2

Probability

2.1. Examples of probability

We learned something about sets in Chapter 1; now we are going to measure them. The most primitive way of measuring is to count the number, so we will begin with such an example.

Example 1. In Example (a′) of §1.1, suppose that the number of rotten apples is 28. This gives a measure to the set A described in (a′), called its size and denoted by $|A|$. But it does not tell anything about the total number of apples in the bushel, namely the size of the sample space Ω given in Example (a). If we buy a bushel of apples we are more likely to be concerned with the relative *proportion* of rotten ones in it rather than their absolute number. Suppose then the total number is 550. If we now use the letter P provisionarily for "proportion," we can write this as follows:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{28}{550}. \tag{2.1.1}$$

Suppose next that we consider the set B of unripe apples in the same bushel, whose number is 47. Then we have similarly

$$P(B) = \frac{|B|}{|\Omega|} = \frac{47}{550}.$$

It seems reasonable to suppose that an apple cannot be both rotten and unripe (this is really a matter of definition of the two adjectives); then the two sets are disjoint so their members do not overlap. Hence the number of “rotten or unripe apples” is equal to the sum of the number of “rotten apples” and the number of “unripe apples”: $28 + 47 = 75$. This may be written in symbols as:

$$|A + B| = |A| + |B|. \tag{2.1.2}$$

If we now divide through by $|\Omega|$, we obtain

$$P(A + B) = P(A) + P(B). \tag{2.1.3}$$

On the other hand, if some apples can be rotten and unripe at the same time, such as when worms got into green ones, then the equation (2.1.2) must be replaced by an inequality:

$$|A \cup B| \leq |A| + |B|,$$

which leads to

$$P(A \cup B) \leq P(A) + P(B). \tag{2.1.4}$$

Now what is the excess of $|A| + |B|$ over $|A \cup B|$? It is precisely the number of “rotten and unripe apples,” that is, $|A \cap B|$. Thus

$$|A \cup B| + |A \cap B| = |A| + |B|,$$

which yields the pretty equation

$$P(A \cup B) + P(A \cap B) = P(A) + P(B). \tag{2.1.5}$$

Example 2. A more sophisticated way of measuring a set is the area of a plane set as in Examples (f) and (f′) of §1.1, or the volume of a solid. It is said that the measurement of land areas was the origin of geometry and trigonometry in ancient times. While the nomads were still counting on their fingers and toes as in Example 1, the Chinese and Egyptians, among other peoples, were subdividing their arable lands, measuring them in units and keeping accounts of them on stone tablets or papyrus. This unit varied a great deal from one civilization to another (who knows the conversion rate of an acre into *mou*’s or hectares?). But again it is often the ratio of two areas that concerns us as in the case of a wild shot that hits the target board. The proportion of the area of a subset A to that of Ω may be written, if we denote the area by the symbol $|\ |$:

$$P(A) = \frac{|A|}{|\Omega|}. \tag{2.1.6}$$

This means also that if we fix the unit so that the total area of Ω is 1 unit, then the area of A is equal to the fraction $P(A)$ in this scale. Formula (2.1.6) looks just like formula (2.1.1) by the deliberate choice of notation in order to underline the similarity of the two situations. Furthermore, for two sets A and B the previous relations (2.1.3) to (2.1.5) hold equally well in their new interpretations.

Example 3. When a die is thrown there are six possible outcomes. If we compare the process of throwing a particular number [face] with that of picking a particular apple in Example 1, we are led to take $\Omega = \{1, 2, 3, 4, 5, 6\}$ and define

$$P(\{k\}) = \frac{1}{6}, \quad k = 1, 2, 3, 4, 5, 6. \tag{2.1.7}$$

Here we are treating the six outcomes as "equally likely," so that the same measure is assigned to all of them, just as we have done tacitly with the apples. This hypothesis is usually implied by saying that the die is "perfect." In reality, of course, no such die exists. For instance, the mere marking of the faces would destroy the perfect symmetry; and even if the die were a perfect cube, the outcome would still depend on the way it is thrown. Thus we must stipulate that this is done in a perfectly symmetrical way too, and so on. Such conditions can be approximately realized and constitute the basis of an assumption of equal likelihood on grounds of symmetry.

Now common sense demands an empirical interpretation of the "probability" given in (2.1.7). It should give a measure of what is *likely* to happen, and this is associated in the intuitive mind with the observable frequency of occurrence . Namely, if the die is thrown a number of times, how often will a particular face appear? More generally, let A be an event determined by the outcome; e.g., "to throw a number not less than 5 [or an odd number]." Let $N_n(A)$ denote the number of times the event A is observed in n throws; then the *relative frequency* of A in these trials is given by the ratio

$$Q_n(A) = \frac{N_n(A)}{n}. \tag{2.1.8}$$

There is good reason to take this Q_n as a measure of A. Suppose B is another event such that A and B are *incompatible* or *mutually exclusive* in the sense that they cannot occur in the same trial. Clearly we have $N_n(A + B) = N_n(A) + N_n(B)$, and consequently

$$\begin{aligned} Q_n(A + B) &= \frac{N_n(A + B)}{n} \\ &= \frac{N_n(A) + N_n(B)}{n} = \frac{N_n(A)}{n} + \frac{N_n(B)}{n} = Q_n(A) + Q_n(B). \end{aligned} \tag{2.1.9}$$

Similarly for any two events A and B in connection with the same game, not necessarily incompatible, the relations (2.1.4) and (2.1.5) hold with the P's there replaced by our present Q_n. Of course, this Q_n depends on n and will fluctuate, even wildly, as n increases. But if you let n go to infinity, will the sequence of ratios $Q_n(A)$ "settle down to a steady value"? Such a question can never be answered empirically, since by the very nature of a limit we cannot put an end to the trials. So it is a mathematical idealization to assume that such a limit does exist, and then write

$$Q(A) = \lim_{n\to\infty} Q_n(A). \tag{2.1.10}$$

We may call this the empirical *limiting frequency* of the event A. If you know how to operate with limits, then you can see easily that the relation (2.1.9) remains true "in the limit." Namely when we let $n \to \infty$ everywhere in that formula and use the definition (2.1.10), we obtain (2.1.3) with P replaced by Q. Similarly, (2.1.4) and (2.1.5) also hold in this context.

But the limit Q still depends on the actual sequence of trials that are carried out to determine its value. On the face of it, there is no guarantee whatever that another sequence of trials, even if it is carried out under the same circumstances, will yield the same value. Yet our intuition demands that a measure of the likelihood of an event such as A should tell something more than the mere record of one experiment. A viable theory built on the frequencies will have to assume that the Q defined above is in fact the same for all similar sequences of trials. Even with the hedge implicit in the word "similar," that is assuming a lot to begin with. Such an attempt has been made with limited success, and has a great appeal to common sense, but we will not pursue it here. Rather, we will use the definition in (2.1.7) which implies that if A is any subset of Ω and $|A|$ its size, then

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{6}. \tag{2.1.11}$$

For example, if A is the event "to throw an odd number," then A is identified with the set $\{1, 3, 5\}$ and $P(A) = 3/6 = 1/2$.

