finger with all other fingers folded is a common gesture among sports fans in the U.S., signifying "We're no. 1!" Inappropriate use of this gesture may be annoying to those around one but is unlikely to cause serious problems. In contrast, in the United States a raised middle finger with all other

fingers folded is used to signify extreme displeasure, and use of this gesture in the wrong situation or wrong company can lead to unfortunate results. Let us assume that every culture possesses exactly one such dangerous hand gesture, selected randomly sometime in the distant past. Now the problem. You have just parachuted into the jungles of New Guinea, and when rescued by the natives you wish to greet them with a friendly wave of your hand. At random, you raise anywhere between zero and five fingers (keeping the rest folded) and extend your hand toward your rescuers. What is the probability that you have just commited a grave social faux pas and are thereby in danger of being shot?

2. You and your spouse intend to have four children. Your mother-in-law contends that because there is equal probability of having a boy or a girl in each birth, it is most probable that you will have an equal number of boys and girls among your four kids. Is she correct? Why or why not?

3. Gambler 1 wins if he scores at least one “1” (an ace) in six throws of a single die. Gambler 2 wins if he gets at least two aces in twelve throws of a single die. Which gambler is more likely to win? It may motivate you as you work through the math to know that this problem was first posed by Samuel Pepys and solved by Sir Isaac Newton in 1693 (Feller 1960).

4. Four deer are captured from a population of N deer, marked, and released back into the population. After a time lapse sufficient to ensure that the marked deer are randomly distributed among the population, five deer are captured at random from the population. What is the probability that *exactly one* of these recaptured deer is marked if

- $N = 8$
- $N = 10$
- $N = 15$
- $N = 20$
- $N = 25$
- $N = 30$

Graph your results of probability versus population size. Can you provide an intuitive explanation for the shape of this curve?

5. A box of one hundred screws contains ten screws that are defective. You pick ten screws at random from the box. What is the probability that all ten screws you have chosen are good? What is the probability that exactly one is defective?

6. How many people do you have to assemble in a room (people picked at random from the general population) before there is at least an even chance ($P = 0.5$) that at least two have the same birthdate (e.g., February 4 or September 9)? Assume that the year has 365 days; that is, don't worry about leap year. *Hint*: Check your answer by using the same method to calculate the probability that two people will have the same birthday if there are 366 people in the room.

7. You are a contestant on a popular game show that allows you to choose from among three doors, each of which hides a prize. Behind one of the doors is the vacation of your dreams. Behind each of the other two doors is a block of moldy cheese. You make your selection and tell the host. In response, the host (who knows what is behind each door) opens one of the two doors you did not choose, revealing some odorous cheese. She then gives you a chance to change your mind and select the unopened door you did not choose originally. Would it be an advantage to switch? In other words, what is your probability of winning the vacation if you keep your original choice? What is your probability of winning if you switch? (*Note*: This puzzle stumped a number of mathematicians when it was posed in a newspaper in 1991; see Tierney 1991; Hoffman 1998.)

8. Suppose the game show in question 7 had four doors instead of three. If all rules of the game are otherwise the same, should you switch doors?

9. You live in a town of $n+1$ people, and are interested in the dynamics of rumors. You start a rumor by telling it to one other person, who then picks a person at random from the town and passes the rumor on. This second person likewise picks a recipient at random, and so forth. What is the probability that the rumor is told k times before it comes full circle and is repeated to you? What is the probability that the rumor is told k times before it is repeated to anyone? Work the problem again, but assume that each time the rumor is passed on, it is told to a group of N randomly chosen individuals. (This problem was borrowed from Feller 1960.)

10. Who's the father? A mare is placed in a corral with two stallions. One of the stallions is a champion thoroughbred racehorse worth millions of dollars. The other stallion looks similar but did not have such a distinguished racing career. The mare becomes pregnant and produces a colt. If the colt was fathered by the thoroughbred, he is worth a lot of money.

a. From the information given so far, what is the probability that the colt was fathered by the thoroughbred?

b. Suppose the thoroughbred has a relatively rare genetic marker on his Y chromosome that only occurs in 2% of horses. You know nothing about the genetics of the second stallion. You test the colt and find he also carries the rare marker. What is the probability the colt is the son of the thoroughbred given that he has the genetic marker?

c. Suppose the mare recently spent time roaming free on the range. During this time she was exposed to 998 other stallions who also could be the father of the colt. Now what is the probability that the colt is the son of the thoroughbred, given that he has the genetic marker?

d. What are the implications of this exercise to human legal trials where genetic markers are used to identify potential suspects?