It is a fundamental proposition in the theory of probability that under certain conditions (repeated *independent* trials with *identical* die), the limiting frequency in (2.1.10) will indeed exist and be equal to $P(A)$ defined in (2.1.11), for "practically all" conceivable sequences of trials. This celebrated theorem, called the *Law of Large Numbers*, is considered to be the cornerstone of all empirical sciences. In a sense it justifies the intuitive foundation of probability as frequency discussed above. The precise statement and derivation will be given in Chapter 7. We have made this early announcement to quiet your feelings or misgivings about frequencies and to concentrate for the moment on sets and probabilities in the following sections.

2.2. Definition and illustrations

First of all, a probability is a number associated with or assigned to a set in order to measure it in some sense. Since we want to consider many sets at the same time (that is why we studied Chapter 1), and each of them will have a probability associated with it, this makes probability a "function of sets." You should have already learned in some mathematics course what a function means; in fact, this notion is used a little in Chapter 1. Nevertheless, let us review it in the familiar notation: a function f defined for some or all real numbers is a rule of association, by which we assign the number $f(x)$ to the number x. It is sometimes written as $f(\cdot)$, or more painstakingly as follows:

$$f : x \to f(x). \tag{2.2.1}$$

So when we say a probability is a function of sets we mean a similar association, except that x is replaced by a set S:

$$P : S \to P(S). \tag{2.2.2}$$

The *value* $P(S)$ is still a number; indeed it will be a number between 0 and 1. We have not been really precise in (2.2.1), because we have not specified the set of x there for which it has a meaning. This set may be the interval (a, b) or the half-line $(0, \infty)$ or some more complicated set called the domain of f. Now what is the domain of our probability function P? It must be a *set of sets* or, to avoid the double usage, a *family* (*class*) of sets. As in Chapter 1 we are talking about subsets of a fixed sample space Ω. It would be nice if we could use the family of *all* subsets of Ω, but unexpected difficulties will arise in this case if no restriction is imposed on Ω. We might say that if Ω is too large, namely when it contains uncountably many points, then it has too many subsets, and it becomes impossible to assign a probability to each of them and still satisfy a basic rule [Axiom (ii*) ahead] governing the assignments. However, if Ω is a finite or countably infinite set, then no such trouble can arise and we may indeed assign a probability to each and all of its subsets. This will be shown at the beginning of §2.4. You are supposed to know what a finite set is (although it is by no means easy to give a logical definition, while it is mere tautology to say that "it has only a finite number of points"); let us review what a countably infinite set is. This notion will be of sufficient importance to us, even if it only lurks in the background most of the time.

A set is countably infinite when it can be put into 1-to-1 correspondence with the set of positive integers. This correspondence can then be exhibited by labeling the elements as $\{s_1, s_2, \dots, s_n, \dots\}$. There are, of course, many ways of doing this, for instance we can just let some of the elements swap labels (or places if they are thought of being laid out in a row). The set of positive rational numbers is countably infinite, hence they can be labeled

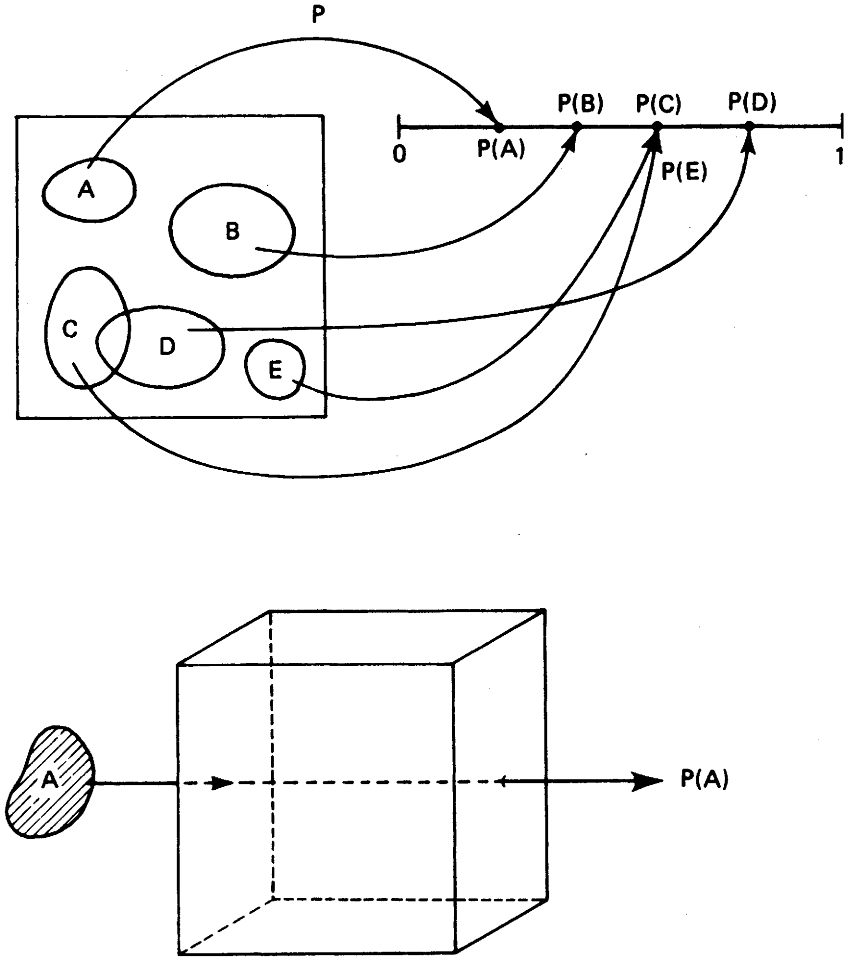

Figure 11

in some way as $\{r_1, r_2, \dots, r_n, \dots\}$, but don't think for a moment that you can do this by putting them in increasing order as you can with the positive integers $1 < 2 < \cdots < n < \cdots$. From now on we shall call a set *countable* when it is either finite or countably infinite. Otherwise it is called *uncountable*. For example, the set of all real numbers is uncountable. We shall deal with uncountable sets later, and we will review some properties of a countable set when we need them. For the present we will assume the sample space Ω to be countable in order to give the following definition in its simplest form, without a diverting complication. As a matter of fact, we could even assume Ω to be finite as in Examples (a) to (e) of §1.1, without losing the essence of the discussion below.

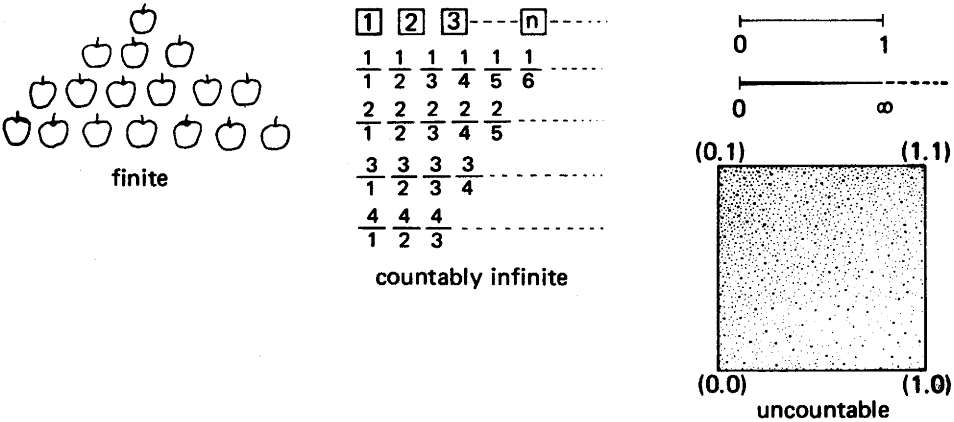

Figure 12

Definition. A *probability measure* on the sample space Ω is a function of subsets of Ω satisfying three axioms:

(i) For every set $A \subset \Omega$, the value of the function is a nonnegative number: $P(A) \geq 0$.

(ii) For any two disjoint sets A and B, the value of the function for their union $A + B$ is equal to the sum of its value for A and its value for B:

$$P(A + B) = P(A) + P(B) \quad \text{provided} \quad AB = \emptyset.$$

(iii) The value of the function for Ω (as a subset) is equal to 1:

$$P(\Omega) = 1.$$

Observe that we have been extremely careful in distinguishing the function $P(\cdot)$ from its values such as $P(A)$, $P(B)$, $P(A + B)$, $P(\Omega)$. Each of these is "a probability," but the function itself should properly be referred to as a "probability measure" as indicated.

Example 1 in §2.1 shows that the proportion P defined there is in fact a probability measure on the sample space, which is a bushel of 550 apples. It assigns a probability to every subset of these apples, and this assignment satisfies the three axioms above. In Example 2 if we take Ω to be all the land that belonged to the Pharaoh, it is unfortunately not a countable set. Nevertheless we can define the area for a very large class of subsets that are called "measurable," and if we restrict ourselves to these subsets only, the "area function" is a probability measure as shown in Example 2 where this restriction is ignored. Note that Axiom (iii) reduces to a convention: the decree of a unit. Now how can a land area not be measurable? While this is a sophisticated mathematical question that we will not go into in this book, it is easy to think of practical reasons for the possibility: the piece of land may be too jagged, rough, or inaccessible (see Fig. 13).

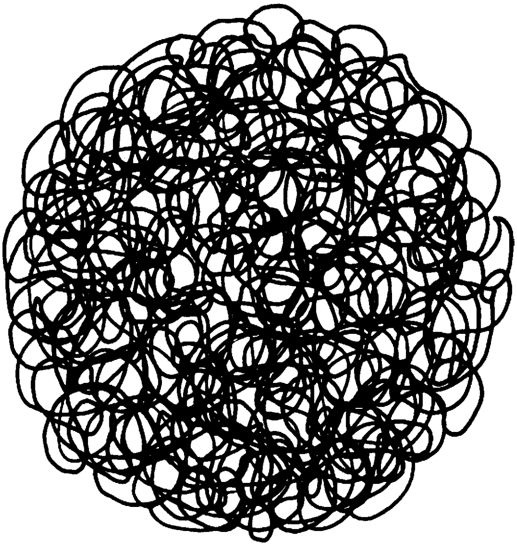

Figure 13

In Example 3 we have shown that the empirical relative frequency is a probability measure. But we will not use this definition in this book. Instead, we will use the first definition given at the beginning of Example 3, which is historically the earliest of its kind. The general formulation will now be given.

Example 4. A classical enunciation of probability runs as follows. The probability of an event is the ratio of the number of cases *favorable* to that event to the total number of cases, provided that all these are *equally likely*

.

To translate this into our language: the sample space is a finite set of possible cases: $\{\omega_1, \omega_2, \dots, \omega_m\}$, each ω_i being a "case." An event A is a subset $\{\omega_{i_1}, \omega_{i_2}, \dots, \omega_{i_n}\}$, each ω_{i_j} being a "favorable case." The probability of A is then the ratio

$$P(A) = \frac{|A|}{|\Omega|} = \frac{n}{m}. \tag{2.2.3}$$

As we see from the discussion in Example 1, this defines a probability measure P on Ω anyway, so that the stipulation above that the cases be equally likely is superfluous from the axiomatic point of view. Besides, what does it really mean? It sounds like a bit of tautology, and how is one going to decide whether the cases are equally likely or not?

A celebrated example will illustrate this. Let two coins be tossed. D'Alembert (mathematician, philosopher, and encyclopedist, 1717–83) argued that there are three possible cases, namely:

(i) both heads, (ii) both tails, (iii) a head and a tail.

So he went on to conclude that the probability of "a head and a tail" is equal to 1/3. If he had figured that this *probability* should have something to do with the experimental *frequency* of the occurrence of the event, he might have changed his mind after tossing two coins more than a few times. (History does not record if he ever did that, but it is said that for centuries people believed that men had more teeth than women because Aristotle had said so, and apparently nobody bothered to look into a few mouths.) The three cases he considered are not equally likely. Case (iii) should be split into two:

(iiia) first coin shows head and second coin shows tail.
(iiib) first coin shows tail and second coin shows head.

It is the four cases (i), (ii), (iiia) and (iiib) that are equally likely by symmetry and on empirical evidence. This should be obvious if we toss the two coins one after the other rather than simultaneously. However, there is an important point to be made clear here. The two coins may be physically indistinguishable so that in so far as actual observation is concerned, D'Alembert's three cases are the only distinct *patterns* to be recognized. In the model of two coins they happen not to be equally likely on the basis of common sense and experimental evidence. But in an analogous model for certain microcosmic particles, called Bose–Einstein statistics (see Exercise 24 of Chapter 3), they are indeed assumed to be equally likely in order to explain some types of physical phenomena. Thus what we regard as "equally likely" is a matter outside the axiomatic formulation. To put it another way, if we use (2.2.3) as our definition of probability then we are in effect treating the ω's as equally likely, in the sense that we count only their numbers and do not attach different weights to them.

Example 5. If six dice are rolled, what is the probability that all show different faces?

This is just Example (e) and (e$'$). It is stated elliptically on purpose to get you used to such problems. We have already mentioned that the total number of possible outcomes is equal to $6^6 = 46656$. They are supposed to be all "equally likely" although we never breathed a word about this assumption. Why, nobody can solve the problem as announced without such an assumption. Other data about the dice would have to be given before we could begin—which is precisely the difficulty when similar problems arise in practice. Now if the dice are all perfect, and the mechanism by which they are rolled is also perfect, which excludes any collusion between the movements of the several dice, then our hypothesis of equal likelihood may be justified. Such conditions are taken for granted in a problem like this when nothing is said about the dice. The solution is then given by (2.2.3) with $n = 6^6$ and $m = 6!$ (see Example 2 in §3.1 for these computations):

$$\frac{6!}{6^6} = \frac{720}{46656} = .015432$$

approximately.

Let us note that if the dice are not distinguishable from each other, then to the observer there is exactly one *pattern* in which the six dice show different faces. Similarly, the total number of different patterns when six dice are rolled is much smaller than 6^6 (see Example 3 of §3.2). Yet when we count the possible outcomes we must think of the dice as distinguishable, as if they were painted in different colors. This is one of the vital points to grasp in the counting cases; see Chapter 3.

In some situations the equally likely cases must be searched out. This point will be illustrated by a famous historical problem called the "problem of points."

Example 6. Two players A and B play a series of games in which the probability of each winning a single game is equal to 1/2, irrespective [independent] of the outcomes of other games. For instance, they may play tennis in which they are equally matched, or simply play "heads or tails" by tossing an unbiased coin. Each player gains a "point" when he wins a game, and nothing when he loses. Suppose that they stop playing when A needs 2 more points and B needs 3 more points to win the stake. How should they divide it fairly?

It is clear that the winner will be decided in 4 more games. For in those 4 games either A will have won ≥ 2 points or B will have won ≥ 3 points, but not both. Let us enumerate all the possible outcomes of these 4 games using the letter A or B to denote the winner of each game:

$$\begin{array}{lllll}
AAAA & AAAB & AABB & ABBB & BBBB \\
 & AABA & ABAB & BABB & \\
 & ABAA & ABBA & BBAB & \\
 & BAAA & BAAB & BBBA & \\
 & & BABA & & \\
 & & BBAA & &
\end{array}$$

These are equally likely cases on grounds of symmetry. There are* $\binom{4}{4} + \binom{4}{3} + \binom{4}{2} = 11$ cases in which A wins the stake; and $\binom{4}{3} + \binom{4}{4} = 5$ cases

*See (3.2.3) for notation used below.

in which B wins the stake. Hence the stake should be divided in the ratio 11:5. Suppose it is \$64000; then A gets \$44000, B gets \$20000. [We are taking the liberty of using the dollar as currency; the United States did not exist at the time when the problem was posed.]

This is Pascal's solution in a letter to Fermat dated August 24, 1654 . [Blaise Pascal (1623–62); Pierre de Fermat (1601–65); both among the greatest mathematicians of all time.] Objection was raised by a learned contemporary (and repeated through the ages) that the enumeration above was not reasonable, because the series would have stopped as soon as the winner was decided and not have gone on through all 4 games in some cases. Thus the real possibilities are as follows:

$$\begin{array}{ll}
AA & ABBB \\
ABA & BABB \\
ABBA & BBAB \\
BAA & BBB \\
BABA & \\
BBAA &
\end{array}$$

But these are not equally likely cases. In modern terminology, if these 10 cases are regarded as constituting the sample space, then

$$\begin{aligned}
P(AA) = \frac{1}{4}, \quad P(ABA) &= P(BAA) = P(BBB) = \frac{1}{8}, \\
P(ABBA) = P(BABA) &= P(BBAA) = P(ABBB) \\
&= P(BABB) = P(BBAB) = \frac{1}{16}
\end{aligned}$$

since A and B are independent events with probability $1/2$ each (see §2.4). If we add up these probabilities we get of course

$$\begin{aligned}
P(A \text{ wins the stake}) &= \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{8} + \frac{1}{16} + \frac{1}{16} = \frac{11}{16}, \\
P(B \text{ wins the stake}) &= \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{8} = \frac{5}{16}.
\end{aligned}$$

Pascal did not quite explain his method this way, saying merely that "it is absolutely equal and indifferent to each whether they play in the natural way of the game, which is to finish as soon as one has his score, or whether they play the entire four games." A later letter by him seems to indicate that he fumbled on the same point in a similar problem with three players. The student should take heart that this kind of reasoning was not easy even for past masters.

2.3. Deductions from the axioms

In this section we will do some simple "axiomatics." That is to say, we shall deduce some properties of the probability measure from its definition, using, of course, the axioms but nothing else. In this respect the axioms of a mathematical theory are like the constitution of a government. Unless and until it is changed or amended, every *law* must be made to follow from it. In mathematics we have the added assurance that there are no divergent views as to how the constitution should be construed.

We record some consequences of the axioms in (iv) to (viii) below. First of all, let us show that a probability is indeed a number between 0 and 1.

(iv) For any set A, we have

$$P(A) \leq 1.$$

This is easy, but you will see that in the course of deducing it we shall use all three axioms. Consider the complement A^c as well as A. These two sets are disjoint and their union is Ω:

$$A + A^c = \Omega. \tag{2.3.1}$$

So far, this is just set theory, no probability theory yet. Now use Axiom (ii) on the left side of (2.3.1) and Axiom (iii) on the right:

$$P(A) + P(A^c) = P(\Omega) = 1. \tag{2.3.2}$$

Finally use Axiom (i) for A^c to get

$$P(A) = 1 - P(A^c) \leq 1.$$

Of course, the first inequality above is just Axiom (i). You might object to our slow pace above by pointing out that since A is *contained in* Ω, it is obvious that $P(A) \leq P(\Omega) = 1$. This reasoning is certainly correct, but we still have to pluck it from the axioms, and that is the point of the little proof above. We can also get it from the following more general proposition.

(v) For any two sets such that $A \subset B$, we have

$$P(A) \leq P(B), \quad \text{and} \quad P(B - A) = P(B) - P(A).$$

The proof is an imitation of the preceding one with B playing the role of Ω. We have

$$B = A + (B - A),$$
$$P(B) = P(A) + P(B - A) \geq P(A).$$

The next proposition is such an immediate extension of Axiom (ii) that we could have adopted it instead as an axiom.

(vi) For any finite number of disjoint sets $A_1, \dots, A_n$, we have

$$P(A_1 + \cdots + A_n) = P(A_1) + \cdots + P(A_n). \tag{2.3.3}$$

This property of the probability measure is called *finite additivity* . It is trivial if we recall what "disjoint" means and use (ii) a few times; or we may proceed by induction if we are meticulous. There is an important extension of (2.3.3) to a countable number of sets later, *not* obtainable by induction!

As already checked in several special cases, there is a generalization of Axiom (ii), hence also of (2.3.3), to sets that are not necessarily disjoint. You may find it trite, but it has the dignified name of *Boole's inequality.* Boole (1815–64) was a pioneer in the "laws of thought" and author of *Theories of Logic and Probabilities.*

(vii) For any finite number of arbitrary sets $A_1, \dots, A_n$, we have

$$P(A_1 \cup \cdots \cup A_n) \le P(A_1) + \cdots + P(A_n). \tag{2.3.4}$$

Let us first show this when $n = 2$. For any two sets A and B, we can write their union as the sum of disjoint sets as follows:

$$A \cup B = A + A^c B. \tag{2.3.5}$$

Now we apply Axiom (ii) to get

$$P(A \cup B) = P(A) + P(A^c B). \tag{2.3.6}$$

Since $A^c B \subset B$, we can apply (v) to get (2.3.4).

The general case follows easily by mathematical induction, and you should write it out as a good exercise on this method. You will find that you need the associative law for the union of sets as well as that for the addition of numbers.

The next question is the difference between the two sides of the inequality (2.3.4). The question is somewhat moot since it depends on what we want to use to express the difference. However, when $n = 2$ there is a clear answer.

(viii) For any two sets A and B, we have

$$P(A \cup B) + P(A \cap B) = P(A) + P(B). \tag{2.3.7}$$

This can be gotten from (2.3.6) by observing that $A^cB = B - AB$, so that we have by virtue of (v):

$$P(A \cup B) = P(A) + P(B - AB) = P(A) + P(B) - P(AB),$$

which is equivalent to (2.3.7). Another neat proof is given in Exercise 12.

We shall postpone a discussion of the general case until §6.2. In practice, the inequality is often more useful than the corresponding identity which is rather complicated.

We will not quit formula (2.3.7) without remarking on its striking resemblance to formula (1.4.8) of §1.4, which is repeated below for the sake of comparison:

$$I_{A \cup B} + I_{A \cap B} = I_A + I_B. \tag{2.3.8}$$

There is indeed a deep connection between the pair, as follows. The probability $P(S)$ of each set S can be obtained from its indicator function I_S by a procedure (operation) called "taking expectation" or "integration." If we perform this on (2.3.8) term by term, their result is (2.3.7). This procedure is an essential part of probability theory and will be thoroughly discussed in Chapter 6. See Exercise 19 for a special case.

To conclude our axiomatics, we will now strengthen Axiom (ii) or its immediate consequence (vi), namely the finite additivity of P, into a new axiom.

(ii*) Axiom of countable additivity . For a countably infinite collection of disjoint sets A_k, $k = 1, 2, \dots$, we have

$$P\left(\sum_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k). \tag{2.3.9}$$

This axiom includes (vi) as a particular case, for we need only put $A_k = \emptyset$ for $k > n$ in (2.3.9) to obtain (2.3.3). The empty set is disjoint from any other set including itself and has probability zero (why?). If Ω is a finite set, then the new axiom reduces to the old one. But it is important to see why (2.3.9) *cannot* be deduced from (2.3.3) by letting $n \to \infty$. Let us try this by rewriting (2.3.3) as follows:

$$P\left(\sum_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k). \tag{2.3.10}$$

Since the left side above cannot exceed 1 for all n, the series on the right side must converge and we obtain

$$\lim_{n\to\infty} P\left(\sum_{k=1}^{n} A_k\right) = \lim_{n\to\infty} \sum_{k=1}^{n} P(A_k) = \sum_{k=1}^{\infty} P(A_k). \tag{2.3.11}$$

Comparing this established result with the desired result (2.3.9), we see that the question boils down to

$$\lim_{n\to\infty} P\left(\sum_{k=1}^{n} A_k\right) = P\left(\sum_{k=1}^{\infty} A_k\right),$$

which can be exhibited more suggestively as

$$\lim_{n\to\infty} P\left(\sum_{k=1}^{n} A_k\right) = P\left(\lim_{n\to\infty}\sum_{k=1}^{n} A_k\right). \qquad (2.3.12)$$

See end of §1.3 (see Fig. 14).

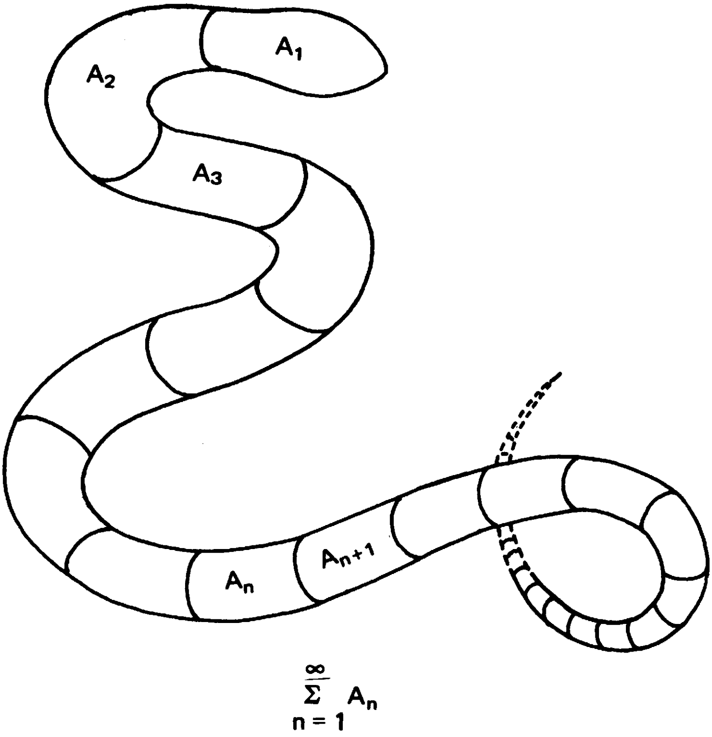

Figure 14

Thus it is a matter of interchanging the two operations "lim" and "P" in (2.3.12), or you may say, "taking the limit inside the probability relation." If you have had enough calculus you know this kind of interchange is often hard to justify and may be illegitimate or even invalid. The new axiom is created to secure it in the present case and has fundamental consequences in the theory of probability.

2.4. Independent events

From now on, a "probability measure" will satisfy Axioms (i), (ii*), and (iii). The subsets of Ω to which such a probability has been assigned will also be called an *event*.

We shall show how easy it is to *construct* probability measures for any countable space $\Omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\}$. To each sample point ω_n let us attach an arbitrary "weight" p_n subject only to the conditions

$$\forall n\colon \quad p_n \geq 0, \ \sum_n p_n = 1. \tag{2.4.1}$$

This means that the weights are positive or zero, and add up to 1 altogether. Now for any subset A of Ω, we define its probability to be the *sum of the weights of all the points in it.* In symbols, we put first

$$\forall n\colon \quad P(\{\omega_n\}) = p_n; \tag{2.4.2}$$

and then for every $A \subset \Omega$:

$$P(A) = \sum_{\omega_n \in A} p_n = \sum_{\omega_n \in A} P(\{\omega_n\}).$$

We may write the last term above more neatly as

$$P(A) = \sum_{\omega \in A} P(\{\omega\}). \tag{2.4.3}$$

Thus P is a function defined for all subsets of Ω and it remains to check that it satisfies Axioms (i), (ii*), and (iii). This requires nothing but a bit of clearheaded thinking and is best done by yourself. Since the weights are quite arbitrary apart from the easy conditions in (2.4.1), you see that probability measures come "a dime a dozen" in a countable sample space. In fact, we can get them all by the above method of construction. For if any probability measure P is given, never mind how, we can define p_n to be $P(\{\omega_n\})$ as in (2.4.2), and then $P(A)$ must be given as in (2.4.3), *because of Axiom* (ii*). Furthermore the p_n's will satisfy (2.4.1) as a simple consequence of the axioms. In other words, any given P is necessarily of the type described by our construction.

In the very special case that Ω is finite and contains exactly m points, we may attach equal weights to all of them, so that

$$p_n = \frac{1}{m}, \quad n = 1, 2, \dots, m.$$

Then we are back to the "equally likely" situation in Example 4 of §2.2. But in general the p_n's need not be equal, and when Ω is countably infinite they cannot all be equal (why?). The preceding discussion shows the degree of arbitrariness involved in the general concept of a probability measure.

An important model of probability space is that of *repeated independent trials* : this is the model used when a coin is tossed, a die thrown, a card drawn from a deck (with replacement) several times. Alternately, we may toss several coins or throw several dice at the same time. Let us begin with an example.

Example 7. First toss a coin, then throw a die, finally draw a card from a deck of poker cards. Each trial produces an event; let

$$\begin{aligned} A &= \text{coin falls heads;} \\ B &= \text{die shows number 5 or 6;} \\ C &= \text{card drawn is a spade.} \end{aligned}$$

Assume that the coin is fair, the die is perfect, and the deck thoroughly shuffled. Furthermore assume that these three trials are carried out "independently" of each other, which means intuitively that the outcome of each trial does not influence that of the others. For instance, this condition is approximately fulfilled if the trials are done by different people in different places, or by the same person in different months! Then all possible joint outcomes may be regarded as equally likely. There are respectively 2, 6, and 52 possible cases for the individual trials, and the total number of cases for the whole set of trials is obtained by multiplying these numbers together: $2 \cdot 6 \cdot 52$ (as you will soon see it is better not to compute this product). This follows from a fundamental rule of counting, which is fully discussed in §3.1 and which you should read now if need be. [In general, many parts of this book may be read in different orders, back and forth.] The same rule yields the numbers of favorable cases to the events A, B, C, AB, AC, BC, ABC given below, where the symbol $|\dots|$ for size is used:

$$\begin{aligned} |A| &= 1 \cdot 6 \cdot 52, \quad &|B| &= 2 \cdot 2 \cdot 52, \quad &|C| &= 2 \cdot 6 \cdot 13, \\ |AB| &= 1 \cdot 2 \cdot 52, \quad &|AC| &= 1 \cdot 6 \cdot 13, \quad &|BC| &= 2 \cdot 2 \cdot 13, \\ |ABC| &= 1 \cdot 2 \cdot 13. \end{aligned}$$

Dividing these numbers by $|\Omega| = 2 \cdot 6 \cdot 52$, we obtain after quick cancellation of factors:

$$\begin{aligned} P(A) &= \frac{1}{2}, \quad &P(B) &= \frac{1}{3}, \quad &P(C) &= \frac{1}{4}, \\ P(AB) &= \frac{1}{6}, \quad &P(AC) &= \frac{1}{8}, \quad &P(BC) &= \frac{1}{12}, \\ P(ABC) &= \frac{1}{24}. \end{aligned}$$

We see at a glance that the following set of equations holds:

$$P(AB) = P(A)P(B), \ P(AC) = P(A)P(C), \ P(BC) = P(B)P(C) \quad (2.4.4)$$
$$P(ABC) = P(A)P(B)P(C).$$

The reader is now asked to convince himself that this set of relations will also hold for any three events A, B, C such that A is determined by the coin, B by the die, and C by the card drawn *alone*. When this is the case we say that these trials are *stochastically independent* as well as the events so produced. The adverb "stochastically" is usually omitted for brevity.

The astute reader may observe that we have not formally defined the word "trial," and yet we are talking about independent trials! A logical construction of such objects is quite simple but perhaps a bit too abstract for casual introduction. It is known as "product space"; see Exercise 29. However, it takes less fuss to define "independent events" and we shall do so at once.

Two events A and B are said to be independent if we have $P(AB) = P(A)P(B)$. Three events A, B, and C are said to be independent if the relations in (2.4.4) hold. Thus independence is a notion relative to a given probability measure (by contrast, the notion of disjointness, e.g., does not depend on any probability). More generally, the n events $A_1, A_2, \dots, A_n$ are independent if the intersection [joint occurrence] of any subset of them has as its probability the product of probabilities of the individual events. If you find this sentence too long and involved, you may prefer the following symbolism. For any subset $(i_1, i_2, \dots, i_k)$ of $(1, 2, \dots, n)$, we have

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}). \quad (2.4.5)$$

Of course, here the indices $i_1, \dots, i_k$ are distinct and $1 \leq k \leq n$.

Further elaboration of the notion of independence is postponed to §5.5, because it will be better explained in terms of random variables. But we shall briefly describe a classical scheme—the grand daddy of repeated trials, and subject of intensive and extensive research by J. Bernoulli, De Moivre, Laplace, ... , Borel, ...

Example 8. (The coin-tossing scheme). A coin is tossed repeatedly n times. The joint outcome may be recorded as a sequence of H's and T's, where $H =$ "head," $T =$ "tail." It is often convenient to *quantify* by putting $H = 1$, $T = 0$; or $H = 1$, $T = -1$; we shall adopt the first usage here. Then the result is a sequence of 0's and 1's consisting of n terms such as 110010110 with $n = 9$. Since there are 2 outcomes for each trial, there are 2^n possible joint outcomes. This is another application of the fundamental rule in §3.1. If all of these are assumed to be equally likely so that each particular joint outcome has probability $1/2^n$, then we can proceed as in Example 7 to verify that the trials are independent and the coin is fair. You will find this a dull exercise, but it is recommended that you go through it in your head if not on paper. However, we will turn the table around here by *assuming at the outset* that the successive tosses do form independent trials. On the other hand, we do not assume the coin to be "fair," but only that the probabilities for head (H) and tail (T) remain constant throughout the trials. Empirically speaking, this is only approximately true since things do not really remain unchanged over long periods of time. Now we need a precise notation to record complicated statements, ordinary words being often awkward or ambiguous. Then let X_i denote the outcome of the ith trial and let ϵ_i denote 0 or 1 for each i, but of course varying with the subscript. Then our hypothesis above may be written as follows:

$$P(X_i = 1) = p; \quad P(X_i = 0) = 1 - p; \quad i = 1, 2, \dots, n; \tag{2.4.6}$$

where p is the probability of heads for each trial. For any particular, namely completely specified, sequence $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ of 0's and 1's, the probability of the corresponding sequence of outcomes is equal to

$$\begin{aligned} P(X_1 = \epsilon_1, X_2 = \epsilon_2, \dots, X_n = \epsilon_n) \\ = P(X_1 = \epsilon_1)P(X_2 = \epsilon_2)\dots P(X_n = \epsilon_n) \end{aligned} \tag{2.4.7}$$

as a consequence of independence. Now each factor on the right side above is equal to p or $1 - p$ depending on whether the corresponding ϵ_i is 1 or 0. Suppose j of these are 1's and $n - j$ are 0's; then the quantity in (2.4.7) is equal to

$$p^j(1 - p)^{n-j}. \tag{2.4.8}$$

Observe that for each sequence of trials, the number of heads is given by the sum $\sum_{i=1}^n X_i$. It is important to understand that the number in (2.4.8) is not the probability of obtaining j heads in n tosses, but rather that of obtaining a specific sequence of heads and tails in which there are j heads. In order to compute the former probability, we must count the total number of the latter sequences since all of them have the same probability given in (2.4.8). This number is equal to the binomial coefficient $\binom{n}{j}$; see §3.2 for a full discussion. Each one of these $\binom{n}{j}$ sequences corresponds to one possibility of obtaining j heads in n trials, and these possibilities are disjoint. Hence it follows from the additivity of P that we have

$$\begin{aligned} P\left(\sum_{i=1}^n X_i = j\right) &= P(\text{exactly } j \text{ heads in } n \text{ trials}) \\ &= \binom{n}{j} P(\text{any specified sequence of } n \text{ trials with exactly } j \text{ heads}) \\ &= \binom{n}{j} p^j(1 - p)^{n-j}. \end{aligned}$$

This famous result is known as *Bernoulli's formula.* We shall return to it many times in the book.

2.5. Arithmetical density*

We study in this section a very instructive example taken from arithmetic.

Example 9. Let Ω be the first 120 *natural numbers* $\{1, 2, \dots, 120\}$. For the probability measure P we use the proportion as in Example 1 of §2.1. Now consider the sets

$$A = \{\omega \mid \omega \text{ is a multiple of } 3\},$$
$$B = \{\omega \mid \omega \text{ is a multiple of } 4\}.$$

Then every third number of Ω belongs to A, and every fourth to B. Hence we get the proportions

$$P(A) = 1/3, \quad P(B) = 1/4.$$

What does the set AB represent? It is the set of integers that are divisible by 3 and by 4. If you have not entirely forgotten your school arithmetic, you know this is just the set of multiples of $3 \cdot 4 = 12$. Hence $P(AB) = 1/12$. Now we can use (viii) to get $P(A \cup B)$:

$$P(A \cup B) = P(A) + P(B) - P(AB) = 1/3 + 1/4 - 1/12 = 1/2. \quad (2.5.1)$$

What does this mean? $A \cup B$ is the set of those integers in Ω which are divisible by 3 or by 4 (or by both). We can count them one by one, but if you are smart you see that you don't have to do this drudgery. All you have to do is to count up to 12 (which is 10% of the whole population Ω), and check them off as shown:

$$\begin{array}{cccccccccccc} 1, & 2, & 3, & 4, & 5, & 6, & 7, & 8, & 9, & 10, & 11, & 12. \\ & & \checkmark & \checkmark & & \checkmark & & \checkmark & \checkmark & & & \checkmark \\ & & & & & & & & & & & \checkmark \end{array}$$

Six are checked (one checked twice), hence the proportion of $A \cup B$ among these 12 is equal to $6/12 = 1/2$ as given by (2.5.1).

An observant reader will have noticed that in the case above we have also

$$P(AB) = 1/12 = 1/3 \cdot 1/4 = P(A) \cdot P(B).$$

*This section may be omitted.

This is true because the two numbers 3 and 4 happen to be *relatively prime*, namely they have no common divisor except 1. Suppose we consider another set:

$$C = \{\omega \mid \omega \text{ is a multiple of } 6\}.$$

Then $P(C) = 1/6$, but what is $P(BC)$ now? The set BC consists of those integers that are divisible by both 4 and 6, namely divisible by their *least common multiple* (remember that?), which is 12 and not the product $4 \cdot 6 = 24$. Thus $P(BC) = 1/12$. Furthermore, because 12 is the least common multiple we can again stop counting at 12 in computing the proportion of the set $B \cup C$. An actual counting gives the answer $4/12 = 1/3$, which may also be obtained from the formula (2.3.7):

$$P(B \cup C) = P(B) + P(C) - P(BC) = 1/4 + 1/6 - 1/12 = 1/3. \quad (2.5.2)$$

This example illustrates a point that arose in the discussion in Example 3 of §2.1. Instead of talking about the proportion of the multiples of 3, say, we can talk about its frequency. Here no rolling of any fortuitous dice is needed. God has given us those natural numbers (a great mathematician Kronecker said so), and the multiples of 3 occur at perfectly regular periods with the frequency 1/3. In fact, if we use $N_n(A)$ to denote the number of natural numbers up to and including n which belong to the set A, it is a simple matter to show that

$$\lim_{n\to\infty} \frac{N_n(A)}{n} = \frac{1}{3}.$$

Let us call this $P(A)$, the limiting frequency of A. Intuitively, it should represent the chance of picking a number divisible by 3, if we can reach into the whole bag of natural numbers as if they were so many indistinguishable balls in an urn. Of course, similar limits exist for the sets B, C, AB, BC, etc. and have the values computed above. But now with this infinite sample space of "all natural numbers," call it Ω^*, we can treat by the same method any set of the form

$$A_m = \{\omega \mid \omega \text{ is divisible by } m\}, \quad (2.5.3)$$

where m is an arbitrary natural number. Why then did we not use this more natural and comprehensive model?

The answer may be a surprise for you. By our definition of probability measure given in §2.2, we should have required that every subset of Ω^* have a probability, provided that Ω^* is countable, which is the case here. Now take for instance the set that consists of the single number $\{1971\}$ or, if you prefer, the set $Z = \{\text{all numbers from 1 to 1971}\}$. Its probability is given by $\lim_{n\to\infty} N_n(Z)/n$ according to the same rule that was applied to the set

A. But $N_n(Z)$ is equal to 1971 for all values of $n \geq 1971$; hence the limit above is equal to 0 and we conclude that every finite set has probability 0 by this rule. If P were to be countably additive as required by Axiom (ii*) in §2.3, then $P(\Omega^*)$ would be 0 rather than 1. This contradiction shows that P cannot be a probability measure on Ω^*. Yet it works perfectly well for sets such as A_m.

There is a way out of this paradoxical situation. We must abandon our previous requirement that the measure be defined for all subsets (of natural numbers). Let a finite number of the sets A_m be given, and let us consider the *composite sets* that can be obtained from these by the operations complementation, union, and intersection. Call this class of sets the class *generated by* the original sets. Then it is indeed possible to define P in the manner prescribed above for all sets in *this* class. A set that is not in the class has no probability at all. For example, the set Z does not belong to the class generated by A, B, C. Hence its probability is *not* defined, rather than zero. We may also say that the set Z is *nonmeasurable* in the context of Example 2 of §2.1. This saves the situation, but we will not pursue it further here except to give another example.

Example 10. What is the probability of the set of numbers divisible by 3, not divisible by 5, and divisible by 4 or 6?

Using the preceding notation, the set in question is $AD^c(B \cup C)$, where $D = A_5$. Using distributive law, we can write this as $AD^cB \cup AD^cC$. We also have

$$(AD^cB)(AD^cC) = AD^cBC = ABC - ABCD.$$

Hence by (v),

$$P(AD^cBC) = P(ABC) - P(ABCD) = \frac{1}{12} - \frac{1}{60} = \frac{1}{15}.$$

Similarly, we have

$$P(AD^cB) = P(AB) - P(ABD) = \frac{1}{12} - \frac{1}{60} = \frac{4}{60} = \frac{1}{15},$$
$$P(AD^cC) = P(AC) - P(ACD) = \frac{1}{6} - \frac{1}{30} = \frac{4}{30} = \frac{2}{15}.$$

Finally we obtain by (viii):

$$\begin{aligned} P(AD^cB \cup AD^cC) &= P(AD^cB) + P(AD^cC) - P(AD^cBC) \\ &= \frac{1}{15} + \frac{2}{15} - \frac{1}{15} = \frac{2}{15}. \end{aligned}$$

You should check this using the space Ω in Example 9.

The problem can be simplified by a little initial arithmetic, because the set in question is seen to be that of numbers divisible by 2 or 3 and not by 5. Now our method will yield the answer more quickly.

Exercises

1. Consider Example 1 in §2.1. Suppose that each good apple costs 1¢ while a rotten one costs nothing. Denote the rotten ones by R, an arbitrary bunch from the bushel by S, and define

$$Q(S) = |S \setminus R| / |\Omega - R|.$$

   Q is the relative value of S with respect to that of the bushel. Show that it is a probability measure.

2. Suppose that the land of a square kingdom is divided into three strips A, B, C of equal area and suppose the value per unit is in the ratio of 1:3:2. For any piece of (measurable) land S in this kingdom, the relative value with respect to that of the kingdom is then given by the formula:

$$V(S) = \frac{P(SA) + 3P(SB) + 2P(SC)}{2}$$

   where P is as in Example 2 of §2.1. Show that V is a probability measure.

*3. Generalizing No. 2, let $a_1, \dots, a_n$ be arbitrary positive numbers and let $A_1 + \cdots + A_n = \Omega$ be an arbitrary partition. Let P be a probability measure on Ω and

$$Q(S) = [a_1 P(SA_1) + \cdots + a_n P(SA_n)] / [a_1 P(A_1) + \cdots + a_n P(A_n)]$$

   for any subset of Ω. Show that P is a probability measure.

4. Let A and B denote two cups of coffee you drank at a lunch counter. Suppose the first cup of coffee costs 15¢, and a second cup costs 10¢. Using P to denote "price," write down a formula like Axiom (ii) but with an inequality (P is "subadditive").

5. Suppose that on a shirt sale each customer can buy two shirts at \$4 each, but the regular price is \$5. A customer bought 4 shirts $S_1, \dots, S_4$. Write down a formula like Axiom (ii) and contrast it with Exercise 3. Forget about sales tax! (P is "superadditive.")

6. Show that if P and Q are two probability measures defined on the same (countable) sample space, then $aP + bQ$ is also a probability measure for any two nonnegative numbers a and b satisfying $a + b = 1$. Give a concrete illustration of such a *mixture*.

*7. If P is a probability measure, show that the function $P/2$ satisfies Axioms (i) and (ii) but not (iii). The function P^2 satisfies (i) and (iii) but not necessarily (ii); give a counterexample to (ii) by using Example 1.

*8. If A, B, C are arbitrary sets, show that
(a) $P(A \cap B \cap C) \leq P(A) \wedge P(B) \wedge P(C)$,
(b) $P(A \cup B \cup C) \geq P(A) \vee P(B) \vee P(C)$.

*9. Prove that for any two sets A and B, we have

$$P(AB) \geq P(A) + P(B) - 1.$$

Give a concrete example of this inequality. [Hint: Use (2.3.4) with $n = 2$ and De Morgan's laws.]

10. We have $A \cap A = A$, but when is $P(A) \cdot P(A) = P(A)$? Can $P(A) = 0$ but $A \neq \emptyset$?

11. Find an example where $P(AB) < P(A)P(B)$.

12. Prove (2.3.7) by first showing that

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

13. Two groups share some members. Suppose that Group A has 123, Group B has 78 members, and the total membership in both groups is 184. How many members belong to both?

14. Groups A, B, C have 57, 49, 43 members, respectively. A and B have 13, A and C have 7, B and C have 4 members in common; and there is a lone guy who belongs to all three groups. Find the total number of people in all three groups.

*15. Generalize Exercise 14 when the various numbers are arbitrary but, of course, subject to certain obvious inequalities. The resulting formula, divided by the total population (there may be any nonjoiners!), is the extension of (2.3.7) to $n = 3$.

16. Compute $P(A \triangle B)$ in terms of $P(A)$, $P(B)$, and $P(AB)$; also in terms of $P(A)$, $P(B)$, and $P(A \cup B)$.

*17. Using the notation (2.5.3) and the probability defined in that context, show that for any two m and n we have

$$P(A_m A_n) \geq P(A_m)P(A_n).$$

When is there equality above?

*18. Recall the computation of plane areas by double integration in calculus; for a nice figure such as a parallelogram, trapezoid, or circle, we have

$$\text{area of } S = \iint\limits_S 1\, dx dy.$$

Show that this can be written in terms of the indicator I_S as

$$A(S) = \iint I_S(x, y)\, dxdy,$$

where Ω is the whole plane and $I_S(x, y)$ is the value of the function I_S for (at) the point (x, y) (denoted by ω in §1.4). Show also that for two such figures S_1 and S_2, we have

$$A(S_1) + A(S_2) = \iint (I_{S_1} + I_{S_2}),$$

where we have omitted some unnecessary symbols.

*19. Now you can demonstrate the connection between (2.3.7) and (2.3.8) mentioned there, in the case of plane areas.

20. Find several examples of $\{p_n\}$ satisfying the conditions in (2.4.1); give at least two in which all $p_n > 0$.

21. Deduce from Axiom (ii) the following two results. (a) If the sets A_n are nondecreasing, namely $A_n \subset A_{n+1}$ for all $n \geq 1$, and $A_\infty = \bigcup_n A_n$, then $P(A_\infty) = \lim_{n\to\infty} P(A_n)$. (b) If the sets A_n are nonincreasing, namely $A_n \supset A_{n+1}$ for all $n \geq 1$, and $A_\infty = \bigcap_n A_n$, then $P(A_\infty) = \lim_{n\to\infty} P(A_n)$. [Hint: For (a), consider $A_1 + (A_2 - A_1) + (A_3 - A_2) + \cdots$; for (b), dualize by complementation.]

22. What is the probability (in the sense of Example 10) that a natural number picked at random is not divisible by any of the numbers 3, 4, 6 but is divisible by 2 or 5?

*23. Show that if $(m_1, \dots, m_n)$ are co-prime positive integers, then the events $(A_{m_1}, \dots, A_{m_n})$ defined in §2.5 are independent.

24. What can you say about the event A if it is independent of itself? If the events A and B are disjoint and independent, what can you say of them?

25. Show that if the two events (A, B) are independent, then so are (A, B^c), (A^c, B), and (A^c, B^c). Generalize this result to three independent events.

26. Show that if A, B, C are independent events, then A and $B \cup C$ are independent, and $A \setminus B$ and C are independent.

27. Prove that

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(AB) - P(AC) - P(BC) + P(ABC)$$

when A, B, C are independent by considering $P(A^cB^cC^c)$. [The formula remains true without the assumption of independence; see §6.2.]

28. Suppose five coins are tossed; the outcomes are independent but the probability of heads may be different for different coins. Write the probability of the specific sequence $HHTHT$ and the probability of exactly three heads.

*29. How would you build a mathematical model for arbitrary repeated trials, namely without the constraint of independence? In other words, describe a sample space suitable for recording such trials. What is the mathematical definition of an event that is determined by one of the trials alone, two of them, etc.? You do not need a probability measure. Now think how you would cleverly construct such a measure over the space in order to make the trials independent. The answer is given in, e.g., [Feller 1, §V.4], but you will understand it better if you first give it a try yourself.